UNIVERSIDAD MIGUEL HERNÁNDEZ

FACULTAD DE CIENCIAS SOCIALES Y JURÍDICAS DE ELCHE

GRADO EN
ESTADÍSTICA EMPRESARIAL

MINERÍA DE DATOS: PRE-PROCESAMIENTO

TRABAJO FIN DE GRADO

2020-2021

ALUMNO: Christian Ledesma Mejías

DIRECTOR: Alex Rabasa

CO-DIRECTORA: Miriam Esteve

ÍNDICE

1.RESUMEN

El estudio del ámbito médico es un tema muy importante y de interés en un gran número de trabajos de investigación, porque gracias a estos y a la aplicación de la Minería de datos en ellos se consigue conocer ciertos puntos como las probabilidades de sufrir ciertas enfermedades dependiendo de las características de cada paciente, conocer posibles consecuencias en pacientes, medicamentos, etc.

En esta investigación se busca el objetivo de aplicar ciertas técnicas de Minería de datos (datos ausentes, datos atípicos, selección de atributos, factorización y árboles de decisión) a distintas bases obtenidas referentes al ámbito médico en concreto dedicadas a las enfermedades cardíacas. La finalidad de aplicar estas técnicas es analizar cómo afectan con respecto a los datos originales, comparando las precisión y exactitud de los distintos modelos obtenidos de una misma base de datos.

Palabras clave: Enfermedades cardíacas**,** Técnicas de Minería de datos, Datos Ausentes, Datos atípicos, Factorización, Selección de atributos, Árboles de decisión.

ABSTRACT

The study of the medical field is a very important interesting topic in a great number of research works, because thanks to these and the application of Data mining in them, it is posSible to know certain point such as the characteristics of each patient, to know posible consequences in patients, medicine, etc.

The objective of this research is to apply certaing data Mining techniques (Missing Data, Atypical Data, Attribute Selection, Factoring and Decision Trees) to different databases obteined from the medical field, specifically dedicated to heart diseases. The purpose of applying these techniques is to analyse how they affect the original data, comparing the precisión and accuracy of the different models obtained from the same databases.

Keywords: Heart Disease, Missing Data, Outliers**,** Data Mining Techniques, Factoring, Attribute Selection, Decision Trees.

2. **INTRODUCCIÓN**

Un accidente cardiovascular o cerebrovascular es aquella situación de peligro que se produce cuando el suministro de sangre de una parte del cerebro se corta o se reduce. Esta situación impide que el tejido cerebral reciba oxígeno, por lo tanto, las células del cerebro mueren, perdiendo su completa función.

Este tipo de accidentes puede afectar a distintas capacidades del cuerpo humano como:

- Dificultad para hablar y entender lo que otras personas están diciendo.
- Parálisis o entumecimiento de la cara, brazo o pierna.
- Problemas para ver
- Dolor de cabeza
- Dificultad para caminar

Según la Organización Mundial de la Salud el accidente cerebrovascular es la segunda causa principal de muerte a nivel mundial, responsable aproximadamente del 11% del total de muertes. Por ello, es muy importante buscar atención médica de inmediato cuando se detecte algún signo o síntoma de esta enfermedad. Estos síntomas pueden ser:

- Caída de un lado de la cara
- Impedimento levantar un brazo o cuando se levanta uno de ellos tiende a caer
- Arrastra palabras o habla de forma extraña

Cabe destacar que la mayoría de estas muertes están producidas en países con ingresos de salarios bajos y medios.

En los accidentes cerebrovasculares hay ciertos factores de riesgo productores de esta enfermedad que no pueden ser modificados, como son:

- Edad
- Raza
- Género
- Historial de accidentes
- Herencia o genética

Pero, por otro lado, hay ciertos factores de riesgo que si se pueden modificar:

- Presión sanguínea
- Colesterol
- Tabaco
- Diabetes
- Sedentarismo
- Obesidad
- Alcohol o drogas
- Ritmo cardíaco
- Lugar de residencia

- Temperatura
- Clima
- Factores económicos

Todos estos factores si se llevan a cabo pueden aumentar la aparición de estos accidentes cerebrovasculares, pero en este estudio se va a intentar ver cómo pueden llegar a afectar alguno de ellos.

Para este estudio se utiliza la Minería de datos. La Minería de datos es un campo de la estadística y es referido a ese proceso de encontrar anomalías, patrones o correlaciones en grandes bases de datos, de manera automática o semiautomática utilizando una variedad de técnicas.

La Minería de datos en el sector de la salud como en otros sectores afecta a distintos ámbitos. El primero es el ámbito científico. Este permite ayudar a determinar causas de ciertas patologías o identificar poblaciones de riesgo. Con esto, por ejemplo, es posible poder detectar de forma precoz enfermedades en los pacientes. El segundo es el ámbito de la gestión de la salud que ayuda en la toma de decisiones, optimizar recursos y a detectar prácticas fraudulentas.

Principalmente de todo lo que abarca la Minería de datos en esta investigación se va a emplear el preprocesamiento de datos. Esto es algo vital en el estudio de un dataset, debido a que, su función es recoger los datos y traducirlos en una información utilizable. Es un proceso muy importante, porque si su procesamiento no se hace correctamente afectará negativamente a los resultados finales obtenidos.

Esta etapa de preprocesamiento es la etapa en la cual se mejoran los datos, porque, los dataset suelen tener datos brutos que limpiar, observaciones con datos faltantes, errores por la recogida de datos, etc.

Por lo tanto, si no se aplica el pre-procesamiento una base de datos puede no estar lista para el estudio y, por lo tanto, disminuirá la calidad de la Minería de datos y la precisión del estudio.

2.1 **OBJETIVOS DEL TFG**

El objetivo principal a conseguir con el actual estudio es la utilización de la Minería de datos, en específico, la aplicación de la etapa de preprocesamiento en el dataset para así influir en los datos y mejorar la calidad de los mismos.

En la etapa de preprocesamiento se incluyen técnicas como son el tratamiento de outliers, tratamiento de valores nulos, discretizaciones de variables numéricas y selección de atributos más relevantes para el estudio de la variable objetivo. Con esto se pretende poder mejorar la calidad del estudio y con ello mejorar la precisión del mismo.

2.2 OBJETIVOS PERSONALES

Los objetivos personales a conseguir en este trabajo son los siguientes:

- Aplicar los distintos conocimientos adquiridos durante la formación en el Grado de Estadística Empresarial en un caso práctico.

3. ESTADO DE LA CUESTIÓN Y MARCO TEÓRICO

3.1 OUTLIERS

El tratamiento de los valores atípicos o denominado outliers en inglés es una de las etapas de pre-procesamiento más utilizadas. Esta consiste en eliminar, sustituir o estudiar de la base de datos las muestras irrelevantes o redundantes consiguiendo de esta manera reducir el coste computacional y aumentar la precisión del modelo.

Los datos atípicos pueden deberse a distintos errores, por ello pueden clasificarse en diversos tipos:

- Observaciones debido a errores en la construcción de la base de datos, errores en codificación o en la entrada de datos. Se tratan como ausentes y se eliminan de la muestra.
- Observaciones que no representan a ningún conjunto de la población, estas son observaciones extraordinarias con explicación, por lo tanto, pueden ser eliminadas.
- Observaciones cuyos valores representan a conjuntos de la población, pero con valores únicos. Estas observaciones habría que estudiarlas.
- Observaciones cuyos valores no pueden explicarse, lo mejor en estos casos es estudiar en análisis con estas observaciones y sin ellas para comprobar el alcance de estas variables.

Debido a que tenemos diversos tipos hay que estudiar minuciosamente su eliminación o modificación, puesto que, si se elimina observaciones no procedentes de errores puede afectar a las inferencias que se realizan en el modelo, produciendo sesgos y dando muestras engañosas.

Uno de los métodos más utilizado en todos los ámbitos es el test de Tukey, este método toma la diferencia entre el primer cuartil Q1 y el tercer cuartil Q3, o lo que es lo mismo el rango intercuartílico. El valor atípico se encuentra a 1.5 veces de distancia de los cuartiles o también podemos encontrarlo a una distancia de 3, esto podemos observarlo en el gráfico de cajas de la Figura 1.

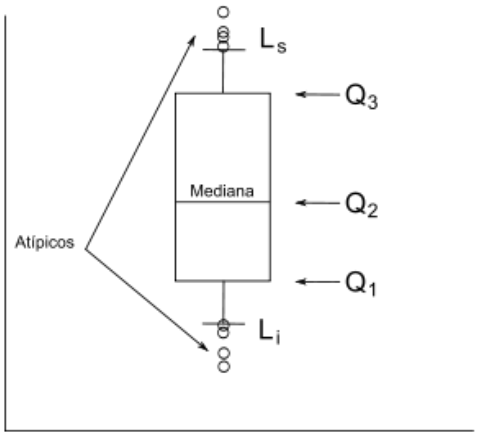

Fig. 1. Gráfico de cajas

Fuente: Elaboracion propia

3.2 DATOS AUSENTES

Los datos faltantes (missing values) se pueden definir como aquellos valores no disponibles en el dataset pero que son importantes para el análisis. Estos suelen aparecer casi siempre en cualquier base de datos.

Es un problema bastante frecuente en los estudios médicos. En los ensayos médicos los missing values introducen sesgos importantes y comprometen la efectividad del estudio. Además, esto influye en la confiabilidad de los resultados, debido a que, los datos faltantes pueden estar relacionados con factores importantes sobre medicamentos del paciente, de enfermedades o pronósticos.

Se pueden encontrar diferentes tipos de datos faltantes y distintas razones por las que ocurren. Además, de que es muy importante comprender la aleatoriedad de los missing values, es decir, los datos faltantes pueden afectar a todas las observaciones por igual o, por otro lado, se crean por una razón específica que afecta solo a unos pocos. Esto introduciría un sesgo en el dataset.

Por ellos, se pueden detectar dos tipos de datos ausentes:

- Prescindibles: Estos son cuando no causan una pérdida representativa en la muestra
- No prescindibles: Estos se producen cuando eliminándolos si que producen una pérdida de representatividad en la muestra.

Los datos prescindibles pueden darse en el caso de observaciones que tenga un alto porcentaje de datos ausentes, en este caso se eliminaría. Pero también tenemos el caso de las variables, esta si también presenta un alto porcentaje de datos ausentes habrá que valorar la importancia que tiene, así como valorar la posibilidad de reemplazar la información faltante con un contenido similar.

Y, en los datos no prescindibles es adecuado estudiar la aleatoriedad de estos:

- Completamente aleatorios: Al presentar un grado tan alto de aleatoriedad el hecho de falta de una observación no estará relacionado con los valores ausentes ni existentes.
- Aleatorios: Los datos ausentes presentan aleatoriedad, pero existen diferencia entre las categorías de una variable. Ciertas características registradas suelen poder explicar la distribución de los datos ausentes.
- No aleatorios: En este caso, nos encontramos cuando se detectan patrones sistemáticos en el proceso de datos ausentes. Probablemente los datos faltantes estén relacionados con datos no observados.

Como lidiar con esta falta de información es un factor importante en el estudio. Podemos actuar omitiendo las variables con los datos faltantes, omitir los individuos con un alto porcentaje de datos ausentes o por último reemplazar los datos ausentes con una estimación.

Si finalmente decidimos no eliminar los casos o variables y decidimos reemplazar los datos ausentes tenemos tres formas:

- En primer lugar, podemos sustituir por la media, mediana o moda. La media se utilizará cuando tenemos unos datos uniformes, la mediana cuando detectamos algún dato atípico que se salga de la muestra y la moda es la única que podría usarse en variables categóricas.

- En segundo lugar, se puede sustituir por una constante, que en este caso debería ser estimada en investigaciones previas o teóricas.

- Y, por último, se puede reemplazar los datos faltantes por regresión.

En todo caso, es conveniente limitar estos métodos cuando son variables con pocos datos ausentes, y cuando el porcentaje de datos ausentes sea mayor decidir si eliminar la variable, reformular o cambiar las variables, aunque estos últimos métodos aumentan el coste.

3.3 DISCRETIZACIONES

La técnica de discretización [1] es un procedimiento de procesamiento de datos en el cual se trata de transformar los datos, es decir, se buscar transformar las variables numéricas en variables categóricas. Esto se debe a que muchos algoritmos están enfocados a tratar con datos cualitativos y aunque muchos tratan con datos cuantitativos, el aprendizaje suele ser menos eficiente y eficaz.

Los algoritmos principales de discretización se pueden dividir en dos bloques, llamados métodos supervisados y métodos no supervisados.

En primer lugar, los métodos de discretización **no supervisados** transforman las variables categóricas en numéricas, pero no tienen en cuenta la información del

objetivo(class). Los dos métodos más comúnmente usados sin supervisión son los de “Mismo ancho” y “Misma frecuencia”.

- **Mismo ancho**: Este algoritmo busca dividir en *k* intervalos del mismo tamaño. El ancho de estos intervalos y los límites se miden con las siguientes ecuaciones:

$$w = (max - min) / k$$

$$min + w, min + 2w, ... , min + (k-1)w$$

- **Misma frecuencia**: Este algoritmo también divide los datos en *k* intervalos, donde cada intervalo tendrá aproximadamente la misma cantidad de valores.

Para ambos métodos la mejor forma de determinar *k* es observando el histograma y determinar los distintos intervalos.

Y, los métodos de discretización **supervisada** transforman las variables numéricas en categóricas, pero en este caso sí que utilizan la referencia de la información objetivo. Algunos de los métodos más usados son los de discretización entrópica, en concreto discretización entrópica de Fayyad y Irani.

- **Discretización entropíca**: La entropía o el contenido de información se calcula a través de la etiqueta de la clase. Con esto se consigue que la mayoría de datos que tienen cierta semejanza estén en la misma etiqueta de clase. Este método tiene la característica de encontrar la separación con la obtención máxima de información.

En el sector de la medicina, se recogen una gran cantidad de atributos numéricos, ya sea referentes a pacientes, enfermedades, medicamentos, etc. Por ello, discretizar variables es muy importante para poder estudiar adecuadamente algunas variables. También dependerá del tipo de modelo que se quiera realizar.

3.4 SELECCIÓN VARIABLES INFLUYENTE

Algo muy habitual e importante en el preprocesamiento de datos es la selección de atributos, es decir, seleccionar aquellos atributos con una mayor importancia para el estudio. Hay muchos mecanismos que consiguen seleccionar los atributos y ordenarlos según el grado de importancia en relación con la variable objetivo, con esto se puede seleccionar los más adecuados para la construcción de los modelos.

Además, esto nos aporta un aprendizaje más rápido, puesto que, muchas variables en estudios son redundantes o no aportan información sobre la variable objetivo y esta es una manera de aumentar la velocidad, simplificar el modelo aportando un mayor conocimiento y entendimiento y, por lo tanto, mejorando el estudio.

A la hora de seleccionar las variables se puede distinguir los atributos entre relevantes, irrelevantes y redundantes. Un atributo se considera irrelevante cuando no aporta nada al estudio de la variable objetivo, por lo tanto, sería innecesario seleccionar este tipo de variables. Los atributos se consideran redundantes cuando se tiene variables muy similares, es decir, cuando dos variables aportan la misma información a la variable objetivo o que es lo mismo, tienen una alta correlación entre ellas se consideran redundantes y se puede eliminar una de ellas.

Y, por último, un atributo es relevante cuando aporta cierta información de importancia para predecir la variable objetivo. Estos atributos tras su eliminación la precisión del modelo disminuye y los resultados empeoran.

Además, existen tres tipos de métodos de selección de atributos que se usa para inferir en el modelo:

- **Técnicas de filtro:** Los métodos de filtro se basan en la relevancia entre los atributos, eliminando aquellos menos relevantes. La selección de los atributos se calcula según sus propiedades sin tener alguna relación con el algoritmo de aprendizaje, son totalmente independientes.
- **Técnicas de envoltorio (wrapper):** Los métodos de envoltorio o envoltura trata de un algoritmo de Minería de datos donde evalúa un subconjunto de atributos, sobre un conjunto de entrenamiento. Con este método se obtiene un subconjunto de atributos adecuados, pero suelen ser muy lentos al tener que ejecutar el algoritmo de aprendizaje varias veces.
- **Técnicas embebidas:** Los métodos embebidos o incrustados son aquellos donde el algoritmo de aprendizaje incluye la propia búsqueda del subconjunto óptimo de variables. Este suele tener un menor coste con respecto al método wrapper.

3.5 CLASIFICACIÓN

Se considera la clasificación como el proceso de identificar una nueva categoría sobre el conjunto de datos que contiene las observaciones. Se busca asociar dichas observaciones a una clasificación predefinida. Este se trata de un proceso supervisado, es decir, es un proceso donde se conoce a priori la clase a la que pertenecen.

En medicina como en otros ámbitos en muchos casos se utiliza el análisis e interpretación de los datos de forma manual, pero esto es una técnica muy lenta, con un alto coste económico y muy subjetiva, de hecho, cuando el volumen de datos es sumamente alto y sobrepasa la capacidad humana es irrealizable sin la ayuda de las herramientas tecnológicas adecuadas.

Para ello, es posible aplicar distintos métodos para solucionar estos problemas. En este caso, los algoritmos, también llamados clasificadores se usan como su nombre indica para ayudar en la clasificación.

Hay una cantidad muy alta de métodos clasificadores, pero los arboles de clasificación son unos de los métodos más utilizados en todos los ámbitos, siendo en la medicina el método más utilizado gracias a su fácil interpretación.

Alguno de los árboles de clasificación [3] más utilizados en medicina son el ID3, J48 y el Naive Bayes. El método ID3 y J48 son muy semejantes, tanto que el J48 es un descendiente del ID3. Estos construyen un árbol a partir de las diferencias existentes entre los datos, maximizando la información obtenida. Por último, el método Naive Bayes es un método más simple de clasificación, que corresponde a un modelo de atributos independientes. Pero, en definitiva, son de los más utilizados debido a su sencillez, precisión y el bajo coste de ejecución.

Una vez se aplican los clasificadores hay que medir su calidad. La matriz de confusión es una herramienta que permite visualizar la exactitud (accuracy) obtenida en el algoritmo. Esta matriz de confusión tiene un aspecto como la vista en la tabla 1.

