UNIVERSIDAD MIGUEL HERNÁNDEZ DE ELCHE
Doctorado en Tecnologías Industriales y de Telecomunicación


# Predicción de la severidad de accidentes de tráfico en la Red de Carreteras de España y Reino Unido mediante modelos estadísticos basados en Random Forest y Regresión Logística

**Autor:** David Úbeda González
**Directores:** Dr. Arturo Gil Aparicio
Dr. Agustín Pérez Martín

**Tesis Doctoral presentada en la Universidad Miguel Hernández de Elche para la obtención del título de Doctor del Programa de Doctorado en Tecnologías Industriales y de Telecomunicación**

**2017**

UNIVERSITAS
Miguel
Hernández

# AUTORIZACIÓN DE PRESENTACIÓN DE TESIS DOCTORAL

Directores: Dr. Arturo Gil Aparicio y Dr. Agustín Pérez Martín

Título de la tesis: ***Predicción de la severidad de accidentes de tráfico en la Red de Carreteras de España y Reino Unido mediante modelos estadísticos basados en Random Forest y Regresión Logística***

Autor: David Úbeda González

Departamento de Ingeniería de Sistemas y Automática
Universidad Miguel Hernández de Elche

Los directores de la tesis reseñada CONFIRMAN QUE HA SIDO REALIZADA BAJO SU DIRECCIÓN POR D. David Úbeda González en el Departamento de Ingeniería de Sistemas y Automática de la Universidad Miguel Hernández de Elche y autorizan su presentación.

En Elche, a de de 2017.

Fdo: Dr. D. Arturo Gil Aparicio

Fdo: Dr. D. Agustín Pérez Martín

UNIVERSITAS
Miguel
Hernández

**Departamento de Ciencia de Materiales, Optica y Tecnología Electrónica**

Campus Elche - Edificio Torrevaillo

**Piedad Nieves De Aza Moya** Directora del Departamento de Ciencia de Materiales, Óptica y Tecnología Electrónica de la Universidad Miguel Hernández:

HACE CONSTAR:

Que da su conformidad a la lectura de la tesis doctoral presentada por

Don **David Úbeda González**, titulada "***Predicción de la severidad de accidentes de tráfico en la Red de Carreteras de España y Reino Unido mediante modelos estadísticos basados en Random Forest y Regresión Logística***" , la cual ha sido desarrollada dentro del Programa de doctorado en Tecnologías Industriales y de Telecomunicación en este Departamento, bajo la dirección del Dr. D. Arturo Gil Aparicio, y el Dr. Agustín Pérez Martín.

Elche, a de julio de 2017.

**Piedad Nieves De Aza Moya**

Catedrática de Ciencia de Materiales e Ingeniería Metalúrgica.
Directora Departamento de Ciencia de Materiales, Óptica y Tecnología Electrónica

UNIVERSITAS
Miguel
Hernández

## Resumen

En este trabajo se propone una nueva técnica de predicción, basada en dos modelos de clasificación y regresión, destinada a guiar al usuario en su comportamiento al volante a través de la dotación de inteligencia vial colectiva al vehículo y a los dispositivos que se usen en él.

En base al análisis previo de las características que envuelven a los accidentes de tráfico y mediante sistemas que sean capaces de aprender de forma autónoma en base a nuevas incidencias, este trabajo será de posible aplicación para evitar situaciones futuras no deseadas.

Se propone para ello un modelo probabilístico de clasificación que permite predecir la severidad de un accidente de tráfico para cada uno de los ocupantes de un vehículo, en caso de que éste ocurriera. A lo largo del documento se presentarán resultados comparando diferentes técnicas de selección de predictores de entre un gran conjunto de variables disponibles, que se aplican sobre un conjunto de datos masivos de accidentes de tráfico.

Finalmente, el modelo de clasificación propuesto, basado conjuntamente en Random Forest y BayesGLM, permitirá inferir las relaciones entre accidentes y sus factores contribuyentes, con el fin de reconocer las causas que determinan los daños asociados a cada víctima, permitiendo así extraer información que puede resultar de suma importancia en la planificación de políticas de reducción de accidentes de tráfico válidas para los gobiernos, pero también y quizás más importante sea la posibilidad de dotar a los vehículos de inteligencia en seguridad vial para obtener rutas más seguras.

Las aportaciones de este trabajo se han aplicado a dos casos prácticos con resultados muy satisfactorios. En ellos se estudió la predicción de daños a los ocupantes de los vehículos mediante el análisis de los accidentes de tráfico ocurridos durante los periodos de 2011 a 2015 en España y entre 2009 a 2014 en Reino Unido. En él se demuestra científicamente que el sistema es capaz de aprender a partir de una serie de datos recogidos sobre accidentes de tráfico y de encontrar tendencias en ellos estadísticamente demostrables y probables a través de escoger correctamente la información y clasificarla mediante técnicas basadas en la supervisión entre algoritmos de regresión o de clasificación.

UNIVERSITAS
Miguel
Hernández

## Abstract

This thesis proposes a regression model that allows to predict the severity of a traffic accident for each of its occupants. The model is intended to provide a collective intelligence to the vehicle and also will be useful to improve the userâĂŹs driving behavior.

For this purpose, a probabilistic classification model is proposed that allows to predict the severity of a traffic accident based on a set of predictors that were involved in the collision. A massive dataset of accidents has been used to test the techniques presented in this study. Different approaches for reducing the number of features involved in each road accident are presented and tested.

Based on the previous analysis of all the characteristics that involve traffic accidents and through autonomous systems that are able to learn from new incidents, this study will be of possible application to avoid future undesirable events.

The model proposed is based on the Random Forest and BayesGLM algorithms and allow us to infer the relationships between accidents and their contributing factors, in order to recognize the causes that determine the physical damages associated to each victim, thus allowing us to extract information that can be of great importance when planning traffic policies by governments. In addition, another important applications involves the possibility of providing vehicles with the intelligence to choose safer routes, thus increasing the overall road safety.

This work has been applied in two practical cases with very satisfactory results. On the one hand, the research includes the prediction of injuries to vehicle occupants by analyzing the traffic accidents that occurred from 2011 to 2015 in Spain and from 2009 to 2014 in the United Kingdom. The results presented demonstrate that the system is able to learn from a series of data collected on traffic accidents and to find trends in them that can be statistically demonstrable. The algorithm proposed is able to process the information and classify it by a supervised combination of regression and classification techniques.

UNIVERSITAS
Miguel
Hernández

# Agradecimientos

Desde abril de 1998, arrastraba la firme convicción de que la medicina sería capaz de obtener una cura a determinadas enfermedades que, como el cáncer, acababan y acaban con la vida de multitud de personas en el mundo. Sin embargo, 15 años después, supe que estaba equivocado: el ser humano sólo ganará la batalla cuando comprenda que debe atender, en mayor medida, a cómo anticipar la muerte, bien sea porque ésta se produce por la aparición de determinadas enfermedades, o por otras causas, como aquellas relacionadas con accidentes de tráfico, tema de investigación de esta tesis doctoral. Desde ese momento, trato de redirigir mi formación y mis conocimientos en esa línea: *la predicción*.

Por todo ello, me gustaría comenzar dedicando este trabajo a la persona que durante estos casi 20 años ha sido mi inspiración desde la lejanía: mi padre. Muchas gracias, papá, estés donde estés.

Martina, hija, han sido muchas las noches que has pasado en mi regazo delante del ordenador mientras redactaba este documento o lanzaba en los servidores alguna nueva ejecución. En todo este proceso aprendiste, con tan sólo algo más de dos años, a pronunciar palabras como '*accidente*', '*tesis*', '*adivinar*', '*futuro*' y '*esfuerzo*'. Sin duda merecías que, durante todas estas noches, hubiera estado a tu lado para contarte un cuento, o quizás tan sólo para ofrecerte sosiego mientras te dormías abrazada a mi. Sin embargo, dos de esas palabras que aprendiste han sido los principales motivos que me han ayudado a emprender esta tarea, en determinadas ocasiones titánica, pero que estoy seguro que con los años descubrirás y valorarás lo que realmente hay detrás de ella. Por todo ello, te dedico esta obra con todo el cariño del mundo, la cual está fundamentada en y desde el esfuerzo, con el ánimo de mejorar aquello que nos rodea.

Como continuación a esta dedicatoria, entenderán que un trabajo de esta envergadura no es obra de una sola persona, por ello, deseo expresar mi más profundo agradecimiento a mi familia, no únicamente por este periodo doctoral, sino a la dedicación de toda una vida a mi desarrollo personal e intelectual. Por todo ello, gracias mamá, gracias iaia, gracias Antonio, gracias de todo corazón.

Edith, amor, lo hemos conseguido... no sólo concluir este trabajo, hemos conseguido algo mucho más importante, y que sin tu paciencia, ayuda y sacrificio no lo hubiéramos sacado adelante de forma tan brillante: a nuestra hija. De veras siento continuamente que lo bueno no ha hecho más que comenzar. Te quiero...

Por supuesto, no hubiera podido ni siquiera dar comienzo a mi carrera investigadora de no haber sido gracias a los integrantes del grupo de investigación ARVC, que desde el minuto cero me acogieron y ayudaron en todo lo que estuvo en su mano. En especial, me gustaría agradecer a uno de sus integrantes, a Arturo, su incondicionalidad y su ayuda personal en todos y cada uno de los retos que me he propuesto en mi carrera docente e investigadora, dejando de lado aquello en lo que estuviera inmerso para ayudarme. Hoy es Director de este trabajo, pero lo más importante es que es un gran amigo. Estoy seguro que juntos alcanzaremos retos profesionales aún más reconfortantes.

También me gustaría agradecer a Agustín, co-director de esta Tesis Doctoral, su ánimo, apoyo y amistad durante este largo proceso, así como la transmisión de sus grandes conocimientos en estadística, a través de pequeñas clases magistrales telefóni-

cas a horas intempestivas de la madrugada, sin importar el hecho de que se encontrara disfrutando de su periodo vacacional.

Por último, y no por ello menos importante, aprovecho para enviar un fuerte abrazo de agradecimiento a todas y cada una de las personas que me han arropado durante todos estos años, o que han colaborado de alguna forma en este trabajo.

Gracias a todos y a todas desde lo más profundo de mi corazón.

*A mi hija*

UNIVERSITAS
Miguel
Hernández

# Índice de tablas

UNIVERSITAS
Miguel
Hernández

# Índice de figuras

# Índice general

UNIVERSITAS
Miguel
Hernández

# 1 Introducción

## 1.1. Preámbulo

A la hora de plantear soluciones en materia de seguridad vial, las fuerzas de seguridad de los distintos países, así como sus agencias de tráfico, se suelen centrar en la resolución de los problemas relacionados con el triángulo básico *infraestructura-conductor-vehículo*, donde el uso de medidas reglamentarias como radares de control de velocidad, controles de drogas y alcohol, o controles del cumplimiento de inspecciones técnicas se consideran actuaciones suficientes para la reducción de los accidentes de tráfico por su parte. Sin embargo, existe otro tipo de acciones que, sin duda, aprovecharían de una forma más eficiente el trabajo que este tipo de organizaciones públicas realizan.

La recolección de todas las circunstancias que envuelven a un accidente de tráfico resulta una labor minuciosa. Su estudio, no obstante, acaba en la puesta en marcha de medidas relativamente obvias y a veces de dudosa efectividad para salvaguardar la seguridad de los ocupantes de los vehículos. El planteamiento de medidas paliativas relativas a la seguridad basadas en soluciones de inteligencia artificial y relacionadas con inferir conocimiento y trasladarlo a la infraestructura o a los automóviles, hasta ahora, no se ha planteado de forma óptima, o bien no se ha deseado asumir por los organismos públicos con competencias en la materia, a pesar de ser las agencias que mayor información poseen al respecto. Por ello, en tiempos en el que el vehículo autónomo ya comienza a ser una realidad, pensamos que es necesario asumir una actitud proactiva en esta línea.

En este trabajo se propone procesar dos conjuntos masivos de accidentes reales con el fin de reconocer las causas que determinan la severidad del mismo y que además

estén asociados a cada víctima. El planteamiento inicial consiste en la búsqueda de una función o clasificador que nos permita predecir el resultado de los ocupantes de un vehículo en un accidente en base a una serie de características.

Como resultado del estudio, se aplicarán modelos estadísticos de regresión y clasificación de la severidad de las lesiones producidas, a todos los ocupantes de un vehículo, en el caso de que se produjera un accidente de tráfico. Para ello, se analizará cada accidente ocurrido en España y Reino Unido en base a un conjunto de variables involucradas en él, donde los modelos estadísticos nos permitirán inferir las relaciones entre los accidentes y sus factores contribuyentes, permitiéndonos de esta forma extraer información que puede ser de gran utilidad en la planificación de políticas de reducción de accidentes de tráfico, propuestas de mejora de las infraestructuras existentes, señalización y comunicaciones con los conductores, entre otras.

## 1.2. Hipótesis y Objetivos de la Tesis Doctoral

La principal hipótesis que se plantea en esta Tesis Doctoral se basa en que existe una dependencia de las variables que rodean al accidente y los daños que puede sufrir cada uno de los ocupantes del vehículo.

Adicionalmente, y como hipótesis derivada de la anterior, se plantea si es posible proponer un modelo estadístico que permita predecir estos daños en el caso de que ocurriera un accidente. Dicho de otra forma, para modelar la mencionada dependencia, en este trabajo de investigación se plantea el uso de técnicas estadísticas que permitan encontrar las relaciones entre la severidad del accidente y otras variables como la temperatura, el tipo de vía o el tipo de vehículo, entre otras.

Por todo ello, el objetivo principal de este trabajo consistirá en el desarrollo de un modelo probabilístico que sea capaz de estimar la probabilidad del tipo de daños causados a los ocupantes de un vehículo como consecuencia de un accidente de circulación. Se pretende pronosticar el evento, esto es, la gravedad o no del accidente y de los ocupantes, así como tratar de evitar las circunstancias que lo produjeron.

De forma paralela se segmenta el objetivo principal en los siguientes objetivos parciales estructurados de forma ordenada con el fin de facilitar la consecución de dicho objetivo:

- Desarrollar una metodología donde se definan todas las fases de la investigación que lleven a la consecución de todos los objetivos.
- Revisión de los modelos y métodos matemáticos previamente utilizados a este trabajo de investigación.
- Revisión del estado del arte de la metodología empleada previamente a este trabajo en el análisis de siniestralidad vial.

- Planteamiento de las aportaciones realizadas mediante esta tesis en el campo de investigación aplicado.
- Definición de la metodología empleada para la reducción y estimación de los predictores necesarios con la finalidad de construir un modelo probabilístico basado en necesidades de temporalidad y acierto como premisas para su aplicación en una infraestructura de Big Data en función del país.
- Definición e implementación de un modelo probabilístico capaz de estimar la probabilidad del tipo de daños causados a los ocupantes de un vehículo como consecuencia de un accidente de circulación.

## 1.3. Organización del trabajo

A modo de breve resumen, el contenido fundamental de esta tesis doctoral está organizado en los siguientes capítulos:

- Capítulo 2. En este capítulo se presenta una revisión de la teoría de modelos de clasificación y regresión alrededor de los sucesos raros y accidentes de trafico.
- Capítulo 3. Revisión del estado del arte en materia de accidentes de tráfico y seguridad vial así como de las investigaciones realizadas hasta la fecha en este ámbito por otros autores.
- Capítulo 4. Se describe una nueva arquitectura de Big Data para el almacenamiento y analítica de datos de incidencias de tráfico en tiempo real.
- Capítulo 5. Se presenta una aplicación real de las técnicas desarrolladas en este trabajo en base a un estudio estadístico de los accidentes de tráfico que se han producido en España entre 2011 y 2015.
- Capítulo 6. Se presenta una aplicación real de las técnicas desarrolladas en este trabajo en base a un estudio estadístico de los accidentes de tráfico que se han producido en Reino Unido entre 2009 y 2014.
- Capítulo 7. Se presentan las principales conclusiones y líneas futuras a partir de este trabajo.

## 1.4. Aportaciones

Una de las principales aportaciones de este trabajo de investigación consiste en la comparativa entre numerosos algoritmos de selección de características y modelos de clasificación o regresión para sets de datos de accidentes de tráfico desbalanceados.

En los capítulos siguientes se estudiarán aportaciones más detalladas al final de cada uno de ellos y que se resumen a continuación:

- Capítulo 2: se repasan los principales métodos y técnicas para selección de características, clasificación y regresión, prestando especial interés en la aplicación que nos ocupa: accidentes de tráfico.
- Capítulo 3: se recogen las principales aportaciones e investigaciones en estudios similares de tráfico realizados por otros investigadores.
- Capítulo 4: se presenta una arquitectura para el almacenamiento masivo y en tiempo real de históricos de datos relativos a incidencias de tráfico a través de un clúster basado en bases de datos NoSQL mediante Apache Cassandra.
- Capítulos 5 y 6: se presentan sendos estudios comparativos de técnicas para el pre-procesamiento de set de datos relativos a accidentes de tráfico ocurridos en España y Reino Unido, respectivamente. Adicionalmente, se presenta un estudio comparativo de reducción de predictores y un nuevo modelo de predicción basado en Random Forest y BayesGLM de aplicación a los accidentes de tráfico ocurridos en dichos países.

### 1.4.1. Publicaciones

Los siguientes trabajos publicados en revistas y congresos de reconocido prestigio apoyan esta investigación y su aplicabilidad en materia de tráfico y seguridad vial.

#### Publicaciones en Revistas

- A study of traffic accidents in Spanish intercity roads by means of feature vectors
  D. Úbeda, A. Gil, L. Payá, O. Reinoso
  International Journal of Design & Nature and Ecodynamics (2016)
  Ed. WIT Press ISSN:1755-7437 DOI: 10.2495/DNE-V11-N3-317-327 - Vol. 11(3)
- Analyzing Traffic Accident Severity in UK from a Classification point of view
  D. Úbeda, A. Gil, A. Pérez Enviada a Journal of Advanced Transportation. Ed. Hindawi - John Wiley & Sons **JCR-SCI Impact Factor: 1.292**, Quartil **Q1**.

#### Registros de la propiedad intelectual

En abril de 2017 se cursa solicitud de registro denominada *'Ecosistema de aplicaciones y servicios para la gestión y el almacenamiento de Big Data en aplicaciones de vehículos'* en el Registro de la Propiedad Intelectual e Industrial con identificación A-1632017. El software registrado nace como consecuencia de distintas aportaciones y desarrollos de esta tesis doctoral. En el anexo J se puede acceder al documento de solicitud de registro.

#### Transferencia de tecnología

En junio de 2017 se aprueba por Consejo de Gobierno y por Consejo Social la creación de una spin-off con la finalidad de transferir una parte de la tecnología desarrollada a través de esta tesis doctoral a la industria de la automoción.

## 1.5. Subvenciones

Parte de este trabajo ha sido subvencionado por la Generalitat Valenciana a través del proyecto GV/2015/031: *Creación de mapas topológicos a partir de la apariencia global de un conjunto de escenas.*

UNIVERSITAS
Miguel
Hernández

# 2 Teoría de algoritmos de regresión y clasificación

## 2.1. Introducción

### 2.1.1. Sucesos raros y su tratamiento

En una distribución binomial, si el número de ensayos $N$ es grande y, al mismo tiempo, la probabilidad $P$ de ocurrencia de un evento es cercano a cero, por lo que $q = 1 - P$ (probabilidad del fracaso) es cercano a 1, el evento es llamado un evento raro. En la práctica consideramos un evento raro si el número de ensayos es al menos 50 ($N \geq 50$); mientras que $NP$ es menor de 5. En tales casos, la distribución binomial es muy cercana a la distribución de *Poisson* con la siguiente media [58]:

$$\lambda = N \cdot \mathcal{P} \tag{2.1}$$

La distribución de *Poisson* o *Ley de los Eventos Raros* es una extensión de la distribución binomial llevada al caso en que, con muestras grandes, las probabilidades de observar un evento de interés son bajas [88], es decir, la distribución de Poisson es el caso límite de la distribución binomial cuando $N$ tiende a $\inf$ y $P$ a cero.

En esta investigación se considera a los accidentes de tráfico *eventos raros*, ya que su frecuencia de ocurrencia en el tiempo es muy baja.

La *Ley de los Eventos Raros* fue demostrada por *Poisson* en su libro '*Recherches sur la Probabilités des Jugements en Matière Criminelle et Matière Civile*', y surge al estudiar la probabilidad de obtener $k$ éxitos en $n$ ensayos *Bernoulli* con parámetro

$p$. Un ensayo Bernoulli es un experimento aleatorio en el que únicamente se pueden obtener dos resultados, por ejemplo, éxito y fracaso. Estos ensayos están modelados por una variabla aleatoria, la cual puede tomar sólo dos valores, 0 y 1. El término raro se refiere a que la probabilidad de éxito $p > 0$ es pequeña, del orden de $1/n$ [47]. El objetivo consistirá en contar el número de eventos raros que ocurren en cada intervalo de tiempo $[0, t]$, donde $N_t$ se corresponde con el número de eventos en el intervalo $[0, t]$.

Partiendo de un instante inicial con $N_0 = 0$, para dos instantes de tiempo $t_1$ y $t_2$, tales que $t_1 < t_2$, entonces $N_{t_1} \leq N_{t_2}$

Se parte de $N_0 = 0$ donde si $t_1 < t_2$, entonces $N_{t1} \leq N_{t2}$, ya que el número de eventos que ocurren en el intervalo $[0, t_1]$ es menor o igual que los que ocurren en $[0, t_2]$.

I $N_0 = 0$

II Si $s < t$, entonces $N_s \leq N_t$

III Para toda $n > 0$ y $0 < t_1 < t_2 ... < t_n$, las variables aleatorias

$$N_{t1}, N_{t2} - N_{t1}, ...N_{tn} - N_{tn-1},$$

son independientes.

IV Para toda $h > 0$ y t $\in \mathbb{R}^+, N_h$ y $N_{t+h} - N_t$ tienen la misma distribución.

A los procesos estocásticos que cumplen la condición (iii) se les conoce como *procesos independientes*, y a los que cumplen (iv) como *estacionarios*.

*Poisson* acuñó el término de '*Ley de los Grandes Números*' considerando que la '*Ley de los Errores de Observación*' no seguía una distribución normal o gaussiana en todos los casos, sobre todo en aquellos eventos poco frecuentes, a los que denominó *eventos raros*. Su intención, por tanto, fue conocer cómo calcular la probabilidad de que tales eventos raros fueran efectivamente observados para los que la distribución binomial no ofrecía la mejor estimación.

En resumen, la distribución de Poisson puede utilizarse para realizar estimaciones puntuales, sin embargo, es fundamental que se cumpla:

- que el tamaño de muestra sea lo suficientemente grande como para poder aplicar la distribución binomial.
- que la probabilidad de observar un evento en un intervalo de tiempo es proporcional al tamaño de dicho intervalo de tiempo.
- que la probabilidad de que dos eventos ocurran en tal intervalo debe ser insignificante.
- que la probabilidad de la ocurrencia del evento sea la misma entre los intervalos.
- que la ocurrencia de tales eventos en un intervalo sea independiente de los ocurridos en intervalos anteriores.

### 2.1.2. Minería de datos, Machine learning y Big Data

El término Big Data habitualmente se relaciona con un gran volumen de datos, como su traducción literal del inglés indica, y es por ello que este gran volumen requiere de nuevas técnicas para procesarlos, de nuevas bases de datos para almacenarlos y de nuevos supercomputadores para analizarlos. Por todo ello, surgen una serie de dificultades al trabajar con estos datos que se han resumido en lo que se conoce como las 6 V's:

- *Volumen*: los dataset que se emplean, al poseer un volumen tan elevado, su análisis en computación provoca un consumo muy elevado en memoria RAM, por lo que resulta importante prever este dato a la hora de realizarlo.
- *Velocidad*: en proyectos de predicción, en la mayoría de ocasiones, resulta imprescindible procesar la información en tiempo real, o cuasi real, tarea compleja por la velocidad de computación requerida para tal fin en comparación con el volumen de datos a analizar.
- *Variedad*: fundamentalmente es en la etapa de preprocesamiento donde lidiamos con estos problemas de variedad en los tipos de datos, lo cual requiere a su vez la aplicación de diversas técnicas de preprocesamiento y análisis para trabajar de forma correcta.
- *Variabilidad*: a medida que se van capturando datos, en función de factores temporales, como la época del año, la hora del día o cualquier otra variación temporal en función de la aplicación, es posible que los datos varíen con bastante frecuencia. Es necesario que los algoritmos que se empleen sean capaces de actuar de forma rápida frente a estos cambios.
- *Veracidad*: la incertidumbre con la que se trabaja con determinadas fuentes de datos hace necesario que haya que eliminarla a la hora de analizarlos.
- *Valor*: resulta relevante obtener conclusiones de interés con los datos. En caso contrario, conviene estudiar por qué y plantear a su eliminación si procede.

Muchos de estos inconvenientes que se plantean a la hora de trabajar con sets de datos muy grandes, se han de trabajar en etapas previas a la analítica mediante técnicas de data mining o machine learning. Fundamentalmente, será en etapas posteriores de procesamiento donde se emplearán técnicas para evitar el desbordamiento de la memoria de los computadores o para ejecutarlos a velocidades mayores en los procesadores.

Habitualmente, se suele trabajar con sistemas distribuidos a la hora de analizar una ingente cantidad de datos, donde el uso de estos tipos de sistemas permiten repartir la carga entre distintos nodos del clúster y procesar la información en paralelo. Gracias a ello, es posible reducir algunos de los problemas que se observaban anteriormente y que, en gran medida, podrían venir dados por trabajar sobre un solo nodo.

De forma paralela, la minería de datos es la disciplina consistente en resolver problemas analizando los datos una vez que los tenemos almacenados en bases de datos. Se define, por tanto, como el proceso de descubrir patrones en los datos. El proceso debe ser lo más automático posible y los patrones descubiertos deben ser significativos, puesto que deben conducir a alguna ventaja, de aqui la *'V'* de *valor*, usualmente económica o, como es el caso de este trabajo, para tratar de salvar vidas humanas.

Habitualmente, como se ha explicado en párrafos anteriores, los datos están presentes en cantidades sustanciales, por tanto, emplear las técnicas adecuadas para advertir estos patrones, resultará fundamental. Sin embargo, la pregunta importante en esta situación es cómo vienen dados los patrones. En ocasiones, éstos vienen dados como una caja negra cuyas entrañas son prácticamente incomprensibles; y en otros casos, como una caja transparente cuya construcción revela la estructura del patrón. Todo ello no quiere decir que las predicciones sean malas o no, en un caso o en otro, lo que viene a decir es si los patrones extraídos se representan o no en términos de una estructura que puede ser examinada, razonada y utilizada para informar decisiones futuras. Ahí radica la verdadera importancia del análisis que efectuemos sobre los datos, ya que nos ayudará a entender por qué ocurren los hechos, no solamente si van a ocurrir o no.

Si la minería de datos consiste en el proceso de descubrir patrones útiles en grandes cantidades de datos de la forma más automatizada posible, se define el concepto de '*Machine Learning*' como el aprendizaje que hace cambiar el comportamiento de las cosas de una manera que los hace funcionar mejor en el futuro [114]. Como se puede observar, el término aprendizaje adquiere un matiz más enfocado al comportamiento que realmente al conocimiento. Adicionalmente, con buen criterio, Witten [114] relaciona el aprendizaje al propósito, de igual manera que aproxima al aprendizaje sin propósito como mero entrenamiento.

### 2.1.3. ¿Predicción vs. Interpretación?

Lo descrito en los párrafos anteriores lleva a la comparación entre dos conceptos fundamentales a la hora de trabajar con algoritmos de minería de datos para *machine learning*: predicción o interpretación.

Teniendo en cuenta que los datos históricos se pueden emplear para tratar de predecir acontecimientos futuros mediante una serie de algoritmos matemáticos, el propósito no es siempre conocer y comprender el por qué van a ocurrir los eventos. En algunas ocasiones, el propósito viene dado por la precisión con la que se adivinará el evento, y no por las circunstancias que lo produjeron.

Sin embargo, en determinados campos se hace necesaria la conjunción de ambos términos, como es el trabajo que nos ocupa, donde se pretende 'adivinar' con la máxima precisión posible qué ocurrirá en un accidente con los ocupantes de un vehículo, pero es tan importante o más saber interpretar el modelo para evitar repetir el resultado del evento. Desafortunadamente, a medida que se tiende a mejorar la exactitud de la predicción, se acaban empleando modelos que resultan difíciles de interpretar.

### 2.1.4. Terminología

Empleando como referencia a Kuhn [76], en este apartado se describen los principales términos que se emplean en los sucesivos capítulos de esta memoria:

- Muestra o datapoint hace referencia a una única e independiente unidad de dato.
- Set de datos de entrenamiento o training dataset es el set de datos empleado en el desarrollo del modelo.
- Set de testeo o testing dataset se usa para evaluar el rendimiento del modelo final escogido.
- Predictores, variables independientes, atributos, características o descriptores que hacen referencia a los datos empleados como entrada para el modelo o ecuación de predicción, y que sirven para evaluar, seleccionar o relacionar con la variable objeto de estudio.
- La variable dependiente, salida, resultado a predecir, clase o respuesta hace referencia al evento resultado del modelo, o también llamada característica de interés.
- Datos continuos o numéricos son aquellos que poseen una escala numérica y que pueden ser alterados con cualquier operación aritmética sin perder una posible interpretación.
- Factores, categorías o datos nominales, hacen referencia a aquellos datos que se agrupan en categorías y que no tienen escala, o que en caso de tenerla no tiene ningún sentido, en términos interpretativos, aplicarles una operación aritmética.

## 2.2. Pre-procesamiento de datos

El pre-procesamiento de datos está relacionado con el tratamiento que se lleva a cabo sobre el set de datos de entrenamiento relativo a eliminar, añadir o transformar variables.

Aunque existen multitud de opciones para obtener o construir un set de datos, en la mayoría de ocasiones resulta necesario *'acondicionar'* los datos debido a que cabe la posibilidad de que nos encontremos, desde valores nulos o valores que falten, hasta datos incorrectos, *outliers* o incluso que surja la necesidad de crear una nueva variable a partir de otras, entre otras posibilidades.

La preparación de los datos de entrenamiento resulta tan importante que puede ser un factor determinante para que el modelo de predicción tenga o no la capacidad predictiva óptima [76]. Adicionalmente, la manera y, en determinadas ocasiones, el orden con que los predictores son introducidos en el modelo es de vital importancia.

Sin embargo, la profundidad con la que abordar el pre-procesamiento dependerá del tipo de modelo que se use. Por ejemplo, los modelos basados en árboles de decisión

no son tan sensibles a las características de los datos de los predictores como los modelos lineales.

En esta sección nos centramos en la transformación de los datos en función del número de predictores. Es decir, en cómo obtener la importancia de cada predictor o cómo seleccionar los predictores en función de si poseemos clases desbalanceadas de forma severa. Adicionalmente, se estudian métodos para obtener los sets de datos para entrenamiento y testeo, así como técnicas de validación cruzada.

### 2.2.1. Centrado y escalado de variables

Habitualmente, se suele centrar la escala de los predictores restándole el valor medio de ese predictor, obteniendo por tanto predictores que poseen una media de valor cero. Adicionalmente, cada valor de los predictores se escala dividiendo por su desviación estándar, con lo que se obliga a que los predictores tengan la misma desviación estándar, es decir, de valor uno. A este proceso se le denomina *estandarización*.

Aunque algunos modelos mejoran en estabilidad en diversos cálculos cuando se tienen todos los predictores en la misma escala, se pierde cierta interpretabilidad de los valores individuales, ya que los datos no se encuentran en las unidades originales [76].

### 2.2.2. Transformación de datos para múltiples predictores

Este tipo de transformaciones están relacionadas con métodos para resolver diversos problemas con los sets de datos, tales como aparición de valores atípicos, reducción de datos o extracción de variables.

Un valor atípico u *outlier* consiste en una muestra que se encuentra numéricamente distante del resto de las muestras. Habitualmente, determinadas características pueden poseer una serie de valores atípicos que conviene estudiar, ya que éstos pueden aparecer al cometer errores en la tarea de adquirir los datos o al trasladarlos a un fichero, por lo que conviene eliminarlos con el fin de que la media de la variable no se vea alterada.

Sin embargo, se ha de ser muy cuidadoso en la elección de aquellos valores a eliminar, puesto que en determinadas ocasiones, pueden poseer un interés especial que evidencie un suceso extraordinario. Es muy importante aplicar el sentido común para tomar una decisión sobre qué hacer con ellos.

Resulta sencillo visualizar gráficamente en un diagrama de caja este tipo de muestras. Atendiendo al método de Tukey [11], se considerará como referencia la diferencia entre el primer y el tercer cuartil. Un valor atípico se encontrará $3/2$ veces dicha referencia si es leve, o más de 3 veces esa distancia si se trata de un valor atípico extremo.

Adicionalmente, existen una serie de modelos predictivos *'resistentes'* a los valores atípicos. Por ejemplo, los modelos de clasificación basados en árbol crean *'splits'* de los datos de entrenamiento y la ecuación de predicción está basada en una serie de condiciones lógicas que hacen que el valor atípico no tenga una influencia desmesurada en el modelo.

## 2.3. Selección de características

En determinadas ocasiones, cuando el set de datos posee un número elevado de características, la realización de un análisis exploratorio de todos los predictores puede ser un arduo proceso, por lo que conviene fijarse en aquellos atributos que poseen una fuerte relación con la variable que deseamos predecir.

Existen distintos métodos que permiten realizar un ranking de atributos en este sentido. Por ejemplo, aquellos modelos basados en árboles de decisión evalúan qué ocurre en cuanto al rendimiento del modelo de predicción a medida que se van añadiendo características al mismo. Por otro lado, los modelos de regresión lineal o logística emplean cuantificaciones probabilísticas basadas en los coeficientes del modelo o medidas estadísticas.

Habitualmente, cuando los modelos escogidos posean métricas que midan la importancia de los predictores, resulta conveniente emplear dichas métricas en lugar de realizar un preprocesamiento previo, ya que ésto puede restar poder de predicción al modelo. Sin embargo, con motivo de reducir tiempos en las ejecuciones de los modelos de predicción, es posible realizar cierto preprocesamiento que haga una primera criba de predictores altamente correlacionados entre sí, y poder eliminarlos para reducir dichos tiempos.

La selección de características en minería de datos intenta reducir la dimensión del set de datos, en cuanto al número de variables se refiere, tratando de analizar el impacto de las características sobre el modelo, y una de las principales razones para medir la relevancia de los predictores viene dada por la posibilidad de escoger cuáles de ellos se deben introducir como entradas en un modelo.

Resulta de suma importancia escoger correctamente los predictores, ya que determinará si el modelo predictivo es lo suficientemente efectivo, o si por el contrario, todavía queda sin explicar cierta variabilidad de la característica de interés. También nos servirá para entender la relevancia de éstos sobre el modelo y la correlación sobre la variable a predecir o de salida. Por ejemplo, en áreas de salud, la reducción de características implica la simplificación y el abaratamiento de análisis cuando se trata de detectar una enfermedad a partir de un conjunto de marcadores.

Adicionalmente, la selección de características se emplea para reducir tiempos de cálculo a través de la eliminación de aquellas que son innecesarias para construir el modelo de predicción. Morales [85] nos muestra, detalladamente, cómo aumenta el coste computacional al aumentar el número de predictores.

Por ello, existen una serie de técnicas de selección de características que es posible dividir en dos enfoques diferentes:

- Clasificación de características
- Selección del subset de datos

En el primer enfoque, las características se clasifican en base a diferentes criterios y, a continuación, se seleccionan aquellas que se encuentran por encima de un umbral definido. Sin embargo, en el segundo enfoque, se busca una serie de subconjuntos de características con el que obtener el subconjunto óptimo. En este segundo caso, podemos tener una serie de aproximaciones:

- Filtrado de características previo para crear un subset que usaremos para insertar como entrada de un algoritmo de clasificación.
- Selección de características como parte del proceso de un algoritmo de clasificación.
- Uso de técnicas *Wrapper*: el algoritmo de clasificación se aplica sobre el conjunto de datos con el fin de identificar las mejores características.

La mayoría de algoritmos de selección de predictores se basan en una serie de puntuaciones otorgadas en función de la relevancia de las mismas. La importancia de una variable está relacionada con la cuantificación de la relación entre el predictor y la variable a predecir, donde las componentes principales (PCA) o el análisis factorial no lo tienen. Sin embargo, resulta complejo que los algoritmos informen acerca de la naturaleza de esa relación. Por ejemplo, que se informe acerca de si incrementando los valores del predictor decrecerán los valores de la variable a predecir o viceversa. Sin embargo, existen modelos que poseen formas paramétricas precisas, como los modelos basados en regresión lineal o logística, que cuantifican mediante sus coeficientes.

Atendiendo al estudio de Granitto [52], la típica estrategia que se emplea de forma incorrecta es realizar un bucle de validación cruzada para seleccionar las características, y luego utilizar un nuevo bucle de validación cruzada, sobre las mismas muestras, para estimar el error de test. Es necesario comprender que al realizar ésto durante el segundo bucle de validación cruzada, cada muestra insertada en el conjunto de test no será completamente independiente del modelo que se está evaluando, porque evidentemente se usó previamente para seleccionar las características [10].

Por todo ello, conviene utilizar métodos de remuestreo junto con la repetición de experimentos con el fin de evitar estos problemas [50]. Esta metodología también proporciona una solución parcial a un problema generalmente descuidado: la selección de características importantes sobre los experimentos replicados no es sencillo, debido a la inestabilidad de los métodos de selección. En esta Tesis, nos basamos en el análisis de probabilidad de selección y co-ocurrencia de ellos en subconjuntos dados a través del uso de métodos de validación cruzada, *k-fold*, así como la repetición de los experimentos $n$ veces.

### 2.3.1. Overfitting en la selección de características

Habitualmente, el set de datos global se divide en dos subconjuntos, un primer subconjunto de entrenamiento o *training* y un segundo conjunto para probar el modelo,

conjunto comúnmente llamado de test o *testing*. La proporción de ambos subconjuntos es típicamente $75\,\%/25\,\%$ u $80\,\%/20\,\%$, respectivamente, manteniendo las frecuencias de clases balanceadas siempre que sea posible.

Sin embargo, cualquier método de selección de características que utilice información sobre la variable a predecir puede llegar a caer en el sobreajuste u *overfitting*, en particular si posee ratios $n/p$ muy bajos, siendo $p$ el número de características y $n$ el número de muestras de cada clase.

El *overfitting* es un efecto no deseado que ocurre cuando un algoritmo se sobreentrena a partir de un set de datos para los que es conocido el resultado de la variable objeto de estudio. Este efecto se descubre cuando se desea predecir el resultado de la variable objeto a través del set de datos que se ha almacenado para testeo. Cuando un modelo se encuentra sobreajustado, no es capaz de predecir a partir de las estas muestras de test, aunque se habían obtenido buenos resultados con los datos de entrenamiento previamente. Esto ocurre debido a que el modelo ha aprendido las características de cada muestra de ruido única.

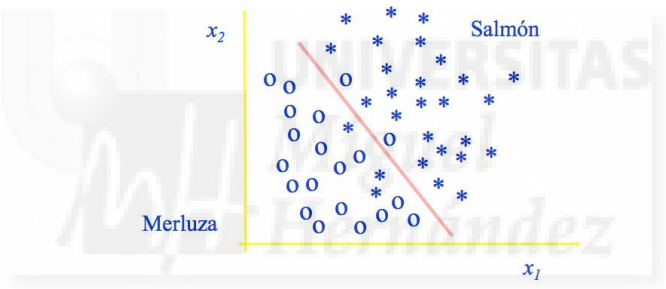



**Figura 2.1:** Función de separación general que separa las clases 'Merluza' y 'Salmón' en base a dos características.

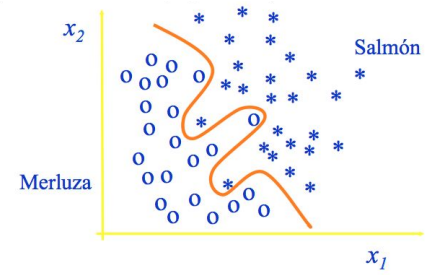



**Figura 2.2:** Función de separación compleja que separa las clases 'Merluza' y 'Salmón' para el conjunto de datos de entrenamiento.

La figura 2.1 ilustra una función de separación muy sencilla y general que separa dos clases de ejemplo '*Merluza*' y '*Salmón*' en base a dos características $x_1$ y $x_2$. Por otra parte, la figura 2.2 representa una función mucho más compleja que funcionará

particularmente bien para el conjunto de datos de entrenamiento, pero contendrá errores en la predicción del conjunto de muestras de test, es decir, poseerá sobreajuste u *overfitting*.

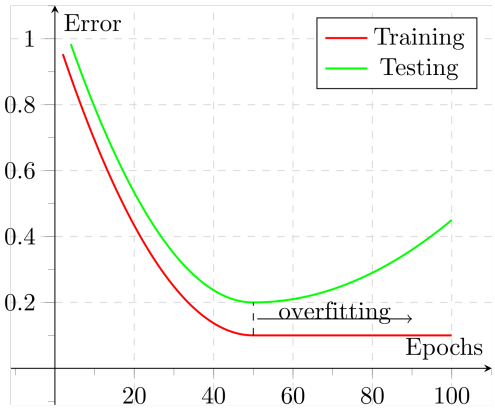



**Figura 2.3:** Ejemplo de un modelo con *overfitting* o sobreajustado. Autor: Martin Thoma

En la figura 2.3 se puede observar el error que se comenta para el modelo con sobreajuste. Para este caso concreto, una vez llegado al 50 % de las muestras de test, el error aumenta considerablemente en el momento que se pretende predecir el valor de la variable objeto de estudio.

Habitualmente, aunque de forma errónea, el set de datos se divide de forma aleatoria en el conjunto de entrenamiento y en el de test, respectivamente. Sin embargo, de esta forma no se controla el porcentaje de muestras de cada una de las clases de los atributos, y si una clase aparece con una frecuencia desproporcionada respecto a las otras, la distribución de la variable objeto de la predicción será diferente para el set de datos de entrenamiento y para el de test.

De forma alternativa, es posible dividir los datos en los dos sets de entrenamiento y test, respectivamente, tomando como base los valores de la variable objeto de la predicción mediante distintos métodos ([113] [29]). En la sección 2.3.3.1 se estudiarán distintas estrategias para evitar el efecto no deseado de *overfitting*.

Con el fin de obtener estimaciones no sesgadas del error de predicción, la selección de características conviene que esté incluida en el modelado, y no tratada como un paso de pre-procesamiento, como se hace a menudo. Sin embargo, cabe la posibilidad de que determinados modelos no incorporen esta característica, por lo que en este caso se habrán de realizar tareas de selección de predictores durante el pre-procesamiento.

### 2.3.2. Consecuencias de usar predictores que no aportan información

En el momento que pretendemos realizar una selección de características, nos debemos basar inicialmente en eliminar aquellos predictores que no aporten información

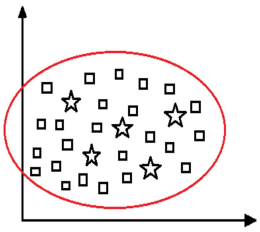


**Figura 2.4:** Gráfico donde se muestra el desbalanceo entre dos clases: cuadrados y estrellas.

o que aporten información de forma redundante.

Se ha de tener en cuenta que muchos modelos, como los basados en regresión, estiman parámetros para cada predictor, por lo que si poseemos información redundante o predictores que no contengan información, es posible que se añada incertidumbre en la predicción y se reduzca efectividad en el modelo, o incluso que dicha información contrarreste hasta el punto de empeorar las predicciones.

En resumen, modelos basados en árboles de regresión, como MARS y Random Forest, no se ven afectados por incluir predictores que no aporten información, sin embargo, modelos estructurados de forma paramétrica como modelos de regresión lineal o redes neuronales son los que resultan más afectados por este tipo de predictores [76].

### 2.3.3. Problemática de los sets de datos no balanceados

El concepto de set de datos no balanceado consiste en el desequilibrio que existe en una muestra de datos cuando el número de instancias de un grupo (clase mayoritaria) es superior al de otros grupos (clases minoritarias), o dicho de otro modo, cuando a la hora de realizar una clasificación, sus categorías no están representadas equitativamente, tal y como podemos ver en el ejemplo de la figura 2.4.

Que las clases no se encuentren balanceadas suele provocar que los resultados tiendan a favorecer a las clases con mayor número de muestras, surgiendo errores en la clasificación precisamente por este sesgo y, por tanto, aparezca un notable deterioro en el *accuracy* de nuestro clasificador. Sin embargo, existe una problemática aun mayor puesto que no existe un umbral que nos indique cuando una base de datos se encuentra o no desbalanceada.

Esta problemática ha sido objeto de numerosos estudios desde hace años [65, 64, 109, 57], y a día de hoy, lidiamos en mayor medida debido a la aparición de sets de datos provenientes de tecnologías derivadas de Internet de las Cosas [13].

Se han establecido numerosos criterios de evaluación para los sets de datos no balanceados, y a su vez, se ha llegado a la conclusión de que la exactitud (*accuracy*) del modelo como medida de rendimiento del mismo se encontraba bastante limitada. Como alternativa se plantea el uso de curvas ROC [57], tal y como se verá más adelante en esta memoria.

Chawla [24] plantea la cuestión acerca de cuál es la distribución correcta para un algoritmo de aprendizaje, y concluye con que a menudo no es una buena idea usar la distribución natural como distribución para que aprenda un clasificador [23].

En la literatura se han utilizado diversas estrategias de re-muestreo tales como el oversampling aleatorio con reemplazo, undersampling aleatorio, oversampling focalizado, undersampling focalizado, oversampling con generación sintética de nuevas muestras basadas en la información conocida y combinaciones de las técnicas anteriores [26].

Otro problema importante que surge debido a la escasez de datos, es la distribución de datos dentro de cada clase [63]. Este problema también está relacionado con la cuestión de las pequeñas disyunciones en el aprendizaje del árbol de decisiones.

#### 2.3.3.1. Estrategias de muestreo de datos

La metodología de *over* y *under-sampling* ha recibido una atención significativa para contrarrestar el efecto de conjuntos de datos no balanceados [102, 25, 112, 66, 14].

Sin embargo, los métodos aleatorios de *undersampling* y *oversampling* tienen diversos inconvenientes. El método de *undersampling* aleatorio puede eliminar ciertos ejemplos potencialmente importantes, y el *oversampling* aleatorio puede conducir a un ajuste excesivo de las funciones del clasificador.

El muestreo de datos, utilizado para mejorar el rendimiento de clasificación principalmente para conjuntos de datos con clases desbalanceadas, también resulta valioso incluso cuando la clase de interés es muy rara ($< 0,1\,\%$ de los datos), hecho que no ocurre en nuestro set de datos de accidentes, ya que la clase desbalanceada representa aproximadamente un 20 % de las muestras totales. De hecho, el rendimiento del muestreo de datos supera en gran medida al *oversampling*, tal como duplicación, cuadruplicación o incluso el aumento del número de instancias minoritarias hasta 16 veces de su valor original. Por lo tanto, el muestreo de datos es una herramienta muy valiosa para mejorar el rendimiento de la clasificación, incluso cuando los datos contienen eventos muy raros [95].

Chawla [24] presentó una extensa revisión de distintos autores, donde destaca la evaluación del efecto de un conjunto de datos desbalanceado que realizó Japkowicz [62]. Según su investigación, se evaluaron varias estrategias, entre ellas sub-muestreo y re-muestreo, y consideró dos métodos de muestreo para el *over* y el *sub-sampling*. El remuestreo aleatorio consistió en el *oversampling aleatorizado* de la clase más pequeña hasta que se generaban tantas muestras como la clase mayoritaria, y el *resampleo enfocado* consistía en el *oversampling* sólo de aquellos ejemplos minoritarios que ocurrieron en el límite entre las clases minoritaria y mayoritaria. El *subsampling* aleatorio involucró el submuestreo de las muestras de la clase mayoritaria al azar hasta que sus números coincidieran con el número de muestras de la clase minoritaria. El *subsampling enfocado* involucró el submuestreo de las muestras de la clase mayoritaria situadas más lejos. El autor señaló que ambos métodos de muestreo eran eficaces, y también observó

que el uso de sofisticadas técnicas de muestreo no aportaba ninguna ventaja clara en el dominio considerado. Sin embargo, sus metodologías de *oversampling* no construyeron nuevos ejemplos.

Por otra parte, el autor también incide en el estudio de Ling [80], donde se combina el muestreo excesivo de la clase minoritaria con sub-muestreo de la clase mayoritaria. Se propuso que los ejemplos de prueba se clasificaran por una medida de confianza y luego se utilizarían como criterios de evaluación. En un experimento, se submuestreó la clase mayoritaria y se observó que el mejor índice se obtiene cuando las clases están representadas equitativamente. En otro experimento, se sometió a sobremuestreo los ejemplos positivos (minoritarios) con sustitución para hacerlos coincidir con el número de ejemplos negativos (mayoritarios). En el estudio presentado se concluye que la combinación de sobre-muestreo y sub-muestreo no proporcionó una mejora significativa.

Otros autores como Batista [14] evaluaron varias metodologías de muestreo en una variedad de sets de datos con diferentes distribuciones de las clases. Incluyeron varios métodos de sobremuestreo y submuestreo, y concluyeron que mediante la creación de muestras artificiales a través de la técnica SMOTE[25] se proporcionan muy buenos resultados para conjuntos de datos con un pequeño número de ejemplos positivos de clase.

Adicionalmente, existe un algoritmo llamado SMOTEBoost que combina SMOTE y un procedimiento de *boosting* [27]. La técnica SMOTE funciona relativamente bien para mejorar el rendimiento sobre la clase minoritaria, y al combinarla con una técnica de *boosting* se mantiene el rendimiento sobre el set de datos completo.

### 2.3.4. Metodología para la selección de predictores

Existen diferentes métodos de selección de predictores en base a su importancia cuando la variable a predecir es del tipo factor.

Esta investigación centrará la atención fundamentalmente en dos:

- Filtrado de características previo para crear un subset que se usará para insertar como entrada de un algoritmo de clasificación.
- Técnicas Wrapper: el algoritmo de clasificación se aplica sobre el conjunto de datos con el fin de identificar las mejores características.

#### 2.3.4.1. Predictores como entradas para la curva ROC

Cuando se trata de dos clases a predecir, es decir, dos posibles valores de la variable objetivo, el área encerrada bajo la curva ROC permite cuantificar la relevancia del predictor. El acrónimo ROC significa *'Receiver Operating Characteristic'*, sin embargo, Swets [106] sitúa sus orígenes en Psicofísica y Teoría de la Señal, donde en la $2^a$ Guerra Mundial los operadores de radar debían tomar las decisiones de lo que era

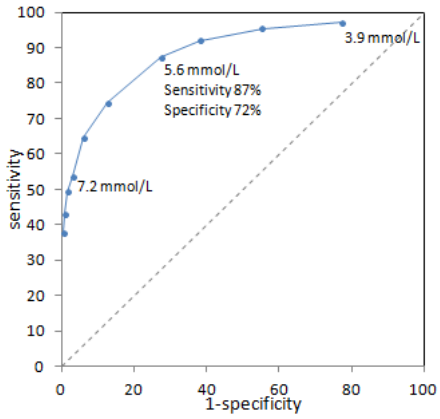



**Figura 2.5:** Representación de la Sensibilidad y Especificidad en una curva ROC en función del umbral o threshold escogido

ruido en la pantalla, o componentes enemigos o amigos. La curva ROC se empleaba, entonces, para la detección de señales con el fin de caracterizar la relación entre la tasa de éxito y la tasa de falsas alarmas en canales ruidosos.

En función del umbral que se escoja en el rango de valores de la variable objetivo, la curva ROC se puede utilizar para detectar valores dicotómicos, es decir, 0 o 1, sí o no, etc. Una prueba puede tener un resultado 'positivo' o 'negativo', aunque, en realidad, los resultados de las pruebas se encuentran a lo largo de un continuo. Por ejemplo [56], los valores de la glucemia en ayunas pueden oscilar entre 50 mg/dl y 250 mg/dl. Se puede escoger 120 mg/dl como punto de corte o *threshold* para una prueba 'positiva', pero claramente es posible elegir otros valores.

A su vez, las curvas ROC representan el rendimiento de un clasificador sin tener en cuenta la distribución de clase o los costes de error. Para construir la curva se representan la tasa de Verdaderos Positivos ($VP$) o *True Positive* ($TP$) en el eje vertical con respecto a la tasa de Verdaderos Negativos ($VN$) o *True Negative* ($TN$) en el eje horizontal. Por tanto, la curva ROC representa la *sensibilidad* frente a la especificidad. En la figura 2.5 se muestra un ejemplo de la representación de una curva ROC.

En este sentido se usan los datos del predictor como entradas en la curva ROC. Si el predictor puede separar perfectamente las clases, existirá un punto de corte para el predictor que consiga una sensibilidad y especificidad de $1$ y el área debajo de la curva será $1$. Por el contrario, si el área de la curva es de aproximadamente $0{,}5$ el predictor no aportará información al modelo.

La definición de VP (verdaderos positivos), FP (falsos positivos), FN (falsos negativos) y VN (verdaderos negativos) se presenta en la figura 2.6, aplicada al caso de un diagnóstico clínico de una enfermedad. Es necesario especificar dos cantidades

| | | Estado real | | |
|---|---|---|---|---|
| | | Positivo | Negativo | Suma |
| Diagnóst. | Positivo | Verdaderos Positivos (VP)<br>a | Falsos Positivos (FP) (Falsas alarmas)<br>b | a+b |
| | Negativo | Falsos Negativos (FN) (omisiones)<br>c | Verdaderos Negativos (VN)<br>d | c+d |
| | Suma | (a+c) | (b+d) | a+b+c+d |

**Figura 2.6:** Clasificación del resultado de una prueba diagnóstica

con motivo de la obtención del comportamiento diagnóstico, para el ejemplo que nos ocupa. Por ello, a continuación se definen dos medidas:

- La *sensibilidad* se define como la capacidad que posee un modelo para detectar los verdaderos positivos (TP). Aplicada esta definición al ejemplo del caso de diagnóstico, se definiría como la proporción de casos diagnosticados como afirmativos, a partir del umbral escogido, en los que se puede observar que, efectivamente, sucede el estado que se pretende diagnosticar.
- La *especificidad* se define como la capacidad de un modelo para detectar los verdaderos negativos (TN). Es decir, en el mismo ejemplo se correspondería con la proporción de casos diagnosticados como negativos, a partir del umbral escogido, para los que es posible observar que, efectivamente, no sucede el estado que se pretende diagnosticar.

Un dato importante respecto a la sensibilidad y la especificidad es que, además de encontrarse relacionadas entre sí, ambas proporciones se pueden obtener para cada punto de corte en la curva ROC, y dependen de éste, por lo que si se fija un umbral muy pequeño, se obtendrá una especificidad alta pero una sensibilidad baja, y viceversa.

$$Sensibilidad = \frac{TP}{TP + FN} \tag{2.2}$$

Siendo $TP$ los *Verdaderos Positivos*, mientras que $(TP + FN)$ son *todos los realmente positivos*.

La sensibilidad es, en resumen, la capacidad de un sistema para detectar un hecho, o si nos centramos en términos de medicina, la probabilidad de que una prueba diagnóstica dé positivo cuando el paciente ha contraído una determinada enfermedad.

$$Sensibilidad = \frac{TN}{TN + FP} \tag{2.3}$$

Siendo $TN$ los *Verdaderos Negativos*, mientras que $(TN + FP)$ son *todos los realmente negativos*.

Para resumir las curvas ROC en una sola cantidad, en determinadas ocasiones se emplea el término '*área bajo la curva (AUC)*' donde, a grandes rasgos, cuanto mayor es el área, mejor es el modelo. El área también tiene una interpretación positiva, como la probabilidad de que el modelo clasifique una instancia positiva elegida aleatoriamente sobre una negativa elegida también aleatoriamente.

Para interpretar el área bajo la curva (AUC) ROC, continuando con el ejemplo en medicina, el valor del AUC ROC de $0{,}91$ significa que un individuo seleccionado aleatoriamente del grupo positivo "*[...] tiene un valor en la prueba mayor que el de un individuo elegido aleatoriamente del grupo negativo un 91 % de las veces [...]*"[119].

Para usar el método basado en el área bajo la curva ROC para cada predictor respecto a las clases con los sets de datos de tráfico empleamos la función de R *filterVarImp*, perteneciente al paquete *Caret*[75] de *R*.

Adicionalmente, también existe la posibilidad de realizar una clasificación a través de un análisis de la curva ROC para cada predictor. Por ejemplo, cuando se tienen modelos de dos clases como variable a obtener en la predicción o clasificación, se aplican una serie de puntos de corte a los datos de los predictores para predecir la clase. La sensibilidad y especificidad se calculan para cada corte y, adicionalmente, se calcula la curva ROC. El área encerrada bajo la curva ROC se utiliza como medida de importancia de cada variable de forma individual. Como se verá más adelante en capítulos experimentales, esta técnica se empleará en nuestra investigación para hallar una clasificación en base a la importancia de los predictores.

Un ejemplo de área encerrada bajo la curva ROC o AUC es el que se observa en la tabla 2.1 para la clase X0.

| | X0 |
|---|---|
| Casualty_Type | 0.6441775 |
| Number_of_Vehicles | 0.6123903 |
| Vehicle_Type | 0.5760128 |
| Engine_Capacity_.CC. | 0.5653059 |

**Tabla 2.1:** Ejemplo de área encerrada bajo la curva ROC para la clase X0

Para finalizar esta sección, se pasan a definir otros parámetros importantes:

- *Accuracy* es la medida que indica las predicciones correctamente y no correctamente acertadas. Viene definido por la ecuación 2.4. Puede ser problemático cuando las clases están desbalanceadas por lo que en este estudio se plantea no usarlo como medida de ajuste del modelo, salvo en alguna excepción.

$$Accuracy = \frac{TP + TN}{NúmeroTotalPredicciones} \tag{2.4}$$

- *Kappa*: el parámetro definido por la ecuación 2.5 es considerado la tasa de error esperada.

$$Kappa = \frac{AccuracyObservado - AccuracyEsperado}{1 - AccuracyEsperado} \tag{2.5}$$

#### 2.3.4.2. Selección de características mediante el algortimo Relief

Otras técnicas válidas para cuantificar la importancia de cada predictor se basan en el denominado algoritmo *Relief* [72]. *Relief* resulta un método simple, rápido y eficaz para obtener el peso de cada predictor. Las estimaciones de *Relief* son mejores que las estimaciones estadísticas de atributo habituales, como la correlación o la covarianza, porque tiene en cuenta las interrelaciones de los atributos.

Una de las mayores ventajas del algoritmo *Relief* es que puede ser usado tanto para predictores continuos como para variables de tipo factor, lo que resulta de vital importancia dados los tipos de datos de las características que se van a manejar en este trabajo. Adicionalmente, este algoritmo puede reconocer relaciones no lineales entre los predictores y la variable a predecir.

Esta técnica emplea una selección de puntos aleatorios y sus vecinos más cercanos para evaluar cada predictor de forma aislada. Para un predictor en particular, la puntuación trata de caracterizar la separación entre las clases en secciones aisladas de los datos. La salida del algoritmo Relief es un peso entre $-1$ y $1$ para cada atributo. A mayores pesos positivos, más atributos predictivos [92]. Los atributos con pesos negativos o los cercanos a cero son aquellos que Relief descarta.

El peso de un atributo se actualiza iterativamente seleccionando una muestra a partir de los datos e identificando la muestra inmediatamente más cercana que pertenece a la misma clase y la muestra inmediatamente más cercana que pertenece a la clase contraria. Un cambio en el valor del atributo acompañado por un cambio en la clase lleva a la ponderación del atributo basado en la intuición de que el cambio del mismo podría ser responsable del cambio de clase. Por otro lado, un cambio en el valor del atributo sin que haya ningún cambio en la clase, conduce a una ponderación a la baja del atributo basado en la observación de que el cambio del mismo no tuvo ningún efecto sobre la clase.

Este procedimiento de actualización del peso del atributo se realiza para un conjunto aleatorio de muestras en los datos o para cada muestra en ellos. Las actualizaciones de peso se promedian entonces para que el peso final esté en el intervalo $[-1, 1]$, como se comentaba anteriormente.

El atributo de peso estimado por *Relief* tiene una interpretación probabilística. Es proporcional a la diferencia entre dos probabilidades condicionales, es decir, la probabilidad de que el valor del atributo sea diferente condicionado por la muestra inmediatamente más cercana que pertenece a la clase contraria y la muestra inmediatamente más cercana que pertenece a la misma clase dados respectivamente [91].

### Función permuteRelief en R

En esta investigación se emplea una variante de *Relief* llamada *permuteRelief*, nombre que recibe igualmente la función en lenguaje *R* para trabajar con dicha variante. Básicamente, los que hace es emplear una aproximación a través de permutación para determinar la magnitud relativa de las puntuaciones otorgadas por *Relief*. ([72] y [73]).

Las puntuaciones de cada predictor se calculan utilizando los datos originales y después de que los datos a partir de la variable a predecir se mezclen aleatoriamente. Se determinan la media y la desviación estándar de los valores permutados y se determina una versión estandarizada de las puntuaciones observadas, restando las medias de los valores permutados de los valores originales y, a continuación, dividiendo cada uno por la desviación estándar correspondiente. En resumen, se trata de una estandarización con permuta.

Los valores aportados son una lista con los siguientes elementos:

- *estandarizados*: un vector con las puntuaciones estandarizadas de los predictores.
- *permutaciones*: los valores de las puntuaciones de los permutados, para trazar con el fin de evaluar la distribución de la permutación.
- *observados*: la puntuación de los valores observados.

Los atributos redundantes tendrán asignados una puntuación menor, sin embargo, una puntuación alta indica una interacción mayor del atributo.

### Algoritmo Relief-F

Adicionalmente al algoritmo *Relief*, existe un algoritmo de ponderación de características basado en filtros, que resulta una versión extendida del primero, y que se denomina *Relief-F* [73]. A diferencia del algoritmo *Relief* original, que únicamente podía abordar conjuntos de datos de dos clases a predecir, el algoritmo *Relief-F* se ocupa de conjuntos de datos de varias clases. La simpleza detrás del algoritmo *Relief-F* es que una buena característica es una característica con poca variación dentro de la clase y muchas variaciones entre clases. Una mala característica se caracteriza por variaciones dentro de clase y entre clases de magnitudes aproximadamente iguales.

#### 2.3.4.3. Métodos Wrapper

Los métodos *Wrapper*, aunques proporcionan resultados más precisos que aquellos que realizan un filtrado de características [81], son computacionalmente mucho más exigentes. Se ha de tener en consideración que estos métodos realizan agrupaciones de características en base a su nivel de predicción, y para la elección de estas agrupaciones realizan una exploración del conjunto de características optimizando el algoritmo que se emplea para la clasificación.

Este tipo de algoritmos se subdividen en tres modalidades: los primeros añaden variable a variable al modelo (*forward*), las eliminan una a una del mismo (*backward*), o se realiza una combinación *forward-backward* llamada pasos sucesivos. Cualquiera de estas dos formas de proceder con el conjunto de variables se emplea para encontrar la combinación óptima que maximice el rendimiento del modelo.

El procedimiento de tipo *forward* consiste en evaluar predictores uno por uno basándose, por ejemplo, en modelos de regresión lineal. Si un predictor posee un p-valor por debajo de un umbral escogido, el predictor asociado con el menor valor se añade al modelo y el proceso vuelve a comenzar. El algoritmo finaliza cuando ninguno de los p-valores de los predictores que quedan son estadísticamente significativos. Sin embargo, este tipo de procedimiento no re-evalúa soluciones antiguas, y usa repetidamente tests de validación de hipótesis, lo que puede suponer que se invaliden muchas de sus propiedades estadísticas, ya que los mismo datos se evalúan muchas veces [76].

Los métodos automáticos son útiles cuando el número de variables explicativas es grande y no es factible, por una cuestión de tiempo de computación y de metodología casi manual, adaptar todos los modelos posibles. En este caso, es más eficiente usar un algoritmo de búsqueda (por ejemplo, selección directa, eliminación *backward* o regresión paso a paso) para encontrar el mejor modelo.

##### a. Backward Selection: Selección de características a partir de ejecuciones recursivas de GLM o BayesGLM

En este trabajo se emplearán algoritmos de selección de tipo *backward* mediante GLM y BayesGLM, así como otro tipo de métodos *wrapper* basados en la eliminación de características de forma recursiva, como el tipo *Recursive Feature Elimination* (RFE).

Para la selección de características a partir de algoritmos basados en la eliminación de ellas (*backward*), como indica Kuhn [76] el modelo inicial contiene todos los $P$ predictores y se van eliminando progresivamente aquellos que no tienen una contribución significativa para el modelo. Para los estudios de esta Tesis Doctoral donde hemos empleado este tipo de selección de características, nos hemos decantado por emplear un modelo de regresión logística, ya que éstos emplean una serie de coeficientes para el modelo que nos permiten cuantificar la importancia de cada predictor.

Para escoger el modelo GLM adecuado, tendremos que escoger aquel que tenga la mayor proporción de la varianza, siendo significativas las características elegidas.

### b. Recursive Feature Elimination (RFE)

Según Granitto [52] existen dos objetivos diferenciados para realizar una selección de características:

- Encontrar un subconjunto de características con el error de generalización mínimo posible.
- Seleccionar el subconjunto más pequeño posible con una capacidad de discriminación dada.

Los algoritmos desarrollados para la Eliminación Recursiva de Características, en adelante RFE, poseen un buen rendimiento con una carga computacional no muy elevada [9]. RFE es básicamente un proceso recursivo que clasifica las características según alguna medida de su importancia. En cada iteración se mide la importancia de cada característica y se elimina la menos relevante. También es posible eliminar varias características al mismo tiempo para que el proceso sea más rápido.

Según Alexandrinis [9], que el proceso sea recursivo es necesario, porque para algunas medidas, la importancia relativa de cada característica puede cambiar sustancialmente cuando se evalúa sobre un subconjunto diferente de características durante el proceso de eliminación gradual. Esto posee mayor relevancia si cabe cuando existen características que están fuertemente correladas.

### c. Random Forest-RFE (RF-RFE)

Las funciones para la evaluación de la importancia de los predictores pueden dividirse en dos grupos: los que usan la información del modelo y los que no. Utilizar un enfoque basado en modelos está más estrechamente vinculado al rendimiento del mismo y puede ser capaz de incorporar la estructura de correlación entre los predictores al calcular su importancia.

Random Forest es un algoritmo multi-clase con una medida interna (imparcial) de importancia de características.

El algoritmo de RFE basado en Random Forest registra para cada árbol la exactitud (parámetro *accuracy*). A continuación, se hace lo mismo después de permutar cada predictor. Seguidamente, la diferencia entre los dos parámetros de *accuracy* se promedia sobre todos los árboles, y se normaliza por el error estándar. Para la regresión, el Error Cuadrático Medio (MSE) se calcula sobre los datos para cada árbol, y luego se vuelve a calcular el MSE después de permutar una variable. La diferencia se promedia y se normaliza por el error estándar. Si el error estándar es igual a $0$ para una variable, la división no se realiza.

En R existe un paquete para calcular la importancia de los predictores llamado *FSelector Package*, con una función llamada *random.forest.importance()*.

#### d. FSelector Package: Estadísticas Chi cuadrado

Los tests estadísticos mediante chi-cuadrado se utilizan para determinar la dependencia de dos variables y se pueden aplicar únicamente sobre datos organizados en categorías o datos nominales. Si por el contrario nuestro set de datos contiene variables continuas, podemos aun así aplicar la técnica chi-cuadrado siempre que se haya discretizado y categorizado previamente dichas variables.

El funcionamiento de dicho test es muy sencillo: dadas las características que definen un set de datos junto con la clase que se encuentra incluida dentro de las características anteriores, y que será la variable a predecir, la técnica chi-cuadrado para seleccionar características de un set de datos consiste en calcular las estadísticas chi-cuadrado entre cada variable del set de datos y la variable a estimar, observando la relación entre dichas variables.

Si la variable que deseamos predecir y la escogida del set de datos son independientes, se puede descartar dicha variable del set de datos. Si por el contrario ambas variables son dependientes, la variable del set de datos escogida se considerará que tiene un peso de importancia alto dentro de la selección y no se descartará.

### 2.3.5. Técnicas de remuestreo: validación cruzada

La Validación Cruzada es una técnica en la que no se incluye todo el set de datos para realizar el ajuste del modelo, sino que los datos se separan en dos subconjuntos, conjunto de *training*, usado para ajustar el modelo, y conjunto de *testing* usado para contrastar la bondad del ajuste del modelo.

La validación cruzada se emplea para garantizar que los datos del subset de entrenamiento o *training* son independientes de los datos del subset de pruebas o *testing*.

Cuando se habla de la validación cruzada de $K$ iteraciones o *K-fold* lo que se hace es particionar el set de entrenamiento en $K$ subconjuntos de aproximadamente el mismo número de muestras aleatoriamente que tenía éste. Un subconjunto se entrena usándolo como datos de prueba y el resto de grupos $(K-1)$ como datos de entrenamiento. Este proceso se repite $K$ veces con cada posible subconjunto de datos de prueba, donde el primer grupo de muestras se devuelve al set de entrenamiento y se repite con el segundo subconjunto la misma operación, y así hasta $K$ veces. El proceso para un ejemplo de 4 subconjuntos o *folds* se puede observar gráficamente en la figura 2.7.

Se realiza la media aritmética y el error estándar de los resultados de cada una de las K iteraciones para obtener un único resultado. De esta forma se logra comprender la relación entre los parámetros de ajuste del modelo junto con su utilidad. Se ha de tener en cuenta que a pesar de ser un método bastante preciso, resulta muy costoso desde el punto de vista computacional.

Habitualmente se suelen escoger valores para $K$ de 5 o 10, aunque a medida que $K$ se escoge mayor, la diferencia de tamaño entre el set de entrenamiento y los

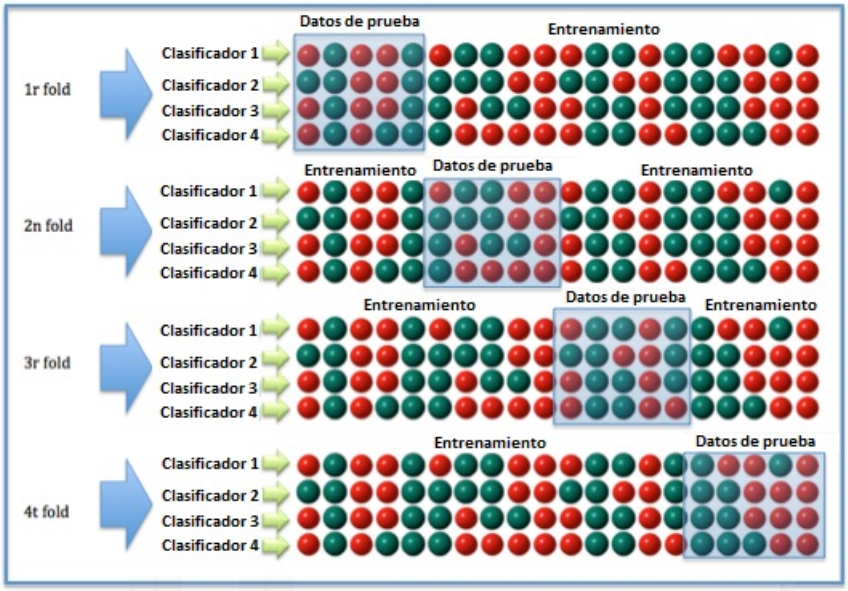



**Figura 2.7:** Validación cruzada de 4 iteraciones y 4 clasificadores (cortesía de Joan Domenech Licencia CC BY-SA 3.0)

subsets de muestras se hace más pequeña, por lo que a medida que la diferencia decrece, el sesgo es menor, es decir, para $K = 10$ el sesgo será menor que para $K = 5$, entendiendo sesgo como la diferencia entre el valor estimado y el verdadero. Para sets de entrenamiento muy grandes, los problemas potenciales con la varianza y el sesgo se vuelven insignificantes.

## 2.4. Modelos de regresión y clasificación

### 2.4.1. Modelos de regresión lineal

En estadística, la palabra *regresión* consiste en el proceso de predicción de una cantidad numérica, mediante esta u otras que pueden ser numéricas o numéricas y cualitativas.

Los modelos basados en regresión lineal poseen tres tipos de componentes:

- Componente aleatoria $(y)$ o variable respuesta junto con su distribución de probabilidad, que en el modelo de regresión lineal es un distribución normal.
- Componentes sistemáticas o predictores, que son variables independientes entre sí y explican el valor de la componente aleatoria.
- Función *link* o función predictora lineal, que se corresponde con el valor que se espera de $Y$ y no es más que una combinación lineal de los predictores.

y dichos modelos lineales se pueden modelar mediante la fórmula o función *link* definida por la ecuación 2.6.

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + ... + b_P x_{iP} + e_i \tag{2.6}$$

donde:

- $y_i$ representa la variable numérica o componente aleatoria para la muestra i.
- $b_0$ representa el *intercept* estimado.
- $b_j$ representa el coeficiente estimado para el predictor j de $X_j$.
- $x_{ij}$ es la componente sistemática y representa el valor del predictor j-ésimo para la muestra i.
- $e_i$ representa el error aleatorio que no puede ser explicado por el modelo en la muestra i.

Cada uno de estos modelos busca encontrar estimaciones de los parámetros para que la suma de los errores al cuadrado se minimice.

La principal ventaja del uso de modelos de regresión lineal es que el resultado es fácilmente interpretable, ya que el coeficiente estimado para cada predictor está directamente relacionado con el incremento o disminución de la respuesta en base al valor del coeficiente.

En los modelos lineales generalizados, con *generalizados* se refiere a la dependencia potencial de que exista más de una variable explicativa, en contraposición con el modelo lineal simple (ecuación 2.7).

$$y_i = b_0 + b_1 x_i + e_i \tag{2.7}$$

Los modelos lineales son más fáciles de visualizar en dos dimensiones, donde son equivalentes a dibujar una línea recta a través de un conjunto de puntos de datos. Para *p* variables se tiene un hiperplano de dimensión *P*.

Adicionalmente, este tipo de modelos presenta la ventaja de poder hallar de forma sencilla los errores estándar de los coeficientes, lo que permite evaluar el significado estadístico de cada predictor en el modelo.

Se asume que los errores $e_i$ son independientes e idénticamente distribuidos de tal forma que:

$$E[e_i] = 0 \tag{2.8}$$

$$var[e_i] = \sigma^2 \tag{2.9}$$

Existen algunas situaciones en las que los modelos lineales generales no son apropiados:

- El rango de $y$ es restringido (por ejemplo, binario)
- La varianza de $y$ depende de la media

En esos casos es cuando debemos trabajar con otros modelos, como por ejemplo modelos basados en regresión logística.

### 2.4.2. Modelos de regresión logística

Es posible determinar los principales objetivos del modelo de regresión logística como:

- determinar la existencia o ausencia de relación entre una o más variables independientes ($X_i$) y una variable dependiente dicotómica ($Y$).
- medir el signo de dicha relación, en caso de que exista.
- estimar la probabilidad de que se produzca el suceso definido como $Y = 1$, en función de los valores que adopten las variables independientes.

Por tanto, si suponemos que:

$$Y_i \sim Binomial(X_i, p_i) \tag{2.10}$$

y considerando que el modelo más sencillo es aquél que incluye una única variable independiente (X):

$$Y = \alpha + \beta X + u \tag{2.11}$$

donde '$\alpha$' es el término independiente o constante; '$\beta$' es el coeficiente de regresión asociado a la variable independiente; y '$u$' es el término de perturbación aleatoria.

Sin embargo, como $Y$ sólo toma el valor $1$ ó $0$ para cada muestra, entonces si $Y = 1$:

$$P(Y = 1) = \alpha + \beta X \tag{2.12}$$

siendo P la probabilidad estimada de que un accidente escogido al azar pueda ser leve.

Sin embargo, debido a que $Y$ se encuentra en un rango real entre 0 y 1, la ecuación 2.12 presenta una serie de limitaciones a la hora de estimar valores en ese rango, por lo que el modelo se ajusta a:

$$P = e^{(\alpha+\beta X)} \tag{2.13}$$

sin embargo, el modelo descrito por la ecuación 2.13 presenta un problema adicional, ya que permite estimar valores de $P > 1$, y la intención del modelo es que se encuentre restringido en el rango $[0, 1]$. Por todo ello, se escogerá la función de probabilidad definida en la ecuación 2.14.

$$P = \frac{e^{(\alpha+\beta X)}}{1 + e^{(\alpha+\beta X)}} \tag{2.14}$$

La función definida por esta ecuación es exponencial aunque puede transformarse, tomando logaritmos neperianos ($Ln$), en una función lineal:

$$Ln[\frac{P}{1-P}] = \alpha + \beta X \tag{2.15}$$

Si $Y = 1$, entonces:

$$Ln[\frac{P(Y=1)}{1-P(Y=1)}] = Ln[\frac{P(Y=1)}{P(Y=0)}] \tag{2.16}$$

Definimos como *odds* el cociente $\frac{P(Y=1)}{P(Y=0)}$, consistente en la razón entre la probabilidad de que se produzca un accidente y la probabilidad de que no se produzca.

La *odds* puede tomar valores desde 0 cuando $P(Y = 1) = 0$ hasta $\infty$ cuando $P(Y = 1) = 1$.

Otro término relevante es *logit* (L), definido en la ecuación 2.17:

$$L = Ln(\textit{odds}) = Ln[\frac{P}{1-P}] \tag{2.17}$$

donde los valores de L van desde $-\infty$ a $+\infty$ si $P(Y = 1) = 0$ y $P(Y = 1) = 1$, respectivamente.

La probabilidad estimada 'P(Y=1)' es una función *sigmoide*, como se observa en la figura 2.8.

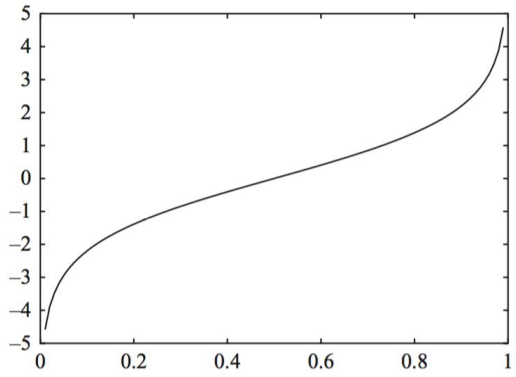



**Figura 2.8:** Función Sigmoide

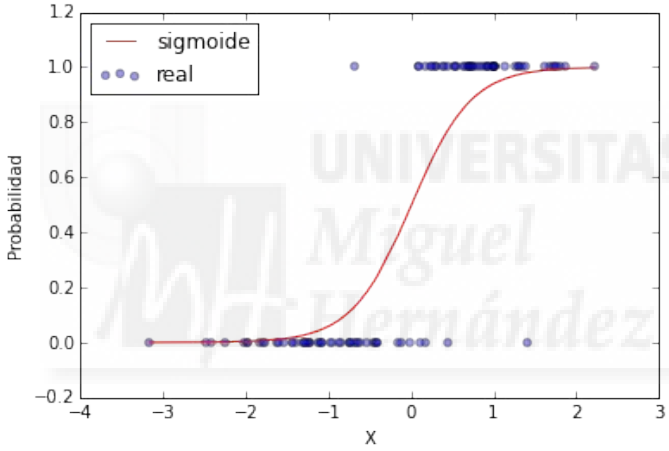



**Figura 2.9:** Ejemplo de clasificación por categorías con un umbral de 0.5 para separar ambas clases

En dicha figura se aprecia claramente la curva en forma de $S$ de la función logística. En caso de que se tuviera un punto con $X = 0$ tendría el 50 % de probabilidades de pertenecer a cualquiera de las dos clases $Y = 0$ ó $Y = 1$.

En caso de desear realizar una clasificación por categorías o clases se debería definir un valor umbral, de tal forma, que cuando la función logística asigne una probabilidad mayor a, por ejemplo, $0{,}5$ entonces se asignaría a esa categoría.

En la figura 2.9 se puede observar claramente este ejemplo para dos clases.

#### 2.4.2.1. Cálculo de la desviación

La *desviación* es una medida de bondad de ajuste de un modelo lineal generalizado. O más bien, es una medida de lo malo que es el ajuste. Los números más altos

indican un ajuste peor.

En R se reportan dos formas de desviación: la desviación nula y la desviación residual.

- La desviación nula muestra cuán bien el modelo predice la variable respuesta que incluye sólo el 'intercept' (media grande, es decir, la media ponderada de todos. Si el intercept es considerablemente $> 0$, hay que estudiarlo).
- La desviación residual muestra como es la respuesta predicha por el modelo cuando los predictores son incluidos.

#### 2.4.2.2. BayesGLM

A pesar de lo explicado anteriormente, el modelo de regresión logística (*GLM*) falla en la separación y proporciona respuestas ruidosas cuando hay escasos datos. Adicionalmente, *GLM* no funciona bien para múltiples predictores. Los algoritmos de aprendizaje no bayesianos subestiman la incertidumbre en las predicciones.

Sin embargo, *BayesGLM* proporciona una inferencia inductiva coherente y consistente, una interpretación intuitiva y permite una clase más amplia de inferencia aplicable que la interpretación de la frecuencia de probabilidad. La información previa puede permitir una inferencia razonable con muestras moderadas.

Por tanto, la mejor ventaja del enfoque bayesiano es el uso de información externa para mejorar las estimaciones de los coeficientes del modelo lineal.

Respecto a la función en lenguaje *R*, *Bayesglm()*, realiza un escalado automático a priori y define los siguientes parámetros de ajuste:

- La columna denominada '*Accuracy*' es la tasa de coincidencia global promediada sobre las iteraciones de validación cruzada que realiza.
- La desviación estándar de dicha coincidencia también se calcula a partir de los resultados de la validación cruzada.
- La columna '*Kappa*' es la estadística de Kappa de Cohen (no ponderada) promediada a través de los resultados del remuestreado.

#### 2.4.2.3. Árboles de regresión

Los modelos basados en árboles son altamente interpretables, además de poder manejar distintos tipos de predictores, desde continuos a nominales, sin necesidad de realizar preprocesamiento alguno. Además, son muy sencillos de implementar y son capaces de manejar sets de datos con valores perdidos. Muchos de estos árboles poseen métodos de selección de variables. La ventaja del árbol es que no necesita realizar ciertas asunciones que se hacen en un modelo de regresión.

Sin embargo, aquellos modelos basados en un solo árbol presentan determinadas debilidades [76] como:

- *Inestabilidad del modelo*: pequeños cambios en los datos pueden hacer que la estructura del árbol cambie, y por ende la interpretación.
- Tendencia a poseer un *error de predicción* elevado si la relación entre predictores y variable a predecir no se puede definir por subespacios rectangulares de los predictores.

De estas debilidades surgen otro tipo de modelos basados en conjuntos de árboles, en lugar de estar compuestos por un solo árbol.

Para entender en qué consiste esta agrupación de árboles de regresión, la cual ha sido empleada en esta tesis, en primer lugar estudiaremos el funcionamiento de los árboles de regresión más básicos.

Estos tipos de árboles básicos dividen los datos en pequeños grupos que son más homogéneos en relación con la variable a estimar. Para ello, los árboles determinan [76]:

- El predictor a dividir y en qué proporción.
- La profundidad del árbol.
- La ecuación de predicción en los nodos terminales.



En modelos de regresión basados en árbol, como el método CART, el modelo comienza con todo el set de datos, al que denominaremos $S$, y busca cada valor distinto de cada predictor para buscar el predictor y valor de corte que divide los datos en dos grupos, $S_1$ y $S_2$ (ec. 2.18) de tal forma que la suma total de los errores cuadráticos se minimizan [20] tal y como se puede observar en la ecuación 2.19.

$$S_1 \cup S_2 = S \tag{2.18}$$

$$SSE = \sum_{i \in S_1} (y_i - \bar{y_1})^2 + \sum_{i \in S_2} (y_i - \bar{y_2})^2 \tag{2.19}$$

donde $\bar{y_1}$ e $\bar{y_2}$ son los valores medios de las variables de salida del set de datos de entrenamiento de los grupos $S_1$ y $S_2$. Este método busca el predictor y el valor de partición de los datos que mejor reduce $SSE$.

### Random Forest para regresión

El ensamble de varios modelos basados en árboles mejora el rendimiento a la hora de predecir con respecto al modelo basado en un sólo árbol, ya que el primero reduce la varianza de la predicción. Sin embargo, los árboles que participan en el ensamble de árboles no son del todo independientes entre sí, debido a que todos los predictores originales se consideran en cada agrupación o *split* de cada árbol. Esto puede ocasionar lo que se conoce como una *correlación entre árboles* [45]. Si se consigue reducir la correlación entre árboles, se mejora considerablemente el rendimiento del ensamble.

Para conseguir reducir la correlación entre predictores, es posible añadir aleatoriedad en la construcción del árbol (muestras bootstrap). Dietterich [39] desarrolló la idea de la selección aleatoria de la división, donde los árboles se construyen usando un subset de datos aleatorio de los $k$ mejores predictores en cada división del árbol. Breiman [19] construyó un algoritmo unificado llamado *Random Forest*.

El algoritmo *Random Forest* trabaja con una extensa colección de árboles de decisión decorrelados, de ahí que hablemos de bosque (*Forest*) puesto que trabaja con numerosos árboles de decisión.

Cada modelo del ensamblaje se usa para generar una predicción para una nueva muestra y estas predicciones se promedian para dar la predicción del conjunto de árboles. Puesto que el algoritmo selecciona aleatoriamente los predictores para cada split de datos, la correlación entre árboles disminuye considerablemente.

Los parámetros para ajustar un modelo basado en Random Forest consiste en:

- *Número de árboles a usar*: este parámetro reduce la varianza. La recomendación está entre 100 y 500 árboles, aunque se recomienda parar en un número de árboles a partir del cual se estabilice la AUC ROC.
- *Número de predictores k* para usar en cada split de datos $(m_{try})$ : reduce la correlación entre los árboles y reduce la varianza. Se recomienda que su valor sea la raíz cuadrada del número de características.
- *Número de hojas* en el nodo terminal (nodesize): reduce el sesgo. El valor por defecto es 1.

El parámetro $k$ se ha de seleccionar aleatoriamente para escoger el número de predictores de cada split de datos. A este parámetro se le conoce como $m_{try}$ y Breiman [19] recomienda que el valor de este parámetro se corresponda con $1/3$ del número de predictores.

Kuhn [76] recomienda que para ajustar el parámetro $m_{try}$, y puesto que el algoritmo de Random Forest requiere muchos recursos computacionales, se comience con $5$ valores de $k$ que se encuentren uniformemente espaciados en el rango de $2$ a $P$. Adicionalmente, se debe elegir el número de árboles que compondrán el bosque.

Breiman [19] probó que el algoritmo random forest se encuentra protegido frente al sobreentrenamiento (*overfitting*), así que no ocurrirá nada si se emplea un número elevado de árboles para componer el bosque. Kuhn [76] recomienda usar 1.000 árboles en un inicio, y si se considera que el rendimiento (ec. 2.4) sigue aumentando con 1.000 árboles, incrementar el número de árboles hasta que los niveles de rendimiento se estabilicen.

El algoritmo de Random Forest reduce los ratios de error al escoger árboles independientes y fuertes ya que son seleccionados independientemente de los árboles por los que se ha optado anteriormente, por lo que Random Forest es robusto frente a respuestas ruidosas. Al mismo tiempo, esta independencia de los árboles puede subentrenar los datos cuando la respuesta no es ruidosa.

Hablando de la eficiencia computacional, Random Rorest es muy eficiente, ya que el proceso de construcción del árbol únicamente necesita evaluar una fracción de los predictores originales en cada split, a pesar de que se requieren más árboles para este algoritmo en comparación, por ejemplo, de un método basado en remuestreo.

Sin embargo, no todo iba a resultar positivo. La propia naturaleza de ensamblaje en Random Forest hace que sea imposible entender la relación entre los predictores y la respuesta. En cualquier caso, sí que es posible cuantificar el impacto de cada predictor en el modelo. De todas formas, hemos de ser cuidadosos a la hora de cuantificar dicho impacto, ya que es posible que si tenemos predictores correlados obtengamos un impacto significativo sobre el modelo [105]. Strobl También demostró que el parámetro $m_{try}$ tiene un gran efecto sobre los valores de importancia de cada predictor en el modelo.

#### 2.4.2.4. Random Forest para clasificación

Para que Random Forest se pueda utilizar para clasificación se requiere un pequeño retoque respecto al algoritmo de regresión: se usa un árbol de clasificación en lugar de un árbol de regresión. Como se explicaba anteriormente, cada árbol en el bosque emite un voto para la clasificación de una nueva muestra y la proporción de los votos en cada clase a lo largo del ensamblaje es el vector de probabilidad predicho [76].

Aunque el tipo de árbol cambie respecto al algoritmo de regresión, el número de predictores a escoger aleatoriamente en cada split es el mismo ($m_{try}$). Sin embargo, Breiman [19] recomienda que para problemas de clasificación se escoja como parámetro $m_{try}$ la raíz cuadrada del número de predictores. El resto de recomendaciones que nos aportaba Kuhn para regresión se mantienen, es decir, usar 1.000 árboles en un inicio y comenzar con $5$ valores de $k$ que se encuentren uniformemente espaciados en el rango de $2$ a $P$, siendo $P$ el número de predictores.

Una de las principales diferencias del uso de Random Forest para clasificación respecto a regresión tiene que ver con la habilidad de evaluar el peso de las clases diferencialmente.

### 2.4.3. Rendimiento de los modelos

#### 2.4.3.1. Overfitting y model tuning

A pesar de que muchos de los modelos de clasificación o de regresión son capaces de extraer relaciones complejas de los set de datos, cabe la posibilidad de que algunos de los patrones detectados y sobre los que el modelo aplica mucha atención, no sean reproducibles. En muchas ocasiones, es posible que cuando estamos entrenando un modelo, éste nos arroje unos resultados con un valor de precisión del modelo o *accuracy* muy elevados, sin embargo, en el momento que pasamos a evaluar el subset de datos de prueba o testing, de repente la precisión cae sustancialmente. Esto es debido a un efecto denominado *overfitting* o sobreentrenamiento del modelo, y que puede hacer que el modelo carezca de validez pese a los buenos resultados obtenidos en el entrenamiento del mismo. Hoy en día, cuando se va a realizar la construcción del modelo, se suele dividir en los dos subconjuntos que hemos comentado previamente: *training* y *testing*. Sin embargo, a su vez, se suelen realizar distintos conjuntos de *training* y distintos conjuntos de *test*, ya que suelen proporcionar mejores resultados al entrenar el modelo además de proporcionar mayor precisión.

Cuando un modelo se encuentra sobreentrenado o con *overfitting*, es porque éste ha sido capaz de aprender el ruido que posee cada una de las muestras de entrenamiento, por lo que, como se comentaba anteriormente, cuando se introduzcan nuevas muestras, por ejemplo del conjunto de *testing*, la precisión bajará bruscamente y no acertará muchas de ellas.

Es posible adoptar distintas soluciones para evitar o combatir el *overfitting*, algunas de las cuales ya se adoptan en el propio modelo, como ocurre con Random Forest, sin embargo, a pesar de ello conviene sintonizar los parámetros de ajuste del mismo, ya que el modelo por sí mismo o con los datos de los que se alimenta, es posible que no sea capaz de encontrar una fórmula de forma analítica que evite el *overfitting*.

En este trabajo, con motivo de ajustar los parámetros del modelo, fundamentalmente, nos basaremos en parámetros tales como el área encerrada bajo la curva ROC o la precisión (*accuracy*) del mismo. Adicionalmente, la técnica que se empleará para combatir el *overfitting* será la validación cruzada mediante múltiples sets de datos de *training* y *testing*, o Validación cruzada de Monte Carlo [117].

#### 2.4.3.2. Factores que pueden afectar al rendimiento del modelo

Aunque en las secciones anteriores nos hemos basado en problemas como poseer clases desbalanceadas para la predicción de un modelo o problemas relativos a *overfitting*, en determinadas ocasiones no existe relación entre el modelo y los factores que afectan a su rendimiento. Algunos de estos factores pueden estar relacionados con la forma de codificar las variables, por ejemplo, para obtener una característica que se haya obtenido derivada de otras o similar. El nombre que recibe esta metodología es: *ingeniería de las características*.

Otro tipo de factores pueden estar relacionados con el ruido o el error, que introduzcamos al modelo y que podemos encontrar de algunas de estas formas, según Kuhn:

- Ruido obtenido por el sistema empleado para obtener los valores de las características.
- Introducción de predictores no informativos, es decir, aquellos que no guardan relación con la respuesta. Algunos modelos son capaces de filtrar información irrelevante en este sentido.
- Ruido obtenido en la medición de la variable de respuesta. Es decir, pongamos el caso que un resultado categórico medido fuera mal etiquetado en los datos de entrenamiento el 10 % del tiempo, es poco probable que cualquier modelo pueda alcanzar verdaderamente más del 90 % de precisión. Este tipo de ruido viene dado por el *Error Cuadrático Medio* (RMSE) y la idea es eliminar este tipo de error durante el proceso de modelado.

Como se puede observar, el tratamiento que se debe realizar sobre las variables es muy importante a la hora de no arrastrar el error a posteriori en el modelo y que éste no nos lastre su rendimiento.

Por otro lado, es muy importante el desarrollo de un modelo que sea capaz de responder a lo que realmente deseamos con la variable respuesta, o por el contrario, se cometerá un nuevo error denominado '*Error Tipo III*' [71], donde en determinadas ocasiones, se suele centrar la solución en la parte técnica, dejando de lado la estrategia global relacionada con el problema en sí mismo.

## 2.5. Conclusiones y aportaciones

### Conclusiones

El propósito no es siempre conocer y comprender el por qué van a ocurrir los eventos. En algunas ocasiones, el propósito viene dado por la precisión con la que se adivinará el evento, y no por las circunstancias que lo produjeron.

Desafortunadamente, a medida que se tiende a mejorar la exactitud de la predicción se acaban empleando modelos que resultan difícil de interpretar. Puesto que en este trabajo de investigación se pretende aunar ambos propósitos, adivinar el evento y tratar de evitar las circunstancias que lo produjeron, se opta por usar dos técnicas de clasificación, Random Forest y BayesGLM, y complementarlas.

Evitar el sobreentrenamiento u *overfitting* de un modelo de clasificación es un factor muy importante a la hora de escoger los algoritmos para trabajar con sets de datos de accidentes de tráfico. Por ello, como se ha mencionado en la sección 2.4.2.3, se ha escogido Random Forest como uno de los algoritmos principales a la hora de

trabajar con la clasificación respecto a la severidad de accidentes de tráfico debido a que incorpora distintos mecanismos para evitar caer en el sobreentrenamiento u *overfitting*. A pesar de que se podría pensar que al usar un número elevado de árboles para componer el bosque sería sencillo caer en este error, Random Forest reduce los ratios de error al escoger árboles independientes y fuertes ya que son escogidos independientemente de los árboles escogidos anteriormente, por lo que Random Forest es robusto frente a respuestas ruidosas.

No es posible decir lo mismo de los algoritmos de regresión logística, como BayesGLM. A pesar de que BayesGLM ofrece una gran cantidad de información y es especialmente útil para modelizar probabilidades y el modelo resultante es fácil de interpretar, sin embargo, no presenta soluciones robustas a la hora de evitar el sobreentrenamiento. Por todo ello, se ha optado por escoger ambos algortimos, BayesGLM y Random Forest, sin llegar a ensamblarlos, con motivo de realizar una supervisión de los resultados y resulte además sencillo de interpretar el modelo, ya que se pretenden aportar distintas consideraciones respecto a la seguridad vial.

Se han estudiado las bases para la validación de los modelos estadísticos que se propondrán en esta tesis. Partiendo de las diferentes formas de validar dichos modelos estadísticos, se ha escogido la validación mediante AUC ROC por considerar que se adecúa en mayor medida a la clasificación/predicción entre clases.

### Aportaciones

En este capítulo se presentan las siguientes aportaciones:

- Se describe la teoría de los eventos raros y cómo aplicarla al estudio de accidentes de tráfico.
- Se ha estudiado la literatura referente a predictores que no aportan información ya que el hecho de usar predictores que no contengan información o esta sea redundante es posible que añadan incertidumbre en la predicción.
- Se definen y exponen distintas metodologías de selección de características.
- Se ha definido el concepto de sets de datos desbalanceados y se ha estudiado la problemática que podría ocasionar el uso de este tipo de datos.
- Se han estudiado diferentes técnicas aportadas en diferentes investigaciones por otros autores para evitar la problemática de este tipo de sets de datos:
    - estrategias para muestreo de datos
    - estrategias para generación de nuevos ejemplos
- En nuestro caso, las técnicas para lidiar con este problema se presentan en los capítulos 5 y 6, donde se aplican los modelos en casos prácticos.

- Se ha descrito el uso de dos técnicas de clasificación complementarias para clasificación de accidentes de tráfico, partiendo del objetivo principal de esta tesis doctoral de adivinar el evento y tratar de evitar las circunstancias que lo produjeron.
- Se presenta el algoritmo Random Forest como uno de los algoritmos principales a la hora de trabajar con la clasificación respecto a la severidad de accidentes de tráfico, debido a que incorpora distintos mecanismos para evitar caer en el sobreentrenamiento u *overfitting*.
- Adicionalmente, se plantea el uso mixto de Random Forest y BayesGLM como selector de características, gracias a la posibilidad que nos ofrecen para ello, como algoritmos de regresión y como algoritmos de clasificación.
- Se presenta mediante el Área Bajo la Curva ROC la validación de los modelos estadísticos empleados para la clasificación y predicción de accidentes de tráfico en base a la severidad del accidente o de los accidentados.

# 3 Revisión del Estado del Arte

## 3.1. Introducción

Conducir un vehículo a día de hoy representa una tarea compleja llevada a cabo en un entorno en constante cambio donde influyen numerosos factores que pueden contribuir a una colisión. Aparte de los factores de ingeniería y obras públicas, como la construcción y mantenimiento de vehículos y carreteras, un gran porcentaje de las lesiones en la carretera se atribuyen al error perceptivo humano o a errores de decisión.

