FACULTAD DE CIENCIAS SOCIALES Y JURÍDICAS DE ELCHE

GRADO EN SEGURIDAD PÚBLICA Y PRIVADA

CURSO 2021/2022

TRABAJO FIN DE GRADO

TÍTULO: EL SESGO EN EL USO DEL BIG DATA APLICADO A LA SEGURIDAD Y DEFENSA PARA LA PREDICCIÓN DE LA COMISIÓN DE DELITOS.

AUTOR: JOSÉ ANTONIO BELMONTE RUIPÉREZ

TUTORA: MARINA LEAL PALAZÓN

Resumen.

Cada día se generan en todo el mundo una inmensa cantidad de datos que crecen de manera exponencial cada año. A estos datos masivos y a su almacenamiento y procesamiento los denominamos Big Data. Esta enorme cantidad de información provienen de nuevas fuentes de datos como, por ejemplo; dispositivos GPS, relojes inteligentes, sensores o dispositivos conectados a Internet (IoT).

Los datos pueden presentarse en distintos formatos, y responden a las características de Volumen, Velocidad, Variedad, Veracidad y Valor. Se trata de una cantidad de datos tan grande y compleja que, para su almacenamiento, gestión, análisis y tratamiento, se requiere de softwares creados específicamente para ello, puesto que los softwares tradicionales empleados para la gestión de datos son ineficientes. Es aquí donde cobra importancia el Machine Learning. Mediante algoritmos de aprendizaje el sistema es capaz de descubrir patrones basados en los datos que recibe y en función de ellos, realizar predicciones.

El uso de Big Data tiene gran valor añadido y puede ser aplicado en numerosos sectores, como es el de la seguridad y la defensa. En este ámbito permite, entre otras aplicaciones, su uso para la predicción de la comisión de delitos mediante el análisis de datos. Machine Bias es un artículo que analiza un caso concreto del uso de métodos de Big Data para la predicción de la reincidencia de criminales detenidos, y muestra el sesgo racial que se produce al aplicar estos métodos. Motivado por este caso, analizaré la relación del Big Data con la Seguridad y la Defensa. Realizaré un estudio práctico utilizando el software estadístico R para la aplicación de métodos de Big Data en predicción de delitos, y determinar si puede producirse sesgo.

Palabras clave: Sesgo, Big Data, Predicción, Delitos.

ÍNDICE

1.- Introducción: ¿Qué es el Big Data?

1.1.- El auge de los datos masivos

1.2.- Las 5 Vs del Big Data

1.3.- Aplicaciones del Big Data

2. El Big Data orientado a la Defensa y la Seguridad

2.1.- La Criminología Computacional

3.- Estructura para el análisis del Big Data

3.1.- Los Árboles de clasificación

4.- Sesgos en el uso del Big Data para la predicción de la delincuencia.

4.1.- Análisis de un caso real "Machine Bias"

5.- Predicciones de reincidencia en crimen. Estudio práctico en R.

6.- Conclusión

7.- Bibliografía

1.- Introducción: ¿Qué es el Big Data?

El siglo XXI se caracteriza por la irrupción de la Revolución Digital, en sectores como la industria, el sistema sanitario, la educación o la seguridad y en general en todos los aspectos que nos afectan en el día a día. A esta revolución se la conoce como La Tercera Revolución Industrial, y ha conducido el avance de la tecnología, promoviendo la transición de un mundo analógico a un mundo digital y tecnológico con inteligencia artificial y aprendizaje automático de máquinas.

Esta digitalización ha facilitado que personas de todo el mundo puedan interactuar y compartir todo tipo de información en Internet y a través de redes sociales. Todo ello mediante el empleo de dispositivos digitales tanto a nivel personal, como en el hogar o en la industria, tales como teléfonos inteligentes, wearables, enchufes wifi, cámaras de video vigilancia, sensores…, la mayoría de ellos conectados a Internet generando datos a gran velocidad.

En el año 2016 **Klaus Schwab**, fundador del Foro Económico Mundial acuñó el término **Cuarta Revolución Industrial** estableciendo que "*genera un mundo en el que los sistemas de fabricación virtuales y físicos cooperan entre sí de una manera flexible a nivel global*". "*Estamos al borde de una revolución tecnológica que modificará la forma en que vivimos, trabajamos y nos relacionamos. En una escala de alcance y complejidad la transformación será diferente a cualquier cosa que el género humano haya experimentado antes*" (K. Schawab, 2016)

En esta cuarta Revolución Industrial los datos y la información son el combustible para que mediante algoritmos se generen decisiones para un mundo más inteligente y eficiente. A esta nueva área de la tecnología se la ha definido como análisis y ciencia del Big Data.