	PREDICCIÓN	
	POSITIVO	**NEGATIVO**
POSITIVO	Verdadero Positivo (Vp)	Falso Positivo (Fn)
NEGATIVO	Falso Negativo (Fn)	Verdadero Negativo (Vn)

Tabla. 1. Matriz de confusión

Fuente: Elaboración propia

La calidad la medimos con el accuracy, mencionado anteriormente. El accuracy es la proporción de acierto en la clasificación con respecto a los datos originales de la base de datos. Por lo tanto, la exactitud la medimos dividiendo las predicciones correctas entre el número total de predicciones.

$$Precision = \frac{Vp + Vn}{Vp + Fp + Vn + Fn}$$

3.4.1 ÁRBOLES DE DECISIÓN

En el caso de la medicina, sufren con una gran cantidad de datos como los síntomas, pacientes, los padecimientos involucrados, etc. Por esto, algo muy importante en este ámbito es contar con herramientas que les permita tratar estos datos sintomatológicos de los pacientes, para así, con todos los datos anteriores más los nuevos análisis poder tomar un diagnostico con una precisión muy superior y poder darle el tratamiento adecuado cuanto antes. Por esto, una herramienta utilizada para la predicción y clasificación de grandes cantidades de datos son los arboles de decisión.

Un árbol de decisión es un modelo de predicción que determina una probabilidad estadística o un curso de acción. Este tipo de modelo se basa en obtener la decisión final siguiendo unas condiciones que comienzan en la raíz de árbol hasta las hojas, pudiendo elegir una de todas ellas. Estos modelos son muy utilizados en economía, inteligencia artificial, procedimientos legales, etc y hasta se usan en procedimientos médicos. Por eso, estos modelos son probablemente el modelo de clasificación más utilizado y popular.

Un árbol de decisión basa su conocimiento a través de un proceso de aprendizaje inductivo. Este se representa gráficamente por un conjunto de nodos, hojas y ramas. Los nodos principales o raíz es el atributo donde se inicia el proceso y está situado arriba del todo, este está conectado con diversos nodos internos que corresponden a distintas variables predictoras. Estos nodos internos pueden ser o decisiones finales, llamados nodo hoja o nodos finales, o un nodo hijo del cual saldrán otros nodos internos. Todo esto se ilustra en la Figura 2.

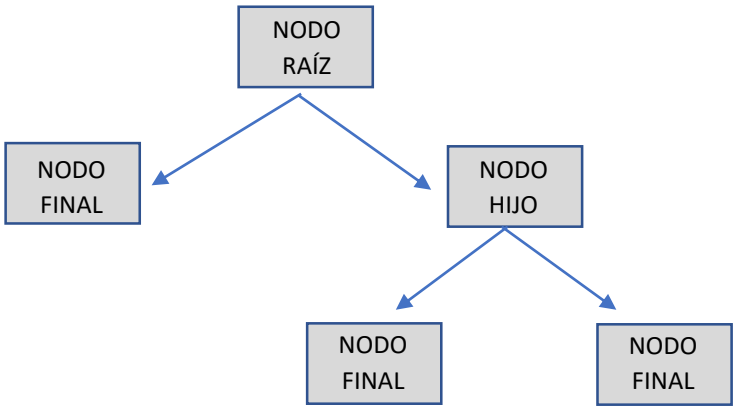

Fig. 2. Modelo árbol

Fuente: Elaboración propia

Este tipo de métodos tiene una serie de ventajas muy evidentes, pero también tiene sus desventajas. En primer lugar, la ventaja principal que destaca es lo simple que es de entender y de interpretar, al ser un gráfico tan simple y obvio facilita mucho el trabajo. Otras ventajas que se pueden destacar son: la validez

que tiene, debido a que, se pueden utilizar tanto variables cualitativas como cuantitativas, la posibilidad de poder agregar en cualquier momento nuevas variables de forma simple, no necesita datos muy complejos, puede obtener múltiples resultados, utiliza un modelo de caja blanca, es decir, la respuesta obtenida es fácilmente justificable. Por otro lado, las desventajas que se pueden destacar son su inestabilidad, es decir, al poder añadir nuevas variables cualquier pequeño cambio supone un árbol de decisión totalmente distinto, no se puede garantizar que el resultado obtenido, el árbol en cuestión, sea el óptimo, además, si se tiene variables categóricas con distintos niveles es posible que la información obtenida se incline por aquellas con mayores niveles.

Dentro de los árboles de decisión podemos destacar dos tipos, los arboles de clasificación y los arboles de regresión. La diferencia entre estos dos modelos es que, los arboles de clasificación son aquellos donde la predicción obtenida pertenece a la clase de los datos mientras que en los arboles de regresión la predicción que se obtiene es un número real, es decir, sus resultados son posibles y continuos. Además, se utilizan los árboles de clasificación cuando nuestra variable objetivo es discreta y, al contrario, se emplean los arboles de regresión cuando es continua.

Hay muchos algoritmos en los arboles de decisión, pero entre los más importantes podemos encontrar el **ID3**, **Cart**, **C4.5** (sucesor de ID3), **ACR** (arboles de clasificación y regresión), **Naive Bayes**, etc.

4. **HIPÓTESIS DE PARTIDA**

En el presente TFG se plantea el efecto que puede tener el preprocesamiento de un dataset en la predicción final del estudio, ya sea, produciéndole una mejoría o por otro lado afectando a los datos. Por esto, surge la pregunta siguiente:

¿El pre-procesamiento mejora la predicción del modelo?

En un dataset podemos encontrar ruido en los datos, datos incompletos, datos inconsistentes, variables poco adecuadas, patrones o reglas poco útiles… Por ello, un preprocesamiento de los datos suele tener una mejoría en la calidad de estos consiguiendo así una mayor resolución en los resultados. Esto suele producir dataset más pequeños que los originales pudiendo mejor su eficacia gracias a la Minería de datos.

Además, antes de analizar los datos en el preprocesamiento también entra la tarea de preparar un buen dataset. Esto genera que tengamos unos datos de calidad, por ello, es muy importante la recogida de los datos. Una buena obtención de los datos puede disminuir los errores y los posteriores métodos para limpiar la base de datos, ahorrando tiempo y coste.

Por todo esto, como dijo Dorian Pyle [2], el propósito fundamental del preprocesamiento de los datos es manipular y transformar los datos sin procesar

para que el contenido de la información incluido en el conjunto de datos pueda exponerse o hacerse más fácilmente.

Por lo tanto, en este TFG se pretende comprobar esta hipótesis de partida y analizar como se ve afectada la precisión y accuracy del modelo por el preprocesamiento efectuado en el dataset.

5. **METODOLOGIA**

Para describir las diferentes técnicas de preprocesamiento vamos a usar la base de datos que se detalla a continuación:

5.1 BASE DE DATOS

Este conjunto de datos "**healthcare-dataset-stroke-data.csv"** trata de estudiar los accidentes cerebrovasculares que pueden sufrir las personas, por ello, el objetivo de este conjunto de datos es intentar predecir según ciertas características específicas de los pacientes si sufren o no un accidente cerebrovascular.

La recopilación de datos es un factor muy importante a la hora de realizar un estudio, debido a que, interfiere directamente con la calidad de los datos. Realizar un enfoque sistemático adecuado te ayuda a reunir y medir información en un contexto más completo y preciso.

La recogida de datos tiene diferentes métodos y técnicas como entrevistas personales o telefónicas, cuestionarios, observación, etc. Esta base de datos es totalmente confidencial con usos meramente académicos.

Entre los atributos recopilados se incluye diversas características de los pacientes tanto de ámbitos sociales, como demográficos, como médicos. Además de la variable objetivo de estudio. La recogida de datos se realizó a un total de 5110 pacientes y los atributos recogidos son los siguientes:

1. **id** – Identificador único para cada paciente (numérico)
2. **género** - género del paciente (nominal: Female = Femenino, Male = Masculino u Other = Otro)
3. **edad** – edad del paciente (numérico: de 0.08 a 82, las edades como 0.64 hacen referencia a 0.64 años, es decir, 234 días dividido 365 días que tiene un año 0,64 años.)
4. **hypertensión** – hipertensión del paciente (binario: "1" si tiene hipertensión, "0" en caso contrario)
5. **cardiopatía** – cardiopatía del paciente (binario:"1" si tiene enfermedad cardiaca, "0" en caso contrario)

6. **Acasado** – si alguna vez ha estado casado/a (binario: “No” o “Sí”)
7. **trabajo** – tipo de trabajo (nominal: children=”niño/a”, Govt_job =”trabajador del gobierno”, Never_worked=”nunca ha trabajado”, Private=”información privada” y Self-employed=”trabajador por cuenta propia”)
8. **residencia** – residencia del paciente (binario: rural o urbano)
9. **Nglucosa** – nivel medio de glucosa en sangre en miligramo/decilitro (numérico: de 55,1 a 272)
10. **bmi** – índice de masa corporal (numérico: de 10 a 97)
11. **Sfumador** – estado fumador del paciente (nominal: “anteriormente fumador”, “nunca fumador”, “fumador” o “desconocido”)

Y, por último, el atributo a predecir:

12. **stroke** – accidente cerebrovascular del paciente (binaria: “1” si ha sufrido un accidente cerebrovascular o “0” en caso contrario)

Tras conocer los distintos atributos que componen el dataset es necesario estudiar las características de los mismos. Una forma de estudiar y conocer las variables con las que estamos tratando es realizando un análisis exploratorio.

Este es un análisis que trata de estudiar los gráficos y los estadísticos de cada una de las variables, esto se puede realizar a variables de forma individual o de forma conjunta con otras variables. Este tipo de análisis nos permite conocer las características de las variables además de su distribución y así poder conocer si el dataset presenta outliers o valores atípicos, valores nulos, patrones en los datos, etc.

Posteriormente, se realiza un análisis descriptivo de todos los atributos. En él, se analizará las relaciones entre las distintas variables y cómo influyen en la variable de estudio *stroke.*

Para llevar a cabo todos los análisis y en general, el presente estudio, se utiliza el software R-Studio (Versión 1.14.1717). Este programa es uno de los más utilizados tanto en universidades como por investigadores, además de por ser gratuito también aporta una curva de aprendizaje sencilla y es un buen programa para realizar análisis estadísticos y gráficos.

Este programa cuenta con una gran cantidad de librerías que sirven de ayuda para realizar distinta función como la importación de datos, ajustes y evaluaciones de modelos, representaciones gráficas, análisis de descriptivos, mejoras visuales, etc. Los distintos paquetes utilizados son los siguientes:

- library(dplyr)
- library(pander)
- library(ggplot2)
- library(kableExtra)

- library(tidyverse)
- library(ggpubr)
- library (xlsx)
- library(arules)
- library(Rcpp)
- library(FSelector)
- library(data.table)
- library(Boruta)
- library(rpart)
- library(rpart.plot)
- library(caret)

Además, el harware empleado en este caso no es muy potente, debido a que, los cálculos necesarios para este estudio no requieren de demasiada potencia. Las especificaciones son las siguientes:

- Intel (R) Core ™ i5-33330 CPU @ 3.00GHz (4 CPUs), ~3.2GHz
- Memoria RAM de 16,0GB
- Sistema operativo Windows 10 Home 64 bits

A continuación, se presentan las distintas técnicas utilizadas para el pre-procesamiento del dataset. Estas técnicas se realizan al dataset completo, es decir, se realiza a las 5110 observaciones registradas.

5.2 DATOS AUSENTES

En primer lugar, se analiza la ausencia de los datos ausentes. Para analizar esta técnica se usa los gráficos por la simplificación y la facilidad del mismo.

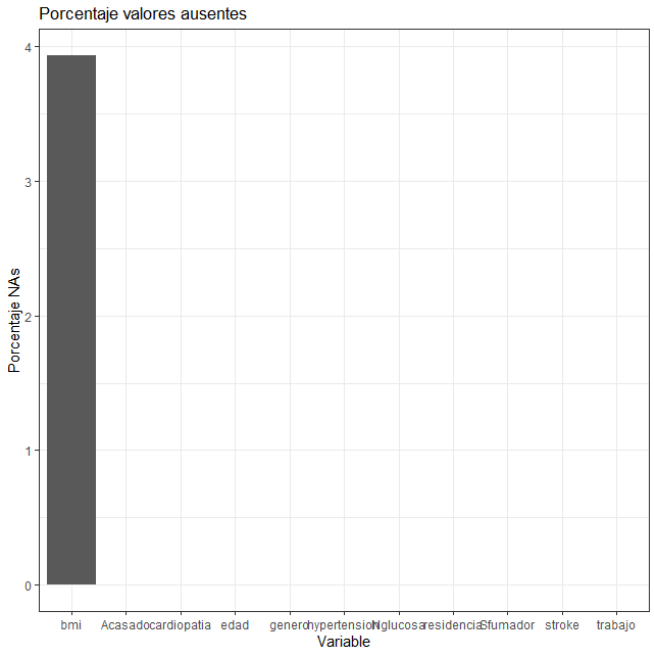

Fig. 3. Porcentaje valores ausentes

Fuente: elaboración propia

DATASET COMPLETO	TOTAL
BMI	4909
NA'S	201

Tabla. 2. Total valores ausentes bmi

Fuente: elaboración propia

Como se observa en la figura 3 la única variable con datos ausentes es "bmi", índice de masa corporal. Esta variable contiene un total de 201 datos ausentes, como se contempla en la tabla 2, de las 5110 observaciones totales. Esta cantidad no es suficiente como para eliminar la variable por no aportar información, pero es conveniente analizar más en profundidad esta ausencia.

MALE	TOTAL
BMI	2011
NA'S	104

Tabla. 3. Total valores ausentes hombres

Fuente: elaboración propia

FEMALE	TOTAL
BMI	2897
NA'S	97

Tabla. 4. Total valores ausentes mujeres

Fuente: elaboración propia

De los 201 datos ausentes observamos en las tablas 3 y 4 como 104 son chicos y 97 son chicas, es decir, no hay prácticamente diferencia entre las clases de una variable. Por lo tanto, se consideran datos ausentes aleatorios y se procede a rellenar esta falta de datos.

En este caso, se selecciona reemplazar los datos ausentes por la media, debido a que, los datos de “bmi” son uniformes y la media va a ser más representante para el dataset. Además, se sustituye a cado uno por su media, es decir, se diferencia entre la media de los chicos con la de las chicas, puesto que, entre sexos puede haber diferencias claras en masa corporal y así se puede conseguir una muestra más representativa.

En el caso de los hombres se sustituye por *28.20338* y en el caso de las chicas por *28.62099.* Por lo tanto, como se observa en la figura 3 se han reemplazado todos los datos ausentes en el dataset.

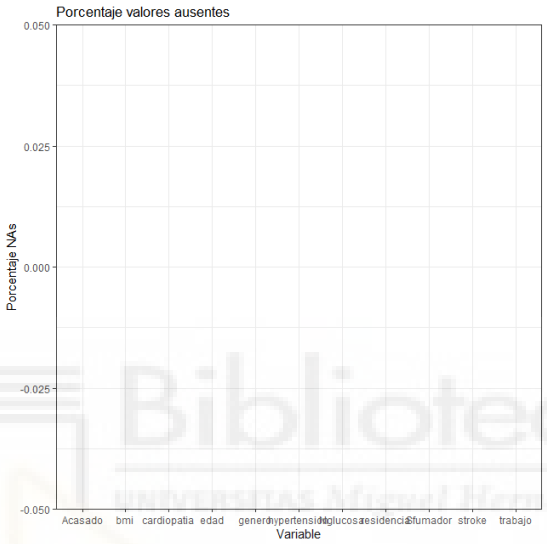

Fig. 4. Porcentaje valores ausentes

Fuente: elaboración propia

5.3 OUTLIERS

A continuación, se procede a estudiar los datos atípicos o outliers. En este caso, se analizan todas las variables numéricas para encontrar los outliers. Las tres variables numéricas disponibles en el dataset son “edad”, “Nglucosa” y “bmi”. Se utiliza para detectar los outliers los gráficos de caja o bigotes.

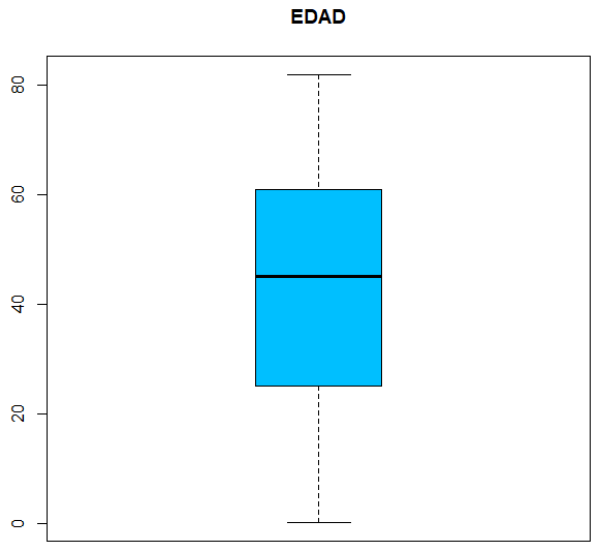

Fig. 5. Detección de outliers en edad

Fuente: elaboración propia

En primer lugar, se analiza la variable “edad”. En la figura 5 se ve como esta no presenta ningún dato atípico, debido a que, no se observa ningún valor fuera del límite inferior ni superior. Además, la mediana corta la caja en dos partes desiguales, es decir, la parte inferior en más grande que la superior, esto nos dice que la cantidad de valores en la parte inferior en mayor que en la superior, hay más personas jóvenes que mayores en el estudio.

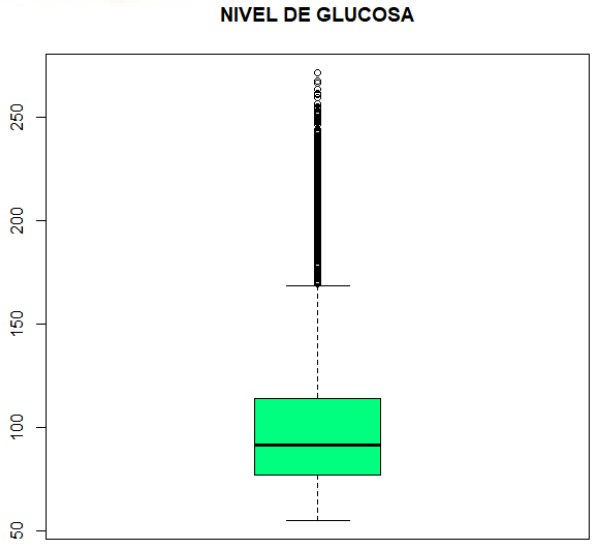

Fig. 6. Detección de outliers en nivel de glucosa

Fuente: elaboración propia

En el gráfico anterior, figura 6, se procede a analizar el nivel de glucosa en sangre. Esta variable presenta una gran cantidad de outliers superando el nivel superior de la caja.