Los accidentes de tráfico son la causa de un elevado número de pérdidas humanas en el mundo e implican un trauma psicológico y físico significativo en aquellos accidentes con heridos graves, o en los propios familiares que sufren día a día las secuelas o las pérdidas de sus familiares. Cada año se producen más de 40.000 heridos en accidentes de vehículos en los estados miembros de la Unión Europea según CARE [7]. Por ejemplo, las estadísticas generales sobre lesiones por accidentes de tráfico en Reino Unido arrojan datos elevados de más de 3.600 muertes (más 230.000 heridos) con datos desde 2001, es decir, un promedio de 60 muertes por millón de personas. Por ello, Reino Unido era considerado en 2001 uno de los países más seguros de Europa, con una tasa de mortalidad en accidentes de tráfico de 6 personas por cada 100.000, cuando la media de la Unión Europea alcanzaba cifras de 11 [53]. Sin embargo, a partir de 2005, Reino Unido logró una disminución de un 46 % [48] la tasa de muertes por accidente de carretera. En promedio, 5 muertes ocurren cada día en sus carreteras y 61 ciudadanos acaban gravemente heridos. Sin embargo, España presenta unas cifras, aunque más esperanzadoras, todavía insuficientes, con 3,6 fallecidos por 100.000 habitantes en 2015, por debajo de la tasa de mortalidad media de la UE registrada en 5,1 fallecidos. Aun con estos resultados, España presenta un mínimo histórico en el

número de víctimas por accidente desde 1960, donde el parque automovilístico era de un millón de vehículos frente a los más de 31 millones actuales.

Adicionalmente, según datos aportados desde la UE [4], la carga económica anual que conllevan estos accidentes se calcula en Europa entre 10.000 y 14.000 millones de euros, resultando el coste de una sola fatalidad entorno a 1 millón de euros. El conjunto de estos hechos ha motivado incluir la mejora de la seguridad vial como uno de los principales retos en el Horizonte 2020, que ha incrementado la inversión en investigación de tráfico.

Por otro lado, una prioridad adicional es la creación y el mantenimiento de plataformas de transparencia para proporcionar a los investigadores datos actualizados. En este sentido, es necesario reconocer el enorme esfuerzo realizado por el gobierno del Reino Unido en materia de transparencia. La gran cantidad de datos abiertos disponibles públicamente ha posibilitado el inicio del estudio presentado en este manuscrito. España, sin embargo, todavía se encuentra alejada de este nivel de transparencia, aunque en estos últimos años se ha de reconocer el esfuerzo a través de la puesta en marcha de ciertas acciones en materia de datos abiertos de tráfico, como ha sido la creación del portal estadístico de datos de la DGT, que hace pensar en una mejora en este campo.

## 3.2. Estadísticas públicas aportadas de los Gobiernos de Reino Unido y España

### 3.2.1. Estadísticas aportadas por Reino Unido en materia de tráfico

#### a. Introducción

En el caso de Reino Unido, la información utilizada para crear las estadísticas es recopilada por las fuerzas policiales, ya sea a través de agentes de las fuerzas de seguridad que acuden a la escena del siniestro, o por aquellos ciudadanos que informan del mismo en las comisarías a posteriori.

Es importante remarcar que al no existir obligación alguna de que las personas denuncien accidentes a la policía es posible que las cifras que maneja estadísticamente el gobierno, por lo tanto, no representen la gama completa de todos los accidentes o víctimas en Gran Bretaña.

Sin embargo, todos los accidentes que fueron reportados a la policía y ocurrieron en una vía pública involucrando al menos un vehículo de motor, jinete o ciclista, y donde al menos una persona resultó lesionada, están incluidos. Los accidentes ocurridos en terrenos privados o aparcamientos no se incluyen en las estadísticas.

Hay una serie de factores que probablemente han contribuido a los cambios en el número de personas muertas o heridas en accidentes de tráfico reportados. Hay evidencia que sugiere que las recesiones económicas han acelerado la disminución de

las muertes por tráfico. Los dos periodos de grandes caídas en las muertes por carretera desde 1979 (1990-94 y 2006-10) coincidieron con las recesiones de 1990-92 y 2008-09. El '*Informe Anual de Seguridad Vial 2015*' del *Foro Internacional del Transporte (ITF)* destaca la relación entre las condiciones económicas y el número de víctimas.

La ITF estima que dos tercios de la reducción de las muertes en los países miembros del IRTAD entre 2008 y 2010 fueron resultado del deterioro de las condiciones económicas en los países. También hay pruebas de que la velocidad media del tráfico en las zonas de libre flujo, así como la proporción de conductores que exceden el límite de velocidad, ha disminuido en la última década. Esto podría no sólo ayudar a los conductores a evitar accidentes por completo, sino que también podría reducir la gravedad y el número de víctimas cuando se producen. Las mejoras tecnológicas y en ingeniería relacionadas con los vehículos y las vías han desempeñado un papel similar tanto en evitar accidentes como en minimizar sus consecuencias.

Es probable que la mejora de la educación y la formación haya producido mejores y más seguros conductores, y es probable que las mejoras en la atención traumatológica (y en particular la creación de centros de traumatismos importantes en Inglaterra) hayan ayudado a mejorar los resultados una vez que se ha producido un accidente.

Sin embargo, Gran Bretaña ha reportado aumentos en las muertes por carretera durante el 2014, pero no es el único país. Como se destaca en el informe PIN del ETSC (*Consejo Europeo de Seguridad en el Transporte*) 2015, aunque las muertes en carretera en la Unión Europea disminuyeron un 0,6 % entre 2013 y 2014, se enmascaran variaciones significativas en toda Europa. Trece países, entre ellos el Reino Unido, Suecia, Alemania, Francia e Irlanda registraron aumentos en las muertes por carretera en 2014 y los Países Bajos se mantuvieron sin cambios desde 2013. Está claro, por tanto, que hay mucho por hacer en materia de seguridad.

### El efecto de la meteorología en el número de víctimas

Es consecuente pensar que el clima posee una influencia significativa en el número de víctimas en la carretera. Los periodos de mal tiempo ejercen una influencia sobre el comportamiento de los conductores y ejercen, por tanto, una influencia también en el número de víctimas [16]:

- En primer lugar, las malas condiciones climatológicas pueden limitar la visibilidad en las carreteras y hacer la superficie de la carretera más resbaladiza. Como consecuencia puede hacer que las condiciones de conducción sean más peligrosas, si los conductores no varían su comportamiento al volante. En ese caso podría aumentar el número de accidentes. El efecto que se observa más a menudo es que los conductores responden positivamente a condiciones adversas, por ejemplo, lluvias fuertes, riesgo de nieve y hielo, etc. En ese caso tienden a disminuir la velocidad y su forma de conducción es más cuidadosa, reduciendo así el riesgo de colisión y reduciendo la severidad de los accidentes en caso de ocurrir.

- En segundo lugar, el mal tiempo posee una fuerte influencia sobre la exposición a accidentes. En condiciones severas, se suelen posponer o cancelar viajes o cambiar a modos de transporte más seguros (por ejemplo, ciclistas y motociclistas que deciden emplear otro tipo de vehículos, o usuarios de carretera cambiando a medios de transporte como el tren en lugar del coche). Claro ejemplo de ello fueron las fuertes nevadas tanto al inicio como al final de 2010 en Reino Unido. Por otro lado, en condiciones climáticas menos severas, pero aún desagradables, como las condiciones de viento o humedad, algunos usuarios, especialmente los usuarios vulnerables, como ciclistas y motociclistas y peatones, tienden a cambiar a vehículos cerrados más seguros, como automóviles y autobuses.

Por lo tanto, cuando hay mal tiempo, siempre existe un equilibrio entre las condiciones de la carretera creando mayores riesgos en ésta y los usuarios tendiendo a reducir su exposición al tráfico, ya sea por no viajar o no realizar cambios de movilidad. Sin embargo, el mal tiempo generalmente acaba en una reducción de accidentes, ya que la reducción de la exposición y la reducción de las velocidades de conducción por lo general superan el aumento del riesgo inherente de malas condiciones de la carretera.

Aquellos períodos de inusual buen tiempo casi siempre tienen un efecto en el aumento de accidentes. Esto se debe a que los períodos de buen tiempo a menudo pueden estimular a realizar viajes, aumentando así la exposición a un accidente. Algunos de estos viajes se realizan en medios de transporte relativamente seguros, principalmente coches. Sin embargo, el buen tiempo tiene una influencia más significativa en el ciclismo y para los usuarios de motocicletas, ya que estos grupos son particularmente sensibles al clima. Las buenas condiciones estimulan más los viajes, especialmente en los meses de primavera y otoño, donde estos usuarios tendrán más probabilidades de decidir qué tipo de transporte utilizarán dependiendo de las condiciones imperantes.

Los últimos años han estado fuertemente influenciados por el clima. Por ejemplo, las nevadas generalizadas que ocurrieron en 2010 tuvieron un efecto de amortiguamiento tan importante en el número de víctimas que hubo aumentos en todas las severidades en 2011. Es muy probable que las cifras de 2010 se decrementaran considerablemente por la nieve y sin ella hubiera habido un aumento de víctimas ese año. Del mismo modo, 2012 fue el segundo año más húmedo registrado. Esto, una vez más, habría suprimido el número de víctimas, especialmente en los grupos vulnerables, como ciclistas o motociclistas.

Los patrones climáticos en 2014 podrían explicar parcialmente algunos de los aumentos en los accidentes. Las cifras de tráfico indican que actividades como el ciclismo y el uso de motocicletas aumentaron durante 2014 en comparación con 2013. A medida que aumenta el número de personas que viajan, es lógico esperar que se produzcan más accidentes. Y mientras que los meses de fuertes lluvias disuadieron algunos viajes, gran parte de la lluvia cayó durante el invierno y al finalizar otoño, cuando menos usuarios vulnerables se encontraban circulando por las carreteras. Es posible, por lo tanto, que la lluvia extra caída durante este período pueda haber conducido al fomento de realización de viajes más peligrosos en vehículos a motor, contribuyendo así al aumento.

### El volumen de tráfico

El volumen de tráfico por carretera es un factor clave que explica el número de accidentes y víctimas. Sin embargo, la relación es aún más compleja.

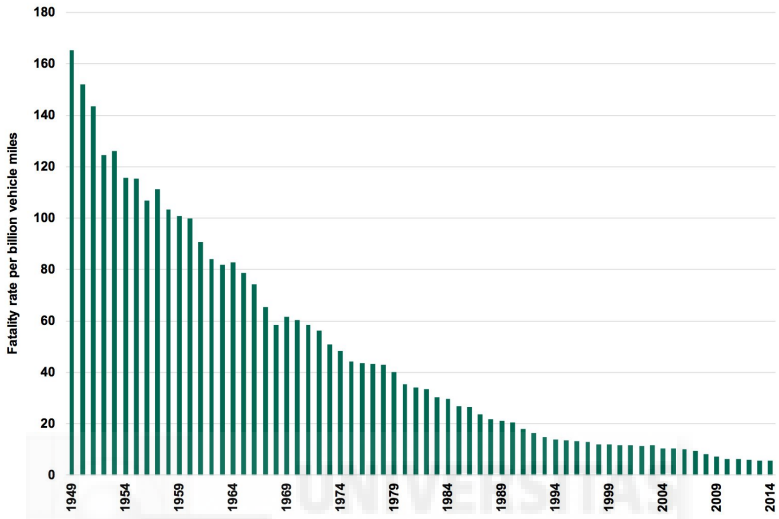



**Figura 3.1:** Fallecimientos en carreteras de Gran Bretaña por cada 1000 millones de millas recorridas (1949-2014). Fuente: Departamento para el Transporte de Reino Unido

En un primer momento, el aumento en el volumen de tráfico debería conducir a más accidentes y víctimas. Resulta evidente que cuanto mayor sea el número de vehículos en la red de carreteras, mayor será el número de interacciones entre vehículos y entre vehículos y peatones, y por tanto mayor será la probabilidad de que ocurran accidentes. En la práctica ésto no siempre sucede. Durante 28 años, tanto el número de muertes como el volumen de tráfico han aumentado, y en 22 de estos años, el tráfico ha crecido más rápido que las muertes. También durante 29 años, el tráfico rodado ha aumentado, pero las muertes han caído. Independientemente a estos años excepcionales, la tasa de mortalidad ha disminuido casi todos los años desde un máximo de 165 muertes por bvm (billion vehicle miles) en 1949 a 5,6 muertes por bvm en 2013 y 5,7 muertes por bvm en 2014.

Aunque el tráfico vial deba tener un efecto en el número de accidentes, no es una relación simple o directa. Sin embargo, a corto plazo, a falta de otros cambios, es probable que los cambios repentinos en los niveles de tráfico, como los períodos de crisis económica o los períodos de mal tiempo, se reflejen de alguna forma en el número de accidentes.

Antes de la recesión económica en 2007, el tráfico había aumentado en casi todos los años desde que comenzaron los registros. Desde 1949, el número de muertes cada año había caído un 0,4 por ciento en promedio, sin embargo, cada año los volúmenes de tráfico han aumentado. En los últimos 20 años, las muertes se han disminuido en un 1,1

por ciento anual en promedio en años de crecimiento del tráfico. En contraste con esto, los años en los que se ha reducido el tráfico han llevado también a una reducción media de la mortalidad del 7,7 por ciento desde 1949 y más del 10 por ciento desde 1994. Está claro, por lo tanto, que las reducciones del volumen de tráfico (probablemente en combinación con otros factores como conducción más lenta para ahorrar combustible) representen una mayor caída en muertes comparativamente con otros períodos.

El volumen de tráfico en 2014 fue de un 2,4 por ciento más alto que en 2013. Este es el mayor crecimiento en el tráfico total desde 1996. Por lo tanto, es probable que al menos parte del aumento de muertes y heridos se relacione con que se ha vuelto a una tendencia al alza en vehículos de motor, que probablemente ha sido impulsada por la mejora de la economía.

### b. Estadísticas por tipo de usuarios

Históricamente y aún en la actualidad, los ocupantes de automóviles constituyen el mayor grupo de usuarios con víctimas en accidentes. Esto se debe a que los automóviles representan casi el 80 por ciento de todo el tráfico conducido en Gran Bretaña. Sin embargo, el número de víctimas por grupo de usuarios no es proporcional a la distancia total que este grupo viaja.

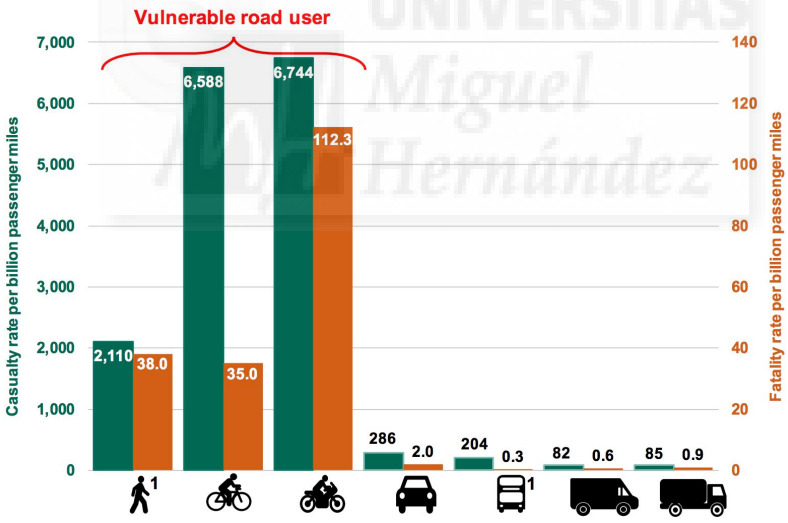



**Figura 3.2:** Tasa de accidentes y fallecimientos por mil millones de pasajeros y millas por tipo de usuario de la vía en 2014. (Fuente: Departamento para el Transporte de Reino Unido)

En cuanto a los accidentados de cualquier tipo de severidad, los ciclistas tienen una tasa similar a los motociclistas, con más de 6.500 accidentados por cada mil millones de millas y pasajeros. La tasa para los peatones es de 2.110 víctimas por mil millones de millas recorridas.

El número total de víctimas de ocupantes de automóviles también aumentó en un 5,2 por ciento a 115.530 personas en 2014. Al igual que ocurre con los heridos

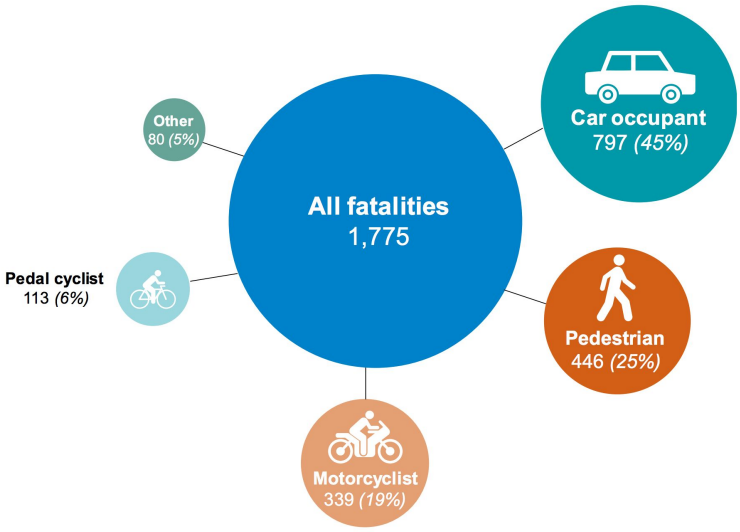



**Figura 3.3:** Muertes reportadas por accidentes de tráfico por tipo de usuarios en Gran Bretaña durante 2014. (Fuente: Departamento para el Transporte de Reino Unido)

graves, 2014 fue el segundo año con menor número de víctimas registrado y el cambio es lo suficientemente grande como para ser estadísticamente significativo.

### Usuarios de coche y taxi

El tráfico de automóviles y taxis en Gran Bretaña aumentó un 1,9 por ciento entre 2013 y 2014. El incremento en el tráfico de automóviles y taxis es una de las razones que pueden llevar a un aumento de accidentes.

### Usuarios de motocicleta

Respecto a los usuarios de motocicleta, hubo un total de 339 usuarios de este tipo de vehículos fallecidos en accidentes de tráfico reportados durante 2014. Aunque representa un aumento del 2,4 por ciento desde 2013, son sólo 8 muertes más, no siendo este cambio estadísticamente significativo. El número de muertes de motociclistas cayó cada año de casi 600 en 2006 a 328 en 2012.

En contraste con los fallecimientos, se ha producido un claro incremento en el número de usuarios de motocicletas que sufrieron lesiones graves. Hubo 5.289 heridos graves en 2014, lo que representa un aumento del 8,7 por ciento desde 2013. El número de lesiones graves se remonta a la media de 2005-09 y se encuentra en el nivel más alto desde 2009.

El número de accidentados en motocicletas con heridas leves también aumentó un 8,7 por ciento en 2014. Por lo tanto, el número total de accidentados en motocicletas

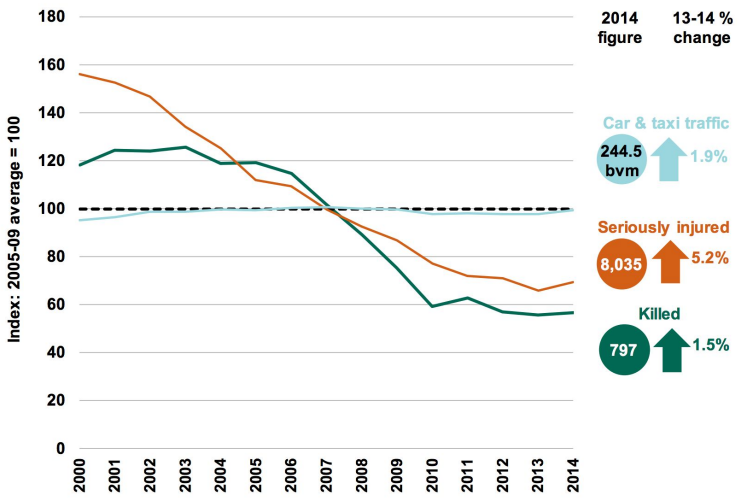



**Figura 3.4:** Número de ocupantes de coches y taxis fallecidos o heridos gravemente, Gran Bretaña: 2000-2014 (Las figuras de 2014 que se muestran en los círculos representan el cambio en 2013-2014 por 1.000 millones de millas recorridas). Fuente: Departamento para el Transporte de Reino Unido

de cualquier tipo de severidad en 2014 fue de 20.366, que es la cifra más alta desde 2009.

El tráfico de motocicletas aumentó en un 3 por ciento a partir de 2013. Esto equivale a 0,6 mil millones de vehículos menos que en 2007. El aumento de los heridos ha superado, por tanto, al aumento del tráfico, lo que indica que no es sólo un aumento de la exposición lo que está impulsando la subida en el número de accidentados.

### Niños (igual o menores de 15 años)

En los últimos cinco años el número de muertes infantiles ha fluctuado entre 48 y 61. Por ejemplo, en 2014 hubo 53 muertes de niños, 5 más que en 2013, lo que sugiere que la cifra ha caído a un nivel estable y los cambios son una función de la variación natural en lugar de ser una tendencia.

Como ha sucedido históricamente, las muertes infantiles ocurren principalmente en las categorías de peatones (29 muertos en 2014) y de ocupantes de vehículos (18 muertes), con un menor número en ciclistas (6 muertes). Esto se debe a que son las formas de transporte más utilizadas por los niños.

A pesar de haber crecido en un 5 %, el número de niños gravemente heridos en accidentes de tráfico en carretera sigue siendo el segundo más bajo de la historia, con 2.029 víctimas. Del mismo modo, el número total de víctimas infantiles de todas las severidades aumentó un 6,2 por ciento a 16.727, pero 2014 ocupa el segundo lugar después de 2013 en relación al total. Incluso después de los aumentos de 2013 a 2014,

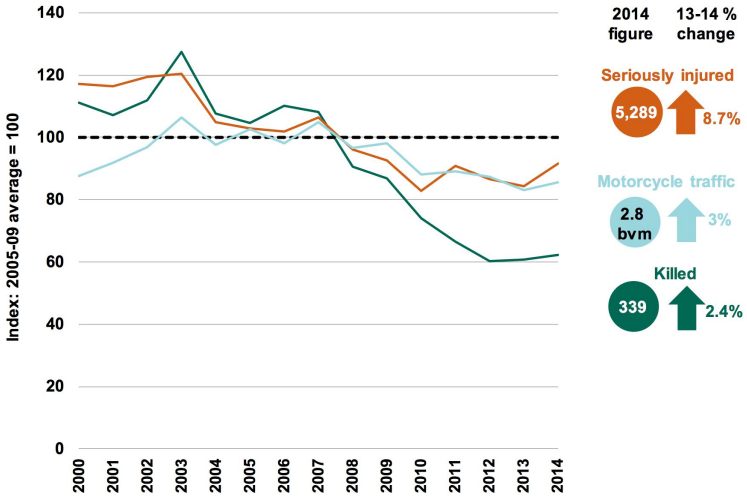



**Figura 3.5:** Número de ocupantes de motocicletas fallecidos o heridos gravemente, Gran Bretaña: 2000-2014 (Las figuras de 2014 que se muestran en los círculos representan el cambio en 2013-2014 por bvm - billion vehicle miles). Fuente: Departamento para el Transporte de Reino Unido

el número de niños muertos o heridos en accidentes de tráfico reportados es un 30 por ciento menor que el promedio de 2005-09, con muertes por un 58 por ciento en el mismo período.

### c. Estadísticas por tipo de vía

El número de personas fallecidas en las carreteras urbanas aumentó en un 9,1 por ciento a 783 muertes en 2014. Esto se relaciona con el aumento de las muertes de peatones (un 15,7 %), ya que la mayoría de las muertes y lesiones de peatones ocurren en carreteras urbanas. El resto de víctimas mortales aumentaron un 4,6 %. El número de heridos graves y heridos leves en vías urbanas aumentó un 4,2 % y un 7,2 %, respectivamente.

El número de muertes en carreteras no urbanizadas se mantuvo constante entre 2013 y 2014, con 896 muertes. Sin embargo, hubo aumentos de un 7,4 por ciento en heridos graves y 2,4 por ciento en heridos leves.

En autopistas, durante 2014 fallecieron 96 personas, 4 menos que en 2013. Este cambio es probable que sea un reflejo de la variación natural en las cifras. El número de heridos graves en las autopistas aumentó por segundo año consecutivo, con un 8,8 por ciento, a 718. Hubo un aumento del 5,3 por ciento en el número de heridos leves. El tráfico en autopistas aumentó en un 1,6 por ciento, en las carreteras generales 'A' en un 2,0 por ciento, en las carreteras urbanas 'A' en un 1,7 por ciento y en el resto de carreteras generales en un 5,5 por ciento.

La mayoría de los heridos ocurrieron en vías urbanizadas (72 por ciento del total de víctimas en 2014). Sin embargo, la mayoría de las muertes ocurrieron en carreteras

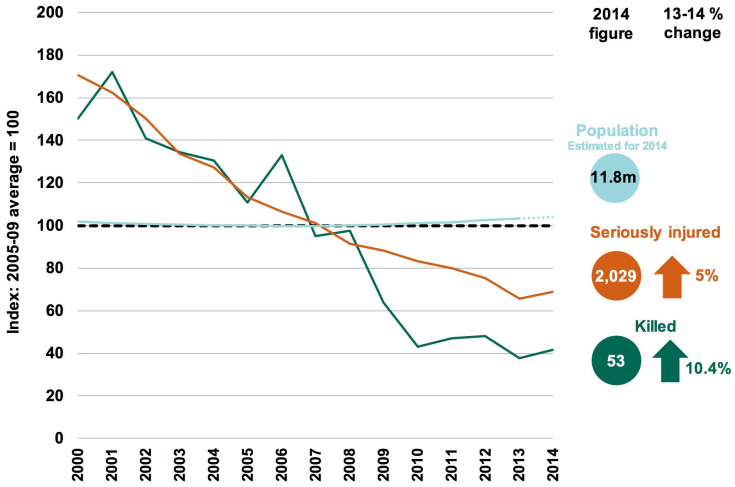



**Figura 3.6:** Número de niños fallecidos o gravemente heridos menores de 15 años, Gran Bretaña: 2000-2014 (Las figuras de 2014 que se muestran en los círculos representan el cambio en 2013-2014 por bvm - billion vehicle miles). Fuente: Departamento para el Transporte de Reino Unido

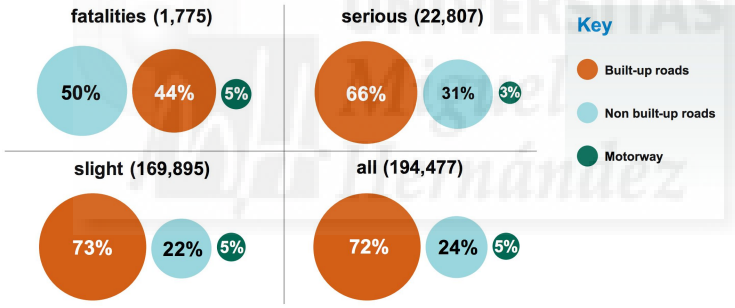



**Figura 3.7:** Número de accidentados por tipo de severidad, Gran Bretaña: 2014. Fuente: Departamento para el Transporte de Reino Unido

no urbanizadas (poco más de la mitad). Aunque las autopistas poseen alrededor del 21 por ciento del tráfico, sólo representan el 5,4 por ciento de las muertes y el 4,7 por ciento de los heridos.

### d. Resumen

Se ha escogido 2014 para analizar las estadísticas de los accidentes ocurridos en Reino Unido y poder así averiguar lo ocurrido de forma anual. Adicionalmente, se ha extraído una comparativa del Departamento para el transporte de Reino Unido, visible en las figuras 3.4, 3.5, 3.6, 3.8 y 3.9.

En el resumen anual, un total de 1.775 personas murieron en accidentes de tráfico en Gran Bretaña durante 2014. Se trata de un aumento de un 4 por ciento

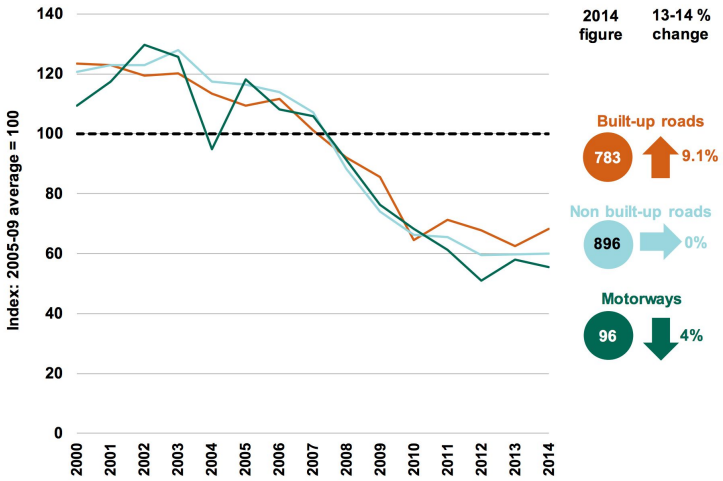



**Figura 3.8:** Número de muertes por tipo de vía, Gran Bretaña: 2000-2014. Fuente: Departamento para el Transporte de Reino Unido

desde 2013. Es el tercer año más bajo registrado después de 2012 y 2013. Hubo un 45 por ciento menos de fallecidos en 2014 que en la década anterior en 2005 y un 37 por ciento menos que el promedio entre 2005-09.

En 2014, se reportaron un total de 22.807 accidentes de tráfico con heridos graves. Esto representa un aumento del 5,3 por ciento desde 2013, aunque es inferior a los 23.039 heridos graves de 2012. Hubo un total de 194.477 accidentes de todas las severidades durante 2014 y ese año representa el segundo con registros más bajos, aunque es un 5,9 por ciento mayor que en 2013. Es el primer año donde aumentan en los accidentados totales desde 1997.

El número de heridas graves en los accidentes de tráfico notificados aumentó en un 5 por ciento a 22.807 en 2014, en comparación con 2013.

Con excepción de 2010 a 2011, que se vio afectada por el mal tiempo, 2014 es el primer aumento de muertes desde 2003. Aumenta por primera vez el número de heridos graves desde 1994.

### 3.2.2. Estadísticas de España

Se ha escogido un año al azar para realizar un análisis en cifras absolutas acerca de los principales indicadores asociados a accidentes de tráfico en las vías españolas.

Según la DGT [35], en 2015 se produjeron un total de 97.756 accidentes con víctimas, de las cuales 63.198 se produjeron en vías urbanas y 34.558 en vías interurbanas. Puesto que el estudio de esta Tesis Doctoral no incluye los accidentes urbanos, únicamente revisaremos datos estadísticos de accidentes producidos en vías interurbanas.

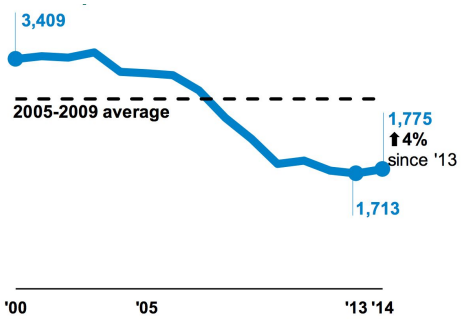



**Figura 3.9:** Reporte de víctimas en Gran Bretaña. Periodo 2000 a 2014. Fuente: Departamento para el Transporte de Reino Unido

#### 3.2.2.1. Resumen de accidentes en vías interurbanas

Como se mencionaba anteriormente, durante 2015 se produjeron un total de 34.558 accidentes de circulación en vías interurbanas, los cuales produjeron 54.028 víctimas.

**Fallecidos**

El número total de fallecidos en las vías interurbanas fue de 1.248 personas (2,3 %), de las cuales:

- el 70,8 % eran los conductores de los vehículos,
- el 19,5 % eran pasajeros y
- el 9,6 % eran peatones.

**Heridos graves**

Por otro lado, el número total de heridos graves fue de 4.744 (8,8 %) de los cuales:

- el 71,3 % eran los conductores de los vehículos,
- el 23,7 % eran pasajeros y
- el 5 % eran peatones.

**Heridos leves**

Por otro lado, el número total de heridos graves fue de 48.036 (89 %) de los cuales:

- el 64,3 % eran los conductores de los vehículos,

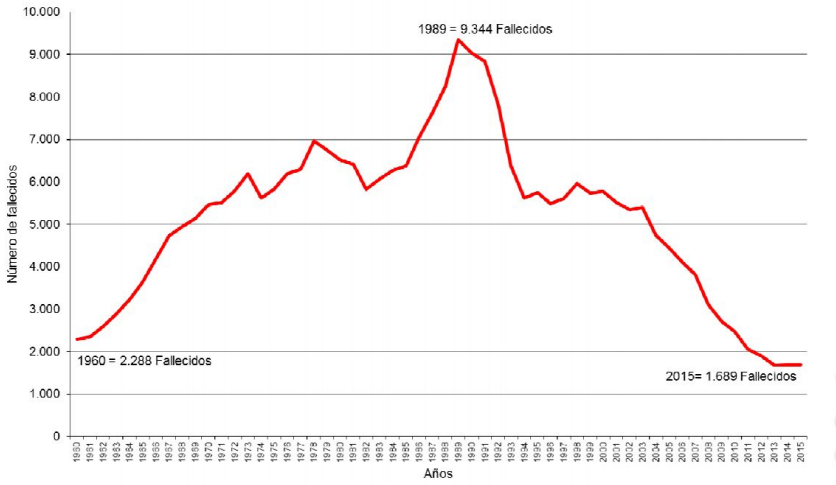



**Figura 3.10:** Reporte de fallecidos en España. Periodo 1960 a 2015. Fuente: DGT España

- el 34,6 % eran pasajeros y
- el 1,1 % eran peatones.

### Conclusiones

Aunque resulta compleja la interpretación de la figura 3.10 debido a que sería necesario considerar datos que apoyen las fases fácilmente visibles en ella, y la posible correlación entre esos datos, sí que es posible observar un aumento considerable en el número de fallecidos entre los años 1960 hasta 1989, aumento debido a la progresión del parque automovilístico, ello unido a la recuperación de la crisis que se produjo a partir de 1985, llevando hasta el año 1989 la necesidad de una revisión imperiosa de las políticas en seguridad vial, que se ven reflejadas en el brusco descenso a partir de este año en el número de fallecidos que se estabiliza entre el periodo 1993 a 2003, donde las políticas en seguridad se vuelven a revisar y coincide además con la crisis de 2007 hasta nuestros días. Actualmente, se puede observar que el número de fallecidos ha entrado en una fase valle donde, en nuestra opinión, resulta necesario tomar nuevas medidas para tratar de disminuir estas muertes.

La evolución de fallecidos, heridos hospitalizados y heridos no hospitalizados fue variando entre 1965 y 2015, donde en 1965 las proporciones eran de un 5 % fallecidos, un 26 % heridos hospitalizados y un 68 % heridos no hospitalizados y se mantienen prácticamente hasta 1998, año en el que esa proporción fue 4 % fallecidos, 24 % heridos hospitalizados y 72 % heridos no hospitalizados. En 2003 la proporción cambia reduciéndose la de fallecidos al 3 % y la de heridos hospitalizados al 17 % y vuelve a cambiar a partir de 2004 descendiendo la proporción de fallecidos y de heridos hospitalizados hasta 2015 en el que ha sido del 1,2 % y del 7,0 % respectivamente.

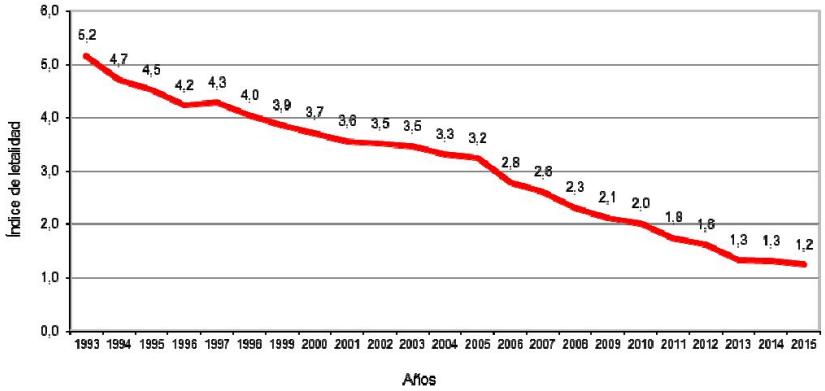



**Figura 3.11:** Evolución de la letalidad en los accidentes de tráfico con víctimas entre 1993 y 2015. Fuente: DGT España

### Evolución de la letalidad en los accidentes de tráfico con víctimas entre 1993 y 2015

La Dirección General de Tráfico, define la letalidad como la razón entre el número de fallecidos y el número de víctimas (ecuación 3.1).

$$Letalidad = \frac{Núm\_Fallecidos}{Núm\_Víctimas} \cdot 100 \tag{3.1}$$

La letalidad ha disminuido en España paulatinamente desde el año 1993, no siempre por la disminución de los fallecidos, sino también debido al aumento de los registros de heridos no hospitalizados.

En la figura 3.11 se puede observar la evolución de la letalidad desde 1993 hasta nuestros días.

En la figura 3.12 se puede observar que los accidentes con víctimas a partir del año 1998 han permanecido prácticamente constantes, sin embargo, en vías interurbanas, se puede ver que desde 1993 hasta 2007 los accidentes han ido en aumento, y sin embargo, desde 2007 hasta la actualidad han ido en descenso teniendo en cuenta que 2015 ha sido el segundo año desde 1993 con menor número de accidentes en este tipo de vías.

En la tabla 3.1, sin embargo, se observa con mayor detalle el desglose en número de fallecidos, heridos graves y heridos leves, donde se puede observar como disminuyen notablemente los fallecidos y heridos graves, sin embargo, el número de heridos leves aumenta casi al doble entre 1993 y 2007, aunque se genera un ligero descenso desde 2008 hasta 2015, probablemente por la crisis en la que nos vemos envueltos todavía, que provoca una reducción en el número de desplazamientos.

| Año | Fallecidos | Heridos graves | Heridos leves |
|---|---|---|---|
| 1993 | 5.236 | 22.276 | 38.249 |
| 1994 | 4.514 | 20.389 | 37.147 |
| 1995 | 4.713 | 21.959 | 40.092 |
| 1996 | 4.464 | 20.680 | 41.574 |
| 1997 | 4.472 | 21.282 | 40.219 |
| 1998 | 4.811 | 22.905 | 49.963 |
| 1999 | 4.709 | 21.309 | 52.849 |
| 2000 | 4.706 | 18.524 | 55.827 |
| 2001 | 4.543 | 18.468 | 56.689 |
| 2002 | 4.435 | 18.225 | 55.857 |
| 2003 | 4.480 | 19.006 | 60.466 |
| 2004 | 3.841 | 14.631 | 56.459 |
| 2005 | 3.652 | 14.920 | 53.869 |
| 2006 | 3.367 | 14.763 | 62.306 |
| 2007 | 3.082 | 13.201 | 63.587 |
| 2008 | 2.466 | 11.077 | 56.222 |
| 2009 | 2.130 | 8.748 | 54.180 |
| 2010 | 1.928 | 7.642 | 52.247 |
| 2011 | 1.603 | 6.825 | 47.692 |
| 2012 | 1.442 | 6.044 | 47.936 |
| 2013 | 1.230 | 5.182 | 51.320 |
| 2014 | 1.247 | 4.834 | 48.693 |
| 2015 | 1.248 | 4.744 | 48.036 |

**Tabla 3.1:** Número de fallecidos, heridos graves y heridos leves en accidentes de tráfico en vías interurbanas entre 1993 y 2015

## 3.3. Revisión del estado del arte relativo a factores y técnicas empleadas en diversos estudios

### 3.3.1. Factores o características estudiadas como causas en los accidentes de tráfico

Desde que los vehículos se convirtieron en un medio de transporte predominante se han realizado diferentes análisis de accidentes de tráfico. En general, las autoridades de tráfico consideran el triángulo conductor-carretera-vehículo en un intento de estudiar los accidentes de tráfico. Sin embargo, la mayoría de los investigadores asumen que existe un conjunto más grande de circunstancias subyacentes que están relacionadas entre sí y que pueden tratarse matemáticamente para modelar la ocurrencia y gravedad de un accidente.

Un paradigma común considera que los accidentes no se encuentran dispersos al azar a lo largo de la red de carreteras y que los conductores no están involucrados en accidentes también al azar y, por lo tanto, existe una relación entre la frecuencia y la gravedad del accidente con un conjunto de variables que lo causaron [74].

Hasta la fecha, se han presentado una gran diversidad de estudios de accidentes de tráfico que difieren sobre el tipo de características utilizadas en él, así como las técnicas de minería de datos para extraer los factores más importantes involucrados en el accidente. Por ejemplo, Sabey y Taylor [93] utilizaron la velocidad del automóvil en el momento del accidente, mientras que Hakim et al. [55] estudió la importancia

| AÑOS | TOTAL | Variación respecto al año anterior | Vías Interurbanas | Variación respecto al año anterior | Vías urbanas | Variación respecto al año anterior |
|---|---|---|---|---|---|---|
| 1993 | 79.925 | -7.368 | 35.814 | -3.307 | 44.111 | -4.061 |
| 1994 | 78.474 | -1.451 | 34.354 | -1.460 | 44.120 | 9 |
| 1995 | 83.586 | 5.112 | 37.217 | 2.863 | 46.369 | 2.249 |
| 1996 | 85.588 | 2.002 | 37.434 | 217 | 48.154 | 1.785 |
| 1997 | 86.067 | 479 | 36.551 | -883 | 49.516 | 1.362 |
| 1998 | 97.570 | 11.503 | 44.388 | 7.837 | 53.182 | 3.666 |
| 1999 | 97.811 | 241 | 44.784 | 396 | 53.027 | -155 |
| 2000 | 101.729 | 3.918 | 44.720 | -64 | 57.009 | 3.982 |
| 2001 | 100.393 | -1.336 | 45.483 | 763 | 54.910 | -2.099 |
| 2002 | 98.433 | -1.960 | 44.871 | -612 | 53.562 | -1.348 |
| 2003 | 99.987 | 1.554 | 47.567 | 2.696 | 52.420 | -1.142 |
| 2004 | 94.009 | -5.978 | 43.787 | -3.780 | 50.222 | -2.178 |
| 2005 | 91.187 | -2.822 | 42.624 | -1.163 | 48.563 | -1.659 |
| 2006 | 99.797 | 8610 | 49.221 | 6597 | 50.576 | 2.013 |
| 2007 | 100.508 | 711 | 49.820 | 599 | 50.688 | 112 |
| 2008 | 93.161 | -7.347 | 43.831 | -5.989 | 49.330 | -1.358 |
| 2009 | 88.251 | -4.910 | 40.789 | -3.042 | 47.462 | -1.868 |
| 2010 | 85.503 | -2.748 | 39.174 | -1.615 | 46.329 | -1.133 |
| 2011 | 83.027 | -2.476 | 35.878 | -3.296 | 47.149 | 820 |
| 2012 | 83.115 | 88 | 35.425 | -453 | 47.690 | 541 |
| 2013 | 89.519 | 6404 | 37.297 | 1872 | 52.222 | 4.532 |
| 2014 | 91.570 | 2.051 | 35.147 | -2.150 | 56.423 | 4.201 |
| 2015 | 97.756 | 6.186 | 34.558 | -589 | 63.198 | 6.775 |

**Figura 3.12:** Evolución de la letalidad en los accidentes de tráfico con víctimas entre 1993 y 2015 por tipos de vía. Fuente: DGT España

del límite de velocidad de la vía. El uso de medidas de seguridad fue una variable en la investigación de Simoncic [100]. La clasificación del conductor se utilizó en otros estudios como una de sus principales características como, por ejemplo, la edad, utilizada por Davison [30], o la experiencia, también estudiado por Simoncic en el estudio anteriormente citado. Otros factores relacionados con la infraestructura también se han incluido en los estudios, como el ancho de la vía [68], la sección de longitud de la misma o el número de carriles [111].

Como se observa en los párrafos anteriores, a lo largo de estos años, se ha analizado el accidente como suceso con el fin de establecer modelos que aporten respuestas a las causas que los produjeron y, por consiguiente, tratar de aportar soluciones que ayuden a prevenir este tipo de sucesos no deseados. Por todo ello, a la hora de seleccionar las características alrededor de dicho suceso, y que hacen que su frecuencia aumente, en la mayoría de los casos se han empleado variables relacionadas con la meteorología, características de la calzada por donde circula el vehículo, o condiciones del conductor a la hora de la conducción, como la conducción bajo los efectos del alcohol o las drogas.

Si se centra la atención en causas relacionadas con los conductores, las características mayormente empleadas en investigación, junto con su fuente, son las siguientes:

- velocidad [97]
- uso del cinturón de seguridad [100]

- edad de los conductores [30] [78]
- experiencia al volante [93] [100]

Aunque éstas suelen ser minoritarias frente a las relacionadas con la infraestructura o el diseño de la vía, como por ejemplo:

- número de carriles ([97], [82], [104], [22], [111], [28])
- anchura de cada carril ([118], [77], [54], [97], [104], [79], [116])
- anchura de los arcenes ([118], [82], [104], [79], [8], [99])
- IMD (Intensidad Diaria Media - veh/día) ([118], [79], [83], [8], [111])
- límites de velocidad de la vía ([55], [49], [97], [79], [40], [99], [111], [28])
- longitud del tramo ([97], [82], [104], [22], [40], [116], [44], [111], [28], [67], [55])

Por otro lado, numerosas investigaciones ([67], [97], [49], [100], [5]) han correlado efectos meteorológicos, como lluvia, niebla o nieve, como causa de la aparición de accidentes de tráfico, pero siempre enfocado desde un punto de vista de la movilidad del usuario.

Es, por tanto, importante identificar las situaciones meteorológicas que incrementan o reducen las salidas en vehículos, y por ello, las variables asociadas. La temperatura es una de ellas puesto que puede representar los cambios en el número de salidas en vehículo al cabo del año, ya que, evidentemente, el buen tiempo suele favorecer la movilidad, sin embargo, el frío la suele reducir.

Adicionalmente, se considera a la meteorología como un factor directamente proporcional con el número de víctimas o accidentes de tráfico, ya que puede ocasionar distintos efectos en según qué tipo de carreteras. Por ejemplo, en Francia durante la última década del siglo XX, cerca del 15 % de los accidentes con lesiones tuvieron lugar con tiempo lluvioso [12].

En otras investigaciones [16] se ha decidido focalizar en tres situaciones meteorológicas: la lluvia como factor de riesgo, el buen tiempo como factor de movilidad y la nieve como factor de riesgo y movilidad al mismo tiempo.

### 3.3.2. Técnicas empleadas para el estudio de accidentes de tráfico

#### Modelos de regresión logística

En relación con las técnicas de minería de datos empleadas, podemos destacar el uso de modelos de regresión logística ([77], [49], [94], [87]) para identificar factores de predicción de accidentes y daños producidos por los mismos. Adicionalmente, Mussone [86] analizó los accidentes de tráfico que ocurrieron en las intersecciones de la ciudad de Milán, Italia. Estos métodos también se aplicaron desde el punto de vista de clasificación para modelar la gravedad del accidente en Corea, y sus resultados muestran que estas técnicas proporcionan información importante para la prevención de accidentes. Por último, Sohn [101] propone la fusión de la información proporcionada por clasificadores individuales para obtener mejores resultados.

El propósito del estudio efectuado por Gomes [51] fue desarrollar modelos de predicción de accidentes para áreas urbanas situadas en Lisboa, Portugal, que describirían el número esperado de accidentes en función de una serie de variables explicativas, como por ejemplo, conteos de tráfico de vehículos y peatones y características de diseño geométrico de carreteras. Con el fin de lograr este objetivo, se desarrolló una base de datos de accidentes posicionados, que permitió una rápida extracción de la información pertinente sobre una base geográfica (sitios incluidos en el conjunto de datos). En lo que se refiere a la tarea de modelado, los coeficientes se estimaron utilizando los enfoques GLM con la distribución de Poisson-Gamma. Se desarrollaron modelos para diferentes tipos de desagregación: según el elemento de la carretera - en las intersecciones (cruces de 3 salidas, 4 salidas y rotondas) y segmentos; según el tipo de accidente - accidentes peatonales y otros accidentes con lesiones; y por tipo y número de variables explicativas relacionadas con el entorno vial - modelos simplificados (sólo con las variables de exposición) y globales (con todas las variables explicativas medidas).

Diversos estudios se han centrado en modelos de regresión binomial negativa para el estudio en tramos de vías [54], algunos basados en su diseño junto con variables meteorológicas [97], donde precisamente estos autores emplean herramientas basadas en modelos de regresión para la predicción de accidentes. Poch [89] analizó los efectos geométricos de diversas vías urbanas en EEUU, y continuando con el análisis de las vías, la investigación de Bauer [15] se centró en la relación entre los accidentes producidos en carriles de aceleración o rampas de intercambio.

Como se puede observar, el uso de Modelos de Regresión Logística se encuentra bastante extendido para el análisis de accidentes de tráfico, destacando muchos otros autores ([6], [104], [97], [40]).

#### Modelos de regresión de Poisson

Diversos estudios han empleado modelos de regresión de Poisson ([28], [68], [107], [104], [111], [67], [78], [59]) con el fin de estudiar las características más relevantes que definen un accidente de tráfico. Adicionalmente, Lee [79] además tiene en cuenta el triángulo empleado en este trabajo: personas, vehículos y vías de circulación.

### Clasificación basada en J48 y/o Naïve Bayes

Otro tipo de investigaciones [110] basadas en Naïve Bayes y árboles de decisión J48 desarrollaron modelos de predicción para clasificar de forma automática la severidad de accidentes de tráfico ocurridos en Hong Kong durante el año 2008. Los resultados de la investigación arrojaron mejores resultados de clasificación mediante el algoritmo J48. Adicionalmente al clasificador, previamente se usaron algoritmos Genéticos para la selección de características, con el fin de reducir la dimensionalidad del set de datos, y dónde únicamente se seleccionaron características relacionadas con la edad y el sexo del conductor, el tipo de vehículo, la severidad del accidente o el año de manufactura.

El estado de la India *Uttar Pradesh*, experimenta la tasa más alta de tales accidentes en este país. En el trabajo desarrollado por Kashyap [69], se aplican tecnologías de minería de datos para vincular las características almacenadas de la vía a la severidad del accidente. El objetivo de este trabajo fue detectar las causas de los accidentes y cómo reducirlas. El clasificador empleado fue Naïve Bayes consiguiendo una precisión (*accuracy*) del modelo de 89,4554 %. En contraste con otros trabajos previos de otros investigadores que se centraron en las características del conductor, Kashyap se centra en la contribución de varios factores relacionados con la vía, las condiciones meteorológicas o las anomalías de los vehículos. Otras investigaciones similares las realizan Mitra [83] y Shively[99].

### Clustering

Beshah [18], sin embargo, centra su investigación en la relación entre la severidad del accidente y determinados factores de la vía mediante el uso del algoritmo PART. En la investigación se obtienen buenos resultados en el área encerrada bajo la curva ROC por el clasificador mediante el algoritmo del K-vecino más cercano. Las características escogidas fueron: Subcity, ParticularArea, RoadSeparation (indica cómo están separados los segmentos de carretera), RoadOrientation, RoadJunction, RoadSurfaceType (asfalto o tierra), RoadSurfCondition (seca, húmeda, etc.), WeatherCondition, LightCondition y AccidentSeverity.

Depaire [37] examina si el análisis a través de cluster podría ser utilizado como una técnica de clasificación de accidentes de tráfico. Se seleccionaron clusters de clase latente como el análisis de cluster aplicado y se lograron encontrar varios clusters en un set de datos heterogéneo de accidentes de tráfico. Se indica, adicionalmente, que los tipos de accidentes de tráfico, identificados por siete clusters, tienen sentido y agregan valor a los análisis de lesiones posteriores. El análisis mediante cluster utiliza el tipo de vehículo como base para la clasificación. Sin embargo, también encuentra variables menos triviales para segmentar los datos, como el tipo de carretera y la edad.

### Redes Bayesianas

Ehsaei [43] demuestra que la probabilidad de que ocurra un accidente mortal o grave en condiciones específicas puede ser reproducida usando Redes Bayesianas. Adicionalmente, demostró que es posible calcular la probabilidad de un accidente en

diferentes puntos de la red de carreteras (como los tipos de cruce) para definir un punto negro para accidentes fatales y serios a través de variables como *junction_detail* o *special_condition*, que veremos más adelante en este documento. Por otro lado, aporta otras conclusiones, relativas a la alta probabilidad de que un accidente ocurra durante el día que durante la noche. Otro hallazgo del estudio es que la posibilidad de accidentes mortales cuando no existe iluminación en la vía, es el doble que la posibilidad de accidentes con lesiones graves. También se encontró que la probabilidad de accidentes graves y mortales ocurridos en superficies húmedas había disminuido durante los últimos cinco años.

Por otro lado, De Oña [33] muestra la posibilidad del uso de este tipo de modelos basados en Redes Bayesianas para clasificar los accidentes en base a su gravedad. La ventaja que aportan en este estudio dichos modelos es la posibilidad de realizar predicciones sin la necesidad de presunciones previas. Adicionalmente, es posible utilizarlas para hacer representaciones gráficas de sistemas complejos con componentes interrelacionados. En este estudio, el autor presenta un análisis de 1.536 accidentes ocurridos en carreteras españolas, donde se utilizaron 18 variables que representan factores contribuyentes para construir tres Redes Bayesianas diferentes, que clasificaron la gravedad de los accidentes para las clases de lesiones leves y muertos o gravemente heridos. En el estudio se concluyó que aquellas variables que mejor identifican los factores asociados con un accidente donde hayan fallecidos o heridos graves son: tipo de accidente, edad del conductor, iluminación y número de lesiones.

Deublein [38] en 2013 presenta una nueva metodología destinada a la predicción de la ocurrencia de accidentes de tráfico. La metodología utiliza una combinación de tres métodos estadísticos:

- actualización *gamma* de las tasas de ocurrencia de lesiones en accidentes y usuarios.
- análisis de regresión multivariable *Poisson-lognormal* teniendo en cuenta las correlaciones entre variables de respuesta del modelo y los efectos de los datos de recuento de los accidentes, por ejemplo, sobredispersión.
- algoritmos de inferencia bayesiana, que se aplican mediante técnicas de minería de datos apoyadas por redes probabilísticas bayesianas para representar la no-linealidad entre las variables indicadoras de riesgo y de respuesta del modelo, así como los diferentes tipos de incertidumbres que podrían estar presentes en el desarrollo de los modelos específicos.

Simoncic [100] extrajo unos años antes un modelo basado en Redes Bayesianas para el estudio relacionado con accidentes únicamente entre dos vehículos.

### Estudios comparativos entre modelos

Tres años más tarde, Khera [70] evalúa un set de características con el fin de averiguar el grado de severidad del accidente de tráfico. En la tabla 3.2 extraída de su contribución se estima el modelo estadístico por orden de precisión de menor a mayor.

| Técnica | Accuracy |
|---|---|
| Árboles de regresión y clasificación | 72,49 |
| Random Forest | 73,00 |
| Árbol ID3 | 77,70 |
| Árbol Funcional | 70,27 |
| Naïve Bayes | 84,66 |
| Part | 84,64 |
| J48 | 85,18 |

**Tabla 3.2:** Resultados de *accuracy* obtenidos por Khera et al. mediante el uso de diferentes técnicas.

### Modelos estadísticos multivariable

Elvik [46] discute en 2011 la aplicación de criterios operacionales de causalidad a modelos estadísticos multivariantes desarrollados para identificar fuentes de variación sistemática en los recuentos de accidentes, en particular los efectos de las variables que representan los tratamientos de seguridad.

### Naïve-Poisson

Soler [103] en 2013 trata de aportar un modelo para el estudio de accidentes de carretera llamado Naïve-Poisson. Presenta la implementación de un sistema basado en modelos de Redes Bayesianas para la estimación de las frecuencias en que suceden los accidentes, aplicables en cualquier situación y para cualquier conjunto de datos disponible.

### Redes neuronales

Dougherty [41], Mussone [86] en la década de los 90 y Xie [116] en 2007 obtienen modelos de accidentes entre vehículos a través de distintos tipos de Redes Neuronales.

### Árboles de regresión y clasificación

Mohaymany [84] basó su investigación mediante árboles de regresión y clasificación obteniendo un modelo de estudio de la severidad del accidente con un *accuracy* de 72,49 %, empleando características muy similares a las empleadas en esta Tesis, aunque introduce varios predictores relacionados con las características de la vía, como superficie de la vía, ocurrencia, tipo de arcén, ancho de arcén. Adicionalmente, tiene en cuenta otro tipo de variables como sexo, edad, cinturón de seguridad, causa del accidente, tipo de colisión, tipo de vehículo, tipo de localización, luminosidad y condiciones meteorológicas.

Una investigación similar realizada por Beshah [17], pero con distintos predictores, genera un modelo basado en árboles de regresión y clasificación para predecir la severidad de un accidente de tráfico. El resultado obtenido para *accuracy* del modelo fue de 87,47 % y las características empleadas fueron: ID_Accidente, edad del conductor, antigüedad del carné de conducir, antigüedad del vehículo, tipo de vehículo, tipo

de superficie de la vía, condiciones de la vía, condiciones meteorológicas, luminosidad, tipo de accidente y causa del accidente.

#### Random Forest

Wu [115] en 2009 estudia la severidad para accidentes de un solo vehículo. A través de un modelo basado en Random Forest obtuvo un *accuracy* de 73 %, y los predictores empleados fueron: condiciones meteorológicas, límite de velocidad de la vía, luminosidad, factores de colisión, sexo, edad, antigüedad del carné de conducir, cinturón de seguridad, tipo de vehículo.

#### C4.5, ID3, CRT, Naïve Bayes y Random Forest

Shanthi [98] tres años más tarde estudió las características más significativas en la clasificación de accidentes de tráfico debido a su gravedad. El resultado óptimo para el *accuracy* fue de un 99,73 % Las características empleadas para el estudio fueron: provincia/estado, país, mes, fecha, hora, día, tipo de accidente, tipo de ocupante, posición del ocupante en el vehículo, edad, sexo, severidad del accidente, airbag, ocupante expulsado del vehículo, trayectoria de la expulsión, año de fallecimiento, mes de fallecimiento, test de alcohol, test de drogas, antecedentes drogas y geolocalización del accidente.

## 3.4. Convenciones de términos asociadas a los accidentes de tránsito

Se cree necesario incluir una serie de términos que despejen cualquier duda que pueda surgir en relación con los tipos de vehículos, personas que intervienen en un accidente, meteorología e infraestructura, entre otros, que son necesarios para comprender exactamente su relación con el accidente, así como la exactitud del dato que representan.

- **Accidente**: se refiere al que produce lesiones personales ocurridas en la vía pública (incluyendo las aceras) en las que al menos un vehículo de carretera está involucrado. Se informa a la policía dentro de los 30 días después de su ocurrencia. Un accidente puede dar lugar a varias bajas.

- **Vehículos implicados en accidentes**: Vehículos cuyos conductores o pasajeros resulten heridos, que golpeen y dañen a un peatón u otro vehículo cuyo conductor o pasajeros se lesione, o que contribuya al accidente. No se incluyen los vehículos que chocan después del accidente inicial que causó lesión, a menos que agraven el grado de lesión o causen más bajas. Incluye ciclos de pedales montados en la pista.

### La severidad o gravedad de los accidentes

En base a las notas, definiciones y convenciones que los Gobiernos de Reino Unido y España realizan [96], la severidad o la gravedad de las víctimas en un accidente de tráfico en estos países se clasifican en tres niveles:

- **Accidentes fatales (Fatal):** La definición internacional habitual de un accidente con una consecuencia fatal, tal como fue adoptada por la Convención de Vienna [42], es: ' *[...] Una víctima humana que muere dentro de los 30 días posteriores a la colisión por lesiones recibidas en el accidente [...]*'.
- **Accidentado grave (Serious injury):** La definición de accidentado grave es menos clara y puede variar con el tiempo y en función del lugar. Por ejemplo, la definición en Reino Unido se refiere a la hospitalización de una persona a causa de las lesiones que le ha producido el accidente: fracturas, contusiones, lesiones internas, quemaduras (excluyendo quemaduras por fricción), cortes severos, shock general severo que requiera tratamiento médico, incluso aunque no dé lugar a hospitalización.
- **Accidentado Leve (Slight injury):** Incluye todo tipo de lesiones leves, como esguinces (incluyendo lesión de latigazo cervical), moratones o cortes, etc. También recoge el shock ligero producido por el accidente que requiera asistencia en carretera.

### Definiciones relativas a vehículos

- Vehículos: Los vehículos (excepto los taxis) se clasifican según su tipo estructural y no según su empleo o categoría de licencia en el momento de un accidente.
- Autobuses y autocares: Autobuses o autocares con aforo para transportar 17 pasajeros o más, independientemente del uso. Se incluyen los autobuses de trabajo.
- Coche: Incluye taxis, triciclos de personas inválidas, coches de tres y cuatro ruedas y minibuses.
- Furgonetas: incluyen principalmente los vehículos del tipo furgoneta construidos a partir de un chasis de coche. Estos se definen como aquellos vehículos que no superen el peso bruto máximo autorizado del vehículo de 3,5 toneladas.
- Motocicletas: Ciclomotores, scooters y vehículos de motor de dos ruedas (incluidas las combinaciones de motocicletas).
- Vehículos de mercancías: Se dividen en dos grupos según el peso del vehículo. Incluyen camiones cisterna, tractores sin sus semirremolques, remolques, vehículos articulados y pick-ups.
- Vehículos pesados de mercancías (HGV): vehículos de mercancías de más de 3,5 toneladas de peso bruto máximo autorizado del vehículo (gvw).

- Vehículos de transporte de mercancías ligeros: vehículos de mercancías, principalmente furgonetas (incluidas las furgonetas de automóviles), no más de 3,5 toneladas de peso bruto máximo autorizado del vehículo.
- Vehículos agrícolas: Comprende principalmente tractores agrícolas (con o sin remolque), pero también incluye excavadoras móviles y dumpers delanteros.
- Scooter de movilidad: Una silla de ruedas accionada o una vespa con un peso máximo sin carga de 150kg y una velocidad máxima de 8mph.
- Otros vehículos: Otros vehículos incluyen ambulancias, camiones de bomberos, tranvías, vehículos de basura, rodillos de carretera, excavadoras, grúas móviles, etc., excepto donde se indique lo contrario. También se incluyen los vehículos no motorizados, incluidos los dibujados por un animal, caballo montado, sillas de ruedas sin motor, callejones, etc, salvo que se indique lo contrario.
- Ciclos de pedales: Incluye tándems, triciclos y ciclos de juguete montados en la calzada. A partir de 1983 la definición incluye un pequeño número de ciclos y triciclos con asistencia de batería con una velocidad máxima de 15mph.
- Taxi: Cualquier vehículo que funcione como un carro de caballos, independientemente de la construcción, y que lleve el consejo de distrito apropiado o las placas del carretón de la autoridad local. También incluye coches privados de alquiler.

### Definiciones relativas a Personas

- Víctima: Una persona fallecida o herida en un accidente. Las víctimas se subdividen en muertos, lesionados gravemente y ligeramente heridos.
- Niños: Personas menores de 16 años de edad.
- Conductores: Personas que conducen vehículos distintos de los de pedales, motocicletas y animales montados. El resto de ocupantes de los vehículos se denominan pasajeros.
- Peatón: Incluye a las personas que montan bicicletas de juguete en las aceras, las personas que empujan las bicicletas, aquellas que empujan o tiran de otros vehículos o manejan vehículos controlados por peatones, los que llevan o pastorean animales, los ocupantes de cochecitos de niños o sillas de ruedas y las personas que salen con seguridad de los vehículos y resultan heridas.
- Pasajeros: Ocupantes de vehículos, distintos de la persona en control (el conductor o jinete).
- Ciclistas: Pilotos de ciclos de pedales, incluidos los pasajeros.
- Usuarios de carretera: peatones y jinetes, conductores y pasajeros.
- Usuarios de un vehículo: Todos los ocupantes, es decir, el conductor (o jinete) y los pasajeros, incluyendo las personas heridas al subir o bajar del vehículo.

#### Definiciones relativas a meteorología o iluminación de la vía

- Oscuridad: Pasados 30 minutos desde la puesta de sol hasta 30 minutos antes de amanecer.
- Luz de día: Todos los tiempos distintos de la oscuridad.
- Accidente fatal: Un accidente en el que al menos una persona fallece.
- Accidente con víctimas: Un accidente que implica lesiones humanas o la muerte.
- Fallecimientos: víctimas humanas que sufrieron lesiones que causaron la muerte menos de 30 días después del accidente. Se excluyen los suicidios confirmados.

#### Definiciones relativas a infraestructuras

A continuación, se detallarán algunos términos relativos a despejar dudas de las infraestructuras de Reino Unido.

- Autopistas: Autopista A y carreteras A (M).
- Otras carreteras: Todas las carreteras B, C y no clasificadas, a menos que se indique lo contrario.
- Caminos rurales: caminos principales y caminos secundarios fuera de áreas urbanas y tener una población de menos de 10 mil.
- Carreteras urbanas: Carreteras mayores y menores dentro de un área urbana con una población de 10 mil o más habitantes. La definición se basa en aquella de asentamientos urbanos de 1991 de la Oficina del Viceprimer Ministro. Las áreas urbanas utilizadas para tablas en los boletines a partir de 2013 se basan en los datos del censo de 2011. Los boletines anteriores se basan en datos del censo de 2001.

## 3.5. Software de análisis estadístico empleado en esta investigación

Durante esta Tesis Doctoral se ha empleado distinto software libre para el análisis de los sets de datos así como para la construcción y el entrenamiento de los distintos modelos testados y utilizados.

El software mayoritariamente empleado ha sido *R* [90]. R es un lenguaje y conjunto de módulos estadísticos que, mediante cualquiera de los interfaces de que dispone, permite realizar análisis de datos y representación de los mismos.

Adicionalmente, se ha empleado el software *Weka* [114]. Weka es un software que incorpora diversas técnicas de *Machine Learning*. Fue creado por profesores de la Universidad de Waikato y su uso, además de para investigación en Machine Learning, es ampliamente utilizado para la docencia.

## 3.6. Conclusiones

En este capítulo se detallan estadísticas globales relativas a accidentes de tráfico con el objetivo de definir el problema de su ocurrencia en Reino Unido y España. Se indican datos macroscópicos, distinguiendo tipos de accidente en función de las vías, etc. y se han interpretado las gráficas aportadas por el Dpto. para Transporte de Reino Unido. Adicionalmente, se detallan estadísticas globales relativas a accidentes de tráfico ocurridos en España, centrando los datos de forma macroscópica y, asimismo, focalizando sobre su ocurrencia en un año escogido aleatoriamente. Se han interpretado las gráficas aportadas por la Dirección General de Tráfico en algunos de sus informes anuales.

A continuación, se realiza una revisión exhaustiva sobre el estado del arte de las características y técnicas empleadas para la clasificación y predicción de accidentes. En concreto se revisan trabajos realizados por otros autores empleando modelos de regresión logística, modelos de regresión de Poisson y clasificadores basados en árboles J48 y modelos bayesianos. Por último, se hace referencia a diversos trabajos realizados por otros autores mediante el uso del algoritmo Random Forest en sus investigaciones.

# 4 Implementación de arquitectura Big Data para captura de datos de tráfico

## 4.1. Introducción

En este capítulo se presenta una arquitectura diseñada para el almacenamiento de datos de tráfico. Para ello se han utilizado métodos y tecnología relacionada con el '*Internet de las cosas*' (IoT) [13] y Big Data para almacenar y procesar en tiempo real grandes cantidades de datos de sensores en sistemas distribuidos.

La idea con la que nace el sistema es capturar los eventos de tráfico que publica la Dirección General de Tráfico. Además, se han diseñado e implementado otros servicios típicos de arquitecturas *PaaS* (Platform as a Service) para visualizar los datos, descargarlos o analizarlos.

En el núcleo principal sobre el que está basado el sistema se ha utilizado una base de datos no relacional: *Apache Cassandra* [61].

Es el momento de justificar que los datos almacenados (datos relativos a eventos y accidentes de tráfico) se pueden considerar como Big Data. En efecto, únicamente la publicación de los eventos de tráfico de España con una periodicidad de unos 5 minutos puede considerarse una cantidad de información asumible por bases de datos, técnicas y métodos SQL. Sin embargo, si, además, para el estudio de tráfico añadimos los sensores aforados para la medida del tráfico en carreteras y las estaciones metereológicas existentes la cantidad de información a almacenar crece de forma ingente. En concreto, las:

- Estaciones aforadas: Están basadas generalmente en una espira de corriente soterrada en el asfalto. Al pasar un vehículo sobre la espira se produce un cambio

en la reluctancia del medio que es posible detectar mediante un sensor. La DGT publica datos de estos sensores soterrados, como:

  - El número de vehículos por hora que pasan sobre el sensor.
  - La ratio de vehículos ligeros/pesados.
  - La utilización de la vía en términos de vehículos por hora sobre la capacidad máxima de la vía.

- Estaciones metereológicas. Estas estaciones metereológicas están ubicadas en puntos estratégicos de las vías y dan datos como:

  - Temperatura.
  - Temperatura de rocío.
  - Estado de la superficie: hielo, nieve, superficie seca o mojada.
  - Velocidad y dirección del viento.

## 4.2. Motivación

Durante el comienzo de la investigación realizada en esta Tesis se planteó el estudio de los accidentes de tráfico en España, teniendo como objetivo obtener un modelo de ellos en base a un conjunto de predictores. Así pues, en ese momento comenzó la búsqueda de fuentes de datos que nos pudieran proporcionar datos de entrada para la realización del estudio. Pronto resultó evidente que sería necesario que nosotros mismos capturásemos esa información para almacenarla y después analizarla. Así pues, en febrero de 2014 comenzó a funcionar el sistema que se describe en este capítulo, capturando la información procedente de http://infocar.dgt.es/etraffic/.

Posteriormente, en julio de 2015, atendimos expectantes a la inauguración del portal estadístico de la DGT. La iniciativa hizo que nos planteáramos la necesidad del sistema de adquisición de datos diseñado por nosotros. Sin embargo, se ha decidido mantener el cluster en funcionamiento por las siguientes razones:

- Localización geográfica exacta de las incidencias de tráfico: En el portal estadístico de la DGT no se muestra la geolocalización exacta de los accidentes de tráfico, y cuanto menos, la del resto de incidencias.
- Periodo de los datos: Hemos hallado en dicho portal de estadísticas datos de accidentes en el periodo de 2011 a 2015. En el momento de lectura de esta Tesis, septiembre de 2017, todavía no se habían actualizado el periodo con datos más recientes.
- Disponibilidad de datos: En el portal de estadísticas nos resultó, a veces, difícil correlar datos de metereología y de afluencia de tráfico con cada accidente. En consecuencia, se decidió mantener nuestro servicio para poder ejecutar las consultas contra nuestro cluster de datos.

## 4.3. Objetivos

Se presentan, a continuación, los objetivos que se marcaron para el diseño del sistema de recopilación de datos que se presenta en este capítulo.

El objetivo principal del sistema consiste en disponer de una plataforma para poder almacenar información relativa a eventos de tráfico, estaciones aforadas y sensores metereológicos. En el caso de los eventos de tráfico nos referimos a:

- Vehículos averiados o volcados en la calzada.
- Retenciones de tráfico.
- Eventos de carácter lúdico/deportivo.
- Restricciones a la circulación de mercancías o vehículos pesados.
- Meteorología adversa.
- Desprendimientos u obstáculos en la calzada.
- Radares: se publican también la posición y ubicación de radares fijos en las vías españolas.
- Obras: se presenta también la ubicación de las obras en las carreteras.
- Cámaras y paneles de tráfico.

Además, para el estudio de los accidentes de tráfico se consideró necesario también disponer de datos de climatología y datos relativos a la carga de tráfico de la carretera en la que ocurrió cada accidente. En consecuencia, con un periodo de 5 minutos se almacena información relativa a unas 1100 estaciones meteorológicas y, aproximadamente, 600 estaciones aforadas. En cada estación meteorológica encontramos:

- Humedad relativa ( %).
- Visibilidad (m).
- Precipitaciones (mm/h). Naturaleza de precipitaciones (nieve, lluvia, aguanieve).
- Velocidad del viento (Km/h). Naturaleza del viento (normal, racheado)
- Dirección el viento (grados norte).
- Presión atmosférica (hPa)
- Tiempo presente: Seco, lluvioso, húmedo, nieve...

- Estado de la superficie de la calzada (nevada, escarchada, mojada, helada, húmeda, seca).
- Radiación atmosférica, global o terrestre ($W/m^2$).
- Salinidad ( %).
- Temperatura del aire (°C).
- Temperatura de congelación (°C).
- Temperatura de rocío (°C).
- Temperatura del subsuelo (°C).
- Temperatura de la superficie (°C).

No todas las estaciones meteorológicas disponen de la misma información. Por otra parte, hay algunos datos que no se han considerado relevantes para el estudio, mientras que otras características atmosféricas (p.e. la visibilidad) se consideró que merecían ser estudiadas para averiguar si existían relaciones directas con los accidentes.

Por otra parte, las estaciones aforadas publican información como la que sigue:

- Intensidad de tráfico (vehículos/hora).
- Velocidad media (km/h).
- Ocupación ( %).
- Porcentaje de vehículos ligeros/pesados ( %).

Finalmente, además del cluster basado en Cassandra para el almacenamiento de toda esta información, se programaron servicios web para la visualización a tiempo real de la información que existía en el sistema y la visualización georeferenciada de históricos de accidentes globales, por vía y por provincia. También se incluyeron servicios automáticos de backup y de descarga de datos para poder ser procesados posteriormente con R.

## 4.4. Estado del arte

En este apartado se describen conceptos que se encuentran en relación con la arquitectura que se ha realizado. Se comienza por definir el concepto de Internet de las Cosas y se plantean los conceptos de 'Big Data'. Finalmente se analizan bases de datos NoSQL, comparando distintas tecnologías existentes.

### 4.4.1. Internet de las cosas

El Internet de las cosas o IoT, por sus siglas en ingles, plantea la conexión de multitud de elementos y dispositivos cotidianos a Internet con el objetivo de compartir, almacenar y, posteriormente, procesar la información. El objetivo principal es procesar la información almacenada para poder extraer información de utilidad. La idea fue propuesta por Kevin Ashton, cofundador y director ejecutivo del Auto-ID Center del MIT, como título de una presentación que dio en Procter & Gamble (P&G) en 1999 [13].

En la actualidad se ha aumentado notablemente la capacidad de conexión de cualquier dispositivo con Internet. Las interfaces bluetooth, Wifi y Ethernet se han simplificado, con lo que se ha incrementado de forma exponencial el número de sensores que se encuentran conectados a Internet.

### 4.4.2. Aplicaciones

Las aplicaciones en el ámbito del Internet de las Cosas se pueden encontrar dispersas en áreas de conocimiento muy diferentes entre sí. Generalmente, podemos distinguir seis tipos distintos de aplicaciones englobadas en 2 categorías: información y análisis y automatización y control.

#### 4.4.2.1. Información y análisis

La información de la que dispone una organización es determinante a la hora de tomar una decisión. Continuamente se buscan medios para conocer y analizar una situación económica, un mercado o cualquier otro factor que pueda otorgar una ventaja competitiva a la empresa. Algunas compañas empiezan a interesarse por desarrollar aplicaciones del Internet de las cosas con el fin de conocer mejor el impacto de sus productos, realizar valoraciones o analizar el mercado.

Hoy en día es posible integrar todo tipo de sensores para tal fin, sin que esto repercuta seriamente en los costes de producción. Conociendo mejor el comportamiento de los consumidores, se conoce mejor el mercado en el que se opera, con lo que se puede mejorar el modelo de negocio. Es algo que empresas que ofrecen servicios en Internet como Google, Yahoo o Facebook llevan realizando desde hace tiempo con buenos resultados. La idea es la misma, pero extendida a todo tipo de productos, no solo software.

#### 4.4.2.2. Automatización y control

Este tipo de aplicaciones eran ya típicas en el pasado en el área de automatización industrial. Ahora, simplemente, se hacen ubicuas y se precisan de nuevas herramientas que puedan procesar la ingente cantidad de información que se recibe. Se plantea:

- La monitorización de sistemas.
- La toma de decisiones basadas en los datos capturados.

- La optimización de recursos y procesos.

En este tipo de aplicaciones se plantea cerrar el bucle de control de un proceso utilizando información proveniente del IoT. Esta información, tiene que ser procesada y traducida a acciones de control que permiten actuar sobre el proceso.

Optimización de procesos: algunas industrias ya estan haciendo uso de del Internet de las cosas para mejorar procesos y la calidad de sus productos o servicios a la vez que se reducen costes. Disponiendo sensores y actuadores conectados en red a lo largo de un proceso es posible ajustar los parámetros implicados en él con el objetivo de corregir pequeñas desviaciones. Hace tiempo que muchas industrias se benecian de este tipo de sistemas para asegurar el perfecto funcionamiento y control de las factorías.

Optimización del consumo de recursos: consiste en hacer uso de redes de sensores y sistemas de información automáticos con el objeto de cambiar los patrones de uso de recursos limitados como el agua o la energía. Las Smart Grids o redes eléctricas inteligentes son un ejemplo de este tipo de aplicaciones. En una Smart Grid se encuentran medidores inteligentes de energía que registran el consumo eléctrico de un hogar o de un tramo de la red eléctrica, permitiendo así definir un perfil de uso de la red eléctrica. Existen ya en el mercado un gran número de empresas que comercializan este tipo de dispositivos que permiten a empresas e instituciones hacer un uso más eficiente de la energía eléctrica y adherirse a planes y tarifas energéticas que les resulten más beneficiosas. La Smart City o ciudad inteligente es una idea que se obtiene extrapolando el concepto de las Smart Grids a todos los sistemas de abastecimiento que conforman una ciudad. Una Smart City consiste en un área urbana que integra las infraestructuras de energía (electricidad, gas y agua), redes de transportes y comunicaciones, entre otros, con el fin de un alcanzar un desarrollo sostenible y eficiente, respondiendo adecuadamente a las necesidades de cada individuo.

### 4.4.3. Big data

El término *Big Data* se emplea para referirse a nuevas técnicas, métodos y herramientas software que deben lidiar con cantidades de datos con un volumen mucho mayor al que estábamos acostumbrados hasta hace unos años. El Big Data plantea el almacenamiento y análisis de esta gran cantidad de información para obtener datos relevantes para empresas, comercios, etc.

El Big Data se alimenta de fuentes de datos muy diversas, estando las más importantes relacionadas con datos de seres humanos: correos electrónicos, mensajes publicados en redes sociales y otros permiten a obtener datos y estadísticas muy importantes para estudios de mercado y definir estrategias de marketing. En el sector bancario se manejan grandes volúmenes de datos relacionados con las transacciones realizadas, tendencias económicas y otros movimientos de los clientes.

### 4.4.4. Bases de datos NoSQL

Según se ha comentado con anterioridad, las bases de datos han evolucionado para ser capaces de almacenar y hacer accesible las grandes cantidades de información de las que estamos hablando. Las bases de datos que siguen el modelo SQL (Structured Query Language) se han considerado poco apropiadas para tareas de Big Data, con lo que las nuevas bases de datos orientadas a Big Data se han denominado NoSQL.

Las bases de datos NoSQL se han diseñado y estructurado para ser capaces de gestionar volúmenes de datos mucho mayores y, además, disponer de mecanismos adicionales para mantener estos datos (p.e. herramientas para mantener la consistencia de los datos y mecanismos para mantener bases de datos redundantes). Todo esto se ha hecho con un coste: estas base de datos no tienen las mismas capacidades para hacer consultas y operaciones que las bases de datos relacionales (RDBMS). En el caso de Apache Cassandra, los desarrolladores han diseñado un lenguaje propio CQL que tiene solamente ciertas similitudes con el lenguaje SQL.

Empresas como Google, Amazon o Facebook han desarrollado y utilizan estas bases de datos NoSQL en aplicaciones Web de tiempo real o para almacenar y analizar grandes cantidades de datos.

### 4.4.5. Teorema CAP

El teorema CAP [21] (por sus siglas en ingles: Consistency, Availability, Partition Tolerance) o teorema de Brewer, plantea que en un sistema de computación distribuida no es posible garantizar simultáneamente las siguientes características:

- *Consistencia*: plantea que toda la información esté totalmente sincronizada en todos los nodos del sistema en todo momento, de tal manera que todos estos nodos tenga la misma información en todo instante. Se dice que un sistema es consistente cuando opera replicando la información en todos los nodos del sistema. De esta forma, operación lectura obtiene la información más reciente introducida en el sistema.
- *Disponibilidad*: hace referencia a que el servicio esté disponible en cualquier momento. Es decir, que se garantice que toda petición reciba una respuesta.
- *Tolerancia al particionado*: que plantea que el sistema funcione correctamente a través de las divisiones físicas de la red, independientemente del hecho de que se pierdan mensajes o se sufran fallos en algunos nodos.

Según comentaremos más adelante en el apartado 4.5, aunque la solución elegida para almacenar los datos de tráfico no es capaz de cumplir, obviamente, con todos los puntos anteriores, desde un punto de vista práctico sí que cumple bien con todos ellos para la aplicación para la que la hemos destinado.

## 4.5. Cassandra

Se eligió utilizar Cassandra por tratarse de una base de datos NoSQL de código abierto, distribuida, escalable, con alta disponibilidad, consistencia ajustable y tolerante a fallos. Cassandra está basada en una arquitectura descentralizada en la cual la carga computacional se distribuye entre todos los nodos del cluster. Cassandra posee una arquitectura distribuida *peer-to-peer*, funcionando en múltiples máquinas al mismo tiempo. Cassandra está diseñada para funcionar en múltiples máquinas y tambien en múltiples centros de datos, e incluso en clusters dispersos por todo el mundo.

Cassandra tiene capacidad para escalarse añadiendo más nodos al sistema para, así, ser capaz de responder a un mayor numero de peticiones sin experimentar una gran degradacion en el rendimiento del servicio. En Cassandra, el crecimiento es lineal. La potencia aumenta linealmente en función del número de máquinas que trabajen en paralelo. Es decir, si dos nodos de Cassandra pueden ejecutar 100.000 operaciones por segundo, cuatro nodos iguales podrán con 200.000 operaciones por segundo.

Actualmente Cassandra es un proyecto de código abierto y principalmente está soportado por DataStax, así como Facebook, LinkedIn, Twitter y nScaled. DataStax es una compaña de software situada en California, cuyo modelo de negocio se centra en proporcionar asesoría y vender herramientas para la base de datos Apache Cassandra.

Cassandra está diseñada para manejar grandes cantidades de datos entre múltiples nodos sin fallos del tipo SPOF (Single Point Of Failure). Es un sistema distribuido con arquitectura *peer-to-peer* donde la información es dividida uniformemente y distribuida automáticamente por todos los nodos que conforman un data center, según una clave o valor único para cada conjunto de datos. En cada nodo, un registro secuencial de *commits* (commit log) captura todas las escrituras para asegurar la durabilidad de los datos. Los datos son entonces indexados y escritos en una estructura en memoria llamada *memtable*, similar a una cache de escritura demorada (write-back cache). Cuando la estructura de memoria se llena, la información es volcada en un archivo de datos SSTable. Una SSTable o *Sorted Strings Table*, es un concepto prestado de la *Google BigTable*, una de las bases de datos NoSQL en las que Cassandra se basa. Las SSTables se escriben en disco (puede ser SSD) y son inmutables, es decir, una vez el contenido ha sido escrito en ella, éste no puede ser modificado. Este concepto es fácil de entender si comprendemos que una SSTable es, básicamente, un fichero escrito en el disco duro de un PC y que estos ficheros pueden tener un tamaño considerable. Imaginemos que deseamos cambiar un dato guardado en un fichero: esta tarea precisa leer el fichero, encontrar el dato a modificar y volver a guardar el fichero. Lógicamente, el proceso resulta bastante ineficiente y lento si pretendemos manejar una cantidad considerable de datos.

Con estas ideas en mente, nos deberíamos preguntar cómo es posible entonces, modificar una entrada en una tabla o bien eliminar un dato. El paradigma que se emplea en Cassandra se puede resumir mediante el siguiente lema: siempre es más rápido escribir que 'buscar, modificar y escribir'. Si deseamos eliminar un dato, Cassandra realmente escribirá ese dato y anotará que está borrado (con un timestamp posterior).

Si deseamos modificar el dato, Cassandra escribirá el nuevo valor del dato y anotará un tiempo. A la hora de consultar un dato Cassandra se queda con las entradas más recientes y descarta el resto. Dicho de otra manera, el borrado de datos en Cassandra es *soft*, es decir los datos no son borrados de la SSTable una vez almacenados, sino que son marcados a través de *tombstones*, de tal manera que éstos no aparecen en los resultados de las consultas. En la figura 4.1 se encuentra un esquema del proceso de escritura de datos en Cassandra.

El proceso planteado hasta el momento requiere, por tanto, escribir siempre los datos, lo que podría ser ineficiente si, en efecto, deseamos eliminar datos de forma periódica. Para resolver esta cuestión, Cassandra implementa un proceso llamado compactación. Mediante un proceso de compactación periódica Cassandra consolida las SSTables periódicamente, descartando datos obsoletos y actualizando nuevos valores. Las entradas marcadas con tombstones son borradas definitivamente en el proceso de compactación cuando el tiempo indicado en gc_grace_seconds (garbage collector grace seconds) ha expirado, el cuál es un parámetro que se define por tabla de datos.

Por otra parte, en Cassandra el factor de replicación es un parámetro que determina cuantas copias de los datos existirán en el cluster. Todas las escrituras son partidas y replicadas en distintos nodos para conseguir una gran disponibilidad y tolerancia a fallos del sistema. Si un nodo cae, la información puede ser extraída de otro nodo que contenga una copia de los datos originales. El factor de replicación debe ser ajustado en función de la probabilidad de que un nodo del sistema caiga. Normalmente se utilizan factores de replicación de 2 o 3 para asegurar la redundancia de los datos y por lo tanto su disponibilidad en caso de fallo de algún nodo.

Al añadir nuevos nodos a Cassandra, cada nuevo nodo se hace responsable de un conjunto de datos denido por un rango de claves específico (véase tolerancia al particionado en el teorema CAP en el apartado 4.4.5). El protocolo de comunicación interno de Cassandra es el encargado de saber qué nodo tiene los datos con la clave a la que se quiere acceder, de tal manera que cuando se desean consultar dichos datos, contenidos en una clave específica, el nodo responsable pueda responder rápidamente con toda la información solicitada. Del mismo modo al escribir, los datos se almacenan en los nodos responsables de la clave indicada.

Al replicar los datos existe la posibilidad de llegar a inconsistencias cuando las replicas poseen estados distintos de las mismas estructuras. Para evitar y poder resolver errores de este tipo, Cassandra hace uso de diferentes mecanismos como anti-entropía, lectura y reparación con el fin de sincronizar correctamente los datos entre los distintos nodos. A su vez, Cassandra ofrece la posibilidad de ajustar el nivel de consistencia en el momento de realizar una lectura o escritura para controlar la disponibilidad frente a la exactitud de los datos (Teorema CAP en el apartado 4.4.5).

Las peticiones de escritura o lectura por parte de los cliente pueden ser enviadas contra cualquier nodo del cluster. Cuando un cliente se conecta a un nodo cualquiera con una petición, ese nodo se convierte en el coordinador de la operación, actuando como un proxy entre la aplicación del cliente y los nodos que poseen los datos solicitados.

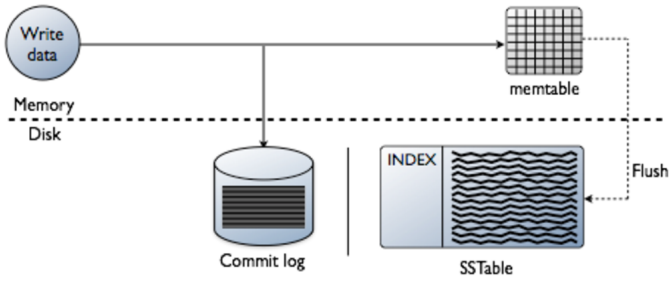



**Figura 4.1:** Se presenta un esquema del proceso de escritura de datos en Cassandra.

El coordinador decide qué nodo ha de recibir la solicitud en base a cómo está configurado el cluster. Este comportamiento está definido por los niveles de consistencia que pueden ser los siguientes:

Para la escritura de un dato:

- ALL: La escritura ha de realizarse en el commit log y la memtable de todos los nodos replica de la partición indicada.
- EACH QUORUM: Indica una fuerte consistencia. La escritura ha de realizarse en el commit log y la memtable de un conjunto de nodos réplica de todos los data centers.
- QUORUM: La escritura ha de realizarse en el commit log y memtable de un conjunto de nodos réplica.
- LOCAL QUORUM:La escritura ha de realizarse en el commit log y memtable de un conjunto de nodos réplica del mismo data center que el nodo coordinador. Evita la latencia en la comunicación interna del centro de datos.
- LOCAL ONE: La escritura ha de alcanzar al menos un nodo réplica en el data center.
- ONE: La escritura ha de realizarse en el commit log y memtable de al menos un nodo réplica.
- TWO: La escritura ha de realizarse en el commit log y memtable de al menos dos nodos réplica.
- THREE: La escritura ha de realizarse en el commit log y memtable de al menos tres nodos réplica.
- ANY: La escritura ha de alcanzar al menos un nodo. Si todos los nodos responsables de la clave indicada no están disponibles, la escritura puede llevarse a cabo por medio de un mecanismo llamado Hinted Handoff.

Para la lectura de un dato:

- ALL: Devuelve el registro después de que todas las réplicas hayan respondido. Si un solo nodo no responde, la operación fallará. Además, Cassandra
- EACH QUORUM: Devuelve el registro obtenido después de que todo un conjunto de réplicas de cada data center haya respondido.
- QUORUM: Devuelve el registro obtenido después de que un conjunto de réplicas de cualquier data center haya respondido.
- LOCAL QUORUM: Devuelve el registro obtenido por el nodo coordinador después de que un conjunto de réplicas del data center actual haya respondido.
- LOCAL ONE: Devuelve la respuesta del nodo réplica más cercano en el data center.
- ONE: Devuelve la respuesta del nodo réplica más cercano determinado por nodo coordinador.
- TWO: Devuelve los datos de los dos nodos réplica más cercanos.
- THREE: Devuelve los datos de los tres nodos réplica más cercanos.

### 4.5.1. Cassandra Query Language

Cassandra ha desarrollado una 'variante' del lenguaje SQL que ha denominado CQL (*Cassandra Query Language*). En este lenguaje están implementadas las limitaciones propias de la base de datos relativas principalmente al tipo de consultas que se pueden hacer sobre las tablas.

### 4.5.2. Keyspace

En Cassandra, las tablas se agrupan en estructuras llamadas keyspaces. Una aplicación accederá a un keyspace para leer o escribir datos. El keyspace puede definirse mediante la siguiente sentencia CQL:

```
CREATE KEYSPACE datosdgt WITH replication =
{'class': 'SimpleStrategy', 'replication_factor': 2};
```

El factor de replicación (RF) de los datos es una propiedad del keyspace que se establece al crearlo. En el caso de que datos de una misma aplicación requieran distintos factores de replicación, estos deberán ir en keyspaces diferentes. El factor de replicación puede ser modificado, aunque esto conlleva un incremento instantáneo de las transacciones de datos entre los nodos, pues si se aumenta el RF es necesario propagar más información a otros nodos que no la tenían.

### 4.5.3. Series de datos temporales

El modelado de los datos de cada sensor (estaciones metereológicas, estaciones aforadas, etc.) que la plataforma almacena se ha realizado siguiendo los principios de diseño que propone Datastax en [60].

Cassandra almacena los datos relativos a secuencias ordenadas en archivos SSTable. Con su diseño se puede acceder posteriormente a los datos de forma rápida y eficiente minimizando las búsquedas en disco. Las series de datos temporales encajan perfectamente para el uso para el que está diseñada Cassandra.

Cassandra es capaz de almacenar hasta dos mil millones de columnas por fila. Esta cantidad puede ser más que suficiente para almacenar todos los datos obtenidos por un sensor a lo largo de toda la vida útil.

```
CREATE TABLE data_points (
device_id uuid,
sensor_name text,
time timestamp,
value text,
type text,
units text,
PRIMARY KEY ((device_id, sensor_name), time) )
WITH CLUSTERING ORDER BY (time DESC);
```

La anterior sentencia de CQL crea una tabla denominada *data_points*. Esta tabla agrupa los datos según un clave de partición compuesta que está formada por el identificador del dispositivo al que pertenece el sensor (de tipo UUID), el nombre del sensor y el tiempo en el que se registran los datos. El valor del sensor es un tipo texto, lo que posibilita almacenar todo tipo de valores, por ejemplo:

- float, int, etc.
- literales: lluvia intensa, vierto fuerte, nieve, etc.

El valor del sensor se obtiene siempre por una operación de *cast* en función de la información que se almacene en el campo *type*.

### 4.5.4. Driver para python

La arquitectura se ha basado en un driver para python que encapsula las comunicaciones con la base de datos y se encarga también de gestionar el modelado de la base de datos. El driver se denomina *CQL Engine* y es uno de los drivers estándar que ofrece Cassandra para desarrollar aplicaciones basadas en ella.

Como ejemplo, la tabla de *data points* definida en el apartado 4.5.3 se puede definir en python como:

```
class DataPoint(Model):
    __table_name__ = 'datapoints'
    name = Text(partition_key=True)
    sensor_name = Text(primary_key=True)
    time = DateTime(primary_key=True, clustering_order='DESC')
    value = Text(required=True)
    type = Text()
    units = Text()
```

### 4.5.5. Cassandra shell

Cassandra dispone de un intérprete de comandos denominado cqlsh. Podemos acceder al intérprete mediante:

```
# cqlsh
Connected to Test Cluster at 127.0.0.1:9042.
[cqlsh 5.0.1 | Cassandra 3.9 | CQL spec 3.4.2 | Native protocol v4]
Use HELP for help.
cqlsh>
```

El intérprete de comandos acepta sentencias de tipo CQL (que recordarán a muchos las sentencias de tipo SQL, pero con las limitaciones indicadas en el apartado 4.5.7). Por ejemplo, si se desean seleccionar todas las columnas de una tabla en concreto y mostrar todos los datos por consola:

```
#cqlsh
Connected to Test Cluster at 127.0.0.1:9042.
[cqlsh 5.0.1 | Cassandra 3.9 | CQL spec 3.4.2 | Native protocol v4]
Use HELP for help.
cqlsh> use carmetryapp ;
cqlsh:carmetryapp> use carmetry;
cqlsh:carmetry> select * from trafficobjectsactivelist ;

 category | road_name | event_id | cause | city | km_end | km_start |
 lat | lng | message_advice_long | message_advice_short | metadata |
 province | severity | text_times | traffic_direction | update_time
```

### 4.5.6. Claves compuestas y *clustering*

En Cassandra se definen una clase especial de claves primarias, denominadas *partition key* o claves de partición. Una clave de partición determina en qué nodo del cluster se almacena un dato en particular. Para ello, Cassandra utiliza el valor que se asigne a la columna que hayamos designado como `partition key` para calcular una función de tipo hash. En base al valor obtenido, Cassandra decide qué nodo(s) debe almacenar el dato. En el siguiente ejemplo, definimos mediante python la siguiente tabla:

```
class Message(Model):
    __table_name__ = 'messages'
    message_source = Text(partition_key=True)
    time = DateTime(primary_key=True, clustering_order='DESC')
    message = Text()
```

Esto crea en Cassandra una tabla denominada 'messages' que está destinada al almacenamiento de mensajes de error a modo de `log` del sistema. La columna denominada `message_source` almacenará el nodo del sistema que generó el error. La columna `time` almacenará el tiempo en el que se produjo el error. En el ejemplo se le indica a Cassandra que almacene el tiempo de forma descendente 'DESC'

Lo que en CQL es equivalente a:

```
cqlsh:datosdgt> describe messages;

CREATE TABLE carmetry.messages (
    message_source text,
    time timestamp,
    message text,
    PRIMARY KEY (message_source, time)
) WITH CLUSTERING ORDER BY (time DESC)
```

Esto nos indica que:

- La columna denominada `message_source` se ha definido como *partition key*.
- La columna indicada como `time` está definida como una *clustering column*. Una columna de este tipo únicamente indica que Cassandra la ordenará en el orden que se indique para poder hacer las búsquedas de forma más rápida.

Para simplificar el diseño de las tablas utilizando el driver de python CQL engine es aconsejable:

- Indicar las columnas que se deseen utilizar como *partition keys* utilizando la opción `partition_key=True`.
- El resto de columnas que se indique como `primary_key=True` serán claves de tipo *clustering*.

En Cassandra es posible utilizar claves de partición compuestas por varias columnas. En realidad el propósito de la clave de partición es hacer que los datos se distribuyan bien entre todos los nodos del cluster. Imaginemos que tenemos una tabla como la siguiente destinada a guardar nombres de personas:

```
class Names(Model):
    __table_name__ = 'names'
    first_name = Text(partition_key=True)
    last_name = Text(primary_key=True, clustering_order='DESC')
    age = Integer()
```

Con esto, estamos eligiendo el *nombre de pila* de una persona como clave de partición y el apellido como clave de cluster. Imaginemos que se desean almacenar muchos nombres y que, además, sabemos que hay nombres de pila mucho más frecuentes que otros. Con esta definición podría ocurrir que algunos nodos del cluster almacenaran más información que otros. Podríamos aumentar la variabilidad con la que se distribuye la información en el cluster haciendo lo siguiente:

```
class Names(Model):
    __table_name__ = 'names'
    first_name = Text(partition_key=True)
    last_name = Text(partition_key=True)
    age = Integer()
```

Con esto Cassandra utilizará una clave de partición compuesta formada por nombre y apellido de la persona. Cassandra utiliza siempre la primera columna que se indique como clave primaria.

### 4.5.7. Sentencias CQL

En esta sección hablaremos de los tipos de sentencias CQL que se pueden realizar en base a las definiciones de tablas posibles. La empresa Datastax no explica del todo claramente qué tipo de *queries* se pueden realizar utilizando la sentencia de CQL SELECT. En las líneas siguientes intentaremos aclararlo y dar unas pinceladas sobre el diseño de tablas para poder hacer las consultas que se requieran. El paradigma usado al trabajar con bases de datos NoSQL suele plantear, como primer paso, lo siguiente:

- Se comienza anotando todas las consultas que se van a realizar sobre los datos.
- Conocidas las consultas se diseñan las tablas, claves de partición y claves de cluster para que se puedan realizar las consultas anteriores.

Por otra parte, y comparando con el lenguaje de consultas SQL, tenemos que en CQL:

- No existen las sentencias de tipo JOIN.
- No existen sentencias de tipo GROUP BY.
- No podemos hacer sentencias WHERE de cualquier tipo.
- No podemos ordenar lo datos devueltos de cualquier manera (ORDER BY).

Básicamente una sentencia WHERE en CQL funciona únicamente en las claves que se hayan indicado como primarias (de tipo cluster o de tipo partición) o bien en columnas que se hayan definido con índices secundarios. Se pueden combinar varias sentencias de tipo WHERE usando AND, pero no OR.

La clave de partición indicada en WHERE indica a Cassandra a qué nodo del cluster debe realizar la consulta, pues este nodo es el que posee los datos. Esto implica una restricción importante, pues, en principio el valor de la clave de partición se debe especificar siempre al hacer una consulta.