En general el Big Data puede ser considerado como la tendencia tecnológica que ha abierto las puertas hacia un enfoque de entendimiento y de toma de decisiones para describir y analizar grandes cantidades de datos, que de otra manera no sería posible analizar.

Existen numerosas definiciones sobre el Big Data entre las que podemos destacar:

"*Conjunto de técnicas que permiten analizar, procesar y gestionar conjuntos de datos extremadamente grandes que pueden ser analizados informáticamente para revelar patrones, tendencias y asociaciones, especialmente en relación con la conducta humana y las interacciones de los usuarios*". (www.dpej.rae.es)

"*Las tecnologías de Big Data describen un nuevo conjunto de tecnologías y arquitecturas, diseñadas para extraer valor y beneficio de grandes volúmenes de datos con una amplia variedad en su naturaleza, mediante procesos que permitan capturar, descubrir y analizar información a alta velocidad y con un coste reducido*." (J. Gantz, & D. Reinsel, 2012).

"*Conjunto de datos cuyo tamaño va más allá de la capacidad de captura, almacenado, gestión y análisis de las herramientas de bases de datos"* (J. Manyika, J. Chui, B. Brown, R. Dobbs, C. Roxburgh, A. Hung Byers, 2011)

"*Big Data son los grandes conjuntos de datos que tienen tres características principales: volumen (cantidad), velocidad (velocidad de creación y utilización) y variedad (tipos de fuentes de datos no estructurados, tales como la interacción social, video, audio, cualquier cosa que se pueda clasificar en una base de datos). Estos factores, naturalmente, conducen a una complejidad extra de los Big Data; en síntesis "Big Data" es un conjunto de datos tan grandes como diversos que rompen las infraestructuras de TI tradicionale*s." (L. Dougas, Consultora Gartner, 2012).

1.1.- El auge de los datos masivos

La progresiva digitalización de nuestro mundo conlleva el crecimiento exponencial en la generación de datos, cada día se generan miles de terabytes de valiosa información. Generamos datos cuando simplemente navegamos con nuestros teléfonos móviles a través de Internet o realizamos búsquedas, cuando interactuamos en las redes sociales o caminamos por la ciudad con el móvil en el bolsillo mediante el rastreo GPS de nuestros movimientos, realizamos descargas o compras on-line.

Los wearables, como los relojes y pulseras inteligentes, monitorizan constantemente, nuestras constantes vitales, miden nuestra actividad física y generan datos sobre el estado de salud de nuestro cuerpo, ritmo cardiaco o nivel

de estrés. Las entidades bancarias, los vehículos conectados, las plataformas de video en *streaming*, las cámaras de seguridad que graban y almacenan miles de horas de video…

Un concepto muy importante que permite la transmisión de datos a través de la red es el de IoT o Internet de las cosas. El término hace referencia a la red de dispositivos físicos “cosas”, que llevan incorporados sensores, software y otras tecnologías con el objetivo principal de conectarse e intercambiar datos con otros dispositivos y sistemas a través de Internet. Esto posibilita una interacción “inteligente” entre estos dispositivos con mínima intervención humana, de manera que, unos pueden llevar a cabo acciones en función de los datos transmitidos por otros tal y como re representa en la Imagen 1. Toda esta información es almacenada en la nube por las empresas, para posteriormente y mediante el análisis de los datos mejorar sus productos.

Imagen 1. Ilustración sobre IoT o Internet de las Cosas. A. Armenta, (2021). Data Lake vs. Big Data for Industrial Applications.

Cada día, en el mundo se generan miles de terabytes de datos, y esto solo es el principio, porque año tras año esa cifra se duplica, así que nos podemos hacer una idea de la importancia de la recolección, almacenamiento y análisis de los mismos.

“Cada día creamos 2,5 quintillones de bytes de datos, de forma que el 90% de los datos del mundo actual se han creado en los últimos dos años. Estos datos proceden de todos los sitios: sensores utilizados para recoger información del clima, entradas (posts) en sitios de medios sociales, imágenes digitales, fotografías y videos, registros de transacciones comerciales y señales GPS de teléfonos celulares, por citar unas pocas referencias. Estos datos, son, según IBM, Big Data.