En la figura 7, se amplía el estudio de la variable “nivel de glucosa”, porque es conveniente analizarla en profundidad para decidir si es buena opción eliminar tantas observaciones o no.

En este caso, se enfrenta a la variable “stroke”, diferenciando entre si tienen un accidente o no. Y como se observa al enfrentarlas la categoría “No” presenta una gran cantidad de datos atípicos. Además, si se decide distinguir entre distintos niveles de glucosa, es decir, por un lado, se considera unos niveles superiores a 100 mg/dl y, inferiores a 150 mg/dl (tabla 5) y por otro lado niveles superiores a 150 mg/dl (tabla 6), esto se puede considerar unos niveles “preocupantes” o “altos”, se ve en las tablas como hay una subida en accidentes cerebrovasculares cuanto mayor es el nivel de glucosa, pero no se considera algo relativamente suficiente como para tener en cuenta. Además, esto es algo que tendría que valorar un experto en la materia para considerar altos niveles de glucosa o no.

En todo caso, se decide eliminar los datos atípicos de la variable nivel de glucosa al ser observaciones que puede interferir y además para que el modelo sea capaz de discriminar mejor.

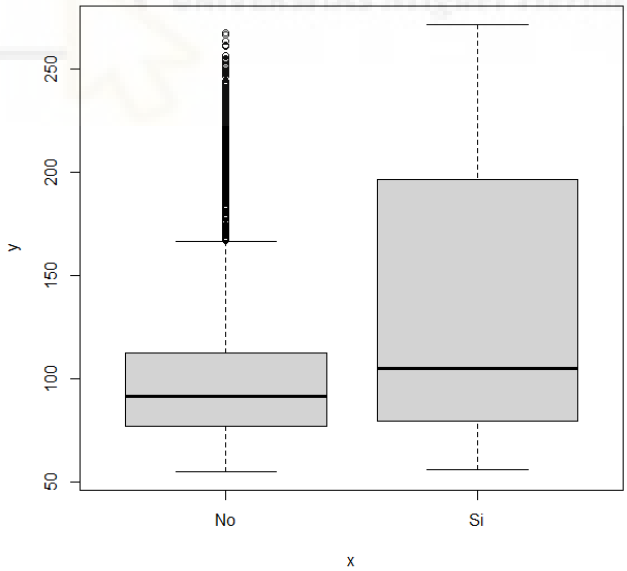

Fig. 7. Outliers en edad

Fuente: elaboración propia

Nglucosa >100 y Nglucosa <=150	TOTAL
Sí	47
No	1201

Tabla. 7. Total nivel de glucosa

Fuente: elaboración propia

Nglucosa >150	TOTAL
Sí	90
No	640

Tabla. 8. Total nivel de glucosa

Fuente: elaboración propia

Tras la eliminación de los datos atípicos para la variable “nivel de glucosa”, se obtiene el siguiente gráfico (figura 7):

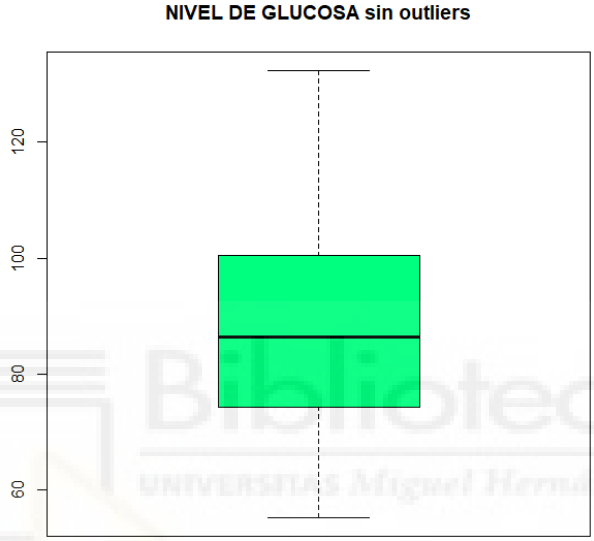

Fig. 8. Gráfico de cajas de nivel de glucosa

Fuente: elaboración propia

Por último, en la figura 8 se analiza el gráfico de cajas para la variable “bmi”. Esta igual que la anterior presenta datos atípicos por la parte superior de la caja, pero en este caso en menor medida. Estas observaciones se proceden a eliminarlas para no estropear la muestra ni el estudio.

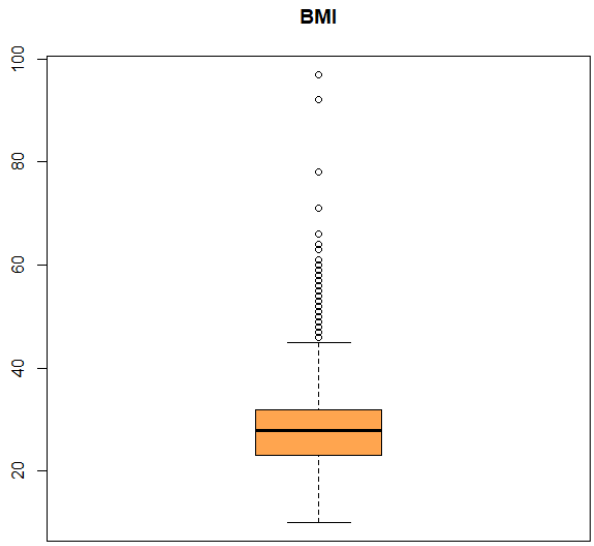

Fig. 9. Outliers en bmi

Fuente: elaboración propia

Tras la eliminación de los datos atípicos, el gráfico de cajas de la variable “bmi” se obtiene como en la figura 9:

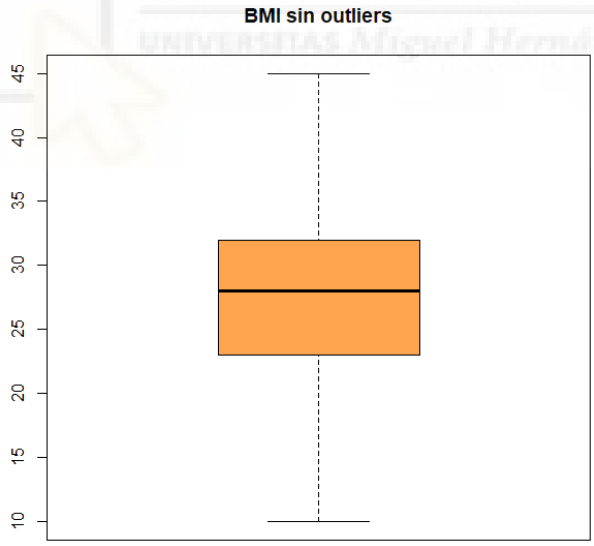

Fig. 10. Gráfico de cajas sin outliers

Fuente: elaboración propia

Tras eliminar los datos atipicos de las variables numéricas Nivel de glucosa(Nglucosa) y indice de masa corporal(bmi) el dataset actual dispone de 4125 observaciones de estudio, de las 5110 iniciales.

5.4 ANÁLISIS DESCRIPTIVO

El análisis descriptivo del dataset ayuda a hacer síntesis de la información para arrojar precisión, sencillez y orden en los datos. Con esto, además, se lleva a cabo el estudio de las relaciones existentes entre las distintas variables que componen el dataset, por lo tanto, es muy necesario del análisis así conoceros como afectan están variables a la variable objetivo “stroke”.

Este análisis es el mejor método para conocer cómo se relacionan las variables y como afectan de verdad, además de ayudar a la comprensión del tema de estudio como facilitar la interpretación de los resultados. Pero, por otro lado, este análisis puede producir sesgos si no se analiza toda la información además de que si la investigación y recogida de información no fue la apropiada esto producirá errores en el resultado general.

En este análisis, se usan tanto medios gráficos como el estudio de los estadísticos de las variables.

A continuación, se presentan las distintas técnicas utilizadas para conocer las características de las variables, así como sus relaciones con las variables objetivo. En primer lugar, se analizarán las variables cualitativas y posteriormente las cuantitativas. Este análisis se realiza a las 4125 observaciones.

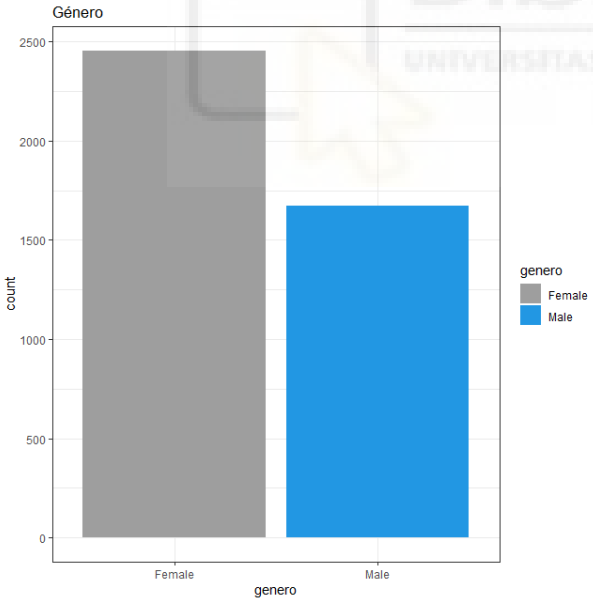

Fig. 12. Número de pacientes según el genero

Fuente: elaboración propia

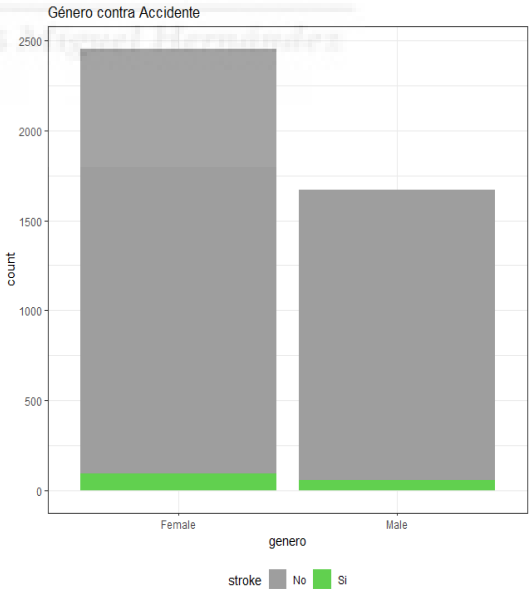

Fig. 13. Número de pacientes con accidente según el género

Fuente: elaboración propia

En la figura 12 analizamos la variable “genero”. Esta variable se divide en dos tipos de clase o categorías (femenino y masculino). Se observa a simple vista como en el estudio hay una gran disposición más de mujeres que hombres.

Por otro lado, en la figura 13, se analiza la variable genero contra la variable de estudio “stroke”. En esta se observa como la relación se mantiene, es decir, al tener más mujeres en el estudio, hay más mujeres con accidentes cerebrovasculares, por lo tanto, en este momento no se puede certificar si influye el hecho del genero del paciente en los accidentes cerebrovasculares.

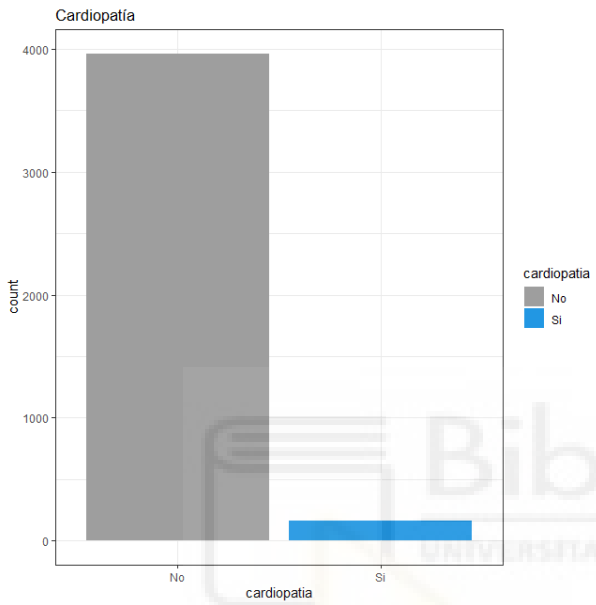

Fig. 14. Número de pacientes según si sufren una cardiopatía

Fuente: elaboración propia

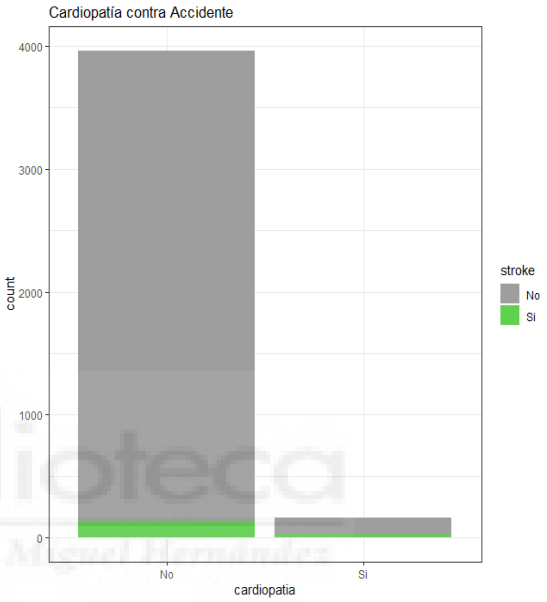

Fig. 15. Número de pacientes con accidente según si sufren una cardiopatía

Fuente: elaboración propia

La siguiente variable, cardiopatía, se observa en la figura 14 como el número de los pacientes del estudio que no sufren una “cardiopatía” es muy superior aquellos que si la sufren. Siendo en total, 3962 pacientes que no contra 163 que sí.

Por otra parte, el porcentaje de accidentes en aquellas personas con cardiopatía es muy superior, porque viendo la figura 15 se observa como de esos 163 pacientes que tienen una cardiopatía 20 sufren un accidente cerebrovascular (el 12.3% sufren un accidente), mientras que de los 3962 que no tienen una cardiopatía solo 133 sufren el accidente (el 3.3% sufren un accidente), por lo tanto, la probabilidad de sufrir un accidente cerebrovascular teniendo una cardiopatía es bastante superior a no sufrir ninguna cardiopatía.

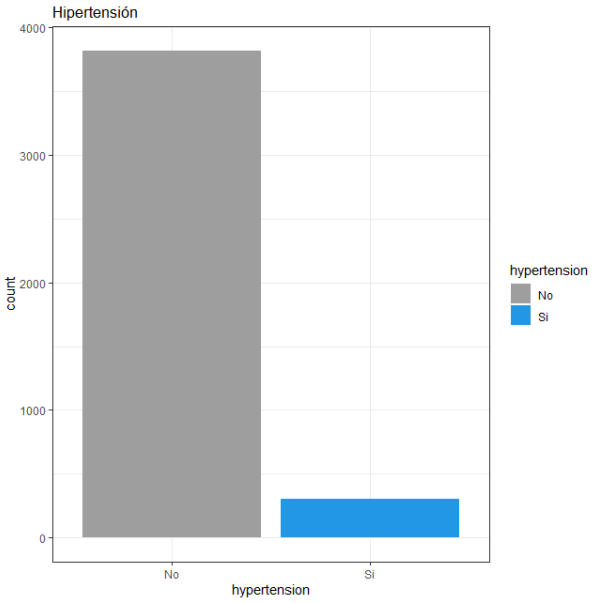

Fig. 16. Número de pacientes en función de hipertensión

Fuente: elaboración propia

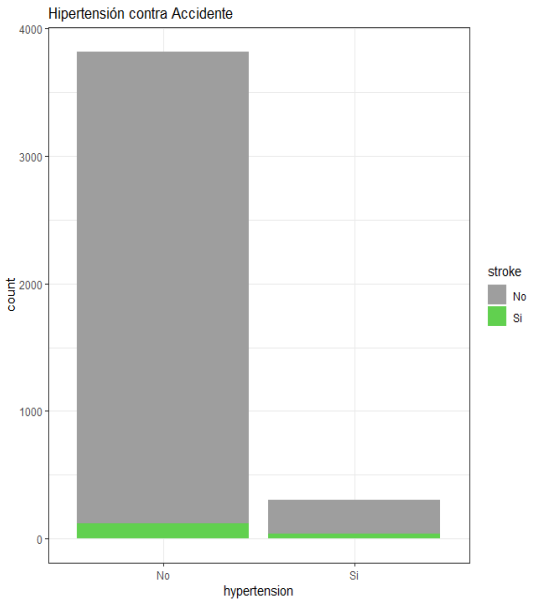

Fig. 17. Número de pacientes con accidente en función de hipertensión

Fuente: elaboración propia

En cuanto a la variable “hipertensión” (figura 16) hay una diferencia clara entre sus clases, La mayoría de pacientes no sufren de hipertensión, en concreto 3819 de los 4125 no sufren hipertensión contra 306 que sí. En contra punto, semejante a la cardiopatía anteriormente analizada, aquellos pacientes que sufren de hipertensión tienen un porcentaje superior de accidentes cerebrovasculares. Como se observa en la figura 17, de los 3819 pacientes que no tienen hipertensión solo 118 sufren un accidente cerebrovascular, es decir, el 3% de las personas que no tienen hipertensión sufren un accidente cerebrovascular. Y, por otro lado, de los 306 que si sufren hipertensión 66 tienen un accidente, es decir, el 11.4% de ellos sufren un accidente cerebrovascular.