Como ejemplo observaremos la tabla donde se almacenan todos los objetos de tráfico:

```
class TrafficObjectAllList(Model):
    __table_name__ = 'trafficobjectsalllist'
    road_name = Text(partition_key=True)
    cause = Text(primary_key=True)
    kind = Text(primary_key=True)
    sub_kind = Text(primary_key=True)
    event_id = Text(primary_key=True)
    lat = Float()
    lng = Float()
    level = Text()
    province = Text()
    city = Text()
    traffic_direction = Text()
    km_start = Float()
    km_end = Float()
    message = Text()
    metadata=Text()
```

El nombre de la carretera (`road_name`) es la clave de partición y observamos que otras columnas se han definido como claves de cluster. Veremos qué operaciones se pueden hacer con CQL y qué operaciones no están permitidas o aconsejadas utilizando una serie de ejemplos con la tabla anterior. Para poder hacer búsquedas eficientes en las columnas podemos hacer las siguientes consultas:

- Búsqueda de todos los accidentes de tráfico:

```
cqlsh:datosdgt> SELECT * FROM trafficobjects;
```

- Búsqueda de todos los accidentes en una carretera:

```
>SELECT * FROM trafficobjects WHERE road_name = 'se-30';

road_name | cause   | kind   | sub_kind    | event_id | city |...
se-30 | averia | incidencia | undefined | 1037339 | sevilla |
se-30 |circulacion | incidencia |restricciones | 921650 | sevilla
se-30 |circulacion | incidencia |restricciones | 921529 | sevilla |
```

- Podemos hacer *queries* para las averías que se hayan dado en la carretera SE-30:

```
>SELECT * FROM trafficobjects WHERE road_name = 'se-30'
AND cause = 'averia';

road_name | cause   | kind   | sub_kind    | event_id | city |...
se-30 | averia | incidencia | undefined | 1037339 | sevilla |
```

- No podemos consultar todas las causas de los objetos de tráfico que correspondan a avería. Como nota: buscaremos siempre de izquierda a derecha, para buscar un dato que se encuentre a la derecha siempre debemos especificar los de la izquierda. Ejemplo:

```
>SELECT * FROM trafficobjects WHERE cause = 'averia';
InvalidRequest: Error from server: code=2200 [Invalid query]
message="Cannot execute this query as it might involve data
filtering and thus may have unpredictable performance.
If you want to execute this query despite the performance
unpredictability, use ALLOW FILTERING
```

  Nótese que si no especificamos un valor para la '*partition key*' estamos obligando a Cassandra a consultar todos los datos en todos los nodos, lo que es totalmente ineficiente. Por esta razón, la consola de comandos devuelve un error y sugiere que se use ALLOW FILTERING, aún a expensas de perder la velocidad que proporciona Cassandra.

- La clausula 'ALLOW FILTERING' permite hacer excepciones al caso anterior, pero está totalmente desaconsejada. Nótese la siguiente consulta:

```
>select * from trafficobjects where kind = 'incidencia'

ALLOW FILTERING;
 road_name | cause           | kind        | sub_kind   | event_id |
cv-240 | meteorologiaadversa | incidencia | undefined | 1108783 |
n-121a | undefined | incidencia | otros | 968057 |
n-121a | undefined | incidencia | undefined |968057 |
a-136  | meteorologiaadversa | incidencia | undefined |1108437 |
a-1511 | meteorologiaadversa | incidencia | undefined |1108772 |
```

Con estas ideas en mente, en ocasiones es necesario crear dos tablas con ordenaciones diferentes de las columnas para poder realizar las consultas deseadas de forma eficiente.

## 4.6. Índices secundarios

Cassandra permite la definición de algunas columnas como índices secundarios sobre los que se pueden realizar consultas. Existen varias contraindicaciones para hacer esto según la documentación de DataStax con lo que es necesario siempre realizar pruebas de funcionamiento. Veremos un ejemplo de definición de columna con un índice secundario en python:

```
class TrafficObjectAllList(Model):
    __table_name__ = 'trafficobjectsalllist'
    road_name = Text(partition_key=True)
    cause = Text(primary_key=True)
    kind = Text(primary_key=True)
    sub_kind = Text(primary_key=True)
    event_id = Text(primary_key=True)
    lat = Float()
    lng = Float()
    level = Text()
    province = Text(index=True)
    city = Text()
    traffic_direction = Text()
    km_start = Float()
    km_end = Float()
    message = Text()
    metadata=Text()
```

En esta tabla se ha especificado que provincia va a ser un índice secundario.

## 4.7. Creación de un cluster para la captura de datos de tráfico

En este apartado se describe el cluster que se creó con el objetivo de almacenar los datos de tráfico necesarios para realizar algunos de los estudios presentados en esta Tesis. Durante los inicios de la investigación no existía ningún portal que permitiera descargar datos abiertos para el estudio de los accidentes de tráfico en España. Este hecho motivó la creación del cluster basado en Cassandra que se describe a continuación.

### 4.7.1. Arquitectura de red

La arquitectura que se muestra a continuación corresponde a un único *data center* formado por un conjunto de nodos que forman un cluster. Un *data center* en Cassandra hace referencia a un grupo de máquinas que se encuentran replicadas. Es decir, los datos se encuentran replicados entre el grupo con el RF determinado (apartado 4.5.2). Para que todo funcione correctamente es necesario que Cassandra esté instalada y corriendo en todos los nodos del cluster. Además, es necesario fijar y configurar una serie de parámetros, principalmente:

- Las direcciones IP de cada nodo para poder enlazarlos todos correctamente.

| Nombre | Dirección IP | Función |
|---|---|---|
| pocoyo | 192.168.1.10 | Captura de datos. |
| elly | 192.168.1.110 | Nodo de Cassandra (seed node) |
| valentina | 192.168.1.220 | Nodo de Cassandra |
| pulpo | 192.168.1.221 | Nodo de Cassandra |
| ballena | 192.168.1.222 | Nodo de Cassandra |
| mariposa | 192.168.1.223 | Nodo de Cassandra (seed node) |
| nube | 192.168.1.224 | Nodo de Cassandra |
| alien | 192.168.1.225 | Nodo de Cassandra |
| orquesta | 192.168.1.226 | Nodo de Cassandra |
| flor | 192.168.1.227 | Nodo de Cassandra |
| pelotas | 192.168.1.228 | Nodo de Cassandra (seed node) |

**Tabla 4.1:** Nodos del cluster

- Los nodos que serán las semillas o *seed nodes*.

En esta aplicación se ha instalado Cassandra en 10 máquinas con las direcciones y funciones especificadas en la tabla 4.1. Cassandra funciona gracias a una máquina virtual de Java (JVM) y necesita ingentes cantidades de memoria RAM. Así pues, resulta recomendable que las máquinas que en las que está instalada Cassandra no tengan otras funciones. En consecuencia, en la arquitectura descrita en la tabla 4.1 existe un PC encargado de realizar peticiones de captura de datos, procesarlas y hacer peticiones a Cassandra para almacenar los datos. El resto de nodos se dedican a ser nodos de Cassandra. Algunos de esos nodos se designan como nodos coordinadores (*seed nodes*) que se encargan de tareas de coordinación dentro del cluster.

### 4.7.2. Arranque de Cassandra

En cada uno de los nodos se arranca automáticamente Cassandra en el inicio o a través de:

```
# service cassandra start
```

Para observar la configuración y estado del cluster podemos utilizar la herramienta nodetool.

```
# nodetool status datosdgt
Datacenter: DC1
===============
Status=Up/Down
|/ State=Normal/Leaving/Joining/Moving
--  Address  Load       Tokens  Owns    Host ID                               Rack
UN  x.110  2,44 GB  256     51,3%  7138ebbb-1f46-41ed-8a4a-c1d3abf42176  RAC1
UN  x.220  2,35 GB  256     49,2%  448fe5cb-60f9-4bff-8f2a-e8cc4b8993c1  RAC1
UN  x.221  2,34 GB  256     51,3%  1a9b1f70-854f-11e7-873d-fa163ea135d4  RAC1
UN  x.222  2,18 GB  256     48,4%  1b8925a8-85e5-a3a0-9bfb-3c075438bda5  RAC1
UN  x.223  2,25 GB  256     50,8%  1c255fd6-8655-e724-7b54-6b075438e814  RAC1
UN  x.224  2,32 GB  256     49,1%  1c6741c6-8690-5e5a-1eba-1c408918cbf8  RAC1
UN  x.225  2,65 GB  256     49,6%  1cd17bcc-8741-ab02-ac3e-968754385da8  RAC1
UN  x.226  2,87 GB  256     50,7%  1d5b6aa8-87ba-f174-1c46-4aa754386eh5  RAC1
UN  x.227  2,29 GB  256     50,8%  640fea6e-da23-eda0-78eb-3f275438b225  RAC1
UN  x.228  2,21 GB  256     48,9%  637ea6f8-54ba-72fa-6ea1-1af75438bdff  RAC1
```

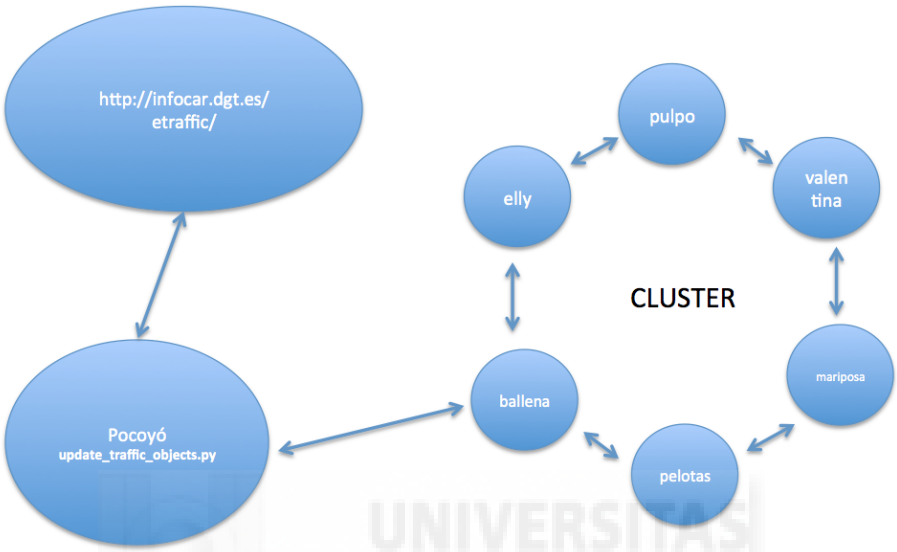



**Figura 4.2:** Se presenta un esquema software y de arquitectura del cluster de Cassandra diseñado.

En el resultado mostrado se han ocultado parcialmente las direcciones IP por claridad. La información de carga (*load*) hace referencia a la cantidad de información que almacena cada nodo. La columna indicada como *owns* indica el porcentaje de datos que almacena cada nodo y está relacionado con el factor de replicación RF. En nuestro caso, dado que se ha elegido RF=2, cada nodo almacena su información y, además, almacenará los datos replicados de otro equipo, de ahí que su valor teórico sea del 50 %. Cada nodo está identificado por un tipo UUID, según se indica en el cuadro.

## 4.8. Diseño de la arquitectura software

En la figura siguiente se presenta un esquema de la arquitectura de red y del software. El software está formado por un cliente python denominado `update_traffic_objects.py` que se ejecuta en la máquina que proporciona los servicios. Para obtener los datos de tráfico se conecta al servicio de la DGT (inforcar.dgt.es).

### 4.8.1. Diseño de tablas y queries

Se presentan, a continuación, todas las tablas diseñadas para almacenar los eventos de tráfico y se especifica la función de cada una.

```
class TrafficObjectAllList(Model):
    __table_name__ = 'trafficobjectsalllist'
    road_name = Text(partition_key=True, default='undefined')
    cause = Text(primary_key=True, default='undefined')
    kind = Text(primary_key=True, default='undefined')
    sub_kind = Text(primary_key=True, default='undefined')
    event_id = Text(primary_key=True, clustering_order='DESC')
    lat = Float()
    lng = Float()
    level = Text()
    province = Text()
    city = Text()
    #text_times=Map(key_type=Text, value_type=Text)
    text_times = Text()
    update_time = DateTime()
    traffic_direction = Text()
    km_start = Float()
    km_end = Float()
    message_advice_short = Text()
    message_advice_long = Text()
    metadata=Text()

class TrafficObjectActiveList(Model):
    __table_name__ = 'trafficobjectsactivelist'
    category = Text(partition_key=True, default='undefined')
    road_name = Text(primary_key=True, default='undefined')
    event_id = Text(primary_key=True, clustering_order='DESC')
    cause = Text()
    lat = Float()
    lng = Float()
    #level = Text()
    severity = Text()
    traffic_direction = Text()
    km_start = Float()
    km_end = Float()
    metadata=Text()
    province = Text()
    city = Text()
    text_times=Text()
    update_time = DateTime()
    message_advice_short = Text()
    message_advice_long = Text()

class TrafficObjectActiveGeolocated(Model):
    __table_name__ = 'trafficobjectsactivegeolocated'
    geolocator = Text(partition_key=True)
    category = Text(primary_key=True,clustering_order='DESC')
    severity = Text(primary_key=True,clustering_order='DESC')
    event_id = Text(primary_key=True,clustering_order='DESC')
    road_name = Text()
    cause = Text()
    lat = Float()
    lng = Float()
    traffic_direction = Text()
    km_start = Float()
    km_end = Float()
    metadata=Text()
    province = Text()
    city = Text()
    text_times=Text()
    update_time = DateTime()
    message_advice_short = Text()
    message_advice_long = Text()

class RoadName(Model):
    __table_name__ = 'roadnames'
    country = Text(partition_key=True, default='undefined')
    road_name = Text(primary_key=True, clustering_order='DESC')

class TrafficObjectCategory(Model):
    __table_name__ = 'trafficobjectscategory'
```

```
category = Text(partition_key=True, default = 'undefined')
```

Describimos cada una de ellas a continuación:

- `trafficobjectsalllist`: Mantiene una lista de todos los eventos de tráfico que se han producido. El script de python `update_traffic_objects.py` se dedica periódicamente (cada 3 minutos) a hacer consultas al servicio web

  `http://infocar.dgt.es/etraffic` y guardar todos los elementos encontrados en dicha tabla.

- `trafficobjectsactivelist`: Mantiene una lista de los eventos que se encuentran vigentes. Para hacer esto empleamos una característica muy práctica de Cassandra que se denomina TTL (*time to live*). Al escribir cualquier dato en la base de datos se indica cuál será su periodo de vigencia. Pasado el periodo de vigencia Cassandra considera que el dato se ha borrado. En este caso, cada 3 minutos (véase el punto anterior) se hacen consultas al sistema y se especifica un TTL para todos los elementos de 4 minutos. Así, cuando un elemento deja de ser anunciado por la DGT, pasados 4 minutos se borra automáticamente de nuestra base de datos.

- `trafficobjectsactivegeolocated`: Se trata de una lista de todos los eventos de tráfico que están activos. En este caso, se utiliza un *geohash* para poder hacer búsquedas georeferenciadas.

- `roadnames`: La lista de carreteras españolas.

- `trafficobjectscategory`: Listado de todas las categorías de eventos de tráfico que se indexan.

## 4.9. Pruebas realizadas

Existe un paquete denominado *cassandra-stress* que permite realizar pruebas de carga para ciertas configuraciones de tablas y parámetros. En nuestro caso y debido a las deficiencias encontradas en el paquete, decidimos realizar las pruebas desarrollando nosotros mismos las definiciones de tablas y las pruebas a realizar mediante scripts de python.

Se realizaron pruebas de dos tipos: pruebas de rendimiento y de tolerancia a fallos. Las primeras se centraron principalmente en intentar calcular la velocidad de respuesta del cluster de Cassandra en operaciones de escritura y de lectura. Por otra parte, las pruebas de tolerancia a fallos buscaron determinar la capacidad de recuperación del sistema cuando alguno de los nodos de nuestro sistema fallaba.

| Nombre | Procesador | RAM |
|---|---|---|
| pocoyó | Intel Quad Core i5 | 8 GB |
| elly | Intel dual Core i3 | 4 GB |
| valentina | Intel dual Core i3 | 4 GB |
| pulpo | Intel dual Core i3 | 4 GB |
| ballena | Intel dual Core i3 | 4 GB |
| mariposa | Intel dual Core i3 | 4 GB |
| nube | Intel dual Core i3 | 4 GB |
| alien | Intel dual Core i3 | 4 GB |
| orquesta | Intel dual Core i3 | 4 GB |
| flor | Intel dual Core i3 | 4 GB |
| pelotas | Intel dual Core i3 | 4 GB |

**Tabla 4.2:** Nodos del cluster

### 4.9.1. Características de los equipos

Durante el transcurso de la investigación llevada a cabo las características de los equipos han ido cambiando con el propósito de probar diferentes configuraciones y necesidades hardware. Finalmente todas las máquinas tienen las características mostradas en la tabla 4.2. Debido a las necesidades de memoria que plantea Cassandra y la JVM, DataStax recomienda un mínimo de 4 GB de memoria RAM y discos duros de estado sólido (SSD). No obstante, en los PC's descritos en la tabla 4.2 se usaron discos duros magnéticos debido a tener un presupuesto limitado. Los resultados obtenidos son aceptables para la aplicación que se describe y se asume que los resultados habrán sido significativamente mejores al haber utilizado discos SSD. La máquina denominada *pocoyo* se encarga de hacer peticiones para recabar datos y se encarga también de realizar peticiones. Por esa razón cuenta con recursos mayores.

### 4.9.2. Alterando el factor de replicación

Se hicieron pruebas de alteración del factor de replicación, lo que simularía el hecho de que, en una aplicación determinada, se precisará una mayor seguridad en los datos. Para ello, Desde cqlsh (o bien usando cqlengine y python) se pueden cambiar "en caliente" muchas características de tablas de cassandra. Por ejemplo, cambiar el *replication factor* de un *keyspace*, o cambiar las opciones de compactación de una tabla. Por ejemplo, para cambiar las opciones de compactación haremos:

```
cqlsh:datosdgt> ALTER TABLE datapoints WITH compaction =
{'class': 'DateTieredCompactionStrategy',
'timestamp_resolution':'MICROSECONDS',
'base_time_seconds':'3600',
'max_sstable_age_days':'365',
'min_threshold': '4',
'max_threshold': '128',
'tombstone_threshold': '0.01'};
```

Por ejemplo, para la tabla de datapoints del keyspace *datosdgt* donde se almacenan los datos meteorológicos y las cargas de cada carretera. Una vez se ha realizado la

operación anterior, en cada nodo del cluster, Cassandra comprueba si el estado de compactación de las sstables es correcto e iniciará tareas de compactación si lo considera necesario.

A continuación indicamos una prueba en la que se alteró el factor de replicación del cluster:

```
cqlsh> ALTER KEYSPACE carmetryapp WITH REPLICATION =
{ 'class' : 'SimpleStrategy',
'replication_factor' : 3 };
```

En este ejemplo, inicialmente, teníamos $RF = 2$ en un cluster de prueba con 4 nodos, en este caso, con el comando ALTER KEYSPACE estamos aumentando la redundancia de datos. En consecuencia, se deberían crear nuevas copias de los datos en el resto de nodos. Después de ejecutar el comando ALTER KEYSPACE es necesario ejecutar una acción de reparación (`nodetool repair`) en cada uno de los nodos para iniciar la propagación de datos desde unos nodos a los nodos que replique los datos. En el `log` de Cassandra vemos que se hace un streaming de datos al resto de nodos. También observamos que se ejecutan tareas de compactación sobre los datos que se están recibiendo.

Antes de aumentar el RF tenemos que:

```
root@pato:~# nodetool status datos dgt
Datacenter: DC1
===============
Status=Up/Down
|/ State=Normal/Leaving/Joining/Moving
--  Address Load  Tokens Owns   Host ID                               Rack
UN  x.90  1,64 GB 256 50,5%  7138ebbb-1f46-41ed-8a4a-c1d3abf42176  RAC1
UN  x.110 1,31 GB 256 51,2%  948fe5cb-60f9-4bff-8f2a-e8cc4b8993c1  RAC1
UN  x.100 1,54 GB 256 51,7%  4d9fbd6d-5a51-491c-9b1b-f79b9e2e855f  RAC1
UN  x.70  1,23 GB 256 46,5%  8461146a-376d-4221-90c1-7b4522bdeeb9  RAC1
```

A continuación, y después de ejecutar el comando ALTER TABLE y el comando `nodetool repair` en todos los nodos, obtenemos la siguiente configuración:

```
root@pato:~# nodetool status carmetryapp
Datacenter: DC1
===============
Status=Up/Down
|/ State=Normal/Leaving/Joining/Moving
--  Address Load  Tokens Owns   Host ID                               Rack
UN  x.90  3,64 GB 256 75,5%  7138ebbb-1f46-41ed-8a4a-c1d3abf42176  RAC1
UN  x.110 3,31 GB 256 71,2%  948fe5cb-60f9-4bff-8f2a-e8cc4b8993c1  RAC1
UN  x.100 3,54 GB 256 75,7%  4d9fbd6d-5a51-491c-9b1b-f79b9e2e855f  RAC1
UN  x.70  3,23 GB 256 76,5%  8461146a-376d-4221-90c1-7b4522bdeeb9  RAC1
```

Vemos, en este caso, que ahora cada nodo tiene una participación de datos del 75 % aproximadamente. Es decir: cada nodo almacena sus datos y los de otros dos nodos, de manera que los datos esté replicados 3 veces en todo el cluster. La cantidad total de datos ha aumentado también. Durante este proceso se observó que el número de sstables en todos los nodos se disparaba ya que Cassandra decide replicar los datos enviando sstables a otros nodos. P

En la siguiente prueba disminuímos el RF. En este caso, pasamos de tener RF=3 a RF=2. Para ello, hacemos:

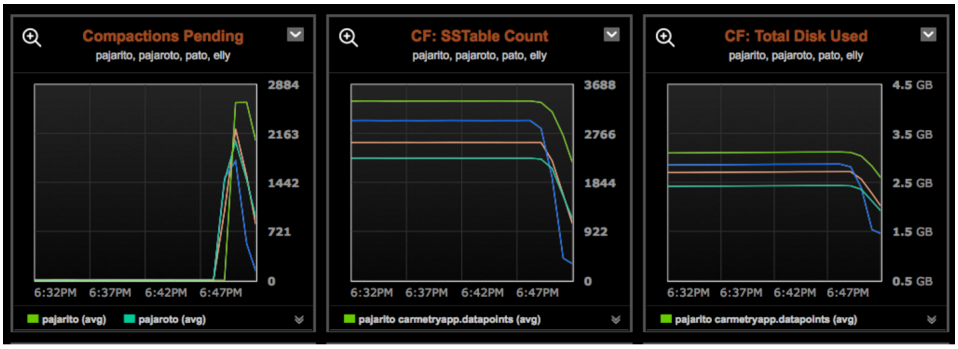



**Figura 4.3:** Se presenta la evolución de las SSTables durante un cambio en el RF del cluster.

```
cqlsh> ALTER KEYSPACE carmetryapp WITH REPLICATION =
  { 'class' : 'SimpleStrategy', 'replication_factor' : 2 };
```

Seguidamente, se debe ejecutar `nodetool cleanup` para quitar las replicas accesorias de datos que se encuentran distribuidas en el cluster. Durante la ejecución del comando `nodetool cleanup` se observó que el número de compactaciones necesarias crecía mientras el número de sstables y el `total disk used` disminuían, según se puede observar en la figura 4.3.

## 4.10. Resultados de las pruebas de carga

Se presenta a continuación los resultados de escritura y lectura de una prueba de larga duración. Simulamos que existen una serie de clientes python (que se conectan directamente al cluster). Con esta prueba pretendemos mostrar que el cluster es estable con un gran número de escrituras/lecturas y que los tiempos de escritura y lectura se degradan de forma lineal. Además, durante la prueba se hace que uno de los nodos caiga y se comprueba que el sistema sigue funcionando con normalidad. En la tabla 4.3 se presentan las condiciones en las que se realizó el test descrito.

Seguidamente, en la tabla 4.4 se presentan los resultados en cuanto a escritura/lectura desde el punto de vista del cliente python. Es importante incidir en que se simula el caso de un único PC (pocoyo) quien proporciona los servicios. Es decir, todas las peticiones de escritura y lectura se hacen a través de él. En el caso máximo, se han simulado 100 procesos al mismo tiempo de python leyendo y escribiendo datos aleatorios en el cluster. Este mismo cliente mide tiempos de escritura y de lectura por datapoint desde el punto de vista de un cliente. Es decir: hay un cliente de python que hace peticiones de escritura y de lectura y nos interesa observar el tiempo que se tardará en servir esos datos. Los resultados se presentan también en la figura 4.4 donde podemos observar que los tiempos de lectura y escritura aumentan linealmente (aproximadamente) con la cantidad de lecturas y escrituras por unidad de tiempo.

| Hora | Proceso | Número de procesos | Comentario |
|---|---|---|---|
| 11:25 | python test_core.py –read –write | 1 | – |
| 11:41 | python test_core.py –read –write | 10 | – |
| 12:25 | python test_core.py –read –write | 50 | – |
| 13:00 | python test_core.py –read –write | 100 | – |
| 13:20 | python test_core.py –read –write | 100 | Se mata un nodo del cluster. |
| 13:50 | python test_core.py –read –write | 100 | Se arranca de nuevo el nodo del cluster. |
| 14:27 | python test_core.py –read –write | 100 | Finaliza el test. |

**Tabla 4.3:** Pruebas de carga del cluster

**Tabla 4.4:** Tiempos de escritura y lectura en el cluster

| Número de procesos | Lectura (ms/datapoint) | Escritura (ms/datapoint) |
|---|---|---|
| 1 | 0.1193 | 0.5168 |
| 10 | 0.1846 | 0.8238 |
| 20 | 0.5303 | 25.8010 |
| 50 | 11.6210 | 13.4227 |
| 100 | 23.9680 | 57.2330 |

Como conclusión a este apartado hemos observado que el cluster de nodos de Cassandra aguanta perfectamente cargas de funcionamiento altas y cumple con las exigencias necesarias para utilizarlo en nuestra aplicación. Debe también tenerse en cuenta que los datos obtenidos están sujetos al tipo de hardware utilizado. Consideramos que los resultados se pueden mejorar bastante si sustituimos todos los soportes de almacenamiento magnético nodos por discos SSD.

Por otra parte hemos considerado que es vital que las máquinas dispongan de 4GB de RAM o más, ya que Cassandra consume una gran cantidad de esta memoria y hemos observado que con cantidades menores de memoria RAM los nodos de Cassandra se caían en condiciones de carga alta.

## 4.11. Conclusiones y aportaciones

Con lo dicho hasta ahora, y según las pruebas realizadas que se mostrarán en el apartado 4.9 Cassandra y CQL son unas herramientas más que adecuadas para tratar con grandes cantidades de datos crecientes en el tiempo.

No obstante y como comentario final, debemos entender que esto se consigue a costa de limitar en gran medida las búsquedas que se pueden realizar sobre los datos y, muchas veces, teniendo que duplicar tablas para poder hacer búsquedas por varias columnas al mismo tiempo.

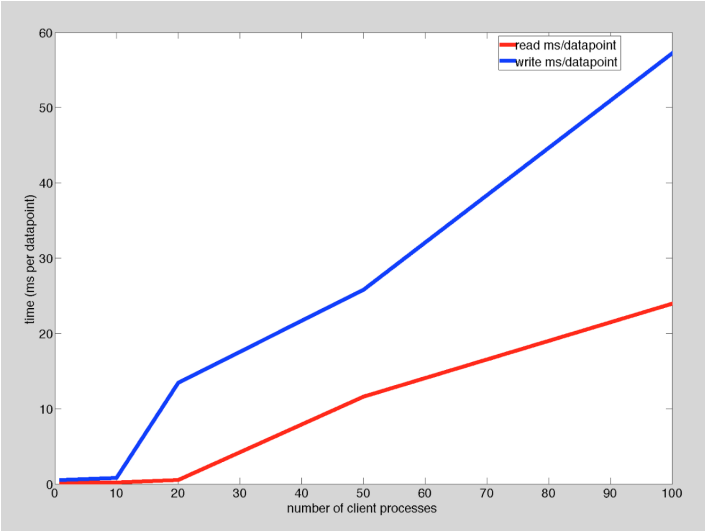



**Figura 4.4:** Se presentan los tiempos de escritura y lectura de un dato durante la prueba realizada.

Se presenta una nueva arquitectura diseñada para el almacenamiento de datos de tráfico.

El objetivo del sistema consiste en capturar los eventos de tráfico que publica la Dirección General de Tráfico de España.

Se han diseñado e implementado otros servicios típicos de arquitecturas *PaaS* (Platform as a Service) para visualizar los datos, descargarlos y analizarlos.

# 5 Estudio de accidentes de tráfico en España

## 5.1. Preámbulo

Se presentan, a continuación, dos estudios estadísticos extraídos de diferentes fuentes con la finalidad de estudiar los accidentes de tráfico que se han producido en España, entre 2011 y 2015. Para el primero de los estudios se utilizarán datos extraídos del '*Portal Estadístico de la Dirección General de Tráfico*' [36]. Para el segundo estudio se utilizarán datos recopilados de la infraestructura para Big Data descrita en el capítulo 4 que incorpora datos extraídos en tiempo real del portal Infocar [34] de la DGT.

Es importante destacar que las competencias en materia de tráfico en Cataluña y País Vasco se encuentran transferidas a las propias Comunidades Autónomas, por lo que la información correspondiente a las provincias de Barcelona, Girona, Lleida y Tarragona es facilitada directamente por la Generalidad de Cataluña, mientras que la información correspondiente a las provincias de Araba/Álava, Bizkaia y Gipuzkoa es facilitada por el Gobierno Vasco. Este detalle resulta de importancia debido a que se ha tenido que adaptar cierta información recopilada del portal Infocar por motivos del tratamiento que ambas Fuerzas y Cuerpos de Seguridad, Ertzaintza y Mossos d'Esquadra, proporcionaban a la información.

La motivación que ha llevado a utilizar distintas fuentes para realizar el análisis estadístico está relacionada con la necesidad de complementar el estudio con descriptores diferentes utilizados por ambas fuentes. Por ello, se han estudiado distintas metodologías estadísticas para encontrar un modelo que sea capaz de clasificar los accidentes de tráfico en base a su severidad o a la severidad de las lesiones de los ocupantes del vehículo.

Se comienza este estudio describiendo unas características de partida proporcionadas por el Gobierno de España, así como las tablas donde éstas se encuentran contenidas. Se proponen, a continuación, las fases fundamentales que se van a llevar a cabo para realizar dicho estudio:

- Fase 1: Acerca de la recopilación de los accidentes de tráfico en España.
- Fase 2: Procesamiento de datos relativos a accidentes de tráfico en España, donde se propondrá una agrupación de características con el fin de reducir el tiempo de cálculo, así como una unificación de tablas.
- Fase 3: Se extraerán nuevas características no contempladas por la DGT, que aunque no se unirán al set de datos total, se usarán para obtener conclusiones acerca de ciertas medidas recomendables a tomar en materia de seguridad vial.
- Fase 4: Se presenta una distribución de accidentes en función de determinadas variables tomadas a partir del portal Infocar [34] de la DGT en tiempo real.
- Fase 5: Se presenta un estudio de diferentes métodos de selección de características.
- Fase 6: Estimación de nuevos modelos de clasificación en función de los predictores seleccionados.
- Fase 7: Desarrollo de conclusiones y recomendaciones.

## 5.2. Recopilación y preprocesamiento de datos de España

En la siguiente sección se explicará el proceso de extracción y preprocesamiento de los datos que posterior se emplean para su estudio estadístico en la sección 5.3. Cabe destacar que este preprocesamiento se ha realizado sobre las dos fuentes antes reflejadas:

- datos estadísticos, entre 2011 y 2015, confeccionados por la DGT y publicados anualmente en su portal estadístico [36].
- datos de la infraestructura de Big Data basada en Cassandra e implementada con motivo de esta tesis doctoral y que incorpora datos extraídos en tiempo real del portal Infocar [34] de la DGT ente otras fuentes.

Es importante indicar que se han tenido que separar ambas investigaciones referentes a estas dos fuentes, la que se ha realizado a partir de los dataset extraídos de la DGT y aquellos que se han generado en la infraestructura de BigData, debido a que no parece haber forma de relacionar el mismo accidente entre las dos plataformas puesto que se emplean IDs distintos entre incidencias y accidentes.

### 5.2.1. Microdatos extraídos del portal estadístico de la DGT

La Dirección General de Tráfico es el órgano de la Administración Pública competente para la elaboración de la estadística de accidentes en base a lo establecido en los artículos 5 y 6 de la Ley sobre Tráfico, Circulación de Vehículos a Motor y Seguridad Vial [32].

La elaboración de la estadística de accidentes de tráfico con víctimas está regulada en la Orden Ministerial de Relaciones con las Cortes y de la Secretaría del Gobierno de 18 de febrero de 1993 [31], donde se especifican las definiciones y procedimientos que se deben aplicar.

La DGT, a través de su portal estadístico [36], almacena y publica anualmente datos estadísticos relativos a los accidentes con víctimas, a los fallecidos y/o heridos ocurridos en esos accidentes, así como aquellos datos relativos a los conductores implicados en los mismos.

En este estudio se ha optado por el análisis de los microdatos que se publican a principios de la anualidad siguiente a su ocurrencia. En la sección 5.2.1.1 se encuentra una descripción más detallada del procesamiento efectuado sobre estos datos que se encuentran almacenados en tres tablas diferenciadas por los datos relativos al accidente, los datos relativos a las víctimas y los datos relativos al vehículo.

#### 5.2.1.1. Agrupación de características

Las tablas de microdatos de accidentes con víctimas (TABLA_ACCVICT_XXX), vehículos (TABLA_VEHIC_XXXX) y personas (TABLA_PERS_XXXX), donde XXXX se refiere al año de ocurrencia de los accidentes, se relacionan entre sí de la siguiente manera:

- Accidentes con Víctimas y Vehículos: relación 1 a n, a través de la variable ID_ACCIDENTE.
- Vehículos y Personas: relación 1 a n, a través de 2 variables, ID_ACCIDENTE e ID_VEHICULO.
- Accidentes con Víctimas y Personas: relación 1 a n, a través de la variable ID_ACCIDENTE.

Esas relaciones se fundamentan en la estructura de la base de datos de accidentes con víctimas en la que para cada registro de la tabla *Accidentes* puede haber 1 o más vehículos implicados. Además para cada vehículo implicado puede haber 1 o más ocupantes del vehículo.

En algunas tablas, es necesario realizar una serie de operaciones para agrupar determinadas características que en el diccionario que realiza la DGT aparecen disociadas. En este diccionario, se relacionan las características, definidas en su mayoría como

factores, con unos determinados valores para cada nivel del factor. La motivación que nos lleva a agrupar estas características es reducir el número de columnas de nuestra matriz para que el cálculo en las ejecuciones sea más rápido. Por ello, a continuación, vamos a comenzar analizando las tablas en las siguientes secciones.

### Tabla Accidentes

En la tabla *Accidentes* correspondiente a los accidentes registrados anualmente por la DGT en su portal estadístico, se almacenan 39 variables descritas en la tabla 5.1.

| Característica |
|---|
| ID_ACCIDENTE |
| ANIO |
| MES |
| HORA |
| DIASEMANA |
| **PROVINCIA** |
| **COMUNIDAD_AUTONOMA** |
| **ISLA** |
| **COD_MUNICIPIO** |
| **MUNICIPIO** |
| **TOT_VICTIMAS** |
| **TOT_VICTIMAS30D** |
| **TOT_MUERTOS** |
| **TOT_MUERTOS30D** |
| **TOT_HERIDOS_GRAVES** |
| **TOT_HERIDOS_GRAVES30D** |
| **TOT_HERIDOS_LEVES** |
| **TOT_HERIDOS_LEVES30D** |
| TOT_VEHICULOS_IMPLICADOS |
| **ZONA** |
| ZONA_AGRUPADA |
| **CARRETERA** |
| **RED_CARRETERA** |
| TIPO_VIA |
| TRAZADO_NO_INTERSEC |
| TIPO_INTERSEC |
| PRIORIDAD_AGENTE |
| PRIORIDAD_SEMAFORO |
| PRIORIDAD_STOP |
| PRIORIDAD_CEDA |
| PRIORIDAD_MARCAS |
| PRIORIDAD_PASO |
| PRIORIDAD_OTRA |
| SUPERFICIE_CALZADA |
| LUMINOSIDAD |
| FACTORES_ATMOSFERICOS |
| VISIBILIDAD_RESTRINGIDA |
| **ACERAS** |
| TIPO_ACCIDENTE |

**Tabla 5.1:** Características nativas de TABLA_ACCVICT

A continuación, se eliminarán aquellas características que no se emplearán como predictores en esta investigación para realizar la clasificación de la gravedad de los accidentes. El motivo de su eliminación viene dado por la necesidad de reducción

del vector de descriptores para reducir tiempos de computación. En la tabla 5.1 se han resaltado en negrita aquellas características que se van a ir eliminando progresivamente.

En primer lugar, las características que se eliminarán serán las referentes a la geografía del accidente, ya que desde el enfoque escogido para esta tesis, no se basará en la localización del mismo, sino en otro tipo de características extrapolables a cualquier situación geográfica. En la tabla 5.2 se muestran las características en base a la posición geográfica que hemos eliminado.

| Característica | Descripción |
|---|---|
| COMUNIDAD_AUTONOMA | Indica la comunidad autónoma donde se produjo el accidente |
| ISLA | Indica si el accidente se produjo en la península o en una isla y en qué isla se produjo |
| COD_MUNICIPIO | Indica el código del municipio donde se produjo el accidente. El código de municipio está normalizado por el INE (5 dígitos alfanuméricos, 2 para provincia +3 para municipio). |
| MUNICIPIO | Indica el municipio donde se produjo el accidente indicando si éste posee menos o más de 5000 habitantes. |
| PROVINCIA | Indica la provincia donde se produjo el accidente |

**Tabla 5.2:** Características desechadas referentes a la posición geográfica del accidente

A continuación, se eliminarán aquellas variables referentes a la vía, donde se muestra una descripción en la tabla 5.3. El motivo de su eliminación viene dado porque no se pretende realizar una segmentación por vías o aceras en este primer estudio en la geografía española.

| Característica | Descripción |
|---|---|
| ZONA | Indica si se produjo en una carretera, una zona urbana, una travesía o una variante. Esta variable se eliminará debido a que existe otra variable que indica si el accidente se produjo en zona urbana o no: ZONA_AGRUPADA |
| CARRETERA | Indica la denominación de la vía donde se produjo el accidente |
| RED_CARRETERA | Indica la titularidad de la vía: Estatal, Autonómica, Provincial, Municipal u otra. |
| ACERAS | Indica si la vía posee o no acera. |

**Tabla 5.3:** Características desechadas referentes a la vía donde se produjo el accidente

Por último, se eliminarán aquellas variables que hacen referencia al número de víctimas, fallecidos y heridos graves o leves (ver tabla 5.4). El motivo de la eliminación de estas variables se debe a que los datos son globales por accidente, y no por persona, que sería lo deseable, ya que se desea clasificar a cada persona en función del resultado del accidente, es decir, si acabó accidentada grave o leve, o si falleció. Por todo ello, atenderemos a los datos que se encuentran en la tabla '*Personas*'.

Seguidamente, será necesario unificar todas las variables relativas a PRIORIDAD en una sola. La tabla 5.5 indica una descripción de dichas variables. Anteriormente a

| Característica | Descripción |
|---|---|
| TOT_VICTIMAS | Indica el número de víctimas a consecuencia del accidente en el periodo de las 24 horas siguientes a que aquel sucediera |
| TOT_VICTIMAS30D | Indica el número de víctimas a consecuencia del accidente en el periodo de los 30 días siguientes a que aquel sucediera |
| TOT_MUERTOS | Indica el número de personas fallecidas a consecuencia del accidente en el periodo de las 24 horas siguientes a que aquel sucediera |
| TOT_MUERTOS30D | Indica el número de personas fallecidas a consecuencia del accidente en el periodo de los 30 días siguientes a que aquel sucediera |
| TOT_HERIDOS_GRAVES | Indica el número de personas heridas grave a consecuencia del accidente en el periodo de las 24 horas siguientes a que aquel sucediera |
| TOT_HERIDOS_GRAVES30D | Indica el número de personas heridas graves a consecuencia del accidente en el periodo de los 30 días siguientes a que aquel sucediera |
| TOT_HERIDOS_LEVES | Indica el número de personas heridas leve a consecuencia del accidente en el periodo de las 24 horas siguientes a que aquel sucediera |
| TOT_HERIDOS_LEVES30D | Indica el número de personas heridas leve a consecuencia del accidente en el periodo de los 30 días siguientes a que aquel sucediera |

**Tabla 5.4:** Características desechadas referentes a las víctimas por accidente

unificarlas en una sola variable, debemos modificar el valor de todas ellas puesto que inicialmente poseen un valor de 1, por lo que no se distinguiría el tipo de prioridad al unificarlas en una variable. Adicionalmente, es necesario observar que ninguna de ellas puede cumplirse al mismo tiempo, tal y como indica la tabla 5.5. La misma tabla indica el valor anterior y el nuevo de todas las variables.

| Característica | Descripción | Valor anterior | Nuevo valor en la característica 'prioridad' |
|---|---|---|---|
| PRIORIDAD_AGENTE | La prioridad la indica un agente de tráfico | 1 | 1 |
| PRIORIDAD_SEMAFORO | La prioridad la indica un semáforo | 1 | 2 |
| PRIORIDAD_STOP | La prioridad la indica una señal de Stop | 1 | 3 |
| PRIORIDAD_CEDA | La prioridad la indica una señal de 'Ceda el paso' | 1 | 4 |
| PRIORIDAD_MARCAS | Sólo existen marcas viales | 1 | 5 |
| PRIORIDAD_PASO | Existe un paso para peatones | 1 | 6 |
| PRIORIDAD_OTRA | La prioridad la indica otra señal | 1 | 7 |

**Tabla 5.5:** Modificación de valores de características relativas a prioridad en la calzada

En la tabla 5.6 también se indica cómo quedarán los valores tabulados para la nueva característica. Seguidamente, se unirán todas ellas en una sola característica, a la que se ha denominado PRIORIDAD, asignando el valor de cada una de ellas cuando

| Nueva_Columna | PRIORIDAD_SEMAFORO=2 | PRIORIDAD_STOP=3 |
|---|---|---|
| 2 | 2 | NA |
| NA | NA | NA |
| 3 | NA | 3 |

**Tabla 5.6:** Valores tabulados que formarán la nueva característica de prioridad

sea distinto de NA, tal y como indica la tabla 5.6 con un ejemplo.

### Tabla Vehículos

En lo referente a la tabla '*Vehículos*' (5.7), la DGT almacena en esta tabla un total de 14 características.

| Característica |
|---|
| ID_ACCIDENTE |
| ID_VEHICULO |
| ANIO_MATRICULA_VEHICULO |
| **MES_MATRICULA_VEHICULO** |
| TIPO_VEHICULO |
| ANOMALIA_NINGUNA |
| **ANOMALIA_NEUMATICO** |
| **ANOMALIA_REVENTON** |
| **ANOMALIA_DIRECCION** |
| **ANOMALIA_FRENOS** |
| NUMERO_OCUPANTES_VEH |
| **MERCANCIAS_PELIGROSAS** |
| **VEHICULO_INCENDIADO** |
| ANIO |

**Tabla 5.7:** Variables nativas almacenadas en la tabla Vehículos

Lo primero que se ha realizado ha sido estudiar la posibilidad de unificar los tipos de anomalías en uno solo, sin embargo, tras su minucioso estudio, se ha identificado que en la variable 'ANOMALIA_NINGUNA' ya se indicaba si el vehículo sufría o no alguna anomalía.

Puesto que el objetivo será eliminar todos los registros de accidente que poseen alguna anomalía, ya que de lo contrario se realizaría un sesgo respecto al resto de accidentes, se procederá a eliminar todos aquellos registros cuyo valor de ANOMALIA_-NINGUNA=2.

En la tabla 5.7 se ha resaltado en negrita aquellas características que se eliminarán de aquí en adelante.

Finalmente, se eliminarán aquellas variables relacionadas con el resto de las anomalías (véanse tablas de 5.8 a 5.12).

El siguiente paso realizado consistirá en eliminar las características que no consideramos necesarias a usar como predictores. En la tabla 5.13 se realiza una breve descripción de todas ellas.

| Característica | Estado del vehículo | Valor | Registros |
|---|---|---|---|
| Anomalía dirección | Anomalías previas en la dirección | 1 | 44 |
| | Se desconoce/Sin especificar | 2 | 148.791 |
| | Sin anomalías o con otras diferentes a las anteriores | 3 | 21.914 |

**Tabla 5.8:** Descripción de la característica referente a anomalías en la dirección

| Característica | Estado del vehículo | Valor | Registros |
|---|---|---|---|
| Anomalía frenos | Anomalías previas en los frenos | 1 | 181 |
| | Se desconoce/Sin especificar | 2 | 148.654 |
| | Sin anomalías o con otras diferentes a las anteriores | 3 | 21.914 |

**Tabla 5.9:** Descripción de la característica referente a anomalías en los frenos

| Característica | Estado del vehículo | Valor | Registros |
|---|---|---|---|
| Anomalía neumáticos | Neumáticos muy desgastados o defectuosos | 1 | 557 |
| | Se desconoce/Sin especificar | 2 | 148.278 |
| | Sin anomalías o con otras diferentes a las anteriores | 3 | 21.914 |

**Tabla 5.10:** Descripción de la característica referente a anomalías en los neumáticos

| Característica | Estado del vehículo | Valor | Registros |
|---|---|---|---|
| Anomalía reventón | Neumáticos muy desgastados o defectuosos | 1 | 185 |
| | Se desconoce/Sin especificar | 2 | 148.650 |
| | Sin anomalías o con otras diferentes a las anteriores | 3 | 21.914 |

**Tabla 5.11:** Descripción de la característica referente a anomalías en los neumáticos respecto a pinchazos

| Característica | Estado del vehículo | Valor | Registros |
|---|---|---|---|
| Anomalía ninguna | Se desconoce/Sin especificar | 1 | 145.889 |
| | Alguna anomalía detectada | 2 | 2943 |
| | Sin anomalías o con otras diferentes a las anteriores | 3 | 21.917 |

**Tabla 5.12:** Descripción de la característica referente a anomalías detectadas

| Característica | Descripción |
|---|---|
| MES_MATRICULA_VEHICULO | Indica el mes en el que se matriculó el vehículo. |
| VEHICULO_INCENDIADO | Indica si el vehículo se incendió en el accidente |
| MERCANCIAS_PELIGROSAS | Indica si el vehículo accidentado transportaba mercancías peligrosas |

**Tabla 5.13:** Características a eliminar de la tabla Vehículos

Seguidamente, se ha calculado la antigüedad del vehículo. El cálculo ha sido posible realizarlo a través de una variable denominada ANIO_MATRICULA_VEHICULO. Puesto que es posible que un coche se matricule varias veces y ello conlleve un sesgo en los datos, se espera que debido a la cantidad de accidentes registrados de los que disponemos, tal sesgo sea prácticamente inapreciable.

Los nuevos valores de antigüedad del vehículo se han insertado en la variable 'EDAD_VEHICULO'.

### Tabla Personas

La tabla '*Personas*' (tabla 5.14) describe todos los datos relativos a los accidentados en el siniestro, independientemente de que fueran conductores, pasajeros o peatones.

La DGT almacena un total de 31 características referentes a los accidentados, que son las siguientes:

| Característica |
|---|
| ID_ACCIDENTE |
| ID_VEHICULO |
| ID_PERSONA |
| **ID_CONDUCTOR** |
| **ID_PASAJERO** |
| **ID_PEATON** |
| EDAD |
| SEXO |
| **ANIO_PERMISO** |
| POSICION |
| **USO_CINTURON** |
| **USO_SRI** |
| **USO_CASCO** |
| **MUERTO_24H** |
| MUERTO_30D |
| **HERIDO_GRAVE_24H** |
| HERIDO_GRAVE30D |
| **HERIDO_LEVE_24H** |
| HERIDO_LEVE30D |
| **DICCIONARIO_MANIOBRAS** |
| MANIOBRAS |
| **INFRACC_VELOCIDAD** |
| **INFRACC_COND** |
| **INFRACC_APERTURA** |
| **INFRACC_ALUMBRADO** |
| **INFRACC_CARGA_VEHICULO** |
| **INFRACC_RESUMEN** |
| **INFRACC_PEATON** |
| **DICCIONARIO_ACCION_PEATON** |
| ACCION_PEATON |
| ANIO |

**Tabla 5.14:** Características nativas de la tabla Personas

Puesto que el estudio que se presenta en esta tesis doctoral está directamente relacionado con lo que le suceda a los ocupantes de los vehículos, no se tendrá en

cuenta lo que le ocurra a los peatones, ya que no serán objeto de estudio en este trabajo.

Por ello, se eliminarán todos los registros relativos a este colectivo de los accidentes. En la tabla 5.15 se enumeran y describen las características que se van a eliminar. Es necesario observar que se eliminarán aquellas relativas a las secuencias de identificación de cada persona, debido a que existen otras características que nos aportan esa información.

| Característica | Descripción |
|---|---|
| ID_CONDUCTOR | Indica un número secuencial de conductor en el accidente o por el contrario indica si no se trataba del conductor sino de un pasajero o peatón. |
| ID_PASAJERO | Indica un número secuencial de pasajero en el accidente o por el contrario indica si no se trataba de un pasajero sino de un conductor o peatón. |
| ID_PEATON | Indica un número secuencial de peatón en el accidente o por el contrario indica si no se trataba de un peatón sino de un pasajero o conductor. |

**Tabla 5.15:** Características a eliminar de la tabla Personas relativas a secuencias de identificación de los miembros involucrados en un accidente

Otra de las características que se eliminará será el año del permiso del conductor, por no poseer suficiente información en el set de datos. Aunque se considera una información importante para este estudio, en la mayoría de los casos no aparece la fecha del permiso de conducir junto al conductor.

Adicionalmente, las características relativas a diccionarios también se han eliminado, tal y como se puede comprobar en la tabla 5.16. El motivo de su eliminación viene dado porque únicamente describían el tipo de maniobras o de acciones del peatón, lo que no aportaba dato alguno para el vector descriptor del accidente.

| Característica | Descripción |
|---|---|
| ANIO_PERMISO | Indica el año en el que el conductor del vehículo obtuvo el permiso de conducir |
| DICCIONARIO_MANIOBRAS | Describe el tipo de maniobra |
| DICCIONARIO_ACCION_PEATON | Describe los tipos de acciones que puede llevar a cabo un peatón |

**Tabla 5.16:** Características a eliminar de la tabla Personas relativas diccionarios

Por otro lado, las características relativas a las infracciones cometidas por los conductores, pasajeros o peatones, tampoco son importantes debido a que el número registrado en el set de datos es muy pobre. Por ello, se eliminarán las características indicadas en la tabla 5.17.

Será necesario generar las clases a predecir o a clasificar donde, como se ha comentado en secciones anteriores, consistirá en las consecuencias para los ocupantes de cada vehículo, es decir, si la persona falleció, fue herida gravemente, fue herida de levedad o resultó ilesa.

| Característica | Descripción |
|---|---|
| INFRACC_VELOCIDAD | Indica si el conductor cometió una infracción que produjo el accidente por velocidad inadecuada, bien sea porque sobrepasó la velocidad o por marcha lenta. |
| INFRACC_COND | Indica si se realizó una conducción distraída o desatenta |
| INFRACC_APERTURA | Describe los tipos de acciones que puede llevar a cabo un peatón |
| INFRACC_ALUMBRADO | El conductor cometió una infracción por incorrecta utilización del alumbrado |
| INFRACC_CARGA_VEHICULO | Indica si se cometió una infracción al cargar el vehículo |
| INFRACC_RESUMEN | Indica un resumen de todas las infracciones |
| INFRACC_PEATON | Indica si el peatón cometió una infracción por no respetar semáforo, agente, no utilizar paso de peatones, por marchar por la calzada o por el arcén de forma antirreglamentaria o por subir o bajar de un vehículo de forma no reglamentaria. |

**Tabla 5.17:** Características a eliminar de la tabla Personas relativas a infracciones

Por ello, en lo que concierne a fallecidos, heridos grave o leve, no se distinguirá en la referencia temporal de cuándo ocurrió, es decir, si fue en las 24 primeras horas transcurrido el accidente o en los 30 días siguientes. Puesto que todos los resultados de los accidentes relativos al desenlace del mismo se encuentran incluidos en las características a 30 días, usaremos éstas para organizar la variable. Por tanto, las variables que aparecen en la tabla 5.18 se eliminarán.

| Característica | Descripción |
|---|---|
| MUERTO_24H | Indica si el accidentado falleció en las 24h siguientes a producirse el accidente |
| HERIDO_GRAVE_24H | Indica si las heridas causadas por el accidente tuvieron como consecuencia un desenlace grave para el conductor en las 24h siguientes a producirse el accidente |
| HERIDO_LEVE_24H | Indica si las heridas causadas por el accidente tuvieron como consecuencia un desenlace leve para el conductor en las 24h siguientes a producirse el accidente |

**Tabla 5.18:** Características a eliminar de la tabla Personas relativas a las consecuencias de las lesiones para los ocupantes de los vehículos

Con respecto a los accesorios de seguridad, puesto que se ha comprobado que existen muy pocos registros de si se empleaban o no en los accidentes, se ha decidido eliminar las características relativas a ellos, tal y como se puede comprobar en la tabla 5.19.

Para finalizar, se ha analizado la columna POSICION con el fin de detectar cuáles de estos registros de accidentes eran peatones. En la tabla 5.20 podemos encontrar una descripción de todos los valores que puede tomar este predictor, y donde eliminaremos aquellos que poseen valor de 99.

| Característica | Descripción |
|---|---|
| USO_CINTURON | Indica si el accidentado llevaba o no el cinturón |
| USO_SRI | Indica si el bebé o menor llevaba los sistemas de retención infantil correctamente abrochados |
| USO_CASCO | Indica si el conductor de motocicleta portaba o no casco |

**Tabla 5.19:** Características a eliminar de la tabla Personas relativas a los accesorios de seguridad

| Característica | Descripción |
|---|---|
| POSICION | Indica qué posición ocupaba el accidentado en el vehículo |
| | 1 CONDUCTOR VEHÍCULO<br>2 PASAJERO DELANTERO<br>3 PASAJERO TRASERO IZQUIERDO<br>4 PASAJERO TRASERO DERECHO<br>5 PASAJERO TRASERO CENTRAL<br>6 CONDUCTOR VEHÍCULO DE DOS RUEDAS<br>7 PASAJERO VEHÍCULO DE DOS RUEDAS<br>8 OTROS PASAJEROS SENTADOS<br>9 OTROS PASAJEROS DE PIE<br>99 NO APLICA (PEATONES) |

**Tabla 5.20:** Posición del accidentado en el vehículo

### Fusión de tablas por año

El siguiente paso consistirá en fusionar las tres tablas cada año desde 2011 a 2015, ambos inclusive. Para ello, se llevarán a cabo los siguientes pasos:

1. Se unirán las columnas *Accident_Index* y *Vehicle_Reference* en una sola.
2. Se realizará un *inner_join* entre las tablas *casualties* y *vehicles*.
3. Se separará la columna 1 de *vehicles.casualties* en dos columnas.
4. Se realizará un *inner_join* entre *vehicles.casualties* y *accidents*. Previamente, se convertirá en carácter el *ID_ACCIDENTE* para poder comparar.
5. Se eliminarán algunas columnas que se repiten en ambas tablas.
6. Se almacenará cada tabla resultado de fusionar las otras tres por año.

Se obtendrán como resultado 5 datasets, uno por año, tal y como se puede observar en la tabla 5.21.

### Unificación de tablas periodo 2011 a 2015

Puesto que es deseable utilizar una única tabla de datos para trabajar con los diversos algoritmos de clasificación que emplearemos en sucesivos capítulos, se unificarán todos los set de datos detallados en la tabla 5.21 en una sola. Para ello, se concatenarán todas las tablas por año.

| Dataset | Descripción |
|---|---|
| vehicles.casualties.accidents_2011 | Dataset relativo a las tablas fusionadas de vehículos, personas y accidentes registradas en 2011 |
| vehicles.casualties.accidents_2012 | Dataset relativo a las tablas fusionadas de vehículos, personas y accidentes registradas en 2012 |
| vehicles.casualties.accidents_2013 | Dataset relativo a las tablas fusionadas de vehículos, personas y accidentes registradas en 2013 |
| vehicles.casualties.accidents_2014 | Dataset relativo a las tablas fusionadas de vehículos, personas y accidentes registradas en 2014 |
| vehicles.casualties.accidents_2015 | Dataset relativo a las tablas fusionadas de vehículos, personas y accidentes registradas en 2015 |

**Tabla 5.21:** Datasets parciales por año construidos a partir de las tablas de vehículo, personas y accidentes

Una vez concatenadas las tablas, el siguiente paso consistirá en estudiar la frecuencia absoluta de cada variable en tanto por cien. Mediante este parámetro se podrán conocer todos los valores que puede tomar cada una de ellas y eliminar las que posean un alto índice de campos sin dato, con NA o valores anómalos.

A través de este estudio, se observa que es necesario eliminar las siguientes características expuestas en la tabla 5.22 al no disponer de los suficientes datos para aportar al mismo.

| Característica | Descripción |
|---|---|
| VISIBILIDAD_RESTRINGIDA | Indica cómo era la visibilidad en el momento del accidente |
| EDAD_VEHICULO | Indica la edad del vehículo |
| ACCION_PEATON | Indica qué acción estaba llevando a cabo el peatón en el momento del accidente |

**Tabla 5.22:** Características a eliminar por no disponer de información suficiente

Adicionalmente, y extraído también del estudio anterior, se eliminarán aquellas filas con valores anómalos de las características de la tabla 5.23.

| Característica | Descripción |
|---|---|
| EDAD | Posee valores anómalos: ej. 999 |
| NUMERO_OCUPANTES_VEH | Posee valores anómalos: ej. 9999 |
| SEXO | Posee valores anómalos: ej. 999 |
| SUPERFICIE_CALZADA | Posee valores anómalos: ej. 999 |
| FACTORES_ATMOSFERICOS | Posee valores anómalos: ej. 999 |

**Tabla 5.23:** Variables con valores anómalos

A continuación, se unificarán las variables TRAZADO_NO_INTERSEC y TIPO_INTERSEC en una única variable, que tomará el nombre de la segunda de ellas, para estudiar todo tipo de trazados. Para ello, inicialmente se modificará el diccionario de valores para la variable TIPO_INTERSEC quedando como indica la tabla 5.24, y con el

fin que no interfiera sobre los valores de la característica TRAZADO_NO_INTERSEC que posee los valores indicados en la tabla 5.25.

| Característica | Descripción | Valor Anterior | Nuevo valor en la característica 'prioridad' |
|---|---|---|---|
| TIPO_INTERSEC | Intersección en T ó Y | 1 | 6 |
| | Intersección en X ó + | 2 | 7 |
| | Enlace de entrada | 3 | 8 |
| | Enlace de salida | 4 | 9 |
| | Giratoria | 5 | 10 |
| | Otro tipo de intersección | 6 | 11 |
| | No es una intersección | 999 | 999 |

**Tabla 5.24:** Modificación de valores de características relativas a tipos de intersección

| Característica | Descripción | Valor Anterior |
|---|---|---|
| TRAZADO_NO_INTERSEC | Tramo en recta | 1 |
| | Tramo en curva suave | 2 |
| | Tramo en curva fuerte sin señalizar | 3 |
| | Tramo en curva fuerte con señal y sin velocidad señalizada | 4 |
| | Tramo en curva fuerte con señal y velocidad señalizada | 5 |

**Tabla 5.25:** Descripción de los posibles valores de la característica TRAZADO_NO_INTERSEC

Finalmente, modificaremos el nombre de esta característica pasando a denominarse 'INTERSECCION'. Los valores que puede tomar esta característica se recogen en la tabla 5.26.

| Característica | Descripción | Valor Anterior |
|---|---|---|
| INTERSECCION | Tramo en recta | 1 |
| | Tramo en curva suave | 2 |
| | Tramo en curva fuerte sin señalizar | 3 |
| | Tramo en curva fuerte con señal y sin velocidad señalizada | 4 |
| | Tramo en curva fuerte con señal y velocidad señalizada | 5 |
| | Intersección en T ó Y | 6 |
| | Intersección en X ó + | 7 |
| | Enlace de entrada | 8 |
| | Enlace de salida | 9 |
| | Giratoria | 10 |
| | Otro tipo de intersección | 11 |

**Tabla 5.26:** Descripción de los posibles valores de la característica INTERSECCION

### 5.2.2. Microdatos integrados mediante Cassandra desde Infocar

A pesar que la DGT publica numerosas variables relacionadas con el accidente, el vehículo y las personas heridas a través de su portal estadístico, y a pesar de que estas características pueden ayudarnos a entender algunas de las causas involucradas en el

accidente, sin embargo, creemos que falta diversa información importante que, sin duda, ayudará a los investigadores a mejorar sus estudios. Entre esta información necesaria, cabría proporcionar la ubicación geográfica exacta de cada accidente ocurrido. Ello ayudaría a obtener otras características derivadas de su situación.

Para solucionar este y otros problemas, comenzamos a estudiar alternativas que complementaran los datos obtenidos a través del portal estadístico de la DGT. Para ello, nuestro enfoque se basa en la indexación de datos en tiempo real de las incidencias de tráfico del portal web *Infocar DGT* [34]. Se ha incorporado a este trabajo de investigación los datos extraídos durante el un periodo comprendido entre finales del año 2015 y mediados del año 2016.

La extracción de datos se realiza secuencialmente a través de web *scraping* de todas las incidencias registradas y, a continuación, todas ellas se van almacenando en el cluster de base de datos Cassandra, tal y como se describió en el capítulo 4.

Adicionalmente, se han almacenado registros de accidentes que se han producido en tramos de carreteras en obras, así como la ubicación geográfica de todos los radares de control de velocidad fijos y móviles, con la finalidad de observar su impacto en la seguridad vial. El resto de datos relativos a los accidentes y otras incidencias que ocurrían en tiempo real también se han almacenado. En la tabla 5.27 se muestran algunos de los datos generales que actualmente se indexan a partir de las incidencias publicadas por la DGT en tiempo real. Como se mencionó, se almacenan en nuestro cluster cada cinco minutos.

| Características indexadas | |
|---|---|
| Causa | Tipo |
| Subtipo | ID de la incidencia |
| Ciudad | Provincia |
| Pto. Km inicial | Pto. Km final |
| Latitud | Longitud |
| Nivel de severidad | Descripción |
| Marca temporal | Dirección (carril) |

**Tabla 5.27:** Características publicadas de cada incidencia por la DGT

#### 5.2.2.1. Extracción de nuevas características no contempladas por la DGT para el estudio de los accidentes de tráfico

Después de la investigación preliminar descrita anteriormente a través de los datasets que la DGT publica en abierto a través de su portal estadístico, a continuación, se definen características más complejas que pueden derivarse a partir de las vistas en la tabla 5.27.

Como se ha comentado, estas características son nuevas y no se han empleado antes en estudios de tráfico, tal y como hemos podido comprobar en el estado del arte. Ello nos permiten analizar el problema a través de un nuevo modelo de análisis avanzado (figura 5.1).

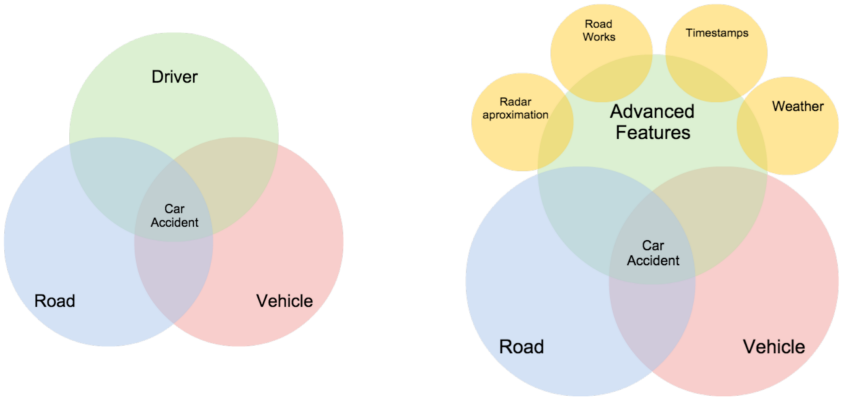



**Figura 5.1:** Análisis clásico del accidente de tráfico frente al análisis avanzado propuesto

#### 5.2.2.2. Análisis de accidentes ocurridos en tramos en obras

Una de las aportaciones de esta investigación consiste en analizar si un accidente ha ocurrido entre dos puntos, inicial y final, de un tramo en obras. Para este análisis, todas las incidencias indexadas como '*roadwork*' se verifican con cada accidente indexado en *Cassandra* para el mismo tramo en el momento en el que se mantiene esa incidencia abierta. Se han programado algunos métodos en *Python* para realizar tales comprobaciones. Se genera una nueva variable en el descriptor del accidente donde se indica si se ha producido o no en un tramo en obras. El nombre de la variable será '*OBRA*'.

#### 5.2.2.3. Análisis de accidentes ocurridos en tramos cercanos a un radar de control de velocidad

Debido a nuestra experiencia personal, determinadas situaciones peligrosas se pueden producir en la proximidad a un radar de control de velocidad, debido a la fuerte desaceleración en cadena de varios vehículos que se aproximan al mismo cuando circulan a una velocidad superior a la permitida.

Para averiguar si el accidente se produjo en la vecindad de un radar, la situación es algo más compleja que la detección de un accidente en un tramo en obras, ya que es necesario calcular la distancia entre dos puntos en una esfera, la Tierra. Para realizar el cálculo, primero se comprobará que el accidente ocurrió en la misma vía donde se encuentra el radar de control de velocidad. A continuación, la distancia entre ellos se calcula a través de la fórmula de distancia euclídea de Haversine (ecuación 5.1), donde $\phi_1$, $\phi_2$ y $\lambda_1$, $\lambda_2$ son la latitud y longitud, ambas expresadas en radianes, de los dos puntos, respectivamente, y $r$ es igual al radio medio de la Tierra. Una vez que conocemos la distancia $d$ entre los dos puntos, se comprueba que esta distancia es inferior a 500 metros antes y después del radar de control de velocidad.

$$d = 2r \cdot \arcsin(\sqrt{\sin^2(\frac{\phi_2 - \phi_1}{2}) + \cos(\phi_1) \cdot \cos(\phi_2) \cdot \sin^2(\frac{\lambda_2 - \lambda_1}{2})}) \quad (5.1)$$

#### 5.2.2.4. Extracción de características de meteorología avanzadas

Debido a que no poseemos datos meteorológicos históricos, si se desea obtener datos concretos de algún parámetro meteorológico es necesario acudir a algún proveedor que nos los proporcione. Para el caso concreto de esta investigación, se decidió acudir a la empresa *Wunderground* [108], la cual posee una API muy completa para la extracción de datos históricos de estaciones meteorológicas dadas la longitud y la latitud, junto con la marca de tiempo de ocurrencia de cada accidente. El mencionado servicio web devuelve los datos relativos a la estación meteorológica más cercana y podemos agregar algunos datos interesantes a nuestro vector de características:

- Presión en mBar
- Temperatura en Celsius
- Humedad Relativa en %
- Velocidad del viento en Km/h
- Dirección del viento (SW, NNE, etc.)
- Visibilidad
- Precipitación en mm

#### 5.2.2.5. Set de datos analizado con características avanzadas

El conjunto de datos empleado contiene 5.157 registros de accidentes de mayo de 2015 a enero de 2016; cada tabla incluye los datos de la carretera y la ubicación geográfica del accidente, los informes meteorológicos y los datos relativos al radar de control de velocidad más cercano o su situación dentro de un tramo en obras.

Adicionalmente, se han analizado 17.573 incidentes en el mismo periodo en que se encontró un tramo en obras en la red de carreteras española y 1.247 radares de control de velocidad localizados por toda la red.

#### 5.2.2.6. Conclusiones de este estudio con características avanzadas

Se ha considerado interesante mostrar algunas de las distribuciones que se han obtenido del análisis preliminar de datos que hemos efectuado. Este estudio ha analizado el número de accidentes de tráfico extraídos del conjunto de datos de 5.176 accidentes.

A continuación, se presentan resultados absolutos que clasifican las provincias en función del número de accidentes, así como el día de la semana y la hora del día.

### Distribución de accidentes por vía

Si se entra en detalles acerca de la distribución de accidentes por vía, tal y como se puede observar en la figura 5.2, es relevante observar que la autovía de circunvalación SE-30, ubicada en Sevilla y de tan solo 22 km de longitud total, es la vía con más accidentes. Resulta relevante observar también que Sevilla es la tercera provincia española con mayor índice de accidentes. Para contextualizar esta vía, es significativo indicar que el tráfico en ella es muy denso, según la DGT, ya que es la principal vía de un área metropolitana de más de 1.500.000 habitantes.

| Nivel de factor | Descripción |
|---|---|
| 1 | VÍAS INTERURBANAS |
| 2 | VÍAS URBANAS |

**Tabla 5.28:** Diccionario del predictor *zona agrupada*. Extraído de la Dirección General de Tráfico

A esta vía, le sigue a la Autovía del Nordeste o A-2, que une Madrid con Barcelona, con un 17 % menos de accidentes, y con una longitud total de la vía de 780 km.

A la Autovía A-2 le sigue otra autovía, la del Mediterráneo o A-7, con un 29 % menos de accidentes que la SE-30, pero con una longitud total de 1.329 km.

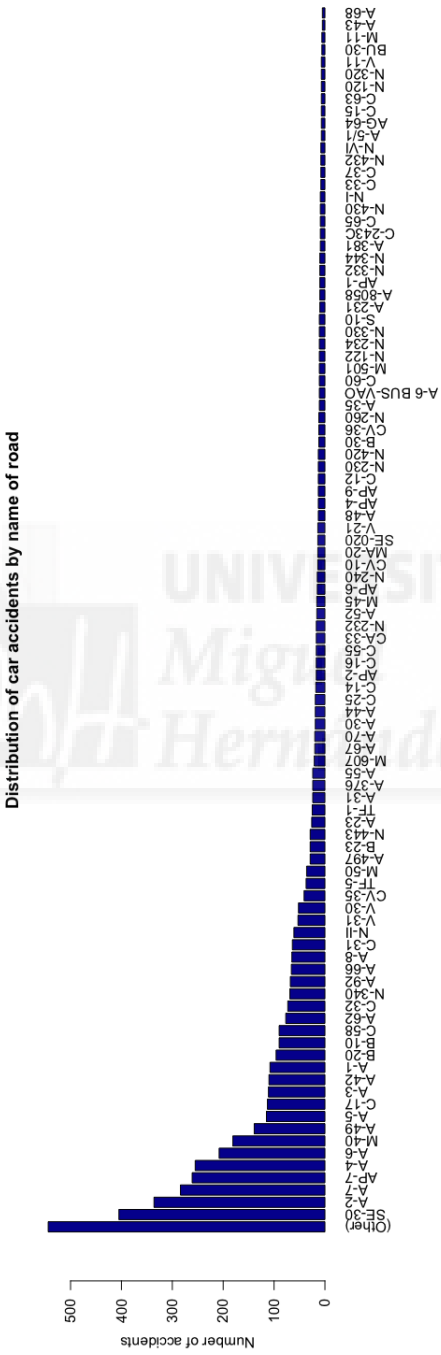



**Figura 5.2:** Distribución de accidentes por vía entre mayo de 2015 y enero de 2016

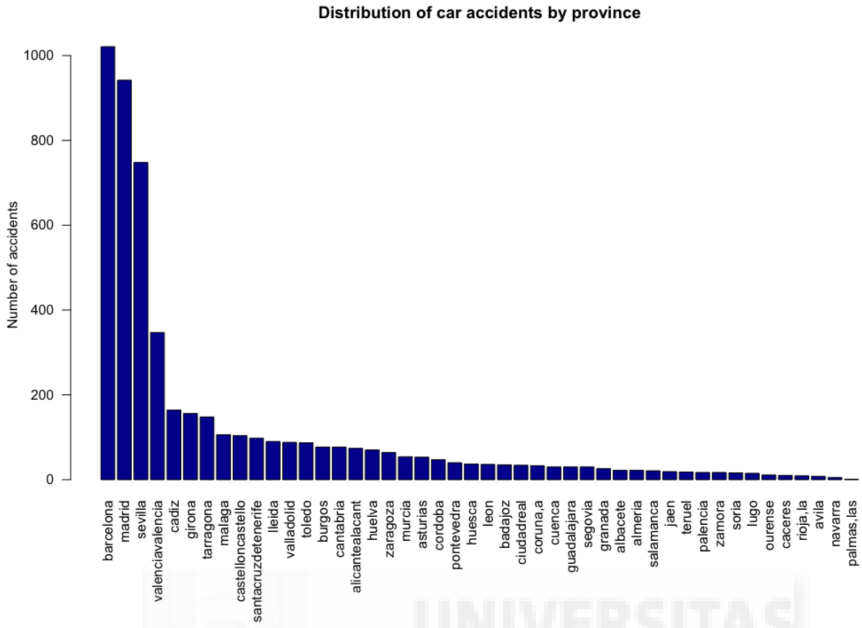



**Figura 5.3:** Distribución de accidentes por provincia entre mayo de 2015 y enero de 2016

En la figura 5.28 se pueden observar los niveles de factor con los que se cuenta en el portal estadístico de la DGT.

### Distribución de accidentes por provincia

Cuando se inició esta investigación, se esperaba encontrar un mayor número de accidentes en provincias con mayores índices de población. Efectivamente, esta suposición se ha confirmado en provincias como Barcelona y Madrid (ver figura 5.4), ambas con el mayor índice de accidentes en todo el territorio español. Sin embargo, parece alarmante el volumen de accidentes que se produjeron en la provincia de Sevilla, en comparación con una provincia de extensión similar como Valencia, que sin embargo es la siguiente en el ranking de accidentes por provincia (ver figura 5.3). En Valencia se producen un 50 % menos accidentes que Sevilla.

### Distribución de accidentes por día de la semana y hora del día

En cuanto a la distribución de accidentes por día de la semana y hora del día, en las figuras 5.5 y 5.6 se pueden observar los accidentes de tráfico acumulados durante todo el período estudiado, ordenados por el día de la semana en que ocurrieron, en función del volumen de accidentes. También se puede observar la distribución de accidentes en función de la hora del día en que ocurrieron. Esto nos permite definir una

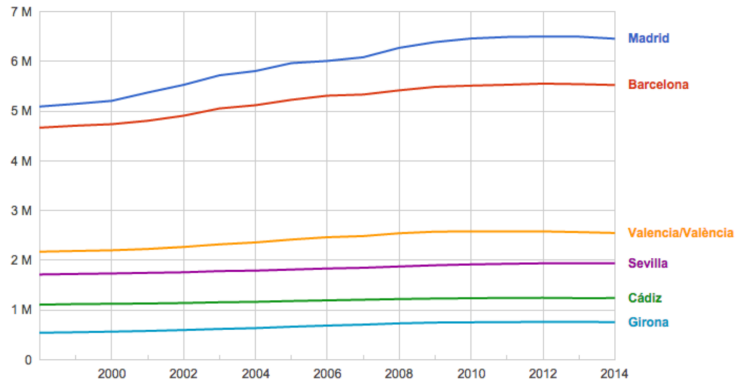



**Figura 5.4:** Número de habitantes por provincia y año

probabilidad total *a priori* de ocurrencia de un accidente en cualquier vía de España dado el día de la semana.

La conclusión principal que podemos observar es que los sábados y los domingos el número de accidentes de tráfico es significativamente bajo. Es notable también que la distribución de accidentes es similar de lunes a jueves, sin embargo, aumenta los viernes, día en el que más accidentes ocurren.

En cuanto a las horas a las que ocurren, se pueden observar tres picos principales:

- entre las 8:00 y las 9:00h,
- entre las 14:00 y las 15:00h,
- entre las 19:00 y las 20:00h.

Esta distribución se puede explicar fácilmente debido a las horas donde se producen mayores desplazamientos con motivo del comienzo o finalización de la jornada laboral, que generalmente coincide con estas horas punta.

### Distribución de accidentes por temperatura

En cuanto a la distribución de accidentes por temperatura, la distribución no es del todo representativa. La figura 5.7 representa el número total de accidentes en un histograma de temperaturas.

Este gráfico presenta una distribución de datos Gaussiana, centrada en 19 grados Celsius. Este resultado preliminar se utilizará como punto de partida para analizar y confirmar la dependencia del clima cálido con la ocurrencia de accidentes, tal y como se ha reflejado en la sección 3.2.1 del capítulo 3.

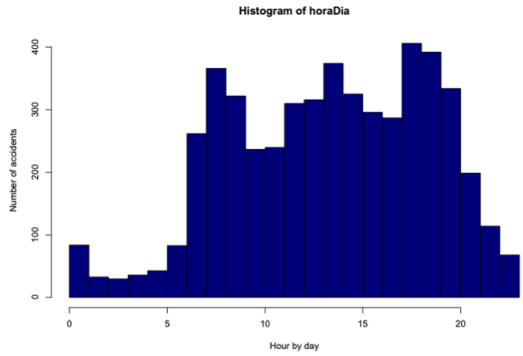



**Figura 5.5:** Distribución de accidentes en función de la hora del día en que ocurrieron entre mayo de 2015 y enero de 2016

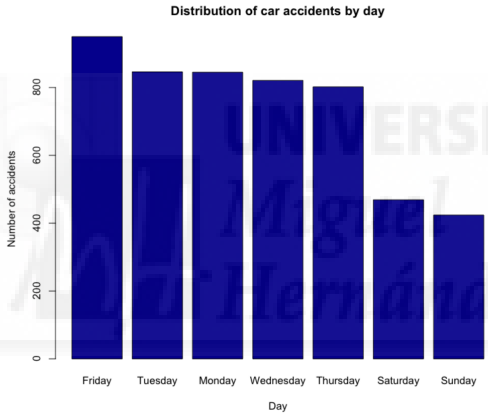



**Figura 5.6:** Distribución de accidentes por día de la semana entre mayo de 2015 y enero de 2016

### Número de accidentes producidos en la cercanía de un radar de control de velocidad

En esta sección se presenta una estadística que representa el número de accidentes producidos en la cercanía de un radar de control de velocidad. Se pretende hacer notar que cabe la posibilidad que la ubicación de determinados radares de control se deba poner en cuestión como una medida de seguridad, con el fin de reducir los accidentes de tráfico mediante el ajuste de la velocidad de los conductores al máximo establecido en la vía.

Nuestra experiencia personal sugiere que los conductores tienden a ajustar bruscamente su velocidad cuando se acercan a estos radares y este hecho puede producir situaciones de riesgo. Para analizar estas situaciones y por ello la suposición de la

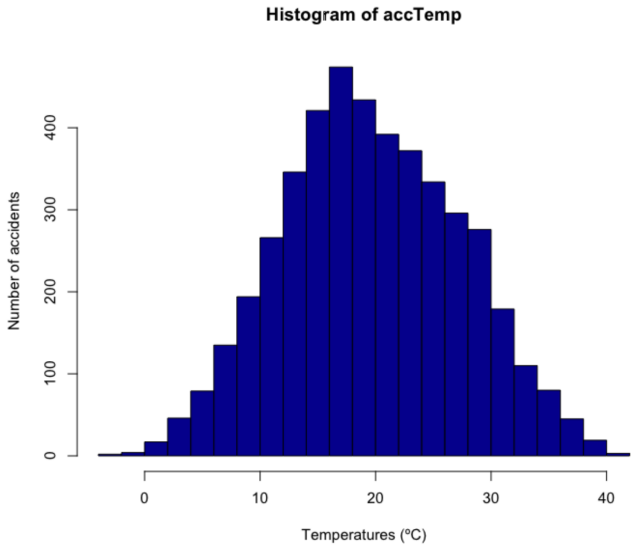



**Figura 5.7:** Distribución de accidentes por temperatura entre mayo de 2015 y enero de 2016

que partíamos, se han tomado dos distancias diferentes (200 y 500 metros) entre la distancia donde se produjeron los accidentes respecto al radar situado en la misma vía.

Se contabilizan para ello el número de accidentes que ocurrieron dentro de estos radios. La principal conclusión que se extrae del estudio, después de procesar el conjunto de datos, es que aproximadamente un 5,35 % del total de accidentes (276 accidentes) ocurrieron en el periodo mencionado (mayo 2015 - enero 2016), considerando un radio de 500 m al radar más cercano. Sin embargo, cuando se emplea un radio de 200 m, se contabiliza un 2,87 % del total de accidentes en la proximidad del radar (un total de 148) en el mismo periodo.

Como conclusión, los resultados sugieren que tal vez algunas medidas de control de velocidad no están funcionando tal y como se desea, y resultaría conveniente realizar un estudio de su ubicación o su verdadero objetivo.

### Número de accidentes producidos en tramos en obras

En este trabajo también nos hemos centramos en averiguar el número de accidentes que ocurrieron en tramos que se encontraban en obras. El gran número de vías que poseían tramos en reforma o construcción sugirió la necesidad de encontrar el número total de accidentes que tuvieron lugar bajo estas condiciones.

Este estudio no revela un dato concluyente debido a que es bastante confuso establecer periodos de obra o tipos en obras, ya que se ha observado que algunas incidencias relacionadas con obras no parece que se cierren cuando estas han finalizado

realmente. En cualquier caso, se ha registrado que un 30 % del total de accidentes que se produjeron en ese periodo coinciden con tramos de las vías cuando se encontraban con obras en proceso. Como se ha comentado, este dato realmente no revela la veracidad de dichos accidentes, y es susceptible de que sea significativamente menor, debido a que no queda claro que el cierre de incidencias de obras por parte de la DGT coincida realmente con la finalización de la obra.

Sin embargo, sí conviene llamar la atención y revisar el proceso que se realiza en cuanto a este tipo de incidencias publicadas a través del portal *Infocar* de la DGT.

## 5.3. Metodología de la investigación en el estudio de accidentes en España

La metodología de la investigación para llevar a cabo la parte experimental de extracción del modelo consistirá en las siguientes fases:

- Fase 1. Selección de características o predictores mediante distintos métodos.
- Fase 2. Comparativa entre los selectores y elección de los que mejores resultados proporcionen, en base al Área Encerrada bajo la Curva ROC.
- Fase 3. Inserción una a una de las características seleccionadas en el punto anterior en un modelo Random Forest (acumulativo).
- Fase 4. Inserción de la elección en la fase 3 de los predictores en un modelo BayesGLM, donde se rechazarán aquellas variables que no sean significativas o porque el signo del coeficiente no corresponda no sea coherente con la realidad. Adicionalmente, se empleará este modelo por la facilidad que ofrece para interpretar los resultados, en contraposición a Random Forest.

Como se ha visto en el capítulo anterior, el dataset que se emplea inicialmente posee un total de 84 características.

Después de preprocesar el conjunto de datos con el fin de unificar algunas de ellas para obtener un único predictor, y de descartar aquellas que en principio podrían no presentar ninguna relación con la investigación, el set de datos ha quedado recortado inicialmente a 23 características:

```
ID_ACCIDENTE, ID_VEHICULO, ID_PERSONA, TIPO_VEHICULO, ANOMALIA_NINGUNA, NUMERO_OCUPANTES_VEH, EDAD, SEXO,
POSICION, Casualty_Severity, ANIO, MES, DIASEMANA, HORA, TOT_VEHICULOS_IMPLICADOS, ZONA_AGRUPADA, TIPO_VIA,
INTERSECCION, SUPERFICIE_CALZADA, LUMINOSIDAD, FACTORES_ATMOSFERICOS, TIPO_ACCIDENTE, MANIOBRAS
```

de estas 23 características prescindiremos de los identificadores de cada tabla, puesto que no aportan información, por lo que el dataset se recortará a 20 características, que son las siguientes:

```
ANIO, EDAD,ANOMALIA_NINGUNA, HORA, MES, DIASEMANA,POSICION, MANIOBRAS, TIPO_VEHICULO, INTERSECCION, TIPO_VIA,
ZONA_AGRUPADA, TIPO_ACCIDENTE, TOT_VEHICULOS_IMPLICADOS, SEXO, NUMERO_OCUPANTES_VEH, LUMINOSIDAD,
SUPERFICIE_CALZADA, FACTORES_ATMOSFERICOS, Casualty_Severity
```

De ellas se extrae la característica de interés, *Casualty_Severity*, que se dividió en dos clases, *clase 0* para los accidentes graves y fallecimientos, y *clase 1* para los accidentes leves, consiguiendo de esta forma una variable dicotómica con la que trabajar.

En las siguientes secciones se estudiarán diferentes métodos para realizar una selección de características que no penalice en exceso el resultado del área encerrada bajo la curva ROC en nuestra clasificación.

## 5.4. Selección de características y ajuste del modelo

### 5.4.1. Uso del método de validación cruzada para generar los splits de datos

Antes de seleccionar características y realizar el ajuste del modelo, se procede a dividir el set de datos en dos subconjuntos: training y testing. Sin embargo, se ha de garantizar que los datos del subset training son independientes de los datos del subset de testing.

Con el fin de asegurar esta independencia, se empleará el método de validación cruzada o K-fold. El método K-fold particionará el set de entrenamiento en K subconjuntos de aproximadamente el mismo número de muestras aleatorias que tenía éste. Un subconjunto se entrena usándolo como datos de prueba y el resto grupos (K-1) como datos de entrenamiento, y repetiremos este proceso K veces con cada subconjunto de prueba, donde el primer grupo de muestras se devolverá al set de entrenamiento. Un ejemplo gráfico se muestra en la figura 2.7 de la sección 2.3.5.

Para trabajar mediante la técnica de validación cruzada, acudiremos al software estadístico R donde, a través de la función *createDataPartition()* del paquete *caret*, conseguiremos tal fin:

```
set.seed(2014)
trainingRows<-createDataPartition(
      vehicles.casualties.accidents_all2$Casualty_Severity,
      p=.80,list=FALSE)
training<-vehicles.casualties.accidents_all2[trainingRows,]
testing<-vehicles.casualties.accidents_all2[-trainingRows,]
```

Será necesario emplear el mismo *seed* para la generación de los números aleatorios.

### 5.4.2. Selección de características mediante Random Forest y BayesGLM

Mediante el uso de selectores de características basados en RFE, y para el caso concreto que nos ocupa, a través de Random Forest, se pretende dar soporte a la selección anterior mediante el uso de este tipo de selectores ya que permiten realizar estimaciones individuales mediante la inserción progresiva de predictores al modelo de clasificación.

Se ha implementado un script en R que permite generar dichas estimaciones y cuyos puntos más importantes se describen a continuación:

```
# Matriz de predictores sin Casualty_Severity
trainX2 = carmetry_train2[ , -15]

set.seed(2014)
# define the control using a random forest selection function
control = rfeControl(functions=rfFuncs,
                                method="cv",
                                number=10,
                                returnResamp="final",
                                verbose = TRUE)

trainctrl = trainControl(classProbs = TRUE,
                                    summaryFunction = twoClassSummary)
#The simulation will fit models with subset sizes of 1 to 14
subsets = c(1:14)

# run the RFE algorithm
  results2 = rfe(trainX2, carmetry_train2$Casualty_Severity,
                        subsets,trcontrol=trainctrl,method="rf",
                        metric = "ROC" ,rfeControl=control)
```

### 5.4.3. Ajuste del modelo de clasificación

En las sucesivas páginas de esta sección se muestra el proceso de ajuste del modelo que incluye el descarte de determinadas características, en algunos casos que BayesGLM o Random Forest consideren que carecen de importancia, y al mismo tiempo, se considere que no es fundamental conservarlas para nuestro estudio.

Se partirá para ello de las 19 características anteriores:

ANIO, EDAD, NOMALIA_NINGUNA, HORA, MES, DIASEMANA, POSICION, MANIOBRAS, TIPO_VEHICULO, INTERSECCION, TIPO_VIA, ZONA_AGRUPADA, TIPO_ACCIDENTE, TOT_VEHICULOS_IMPLICADOS, SEXO, NUMERO_OCUPANTES_VEH, LUMINOSIDAD, SUPERFICIE_CALZADA, FACTORES_ATMOSFERICOS

junto con la clase: *Casualty_Severity*

#### Test 1. Estimación de la importancia del año de ocurrencia del accidente

Se realiza inicialmente un análisis acerca de la importancia de las variables mediante Random Forest. Para ello, se entrenará un modelo que realice una estimación de la distancia menor en la curva ROC al vértice (1,1) de sensibilidad y especificidad.

Atendiendo a la figura 3.10, se puede observar que para la serie 2011 a 2015, años que alberga el estudio que nos ocupa, hay una variación de entorno al 23 % en el número de fallecidos en vías interurbanas (ver tabla 5.29), por lo que ya se prevé que la importancia de esta variable será elevada, sin embargo, se deberá descartar más adelante puesto que podrá sesgar nuestro estudio predictivo.

Al hablar de heridos de gravedad, la variación es aun mayor, alcanzando diferencias del 30 %, aproximadamente, en función del año (ver tabla 5.30).

Los heridos leves, sin embargo, se mantienen ligeramente constantes, con una variación no superior al 7 % entre los años del periodo 2011 a 2015, tal y como se puede observar en la tabla 5.31.

Para realizar la selección de características, se procederá a seleccionar los siguientes parámetros del modelo:

- library: *randomForest*
- type: *Classification*
- parameters: *mtry, threshold, ROC*

donde el parámetro $m_{try}$ indicará en el número de predictores que se seleccionan aleatoriamente, mientras que el parámetro *threshold* indicará el umbral o punto de corte en la curva ROC, cuya distancia al vértice $(1, 1)$ de sensibilidad y especificidad sea la menor.

Se escogerá la secuencia entre 0,01 y 0,99 para el valor de del úmbral o *threshold* y la raíz cuadrada del número total de predictores para el valor del parámetro $m_{try}$, con el fin de reducir la correlación entre árboles y reducir la varianza.

| Año | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|
| Total fallecidos | 1.603 | 1.442 | 1.230 | 1.247 | 1.248 |

**Tabla 5.29:** Número de fallecidos en vías interurbanas entre 2011 y 2015. (Fuente: Dirección General de Tráfico)

| Año | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|
| Total heridos graves | 6.825 | 6.044 | 5.182 | 4.834 | 4.744 |

**Tabla 5.30:** Número de heridos graves en vías interurbanas entre 2011 y 2015. (Fuente: Dirección General de Tráfico)

La matriz de confusión vendrá dada por la tabla 5.33 para el modelo en el que se incluye el año como predictor, y la tabla 5.34 muestra la matriz de confusión para el modelo que no incluye el año de ocurrencia del accidente como predictor. La primera

| Año | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|
| Total heridos leves | 47.692 | 47.936 | 51.320 | 48.693 | 48.036 |

**Tabla 5.31:** Número de heridos leves en vías interurbanas entre 2011 y 2015. (Fuente: Dirección General de Tráfico)

| threshold | ROC | Sens | Spec | Dist |
|---|---|---|---|---|
| 0,01000000 | 0,8666775 | 0,8776770560 | 0,6790375 | 0,3435253 |
| 0,06157895 | 0,8666775 | 0,5190435095 | 0,9317866 | 0,4857742 |
| 0,11315789 | 0,8666775 | 0,3541916952 | 0,9720465 | 0,6464138 |
| 0,16473684 | 0,8666775 | 0,2610731411 | 0,9858932 | 0,7390618 |
| 0,21631579 | 0,8666775 | 0,1991054952 | 0,9922240 | 0,8009324 |
| 0,26789474 | 0,8666775 | 0,1599536292 | 0,9952811 | 0,8400597 |
| 0,31947368 | 0,8666775 | 0,1298370618 | 0,9971235 | 0,8701677 |
| 0,37105263 | 0,8666775 | 0,1092116366 | 0,9981514 | 0,8907903 |
| 0,42263158 | 0,8666775 | 0,0921150895 | 0,9987588 | 0,9078858 |
| 0,47421053 | 0,8666775 | 0,0782126578 | 0,9991559 | 0,9217877 |
| 0,52578947 | 0,8666775 | 0,0658616170 | 0,9994300 | 0,9341386 |
| 0,57736842 | 0,8666775 | 0,0545755250 | 0,9996309 | 0,9454246 |
| 0,62894737 | 0,8666775 | 0,0410989950 | 0,9997835 | 0,9589010 |
| 0,68052632 | 0,8666775 | 0,0272573165 | 0,9998894 | 0,9727427 |
| 0,73210526 | 0,8666775 | 0,0186177465 | 0,9999361 | 0,9813823 |
| 0,78368421 | 0,8666775 | 0,0115600830 | 0,9999642 | 0,9884399 |
| 0,83526316 | 0,8666775 | 0,0062058785 | 0,9999813 | 0,9937941 |
| 0,88684211 | 0,8666775 | 0,0030725349 | 0,9999907 | 0,9969275 |
| 0,93842105 | 0,8666775 | 0,0004563247 | 1,0000000 | 0,9995437 |
| 0,99000000 | 0,8666775 | 0,0000000000 | 1,0000000 | 1,0000000 |

**Tabla 5.32:** Parámetros de ajuste para modelo Random Forest de 19 predictores

de ellas se estudia a partir de los resultados devueltos por el modelo Random Forest mediante un re-muestreo de resultados a través de los parámetros de ajuste, tal y como se puede observar en la tabla 5.32.

Para entender el significado del parámetro '*threshold*', primero observaremos la figura 5.8 perteneciente a un ejemplo obtenido de un set de datos distinto al que actualmente estamos empleando, pero que puede ayudarnos a entender el significado de este parámetro. Dicha figura corresponde al comando '*ggplot(mod1)*' que se emplea para visualizar gráficamente el perfil de ajuste de un modelo. En ella se muestra un '*plot*' de la sensibilidad, especificidad y distancia al modelo perfecto, es decir, al vértice (1,1). Se muestra que a medida que aumentamos el punto de corte para la probabilidad (eje x) se ve claramente que cada una de las muestras se predecirá como perteneciente a la primera clase. Sin embargo, existe un punto de corte en el que la sensibilidad y la especificidad alcanzan el mismo valor, que en el caso de la figura es de 0,887, y que se corresponderá con el punto óptimo.

Ese punto de corte que se comentaba es el denominado '*threshold*' y es posible variarlo en función de nuestras necesidades para obtener un determinado valor para la sensibilidad y especificidad.

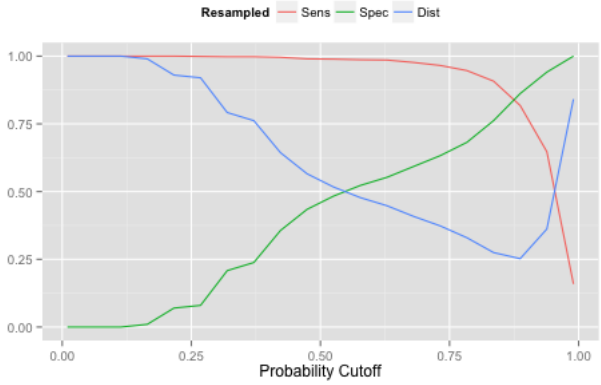



**Figura 5.8:** Ejemplo de ggplot para visualizar gráficamente el perfil de ajuste de un modelo

| Predicción \ Referencia | X0 | X1 |
|---|---|---|
| X0 | 7111 | 50559 |
| X1 | 1106 | 109967 |

**Tabla 5.33:** Matriz de confusión del dataset de test mediante Random Forest con año del siniestro para umbral=0,01

- Accuracy : 0,6938
- Kappa : 0,1428

Donde el valor de kappa se obtiene de la ecuación 5.2, mientras que el valor de accuracy se obtiene de la ecuación 5.4.

$$Kappa = \frac{observedaccuracy - expectedaccuracy}{1 - expectedaccuracy} \tag{5.2}$$

$$ExpectedAccuracy = \frac{\left(\frac{(TP+FN)(TP+FP)}{Total}\right) + \left(\frac{(FP+TN)(FN+TN)}{Total}\right)}{Total} \tag{5.3}$$

$$ObservedAccuracy = \frac{(TP+TN)}{Total} \tag{5.4}$$

- Accuracy : 0,7165

| Predicción \ Referencia | X0 | X1 |
|---|---|---|
| X0 | 6908 | 46537 |
| X1 | 1309 | 113989 |

**Tabla 5.34:** Matriz de confusión del dataset de test mediante Random Forest sin año del siniestro para umbral=0.01

- Kappa : 0,1525

El modelo obtenido mediante Random Forest para los 19 predictores de la tabla 5.35 se puede observar en la tabla 5.36.

| Predictores | |
|---|---|
| TIPO_VEHICULO | ANOMALIA_NINGUNA |
| NUMERO_OCUPANTES_VEH | EDAD |
| SEXO | POSICION |
| ANIO | MES |
| DIASEMANA | HORA |
| TOT_VEHICULOS_IMPLICADOS | ZONA_AGRUPADA |
| TIPO_VIA | INTERSECCION |
| SUPERFICIE_CALZADA | LUMINOSIDAD |
| FACTORES_ATMOSFERICOS | TIPO_ACCIDENTE |
| MANIOBRAS | |

**Tabla 5.35:** Predictores iniciales empleados para modelo Random Forest

| Predictores | Modelo | ROC Train | ROC Test | Sens | Spec | Threshold | K-fold |
|---|---|---|---|---|---|---|---|
| 19 | RF | 0,866 | - | 0,877677056 | 0,6790375 | 0,01 | 10 |

**Tabla 5.36:** Modelo obtenido mediante Random Forest con 19 predictores

El parámetro de entrenamiento $m_{try}$ se mantuvo constante a un valor de 14.

El listado de las 20 clases de los factores o características más importantes de los predictores, por orden de importancia, se puede observar en la tabla 5.37.

En este listado se confirma el hecho de que, efectivamente, la variable ANIO ocupa una posición de importancia en él, según el modelo entrenado de Random Forest. Puesto que no resulta un objetivo incluido en esta investigación la estimación del año de ocurrencia de cada accidente, se deja como línea futura la implementación de un modelo mixto donde el año sea una variable aleatoria a predecir.