En el año 2000, se almacenaron en el mundo 800.000 petabytes. Se espera que en el año 2020 se alcance los 35 Zettabytes (ZB). Solo Twitter genera más de 9 Terabytes (TB) de datos cada día. Facebook, 10TB; y algunas empresas genera terabytes de datos cada hora de cada día del año.” (L. Joyanes Aguilar, 2013)

Según los tipos de datos, estos pueden ser estructurados, semiestructurados y no estructurados. La mayoría de las fuentes de datos tradicionales son datos estructurados, los cuales tienen un formato o esquema fijo con campos fijos, hojas de cálculo y archivos fundamentalmente. En estas fuentes, los datos vienen en un formato bien definido que se especifica en detalle, y que conforma las bases de datos relacionales. Los datos estructurados se componen de piezas de información que se conocen de antemano, tienen un formato especificado, y se producen en un orden especificado. Un ejemplo son las fechas de nacimiento (DD, MM, AA); documento nacional de identidad o pasaporte (8 dígitos y una letra); un número de la cuenta corriente de un bajo (20 dígitos), etc.

Los datos semiestructurados tienen un flujo lógico y un formato que puede ser definido, pero no es fácil su comprensión por el usuario. Se trata de datos sin formatos fijos, pero que contienen etiquetas y otros marcadores que permiten separar los elementos. La lectura de datos semiestructurados requiere el uso de reglas complejas que determinan cómo proceder después de la lectura de cada pieza de información. Un ejemplo de datos semiestructurados son los registros Web logs de las conexiones a Internet. Un Web log se compone de diferentes piezas de información, cada una de las cuales sirve para un propósito específico. Un ejemplo es el texto de etiquetas de lenguajes XML y HTML.

Por último, los datos no estructurados son datos sin tipos predefinidos. Se almacenen como documentos u objetos sin estructura uniforme, y se tiene poco

o ningún control sobre ellos. Son datos de texto, video, audio, fotografía… Como ejemplo las imágenes se clasifican por su resolución en píxeles. Datos que no tienen campos fijos como, por ejemplo: audio, vídeo, fotografías, documentos impresos, cartas, hojas electrónicas, imágenes digitales, formularios especiales, mensajes de correo electrónico y de texto, formatos de texto libre como correos electrónicos, mensajes instantáneos SMS, artículos, libros mensajes de mensajería instantánea tipo WhatsApp o WeChat. Al menos el 80% de la información de las organizaciones reside en las bases de datos relacionales o archivos de datos, sino que se encuentra esparcido a lo largo y ancho de la organización y esto es a lo que llamamos datos no estructurados.

El primer paso para aplicar el análisis en el Big Data es la recolección de datos, después estos deben ser preparados para su análisis.

En ocasiones, los datos pueden necesitar ser procesados para eliminar aquellos que puedan inducir a error porque falten valores, contengan datos corruptos, duplicados, inconsistentes o porque tengan un formato incorrecto, es por ello, que han de someterse a un proceso de depuración para detectar y resolver estos problemas. Sirva de ejemplo, utilizar archivos de texto con diferentes fuentes que si no procesamos correctamente pueden llevar a inconsistencias por los separados de campo utilizados en los diferentes archivos. En algunos, el separador puede ser la coma y en otros el tabulador, mediante un proceso de depuración es posible eliminar estas inconsistencias. Lo mismo ocurre si recogemos información de diferentes estaciones meteorológicas con información en grados Celsius y Fahrenheit.

Algunos ejemplos de datos sin procesar son: los archivos *.log* generados por aplicaciones y servicios de monitorización, datos generados por aplicaciones de comercio electrónico, bancario y financiero, datos generados por redes sociales, datos generados por los sensores de los dispositivos conectados a Internet IoT, datos generados por los flujos de *clics* en las aplicaciones y páginas webs que pueden ser usados parta analizar patrones de conducta de los usuarios, datos recogidos de los sistemas de vigilancia…

Uno de los retos a que se enfrenta el Big Data, es dónde almacenar toda esta enorme cantidad de información y de qué manera procesarla. “*El volumen de*

datos que se mueve en el mundo bate récords día a día. El Complete Forecast Update, 2017–2022 de Cisco Systems prevé que en 2022 el tráfico web registrado en todo el planeta alcanzará los 4,8 zetabytes anuales, lo que supone pasar de 122 exabytes mensuales en 2017 a 396 exabytes dentro de apenas dos años. No es de extrañar que, según subraya Statista, el valor de mercado del big data en el mundo vaya a multiplicarse por dos en tan solo siete años: de 55.000 millones de dólares en 2020 a 103.000 millones en 2027. Todo un desafío para las soluciones de almacenamiento de datos, que se ven obligadas a adaptarse a marchas forzadas para dar respuesta a la demanda de las organizaciones." (IT Solutions de BETWEEN, 2020)