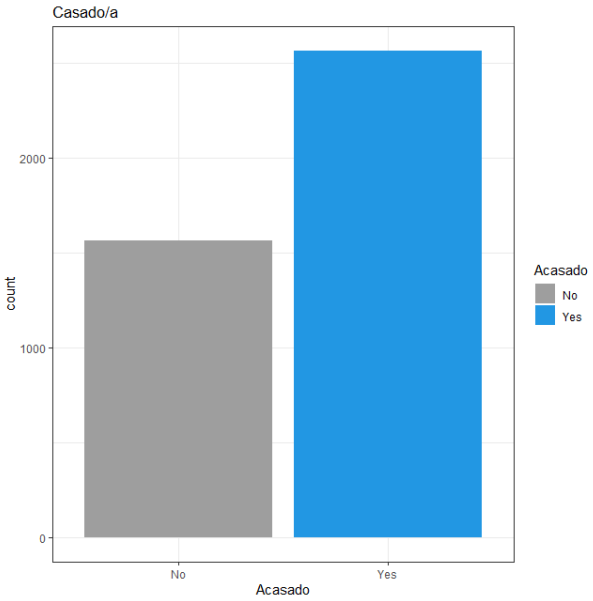

Fig. 18. Número de pacientes en función de su estado civil

Fuente: elaboración propia

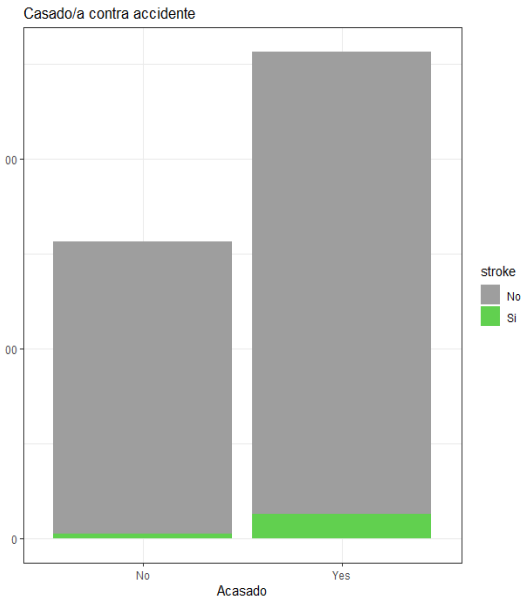

Fig. 19. Número de pacientes con accidente en función de su estado civil

Fuente: elaboración propia

En la figura 18 se analiza la variable “Acasado/a”, en esta se observa como la mayoría de pacientes están casados, en total 2563 pacientes están casados contra 1562 que no lo están.

Al enfrentar esta variable contra la variable de estudio “stroke” (figura 19) se contempla como no se detecta ninguna implicación en los accidentes el que un paciente este casado o no. Se obtiene que de los 2563 pacientes casados 129 sufren un accidente cerebrovascular, en total el 5% de los pacientes, mientras que aquellos que no están casados (1562 pacientes) solo 24 si tienen accidente, es decir, el 1.5%.

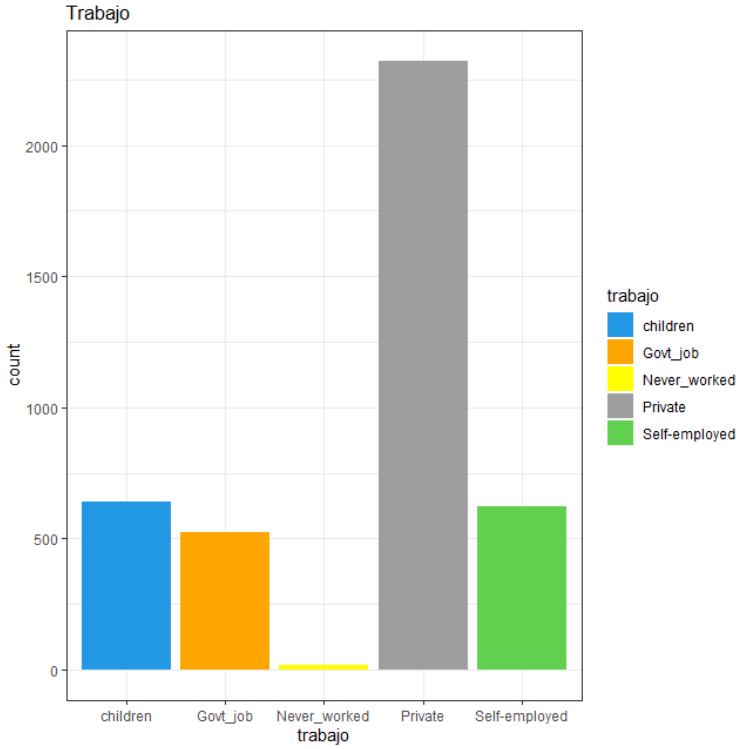

Fig. 20. Número de pacientes en función de su oficio

Fuente: elaboración propia

TRABAJO	TOTAL
Niño	639
Gobierno	524
Trabajo	19
Privado	2321
Cuenta propia	622

Tabla. 9. Número total por oficio

Fuente: elaboración propia

TRABAJO	Accidente
Niño	2
Gobierno	20
Trabajo	0
Privado	87
Cuenta propia	44

Tabla. 9. Número total de accidentes por oficio

Fuente: elaboración propia

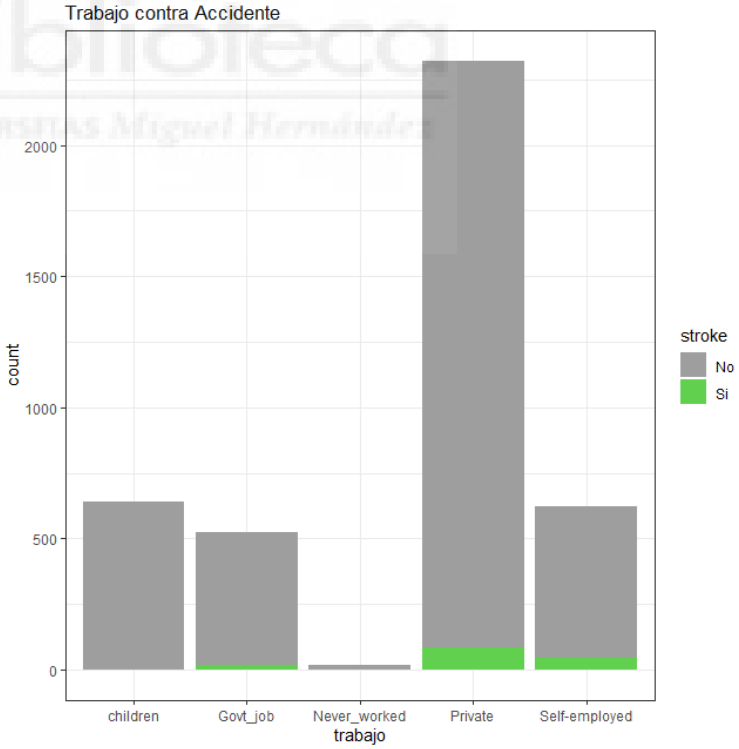

Fig. 21. Número de pacientes con accidente en función de su oficio

Fuente: elaboración propia

En las figuras 20 y 21 se analizan los trabajos de los pacientes. Esta variable se divide en cuatro clases (niños, trabajador del gobierno, nunca ha trabajado, sector privado, trabajador por cuenta propia). En primer lugar, en la figura 20 se observa como la gran mayoría de los pacientes trabaja en el sector privado, luego hay dos puestos prácticamente similares como son “niños”, “trabajadores por cuenta propia”. Luego, por otro lado, siguen los trabajadores del sector público y, por último, unos pocos que “nunca han trabajado”.

Y, al analizar la figura 21 donde se enfrenta la variable “trabajo” frente a “stroke” lo único a destacar es el bajo porcentaje de niños que sufren accidentes cerebro vasculares, de 639 niños solo 2 de ellos han sufrido un accidente cerebro vascular, por lo tanto, es algo a destacar. Por otro lado, el resto de trabajos tienen unos accidentes similares siendo aquel con un mayor porcentaje entre trabajadores y accidentes son aquellos que trabajan por cuenta propia con un porcentaje de 7%.

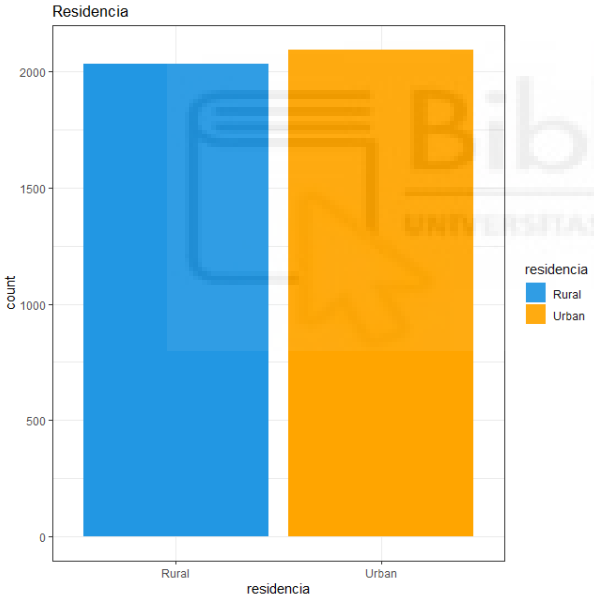

Fig. 22. Número de pacientes en función de su residencia

Fuente: elaboración propia

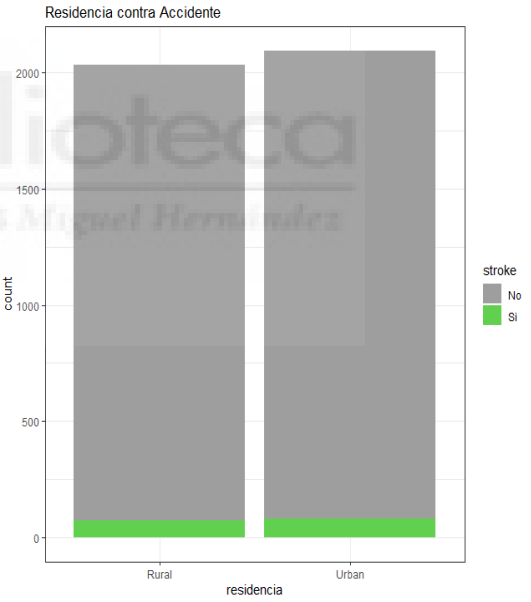

Fig. 23. Número de pacientes con accidente en función de su residencia

Fuente: elaboración propia

La siguiente variable (residencia) se analiza en las figuras 22 y 23. En este caso, se divide la variable entre “zona rural” y “zona urbana”. Hay poco que destacar, prácticamente el 50% de los pacientes viven en zona urbana y el otro 50% en zona rural. Y al enfrentarla con la variable objetivo “stroke” vemos como prácticamente los resultados son similares, por lo tanto, no se detecta ninguna relevancia entre el lugar de residencia en los accidentes cerebrovasculares.

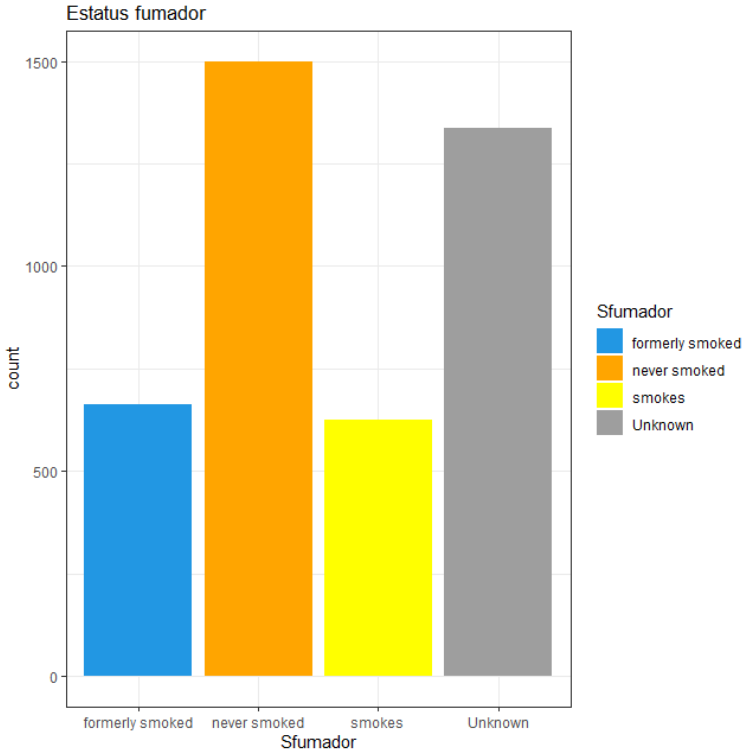

Fig. 24. Número de pacientes en función de su estado fumador

Fuente: elaboración propia

Como última variable cualitativa (Sfumador) se analiza en las figuras 24 y 25, en este caso, se divide en cuatro clases como “anteriormente fumador”, “nunca fumador”, “fumador” o “desconocido”. En primer lugar, la clase donde hay más pacientes es en la de “nunca fumador”, seguido de “desconocido” y, por último, prácticamente igualitarias “fumador” y “anteriormente fumador”.

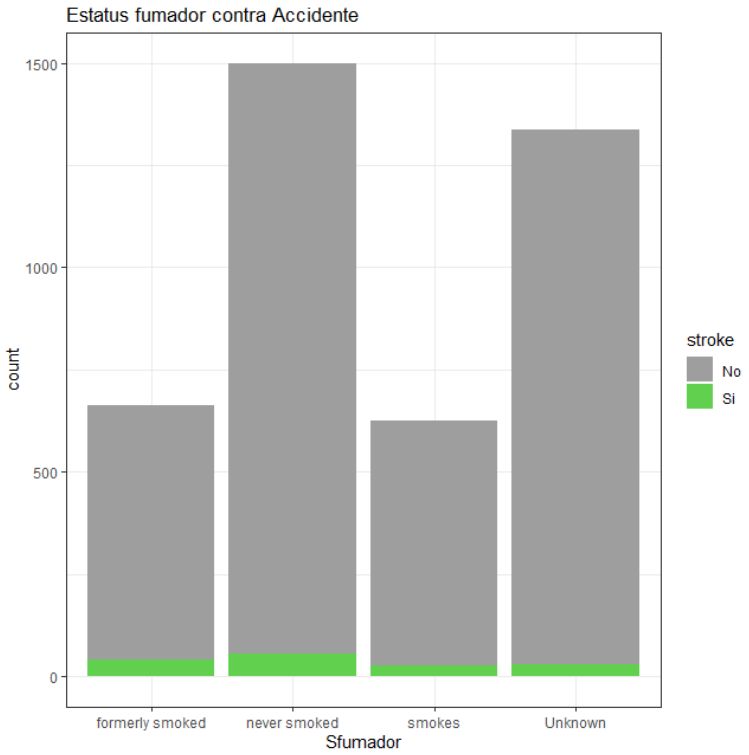

Fig. 25. Número de pacientes con accidente en función de su estado fumador

Fuente: elaboración propia

Por otro lado, al enfrentarlo contra “stroke”, se observa como la acción de fumar afecta a los accidentes cerebrovasculares, porque los que nunca han fumado de 1499 solamente 56 han sufrido un accidente cerebro vascular contra 663 pacientes que anteriormente fumaban y 42 que han sufrido un accidente cerebrovascular y 625 personas que fuman y 25 accidentes cerebrovasculares. Es decir, los que nunca han fumado tienen un 3.8% de pacientes con accidente cerebrovascular, anteriormente fumado tiene 8,6% de accidente cerebrovascular, fumador tiene un 6.8% de accidente cerebrovascular y por último, los pacientes que se desconoce su estado fumador tiene un 2.3% de accidente cerebrovascular.

A continuación, se analizan las variables cuantitativas. Estas son “edad”, “nivel de glucosa” y “bmi”.

MEDIA	DESVIACIÓN	MIN	MAX	N	NAs
40.88	22.59	0.08	82	4215	0

Tabla. 10. Estadísticos edad

Fuente: elaboración propia

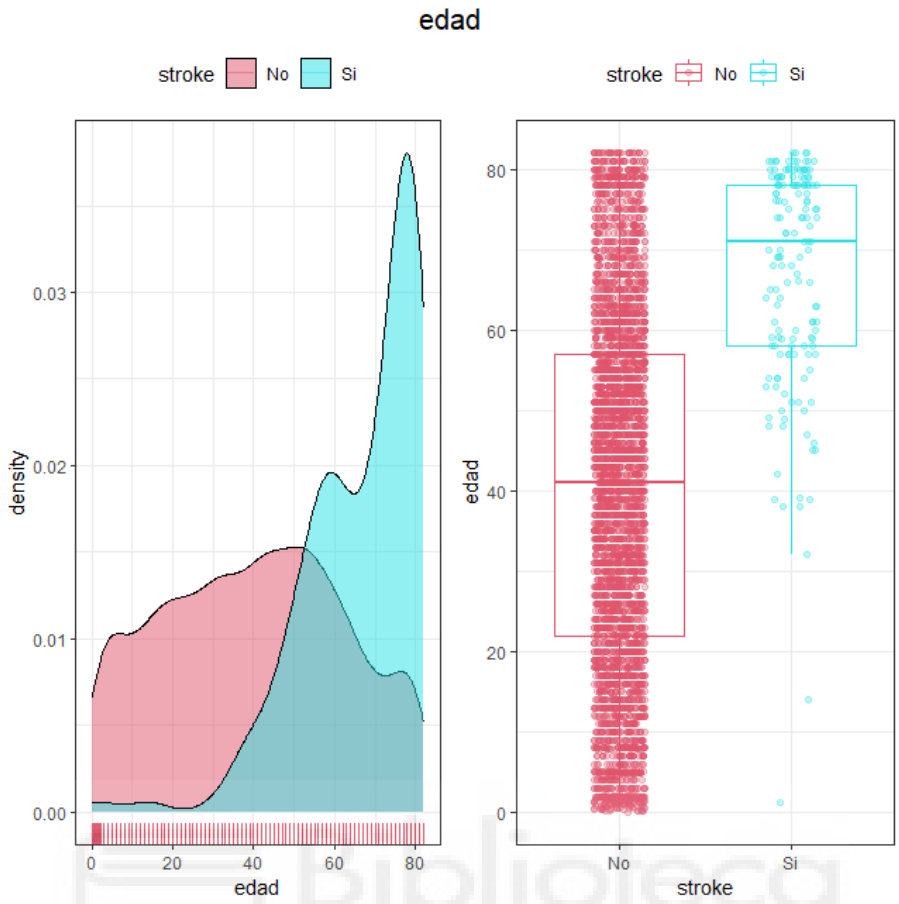

Fig. 26. Distribución de la edad de los pacientes en función del accidente cerebrovascular

Fuente: elaboración propia

En primer lugar, en la tabla 10 se contempla la variable edad. En esta se observa como el valor mínimo de edad es de 0.08, este se traduce en 0.08 años o lo que es lo mismo 30 días dividido entre 365 días que tiene un año. Y el valor máximo de 84 años. Por otro lado, el promedio de edad de los pacientes es de 40.88 años y como vemos la desviación en edad es de 22.59.