Adicionalmente, se puede observar en la figura 5.9 las curvas ROC con la variable año y sin ella, donde no se aprecia apenas variación alguna debido a que el área encerrada bajo la curva ROC con la variable año fue de $0,866$ (tabla 5.36), en contraposición con aquella estimada sin la variable año que fue de $0,857$ (tabla 5.38) tal y como se observará en la siguiente sección, donde se estudiará el efecto de la eliminación de dicha variable sobre el modelo.

| Predictor | Overall |
|---|---|
| EDAD | 100,00 |
| HORA | 73,35 |
| ANOMALIA_NINGUNAX3 | 69,71 |
| MES | 59,21 |
| DIASEMANA | 47,35 |
| ANIO | 42,73 |
| POSICIONX6 | 38,67 |
| MANIOBRASX52 | 33,75 |
| INTERSECCIONX2 | 26,76 |
| TIPO_VEHICULOX4 | 25,68 |
| TIPO_VIAX3 | 18,10 |
| TIPO_VEHICULOX6 | 17,90 |
| TIPO_VIAX8 | 15,69 |
| ZONA_AGRUPADAX2 | 14,41 |
| TOT_VEHICULOS_IMPLICADOSX2 | 14,17 |
| SEXOX2 | 13,85 |
| TIPO_ACCIDENTEX4 | 13,77 |
| MANIOBRASX72 | 13,65 |
| TIPO_ACCIDENTEX16 | 13,38 |
| NUMERO_OCUPANTES_VEHX1 | 12,03 |

**Tabla 5.37:** Orden de importancia de las primeras 20 clases de los predictores obtenido mediante Random Forest

### Test 2. Construcción de modelo independiente del año de ocurrencia del accidente

A continuación, se va a entrenar un modelo prescindiendo de la variable *ANIO* con el fin de no sesgar la predicción, tal y como se comentaba anteriormente. Los parámetros de entrenamiento empleados serán los mismos que en el apartado anterior.

En la tabla 5.38 se puede ver que el valor de threshold escogido en el entrenamiento con el fin de minimizar la distancia al vértice (1,1) de la curva ROC ha sido de:

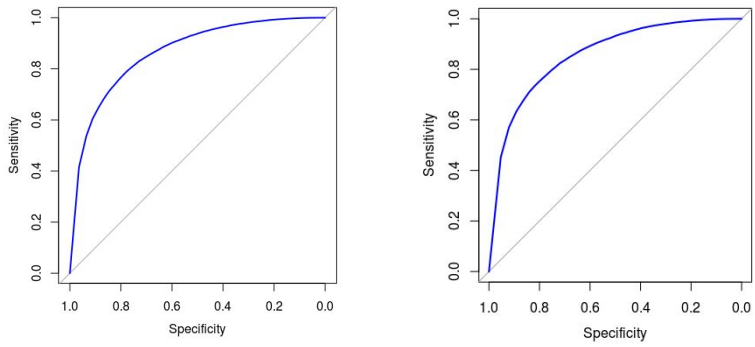



**Figura 5.9:** Curva ROC con la variable ANIO (izquierda) y sin la variable ANIO (derecha)

- $threshold = 0,01$

| threshold | ROC | Sens | Spec | Dist |
|---|---|---|---|---|
| 0,01000000 | 0,8572208 | 0,8473473341 | 0,7050847 | 0,3321363 |
| 0,06157895 | 0,8572208 | 0,4773667019 | 0,9395829 | 0,5261166 |
| 0,11315789 | 0,8572208 | 0,3290335758 | 0,9749433 | 0,6714347 |
| 0,16473684 | 0,8572208 | 0,2427594568 | 0,9872700 | 0,7573477 |
| 0,21631579 | 0,8572208 | 0,1891577600 | 0,9928002 | 0,8108743 |
| 0,26789474 | 0,8572208 | 0,1537478524 | 0,9955833 | 0,8462637 |
| 0,31947368 | 0,8572208 | 0,1257910596 | 0,9971796 | 0,8742135 |
| 0,37105263 | 0,8572208 | 0,1062000215 | 0,9981545 | 0,8938019 |
| 0,42263158 | 0,8572208 | 0,0908070816 | 0,9987136 | 0,9091938 |
| 0,47421053 | 0,8572208 | 0,0763874609 | 0,9991341 | 0,9236130 |
| 0,52578947 | 0,8572208 | 0,0649793998 | 0,9994331 | 0,9350208 |
| 0,57736842 | 0,8572208 | 0,0531152453 | 0,9996091 | 0,9468848 |
| 0,62894737 | 0,8572208 | 0,0402167501 | 0,9997633 | 0,9597833 |
| 0,68052632 | 0,8572208 | 0,0267402016 | 0,9998754 | 0,9732598 |
| 0,73210526 | 0,8572208 | 0,0193174542 | 0,9999221 | 0,9806825 |
| 0,78368421 | 0,8572208 | 0,0116209287 | 0,9999595 | 0,9883791 |
| 0,83526316 | 0,8572208 | 0,0054149947 | 0,9999891 | 0,9945850 |
| 0,88684211 | 0,8572208 | 0,0019165210 | 0,9999969 | 0,9980835 |
| 0,93842105 | 0,8572208 | 0,0001216823 | 1,0000000 | 0,9998783 |
| 0,99000000 | 0,8572208 | 0,0000000000 | 1,0000000 | 1,0000000 |

**Tabla 5.38:** Parámetros de ajuste para modelo Random Forest de 18 predictores

En la figura 5.10 se observa la densidad respecto a los resultados devueltos por la tabla 5.38 en función del área bajo la curva ROC, Sensibilidad, Especificidad y Distancia al vértice (1,1).

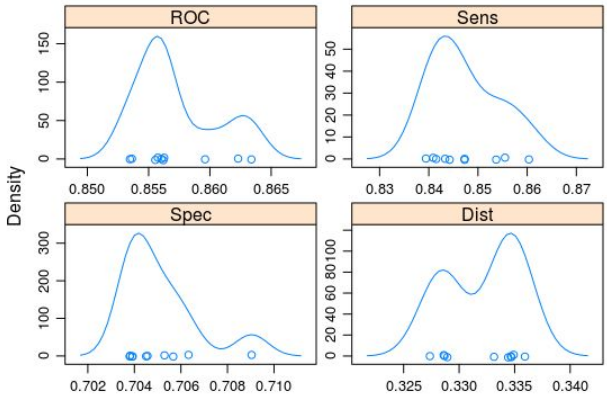



**Figura 5.10:** Gráfico donde se expresa la densidad en función del área encerrada bajo la curva ROC, Sensibilidad, Especificidad y Distancia

Adicionalmente, se puede observar en la tabla 5.39 que la clasificación de importancia de los predictores para este caso es similar al anterior, donde continúa siendo la

variable EDAD la más importante, aunque en este caso, la variable HORA ha pasado de la segunda a la tercera posición en beneficio de ANOMALIA_NINGUNAX3.

En la figura 5.11 se puede observar de forma gráfica la representación de la distribución en porcentaje de las 20 primeras clases en orden de importancia después de entrenar el modelo anterior. Como se puede observar en dicha figura, las marcas de tiempo son importantes a la hora de estudiar un accidente. Prueba de ello es que los predictores asociados a la hora, mes y día de la semana de ocurrencia de los accidentes se encuentran entre las variables más importantes en este gráfico basado en Random Forest. Más adelante, en este capítulo, se verá si es posible prescindir de alguna de ellas con motivo de obtener una generalización en la obtención del modelo al menos no supeditada al día de la semana.

Anteriormente, en la sección 3.2.1, se estudió que la meteorología jugaba un factor importante para que aumentara el número de personas que decidieran coger su vehículo o no. Se vio que debido a unas malas condiciones climatológicas era posible que la visibilidad en las carreteras estuviera mermada o provocara que la superficie de éstas fuera más resbaladiza. Por el contrario, también se estudió que aquellos períodos de inusual buen tiempo casi siempre tienen un efecto en el aumento de accidentes debido a que pueden estimular a la realización de viajes, aumentando así la exposición a un accidente.

Por otro lado, la hora y el día de la semana tiene su importancia debido a que se producen mayores desplazamientos por el comienzo o finalización de la jornada laboral, o por el inicio o finalización del horario escolar, tal y como se observó en el caso de Reino Unido (sección 6.3.1), o en el caso de España (sección 5.2.2.6).

| Predictor | Overall |
|---|---|
| EDAD | 100,00 |
| ANOMALIA_NINGUNAX3 | 74,05 |
| HORA | 72,55 |
| MES | 58,20 |
| DIASEMANA | 46,58 |
| POSICIONX6 | 39,28 |
| MANIOBRASX52 | 35,87 |
| TIPO_VEHICULOX4 | 28,24 |
| INTERSECCIONX2 | 27,27 |
| TIPO_VEHICULOX6 | 18,77 |
| TIPO_VIAX3 | 17,99 |
| TIPO_VIAX8 | 16,13 |
| ZONA_AGRUPADAX2 | 15,47 |
| MANIOBRASX72 | 15,02 |
| TIPO_ACCIDENTEX4 | 14,78 |
| TOT_VEHICULOS_IMPLICADOSX2 | 14,17 |
| SEXOX2 | 13,61 |
| TIPO_ACCIDENTEX16 | 13,55 |
| NUMERO_OCUPANTES_VEHX1 | 12,04 |
| TIPO_ACCIDENTEX2 | 10,78 |

**Tabla 5.39:** Orden de importancia de las primeras 20 clases de los predictores obtenido mediante Random Forest

### Test 3: Modelo basado en BayesGLM con 18 predictores

En este test se parte de los mismos 18 predictores obtenidos mediante la selección realizada por Random Forest, donde únicamente se ha eliminado el parámetro relativo al año de ocurrencia del accidente por los motivos anteriormente comentados. Se obtiene el área encerrada bajo la curva ROC para este modelo obteniendo un buen resultado, tal y como se puede observar en la tabla 5.40.

| Test | Predictores | Modelo | Vars. elim. | ROC Train | ROC Test | Sens | Spec | Threshold | K-fold |
|---|---|---|---|---|---|---|---|---|---|
| 3 | 18 | BayesGLM | ANIO | - | 0,8727 | - | - | - | 10 |

**Tabla 5.40:** Modelo obtenido mediante BayesGLM con 18 predictores

El tiempo que se ha tomado el sistema para la ejecución de este modelo ha sido bastante elevado:

- Tiempo empleado en la ejecución: 25.644 seg ($\approx ,7h$).

Se pretende estimar un nuevo modelo con un menor número de predictores que permita en un futuro la creación de un modelo dinámico en función del almacenamiento en tiempo real de los accidentes de tráfico.

Aunque el *summary()* del modelo lo podemos encontrar en el anexo I, el área encerrada bajo la curva ROC se puede observar en la figura 5.12.

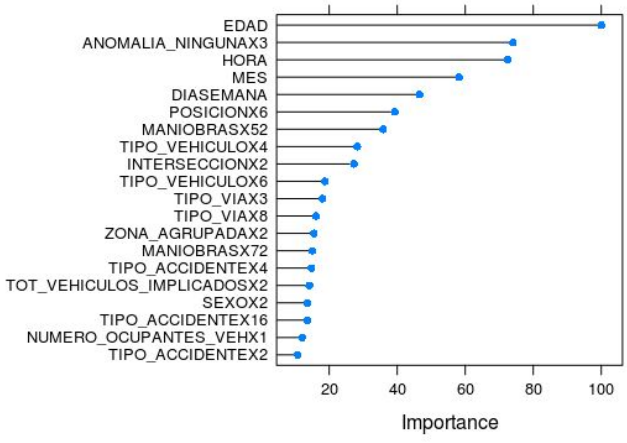



**Figura 5.11:** Gráfico donde se expresa la importancia de cada variable según el modelo basado en Random Forest para 18 predictores

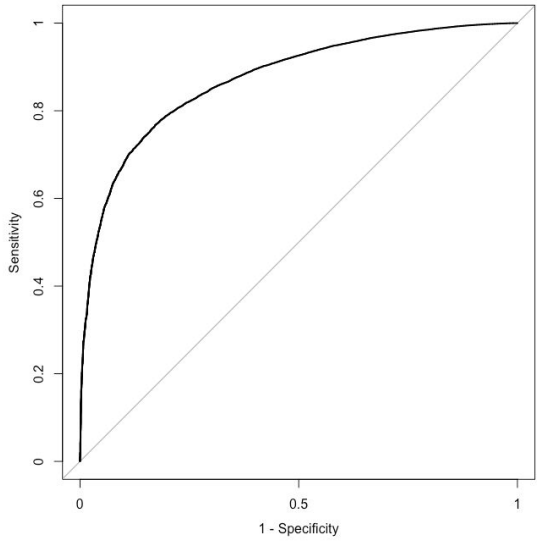



**Figura 5.12:** Área encerrada bajo la curva ROC del modelo BayesGLM de 18 predictores (test 3)

### Test 4: Modelo basado en RFE-RF con 18 predictores

Los dos tests que se van a llevar a cabo a continuación corresponden con la estimación de la importancia de la variable *ANOMALIA_NINGUNA*. El motivo de la selección de esta variable con la finalidad de someterla a estudio no es casual, ya que como se estudió en la subsección 5.2.1.1 sería necesario conocer si esta variable se debe o no incorporar al modelo, ya que existen muy pocos accidentes donde se ha detectado alguna anomalía previa en el vehículo, por lo que en un principio podría parecer que puede crear un sesgo, aunque más adelante se descubrirá que no es así.

Puesto que se han obtenido mejores resultados a través una clasificación previa mediante Random Forest, se tratará de estudiar la importancia de esta variable a través de RFE-RF. En primer lugar, se va a realizar un test conservando dicha variable para, a continuación, eliminar la variable y realizar un nuevo test, ambos bajo el selector de características RFE-RF, tal y como se ha comentado. Ambos tests se han denominado test 4 y test 5, respectivamente.

El tiempo de ejecución para la selección mediante RFE-RF, como se puede observar, es demasiado elevado:

- Tiempo empleado en la ejecución: 126.805 seg ($\approx ,35h$)

Sin embargo, al contrario de lo que podría ocurrir con BayesGLM, nos aseguramos que con este método no se tendrán problemas de *overfitting*, donde internamente en el selector se ha realizado una validación cruzada de muestras o *K-fold*, para $K = 10$.

El resultado de las variables más importantes lo podemos encontrar en la tabla 5.41.

| Variables | Accuracy | Kappa |
|---|---|---|
| 1 | 0,9513 | 0,00000 |
| 2 | 0,9513 | 0,00000 |
| 3 | 0,9513 | 0,00000 |
| 4 | 0,9192 | 0,07816 |
| 5 | 0,9255 | 0,09647 |
| 6 | 0,9416 | 0,13491 |
| 7 | 0,9445 | 0,13957 |
| 8 | 0,9449 | 0,13891 |
| 9 | 0,9371 | 0,16168 |
| 10 | 0,9369 | 0,16815 |
| 11 | 0,9361 | 0,18239 |
| 12 | 0,9356 | 0,18931 |
| 13 | 0,9372 | 0,18847 |
| 14 | 0,9376 | 0,19268 |
| 15 | 0,9386 | 0,18821 |
| 16 | 0,9031 | 0,20217 |

**Tabla 5.41:** Resultados de la estimación del *Accuracy* de las variables más importantes mediante RFE-RF

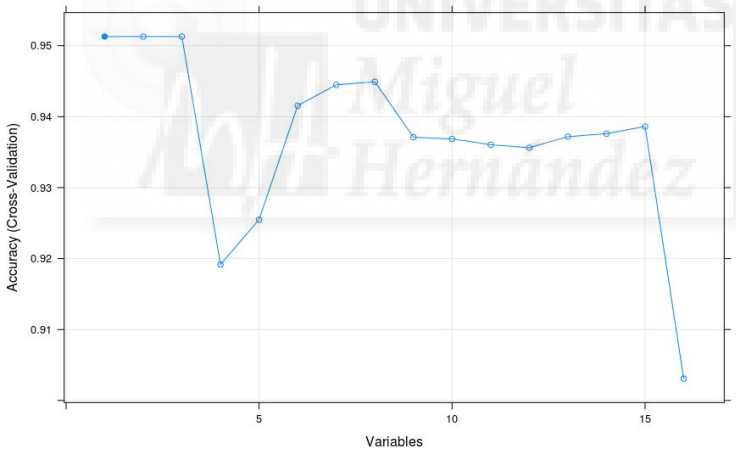



**Figura 5.13:** Determinación mediante RFE-RF del *Accuracy* del modelo RF en función del número de predictores

### Test 5: Modelo basado en RFE-RF con 17 predictores

En el siguiente test se realizará un nuevo test de selección mediante RFE-RF donde se eliminará la variable *ANOMALIA_NINGUNA* y se estimará la evolución del resto de variables en función de su *Accuracy*.

- Tiempo empleado en la ejecución: 109.181 seg ($\approx 30h$)

El orden de importancia obtenido de predictores vendrá dado por la tabla 5.42.

| | |
|---|---|
| 1 | MANIOBRAS |
| 2 | EDAD |
| 3 | SUPERFICIE_CALZADA |
| 4 | HORA |
| 5 | POSICION |
| 6 | FACTORES_ATMOSFERICOS |
| 7 | TIPO_VEHICULO |
| 8 | ZONA_AGRUPADA |
| 9 | MES |
| 10 | LUMINOSIDAD |
| 11 | TIPO_VIA |
| 12 | DIASEMANA |
| 13 | SEXO |
| 14 | NUMERO_OCUPANTES_VEH |

**Tabla 5.42:** Resultados de la clasificación de predictores en función del *Accuracy* mediante RFE-RF

No obstante, se considera necesario obtener el área encerrada bajo la curva ROC sin el predictor *ANOMALIA_NINGUNA* con el fin de poder estimar este parámetro y compararlo con el *accuracy* devuelto. Recordemos que nos basaremos en el área encerrada bajo la curva ROC para validar el modelo, aunque será posible obtener apoyo a través de la precisión o *accuracy* del modelo.

Para ello, se lanza el test 5 (tabla 5.43), pero obteniendo como parámetro de entrenamiento del modelo el área encerrada bajo la curva ROC.

| Test | Predictores | Modelo | Vars. elim. | ROC Train | ROC Test | Sens | Spec | Threshold | K-fold |
|---|---|---|---|---|---|---|---|---|---|
| 5 | 17 | RF | ANIO, ANOMALIA_-NINGUNA | 0,82974 | 0,8253 | 0,8036 | 0,707 | 0,01 | 10 |

**Tabla 5.43:** Modelo obtenido mediante RFE-RF con 17 predictores

Por algún motivo aparentemente ajeno a la ejecución del algoritmo, el tiempo casi se ha triplicado, llegando a más de 80 horas. Se considera un tiempo anómalo y no se tiene en cuenta para comparar con el resto:

- Tiempo empleado en la ejecución: 294.054 seg ($\approx 81h$)
- Parámetro $m_{try}$ : 14

En la figura 5.14 se puede observar el gráfico perteneciente a la curva ROC de este modelo.

La clasificación de los niveles de los predictores en orden de importancia vendrá dado por la tabla 5.44.

Adicionalmente, en la figura 5.15 se puede observar cómo decae bruscamente el parámetro *accuracy* en la estimación del modelo a medida que se insertan predictores en él. Se puede observar claramente que el predictor 9 (mes) hace decaer esta medida en casi tres puntos, lo cual será objeto de estudio más adelante.

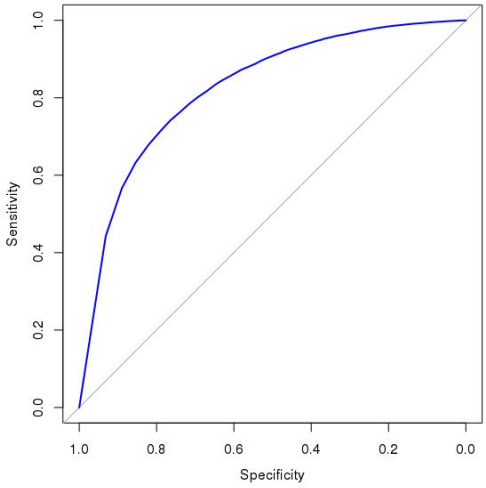



**Figura 5.14:** Área encerrada bajo la curva ROC del modelo Random Forest de 17 predictores (test 5)

| Predictor | Importance |
|---|---|
| EDAD | 100 |
| HORA | 72,73 |
| MES | 57,60 |
| DIASEMANA | 45,90 |
| POSICIONX6 | 37,88 |
| MANIOBRASX52 | 29,40 |
| INTERSECCIONX2 | 28,53 |
| TIPO_VEHICULOX4 | 27,64 |
| TIPO_VEHICULOX6 | 18,77 |
| TIPO_VIAX3 | 17,57 |
| TIPO_ACCIDENTEX4 | 15,60 |
| TOT_VEHICULOS_IMPLICADOSX2 | 14,06 |
| TIPO_ACCIDENTEX20 | 13,91 |
| TIPO_VIAX8 | 13,83 |
| ZONA_AGRUPADAX2 | 13,79 |
| SEXOX2 | 13,63 |
| TIPO_ACCIDENTEX16 | 13,21 |
| NUMERO_OCUPANTES_VEHX1 | 11,53 |
| TIPO_ACCIDENTEX2 | 10,85 |
| TIPO_ACCIDENTEX12 | 10,40 |

**Tabla 5.44:** Orden de importancia de los predictores obtenido mediante Random Forest test 5

### Test 6: Modelo basado en BayesGLM con 17 predictores

En este test se eliminará la variable NUMERO_OCUPANTES_VEH debido a que anteriormente se ha comprobado que Random Forest indicaba que no era importante, cosa que se corrobora en el análisis efectuado en el test anterior mediante BayesGLM.

El tiempo empleado en la ejecución se ha reducido a menos de la mitad que para el

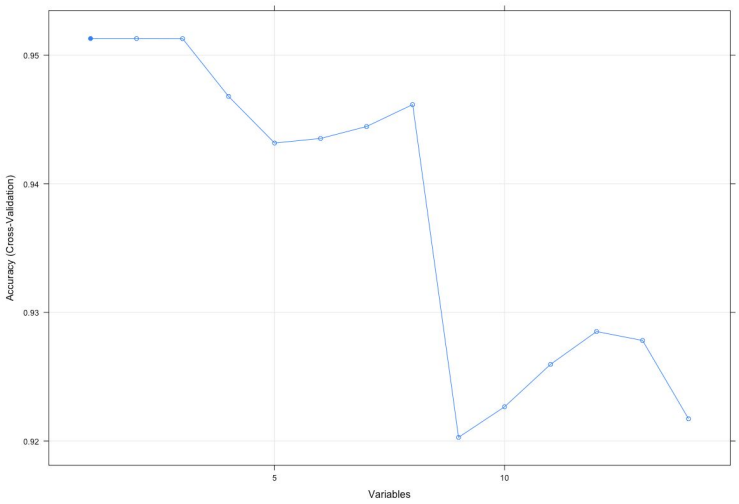



**Figura 5.15:** Determinación mediante RFE-RF del *Accuracy* del modelo RF en función del número de predictores

| Test | Predictores | Modelo | Vars. elim. | ROC Train | ROC Test | Sens | Spec | Threshold | K-fold |
|---|---|---|---|---|---|---|---|---|---|
| 6 | 17 | BayesGLM | ANIO, NUMERO_OCUPANTES_VEH | 0,870525 | 0,8721 | 0,7733 | 0,8203 | 0,0505 | 10 |

**Tabla 5.45:** Modelo obtenido mediante BayesGLM con 17 predictores

test 3.

- Tiempo empleado en la ejecución: 12.104 seg ($\approx 3h$).

El área encerrada bajo la curva ROC se puede observar en la figura 5.16. Adicionalmente, se puede observar que se ha obtenido el área encerrada bajo la curva ROC para ambos set de datos: training y test. Es importante señalar que debido al método de validación cruzada (K-fold) y a una selección de características mediante Random Forest, el modelo no cae en overfitting, tal y como podemos comprobar al no caer el valor de la curva ROC con el dataset de testing respecto a aquella estimada para el dataset de training.

### Test 7: Modelo basado en BayesGLM con 16 predictores

En este test se eliminará la variable *TOT_VEHICULOS_IMPLICADOS* debido a que Random Forest indicaba que poseía una importancia baja en el conjunto de datos, hecho que se corrobora en el análisis efectuado en el párrafo anterior mediante BayesGLM.

Como se puede comprobar, el tiempo el empleado en la ejecución ha disminuido en una hora adicional, reduciéndose a 2 horas:

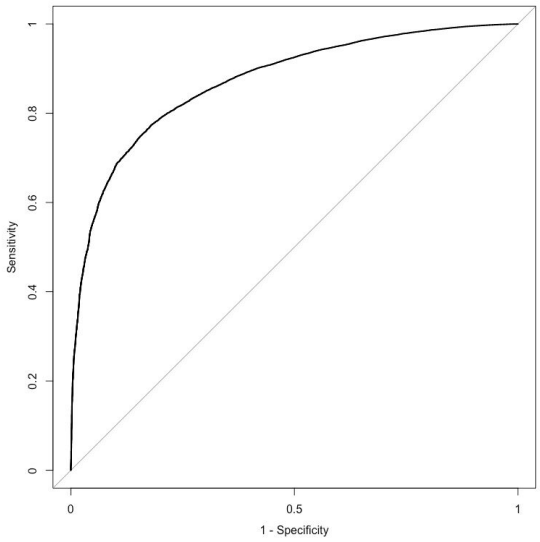



**Figura 5.16:** Área encerrada bajo la curva ROC del modelo BayesGLM de 17 predictores (test 6)

- Tiempo empleado en la ejecución: 6.685 seg ($\approx 2h$)

### Test 8: Modelo basado en BayesGLM con 15 predictores

En este test se eliminará la variable *FACTORES_ATMOSFERICOS* debido a que, una vez más, Random Forest indicaba que no era excesivamente importante. La ejecución lanzada sobre BayesGLM incluyendo esta variable arroja unos resultados similares.

El tiempo de ejecución se sigue reduciendo, esta vez en 30 minutos, aproximadamente:

- Tiempo empleado en la ejecución: 5.517 seg ($\approx 1,5h$).

El área encerrada bajo la curva ROC se puede observar en la figura 5.17.

| Test | Predictores | Modelo | Vars. elim. | ROC Train | ROC Test | Sens | Spec | Threshold | K-fold |
|---|---|---|---|---|---|---|---|---|---|
| 7 | 16 | BayesGLM | ANIO, NUMERO_-OCUPANTES_VEH, TOT_VEHICU-LOS_IMPLICADOS | 0,8700 | 0,8715 | 0,775 | 0,815 | 0,0516 | 10 |

**Tabla 5.46:** Modelo obtenido mediante BayesGLM con 16 predictores

| Test | Predictores | Modelo | Vars. elim. | ROC Train | ROC Test | Sens | Spec | Threshold | K-fold |
|---|---|---|---|---|---|---|---|---|---|
| 8 | 15 | BayesGLM | ANIO, NUMERO_OCUPANTES_VEH, TOT_VEHICULOS_IMPLICADOS, FACTORES_ATMOSFERICOS | 0,8706 | 0,8714 | 0,773 | 0,8185 | 0,0507 | 10 |

**Tabla 5.47:** Modelo obtenido mediante BayesGLM con 15 predictores

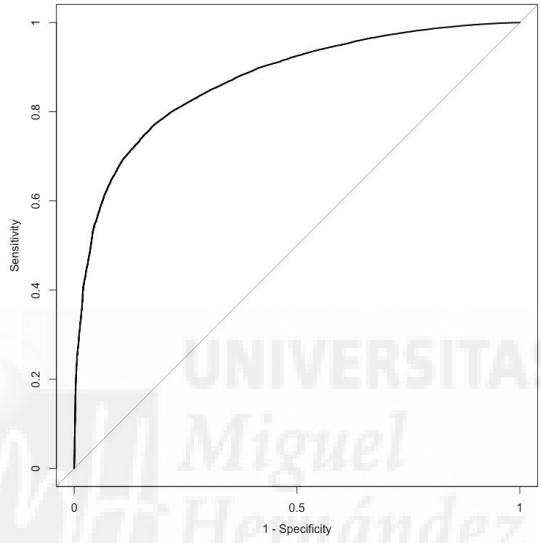



**Figura 5.17:** Área encerrada bajo la curva ROC del modelo BayesGLM de 15 predictores (test 8)

### Test 9: Modelo basado en Random Forest con 15 predictores

A continuación se va a estimar un nuevo modelo basado en Random Forest a partir de 15 predictores (tabla 5.48).

| Test | Predictores | Modelo | Vars. elim. | ROC Train | ROC Test | Sens | Spec | Threshold | K-fold |
|---|---|---|---|---|---|---|---|---|---|
| 9 | 15 | RF | ANIO, NUM_OCUP, TOT_VEHIC_IMP, TIPO_VIA | 0,8532 | 0,8519 | 0,721 | 0,829 | 0,09 | 10 |

**Tabla 5.48:** Modelo obtenido mediante RF con 15 predictores

Como novedad en este test, se va a obtener la matriz de confusión con motivo de observar con mayor detenimiento la representación del modelo para cada una de las clases. Ello aportará una idea del número de verdaderos positivos (TP) que vamos obteniendo, y por tanto, el número de fallecimientos y accidentes graves que se podrían haber detectado.

Por tanto, la matriz de confusión devuelta por el modelo Random Forest con el *threshold* escogido se puede observar en la tabla 5.49

| Predicción \ Referencia | X0 | X1 |
|---|---|---|
| X0 | 6814 | 44838 |
| X1 | 1403 | 115700 |

**Tabla 5.49:** Matriz de confusión mediante Random Forest para el test 9

La gráfica del área encerrada bajo la curva ROC se puede observar en la figura 5.18.

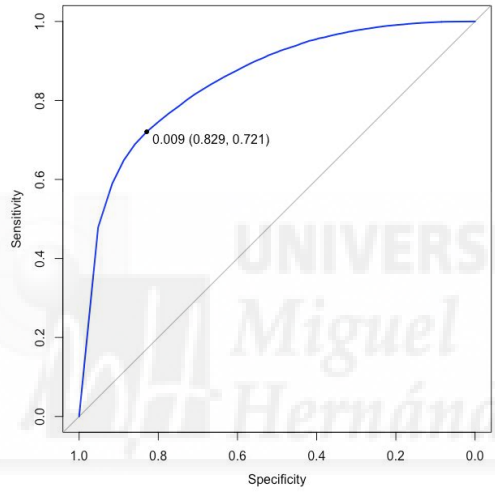



**Figura 5.18:** Determinación mediante Random Forest del área encerrada bajo la curva ROC en función del número de predictores del modelo

La clasificación de los predictores en orden de importancia vendrá dado por la tabla 5.50.

### Test 10: Modelo basado en BayesGLM con 14 predictores

En este test se eliminará la variable *TIPO_VIA* debido a que Random Forest indicaba que su importancia no era muy elevada. Al mismo tiempo, BayesGLM corroboraba lo anterior a través de unos resultados similares con dicha variable.

El tiempo empleado para este modelo, sin embargo, se ha mantenido constante respecto al experimento anterior:

- Tiempo empleado en la ejecución: 5.163 seg ($\approx 1:30h$)

El área encerrada bajo la curva ROC se puede observar en la figura 5.19.

| Predictor | Importance |
|---|---|
| EDAD | 100,00 |
| HORA | 70,404 |
| ANOMALIA_NINGUNAX3 | 61,529 |
| MES | 58,525 |
| DIASEMANA | 44,912 |
| POSICIONX6 | 32,020 |
| MANIOBRASX52 | 31,777 |
| ZONA_AGRUPADAX2 | 29,513 |
| INTERSECCIONX2 | 26,173 |
| TIPO_VEHICULOX4 | 23,920 |
| MANIOBRASX72 | 18,883 |
| TIPO_VEHICULOX6 | 14,859 |
| TIPO_ACCIDENTEX4 | 12,875 |
| TIPO_ACCIDENTEX16 | 12,237 |
| SEXOX2 | 12,161 |
| SUPERFICIE_CALZADAX3 | 10,596 |
| TIPO_ACCIDENTEX2 | 9,806 |
| TIPO_ACCIDENTEX20 | 9,265 |
| TIPO_ACCIDENTEX12 | 9,095 |
| LUMINOSIDADX4 | 8,795 |

**Tabla 5.50:** Orden de importancia de los predictores obtenido mediante Random Forest (test 5)

### Test 11: Modelo basado en Random Forest con 14 predictores

En este test se ha modificado el algoritmo empleado con motivo de comparar resultados entre BayesGLM y Random Forest. Para ello, se ha lanzado un ensayo con los mismos 14 predictores que para el test 10, pero se ha empleado Random Forest como clasificador.

En la tabla 5.52 se puede cotejar el resultado respecto al área encerrada bajo la curva ROC para el set de datos de *testing*, así como la sensibilidad y especificidad devueltas. En ella, se puede observar que el área bajo la curva es muy similar, aunque es sensiblemente mejor para el clasificador basado en BayesGLM.

El handicap de este modelo radica en el tiempo empleado para su ejecución, de más de un día:

- Tiempo empleado en la ejecución: 113.853 seg ($\approx 31h$).

| Test | Predictores | Modelo | Vars. elim. | ROC Train | ROC Test | Sens | Spec | Threshold | K-fold |
|---|---|---|---|---|---|---|---|---|---|
| 10 | 14 | BayesGLM | ANIO, NUMERO_OCUPANTES_VEH, TOT_VEHICULOS_IMPLICADOS, FACTORES_ATMOSFERICOS, TIPO_VIA | 0,869 | 0,871 | 0,7705 | 0,8203 | 0,0501 | 10 |

**Tabla 5.51:** Modelo obtenido mediante BayesGLM con 14 predictores

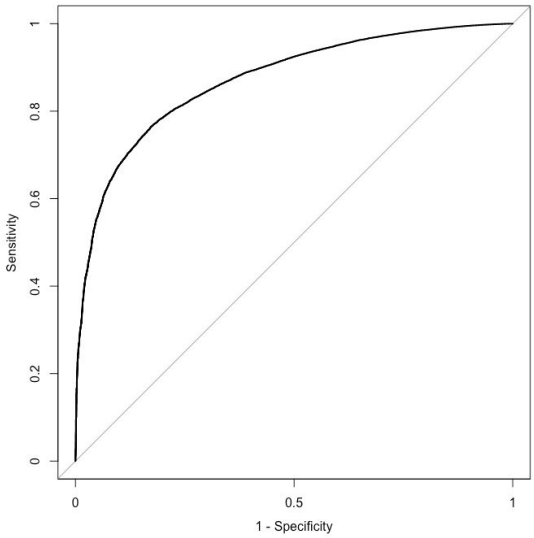



**Figura 5.19:** Área encerrada bajo la curva ROC del modelo BayesGLM de 14 predictores (test 10)

Adicionalmente, la matriz de confusión devuelta por el modelo Random Forest con el *threshold* escogido es:

Algunos valores adicionales del modelo, como la precisión, son los siguientes:

- Accuracy: $0,726$
- Kappa: $0,1568$

Como se puede observar, la precisión (*accuracy*) del modelo no es muy elevada respecto a otros modelos anteriores, sin embargo, no nos resultará tan importante debido a que nosotros basaremos nuestro modelo en el área encerrada bajo la curva ROC, ya que pretendemos proporcionar estimar una probabilidad mayor de acierto para la clase de $X0$ de la variable *Casualty_Severity*, correspondiente a aquellos accidentes donde la

| Test | Predictores | Modelo | Vars. elim. | ROC Train | ROC Test | Sens | Spec | Threshold | K-fold |
|---|---|---|---|---|---|---|---|---|---|
| 11 | 14 | RF | ANIO, NUMERO_OCUPANTES_VEH, TOT_VEHICULOS_IMPLICADOS, FACTORES_ATMOSFERICOS, TIPO_VIA | 0,853 | 0,8519 | 0,829 | 0,72 | 0,01 | 10 |

**Tabla 5.52:** Modelo obtenido mediante Random Forest con 14 predictores

| Predicción \ Referencia | X0 | X1 |
|---|---|---|
| X0 | 6814 | 44838 |
| X1 | 1403 | 115700 |

**Tabla 5.53:** Matriz de confusión mediante Random Forest para el test 11

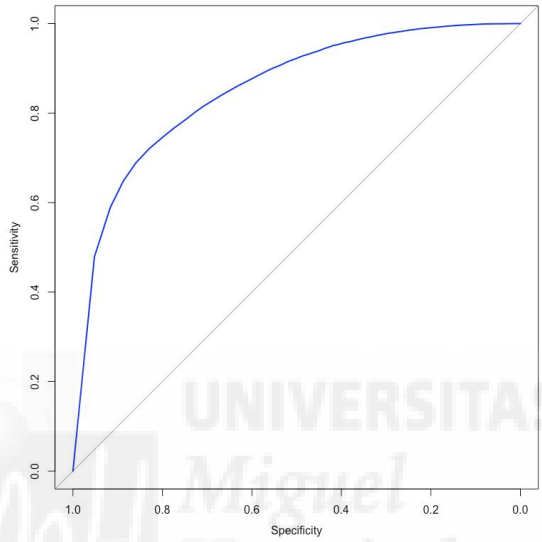



**Figura 5.20:** Área encerrada bajo la curva ROC del modelo Random Forest de 14 predictores (test 11)

víctima fallece o es herida gravemente. La curva ROC para este test se puede observar en la figura 5.20.

En este momento, se replantea de nuevo la importancia de cada predictor. Para ello, se obtiene un nuevo ranking en orden de importancia en tanto por cien de cada uno de ellos, a partir del modelo Random Forest con el *threshold* obtenido, como se puede observar en la tabla 5.54.

### Test 12: Modelo basado en RFE- Random Forest con 14 predictores. Estimación del parámetro *accuracy* del modelo como ayuda para la selección de características

Se pretende comparar el orden devuelto por RFE-RF en función del parámetro *accuracy* que aporta cada predictor al modelo mediante la estimación del área encerrada bajo la curva ROC, obtenida igualmente a través de RFE-RF. La estimación de esta medida ha resultado algo compleja y tediosa debido a que ha sido necesario introducir variable a variable de forma manual en el selector y lanzar su ejecución cada vez, lo que ha conllevado el lanzamiento de un total de 14 ejecuciones.

| Predictor | Overall |
|---|---|
| EDAD | 100,00 |
| HORA | 70,40 |
| ANOMALIA_NINGUNAX3 | 61,53 |
| MES | 58,53 |
| DIASEMANA | 44,91 |
| POSICIONX6 | 32,02 |
| MANIOBRASX52 | 29,51 |
| INTERSECCIONX2 | 26,17 |
| TIPO_VEHICULOX4 | 23,92 |
| MANIOBRASX72 | 18,88 |
| TIPO_VEHICULOX6 | 14,85 |
| TIPO_ACCIDENTEX4 | 12,87 |
| TIPO_ACCIDENTEX16 | 12,24 |
| SEXOX2 | 12,16 |
| SUPERFICIE_CALZADAX3 | 10,60 |
| TIPO_ACCIDENTEX2 | 9,80 |
| TIPO_ACCIDENTEX20 | 9,265 |
| TIPO_ACCIDENTEX12 | 9,09 |
| LUMINOSIDADX4 | 8,79 |

**Tabla 5.54:** Orden de importancia de las primeras 20 clases de los predictores obtenido mediante Random Forest (test 11)

Las variables introducidas en orden han sido las siguientes:

```
'MANIOBRAS', 'EDAD', 'SUPERFICIE_CALZADA', 'HORA','POSICION',
'FACTORES_ATMOSFERICOS', 'TIPO_VEHICULO','ZONA_AGRUPADA',
'MES', 'LUMINOSIDAD', 'TIPO_VIA', 'DIASEMANA', 'SEXO',
'NUMERO_OCUPANTES_VEH'
```

El resultado se puede observar en la figura 5.21.

En el anexo F se encuentra el script implementado con todas las secuencias de control para su consulta, y en el anexo E es posible seguir todo el proceso.

### Test 13: Modelo basado en BayesGLM con 14 predictores

En este test se ha implementado un modelo de 14 predictores también, sin embargo las variables escogidas han sido otras en función del orden de importancia estimado por Random Forest.

La tabla 5.55 muestra la obtención de un área bajo la curva ROC elevada, lo que indica que el modelo aparentemente está correctamente ajustado. Se observa adicionalmente que el modelo no posee *overfitting* debido a que el área obtenida es muy similar para el set de datos de training y testing. Sin embargo, no se consigue obtener una sensibilidad (predicción de clase de accidente grave o fallecimiento) elevada sin penalizar en exceso sobre la clase de especificidad (predicción de clase de accidente leve).

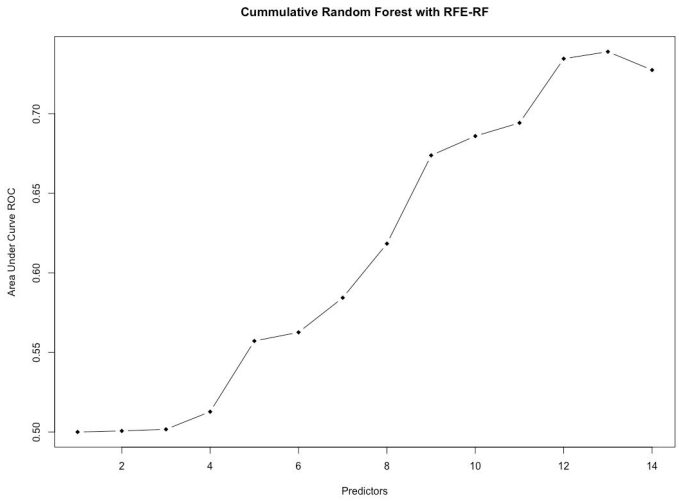



**Figura 5.21:** Determinación mediante RFE-RF del área encerrada bajo la curva ROC en función del número de predictores del modelo

### Test 14: Modelo basado en BayesGLM con 14 predictores

El test 14 es muy similar, en cuanto a descriptores escogidos, al test 13. En él se ha implementado un modelo de 14 predictores también, sin embargo, se ha eliminado la variable referente a factores atmosféricos y se ha conservado aquella referente al número de ocupantes del vehículo. Se puede observar que gracias al uso mixto de Random Forest y BayesGLM los modelos que se obtienen no poseen *overfitting* y son bastante robustos.

Aunque en este test se ha conseguido una mayor sensibilidad, sigue sin obtenerse un ajuste adecuado aunque se modifique el umbral de corte entre sensibilidad y especificidad.

### Test 15: Modelo basado en RFE- Random Forest con 13 predictores

Eliminar la variable 9, 'mes', que provoca una caída en el *accuracy* del modelo, se comprueba en el test 15 que no resultará una buena idea, ya que como se puede apreciar en la figura 5.21, esta variable provoca un aumento muy importante en el área

| Test | Predictores | Modelo | Vars. elim. | ROC Train | ROC Test | Sens | Spec | Threshold | K-fold |
|---|---|---|---|---|---|---|---|---|---|
| 13 | 14 | BayesGLM | ANIO ,NUM_OCUP, TOT_VEHIC_IMP, TIPO_VIA, TIPO_ACC | 0,841 | 0,8459 | 0,730438 | 0,801996 | - | 10 |

**Tabla 5.55:** Modelo obtenido mediante BayesGLM con 14 predictores

| Test | Predictores | Modelo | Vars. elim. | ROC Train | ROC Test | Sens | Spec | Threshold | K-fold |
|---|---|---|---|---|---|---|---|---|---|
| 14 | 14 | BayesGLM | ANIO, TOT_VEHIC_IMP, FAC_ATM, TIPO_VIA, TIPO_ACC | 0,8416856 | 0,8463 | 0,73520288 | 0,79591092 | - | 10 |

**Tabla 5.56:** Modelo obtenido mediante BayesGLM con 14 predictores

bajo la curva ROC. En la tabla 5.57 se muestra un resumen donde se observa la caída del área encerrada bajo la curva ROC.

| Test | Predictores | Modelo | Vars. elim. | ROC Train | ROC Test | Sens | Spec | Threshold | K-fold |
|---|---|---|---|---|---|---|---|---|---|
| 15 | 13 | Random Forest | ANIO, ANOMALIA_NINGUNA,INTERSECCION, TIPO_ACCIDENTE, TOT_VEHICULOS_IMPLICADOS, MES | 0,685274 | - | 0,37503 | 0,87208 | 0,01 | 10 |

**Tabla 5.57:** Modelo obtenido mediante Random Forest con 13 predictores

### Test 16: Modelo basado en RFE- Random Forest con 13 predictores. Estimación de la importancia de la variable *Numero_Ocupantes_Veh*

Puesto que los resultados sin la variable 9, mes, empeoran el modelo, se decide realizar dos tests adicionales:

- Test 12: se mantienen la variable 9 (mes) y la 14 (Numero_Ocupantes_Veh), a pesar de que provocan una disminución del *accuracy*.
- Test 16: se mantiene la variable 9 (mes) y se elimina la 14 (Numero_Ocupantes_Veh) que provocaba una disminución del *accuracy* y del área encerrada bajo la curva ROC, simultáneamente.

Los resultados de ambos tests se pueden encontrar en el anexo E en el momento de la inserción de 14 y 13 variables, respectivamente.

En las tablas 5.58 y 5.59 se muestra el resumen.

| Test | Predictores | Modelo | Vars. elim. | ROC Train | ROC Test | Sens | Spec | Threshold | K-fold |
|---|---|---|---|---|---|---|---|---|---|
| 12 | 14 | RFE-RF | ANIO, ANOMALIA_NINGUNA, INTERSECCION, TIPO_ACCIDENTE, TOT_VEHICULOS_IMPLICADOS | 0,7274 | - | 0,5240 | 0,8038 | 0,01 | 10 |

**Tabla 5.58:** Modelo obtenido mediante RFE-RF con 14 predictores

### Test 17: Modelo basado en BayesGLM con 13 predictores

Se realiza un nuevo test donde se elimina la variable *TIPO_ACCIDENTE* para implementar un nuevo modelo mediante BayesGLM.

En la tabla 5.60 se puede observar el resultado obtenido, donde el área bajo la curva ROC no se ha visto prácticamente mermada en relación al test 10. Por ello, se decide prescindir de dicha variable.

En este test, el tiempo empleado en la ejecución se ha conseguido reducir en 30 minutos respecto al test 10:

- Tiempo empleado en la ejecución: 3.839 seg ($\approx 1h$)

El área encerrada bajo la curva ROC se puede observar en la figura 5.22.

En este momento se replantea de nuevo la importancia de cada predictor. Para ello, se obtiene la importancia en tanto por cien de cada uno de ellos a partir del modelo BayesGLM con el threshold obtenido, como se puede observar en la tabla 5.61.

### Test 18: Modelo basado en BayesGLM con 12 predictores

- Tiempo empleado en la ejecución: 3.504 seg ($\approx 1h$)

La gráfica del área encerrada bajo la curva ROC se puede observar en la figura 5.23.

Parámetros adicionales devueltos por el modelo:

| Test | Predictores | Modelo | Vars. elim. | ROC Train | ROC Test | Sens | Spec | Threshold | K-fold |
|---|---|---|---|---|---|---|---|---|---|
| 16 | 13 | RFE-RF | ANIO, ANOMALIA_-NINGUNA, INTER-SECCION, TIPO_-ACCIDENTE, TOT_-VEHICULOS_IMPLI-CADOS, NUM_OCU-PANTES_VEHIC | 0,7389 | - | 0,55819 | 0,7906 | 0,01 | 10 |

**Tabla 5.59:** Modelo obtenido mediante RFE-RF con 13 predictores

| Test | Predictores | Modelo | Vars. elim. | ROC Train | ROC Test | Sens | Spec | Threshold | K-fold |
|---|---|---|---|---|---|---|---|---|---|
| 17 | 13 | BayesGLM | ANIO, NUMERO_-OCUPANTES_VEH, TOT_VEHICULOS_-IMPLICADOS, FACTO-RES_ATMOSFERICOS, TIPO_VIA, TIPO_AC-CIDENTE | 0,841 | 0,846 | 0,732 | 0,798 | 0,0488 | 10 |

**Tabla 5.60:** Modelo obtenido mediante BayesGLM con 13 predictores

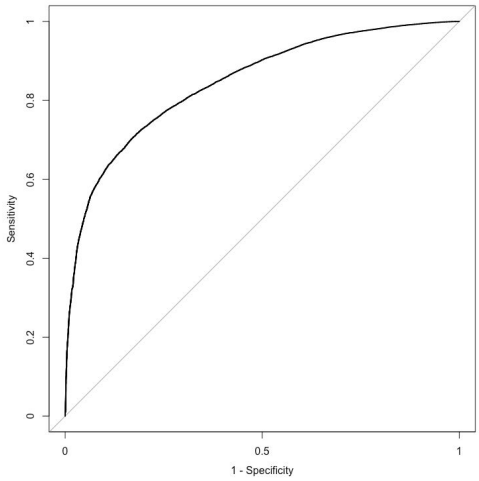



**Figura 5.22:** Área encerrada bajo la curva ROC del modelo BayesGLM de 13 predictores (test 17)

| Predictor | Importance |
|---|---|
| MANIOBRAS | 100,00 |
| TIPO_VEHICULO | 98,28 |
| ZONA_AGRUPADA | 95,84 |
| SEXO | 73,66 |
| INTERSECCION | 67,25 |
| HORA | 65,04 |
| SUPERFICIE_CALZADA | 56,02 |
| MES | 52,91 |
| LUMINOSIDAD | 44,33 |
| EDAD | 37,11 |
| DIASEMANA | 36,47 |
| ANOMALIA_NINGUNA | 32,21 |
| POSICION | 0,00 |

**Tabla 5.61:** Orden de importancia de los predictores obtenido mediante BayesGLM (test 17)

- Accuracy : 0,7199
- Kappa : 0,1498

La clasificación de los predictores en orden de importancia viene dado en la tabla 5.64.

| Test | Predictores | Modelo | Vars. elim. | ROC Train | ROC Test | Sens | Spec | Threshold | K-fold |
|---|---|---|---|---|---|---|---|---|---|
| 18 | 12 | BayesGLM | ANIO,NUM_OCUP, TOT_VEHIC_IMP, FAC_ATM, TIPO_VIA, TIPO_ACC, POSICION | 0,8401 | 0,8449 | 0,820 | 0,714 | 0,046 | 10 |

**Tabla 5.62:** Modelo obtenido mediante BayesGLM con 12 predictores

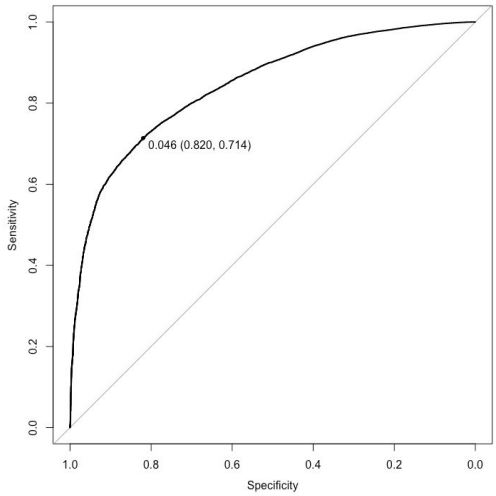



**Figura 5.23:** Determinación mediante BayesGLM del área encerrada bajo la curva ROC en función del número de predictores del modelo

| Predicción \ Referencia | X0 | X1 |
|---|---|---|
| X0 | 6718 | 45764 |
| X1 | 1499 | 114774 |

**Tabla 5.63:** Matriz de confusión mediante BayesGLM para el test 18

| Predictor | Importance |
|---|---|
| MANIOBRAS | 100,000 |
| TIPO_VEHICULO | 97,455 |
| ZONA_AGRUPADA | 93,870 |
| SEXO | 61,144 |
| INTERSECCION | 51,690 |
| HORA | 48,429 |
| SUPERFICIE_CALZADA | 35,128 |
| MES | 30,532 |
| LUMINOSIDAD | 17,883 |
| EDAD | 7,236 |
| DIASEMANA | 6,291 |
| ANOMALIA_NINGUNA | 0,000 |

**Tabla 5.64:** Orden de importancia de los predictores obtenido mediante BayesGLM (test 18)

### Test 19: Modelo basado en BayesGLM con 12 predictores

En el test 18 se podía observar que eliminando la variable posición se penalizaba en una cantidad muy pequeña el área que se obtenía bajo la curva ROC, por lo que parecía coherente su eliminación con la finalidad de reducir tiempo de ejecución del modelo. Sin embargo, para este estudio, la necesidad de trabajar con este predictor

resultará fundamental puesto que la severidad del accidente se estima para cada persona accidentada, por lo que conocer su posición en el vehículo es importante.

Por todo ello, en este nuevo test se conservará este predictor y se eliminará aquel cuyo nivel de importancia era cero en la tabla 5.64, es decir, 'ANOMALIA_NINGUNA'.

| Test | Predictores | Modelo | Vars. elim. | ROC Train | ROC Test | Sens | Spec | Threshold | K-fold |
|---|---|---|---|---|---|---|---|---|---|
| 19 | 12 | BayesGLM | ANIO, NUM_OCUP, TOT_VEHIC_IMP, FAC_ATM, TIPO_VIA, TIPO_ACC, ANOMALIA_NINGUNA | 0,809614 | 0,8114 | 0,777 | 0,702 | 0,049 | 10 |

**Tabla 5.65:** Modelo obtenido mediante BayesGLM con 12 predictores

- Tiempo empleado en la ejecución: 3.845 seg ($\approx 1h$)

La gráfica del área encerrada bajo la curva ROC se puede observar en la figura 5.24.

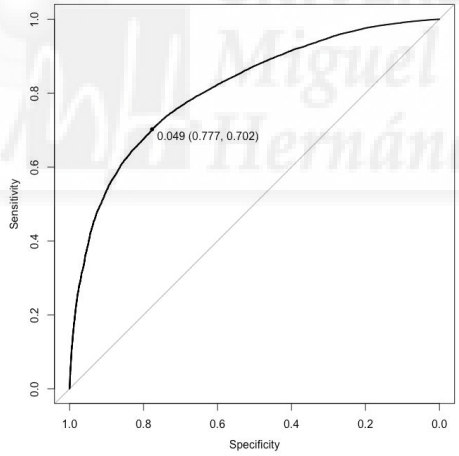



**Figura 5.24:** Determinación mediante BayesGLM del área encerrada bajo la curva ROC en función del número de predictores del modelo

| Predicción \ Referencia | X0 | X1 |
|---|---|---|
| X0 | 6362 | 47505 |
| X1 | 1855 | 113033 |

**Tabla 5.66:** Matriz de confusión mediante BayesGLM para el test 19

Parámetros adicionales devueltos por el modelo:

- Accuracy : 0,7075
- Kappa : 0,1316

Como se puede observar en la tabla resumen del modelo 5.65, la penalización por eliminar este predictor es muy elevada, por lo que se toma la decisión de mantenerlo en el set de datos.

### Test 20: Modelo basado en BayesGLM con 12 predictores

Al obtener una penalización excesiva por la eliminación del predictor del test 19, se toma otra decisión: se eliminará el siguiente predictor de menor orden de importancia obtenido en la tabla 5.64 y se mantiene 'ANOMALIA_NINGUNA'. Este predictor eliminado es 'DIASEMANA'.

La tabla 5.67 muestra un resumen del modelo.

| Test | Predictores | Modelo | Vars. elim. | ROC Train | ROC Test | Sens | Spec | Threshold | K-fold |
|---|---|---|---|---|---|---|---|---|---|
| 20 | 12 | BayesGLM | ANIO,NUM_OCUP, TOT_VEHIC_IMP, FAC_ATM, TIPO_VIA, TIPO_ACC, DIASEMANA | 0,8404322 | 0,8454 | 0,838 | 0,695 | 0,043 | 10 |

**Tabla 5.67:** Modelo obtenido mediante BayesGLM con 12 predictores

- Tiempo empleado en la ejecución: 4.185 seg ($\approx 1:15h$)

La gráfica del área encerrada bajo la curva ROC se puede observar en la figura 5.25.

| Predicción \ Referencia | X0 | X1 |
|---|---|---|
| X0 | 6868 | 48781 |
| X1 | 1349 | 111757 |

**Tabla 5.68:** Matriz de confusión mediante BayesGLM para el test 20

Parámetros adicionales devueltos por el modelo:

- Accuracy : 0,7029
- Kappa : 0,1423

La clasificación de los predictores en orden de importancia vendrá dado por la tabla 5.69.

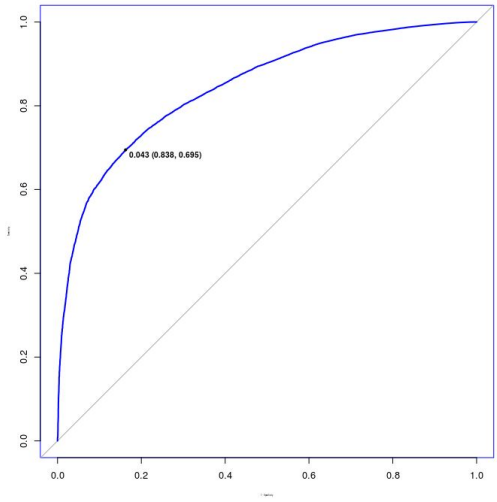



**Figura 5.25:** Determinación mediante BayesGLM del área encerrada bajo la curva ROC en función del número de predictores del modelo

| Predictor | Importance |
|---|---|
| MANIOBRAS | 100,00 |
| TIPO_VEHICULO | 98,28 |
| ZONA_AGRUPADA | 95,84 |
| SEXO | 73,66 |
| INTERSECCION | 67,25 |
| HORA | 65,04 |
| SUPERFICIE_CALZADA | 56,02 |
| MES | 52,91 |
| LUMINOSIDAD | 44,33 |
| EDAD | 37,11 |
| ANOMALIA_NINGUNA | 32,21 |
| POSICION | 0,00 |

**Tabla 5.69:** Orden de importancia de los predictores obtenido mediante BayesGLM (test 20)

### Test 21: Modelo basado en BayesGLM con 11 predictores

A pesar de que se ha obtenido un modelo en el test anterior con un área bajo la curva ROC estable y elevada, y no posee *overfitting*, se tratará de ajustar algo más mediante la eliminación de un nuevo predictor: 'EDAD'.

El resumen del modelo se puede observar en la tabla 5.70. En ella se observa que la eliminación de esta característica penaliza levemente sobre la curva del modelo.

- Tiempo empleado en la ejecución: 3.703 seg ($\approx 1h$)

Adicionalmente, la gráfica del área encerrada bajo la curva ROC se puede observar en la figura 5.26.

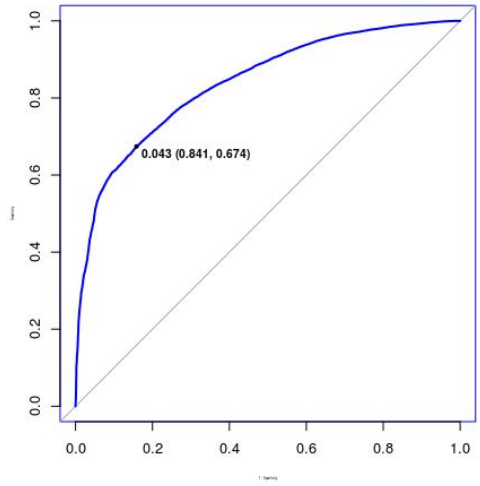



**Figura 5.26:** Determinación mediante BayesGLM del área encerrada bajo la curva ROC en función del número de predictores del modelo

Los parámetros adicionales devueltos por el modelo son:

- Accuracy : 0,6815
- Kappa : 0,1303

La clasificación de los predictores en orden de importancia vendrá dado por la tabla 5.72.

Este test presenta una nueva reflexión acerca de si se mantiene o no el predictor de EDAD, y en vista de que los tiempos de cálculo no aumentan demasiado (1.800 frente

| Test | Predictores | Modelo | Vars. elim. | ROC Train | ROC Test | Sens | Spec | Threshold | K-fold |
|---|---|---|---|---|---|---|---|---|---|
| 21 | 11 | BayesGLM | ANIO,NUM_OCUP, TOT_VEHIC_IMP, FAC_ATM, TIPO_VIA, TIPO_ACC, DIASEMANA, EDAD | 0,8345614 | 0,8397 | 0,841 | 0,674 | 0,043 | 10 |

**Tabla 5.70:** Modelo obtenido mediante BayesGLM con 11 predictores

| Predicción \ Referencia | X0 | X1 |
|---|---|---|
| X0 | 6915 | 52444 |
| X1 | 1302 | 108094 |

**Tabla 5.71:** Matriz de confusión mediante BayesGLM para el test 21

| Predictor | Importance |
|---|---|
| MANIOBRAS | 100,00 |
| TIPO_VEHICULO | 98,28 |
| ZONA_AGRUPADA | 95,84 |
| SEXO | 73,66 |
| INTERSECCION | 67,25 |
| HORA | 65,04 |
| SUPERFICIE_CALZADA | 56,02 |
| MES | 52,91 |
| LUMINOSIDAD | 44,33 |
| ANOMALIA_NINGUNA | 32,21 |
| POSICION | 0,00 |

**Tabla 5.72:** Orden de importancia de los predictores obtenido mediante BayesGLM (test 21)

a unos 2.000 segundos), se toma la decisión de mantener este predictor en el set de datos.

### Test 22: Modelo basado en BayesGLM con 11 predictores

Se realizará un nuevo experimento donde, partiendo del test 21 (tabla 5.70) y del test 18 (tabla 5.62) , se eliminará de nuevo la característica 'ANOMALIA_NINGUNA'.

El resumen del modelo se puede observar en la tabla 5.73.

| Test | Predictores | Modelo | Vars. elim. | ROC Train | ROC Test | Sens | Spec | Threshold | K-fold |
|---|---|---|---|---|---|---|---|---|---|
| 22 | 11 | BayesGLM | ANIO,NUM_OCUP, TOT_VEHIC_IMP, FAC_ATM, TIPO_VIA, TIPO_ACC, POSICION, ANOMALIA_NINGUNA | 0,8086 | 0,8103 | 0,782 | 0,697 | 0,048 | 10 |

**Tabla 5.73:** Modelo obtenido mediante BayesGLM con 11 predictores

- Tiempo empleado en la ejecución: 3.461 seg ($\approx 1h$)

La figura 5.27 muestra una comparativa entre las dos curvas ROC de los tests 18 y 22. En color rojo, la gráfica de la curva ROC obtenida para el test 22, donde se ha eliminado la característica *ANOMALIA_NINGUNA*, mientras que la curva dibujada en color azul corresponde al test 18. En esta figura se observa claramente que el área obtenida bajo la curva es menor para el test 22 que para el 18.

Como se puede observar, la penalización por eliminar este predictor del modelo es elevada. Se concluye, por tanto, que su inclusión es necesaria.

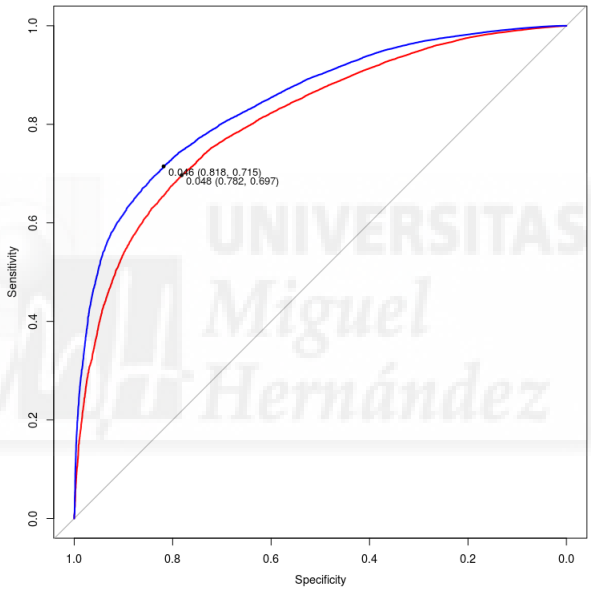



**Figura 5.27:** Determinación mediante BayesGLM del área encerrada bajo la curva ROC en función del número de predictores del modelo

| Test | Preds | Modelo | Vars. Eliminadas | AUC/ROC Train | AUC/ROC Test | Sens (X0) | Spec (X1) | Threshold | Distancia | K-Fold |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 19 | RF | - | 0.866 | - | 0.877 | 0.679 | 0.01 | 0.3435 | 10 |
| 2 | 18 | RF | ANIO | 0.857 | - | 0.847 | 0.705 | 0.01 | 0.3321 | 10 |
| 3 | 18 | BayesGLM | ANIO | - | 0.8727 | - | - | - | - | 10 |
| 4 | 18 | RF | ANIO | 0.829 | - | 0.803 | 0.707 | - | - | 10 |
| 5 | 17 | RF | ANIO, ANOMALIA_NINGUNA | 0.82974 | 0.8253 | 0.8036 | 0.707 | 0.01 | 0.3525024 | 10 |
| 6 | 17 | BayesGLM | ANIO,NUM_OCUP | - | 0.8721 | - | - | 0.0505 | - | 10 |
| 7 | 16 | BayesGLM | ANIO,NUM_OCUP, TOT_VEHIC_IMP | 0.8700 | 0.8715 | 0.775 | 0.815 | 0.0516 | - | 10 |
| 8 | 15 | BayesGLM | ANIO,NUM_OCUP, TOT_VEHIC_IMP, FAC_ATM | 0.8706 | 0.8714 | 0.773 | 0.8185 | 0.0507 | - | 10 |
| 9 | 15 | RF | ANIO,NUM_OCUP, TOT_VEHIC_IMP, TIPO_VIA | 0.8532972 | 0.8519 | 0.721 | 0.829 | 0.09 | 0.3289890 | 10 |
| 10 | 14 | BayesGLM | ANIO,NUM_OCUP, TOT_VEHIC_IMP, FAC_ATM, TIPO_VIA | 0.869 | 0.871 | 0.7705 | 0.8203 | 0.0501 |  | 10 |
| 11 | 14 | RF | ANIO,NUM_OCUP, TOT_VEHIC_IMP, FAC_ATM, TIPO_VIA | 0.853 | 0.8519 | 0.829 | 0.72 | 0.01 | 0.32898 | 10 |
| 12 | 14 | RF | ANIO, ANOMALIA_NINGUNA,INTERSECCION, TIPO_ACCIDENTE, TOT_VEHICULOS_IMPLICADOS | 0.7274284 | - | 0.5240632104 | 0.8038025 | 0.010 | 0.5147993 | 10 |
| 13 | 14 | BayesGLM | ANIO,NUM_OCUP, TOT_VEHIC_IMP, TIPO_VIA, TIPO_ACC | 0.841 | 0.8459 | 0.73043765 | 0.80199586 | - | - | 10 |
| 14 | 14 | BayesGLM | ANIO, TOT_VEHIC_IMP, FAC_ATM, TIPO_VIA, TIPO_ACC | 0.8416856 | 0.8463 | 0.73520288 | 0.79591092 | - | - | 10 |
| 15 | 13 | RF | ANIO, ANOMALIA_NINGUNA,INTERSECCION, TIPO_ACCIDENTE, TOT_VEHICULOS_IMPLICADOS,MES | 0.685274 | - | 0.3750304506 | 0.8720821 | 0.0100 | 0.6379293 | 10 |
| 16 | 13 | RF | ANIO, ANOMALIA_NINGUNA,INTERSECCION, TIPO_ACCIDENTE, TOT_VEHICULOS_IMPLICADOS,NUM_OCUPANTES_VEHIC | 0.7389 | - | 0.55819 | 0.7906 | 0.0100 | 0.4889072 | 10 |
| 17 | 13 | BayesGLM | ANIO,NUM_OCUP, TOT_VEHIC_IMP, FAC_ATM, TIPO_VIA, TIPO_ACC | 0.841 | 0.846 | 0.732 | 0.798 | 0.04885 | - | 10 |
| 18 | 12 | BayesGLM | ANIO,NUM_OCUP, TOT_VEHIC_IMP, FAC_ATM, TIPO_VIA, TIPO_ACC, POSICION | 0.8401795 | 0.8449 | 0.820 | 0.714 | 0.046 | - | 10 |
| 19 | 12 | BayesGLM | ANIO,NUM_OCUP, TOT_VEHIC_IMP, FAC_ATM, TIPO_VIA, TIPO_ACC, ANOMALIA_NINGUNA | 0.809614 | 0.8114 | 0.777 | 0.702 | 0.049 | - | 10 |
| 20 | 12 | BayesGLM | ANIO,NUM_OCUP, TOT_VEHIC_IMP, FAC_ATM, TIPO_VIA, TIPO_ACC, DIASEMANA | 0.8404322 | 0.8454 | 0.838 | 0.695 | 0.043 | - | 10 |
| 21 | 11 | BayesGLM | ANIO,NUM_OCUP, TOT_VEHIC_IMP, FAC_ATM, TIPO_VIA, TIPO_ACC, DIASEMANA, EDAD | 0.8345614 | 0.8397 | 0.841 | 0.674 | 0.043 | - | 10 |
| 22 | 11 | BayesGLM | ANIO,NUM_OCUP, TOT_VEHIC_IMP, FAC_ATM, TIPO_VIA, TIPO_ACC, POSICION, ANOMALIA_NINGUNA | 0.8086628 | 0.8103 | 0.778 | 0.701 | 0.049 | - | 10 |
| 23 | 11 | RF | ANIO, ANOMALIA_NINGUNA, INTERSECCION, TIPO_ACCIDENTE, TOT_VEHICULOS_IMPLICADOS, LUMINOSIDAD, SUPERFICIE_CALZADA, FACTORES_ATMOSFERICOS, | 0.6832 | - | 0.325 | 0.895 | 0.010 | 0.682576 | 10 |

**Figura 5.28:** Figura resumen de todos los tests realizados para España

### 5.4.4. Descripción del modelo. Curva ROC y matriz de confusión

Tras haber realizado todas las ejecuciones finales con motivo de simplificar el modelo lo máximo posible, se eliminarán parámetros o niveles de un factor que sean redundantes. El procedimiento para la simplificación del modelo se realizará a través de los siguientes pasos.

1. Eliminando las variables explicativas que no sean significativas.
2. Unificando niveles de factores no significativos con el nivel del factor equivalente al Intercept (término independiente de la ecuación de regresión).

Una vez finalizado todo el proceso, la ecuación que define nuestro modelo vendrá dada por el *logit* obtenido:

$$logit(\pi(\hat{x})) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \tag{5.5}$$

Sustituyendo cada coeficiente obtenemos la siguiente ecuación de nuestro modelo:

$$\begin{aligned} logit(\pi(\hat{x})) = {} & 3,646457 - 0,015898 \cdot EDAD - 2,603823 \cdot ANOMALIA\_NINGUNAX3 \\ & + 0,021489 \cdot HORA + 0,006960 \cdot MES - 0,204760 \cdot POSICIONX2 - 0,291019 \cdot POSICIONX3 \\ & - 0,290979 \cdot POSICIONX4 - 0,360185 \cdot POSICIONX5 - 0,376290 \cdot POSICIONX8 \\ & - 1,359436 \cdot POSICIONX9 + 0,277803 \cdot MANIOBRASX11 + 0,594241 \cdot MANIOBRASX12 \\ & + 0,406582 \cdot MANIOBRASX13 + 1,555601 \cdot MANIOBRASX15 + 0,502247 \cdot MANIOBRASX17 \\ & + 0,441808 \cdot MANIOBRASX18 + 0,604077 \cdot MANIOBRASX19 + 2,196401 \cdot MANIOBRASX20 \\ & + 0,281859 \cdot MANIOBRASX21 - 0,121468 \cdot MANIOBRASX22 + 0,762910 \cdot MANIOBRASX23 \\ & + 0,900468 \cdot MANIOBRASX24 + 0,308351 \cdot MANIOBRASX25 + 0,587099 \cdot MANIOBRASX26 \\ & + 0,906092 \cdot MANIOBRASX29 - 0,078243 \cdot MANIOBRASX3 + 0,789679 \cdot MANIOBRASX31 \\ & + 0,165768 \cdot MANIOBRASX41 + 0,961114 \cdot MANIOBRASX42 + 0,755005 \cdot MANIOBRASX43 \\ & - 0,226516 \cdot MANIOBRASX5 + 1,763338 \cdot MANIOBRASX51 + 1,839539 \cdot MANIOBRASX52 \\ & + 0,784178 \cdot MANIOBRASX6 + 0,414279 \cdot MANIOBRASX71 + 2,673937 \cdot MANIOBRASX72 \\ & + 0,477156 \cdot MANIOBRASX77 + 1,469989 \cdot MANIOBRASX8 - 0,661662 \cdot TIPO\_VEHICULOX13 \\ & - 0,619892 \cdot TIPO\_VEHICULOX15 + 0,549678 \cdot TIPO\_VEHICULOX16 \\ & + 0,665469 \cdot TIPO\_VEHICULOX17 - 1,899010 \cdot TIPO\_VEHICULOX18 \\ & - 2,000156 \cdot TIPO\_VEHICULOX2 - 1,096894 \cdot TIPO\_VEHICULOX20 \\ & - 1,046202 \cdot TIPO\_VEHICULOX21 - 2,212017 \cdot TIPO\_VEHICULOX3 \\ & - 2,283154 \cdot TIPO\_VEHICULOX4 - 0,287688 \cdot TIPO\_VEHICULOX6 \\ & - 0,551663 \cdot TIPO\_VEHICULOX7 - 1,365981 \cdot TIPO\_VEHICULOX8 \\ & - 0,288959 \cdot TIPO\_VEHICULOX9 - 0,570766 \cdot INTERSECCIONX2 \\ & + 0,126221 \cdot INTERSECCIONX6 + 0,056589 \cdot INTERSECCIONX7 \\ & + 0,623237 \cdot INTERSECCIONX8 + 0,267545 \cdot INTERSECCIONX9 \\ & + 0,185341 \cdot INTERSECCIONX999 + 1,034802 \cdot ZONA\_AGRUPADAX2 \\ & + 0,204573 \cdot SEXOX2 - 0,243769 \cdot LUMINOSIDADX2 - 0,262164 \cdot LUMINOSIDADX3 \\ & - 0,553158 \cdot LUMINOSIDADX4 + 0,302556 \cdot SUPERFICIE\_CALZADAX3 \\ & + 0,442031 \cdot SUPERFICIE\_CALZADAX4 + 0,610606 \cdot SUPERFICIE\_CALZADAX5 \\ & + 1,527321 \cdot SUPERFICIE\_CALZADAX6 \end{aligned} \tag{5.6}$$

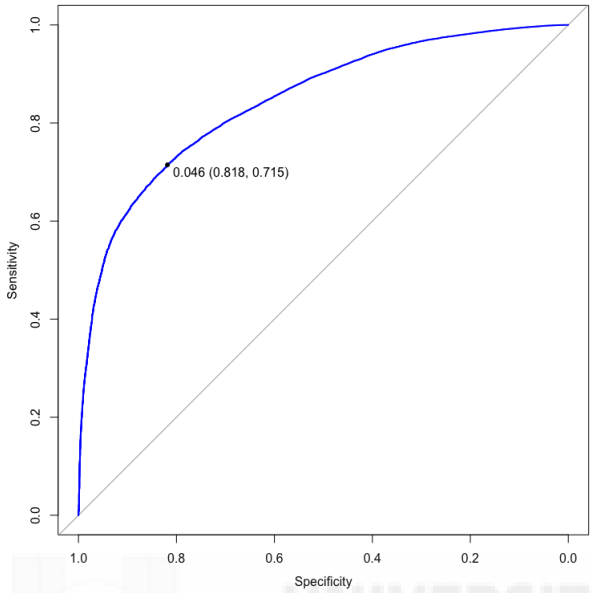



**Figura 5.29:** Área encerrada bajo la curva ROC del modelo obtenido

| Predicción \ Referencia | X0 | X1 |
|---|---|---|
| X0 | 6713 | 45679 |
| X1 | 1504 | 114859 |

**Tabla 5.74:** Matriz de confusión del modelo final mediante BayesGLM con dataset de test para $umbral = 0,046$

## 5.5. Interpretación del modelo

Se propone el uso de un clasificador supervisado mixto basado en Random Forest y regresión logística bayesiana (BayesGLM). La selección global de características vendrá dada por el primero de ellos a través de diversos árboles, para reducir la posibilidad de *overfitting*, y mediante la técnica de Recursive Feature Elimination (RFE). El uso del clasificador lineal bayesiano se realizará debido a la posibilidad de interpretar los resultados de forma sencilla, así como de obtener distintas probabilidades de *'Casualty_Severity'* a través del modelo.

Para el modelo que nos ocupa se han efectuado 5 ejecuciones con 8 iteraciones en cada ejecución, hasta obtener únicamente las variables que eran relevantes, es decir, se han eliminado aquellas con el término menos significativo $(p-valor > 0{,}1)$. Adicionalmente, se han unificado aquellos niveles de factores no significativos con el nivel de factor equivalente al *Intercept*, tal y como se comentaba anteriormente.

Teniendo en cuenta que el valor de la variable independiente *Casualty_Severity* es dicotómica, la probabilidad de ocurrencia de cada uno de sus dos valores viene dada por:

$$Prob = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k}} \text{ , si Casualty\_Severity=1} \quad (5.7)$$

$$Prob = \frac{1}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k}} \text{ , si Casualty\_Severity=0} \quad (5.8)$$

En base a nuestro modelo, la probabilidad inicial de que el accidente sea grave o provoque un fallecimiento sin que influya ninguna variable es:

$$Prob = \frac{1}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k}} = \frac{1}{1+e^{3,64}} = 0,02558079 \quad (5.9)$$

Es decir, la probabilidad inicial de que el índice sea grave, siendo cero el resto de variables, es del 2.56 %.

Por otro lado, la probabilidad de que el accidente sea leve es de:

$$Prob = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k}} = \frac{e^{3,64}}{1+e^{3,64}} = 0,9744192 \quad (5.10)$$

Es decir, la probabilidad inicial de que el índice sea leve, sin tener en cuenta el resto de variables, es del 97,44 %.

Para cada nivel de predictor del modelo escogido, se calcularán dos nuevos parámetros relativos a la probabilidad de aumento de la gravedad:

- Probabilidad estimada de gravedad o fallecimiento del nivel (Prob. grave): consistirá en la estimación de la probabilidad del *Intercept* junto con el nivel del factor o variable para cada caso.
- Probabilidad estimada de incremento de la gravedad (Incr. grave): consistirá en la estimación del aumento de la gravedad a partir de la probabilidad de gravedad o fallecimiento. El cálculo viene definido por la ecuación 5.11.

$$Incr.grave = ProbGraveNivel - ProbGraveIntercept \quad (5.11)$$

| | Estimate | Std. Error | z value | Pr(>\|z\|) | Prob. grave | Incr. grave |
|---|---|---|---|---|---|---|
| Edad | -0,015898 | 0,000377 | -42,167 | <2e-16 | 2,5817 % | 1,56 % |

**Tabla 5.75:** Descripción de la variable edad en el modelo estimado

### 5.5.1. Influencia de la edad

La variable edad posee los coeficientes en el modelo indicados por la tabla 5.75.

En resumen, esta variable influye negativamente a medida que aumenta, es decir, a mayor edad, mayor riesgo se posee de fallecer:

- -0,015898 es el cambio esperado en el logit al aumentar una unidad la edad, supuestas estables el resto de las variables en el modelo.
- exp(0,015898) es la razón de Odds al aumentar una unidad la edad, supuestas estables el resto de las variables en el modelo.

Es decir, al aumentar una unidad la edad, la Odds o ventaja de accidente grave o fallecimiento se multiplica por $exp(0,015898) = 1,016025$;

### 5.5.2. Influencia del tipo de maniobra

Como se puede observar en la tabla 5.76, las tres primeras maniobras hacen que aumente la probabilidad de que la vida de los ocupantes del vehículo se ponga en peligro. Sin embargo, la DGT no tiene definida la primera de ellas en su diccionario. Cruzar una intersección es la segunda maniobra que hace que aumente el riesgo de accidente con fallecimiento o gravedad de los ocupantes. Los adelantamientos por la izquierda son el tercer tipo de maniobras que aumentan dicho riesgo.

La tabla 5.77 muestra la correspondencia del nivel del factor con su significado.

Dicha tabla muestra la importancia de las maniobras que harán que aumente o disminuya la probabilidad de que alguno de los ocupantes fallezca o tenga un accidente grave.

Se puede observar lo que se comentaba anteriormente: la maniobra *'22. Cruzando intersección'* es una de las que poseen mayor importancia debido a la probabilidad de que se ponga en riesgo la vida de los ocupantes. La existencia o no de esta maniobra provocará que se sume el término de ventaja de accidente grave o fallecimiento, cuyo valor es: $exp(0,121468) = 1,129153$.

La realización de la maniobra *'3. Adelantamiento por la izquierda'* provocará que se sume el término de ventaja de accidente grave o fallecimiento, cuyo valor es: $exp(0,078243) = 1,081385$.

| | Estimate | Std. Error | z value | Pr(>\|z\|) | Prob. grave | Incr. grave |
|---|---|---|---|---|---|---|
| MANIOBRASX5 | -0,226516 | 0,060242 | -3,760 | 0,000170 *** | 3,17 % | 24,6 % |
| MANIOBRASX22 | -0,121468 | 0,039335 | -3,088 | 0,002015 ** | 2,86 % | 12,5 % |
| MANIOBRASX3 | -0,078243 | 0,027404 | -2,855 | 0,004301 ** | 2,74 % | 7,9 % |
| MANIOBRASX72 | 2,673937 | 0,038912 | 68,718 | <2e-16 *** | 0,18 % | -92 % |
| MANIOBRASX20 | 2,196401 | 0,176201 | 12,465 | <2e-16 *** | 0,28 % | -88,6 % |
| MANIOBRASX52 | 1,839539 | 0,025559 | 71,972 | <2e-16 *** | 0,41 % | -83,8 % |
| MANIOBRASX51 | 1,763338 | 0,088745 | 19,870 | <2e-16 *** | 0,44 % | -82,4 % |
| MANIOBRASX15 | 1,555601 | 0,188999 | 8,231 | <2e-16 *** | 0,54 % | -78,5 % |
| MANIOBRASX42 | 0,961114 | 0,232957 | 4,126 | 3,70e-05 *** | 0,99 % | -61,1 % |
| MANIOBRASX24 | 0,900468 | 0,092167 | 9,770 | <2e-16 *** | 1,05 % | -58,7 % |
| MANIOBRASX31 | 0,789679 | 0,167580 | 4,712 | 2,45e-06 *** | 1,17 % | -53,96 % |
| MANIOBRASX6 | 0,784178 | 0,177643 | 4,414 | 1,01e-05 *** | 1,17 % | -53,71 % |
| MANIOBRASX19 | 0,604077 | 0,083803 | 7,208 | 5,66e-13 *** | 1,40 % | -44,70 % |
| MANIOBRASX23 | 0,762910 | 0,119136 | 6,404 | 1,52e-10 *** | 1,20 % | -52,72 % |
| MANIOBRASX43 | 0,755005 | 0,103323 | 7,307 | 2,73e-13 *** | 1,21 % | -52,36 % |
| MANIOBRASX12 | 0,594241 | 0,050725 | 11,715 | <2e-16 *** | 1,42 % | -44,17 % |
| MANIOBRASX17 | 0,502247 | 0,087648 | 5,730 | 1,00e-08 *** | 1,55 % | -38,86 % |
| MANIOBRASX77 | 0,477156 | 0,085531 | 5,579 | 2,42e-08 *** | 1,59 % | -37,34 % |
| MANIOBRASX18 | 0,441808 | 0,108457 | 4,074 | 4,63e-05 *** | 1,65 % | -35,12 % |
| MANIOBRASX71 | 0,414279 | 0,043521 | 9,519 | <2e-16 *** | 1,69 % | -33,34 % |
| MANIOBRASX13 | 0,406582 | 0,120547 | 3,373 | 0,000744 *** | 1,71 % | -32,83 % |
| MANIOBRASX11 | 0,277803 | 0,056027 | 4,958 | 7,11e-07 *** | 1,93 % | -23,78 % |
| MANIOBRASX21 | 0,281859 | 0,043528 | 6,475 | 9,46e-11 *** | 1,94 % | -24,08 % |
| MANIOBRASX8 | 1,469989 | 0,469554 | 3,131 | 0,001744 ** | 0,60 % | -76,55 % |
| MANIOBRASX29 | 0,906092 | 0,276193 | 3,281 | 0,001036 ** | 1,04 % | -58,97 % |
| MANIOBRASX25 | 0,308351 | 0,139328 | 2,213 | 0,026888 * | 1,88 % | -26,03 % |
| MANIOBRASX41 | 0,165768 | 0,069001 | 2,402 | 0,016288 * | 2,16 % | -14,95 % |
| MANIOBRASX26 | 0,587099 | 0,302495 | 1,941 | 0,052276 , | 1,43 % | -43,77 % |

**Tabla 5.76:** Descripción de la variable *Maniobras* en el modelo estimado

### 5.5.3. Influencia de las anomalías en el vehículo

Como se puede observar en la tabla 5.78 el poseer alguna anomalía en el vehículo no es un factor excesivamente importante o al menos no es determinante para que el accidente pueda ser grave, ya que el modelo devuelve únicamente como importante el nivel *'ANOMALIA_NINGUNAX3'*. Como el lector podía imaginar, no poseer ninguna anomalía en el vehículo no exime del riesgo de que el ocupante sufra un accidente grave o pueda producirle incluso la muerte, tal y como podemos ver en dicha tabla.

### 5.5.4. Influencia de la oposición de los ocupantes del vehículo

En la tabla 5.79 se puede observar la influencia que posee ocupar una determinada posición en un vehículo en relación con que el ocupante pueda fallecer o acabar herido grave. Los tres asteriscos por nivel del factor indican que ese nivel posee bastante importancia dentro del modelo.

Cuando la posición del ocupante de un vehículo es *'9. PASAJEROS DE PIE'*, por ejemplo en un autobús, dicha posición provocará que se sume el término de ventaja de accidente grave o fallecimiento, cuyo valor es:

$$exp(1,359436) = 3,893996$$

| Nivel de factor | Descripción |
|---|---|
| 0 | SIN DATO |
| 1 | SIGUIENDO LA RUTA |
| 2 | ADELANTANDO POR LA DERECHA |
| 3 | ADELANTANDO POR LA IZQUIERDA |
| 11 | GIRANDO O SALIENDO HACIA OTRA VÍA O ACCESO POR LA DERECHA |
| 12 | GIRANDO O SALIENDO HACIA OTRA VÍA O ACCESO POR LA IZQUIERDA |
| 13 | GIRANDO EN 'U' |
| 21 | INCORPORÁNDOSE DESDE OTRA VÍA O ACCESO |
| 22 | CRUZANDO INTERSECCIÓN |
| 23 | ESTACIONANDO O SALIENDO DEL ESTACIONAMIENTO |
| 31 | CIRCULANDO HACIA ATRÁS |
| 41 | MANIOBRA SÚBITA PARA SALVAR OBSTÁCULO O VEHÍCULO |
| 42 | MANIOBRA SÚBITA PARA SALVAR PEATÓN AISLADO O EN GRUPO |
| 43 | BRUSCA REDUCCIÓN DE VELOCIDAD |
| 51 | RETENCIÓN POR IMPERATIVO DE LA CIRCULACIÓN |
| 52 | PARADO O ESTACIONADO |
| 61 | FUGADO |
| 71 | OTRA |
| 72 | SE IGNORA |

**Tabla 5.77:** Diccionario del predictor *maniobras*. Extraído de la Dirección General de Tráfico

| | Estimate | Std. Error | z value | Pr(>\|z\|) | Prob. grave | Incr. grave |
|---|---|---|---|---|---|---|
| ANOMALIA_NINGUNAX3 | -2,603823 | 0,022109 | -117,771 | <2e-16 *** | 26,06 % | 925 % |

**Tabla 5.78:** Descripción de la variable *anomalia_ninguna* en el modelo estimado

| | Estimate | Std. Error | z value | Pr(>\|z\|) | Prob. grave | Incr. grave |
|---|---|---|---|---|---|---|
| POSICIONX9 | -1,359436 | 0,174271 | -7,801 | 6,16e-15 *** | 9,22 % | 262,0 % |
| POSICIONX3 | -0,291019 | 0,038334 | -7,592 | 3,16e-14 *** | 3,37 % | 32,6 % |
| POSICIONX4 | -0,290979 | 0,035001 | -8,313 | <2e-16 *** | 3,37 % | 32,6 % |
| POSICIONX5 | -0,360185 | 0,061922 | -5,817 | 6,00e-09 *** | 3,60 % | 41,8 % |
| POSICIONX8 | -0,376290 | 0,069034 | -5,451 | 5,01e-08 *** | 3,66 % | 44,0 % |
| POSICIONX2 | -0,204760 | 0,021050 | -9,727 | <2e-16 *** | 3,10 % | 22,0 % |

**Tabla 5.79:** Descripción de la variable *posición* en el modelo estimado

A continuación, se muestran ordenadas las posiciones de los ocupantes de un vehículo de más peligrosas a menos en cuanto a padecer un accidente grave o fallecer:

1. PASAJEROS DE PIE (9)
2. OTROS PASAJEROS SENTADOS (8)
3. PASAJERO TRASERO CENTRAL (5)

4. PASAJERO TRASERO IZQUIERDO (3)
5. PASAJERO TRASERO DERECHO (4)
6. PASAJERO DELANTERO (2)
7. CONDUCTOR (1)

### 5.5.5. Influencia del tipo de vehículo

En la tabla 5.80 se puede observar la influencia que posee el tipo de vehículo en cuanto a que el ocupante del vehículo sufra lesiones graves o incluso fallezca. La tabla 5.81 muestra el significado de cada nivel.

| | Estimate | Std. Error | z value | Pr(>\|z\|) | Prob. grave | Incr. grave |
|---|---|---|---|---|---|---|
| TIPO_VEHICULOX4 | -2,283154 | 0,045822 | -49,827 | <2e-16 *** | 20,37 % | 701,3 % |
| TIPO_VEHICULOX3 | -2,212017 | 0,049419 | -44,760 | <2e-16 *** | 19,24 % | 656,9 % |
| TIPO_VEHICULOX2 | -2,000156 | 0,049953 | -40,041 | <2e-16 *** | 16,16 % | 535,7 % |
| TIPO_VEHICULOX18 | -1,899010 | 0,101980 | -18,621 | <2e-16 *** | 14,84 % | 483,7 % |
| TIPO_VEHICULOX8 | -1,365981 | 0,095590 | -14,290 | <2e-16 *** | 9,27 % | 264,9 % |
| TIPO_VEHICULOX20 | -1,096894 | 0,082817 | -13,245 | <2e-16 *** | 7,24 % | 185,0 % |
| TIPO_VEHICULOX21 | -1,046202 | 0,189443 | -5,523 | 3,34e-08 *** | 6,91 % | 171,9 % |
| TIPO_VEHICULOX13 | -0,661662 | 0,099861 | -6,626 | 3,45e-11 *** | 4,81 % | 89,3 % |
| TIPO_VEHICULOX15 | -0,619892 | 0,072670 | -8,530 | <2e-16 *** | 4,62 % | 81,9 % |
| TIPO_VEHICULOX7 | -0,551663 | 0,124586 | -4,428 | 9,51e-06 *** | 4,33 % | 70,4 % |
| TIPO_VEHICULOX9 | -0,288959 | 0,051876 | -5,570 | 2,54e-08 *** | 3,37 % | 32,4 % |
| TIPO_VEHICULOX6 | -0,287688 | 0,044744 | -6,430 | 1,28e-10 *** | 3,36 % | 32,2 % |
| TIPO_VEHICULOX16 | 0,549678 | 0,096915 | 5,672 | 1,41e-08 *** | 1,48 % | -41,6 % |
| TIPO_VEHICULOX17 | 0,665469 | 0,333851 | 1,993 | 0,046227 * | 1,32 % | -47,9 % |

**Tabla 5.80:** Descripción de la variable *tipo de vehículo* en el modelo estimado

Las 5 primeras filas hacen referencia a motocicletas y ciclomotores en función de su cilindrada, no especificada por la DGT en el diccionario donde define la correspondencia valor-descripción de las variables (tabla 5.81). Éstas hacen que el ocupante tenga mayor riesgo de lesiones graves o de fallecimiento.

Los siguientes tipos de vehículos que incrementan la probabilidad de fallecimiento son los turismos y los turismos de Servicio Público de hasta 9 plazas, respectivamente.

### 5.5.6. Influencia de la iluminación de la vía

La probabilidad de que las lesiones de los ocupantes del vehículo sean graves o incluso que provoquen la muerte se ven incrementadas, claramente, cuando la iluminación nocturna es insuficiente (ver tabla 5.82). Adicionalmente, se puede comprobar que, cuando la iluminación es suficiente en la noche o en momento de crepúsculo, ambas probabilidades son similares aunque el incremento de la gravedad del accidente es algo mayor que en condiciones lumínicas de pleno día.

La tabla 5.83 muestra un diccionario para hacer corresponder el nivel del factor con el significado.

| Nivel de factor | Descripción |
|---|---|
| 1 | Bicicleta o triciclo sin motor |
| 2 | Ciclomotor |
| 10 | Coche de Minusválido |
| 11 | Motocicleta |
| 21 | Turismo de SP hasta 9 plazas |
| 22 | Turismo sin remolque |
| 23 | Turismo con remolque |
| 24 | Ambulancia |
| 30 | Maquinaria de obras y agrícola |
| 31 | Tractor agrícola sin remolque |
| 32 | Tractor agrícola con remolque |
| 41 | Camión (PM <= 3500 K) sin remolque |
| 42 | Camión (PM <= 3500 K) con remolque |
| 43 | Furgoneta |
| 51 | Camión (PM >3500 K) sin remolque |
| 52 | Camión (PM >3500 K) con remolque |
| 53 | Camión cisterna sin remolque |
| 54 | Camión cisterna con remolque |
| 55 | Vehículo articulado |
| 61 | Autobús de línea regular |
| 62 | Autobús escolar |
| 63 | Otro autobús |
| 70 | Tren |
| 80 | Carro |
| 81 | Otros Vehículos |
| 82 | Cuadriciclo |
| 90 | Desconocido |

**Tabla 5.81:** Diccionario del predictor *tipo de vehículo*. Extraído de la Dirección General de Tráfico

| Nivel de factor | $e^{logits}$ | Odds | Prob. grave | Incr. grave |
|---|---|---|---|---|
| Crepúsculo | $exp(-0,243769)$ | 0,7836686 | 3,22 % | 26,72 % |
| Noche: iluminación suficiente | $exp(-0,262164)$ | 0,7693848 | 3,28 % | 28,99 % |
| Noche: iluminación insuficiente | $exp(-0,553158)$ | 0,5751307 | 4,34 % | 70,67 % |

**Tabla 5.82:** Descripción de la variable *iluminación de la vía* en el modelo estimado

| Nivel de factor | Descripción |
|---|---|
| 1 | PLENO DÍA |
| 2 | CREPÚSCULO |
| 3 | NOCHE: ILUMINACIÓN SUFICIENTE |
| 4 | NOCHE: ILUMINACIÓN INSUFICIENTE |
| 5 | NOCHE: SIN ILUMINACIÓN |

**Tabla 5.83:** Diccionario del predictor *luminosidad*. Extraído de la Dirección General de Tráfico

### 5.5.7. Influencia del estado de la superficie de la calzada

La tabla 5.84 muestra qué influencia tiene el estado de la superficie de la calzada en la severidad del accidentado. Como se puede observar, la probabilidad de que el estado del accidentado sea grave o incluso que fallezca, se reduce en vías con hielo o nieve respecto a aquellas mojadas. Todo ello viene a confirmar lo estudiado en la sección 3.2.1, debido

a que estas condiciones climatológicas ejercen una influencia sobre el comportamiento de los conductores, y éstos conducen más despacio, por lo que aunque la probabilidad de que el accidente sea leve aumente, la de grave o fallecimiento de alguno de los ocupantes se ve reducida significativamente respecto a las vías con superficie seca y limpia.

La tabla 5.85 muestra a su vez todos los posibles niveles del factor superficie de la calzada, donde se puede observar que los niveles para vías umbrías, con gravilla suelta, aceite o de otro tipo distinto a los descritos en la tabla 5.84 no se consideran importantes para que favorezcan accidentes.

| Nivel del factor | Descripción | Estimate | Std. Error | z value | Pr(>\|z\|) | Prob. grave | Incr. grave |
|---|---|---|---|---|---|---|---|
| SUPERFICIE_CALZADAX3 | Mojada | 0,302556 | 0,020500 | 14,759 | <2e-16 *** | 1,89 % | -25,61 % |
| SUPERFICIE_CALZADAX4 | Helada | 0,442031 | 0,135787 | 3,255 | 0,001133 ** | 1,65 % | -35,14 % |
| SUPERFICIE_CALZADAX5 | Nevada | 0,610606 | 0,161478 | 3,781 | 0,000156 *** | 1,40 % | -45,06 % |
| SUPERFICIE_CALZADAX6 | Barrillo | 1,527321 | 0,165806 | 9,211 | <2e-16 *** | 0,56 % | -77,85 % |

**Tabla 5.84:** Descripción de la variable *superficie de la calzada* en el modelo estimado

| Nivel de factor | Descripción |
|---|---|
| 1 | SECA Y LIMPIA |
| 2 | UMBRÍA |
| 3 | MOJADA |
| 4 | HELADA |
| 5 | NEVADA |
| 6 | BARRILLO |
| 7 | GRAVILLA SUELTA |
| 8 | ACEITE |
| 9 | OTRO TIPO |

**Tabla 5.85:** Diccionario del predictor *superficie de la calzada*. Extraído de la Dirección General de Tráfico

## 5.6. Conclusiones y aportaciones

En este capítulo se presenta un estudio de accidentes de tráfico ocurridos en las carreteras españolas durante el periodo 2011-2015, cuyos descriptores han sido obtenidos a partir de un set de datos publicado en abierto por la DGT.

Cada accidente ha sido considerado como un evento aleatorio discreto e independiente cuya probabilidad de ocurrencia puede ser modelada por un gran número de características, como las condiciones meteorológicas, obras en la calzada, la ubicación geográfica de los radares de control de velocidad, la infraestructura de la carretera, etc.

El número total de accidentes en el período estudiado no conforma un gran problema de datos por sí mismo, sin embargo, el alto número de variables involucradas implica que han de desplegarse técnicas de Big Data para extraer y almacenar los vectores de características.

Se propone un nuevo set de datos para el estudio de los accidentes de tráfico en España, con nuevos descriptores creados a partir de la información extraída de los sets de datps publicados por el Ministerio del Interior.

Adicionalmente, en este capítulo se aporta un nuevo modelo más complejo para el estudio de las causas de los accidentes de tráfico que se separa del triángulo clásico que relaciona conductor, infraestructura y accidente. Se incluyen una serie de características que, unidas al gran número de aquellas que ya se emplean históricamente, permiten analizar los accidentes en profundidad.

En la sección 5.2 se han indicado nuevas características que podrían intervenir como causas en los accidentes de tráfico y que no se han empleado anteriormente en estudios en la materia. Dichas características nos permiten analizar el problema a través de un nuevo modelo de análisis avanzado. Entre ellas, se presentan dos características novedosas:

- característica que indica si el accidente se ha producido en la vecindad de un radar de control de velocidad y la distancia euclídea a él.
- característica que indica si el accidente se ha producido en un tramo que se encontraba en obras.

Se propone el uso de técnicas basadas en k-fold para el muestreo de accidentes de tráfico cuando se posee una clase minoritaria y una mayoritaria. En nuestro caso contamos con una proporción 20/80 lo que provoca cierto desbalanceo entre ellas y la posibilidad de ocurrencia de *overfitting* si el muestreo fuera incorrecto.

Adicionalmente, se han presentado resultados en términos absolutos sobre las vías de España con más accidentes durante el período estudiado. Además, se muestra una distribución de los accidentes con respecto a la temperatura, así como la relación del número de accidentes con respecto al día de la semana en que éstos se producen.