El almacenamiento de datos para Big Data debe estar preparado no solo para albergar gran cantidad de información, sino también para dar respuesta a determinadas necesidades inherentes al mismo como poder ser modulable respecto al almacenamiento exponencial de datos heterogéneos, ser compatible con datos estructurados y no estructurados, tener una baja latencia ante las solicitudes que se realicen, estar protegidos de potenciales ciber amenazas y ser de fácil acceso. Entre las distintas tecnologías de almacenamiento de datos se encuentran:

Data lakes o lagos de datos. Son repositorios que admiten datos estructurados y no estructurados procedentes de fuentes muy diversas y en bruto, sin necesidad de tratamiento previo antes de su inclusión. (IT Solutions de BETWEEN, 2020)

Edge computing. Útil cuando se trata de recolectar datos de millares de sensores del Internet de las Cosas, de manera que el almacenamiento y el procesamiento de la información se lleva a cabo cerca del punto de recolección, reduciendo así la latencia en la toma de decisiones. (IT Solutions de BETWEEN, 2020)

Cloud híbrida. Se basa en aprovechar las ventajas de combinar el uso de una nube pública, como Amazon Web Services o Microsoft Azure, y el de una nube privada, con una configuración a medida para los miembros de la organización. (IT Solutions de BETWEEN, 2020)

1.2.- Las 5 Vs del Big Data

Imagen 2. Las 5 Vs del Big Data. https://isarq.com (2018), ¿De qué hablamos cuando hablamos de Big Data?

Las características del Big Data se pueden resumir en lo que se conoce como las 5Vs (Imagen 2).

1) Volumen: en el Big Data el volumen de datos es tan grande que no cabría en una sola máquina, por lo tanto, se requieren herramientas y marcos especializados para almacenar, procesar y analizar dichos datos. Por ejemplo, las aplicaciones de redes sociales procesan miles de millones de mensajes todos los días, las industrias y los sistemas de energía pueden generar terabytes de datos de sensores todos los días, etc. Los volúmenes de datos generados por industria, salud, Internet de las Cosas y otros sistemas están creciendo exponencialmente impulsada por la reducción de los costes de almacenamiento de datos y arquitecturas de procesamiento y la necesidad de extraer información valiosa de los datos para mejorar los procesos comerciales, la eficiencia y el servicio a los consumidores. Aunque no existe un umbral fijo para que el volumen de datos se considere Big Data, sin embargo, por lo general, este término se usa para datos de escala masiva que son difíciles de almacenar, administrar y

procesar utilizando bases de datos tradicionales y arquitecturas de procesamiento de datos clásicas.

2) Velocidad: La velocidad de los datos se refiere a cómo de rápido se generan estos. Los datos se pueden generar a velocidades muy altas, por ejemplo, datos de redes sociales o datos de sensores. La alta velocidad de generación de los datos es otra de las características importantes de Big Data y la razón principal del crecimiento exponencial de datos. La alta velocidad de los datos hace que el volumen de datos acumulados se convierta en muy grande, en poco tiempo. Algunas aplicaciones pueden tener plazos estrictos para el análisis de datos (como el comercio o la detección de fraudes en línea) y los datos deben analizarse en tiempo real. Se requieren herramientas especializadas para ingerir datos de alta velocidad en la infraestructura de Big Data y analizar los datos en tiempo real.

3) Variedad: La variedad se refiere a las **formas de los datos**. Tal y como ya se ha explicado anteriormente, los grandes datos **se presentan en diferentes formas, como estructurados, no estructurados o semiestructurados, incluidos datos de texto, imagen, audio, video y datos de sensores**. Los sistemas de datos deben ser lo suficientemente flexibles para manejar tal variedad de datos.

4) Veracidad: La veracidad se refiere a cómo de precisos son los datos. Para extraer valor de los datos, los datos deben ser limpiados para eliminar el ruido. Las aplicaciones basadas en datos solo pueden aprovechar los beneficios del Big Data cuando los datos son significativos y precisos. Por lo tanto, la limpieza de los datos es importante para que los datos incorrectos y defectuosos se pueden filtrar.