Siguiendo con la figura 26 se analiza la variable edad gráficamente, en esta se observa como está muy descompensada entre pacientes que han sufrido un accidente cerebrovascular y aquellos que no, es decir, los pacientes observados a partir de los 40 años, aproximadamente, comienza a subir de una forma importante los accidentes cerebrovasculares. Esto hace que sea conveniente discretizar la variable edad y tratar con intervalos de edad, debido a que, antes de los 40 años prácticamente no hay accidentes cerebrovasculares y a partir de esa edad se disparan.

Además, los gráficos de cajas apoyan esto, porque viendo la gráfica de los “Sí” observamos como la mayoría de pacientes con accidentes están en las edades más altas.

MEDIA	DESVIACIÓN	MIN	MAX	N	NAs
87.93	18.07	55.12	132.5	4125	0

Tabla. 11. Estadísticos nivel de glucosa

Fuente: elaboración propia

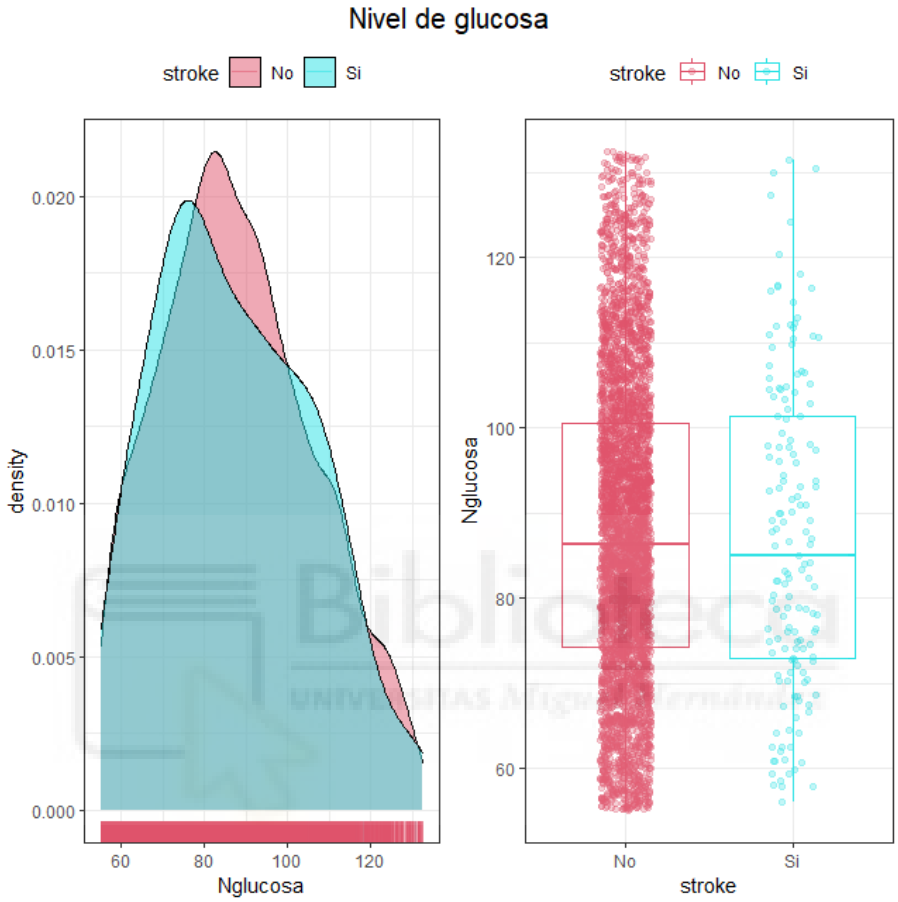

Fig. 27. Distribución del nivel de glucosa de los pacientes en función del accidente cerebrovascular

Fuente: elaboración propia

Para la variable “nivel de glucosa” en la tabla 11, se observa como esta va desde el nivel mínimo de 55.12 y el nivel máximo 132.5. Además, de una media de la variable en 87.93 con una desviación de los datos de 18.07.

Observando la figura 27, se analiza como la variable nivel de glucosa no tiene unos niveles donde realmente se vea que afecta a los accidentes cerebrovasculares, además, en el gráfico de cajas analizando el level “SI” se ve como las observaciones están muy repartidas entre los distintos niveles. Aun así, sería de gran ayuda discretizar para poder observar adecuadamente como están repartido los accidentes cerebrovasculares y si de verdad tienen alguna consecuencia un mayor nivel de glucosa.

MEDIA	DESVIACIÓN	MIN	MAX	N	NAs
27.3	6.51	10	45	4125	0

Tabla. 12. Estadísticos bmi

Fuente: elaboración propia

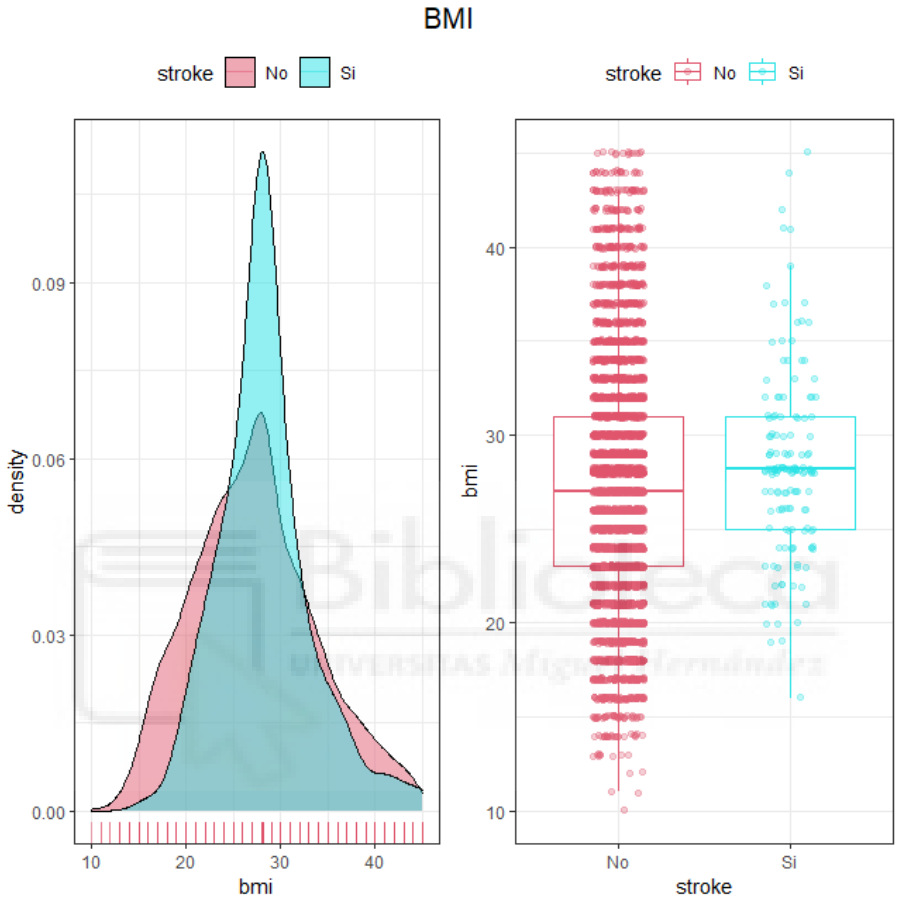

Fig. 28. Distribución del índice de masa muscular de los pacientes en función del accidente cerebrovascular

Fuente: elaboración propia

Y, por último, la variable “bmi” donde se analiza en la tabla 12. Esta va desde un mínimo de 10 hasta un 45 de índice de masa corporal. La media de la variable está en 27.3 y con una desviación de 6.51, esta es la variable con la desviación más baja, es decir, la mayoría de datos están en torno a los mismos valores. Además, esto también se puede observar gráficamente en la figura 28, donde se aprecia como los datos están cercanos a la misma franja. También, cabe añadir como la mayoría de “SI” en accidente cerebrovascular están entorno a la media de la variable, donde también se encuentran la mayoría de datos.

Observando el gráfico, sería de ayuda discretizar debido a los tramos donde apenas encontramos accidentes cerebrovasculares y viendo el pico claro donde se concentran prácticamente todos los accidentes cerebrovasculares.

Para terminar de analizar las variables presentes en el estudio, aunque se ha ido tratando a lo largo del análisis de las variables frente a cada una es conveniente conocer la variable objetico “stroke”, es decir, si los pacientes tienen accidente cerebrovascular o no por su cuenta.

En este caso, como se observa en la figura 29 el estudio cuenta con un total de 3972 pacientes que no han sufrido un accidente cerebrovascular, contra un total de 153 pacientes que si han llegado a sufrirlo. Para facilitar la información, este se traduce en 96.3% de pacientes que no lo han sufrido contra un simple 3.7% de que si (véase fig. 30). La gran mayoría de los pacientes no han sufrido un accidente cerebrovascular.

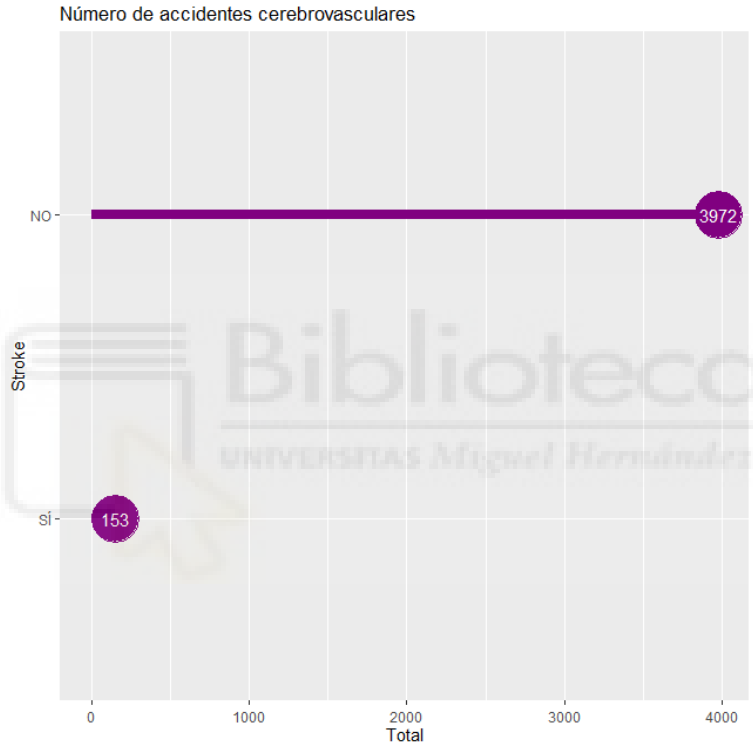

Fig. 29. Número de pacientes según si han sufrido o no un accidente cerebrovascular

Fuente: elaboración propia

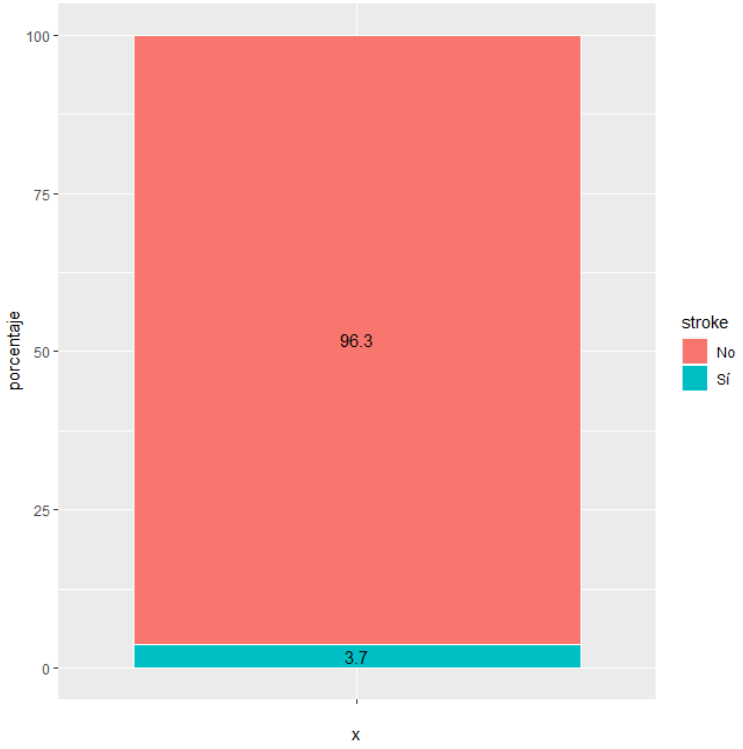

Fig. 30. Porcentaje de pacientes según si han sufrido o no un accidente cerebrovascular

Fuente: elaboración propia

5.5 DISCRETIZACIONES

Tras observar la situación de las variables cuantitativas del dataset se procede a discretizar dichas variable, porque como se ha analizado anteriormente esto mejoraría su aportación al estudio haciéndolas mucho más sencillas de analizar a la vez de añadirles mayor significado con respecto a la variable objetivo.

Además, en el posterior modelo se va a trabajar con modelos de árboles de clasificación, por lo tanto, buscamos tener variables discretas en el modelo y tanto “Nglucosa” como “bmi” no cumplen con esto. Por ello, discretizar va a ayudar a trabajar con las variables numéricas.

Para poder discretizar las variables numéricas se trabaja con la librería ‘arules’ en RStudio, esta librería implementa un algoritmo para la identificación de “ítems” frecuentes ayudando así a discretizar. Pero, en específico, se emplea la función “discretize” que incorpora la librería “arules”, con esta función se obtiene la discretización de la variable deseada en los niveles que considera oportuno, pero esto se puede modificar en caso que se quiera discretizar en más factores.

En primer lugar, se analiza la variable edad. En la figura 31 se puede apreciar la distribución de esta variable viéndose como hay una gran variedad de edad siendo entre los 40 y 50 años donde se aprecia una mayor frecuencia.

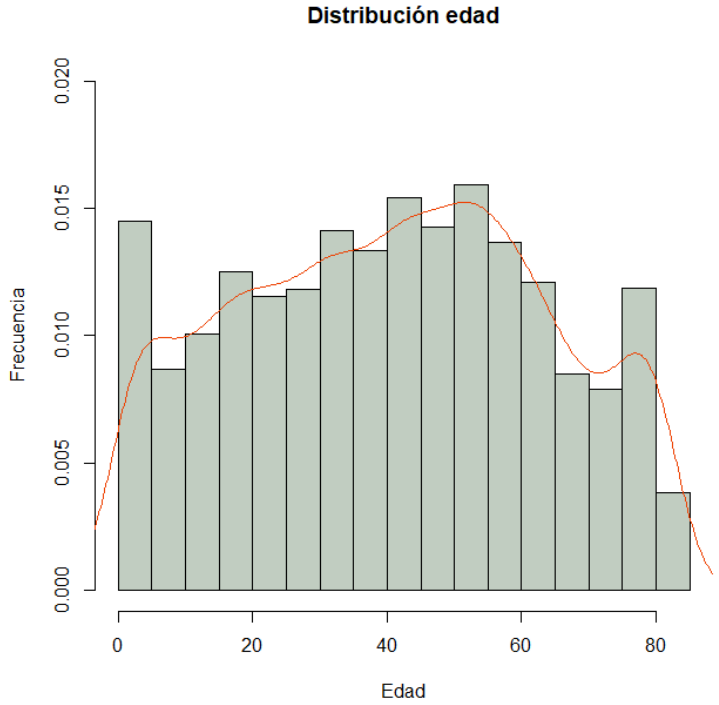

Fig. 31. Distribución de la edad de los pacientes

Fuente: elaboración propia

Tras ver y conocer la variedad de esta variable y tras anteriormente analizar que era adecuado discretizar se procede a utilizar la librería de RStudio para discretizar la variable. En este caso, como se puede observar en la tabla 5 la función discretizada la variable edad en 3 categorías de la siguiente manera:

EDAD	VALOR
[0.08,29.7)	Adolescentes
[29.7,53)	Adultos
[53,82)	Mayores

Tabla. 13. Discretización en tres categorías del atributo edad

Fuente: elaboración propia

Como RStudio ha devuelto una discretización en tres categorías se analiza la posibilidad de discretizar en dos categorías. Obteniendo la variable de la siguiente manera (tabla 6):

EDAD	VALOR
[0.08,42)	Jóvenes
[42,82)	Mayores

Tabla. 14. Discretización en dos categorías del atributo edad

Fuente: elaboración propia

La siguiente variable tratada y discretizada es “nivel de glucosa”. En la figura 32 se observa igual que en la variable anterior la distribución de la variable.

En esta podemos observar como hay unos niveles de glucosa donde claramente están la mayoría de los pacientes, siendo esto en los valores cercanos a 80dl/gr. Por otro lado, hay una gran variedad de distintos niveles de glucosa. Esto hace que sea muy oportuno discretizar la variable, ya que puede solucionar muchos problemas.

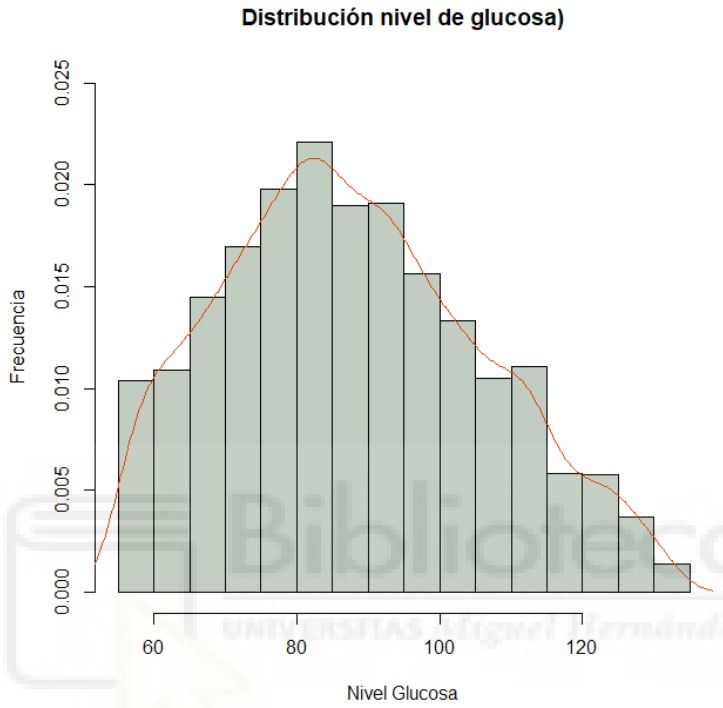

Fig. 32. Distribución del nivel de glucosa de los pacientes

Fuente: elaboración propia

Al discretizar esta variable en RStudio se obtiene una discretización en tres categorías, donde se dividen como se puede ver en la tabla 7.