Una conclusión importante de nuestro estudio es que a la hora de estudiar accidentes de tráfico sería conveniente el uso de datos más precisos extraídos de fuentes oficiales a través de los cuales se podría modelar el accidente de una manera más concreta. Especialmente, hemos encontrado grandes problemas en la precisión GPS de la situación de los accidentes servidos por Infocar. Quizás la DGT debería revisar los procedimientos de captura precisa del lugar del accidente a través de los Cuerpos de Seguridad del Estado o servicios de sanitarios o de ayuda en carretera.

Por último, y no por ello menos importante, se han tenido que separar ambas investigaciones, la que se ha realizado a partir de los dataset extraídos de la DGT y aquellos que se han generado en la infraestructura de BigData, debido a que no parece haber forma de relacionar el mismo accidente entre las dos plataformas puesto que se emplean IDs distintos entre incidencias y accidentes.

En este sentido, el Departamento para el Transporte en Reino Unido ha trabajado para aportar la localización geográfica de cada accidente en todos sus sets de datos.

Se propone el uso de un clasificador supervisado mixto basado en Random Forest y regresión logística bayesiana (BayesGLM). La selección global de características vendrá dada por el primero de ellos a través de diversos árboles, para reducir la posibilidad de *overfitting*, mediante la técnica de Recursive Feature Elimination (RFE). El uso de un ensamble lineal con técnicas bayesianas permite seleccionar una distancia entre clases de forma más óptima que con métodos lineales puros. Por último, se ha seleccionado un clasificador lineal bayesiano debido a la posibilidad que ofrece de interpretar los resultados de forma sencilla, así como de obtener distintas probabilidades a través del modelo.

Se presenta un modelo de predicción probado con el set de datos de prueba arrojando los siguientes resultados:

- Predicción de accidentados graves o fallecimientos: $82\,\%$.
- Predicción de accidentados leves: $72\,\%$.

Se estudia la influencia de cada característica del vector escogido para generar el modelo aportando unas conclusiones sobre ellas que se deben de considerar a la hora de proponer medidas de seguridad vial, así como de dotar de inteligencia artificial a la tecnología empleada por los usuarios en forma de automóvil o de infraestructura vial.

UNIVERSITAS
Miguel
Hernández

# 6 Estudio de Accidentes de Tráfico en Reino Unido

## 6.1. Recopilación de datos de tráfico de Reino Unido

En Reino Unido cada accidente es registrado por la policía utilizando el sistema STATS19 [3]. STATS19 hace referencia al formulario donde se describe la información que debe recopilarse cada vez que un accidente es reportado. También se utiliza frecuentemente para referirse a las estadísticas oficiales de accidentes de tráfico de Gran Bretaña, que se derivan de las estadísticas de STATS19 de la Policía y compiladas por el Departamento de Transporte [2].

La información que proporciona el sistema STATS19 es de gran valor para los profesionales de la seguridad vial, aunque como es de imaginar, no es posible hacer pública la totalidad de los datos que contiene por razones de confidencialidad de las víctimas. Sin embargo, la información derivada del análisis de los datos de STATS19 constituye la base para la publicación anual del informe '*Road Casualties in Great Britain*' [1], y es también un componente clave para el análisis de los datos.

Prácticamente la totalidad de los accidentes recogidos en las estadísticas de Reino Unido están relacionadas con aquellos producidos en carreteras públicas y, evidentemente, han sido informados a la policía. De forma general, toda la información referente al accidente es recogida por las Fuerzas de Seguridad en la escena del mismo, aunque también se puede dar la posibilidad de que ésta sea reportada por algún particular en los 30 primeros días desde que ocurrió el mismo. La policía comprueba y analiza todos los datos antes de registrarlos a través del formulario estándar STATS19.

En cuanto a los datos almacenados, basándonos en el número de fallecimientos registrados en accidentes de tráfico, se puede concluir que son bastante fiables, ya que

cotejando estadísticas del registro de defunciones de Reino Unido, se demuestra que la totalidad de fallecimientos en accidentes de tráfico son reportados por la policía al registro.

Sin embargo, desafortunadamente, no podemos decir lo mismo de las víctimas no mortales, ya que los datos de los hospitales, las encuestas y las reclamaciones de indemnización a las aseguradoras dan lugar a un mayor número de víctimas que las notificadas.

Por lo tanto, los datos de STATS19 no representan un registro completo de todos los accidentes con lesiones, y ésto debe tenerse en cuenta al utilizar y analizar los datos contenidos en él. Sin embargo, siguen siendo la fuente de información más completa y fiable sobre víctimas de carretera que abarca todo Reino Unido.

## 6.2. Descripción de los dataset

Para realizar un análisis de los accidentes de tráfico en Reino Unido, se ha empleado un conjunto de datos formado por un gran número de accidentes registrados entre 2009 y 2014 en este país e incluyen la severidad del accidente y de los ocupantes, así como un gran número de variables que describen las causas alrededor de él, entre ellas, las relativas a infraestructura vial, condiciones climáticas y otras.

Todos los datos han sido extraídos del departamento de transporte del Reino Unido en el periodo de 2009 a 2014, y el dataset completo se encuentra dividido en tres tablas:

- Tabla de vehículos (24 características)
- Tabla de accidentados (14 características)
- Tabla de accidentes (32 características)

cada una de ellas almacenadas en formato CSV y relacionadas entre sí por una serie de identificadores, como son:

- el identificador de accidentado
- el identificador de accidente
- el identificador de vehículo

### 6.2.1. Introducción y descripción de las clases

Puesto que este trabajo pretende aportar nuevas técnicas en el estudio y análisis acerca de las víctimas de los accidentes de tráfico, el planteamiento inicial de este trabajo consistirá en la búsqueda de una función de probabilidad o clasificador que nos permita obtener la probabilidad de que los accidentados posean un pronóstico leve, grave, o incluso que pueda fallecer, en base a una serie de características.

Para ello, en función de la selección de características en cada experimento, se obtendrán una serie de variables predictoras y una clase dicotómica, variable objetivo *Casualty_Severity*.

Se ha empleado una selección de modelos de clasificación con el fin de tratar de obtener el mejor resultado tras su entrenamiento, y para ello, se ha optado por unir los accidentes *Graves* y los accidentes *Fatales* de tal forma que se obtiene una clase binomial a predecir y que nos da pie a simplificar el problema optando por usar modelos basados en clases binomiales en lugar de tener que emplear modelos de clasificación basados en clases multinomiales.

En resumen, la modificación de la variable *Casualty_Severity* se representa de la siguiente forma:

- Casualty_Severity=0 (accidente grave o fatal)
- Casualty_Severity=1 (accidente leve)

### 6.2.2. Introducción y descripción del dataset

En esta sección se realizará una breve descripción de todas las variables estudiadas divididas por tablas.

#### 6.2.2.1. Tabla *Vehicles*

Esta primera tabla contiene indexadas todas las características relativas a los vehículos implicados en cada accidente, así como otras de interés para el estudio como, por ejemplo, las maniobras que estaban efectuando, o los datos relativos al conductor y ocupantes.

En la tabla 6.1 se enumeran todas las características contenidas en dicha tabla *Vehicles*. Adicionalmente, se ha añadido una breve descripción de cada una de ellas con algún ejemplo de sus posibles valores.

#### 6.2.2.2. Tabla *Accidents*

Esta tabla contiene indexadas todas las características relativas al accidente en sí, y registra variables relativas al número de vehículos implicados, tipo de vía o condiciones meteorológicas, entre otras.

En la tabla 6.2 se enumeran todas las características contenidas en la tabla *Accidents*. Adicionalmente, se ha añadido una breve descripción de cada característica con algún ejemplo de sus posibles valores.

| Característica | Descripción |
|---|---|
| Accident Index | Representa el índice del accidente |
| Vehicle Reference | Representa el índice del vehículo relativo al accidente |
| Vehicle Type | Tipo de vehículo referente a la variable anterior: motocicleta, coche, etc. |
| Towing and Articulation | Indica si el vehículo es articulado (trailer) o posee remolque, caravana, etc |
| Vehicle Manoeuvre | Referente a la maniobra que se encontraba realizando el vehículo en el momento del accidente: aparcado, girando a la izquierda, cambiando de carril, etc. |
| Vehicle Location-Restricted Lane | Dónde se encontraba situado el vehículo en el momento del accidente: carril bus, carril ciclista, etc. |
| Junction Location | Indica si el vehículo se encontraba acercándose a un cruce, entrando a una rotonda, etc. |
| Skidding and Overturning | Indica si el vehículo derrapó, volcó, etc. |
| Hit Object in Carriageway | Indica contra qué chocó el vehículo en la calzada: contra una obra, contra un animal, contra una puerta de un coche abierta, ... |
| Vehicle Leaving Carriageway | Indica qué pasó con el vehículo respecto a la calzada: fue a la mediana, cruzó toda la calzada, etc. |
| Hit Object off Carriageway | Indica contra qué chocó el vehículo al salir de la vía: contra una farola, señal de tráfico, poste eléctrico, etc. |
| 1st Point of Impact | Indica cuál fue el primer punto de impacto cuando chocó el vehículo: Front, Back, etc. |
| Was Vehicle Left Hand Drive | Indica si el volante se encontraba a la izquierda: No, Yes |
| Journey Purpose of Driver | Indica cuál era el objeto del trayecto de ese vehículo: work, school, etc. |
| Sex of Driver | Sexo del conductor |
| Age of Driver | Edad del conductor |
| Age Band of Driver | Rangos de edad del conductor |
| Engine Capacity | Motorización del vehículo |
| Vehicle Propulsion Code | Tipo de motor: Gasolina, eléctrico, etc. |
| Age of Vehicle (manufacture) | Año de manufactura del vehículo |
| Driver IMD Decile | Más necesitados, menos necesitados, etc. |
| Driver Home Area Type | Zona del accidente (urbana, pueblo pequeño, rural) |
| Make | Marca del vehículo |

**Tabla 6.1:** Descripción de todas las características contenidas en la tabla *Vehicles*. Tabla extraída del Departamento para el Transporte del Reino Unido.

#### 6.2.2.3. Tabla *Casualties*

La tabla relativa a los accidentados es la más extensa de todo el set de datos, ya que recoge todos los datos de cada ocupante de aquellos vehículos implicados. Por ejemplo, recoge desde las edades de las víctimas, hasta la posición que ocupaban en el interior del vehículo.

En la tabla 6.3 se enumeran todas las características contenidas en la tabla *Casualties*. Adicionalmente, se ha añadido una breve descripción de cada una de ellas con algún ejemplo de sus posibles valores, si corresponde.

## 6.3. Primera aproximación a la elección de predictores

El primer paso será fusionar las tres tablas relativas a accidentes, accidentados y vehículos, para formar una única tabla. Para ello, se parte de la tabla accidentados, añadiendo

| Característica | Descripción |
|---|---|
| Accident Index | Representa el índice del accidente |
| Police Force | Indica qué policía acudió o registró el accidente: policía metropolitana, Cleveland, Leicester,etc. |
| Accident Severity | Indica la severidad del accidente: Fatal, serious, slight |
| Number of Vehicles | Alude al número de vehículos implicados |
| Number of Casualties | Número de víctimas |
| Date | Fecha del accidente |
| Day of Week | Día de la semana en formato numérico, 1 Sunday, 2 Monday, etc. |
| Time | Hora a la que ocurrió el accidente |
| Location Easting OSGR | |
| Location Northing OSGR | |
| Longitude | Longitud |
| Latitude | Latitud |
| Local Authority (District) | Distrito en el que ocurrió el accidente: Westminster, Camden, etc. |
| Local Authority (Highway Authority - ONS code) | Birmingham, Bolton, etc. |
| 1st Road Class | Tipo de vía: Motorway, A-M, A, B, C |
| 1st Road Number | |
| Road Type | Dual carriageway, Single carriageway, etc. |
| Speed limit | Límite de velocidad en la vía |
| Junction Detail | Detalle del cruce donde se produjo el accidente: Crossroads, Private drive or entrance, etc. |
| Junction Control | Tipo de regulación del cruce: Auto traffic signal, etc. |
| 2nd Road Class | Motorway, A-M, etc. |
| 2nd Road Number | |
| Pedestrian Crossing-Human Control | None within 50 metres, Control by school crossing patrol, etc. |
| Light Conditions | Daylight, Darkness lights lit, etc. |
| Weather Conditions | Condiciones meteorológicas: Fine no high winds, Raining no high winds, etc. |
| Road Surface Conditions | Condiciones de la vía: Dry, Wet or damp, etc. |
| Special Conditions at Site | Auto traffic signal out, Oil or diesel, Roadworks, etc. |
| Carriageway Hazards | Vehicle load on road, Previous accident, etc. |
| Urban or Rural Area | rural, urban, etc. |
| Did Police Officer Attend Scene of Accident | Yes, no |
| Carriageway Hazards | Vehicle load on road, Previous accident, etc. |
| Lower Super Output Area of Accident_Location (England and Wales only) | |

**Tabla 6.2:** Descripción de todas las características contenidas en la tabla *Accidents*. Tabla extraída del Departamento para el Transporte del Reino Unido.

a cada uno de los registros almacenados en ella los datos relativos a los vehículos donde viajaba cada víctima, así como los datos relativos al accidente en sí mismo. Este proceso ha sido realizado mediante la programación de un script en lenguaje *R*.

A continuación, ha sido necesario comenzar a estudiar variable a variable de las 50 que contenía el dataset ya unificado, y considerar si era necesario incluirlas todas en el estudio, o si por el contrario se podía prescindir de algunas de ellas, con el fin de aligerar los cálculos, ya que era posible que muchas de ellas fueran variables redundantes, *dummy* o que no tuvieran repercusión en la búsqueda del objetivo para este trabajo de investigación, que consistirá en la clasificación de los accidentados.

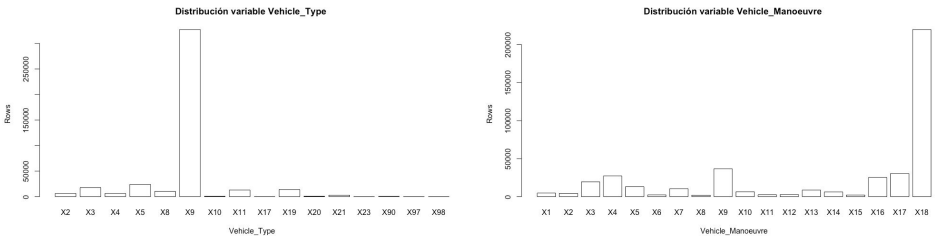



**Figura 6.1:** Distribución de variables Tipo de vehículo y Maniobra del vehículo

### 6.3.1. Distribución individual de cada variable

Como se puede observar en la figura 6.1, cotejando los valores con la tabla 6.4, la frecuencia de accidentes de coches es la mayoritaria. Adicionalmente, se puede observar que la frecuencia de ocurrencia en orden descendente corresponde a accidentes de motocicletas con una cilindrada mayor de 500cc, seguida de accidentes de motocicletas entre 125cc y 50cc. Los accidentes correspondientes a autobuses o autocares, con un aforo igual o mayor a 17 asientos, también ocupan un lugar similar al de los accidentes de furgonetas para una TARA igual o inferior a 3,5 toneladas.

En la tabla 6.4 se identifican todos los posibles tipos de vehículos que el DfT de Reino Unido contempla. Por otro lado, las distribuciones de los posibles tipos de maniobra que se encontraba efectuando el vehículo en el momento del accidente, se muestran en la figura 6.1 derecha. En ella se puede observar que la maniobra mayoritaria en accidentes

| Característica | Descripción |
|---|---|
| Accident Index | Representa el índice del accidente |
| Vehicle Reference | Representa el índice del vehículo relativo al accidente |
| Casualty Reference | Representa el índice de la víctima relativo al accidente |
| Casualty Class | Tipo de víctima: conductor, pasajero, peatón |
| Sex of Casualty | Sexo de la víctima: male, female |
| Age of Casualty | Edad de la víctima: 0-120 |
| Age Band of Casualty | Banda de edad de la víctima |
| Casualty Severity | Severidad de la víctima: Fatal, serious, slight |
| Pedestrian Location | Localización en el momento del accidente del peatón: cruzando por un paso de peatones, etc. |
| Pedestrian Movement | Acción que iba realizando el peatón: cruzando por el lado del conductor, etc. |
| Car Passenger | Lugar que ocupaba el pasajero del vehículo accidentado: front seat passenger, rear seat passenger, not car passenger |
| Bus or Coach Passenger | Lugar que ocupaba la víctima en el bus: Boarding, seated, etc. |
| Pedestrian Road Maintenance Worker (From 2011) | Sí, no conocido, no aplicable |
| Casualty Type | Tipo de víctima: peatón, ciclista, motocicleta 50cc, de 125cc, taxi, coche, etc. |
| Casualty IMD Decile | Más necesitados, menos necesitados, etc. |
| Casualty Home Area Type | zona urbana, pueblo pequeño, rural. |

**Tabla 6.3:** Descripción de todas las características contenidas en la tabla *Casualties*. Tabla extraída del Departamento para el Transporte del Reino Unido.

| Código | Tipo de Vehículo |
|---|---|
| 1 | Pedal cycle |
| 2 | Motorcycle 50cc and under |
| 3 | Motorcycle 125cc and under |
| 4 | Motorcycle over 125cc and up to 500cc |
| 5 | Motorcycle over 500cc |
| 8 | Taxi/Private hire car |
| 9 | Car |
| 10 | Minibus (8 - 16 passenger seats) |
| 11 | Bus or coach (17 or more pass seats) |
| 16 | Ridden horse |
| 17 | Agricultural vehicle |
| 18 | Tram |
| 19 | Van / Goods 3.5 tonnes mgw or under |
| 20 | Goods over 3.5t. and under 7.5t |
| 21 | Goods 7.5 tonnes mgw and over |
| 22 | Mobility scooter |
| 23 | Electric motorcycle |
| 90 | Other vehicle |
| 97 | Motorcycle - unknown cc |
| 98 | Goods vehicle - unknown weight |
| -1 | Data missing or out of range |

**Tabla 6.4:** Tipos de vehículos contemplados por el Departamento para Tráfico de Reino Unido.

indexados corresponde con el tipo '*Yendo hacia adelante (otras)*'. La maniobra que le sigue en orden de frecuencia mayoritaria se corresponde con los '*giros a la derecha* y, a continuación, '*marcha adelante con curva a la derecha* o '*aminorando o parando*, respectivamente.

| Código | Maniobra |
|---|---|
| 1 | Marcha atrás |
| 2 | Aparcado |
| 3 | Esperando para salir |
| 4 | Aminorando o parando |
| 5 | Moviéndose |
| 6 | Cambiando de sentido |
| 7 | Girando a la izquierda |
| 8 | Esperando para girar a la izquierda |
| 9 | Girando a la derecha |
| 10 | Esperando para girar a la derecha |
| 11 | Cambiando al carril de la izquierda |
| 12 | Cambiando al carril de la derecha |
| 13 | Adelantando a un vehículo en movimiento - offside |
| 14 | Adelantando a un vehículo parado - offside |
| 15 | Adelantando - nearside |
| 16 | Yendo hacia atrás para girar a mano izquierda |
| 17 | Yendo hacia delante para girar a mano derecha |
| 18 | Yendo hacia adelante (otras) |
| -1 | Datos perdidos o fuera de rango |

**Tabla 6.5:** Tipos de maniobras recogidos por el Departamento para Tráfico de Reino Unido

En la tabla 6.5 podemos observar todos los tipos de maniobras recogidos por la DfT. Esta variable será clave en el diseño de la infraestructura así como en el diseño de

navegadores o aplicaciones de navegación. Adicionalmente, debería ser considerada por los diseñadores de vehículos autónomos puesto que puede ser de gran ayuda a la hora de tomar decisiones al escoger rutas.

La figura 6.2 representa la distribución absoluta de accidentes por marca de vehículo. En ella podemos observar que la marca que posee mayor numero de accidentes registrados es Vauxhall, seguida por Ford en segunda posición, y por Peugeot en tercera posición.

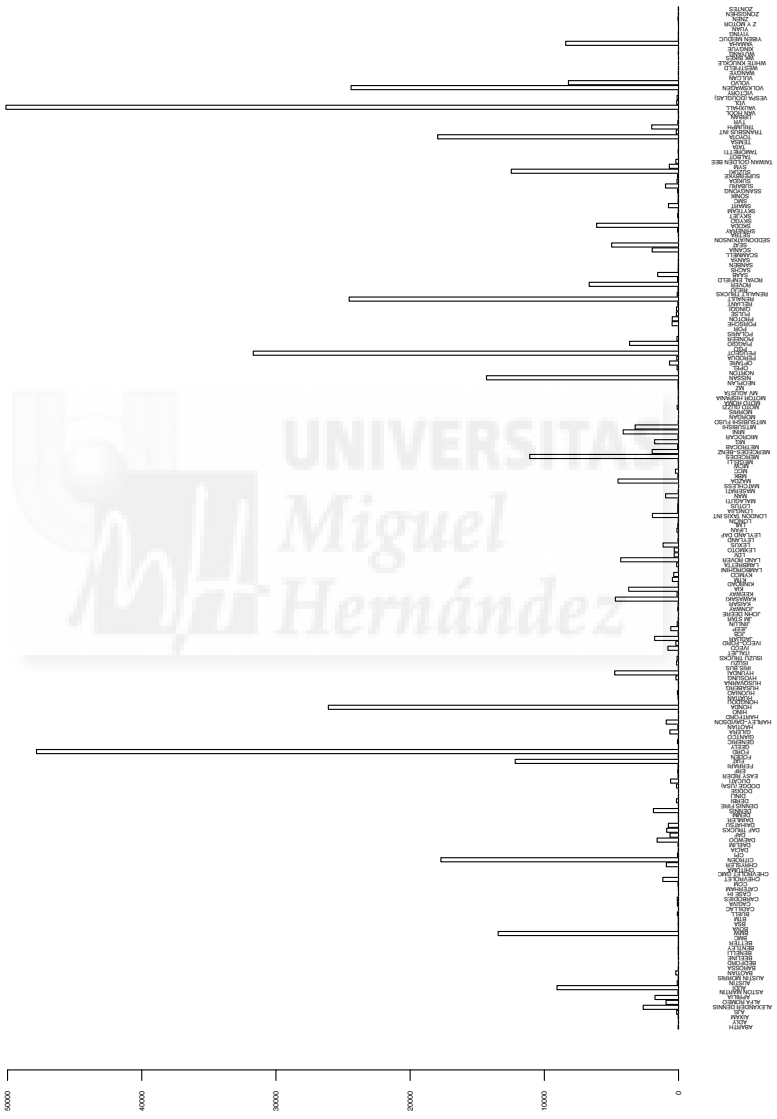



**Figura 6.2:** Distribución de la variable Marca del Vehículo


**Figura 6.3:** Distribución de variable Antigüedad del Vehículo

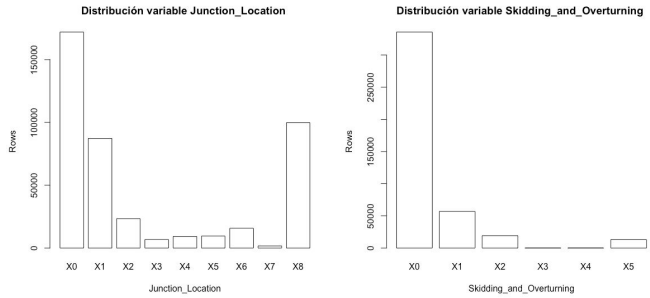



**Figura 6.4:** Distribución de variables Localización del cruce y Derrape y volcado

Respecto a la antigüedad del vehículo, en la figura 6.3 se puede observar que la mayor distribución de accidentes oscila entre el primer y el decimoprimer año con distribuciones cercanas a 35.000 accidentes en cada categoría de antigüedad del vehículo. A partir del decimosegundo año disminuye hasta los 50.000 accidentes por año hasta una distribución menor a partir del vigésimo segundo año.

La variable *Junction_Location* hace referencia a la localización de la unión de un cruce de dos o más vías. La clase mayoritaria en este caso corresponde a los accidentes que se producen fuera de un cruce o en los 20m próximos a éste. Le siguen aquellos accidentes que se producen entrando a un cruce o en el interior de una rotonda.

En la tabla 6.6 podemos observar los distintos tipos de maniobras en los cruces incluidos en el set de datos. Esta variable unida a la recogida en la tabla 6.5 serán importantes a la hora del estudio de un mejor diseño de la infraestructura, tal y como se comentaba anteriormente.

| Código | Maniobra |
|---|---|
| 0 | Not at or within 20 metres of junction |
| 1 | Approaching junction or waiting/parked at junction approach |
| 2 | Cleared junction or waiting/parked at junction exit |
| 3 | Leaving roundabout |
| 4 | Entering roundabout |
| 5 | Leaving main road |
| 6 | Entering main road |
| 7 | Entering from slip road |
| 8 | Mid Junction - on roundabout or on main road |
| -1 | Data missing or out of range |

**Tabla 6.6:** Tipos de maniobras incluidos en el set de datos

Las distribuciones de las variables *Hit_Object_Carriageway* y *First_Point_of_Impact* pueden observarse en la figura 6.5. En ella se representa el golpeo contra objetos fuera de la vía y el primer punto de impacto en el vehículo, respectivamente. Los árboles son el objeto, identificado en el parte de accidente, más frecuente contra el que se suele golpear el vehículo cuando se sale de la vía.

| Código | Maniobra |
|---|---|
| 0 | None |
| 1 | Skidded |
| 2 | Skidded and overturned |
| 3 | Jackknifed |
| 4 | Jackknifed and overturned |
| 5 | Overturned |
| -1 | Data missing or out of range |

**Tabla 6.7:** Variable que indica si el vehículo ha derrapado y/o volcado

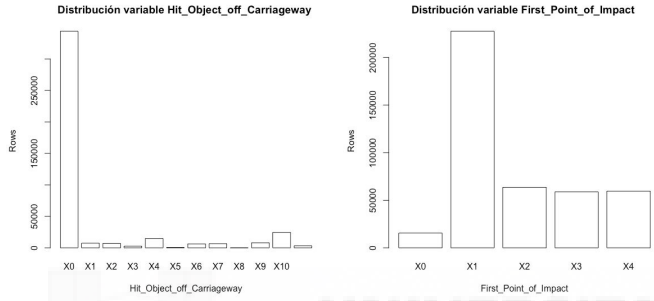



**Figura 6.5:** Distribución de variables Golpe contra objeto fuera de la vía y Primer punto de impacto

El primer punto de impacto en el vehículo cuando se sufre un accidente es la parte delantera del mismo, con $2/3$ de diferencia respecto al resto de partes, que se suelen golpear con igual frecuencia. A simple vista, parece sencillo pensar que los fabricantes de vehículos estarán pendientes de este punto de impacto a la hora de fabricarlos.

| Código | Maniobra |
|---|---|
| 0 | None |
| 1 | Road sign or traffic signal |
| 2 | Lamp post |
| 3 | Telegraph or electricity pole |
| 4 | Tree |
| 5 | Bus stop or bus shelter |
| 6 | Central crash barrier |
| 7 | Near/Offside crash barrier |
| 8 | Submerged in water |
| 9 | Entered ditch |
| 10 | Other permanent object |
| 11 | Wall or fence |
| -1 | Data missing or out of range |

**Tabla 6.8:** Variable que indica contra qué objeto golpeó cuando se salió de la vía

La movilidad durante la jornada laboral es el tipo de accidente que ocurre con mayor frecuencia, si nos centramos en la variable '*Propósito del viaje*'. Sin embargo, acudir o volver del trabajo también posee una frecuencia elevada de ocurrencia de accidentes de tráfico, incluso muy por encima de llevar o recoger a las/los hijas/os del colegio, tal

| Código | Maniobra |
|---|---|
| 0 | Did not impact |
| 1 | Front |
| 2 | Back |
| 3 | Offside |
| 4 | Nearside |
| -1 | Data missing or out of range |

**Tabla 6.9:** Variable que indica cuál fue el primer punto de impacto en el vehículo

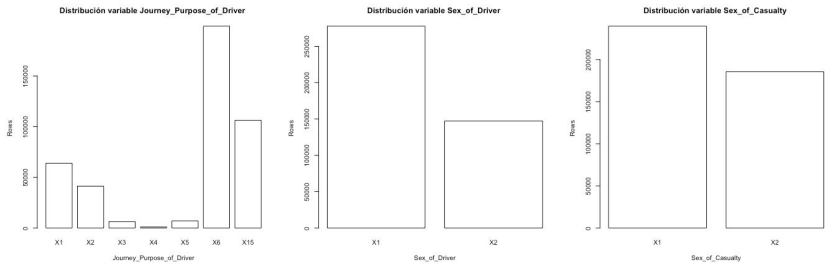



**Figura 6.6:** Distribución de variables Propósito del viaje, Sexo del conductor, Sexo de la víctima

y como se puede observar en la figura 6.6.

En Gran Bretaña, a la finalización de la fecha del estudio en 2014, el número de conductores con licencia para conducir se detalla en la tabla 6.10. En ella se puede comprobar que el 53 % de los conductores son varones, mientras que el 47 % son mujeres. Sin embargo, los hombres tienen aproximadamente el doble de accidentes que las mujeres, tal y como podemos ver en la figura 6.6.

| | Male | Female | Total |
|---|---|---|---|
| Número de personas con licencia | 24.446.143 | 21.667.400 | 46.113.543 |

**Tabla 6.10:** Variable que indica cuál fue el primer punto de impacto en el vehículo

En cuanto al sexo de la víctima, los hombres suelen ser de nuevo mayoría, sin embargo, en este caso las mujeres están próximas, tal y como podemos ver en la gráfica 6.6.

El rango de edad de conductores que posee una frecuencia mayor de accidentes oscila entre los 26 a los 45, siendo la edad comprendida entre los 26 y los 35 años la que presenta el pico de concentración de accidentes anuales, siendo de unos 90.000 accidentes aproximadamente el total en los 5 años del estudio.

En cuanto a la edad de la población de accidentados, resultará evidente que una vez más existirá un pico entre los 26 y los 35 años, alcanzando el máximo total, sin embargo, resulta curioso el ascenso del rango de edades entre los 16 y los 20 años de este colectivo de accidentados. En la figura 6.7 se puede observar todas estas frecuencias.

| Código | Maniobra |
|---|---|
| 1 | Journey as part of work |
| 2 | Commuting to/from work |
| 3 | Taking pupil to/from school |
| 4 | Pupil riding to/from school |
| 5 | Other |
| 6 | Not known |
| 15 | Other/Not known (2005-10) |
| -1 | Data missing or out of range |

**Tabla 6.11:** Variable que indica cuál fue el destino de la ruta

| Código | Maniobra |
|---|---|
| 1 | Male |
| 2 | Female |
| 3 | Not known |
| -1 | Data missing or out of range |

**Tabla 6.12:** Variable que indica el sexo del conductor

| Código | Maniobra |
|---|---|
| 1 | Male |
| 2 | Female |
| -1 | Data missing or out of range |

**Tabla 6.13:** Variable que indica el sexo de la víctima

En el momento de abandonar la vía, conviene observar en la figura 6.8 que la mayoría de accidentes se producen cuando no existe abandono de la vía, aunque algunos de ellos se producen al abandonar la vía por la izquierda, como resulta normal al conducir por el lado izquierdo en Reino Unido.

En la figura 6.9 se puede observar la frecuencia de ocurrencia de accidentes en función de la clase de víctima, donde se presenta una tasa que se podría considerar normal, aunque la media de peatones atropellados es de 10.000/año, algo elevada respecto al número de conductores.

Las víctimas, graves, leves o fallecidos, respecto a la posición que ocupan en el vehículo, está bastante equilibrada en cuanto a las que ocupan lugares en el asiento de atrás, con respecto a las que se encuentran sentadas en la parte delantera, tal y como se puede observar en la figura 6.10.

En cuanto a los tipos principales de vías donde se produjeron el mayor número de accidentes fueron las vías de tipo A, correspondientes a las carreteras principales, a veces de doble calzada parecidas a las autovías o autopistas. y en las de tipo B, correspondientes a las carreteras secundarias. Toda la información se muestra en la figura 6.11. Las vías de simple carril son las que ostentan el mayor número de accidentes, seguidas de las de doble carril con 75 % menos de accidentes que en las anteriores, tal y como se puede comprobar en la figura 6.13.

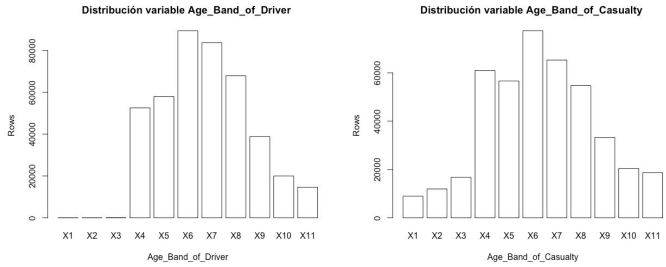



**Figura 6.7:** Distribución de variables Rango de edad del conductor y Rango de edad de la víctima

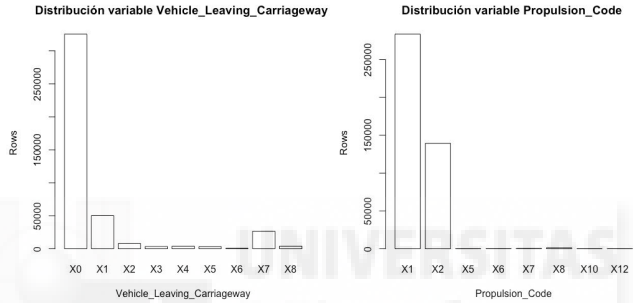



**Figura 6.8:** Distribución de variables El vehículo abandonaba la vía y Motorización

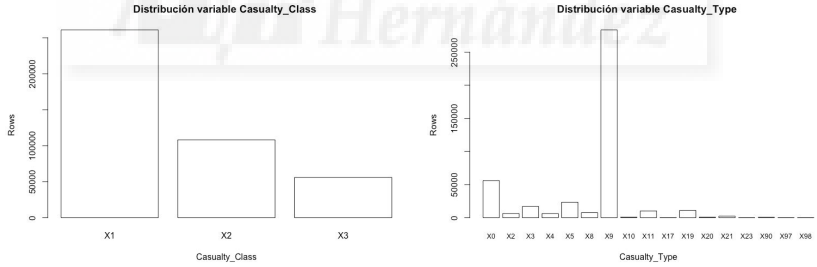



**Figura 6.9:** Distribución de variables Clase de víctima y Tipo de víctima

En cuanto al límite de velocidad, se puede observar que aquellas vías restringidas a 30 mph son las que presentan mayor número de accidentes, con más de 250.000 en 5 años.

Al igual que ocurre en España, el día de la semana donde se producen más accidentes es en viernes, mientras que el pico de accidentes lo encontramos entre las 15 y las 17h, seguido por otro pico a las 8h, coincidente con el comienzo de la jornada laboral (figura 6.12).

El detalle de los cruces (ver tabla 6.14) también dice mucho del tráfico en este país.

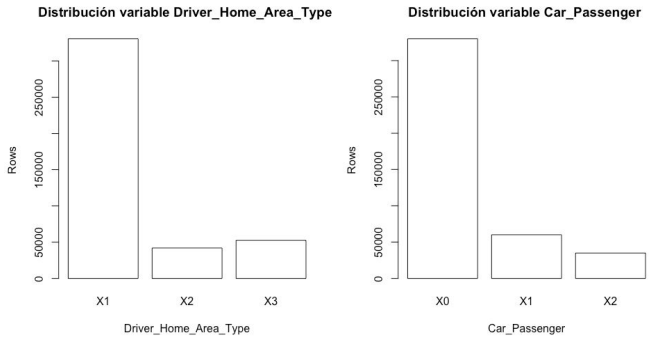



**Figura 6.10:** Distribución de variables Tipo de área cercana al hogar y Tipo de pasajero

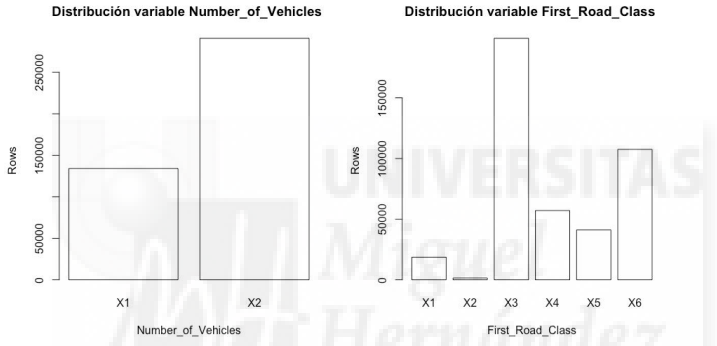



**Figura 6.11:** Distribución de variables Número de vehículos y Clase de vía principal

Aquellos tramos con cruces en T poseen registrados más de 120.000 accidentes en 5 años.

Respecto a la distribución de los accidentes por el tipo de iluminación de la vía (ver tabla 6.15), debido a que la mayoría de tráfico es con luz del día, ésta será la categoría con mayor número de accidentes, seguida de aquellas vías que presentan algún tipo de iluminación. Es significativo comprobar la elevada siniestralidad en vías sin iluminación a pesar de que a esas horas el tráfico es muy bajo. La figura 6.15 presenta un breve resumen de la distribución de accidentes por tipo de iluminación de la vía.

Las condiciones de la vía (figura 6.16), como es habitual, suelen ser en muchas ocasiones una causa adicional para que los conductores puedan sufrir un accidente. Sin embargo, el número de accidentes que se producen en vías secas es más del doble de aquellos que se producen en vías mojadas. Teniendo en cuenta que la media de días con lluvia al año en Reino Unido es de 148 días, es decir, un 40 % del total de días, la distribución se ajusta aproximadamente a dicha media.

Curiosamente, los accidentes entre dos coches no son los habituales, aunque sí los que ocupan la segunda posición en la distribución de la variable 'other_vehicle'. Aquellos

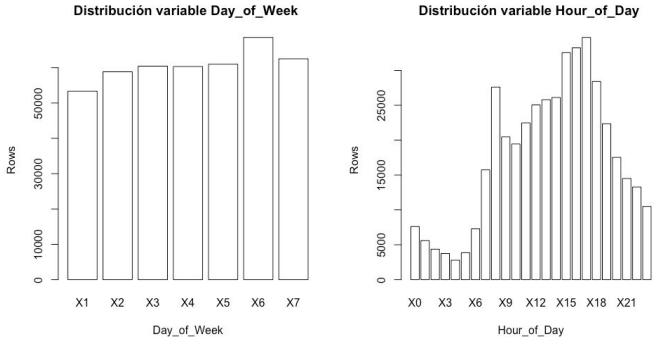



**Figura 6.12:** Distribución de variables Día de la semana y Hora del día

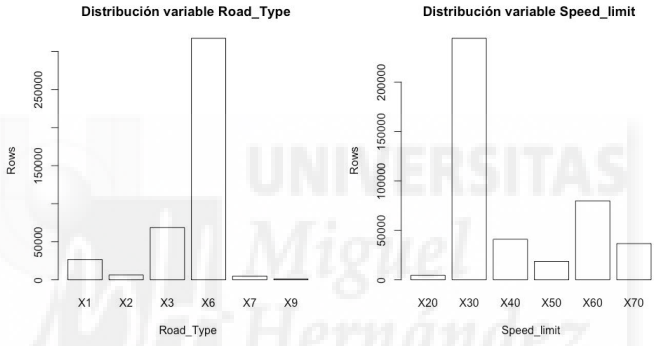



**Figura 6.13:** Distribución de variables Tipo de carretera y Límites de velocidad

accidentes que se producen sin que intervenga ningún otro vehículo son los más frecuentes con más de 350.000 accidentes en 5 años, muy por encima de los accidentes entre coches que ascienden a unos 60.000, aproximadamente. En la figura 6.17 se puede encontrar un resumen de la distribución de accidentes en base a los tipos de vehículo con los que se choca.

### 6.3.2. Pre-procesamiento de variables

Se ha realizado un trabajo de pre-procesamiento con los siguientes objetivos principales:

- Conversión a tipo de dato factor de aquellas variables que deseemos agrupar como tal.
- Eliminar aquellos registros con valores ausentes (NA o $-1$) o la variable completa si fuera necesario.
- Eliminar aquellas variables cuyo porcentaje de repetición de un solo valor está por encima de un 90 %.

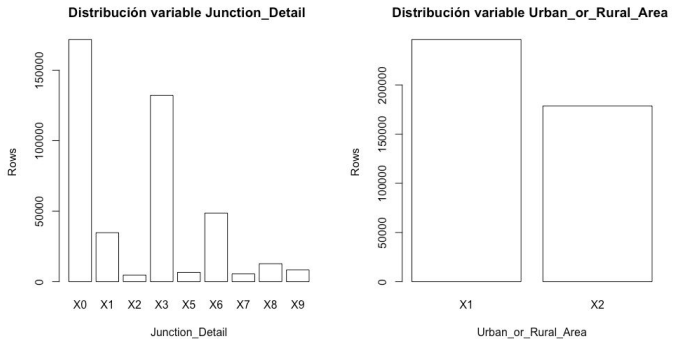



**Figura 6.14:** Distribución de variables Detalle del cruce y Tipo de Área: Rural o urbana

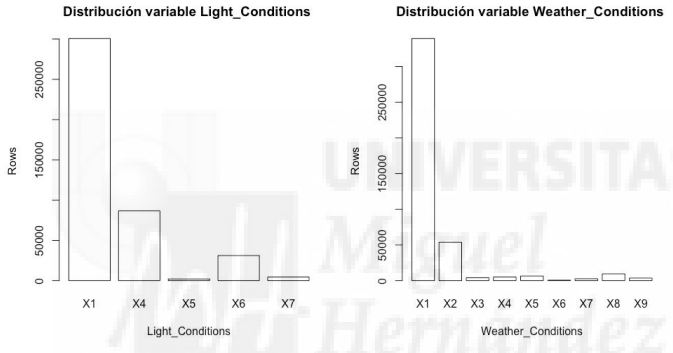



**Figura 6.15:** Distribución de variables Condiciones de luminosidad y Condiciones meteorológicas

En primer lugar, se ha comenzado eliminando las variables cuyo número de registros con valor $-1$ era muy elevado. Es necesario tener en cuenta que si se hubiera optado por eliminar los registros en lugar de las variables, habría conllevado un recorte significativo del set de datos, por lo que ha sido preferible prescindir de las siguientes variables por no contener información útil para nuestro estudio:

- Junction_Control
- X2nd_Road_Class

Adicionalmente, se optó por eliminar otra variable relativa a la referencia del vehículo en el accidente:

- Vehicle_Reference

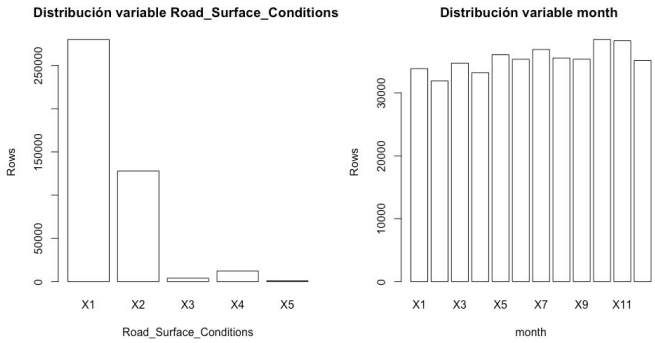



**Figura 6.16:** Distribución de variables Condiciones de la superficie de la vía y Mes

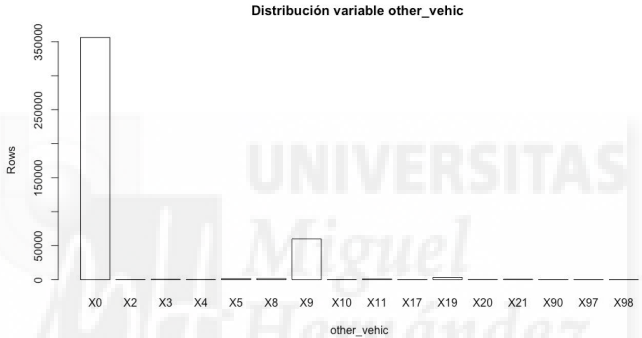



**Figura 6.17:** Distribución de variable other vehicle

Aunque esta variable resultó de gran utilidad a la hora de realizar la fusión de tablas, su aportación en la búsqueda del objetivo no es significativa.

Finalmente, se partirán de las 47 variables indicadas en la tabla 6.14.

De todas ellas, se decide realizar una segunda criba para eliminar otra serie de variables que no eran interesantes para el cálculo. Estas variables eliminadas y su motivación se encuentran en la tabla 6.15.

El dataset resultante por tanto está compuesto de 35 variables.

Adicionalmente, se ha generado una nueva variable a la que se ha denominado:

- other_vehic

Esta variable ha sido generada a posteriori mediante una programación en lenguaje *R* con motivo de conocer si existía alguna dependencia entre el tipo de vehículo contra

| Núm. de variable | Nombre |
|---|---|
| 1 | Acc_Index |
| 2 | Vehicle_Index |
| 3 | accyr |
| 4 | Vehicle_Type |
| 5 | Vehicle_Manoeuvre |
| 6 | Junction_Location |
| 7 | Skidding_and_Overturning |
| 8 | Vehicle_Leaving_Carriageway |
| 9 | Hit_Object_off_Carriageway |
| 10 | X1st_Point_of_Impact |
| 11 | Journey_Purpose_of_Driver |
| 12 | Sex_of_Driver |
| 13 | Age_Band_of_Driver |
| 14 | Engine_Capacity_.CC. |
| 15 | Propulsion_Code |
| 16 | Age_of_Vehicle |
| 17 | Driver_Home_Area_Type |
| 18 | make |
| 19 | Casualty_Reference |
| 20 | Casualty_Class |
| 21 | Sex_of_Casualty |
| 22 | Age_Band_of_Casualty |
| 23 | Casualty_Severity |
| 24 | Car_Passenger |
| 25 | Casualty_Type |
| 26 | Location_Easting_OSGR |
| 27 | Location_Northing_OSGR |
| 28 | Police_Force |
| 29 | Accident_Severity |
| 30 | Number_of_Vehicles |
| 31 | Day_of_Week |
| 32 | Hour_of_Day |
| 33 | Local_Authority_.District. |
| 34 | Local_Authority_.Highway. |
| 35 | X1st_Road_Class |
| 36 | X1st_Road_Number |
| 37 | Road_Type |
| 38 | Speed_limit |
| 39 | Junction_Detail |
| 40 | X2nd_Road_Number |
| 41 | Light_Conditions |
| 42 | Weather_Conditions |
| 43 | Road_Surface_Conditions |
| 44 | Urban_or_Rural_Area |
| 45 | Did_Police_Officer_Attend_Scene_of_Accident |
| 46 | LSOA_of_Accident_Location |
| 47 | month |

**Tabla 6.14:** Variables escogidas en primera instancia

| Núm. de variable | Nombre | Motivación |
|---|---|---|
| 3 | accyr | Posibilidad de sesgo del clasificador |
| 19 | Casualty_Reference | Variable identificadora del accidentado individualmente |
| 26 | Location_Easting_-OSGR | No usaremos localización en esta etapa |
| 27 | Location_Northing_-OSGR | No usaremos localización en esta etapa |
| 28 | Police_Force | Carece de importancia la patrulla que acudió a la escena del siniestro |
| 29 | Accident_Severity | Variable redundante respecto a la clase con la que clasificar Casualty_Severity |
| 33 | Local_Authority_.District. | Carece de importancia la autoridad local para la clasificación |
| 34 | Local_Authority_.Highway. | Carece de importancia la autoridad local para la clasificación |
| 36 | X1st_Road_Number | Carece de importancia la numeración de la vía en esta etapa |
| 40 | X2nd_Road_Number | Carece de importancia la numeración de la vía en esta etapa |
| 45 | Did_Police_Officer_Attend_Scene_of_Accident | Carece de importancia si la patrulla acudió a la escena del siniestro |
| 46 | LSOA_of_Accident_Location | No usaremos localización en esta etapa |

**Tabla 6.15:** Variables eliminadas en la segunda criba junto con su motivación

el cual se colisionaba y la severidad de los accidentados. Para construirla, se hubo de eliminar aquellos accidentes que contenían una colisión entre más de dos vehículos, ya que de lo contrario era imposible conocer contra qué vehículo impactó cada uno de ellos.

El script, por tanto, evaluó vehículo a vehículo y extrajo el tipo de vehículo contra el que impactó cada uno de los ellos. En la sección 6.9 se aportan detalles de la distribución de accidentes y su relación referente a los tipos de vehículo contra los que se impacta.

Seguidamente, con las 36 variables definitivas, se estudió qué variables poseían valores ausentes (NA o $-1$), y se procedió a eliminar los registros que presentaban esta peculiaridad.

A continuación, se muestra una porción del código en R empleado:

```
myData=subset(myData, myData$Journey_Purpose_of_Driver!=-1)
myData=subset(myData, myData$Age_Band_of_Driver!=-1)
myData=subset(myData, myData$Engine_Capacity_.CC.!=-1)
myData=subset(myData, myData$Propulsion_Code!=-1)
myData=subset(myData, myData$Age_of_Vehicle!=-1)
myData=subset(myData, myData$Driver_Home_Area_Type!=-1)
myData=subset(myData, myData$Road_Surface_Conditions!=-1)
```

Por último, se convierten todos los campos que se encuentran vacíos a valores '*NA*' reconocibles por R y con los que resulta más sencillo trabajar:

```
myData[myData==""]<-NA
```

y se borran todos aquellos valores que poseen valores '*NA*':

```
myData<-na.omit(myData)
```

Finalmente, nos queda un dataset de 36 variables libres de valores ausentes, tal y como se puede ver en la tabla 6.16 junto con el tipo de dato al que se debe transformar.

| Núm. de variable | Nombre | Tipo de dato |
|---|---|---|
| 1 | Acc_Index | Factor (323803 niveles) |
| 2 | Vehicle_Index | int |
| 4 | Vehicle_Type | Factor (16 niveles) |
| 5 | Vehicle_Manoeuvre | Factor (18 niveles) |
| 6 | Junction_Location | Factor (9 niveles) |
| 7 | Skidding_and_Overturning | Factor (6 niveles) |
| 8 | Vehicle_Leaving_Carriageway | Factor (9 niveles) |
| 9 | Hit_Object_off_Carriageway | Factor (12 niveles) |
| 10 | X1st_Point_of_Impact | Factor (5 niveles) |
| 11 | Journey_Purpose_of_Driver | Factor (7 niveles) |
| 12 | Sex_of_Driver | Factor (2 niveles) |
| 13 | Age_Band_of_Driver | Factor (11 niveles) |
| 14 | Engine_Capacity_.CC. | int |
| 15 | Propulsion_Code | Factor (8 niveles) |
| 16 | Age_of_Vehicle | int |
| 17 | Driver_Home_Area_Type | Factor (3 niveles) |
| 18 | make | Factor (331 niveles) |
| 19 | Casualty_Reference | Factor (16 niveles) |
| 20 | Casualty_Class | Factor (3 niveles) |
| 21 | Sex_of_Casualty | Factor (2 niveles) |
| 22 | Age_Band_of_Casualty | Factor (11 niveles) |
| 23 | Casualty_Severity | Factor (2 niveles) |
| 24 | Car_Passenger | Factor (3 niveles) |
| 25 | Casualty_Type | Factor (17 niveles) |
| 30 | Number_of_Vehicles | Factor (2 niveles) |
| 31 | Day_of_Week | int |
| 32 | Hour_of_Day | int |
| 35 | X1st_Road_Class | Factor (6 niveles) |
| 37 | Road_Type | Factor (6 niveles) |
| 38 | Speed_limit | int |
| 39 | Junction_Detail | Factor (9 niveles) |
| 41 | Light_Conditions | Factor (5 niveles) |
| 42 | Weather_Conditions | Factor (9 niveles) |
| 43 | Road_Surface_Conditions | Factor (5 niveles) |
| 44 | Urban_or_Rural_Area | Factor (2 niveles) |
| 47 | month | int |
| 48 | other_vehic | Factor (16 niveles) |

**Tabla 6.16:** Características escogidas junto con el tipo de dato

### 6.3.3. Estudio de valores atípicos (outliers)

El siguiente paso que se realizó consistió en observar qué muestras se encontraban numéricamente distantes del resto de los datos, es decir, observar valores atípicos o del inglés *outliers*. Sin embargo, con el fin de no eliminar valores atípicos útiles que

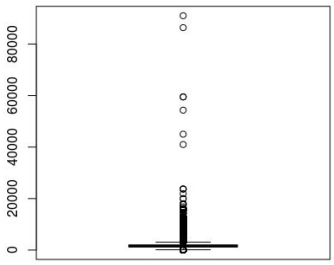



**Figura 6.18:** Representación mediante diagrama de cajas de la variable Engine_Capacity_.CC.

queramos considerar, se ha resuelto que es necesario obtener inicialmente una visión gráfica de cada variable mediante diagramas de caja o *boxplots*.

### Predictor Engine_Capacity_.CC.

Se ha representado mediante un diagrama de cajas o bloxplot la variable Engine_-Capacity_.CC., tal y como se puede observar en la figura 6.18.

```
>boxplot(datos2$Engine_Capacity_.CC.)
```

Tras analizar gráficamente los posibles valores máximo y mínimo que puede adoptar dicha variable, se presenta un inconveniente: al eliminar los *outliers* se eliminarían las motorizaciones de ciclomotores (49c.c.) o vehículos de alta gama, con motorizaciones por encima de los 3.000 c.c., junto con camiones y autobuses.

Por ello, el siguiente paso será estudiar cuál es la distribución de esos vehículos con motivo de conocer cuántas muestras aportan al set de datos.

Es posible mostrar los valores clasificados como extremos a través de la función *boxplot.stats* de *R*:

```
> boxplot.stats(datos2$Age_of_Vehicle)
```

los cuales aparecerán en el argumento out.

Con el fin de estudiar correctamente la distribución, en primer lugar, se almacenarán los *outliers* de dicha variable en una nueva denominada *engine_out* y a continuación se procederá a convertir en factor para establecer el número de accidentes registrados por motorización:

Se ha establecido un criterio de 20 accidentes como máximo para considerarlo realmente un valor atípico, tal y como se puede observar en la programación. Sin embargo, no se han encontrado motorizaciones con un número de accidentes igual o menor de 20, por tanto, puesto que no se cumple el criterio que se ha impuesto, se considerará que no se debe descartar ningún valor, es decir, no se considerarán valores atípicos para este predictor.

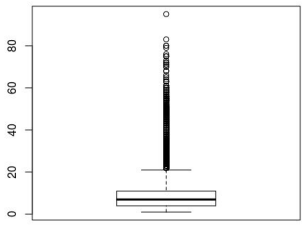



**Figura 6.19:** Representación mediante diagrama de cajas de la variable Age_of_Vehicle

### Predictor Age_of_Vehicle

La forma de analizar este predictor será de forma similar al anterior, es decir, se representará la variable mediante un diagrama de cajas y se analizarán gráficamente los niveles que se encuentren representados por este diagrama.

En la figura 6.19 se puede observar la representación.

Una vez representada dicha variable, se escogerá un criterio algo menos restrictivo que el empleado con la variable anterior, eliminando aquellos *outliers* cuya distribución sobre el total de accidentes sea menor o igual a 10. Para ello, buscaremos aquellos niveles con dicha distribución:

```
> age_out<-boxplot.stats(datos2$Age_of_Vehicle)$out
> age_out <- as.factor(age_out)

> summary(age_out)<=10
   22    23    24    25    26    27    28    29    30    31    32    33    34    35    36
FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
   37    38    39    40    41    42    43    44    45    46    47    48    49    50    51
FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
   52    53    54    55    56    57    58    59    60    61    63    64    65    66    68
TRUE TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
   70    71    72    73    75    76    79    80    83    95
TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

Como se puede observar, se han localizado valores de la variable cuya distribución sobre el total de accidentes es menor o igual a 10, que se corresponderán con todos aquellos accidentes donde han intervenido vehículos cuya edad era mayor o igual de 56 años.

Se procederá, por tanto, a eliminarlos:

```
> datos3 <- datos3[datos3$Age_of_Vehicle>=56,]
```

### Predictor Speed_limit

Otro predictor que resulta interesante analizar tras visualizar el summary de niveles de todos ellos en el apartado anterior es *Speed_limit*.

La forma de analizar este predictor será exactamente igual a los anteriores, es decir, se representará la variable mediante un diagrama de cajas y se analizarán gráficamente los niveles que se encuentren representados por este diagrama.

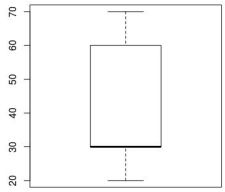



**Figura 6.20:** Representación mediante diagrama de cajas de la variable Speed_limit

```
> boxplot(datos2$Speed_limit )
```

Sin embargo, tal y como se puede observar en la figura 6.20, no existe ningún valor atípico que estudiar, por lo tanto, no se realizará acción alguna de preprocesamiento sobre este predictor.

### 6.3.4. Niveles de características con pocas muestras

La siguiente acción de preprocesamiento de variables consistirá en analizar y eliminar si procede, aquellos niveles de los predictores que posean pocas muestras y que puedan interferir en clasificadores de tipo lineal.

Para ello, analizaremos variable a variable visualizando el número de muestras por nivel y eliminando aquellas que posean un número inferior a 10 muestras.

#### Predictor Vehicle_Type

Realizando un summary de dicha variable, se puede observar que posee 16 niveles:

```
summary(datos3$Vehicle_Type)
```

| X2 | X3 | X4 | X5 | X8 | X9 | X10 | X11 | X17 | X19 | X20 | X21 | X23 | X90 | X97 | X98 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 6414 | 17709 | 6399 | 24168 | 10406 | 327382 | 843 | 12963 | 95 | 13968 | 972 | 2960 | 2 | 667 | 52 | 13 |

**Tabla 6.17:** Niveles de la variable Vehicle_Type

de los cuales, los registros X23, X97 y X98 poseen pocas muestras, pero únicamente el nivel X23 posee un número por debajo de 10.

Se eliminarán, por tanto, los accidentes para los tipos de vehículo 23.

```
> datos2<-datos2[datos2$Vehicle_Type!=X23,]

indices_a_eliminar<-subset(datos3, Vehicle_Type=='X23',
                           select=c(Acc_Index))

for (i in indices_a_eliminar[,1])
  datos4<-subset(datos4, !grepl(i, datos4$Acc_Index))
```

### 6.3.5. Reducción del conjunto de variables

#### 6.3.5.1. Variables cuyas observaciones poseen un único valor

Como se ha observado en el capítulo 2, es necesario tener en cuenta que muchos modelos, como los basados en regresión o clasificación estiman parámetros para cada predictor, por lo que si poseemos información redundante o predictores que no contengan información es posible que se añada incertidumbre en la predicción y se reduzca efectividad en el modelo.

Por ello, se ha decidido realizar una analítica previa acerca de aquellas variables que poseen varianza casi cero y estimar su eliminación, ya que podría llevar a que la efectividad del modelo sea más baja de lo que sería sin esa variable.

Se realizarán las siguientes etapas mediante un script en R:

- Comprobación de zero '*variance predictors*' o '*near-zero variance predictors*' mediante la función '*nearZeroVar()*'
- Se pueden detectar comparando las frecuencias de la $1^a$ y $2^a$ categorías más numerosas. También se debe comprobar que el número de valores distintos es pequeño

Se usa la función '*nearZeroVar()*' del paquete caret en R para buscar las variables que tengan 'varianza casi cero', aquellas que entre las dos categorías más numerosas tengan una relación de 95/5 y que tengan menos de 10 valores únicos:

```
ind.zero.var=nearZeroVar(training, freqCut = 95/5, uniqueCut = 10)
```

Al imprimir la salida de la función, se observa que no existe ninguna variable que cumpla el criterio.

Se prueba con un criterio menos estricto, con una relación 90/10 entre el primer y segundo nivel de las variables:

```
ind.zero.var = nearZeroVar ( training, freqCut = 90/10, uniqueCut = 10 )
```

En este caso, la salida de la función arroja un cumplimiento del criterio por parte de las variables 4 y 9.

A continuación, se comprueba la distribución de dichas variables:

```
> table(training[ ,4] ) # variable Vehicle_Type
   X10    X11    X17    X19     X2    X20    X21    X23     X3     X4     X5     X8     X9
X90    X97
   714  10527     68  11353   5586    783   2471      1  14159   5395  20250   8392 272158
542      2
```

Como se puede observar, la categoría o nivel de la variable Vehicle_Type que aparece con mayor frecuencia es la X9 con 272.158 registros, mientras que la siguiente que le sigue es la X5 con 20.250.

Continuando con la siguiente variable que cumplía dicho criterio, Hit_Object_off_-Carriageway, se observa que el nivel que aparece con mayor frecuencia es el X0 con 284.996 registros, mientras que el segundo en número de registros es el X10 con 21.367.

```
> table(training[ ,9] ) # variable Hit_Object_off_Carriageway
    X0     X1     X2     X3     X4     X5     X6     X7     X8     X9    X10    X11
284996   6103   5853   2301  12310    446   5274   5573     82   6634  21367   1462
```

Finalmente, se toma decisión de no eliminar ninguna de las dos variables puesto que es necesario distinguir entre el tipo de vehículo o contra qué chocó el vehículo fuera de la calzada, para estimar soluciones en base al resultado del accidente debido al tipo de predictor.

#### 6.3.5.2. Correlación de variables

Algunos conjuntos de datos contienen variables muy correlacionadas, algunas de las cuáles se podrían eliminar:

- Se pretende eliminar el mínimo número de predictores de tal forma que todas las correlaciones entre todas las variables restantes estén por debajo de un determinado valor de corte o umbral prefijado (threshold).
- Se van eliminando de forma iterativa mientras que queden variables con correlaciones por encima del threshold. Se van seleccionando las dos variables con mayor correlación y entre ellas se elimina la más correlacionadas con el resto.

Vamos a situar un umbral de $\pm 0,8$, es decir, se va a considerar que los predictores con una correlación $> 0,8$ serán dependientes.

```
datos_corr<-datos3
datos_corr <-datos_corr[,-2:-3]
w.corr<-hetcor(datos_corr) # 1
w.corr
```

Los datos de correlación se encuentran en el apéndice B.

Tras analizar los resultados, es posible concluir de la siguiente forma:

- Los predictores Hit_Object_off_Carriageway y Vehicle_Leaving_Carriageway poseen un índice de correlación de $0,82$.
- Las variables Sex_of_Casualty y Sex_of_Driver poseen un índice de correlación de $0,828$.

| | ROC | Accuracy | Kappa |
|---|---|---|---|
| Modelo con ambas variables | 0,7781107 | 0,8692888 | 0,119709 |

**Tabla 6.18:** Características escogidas junto con el tipo de dato

Como se puede observar, aunque aparecen correlados los predictores Sex_of_Casualty y Sex_of_Driver, si eliminamos una de las dos perderemos información acerca de:

- El sexo de la persona que conducía en el momento del accidente
- El sexo de cada uno de los ocupantes del vehículo siniestrado en el accidente

Por ello, se decide no eliminar dicho predictor, puesto que puede ser de utilidad conocer el valor de estas variables desde un punto de vista estadístico.

Por otro lado, respecto a los predictores *Hit_Object_off_Carriageway* y *Vehicle_Leaving_Carriageway*, se toma otra decisión, que será la de eliminar uno de los dos.

Previamente a eliminar uno de ellos, se realizará una prueba que consistirá en estudiar el área encerrada bajo la curva ROC (AUC) así como el *accuracy* de un modelo de tipo BayesGLM manteniendo ambas variables primero y, a continuación, eliminando la variable Vehicle_Leaving_Carriageway. El resultado se muestra en la tabla 6.18.

## 6.4. Metodología de la investigación de la accidentalidad en Reino Unido

La metodología de la investigación para llevar a cabo la parte experimental de extracción del modelo consistirá en las siguientes fases:

1. Selección de características o predictores mediante distintos métodos
2. Comparativa entre los selectores y elección de los que mejores resultados en base al Área Encerrada bajo la Curva ROC proporcionen
3. Random Forest (acumulativo) en base a las características escogidas por cada selector
4. Inserción de la elección en la fase 3 de predictores en modelo BayesGLM, donde se rechazarán aquellas variables que no sean significativas o porque el signo del coeficiente no corresponda al tipo de efecto esperado.

## 6.5. Características de partida

Como se ha visto en el capítulo anterior, el dataset que se emplea inicialmente posee un total de 70 características repartidas en las siguientes tablas:

- Tabla de vehículos (24 características)
- Tabla de accidentados (14 características)
- Tabla de accidentes (32 características)

Después de unificar algunas características para obtener un único predictor, y de descartar aquellas que, en principio, podrían no presentar ninguna relación con la investigación a realizar con este trabajo, el set de datos ha quedado recortado inicialmente a características:

```
Vehicle_Type, Vehicle_Manoeuvre, Junction_Location, Skidding_and_Overturning,
Vehicle_Leaving_Carriageway, Hit_Object_off_Carriageway, First_Point_of_Impact,
Journey_Purpose_of_Driver, Sex_of_Driver, Age_Band_of_Driver,
Engine_Capacity_.CC., Propulsion_Code, Age_of_Vehicle, Driver_Home_Area_Type,
make, Casualty_Severity, Casualty_Class, Sex_of_Casualty, Age_Band_of_Casualty,
Car_Passenger, Casualty_Type, Number_of_Vehicles, Day_of_Week, Hour_of_Day,
First_Road_Class, Road_Type, Speed_limit, Junction_Detail, Light_Conditions,
Weather_Conditions, Road_Surface_Conditions, Urban_or_Rural_Area, month, other_vehic
```

De estas 34 características, ya se ha prescindido previamente de los identificadores de cada tabla, por lo que el dataset con el que trabajaremos desde el inicio será este. Se extrae la característica de interés, *Casualty_Severity*, que se divide en dos clases, clase 0 para los accidentes graves y fallecimientos, y clase 1 para los accidentes leves, consiguiendo de esta forma una variable dicotómica con la que trabajar.

En las siguientes secciones se estudiarán diferentes métodos para realizar una selección de características que no penalice en exceso el resultado del área encerrada bajo la curva ROC de nuestra clasificación.

## 6.6. Selección de características

Se pretenden comparar algunos de los métodos de selección de características que se han detallado en el capítulo 2.

Concretamente, se van a emplear los siguientes métodos de selección de predictores:

- Selección mediante el área encerrada bajo la curva ROC por cada predictor individualmente.
- Método ReliefF.
- Método Backward BayesGLM.

- RFE-Random Forest.
- RFE-$Chi^2$.

Para ello, inicialmente, se van a realizar una serie de desarrollos en lenguaje R con el fin de obtener los datos individuales a través de cada uno de los métodos anteriormente enumerados.

Vamos a diseñar un script en R llamado Metodos_seleccion_caracteristicas.R, para aplicar cada uno de los métodos. El script en R empleado se encuentra en el anexo D.

Comenzaremos con el cálculo del área debajo de la curva ROC cuando la variable a predecir es un factor. En primer lugar, se deben leer los datos preprocesados anteriormente y convertir el tipo de dato de algunos predictores a tipo de dato factor y mantener el resto como tipo de dato numérico. En las siguientes líneas de pseudocodigo se puede apreciar con mayor detalle lo que se comenta:

```
INICIO
    LEER 'datos_other_vehic_filtrada.csv'
    VARIABLES
            FACTOR 'predictores_factor' as factor
            INT 'predictores_num'
    ELIMINAR columnas de indices no útiles
            datos <-datos[,-1:-2]
FIN
```

### 6.6.1. Uso del método de validación cruzada para generar los splits de datos

Antes de comenzar a seleccionar características y realizar el ajuste del modelo, se procede a dividir el set de datos en dos subconjuntos: *training* y *testing*. Sin embargo, se ha de garantizar que los datos del subset *training* son independientes de los datos del subset de *testing* respecto a la clase '*Casualty_Severity*'.

Con el fin de asegurar esta independencia, se empleará el método de validación cruzada o K-fold. El método K-fold particionará el set de entrenamiento en K subconjuntos de aproximadamente el mismo número de muestras aleatoriamente que tenía éste. Un subconjunto se entrena usándolo como datos de prueba y el resto grupos (K-1) como datos de entrenamiento, y repetiremos este proceso K veces con cada subconjunto de prueba, donde el primer grupo de muestras se devolverá al set de entrenamiento.

### 6.6.2. Selección de predictores mediante su Área Encerrada bajo la Curva ROC

Para usar el método basado en el área bajo la curva ROC para cada predictor respecto a las clases, emplearemos la función de R *filterVarImp*. *filterVarImp* usa una metodología de selección basada en filtro a través de métodos de selección lineales.

En cuanto a su programación en R, es bastante sencilla, ya que únicamente se ha de pasar el set de datos, sin la variable a predecir, como primer parámetro, y a continuación, la variable a predecir como segundo parámetro:

```
rocValues<-filterVarImp(x=datos[,-15],
                        y=datos$Casualty_Severity)
```

El test 6.19 por tanto cuenta con los siguientes datos:

| Test | Predictores | Selector |
|---|---|---|
| 1 | 32 | Filter |

**Tabla 6.19:** Test de selección de predictores mediante su Área Encerrada bajo la Curva ROC

El resultado acerca de la importancia de los predictores arrojado por el script para cada una de las clases X0 (Fallecimiento o víctima grave) y X1 (víctima leve) se indica en la tabla 6.20.

| | X0 | X1 |
|---|---|---|
| Casualty_Type | 0,6441775 | 0,6441775 |
| Number_of_Vehicles | 0,6123903 | 0,6123903 |
| Vehicle_Type | 0,5760128 | 0,5760128 |
| Engine_Capacity_.CC. | 0,5653059 | 0,5653059 |
| Sex_of_Casualty | 0,5646086 | 0,5646086 |
| Junction_Location | 0,5621597 | 0,5621597 |
| Sex_of_Driver | 0,5610133 | 0,5610133 |
| Junction_Detail | 0,5551340 | 0,5551340 |
| Car_Passenger | 0,5490011 | 0,5490011 |
| First_Point_of_Impact | 0,5412086 | 0,5412086 |
| Propulsion_Code | 0,5345383 | 0,5345383 |
| other_vehic | 0,5230773 | 0,5230773 |
| Weather_Conditions | 0,5169330 | 0,5169330 |
| Road_Surface_Conditions | 0,5162206 | 0,5162206 |
| Age_Band_of_Driver | 0,5080578 | 0,5080578 |
| Day_of_Week | 0,5011997 | 0,5011997 |
| month | 0,5011133 | 0,5011133 |
| Hour_of_Day | 0,4997856 | 0,4997856 |
| First_Road_Class | 0,4997182 | 0,4997182 |
| make | 0,4953184 | 0,4953184 |
| Journey_Purpose_of_Driver | 0,4907514 | 0,4907514 |
| Age_of_Vehicle | 0,4806172 | 0,4806172 |
| Driver_Home_Area_Type | 0,4786997 | 0,4786997 |
| Road_Type | 0,4758606 | 0,4758606 |
| Light_Conditions | 0,4742810 | 0,4742810 |
| Age_Band_of_Casualty | 0,4712232 | 0,4712232 |
| Hit_Object_off_Carriageway | 0,4656716 | 0,4656716 |
| Speed_limit | 0,4638037 | 0,4638037 |
| Urban_or_Rural_Area | 0,4584339 | 0,4584339 |
| Skidding_and_Overturning | 0,4576850 | 0,4576850 |
| Casualty_Class | 0,4459771 | 0,4459771 |
| Vehicle_Manoeuvre | 0,4327942 | 0,4327942 |

**Tabla 6.20:** Importancia de los predictores mediante Área Encerrada bajo la Curva ROC

A la vista del resultado del área bajo la curva ROC para cada predictor, se puede observar que las variables que según este método son más importantes corresponden a *Casualty_Type* y *Number_of_Vehicles*. Por el contrario entre aquellas con menor AUC ROC son *Vehicle_Manoeuvre*, *Casualty_Class*, *Skidding_and_Overturning* o *Urban_or_Rural_Area*.

### 6.6.3. Selección de predictores mediante el algoritmo ReliefF

Para usar el método Relief, emplearemos la función de R *attrEval* a la cual hemos de pasarle el dataset indicando cuál es la variable a predecir, el estimador a utilizar será ReliefFequalK, el cual se trata de un algortitmo de ReliefF donde las k instancias más cercanas tienen el mismo peso. *attrEval* calculará varias versiones de ReliefF empleando dicho estimador:

```
reliefValues<-attrEval(Casualty_Severity ~ ., data=datos,
                       estimator="ReliefFequalK",
                       ReliefIterations=33)
```

Este test se resume en la tabla 6.21.

| Test | Predictores | Selector | Estimador |
|---|---|---|---|
| 2 | 32 | ReliefF | ReliefFequalK |

**Tabla 6.21:** Test de selección de predictores mediante ReliefF

El resultado acerca de la importancia de los predictores arrojado por el script para cada una de las clases X0 y X1 se puede observar en la tabla 6.22.

Con motivo de ordenar los predictores del más importante al menos importante, debemos tener en cuenta que los valores se encuentran el intervalo [-1,1] , siendo 1 el que posee más importancia y -1 el que menos.

Con el fin de usar una aproximación basada en la permuta para investigar los valores observados de la estadística devuelta por ReliefF, podemos emplear la función *permuteRelief*.

En esta función, las puntuaciones de cada predictor se calculan utilizando los datos originales y después de que los datos resultado aleatoriamente se mezclan (*nperm* veces). Se determinan la media y la desviación estándar de los valores permutados y se determina una versión estandarizada de las puntuaciones observadas restando las medias permutados de los valores originales y dividiendo cada uno por la desviación estándar correspondiente.

Se puede ver, a continuación, la parte de código en R que hace uso de dicha función:

```
perm<-permuteRelief(x=datos[,-15],
                    y=datos$Casualty_Severity,
                    nperm=500,
                    estimator="ReliefFequalK",
                    ReliefIterations=33)
```

| Predictor | Importancia |
|---|---|
| Hour_of_Day: | 1,458937e-01 |
| Age_Band_of_Driver: | 1,090909e-01 |
| First_Point_of_Impact: | 1,030303e-01 |
| make: | 8,484848e-02 |
| Speed_limit: | 7,878788e-02 |
| Age_Band_of_Casualty: | 6,363636e-02 |
| Age_of_Vehicle: | 6,273850e-02 |
| month: | 5,968779e-02 |
| Vehicle_Manoeuvre: | 5,454545e-02 |
| First_Road_Class: | 5,454545e-02 |
| Propulsion_Code: | 5,151515e-02 |
| Junction_Location: | 5,151515e-02 |
| Urban_or_Rural_Area: | 5,151515e-02 |
| Journey_Purpose_of_Driver: | 3,939394e-02 |
| Hit_Object_off_Carriageway: | 2,727273e-02 |
| Skidding_and_Overturning: | 2,727273e-02 |
| Day_of_Week: | 2,727273e-02 |
| Vehicle_Type: | 2,424242e-02 |
| Road_Surface_Conditions: | 2,424242e-02 |
| Number_of_Vehicles: | 2,121212e-02 |
| other_vehic: | 2,121212e-02 |
| Weather_Conditions: | 1,515152e-02 |
| Driver_Home_Area_Type: | 1,212121e-02 |
| Sex_of_Driver: | 1,212121e-02 |
| Engine_Capacity_.CC.: | 3,680937e-03 |
| Light_Conditions: | 3,030303e-03 |
| Road_Type: | 3,030303e-03 |
| Sex_of_Casualty: | 3,030303e-03 |
| Casualty_Type: | -2,018587e-17 |
| Car_Passenger: | -2,691450e-17 |
| Junction_Detail: | -9,090909e-03 |
| Casualty_Class: | -2,424242e-02 |

**Tabla 6.22:** Importancia de los predictores mediante ReliefF

| Test | Predictores | Selector | Estimador |
|---|---|---|---|
| 3 | 32 | permuteRelief | ReliefFequalK |

**Tabla 6.23:** Test de selección de predictores mediante permuteRelief

El test 6.23 por tanto cuenta con los siguientes datos:

Si ordenamos la clasificación que nos devuelve la función permuteRelief en su opción *'observed'*, obtenemos la siguiente clasificación de mejor predictor a peor (tabla 6.24), donde los valores más alejados de cero serían los mejores predictores, mientras que los más cercanos a cero serían los peores.

En la figura 6.21, se muestra un histograma donde se puede evaluar la distribución de las permutaciones de las puntuaciones de los predictores cuando no existe relación entre los predictores y las clases.

| Característica | Puntuación |
|---|---|
| First_Point_of_Impact: | 0,142424242 |
| Journey_Purpose_of_Driver: | 0,136363636 |
| Age_Band_of_Driver: | 0,133333333 |
| Age_Band_of_Casualty: | 0,106060606 |
| Hour_of_Day: | 0,090425999 |
| month: | 0,089164371 |
| Junction_Detail: | 0,075757576 |
| Speed_limit: | 0,072727273 |
| First_Road_Class: | 0,069696970 |
| Junction_Location: | 0,060606061 |
| Vehicle_Manoeuvre: | 0,054545455 |
| Propulsion_Code: | 0,048484848 |
| Vehicle_Type: | 0,039393939 |
| Sex_of_Driver: | 0,039393939 |
| Road_Surface_Conditions: | 0,036363636 |
| Skidding_and_Overturning: | 0,033333333 |
| Hit_Object_off_Carriageway: | 0,030303030 |
| Road_Type: | 0,030303030 |
| Casualty_Type: | 0,027272727 |
| Age_of_Vehicle: | 0,024990647 |
| Day_of_Week: | 0,024242424 |
| Urban_or_Rural_Area: | 0,018181818 |
| Light_Conditions: | 0,018181818 |
| Number_of_Vehicles: | 0,015151515 |
| Driver_Home_Area_Type: | 0,015151515 |
| Car_Passenger: | 0,009090909 |
| make: | 0,009090909 |
| other_vehic: | 0,006060606 |
| Weather_Conditions: | 0,006060606 |
| Engine_Capacity_.CC.: | -0,001641732 |
| Casualty_Class: | -0,003030303 |
| Sex_of_Casualty: | -0,027272727 |

**Tabla 6.24:** Importancia de los predictores mediante ReliefF

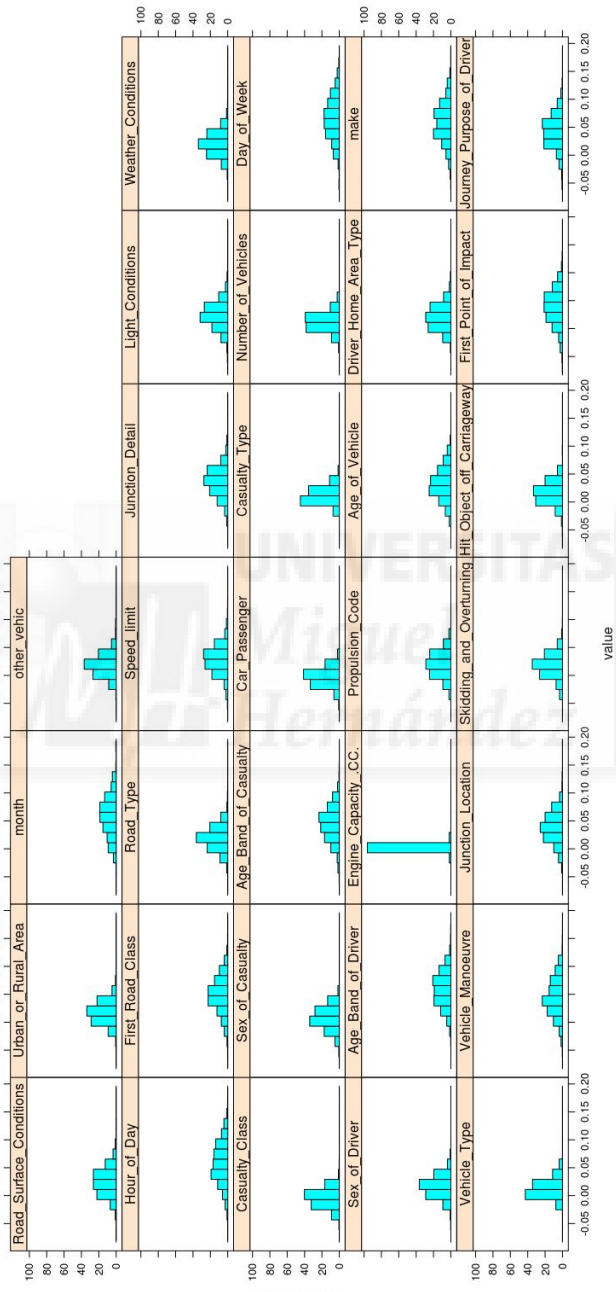



**Figura 6.21:** Aplicación de la puntuación ReliefF para todos los predictores. Se muestra la distribución de permutación de las puntuaciones cuando no existe relación entre los predictores y las clases

### 6.6.4. Selección de características mediante métodos Wrapper

#### 6.6.4.1. Selección de características mediante ejecuciones recursivas de BayesGLM

Para este experimento partimos de las siguientes 33 características

```
Vehicle_Type, Vehicle_Manoeuvre, Junction_Location, Skidding_and_Overturning,
Vehicle_Leaving_Carriageway, Hit_Object_off_Carriageway, First_Point_of_Impact,
Journey_Purpose_of_Driver Sex_of_Driver Age_Band_of_Driver, Engine_Capacity_.CC.,
Propulsion_Code, Age_of_Vehicle, Driver_Home_Area_Type, make, Casualty_Severity,
Casualty_Class, Sex_of_Casualty, Age_Band_of_Casualty, Car_Passenger, Casualty_Type,
Number_of_Vehicles, Day_of_Week, Hour_of_Day, First_Road_Class, Road_Type,
Speed_limit, Junction_Detail, Light_Conditions, Weather_Conditions, Road_Surface_Conditions,
Urban_or_Rural_Area, month, other_vehic
```

El siguiente test cuenta con los siguientes datos que se muestran en la tabla 6.25.

| Test | Predictores | Selector |
|---|---|---|
| 4 | 32 | BayesGLM |

**Tabla 6.25:** Test de selección de predictores mediante BayesGLM

A continuación, en la tabla 6.26 se muestra un breve resumen de cada una de las ejecuciones. Es necesario indicar que, aunque no se ha reflejado en esta tabla, tras cada ejecución, para aquellas clases de variables cuyo p-valor no era trascendente, se han fusionado con aquella clase de la característica que coincidían con el valor del *intercept*.

| Orden de ejecución | Predictores | Tiempo ejecución (s) | AUC ROC |
|---|---|---|---|
| Ejecución 1 | 32 | 493392,852 | 0,778192 |
| Ejecución 2 | 31 | 9371,572 | 0,7782545 |
| Ejecución 3 | 31 | 7843,596 | 0,7782458 |
| Ejecución 4 | 31 | 11140,768 | 0,7782421 |
| Ejecución 5 | 31 | 7750,028 | 0,7782394 |

**Tabla 6.26:** Resumen de cada una de las ejecuciones mediante BayesGLM del test 4

Una vez realizadas todas las ejecuciones, el orden devuelto de mayor a menor importancia se refleja en la tabla 6.27.

La predicción con el dataset de testing tras la ejecución 5 se puede observar también en la tabla 6.27.

La figura 6.22 muestra la curva ROC.

El *summary* de la última ejecución de este experimento, lo podemos encontrar en el anexo G. Finalmente, después de 5 ejecuciones el modelo se seleccionó con: 31 predictores.

| Característica | Puntuación |
|---|---|
| Number_of_Vehicles | 100,000000 |
| Vehicle_Type | 89,144515 |
| Vehicle_Manoeuvre | 79,079145 |
| Engine_Capacity_.CC. | 75,747445 |
| Sex_of_Casualty | 75,707623 |
| Sex_of_Driver | 73,869103 |
| Junction_Detail | 71,179856 |
| First_Point_of_Impact | 66,453336 |
| Propulsion_Code | 60,460246 |
| Junction_Location | 55,499340 |
| other_vehic | 55,342361 |
| Car_Passenger | 52,209585 |
| Weather_Conditions | 51,004011 |
| Road_Surface_Conditions | 50,952580 |
| month | 44,119952 |
| Hour_of_Day | 43,193619 |
| First_Road_Class | 42,758930 |
| Age_Band_of_Casualty | 42,632119 |
| make | 41,349274 |
| Journey_Purpose_of_Driver | 37,119789 |
| Age_of_Vehicle | 33,576902 |
| Driver_Home_Area_Type | 32,711470 |
| Light_Conditions | 30,131260 |
| Road_Type | 29,590050 |
| Hit_Object_off_Carriageway | 25,239449 |
| Speed_limit | 25,238209 |
| Urban_or_Rural_Area | 22,087456 |
| Skidding_and_Overturning | 21,958264 |
| Vehicle_Leaving_Carriageway | 18,925325 |
| Casualty_Class | 2,983132 |
| Casualty_Type | 0,000000 |

**Tabla 6.27:** Importancia de los predictores mediante BayesGLM

| Threshold | Sensitivity | Specificity |
|---|---|---|
| 0,1416629 | 0,6973171 | 0,7134452 |

**Tabla 6.28:** Predicción con testing tras ejecución 5

#### 6.6.4.2. Selección de características mediante Recursive Feature Elimination (rfe)

**Random Forest-RFE (RF-RFE)**

El principal inconveniente que posee el uso de un selector de características basado en Recursive Feature Elimination es que no es capaz de manejar características de tipo factor con más de 53 niveles, sin embargo, la característica 'make' supera con creces esta limitación, ya que posee más de 300 niveles.

Por ello, se ha tomado el criterio de escoger únicamente aquellos niveles de la variable

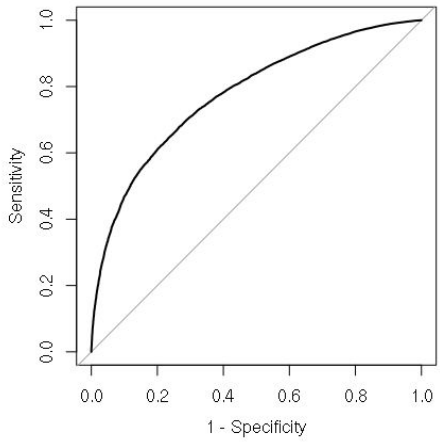



**Figura 6.22:** Curva ROC tras la última ejecución del test 3

'make' que tras la primera ejecución de BayesGLM presentaba algún tipo de importancia para el modelo según el p-valor obtenido.

Tanto la secuencia de control de entrenamiento donde se define el método de Validación Cruzada, como la secuencia de entrenamiento del modelo RFE, se puede observar a continuación, como información adicional al lector para su reutilización si lo considera necesario:

```
control <- rfeControl(functions = rfFuncs,
                      method = "repeatedcv",
                      repeats = 5,
                      verbose = FALSE)
subsets <- c(1:32)
results <- rfe(trainX, training$Casualty_Severity, subsets, rfeControl=control)
```

En la tabla 6.29 se pueden observar los datos con los que se ha realizado.

| Test | Predictores | Selector | Tiempo ejecución (horas) |
|---|---|---|---|
| 4 | 32 | RFE-RF | 117 |

**Tabla 6.29:** Test de selección de predictores mediante RFE-RF

En la tabla 6.30 se muestra el orden de mayor a menor importancia devuelto por el selector.

El valor del 'Accuracy' devuelto por RFE-RF para cada una de las variables, se presenta en la figura 6.23.

| Característica |
|---|
| Hit_Object_off_Carriageway |
| Vehicle_Manoeuvre |
| Age_Band_of_Casualty |
| Hour_of_Day |
| First_Point_of_Impact |
| Speed_limit |
| First_Road_Class |
| Age_Band_of_Driver |
| Skidding_and_Overturning |
| Engine_Capacity_.CC. |
| Light_Conditions |
| Age_of_Vehicle |
| Urban_or_Rural_Area |
| Number_of_Vehicles |
| Journey_Purpose_of_Driver |
| Road_Type |
| Casualty_Type |
| Junction_Detail |
| month |
| Road_Surface_Conditions |
| Weather_Conditions |
| Junction_Location |
| Day_of_Week |
| Driver_Home_Area_Type |
| other_vehic |
| Vehicle_Type |
| Casualty_Class |
| Propulsion_Code |
| Sex_of_Driver |
| Sex_of_Casualty |
| make |
| Car_Passenger |

**Tabla 6.30:** Importancia de los predictores mediante RFE-RF

### Uso del paquete FSelector para RF-RFE

A través del paquete FSelector, es posible obtener también la importancia de cada variable mediante la función *random.forest.importance()*.

Hemos realizado una nueva batería de tests en R para extraer la importancia a través del paquete FSelector, no solamente para RF-RFE, sino para más tipos de selectores que nos permitan extraerlas.

En este test, se usará la función *random.forest.importance()* para obtener la importancia de cada variable mediante una puntuación.

Adicionalmente, la función *cutoff.biggest.diff* permite identificar de forma automática aquellas características que poseen una importancia elevada tras el ranking.

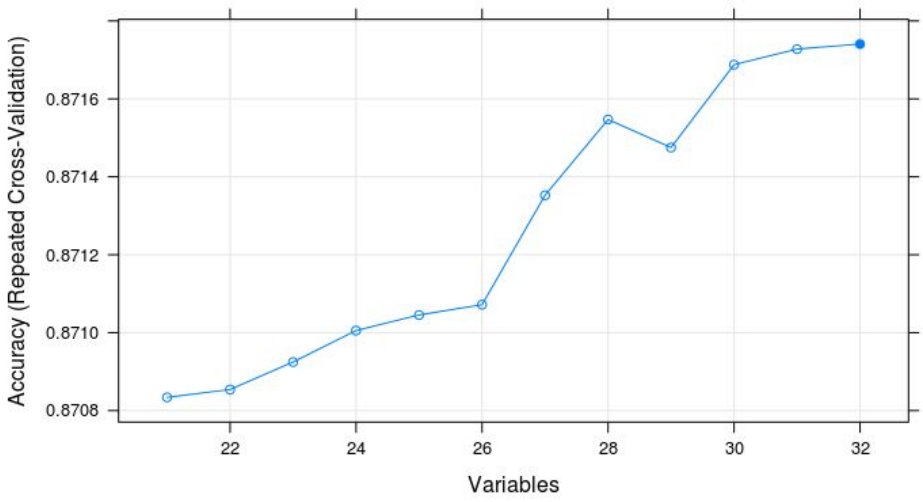



**Figura 6.23:** Representación de la precisión (accuracy) devuelta por RFE en función del número y orden de variables utilizadas

Además, *cutoff.k* permite devolver las k características que ha estimado como más importantes, y la función *cutoff.k.percent* devuelve el k porcentaje de las características de mayor relevancia:

El test 6.31 cuenta con los siguientes datos:

| Test | Predictores | Selector | Tiempo ejecución (horas) |
|---|---|---|---|
| 5 | 32 | RFE-RF | 2 |

**Tabla 6.31:** Test de selección de predictores mediante RFE-RF a través de la librería Fselector

Como se puede observar, el tiempo de ejecución se ha reducido considerablemente pasando de más de 100 horas (tabla 6.30) a sólo 2.

Adicionalmente, en la figura 6.24 se puede observar el orden de importancia de los predictores devuelto por el modelo.

A continuación, se va a proceder a insertar variable a variable de nuevo a Random Forest en el orden que me ha proporcionado el selector y se va a estimar el Área Bajo la curva ROC con cada variable de forma acumulada de una en una.

El procedimiento de la inserción, así como el resultado en cada iteración, se puede encontrar en el anexo H.

A continuación, se generará una gráfica de la evolución del valor de AUC ROC respecto a la inserción de cada variable, la cual se puede observar en la figura 6.25.

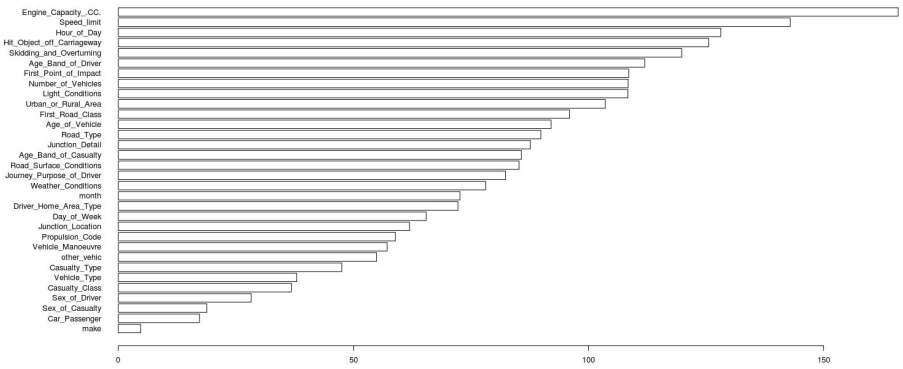



**Figura 6.24:** Representación del orden de importancia de los predictores devuelto por RFE-RF mediante la librería Fselector

| Test | Predictores | Selector | Tiempo ejecución (horas) |
|---|---|---|---|
| 6 | 32 | RFE-RF | 360 |

**Tabla 6.32:** Test de selección de predictores mediante RFE-RF a través de la librería Fselector

Como resultado de la figura anterior, es posible valorar la eliminación de aquellas variables que han resultado últimas en la clasificación y que no generan ninguna mejora en la curva, es decir, de la variable 27 en adelante. En la siguiente tabla se puede observar el orden devuelto:

```
 [1] "Engine_Capacity_.CC."        "Speed_limit"                 "Hour_of_Day"
 [4] "Hit_Object_off_Carriageway" "Skidding_and_Overturning"    "Age_Band_of_Driver"
 [7] "First_Point_of_Impact"       "Number_of_Vehicles"          "Light_Conditions"
[10] "Urban_or_Rural_Area"         "First_Road_Class"            "Age_of_Vehicle"
[13] "Road_Type"                   "Junction_Detail"             "Age_Band_of_Casualty"
[16] "Road_Surface_Conditions"     "Journey_Purpose_of_Driver"   "Weather_Conditions"
[19] "month"                       "Driver_Home_Area_Type"       "Day_of_Week"
[22] "Junction_Location"           "Propulsion_Code"             "Vehicle_Manoeuvre"
[25] "other_vehic"                 "Casualty_Type"               "Vehicle_Type"
[28] "Casualty_Class"              "Sex_of_Driver"               "Sex_of_Casualty"
[31] "Car_Passenger"               "make"
```

### Uso del paquete FSelector para Chi$^2$

Se ha empleado el selector Chi$^2$ para obtener otra clasificación adicional mediante la librería Fselector.

En este caso, el orden de importancia devuelto por el selector será el que se refleja en la tabla 6.33.

Con esta salida vamos a ir probando a ir añadiendo variable a variable en orden de importancia tras el test de Chi Cuadrado hasta haber computado todas ellas

En la figura 6.26 se puede observar la evolución del Área Bajo la Curva ROC a medida que se insertaban características.

## 6.7. Creación de un modelo lineal mediante BayesGLM

En esta sección se van a realizar distintos tests a partir de las clasificaciones que se han realizado a través de diversos clasificadores. Para ello, se entrenará un clasificador de tipo BayesGLM para obtener un modelo que aporte buenos resultados de sensibilidad y especificidad.

### 6.7.1. Modelo BayesGLM basado en el selector de predictores bajo la curva ROC

Se realiza un test donde se eliminan las tres últimas variables de las cuales hemos obtenido peor resultado en cuanto a su área encerrada bajo la curva ROC respecto a las clases `Casualty_Severity=0` y `Casualty_Severity=1` a través de la selección de predictores realizada anteriormente en base al Área Encerrada bajo la Curva ROC de cada predictor.

Dichas características son:

- Skidding_and_Overturning

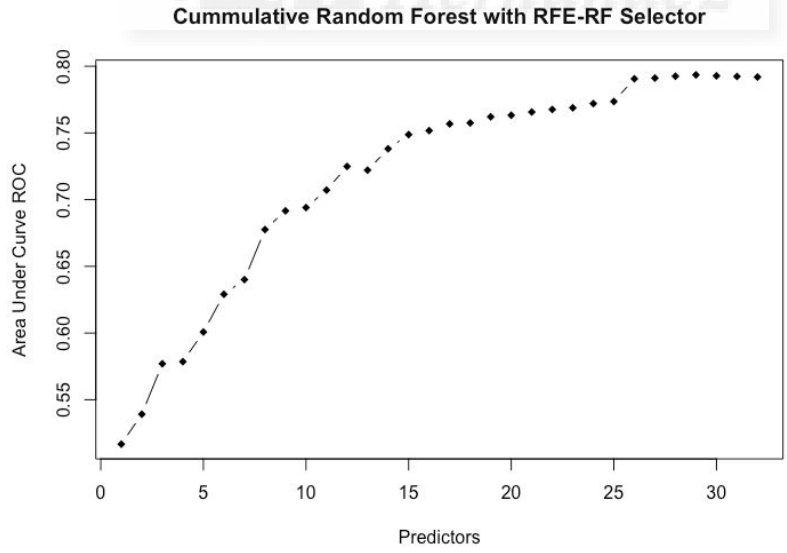



**Figura 6.25:** Representación del valor de AUC ROC devuelta por RFE-RF en función del número y orden de variables utilizadas

| Característica | Importancia |
|---|---|
| Casualty_Type | 0,26472508 |
| Engine_Capacity_.CC. | 0,19339291 |
| Vehicle_Type | 0,19273227 |
| Casualty_Class | 0,16611698 |
| Number_of_Vehicles | 0,16425898 |
| Vehicle_Manoeuvre | 0,13737229 |
| First_Point_of_Impact | 0,12226018 |
| Junction_Detail | 0,09003535 |
| Junction_Location | 0,08920530 |
| Sex_of_Casualty | 0,08824004 |
| Sex_of_Driver | 0,08797660 |
| Hit_Object_off_Carriageway | 0,08382866 |
| Car_Passenger | 0,07904361 |
| Age_Band_of_Casualty | 0,07869042 |
| Skidding_and_Overturning | 0,07222237 |
| Speed_limit | 0,07061335 |
| Hour_of_Day | 0,05769409 |
| Urban_or_Rural_Area | 0,05759697 |
| Age_of_Vehicle | 0,05621477 |
| Road_Type | 0,05202013 |
| Light_Conditions | 0,05022106 |
| other_vehic | 0,05002786 |
| Propulsion_Code | 0,04857275 |
| Age_Band_of_Driver | 0,04722078 |
| Driver_Home_Area_Type | 0,03446762 |
| Weather_Conditions | 0,03286249 |
| Journey_Purpose_of_Driver | 0,02764269 |
| Road_Surface_Conditions | 0,02614211 |
| First_Road_Class | 0,02565143 |
| Day_of_Week | 0,02376703 |
| month | 0,01563582 |

**Tabla 6.33:** Importancia de los predictores mediante $Chi^2$

- Casualty_Class
- Vehicle_Manoeuvre

A continuación, se utilizan el resto de predictores para entrenar un modelo BayesGLM con el fin de aumentar el valor del área encerrada bajo la curva ROC tras el entrenamiento.

| Test | Predictores | Selector | Tiempo ejecución (horas) | AUC ROC |
|---|---|---|---|---|
| 7 | 29 | Curva ROC | 3,5 | 0,7722236 |

**Tabla 6.34:** Test de entrenamiento de modelo basado en BayesGLM con selector AUC ROC

Tras la primera ejecución, se puede comprobar que eliminando esos 3 predictores y pasando a tener únicamente 29 predictores, el valor de la curva ROC no sólamente no ha mejorado sino que ha empeorado, tal y como se puede comprobar en la tabla 6.34.

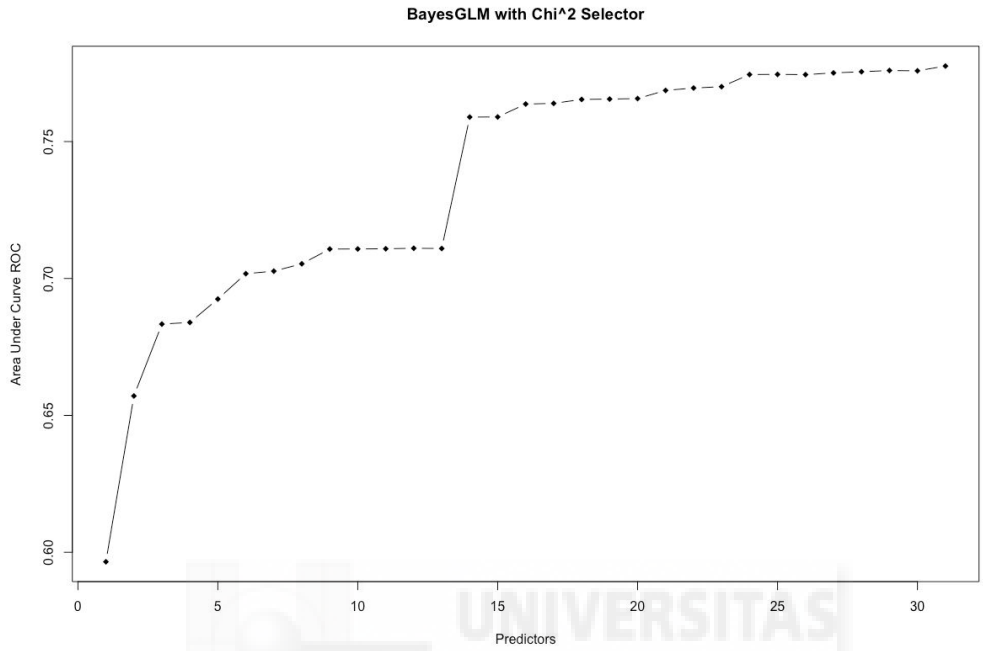



**Figura 6.26:** Representación del valor de AUC ROC devuelta por BayesGLM en función del número y orden de variables devueltas por el clasificador $Chi^2$

### 6.7.2. Modelo BayesGLM basado en el selector de predictores ReliefF

Para el selector ReliefF se eliminan aquellas características más alejadas de cero por la parte negativa, es decir:

- Junction_Detail: -9,090909e-03
- Casualty_Class: -2,424242e-02

En el siguiente test (tabla 6.35) se van insertando al modelo BayesGLM el resto de predictores en orden con el fin de entrenarlo y obtener el Área Bajo la Curva ROC.

| Test | Predictores | Selector | Tiempo ejecución (horas) | AUC ROC |
|---|---|---|---|---|
| 8 | 30 | ReliefF | 3,7 | 0,7757633 |

**Tabla 6.35:** Test de entrenamiento de modelo basado en BayesGLM con selector ReliefF

Como se puede observar, el área encerrada bajo la curva ROC ha mejorado respecto al test anterior.

A la vista de los resultados, se aumentará el número de características a eliminar, aunque esta vez se atenderán a resultados distintos en el clasificador.

Para este test, el número de características a eliminar es de 8, fijando para ello un umbral de puntuación de 0,02. Las 22 características situadas por encima de dicho umbral de clasificación se han mantenido.

Las variables eliminadas del set de datos serán, por tanto:

- Hour_of_Day
- Speed_limit
- Number_of_Vehicles
- Sex_of_Casualty
- Age_of_Vehicle
- Engine_Capacity_.CC.
- Car_Passenger
- Casualty_Class

A continuación, se añadirá una a una y por el orden del ranking establecido por ReliefF a un entrenamiento mediante BayesGLM:

| Test | Predictores | Selector | Tiempo ejecución (horas) | AUC ROC |
|---|---|---|---|---|
| 9 | 24 | BayesGLM | 3,6 | 0,7734127 |

**Tabla 6.36:** Test de entrenamiento de modelo basado en BayesGLM con selector ReliefF

Como se puede observar, mediante este modelo se ha mejorado sensiblemente el AUC ROC respecto al método de selección basado en el área de los predictores bajo la curva ROC en base a las clases definidas por `Casualty_Severity`.

### 6.7.3. Modelo BayesGLM basado en el selector de predictores $Chi^2$

Para la construcción de este modelo basado en BayesGLM, se ha insertado en el orden que el selector basado en RFE-$Chi^2$ ha devuelto en su ranking para estimar el AUC acumulado al insertar variable a variable de esta forma.

En la tabla 6.37 se puede ver el resultado.

### 6.7.4. Comparativa entre selectores de características

En las figuras 6.28, 6.29 y 6.30 se muestra un cuadro resumen acerca del estudio comparativo efectuado entre 9 selectores de características empleando el área bajo la curva ROC proporcionada al insertarlas a un clasificador logístico bayesiano.

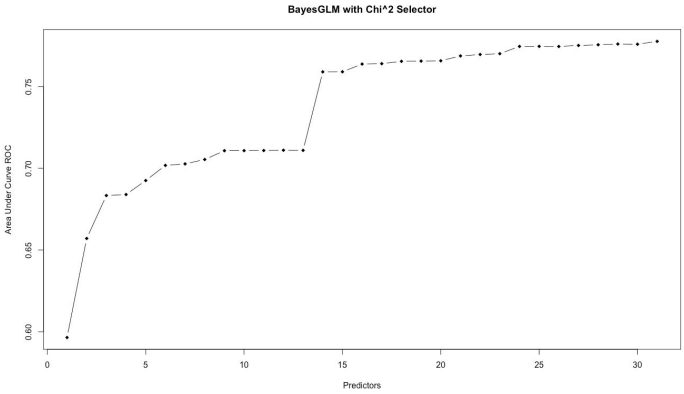



**Figura 6.27:** AUC ROC devuelta al introducir cada predictor en el orden de importancia devuelto por el selector $Chi^2$ en un clasificador de tipo BayesGLM

| Orden predictor según $Chi^2$ | Predictor | $Chi^2$ Score | Cummulative AUC ROC |
|---|---|---|---|
| 1 | Casualty_Type | 0,26472508 | 0,5965219 |
| 2 | Engine_Capacity_.CC. | 0,19339291 | 0,6570863 |
| 3 | Vehicle_Type | 0,19273227 | 0,6833308 |
| 4 | Casualty_Class | 0,16611698 | 0,6833308 |
| 5 | Number_of_Vehicles | 0,16425898 | 0,6924821 |
| 6 | Vehicle_Manoeuvre | 0,13737229 | 0,7017479 |
| 7 | First_Point_of_Impact | 0,12226018 | 0,7026689 |
| 8 | Junction_Detail | 0,09003535 | 0,7053512 |
| 9 | Junction_Location | 0,08920530 | 0,7107355 |
| 10 | Sex_of_Casualty | 0,08824004 | 0,7107727 |
| 11 | Sex_of_Driver | 0,08797660 | 0,7108414 |
| 12 | Hit_Object_off_Carriageway | 0,08382866 | 0,7110261 |
| 13 | Car_Passenger | 0,07904361 | 0,7109383 |
| 14 | Age_Band_of_Casualty | 0,07869042 | 0,7589404 |
| 15 | Skidding_and_Overturning | 0,07222237 | 0,7589836 |
| 16 | Skidding_and_Overturning | 0,07061335 | 0,7637074 |
| 17 | Hour_of_Day | 0,05769409 | 0,7639382 |
| 18 | Urban_or_Rural_Area | 0,05759697 | 0,7653891 |
| 19 | Age_of_Vehicle | 0,05621477 | 0,7655199 |
| 20 | Road_Type | 0,05202013 | 0,7657007 |
| 21 | Light_Conditions | 0,05022106 | 0,7686724 |
| 22 | other_vehic | 0,05002786 | 0,7695583 |
| 23 | Propulsion_Code | 0,04857275 | 0,7700448 |
| 24 | Age_Band_of_Driver | 0,04722078 | 0,7744793 |
| 25 | Driver_Home_Area_Type | 0,03446762 | 0,7745276 |
| 26 | Weather_Conditions | 0,03286249 | 0,7744394 |
| 27 | Journey_Purpose_of_Driver | 0,02764269 | 0,7750774 |
| 28 | Road_Surface_Conditions | 0,02614211 | 0,7755032 |
| 29 | First_Road_Class | 0,02565143 | 0,7759555 |
| 30 | Day_of_Week | 0,02376703 | 0,7758362 |
| 31 | month | 0,01563582 | 0,7775762 |

**Tabla 6.37:** AUC ROC devuelta al introducir cada predictor en el orden de importancia devuelto por el selector $Chi^2$ en un clasificador de tipo BayesGLM

| Método | AUC ROC | ReliefFequalK (attrEval) -R- | ReliefFequalK (permuteRelief) | Relief (attrEval) sin predictor Vehicle_Leaving_Carriageway -weka- | Relief (attrEval) con predictor Vehicle_Leaving_Carriageway -weka- | Backward BayesGLM | RFE - Random Forest | RFE - Random Forest (FSelector) | RFE-Chi Square (FSelector) |
|---|---|---|---|---|---|---|---|---|---|
| Función / Método | | The method evaluates the quality of the features/attributes/dependent variables specified by the formula with the selected heuristic method (ReliefFequalK) | This function uses a permutation approach to deteemining the relative magnitude of Relief scores (Kira and Rendell, 1992 and Kononenko, 1994) | | | | | random.forest.importance() | |
| Servidor | helmantica.umh.es | helmantica.umh.es | helmantica.umh.es | | | helmantica | 10.1.57.29 | | |
| Tiempo ejecución (seg.) | | | | | | 529498.816 | 426435.091 | 5003.842 | |
| Clasificación de características (Orden descendente) | Casualty_Type | Hour_of_Day | First_Point_of_Impact | First_Point_of_Impact | Hour_of_Day | Number_of_Vehicles | Hit_Object_off_Carriageway | Engine_Capacity_.CC. | Casualty_Type |
| | Number_of_Vehicles | Age_Band_of_Driver | Journey_Purpose_of_Driver | make | First_Point_of_Impact | Vehicle_Type | Vehicle_Manoeuvre | Speed_limit | Engine_Capacity_.CC. |
| | Vehicle_Type | First_Point_of_Impact | Age_Band_of_Driver | Age_Band_of_Driver | Age_Band_of_Driver | Vehicle_Manoeuvre | Age_Band_of_Casualty | Hour_of_Day | Vehicle_Type |
| | Engine_Capacity_.CC. | make | Age_Band_of_Casualty | Vehicle_Manoeuvre | make | Engine_Capacity_.CC. | Hour_of_Day | Hit_Object_off_Carriageway | Casualty_Class |
| | Sex_of_Casualty | Speed_limit | Hour_of_Day | Junction_Location | Vehicle_Manoeuvre | Sex_of_Casualty | First_Point_of_Impact | Skidding_and_Overturning | Number_of_Vehicles |
| | Junction_Location | Age_Band_of_Casualty | month | Age_Band_of_Casualty | Age_Band_of_Casualty | Sex_of_Driver | Speed_limit | Age_Band_of_Driver | Vehicle_Manoeuvre |
| | Sex_of_Driver | Age_of_Vehicle | Junction_Detail | Junction_Detail | Junction_Location | Junction_Detail | First_Road_Class | Light_Conditions | First_Point_of_Impact |
| | Junction_Detail | month | Speed_limit | First_Road_Class | Junction_Detail | First_Point_of_Impact | Age_Band_of_Driver | First_Point_of_Impact | Junction_Detail |

**Figura 6.28:** Resumen de la comparativa efectuada entre selectores de características (parte 1)

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Car_Passenger | Vehicle_Manoeuvre | First_Road_Class | Journey_Purpose_of_Driver | First_Road_Class | Propulsion_Code | Skidding_and_Overturning | Urban_or_Rural_Area | Junction_Location |
| | First_Point_of_Impact | First_Road_Class | Junction_Location | Road_Type | Journey_Purpose_of_Driver | Junction_Location | Engine_Capacity_.CC. | Number_of_Vehicles | Sex_of_Casualty |
| | Propulsion_Code | Propulsion_Code | Vehicle_Manoeuvre | Propulsion_Code | Speed_limit | other_vehic | Light_Conditions | First_Road_Class | Sex_of_Driver |
| | other_vehic | Junction_Location | Propulsion_Code | Road_Surface_Conditions | Road_Type | Car_Passenger | Age_of_Vehicle | Age_of_Vehicle | Hit_Object_off_Carriageway |
| | Weather_Conditions | Urban_or_Rural_Area | Vehicle_Type | Light_Conditions | Propulsion_Code | Weather_Conditions | Urban_or_Rural_Area | Road_Type | Car_Passenger |
| | Road_Surface_Conditions | Journey_Purpose_of_Driver | Sex_of_Driver | other_vehic | Road_Surface_Conditions | Road_Surface_Conditions | Number_of_Vehicles | Junction_Detail | Age_Band_of_Casualty |
| | Age_Band_of_Driver | Hit_Object_off_Carriageway | Road_Surface_Conditions | Hit_Object_off_Carriageway | Light_Conditions | month | Journey_Purpose_of_Driver | Age_Band_of_Casualty | Skidding_and_Overturning |
| | Day_of_Week | Skidding_and_Overturning | Skidding_and_Overturning | Day_of_Week | other_vehic | Hour_of_Day | Road_Type | Journey_Purpose_of_Driver | Skidding_and_Overturning |
| | month | Day_of_Week | Hit_Object_off_Carriageway | Weather_Conditions | Day_of_Week | First_Road_Class | Casualty_Type | Road_Surface_Conditions | Hour_of_Day |
| | Hour_of_Day | Vehicle_Type | Road_Type | month | Weather_Conditions | Age_Band_of_Casualty | Junction_Detail | Weather_Conditions | Urban_or_Rural_Area |
| | First_Road_Class | Road_Surface_Conditions | Casualty_Type | Skidding_and_Overturning | Casualty_Type | make | month | Driver_Home_Area_Type | Age_of_Vehicle |
| | make | Number_of_Vehicles | Age_of_Vehicle | Driver_Home_Area_Type | month | Journey_Purpose_of_Driver | Road_Surface_Conditions | month | Road_Type |
| | Journey_Purpose_of_Driver | other_vehic | Day_of_Week | Sex_of_Driver | Vehicle_Type | Age_of_Vehicle | Weather_Conditions | Day_of_Week | Light_Conditions |
| | Age_of_Vehicle | Weather_Conditions | Urban_or_Rural_Area | Casualty_Type | Skidding_and_Overturning | Driver_Home_Area_Type | Junction_Location | Junction_Location | other_vehic |

**Figura 6.29:** Resumen de la comparativa efectuada entre selectores de características (parte 2)

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Driver_Home_Area_Type | Driver_Home_Area_Type | Light_Conditions | Urban_or_Rural_Area | Sex_of_Driver | Light_Conditions | Day_of_Week | Vehicle_Manoeuvre | Propulsion_Code |
| | Road_Type | Sex_of_Driver | Number_of_Vehicles | Vehicle_Type | Driver_Home_Area_Type | Road_Type | Driver_Home_Area_Type | other_vehic | Age_Band_of_Driver |
| | Light_Conditions | Engine_Capacity_.CC. | Driver_Home_Area_Type | Hour_of_Day | Hit_Object_off_Carriageway | Hit_Object_off_Carriageway | other_vehic | Propulsion_Code | Driver_Home_Area_Type |
| | Age_Band_of_Casualty | Light_Conditions | Car_Passenger | Speed_limit | Vehicle_Leaving_Carriageway | Speed_limit | Vehicle_Type | Casualty_Type | Weather_Conditions |
| | Hit_Object_off_Carriageway | Road_Type | make | Number_of_Vehicles | Urban_or_Rural_Area | Urban_or_Rural_Area | Casualty_Class | Vehicle_Type | Journey_Purpose_of_Driver |
| | Speed_limit | Sex_of_Casualty | other_vehic | Sex_of_Casualty | Number_of_Vehicles | Skidding_and_Overturning | Propulsion_Code | Casualty_Class | Road_Surface_Conditions |
| | Urban_or_Rural_Area | Casualty_Type | Weather_Conditions | Age_of_Vehicle | Sex_of_Casualty | Vehicle_Leaving_Carriageway | Sex_of_Driver | Sex_of_Driver | First_Road_Class |
| | Skidding_and_Overturning | Car_Passenger | Engine_Capacity_.CC. | Engine_Capacity_.CC. | Age_of_Vehicle | Casualty_Class | Sex_of_Casualty | Sex_of_Casualty | Day_of_Week |
| | Casualty_Class | Junction_Detail | Casualty_Class | Car_Passenger | Engine_Capacity_.CC. | Casualty_Type | make | Car_Passenger | month |
| | Vehicle_Manoeuvre | Casualty_Class | Sex_of_Casualty | Casualty_Class | Car_Passenger | | Car_Passenger | | |
| | | | | | Casualty_Class | | | | |

**Figura 6.30:** Resumen de la comparativa efectuada entre selectores de características (parte 3)

## 6.8. Selección y ajuste del modelo

En la subsección 6.7.4 se muestra una comparativa realizada con 9 selectores de características que se adaptaban a los tipos de datos manejados para Reino Unido. Mediante esta selección, se han realizado una serie de tests con el fin de estimar el mejor de ellos que llevara a eliminar el *overfitting* provocado por las ejecuciones de BayesGLM.

El área bajo la curva ROC máxima lograda ha sido de $0,7775762$, mediante una ejecución de BayesGLM recursivo insertando 31 características en un orden determinado y mediante el uso de un selector $Chi^2$. Sin embargo, el algoritmo *Random Forest* ha alcanzado un área mayor a través de una ejecución recursiva con las siguientes 29 variables:

```
Vehicle_Type, Vehicle_Manoeuvre, Junction_Location, Skidding_and_Overturning,
Hit_Object_off_Carriageway, First_Point_of_Impact, Journey_Purpose_of_Driver,
Sex_of_Driver, Age_Band_of_Driver, Engine_Capacity_.CC., Propulsion_Code,
Age_of_Vehicle, Driver_Home_Area_Type, make, Casualty_Class, Sex_of_Casualty,
Age_Band_of_Casualty, Car_Passenger, Casualty_Type, Number_of_Vehicles,
Day_of_Week, Hour_of_Day, First_Road_Class, Road_Type, Speed_limit,
Junction_Detail, Light_Conditions, Weather_Conditions
```

donde se ha alcanzado un AUC ROC de $0,7935291$. Todos estos datos se pueden observar en la tabla 6.38.

| Num. Vars. | Modelo | Vars. eliminadas | Selector Caract. | AUC ROC | Tiempo ejec (seg.) | K-Fold |
|---|---|---|---|---|---|---|
| 29 | BayesGLM | Vehicle_Manoeuvre, Casualty_Class, Skidding_and_Overturning | Curva ROC | 0,7722236 | 13312,6 | 5 |
| 30 | BayesGLM | Junction_Detail, Casualty_Class | ReliefF | 0,775763 | 14600,408 | 5 |
| 24 | BayesGLM | Hour_of_Day, Speed_limit, Number_of_Vehicles, Sex_of_Casualty, Age_of_Vehicle, Engine_Capacity_.CC., Car_Passenger, Casualty_Class | ReliefF | 0,7734127 | 14921,018 | 5 |
| 31 | BayesGLM | - | $Chi^2$ | 0,7775762 | - | 5 |
| 29 | RandomForest | Road_Surface_Conditions, Urban_or_Rural_Area, month, other_vehic | RFE-RF | 0,7935291 | 255967,486 | 10 |
| 32 | BayesGLM | | RFE-RF | 0,7760994 | - | 10 |

**Tabla 6.38:** Tabla resumen

### 6.8.1. Ajuste del modelo

#### 6.8.1.1. Parámetros de ajuste del modelo

Tras elegir el modelo de 29 características basado en Random Forest, se extraen los parámetros de ajuste plasmados en la tabla 6.39.

Como se puede comprobar en dicha tabla, a pesar de obtener un buen AUC ROC, la sensibilidad y la especificidad no se encontraban niveladas. Se procede, por tanto,

| Threshold | Sens | Spec | Dist | $m_{try}$ |
|---|---|---|---|---|
| 0,010 | 0,98 | 0,11 | 0,88 | 12 |
| 0,255 | 0,40 | 0,90 | 0,60 | 12 |

**Tabla 6.39:** Tabla de parámetros de ajuste del modelo basado en Random Forest

a buscar un equilibrio tanto de los parámetros que deben ajustarse, como los que necesitan ser escalados o adaptados de forma más general para el conjunto de datos. Por ello, a partir de los resultados extraídos del test realizado con Random Forest como algoritmo escogido, se ha pasado a afinar más en el ajuste de la sensibilidad y especificidad para alcanzar el objetivo deseado, tratando de nivelar ambos parámetros otorgando prioridad a la sensibilidad por tratarse del parámetro que ajusta el número de accidentes graves o fallecimientos clasificados correctamente.

Para ello se ha lanzado el mismo test de nuevo para buscar en más puntos de corte entre 0,010 y 0,255. Los resultados se pueden observar en la tabla 6.40.

En ella se puede observar que a través de $threshold = 0,06444444$ la sensibilidad y la especificidad quedan niveladas a un 70 y 71 % respectivamente, correspondientes a la probabilidad de acierto de accidentado grave y leve, respectivamente.

| Threshold | Sens | Spec | Dist |
|---|---|---|---|
| 0,01000000 | 0,9719999 | 0,2095759 | 0,7909235 |
| 0,03722222 | 0,8461201 | 0,5453665 | 0,4799863 |
| 0,06444444 | 0,7017202 | 0,7123543 | 0,4143936 |
| 0,09166667 | 0,5723649 | 0,8098608 | 0,4680111 |
| 0,11888889 | 0,4549976 | 0,8757702 | 0,5589887 |
| 0,14611111 | 0,3643323 | 0,9167644 | 0,6410982 |
| 0,17333333 | 0,2988519 | 0,9414961 | 0,7035864 |
| 0,20055556 | 0,2456889 | 0,9583975 | 0,7554585 |
| 0,22777778 | 0,2063170 | 0,9687939 | 0,7942970 |
| 0,25500000 | 0,1705525 | 0,9768180 | 0,8297719 |

**Tabla 6.40:** Tabla de parámetros de ajuste del modelo basado en Random Forest

#### 6.8.1.2. Matriz de confusión y Curva ROC

Mediante la curva ROC (figura 6.31), se ha hallado un nuevo punto de corte óptimo que maximiza ligeramente la sensibilidad a costa de reducir, ligeramente también, la especificidad. De esta forma será posible obtener mejores predicciones de accidentados graves o fallecidos respecto al anterior punto de corte de $0,06444444$.

Por tanto, se escogerá el *threshold* de $0,059$. La tabla 6.41 muestra la matriz de confusión obtenida mediante este umbral.

## 6.9. Interpretación del modelo

Se propone el uso de un clasificador supervisado mixto basado en Random Forest y regresión logística bayesiana (BayesGLM). Tanto para la selección global de características así como para la predicción de la severidad del accidentado, en caso de producirse

| Predicción \ Referencia | X0 | X1 |
|---|---|---|
| X0 | 8004 | 20719 |
| X1 | 3371 | 52831 |

**Tabla 6.41:** Matriz de confusión del modelo final mediante Random Forest con dataset de test para $umbral = 0,059$

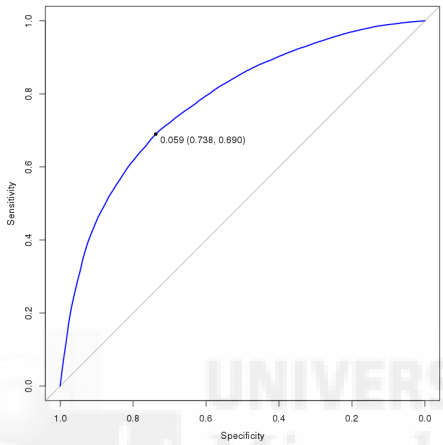



**Figura 6.31:** Área encerrada bajo la curva ROC del modelo obtenido

un accidente, vendrá dada por el primero de ellos. A través de este modelo no se produce *overfitting*, debido a la robustez del algoritmo.

Sin embargo, el modelo Random Forest no permite la interpretación de resultados con el fin de estimar qué niveles de características es necesario vigilar por producir un aumento de la gravedad en el accidente. Por ello, se propone combinarlo con un clasificador lineal bayesiano que permite interpretar los resultados de forma sencilla, así como de obtener distintas probabilidades de gravedad de cada nivel.

En la tabla 6.42 se muestran distintas probabilidades estimadas de la gravedad del accidente para las personas que ocupan el vehículo. Es importante señalar que estas probabilidades serán orientativas, con motivo de interpretar la importancia relativa de cada predictor seleccionado por Random Forest.

Cabe señalar que el modelo obtenido mediante BayesGLM ha descartado 5 características por no considerarlas de importancia y, adicionalmente, se han agrupado aquellos niveles indicados por el modelo como no significativos. Los predictores eliminados son:

```
make, Sex_of_Casualty, Car_Passenger, Number_of_Vehicles, Day_of_Week
```

| $N^o$ | Nivel de factor | Descripción | Estimate | Std. Error | z value | Pr(>\|z\|) | Prob. grave | Incr. grave |
|---|---|---|---|---|---|---|---|---|
| | INTERCEPT | | 4,079e+00 | 7,671e-02 | 53,173 | <2e-16 *** | 1,66 % | 0 % |
| 1 | Vehicle_TypeX2 | Motocicleta 50cc e inferior | 4,870e-01 | 1,801e-01 | 2,704 | 0,006844 ** | 1,03 % | -38,16 % |
| 2 | Vehicle_TypeX21 | Mercancías 7.5 ton. y superior | -8,171e-01 | 1,004e-01 | -8,139 | 4,00e-16 *** | 3,69 % | 121,73 % |
| 3 | Vehicle_TypeX3 | Motocicleta 125cc e inferior | 6,420e-01 | 9,715e-02 | 6,608 | 3,90e-11 *** | 0,88 % | -46,96 % |
| 4 | Vehicle_TypeX4 | Motocicleta entre 125cc y 500cc | 5,443e-01 | 1,731e-01 | 3,144 | 0,001667 ** | 0,97 % | -41,57 % |
| 5 | Vehicle_TypeX5 | Motocicleta 500cc | 4,473e-01 | 9,430e-02 | 4,744 | 2,10e-06 *** | 1,07 % | -35,68 % |
| 6 | Vehicle_TypeX9 | Coche | 2,382e-01 | 2,638e-02 | 9,031 | <2e-16 *** | 1,32 % | -20,92 % |
| 7 | Vehicle_ManoeuvreX3 | Esperando para salir | 7,878e-01 | 6,479e-02 | 12,159 | <2e-16 *** | 0,76 % | -54,10 % |
| 8 | Vehicle_ManoeuvreX4 | Disminuyendo velocidad o parando | 4,782e-01 | 4,035e-02 | 11,852 | <2e-16 *** | 1,04 % | -37,62 % |
| 9 | Vehicle_ManoeuvreX5 | Moviéndose | 3,107e-01 | 4,443e-02 | 6,993 | 2,69e-12 *** | 1,23 % | -26,38 % |
| 10 | Vehicle_ManoeuvreX7 | Girando a la izquierda | 1,561e-01 | 4,661e-02 | 3,348 | 0,000814 *** | 1,43 % | -14,25 % |
| 11 | Vehicle_ManoeuvreX8 | Esperando para girar a la izquierda | 8,427e-01 | 1,788e-01 | 4,714 | 2,43e-06 *** | 0,72 % | -56,53 % |
| 12 | Vehicle_ManoeuvreX9 | Girando a la derecha | 1,157e-01 | 3,593e-02 | 3,219 | 0,001286 ** | 1,49 % | -10,76 % |
| 13 | Vehicle_ManoeuvreX10 | Esperando para girar a la derecha | 5,003e-01 | 9,267e-02 | 5,399 | 6,71e-08 *** | 1,02 % | -38,97 % |
| 14 | Vehicle_ManoeuvreX11 | Cambiando al carril izquierdo | -1,542e-01 | 7,650e-02 | -2,016 | 0,043840 * | 1,94 % | 16,35 % |
| 15 | Vehicle_ManoeuvreX12 | Cambiando al carril derecho | 1,925e-01 | 8,405e-02 | 2,290 | 0,022005 * | 1,38 % | -17,27 % |
| 16 | Vehicle_ManoeuvreX13 | Adelantando a un vehículo en movimiento por el carril izquierdo | -1,061e-01 | 4,210e-02 | -2,521 | 0,011697 * | 1,85 % | 10,99 % |
| 17 | Vehicle_ManoeuvreX15 | Adelantando por el carril izquierdo | 3,085e-01 | 7,815e-02 | 3,947 | 7,90e-05 *** | 1,23 % | -26,22 % |
| 18 | Vehicle_ManoeuvreX16 | Avanzando a mano izquierda | -3,035e-01 | 3,442e-02 | -8,818 | <2e-16 *** | 2,24 % | 34,66 % |
| 19 | Vehicle_ManoeuvreX17 | Avanzando a mano izquierda | -1,657e-01 | 3,412e-02 | -4,855 | 1,20e-06 *** | 1,96 % | 17,67 % |
| 20 | Vehicle_ManoeuvreX18 | Avanzando (hacia otra dirección) | -1,630e-01 | 2,813e-02 | -5,793 | 6,91e-09 *** | 1,95 % | 17,36 % |
| 21 | Junction_LocationX1 | Aproximándose hacia un cruce o esperando/aparcado en la cercanía de un cruce | -6,279e-02 | 1,810e-02 | -3,470 | 0,000521 *** | 1,77 % | 6,37 % |
| 22 | Junction_LocationX2 | Cruce despejado o esperando/aparcado en la salida de un cruce | -1,479e-01 | 2,520e-02 | -5,868 | 4,42e-09 *** | 1,92 % | 15,63 % |
| 23 | Junction_LocationX4 | Incorporándose a una rotonda | -1,499e-01 | 5,382e-02 | -2,785 | 0,005351 ** | 1,93 % | 15,86 % |
| 24 | Junction_LocationX6 | Incorporándose a la vía principal | -1,863e-01 | 3,522e-02 | -5,291 | 1,22e-07 *** | 2,00 % | 20,07 % |
| 25 | other_vehicX11 | Autobús o autocar (17 o más asientos) | -9,410e-01 | 1,189e-01 | -7,915 | 2,48e-15 *** | 4,16 % | 149,76 % |
| 26 | other_vehicX17 | Vehículo Agrícola | -1,810e+00 | 5,326e-01 | -3,398 | 0,000679 *** | 9,37 % | 463,15 % |
| 27 | other_vehicX19 | Furgoneta / Camión de mercancías igual o por debajo de 3.5t. | -5,804e-01 | 6,277e-02 | -9,247 | <2e-16 *** | 2,94 % | 76,37 % |
| 28 | other_vehicX20 | Mercancías por encima de 3.5t. y por debajo de 7.5t | -9,258e-01 | 1,910e-01 | -4,848 | 1,25e-06 *** | 4,10 % | 146,15 % |
| 29 | other_vehicX21 | Mercancías hasta/-por encima de 7.5 toneladas | -1,478e+00 | 1,069e-01 | -13,824 | <2e-16 *** | 6,91 % | 315,04 % |
| 30 | other_vehicX4 | Motocicleta por encima de 125cc hasta 500cc | 7,465e-01 | 3,499e-01 | 2,133 | 0,032905 * | 0,80 % | -52,18 % |
| 31 | other_vehicX8 | Taxi o alquiler de coche privado | -3,639e-01 | 1,107e-01 | -3,287 | 0,001014 ** | 2,38 % | 42,85 % |
| 32 | other_vehicX9 | Coche | -3,366e-01 | 1,924e-02 | -17,491 | <2e-16 *** | 2,31 % | 39,09 % |

Sigue en la página siguiente.

| $N^o$ | Nivel de factor | Descripción | Estimate | Std. Error | z value | Pr(>\|z\|) | Prob. grave | Incr. grave |
|---|---|---|---|---|---|---|---|---|
| 33 | other_vehicX90 | Otro tipo de vehículo | -8,333e-01 | 3,152e-01 | -2,644 | 0,008199 ** | 3,75 % | 125,21 % |
| 34 | Hit_Object_off_-CarriagewayX2 | Señal de tráfico | -6,107e-01 | 4,155e-02 | -14,698 | <2e-16 *** | 3,02 % | 81,63 % |
| 35 | Hit_Object_off_-CarriagewayX1 | Farola | -4,267e-01 | 4,053e-02 | -10,528 | <2e-16 *** | 2,53 % | 51,87 % |
| 36 | Hit_Object_off_-CarriagewayX3 | Poste de electricidad o de telégrafos | -4,998e-01 | 6,055e-02 | -8,255 | <2e-16 *** | 2,71 % | 63,08 % |
| 37 | Hit_Object_off_-CarriagewayX4 | Árbol | -1,023e+00 | 2,612e-02 | -39,155 | <2e-16 *** | 4,50 % | 170,14 % |
| 38 | Hit_Object_off_-CarriagewayX7 | Quitamiedos | -2,839e-01 | 4,491e-02 | -6,323 | 2,56e-10 *** | 2,20 % | -32,11 % |
| 39 | Hit_Object_off_-CarriagewayX8 | Sumergido en agua | -1,339e+00 | 2,394e-01 | -5,594 | 2,22e-08 *** | 6,07 % | 264,45 % |
| 40 | Hit_Object_off_-CarriagewayX9 | Introducido en una zanja | -2,010e-01 | 3,859e-02 | -5,209 | 1,90e-07 *** | 2,03 % | 21,81 % |
| 41 | First_Point_of_-ImpactX1 | Parte delantera | -1,311e-01 | 1,196e-02 | -10,958 | <2e-16 *** | 1,89 % | 13,74 % |
| 42 | First_Point_of_-ImpactX2 | Parte trasera | 6,674e-01 | 2,860e-02 | 23,334 | <2e-16 *** | 0,86 % | -48,28 % |
| 43 | Number_of_-VehiclesX2 | Número de vehículos | 9,528e-02 | 1,677e-02 | 5,683 | 1,32e-08 *** | 1,52 % | -8,95 % |
| 44 | Skidding_and_-OverturningX1 | Derrapó | -1,983e-01 | 1,627e-02 | -12,190 | <2e-16 *** | 2,02 % | 21,49 % |
| 45 | Skidding_and_-OverturningX2 | Derrapó y volcó | -2,150e-01 | 2,574e-02 | -8,353 | <2e-16 *** | 2,06 % | 23,49 % |
| 46 | Skidding_and_-OverturningX5 | Volcó | -2,989e-01 | 2,924e-02 | -10,223 | <2e-16 *** | 2,23 % | 34,06 % |
| 47 | Journey_Purpose_of_DriverX2 | Yendo o volviendo del trabajo | -1,144e-01 | 2,432e-02 | -4,703 | 2,56e-06 *** | 1,86 % | 11,89 % |
| 48 | Journey_Purpose_of_DriverX5 | Otros | -4,013e-01 | 4,355e-02 | -9,214 | <2e-16 *** | 2,47 % | 48,16 % |
| 49 | Journey_Purpose_of_DriverX6 | Desconocido | -1,876e-01 | 1,944e-02 | -9,650 | <2e-16 *** | 2,00 % | 20,22 % |
| 50 | Journey_Purpose_of_DriverX15 | Otros/Desconocido (2005-10) | -2,313e-01 | 2,079e-02 | -11,124 | <2e-16 *** | 2,09 % | 25,48 % |
| 51 | Sex_of_DriverX2 | Mujer | 1,793e-01 | 1,338e-02 | 13,400 | <2e-16 *** | 1,39 % | -16,19 % |
| 52 | Age_Band_of_-DriverX11 | Por encima de 75 | -8,682e-02 | 4,208e-02 | -2,063 | 0,039111 * | 1,81 % | 8,91 % |
| 53 | Age_Band_of_-DriverX4 | 16 - 20 | -2,407e-01 | 3,873e-02 | -6,215 | 5,13e-10 *** | 2,11 % | 26,64 % |
| 54 | Age_Band_of_-DriverX5 | 21 - 25 | -2,922e-01 | 3,680e-02 | -7,941 | 2,01e-15 *** | 2,22 % | 33,18 % |
| 55 | Age_Band_of_-DriverX6 | 26 - 35 | -1,274e-01 | 3,367e-02 | -3,784 | 0,000154 *** | 1,89 % | 13,33 % |
| 56 | Age_Band_of_-DriverX7 | 36 - 45 | -1,187e-01 | 3,448e-02 | -3,442 | 0,000578 *** | 1,87 % | 12,37 % |
| 57 | Age_Band_of_-DriverX8 | 46 - 55 | -1,053e-01 | 3,497e-02 | -3,011 | 0,002605 ** | 1,85 % | 10,90 % |
| 58 | Age_Band_of_-DriverX9 | 56 - 65 | -6,652e-02 | 3,658e-02 | -1,818 | 0,069015 . | 1,78 % | 6,76 % |
| 59 | Engine_Capacity_.CC, | Motorización en CC | -3,176e-05 | 6,227e-06 | -5,101 | 3,38e-07 *** | 1,66 % | 0,00 % |
| 60 | Propulsion_CodeX2 | Heavy oil | 9,467e-02 | 1,535e-02 | 6,166 | 7,01e-10 *** | 1,52 % | -8,90 % |
| 61 | Propulsion_CodeX8 | Híbrido o eléctrico | 3,953e-01 | 1,066e-01 | 3,709 | 0,000208 *** | 1,13 % | -8,90 % |
| 62 | Age_of_Vehicle | Edad en años | -9,794e-03 | 1,106e-03 | -8,858 | <2e-16 *** | 1,68 % | 0,97 % |
| 63 | Driver_Home_-Area_TypeX2 | Pueblo pequeño | -4,319e-02 | 1,837e-02 | -2,351 | 0,018698 * | 1,74 % | 4,34 % |
| 64 | Driver_Home_-Area_TypeX3 | Rural | -6,122e-02 | 1,692e-02 | -3,619 | 0,000296 *** | 1,77 % | 6,20 % |
| 65 | Road_Surface_-ConditionsX2 | Húmedo | 2,653e-02 | 1,553e-02 | 1,708 | 0,087706 . | 1,62 % | -2,58 % |
| 66 | Road_Surface_-ConditionsX3 | Nieve | 6,664e-01 | 6,870e-02 | 9,700 | <2e-16 *** | 0,86 % | -48,23 % |
| 67 | Road_Surface_-ConditionsX4 | Hielo | 4,467e-01 | 3,713e-02 | 12,031 | <2e-16 *** | 1,07 % | -35,64 % |
| 68 | Road_Surface_-ConditionsX5 | Inundado por encima de 3cm | 3,545e-01 | 1,249e-01 | 2,837 | 0,004552 ** | 1,17 % | -29,50 % |
| 69 | Casualty_ClassX3 | Peatón | -1,919e+00 | 2,218e-02 | -86,513 | <2e-16 *** | 10,34 % | 521,30 % |

Sigue en la página siguiente.

| $N^o$ | Nivel de factor | Descripción | Estimate | Std. Error | z value | Pr(>\|z\|) | Prob. grave | Incr. grave |
|---|---|---|---|---|---|---|---|---|
| 70 | Age_Band_of_-CasualtyX4 | 16 - 20 | 1,455e-01 | 2,587e-02 | 5,623 | 1,88e-08 *** | 1,44 % | -13,35 % |
| 71 | Age_Band_of_-CasualtyX5 | 21 - 25 | 1,521e-01 | 2,530e-02 | 6,011 | 1,85e-09 *** | 1,43 % | -13,91 % |
| 72 | Age_Band_of_-CasualtyX7 | 36 - 45 | -9,495e-02 | 2,338e-02 | -4,062 | 4,87e-05 *** | 1,83 % | 9,78 % |
| 73 | Age_Band_of_-CasualtyX8 | 46 - 55 | -2,516e-01 | 2,481e-02 | -10,143 | <2e-16 *** | 2,13 % | 28,00 % |
| 74 | Age_Band_of_-CasualtyX9 | 56 - 65 | -4,565e-01 | 2,808e-02 | -16,261 | <2e-16 *** | 2,60 % | 56,35 % |
| 75 | Age_Band_of_-CasualtyX10 | 66 - 75 | -7,725e-01 | 3,174e-02 | -24,337 | <2e-16 *** | 3,53 % | 112,40 % |
| 76 | Age_Band_of_-CasualtyX11 | Por encima de 75 | -1,138e+00 | 3,139e-02 | -36,240 | <2e-16 *** | 5,02 % | 201,41 % |
| 77 | Casualty_TypeX10 | Minibús (8 - 16 asientos) | 3,566e-01 | 1,684e-01 | 2,118 | 0,034162 * | 1,17 % | -29,64 % |
| 78 | Casualty_TypeX11 | Autobús o autocar (17 o más asientos) | 2,322e-01 | 5,891e-02 | 3,942 | 8,09e-05 *** | 1,32 % | -20,45 % |
| 79 | Casualty_TypeX2 | Motocicleta igual o inferior a 50cc conductor o pasajero | -1,803e+00 | 1,821e-01 | -9,900 | <2e-16 *** | 9,31 % | 459,59 % |
| 80 | Casualty_TypeX21 | Vehículo de mercancías (7.5 tonnes mgw and over) occupant | 8,220e-01 | 1,137e-01 | 7,230 | 4,84e-13 *** | 0,74 % | -55,63 % |
| 81 | Casualty_TypeX3 | Motocicleta 125cc e inferiores, conductor o pasajero | -2,138e+00 | 9,580e-02 | -22,316 | <2e-16 *** | 12,55 % | 654,31 % |
| 82 | Casualty_TypeX4 | Motocicleta de más de 125cc e inferior a 500cc conductor o pasajero | -2,153e+00 | 1,744e-01 | -12,347 | <2e-16 *** | 12,72 % | 664,26 % |
| 83 | Casualty_TypeX5 | Motocicleta hasta 500cc conductor o pasajero | -2,286e+00 | 9,220e-02 | -24,796 | <2e-16 *** | 14,27 % | 757,46 % |
| 84 | Hour_of_Day | $UTC$ | 1,133e-02 | 1,020e-03 | 11,107 | <2e-16 *** | 1,65 % | -1,11 % |
| 85 | First_Road_-ClassX3 | A | -3,681e-01 | 3,287e-02 | -11,198 | <2e-16 *** | 2,39 % | 43,44 % |
| 86 | First_Road_-ClassX4 | B | -3,464e-01 | 3,673e-02 | -9,430 | <2e-16 *** | 2,34 % | 40,43 % |
| 87 | First_Road_-ClassX5 | C | -2,117e-01 | 3,815e-02 | -5,551 | 2,84e-08 *** | 2,05 % | 23,09 % |
| 88 | First_Road_-ClassX6 | Sin clasificar | -2,790e-01 | 3,613e-02 | -7,723 | 1,14e-14 *** | 2,19 % | 31,48 % |
| 89 | Road_TypeX3 | 2 carriles | -1,278e-01 | 3,155e-02 | -4,050 | 5,13e-05 *** | 1,89 % | 13,38 % |
| 90 | Road_TypeX6 | 1 carril | -1,984e-01 | 2,865e-02 | -6,926 | 4,34e-12 *** | 2,02 % | 21,50 % |
| 91 | Speed_Limit | mph | -1,408e-02 | 6,225e-04 | -22,619 | <2e-16 *** | 1,69 % | 1,39 % |
| 92 | Junction_DetailX1 | Rotonda | 4,578e-01 | 3,446e-02 | 13,284 | <2e-16 *** | 1,06 % | -36,34 % |
| 93 | Junction_DetailX2 | Mini-rotonda | 3,727e-01 | 6,992e-02 | 5,330 | 9,82e-08 *** | 1,15 % | -30,75 % |
| 94 | Junction_DetailX3 | Cruce en T or staggered junction | 2,329e-01 | 1,796e-02 | 12,969 | <2e-16 *** | 1,32 % | -20,50 % |
| 95 | Junction_DetailX5 | Slip road | 2,657e-01 | 5,069e-02 | 5,241 | 1,60e-07 *** | 1,28 % | -23,03 % |
| 96 | Junction_DetailX6 | Crossroads | 2,787e-01 | 2,273e-02 | 12,262 | <2e-16 *** | 1,26 % | -24,02 % |
| 97 | Junction_DetailX7 | More than 4 arms (not roundabout) | 2,406e-01 | 5,444e-02 | 4,419 | 9,91e-06 *** | 1,31 % | -21,10 % |
| 98 | Junction_DetailX8 | Private drive or entrance | 2,179e-01 | 3,477e-02 | 6,266 | 3,69e-10 *** | 1,34 % | -19,32 % |
| 99 | Junction_DetailX9 | Other junction | 2,649e-01 | 4,339e-02 | 6,105 | 1,03e-09 *** | 1,28 % | -22,97 % |
| 100 | Light_ConditionsX4 | Oscuridad - Alumbrado encendido | -3,645e-01 | 1,485e-02 | -24,554 | <2e-16 *** | 2,38 % | 42,93 % |
| 101 | Light_ConditionsX5 | Oscuridad - Alumbrado apagado | -2,309e-01 | 7,518e-02 | -3,071 | 0,002134 ** | 2,09 % | 25,43 % |
| 102 | Light_ConditionsX6 | Oscuridad - Sin alumbrado - no lighting | -3,203e-01 | 2,054e-02 | -15,592 | <2e-16 *** | 2,28 % | 36,89 % |
| 103 | Light_ConditionsX7 | Oscuridad - se desconoce alumbrado | -2,515e-01 | 5,348e-02 | -4,701 | 2,58e-06 *** | 2,13 % | 27,99 % |

Sigue en la página siguiente.

| $N^o$ | Nivel de factor | Descripción | Estimate | Std. Error | z value | Pr(>\|z\|) | Prob. grave | Incr. grave |
|---|---|---|---|---|---|---|---|---|
| 104 | Weather_ConditionsX2 | Lluvia sin viento fuerte | 1,690e-01 | 2,104e-02 | 8,033 | 9,52e-16 *** | 1,41 % | -15,33 % |
| 105 | Weather_ConditionsX5 | Lluvia con viento fuerte | 1,467e-01 | 4,582e-02 | 3,201 | 0,001369 ** | 1,44 % | -13,45 % |
| 106 | Weather_ConditionsX7 | Niebla o bruma | 1,562e-01 | 6,845e-02 | 2,282 | 0,022470 * | 1,43 % | -14,25 % |
| 107 | Weather_ConditionsX8 | Otras | 1,821e-01 | 4,150e-02 | 4,388 | 1,14e-05 *** | 1,39 % | -16,42 % |
| 108 | Weather_ConditionsX9 | Desconocida | 1,906e-01 | 6,658e-02 | 2,863 | 0,004201 ** | 1,38 % | -17,11 % |

**Tabla 6.42:** Descripción de los predictores en el modelo estimado

En base a nuestro modelo estimado, la probabilidad inicial de que el accidente sea grave o provoque un fallecimiento sin que influya ninguna variable es:

$$Prob = \frac{1}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k}} = \frac{1}{1 + e^{4,079}} = 0,01664271 \qquad (6.1)$$

Es decir, la probabilidad inicial de que el accidente sea grave es del 1,66 %, siendo cero o no existiendo el resto de variables y niveles de los factores.

Por otro lado, la probabilidad de que el accidente sea leve es de:

$$Prob = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k}} = \frac{e^{3,64}}{1 + e^{3,64}} = 0,9833573 \qquad (6.2)$$

Es decir, la probabilidad inicial de que el índice sea leve, sin tener en cuenta el resto de variables, es del 98,34 %.

Como revela la tabla 6.42, aquellos accidentes que se producen cuando se impacta contra un vehículo agrícola (línea 26 de la tabla), hacen aumentar considerablemente el riesgo de que el pronóstico de los accidentados sea grave o incluso fallezcan, cifra similar al impactar contra un camión de mercancías de 7.5 toneladas o superior (línea 29).

Es posible entrar más en detalle acerca de la distribución de accidentes que posee este tipo de camiones (nivel X21) en función del tipo de vía por la que los sufre. La figura 6.32 muestra dicha distribución como una nube de puntos correspondientes a accidentes. En naranja se representan los accidentes leves, mientras que en azul se representan los accidentes graves. De izquierda a derecha, la primera nube corresponde con el nivel de factor X1, hasta llegar a X6 (ver filas 85 a 88 de la tabla 6.42 para una descripción de cada tipo de vía) en la última posición a la derecha. Como se puede observar en la figura, los accidentes graves de este tipo de vehículos se producen mayoritariamente en las vías de tipo 'A' y, a continuación, en las vías de tipo 'C'.

La figura 6.33 muestra la distribución de accidentes en forma de nube en función del mismo tipo de vehículo, camión de mercancías de 7.5 toneladas o superior (nivel de

factor X21), y de la hora en la que ocurren. En naranja se representan los accidentes leves, mientras que en azul se representan los accidentes graves. Como se puede observar en la figura, los accidentes graves mayoritarios de este tipo de vehículos se producen entre las 8 y las 11h, entre las 16 y las 17h y existe otro pico a las 20h. Quizás se pueda reducir la siniestralidad con este tipo de vehículos abriendo o cerrando carriles alternativos para este tipo de vehículos en esas vías y a esas horas.

**Figura 6.32:** Nube de accidentes tipo de coche vs. clase de vía

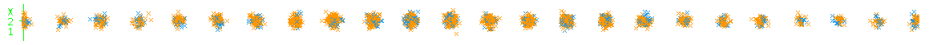


**Figura 6.33:** Nube de accidentes tipo de coche vs. hora del día

En la misma tabla, se puede observar que un accidente contra un autobús o autocar, o contra camiones de entre 3.5 y más de 7.5 toneladas también hacen aumentar el riesgo de que los accidentados acaben gravemente heridos o fallezcan.

Impactar contra una motocicleta de entre 125 y 500 c.c. provoca un aumento de la probabilidad de que los accidentados del vehículo que impacta contra la motocicleta acaben con un pronóstico leve.

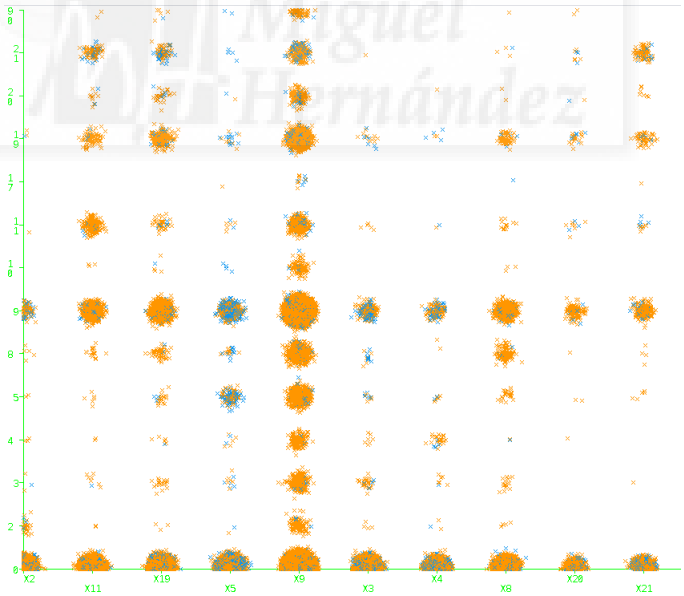



**Figura 6.34:** Nube de accidentes referente a tipo de vehículo (eje X) frente a otros vehículos (eje Y)

En la figura 6.34 se representa la distribución de accidentes, en forma de nube, de la variable *Vehicle_Type*, en el eje X, frente a *other_vehic*, en el eje Y, es decir, muestra

la gravedad del accidente por ocupante cuando chocan ambos vehículos entre ellos. En dicha figura se puede observar que en su mayoría, cuando se produce un accidente de una motocicleta de 500cc. o superior (X5) frente a un coche (X9), el resultado es que el piloto de la motocicleta acaba falleciendo o gravemente herido.

Con los datos de la tabla 6.42 aplicados a estos niveles en la ecuación 6.1 se puede conocer además la probabilidad de ser grave o fallecer.

Es necesario cuidar los objetos que se encuentran en los márgenes de las vías. Por ejemplo, el choque contra un árbol (fila 37 de la tabla 6.42) o contra una señal de tráfico (fila 34 de la misma tabla) provoca un aumento de la gravedad de las heridas de hasta un 170 % respecto a la probabilidad inicial de accidente grave. Resulta curioso que el impacto contra un árbol posea un índice similar de gravedad a la de aquellos vehículos que puedan acabar sumergidos en el agua.

La edad del herido también influye en la probabilidad de fallecimiento o gravedad del ocupante. Aquellos accidentados por encima de 56 años (filas de 74 a 76 tabla 6.42) poseen una probabilidad mayor de riesgo de fallecimiento o resultar gravemente herido, frente a una persona de entre 21 o 25 años o inferior (filas 70 y 71).

Un dato curioso extraído del estudio es que aquellos conductores a bordo de una motocicleta, independientemente de la cilindrada, poseen un riesgo muy similar a los de un coche, reduciendo la probabilidad de gravedad del accidente. Se puede concluir, por tanto, que el tipo de vehículo no siempre es un factor fundamental para que el usuario fallezca o resulte herido leve, salvo para los ocupantes de un vehículo de mercancías de 7.5 toneladas o superior, donde la probabilidad de aumento de la gravedad del accidentado es del 121 % respecto a la probabilidad inicial de gravedad.

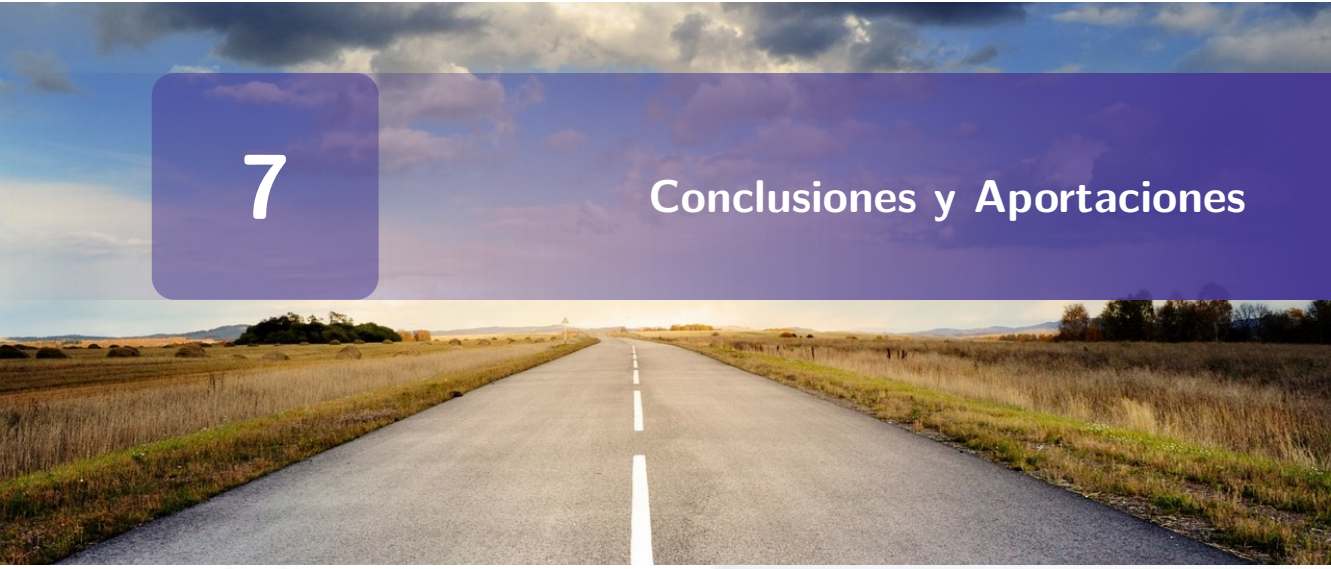
De la misma tabla se deduce que el impacto en la parte trasera del vehículo (fila 42 de la tabla 6.42) reduce notablemente la probabilidad de fallecimiento o heridas graves, frente a un impacto frontal (fila 41 de la tabla) , que hace crecer, en menor medida que las hace descender el anterior, las probabilidades de sufrir un accidente grave.

## 6.10. Conclusiones y aportaciones

- Se ha estudiado un conjunto de datos formado por un gran número de accidentes registrados entre 2009 y 2014 en Reino Unido donde se predice la severidad del accidente respecto a las víctimas. Se emplean para ello un gran número de variables que describen las causas que lo rodearon, entre ellas, las relativas a infraestructura vial, condiciones climáticas y otras.
- Se ha descrito el set de datos completo durante el periodo entre 2009 y 2014, realizando un estudio de cada variable con la finalidad de reducir el conjunto de datos y emplear el dataset resultante en el estudio probabilístico.

- Se ha creado, por tanto, un nuevo dataset a partir del que posee publicado el Departamento para el Transporte de Reino Unido, resultante de fusionar las tablas relativas a accidentes, accidentados y vehículos, para formar una única tabla que explicara, de forma unificada, todas las características que envolvían el accidente.
- Se ha estudiado la frecuencia de ocurrencia de accidentes en base a cada una de las características.
- Se aporta una nueva variable al dataset correspondiente al tipo de vehículo contra el cual se colisionó y la severidad de los accidentados, así como un estudio minucioso de los valores atípicos.
- Se ha realizado un estudio de los niveles de los factores de características con pocas muestras y un análisis de variables que poseen varianza casi cero con el fin de eliminarlas.
- Se aporta un estudio acerca de la existencia de correlación de variables y su motivación para conservarlas o eliminarlas del set de datos.
- Se aporta un completo estudio mediante distintas metodologías acerca de la selección de predictores para el estudio de accidentes de tráfico, focalizando en aquellos ocurridos en Reino Unido.
- Se aporta una comparativa entre los selectores estudiados y una preselección de aquellos con mejores resultados obtenidos en base al área encerrada bajo la curva ROC al introducir en orden las características en un clasificador logístico bayesiano.
- Se garantiza la independencia de los sets de datos para entrenamiento y test a través del uso de la técnica de validación cruzada o K-fold.
- Como una de las aportaciones principales de esta tesis, se aporta un modelo para el estudio de accidentes de tráfico en Reino Unido basado en Random Forest de 29 características seleccionadas con una capacidad de predicción demostrada en base a la sensibilidad y especificidad obtenida con el.
- Se aporta una interpretación del modelo de clasificación seleccionado a través de un estudio de los predictores mediante un algoritmo logístico lineal bayesiano.
- Se aporta un estudio de la influencia probabilística de cada predictor en cuanto a la ocurrencia de un accidente con severidad de los ocupantes grave o incluso que pueda acabar en fallecimiento.
- Se presenta una agrupación de predictores sobre los que cabe prestar atención con la finalidad de aplicar una serie de acciones para reducir el número de fallecimientos o accidentes graves.

UNIVERSITAS
Miguel
Hernández



# 7 Conclusiones y Aportaciones

## 7.1. Conclusiones generales

La mayoría de soluciones en materia de seguridad vial están encaminadas en la resolución de los problemas relacionados con el triángulo básico *infraestructura-conductor-vehículo* donde se incide, mayoritariamente, en el cumplimiento de las medidas reglamentarias adoptadas y que se consideran actuaciones suficientes para la reducción de los accidentes de tráfico por parte de los organismos públicos con competencias en tráfico.

Sin embargo, los Cuerpos y Fuerzas de Seguridad del Estado invierten una gran cantidad de minucioso trabajo en la recolección de todas las circunstancias que envuelven a un accidente de tráfico. Su estudio, no obstante, no siempre acaba en la puesta en marcha de medidas efectivas para salvaguardar la seguridad de los ocupantes de los vehículos.

Por otro lado, el planteamiento del empleo de medidas paliativas relativas a la seguridad basadas en soluciones de inteligencia artificial y relacionadas con inferir conocimiento y trasladarlo a la infraestructura o a los automóviles, hasta ahora, no se ha explorado de forma amplia. No obstante, los Ministerios y Departamentos encargados de la seguridad vial, poseen todos los datos recopilados acerca de las causas de los accidentes de tráfico, por lo que estudios como el que se presenta en esta tesis doctoral pueden ser de gran ayuda para aplicarlos para salvar vidas.

En este trabajo se proponen una serie de métodos encaminados a guiar al usuario en su comportamiento al volante a través de la dotación de inteligencia vial colectiva a dispositivos como el teléfono móvil, el ordenador de abordo del vehículo o el sistema de

navegación del mismo, a través de la actuación en tiempo real en base al análisis previo de situaciones colectivas pasadas y mediante sistemas que sean capaces de aprender de forma autónoma para predecir nuevas posibles situaciones futuras.

Para ello, se ha trabajado inicialmente en dos vías fundamentales:

1. El aprendizaje relacionado con la seguridad vial y los accidentes de tráfico.
2. Poner en cuestión la metodología actual y plantear métodos complementarios en materia de predicción de accidentes de tráfico basados en técnicas de inteligencia artificial.

Se ha propuesto para ello un sistema de predicción de la severidad del accidente de tráfico así como de las lesiones de los propios ocupantes en caso de sufrir una colisión. Para ello se ha inferido conocimiento al sistema a través de información publicada por las agencias de seguridad vial y de los gobiernos, principalmente de España y Reino Unido, empleando dichos países como modelos de aplicación del sistema objeto de este trabajo de investigación.

En él se demuestra científicamente que a partir de una serie de sets de datos el sistema es capaz de alimentarse de esta información y encontrar tendencias en ellos estadísticamente demostrables y probables, a través de escoger correctamente la información y clasificarla mediante técnicas basadas en la supervisión entre algoritmos de regresión o de clasificación.

Se ha hecho frente en este trabajo a distintas problemáticas que surgen de cómo se encuentran almacenados los datos y de su propia naturaleza, y que motivan realizar aportaciones en esta materia a través del uso de algoritmos robustos pero a la vez interpretables con el fin de entender la información soslayada entre los datos.

Se ha conseguido, por tanto, partiendo de casi un centenar de variables de datos validar modelos con apenas el uso de una decena de estas características en países como España, o de una veintena en países como Reino Unido, donde los tipos de datos a los que nos hemos enfrentado poseían una distribución diferente debido entre otros factores a su forma de conducción y a sus propias normas de circulación.

## 7.2. Aportaciones

En esta sección se enumeran las principales aportaciones realizadas separadas por cada uno de los capítulos.

### 7.2.1. Aportaciones y conclusiones específicas de la estudio de la teoría de clasificadores

#### Conclusiones

El propósito no es siempre conocer y comprender el por qué van a ocurrir los eventos. En algunas ocasiones, el propósito viene dado por la precisión con la que se adivinará el evento, y no por las circunstancias que lo produjeron. Se estudian por tanto algoritmos robustos, interpretables y precisos en la predicción que puedan ayudar en el trascurso del trabajo de investigación a la aplicación de medidas en base a sus resultados.

Se han definido distintas metodologías de selección de características, incidiendo en el estudio a través de la literatura de las consecuencias del uso de predictores que no aportan información y de cómo identificarlos para evitar añadir incertidumbre en la predicción.

Se ha definido el concepto de datos desbalanceados y se ha estudiado la problemática que podría ocasionar, y se han estudiado diferentes soluciones aportadas por otros investigadores en distintas materias. Algunas de ellas han sido la búsqueda de estrategias para muestreo de datos, o estrategias para generación de nuevos ejemplos, y sus consecuencias.

Se han estudiado las bases para la validación de los modelos estadísticos que se propondrán en esta tesis. Partiendo de las diferentes formas de validar dichos modelos estadísticos, se ha escogido la validación mediante el área encerrada bajo la curva ROC por considerar que se adecúa en mayor medida al objetivo final basado en la clasificación/predicción entre clases dicotómicas pudiendo balancear una respecto a la otra para predecir qué ocurrirá con los ocupantes de un vehículo en caso de accidente.

Se plantea el uso de un clasificador de tipo Random Forest como selector de características, gracias a la posibilidad que nos ofrece para ello, y como algoritmo de clasificación, complementándolo con un algoritmo de regresión logística bayesiano por la facilidad que éste ofrece en su interpretación en base a la información escogida por el primero.

En el ajuste del modelo, dada la naturaleza de los datos, evitar el sobreentrenamiento del mismo era un factor clave necesario de resolver. Por ello, como se ha estudiado en la sección 2.4.2.3, se ha escogido también Random Forest debido a que incorpora distintos mecanismos para evitar caer en el sobreentrenamiento u *overfitting*, puesto que, bien entrenado, es capaz de reducir los ratios de error al elegir árboles independientes y fuertes escogidos con independencia de los árboles seleccionados anteriormente, por lo que Random Forest es robusto frente a respuestas ruidosas.

#### Aportaciones

En este capítulo se presentan las siguientes aportaciones:

- Se describe la teoría de los eventos raros y cómo aplicarla al estudio de accidentes de tráfico.

- Se ha estudiado la literatura referente a predictores que no aportan información con el fin de evitar añadir incertidumbre en la predicción.
- Se definen y exponen distintas metodologías de selección de características primando la robustez y simplicidad en su interpretación.
- Se ha definido el concepto de sets de datos desbalanceados y se ha estudiado la problemática que podría ocasionar el uso de este tipo de datos.
- Se han estudiado diferentes técnicas aportadas por otros autores para evitar la problemática de este tipo de sets de datos:
  - estrategias para muestreo de datos
  - estrategias para generación de nuevos ejemplos
- Se ha descrito el uso de dos técnicas de clasificación complementarias para clasificación de accidentes de tráfico, partiendo del objetivo principal de esta tesis doctoral de adivinar el evento y tratar de evitar las circunstancias que lo produjeron.
- Se presenta el algoritmo Random Forest como uno de los algoritmos principales a la hora de trabajar con la clasificación respecto a la severidad de accidentes de tráfico, debido a que incorpora distintos mecanismos para selección de características y evitar caer en el sobreentrenamiento u *overfitting*.
- Adicionalmente, se plantea complementar Random Forest con un clasificador de regresión logísitica bayesiano, como es BayesGLM.
- Se presenta mediante el Área Bajo la Curva ROC la validación de los modelos estadísticos empleados para la clasificación y predicción de accidentes de tráfico en base a la severidad del accidente o de los accidentados.

### 7.2.2. Conclusiones específicas y aportaciones en base a la revisión del estado del arte

En el capítulo 3 se detallan estadísticas globales relativas a accidentes de tráfico con el objetivo de definir el problema de su ocurrencia en Reino Unido y Gran Bretaña. Se indican datos macroscópicos distinguiendo tipos de accidente en función de las vías, etc. y se han interpretado las gráficas aportadas por el Departamento para Transporte de Reino Unido.

Adicionalmente, se detallan estadísticas globales relativas a accidentes de tráfico ocurridos en España, centrando los datos de forma macroscópica y focalizando, asimismo, sobre su ocurrencia en un año escogido aleatoriamente. Se han interpretado las gráficas aportadas por la Dirección General de Tráfico en algunos de sus informes anuales.

A continuación, se realiza una revisión exhaustiva sobre el estado del arte de las características y técnicas empleadas para la clasificación y predicción de accidentes. En

concreto se revisan trabajos realizados por otros autores empleando modelos de regresión logística, modelos de regresión de Poisson y clasificadores basados en árboles J48 o modelos bayesianos. Por último, se hace referencia a diversos trabajos realizados por otros autores mediante el uso del algoritmo Random Forest en sus investigaciones.

### 7.2.3. Conclusiones específicas y aportaciones en base al diseño e implementación de una infraestructura basada en Big Data para el almacenamiento y análisis de incidencias de tráfico en tiempo real

Con lo dicho hasta ahora, y según las pruebas realizadas que se mostrarán en el apartado 4.9 que Cassandra y CQL son unas herramientas más que adecuadas para tratar con grandes cantidades de datos crecientes en el tiempo.

No obstante y como comentario final, debemos entender que esto se consigue a costa de limitar en gran medida las búsquedas que se pueden realizar sobre los datos y, muchas veces, teniendo que duplicar tablas para poder hacer búsquedas por varias columnas al mismo tiempo.

### 7.2.4. Aportaciones y conclusiones específicas en el estudio de accidentes ocurridos en España entre 2011 y 2015

#### Conclusiones

Se ha presentado un estudio estadístico de los accidentes de tráfico que se han producido en España entre 2011 y 2015 extraídos del Portal Estadístico de la Dirección General de Tráfico.

Se han estudiado distintas metodologías estadísticas para encontrar un modelo que sea capaz de clasificar estos accidentes de tráfico en base a su severidad o a la severidad de las lesiones de los ocupantes del vehículo

Se ha aportado un conjunto de datos de 5.157 registros de accidentes entre mayo de 2015 a enero de 2016, donde cada tabla incluye los datos de la vía y la ubicación geográfica del accidente, los informes meteorológicos y los datos relativos al radar de control de velocidad más cercano o su situación dentro de un tramo en obras.

Adicionalmente, se han analizado 17.573 incidentes en el mismo periodo en que se encontró un tramo en obras en la red de carreteras española y 1.247 radares de control de velocidad localizados por toda la red.

Es relevante observar que la siniestralidad en la provincia de Sevilla y más específicamente en la autovía de circunvalación SE-30, de tan solo 22 km de longitud total, pero con la tasa de accidentes registrados mas elevada de España. La Autovía del Nordeste o A-2, que une Madrid con Barcelona, y la del Mediterráneo o A-7, le siguen con un 17 % y con un 29 % menos de accidentes, respectivamente y con una longitud total de 780 km y de 1.329 km.

En cuanto al número de accidentes en función del día de la semana, se observa que los sábados y los domingos presentan el índice más bajo de accidentes de tráfico siendo el domingo significativamente bajo. El pico de accidentes en viernes es otra de los datos significativos a estudiar.

Se ha analizado el número de accidentes producidos en la cercanía de un radar de control de velocidad donde un 5,35 % del total de accidentes (276 accidentes) ocurrieron en el periodo de mayo 2015 a enero 2016, considerando un radio de 500 m al radar más cercano y un 2,87 % (un total de 148) rebajando el radio a 200m. Estos resultados sugieren que tal vez algunas medidas de control de velocidad no están funcionando tal y como se desea, y resultaría conveniente realizar un estudio de su ubicación o de su verdadero objetivo.

Se valida el uso de validación cruzada como método para la generación de los conjuntos de datos de entrenamiento, test y en el ajuste del propio modelo.

Se han estudiado diferentes selectores de características en base a la naturaleza de los sets de datos validando el uso de Random Forest y BayesGLM como soporte a la estimación individual de selección de predictores para su inserción progresiva al modelo de clasificación basado a su vez en los mismos clasificadores.

Se ha realizado un estudio exhaustivo del ajuste del modelo de clasificación a base de la estimación de la importancia de cada predictor.

Se valida el uso de un clasificador supervisado mixto basado en Random Forest y regresión logística bayesiana (BayesGLM).

### Aportaciones

Se propone una única tabla de datos para el estudio de los accidentes de tráfico en España, donde se generan nuevas características creadas a partir de la información extraída de otras de ellas.

Se proponen una nuevas variables en el estudio de los accidentes de tráfico para obtener su interacción con otras medidas de seguridad impuestas, como distancia a los radares de control de velocidad o detección de accidentes en tramos en obras.

Se propone el uso de técnicas basadas en k-fold para el muestreo de accidentes de tráfico donde se posee una clase minoritaria y una mayoritaria con una proporción 20/80 lo que provoca cierto desbalanceo entre ellas y la posibilidad de ocurrencia de overfitting si el muestreo fuera incorrecto.

A pesar de que el selector de características empleado ha sido Random Forest basado en RFE, entre otros analizados que aportaban peores resultados, éste no aporta la posibilidad de la interpretación de los resultados debido a que el propio algoritmo se basa en diversos tipos de árboles para generar el bosque adecuado. Por ello, se ha decidido en base a esta aportacion del algortimo realizar una interpretación de los resultado a partir de un método lineal ensamblado con técnicas bayesianas que permiten

trabajar con una distancia entre clases mejor seleccionada que con métodos lineales puros. Dicho método ha sido BayesGLM. Aunamos de esta forma el poder clasificatorio de random forest junto con la facilidad de interpretación que nos aporta BayesGLM.

En base a esta aportación principal del uso de RF y BayesGLM se extraen las conclusiones que se detallarán a continuación:

- Se presenta un modelo de predicción probado con el set de datos de prueba arrojando los siguientes resultados:
  - Predicción de accidentes graves o fallecimientos: $82\,\%$.
  - Predicción de accidentes leves: $72\,\%$.
- Se estudia la influencia de cada característica del vector escogido para generar el modelo aportando unas conclusiones sobre ellas que se deben de considerar a la hora de proponer medidas de seguridad vial, así como de dotar de inteligencia artificial a la tecnología empleada por los usuarios en forma de automóvil o de infraestructura vial.

### 7.2.5. Aportaciones y conclusiones específicas en el estudio de accidentes ocurridos en Reino Unido entre 2009 y 2014

Se ha estudiado un conjunto de datos formado por un gran número de accidentes registrados entre 2009 y 2014 en Reino Unido donde se predice la severidad del accidente respecto a las víctimas. Se emplean para ello un gran número de variables que describen las causas que lo rodearon, entre ellas, las relativas a infraestructura vial, condiciones climáticas y otras.

Se ha descrito el set de datos completo durante el periodo entre 2009 y 2014, realizando un estudio de cada variable con la finalidad de reducir el conjunto de datos y emplear el dataset resultante en el estudio probabilístico.

Se ha creado, por tanto, un nuevo dataset a partir del que posee publicado el Departamento para el Transporte de Reino Unido, resultante de fusionar las tablas relativas a accidentes, accidentados y vehículos, para formar una única tabla que explicara, de forma unificada, todas las características que envolvían el accidente.

Se ha estudiado la frecuencia de ocurrencia de accidentes en base a cada una de las características.

Se aporta una nueva variable al dataset correspondiente al tipo de vehículo contra el cual se colisionó y la severidad de los accidentados, así como un estudio minucioso de los valores atípicos.

Se ha realizado un estudio de los niveles de los factores de características con pocas muestras y un análisis de variables que poseen varianza casi cero con el fin de eliminarlas.

Se aporta un estudio acerca de la existencia de correlación de variables y su motivación para conservarlas o eliminarlas del set de datos.

Se aporta un completo estudio mediante distintas metodologías acerca de la selección de predictores para el estudio de accidentes de tráfico, focalizando en aquellos ocurridos en Reino Unido.

Se aporta una comparativa entre los selectores estudiados y una preselección de aquellos con mejores resultados obtenidos en base al área encerrada bajo la curva ROC al introducir en orden las características en un clasificador logístico bayesiano.

Se garantiza la independencia de los sets de datos para entrenamiento y test a través del uso de la técnica de validación cruzada o K-fold.

Como una de las aportaciones principales de esta tesis, se aporta un modelo para el estudio de accidentes de tráfico en Reino Unido basado en bosque, de 29 características seleccionadas, con una capacidad de predicción demostrada en base a la sensibilidad y especificidad obtenida con él.

Se aporta una interpretación del modelo de clasificación seleccionado a través de un estudio de los predictores mediante un algoritmo logístico lineal bayesiano.

Se aporta un estudio de la influencia probabilística de cada predictor en cuanto a la ocurrencia de un accidente con severidad de los ocupantes grave o incluso que pueda acabar en fallecimiento.

## 7.3. Líneas futuras de investigación

El estudio estadístico de la influencia de determinadas características en la ocurrencia de un accidente, o del por qué éste fue leve o grave para los pasajeros de un vehículo, aunque es una línea de investigación apasionante y creemos que necesaria para continuar mejorando la seguridad vial y salvar más vidas, resulta imposible abarcar en este trabajo de investigación toda la amplitud que conlleva su estudio. Sin embargo, esta tesis doctoral aporta una serie de modelos basados en árboles de decisión y de regresión logística-bayesiana, que pueden abrir paso a una serie de mejoras en la predicción de la severidad del accidente respecto a los ocupantes de un vehículo en caso de que éste ocurriera. Aun así, creemos necesario continuar en este línea y para ello se proponen las siguientes investigaciones:

- Búsqueda de un modelo mixto donde predictores como el año de ocurrencia del accidente sean efectos aleatorios en lugar de efectos fijos. Con ello se cree que se reduciría algún sesgo, mejorando los pronósticos tanto en sensibilidad como en especificidad.
- Aplicación de los modelos presentados en esta investigación para la aplicación sobre tramos geoposicionados en determinadas vías.

- Búsqueda de nuevos modelos de regresión y clasificación cuya aplicabilidad sea única indistintamente del país donde se emplee.
- Buscar un nuevo modelo de clasificación a partir de los datos obtenidos del modelo probabilístico. Con este nuevo modelo Ad-hoc se podría afinar algo más en la obtención de la sensibilidad y especificidad en la curva ROC.
- Convendría buscar nuevos selectores de características que permitan reducir el coste computacional sin penalizar en la bondad de su aplicación a los modelos de regresión o clasificación.

Sería importante cuantificar el no-accidente en las vías de Reino Unido o España con motivo de poder calcular las probabilidades de que se produzca un accidente. En la actualidad, ya hemos abordado esta línea de investigación a través del acceso a los datos de las estaciones de aforamiento instaladas en distintas vías principales en España. Convendría ampliar estar red de sensores a otra vías de la Red de Carreteras del Estado, con la finalidad de poseer datos adicionales de paso de vehículos en otras vías donde se producen numerosos accidentes.

UNIVERSITAS
Miguel
Hernández

## Apéndice: Conjunto de Publicaciones

La mayoría de implementaciones y contribuciones realizadas en esta tesis doctoral están respaldadas por un conjunto de publicaciones en revistas. Los siguientes artículos apoyan el trabajo plasmado en este documento:

Artículo Revista 1

*A study of traffic accidents in Spanish intercity roads by means of feature vectors*.
D. Úbeda, A. Gil, L. Payá, O. Reinoso
**International Journal of Design & Nature and Ecodynamics**. Vol. 11 (3). 2016
ISSN: 1755-7437. DOI: 10.2495/DNE-V11-N3-317-327. Ed. WIT Press.

Artículo Revista 2

*Analyzing Traffic Accident Severity in UK from a Classification point of view*.
D. Úbeda, A. Gil, A. Pérez
**Enviada a Journal of Advanced Transportation**
**JCR-SCI Impact Factor: 1.292**, Quartile **Q1**.
Ed. Hindawi - John Wiley & Sons

Las reproducciones de las publicaciones se adjuntan a continuación:

UNIVERSITAS
Miguel
Hernández

# A study of traffic accidents in Spanish intercity roads by means of feature vectors

D. Úbeda, A. Gil, L. Payá, O. Reinoso
*Department of Systems Engineering and Automation,*
*University Miguel Hernández de Elche, Spain*

## Abstract

Frequently, road traffic accidents are modelled as a discrete and independent random and rare events, which possess a low probability of occurrence through time. Nevertheless, in order to study each accident separately it is necessary to obtain a number of characteristics that surround it, which may be correlated with each other. In this paper we propose to associate the probability of occurrence of an accident with a high number of features such as weather conditions, incidents caused by the start and end of a roadwork, geographical location of speed control radars, roadway infrastructure, etc. The influence of these features is significant and should be taken into account when proposing measures to help alleviate this type of undesirable events. The big data methods employed to extract the variables or features allow us to compose a series of vectors that will serve as a basis to study road accident distributions on them.

*Keywords: road traffic accident, weather features vectors, road traffic data mining*

## 1 Introduction

Road accidents are the cause of a high number of human losses in the world and also imply a significant psychological and physical trauma. In addition, according to the European Union [5], the economic annual burden of road accidents is estimated in Europe between EUR 10 and 14 billion and the cost of a single fatality may be as high as EUR 1 million.

In this paper we consider road accidents as rare events, that is, their frequency of occurrence in time is low. In order to be able to study them, we propose a multivariable system whose many characteristics have been obtained and analysed by big data techniques. Big data methods are employed to extract the

variables or features that allow us to compose a feature vector associated to each accident. These vectors are formed by a set of variables that were involved in the accident. These variables include weather conditions, traffic sensors and more and serve as a basis to study road accident distributions on them.

### 1.1 Studying accidents as rare but massive data events

The treatment of rare events, that is, which occurs with a low probability, is a complex problem. The rare event law or Poisson process consists in an stochastic process defined in a continuous time and is substantiated on counting rare events that happen during time. The time between each pair of consecutive events possesses an exponential probability distribution with a $\lambda$ parameter, and differences between different pairs of consecutive events are considered independent.

Recently these kind of problems based on Poisson processes have been tackled by means of data analysis and data modelling techniques derived from the artificial intelligence field. These events are treated as massive data (although infrequent) due to the high quantity of variables involved in each one of them. One could imagine, for example, that the probability of occurrence of an accident is greater in a rainy day than in a sunny day. In addition it could be reasonable to think that the risk of suffering an accident is greater during rush hours compared with the situation when only few vehicles are circulating. As a result, there is a need to store a high quantity of data associated to each accident in order to be able to study them later and try to find important relations between them. In consequence, big data techniques need to be applied to the problem, since the problem requires storing a great quantity of information associated to each event.

In this paper, in addition to the features mentioned above, other features have been analysed, whose influence in the accident has been deemed useful to include. The use of these new features is the main contribution of this paper, which, as far as the authors know, have not been employed before in the analysis of traffic accidents. One of these most relevant variables is the influence of the proximity of fixed speed limit radars in the driver's behaviour, or the circulation in a certain route of vehicles through a section under roadwork.

## 2 Feature selection and Open Data in Spain

### 2.1 Factors that influence in the probability of occurrence of road accidents

During the last years, road accidents have been analysed by different researchers with the objective to establish models that will allow finding answers to the causes that produced each accident, and, as well, solutions that may alleviate the frequency of these undesired events. Therefore, typically, authors have selected features that could be considered in the frequency or appearance of road incidents. In most cases have been used variables relating to weather, characteristics of the road, or driver conditions.

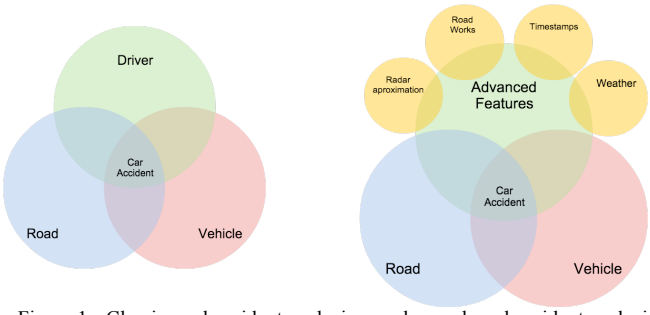



Figure 1: Classic road accident analysis vs. advanced road accident analysis

Driving under the influence of alcohol or drugs is unquestionably dangerous. We call it "*Classic Road Accident Analysis*", fig. 1. The most relevant features studied during the last years are listed in the next paragraphs.

### 2.1.1 Driver's behaviour features

Attending to the literature, the authors have typically associated some features to the accident. Sabey and Taylor [8] used car speed at the time of the accident, whereas, Hakim *et al.* [6] and Wang *et al.* [11] studied the influence of the speed limit of the road. The use of safety measures was a variable in Simoncic [10] research, among others. Driver classification was used in other studies as one of their main features, i.e. age, used by Davison [2], or experience, also studied by Simoncic. Over the years, these features have evolved and also other new features have come upon, such as the use of mobile phones or other electronic devices when driving.

### 2.1.2 Features related to the design of the roadway

To date, the characteristics related to the design of the road have been considered very important compared to other possible variables involved in an accident. This can be explained by the historical sought of the causes of accidents which did originate a true interest to improve the road infrastructure, thus reducing the accident rate.

In the current state of knowledge, it is possible to highlight some of the following features:

- Lane width and Average Daily Traffic expressed in veh./day (Zegeer *et al.* [13])
- Width of shoulder of the road (Kalokota *et al.* [7])
- Number of lanes (Wang *et al.* [11])
- Section length (Shankar *et al.* [9])

### 2.1.3 Weather patterns

Extensive research has correlated weather variables such as rain, fog or snow, as the cause of the occurrence of accidents, but always approached from the point of view of user *mobility*. It's important to identify weather situations that increase or reduce mobility. For instance, inclement weather reduces mobility, so it's important to consider temperature as an appropriate feature in mobility.

In addition, mobility is strongly affected by weather conditions and also influences the number of road accidents and casualties significantly. The causes of the accidents do also depend on the type of road (motorways, rural roads or urban roads). Moreover, as said by others, since the weather affects mobility, there is a direct dependence between their effects and the number of injury accidents and casualties.

A two-stage approach was adopted in order to take account for the climatic factor: first we seek variables that were open access and then find a correlation of these features with the occurrence of accidents.

## 2.2 Current situation of open data related to road traffic in Spain

### 2.2.1 Accidents datasets in Spain

Currently, the Spanish Traffic Department (DGT) publishes in its "open data historical statistical website (2008-2013)" [3] data related to accidents, roads, injured people, and, as well, information concerning each vehicle involved.

Table 1: DGT public features of an accident

| Field names | | |
|---|---|---|
| Accident ID | Priority | Type of intersection |
| Road | Lightly injured | Vehicles involved |
| City | Day of week | Section of road without intersection |
| Isle | Road type | Seriously injured |
| weather | luminosity | Road work |
| Province | sidewalk | Type of accident |
| Zone | Restricted visibility | Traffic Volume |
| year | Road network | Specific measures |
| month | Autonomous region | Road surface |
| hour | Total victims | Grouped zone |

Table 2: DGT public features of the vehicle involved in an accident

| Field names | |
|---|---|
| type | Vehicle registration year |
| state | Vehicle registration month |
| occupants | Dangerous goods |
| year | Burnt car |

Table 3: DGT public features of injured people

| Field names | | |
|---|---|---|
| Age | Gender | Driver's year license |
| Position | Security accessories | Adverse effects |
| year | manoeuvre | Speed infraction |
| Pedestrian action | Pedestrian infraction | |

As can be appreciated in tables 1, 2 and 3, the Traffic Department publishes numerous variables related to the accident, vehicle, and injured people. These features can help us understand some of the causes involved in the accident. However, we believe that there is a lot of important information missing, that will undoubtedly help researchers improve their studies. For instance, accurate geographical location of each accident is not provided. Lane information is also missing.

### 2.2.2 Creating a database suited for our specific needs

As we mentioned in the previous paragraph, we consider of uttermost importance to locate the accident in order to get other features derived from its situation. In order to solve this, we started to study alternatives to the Spanish statistical website.

Our approach is based on real time data indexing of traffic incidences from Infocar DGT website [4]. Data extraction is performed sequentially through real time web scraping of all registered incidences. After this, all of them are stored in a Cassandra [1] ten node cluster.

The data extracted from Infocar during last year has been used in this research. We additionally stored incidences caused by the start and end of a roadwork, geographical location of speed control radars, and of course, the data relating to a recent accident. In the table 4 is shown some of the general data we are currently indexing from the published incidences. As mentioned, they are stored on our cluster every five minutes.

Table 4: DGT public features of all incidences

| | |
|---|---|
| Cause | Type |
| Subtype | Incidence ID |
| City | Province |
| Initial km point | Final km point |
| Latitude | Longitude |
| Level (severity) | Description |
| Timestamps | Lane direction |

## 2.3 Feature extraction

After the preliminary research described before, we now define more complex features that can be derived from the previous simple features from table 4. As said before, these features are new and have not been employed before in traffic studies and allow us analyse the problem through a new advanced analysis model, fig. 1.

### 2.3.1 Case 1: Analysing accidents at section in roadwork

The originality of this research consists in analysing if an accident has happened between two points, initial and final, of a roadwork. For this analysis, all incidences indexed as ‘roadwork’ are checked with every accident indexed in Cassandra for the same road. We have coded some functions in Python to check this. Of course, timestamps from roadwork and accidents are also checked.

### 2.3.2 Case 2: Analysing accidents next to a speed control radar

Due to our personal experience dangerous situations occur in the proximity of radars due to the sharp slowdown in chain by multiple vehicles approaching it when driving with a higher speed than allowed.

To find out whether if the accident is close to a radar, the situation is more complex, because it is necessary to calculate the distance between two points on a sphere. In order to calculate this, first of all we check that the accident happened in the same road where the speed control radar is placed. Then, the distance between them is calculated through the Haversine formula, eqn. (1),

$$d = 2r \cdot \arcsin\left(\sqrt{\sin^2\left(\frac{\phi_2-\phi_1}{2}\right) + \cos(\phi_1)\cdot\cos(\phi_2)\cdot sin^2\left(\frac{\lambda_2-\lambda_1}{2}\right)}\right) \quad (1)$$

where $\phi_1$, $\phi_2$ and $\lambda_1$, $\lambda_2$ refer to the latitude and longitude, both expressed in radians, of the two points respectively, and r is equal to the mean radius of the Earth. Once we know the distance *d* between the two points, it is checked that this distance is less than 500 meters before and after the speed control radar.

### 2.3.3 Extracting weather features

Because we have no historical weather data, we obtain it from the Wunderground API [12]. In order to do this, we post the geographical position and occurrence timestamp of each event. The mentioned web service returns the data relative to the closest meteorological station, and we are able to add interesting data to our feature vector:

- Pressure in mBar
- Temperature in Celsius
- Relative humidity %
- Wind speed in km/h
- Wind direction description (i.e., SW, NNE)
- Visibility
- Precipitation in mm

## 2.4 Dataset analysed

The dataset used contains 5157 records of accidents from May 2015 to January 2016; each table includes data from the road and geographical location of the accident, weather reports, and data relative to the closest speed control radar or its situation inside a section in roadwork.

| Attributes | via | poblacion | provincia | dayOfWeek | dayOfMonth | hh_day | temp | w_speed | visibility | radar | obra |
|---|---|---|---|---|---|---|---|---|---|---|---|
| via | 1 | 0.384 | 0.204 | -0.036 | -0.001 | 0.011 | -0.001 | 0.008 | -0.003 | 0.107 | -0.100 |
| poblacion | 0.384 | 1 | 0.193 | 0.028 | -0.012 | -0.009 | -0.027 | -0.020 | 0.170 | -0.112 | 0.015 |
| provincia | 0.204 | 0.193 | 1 | -0.005 | 0.025 | -0.052 | -0.002 | -0.010 | -0.062 | 0.119 | -0.009 |
| dayOfWeek | -0.036 | 0.028 | -0.005 | 1 | -0.048 | -0.004 | -0.014 | -0.015 | 0.032 | -0.012 | 0.026 |
| dayOfMonth | -0.001 | -0.012 | 0.025 | -0.048 | 1 | 0.019 | 0.014 | -0.003 | -0.051 | 0.015 | -0.020 |
| hh_day | 0.011 | -0.009 | -0.052 | -0.004 | 0.019 | 1 | -0.004 | 0.009 | 0.040 | -0.016 | -0.005 |
| temp | -0.001 | -0.027 | -0.002 | -0.014 | 0.014 | -0.004 | 1 | 0.801 | -0.019 | -0.007 | 0.017 |
| w_speed | 0.008 | -0.020 | -0.010 | -0.015 | -0.003 | 0.009 | 0.801 | 1 | 0.004 | -0.009 | 0.008 |
| visibility | -0.003 | 0.170 | -0.062 | 0.032 | -0.051 | 0.040 | -0.019 | 0.004 | 1 | -0.078 | -0.054 |
| radar | 0.107 | -0.112 | 0.119 | -0.012 | 0.015 | -0.016 | -0.007 | -0.009 | -0.078 | 1 | -0.098 |
| obra | -0.100 | 0.015 | -0.009 | 0.026 | -0.020 | -0.005 | 0.017 | 0.008 | -0.054 | -0.098 | 1 |

Figure 2: Correlation Matrix

In addition, we have studied 17573 incidents in the same period where a roadwork section was found, and 1247 speed control radars located in the Spanish geography.

## 3 Results

### 3.1 Single variable correlations

Our first step in the data analysis is to discard those variables that are not useful in our study. In order to do this, we obtain the matrix correlation shown in fig. 2. It can be observed that a positive strong correlation exists between wind speed (w_speed) and temperature (temp) with a correlation index (r=0.384).

The rest of the features are not correlated, making it possible to conclude that they are independent of one another.

### 3.2 Distribution of the most relevant features

It has been considered interesting to show some of the distributions that have been obtained from our preliminary data analysis. This study has analysed the number of accidents per road extracted from the dataset of 5176 accidents. Next, we present absolute results that classify provinces according to the number of accidents, as well as day of the week and hour of the day.

#### 3.2.1 Accident distribution by road

If we come into detail on the distribution of accidents by road, fig. 3, it is relevant to observe that the SE-30 is the motorway with more accidents, and it is located in Seville, third Spanish province with higher rate of accidents. It is followed by A-2 motorway with a 17% less number of accidents, and the A-7 with a 29% less.

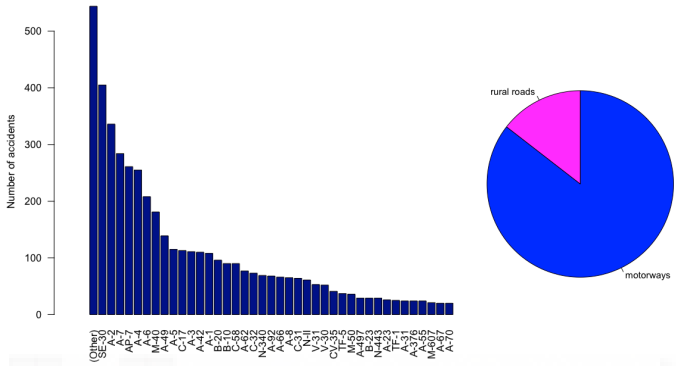



Figure 3: Accident distribution by road and by road type

### 3.2.2 Accident distribution by road type

The number of accidents detected by type of road has been much higher on motorways than on secondary roads, representing a 85% compared to a mere 15% respectively, fig. 3.

### 3.2.3 Accident distribution by province

When this research was started, it was expected to find a greater number of accidents in provinces with higher indexes of population. This has been proved in the provinces of Barcelona and Madrid both of them with the higher rate of accidents around the Spanish territory, fig. 4 and 5. However, we found very curious the volume of accidents of the Seville province compared to a similar extension province like Valencia, next in the ranking by accident records, but with a 50% less accidents than Seville.

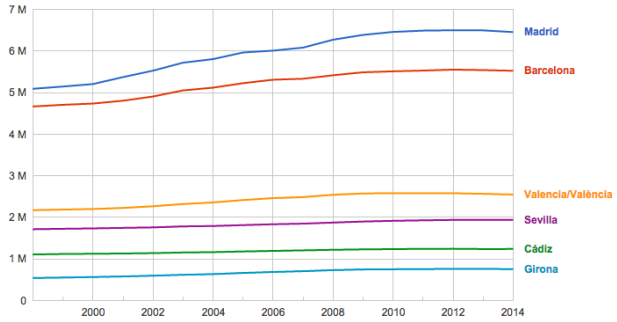



Figure 4: Population by province and year

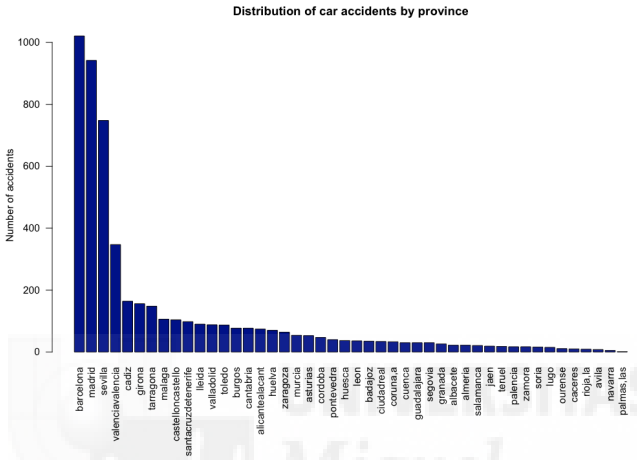



Figure 5: Distribution of car accidents by province

### 3.2.4 Accident distribution by day of week and by hour of the day

Fig. 6 represents the accumulated traffic accidents during the whole period studied ordered by the day of the week when they happened, and also represents the distribution of accidents during the hour of the day.

This allows us to define a total *a priori* probability of accident in any road in Spain given the day of the week. The main conclusion that we can observe is that on Saturday and Sundays the number of traffic accidents is significantly low. It is remarkable that the distribution is similar for the rest of the days but increases on Fridays.

Also, we observe three main peaks at 8:00-9:00h, 14:00-15:00h and 19:00-20:00h. This distribution can be easily explained due to the Spanish working hours that generally coincide with these rush hours.

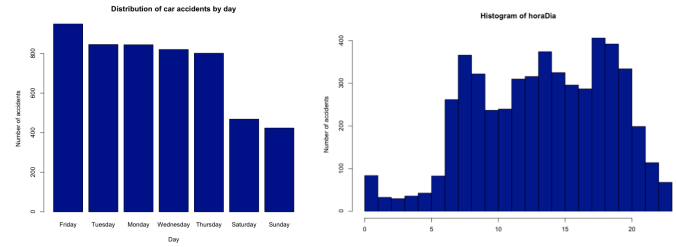



Figure 6: Distribution of car accidents by day of week and by hour of the day

### 3.2.5 Accident distribution by temperatures

Fig. 7 represents the total number of accidents in a histogram of temperatures. This graph presents a distribution that could seem like a Gaussian centered at 19 degrees Celsius. This preliminary result will be used as a starting point to further analyse the dependence of warm weather with the occurrence of accidents.

### 3.2.6 Number of accident produced near a speed control radar

In this Section we present the validity of the speed radar control cameras as a safety measure to reduce traffic accidents by means of adjusting the speed of drivers to the maximum established in the road. Our personal experience suggested that drivers tend to abruptly adjust their speed when suddenly observing these radars and this fact may produce risk situations.

In order to analyse these situations, two different distances have taken (200 and 500 meters) of the accidents to the nearest radar situated in the same road. The accidents that take place inside these radiuses are accounted for the statistics. The main conclusion of the study after processing the whole dataset is that approximately a 5,35% of the accidents took place inside a radius of 500 m of the nearest radar (a total of 276 accidents). When a radius of 200 m is employed a 2,87% of accidents occur in the nearing of the radar (a total of 148).

As a conclusion, the results suggest that maybe some speed control measures are not functioning as desired and it yields convenient to perform a study of their location or true objective.

### 3.2.7 Number of accidents observed in a section under roadwork

In this paper we also focus on the number of accidents that take place in a section under roadwork. The great number of roads that had sections under renovation or construction suggested the need to find the total number of accidents that took place under these conditions. This study reveals that a total of a 30% of accidents occurred at a roadwork section when it was in progress.

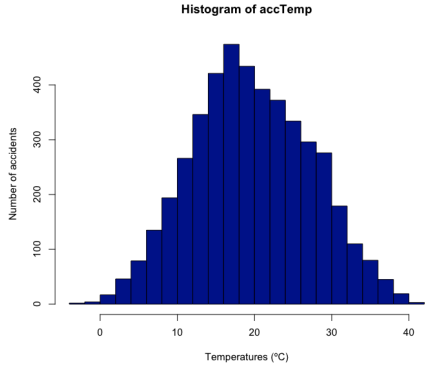



Figure 7: Distribution of car accidents by temperature

## 4 Conclusion

This paper presents a study of road traffic accidents in Spanish roads. Each traffic accident has been considered as a discrete and independent random event. In order to study the set of accidents we have obtained a number of variables associated to each one from several open data sources. Thus, the probability the probability of occurrence of an accident can be modelled by a high number of features, such as weather conditions, incidents caused by the start and end of a roadwork, geographical location of speed control radars, roadway infrastructure, etc. The total number of accidents in the period studied does not conform a big data problem by itself, however, the high number of variables involved imply that big data techniques must be deployed to extract and store the feature vectors.

In this study a more complex model is presented that separates from the classical triangle that relates driver, road and accident. On the contrary, a high number of other features is included and allow analysing accidents more deeply. In addition, two novel features are presented, such as the proximity to speed control radars of accidents and the accidents occurred in roadwork sections. The results have presented in absolute terms the roads in Spain with more accidents during the period studied. In addition, a distribution of accidents with respect to temperature has been presented. Also, the relation of the number of accidents with respect to the day of the week has been shown.

An important conclusion of our study is the need of more precise data from official sources that could model the accident. Specially, we have found great problems in the GPS precision of the accidents. As well, the period of study of accidents had to be restricted since we had no possible choices to correlate to datasets published in the past.

The results presented in this paper will allow us to further focus our research on prediction models that could suggest the probability of occurrence of accidents in roads depending on current variables and, as well, the need of changes in the infrastructure to prevent them.

## Acknowledgements

This work has been supported by the Generalitat Valenciana through the project GV/2015/031: Creación de mapas topológicos a partir de la apariencia global de un conjunto de escenas.

## References

[1] Apache Cassandra, http://cassandra.apache.org/
[2] Davison, P.A., Interrelationships between british drivers' visual Abilities, age and road accidents histories. *Ophthalmic and Physiological Optics*, ed. Oxford Pergamon Press, pp. 195-204, 1995
[3] DGT Statistical website, https://sedeapl.dgt.gob.es/WEB_IEST_CONSULTA/
[4] DGT Infocar, http://infocar.dgt.es/etraffic/

[5] European Union, Smart seat and seatbelt to help sleepy drivers stay alert. *Research*eu Results Magazine,* **42**, pp. 6-7, 2015.
[6] Hakim, S., Shefer, D., Hakkert, A., & Hocherman, I., A critical review of macro models for road accidents. *Accident Analysis and Prevention*, ed. Abdel-Aty, M., Elsevier, pp. 379-400, 1991
[7] Kalokota, K., Seneviratne, P.N., y Center, U.T., *Accident prediction models for two-lane rural highways*, Utah Transportation Center, 1994
[8] Sabey, B. E. & Taylor, H., The known risks we run: the highway. *Societal Risk Assessment*, ed. Schwing, R. C. and Albers, W. A., Springer US: Boston, pp. 43-70, 1980.
[9] Shankar, V., Milton, J. & Mannering, F., Modeling frequencies as zero-altered probability processes: an empirical inquiry. *Accident Analysis and Prevention*, ed. Abdel-Aty, M., Elsevier, pp. 829-837, 1997
[10] Simoncic, M., A bayesian network model of two-car accidents. *Journal of Transportation and Statistics*, ed. Jeeves, A., United States Bureau of Transportation Statistics, pp. 13-25, 2004
[11] Wang, C., Quddus, M.A., & Ison, S.G., Predicting accident frequency at their security levels and its application in site ranking using a two-stage mixed multivariable model. *Accident Analysis and Prevention*, ed. Abdel-Aty, M., Elsevier, pp. 1979-1990, 2011
[12] Wunderground, http://www.wunderground.com/
[13] Zegeer, C.V., Reinfurt, D., Hummer, J., Herf, L. & Hunter, W., Safety effects of cross-section design for two-lane roads, *Transportation Research Board*, ed. US Department of Transportation, 1987

# Analyzing Traffic Accident Severity in UK from a Classification point of view

D. Úbeda[1], A. Gil[1], A. Pérez[2], L. Payá[1]

*[1]Department of Systems Engineering and Automation*
*[2]Department of Economics and Financial Studies*
*Miguel Hernández University,*
*Avenida de la Universidad s/n, Elche, 03202 Alicante, Spain*

## Abstract

Road accidents are the cause of a high number of human losses in the world and also imply a significant psychological and physical trauma. We propose the use of different classification methods in order to study the underlying causes that caused the accident. In addition, we propose the use of different methods to infer the severity of each accident in a massive collection of UK accident data. In this paper, the large set of accidents is used to trace the most important factors that appear associated to each casualty. The influence of these features in the severity of an accident is significant and should be taken into account when proposing measures to help alleviate this type of undesirable events. The results show that some variables included in the dataset have a higher influence on accident severity and thus allow to further propose countermeasures to reduce injuries and deaths in roads.

# 1. Introduction

In this paper we present a study of traffic accident data using of a collection of data mining techniques. We propose the use of different methods to infer knowledge from a massive collection of UK accident data. The dataset is formed by a large number of accidents recorded from 2009 to 2014 that includes the severity of the accident and a large number of variables describing the causes that surrounded it, which includes road infrastructure, weather conditions and others.

Road accidents are the cause of a high number of human losses in the world and also imply a significant psychological and physical trauma. In addition, according to the European Union [4], the economic annual burden of road accidents is estimated in Europe between EUR 10 and 14 billion and the cost of a single fatality may be as high as EUR 1 million. These facts have motivated to include the improvement of road safety as one of the main challenges in the horizon 2020, which has increased the investment in traffic research. In addition, transparency platforms must be created and maintained in order to provide researchers with up-to-date data. It is worth recognising the huge effort carried out by the UK government in this direction. The high amount of open data publicly available has motivated the study presented in this manuscript.

The fatality rate in UK has decreased in a 46% since 2005 [2]. In average, 5 fatalities happen each day in UK roads and 61 citizens are seriously injured. Thus, a high effort must be still done to improve these rates.

In the dataset used, each accident is recorded by the police using the STATS 19 system. We propose to process the massive set of accidents in order to trace the most important factors that appear associated to each casualty. As a result, we can obtain a probability model that is able to classify the severity of each accident based on the variables involved in it. The models allow us to infer hidden relationships between accidents and their contributory factors, thus allowing us to extract useful information that can be of uttermost importance when planning policies to reduce traffic accidents.

The rest of the paper is structured as follows: Section 2 includes a state of the art in relation with traffic accident analysis. Next, Section 3 describes the dataset used in the study along with the variables involved. In addition the main experimental setup is described. Section 4 presents the main results obtained. Finally, the main conclusions are summarized.

# 2. State of the Art

Traffic accident analyses have been undertaken since vehicles became a predominant means of transportation. Generally, traffic authorities consider the driver-road-vehicle triangle in an attempt to study road accidents. However, most researchers assume that there exists a larger set of underlying circumstances that are related to each other and can be treated mathematically to model accident occurrence.

A common paradigm considers that accidents are not randomly scattered along the road network, and that drivers are not involved in accidents at random and thus there exists relationship between the frequency and severity of the accident with a set of variables that caused it [7].

Thus, to date, a high diversity of traffic accidents studies have been presented mostly differing on the kind of features used in the study and the data mining techniques to extract the most important factors involved in the accident. For example, Sabey and Taylor [13] used car speed at the time of the accident, whereas, Hakim et al. [5] studied the importance of the speed limit of the road. The use of security measures was a variable in Simoncic [15] research. Driver classification was used in other studies as one of their main features, i.e. age, used by Davison [1], or experience, also studied in [15]. Other factors involving the infrastructure have also been included in the studies, such as roadside width [6], section length or number of lanes [17].

In relation with the data mining techniques used we can highlight the use of logistic regression models in [12] to identify prediction factors of crashes and crash-related injuries. Also, [10] used neural networks to analyze vehicle accidents that occurred at city intersections in Milan, Italy. The methods are applied from a classification point of view to model accident severity in Korea and their results show that this technique provides crucial information for accident prevention. Finally, others propose the fusion of the information provided by individual classifiers to obtain better results [16].

# 3. Experimental design: Real world road accidents dataset

## 3.1. Data selection

As we mentioned in the introduction, during years the UK government has carried out a huge effort to create and maintain transparency platforms to helps researchers access up-to-date data. In this paper, all the dataset for road traffic accidents was obtained from the UK data archive "*Road Safety Data*" [3].

In Great Britain, the police records the details of every accident on public roads and the Local Processing Authority (LPA) validates and provides them first to the Department for Transport (DfT), and next to other authorities and UK Data Archive (UKDA), where datasets can be accessed via web by the researchers.

For each year, three different datasets of data are available:

- *Accident:* Includes timestamps of accident, severity, light and road conditions, and other important features involved in the accident.
- *Vehicle:* This dataset includes information of vehicle type, first point of impact, sex of the driver, and other features related to the vehicle and the driver.
- *Casualty:* severity of casualty, age of casualty, etc.

A more detailed information of every table is described in Table 1.

In this paper, these three datasets have been joined for every year and five tables have been constructed, one per year (2009 to 2014). Finally, these tables have been joined to produce a 'big' table with more than 500.000 instances.

As presented in Table 1, the 'big' dataset (DS1) contains only one class called "Casualty_Severity", this variable describes the effect of the accident on drivers, occupants or pedestrians. This class can have three different values: SLIGHT, SERIOUS or FATAL. Due to the nature of the data being analyzed, the three classes are highly unbalanced. We decided to analyze the results using different classification techniques and, as well, using balanced and unbalanced data. Dataset DS1 records Casualty_Severity with unbalanced data.

The simplified dataset DS2 derived from DS1 includes 20 variables. Finally, the third dataset (DS3) is derived from DS2 but only selecting those features that are used in the EURO NCAP tests [14]. Thus, the features selected were:

- Age_Band_of_Driver (<=25 years, 26-59 years, >= 60 years)

- Sex_of_Driver
- junction_detail (intersection; non-intersection)
- X1st_point_of_impact
- Speed_limit (of the crash site) (<40mph; 41-59 mph; >=60mph)
- year
- Number_of_Vehicles

Finally, last dataset (DS4) was selected in order to compare the results obtained on the five-year period analyses with a with a single-year analysis.

Table 1. Detailed information of every dataset employed in this study

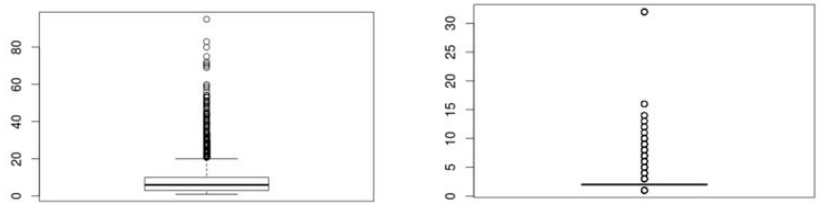
| Ref. | Description | Instances | Class Casualty_Serverity | | | Features |
|---|---|---|---|---|---|---|
| | | | Fatal | Serious | Slight | |
| DS1 | 2009-2014 | 511747 | 5647 | 58949 | 447151 | 47 |
| DS2 | 2009-2014 (simplified) | 64596 | 5647 | 58949 | - | 27 |
| DS3 | 2009-2014 (DS2 - EURO NCAP) | 64596 | 5647 | 58949 | - | 7 |
| DS4 | 2014 | 11100 | 957 | 10143 | - | 44 |

For the DS2, DS3 and DS4 datasets, slight accidents have been removed from the dataset since the main objective of this paper is to present solutions that are capable of reducing the occurrence fatal and serious accidents.

## 3.2 Data exploration: Preparing datasets for data mining

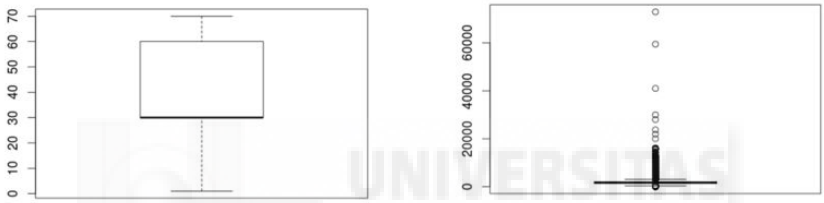
### 3.2.1 Type of data selection and anomalous value detection

For these datasets, two types of data have been used: *nominal* and *numerical* types. Weka [9] has an useful function to convert from csv format, with untyped data, to nominal. This function is called '*NumericToNominal*'. Also, R [19] has been used to detect outliers in order to remove them from the dataset. It is available by Box and Whiskers plots. This kind of plots draws outliers as points beneath and above the whiskers. Later, the records of these points are removed from the dataset. Figures 1 (cases (a) and (b)) and 2(b) illustrate these outlier points. On the other hand, Figure 2(a) does not contain any outlier.




a) Age of Vehicle boxplot b) Number of Vehicles boxplot

Figure 1: Boxplots obtained in R for Age of Vehicle and Number of Vehicles features




a) Speed Limits boxplot b) Engine Capacity boxplot

Figure 2: Boxplots obtained in R for Speed Limits and Engine Capacity features

### 3.2.2 Unbalanced datasets problem: Oversampling and downsampling

Real-world datasets are often heavily unbalanced, thus the classes do not appear with equal frequency in the dataset. This fact poses a problem in most classification or learning algorithms. Often the classifier turns their decision towards the most probable class.

Oversampling and undersampling in data analysis are techniques used to adjust the class distribution on the dataset. There are many options [8] to balance a dataset mostly involving oversampling and undersampling techniques. For example, algorithms like SMOTE [11] perform oversampling by including new synthetic samples that are randomly generated from examples in the dataset. Others obtain different copies of instances from the under-represented class.

In this paper the instances in the data were balanced so that each class has the same total weight. The total sum of weights across all instances were maintained and only the weights in the first batch of data received by this filter were changed.

In order to compare the results between balanced and unbalanced data, and the differences found in attribute selection and raw dataset, a total of four experiments (Figure 3) have been designed:

- ***Experiment 1*:** Figure 3(a) shows the flowchart for this experiment. Only data mining with a training and testing datasets (cross validation) was performed. Attribute selection and class balancing were not used.
- ***Experiment 2*:** Figure 3(b) shows the flowchart for this experiment. The experiment presents and initial attribute-selection stage. Second, data mining with a training and a testing datasets (cross validation) is carried out. Class balancing is not used.
- ***Experiment 3*:** Figure 3(c) shows the flowchart for this experiment. Attribute selection is performed first, second, class balancing is carried out and finally a data mining technique is applied.
- ***Experiment 4*:** Figure 3(d) shows the flowchart for this experiment. Class balancing is applied and next data mining with a training and a testing datasets (cross validation). Attribute selection is not considered in this experiment.

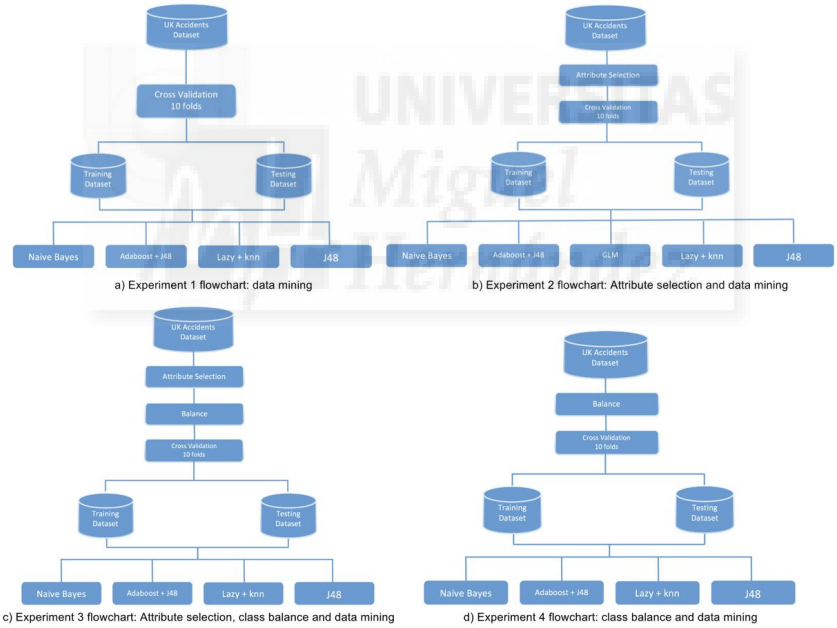



Figure 3: Flowcharts designed for every experiment in this study.

# 4. Experimental Results

## 4.1 Table Description

The results are presented using four different parameters which are described next:

**Algorithm**
Represents one of the five algorithms that have been tested with all datasets. Each dataset has been preprocessed in different ways in order to observe the influence on the results. We have applied the following techniques to the resulting datasets: A Naive Bayes classifier, a J48 decision tree, a lazy classificator with k-nearest neighbours, an Adaboost combined with a J48 classifier and General Linear Models.

**ROC curve**
The ROC curve is a graphical plot that illustrates the performance of a binary classifier system when its discrimination threshold is varied. The best possible prediction method is represented with False Positives (FP) and False Negatives (FN). In these tables the ROC curve is represented with values between 0 (wrong classification) and 1 (correct classification). For the sake of brevity, this graphical results are substituted by the ROC number.

**Casualty_severity**
Represents the gravity of the accident for every casualty. For this class there are three possible values:

1. *Fatal:* includes cases where death occurs in less than 30 days as a result of the accident
2. *Serious:* includes fractures, internal injury, severe cuts, crushing, burns, concussion, severe shock requiring hospital treatment, detention in hospital as an inpatient immediately or at a later day and injuries derived from the crash resulting in death 30 days or more after the crash.
3. *Slight:* includes sprains or whiplash not necessarily requiring medical treatment, bruises, slight cuts, slight shock requiring roadside attention.

**TP Rate**
Represents the true positive (TP) rate. TP measures the proportion of positives that are correctly identified as such.

**FP Rate**
Represents false positive rate. False positives and false negatives are concepts analogous to *type I and type II errors.* The False positive rate is the proportion of all negatives that still yield positive test outcomes, i.e., the conditional probability of a positive test result given an event that was not present [18].

**RMSE**
Root Mean Squared Error measures the differences between values predicted by a model or an estimator and the values actually observed. The RMSD represents the sample standard deviation of the differences between predicted values and observed values.

**Total cases**
Represents the total number of cases for every instance of casualty severity class.

**Good/Bad**
According to the confusion matrix, this is the number of correctly identified instances.

**No. Feat.**

Number of original or chosen features of the dataset.

## 4.2 Experiment 1: Data mining

### 4.2.1 DS2 - Subset 2009-2014 simplified

In this experiment we apply four different techniques to the DS2 data subset. According to the results presented in Table 2, the Naive Bayes classifier outperforms the rest in terms correctly classifying serious accidents. However, the Naive Bayes classifier performs poorly when recognising the fatal class. In addition, the FP Rate for the serious class is very high, although the number of good classified instances is significant.

It is also noticeable that the J48 decision tree is unable to obtain a good classification algorithm and is not capable to recognise the fatal class.

Table 2: Data mining results for subset DS2 with Naive Bayes, J48, Lazy + knn and Adaboost + J48 algorithms

| Algorithm | Casualty_ severity | TP Rate | FP Rate | ROC Area | RMSE | Total cases | Good/Bad | No. Attr. |
|---|---|---|---|---|---|---|---|---|
| **Naive Bayes** | **1 Fatal** | **0.118** | **0.046** | **0.693** | **0.316** | **5647** | **667/4980** | 27 |
| | **2 Serious** | **0.954** | **0.882** | **0.693** | | **58949** | **56236/2713** | |
| Lazy + Knn | 1 Fatal | 0.194 | 0.083 | 0.558 | 0.3806 | 5647 | 1097/4550 | |
| | 2 Serious | 0.917 | 0.886 | 0.558 | | 58949 | 5450/4899 | |
| Tree J48 confidence= 0.25 | 1 Fatal | 0 | 0 | 0.500 | 0.2824 | 5647 | 0/5647 | |
| | 2 Serious | 1 | 1 | 0.500 | | 58949 | 58949/0 | |
| Adaboost + J48 | 1 Fatal | 0.089 | 0.029 | 0.648 | 0.316 | 5647 | 504/5143 | |
| | 2 Serious | 0.971 | 0.911 | 0.648 | | 58949 | 57230/1719 | |
| SVM | 1 Fatal | 0 | 0 | | | 5646 | 0/5646 | 33 |
| | 2 Serious | 1 | 1 | | | 58933 | 58924/9 | |

### 4.2.2 DS3 - Subset 2009-2014 EURO NCAP

The classifiers tested with this subset behave in a similar manner. Serious accidents are classified correctly, but they are not able to classify fatal severity. The Naive Bayes algorithm performs slightly better than the rest since it shows a better ROC curve and the lowest RMSE (Table 3).

Table 3: Data mining results for subset DS3 with Naive Bayes, J48, Lazy + knn and Adaboost + J48 algorithms

| Algorithm | Casualty_ severity | TP Rate | FP Rate | ROC Area | RMSE | Total cases | Good/Bad | No. Attr. |
|---|---|---|---|---|---|---|---|---|

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Naive Bayes** | **1 Fatal** | **0** | **0** | **0.649** | **0.2798** | **5647** | **1/5646** | 7 |
| | **2 Serious** | **1** | **1** | **0.649** | | **58949** | **58944/5** | |
| Lazy + Knn | 1 Fatal | 0.023 | 0.010 | 0.625 | 0.2895 | 5647 | 0/5647 | |
| | 2 Serious | 0.990 | 0.977 | 0.625 | | 58949 | 58949/0 | |
| Tree J48 confidence= 0.25 | 1 Fatal | 0 | 0 | 0.5 | 0.2824 | 5647 | | |
| | 2 Serious | 1 | 1 | 0.5 | | 58949 | | |
| Adaboost + J48 | 1 Fatal | 0.004 | 0.001 | 0.641 | 0.2815 | 5647 | 20/5627 | |
| | 2 Serious | 0.999 | 0.996 | 0.641 | | 58949 | 58913/36 | |

### 4.2.3 Selected algorithms

In the Table 4, there is a summary where the algorithms performing best for every dataset are presented.

Table 4: Selected algorithms from data mining results for subsets DS1-4

| Subset | Feat. | Alg. | Severity | TP | TN | ROC | RMSE |
|---|---|---|---|---|---|---|---|
| **Subset 2009-2014 simplificado** | 27 | **Naive Bayes** | **1 Fatal** | **0.118** | **0.046** | **0.693** | **0.316** |
| | | | **2 Serious** | **0.954** | **0.882** | **0.693** | |
| Subset 2009-2014 EURONCAP | 7 | Naive Bayes | 1 Fatal | 0 | 0 | 0.649 | 0.2798 |
| | | | 2 Serious | 1 | 1 | 0.649 | |

## 4.3 Experiment 2: Attribute selection and Data mining

### 4.3.1 DS1 - Subset 2009-2014

#### A) Attribute selection

The attribute selector *cfsubseteval* was run on the whole dataset and suggested the use of the following features:

> *Vehicle_Type, Vehicle_Manoeuvre, X1st_Point_of_Impact, Sex_of_Driver, Engine_Capacity_.CC., Casualty_Class, Casualty_Severity, Light_Conditions, Road_Surface_Conditions, Urban_or_Rural_Area*

Figure 4(b) and (c) shows both Naive Bayes algorithm and J48 tree respectively. As shown in Figure 4(c) J48 had a poor classification of fatal and serious accidents.

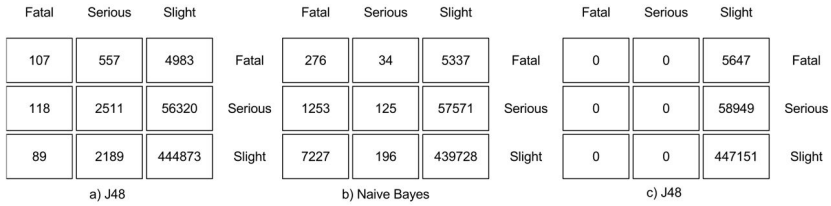



Figure 4: Confusion matrices obtained for DS1 with J48 and Naive Bayes algorithms

### 4.3.2 DS2 - Subset 2009-2014 simplified

#### A) Attribute selection

The criterion here was to select those characteristics that were included at least in half of the ten folds indicated by *SubsetEval + Bestfirst* algorithm:

> *Vehicle_Manoeuvre, Vehicle_Leaving_Carriageway, Sex_of_Driver, Light_Conditions, Engine_Capacity, Age_Band_of_Casualty, X1st_Road_Class, Speed_limit, Junction_Detail, Urban_or_Rural_Area*

#### B) Mining

If we pay attention to Table 5 a slight improvement has been achieved in the classification of fatals compared to Table 4. In particular the good/bad ratio in Table 5 is increased in 0.2 when the Lazy+knn algorithm is used.

Table 5: Data mining results for subset DS2 with Naive Bayes, J48 and Lazy + knn algorithms

| Algorithm | Casualty_ severity | TP Rate | FP Rate | ROC Area | RMSE | Total cases | Good/Bad | No. Attr. |
|---|---|---|---|---|---|---|---|---|
| Naive Bayes | 1 Fatal | 0.065 | 0.028 | 0.683 | 0.3081 | 5647 | 366/5281 | 10 |
| | 2 Serious | 0.972 | 0.935 | 0.683 | | 58949 | 57295/1654 | |
| Lazy + Knn | 1 Fatal | 0.172 | 0.084 | 0.552 | 0.3728 | 5647 | 974/4673 | |
| | 2 Serious | 0.916 | 0.828 | 0.552 | | 58949 | 53973/4976 | |
| Tree J48 confidence= 0.25 | 1 Fatal | 0 | 0 | 0.5 | 0.2824 | 5647 | 0/5647 | |
| | 2 Serious | 1 | 1 | 0.5 | | 58949 | 58949/0 | |
| **GLM, binomial (logit)** | **1 Fatal** | **0.811** | **0.499** | | **1.987** | **5646** | **4581/1065** | **16** |
| | **2 Serious** | **0.500** | **0.189** | | | **5646** | **4581/1065** | |

### 4.3.3 DS4 - Subset 2014

#### A) Attribute selection

The selected attributes from the *SubsetEval + Bestfirst* algorithm were:

> *Vehicle_Type, Junction_Location, Hit_Object_off_Carriageway, Engine_Capacity_.CC., Hour_of_Day, Local_Authority_.District., Speed_limit, Light_Conditions, Urban_or_Rural_Area*

#### B) Data Mining

The best results were obtained by the GLM algorithm (Table 6). For GLM we didn't used *SubsetEval + Bestfirst* algorithm. In this case a reduce data training set has been used. A selection of the attributes based on probabilistic significance was done:

> *Vehicle_Type, Vehicle_Manoeuvre, Junction_Location, Vehicle_Leaving_Carriageway, Hit_Object_off_Carriageway, Sex_of_Driver, Age_Band_of_Casualty, Casualty_Type, Police_Force, X1st_Road_Class, Road_Type, Speed_limit, Light_Conditions, Road_Surface_Conditions, Urban_or_Rural_Area, month*

Table 6: Data mining results for subset DS4 with Naive Bayes, J48, Adaboost + J48 and GLM algorithms.

| Algorithm | Casualty_ severity | TP Rate | FP Rate | ROC Area | RMSE | Total cases | Good/Bad | No. Attr. |
|---|---|---|---|---|---|---|---|---|
| Naive Bayes | 1 Fatal | 0.089 | 0.033 | 0.675 | 0.3076 | 957 | 85/872 | 9 |
| | 2 Serious | 0.967 | 0.911 | 0.675 | | 10143 | 9813/330 | |
| Tree J48 confidence= 0.25 | 1 Fatal | 0 | 0 | 0.499 | 0.2807 | 957 | 0/957 | |
| | 2 Serious | 1 | 1 | 0.499 | | 10143 | 10143/0 | |
| Adaboost + J48 | 1 Fatal | 0.180 | 0.069 | 0.591 | 0.3456 | 957 | 172/785 | |
| | 2 Serious | 0.931 | 0.820 | 0.591 | | 10143 | 9448/695 | |
| **GLM, binomial (logit)** | **1 Fatal** | **0.676** | **0.323** | **-** | **-** | **637** | **431/206** | 16 |
| | **2 Serious** | **0.641** | **0.359** | **-** | | **6763** | **2430/4333** | |

### 4.3.4 Selected algorithms

Table 7 summarizes the performance of the best algorithms found for every dataset analyzed in this experiment. According to the DS4, the algorithm which has obtained best results has been GLM. Opposite, according to subset with years from 2009 to 2014, the

Naive Bayes and Lazy + kNN have obtained similar results. Lamentably the NB algorithm was unable to classify correctly any fatal accident.

Table 7: Selected algorithms from data mining results for subsets DS1-2,4

| Subset | Feat. | New Feat. | Alg | Severity | TP | TN | ROC | RMSE |
|---|---|---|---|---|---|---|---|---|
| **Subset 2014** | **44** | **16** | **GLM, binomial (logit)** | **1 Fatal** | **0.676** | **0.323** | **-** | **-** |
| | | | | **2 Serious** | **0.641** | **0.359** | **-** | |
| **Subset 2009-2014 simplificado** | **27** | **10** | **GLM, binomial (logit)** | **1 Fatal** | **0.811** | **0.499** | | **1.987** |
| | | | | **2 Serious** | **0.500** | **0.189** | | |
| Subset 2009-2014 EURONCAP | 7 | 7 | Naive Bayes | 1 Fatal | 0 | 0 | 0.649 | 0.2798 |
| | | | | 2 Grave | 1 | 1 | 0.649 | |

## 4.4 Experiment 3: Attribute selection + Balance + Data mining

### 4.4.1 DS1 - Subset 2009-2014

#### A) Attribute selection

For this subset the same features as in Experiment 1 were chosen.

#### B) Data balance

Next, the subset was balanced for the *casualty_severity* class. To achieve this, the *ClassBalancer* function from Weka was used. This function resamples the instances in the data so that each class has the same number of instances. The total sum of weights across all instances will be maintained. Only the weights in the first batch of data received by this filter are changed.
Once the dataset was balanced, the J48 decision tree was run. The resulting confusion matrix is shown in Figure 5(a).

#### C) Data mining

Balancing the data has resulted in an improvement of the results. However the error is still high and the shape of the tree does not provide valid conclusions. Naive Bayes algorithm, Figure 5(b), was a new option.


Figure 5: Confusion matrices obtained for DS1 with J48 and Naive Bayes algorithms

Finally, new features were included with the purpose of improving the results in this balanced dataset. The following attributes were used:

> *Day of week, Type of road, Light conditions, Police force, Vehicle manoeuvre, Towing and articulation, Skidding and overturning, Sex of driver*

Naive Bayes, Figure 6(a), and J48, Figure 6(b) confusion matrices are shown.

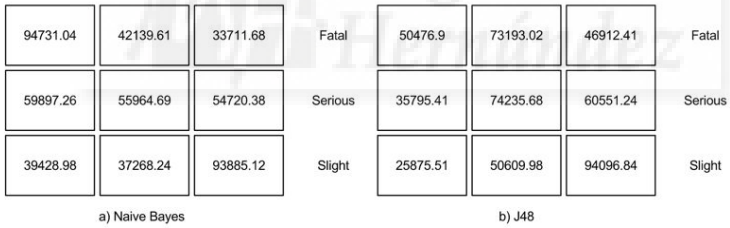



Figure 6: Confusion matrices obtained for DS1 with J48 and Naive Bayes algorithms with new attributes.

For this subset (Table 8), the J48 tree and Naive Bayes, gives very similar results for fatal severity class. Nevertheless, the serious class presents TP values that are significantly smaller when compared to the Naive Bayes (11 features) and J48 tree (7 features).

The Naive Bayes with 7 features and J48 with six features, present similar results, being the J48 slightly better.

Table 8: Data mining results for subset DS1 with Naive Bayes and J48 algorithms.

| Subset | Feat. | New Feat. | Alg | Severity | **TP** | **TN** | **ROC** | RMSE |
|---|---|---|---|---|---|---|---|---|
| **Subset 2009-2014** | **47** | **6** | **J48** | **1 Fatal** | **0.564** | **0.292** | **0.685** | **0.4469** |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | **2 Serious** | **0.330** | **0.172** | **0.639** | |
| | | | | **3 Slight** | **0.646** | **0.266** | **0.757** | |
| Subset 2009-2014 | 47 | 6 | Naive Bayes | 1 Fatal | 0.533 | 0.292 | 0.678 | 0.4525 |
| | | | | 2 Serious | 0.235 | 0.111 | 0.635 | |
| | | | | 3 Slight | 0.715 | 0.356 | 0.751 | |
| Subset 2009-2014 | 47 | 11 | Naive Bayes | 1 Fatal | 0.430 | 0.176 | 0.719 | 0.4543 |
| | | | | 2 Serious | 0.409 | 0.208 | 0.642 | |
| | | | | 3 Slight | 0.715 | 0.338 | 0.763 | |
| Subset 2009-2014 | 47 | 7 | Naive Bayes | 1 Fatal | 0.555 | 0.291 | 0.693 | 0.4522 |
| | | | | 2 Serious | 0.328 | 0.233 | 0.598 | |
| | | | | 3 Slight | 0.550 | 0.259 | 0.703 | |
| Subset 2009-2014 | 47 | 7 | J48 | 1 Fatal | 0.296 | 0.181 | 0.548 | |
| | | | | 2 Serious | 0.435 | 0.363 | 0.548 | |
| | | | | 3 Slight | 0.552 | 0.315 | 0.646 | |

### 4.4.2 DS2 - Subset 2009-2014

#### A) Attribute selection

The attribute selection made on past sections for this subset was the same for this experiment.

#### B) Balance

For the Subset 2009-2014 with 27 instances, the number of new instances for the class *Casualty_severity* for both fatal and serious was 32298.

#### C) Data Mining

According to Table 9, it is interesting to highlight that the Naive Bayes algorithm obtains better results for the TP Rate for fatal severity than the rest of algorithms. The TP Rate for serious severity presents poorer results than others, but the global results are improved due to the high TP rate in the classification of fatal accidents.

Table 9: Data mining results for subset DS2 with Naive Bayes, Lazy+knn, J48 and Adaboost+J48 algorithms.

| Algorithm | Casualty_ severity | TP Rate | FP Rate | ROC Area | RMSE | Total cases | Good/Bad | No. Attr. |
|---|---|---|---|---|---|---|---|---|

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Naive Bayes** | **1 Fatal** | **0.599** | **0.340** | **0.683** | **0.2798** | **32297** | **19337/12960** | 10 |
| | **2 Serious** | **0.660** | **0.401** | **0.683** | | **32297** | **21301/10996** | |
| Lazy + Knn | 1 Fatal | 0.183 | 0.096 | 0.556 | 0.6734 | 32297 | 5696/26401 | |
| | 2 Serious | 0.904 | 0.817 | 0.556 | | 32297 | 29200/3097 | |
| Tree J48 confidence= 0.25 | 1 Fatal | 0.399 | 0.229 | 0.566 | 0.6162 | 32297 | 12886/19411 | |
| | 2 Serious | 0.771 | 0.601 | 0.566 | | 32297 | 24897/7400 | |
| Adaboost + J48 | 1 Fatal | 0.187 | 0.092 | 0.621 | 0.6526 | 32297 | 6039/26258 | |
| | 2 Serious | 0.908 | 0.813 | 0.621 | | 32297 | 29330/2967 | |

### 4.4.3 DS4 - Subset 2014

#### A) Attribute selection

The attribute selection made on 4.3.3 is going to be the same for this experiment.

#### B) Balance

Nevertheless, for the subset 2014 (44 instances) the number of new instances for the same class for both fatal and serious was 5550.

#### C) Data Mining

According to Table 10, again it is noticeable how the Naive Bayes algorithm has better results for the TP Rate for fatal severity than the rest of algorithms. On this occasion, the TP Rate for serious severity has practically the same results than the other algorithms.

Table 10: Data mining results for subset DS4 with Naive Bayes, J48 and Adaboost+J48 algorithms.

| Algorithm | Casualty_ severity | TP Rate | FP Rate | ROC Area | RMSE | Total cases | Good/Bad | No. Attr. |
|---|---|---|---|---|---|---|---|---|
| **Naive Bayes** | **1 Fatal** | **0.518** | **0.282** | **0.661** | **0.5113** | **5550** | **2876/2673** | 9 |
| | **2 Serious** | **0.718** | **0.482** | **0.661** | | **5550** | **3983/1566** | |
| Tree J48 confidence= 0.25 | 1 Fatal | 0.353 | 0.188 | 0.566 | 0.6223 | 5550 | 1960/3589 | |
| | 2 Serious | 0.812 | 0.647 | 0.566 | | 5550 | 4508/1041 | |
| Adaboost + J48 | 1 Fatal | 0.231 | 0.105 | 0.591 | 0.6472 | 5550 | 1281/4268 | |
| | 2 Serious | 0.895 | 0.769 | 0.591 | | 5550 | 4968/581 | |

### 4.4.4 Selected algorithms

Table 11 shows the best algorithms for every dataset analyzed.

Table 11: Selected algorithms from data mining results for subsets DS1-4

| Subset | Feat. | New Feat. | Alg | Severity | **TP** | **TN** | **ROC** | RMSE |
|---|---|---|---|---|---|---|---|---|
| Subset 2009-2014 | 47 | 6 | J48 | 1 Fatal | 0.564 | 0.292 | 0.685 | 0.4469 |
| | | | | 2 Serious | 0.330 | 0.172 | 0.639 | |
| | | | | 3 Slight | 0.646 | 0.266 | 0.757 | |
| **Subset 2014** | **44** | **9** | **Naive Bayes** | **1 Fatal** | **0.518** | **0.282** | **0.661** | **0.5113** |
| | | | | **2 Serious** | **0.718** | **0.482** | **0.661** | |
| **Subset 2009-2014 simplificado** | **27** | **10** | **Naive Bayes** | **1 Fatal** | **0.599** | **0.340** | **0.683** | **0.2798** |
| | | | | **2 Serious** | **0.660** | **0.401** | **0.683** | |
| Subset 2009-2014 EURONCAP | 7 | 7 | Tree J48 confidence =0.25 | 1 Fatal | 0.620 | 0.398 | 0.635 | 0.5027 |
| | | | | 2 Serious | 0.602 | 0.380 | 0.635 | |

## 4.5 Experiment 4: Balance + Data mining

### 4.5.1 DS2 - Subset of 2009-2014

#### A) Balance

The balance after applying ClassBalancer function from Weka was 32298 for both classes, Fatal and Serious.

#### B) Data mining

As said before, this dataset has 27 variables that were chosen manually by the authors of this paper. Those variables that were thought to be unrelated with casualty severity were discarded. Starting at this point, the results obtained with the algorithms from Table 12 must be considered as conclusive results for this dataset.

Once again, Naive Bayes gives better results with the need to balance both classes, Fatal and Serious. The TP rate here is very similar for both classes, 0.621 and 0.659 respectively.

Table 12: Data mining results for subset DS2 with Naive Bayes, J48, Adaboost+J48 and Lazy+knn algorithms.

| Algorithm | Casualty_severity | TP Rate | FP Rate | ROC Area | Root mean squared error |
|---|---|---|---|---|---|
| **Naive Bayes** | **1 Fatal** | **0.621** | **0.341** | **0.693** | **0.4976** |
| | **2 Serious** | **0.659** | **0.379** | **0.693** | |
| Lazy + Knn | 1 Fatal | 0.195 | 0.084 | 0.558 | 0.662 |
| | 2 Serious | 0.916 | 0.805 | 0.558 | |

| | | | | | |
|---|---|---|---|---|---|
| Tree J48 confidence=0.25 | 1 Fatal | 0.364 | 0.194 | 0.571 | 0.6242 |
| | 2 Serious | 0.806 | 0.636 | 0.571 | |
| Adaboost + J48 | 1 Fatal | 0.132 | 0.045 | 0.646 | 0.6638 |
| | 2 Serious | 0.955 | 0.524 | 0.646 | |

### 4.5.2 DS3 - Subset 2009-2014 EURONCAP

#### A) Balance

For this dataset the attribute selection taken from the EURO NCAP tests has 7 features, and the balance after applying ClassBalancer was 32298 for both classes, Fatal and Serious.

#### B) Data Mining

According to the Table 13, on this occasion, the J48 tree has been selected because the success rate for the fatals is slightly higher than the rest of the tests with different algorithms.

Table 13: Data mining results for subset DS3 with Naive Bayes, J48, Adaboost+J48 and Lazy+knn algorithms.

| Algorithm | Casualty_severity | TP Rate | FP Rate | ROC Area | Root mean squared error |
|---|---|---|---|---|---|
| Naive Bayes | 1 Fatal | 0.606 | 0.381 | 0.648 | 0.4838 |
| | 2 Grave | 0.619 | 0.394 | 0.648 | |
| Lazy + Knn | 1 Fatal | 0.582 | 0.385 | 0.625 | 0.5243 |
| | 2 Grave | 0.615 | 0.418 | 0.625 | |
| **Tree J48 confidence= 0.25** | **1 Fatal** | **0.620** | **0.398** | **0.635** | **0.5027** |
| | **2 Grave** | **0.602** | **0.380** | **0.635** | |
| Adaboost + J48 | 1 Fatal | 0.592 | 0.382 | 0.639 | 0.5015 |
| | 2 Grave | 0.618 | 0.408 | 0.639 | |

### 4.5.3 Selected algorithms

Table 14 presents the best algorithms for every dataset analyzed in this experiment.

Table 14: Selected algorithms from data mining results for subsets DS1-4

| Subset | Feat. | Alg | Severity | TP | TN | ROC | RMSE |
|---|---|---|---|---|---|---|---|

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Subset 2009-2014 | 47 | J48 | 1 Fatal | 0.967 | 0 | 0.988 | 0.1349 |
| | | | 2 Grave | 0.950 | 0.015 | 0.954 | |
| | | | 3 Slight | 1 | 0.028 | 0.942 | |
| Subset 2014 | 44 | J48 | 1 Fatal | 0.971 | 0 | 0.984 | 0.1198 |
| | | | 2 Grave | 1 | 0.029 | 0.984 | |
| **Subset 2009-2014 simplificado** | **27** | **Naive Bayes** | **1 Fatal** | **0.621** | **0.341** | **0.693** | 0.4976 |
| | | | **2 Grave** | **0.659** | **0.379** | **0.693** | |
| Subset 2009-2014 EURONCAP | 7 | Naive Bayes | 1 Fatal | 0.620 | 0.398 | 0.635 | 0.5027 |
| | | | 2 Grave | 0.602 | 0.380 | 0.635 | |

## 4.6 Experimental results conclusions

Final results for all experiments for DS1-3 (Table 15) and DS4 (Table 16) are shown. On the one hand, according to experiment 3 with J48 and the subset 2009-2014 with 6 variables, the most frequent vehicle manoeuvre of a motorcycle between 125 cc and 500 cc or more that produces a fatal consequence is going ahead other. Also going ahead left-hand bend and Overtaking a moving vehicle - offside are frequent manoeuvres with fatal consequences. Finally, overtaking a static vehicle - offside and moving off produces fatal consequences for a driver of a motorcycle over 500 cc.

According to private hire cars or Taxis, going ahead other with the impact on the front of the car is also the main manoeuvre that causes a final fatal to the occupants of the taxi. An offside impact while going ahead other is another main fatal accident for these type of cars.

When talking about private cars, the main cause of a fatal consequence for an occupant of a car is a front impact going ahead other. The second fatal cause is also a front impact going ahead left-hand bend between two cars. A front first point of impact overtaking moving vehicle - offside is the next cause of this kind of catastrophic accidents. Nearside first impact going ahead other in a two-vehicle accident is the main cause of a serious impact in the UK. Next, with the same manoeuvre, but itself in an offside impact is the second cause of serious accidents.

On the other hand, according to Naive Bayes classification of the 2009-2014 simplified with a manual election of 27 features (Table 16), about a 70,5% of the total accidents (fatal + serious), an offside - crossed central reservation finished with a fatal ending. Age band of casualties is another significant feature that needs to be analyzed because as age is increased the probability of a fatal accidents increases too. About a 67,7% of total accidents had the worst consequences for occupants over 75 years old. The next range of age for these kind of accidentes is between 66 and 75 years old.

Finally, darkness is the next consequence of fatal accidents occurred at roads without lighting with a 67,1% of deaths. When the road has lights unlit, the probability of having an accident with a fatal ending is about 59% for the total of accidents occurred (fatal + serious). According to the sex of drivers, 52,3% of casualties occurred when a man was driving.
The mean of the engine capacity of the vehicles analyzed is 2000cc for the fatals, and 1600 for the serious. The majority of the fatal accidents occurred in a road type A(M), with a 65,4% of death rate closely followed by motorways with a 65,4%.

Table 15: Selected algorithms from data mining results for subsets DS4

| Method | Subset | Feat. | New Feat. | Alg. | Severity | **TP** | **TN** | **ROC** | RMSE |
|---|---|---|---|---|---|---|---|---|---|
| 2 | Subset 2014 | 44 | 16 | GLM, binomial (logit)* | 1 Fatal | 0.676 | 0.323 | - | - |
| | | | | | 2 Serious | 0.641 | 0.359 | - | |
| 3 | Subset 2014 | 44 | 9 | Naive Bayes | 1 Fatal | 0.518 | 0.282 | 0.661 | 0.5113 |
| | | | | | 2 Serious | 0.718 | 0.482 | 0.661 | |

Table 16: Selected algorithms from data mining results for subsets DS1-3

| Method | Subset | Feat. | New Feat. | Alg. | Severity | **TP** | **TN** | **ROC** | RMSE |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Subset 2009-2014 simplificado | 27 | 27 | Naive Bayes | 1 Fatal | 0 | 0 | 0.649 | 0.2798 |
| | | | | | 2 Serious | 1 | 1 | 0.649 | |
| 2 | Subset 2009-2014 simplificado | 27 | 16 | GLM, binomial (logit) | 1 Fatal | 0.811 | 0.499 | | 1.987 |
| | | | | | 2 Serious | 0.500 | 0.189 | | |
| 3 | Subset 2009-2014 simplificado | 27 | 10 | Naive Bayes | 1 Fatal | 0.599 | 0.340 | 0.683 | 0.2798 |
| | | | | | 2 Serious | 0.660 | 0.401 | 0.683 | |
| 3 | Subset 2009-2014 | 47 | 6 | J48 | 1 Fatal | 0.564 | 0.292 | 0.685 | 0.4469 |
| | | | | | 2 Serious | 0.330 | 0.172 | 0.639 | |
| | | | | | 3 Slight | 0.646 | 0.266 | 0.757 | |
| 4 | Subset 2009-2014 simplificado | 27 | 27 | Naive Bayes | 1 Fatal | 0.621 | 0.341 | 0.693 | 0.4976 |
| | | | | | 2 Serious | 0.659 | 0.379 | 0.693 | |

## 5. Conclusions

In this paper, we proposed the use of different methods to infer knowledge from a massive collection of UK accident data. We propose the use of different classification methods and, as well, compare the results obtained with different pre-processing techniques.

As presented in the experiments, the results obtained are highly dependant on the input dataset and the pre-processing techniques used. In addition, the different algorithms tested in this paper have presented dissimilar behaviours, being the Naive Bayes algorithm surprisingly accurate in most of the experiments performed.

Moreover, the experiments have shown that the classifiers presented allow to extract the most important variables involved in traffic accidents in UK and, as well, propose political actions aimed at reducing accident fatalities on roads.

Finally, as a future work we plan to combine different classifiers into a single one in order to enhance the results obtained globally. In our opinion, by doing this, the capability to extract important information from the datasets will be improved.

## Competing interests

The authors declare that there are no competing interests regarding the publication of this paper.

## References


[1] Davison, P.A., Interrelationships between british drivers' visual abilities, age and road accidents histories. *Ophthalmic and Physiological Optics*, ed. Oxford Pergamon Press, pp. 195-204, 1995

[2] Department for Transport. "Reported road casualties in Great Britain 2015", *UK Government*, https://www.gov.uk/government/statistics/reported-road-casualties-in-great-britain-main-results-2015

[3] Department for Transport. Road Accident Statistics Branch, 2013, *Road Accident Data, 2012*, UK Data Service, SN: 7431, http://dx.doi.org/10.5255/UKDA-SN-7431-1

[4] European Union, Smart seat and seatbelt to help sleepy drivers stay alert. Research*eu Results Magazine, 42, pp. 6-7, 2015.

[5] Hakim, S., Shefer, D., Hakkert, A., & Hocherman, I., A critical review of macro models for road accidents. *Accident Analysis and Prevention*, ed. Abdel-Aty, M., Elsevier, pp. 379-400, 1991

[6] Kalokota, K., Seneviratne, P.N., y Center, U.T., *Accident prediction models for two-lane rural highways*, Utah Transportation Center, 1994

[7] Kononov, J. and B. N. Janson (2002). "Diagnostic Methodology for the Detection of Safety Problems at Intersections " Journal Transportation Research Record 1784: 51-56, 2002

[8] M. R. Anderson and M. Cafarella, "Input selection for fast feature engineering," *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, Helsinki, pp. 577-588, 2016. doi: 10.1109/ICDE.2016.7498272

[9] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten, "The WEKA Data Mining Software: An Update", *SIGKDD Explorations*, Volume 11, Issue 1, 2009.

[10] Mussone, L., A. Ferrari, et al., "An analysis of urban collisions using an artificial intelligence model." Accident Analysis and Prevention 31:705-718, 1999

[11] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique", Volume 16, pages 321-357, 2002. doi:10.1613/jair.953

[12] Ossenbruggen, P. J., J. Pendharkar, et al. "Roadway safety in rural and small urbanized areas." Accidents Analysis and Prevention 33(4):485-498, 2001
[13] Sabey, B. E. & Taylor, H., The known risks we run: the highway. *Societal Risk Assessment*, ed. Schwing, R. C. and Albers, W. A., Springer US: Boston, pp. 43-70, 1980.
[14] Segui-Gomez M, Lopez-Valdes FJ, Frampton R. "An Evaluation of the Euroncap Crash Test Safety Ratings in the Real World", *Annual Proceedings / Association for the Advancement of Automotive Medicine*, no. 51, pp. 281-298, 2007.
[15] Simoncic, M., A bayesian network model of two-car accidents. *Journal of Transportation and Statistics*, ed. Jeeves, A., United States Bureau of Transportation Statistics, pp. 13-25, 2004
[16] Sohn, S. and S. Lee, "Data fusion, ensemble and clustering to improve the classification accuracy for the severity of road traffic accidents in Korea. ." Safety Science 41(1): 1-14, 2002
[17] Wang, C., Quddus, M.A., & Ison, S.G., Predicting accident frequency at their security levels and its application in site ranking using a two-stage mixed multivariable model. *Accident Analysis and Prevention*, ed. Abdel-Aty, M., Elsevier, pp. 1979-1990, 2011
[18] Wikipedia contributors. "False positives and false negatives." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 4 Jul. 2016. Web. 31 Jul. 2016.
[19] R: The R Project for Statistical Computing. https://www.r-project.org/

UNIVERSITAS
Miguel
Hernández

# Apéndice: Matriz de correlación de predictores

Two-Step Estimates

Correlations/Type of Correlation:

| | accyr | Vehicle_Type | Vehicle_Manoeuvre | Junction_Location | Skidding_and_Overturning | Vehicle_Leaving_Carriageway | Hit_Object_off_Carriageway | First_Point_of_Impact | Journey_Purpose_of_Driver |
|---|---|---|---|---|---|---|---|---|---|
| accyr | 1 | Polyserial | Polyserial | Polyserial | Polyserial | Polyserial | Polyserial | Polyserial | Polyserial |
| Vehicle_Type | -0.01787 | 1 | Polychoric | Polychoric | Polychoric | Polychoric | Polychoric | Polychoric | Polychoric |
| Vehicle_Manoeuvre | -0.0002387 | -0.02492 | 1 | Polychoric | Polychoric | Polychoric | Polychoric | Polychoric | Polychoric |
| Junction_Location | 0.02083 | -0.07395 | -0.01621 | 1 | Polychoric | Polychoric | Polychoric | Polychoric | Polychoric |
| Skidding_and_Overturning | -0.04166 | 0.04207 | 0.1884 | -0.3517 | 1 | Polychoric | Polychoric | Polychoric | Polychoric |
| Vehicle_Leaving_Carriageway | -0.01657 | 0.2134 | 0.1747 | -0.3231 | 0.5779 | 1 | Polychoric | Polychoric | Polychoric |
| Hit_Object_off_Carriageway | -0.002944 | 0.2288 | 0.1715 | -0.3228 | 0.5603 | 0.82 | 1 | Polychoric | Polychoric |
| First_Point_of_Impact | 0.001565 | 0.1335 | -0.1197 | 0.09011 | -0.03133 | -0.01231 | -0.02277 | 1 | Polychoric |
| Journey_Purpose_of_Driver | -0.4197 | 0.4017 | 0.02994 | -0.0001691 | 0.04366 | 0.07318 | 0.0674 | 0.02702 | 1 |
| Sex_of_Driver | 0.009592 | 0.5627 | -0.05631 | 0.01144 | -0.08334 | -0.03529 | -0.02287 | 0.05845 | 0.1087 |
| Age_Band_of_Driver | 0.04776 | 0.01372 | -0.03167 | 0.03811 | -0.2182 | -0.1336 | -0.1329 | 0.01329 | -0.05337 |
| Engine_Capacity_.CC. | -0.001534 | <NA> | <NA> | <NA> | <NA> | <NA> | <NA> | <NA> | <NA> |
| Propulsion_Code | 0.06984 | -0.2603 | -0.05405 | -0.02052 | -0.1253 | -0.08334 | -0.07337 | -0.01089 | -0.3168 |
| Age_of_Vehicle | 0.08052 | 0.08892 | 0.03569 | -0.02998 | 0.1083 | 0.0761 | 0.08099 | -0.02441 | 0.0605 |
| Driver_Home_Area_Type | -0.001432 | 0.1439 | 0.03551 | -0.1765 | 0.1959 | 0.2033 | 0.1753 | -0.03117 | 0.05578 |
| make | -0.002381 | -0.01634 | 0.003451 | 0.01033 | 0.002128 | -0.01248 | -0.01346 | -0.003468 | 0.004285 |
| Casualty_Severity | -0.0005791 | 0.1906 | -0.1568 | 0.1349 | -0.136 | -0.1477 | -0.1167 | 0.07592 | -0.02147 |
| Casualty_Class | 0.001325 | 0.02776 | -0.01486 | -0.06183 | -0.221 | -0.2134 | -0.1907 | -0.0171 | -0.03668 |
| Sex_of_Casualty | 0.003519 | 0.3338 | -0.09318 | 0.02807 | -0.1154 | -0.06256 | -0.04742 | 0.01746 | 0.067 |
| Age_Band_of_Casualty | 0.03259 | -0.04194 | -0.08429 | 0.06065 | -0.1114 | -0.04221 | -0.04743 | -0.01055 | -0.03905 |
| Car_Passenger | -0.02309 | 0.452 | -0.02159 | 0.001501 | 0.07004 | 0.1191 | 0.1259 | 0.05592 | 0.1774 |
| Casualty_Type | -0.01949 | <NA> | -0.03572 | 0.0071 | 0.2421 | 0.3842 | 0.3891 | 0.08042 | 0.2545 |
| Number_of_Vehicles | 0.007111 | 0.05465 | -0.1223 | 0.3778 | -0.4359 | -0.489 | -0.5403 | 0.1546 | 0.04549 |
| Day_of_Week | -0.006714 | 0.001355 | -0.007633 | 0.01067 | -0.02236 | -0.02242 | -0.02169 | -0.002059 | 0.009038 |
| Hour_of_Day | -0.005675 | 0.04768 | -0.00317 | 0.04198 | -0.06403 | -0.07151 | -0.08015 | 0.01132 | 0.1513 |
| First_Road_Class | 1.519e-05 | 0.03081 | -0.03626 | 0.06381 | -0.09299 | -0.08433 | -0.0711 | -0.03867 | 0.01232 |
| Road_Type | -0.00185 | -0.01318 | -0.001496 | -0.09436 | -0.03544 | -0.01131 | -0.02464 | -0.09382 | 0.01823 |
| Speed_limit | -0.009137 | 0.1316 | 0.08359 | -0.3495 | 0.374 | 0.3662 | 0.3241 | -0.006281 | 0.01996 |
| Junction_Detail | 0.001874 | -0.05292 | -0.05899 | 0.7672 | -0.335 | -0.3055 | -0.3077 | 0.05227 | -0.004494 |
| Light_Conditions | -0.0006144 | 0.1084 | 0.109 | -0.1023 | 0.1978 | 0.2264 | 0.2276 | -0.03751 | 0.05495 |
| Weather_Conditions | -0.0242 | 0.0872 | 0.01586 | -0.07175 | 0.1607 | 0.1008 | 0.1079 | -0.01982 | -0.0062 |
| Road_Surface_Conditions | -0.03586 | 0.1212 | 0.03754 | -0.1338 | 0.2824 | 0.2035 | 0.2039 | -0.02861 | -0.02749 |
| Urban_or_Rural_Area | -0.003129 | 0.1655 | 0.08372 | -0.3967 | 0.4799 | 0.4762 | 0.4221 | -0.03583 | 0.03776 |
| month | 0.003385 | 0.004639 | 0.004445 | -0.00203 | 0.002648 | -0.00256 | 0.0002018 | -0.004999 | 0.01003 |
| other_vehic | 0.04437 | 0.1919 | 0.02279 | 0.13 | -0.1847 | -0.1668 | -0.2379 | -0.03415 | 0.005971 |

**Figura B.1:** Matriz de correlación

| | Sex_of_Driver | Age_Band_of_Driver | Engine_Capacity_.CC. | Propulsion_Code | Age_of_Vehicle | Driver_Home_Area_Type | make | Casualty_Severity | Casualty_Class | Sex_of_Casualty | Age_Band_of_Casualty |
|---|---|---|---|---|---|---|---|---|---|---|---|
| accyr | Polyserial | Polyserial | Pearson | Polyserial | Pearson | Polyserial | Polyserial | Polyserial | Polyserial | Polyserial | Polyserial |
| Vehicle_Type | Polychoric | Polychoric | Polyserial | Polychoric | Polyserial | Polychoric | Polychoric | Polychoric | Polychoric | Polychoric | Polychoric |
| Vehicle_Manoeuvre | Polychoric | Polychoric | Polyserial | Polychoric | Polyserial | Polychoric | Polychoric | Polychoric | Polychoric | Polychoric | Polychoric |
| Junction_Location | Polychoric | Polychoric | Polyserial | Polychoric | Polyserial | Polychoric | Polychoric | Polychoric | Polychoric | Polychoric | Polychoric |
| Skidding_and_Overturning | Polychoric | Polychoric | Polyserial | Polychoric | Polyserial | Polychoric | Polychoric | Polychoric | Polychoric | Polychoric | Polychoric |
| Vehicle_Leaving_Carriageway | Polychoric | Polychoric | Polyserial | Polychoric | Polyserial | Polychoric | Polychoric | Polychoric | Polychoric | Polychoric | Polychoric |
| Hit_Object_off_Carriageway | Polychoric | Polychoric | Polyserial | Polychoric | Polyserial | Polychoric | Polychoric | Polychoric | Polychoric | Polychoric | Polychoric |
| First_Point_of_Impact | Polychoric | Polychoric | Polyserial | Polychoric | Polyserial | Polychoric | Polychoric | Polychoric | Polychoric | Polychoric | Polychoric |
| Journey_Purpose_of_Driver | Polychoric | Polychoric | Polyserial | Polychoric | Polyserial | Polychoric | Polychoric | Polychoric | Polychoric | Polychoric | Polychoric |
| Sex_of_Driver | 1 | Polychoric | Polyserial | Polychoric | Polyserial | Polychoric | Polychoric | Polychoric | Polychoric | Polychoric | Polychoric |
| Age_Band_of_Driver | -0.01933 | 1 | Polyserial | Polychoric | Polyserial | Polychoric | Polychoric | Polychoric | Polychoric | Polychoric | Polychoric |
| Engine_Capacity_.CC. | <NA> | <NA> | 1 | Polyserial | Pearson | Polyserial | Polyserial | Polyserial | Polyserial | Polyserial | Polyserial |
| Propulsion_Code | -0.2058 | 0.1983 | <NA> | 1 | Polyserial | Polychoric | Polychoric | Polychoric | Polychoric | Polychoric | Polychoric |
| Age_of_Vehicle | -0.04538 | -0.05449 | -0.04312 | -0.2921 | 1 | Polyserial | Polyserial | Polyserial | Polyserial | Polyserial | Polyserial |
| Driver_Home_Area_Type | 0.1052 | 0.03292 | <NA> | 0.02587 | 0.03266 | 1 | Polychoric | Polychoric | Polychoric | Polychoric | Polychoric |
| make | -0.01264 | -0.01441 | <NA> | -0.04004 | 0.02446 | -0.01154 | 1 | Polychoric | Polychoric | Polychoric | Polychoric |
| Casualty_Severity | 0.1855 | -0.03895 | <NA> | 0.1028 | -0.05474 | -0.07139 | -0.0212 | 1 | Polychoric | Polychoric | Polychoric |
| Casualty_Class | -0.09639 | 0.07355 | <NA> | 0.1998 | -0.02844 | -0.1392 | -0.01528 | -0.1622 | 1 | Polychoric | Polychoric |
| Sex_of_Casualty | 0.828 | 0.07453 | <NA> | -0.0389 | -0.04193 | 0.05417 | -0.01931 | 0.1795 | 0.1788 | 1 | Polychoric |
| Age_Band_of_Casualty | -0.04602 | 0.641 | <NA> | 0.08415 | -0.04106 | 0.05517 | -0.007235 | -0.0612 | -0.2499 | 0.04048 | 1 |
| Car_Passenger | -0.02172 | -0.07167 | <NA> | 0.04104 | 0.03508 | -0.006508 | -0.001092 | 0.182 | 0.5897 | 0.2563 | -0.2824 |
| Casualty_Type | 0.3291 | -0.06177 | <NA> | -0.1694 | 0.07684 | 0.1904 | -0.01565 | 0.3403 | -0.7188 | 0.2386 | 0.1173 |
| Number_of_Vehicles | 0.1118 | 0.07252 | <NA> | -0.01717 | -0.03413 | -0.03068 | 0.01209 | 0.3133 | -0.5974 | 0.07279 | 0.1626 |
| Day_of_Week | -0.0004743 | 0.0008221 | 0.001932 | 0.004715 | -0.004926 | 0.00727 | -0.003348 | 0.00316 | 0.01758 | 0.002705 | 0.0008962 |
| Hour_of_Day | 0.00494 | -0.04281 | -0.04589 | -0.05339 | 0.01402 | -0.01645 | 0.002293 | 0.01311 | 0.04911 | 0.01585 | -0.0486 |
| First_Road_Class | 0.03351 | -0.03741 | <NA> | -0.03768 | 0.03503 | -0.01065 | 0.0105 | -0.004814 | 0.1373 | 0.01408 | -0.08052 |
| Road_Type | -0.006709 | 0.001885 | <NA> | -0.01687 | 0.02911 | 0.1173 | 0.007097 | -0.07545 | 0.07571 | -0.01464 | -0.02756 |
| Speed_limit | 0.01946 | -0.01269 | 0.006797 | 0.01372 | 0.0242 | 0.3005 | -0.01287 | -0.07298 | -0.2397 | -0.003086 | 0.04327 |
| Junction_Detail | 0.01766 | 0.04844 | <NA> | -0.0001394 | -0.03045 | -0.1339 | 0.007025 | 0.1148 | -0.004757 | 0.03704 | 0.05607 |
| Light_Conditions | -0.1369 | -0.2024 | <NA> | -0.004569 | 0.03728 | 0.03232 | -0.001427 | -0.07626 | 0.01048 | -0.1266 | -0.1192 |
| Weather_Conditions | 0.03012 | -0.05409 | <NA> | -0.003213 | 0.0133 | 0.04642 | -0.003784 | 0.06283 | -0.05103 | 0.02078 | -0.02079 |
| Road_Surface_Conditions | 0.04242 | -0.07852 | <NA> | -0.007764 | 0.02274 | 0.1098 | -0.002557 | 0.04758 | -0.09329 | 0.02504 | -0.02684 |
| Urban_or_Rural_Area | 0.04099 | -0.002336 | <NA> | -0.001053 | 0.05112 | 0.5959 | -0.01198 | -0.1129 | -0.2955 | 0.004417 | 0.0616 |
| month | 0.0008983 | 0.004677 | -0.003883 | 0.006869 | -0.006265 | 0.009913 | -0.0008874 | 0.001723 | 0.001727 | 0.001471 | 0.009686 |
| other_vehic | 0.06285 | 0.05938 | <NA> | 0.1034 | -0.01462 | 0.07418 | -0.0148 | 0.1116 | -0.2334 | 0.05233 | 0.07424 |

| | Car_Passenger | Casualty_Type | Number_of_Vehicles | Day_of_Week | Hour_of_Day | First_Road_Class | Road_Type | Speed_limit | Junction_Detail | Light_Conditions | Weather_Conditions | Road_Surface_Conditions |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| accyr | Polyserial | Polyserial | Polyserial | Pearson | Pearson | Polyserial | Polyserial | Pearson | Polyserial | Polyserial | Polyserial | Polyserial |
| Vehicle_Type | Polychoric | Polychoric | Polychoric | Polyserial | Polyserial | Polychoric | Polychoric | Polyserial | Polychoric | Polychoric | Polychoric | Polychoric |
| Vehicle_Manoeuvre | Polychoric | Polychoric | Polychoric | Polyserial | Polyserial | Polychoric | Polychoric | Polyserial | Polychoric | Polychoric | Polychoric | Polychoric |
| Junction_Location | Polychoric | Polychoric | Polychoric | Polyserial | Polyserial | Polychoric | Polychoric | Polyserial | Polychoric | Polychoric | Polychoric | Polychoric |
| Skidding_and_Overturning | Polychoric | Polychoric | Polychoric | Polyserial | Polyserial | Polychoric | Polychoric | Polyserial | Polychoric | Polychoric | Polychoric | Polychoric |
| Vehicle_Leaving_Carriageway | Polychoric | Polychoric | Polychoric | Polyserial | Polyserial | Polychoric | Polychoric | Polyserial | Polychoric | Polychoric | Polychoric | Polychoric |
| Hit_Object_off_Carriageway | Polychoric | Polychoric | Polychoric | Polyserial | Polyserial | Polychoric | Polychoric | Polyserial | Polychoric | Polychoric | Polychoric | Polychoric |
| First_Point_of_Impact | Polychoric | Polychoric | Polychoric | Polyserial | Polyserial | Polychoric | Polychoric | Polyserial | Polychoric | Polychoric | Polychoric | Polychoric |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Journey_Purpose_of_Driver | Polychoric | Polychoric | Polychoric | Polyserial | Polyserial | Polychoric | Polychoric | Polyserial | Polychoric | Polychoric | Polychoric | Polychoric |
| Sex_of_Driver | Polychoric | Polychoric | Polychoric | Polyserial | Polyserial | Polychoric | Polychoric | Polyserial | Polychoric | Polychoric | Polychoric | Polychoric |
| Age_Band_of_Driver | Polychoric | Polychoric | Polychoric | Polyserial | Polyserial | Polychoric | Polychoric | Polyserial | Polychoric | Polychoric | Polychoric | Polychoric |
| Engine_Capacity_.CC. | Polyserial | Polyserial | Polyserial | Pearson | Pearson | Polyserial | Polyserial | Pearson | Polyserial | Polyserial | Polyserial | Polyserial |
| Propulsion_Code | Polychoric | Polychoric | Polychoric | Polyserial | Polychoric | Polychoric | Polychoric | Polyserial | Polychoric | Polychoric | Polychoric | Polychoric |
| Age_of_Vehicle | Polyserial | Polyserial | Polyserial | Pearson | Pearson | Polyserial | Polyserial | Pearson | Polyserial | Polyserial | Polyserial | Polyserial |
| Driver_Home_Area_Type | Polychoric | Polychoric | Polychoric | Polyserial | Polyserial | Polychoric | Polychoric | Polyserial | Polychoric | Polychoric | Polychoric | Polychoric |
| make | Polychoric | Polychoric | Polychoric | Polyserial | Polyserial | Polychoric | Polychoric | Polyserial | Polychoric | Polychoric | Polychoric | Polychoric |
| Casualty_Severity | Polychoric | Polychoric | Polychoric | Polyserial | Polyserial | Polychoric | Polychoric | Polyserial | Polychoric | Polychoric | Polychoric | Polychoric |
| Casualty_Class | Polychoric | Polychoric | Polychoric | Polyserial | Polyserial | Polychoric | Polychoric | Polyserial | Polychoric | Polychoric | Polychoric | Polychoric |
| Sex_of_Casualty | Polychoric | Polychoric | Polychoric | Polyserial | Polyserial | Polychoric | Polychoric | Polyserial | Polychoric | Polychoric | Polychoric | Polychoric |
| Age_Band_of_Casualty | Polychoric | Polychoric | Polychoric | Polyserial | Polyserial | Polychoric | Polychoric | Polyserial | Polychoric | Polychoric | Polychoric | Polychoric |
| Car_Passenger | 1 | Polychoric | Polychoric | Polyserial | Polyserial | Polychoric | Polychoric | Polyserial | Polychoric | Polychoric | Polychoric | Polychoric |
| Casualty_Type | 0.6039 | 1 | Polychoric | Polyserial | Polyserial | Polychoric | Polychoric | Polyserial | Polychoric | Polychoric | Polychoric | Polychoric |
| Number_of_Vehicles | 0.1869 | 0.6011 | 1 | Polyserial | Polyserial | Polychoric | Polychoric | Polyserial | Polychoric | Polychoric | Polychoric | Polychoric |
| Day_of_Week | -0.00262 | -0.01502 | 0.007747 | 1 | Pearson | Polyserial | Polyserial | Pearson | Polyserial | Polyserial | Polyserial | Polyserial |
| Hour_of_Day | 0.05233 | 0.002758 | 0.06422 | 0.03459 | 1 | Polyserial | Polyserial | Pearson | Polyserial | Polyserial | Polyserial | Polyserial |
| First_Road_Class | -0.04998 | -0.1462 | -0.1014 | 0.01101 | 0.03838 | 1 | Polychoric | Polyserial | Polychoric | Polychoric | Polychoric | Polychoric |
| Road_Type | -0.05095 | -0.1049 | -0.08834 | 0.001721 | 0.01926 | 0.3616 | 1 | Polyserial | Polychoric | Polychoric | Polychoric | Polychoric |
| Speed_limit | 0.09448 | 0.3403 | -0.0347 | -0.02017 | -0.04601 | -0.4296 | -0.1906 | 1 | Polyserial | Polyserial | Polyserial | Polyserial |
| Junction_Detail | -0.003774 | -0.02895 | 0.3196 | 0.01301 | 0.03869 | 0.09579 | 0.1105 | -0.337 | 1 | Polychoric | Polychoric | Polychoric |
| Light_Conditions | 0.1212 | 0.09873 | -0.2022 | 0.0139 | 0.1949 | -0.04475 | 0.003613 | 0.1054 | -0.1094 | 1 | Polychoric | Polychoric |
| Weather_Conditions | 0.02914 | 0.1024 | -0.06319 | -0.0003494 | -0.002772 | 0.004286 | 0.000123 | 0.06291 | -0.05383 | 0.2223 | 1 | Polychoric |
| Road_Surface_Conditions | 0.03854 | 0.1605 | -0.1202 | -0.005806 | -0.06592 | 0.03523 | 0.04131 | 0.1275 | -0.1239 | 0.2784 | 0.6954 | 1 |
| Urban_or_Rural_Area | 0.1018 | 0.3943 | -0.05775 | -0.0242 | -0.05431 | -0.302 | -0.03347 | 0.7389 | -0.3666 | 0.1147 | 0.08327 | 0.1849 |
| month | 0.007033 | 0.003777 | 0.004931 | 0.002442 | 0.00685 | -0.003365 | -0.003005 | 0.005682 | -0.003337 | 0.06544 | 0.05908 | 0.06601 |
| other_vehic | 0.09835 | 0.4144 | <NA> | -0.0005498 | 0.04545 | -0.02549 | 0.04499 | 0.07172 | 0.1225 | -0.05514 | -0.004438 | -0.003059 |

| | Urban_or_Rural_Area | month | other_vehic |
|---|---|---|---|
| accyr | Polyserial | Pearson | Polyserial |
| Vehicle_Type | Polychoric | Polyserial | Polychoric |
| Vehicle_Manoeuvre | Polychoric | Polyserial | Polychoric |
| Junction_Location | Polychoric | Polyserial | Polychoric |
| Skidding_and_Overturning | Polychoric | Polyserial | Polychoric |
| Vehicle_Leaving_Carriageway | Polychoric | Polyserial | Polychoric |
| Hit_Object_off_Carriageway | Polychoric | Polyserial | Polychoric |
| First_Point_of_Impact | Polychoric | Polyserial | Polychoric |
| Journey_Purpose_of_Driver | Polychoric | Polyserial | Polychoric |
| Sex_of_Driver | Polychoric | Polyserial | Polychoric |
| Age_Band_of_Driver | Polychoric | Polyserial | Polychoric |
| Engine_Capacity_.CC. | Polyserial | Pearson | Polyserial |
| Propulsion_Code | Polychoric | Polyserial | Polychoric |
| Age_of_Vehicle | Polyserial | Pearson | Polyserial |
| Driver_Home_Area_Type | Polychoric | Polyserial | Polychoric |
| make | Polychoric | Polyserial | Polychoric |
| Casualty_Severity | Polychoric | Polyserial | Polychoric |

| | | | |
|---|---|---|---|
| Casualty_Class | Polychoric | Polyserial | Polychoric |
| Sex_of_Casualty | Polychoric | Polyserial | Polychoric |
| Age_Band_of_Casualty | Polychoric | Polyserial | Polychoric |
| Car_Passenger | Polychoric | Polyserial | Polychoric |
| Casualty_Type | Polychoric | Polyserial | Polychoric |
| Number_of_Vehicles | Polychoric | Polyserial | Polychoric |
| Day_of_Week | Polyserial | Pearson | Polyserial |
| Hour_of_Day | Polyserial | Pearson | Polyserial |
| First_Road_Class | Polychoric | Polyserial | Polychoric |
| Road_Type | Polychoric | Polyserial | Polychoric |
| Speed_limit | Polyserial | Pearson | Polyserial |
| Junction_Detail | Polychoric | Polyserial | Polychoric |
| Light_Conditions | Polychoric | Polyserial | Polychoric |
| Weather_Conditions | Polychoric | Polyserial | Polychoric |
| Road_Surface_Conditions | Polychoric | Polyserial | Polychoric |
| Urban_or_Rural_Area | 1 | Polyserial | Polychoric |
| month | 0.006507 | 1 | Polyserial |
| other_vehic | 0.1097 | -0.009629 | 1 |

# Apéndice: Script de preprocesamiento

```
require(corrplot)
require(ggplot2)
require(caret)
require(pROC)
library(ROCR)
require(arm)
library(doParallel)

# Lectura del set de datos final
datos <- read.table("datos_other_vehic.csv",header=T, dec=".", sep=",")

# Después de estimar los valores de correlación, eliminamos la variable 7
datos <-datos[,-7]

# Se elimina un accidente debido a que tiene indexados 2 vehículos sin embargo sólo
# aparece uno con índice 3

datos<-datos[datos$Acc_Index!=2014450012572,]

# Se eliminan los accidentes con más de 2 implicados para poder escoger el tipo de
# vehículo contrario y crear un nuevo campo

datos <- datos[datos$Number_of_Vehicles<3,]


# Se genera un subset para eliminar sexos anómalos distintos de 1 y 2.

datos <- datos[datos$Sex_of_Driver!=3,]

# Se eliminan vehículos con una edad >= 56 años

indices_a_eliminar<-subset(datos, Age_of_Vehicle>=56,
       select=c(Acc_Index))

for (i in indices_a_eliminar[,1])
  datos<-subset(datos, !grepl(i, datos$Acc_Index))


# Eliminamos los accidentes donde intervienen vehículos del tipo Vehicle_Type==X23
indices_a_eliminar<-subset(datos, Vehicle_Type=='X23',
                           select=c(Acc_Index))

for (i in indices_a_eliminar[,1])
  datos<-subset(datos, !grepl(i, datos$Acc_Index))

datos$Vehicle_Type<-as.character(datos$Vehicle_Type)
datos$Vehicle_Type<-as.factor(datos$Vehicle_Type)


######  Eliminamos vehículos Age_Band_of_Driver X1 y X2
datos<-datos[datos$Age_Band_of_Driver!=X1,]
datos<-datos[datos$Age_Band_of_Driver!=X2,]
```

```
# Eliminamos los accidentes donde el Age_Band_of_Driver==X1 o X2
indices_a_eliminar<-subset(datos, Age_Band_of_Driver=='X1' |
Age_Band_of_Driver=='X2',
                           select=c(Acc_Index))

for (i in indices_a_eliminar[,1])
  datos<-subset(datos, !grepl(i, datos$Acc_Index))

datos$Age_Band_of_Driver<-as.character(datos$Age_Band_of_Driver)
datos$Age_Band_of_Driver<-as.factor(datos$Age_Band_of_Driver)


# Eliminamos vehículos Propulsion_Code 10 y 12
datos<-datos[datos$Propulsion_Code!=X10,]
datos<-datos[datos$Propulsion_Code!=X12,]

indices_a_eliminar<-subset(datos, Propulsion_Code=='X10' |  Propulsion_Code=='X11',
                           select=c(Acc_Index))

for (i in indices_a_eliminar[,1])
  datos<-subset(datos, !grepl(i, datos4$Acc_Index))

datos$Propulsion_Code<-as.character(datos$Propulsion_Code)
datos$Propulsion_Code<-as.factor(datos$Propulsion_Code)

#  Eliminamos vehículos Casualty_Type 23
datos<-datos[datos$Casualty_Type!=X23,]

# Eliminamos los accidentes que intervengan coches con Casualty_Type==X23
indices_a_eliminar<-subset(datos, Casualty_Type=='X23',
                           select=c(Acc_Index))

for (i in indices_a_eliminar[,1])
  datos<-subset(datos, !grepl(i, datos$Acc_Index))

datos$Casualty_Type<-as.character(datos$Casualty_Type)
datos$Casualty_Type<-as.factor(datos$Casualty_Type)


# Eliminamos accidentes que han chocado contra X97 u X98
datos<-datos[datos$Acc_Index!='2014621400560',]
datos<-datos[datos$Acc_Index!='201443P019074',]
datos<-datos[datos$Acc_Index!='2014320038567',]
datos<-datos[datos$Acc_Index!='201442I284207',]

### Seleccionamos un subset de datos
datos2<-na.omit(subset(datos,select=c(Acc_Index,Vehicle_Index,
Vehicle_Type,Vehicle_Manoeuvre ,
                                      Junction_Location,
                                      Skidding_and_Overturning ,
                                      Hit_Object_off_Carriageway ,
                                      First_Point_of_Impact ,
                                      Journey_Purpose_of_Driver ,
                                      Sex_of_Driver ,
```

```
                                    Age_Band_of_Driver,
                                    Engine_Capacity_.CC.,
                                    Propulsion_Code ,
                                    Age_of_Vehicle ,
                                    Driver_Home_Area_Type ,
                                    make,
                                    Casualty_Severity,
                                    Casualty_Class ,
                                    Sex_of_Casualty ,
                                    Age_Band_of_Casualty ,
                                    Car_Passenger ,
                                    Casualty_Type ,
                                    Number_of_Vehicles ,
                                    Day_of_Week ,
                                    Hour_of_Day ,
                                    First_Road_Class ,
                                    Road_Type ,
                                    Speed_limit ,
                                    Junction_Detail,
                                    Light_Conditions ,
                                    Weather_Conditions ,
                                    Road_Surface_Conditions ,
                                    Urban_or_Rural_Area ,
                                    month,
                                    other_vehic)))

############################
### DATA SPLITTING ########
############################

set.seed(1)
trainingRows<-createDataPartition(datos2$Casualty_Severity,p=.80,list=FALSE)
training<-datos2[trainingRows,]
testing<-datos2[-trainingRows,]


#############################################################################
### eliminamos los accidentes de training que no se encuentran en testing ####
#############################################################################

# # Comprobamos todas las combinaciones
table(training$make,training$Casualty_Severity)
table(training$Vehicle_Type,training$Casualty_Severity)
table(training$Casualty_Type,training$Casualty_Severity)
table(training$other_vehic,training$Casualty_Severity)

# a. Nos quedamos con un vector con los nombres de las marcas a eliminar
tabla<-table(training$make,training$Casualty_Severity)
tabla2<-tabla[tabla[,'0']==0,]
marcas_elim<-rownames(tabla2[tabla2[,'1']==0,])
#eliminamos los espacios por si acaso el grepl no funciona bien
sin_esp<-gsub(" ", "", marcas_elim)
# Eliminamos las marcas que no aparecen en training del global datos2
for (marca in sin_esp){
```

```
  training<-subset(training, !grepl(marca, training$make))
  testing<-subset(testing, !grepl(marca, testing$make))
}

training$make<-as.character(training$make)
training$make<-as.factor(training$make)
testing$make<-as.character(testing$make)
testing$make<-as.factor(testing$make)

########## Convertimos en factor aquellas variables que deben permanecer como tal
########## para training

training$Vehicle_Type <- as.factor(training$Vehicle_Type)
levels(training$Vehicle_Type) <- make.names(levels(factor(training$Vehicle_Type)))

training$Vehicle_Manoeuvre <- as.factor(training$Vehicle_Manoeuvre)
levels(training$Vehicle_Manoeuvre) <-
make.names(levels(factor(training$Vehicle_Manoeuvre)))

training$Junction_Location <- as.factor(training$Junction_Location)
levels(training$Junction_Location) <-
make.names(levels(factor(training$Junction_Location)))

training$Skidding_and_Overturning <- as.factor(training$Skidding_and_Overturning)
levels(training$Skidding_and_Overturning) <-
make.names(levels(factor(training$Skidding_and_Overturning)))

training$Hit_Object_off_Carriageway <-
as.factor(training$Hit_Object_off_Carriageway)
levels(training$Hit_Object_off_Carriageway) <-
make.names(levels(factor(training$Hit_Object_off_Carriageway)))

training$First_Point_of_Impact <- as.factor(training$First_Point_of_Impact)
levels(training$First_Point_of_Impact) <-
make.names(levels(factor(training$First_Point_of_Impact)))

training$Journey_Purpose_of_Driver <- as.factor(training$Journey_Purpose_of_Driver)
levels(training$Journey_Purpose_of_Driver) <-
make.names(levels(factor(training$Journey_Purpose_of_Driver)))

training$Sex_of_Driver <- as.factor(training$Sex_of_Driver)
levels(training$Sex_of_Driver) <-
make.names(levels(factor(training$Sex_of_Driver)))

training$Age_Band_of_Driver <- as.factor(training$Age_Band_of_Driver)
levels(training$Age_Band_of_Driver) <-
make.names(levels(factor(training$Age_Band_of_Driver)))

training$Propulsion_Code <- as.factor(training$Propulsion_Code)
levels(training$Propulsion_Code) <-
make.names(levels(factor(training$Propulsion_Code)))

training$Driver_Home_Area_Type <- as.factor(training$Driver_Home_Area_Type)
```

```
levels(training$Driver_Home_Area_Type) <-
make.names(levels(factor(training$Driver_Home_Area_Type)))

training$make <- as.factor(training$make)
levels(training$make) <- make.names(levels(factor(training$make)))

training$Casualty_Class <- as.factor(training$Casualty_Class)
levels(training$Casualty_Class) <-
make.names(levels(factor(training$Casualty_Class)))

training$Sex_of_Casualty <- as.factor(training$Sex_of_Casualty)
levels(training$Sex_of_Casualty) <-
make.names(levels(factor(training$Sex_of_Casualty)))

training$Age_Band_of_Casualty <- as.factor(training$Age_Band_of_Casualty)
levels(training$Age_Band_of_Casualty) <-
make.names(levels(factor(training$Age_Band_of_Casualty)))

training$Casualty_Severity <- as.factor(training$Casualty_Severity)
levels(training$Casualty_Severity) <-
make.names(levels(factor(training$Casualty_Severity)))

training$Car_Passenger <- as.factor(training$Car_Passenger)
levels(training$Car_Passenger) <-
make.names(levels(factor(training$Car_Passenger)))

training$Casualty_Type <- as.factor(training$Casualty_Type)
levels(training$Casualty_Type) <-
make.names(levels(factor(training$Casualty_Type)))

training$Number_of_Vehicles <- as.factor(training$Number_of_Vehicles)
levels(training$Number_of_Vehicles) <-
make.names(levels(factor(training$Number_of_Vehicles)))

training$First_Road_Class <- as.factor(training$First_Road_Class)
levels(training$First_Road_Class) <-
make.names(levels(factor(training$First_Road_Class)))

training$Road_Type <- as.factor(training$Road_Type)
levels(training$Road_Type) <- make.names(levels(factor(training$Road_Type)))

training$Junction_Detail <- as.factor(training$Junction_Detail)
levels(training$Junction_Detail) <-
make.names(levels(factor(training$Junction_Detail)))

training$Light_Conditions <- as.factor(training$Light_Conditions)
levels(training$Light_Conditions) <-
make.names(levels(factor(training$Light_Conditions)))

training$Weather_Conditions <- as.factor(training$Weather_Conditions)
levels(training$Weather_Conditions) <-
make.names(levels(factor(training$Weather_Conditions)))

training$Road_Surface_Conditions <- as.factor(training$Road_Surface_Conditions)
```

```
levels(training$Road_Surface_Conditions) <-
make.names(levels(factor(training$Road_Surface_Conditions)))

training$Urban_or_Rural_Area <- as.factor(training$Urban_or_Rural_Area)
levels(training$Urban_or_Rural_Area) <-
make.names(levels(factor(training$Urban_or_Rural_Area)))

training$other_vehic <- as.factor(training$other_vehic)
levels(training$other_vehic) <- make.names(levels(factor(training$other_vehic)))

training <-training[,-1:-2]
```

# Apéndice: Script de selección de predictores

**Script selectores Curva ROC de cada predictor, Relief, permuteRelief, RFE-RF**

```
########################
### Método filterVarImp
########################

#Eliminamos las columnas de indices que no nos sirven
datos <-datos[,-1:-2]
rocValues<-filterVarImp(x=datos[,-15],
                        y=datos$Casualty_Severity)



########################
### Método Relief
########################
reliefValues<-attrEval(Casualty_Severity ~ ., data=datos,
                       estimator="ReliefFequalK",
                       ReliefIterations=33)

reliefValues
perm<-permuteRelief(x=datos[,-15],
                    y=datos$Casualty_Severity,
                    nperm=500,
                    estimator="ReliefFequalK",
                    ReliefIterations=33)
perm$permutations
histogram(~ value|Predictor,
          data=perm$permutations)

################################################
# Recursive Feature Elimination using Random Forest
### Matriz de predictores sin Casualty_Severity
trainX <- training[,-15]

# Puesto que rfe no puede manejar factores con más de 53 categorias, vamos a usar
# el valor devuelto de marcas por BayesGLM después de la primera ejecución
trainX$make <- training2$make

set.seed(123)
control <- rfeControl(functions = rfFuncs,
                      method = "repeatedcv",
                      repeats = 5,
                      verbose = FALSE)

#The simulation will fit models with subset sizes of 21 to 32
subsets <- c(21:32)

system.time(
  results <- rfe(trainX, training$Casualty_Severity, subsets, rfeControl=control)
```

### Script Selector ReliefF mediante Weka

```
=== Run information ===

Evaluator:    weka.attributeSelection.ReliefFAttributeEval -M -1 -D 1 -K 10
Search:       weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation:
datos_other_vehic_filtrada-weka.filters.unsupervised.attribute.Remove-R1-2-weka.fil
ters.unsupervised.attribute.NumericToNominal-R1-9-weka.filters.unsupervised.attribu
te.NumericToNominal-R11-weka.filters.unsupervised.attribute.NumericToNominal-R13-21
-weka.filters.unsupervised.attribute.NumericToNominal-R24-25-weka.filters.unsupervi
sed.attribute.NumericToNominal-R27-31-weka.filters.unsupervised.attribute.NumericTo
Nominal-R33
Instances:    424626
Attributes:   33
              Vehicle_Type
              Vehicle_Manoeuvre
              Junction_Location
              Skidding_and_Overturning
              Hit_Object_off_Carriageway
              First_Point_of_Impact
              Journey_Purpose_of_Driver
              Sex_of_Driver
              Age_Band_of_Driver
              Engine_Capacity_.CC.
              Propulsion_Code
              Age_of_Vehicle
              Driver_Home_Area_Type
              make
              Casualty_Severity
              Casualty_Class
              Sex_of_Casualty
              Age_Band_of_Casualty
              Car_Passenger
              Casualty_Type
              Number_of_Vehicles
              Day_of_Week
              Hour_of_Day
              First_Road_Class
              Road_Type
              Speed_limit
              Junction_Detail
              Light_Conditions
              Weather_Conditions
              Road_Surface_Conditions
              Urban_or_Rural_Area
              month
              other_vehic
Evaluation mode:    evaluate on all training data


=== Attribute Selection on all input data ===
```

```
Search Method:
       Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 15 Casualty_Severity):
       ReliefF Ranking Filter
       Instances sampled: all
       Number of nearest neighbours (k): 10
       Equal influence nearest neighbours

Ranked attributes:
 0.093911    6 First_Point_of_Impact
 0.087387   14 make
 0.08479     9 Age_Band_of_Driver
 0.081715    2 Vehicle_Manoeuvre
 0.073194    3 Junction_Location
 0.070507   18 Age_Band_of_Casualty
 0.070362   27 Junction_Detail
 0.067388   24 First_Road_Class
 0.061769    7 Journey_Purpose_of_Driver
 0.041965   25 Road_Type
 0.036141   11 Propulsion_Code
 0.035092   30 Road_Surface_Conditions
 0.031431   28 Light_Conditions
 0.029465   33 other_vehic
 0.029322    5 Hit_Object_off_Carriageway
 0.027237   22 Day_of_Week
 0.026542   29 Weather_Conditions
 0.02472    32 month
 0.023867    4 Skidding_and_Overturning
 0.021963   13 Driver_Home_Area_Type
 0.021495    8 Sex_of_Driver
 0.02129    20 Casualty_Type
 0.021108   31 Urban_or_Rural_Area
 0.020002    1 Vehicle_Type
 0.019605   23 Hour_of_Day
 0.017298   26 Speed_limit
 0.013947   21 Number_of_Vehicles
 0.009883   17 Sex_of_Casualty
 0.005775   12 Age_of_Vehicle
 0.000741   10 Engine_Capacity_.CC.
-0.000713   19 Car_Passenger
-0.002136   16 Casualty_Class

Selected attributes:
6,14,9,2,3,18,27,24,7,25,11,30,28,33,5,22,29,32,4,13,8,20,31,1,23,26,21,17,12,10,19
,16 : 32
```

**Script obtención de curva ROC por inserción acumulativa de características**

Entrenamiento sin obtener ROC acumulativa por introducción de predictores en el orden que los proporciona el selector:

```
set.seed(2014)
mod1 <- train(Casualty_Severity ~ ., data = training,
              ## 'modelInfo' is a list object found in the linked
              ## source code
              method = modelInfo,
              ## Minimize the distance to the perfect model
              metric = "Dist",
              maximize = FALSE,
              tuneLength = 20,
              trControl = trainControl(method = "cv",
                                       classProbs = TRUE,
                                       summaryFunction = fourStats))
```

```
print(mod1)
prediction <- predict(mod1, testing)
confusionMatrix(prediction,testing$Casualty_Severity)

mod1.probs <- predict(mod1, testing,type = "prob")
mod1.ROC <- roc(response = testing$Casualty_Severity,
                predictor = mod1.probs$X0,
                levels = levels(testing$Casualty_Severity))
plot(mod1.ROC, add=TRUE, col="yellow")
```

**Script de entrenamiento para obtener AUC ROC acumulativa por introducción de predictores en el orden que los proporciona el selector**

```
set.seed(2014)
variables_RFE_RF<-c("MANIOBRAS", "EDAD", "SUPERFICIE_CALZADA", "HORA","POSICION",
"FACTORES_ATMOSFERICOS", "TIPO_VEHICULO","ZONA_AGRUPADA", "MES", "LUMINOSIDAD",
"TIPO_VIA", "DIASEMANA", "SEXO", "NUMERO_OCUPANTES_VEH")
```

```
for (i in 1:14){
system.time(
  mod2 <- train(Casualty_Severity ~ ., data = carmetry_train2[ ,
c(variables_RFE_RF[1:i], "Casualty_Severity")],
                ## 'modelInfo' is a list object found in the linked
                ## source code
                method = modelInfo,
                ## Minimize the distance to the perfect model
                metric = "Dist",
                maximize = FALSE,
                tuneLength = 5,
                trControl = trainControl(method = "cv",
                                         classProbs = TRUE,
                                         summaryFunction = fourStats))
  )
```

```
  print(mod2)
}
```

UNIVERSITAS
Miguel
Hernández

# Apéndice: Salida del selector RFE-RF para España añadiendo predictores de forma incremental

UNIVERSITAS
Miguel
Hernández

**Salida del selector RFE-RF para España añadiendo predictores de forma incremental**

```
674975 samples
     1 predictor
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 607477, 607476, 607477, 607478, 607478, 607477, ...
Resampling results across tuning parameters:

  threshold  ROC  Sens  Spec  Dist
  0.010      0.5  0     1     1
  0.255      0.5  0     1     1
  0.500      0.5  0     1     1
  0.745      0.5  0     1     1
  0.990      0.5  0     1     1

Tuning parameter 'mtry' was held constant at a value of 6
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 6 and threshold = 0.01.
rf

674975 samples
     2 predictor
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 607477, 607477, 607478, 607478, 607478, 607478, ...
Resampling results across tuning parameters:

  threshold  ROC        Sens          Spec       Dist
  0.010      0.5006619  0.0006996892  0.9996325  0.9993004
  0.255      0.5006619  0.0000000000  0.9999984  1.0000000
  0.500      0.5006619  0.0000000000  1.0000000  1.0000000
  0.745      0.5006619  0.0000000000  1.0000000  1.0000000
  0.990      0.5006619  0.0000000000  1.0000000  1.0000000

Tuning parameter 'mtry' was held constant at a value of 6
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 6 and threshold = 0.01.
rf

674975 samples
     3 predictor
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 607477, 607477, 607477, 607477, 607478, 607478, ...
Resampling results across tuning parameters:

  threshold  ROC        Sens         Spec       Dist
```

```
  0.010      0.5016761  0.001977367  0.9988460  0.9980233
  0.255      0.5016761  0.000000000  0.9999984  1.0000000
  0.500      0.5016761  0.000000000  1.0000000  1.0000000
  0.745      0.5016761  0.000000000  1.0000000  1.0000000
  0.990      0.5016761  0.000000000  1.0000000  1.0000000

Tuning parameter 'mtry' was held constant at a value of 6
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 6 and threshold = 0.01.
rf

674975 samples
     4 predictor
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 607477, 607477, 607478, 607477, 607478, 607478, ...
Resampling results across tuning parameters:

  threshold  ROC        Sens        Spec       Dist
  0.010      0.5127387  0.01387195  0.9937518  0.9861479
  0.255      0.5127387  0.00000000  0.9999938  1.0000000
  0.500      0.5127387  0.00000000  1.0000000  1.0000000
  0.745      0.5127387  0.00000000  1.0000000  1.0000000
  0.990      0.5127387  0.00000000  1.0000000  1.0000000

Tuning parameter 'mtry' was held constant at a value of 6
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 6 and threshold = 0.01.
rf

674975 samples
     5 predictor
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 607477, 607476, 607478, 607478, 607478, 607477, ...
Resampling results across tuning parameters:

  threshold  ROC        Sens          Spec       Dist
  0.010      0.5571742  7.593067e-02  0.9762795  0.9243739
  0.255      0.5571742  7.605349e-04  0.9999097  0.9992395
  0.500      0.5571742  6.084576e-05  0.9999891  0.9999392
  0.745      0.5571742  0.000000e+00  1.0000000  1.0000000
  0.990      0.5571742  0.000000e+00  1.0000000  1.0000000

Tuning parameter 'mtry' was held constant at a value of 7
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 7 and threshold = 0.01.
rf

674975 samples
```

```
     6 predictor
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 607477, 607478, 607478, 607477, 607478, 607477, ...
Resampling results across tuning parameters:

  threshold  ROC        Sens        Spec       Dist
  0.010      0.5626168  0.09190191  0.9702493  0.9085854
  0.255      0.5626168  0.00121686  0.9998972  0.9987831
  0.500      0.5626168  0.00000000  0.9999953  1.0000000
  0.745      0.5626168  0.00000000  1.0000000  1.0000000
  0.990      0.5626168  0.00000000  1.0000000  1.0000000

Tuning parameter 'mtry' was held constant at a value of 7
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 7 and threshold = 0.01.
rf

674975 samples
     7 predictor
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 607476, 607478, 607477, 607477, 607478, 607478, ...
Resampling results across tuning parameters:

  threshold  ROC        Sens          Spec       Dist
  0.010      0.5843394  0.1405148628  0.9509923  0.8608830
  0.255      0.5843394  0.0029508341  0.9997835  0.9970492
  0.500      0.5843394  0.0001825373  0.9999860  0.9998175
  0.745      0.5843394  0.0000000000  1.0000000  1.0000000
  0.990      0.5843394  0.0000000000  1.0000000  1.0000000

Tuning parameter 'mtry' was held constant at a value of 9
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 9 and threshold = 0.01.
rf

674975 samples
     8 predictor
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 607478, 607478, 607477, 607477, 607478, 607477, ...
Resampling results across tuning parameters:

  threshold  ROC        Sens          Spec       Dist
  0.010      0.6183383  1.953642e-01  0.9409223  0.8068025
  0.255      0.6183383  4.715333e-03  0.9996496  0.9952847
  0.500      0.6183383  9.126863e-05  0.9999891  0.9999087
```

```
  0.745      0.6183383  0.000000e+00  1.0000000  1.0000000
  0.990      0.6183383  0.000000e+00  1.0000000  1.0000000

Tuning parameter 'mtry' was held constant at a value of 9
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 9 and threshold = 0.01.
rf

674975 samples
     9 predictor
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 607477, 607478, 607477, 607477, 607478, 607478, ...
Resampling results across tuning parameters:

  threshold  ROC        Sens         Spec       Dist
  0.010      0.6737921  0.317838919  0.8985101  0.6896743
  0.255      0.6737921  0.005779949  0.9996293  0.9942201
  0.500      0.6737921  0.000182528  0.9999922  0.9998175
  0.745      0.6737921  0.000000000  1.0000000  1.0000000
  0.990      0.6737921  0.000000000  1.0000000  1.0000000

Tuning parameter 'mtry' was held constant at a value of 9
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 9 and threshold = 0.01.
rf

674975 samples
    10 predictor
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 607478, 607478, 607477, 607478, 607477, 607477, ...
Resampling results across tuning parameters:

  threshold  ROC        Sens          Spec       Dist
  0.010      0.6859189  0.3645355108  0.8771272  0.6472407
  0.255      0.6859189  0.0078790535  0.9994425  0.9921211
  0.500      0.6859189  0.0001216915  0.9999907  0.9998783
  0.745      0.6859189  0.0000000000  1.0000000  1.0000000
  0.990      0.6859189  0.0000000000  1.0000000  1.0000000

Tuning parameter 'mtry' was held constant at a value of 9
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 9 and threshold = 0.01.