5) Valor: El valor de los datos se refiere a la utilidad de los datos para el propósito previsto. El objetivo final de cualquier sistema de análisis de Big Data es extraer valor de los datos.

1.3.- Aplicaciones del Big Data

El constante crecimiento en la generación de datos ha provocado que el ámbito de aplicación del Big Data sea inmenso. Como muestra, enumeraré algunos

ejemplos de aplicaciones extraídos del libro Big Data Analytics: A hands-On Approach (A. Bahga y V. Madisetti, 2019)

Web

Análisis Web

El análisis web se ocupa de la recopilación y el análisis de datos del usuario sobre visitas en sitios web y aplicaciones en la nube. El análisis de estos datos puede proporcionar información sobre la participación del usuario y el seguimiento del rendimiento de las campañas publicitarias en línea. Para recopilar datos sobre las visitas de los usuarios, se utilizan dos enfoques. En el primer enfoque, las visitas de los usuarios se registran en el servidor web que recopila datos como la fecha y la hora de visita, recurso solicitado, dirección IP del usuario, código de estado HTTP, por ejemplo. El segundo enfoque, llamado etiquetado de página, utiliza un JavaScript que está incrustado en la página web. Cada vez que un usuario visita una página web, JavaScript recopila datos del usuario y los envía a un servidor de recopilación de datos de terceros. Se asigna una cookie al usuario que identifica al usuario durante esa visita y las visitas posteriores. El beneficio de este etiquetado de página es que facilita la recopilación y el análisis de datos en tiempo real. Este enfoque permite a servicios de terceros, que no tienen acceso al servidor web recopilar y procesar los datos. Estos proveedores de servicios de análisis especializados (como Google Analytics) ofrecen análisis avanzados e informes resumidos. Los informes incluyen sesiones de usuario, visitas a la página, páginas de entrada y salida principales, tasa de rebote, página más visitada, tiempo dedicado a cada página, número de visitantes únicos, número de visitantes repetidos, etc.

Supervisión del rendimiento

Aplicaciones web y en la nube de comercio electrónico, atención sanitaria, banca y finanzas, comercio minorista y aplicaciones de redes sociales pueden experimentar cambios rápidos en sus cargas de trabajo. Para garantizar el buen rendimiento de tales aplicaciones es necesario proporcionar los recursos adecuados para que las aplicaciones puedan satisfacer las demandas y los niveles de carga de trabajo especificados y al mismo tiempo garantizar que se cumplan los acuerdos de nivel de servicio.

El aprovisionamiento y la planificación de la capacidad es una tarea compleja, pues el sobre aprovisionamiento con antelación para tales sistemas no es económicamente factible. La computación en la nube proporciona un enfoque prometedor para aumentar o reducir dinámicamente la capacidad basada en la carga de trabajo de la aplicación. Para la gestión de recursos y las decisiones de planificación de la capacidad es importante comprender las características de la carga de trabajo de tales sistemas, medir la sensibilidad del rendimiento de la aplicación a los atributos de la carga de trabajo y detectar cuellos de botella en los sistemas. Pruebas de rendimiento de las aplicaciones antes de la implementación puede revelar cuellos de botella en el sistema y respaldar el aprovisionamiento y la capacidad decisiones de planificación. Los sistemas de Big Data se pueden utilizar para analizar los datos generados por tales pruebas, para predecir el rendimiento de la aplicación bajo cargas de trabajo pesadas e identificar cuellos de botella en el sistema para que los fallos puedan ser previstos. Los cuellos de botella, una vez detectados, pueden ser resuelto mediante el aprovisionamiento de recursos informáticos adicionales.

Orientación y análisis de anuncios

Los anuncios gráficos y de búsqueda son los tipos de publicidad que más se utilizan para la publicidad en Internet. En la publicidad basada en búsquedas, se muestra a los usuarios anuncios junto con los resultados de su búsqueda basada en palabras clave específicas en un motor de búsqueda. Los anunciantes pueden crear anuncios usando las herramientas proporcionadas por los motores de búsqueda o las redes sociales. Estos anuncios están configurados para palabras clave específicas relacionadas con el producto o servicio que se anuncia. Los anuncios gráficos son otro tipo de publicidad en Internet, en la que se muestran los anuncios dentro de sitios web, videos y aplicaciones móviles. La red publicitaria muestra los anuncios en función del contenido del sitio web, video o aplicación móvil. El método de pago más utilizado para los anuncios en Internet es el pago por clic, en el que los anunciantes pagan cada vez que un usuario hace clic en un anuncio. Las redes publicitarias utilizan grandes sistemas de datos para mostrar los anuncios y generar informes de estadísticas de anuncios. Los anunciantes pueden usar herramientas de Big Data para hacer seguimiento del rendimiento de los anuncios, optimizar las ofertas de pago por clic, rastrear

qué palabras clave se vinculan más a las páginas de destino de la publicidad u optimizar la asignación de presupuesto a varias campañas publicitarias.