Nivel de glucosa	Valor
[55.1,78.7)	Bajo
[78.7,95.1)	Intermedio
[95.1,132)	Alto

Tabla. 15. Discretización en tres categorías del atributo nivel de glucosa

Fuente: elaboración propia

Y, por otro lado, también se procede a discretizar en dos categorías como se ha hecho en la variable anterior, para ver como se comportan así.

Nivel de glucosa	Valor
[55.1,86.3)	Bajo
[86.3,132)	Intermedio/Alto

Tabla. 16. Discretización en dos categorías del atributo nivel de glucosa

Fuente: elaboración propia

Y, por último, se analiza la variable “bmi”. Esta igual que en las anteriores en la figura 23 se puede observar su distribución. En este caso la mayoría de pacientes tienen unos índices de grasa entre 20 y 40.

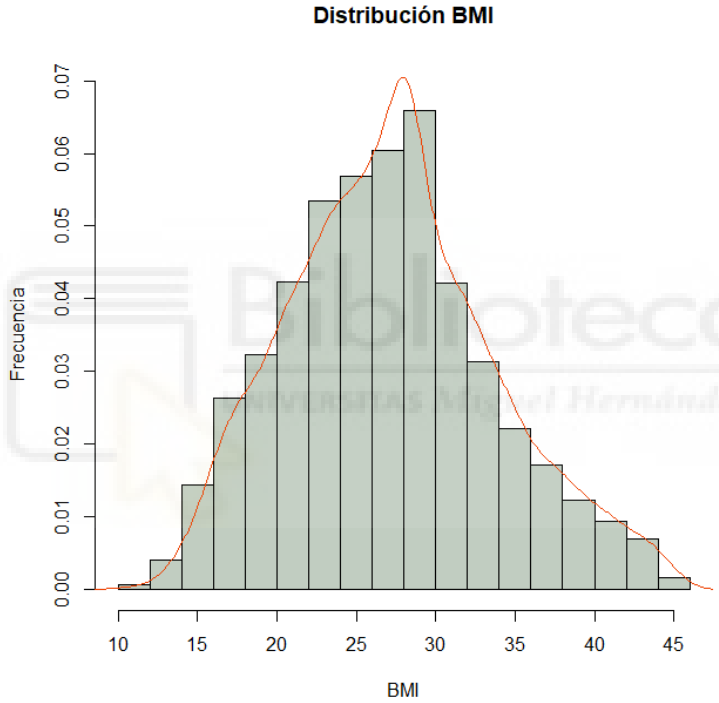

Fig. 33. Distribución del índice de masa corporal de los pacientes

Fuente: elaboración propia

En este caso, al implementar la función discretize también como en las anteriores, discretiza la variable “bmi” en tres categorías, viendo como en la tabla siguiente (tabla8):

BMI	Valor
[10,24)	Normal
[24,29)	Elevado
[29,45)	Muy elevado

Tabla. 17. Discretización en tres categorías del atributo bmi

Fuente: elaboración propia

Y, en la siguiente tabla se puede observar la discretización de la variable en dos categorías (tabla 9):

BMI	Valor
[10,27)	Normal
[27,45)	Elevado

Tabla. 17. Discretización en dos categorías del atributo bmi

Fuente: elaboración propia

5.6 VARIABLES INFLUYENTES

La selección de los atributos más influyentes para la variable objetivo es una de las etapas más importantes de la Minería de datos y, por lo tanto, tiene gran influencia posteriormente en el modelo y en su aprendizaje. En este proceso, se utilizan algoritmos automáticos para la selección y evaluación de las características.

En este caso, para su análisis se ha empleado una librería denominada "Boruta" en RStudio. Esta librería utiliza un algoritmo de envoltura para la selección de las características relevantes. Este algoritmo encuentra dichas características relevantes comparando su importancia con los atributos originales y así se seleccionan las más relevantes para la variable objetivo.

Además, con esta librería se obtiene la importancia de cada una de una forma fácil, simple e intuitiva, siendo así una función de RStudio muy empleada.

attribute	meanImp	medianImp	minImp	maxImp	normHits	decision
edad	23.5590828	23.9305234	20.6704998	26.1378529	1	Confirmed
Acasado	17.3243415	17.3022985	14.3161764	22.3807468	1	Confirmed
bmi	10.9906273	10.6787774	8.27302635	14.4616007	1	Confirmed
cardiopatia	6.43659935	6.76453627	2.78145826	9.14762157	1	Confirmed
hypertension	6.37512755	6.48554506	3.66167225	8.93200591	1	Confirmed
trabajo	3.64897774	3.64240767	0.65260986	6.75879314	0.76086957	Confirmed
Sfumador	1.3814454	1.47330742	-1.76519124	3.64281963	0.26086957	Rejected
Nglucosa	1.26395913	1.11463975	-1.01868526	4.70752773	0.2173913	Rejected
genero	0.35203957	0.60804882	-1.28305457	1.43502763	0	Rejected
residencia	-0.13683592	-0.09403941	-2.43029825	1.47061977	0	Rejected

Tabla. 18. Ranking de importancia de los atributos en función de Boruta

Fuente: elaboración propia

Una vez realizado el análisis con la librería “Boruta”, RStudio devuelve la importancia de las variables respecto a la variable objetivo. El resultado obtenido nos dice que en el dataset se dispone de seis atributos importantes y cuatro atributos no importantes. Entre los importantes como se observa en la figura 28 están “edad”, “hipertensión”, “cardiopatía”, “Acasado”, “trabajo” y “bmi”. Mientras que entre los atributos que considera no importantes se componen de “genero”, “Nglucosa”, “residencia” y “Sfumador”.

En este caso, no hay ninguna variable en duda, llamado atributo tentativo, por lo tanto, no hace falta refinar el modelo boruta.

Tras observar la importancia de los atributos sin discretizar se analizan de nuevo, pero en este caso se añaden todos los atributos, es decir, todas las variables cualitativas, las cuantitativas sin discretizar y las cuantitativas discretizadas (se utilizan las discretizadas en tres categorías).

En este caso, se analizan con la función Boruta un total de catorce atributos. Al añadir los nuevos atributos la importancia de las variables se ve afectada, aumentando el número de estas. A los atributos anteriormente anotados como importante se añade “cardiopatía”, “Nglucosa” como también las discretizaciones de “edad”, “bmi” y como “Nglucosa”. Además, el atributo “Sfumador” se queda como tentativos, teniendo que refinar el modelo boruta.

Finalmente, tras refinar el modelo se obtiene la siguiente selección de atributos importantes para la variable objetivo stroke. Dejando esto como la tabla siguiente (tabla 19):

attribute	meanImp	medianImp	minImp	maxImp	normHits	decision
Acasado	18.7118266	18.6738911	15.3069147	22.5719968	1	**Confirmed**
edad	16.3656695	16.3859587	12.5093408	18.8018741	1	**Confirmed**
bmi	15.6862883	15.7091062	13.2905883	18.2742898	1	**Confirmed**
bmi1	13.7366148	13.857041	10.8381883	15.8473871	1	**Confirmed**
edad1	12.5072242	12.6885982	10.1140845	15.270896	1	**Confirmed**
Nglucosa	6.46353885	6.42010007	3.85648764	9.23217686	1	**Confirmed**
Nglucosa1	6.35656251	6.31217175	4.43087591	9.49129623	1	**Confirmed**
hypertension	4.3146168	4.35852519	0.93987787	7.66348437	0.87878788	**Confirmed**
cardiopatia	3.77600375	3.91168443	0.61948053	6.02168884	0.8989899	**Confirmed**
trabajo	2.83071945	2.86194935	-0.38128146	5.93829844	0.71717172	**Confirmed**
Sfumador	1.97615879	1.76297939	-0.7474173	6.16546771	0.42424242	**Confirmed**
genero	0.18873931	-0.07816407	-1.54815474	2.95613992	0.01010101	**Rejected**
residencia	-0.04526588	-0.3398435	-1.6374712	3.1067915	0.01010101	**Rejected**

Tabla. 19. Ranking de importancia de los atributos tras discretizar en función de Boruta

Fuente: elaboración propia

6. **APLICACIÓN DE MODELO DE CLASIFICACIÓN**

Una vez se han analizado las distintas variables que componen el dataset, así como el pre-procesamiento de las mismas es momento de abordar el modelo predictivo.

En este caso, se utiliza el modelo llamado “árbol de clasificación” o “árbol de decisiones”. Al aplicar este tipo de modelo nos permite clasificar las instancias siguiendo unas reglas a través de los nodos que formal el árbol.

Con este modelo se obtiene una precisión en concreto, es decir, se consigue medir la dispersión de los valores obtenidos, ya sea de forma de repetitividad, reproducibilidad o intermedia. Pero, por otro lado, no podemos confundir el término de precisión con el de “accuracy”.

Accuracy se refiere a como de cerca están los valores del valor medio, es decir, si la medición obtenida está más próxima al valor denominado como verdadero significará que será una medición exacta. En definitiva, accuracy se puede definir como el porcentaje total de aciertos que tiene nuestro modelo.

Para aplicar el modelo de árbol de clasificación se va a utilizar una librería de RStudio denominada “Rpart” (Recursive Partitioning and Regression Trees). Esta es una librería muy potente y muy usada en Machine Learning para R, se utiliza para la creación de árboles de clasificación y regresión, además, implementa particiones recursivas. Es una librería muy fácil de implementar, por ello es una de las más usadas para el modelo de árbol.

Dado que la muestra empleada está muy desbalanceada respecto a su variable objetivo (con muy pocos casos de accedente cerebrovascular) los modelos predictivos tienen un sobreajuste hacia el consecuente no-accidente. Por ello, para probar la generación de modelos predictivo, se recurre a otros datasets del ámbito médico que, además, se analizarán ligeramente y se emplearán las técnicas de preprocesamiento explicadas anteriormente. Estos nuevos datasets son los siguientes (tabla 20):

						PREPROCESAMIENTO				MODELOS
dataset	filas	col	discretas	numéricas	binarias	V.ausentes	Outliers	Discretizaciones	V. influyentes	Arboles Clasificación
Heart_disease	303	14	4	4	6	Sí	No	Sí	Sí	Sí
Heart_failure	299	13	1	6	6	No	No	Sí	Sí	Sí

Tabla. 20. Características datasets

Fuente: elaboración propia

6.1 DATASET 1: HEART_DISEASE

En primer lugar, “heart_disease” es una base de datos de Cleveland con el objetivo de estudiar la presencia de una enfermedad cardiaca.

A este dataset se le realizan unos preprocesamientos como se indica en la tabla anterior (tabla 20) y después una vez este el dataset listo se utiliza un modelo de árbol de decisión. En este caso, se procede a comparar la base de datos sin tratar, es decir, si recibir ningún tipo de cambio contra el dataset con los preprocesamientos indicados.

Heart_disease contiene un total de 303 filas, en los cuales hay ciertos errores a la hora de recoger los datos:

- La variable “thal” contiene dos observaciones con resultado igual a 0. Esto es un error, debido a que, la variable thal solo puede obtener valores de 1,2 o 3. Por lo tanto, esto dos valores son errores a la hora de recoger los datos, en este caso, se tienen que tratar como “Na” en el dataset que supone alrededor de un 0.6% de los datos.
- Y, en la variable “nvasos” sucede algo similar. En el dataset se encuentra 5 observaciones con datos nvasos = 4, cuando la variable nvasos solo puede coger valores de 1, 2 o 3. En este caso, también se sustituyen por “Na” en el dataset que supone un poco más del 1,5% de valores ausente.

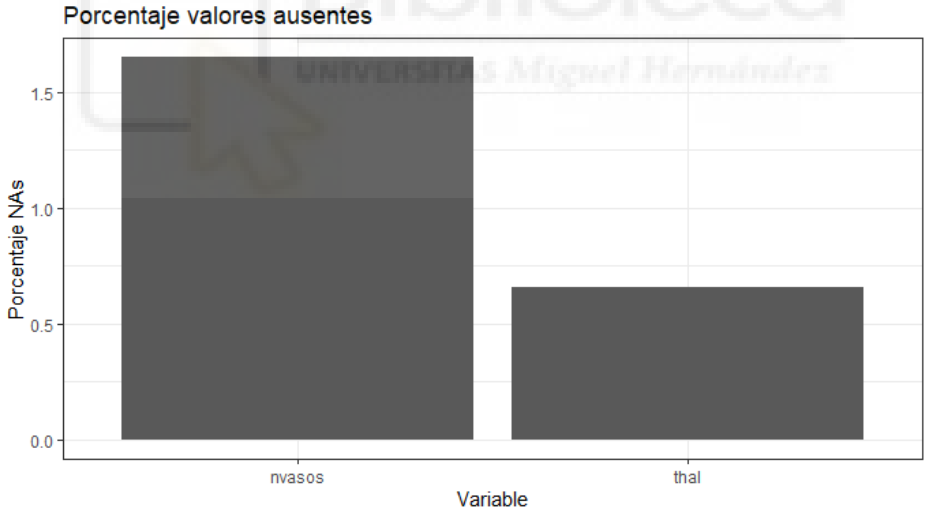

Fig. 34. Gráfico de barras de valores ausentes

Fuente: elaboración propia

Tras tener esto claro, se tratan estos valores ausentes y se sustituyen por el valor de la media. Con esto se busca conseguir que los valores ausentes se asemejen lo máximo posible a la realidad.

Por otro lado, se detectan distintos outliers en el dataset como son en las variables “presión”, “colesterol”, “frec_cardiaca” y “oldpeak”. Al tener un dataset con pocas observaciones (303) y observar cómo afectaría la eliminación de estos

se deciden mantener. Además, de al no ser expertos en la materia se desconoce el alcance que tienen estos valores sobre la variable objetivo.

Tras tener todo esto claro se utiliza la librería Boruta para encontrar cuales son las variables influyentes en el modelo. Obteniendo la siguiente selección:

attribute	meanImp	medianImp	minImp	maxImp	normHits	decision
nvasos	23.27077852	23.28268928	20.37587	26.15122	1	**Confirmed**
thal	2032752893	20.23181237	17.14004	24.10265	1	**Confirmed**
tipo_dolor	20318870195	20.3010476	17.49893	22.97948	1	**Confirmed**
oldpeak	16332725007	16.42425827	13.44315	18.56071	1	**Confirmed**
frec_cardiaca_max	12.24245423	12.29978501	9.701324	14.24507	1	**Confirmed**
angina	11.46139762	11.53408323	9.356708	14.24231	1	**Confirmed**
sexo	10.63393856	10.5448821	8.237427	12.87752	1	**Confirmed**
pendiente	10.12086135	10.04411088	7.696145	13.25854	1	**Confirmed**
edad	7.885030776	7.893415351	4.68115	10.5701	0.989899	**Confirmed**
presion	2.290657853	2.323843415	-1.00629	5.147481	0.444444	**Rejected**
electro	1.053111843	1.330452242	-1.18488	2.630499	0.030303	**Rejected**
azucar	0.015463637	0.13581759	-1.8235	1.350898	0	**Rejected**
colesterol	-1.007029318	-1.155505797	-3.58044	2.059027	0	**Rejected**

Tabla. 21. Selección de atributos

Fuente: elaboración propia

De las 13 variables que completan el dataset 4 de ellas (presión, electro, azúcar y colesterol) no son importante o no son influyentes en la variable de estudio, por lo tanto, estas 4 variables no se añadirán al futuro modelo.

El siguiente preprocesamiento incluido son las discretizaciones. Se analizan gráficamente aquellas variables susceptibles de discretizar, pero en cualquier caso se va a realizar dos tipos de discretizaciones, en dos intervalos y en tres intervalos. Posteriormente, se emplearán aquellas variables que aporten una mayor representatividad para el modelo.

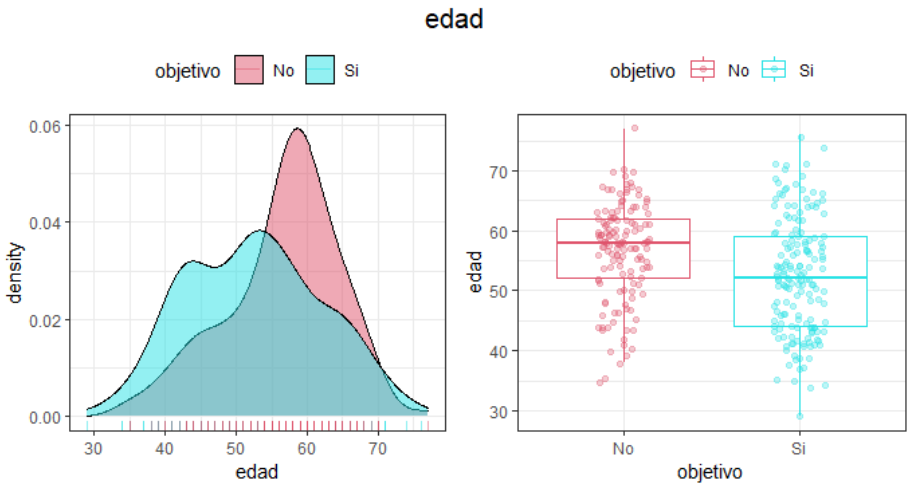

Fig. 35. Distribución de la edad de los pacientes en función de enfermedad de corazón

Fuente: elaboración propia

Como se observa en la figura 35 la variable edad a partir de los 50 años, aproximadamente, es donde más cantidad de pacientes se encuentran. Además, alrededor de los años sesenta se encuentra un gran pico de pacientes con no-enfermedad y sería importante estudiar este comportamiento. Por ello, discretizar podría ayudar a comprender como afecta mejor la variable edad a la variable objetivo.