674975 samples
    11 predictor
     2 classes: 'X0', 'X1'

No pre-processing
```

```
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 607477, 607477, 607477, 607477, 607478, 607478, ...
Resampling results across tuning parameters:

  threshold  ROC        Sens          Spec       Dist
  0.010      0.6942193  0.3893286204  0.8698184  0.6243959
  0.255      0.6942193  0.0098260252  0.9993054  0.9901742
  0.500      0.6942193  0.0005171519  0.9999798  0.9994828
  0.745      0.6942193  0.0000000000  1.0000000  1.0000000
  0.990      0.6942193  0.0000000000  1.0000000  1.0000000

Tuning parameter 'mtry' was held constant at a value of 9
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 9 and threshold = 0.01.

674975 samples
    12 predictor
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 607478, 607478, 607478, 607478, 607478, 607476, ...
Resampling results across tuning parameters:

  threshold  ROC        Sens          Spec       Dist
  0.010      0.7345369  0.5265883093  0.8134910  0.5088293
  0.255      0.7345369  0.0122595778  0.9990843  0.9877409
  0.500      0.7345369  0.0006084205  0.9999891  0.9993916
  0.745      0.7345369  0.0000000000  1.0000000  1.0000000
  0.990      0.7345369  0.0000000000  1.0000000  1.0000000

Tuning parameter 'mtry' was held constant at a value of 9
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 9 and threshold = 0.01.

674975 samples
    13 predictor
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 607478, 607478, 607478, 607478, 607477, 607477, ...
Resampling results across tuning parameters:

  threshold  ROC        Sens          Spec       Dist
  0.010      0.7389465  0.5581949407  0.7906660  0.4889072
  0.255      0.7389465  0.0137199315  0.9990469  0.9862805
  0.500      0.7389465  0.0005171797  0.9999860  0.9994828
  0.745      0.7389465  0.0000000000  1.0000000  1.0000000
  0.990      0.7389465  0.0000000000  1.0000000  1.0000000

Tuning parameter 'mtry' was held constant at a value of 9
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 9 and threshold = 0.01.
```

```
674975 samples
    14 predictor
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 607477, 607477, 607477, 607478, 607477, 607478, ...
Resampling results across tuning parameters:

  threshold  ROC        Sens          Spec       Dist
  0.010      0.7274284  0.5240632104  0.8038025  0.5147993
  0.255      0.7274284  0.0098867784  0.9994471  0.9901134
  0.500      0.7274284  0.0006388342  0.9999875  0.9993612
  0.745      0.7274284  0.0000000000  1.0000000  1.0000000
  0.990      0.7274284  0.0000000000  1.0000000  1.0000000

Tuning parameter 'mtry' was held constant at a value of 12
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 12 and threshold = 0.01.
```

# Apéndice: Script de creación de los modelos

### Script BayesGLM con inserción de predictores de forma recursiva

```
for (i in 1:31){
  system.time(
    bayesFit2 <- train(Casualty_Severity ~ .,
       data = training2[ , c(weights[1:i, 2], "Casualty_Severity")],
       method = "bayesglm", metric='ROC', trControl = cv.ctrl)
  )
  print(bayesFit2)
}
```

### Script para obtener modelo mediante Random Forest

```
###################################
#####   Modelo RandomForest #######
###################################
set.seed(2014)

modelInfo <- list(label = "rf",
                  library = c("randomForest"),
                  type = c("Classification"),
                  parameters  = data.frame(parameter = c("mtry", "threshold"),
                                           class = c("numeric", "numeric"),
                                           label = c("#Randomly Selected Predictors",
                                                     "Probability Cutoff")),
                  grid = function(x, y, len = NULL,search = NULL) {
                    p <- ncol(x)
                    expand.grid(mtry = floor(sqrt(p)),
                                threshold = seq(.01, .99, length = len))
                  },
                  loop = function(grid) {
                    library(plyr)
                    loop <- ddply(grid, c("mtry"),
                                  function(x) c(threshold = max(x$threshold)))
                    submodels <- vector(mode = "list", length = nrow(loop))
                    for(i in seq(along = loop$threshold)) {
                      index <- which(grid$mtry == loop$mtry[i])
                      cuts <- grid[index, "threshold"]
                      submodels[[i]] <- data.frame(threshold = cuts[cuts !=
loop$threshold[i]])
                    }
                    list(loop = loop, submodels = submodels)
                  },
                  fit = function(x, y, wts, param, lev, last, classProbs, ...) {
                    if(length(levels(y)) != 2)
                      stop("This works only for 2-class problems")
                    randomForest(x, y, mtry = param$mtry, ...)
                  },
                  predict = function(modelFit, newdata, submodels = NULL) {
                    class1Prob <- predict(modelFit,
                                          newdata,
                                          type = "prob")[, modelFit$obsLevels[1]]
                    ## Raise the threshold for class #1 and a higher level of
                    ## evidence is needed to call it class 1 so it should
                    ## decrease sensitivity and increase specificity
                    out <- ifelse(class1Prob >= modelFit$tuneValue$threshold,
                                  modelFit$obsLevels[1],
                                  modelFit$obsLevels[2])
```

```
    if(!is.null(submodels))
    {
      tmp2 <- out
      out <- vector(mode = "list", length = length(submodels$threshold))
      out[[1]] <- tmp2
      for(i in seq(along = submodels$threshold)) {
        out[[i+1]] <- ifelse(class1Prob >= submodels$threshold[[i]],
                             modelFit$obsLevels[1],
                             modelFit$obsLevels[2])
      }
    }
    out
  },
  prob = function(modelFit, newdata, submodels = NULL) {
    out <- as.data.frame(predict(modelFit, newdata, type = "prob"))
    if(!is.null(submodels))
    {
      probs <- out
      out <- vector(mode = "list", length = length(submodels$threshold)+1)
      out <- lapply(out, function(x) probs)
    }
    out
  },
  predictors = function(x, ...) {
    ## After doing some testing, it looks like randomForest
    ## will only try to split on plain main effects (instead
    ## of interactions or terms like I(x^2).
    varIndex <- as.numeric(names(table(x$forest$bestvar)))
    varIndex <- varIndex[varIndex > 0]
    varsUsed <- names(x$forest$ncat)[varIndex]
    varsUsed
  },
  varImp = function(object, ...){
    varImp <- randomForest::importance(object, ...)
    if(object$type == "regression")
      varImp <- data.frame(Overall = varImp[,"%IncMSE"])
    else {
      retainNames <- levels(object$y)
      if(all(retainNames %in% colnames(varImp))) {
        varImp <- varImp[, retainNames]
      } else {
        varImp <- data.frame(Overall = varImp[,1])
      }
    }

    out <- as.data.frame(varImp)
    if(dim(out)[2] == 2) {
      tmp <- apply(out, 1, mean)
      out[,1] <- out[,2] <- tmp
    }
    out
  },
  levels = function(x) x$classes,
  tags = c("Random Forest", "Ensemble Model", "Bagging", "Implicit Feature
Selection"),
  sort = function(x) x[order(x[,1]),])



fourStats <- function (data, lev = levels(data$obs), model = NULL) {
  ## This code will get use the area under the ROC curve and the
```

```
  ## sensitivity and specificity values using the current candidate
  ## value of the probability threshold.
  out <- c(twoClassSummary(data, lev = levels(data$obs), model = NULL))

  ## The best possible model has sensitivity of 1 and specifity of 1.
  ## How far are we from that value?
  coords <- matrix(c(1, 1, out["Spec"], out["Sens"]),
                   ncol = 2,
                   byrow = TRUE)
  colnames(coords) <- c("Spec", "Sens")
  rownames(coords) <- c("Best", "Current")
  c(out, Dist = dist(coords)[1])
}
```

Entrenamiento de entrenamiento sin obtener ROC acumulativa por introducción de predictores en el orden que los proporciona el selector:

```
set.seed(2014)
modl <- train(Casualty_Severity ~ ., data = training,
              ## 'modelInfo' is a list object found in the linked
              ## source code
              method = modelInfo,
              ## Minimize the distance to the perfect model
              metric = "Dist",
              maximize = FALSE,
              tuneLength = 20,
              trControl = trainControl(method = "cv",
                                       classProbs = TRUE,
                                       summaryFunction = fourStats))


print(modl)
prediction <- predict(modl, testing)
confusionMatrix(prediction,testing$Casualty_Severity)

modl.probs <- predict(modl, testing,type = "prob")
modl.ROC <- roc(response = testing$Casualty_Severity,
                predictor = modl.probs$X0,
                levels = levels(testing$Casualty_Severity))
plot(modl.ROC, add=TRUE, col="yellow")
```

Script de entrenamiento para obtener ROC acumulativa por introducción de predictores en el orden que los proporciona el selector:

```
set.seed(2014)
variables_RFE_RF<-c("MANIOBRAS", "EDAD", "SUPERFICIE_CALZADA", "HORA","POSICION",
"FACTORES_ATMOSFERICOS", "TIPO_VEHICULO","ZONA_AGRUPADA", "MES", "LUMINOSIDAD",
"TIPO_VIA", "DIASEMANA", "SEXO", "NUMERO_OCUPANTES_VEH")


for (i in 1:14){
system.time(
  mod2 <- train(Casualty_Severity ~ ., data = carmetry_train2[ ,
c(variables_RFE_RF[1:i], "Casualty_Severity")],
```

```
                ## 'modelInfo' is a list object found in the linked
                ## source code
                method = modelInfo,
                ## Minimize the distance to the perfect model
                metric = "Dist",
                maximize = FALSE,
                tuneLength = 5,
                trControl = trainControl(method = "cv",
                                         classProbs = TRUE,
                                         summaryFunction = fourStats))
  )
  print(mod2)
}
```

# Apéndice: Selección de características mediante BayesGLM

## Selección de características mediante BayesGLM

### Ejecución 5

#### Tiempo de ejecución

```
    user    system   elapsed
 7750.028   525.004 48337.748
```

#### Resumen del modelo

```
> summary(bayesFit2)

Call:
NULL

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3402   0.2168   0.3539   0.5655   2.0858

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                     4.372e+00  7.864e-02  55.592  < 2e-16 ***
Vehicle_TypeX11                -4.410e-01  7.911e-02  -5.575 2.48e-08 ***
Vehicle_TypeX19                -4.928e-01  6.213e-02  -7.931 2.18e-15 ***
Vehicle_TypeX20                -5.155e-01  1.222e-01  -4.219 2.46e-05 ***
Vehicle_TypeX21                -1.348e+00  1.223e-01 -11.025  < 2e-16 ***
Vehicle_TypeX8                 -3.934e-01  6.646e-02  -5.920 3.22e-09 ***
Vehicle_TypeX9                 -2.760e-01  5.400e-02  -5.112 3.19e-07 ***
Vehicle_TypeX90                -1.081e+00  1.297e-01  -8.331  < 2e-16 ***
Vehicle_ManoeuvreX10            4.583e-01  9.006e-02   5.089 3.60e-07 ***
Vehicle_ManoeuvreX15            2.694e-01  7.302e-02   3.690 0.000224 ***
Vehicle_ManoeuvreX16           -2.690e-01  2.836e-02  -9.486  < 2e-16 ***
Vehicle_ManoeuvreX17           -1.457e-01  2.807e-02  -5.192 2.08e-07 ***
Vehicle_ManoeuvreX18           -1.466e-01  2.074e-02  -7.071 1.54e-12 ***
Vehicle_ManoeuvreX3             7.127e-01  6.121e-02  11.643  < 2e-16 ***
Vehicle_ManoeuvreX4             4.793e-01  3.616e-02  13.256  < 2e-16 ***
Vehicle_ManoeuvreX5             3.310e-01  4.079e-02   8.115 4.87e-16 ***
Vehicle_ManoeuvreX7             1.799e-01  4.317e-02   4.167 3.09e-05 ***
Vehicle_ManoeuvreX8             8.118e-01  1.759e-01   4.616 3.90e-06 ***
Vehicle_ManoeuvreX9             1.321e-01  3.101e-02   4.259 2.06e-05 ***
Junction_LocationX1            -5.901e-02  1.815e-02  -3.251 0.001149 **
Junction_LocationX2            -1.548e-01  2.515e-02  -6.156 7.47e-10 ***
Junction_LocationX4            -9.886e-02  5.395e-02  -1.832 0.066900 .
Junction_LocationX6            -1.647e-01  3.539e-02  -4.654 3.26e-06 ***
Skidding_and_OverturningX1     -1.769e-01  1.629e-02 -10.860  < 2e-16 ***
Skidding_and_OverturningX2     -1.744e-01  2.581e-02  -6.758 1.40e-11 ***
Skidding_and_OverturningX3      5.234e-01  2.823e-01   1.854 0.063748 .
Skidding_and_OverturningX5     -2.451e-01  2.951e-02  -8.303  < 2e-16 ***
Vehicle_Leaving_CarriagewayX1 -3.965e-01  2.220e-02 -17.856  < 2e-16 ***
Vehicle_Leaving_CarriagewayX2 -3.244e-01  4.058e-02  -7.993 1.31e-15 ***
Vehicle_Leaving_CarriagewayX3 -7.555e-01  5.650e-02 -13.372  < 2e-16 ***
```

```
Vehicle_Leaving_CarriagewayX4 -3.299e-01  6.689e-02  -4.932 8.16e-07 ***
Vehicle_Leaving_CarriagewayX5 -4.504e-01  8.023e-02  -5.614 1.98e-08 ***
Vehicle_Leaving_CarriagewayX6 -7.881e-01  1.143e-01  -6.896 5.36e-12 ***
Vehicle_Leaving_CarriagewayX7 -4.776e-01  2.556e-02 -18.686  < 2e-16 ***
Vehicle_Leaving_CarriagewayX8 -4.216e-01  5.179e-02  -8.140 3.94e-16 ***
Hit_Object_off_CarriagewayX1  -1.315e-01  4.278e-02  -3.075 0.002107 **
Hit_Object_off_CarriagewayX10 -1.691e-01  2.695e-02  -6.274 3.52e-10 ***
Hit_Object_off_CarriagewayX11 -2.281e-01  5.847e-02  -3.901 9.57e-05 ***
Hit_Object_off_CarriagewayX2  -3.202e-01  4.360e-02  -7.343 2.08e-13 ***
Hit_Object_off_CarriagewayX3  -1.715e-01  6.240e-02  -2.749 0.005978 **
Hit_Object_off_CarriagewayX4  -6.974e-01  2.909e-02 -23.970  < 2e-16 ***
Hit_Object_off_CarriagewayX6   2.399e-01  6.582e-02   3.645 0.000268 ***
Hit_Object_off_CarriagewayX8  -9.797e-01  2.456e-01  -3.990 6.62e-05 ***
Hit_Object_off_CarriagewayX9   1.180e-01  4.077e-02   2.893 0.003818 **
First_Point_of_ImpactX1       -1.360e-01  1.197e-02 -11.361  < 2e-16 ***
First_Point_of_ImpactX2        6.684e-01  2.819e-02  23.709  < 2e-16 ***
Journey_Purpose_of_DriverX15  -2.053e-01  2.085e-02  -9.849  < 2e-16 ***
Journey_Purpose_of_DriverX2   -9.774e-02  2.441e-02  -4.004 6.22e-05 ***
Journey_Purpose_of_DriverX5   -3.983e-01  4.326e-02  -9.207  < 2e-16 ***
Journey_Purpose_of_DriverX6   -1.757e-01  1.946e-02  -9.025  < 2e-16 ***
Sex_of_DriverX2                1.137e-01  1.521e-02   7.474 7.77e-14 ***
Engine_Capacity_.CC.          -3.009e-05  7.372e-06  -4.082 4.46e-05 ***
Propulsion_CodeX2              1.063e-01  1.531e-02   6.945 3.79e-12 ***
Propulsion_CodeX8              3.097e-01  1.053e-01   2.941 0.003277 **
Age_of_Vehicle                -1.091e-02  1.086e-03 -10.047  < 2e-16 ***
Driver_Home_Area_TypeX2       -3.496e-02  1.837e-02  -1.903 0.057062 .
Driver_Home_Area_TypeX3       -5.313e-02  1.691e-02  -3.142 0.001676 **
makeAJP.....................  -2.768e+00  1.577e+00  -1.756 0.079104 .
makeAJS.....................  -6.042e-01  2.048e-01  -2.951 0.003169 **
makeAVIA....................  -3.206e+00  1.765e+00  -1.817 0.069239 .
makeBARON...................  -3.386e+00  1.783e+00  -1.899 0.057525 .
makeBUELL...................  -4.720e-01  2.445e-01  -1.930 0.053610 .
makeCADILLAC................  -1.675e+00  7.978e-01  -2.099 0.035804 *
makeDERBI...................  -6.403e-01  2.021e-01  -3.169 0.001532 **
makeHARTFORD................  -1.405e+00  7.493e-01  -1.875 0.060755 .
makeHINO....................   2.885e+00  1.511e+00   1.909 0.056230 .
makeHUONIAO.................  -5.161e-01  2.666e-01  -1.936 0.052872 .
makeHUSABERG................  -1.152e+00  6.458e-01  -1.784 0.074401 .
makeKTM.....................  -3.618e-01  1.123e-01  -3.221 0.001278 **
makeLIFAN...................  -7.932e-01  2.214e-01  -3.583 0.000340 ***
makeLML.....................  -7.843e-01  3.095e-01  -2.534 0.011287 *
makeMETROCAB................   1.030e+00  5.658e-01   1.821 0.068631 .
makeMOTOR.HISPANIA..........  -9.280e-01  4.059e-01  -2.286 0.022237 *
makeMV.AGUSTA...............  -1.245e+00  5.270e-01  -2.362 0.018188 *
makePROTON..................   4.071e-01  2.010e-01   2.025 0.042824 *
makeSANBEN..................  -1.099e+00  4.836e-01  -2.273 0.023048 *
makeSCAMMELL................  -2.202e+00  9.973e-01  -2.208 0.027231 *
makeSMC.....................  -1.343e+00  4.281e-01  -3.136 0.001711 **
makeSUBARU..................  -4.237e-01  9.884e-02  -4.286 1.82e-05 ***
makeWANGYE..................  -1.176e+00  5.515e-01  -2.132 0.033042 *
makeWESTFIELD...............  -2.650e+00  7.680e-01  -3.450 0.000560 ***
makeZHENHUA.................  -1.982e+00  1.052e+00  -1.884 0.059527 .
Casualty_ClassX3              -1.999e+00  2.212e-02 -90.370  < 2e-16 ***
Sex_of_CasualtyX2              1.007e-01  1.386e-02   7.268 3.65e-13 ***
```

| | | | | | |
|---|---|---|---|---|---|
| Age_Band_of_CasualtyX10 | -7.642e-01 | 2.506e-02 | -30.491 | < 2e-16 | *** |
| Age_Band_of_CasualtyX11 | -1.164e+00 | 2.400e-02 | -48.479 | < 2e-16 | *** |
| Age_Band_of_CasualtyX3 | -8.185e-02 | 2.740e-02 | -2.988 | 0.002810 | ** |
| Age_Band_of_CasualtyX6 | -6.216e-02 | 1.710e-02 | -3.634 | 0.000279 | *** |
| Age_Band_of_CasualtyX7 | -1.478e-01 | 1.827e-02 | -8.085 | 6.20e-16 | *** |
| Age_Band_of_CasualtyX8 | -2.961e-01 | 1.894e-02 | -15.630 | < 2e-16 | *** |
| Age_Band_of_CasualtyX9 | -4.882e-01 | 2.183e-02 | -22.368 | < 2e-16 | *** |
| Car_PassengerX2 | -8.823e-02 | 2.458e-02 | -3.590 | 0.000331 | *** |
| Casualty_TypeX2 | -1.813e+00 | 6.711e-02 | -27.013 | < 2e-16 | *** |
| Casualty_TypeX21 | 8.414e-01 | 1.155e-01 | 7.287 | 3.18e-13 | *** |
| Casualty_TypeX3 | -2.027e+00 | 5.891e-02 | -34.414 | < 2e-16 | *** |
| Casualty_TypeX4 | -2.121e+00 | 6.395e-02 | -33.169 | < 2e-16 | *** |
| Casualty_TypeX5 | -2.364e+00 | 5.673e-02 | -41.675 | < 2e-16 | *** |
| Number_of_VehiclesX2 | 6.006e-02 | 1.672e-02 | 3.592 | 0.000328 | *** |
| Hour_of_Day | 1.136e-02 | 1.020e-03 | 11.131 | < 2e-16 | *** |
| First_Road_ClassX3 | -3.518e-01 | 3.294e-02 | -10.681 | < 2e-16 | *** |
| First_Road_ClassX4 | -3.436e-01 | 3.664e-02 | -9.376 | < 2e-16 | *** |
| First_Road_ClassX5 | -1.956e-01 | 3.803e-02 | -5.143 | 2.70e-07 | *** |
| First_Road_ClassX6 | -2.577e-01 | 3.602e-02 | -7.155 | 8.34e-13 | *** |
| Road_TypeX6 | -9.844e-02 | 1.783e-02 | -5.523 | 3.34e-08 | *** |
| Road_TypeX9 | 2.965e-01 | 1.160e-01 | 2.555 | 0.010623 | * |
| Speed_limit | -1.272e-02 | 6.202e-04 | -20.515 | < 2e-16 | *** |
| Junction_DetailX1 | 5.267e-01 | 3.101e-02 | 16.983 | < 2e-16 | *** |
| Junction_DetailX2 | 3.206e-01 | 6.678e-02 | 4.801 | 1.58e-06 | *** |
| Junction_DetailX3 | 2.454e-01 | 1.801e-02 | 13.624 | < 2e-16 | *** |
| Junction_DetailX5 | 2.642e-01 | 5.062e-02 | 5.220 | 1.79e-07 | *** |
| Junction_DetailX6 | 2.926e-01 | 2.279e-02 | 12.836 | < 2e-16 | *** |
| Junction_DetailX7 | 2.733e-01 | 5.544e-02 | 4.929 | 8.26e-07 | *** |
| Junction_DetailX8 | 2.638e-01 | 3.500e-02 | 7.536 | 4.84e-14 | *** |
| Junction_DetailX9 | 3.185e-01 | 4.388e-02 | 7.257 | 3.96e-13 | *** |
| Light_ConditionsX4 | -3.717e-01 | 1.488e-02 | -24.974 | < 2e-16 | *** |
| Light_ConditionsX5 | -2.279e-01 | 7.540e-02 | -3.023 | 0.002504 | ** |
| Light_ConditionsX6 | -3.282e-01 | 2.055e-02 | -15.976 | < 2e-16 | *** |
| Light_ConditionsX7 | -2.758e-01 | 5.348e-02 | -5.157 | 2.51e-07 | *** |
| Weather_ConditionsX2 | 1.700e-01 | 2.107e-02 | 8.071 | 7.00e-16 | *** |
| Weather_ConditionsX5 | 1.186e-01 | 4.566e-02 | 2.597 | 0.009407 | ** |
| Weather_ConditionsX7 | 1.489e-01 | 6.794e-02 | 2.192 | 0.028367 | * |
| Weather_ConditionsX8 | 2.080e-01 | 4.163e-02 | 4.996 | 5.84e-07 | *** |
| Weather_ConditionsX9 | 1.756e-01 | 6.653e-02 | 2.640 | 0.008303 | ** |
| Road_Surface_ConditionsX2 | 2.979e-02 | 1.554e-02 | 1.916 | 0.055343 | . |
| Road_Surface_ConditionsX3 | 5.847e-01 | 6.592e-02 | 8.870 | < 2e-16 | *** |
| Road_Surface_ConditionsX4 | 4.303e-01 | 3.688e-02 | 11.667 | < 2e-16 | *** |
| Road_Surface_ConditionsX5 | 2.201e-01 | 1.215e-01 | 1.811 | 0.070144 | . |
| Urban_or_Rural_AreaX2 | -2.714e-01 | 1.751e-02 | -15.495 | < 2e-16 | *** |
| month | 4.627e-03 | 1.590e-03 | 2.910 | 0.003616 | ** |
| other_vehicX10 | -6.491e-01 | 2.990e-01 | -2.171 | 0.029931 | * |
| other_vehicX11 | -8.227e-01 | 1.229e-01 | -6.694 | 2.17e-11 | *** |
| other_vehicX17 | -1.298e+00 | 5.074e-01 | -2.558 | 0.010531 | * |
| other_vehicX19 | -6.042e-01 | 6.294e-02 | -9.600 | < 2e-16 | *** |
| other_vehicX20 | -8.431e-01 | 2.042e-01 | -4.128 | 3.65e-05 | *** |
| other_vehicX21 | -1.377e+00 | 1.084e-01 | -12.702 | < 2e-16 | *** |
| other_vehicX4 | 6.436e-01 | 3.300e-01 | 1.951 | 0.051113 | . |
| other_vehicX8 | -3.309e-01 | 1.118e-01 | -2.961 | 0.003071 | ** |
| other_vehicX9 | -3.074e-01 | 1.942e-02 | -15.828 | < 2e-16 | *** |

```
other_vehicX90                -7.209e-01  3.216e-01  -2.242 0.024991 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 267828  on 340010  degrees of freedom
Residual deviance: 226027  on 339872  degrees of freedom
AIC: 226305

Number of Fisher Scoring iterations: 11
```

# Apéndice: Modelo Random Forest a partir de inserción recursiva de variables

UNIVERSITAS
Miguel
Hernández

**Modelo Random Forest a partir de inserción recursiva de variables**

```
Loading required package: randomForest
randomForest 4.6-12
Type rfNews() to see new features/changes/bug fixes.

339701 samples
     1 predictors
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 305731, 305731, 305731, 305731, 305731, 305732, ...
Resampling results across tuning parameters:

  threshold  ROC        Sens          Spec       Dist
  0.010      0.5167088  0.0349286394  0.9884583  0.9651407
  0.255      0.5167088  0.0071704815  0.9976006  0.9928325
  0.500      0.5167088  0.0029253780  0.9989600  0.9970752
  0.745      0.5167088  0.0007038333  0.9997451  0.9992962
  0.990      0.5167088  0.0000000000  1.0000000  1.0000000

Tuning parameter 'mtry' was held constant at a value of 1
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 1 and threshold = 0.01.
rf

339701 samples
     2 predictors
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 305730, 305731, 305730, 305731, 305731, 305731, ...
Resampling results across tuning parameters:

  threshold  ROC        Sens         Spec       Dist
  0.010      0.5391448  0.079557389  0.9822150  0.9206147
  0.255      0.5391448  0.005894804  0.9992149  0.9941055
  0.500      0.5391448  0.000000000  1.0000000  1.0000000
  0.745      0.5391448  0.000000000  1.0000000  1.0000000
  0.990      0.5391448  0.000000000  1.0000000  1.0000000
```

```
Tuning parameter 'mtry' was held constant at a value of 1
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 1 and threshold = 0.01.
rf

339701 samples
     3 predictors
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 305731, 305731, 305732, 305730, 305731, 305731, ...
Resampling results across tuning parameters:

  threshold  ROC        Sens          Spec       Dist
  0.010      0.5769561  0.1671872289  0.9450953  0.8346223
  0.255      0.5769561  0.0174424506  0.9968801  0.9825627
  0.500      0.5769561  0.0009457801  0.9998131  0.9990542
  0.745      0.5769561  0.0000000000  1.0000000  1.0000000
  0.990      0.5769561  0.0000000000  1.0000000  1.0000000

Tuning parameter 'mtry' was held constant at a value of 1
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 1 and threshold = 0.01.
rf

339701 samples
     4 predictors
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 305731, 305731, 305730, 305730, 305731, 305731, ...
Resampling results across tuning parameters:

  threshold  ROC        Sens         Spec       Dist
  0.010      0.5785465  0.149260614  0.9544925  0.8519565
  0.255      0.5785465  0.022897133  0.9969480  0.9771078
  0.500      0.5785465  0.003563243  0.9997077  0.9964368
  0.745      0.5785465  0.000000000  1.0000000  1.0000000
  0.990      0.5785465  0.000000000  1.0000000  1.0000000

Tuning parameter 'mtry' was held constant at a value of 3
Dist was used to select the optimal model using  the smallest value.
```

```
The final values used for the model were mtry = 3 and threshold = 0.01.
rf

339701 samples
     5 predictors
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 305731, 305730, 305731, 305731, 305731, 305731, ...
Resampling results across tuning parameters:

  threshold  ROC        Sens         Spec         Dist
  0.010      0.6007683  0.237616758  0.9084785    0.7678584
  0.255      0.6007683  0.027736190  0.9961867    0.9722714
  0.500      0.6007683  0.006532611  0.9994222    0.9934676
  0.745      0.6007683  0.000000000  1.0000000    1.0000000
  0.990      0.6007683  0.000000000  1.0000000    1.0000000


Tuning parameter 'mtry' was held constant at a
 value of 4
Dist was used to select the optimal model using
 the smallest value.
The final values used for the model were mtry = 4
 and threshold = 0.01.
rf

339701 samples
     6 predictors
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 305731, 305731, 305730, 305731, 305732, 305730, ...
Resampling results across tuning parameters:

  threshold  ROC        Sens          Spec       Dist
  0.010      0.6290704  0.3896933248  0.8140377  0.6380269
  0.255      0.6290704  0.0447388477  0.9932571  0.9552851
  0.500      0.6290704  0.0081162906  0.9992251  0.9918840
  0.745      0.6290704  0.0001099771  0.9999966  0.9998900
  0.990      0.6290704  0.0000000000  1.0000000  1.0000000
```

```
Tuning parameter 'mtry' was held constant at a value of 5
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 5 and threshold = 0.01.
rf

339701 samples
     7 predictors
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 305732, 305731, 305731, 305730, 305731, 305731, ...
Resampling results across tuning parameters:

  threshold  ROC        Sens         Spec       Dist
  0.010      0.6400499  0.447299254  0.7691113  0.5990051
  0.255      0.6400499  0.055142685  0.9917074  0.9448939
  0.500      0.6400499  0.007170520  0.9993441  0.9928297
  0.745      0.6400499  0.000109982  1.0000000  0.9998900
  0.990      0.6400499  0.000000000  1.0000000  1.0000000

Tuning parameter 'mtry' was held constant at a value of 5
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 5 and threshold = 0.01.
rf

339701 samples
     8 predictors
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 305731, 305732, 305731, 305730, 305731, 305731, ...
Resampling results across tuning parameters:

  threshold  ROC       Sens          Spec       Dist
  0.010      0.677573  5.312113e-01  0.7315701  0.5402212
  0.255      0.677573  6.367659e-02  0.9899095  0.9363781
  0.500      0.677573  6.840603e-03  0.9993101  0.9931597
  0.745      0.677573  8.797493e-05  1.0000000  0.9999120
  0.990      0.677573  0.000000e+00  1.0000000  1.0000000

Tuning parameter 'mtry' was held constant at a value of 5
Dist was used to select the optimal model using  the smallest value.
```

```
The final values used for the model were mtry = 5 and threshold = 0.01.
rf

339701 samples
     9 predictors
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 305731, 305732, 305731, 305731, 305731, 305731, ...
Resampling results across tuning parameters:

  threshold  ROC        Sens         Spec       Dist
  0.010      0.6916015  0.639714703  0.6425025  0.5075877
  0.255      0.6916015  0.103092528  0.9811886  0.8971052
  0.500      0.6916015  0.016738583  0.9981512  0.9832632
  0.745      0.6916015  0.001077774  0.9999626  0.9989222
  0.990      0.6916015  0.000000000  1.0000000  1.0000000

Tuning parameter 'mtry' was held constant at a value of 6
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 6 and threshold = 0.01.
rf

339701 samples
    10 predictors
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 305731, 305731, 305731, 305731, 305732, 305730, ...
Resampling results across tuning parameters:

  threshold  ROC        Sens         Spec       Dist
  0.010      0.6940533  0.662369967  0.6216010  0.5071384
  0.255      0.6940533  0.112770631  0.9796083  0.8874638
  0.500      0.6940533  0.018828192  0.9978963  0.9811741
  0.745      0.6940533  0.001187732  0.9999456  0.9988123
  0.990      0.6940533  0.000000000  1.0000000  1.0000000

Tuning parameter 'mtry' was held constant at a value of 6
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 6 and threshold = 0.01.
rf
```

```
339701 samples
    11 predictors
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 305732, 305731, 305730, 305731, 305731, 305732, ...
Resampling results across tuning parameters:

  threshold  ROC        Sens          Spec       Dist
  0.010      0.7071344  0.7462609857  0.5515010  0.5153208
  0.255      0.7071344  0.1244941647  0.9768044  0.8758136
  0.500      0.7071344  0.0186301484  0.9978895  0.9813722
  0.745      0.7071344  0.0009677871  0.9999728  0.9990322
  0.990      0.7071344  0.0000000000  1.0000000  1.0000000

Tuning parameter 'mtry' was held constant at a value of 6
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 6 and threshold = 0.01.
rf

339701 samples
    12 predictors
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 305730, 305731, 305730, 305731, 305731, 305732, ...
Resampling results across tuning parameters:

  threshold  ROC        Sens         Spec       Dist
  0.010      0.7249004  0.858239605  0.4319953  0.5854591
  0.255      0.7249004  0.157619224  0.9696469  0.8429287
  0.500      0.7249004  0.025954621  0.9972267  0.9740494
  0.745      0.7249004  0.001627679  0.9999456  0.9983723
  0.990      0.7249004  0.000000000  1.0000000  1.0000000

Tuning parameter 'mtry' was held constant at a value of 6
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 6 and threshold
 = 0.01.
rf
```

```
339701 samples
    13 predictors
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 305732, 305730, 305731, 305731, 305730, 305731, ...
Resampling results across tuning parameters:

  threshold  ROC        Sens         Spec       Dist
  0.010      0.7220849  0.853180406  0.4328687  0.5858507
  0.255      0.7220849  0.148315100  0.9722163  0.8521383
  0.500      0.7220849  0.022721193  0.9974884  0.9772821
  0.745      0.7220849  0.001209773  0.9999830  0.9987902
  0.990      0.7220849  0.000000000  1.0000000  1.0000000

Tuning parameter 'mtry' was held constant at a value of 6
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 6 and threshold
 = 0.01.
rf

339701 samples
    14 predictors
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 305730, 305732, 305731, 305732, 305731, 305730, ...
Resampling results across tuning parameters:

  threshold  ROC        Sens         Spec       Dist
  0.010      0.7381146  0.925655390  0.3261385  0.6779729
  0.255      0.7381146  0.195912855  0.9592200  0.8051210
  0.500      0.7381146  0.036402487  0.9956396  0.9636074
  0.745      0.7381146  0.002309515  0.9999150  0.9976905
  0.990      0.7381146  0.000000000  1.0000000  1.0000000

Tuning parameter 'mtry' was held constant at a value of 7
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 7 and threshold
 = 0.01.
rf
```

```
339701 samples
    15 predictors
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 305732, 305731, 305731, 305731, 305730, 305732, ...
Resampling results across tuning parameters:

  threshold  ROC        Sens        Spec       Dist
  0.010      0.7487681  0.95506329  0.2460058  0.7553387
  0.255      0.7487681  0.25750048  0.9410407  0.7448388
  0.500      0.7487681  0.04245112  0.9950346  0.9575618
  0.745      0.7487681  0.00217755  0.9998810  0.9978225
  0.990      0.7487681  0.00000000  1.0000000  1.0000000

Tuning parameter 'mtry' was held constant at a value of 8
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 8 and threshold
 = 0.255.
rf

339701 samples
    16 predictors
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 305731, 305730, 305730, 305731, 305731, 305732, ...
Resampling results across tuning parameters:

  threshold  ROC        Sens         Spec       Dist
  0.010      0.7517388  0.960694065  0.2276634  0.7733458
  0.255      0.7517388  0.263395343  0.9404800  0.7390059
  0.500      0.7517388  0.046498400  0.9946948  0.9535165
  0.745      0.7517388  0.002485494  0.9998505  0.9975145
  0.990      0.7517388  0.000000000  1.0000000  1.0000000

Tuning parameter 'mtry' was held constant at a value of 8
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 8 and threshold = 0.255.
rf

339701 samples
```

```
    17 predictors
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 305730, 305732, 305731, 305732, 305731, 305730, ...
Resampling results across tuning parameters:

  threshold  ROC        Sens         Spec       Dist
  0.010      0.7567785  0.971032113  0.1929227  0.8076014
  0.255      0.7567785  0.274964672  0.9371459  0.7277554
  0.500      0.7567785  0.046212279  0.9948069  0.9538019
  0.745      0.7567785  0.002375463  0.9998946  0.9976245
  0.990      0.7567785  0.000000000  1.0000000  1.0000000

Tuning parameter 'mtry' was held constant at a value of 8
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 8 and threshold = 0.255.
rf

339701 samples
    18 predictors
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 305732, 305730, 305732, 305730, 305730, 305731, ...
Resampling results across tuning parameters:

  threshold  ROC        Sens         Spec       Dist
  0.010      0.7574966  0.974485360  0.1740400  0.8263560
  0.255      0.7574966  0.302569211  0.9277827  0.7011622
  0.500      0.7574966  0.053998624  0.9934237  0.9460243
  0.745      0.7574966  0.003695203  0.9997893  0.9963048
  0.990      0.7574966  0.000000000  1.0000000  1.0000000

Tuning parameter 'mtry' was held constant at a value of 9
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 9 and threshold = 0.255.
rf

339701 samples
    19 predictors
     2 classes: 'X0', 'X1'
```

```
No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 305730, 305730, 305732, 305732, 305732, 305730, ...
Resampling results across tuning parameters:

  threshold  ROC        Sens         Spec       Dist
  0.010      0.7620886  0.982733640  0.1404412  0.8597329
  0.255      0.7620886  0.318823635  0.9217977  0.6856530
  0.500      0.7620886  0.057606051  0.9929853  0.9424202
  0.745      0.7620886  0.004795018  0.9997485  0.9952050
  0.990      0.7620886  0.000000000  1.0000000  1.0000000

Tuning parameter 'mtry' was held constant at a value of 9
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 9 and threshold = 0.255.
rf

339701 samples
    20 predictors
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 305731, 305730, 305731, 305731, 305731, 305731, ...
Resampling results across tuning parameters:

  threshold  ROC        Sens        Spec       Dist
  0.010      0.7632708  0.98378931  0.1340280  0.8661254
  0.255      0.7632708  0.32671970  0.9192624  0.6781074
  0.500      0.7632708  0.05852976  0.9928391  0.9414975
  0.745      0.7632708  0.00483894  0.9997077  0.9951611
  0.990      0.7632708  0.00000000  1.0000000  1.0000000

Tuning parameter 'mtry' was held constant at a value of 9
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 9 and threshold = 0.255.
rf

339701 samples
    21 predictors
     2 classes: 'X0', 'X1'

No pre-processing
```

```
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 305731, 305730, 305731, 305730, 305731, 305732, ...
Resampling results across tuning parameters:

  threshold  ROC        Sens         Spec       Dist
  0.010      0.7656845  0.987242644  0.1139660  0.8861282
  0.255      0.7656845  0.339235134  0.9152282  0.6661845
  0.500      0.7656845  0.060685246  0.9922545  0.9393469
  0.745      0.7656845  0.005476844  0.9996262  0.9945232
  0.990      0.7656845  0.000000000  1.0000000  1.0000000

Tuning parameter 'mtry' was held constant at a value of 9
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 9 and threshold = 0.255.
rf

339701 samples
    22 predictors
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 305730, 305731, 305730, 305731, 305732, 305730, ...
Resampling results across tuning parameters:

  threshold  ROC        Sens         Spec       Dist
  0.010      0.7676175  0.988694392  0.1063870  0.8936853
  0.255      0.7676175  0.337123988  0.9176888  0.6679682
  0.500      0.7676175  0.057870164  0.9927032  0.9421581
  0.745      0.7676175  0.005124988  0.9996941  0.9948751
  0.990      0.7676175  0.000000000  1.0000000  1.0000000

Tuning parameter 'mtry' was held constant at a value of 9
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 9 and threshold = 0.255.
rf

339701 samples
    23 predictors
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 305731, 305730, 305731, 305730, 305731, 305731, ...
```

```
Resampling results across tuning parameters:

  threshold  ROC        Sens         Spec        Dist
  0.010      0.7688288  0.990190082  0.09892704  0.9011274
  0.255      0.7688288  0.355314503  0.91055169  0.6508649
  0.500      0.7688288  0.064160770  0.99197926  0.9358736
  0.745      0.7688288  0.006598608  0.99962275  0.9934015
  0.990      0.7688288  0.000000000  1.00000000  1.0000000

Tuning parameter 'mtry' was held constant at a value of 10
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 10 and threshold = 0.255.
rf

339701 samples
    24 predictors
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 305731, 305731, 305730, 305730, 305732, 305732, ...
Resampling results across tuning parameters:

  threshold  ROC        Sens         Spec        Dist
  0.010      0.7720257  0.991003912  0.09586491  0.9041806
  0.255      0.7720257  0.343656497  0.91848749  0.6613905
  0.500      0.7720257  0.062005049  0.99247205  0.9380253
  0.745      0.7720257  0.005894785  0.99968053  0.9941053
  0.990      0.7720257  0.000000000  1.00000000  1.0000000

Tuning parameter 'mtry' was held constant at a value of 10
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 10 and threshold = 0.255.
rf

339701 samples
    25 predictors
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 305731, 305731, 305732, 305732, 305731, 305730, ...
Resampling results across tuning parameters:
```

```
  threshold  ROC        Sens         Spec        Dist
  0.010      0.7735766  0.991157816  0.09250366  0.9075403
  0.255      0.7735766  0.342908452  0.91919777  0.6620438
  0.500      0.7735766  0.063698637  0.99236670  0.9363327
  0.745      0.7735766  0.005828768  0.99968053  0.9941713
  0.990      0.7735766  0.000000000  1.00000000  1.0000000

Tuning parameter 'mtry' was held constant at a value of 11
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 11 and threshold = 0.255.
rf

339701 samples
    26 predictors
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 305730, 305731, 305731, 305731, 305732, 305731, ...
Resampling results across tuning parameters:

  threshold  ROC        Sens         Spec       Dist
  0.010      0.7906044  9.885404e-01  0.1175889  0.8824878
  0.255      0.7906044  4.113364e-01  0.9055319  0.5961992
  0.500      0.7906044  1.060398e-01  0.9883733  0.8940360
  0.745      0.7906044  9.018119e-03  0.9995038  0.9909820
  0.990      0.7906044  4.399472e-05  1.0000000  0.9999560

Tuning parameter 'mtry' was held constant at a value of 12
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 12 and threshold = 0.255.
rf

339701 samples
    27 predictors
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 305730, 305731, 305730, 305731, 305732, 305731, ...
Resampling results across tuning parameters:

  threshold  ROC        Sens          Spec       Dist
  0.010      0.7911435  9.876605e-01  0.1262825  0.8738052
```

```
  0.255      0.7911435  3.990189e-01  0.9115815  0.6074532
  0.500      0.7911435  1.058860e-01  0.9889851  0.8941820
  0.745      0.7911435  7.720396e-03  0.9995718  0.9922797
  0.990      0.7911435  2.199252e-05  1.0000000  0.9999780

Tuning parameter 'mtry' was held constant at a value of 12
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 12 and threshold = 0.255.
rf

339701 samples
    28 predictors
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 305730, 305731, 305731, 305731, 305731, 305731, ...
Resampling results across tuning parameters:

  threshold  ROC        Sens          Spec       Dist
  0.010      0.7925346  9.888043e-01  0.1185745  0.8814977
  0.255      0.7925346  3.995468e-01  0.9111974  0.6069866
  0.500      0.7925346  1.005190e-01  0.9903139  0.8995335
  0.745      0.7925346  7.874334e-03  0.9996771  0.9921257
  0.990      0.7925346  2.199252e-05  1.0000000  0.9999780

Tuning parameter 'mtry' was held constant at a value of 12
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 12 and threshold
 = 0.255.
rf

339701 samples
    29 predictors
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 305731, 305731, 305731, 305732, 305731, 305730, ...
Resampling results across tuning parameters:

  threshold  ROC        Sens         Spec       Dist
  0.010      0.7935291  0.989882114  0.1111349  0.8889241
  0.255      0.7935291  0.406364847  0.9094709  0.6005075
```

```
  0.500      0.7935291  0.100519045  0.9902901  0.8995335
  0.745      0.7935291  0.007324473  0.9996601  0.9926756
  0.990      0.7935291  0.000000000  1.0000000  1.0000000

Tuning parameter 'mtry' was held constant at a value of 12
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 12 and threshold
 = 0.255.

339701 samples
    30 predictors
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 305732, 305731, 305730, 305731, 305730, 305730, ...
Resampling results across tuning parameters:

  threshold  ROC        Sens          Spec       Dist
  0.010      0.7927882  9.908279e-01  0.1037327  0.8963149
  0.255      0.7927882  4.060577e-01  0.9094098  0.6008157
  0.500      0.7927882  9.288653e-02  0.9907931  0.9071604
  0.745      0.7927882  6.862552e-03  0.9997451  0.9931375
  0.990      0.7927882  4.398988e-05  1.0000000  0.9999560

Tuning parameter 'mtry' was held constant at a value of 12
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 12 and threshold = 0.255.
rf

339701 samples
    31 predictors
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 305731, 305732, 305732, 305730, 305730, 305731, ...
Resampling results across tuning parameters:

  threshold  ROC        Sens          Spec        Dist
  0.010      0.7923381  9.913997e-01  0.09727193  0.9027698
  0.255      0.7923381  3.999429e-01  0.91076922  0.6066583
  0.500      0.7923381  9.064315e-02  0.99154423  0.9093963
  0.745      0.7923381  6.290727e-03  0.99976550  0.9937093
```

```
  0.990      0.7923381  6.598241e-05  1.00000000  0.9999340

Tuning parameter 'mtry' was held constant at a value of 12
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 12 and threshold = 0.255.

339701 samples
    32 predictors
     2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 305731, 305730, 305732, 305730, 305731, 305731, ...
Resampling results across tuning parameters:

  threshold  ROC        Sens         Spec        Dist
  0.010      0.7918988  0.991157855  0.09971214  0.9003328
  0.255      0.7918988  0.388285794  0.91485097  0.6176147
  0.500      0.7918988  0.089081637  0.99178553  0.9109555
  0.745      0.7918988  0.005498832  0.99975190  0.9945012
  0.990      0.7918988  0.000000000  1.00000000  1.0000000

Tuning parameter 'mtry' was held constant at a value of 13
Dist was used to select the optimal model using  the smallest value.
The final values used for the model were mtry = 13 and threshold = 0.255.
```

# Apéndice: Resumen estadístico del modelo BayesGLM del set de datos de España

**Resumen estadístico del modelo BayesGLM del set de datos de España**

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.7187   0.1238   0.1888   0.3196   2.3633

Coefficients:
                      Estimate Std. Error  z value Pr(>|z|)
(Intercept)           3.646457   0.051539   70.752  < 2e-16 ***
EDAD                 -0.015898   0.000377  -42.167  < 2e-16 ***
ANOMALIA_NINGUNAX3   -2.603823   0.022109 -117.771  < 2e-16 ***
HORA                  0.021489   0.001099   19.555  < 2e-16 ***
MES                   0.006960   0.001836    3.791 0.000150 ***
POSICIONX2           -0.204760   0.021050   -9.727  < 2e-16 ***
POSICIONX3           -0.291019   0.038334   -7.592 3.16e-14 ***
POSICIONX4           -0.290979   0.035001   -8.313  < 2e-16 ***
POSICIONX5           -0.360185   0.061922   -5.817 6.00e-09 ***
POSICIONX8           -0.376290   0.069034   -5.451 5.01e-08 ***
POSICIONX9           -1.359436   0.174271   -7.801 6.16e-15 ***
MANIOBRASX11          0.277803   0.056027    4.958 7.11e-07 ***
MANIOBRASX12          0.594241   0.050725   11.715  < 2e-16 ***
MANIOBRASX13          0.406582   0.120547    3.373 0.000744 ***
MANIOBRASX15          1.555601   0.188999    8.231  < 2e-16 ***
MANIOBRASX17          0.502247   0.087648    5.730 1.00e-08 ***
MANIOBRASX18          0.441808   0.108457    4.074 4.63e-05 ***
MANIOBRASX19          0.604077   0.083803    7.208 5.66e-13 ***
MANIOBRASX20          2.196401   0.176201   12.465  < 2e-16 ***
MANIOBRASX21          0.281859   0.043528    6.475 9.46e-11 ***
MANIOBRASX22         -0.121468   0.039335   -3.088 0.002015 **
MANIOBRASX23          0.762910   0.119136    6.404 1.52e-10 ***
MANIOBRASX24          0.900468   0.092167    9.770  < 2e-16 ***
MANIOBRASX25          0.308351   0.139328    2.213 0.026888 *
MANIOBRASX26          0.587099   0.302495    1.941 0.052276 .
MANIOBRASX29          0.906092   0.276193    3.281 0.001036 **
MANIOBRASX3          -0.078243   0.027404   -2.855 0.004301 **
MANIOBRASX31          0.789679   0.167580    4.712 2.45e-06 ***
MANIOBRASX41          0.165768   0.069001    2.402 0.016288 *
MANIOBRASX42          0.961114   0.232957    4.126 3.70e-05 ***
MANIOBRASX43          0.755005   0.103323    7.307 2.73e-13 ***
MANIOBRASX5          -0.226516   0.060242   -3.760 0.000170 ***
MANIOBRASX51          1.763338   0.088745   19.870  < 2e-16 ***
MANIOBRASX52          1.839539   0.025559   71.972  < 2e-16 ***
MANIOBRASX6           0.784178   0.177643    4.414 1.01e-05 ***
MANIOBRASX71          0.414279   0.043521    9.519  < 2e-16 ***
MANIOBRASX72          2.673937   0.038912   68.718  < 2e-16 ***
MANIOBRASX77          0.477156   0.085531    5.579 2.42e-08 ***
MANIOBRASX8           1.469989   0.469554    3.131 0.001744 **
TIPO_VEHICULOX13     -0.661662   0.099861   -6.626 3.45e-11 ***
TIPO_VEHICULOX15     -0.619892   0.072670   -8.530  < 2e-16 ***
TIPO_VEHICULOX16      0.549678   0.096915    5.672 1.41e-08 ***
TIPO_VEHICULOX17      0.665469   0.333851    1.993 0.046227 *
TIPO_VEHICULOX18     -1.899010   0.101980  -18.621  < 2e-16 ***
TIPO_VEHICULOX2      -2.000156   0.049953  -40.041  < 2e-16 ***
TIPO_VEHICULOX20     -1.096894   0.082817  -13.245  < 2e-16 ***
TIPO_VEHICULOX21     -1.046202   0.189443   -5.523 3.34e-08 ***
TIPO_VEHICULOX3      -2.212017   0.049419  -44.760  < 2e-16 ***
TIPO_VEHICULOX4      -2.283154   0.045822  -49.827  < 2e-16 ***
TIPO_VEHICULOX6      -0.287688   0.044744   -6.430 1.28e-10 ***
TIPO_VEHICULOX7      -0.551663   0.124586   -4.428 9.51e-06 ***
TIPO_VEHICULOX8      -1.365981   0.095590  -14.290  < 2e-16 ***
TIPO_VEHICULOX9      -0.288959   0.051876   -5.570 2.54e-08 ***
```

```
INTERSECCIONX2       -0.570766   0.016176  -35.285  < 2e-16 ***
INTERSECCIONX6        0.126221   0.021889    5.766 8.10e-09 ***
INTERSECCIONX7        0.056589   0.023049    2.455 0.014081 *
INTERSECCIONX8        0.623237   0.030200   20.637  < 2e-16 ***
INTERSECCIONX9        0.267545   0.036943    7.242 4.42e-13 ***
INTERSECCIONX999      0.185341   0.035052    5.288 1.24e-07 ***
ZONA_AGRUPADAX2       1.034802   0.017426   59.381  < 2e-16 ***
SEXOX2                0.204573   0.015090   13.557  < 2e-16 ***
LUMINOSIDADX2        -0.243769   0.028046   -8.692  < 2e-16 ***
LUMINOSIDADX3        -0.262164   0.018143  -14.449  < 2e-16 ***
LUMINOSIDADX4        -0.553158   0.019917  -27.773  < 2e-16 ***
SUPERFICIE_CALZADAX3  0.302556   0.020500   14.759  < 2e-16 ***
SUPERFICIE_CALZADAX4  0.442031   0.135787    3.255 0.001133 **
SUPERFICIE_CALZADAX5  0.610606   0.161478    3.781 0.000156 ***
SUPERFICIE_CALZADAX6  1.527321   0.165806    9.211  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 262803  on 675025  degrees of freedom
Residual deviance: 208950  on 674958  degrees of freedom
AIC: 209086

Number of Fisher Scoring iterations: 8
```

**Matriz de confusión**

```
Confusion Matrix and Statistics

          Reference
Prediction     X0     X1
        X0   6713  45679
        X1   1504 114859

               Accuracy : 0.7204
                 95% CI : (0.7183, 0.7225)
    No Information Rate : 0.9513
    P-Value [Acc > NIR] : 1

                  Kappa : 0.15
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.81696
            Specificity : 0.71546
         Pos Pred Value : 0.12813
         Neg Pred Value : 0.98707
             Prevalence : 0.04869
         Detection Rate : 0.03978
   Detection Prevalence : 0.31046
      Balanced Accuracy : 0.76621

       'Positive' Class : X0
```

# Apéndice: Justificante solicitud de registro de software

GENERALITAT VALENCIANA

# SOL·LICITUD REGISTRE PROPIETAT INTEL·LECTUAL

## *SOLICITUD REGISTRO PROPIEDAD INTELECTUAL*

**MODEL A / *MODELO A***

NÚM. INSC. / *Nº INSC.* 09 / _____ / _____

NÚM. SOL·LICITUD / *SOLICITUD* A-163-2017

## A DADES DEL SOL·LICITANT / *DATOS DEL SOLICITANTE*

(En cas de representació ha d'acreditar-la per mitjà d'autorització de l'autor/a i còpia del DNI de la persona sol·licitant) / *(En caso de representación debe acreditar ésta mediante autorización del autor/a y copia del DNI de la persona solicitante)*

| COGNOMS / *APELLIDOS* | NOM / *NOMBRE* | DNI / PASSAPORT / *DNI / PASAPORTE* | NACIONALITAT / *NACIONALIDAD* |
|---|---|---|---|
| Jordan Vidal | Manuel Miguel | 18965995B | Española |

| ADREÇA (CARRER/PLAÇA, NÚMERO I PORTA) / *DOMICILIO (CALLE/PLAZA, NÚMERO Y PUERTA)* | CP |
|---|---|
| Avda. de la Universidad s/n Edif. Rectorado y Consejo Social | 03202 |

| LOCALITAT / *LOCALIDAD* | PROVÍNCIA / *PROVINCIA* | TELÈFON / *TELÉFONO* | CORREU ELECTRÒNIC / *CORREO ELECTRÓNICO* |
|---|---|---|---|
| Elche | Alicante | 966658620 | otri@umh.es |

## B DADES DE L'OBRA / *DATOS DE LA OBRA*

TÍTOL DE L'OBRA (HAURÀ DE SER IDÈNTIC AL QUE FIGURA EN L'EXEMPLAR APORTAT) / *TÍTULO DE LA OBRA (DEBERÁ SER IDÉNTICO AL QUE FIGURA EN EL EJEMPLAR APORTADO)*

Ecosistema de aplicaciones y servicios para la gestión y almacenamiento de Big Data en aplicaciones de vehículos

| CLASSE D'OBRA / *CLASE DE OBRA* | COL·LECCIÓ DE / *COLECCIÓN DE* _____ OBRES / *OBRAS* | HA SIGUT DIVULGADA? / *¿HA SIDO DIVULGADA?* |
|---|---|---|
| Programa de ordenador | | ☐ SÍ ☒ NO |
| DATA I LLOC DE DIVULGACIÓ / *FECHA Y LUGAR DE DIVULGACIÓN* | NÚM. DE DEPÒSIT LEGAL / *Nº DE DEPÓSITO LEGAL* | NÚM. D'ISBN/ISMN / *Nº DE ISBN/ISMN* |
| | | |

## C AUTOR/S / *AUTOR/ES*

☐ UN. / *UNO.* PSEUDÒNIM: / *SEUDÓNIMO:* _____

☐ DIVERSOS. / *VARIOS.* Indiqueu-ne nre.: / *Indicar nº.:* _____ (S'inclouran en Model A -2, aportant la fotocòpia del DNI i l'autorització de cada un per a sol·licitar la inscripció) / *(Se incluirán en Modelo A -2, aportando la fotocopia del DNI y la autorización de cada uno para solicitar la inscripción)*

| COGNOMS / *APELLIDOS* | NOM / *NOMBRE* | DNI / PASSAPORT / *DNI / PASAPORTE* | NACIONALITAT / *NACIONALIDAD* |
|---|---|---|---|
| | | | |

| ADREÇA (CARRER/PLAÇA, NÚM. I PORTA) / *DOMICILIO (CALLE/PLAZA, Nº Y PUERTA)* | CP | LOCALITAT / *LOCALIDAD* | PROVÍNCIA / *PROVINCIA* |
|---|---|---|---|
| | | | |

| TELÈFON / *TELÉFONO* | PART DE L'OBRA DE LA QUAL ÉS AUTOR / *PARTE DE LA OBRA DE LA QUE ES AUTOR* | CORREU ELECTRÒNIC / *CORREO ELECTRÓNICO* |
|---|---|---|
| | | |

## D TITULAR DEL DRET / *TITULAR DEL DERECHO*

☐ Autor/a / *Autor/a*

☒ Altre (*): / *Otro (*):* ☒ Relació laboral / *Relación laboral* ☐ Transmissió intervius / *Transmisión inter-vivos* ☐ Transmissió mortis causa / *Transmisión mortis causa*

(*) S'aportarà DNI del titular de drets, i en cas de persones jurídiques, el títol que acredite la seua personalitat jurídica, el poder de representació i el CIF, a més d'aquella documentació exigida per a acreditar la cessió. En cas de diversos titulars dels drets diferents als autors s'inclouran en el Model A-3.
*Se aportará DNI del titular de derechos, y en caso de personas jurídicas, el título que acredite su personalidad jurídica, el poder de representación y el CIF, además de aquella documentación exigida para acreditar la cesión. En caso de diversos titulares de derechos distintos a los autores se incluirán en Modelo A-3.*

| COGNOMS I NOM O RAÓ SOCIAL / *APELLIDOS Y NOMBRE O RAZÓN SOCIAL* | DNI / PASSAPORT / *DNI / PASAPORTE* | NACIONALITAT / *NACIONALIDAD* |
|---|---|---|
| UNIVERSIDAD MIGUEL HERNANDEZ DE ELCHE | Q5350015C | ESPAÑOLA |

| ADREÇA (CARRER/PLAÇA, NÚM. I PORTA) / *DOMICILIO (CALLE/PLAZA, Nº Y PUERTA)* | CP | LOCALITAT / *LOCALIDAD* | PROVÍNCIA / *PROVINCIA* |
|---|---|---|---|
| AVDA. DE LA UNIVERSIDAD S/N | 03202 | ELCHE | ALICANTE |

| TELÈFON / *TELÉFONO* | PART DE L'OBRA DE LA QUAL ÉS TITULAR DEL DRET / *PARTE DE LA OBRA DE LA QUE ES TITULAR DEL DERECHO* | CORREU ELECTRÒNIC / *CORREO ELECTRÓNICO* |
|---|---|---|
| 966658782 | | otri@umh.es |

## E COMUNICACIÓ / *COMUNICACIÓN*

El termini màxim per a resoldre i efectuar la notificació corresponent serà de sis mesos comptats des de la data d'entrada de la present sol·licitud a registre territorial competent per a resoldre segons l'art. 24.1 del Reglament del Registre (R.D. 281/2003, de 7 de març. BOE del 28 de març). Transcorregut este termini sense que s'haja dictat resolució expressa, s'entendrà estimada la sol·licitud d'acord amb el que disposa l'art. 43 de la Llei 30/1992, de 26 de novembre, de Règim Jurídic de les Administracions Públiques i del Procediment Administratiu Comú segons redacció donada per la Llei 4/1999, de 13 de gener.

*El plazo máximo para resolver y efectuar la notificación correspondiente será de seis meses contados desde la fecha de entrada de la presente solicitud en el registro territorial competente para resolver según el art. 24.1 del Reglamento del Registro (R.D. 281/2003, de 7 de marzo. BOE del 28 de marzo). Transcurrido dicho plazo sin que se haya dictado resolución expresa, se entenderá estimada la solicitud de acuerdo con lo dispuesto en el art. 43 de la Ley 30/1992, de 26 de noviembre, de Régimen Jurídico de las Administraciones Públicas y del Procedimiento Administrativo Común según redacción dada por la Ley 4/1999, de 13 de enero.*

## F SOL·LICITUD / *SOLICITUD*

La inscripció al Registre de la Propietat Intel·lectual de l'obra el títol i les circumstàncies de la qual s'expressen.
La persona signant declara que són certes les dades que figuren en esta sol·licitud i en la documentació que s'adjunta, assumint, en cas contrari, les responsabilitats legals que pogueren derivar-se.

*La inscripción en el Registro de la Propiedad Intelectual de la obra cuyo título y circunstancias se expresan.*
*La persona que firma declara que son ciertos los datos que figuran en esta solicitud y en la documentación adjunta, asumiendo, en caso contrario, las responsabilidades legales que pudieran derivarse.*

Alacant, 7 d'Abril de 2017

Nom / Firma / DNI:
*Nombre / Firma / DNI:* _____

Les dades personals que conté l'imprés podran ser incloses en un fitxer per al tractament per la Conselleria d'Educació, Cultura i Esport, en l'ús de les funcions pròpies que té atribuïdes en l'àmbit de les seues competències, i es podrà dirigir a qualsevol òrgan d'esta per a exercir els drets d'accés, rectificació, cancel·lació i oposició, segons disposa la Llei Orgànica 15/1999, de 13 de desembre, de Protecció de Dades de Caràcter Personal (BOE núm. 298, de 14/12/99).

*Los datos personales contenidos en este impreso podrán ser incluidos en un fichero para su tratamiento por la Conselleria de Educación, Cultura y Deporte, en el uso de las funciones propias que tiene atribuidas en el ámbito de sus competencias, pudiendo dirigirse a cualquier órgano de la misma para ejercitar los derechos de acceso, rectificación, cancelación y oposición, según lo dispuesto en la Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal (BOE nº 298, de 14/12/99).*

REGISTRE D'ENTRADA / *REGISTRO DE ENTRADA*
GENERALITAT VALENCIANA · REGISTRE PROPIETAT INTEL·LECTUAL · REGISTRO PROPIEDAD INTELECTUAL · SERVICI TERRITORIAL DE CULTURA I ESPORT D'ALACANT · CONSELLERIA D'EDUCACIÓ, INVESTIGACIÓ, CULTURA I ESPORT

DATA D'ENTRADA EN L'ÒRGAN COMPETENT: data, hora i minut
*FECHA ENTRADA EN ÓRGANO COMPETENTE: fecha, hora y minuto*
7-4-2017

(3/3) EXEMPLAR PER A LA PERSONA INTERESSADA / *EJEMPLAR PARA LA PERSONA INTERESADA*

CHAP - IAC · DIN - A4 · IA - 14019 - 01 - E



10h 15m

GENERALITAT VALENCIANA

# OBJECTE/S DE PROPIETAT INTEL·LECTUAL QUE DESITJA PROTEGIR

## *OBJETO/S DE PROPIEDAD INTELECTUAL QUE DESEA PROTEGER*

**MODEL B / *MODELO B***

NÚM. SOL·LICITUD / *Nº SOLICITUD*: A-163-2017

TÍTOL DE L'OBRA / *TÍTULO DE LA OBRA*

Ecosistema de aplicaciones y servicios para la gestión y almacenamiento de Big Data en aplicaciones de vehículos

**A** **OBJECTES DE PROPIETAT INTEL·LECTUAL** (Assenyale amb una creu a la casella corresponent. Les obres hauran de presentar-se segons la informació al dors)
***OBJETOS DE PROPIEDAD INTELECTUAL*** *(Señale con una cruz en la casilla correspondiente. Las obras deberán presentarse según la información al dorso)*

1. ☐ **Obra literària, científica o dramàtica (text)** / ***Obra literaria, científica o dramática (texto)***
   ☐ Amb il·lustracions de l'autor / *Con ilustraciones del autor* ☐ Amb il·lustracions alienes (indiqueu pàgina en observacions) / *Con ilustraciones ajenas (indicar página en observaciones)*
   Nre. de pàgines / *Nº de páginas* ____ Nre. de volums / *Nº de volúmenes* ____ Format / *Formato* ____ En obres dramàtiques duració aproximada / *En obras dramáticas duración aproximada* ____
2. ☐ **Composició musical** / ***Composición musical*** ☐ Sense lletra / *Sin letra* ☐ Amb lletra / *Con letra*
   Gènere musical / *Género musical* ____ Duració aprox. / *Duración aprox.* ____ Nre. de compassos / *Nº de compases* ____
   Plantilla instrumental (i vocal, si escau) / *Plantilla instrumental (y vocal, en su caso)* ____
3. ☐ **Coreografia o pantomima** / ***Coreografía o pantomima*** ☐ Gravació / *Grabación* ☐ Descripció moviment escènic / *Descripción movimiento escénico*
4. ☐ **Obra cinematogràfica i la resta d'obres audiovisuals** / ***Obra cinematográfica y demás obras audiovisuales***
   Minutatge / *Minutaje* ____ Idioma original ____ ☐ Extracte-resum / *Extracto-resumen*
5. ☐ **Escultura, dibuix i pintura i la resta d'obres plàstiques siguen o no aplicades** / ***Escultura, dibujo y pintura y demás obras plásticas sean o no aplicadas***
   Material emprat / *Material empleado* ____ Tipus de suport / *Tipo de soporte* ____
   Tècnica emprada / *Técnica empleada* ____ Dimensions / *Dimensiones* ____
6. ☐ **Gravat i litografia** / ***Grabado y litografía*** Tècnica de gravació / *Técnica de grabación* ____ Material del suport / *Material del soporte* ____
7. ☐ **Tebeo i còmic** / ***Tebeo y cómic*** Nre. de pàgines o fulls / *Nº de páginas u hojas* ____ Nre. de volums / *Nº de volúmenes* ____ Format / *Formato* ____
8. ☐ **Obra fotogràfica** / ***Obra fotográfica*** Suport / *Soporte* ____
9. ☐ **Arquitectura i enginyeria** / ***Arquitectura e ingeniería*** ☐ Projecte / *Proyecto* ☐ Pla / *Plano* ☐ Disseny / *Diseño*
10. ☐ **Maquetes** / ***Maquetas*** Escala ____ Dimensions / *Dimensiones* ____
11. ☐ **Topografia, geografia, ciència en general** / ***Topografía, geografía, ciencia en general*** ☐ Gràfic / *Gráfico* ☐ Mapa ☐ Disseny / *Diseño*
12. ☒ **Programa d'ordinador** / ***Programa de ordenador*** ☐ Executable / *Ejecutable* ☐ Codi font / *Código fuente* ☐ Paper / *Papel* ☒ Suport digital / *Soporte digital*
13. ☐ **Estructura i disposició de base de dades** / ***Estructura y disposición de base de datos***
14. ☐ **Continguts originals de pàgina web o multimèdia** / ***Contenidos originales de página web o multimedia***
    ☐ Text / *Texto* ☐ Imatge / *Imagen* ☐ Audiovisuals / *Audiovisuales* ☐ Multimèdia / *Multimedia* ☐ So / *Sonido* ☐ Codi font / *Código fuente*
15. ☐ **Actuacions d'artistes, intèrprets o executants** / ***Actuaciones de artistas, intérpretes o ejecutantes***
16. ☐ **Produccions fonogràfiques** / ***Producciones fonográficas*** ☐ **Produccions audiovisuals** / ***Producciones audiovisuales***
17. ☐ **Meres fotografies** / ***Meras fotografías*** Data de realització / *Fecha de realización* ____ Col·lecció de fotografies. Indiqueu-ne nre.: / *Colección de fotografías. Indicar nº:* ____
18. ☐ **Produccions editorials art. 129 TRLPI** / ***Producciones editoriales art. 129 TRLPI***
19. ☐ **Un altre (indiqueu-ne quin)** / ***Otro (indicar cual)*** ____

La persona signant declara que són certes les dades que figuren en esta sol·licitud i en la documentació que s'adjunta, assumint, en cas contrari, les responsabilitats legals que pogueren derivar-se.

*La persona que firma declara que son ciertos los datos que figuran en esta solicitud y en la documentación adjunta, asumiendo, en caso contrario, las responsabilidades legales que pudieran derivarse.*

Alicante, 7 de Octubre de 2017

Nom / Firma / DNI:
*Nombre / Firma / DNI:* ____

REGISTRE DE LA PROPIETAT INTEL·LECTUAL DE LA CV
*REGISTRO DE LA PROPIEDAD INTELECTUAL DE LA CV*

GENERALITAT VALENCIANA · REGISTRE PROPIETAT INTEL·LECTUAL · CONSELLERIA D'EDUCACIÓ, INVESTIGACIÓ, CULTURA I ESPORT · SERVICIO TERRITORIAL DE CULTURA I ESPORT D'ALACANT

(3/3) EXEMPLAR PER A LA PERSONA INTERESSADA / *EJEMPLAR PARA LA PERSONA INTERESADA*

CHAP - IAC

DIN - A4

IA - 14123 - 01 - E

04/09/15

GENERALITAT VALENCIANA

# RELACIÓ D'AUTORS

## *RELACIÓN DE AUTORES*

MODEL A2 / *MODELOA2*

NÚM. SOL·LICITUD / *Nº SOLICITUD* A-163-2017

TÍTOL DE L'OBRA / *TÍTULO DE LA OBRA*

## A DADES DELS AUTORS / *DATOS DE LOS AUTORES*

Autor núm. / *Autor nº* 1

| COGNOMS / *APELLIDOS* | NOM / *NOMBRE* | DNI/PASSAPORT *DNI/PASAPORTE* | FIRMA |
|---|---|---|---|
| GIL APARICIO | ARTURO | 74224107W | |

| NACIONALITAT / *NACIONALIDAD* | ADREÇA (CARRER/PLAÇA, NÚM. I PORTA) / *DOMICILIO (CALLE/PLAZA, Nº Y PUERTA)* | CP |
|---|---|---|
| ESPAÑOLA | AVDA. DE LA UNIVERSIDAD S/N EDIF. INNOVA | 03202 |

| LOCALITAT / *LOCALIDAD* | PROVÍNCIA / *PROVINCIA* | TELÈFON / *TELÉFONO* |
|---|---|---|
| ELCHE | ALICANTE | 966658782 |

| PART DE L'OBRA DE LA QUAL ÉS AUTOR / *PARTE DE LA OBRA DE LA QUE ES AUTOR* | CORREU ELECTRÒNIC / *CORREO ELECTRÓNICO* |
|---|---|
| 50% | arturo.gil@umh.es |

Autor núm. / *Autor nº* 2

| COGNOMS / *APELLIDOS* | NOM / *NOMBRE* | DNI/PASSAPORT *DNI/PASAPORTE* | FIRMA |
|---|---|---|---|
| UBEDA GONZALEZ | DAVID | 74222911W | |

| NACIONALITAT / *NACIONALIDAD* | ADREÇA (CARRER/PLAÇA, NÚM. I PORTA) / *DOMICILIO (CALLE/PLAZA, Nº Y PUERTA)* | CP |
|---|---|---|
| ESPAÑOLA | AVDA. DE LA UNIVERSIDAD S/N EDIF. TORREPINET | 03202 |

| LOCALITAT / *LOCALIDAD* | PROVÍNCIA / *PROVINCIA* | TELÈFON / *TELÉFONO* |
|---|---|---|
| ELCHE | ALICANTE | 966652403 |

| PART DE L'OBRA DE LA QUAL ÉS AUTOR / *PARTE DE LA OBRA DE LA QUE ES AUTOR* | CORREU ELECTRÒNIC / *CORREO ELECTRÓNICO* |
|---|---|
| 50% | ubeda@umh.es |

Autor núm. / *Autor nº* ___

| COGNOMS / *APELLIDOS* | NOM / *NOMBRE* | DNI/PASSAPORT *DNI/PASAPORTE* | FIRMA |
|---|---|---|---|
| | | | |

| NACIONALITAT / *NACIONALIDAD* | ADREÇA (CARRER/PLAÇA, NÚM. I PORTA) / *DOMICILIO (CALLE/PLAZA, Nº Y PUERTA)* | CP |
|---|---|---|
| | | |

| LOCALITAT / *LOCALIDAD* | PROVÍNCIA / *PROVINCIA* | TELÈFON / *TELÉFONO* |
|---|---|---|
| | | |

| PART DE L'OBRA DE LA QUAL ÉS AUTOR / *PARTE DE LA OBRA DE LA QUE ES AUTOR* | CORREU ELECTRÒNIC / *CORREO ELECTRÓNICO* |
|---|---|
| | |

Autor núm. / *Autor nº* ___

| COGNOMS / *APELLIDOS* | NOM / *NOMBRE* | DNI/PASSAPORT *DNI/PASAPORTE* | FIRMA |
|---|---|---|---|
| | | | |

| NACIONALITAT / *NACIONALIDAD* | ADREÇA (CARRER/PLAÇA, NÚM. I PORTA) / *DOMICILIO (CALLE/PLAZA, Nº Y PUERTA)* | CP |
|---|---|---|
| | | |

| LOCALITAT / *LOCALIDAD* | PROVÍNCIA / *PROVINCIA* | TELÈFON / *TELÉFONO* |
|---|---|---|
| | | |

| PART DE L'OBRA DE LA QUAL ÉS AUTOR / *PARTE DE LA OBRA DE LA QUE ES AUTOR* | CORREU ELECTRÒNIC / *CORREO ELECTRÓNICO* |
|---|---|
| | |

Autor núm. / *Autor nº* ___

| COGNOMS / *APELLIDOS* | NOM / *NOMBRE* | DNI/PASSAPORT *DNI/PASAPORTE* | FIRMA |
|---|---|---|---|
| | | | |

| NACIONALITAT / *NACIONALIDAD* | ADREÇA (CARRER/PLAÇA, NÚM. I PORTA) / *DOMICILIO (CALLE/PLAZA, Nº Y PUERTA)* | CP |
|---|---|---|
| | | |

| LOCALITAT / *LOCALIDAD* | PROVÍNCIA / *PROVINCIA* | TELÈFON / *TELÉFONO* |
|---|---|---|
| | | |

| PART DE L'OBRA DE LA QUAL ÉS AUTOR / *PARTE DE LA OBRA DE LA QUE ES AUTOR* | CORREU ELECTRÒNIC / *CORREO ELECTRÓNICO* |
|---|---|
| | |

La persona signant declara que són certes les dades que figuren en esta sol·licitud i en la documentació que s'adjunta, assumint, en cas contrari, les responsabilitats legals que pogueren derivar-se.

*La persona que firma declara que son ciertos los datos que figuran en esta solicitud y en la documentación adjunta, asumiendo, en caso contrario, las responsabilidades legales que pudieran derivarse.*

Alacant, 7 d'abril de 2017

Nom / Firma / DNI:
*Nombre / Firma / DNI:*

GENERALITAT VALENCIANA · REGISTRE DE PROPIETAT INTEL·LECTUAL · CONSELLERIA D'EDUCACIÓ, INVESTIGACIÓ, CULTURA I ESPORT · SERVICI TERRITORIAL DE CULTURA I ESPORT D'ALACANT

(3/3). EXEMPLAR PER A LA PERSONA SOL·LICITANT / *EJEMPLAR PARA LA PERSONA SOLICITANTE*

CHAP - IAC

DIN - A4

IA - 14124 - 01 - E

| | OBJECTE/S DE PROPIETAT INTEL·LECTUAL QUE DESITJA PROTEGIR<br>*OBJETO/S DE PROPIEDAD INTELECTUAL QUE DESEA PROTEGER* | MODEL B / *MODELO B*<br>NÚM. SOL·LICITUD / *Nº SOLICITUD*<br>A- 163-2017 |
|---|---|---|

TÍTOL DE L'OBRA / *TÍTULO DE LA OBRA*

Ecosistema de aplicaciones y servicios para la gestión y almacenamiento de Big Data en aplicaciones de vehículos

B **OBSERVACIONS / *OBSERVACIONES***

GENERALITAT VALENCIANA · REGISTRE PROPIETAT INTEL·LECTUAL · SERVICI TERRITORIAL DE CULTURA I ESPORT D'ALACANT · CONSELLERIA D'EDUCACIÓ, INVESTIGACIÓ, CULTURA I ESPORT

El título completo del programa de ordenador es "Ecosistema de aplicaciones y servicios para la gestión y almacenamiento de Big Data en aplicaciones de vehículos y análisis de conducción" pero en el apartado de título no hay espacio suficiente para ponerlo completo

La persona signant declara que són certes les dades que figuren en esta sol·licitud i en la documentació que s'adjunta, assumint, en cas contrari, les responsabilitats legals que pogueren derivar-se.

*La persona que firma declara que son ciertos los datos que figuran en esta solicitud y en la documentación adjunta, asumiendo, en caso contrario, las responsabilidades legales que pudieran derivarse.*

Alacant, 7 d'abril de 2017

Nom / Firma / DNI:
*Nombre / Firma / DNI:*

REGISTRE DE LA PROPIETAT INTEL·LECTUAL DE LA CV
*REGISTRO DE LA PROPIEDAD INTELECTUAL DE LA CV*

04/09/15

(3/3) EXEMPLAR PER A LA PERSONA INTERESSADA / *EJEMPLAR PARA LA PERSONA INTERESADA*

IA - 14123 - 02 - E DIN - A4 CHAP - IAC

# Bibliografía


[1] Road Casualties Great Britain, 2013 department for transport, road accidents and safety statistics. https://www.gov.uk/government/collections/road-accidents-and-safety-statistics. Accessed: 2016-09-30. 163

[2] Series Road accidents and safety statistics department for transport. http://www.dft.gov.uk/statistics/series/road-accidents-and-safety/. Accessed: 2016-10-07. 163

[3] Instructions for the completion of road accident reports with effect from 1 january 2005, 2004. 163

[4] European union, smart seat and seatbelt to help sleepy drivers stay alert. *Research*eu Results Magazine*, 42:6–7, 2015. 42

[5] Khaled A Abbas. Traffic safety assessment and development of predictive models for accidents on rural roads in egypt. *Accident Analysis & Prevention*, 36(2):149–163, 2004. 57

[6] Mohamed A Abdel-Aty and A Essam Radwan. Modeling traffic accident occurrence and involvement. *Accident Analysis & Prevention*, 32(5):633–642, 2000. 58

[7] CARE Community Road Accident. Care community road accident database, 2006. Accedida en febrero de 2016. 41

[8] Jonathan Aguero-Valverde and Paul Jovanis. Analysis of road crash frequency with spatial models. *Transportation Research Record: Journal of the Transportation Research Board*, (2061):55–63, 2008. 57

[9] Alex Alexandridis, Panagiotis Patrinos, Haralambos Sarimveis, and George Tsekouras. A two-stage evolutionary algorithm for variable selection in the development of rbf neural network models. *Chemometrics and intelligent laboratory systems*, 75(2):149–162, 2005. 26

[10] Christophe Ambroise and Geoffrey J McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the national academy of sciences*, 99(10):6562–6566, 2002. 14

[11] FRED J Anscombe and John W Tukey. The examination and analysis of residuals. *Technometrics*, 5(2):141–160, 1963. 12

[12] Maurice Aron, Ruth Bergel-Hayat, and Eric Violette. Added risk by rainy weather on the roads of normandie-centre region in france. In *11th World Conference on Transport Research*, 2007. 57

[13] Kevin Ashton. That internet of things thing, 1999. 17, 67, 71

[14] Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM Sigkdd Explorations Newsletter*, 6(1):20–29, 2004. 18, 19

[15] KM Bauer and Douglas W Harwood. Statistical models of at-grade intersection accidents. Technical report, 1996. 58

[16] Ruth Bergel-Hayat, Mohammed Debbarh, Constantinos Antoniou, and George Yannis. Explaining the road accident risk: Weather effects. *Accident Analysis & Prevention*, 60:456–465, 2013. 43, 57

[17] Tibebe Beshah, Dejene Ejigu, Ajith Abraham, Vaclav Snasel, and Pavel Kromer. Mining pattern from road accident data: Role of road users behaviour and implications for improving road safety. *International journal of tomography and simulation*, 22(1):73–86, 2013. 61

[18] Tibebe Beshah and Shawndra Hill. Mining road traffic accident data to improve safety: Role of road-related factors on accident severity in ethiopia. In *AAAI Spring Symposium: Artificial Intelligence for Development*, 2010. 59

[19] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. 35, 36

[20] Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. Classification and regression trees. wadsworth & brooks. *Monterey, CA*, 1984. 34

[21] Julian Browne. Brewers cap theorem., 2009. 73

[22] Li-Yen Chang and Wen-Chieh Chen. Data mining of tree-based models to analyze freeway accident frequency. *Journal of safety research*, 36(4):365–375, 2005. 57

[23] Nitesh V Chawla. C4. 5 and imbalanced data sets: investigating the effect of sampling method, probabilistic estimate, and decision tree structure. In *Proceedings of the ICML*, volume 3, 2003. 18

[24] Nitesh V Chawla. Data mining for imbalanced datasets: An overview. In *Data mining and knowledge discovery handbook*, pages 853–867. Springer, 2005. 18

[25] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002. 18, 19

[26] Nitesh V Chawla, Nathalie Japkowicz, and Aleksander Kotcz. Editorial: special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter*, 6(1):1–6, 2004. 18

[27] Nitesh V Chawla, Aleksandar Lazarevic, Lawrence O Hall, and Kevin W Bowyer. Smoteboost: Improving prediction of the minority class in boosting. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 107–119. Springer, 2003. 19

[28] Yu-Chiun Chiou and Chiang Fu. Modeling crash frequency and severity using multinomial-generalized poisson model with error components. *Accident Analysis & Prevention*, 50:73–82, 2013. 57, 58

[29] Robert D Clark. Optisim: an extended dissimilarity selection method for finding diverse representative subsets. *Journal of Chemical Information and Computer Sciences*, 37(6):1181–1188, 1997. 16

[30] Peter A Davison. Inter-relationships between british drivers' visual abilities, age and road accident histories. *Ophthalmic and Physiological Optics*, 5(2):195–204, 1985. 56, 57

[31] Ministerio de Interior. Gobierno de España. Orden del ministerio de relaciones con las cortes y de la secretaría del gobierno de 18 de febrero de 1993, 1993. 95

[32] Ministerio de Interior. Gobierno de España. Ley sobre tráfico, circulación de vehículos a motor y seguridad vial, 2015. 95

[33] Juan de Oña, Randa Oqab Mujalli, and Francisco J Calvo. Analysis of traffic accident injury severity on spanish rural highways using bayesian networks. *Accident Analysis & Prevention*, 43(1):402–411, 2011. 60

[34] Dirección General de Tráfico. Información de carreteras: Infocar, 2013. 93, 94, 107

[35] Dirección General de Tráfico. Anuario estadístico de accidentes, 2015. 51

[36] Dirección General de Tráfico. Portal estadístico de la dgt, 2015. 93, 94, 95

[37] Benoît Depaire, Geert Wets, and Koen Vanhoof. Traffic accident segmentation by means of latent class clustering. *Accident Analysis & Prevention*, 40(4):1257–1266, 2008. 59

[38] Markus Deublein, Matthias Schubert, Bryan T Adey, Jochen Köhler, and Michael H Faber. Prediction of road accidents: A bayesian hierarchical approach. *Accident Analysis & Prevention*, 51:274–291, 2013. 60

[39] Thomas G Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning*, 40(2):139–157, 2000. 35

[40] Eric T Donnell and John M Mason. Predicting the frequency of median barrier crashes on pennsylvania interstate highways. *Accident Analysis & Prevention*, 38(3):590–599, 2006. 57, 58

[41] Mark Dougherty. A review of neural networks applied to transport. *Transportation Research Part C: Emerging Technologies*, 3(4):247–260, 1995. 61

[42] Comisión económica para Europa. Convention on road traffic, 1968. Accedida en enero de 2017. 63

[43] Amirhossein Ehsaei and Harry Evdorides. Temporal variation of road accident data caused by road infrastructure. In *3rd International Conference of Road Safety and Simulation, September*, pages 14–16, 2011. 59

[44] Karim El-Basyouny and Tarek Sayed. Accident prediction models with random corridor parameters. *Accident Analysis & Prevention*, 41(5):1118–1123, 2009. 57

[45] Jane Elith, John R Leathwick, and Trevor Hastie. A working guide to boosted regression trees. *Journal of Animal Ecology*, 77(4):802–813, 2008. 35

[46] Rune Elvik. Assessing causality in multivariate accident models. *Accident Analysis & Prevention*, 43(1):253–264, 2011. 61

[47] Begoña Fernández. La ley de los eventos raros, legado de siméon denis poisson. *Memorias Escuela Regional de Probabilidad y Estadística, Villahermosa, UJAT*, 2008. 8

[48] Department for transport. *Reported road casualties in Great Britain 2015*. https://www.gov.uk/government/statistics/reported-road-casualties-in-great-britain-main-results-2015 (accessed January 2017), 2015. 41

[49] Lasse Fridstrøm, Jan Ifver, Siv Ingebrigtsen, Risto Kulmala, and Lars Krogsgård Thomsen. Measuring the contribution of randomness, exposure, weather, and daylight to the variation in road accident counts. *Accident Analysis & Prevention*, 27(1):1–20, 1995. 57, 58

[50] Cesare Furlanello, Maria Serafini, Stefano Merler, and Giuseppe Jurman. Entropy-based gene ranking without selection bias for the predictive classification of microarray data. *BMC bioinformatics*, 4(1):54, 2003. 14

[51] Sandra Vieira Gomes. The influence of the infrastructure characteristics in urban road accidents occurrence. *Accident Analysis & Prevention*, 60:289–297, 2013. 58

[52] Pablo M Granitto, Cesare Furlanello, Franco Biasioli, and Flavia Gasperi. Recursive feature elimination with random forest for ptr-ms analysis of agroindustrial products. *Chemometrics and Intelligent Laboratory Systems*, 83(2):83–90, 2006. 14, 26

[53] Road Research Group. *OECD RRG Road and Traffic accident database*. Organisation for Economic Cooperation and Development, Paris, 2001. 41

[54] Mohammed A Hadi, Jacob Aruldhas, Lee-Fang Chow, and Joseph A Wattleworth. Estimating safety effects of cross-section design for various highway types using negative binomial regression. *Transportation Research Record*, 1500:169, 1995. 57, 58

[55] Simon Hakim, Daniel Shefer, Alfred-Shalom Hakkert, and Irit Hocherman. A critical review of macro models for road accidents. *Accident Analysis & Prevention*, 23(5):379–400, 1991. 55, 57

[56] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982. 20

[57] José Hernández-Orallo, Cèsar Ferri, Nicolas Lachiche, and Peter Flach. Roc analysis in artificial intelligence. In *1st International Workshop, ROCAI-2004, Valencia, Spain*, 2004. 17

[58] Emil. Hernández-Arroyo. *Manual de estadística*. Editorial EDUCC, Bucaramanga, Colombia, 2006. 7

[59] Lena Winslott Hiselius. Estimating the relationship between accident frequency and homogeneous and inhomogeneous traffic flows. *Accident analysis & prevention*, 36(6):985–992, 2004. 58

[60] DataStax Inc. Datastax documentation, 2017. 78

[61] DataStax Inc. Apache cassandra, cassandra2014. Accedida en agosto de 2015. 67

[62] Nathalie Japkowicz. The class imbalance problem: Significance and strategies. In *Proc. of the Int. Conf. on Artificial Intelligence*, 2000. 18

[63] Nathalie Japkowicz. Concept-learning in the presence of between-class and within-class imbalances. In *Conference of the Canadian Society for Computational Studies of Intelligence*, pages 67–77. Springer, 2001. 18

[64] Nathalie Japkowicz. Class imbalances: are we focusing on the right issue. In *Workshop on Learning from Imbalanced Data Sets II*, volume 1723, page 63, 2003. 17

[65] Nathalie Japkowicz et al. Learning from imbalanced data sets: a comparison of various strategies. In *AAAI workshop on learning from imbalanced data sets*, volume 68, pages 10–15. Menlo Park, CA, 2000. 17

[66] Taeho Jo and Nathalie Japkowicz. Class imbalances versus small disjuncts. *ACM Sigkdd Explorations Newsletter*, 6(1):40–49, 2004. 18

[67] Paul P Jovanis and Hsin-Li Chang. Modeling the relationship of accidents to miles traveled. *Transportation Research Record*, 1068:42–51, 1986. 57, 58

[68] KR Kalokota and Prianka N Seneviratne. Accident prediction models for two-lane rural highways. Technical report, Mountain-Plains Consortium, 1994. 56, 58

[69] Jaideep Kashyap and Asst Prof Chandra Prakash Singh. Mining road traffic accident data to improve safety on road-related factors for classification and prediction of accident severity. 2016. 59

[70] Dheeraj Khera and Williamjeet Singh. A review on injury severity in traffic system using various data mining techniques. *International Journal of Computer Applications*, 100(3), 2014. 60

[71] AW Kimball. Errors of the third kind in statistical consulting. *Journal of the American Statistical Association*, 52(278):133–142, 1957. 38

[72] Kenji Kira and Larry A Rendell. The feature selection problem: Traditional methods and a new algorithm. In *AAAI*, volume 2, pages 129–134, 1992. 23, 24

[73] Igor Kononenko. Estimating attributes: analysis and extensions of relief. In *European conference on machine learning*, pages 171–182. Springer, 1994. 24

[74] Jake Kononov and Bruce Janson. Diagnostic methodology for the detection of safety problems at intersections. *Transportation Research Record: Journal of the Transportation Research Board*, (1784):51–56, 2002. 55

[75] Max Kuhn. Caret package. *Journal of Statistical Software*, 28(5):1–26, 2008. 22

[76] Max Kuhn and Kjell Johnson. *Applied predictive modeling*, volume 26. Springer, 2013. 11, 12, 17, 25, 33, 34, 35, 36

[77] Risto Kulmala. *SAFETY AT RURAL THREE-AND FOUR-ARM JUNCTIONS. DEVELOPMENT AND APPLICATION OF ACCIDENT PREDICTION MODELS.*, volume 233. 1995. 57, 58

[78] Rebecca Lawton, Dianne Parker, Antony SR Manstead, and Stephen G Stradling. The role of affect in predicting social behaviors: The case of road traffic violations. *Journal of applied social psychology*, 27(14):1258–1276, 1997. 57, 58

[79] Jinsun Lee and Fred Mannering. Impact of roadside features on the frequency and severity of run-off-roadway accidents: an empirical analysis. *Accident Analysis & Prevention*, 34(2):149–161, 2002. 57, 58

[80] Charles X Ling and Chenghui Li. Data mining for direct marketing: Problems and solutions. In *KDD*, volume 98, pages 73–79, 1998. 19

[81] Sebastián Maldonado and Richard Weber. A wrapper method for feature selection using support vector machines. *Information Sciences*, 179(13):2208–2217, 2009. 25

[82] John Milton and Fred Mannering. The relationship among highway geometrics, traffic-related elements and motor-vehicle accident frequencies. *Transportation*, 25(4):395–413, 1998. 57

[83] Sudeshna Mitra and Simon Washington. On the nature of over-dispersion in motor vehicle crash prediction models. *Accident Analysis & Prevention*, 39(3):459–468, 2007. 57, 59

[84] Afshin Shariat Mohaymany, Ali Tavakoli Kashani, and Andishe Ranjbari. Identifying driver characteristics influencing overtaking crashes. *Traffic injury prevention*, 11(4):411–416, 2010. 61

[85] D Morales, A Pérez-Martın, and M Vaca. Monte carlo simulation study of regression models used to estimate the credit banking risk in home equity loans. *WIT Transactions on Information and Communication Technologies*, 45:141–153, 2013. 13

[86] Lorenzo Mussone, Andrea Ferrari, and Marcello Oneta. An analysis of urban collisions using an artificial intelligence model. *Accident Analysis & Prevention*, 31(6):705–718, 1999. 58, 61

[87] Paul J Ossenbruggen, Jyothi Pendharkar, and John Ivan. Roadway safety in rural and small urbanized areas. *Accident Analysis & Prevention*, 33(4):485–498, 2001. 58

[88] Carlos Javier Vilalta Perdomo. *Análisis de datos*. CIDE, 2016. 7

[89] Mark Poch and Fred Mannering. Negative binomial analysis of intersection-accident frequencies. *Journal of transportation engineering*, 122(2):105–113, 1996. 58

[90] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0. 65

[91] Marko Robnik-Šikonja and Igor Kononenko. Theoretical and empirical analysis of relieff and rrelieff. *Machine learning*, 53(1-2):23–69, 2003. 24

[92] S Francisca Rosario and K Thangadurai. Relief: Feature selection approach. *International Journal of Innovative Research and Development|| ISSN 2278–0211*, 4(11), 2015. 23

[93] Barbara E Sabey and Harold Taylor. The known risks we run: the highway. In *Societal risk assessment*, pages 43–70. Springer, 1980. 55, 57

[94] Tarek Sayed and Felipe Rodriguez. Accident prediction models for urban unsignalized intersections in british columbia. *Transportation Research Record: Journal of the Transportation Research Board*, (1665):93–99, 1999. 58

[95] Chris Seiffert, Taghi M Khoshgoftaar, Jason Van Hulse, and Amri Napolitano. Mining data with rare events: a case study. In *Tools with Artificial Intelligence, 2007. ICTAI 2007. 19th IEEE International Conference on*, volume 2, pages 132–139. IEEE, 2007. 18

[96] Government Digital Service. Reported road casualties in great britain: notes, definitions, symbols and conventions, 2009. Accedida en septiembre de 2016. 63

[97] Venky Shankar, John Milton, and F Mannering. Modeling accident frequencies as zero-altered probability processes: an empirical inquiry. *Accident Analysis & Prevention*, 29(6):829–837, 1997. 56, 57, 58

[98] S Shanthi and R Geetha Ramani. Feature relevance analysis and classification of road traffic accident data through data mining techniques. In *Proceedings of the World Congress on Engineering and Computer Science*, volume 1, pages 24–26, 2012. 62

[99] Thomas S Shively, Kara Kockelman, and Paul Damien. A bayesian semi-parametric model to estimate relationships between crash counts and roadway characteristics. *Transportation research part B: methodological*, 44(5):699–715, 2010. 57, 59

[100] Marjan Simoncic. A bayesian network model of two-car accidents. *Journal of transportation and Statistics*, 7(2/3):13–25, 2004. 56, 57, 60

[101] So Young Sohn and Sung Ho Lee. Data fusion, ensemble and clustering to improve the classification accuracy for the severity of road traffic accidents in korea. *Safety Science*, 41(1):1–14, 2003. 58

[102] AH Schistad Solberg and Rune Solberg. A large-scale evaluation of features for automatic detection of oil spills in ers sar images. In *Geoscience and Remote Sensing Symposium, 1996. IGARSS'96.'Remote Sensing for a Sustainable Future.', International*, volume 3, pages 1484–1486. IEEE, 1996. 18

[103] Francisco Soler-Flores. Expert system for road accidents frequency estimation based in naïve-poisson. In *Proceedings in GV-Global Virtual Conference*, number 1, 2013. 61

[104] James G Strathman, Kenneth Dueker, Jihong Zhang, and Timothy Williams. Analysis of design attributes and crashes on the oregon highway system. 2001. 57, 58

[105] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, 8(1):25, 2007. 36

[106] John A Swets. Form of empirical rocs in discrimination and diagnostic tasks: Implications for theory and measurement of performance. *Psychological bulletin*, 99(2):181, 1986. 19

[107] Marie C Taylor, DA Lynam, and A Baruya. *The effects of drivers' speed on the frequency of road accidents*. Transport Research Laboratory Crowthorne, 2000. 58

[108] LLC The Weather Company. Wunderground, 2015. 109

[109] P Turney. Types of cost in inductive concept learning. in: Workshop on cost-sensitive learning. In *Seventeenth International Conference on Machine Learning University*, pages 15–25, 2000. 17

[110] S Vigneswaran, A Arun Joseph, and E Rajamanickam. Efficient analysis of traffic accident using mining techniques. *International Journal of Software and Hardware research in engineering*, 2(3):110–118, 2014. 59

[111] Chao Wang, Mohammed A Quddus, and Stephen G Ison. Predicting accident frequency at their severity levels and its application in site ranking using a two-stage mixed multivariate model. *Accident Analysis & Prevention*, 43(6):1979–1990, 2011. 56, 57, 58

[112] Gary M Weiss and Foster Provost. Learning when training data are costly: the effect of class distribution on tree induction. *Journal of Artificial Intelligence Research*, 19:315–354, 2003. 18

[113] Peter Willett. Dissimilarity-based algorithms for selecting structurally diverse sets of compounds. *Journal of Computational Biology*, 6(3-4):447–457, 1999. 16

[114] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016. 10, 65

[115] Chaozhong Wu, Hu Lei, Ming Ma, and Xinping Yan. Severity analyses of single-vehicle crashes based on rough set theory. In *Computational Intelligence and Natural Computing, 2009. CINC'09. International Conference on*, volume 2, pages 59–62. IEEE, 2009. 62

[116] Yuanchang Xie, Dominique Lord, and Yunlong Zhang. Predicting motor vehicle collisions using bayesian neural network models: An empirical analysis. *Accident Analysis & Prevention*, 39(5):922–933, 2007. 57, 61

[117] Qing-Song Xu, Yi-Zeng Liang, and Yi-Ping Du. Monte carlo cross-validation for selecting a model and estimating the prediction error in multivariate calibration. *Journal of Chemometrics*, 18(2):112–120, 2004. 37

[118] Charles V Zegeer, Donald W Reinfurt, Joseph Hummer, Lynne Herf, and William Hunter. Safety effects of cross-section design for two-lane roads. *Transportation Research Record*, (1195), 1988. 57

[119] Mark H Zweig and Gregory Campbell. Receiver-operating characteristic (roc) plots: a fundamental evaluation tool in clinical medicine. *Clinical chemistry*, 39(4):561–577, 1993. 22