Recomendación de contenido

Las aplicaciones de contenido (como aplicaciones de transmisión de música y video), recopilan varios tipos de datos como patrones de búsqueda, historial de navegación, historial de contenido consumido y calificaciones de los usuarios. Estas aplicaciones pueden aprovechar los grandes sistemas de datos para recomendar nuevos contenidos a sus usuarios en función de sus preferencias e intereses.

Finanzas

Modelos de riesgo de crédito

Las instituciones bancarias y financieras utilizan modelos de riesgo de crédito para calificar las solicitudes de crédito y predecir si un prestatario incumplirá o no en el acuerdo. Los modelos de riesgo crediticio se crean a partir de los datos históricos de clientes (propios o de otras agencias de crédito) utilizando información como el historial de crédito, datos de saldo de cuenta, transacciones de la cuenta, patrones de gasto, etc. Dado que la información de clientes puede ser obtenida de múltiples fuentes la cantidad de datos que se manejan pueden ser masivos, por ello el Big Data puede ayudar a la construcción de modelos de crédito.

Detección de Fraude

Las instituciones bancarias y financieras pueden aprovechar los sistemas de Big Data para detectar fraudes tales como fraudes con tarjetas de crédito, lavado de dinero y fraudes en reclamaciones de seguros. El análisis en tiempo real puede ayudar a analizar datos de fuentes dispares y etiquetas de transacciones en tiempo real. Los modelos de aprendizaje automático pueden construirse para detectar anomalías en las transacciones y detectar actividades fraudulentas o utilizar datos históricos sobre el cliente para buscar patrones que indiquen fraude.

Sistemas de Salud

El sistema sanitario consta de numerosas entidades, incluidos proveedores de atención médica (médicos de atención primaria, especialistas u hospitales), pagadores (gobierno, salud privada compañías de seguros), compañías farmacéuticas, etc. El proceso de prestación de servicios de salud implica la generación masiva de datos masivos de atención médica en diferentes formatos y almacenados en fuentes de datos dispares. Para promover una mayor coordinación de la atención entre los múltiples proveedores involucrados con los pacientes, su información clínica se agrega cada vez más a los sistemas de Registro Electrónico de Salud (RES). Los RES capturan y almacenan información sobre salud del paciente: resultados, diagnósticos, tratamientos, datos demográficos, etc. Aunque el uso principal de los RES es tener junta toda la información de un paciente individual y proporcionar acceso eficiente a los datos almacenados en el punto de atención, los RES pueden ser la fuente de valiosa información agregada sobre las poblaciones. Los sistemas de análisis de Big Data permiten el análisis de datos clínicos a gran escala y facilitan el desarrollo de una asistencia sanitaria más eficiente, mejoras en la precisión de las predicciones y ayuda en la toma de decisiones correctas. Algunas aplicaciones concretas son: vigilancia epidemiológica, aplicación de inteligencia de decisiones basada en la similitud del paciente, detección de anomalías en las reclamaciones, monitoreo de salud en tiempo real, etc.

Internet de las Cosas (Internet of Things (IoT))

Los sistemas IoT pueden aprovechar las tecnologías de big data para el almacenamiento y análisis de datos. Algunas aplicaciones que se pueden beneficiar de estos sistemas son:

<u>Detección de intrusiones</u>

Los sistemas de detección de intrusos utilizan cámaras y sensores de seguridad (como sensores PIR y sensores de puertas) para detectar intrusiones y generar alertas. Las alertas pueden ser en forma de un SMS o un correo electrónico enviado al usuario. Los sistemas avanzados pueden incluso enviar alertas detalladas, como una captura de imagen o un video corto enviado como un archivo adjunto de correo electrónico.

Carreteras inteligentes

Las carreteras inteligentes equipadas con sensores pueden proporcionar información sobre la conducción, estimaciones de tiempo de viaje, alertas en caso de malas condiciones de conducción, congestiones de tráfico o accidentes. Esta información puede ayudar a que las carreteras sean más seguras y

ayudar a reducir los atascos de tráfico. La información detectada desde las carreteras se puede comunicar a través de Internet a aplicaciones de análisis de Big Data en la nube. Los resultados del análisis pueden ser difundida a los conductores que se suscriban a dichas aplicaciones o a controladores de tráfico.