Al discretizar en 2 intervalos y en 3 intervalos se obtiene como la discretización en 3 intervalos no es buena para esta variable, debido a que, se pierde mucha información y acaba aportando de forma negativa el futuro modelo. La discretización en 2 intervalos es la adecuada para esta variable (tabla 22):

EDAD	VALOR
[29,55)	Adultos
[55,77)	Ancianos

Tabla. 22. Discretización de la variable edad

Fuente: elaboración propia

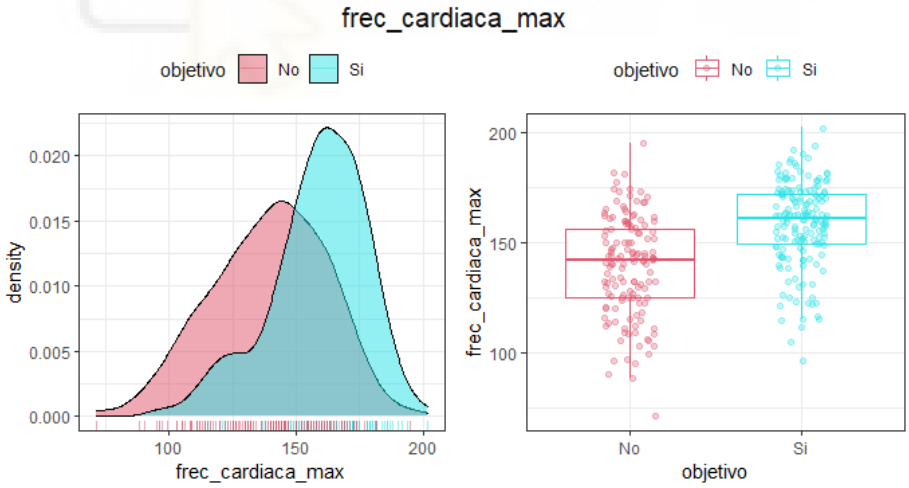

Fig. 36. Distribución de la frecuencia cardiaca de los pacientes en función de enfermedad de corazón

Fuente: elaboración propia

En la figura 36 se encuentra la variable referente a la frecuencia cardiaca, en esta es obvia como aquellos con una mayor frecuencia cardiaca tienen como resultado si-enfermedad, sufriendo un gran pico de casos cercano a la frecuencia de 150, por lo tanto, discretizar podría ayudar a que esta variable aporte una mayor información al modelo.

Igual que anteriormente se discretiza en 2 y 3 intervalos. Al discetizar en 2 intervalos se pierde información de la variable y sería negativo para el modelo, por lo tanto, discretizar en 3 intervalos es aquella que más aporta al modelo. Siendo la siguiente (tabla 23):

F.CARDIACA	VALOR
[71,143)	Baja
[143,162)	Normal
[162,202)	Alta

Tabla. 23. Discretización frecuencia cardiaca

Fuente: elaboración propia

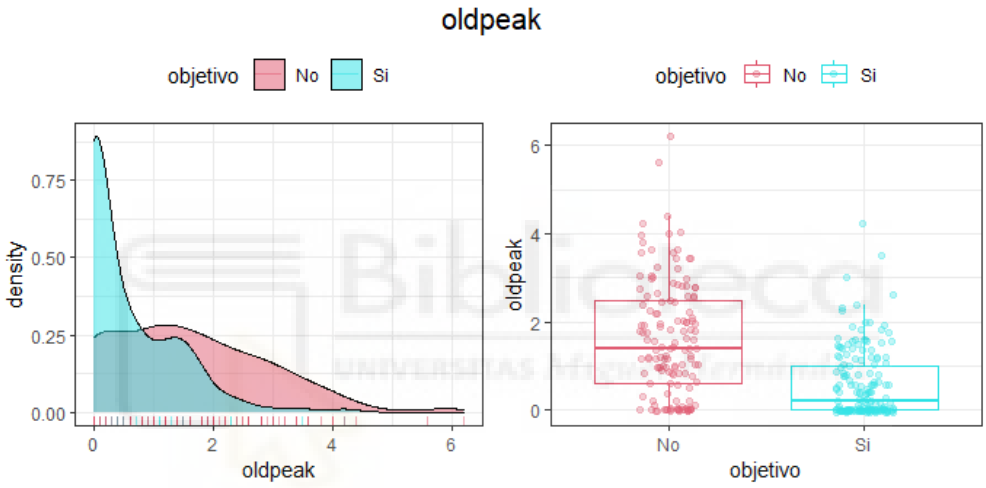

Fig. 37. Distribución de oldpeak de los pacientes en función de enfermedad de corazón

Fuente: elaboración propia

Por último, se discretiza la variable oldpeak (figura 37). En esta se observa como hay claras diferencias entre los distintos niveles que tiene la variable. En los valores más bajos prácticamente todo son casos de si-enfermedad y a medida que aumenta el oldpeak pasan a ser todos los casos no-enfermedad.

Al discretizar en 2 y 3 la variable oldpeak no influye en gran medida sobre la variable objetivo. Por lo tanto, se discretiza la variable en 3 intervalos (tabla 24), para posteriormente ver si es necesaria para el modelo o no:

OLDPEAK	VALOR
[0,0.1)	Baja
[0.1,1.4)	Normal
[1.4,6.2)	Alta

Tabla. 24. Distribución de la variable oldpeak

Fuente: elaboración propia

Tras realizar todo el preprocesamiento se procede a analizar los arboles de decisión del modelo. En este punto se enfrenta un árbol de decisión con la base de datos obtenida, es decir, el dataset inicial sin ningún tipo de modificación contra el dataset con los preprocesamientos utilizados. Posteriormente, se obtendrá la matriz de confusión de cada modelo.

En primer lugar, el dataset original sin ninguna modificación. El modelo utilizado es el siguiente:

$$objetivo \sim edad + sexo + tipo_{dolor} + presion + colesterol + azucar + electro + f_{cardiaca} + angina + oldpeak + pendiente + nvasos + thal$$

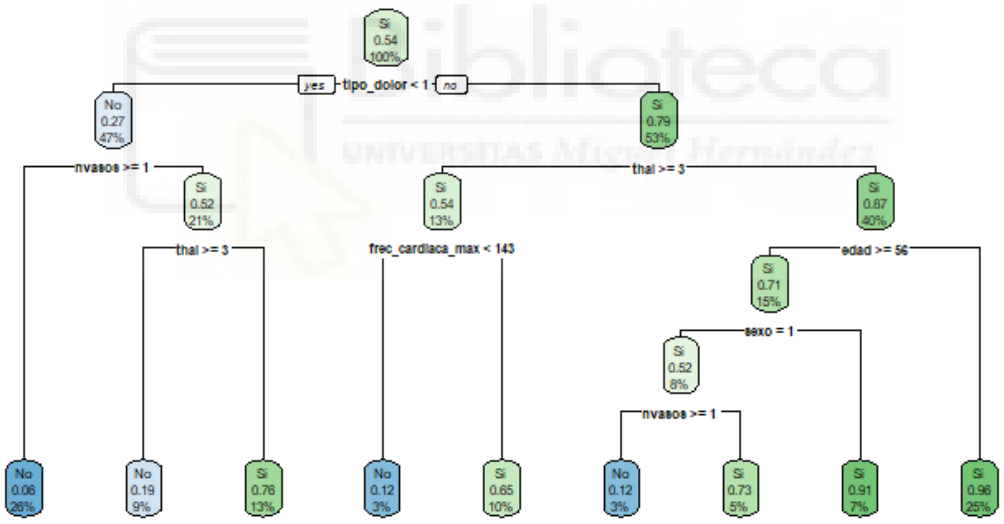

Fig. 39. Árbol de decisión del modelo 1 en heart_disease

Fuente: elaboración propia

Por la rama de la izquierda (figura 39) se puede concluir que aquellos pacientes con un tipo de dolor menor que 1 y con un numero de vasos superior o igual a 1 tiene una probabilidad del 94% de tener una enfermedad de corazón. Por otro lado, siguiendo la rama de la derecha, se obtiene como aquellos pacientes que no tienen un tipo de dolor menor que 1, que tienen un tejido cardiaco superior o igual a 3 y que no tienen una edad superior o igual a 56 años tienen una probabilidad del 96% de tener una enfermedad del corazón.

	Predicción del modelo	
REAL	**NO-ENFERMEDAD**	**ENFERMEDAD**
NO-ENFERMEDAD	109	29
ENFERMEDAD	12	153

Tabla. 25. Matriz de confusión modelo 1

Fuente: elaboración propia

Con este modelo sin tratamiento, se consigue esta matriz de confusión (tabla 25), obteniendo que el modelo clasifica los casos negativos con una probabilidad del 92.7% y, por otro lado, clasifica los casos positivos con una probabilidad del 79%. El modelo está prediciendo más falsos positivos cuando en realidad los pacientes no tienen enfermedad.

El modelo tiene como resultado final una precisión del 84% y una accuracy de 0.86468, es decir, del **86.5%**. Es una exactitud bastante alta teniendo en cuenta que los datos no han sido tratados.

El siguiente modelo es aquel obtenido tras realizar los distintos preprocesamientos recientemente comentados.

$$objetivo \sim edad1 + sexo + tipo_{dolor} + f_{cardiaca1} + angina + oldpeak1 + pendiente + nvasos + thal$$

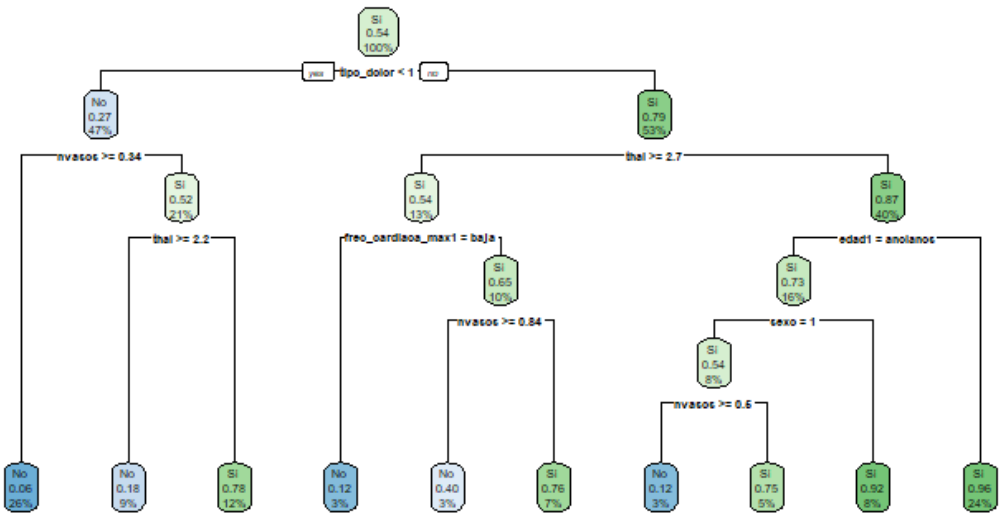

Fig. 40. Árbol de decisión del modelo 2 en heart_disease

Fuente: elaboración propia

Obtenido el modelo (figura 40), siguiendo la rama de la izquierda se obtiene que aquellos pacientes con un tipo de dolor menor a uno y con un número de vasos sanguíneos superior a 0.54 tiene una probabilidad de sufrir una enfermedad en el corazón del 94% (1-0.26), siendo esta similar a la anterior. Por otro lado, siguiendo la rama de la derecha, el paciente que tiene un tipo de dolor superior a 1, que el tejido cardiaco no sea superior o igual a 2.7 y que no estén dentro de la categoría “ancianos” tienen un 96% de probabilidades de tener una enfermedad en el corazón.

	Predicción del modelo	
REAL	**NO-ENFERMEDAD**	**ENFERMEDAD**
NO-ENFERMEDAD	116	22
ENFERMEDAD	16	149

Tabla. 26. Matriz de confusión modelo 2

Fuente: elaboración propia

La matriz de confusión obtenida con este modelo (tabla 26) se aprecia como clasifica los casos positivos con una probabilidad del 84% siendo en el anterior modelo del 79%. Por otra parte, los casos negativos se clasifican con una probabilidad del 90.3%, bajando respectivamente con el anterior. Esto se debe a que este modelo ha predicho más casos como no-enfermedad cuando son si-enfermedad. Aun teniendo esto el modelo consigue una precisión del 87.1% siendo un 3% superior, además, la exactitud de este modelo consigue un **87.4%** un 1% superior que en el ámbito médico puede ser muy importante.

6.2 DATASET 2: HEART_FAILURE

En el dataset 2, “heart_failure” únicamente se realizan los preprocesamiento de selección de variables influyentes y discretización de variables. Esto se debe a que, en primer lugar, el dataset no presenta datos ausentes, está completo, y tampoco hay posibles errores a la hora de recolectar y archivar los datos.

Por otro lado, sí que se encuentran outliers en algunas variables, como son creatinina, plaquetas, suero_creatinina se encuentran una gran cantidad de outliers. Sodio_serico y frac_eyección presentan outliers pero en menor medida.

Se decide no eliminar los outilers por el “bajo” nivel de observaciones (299), ya que, la base de datos se vería disminuida en una cantidad de datos demasiado grande. Además, se desconoce si estos casos fuera de los rangos son aquellos que pueden influir en la variable objetivo. Por lo tanto, se decide mantener los outliers del dataset.

Antes de discretizar variables se utiliza la función Boruta para conocer las variables más influyentes para la variable objetivo. En la tabla siguientes (tabla) están las variables seleccionadas y aquellas con las que se va a trabajar en el modelo.

attribute	meanImp	medianImp	minImp	maxImp	normHits	decision
tiempo	46.3638347	40.5836587	37.0534712	60.0339634	1	Confirmed
frac_eyeccion	18.2169901	17.4130899	16.0859436	22.419036	1	Confirmed
suero_creatinina	16.8007809	16.0726311	14.6209719	21.1224029	1	Confirmed
edad	7.04827793	6.37112851	5.37824399	10.4354293	1	Confirmed
sodio_serico	5.00054421	4.98784595	2.93641875	7.18686357	0.85714286	Confirmed
creatinina	0.41042551	-0.14448689	-1.03974848	2.46794205	0	Rejected
sexo	0.21865226	0.25653233	-1.02893463	1.6604871	0	Rejected
anemia	-0.12398414	-0.38894676	-1.32028371	1.08430884	0	Rejected
fumar	-0.20690692	0.1816459	-2.5218327	0.85015537	0	Rejected
plaquetas	-0.21551775	-0.08096797	-1.55584357	1.89263149	0	Rejected
diabetes	-0.34047579	-0.21771359	-1.65645329	1.23562624	0	Rejected
presion_sang	-0.5953416	-0.50974151	-2.06572836	0.71718977	0	Rejected

Tabla. 27. Selección de atributos heart_failure

Fuente: elaboración propia

Como se observa en la selección de variables (tabla 27) de las 12 variables restantes (quitando la variable objetivo) solo son influyentes 5 de ellas (tiempo, frac_eyeccion, suero_creatinina, edad, sodio_serico). A continuación, se decide discretizar algunas de estas variables, debido a que, así se conseguirá una mayor representación de las variables hacia la predicción del modelo.

En todos los casos se va a discretizar de dos maneras, en 2 intervalos y en 3 intervalos para observar como funcionan las variables de las dos maneras. Pero antes, se analizan estas variables para conocer como se distribuyen y si es adecuado o no la discretización.

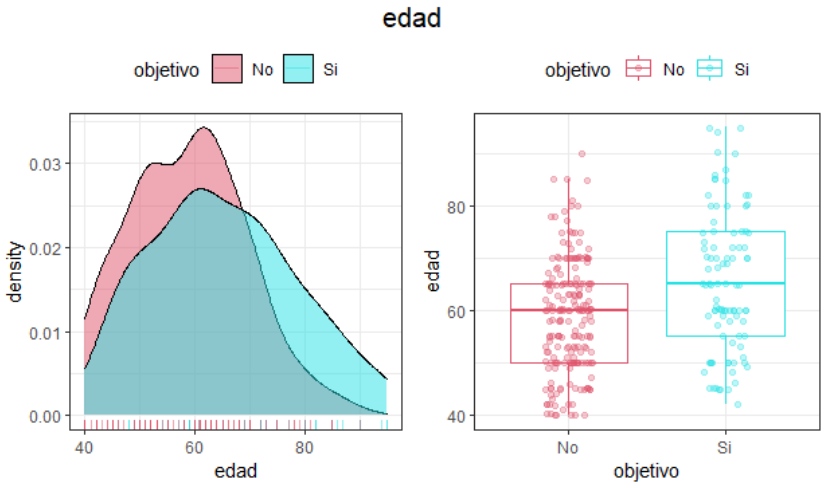

Fig. 41. Distribución de edad de los pacientes en función de insuficiencia cardiaca

Fuente: elaboración propia

En primer lugar, se encuentra la variable edad (figura 41). En esta se observa como los pacientes que son más jóvenes, que también es donde más cantidad de pacientes hay, se observa una mayor cantidad de no-insuficiencia cardiaca y si observamos en las edades más altas hay una mayor cantidad de si-insuficiencia que de no-insuficiencia. Por lo tanto, una discretización podría ayudar al modelo a diferenciar mejor entre las distintas edades y como afectan estas a la variable de estudio.

Tras discretizar en 2 y en 3 intervalos con la función “discretize” la variable edad se distribuye de las siguientes maneras (tablas 28 y 29):

EDAD	VALOR
[40,55)	Adultos
[55,65)	Mayores
[65,95)	Ancianos

Tabla. 28. Discretización edad en 3 tramos

Fuente: elaboración propia

EDAD	VALOR
[40,60)	Adultos
[60,95)	Ancianos

Tabla. 29. Discretización edad en 2 tramos

Fuente: elaboración propia

La siguiente variable “porcentaje de sangre”, se analiza en la figura 42. Esta variable visualmente presenta un patrón teniendo dos picos de no-insuficiencia en dos medidas distintas, por otro lado, en los valores más bajos es donde se concentra la mayoría de casos en si-insuficiencia, por ello una discretización podría diferenciar bien entre los Si y No para la variable objetivo. Las discretizaciones en 2 y 3 intervalos quedan de la siguiente manea (tablas 30 y 31):

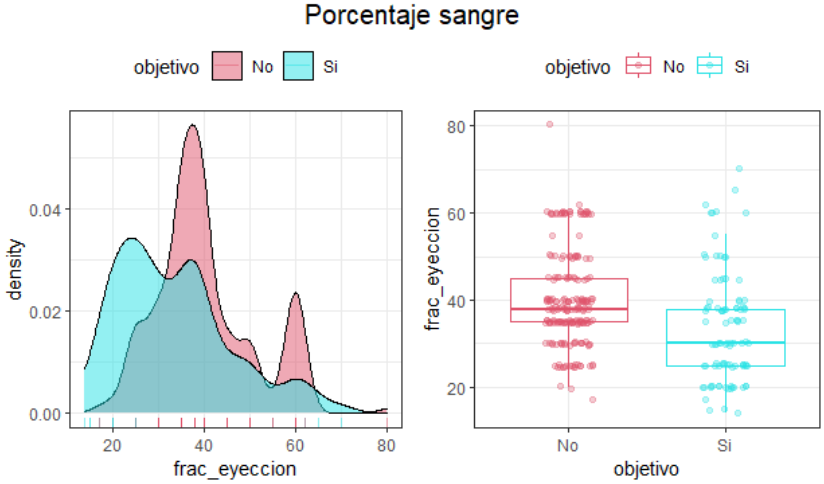

Fig. 42. Distribución de porcentaje de sangre de los pacientes en función de insuficiencia cardiaca

Fuente: elaboración propia

P.SANGRE	VALOR
[14,35)	Bajo
[35,40)	Medio
[40,80)	Alto

Tabla. 30. Discretización de porcentaje de sangre en 3 tramos

Fuente: elaboración propia

P:SANGRE	VALOR
[15,38)	Bajo
[38,80)	Medio

Tabla. 31. Discretización porcentaje de sangre en 2 tramos

Fuente: elaboración propia

La siguiente variable, suero de creatinina (figura 43) presenta como todos los datos están en los valores inferiores, siendo los pacientes con no-insuficiencia aquellos que prevalecen en estos niveles, pero cuando estos valores aumentan ligeramente los pacientes con no-insuficiencia descienden dejando por encima a aquellos con si-insuficiencia. La discretización de esta variable se observa en la tablas 32 y 33.