Parkings inteligentes

Los estacionamientos inteligentes facilitan la búsqueda de plazas de aparcamiento a los conductores, reciben datos de sistemas IoT que detectan el número de plazas de aparcamiento vacías y envían la información a través de Internet. Los conductores pueden acceder a estas aplicaciones desde teléfonos inteligentes, tabletas y sistemas de navegación para automóviles. En un estacionamiento inteligente, un sensor colocado en cada plaza detecta si se encuentra ésta se encuentra libre u ocupada. Esta información es administrada por un controlador inteligente y enviada posteriormente a través de internet para su procesado en la nube

Medio Ambiente

Los sistemas de monitoreo ambiental generan un gran volumen de datos a alta velocidad. Estos datos pueden ayudar a comprender el estado actual del medio ambiente y también a predecir tendencias ambientales. Algunos sistemas de monitoreo ambiental que pueden beneficiarse de los sistemas de Big Data son:

Monitoreo meteorológico

Los sistemas de monitoreo del clima pueden recopilar datos como temperatura, humedad o presión y enviarlos. Estos datos pueden ser luego analizados y visualizado para monitorear el clima y generar alertas meteorológicas.

Monitorización de la contaminación del aire

Los sistemas de monitoreo de la contaminación del aire pueden controlar las emisiones de gases nocivos por fábricas y automóviles que utilizan sensores gaseosos y meteorológicos. Los datos recopilados se pueden analizar para tomar decisiones informadas sobre enfoques de control de la contaminación.

Monitorización de la contaminación acústica

Debido al creciente desarrollo urbano, los niveles de ruido en las ciudades han aumentado e incluso se han vuelto alarmantemente altos en algunas ciudades. La contaminación acústica puede causar riesgos para la salud de los seres humanos debido a la interrupción del sueño y al estrés. El monitoreo de la contaminación acústica puede ayudar a generar mapas de ruido para las ciudades. Los mapas de ruido urbano pueden ayudar los responsables políticos o de mantener el orden en decisiones de planificación urbana, de elaboración de políticas para controlar los niveles de ruido cerca zonas residenciales, escuelas o parques, en detección y freno inmediato de niveles altos, etc.

Detección de incendios forestales

Los incendios forestales pueden causar daños a los recursos naturales, la propiedad y la vida humana. La detección temprana de estos incendios puede ayudar a minimizar el daño. Sistemas de detección de incendios forestales

usan sistemas de monitoreo implementados en diferentes ubicaciones en un bosque. Sistemas de predicción basados en Big Data que incluyan elementos como las condiciones ambientales, la temperatura, humedad, o niveles de luz por ejemplo pueden ayudar a la detección temprana.

Otras **aplicaciones de monitoreo** que comúnmente sacan partido del Big Data pueden ser las aplicaciones para la Detección de Desbordamiento de Ríos o aplicaciones para la Monitorización de la Calidad del Agua.

Logística y Transporte

Las herramientas del Big Data pueden ser también de gran utilidad y aplicación en temas relacionados con la Logística y el Transporte. Algunas aplicaciones en las que actualmente se está utilizando y sacando partido de ello en este contexto son: seguimiento de vehículos o flotas en tiempo real, rastreo de envíos,

diagnóstico remoto de problemas en vehículos, generación de rutas y horarios, etc.

Como ya he dicho, son solo algunas aplicaciones generales del Big Data en la actualidad, existen muchas más en los campos descritos anteriormente u otros como la industria o las ventas.

2. El Big Data orientado a la Defensa y la Seguridad

De manera general se puede entender que la aplicación de Big Data a defensa y seguridad persigue capturar y utilizar grandes cantidades de datos para poder aunar sensores, percepción y decisión en sistemas autónomos, y para incrementar el entendimiento de la situación y contexto del analista y el combatiente.

Aplicaciones concretas puede ser la vigilancia y seguridad en fronteras, la ciberdefensa / ciberseguridad, la lucha contra terroristas o crimen organizado, la lucha contra el fraude, la inteligencia militar o el planeamiento táctico de misiones, y por supuesto la seguridad ciudadana, siendo este último campo de aplicación donde estudiaremos un caso concreto y el sesgo producido por el aprendizaje de la máquina o machine bias.