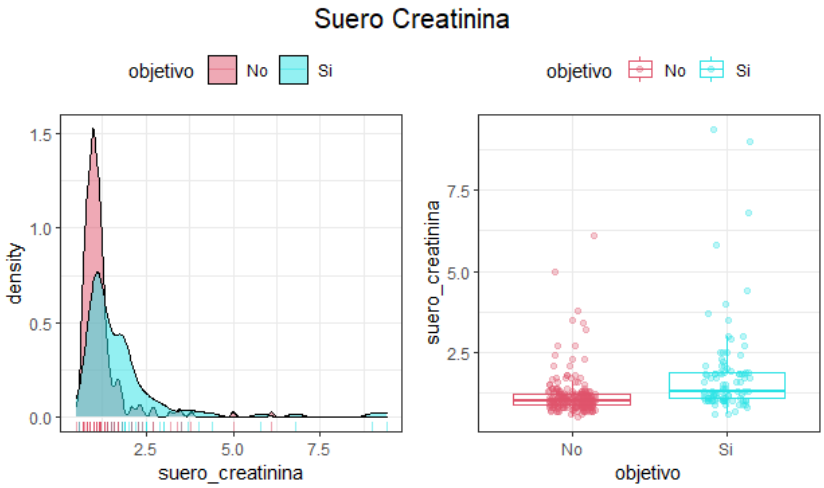

Fig. 43. Distribución de suero de creatina de los pacientes en función de insuficiencia cardiaca

Fuente: elaboración propia

S. CREATININA	VALOR
[0.5,1)	Bajo
[1,1.3)	Medio

[1.3,9.4)	Alto

Tabla. 32. Discretización suero creatinina en 3 tramos

Fuente: elaboración propia

S. CREATININA	VALOR
[0,1.1)	Bajo
[1.1,9.4)	Medio

Tabla. 33. Discretización suero creatinina en 2 tramos

Fuente: elaboración propia

Por último, se analiza la variable “socio serico” en la figura 44. Todos los datos están comprendidos entre los mismos valores, ubicando una pequeña separación entre los si-insuficiencia y los no-insuficiencia. Se discretiza en 2 y 3 intervalos de la siguiente forma (tablas 34 y 35):

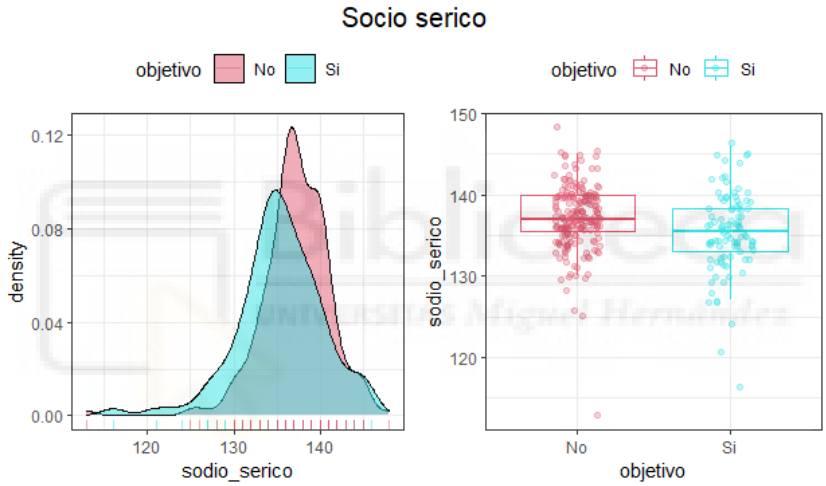

Fig. 44. Distribución de sodio sérico sangre de los pacientes en función de insuficiencia cardiaca

Fuente: elaboración propia

SODIO SERICO	VALOR
[113,136)	Bajo
[136,138)	Medio
[138,148)	Alto

Tabla. 34. Discretización sodio sérico en 3 tramos

Fuente: elaboración propia

SODIO SERICO	VALOR
[113,137)	Bajo

[137,148)	Medio

Tabla. 35. Discretización sodio sérico en 2 tramos

Fuente: elaboración propia

Tras discretizar todas las variables necesarias se realiza los modelos de árboles de decisión. En este caso, se realiza un modelo del dataset original igual que en el anterior, en este modelo no se ha modificado nada después de la recolección de los datos. Una vez este el modelo se creará una matriz de confusión para comprobar como funciona el modelo creado. Posteriormente, se realiza los mismos pasos, pero con el dataset tratado.

El modelo 1, tras no haber sido modificado y utilizando todas las variables en el dataset presenta la siguiente forma:

$$objetivo \sim edad + anemia + creatinina + diabetes + frac_{eyeccion} + presion_{sang} + plaquetas + suero_{creatinina} + sodio_{serico} + sexo + fumar + tiempo$$

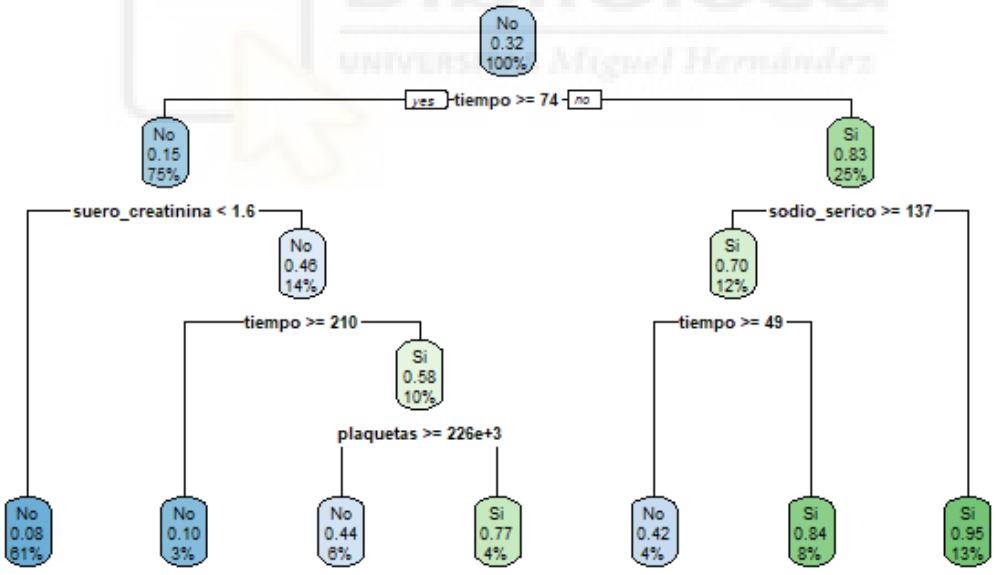

Fig. 45. Árbol de decisión modelo 1 en heart_failure

Fuente: elaboración propia

Si analizamos el árbol de decisión obtenido para el modelo 1 (figura 45), siguiendo la rama de la izquierda tenemos aquellos pacientes que llevan un tiempo superior o igual a 74 días en el hospital y que tienen un suero de creatinina inferior a 1.6 tienen una probabilidad de sufrir una insuficiencia cardiaca del 92% (1-0.08). Por otro lado, en la rama de la derecha encontramos aquellos pacientes que no llevan 74 días o más en el hospital y que no tienen un

sodio sérico de 137 o más tienen una probabilidad de sufrir una insuficiencia cardiaca del 95%.

	Predicción del modelo	
REAL	**NO-ENFERMEDAD**	**ENFERMEDAD**
NO-ENFERMEDAD	194	9
ENFERMEDAD	28	68

Tabla. 36. Matriz de confusión modelo 1 heart_failure

Fuente: elaboración propia

Con este modelo sin realizarle ningún tipo de preprocesamiento, como se observa en la tabla 36, se consigue un accuracy del **87.6%,** es una exactitud bastante alta. Además, consigue una precisión del 88.3%.

En el modelo número 2, tras analizar el dataset y aplicarle los preprocesamiento el modelo queda de la siguiente manera. Las discretizaciones finales usadas en el modelo son todas aquellas correspondientes a los 3 intervalos, debido a que, al utilizar la discretización en 2 intervalo el modelo se veía perjudicado por la pérdida de información al discretizar en tan pocos niveles.

$$objetivo \sim edad1 + frac_{eyeccion1} + suero_{creatinina1} + sodio_{serico1} + tiempo$$

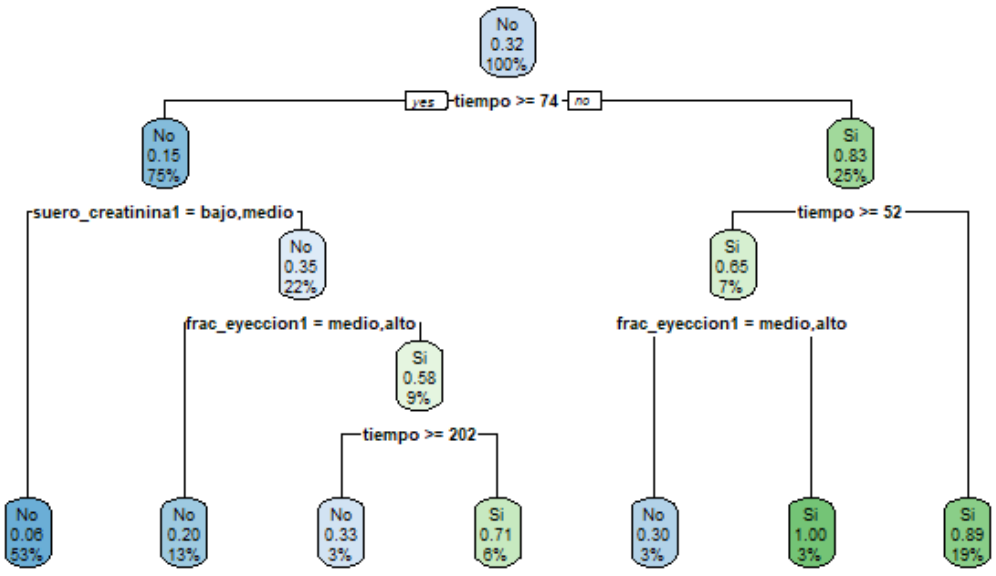

Fig. 46. Árbol de decisión modelo 2 en heart_failure

Fuente: elaboración propia

En este árbol de decisión (figura 46), siguiendo la rama de la izquierda encontramos aquellos pacientes con un tiempo igual o superior a 74 días y que tienen suero creatinina al nivel bajo o medio tienen una probabilidad de sufrir una insuficiencia cardiaca del 94%. Por la otra parte del árbol encontramos a los pacientes que no llevas 74 días o más en el hospital y que tampoco llevan 52 días o más tienen una probabilidad del 89% de sufrir una insuficiencia cardiaca.

	Predicción del modelo	
REAL	**NO-ENFERMEDAD**	**ENFERMEDAD**
NO-ENFERMEDAD	192	11
ENFERMEDAD	24	72

Tabla. 37. Matriz de confusión modelo 2 heart_failure

Fuente: elaboración propiac

Con este modelo se consigue la matriz de confusión anterior (tabla 37), en esta se aprecia como los casos positivos alcanzan una probabilidad del 94.5% y los casos negativos siendo donde falla la predicción del modelo alcanzando el 75%. En el modelo anterior la precisión alcanzaba el 88.3% de acierto y en este modelo disminuye hasta el 86.7% pero aumenta la exactitud del modelo, subiendo desde un 87.6% hasta un **88.3%**.

7. **INTERPRETACIÓN DE RESULTADOS**

	Modelo original	Modelo preprocesado
Heart_disease	86.5%	87.4%
Heart_failure	87.6%	88.3%

Tabla. 38. Accuracy de los cuatro modelos

Fuente: elaboración propia

Los resultados obtenidos (tabla 38) muestran como en el dataset "heart_disease" los preprocesamientos utilizados ayudan a mejorar el modelo. Al sustituir las variables ausentes estamos consiguiendo una muestra más representativa, más cercana a la realidad, con la selección de variables influyentes nos aseguramos de utilizar un modelo simplificado y a la vez más enfocado hacia la variable objetivo y, por último, la discretización nos ayuda a conseguir que esas variables numéricas sean más simples y ayuden a construir mejor el modelo. Por esto, se consigue que el modelo a pesar de ser un dataset bajo en cantidad de datos y muy poco margen para mejorar empleando estas técnicas se consigue mejorar la acurracy en casi un 1%.

En cuanto al dataset "heart_failure" únicamente se emplea la selección de variables influyentes y discretizaciones. En este caso, se decidió no eliminar outliers por las pocas observaciones que componían el dataset y la eliminación de estos valores restaría mucha información al modelo, además de perder datos relevantes para la variable objetivo. Con la selección de las variables influyentes se simplifico el modelo y en las discretizaciones se emplearon aquellas con correspondientes a tres tramos. Por lo tanto, este dataset con el preprocesamiento realizado conseguía que el modelo predijera mejor la variable objetivo con un 88.3%, mejorando ligeramente el dataset original. La mejoría es mínima, del 0.6%, pero teniendo en cuenta que el dataset original tenía bajas observaciones y estaba en plenas condiciones para trabajar es una mejoría importante.

En ambos modelos se probaron discretizaciones en dos y tres tramos y finalmente se empleó aquellas que más ayudaban al modelo a entender mejor las variables. En la siguiente tabla se observa el accuracy de los dataset con las discretizaciones no empleadas, en el dataset 1 con edad (tres tramos), frec_cardiaca (dos tramos) y oldpeak (tres tramos) y en el dataset 2 con todas las variables en tres tramos. En ambos modelos el accuracy baja considerablemente.

	Accuracy
Heart_disease	84.8%
Heart_failure	85.2%

Tabla. 38. Accuracy de las discretizaciones descartadas

Fuente: elaboración propia

8. **CONCLUSIONES**

Tras analizar los distintos dataset planteados y obtener conclusiones en dos de ellos podemos responder a la pregunta efectuada en el punto hipótesis de partida "¿El preprocesamiento mejora la predicción del modelo?".

Gracias a el tratamiento realizado en apartados anteriores queda claro que un procesamiento de los datos es de vital importancia, debido a que, gracias a esto los algoritmos consiguen unos datos más simples y útiles con los que trabajar y obtener mejores resultados. En este caso se aplica en unas muestras encontradas con un volumen bajo de datos, pero de esto se pude extrapolar la importancia de estas técnicas. Así mismo, los modelos resultantes tras el preprocesamiento también son más fáciles de interpretar, gracias a las discretizaciones y selección de variables que intervienen en el problema

Si se extrapola y nos lo llevamos a como es en la actualidad, es decir, el crecimiento masivo y constante en los datos, el gran volumen de estos que se mueve en la actualidad es de vital importancia proceder sobre ellos con técnicas de preprocesamiento para así poder tener la mayor calidad posible en los datos.

Además, si lo tratamos desde el punto de vista del actual documento, es decir, del ámbito médico, la importancia de estas técnicas aumenta considerablemente. En la medicina el número de datos, de información disponible es enorme, desde pacientes, enfermedades, variantes de estas y un largo etcétera. Por ello, es muy importante tratar estos datos para conseguir la calidad adecuada de estos para que en su futuro sirvan de entrada para los algoritmos de Minería de datos y así poder conseguir unos resultados mucho más precisos a los que se conseguiría sin emplear estas técnicas.

9. **BIBLIOGRAFÍA**

[1] Ying Yang, Geoffrey I. Webb and Xindong Wu. (2010). Discretization Methods.

[2] Pyle, D. (1999). Data Preparation for Data Mining (Pap/Cdr ed.). Morgan Kaufmann Pub.

[3] Quinlan, R., & Kohavi, R. (1999). Decision Tree Discovery. IN HANDBOOK OF DATA MINING AND KNOWLEDGE DISCOVERY.

Departamento de Ciencias de la Computación e Inteligencia Artificial, Universidad de Granada (España), García, S., Ramírez-Gallego, S., Luengo, J., & Herrera, F. (2016). Big Data: Preprocesamiento y calidad de datos.

Joaquín Amat Rodrigo - Machine Learning con R y caret. (2018, 25 abril). rpubs. https://rpubs.com/Joaquin_AR/383283

RPubs - Métodos de Clasificación. (2020, 9 agosto). RPUBS. https://rpubs.com/Mabe1995/647359

Juan Bosco Mendoza - Arboles de decisión con R - Clasificación. (2018, 24 abril). Vega. https://rpubs.com/jboscomendoza/arboles_decision_clasificacion

Miguel Arquez Abdala - regression trees and bagging. (2020, 30 marzo). https://rpubs.com/arquez9512/592192

J., & JayJay, V. A. P. B. (2018, 6 abril). Sin supervision. Data Science. http://datascience.esy.es/wiki/sin-supervision/

J., & JayJay, V. A. P. B. (2018b, abril 6). Supervisada. Data Science. http://datascience.esy.es/wiki/supervisada/

Accidente cerebrovascular - Síntomas y causas - Mayo Clinic. (2021, 9 febrero). https://www.mayoclinic.org/es-es/diseases-conditions/stroke/symptoms-causes/syc-20350113

www.sdelsol.com. (2017, 10 agosto). La Minería de datos aplicada a la salud - Blog - Stimulus | APP profesional de estimulación cognitiva. STIMULUS. https://stimuluspro.com/blog/la-mineria-de-datos-aplicada-a-la-salud/