De la generación masiva de datos también se pueden aprovechar los ámbitos de Seguridad y Defensa. A continuación, se detallan algunas de las aplicaciones del Big Data en entornos de Seguridad y Defensa, extraídas del informe *“Big Data en los Entornos de Defensa y Seguridad” del grupo de trabajo sobre Big Data, comisión de investigación de Nuevas Tecnologías del Centro Superior de Estudios de la Defensa Nacional*. (J.A. Carrillo Ruiz, J.E. Marco De Lucas, J.C. Dueñas López, F. Cases Vega, J. Cristino Fernández, G. González Muñoz de Morales, L.F. Pereda Laredo, 2013).

Detección de intrusión física en grandes espacios o infraestructura abierta.

La seguridad de instalaciones o infraestructuras que cubren una gran extensión es de especial interés por las elevadas dificultades técnicas o de costes que supone.

La protección de estos espacios conlleva la detección de posibles amenazas, su clasificación, y en el caso necesario, su localización y seguimiento. El incremento del área que ocupa implica o más sensores o mejores resoluciones en estos sensores, o en muchas ocasiones una combinación de ambos, lo que en cualquier modo se traduce en una explosión de datos. Esta cantidad de datos, y la necesidad de monitorizar en tiempo real de manera automática hace que este problema sea un candidato idóneo para Big Data.

Un ejemplo concreto y real de esto es el caso de IBM, con la puesta en marcha del proyecto Terraechos para el Laboratorio Nacional del Departamento de Energía de EE.UU. El objetivo del sistema era analizar de manera continua grandes cantidades de datos acústicos provenientes de los objetos en movimiento del entorno del perímetro y área a vigilar, tanto sobre la superficie como debajo de la superficie. Para cumplir con su cometido el sistema debe discriminar en tiempo real sonidos provenientes de animales, arboles, climatología, de posibles intrusos. Para ello, implementaron una red de sensores acústicos provenientes de la US Navy, que alimentaban continuamente con terabytes de información un procesador (IBM InfoSphere Streams) que analiza los datos acústicos.

Computación sobre información cifrada.

Actualmente muchos de los datos de interés para una aplicación residen en servidores deslocalizados físicamente de donde se está accediendo y necesitando esos datos e información. Un ejemplo muy actual son los servidores en la “nube”. En el ámbito de seguridad y defensa también se da esta configuración y en ellos, como es lógico, la seguridad de la información es un tema especialmente relevante.

Algunos problemas de la encriptación son que enlentecen la computación y que requieren de altas capacidades de computación. Por tanto, la aplicación del Big Data en este entorno tiene por objetivo conseguir que la computación sobre información cifrada sea práctica.

Una manera de abordar este desafío, es poder operar, computar, sobre esos datos sin necesidad de desencriptarlos previamente. Mediante este planteamiento la seguridad de estos depende de la robustez del cifrado, pero no

se generarían debilidades en el acceso y manejo de los datos para la intercepción de la información por un tercero.

Un ejemplo real de trabajo en esta área es el proyecto PROCEED (Programming Computation on Encrypted Data) de la agencia DARPA (Defense Advanced Research Projects Agency) de EE.UU. DARPA es “una agencia gubernamental estadounidense que tiene el objetivo de prevenir los impactos estratégicos de los posibles enemigos de EE.UU. mediante el desarrollo de actividades que permitan mantener la superioridad tecnológica y militar”.

El Big Data para Computación sobre Información Cifrada es útil también para comunicaciones seguras y para bancos de datos para los ámbitos financieros, seguridad interior, inteligencia o defensa.

Análisis automático de vulnerabilidades de red.

La ciberseguridad y ciberdefensa, y la protección de redes de telecomunicación son cada vez más importantes en un mundo tan tecnológico como el que vivimos. Dos de los objetivos principales de la aplicación del Big Data en este mundo de ciberseguridad y ciberdefensa son:

- Reducir la cantidad de tiempo que los analistas pasan descubriendo ciberataques al agrupar y correlacionar fuentes de redes de datos dispares.
- Incrementar la precisión, tasa y velocidad de detección de ciber-amenazas a redes de ordenadores.

Para ello es necesario cambiar el modo en que la información relativa a las redes es adquirida, procesada y puesta a disposición de los ciber-defensores. Procurando proporcionarles datos conectados y correlacionados, para que puedan abordar directamente el problema del orden de escala de los datos asociados a la seguridad de las redes.

Dos ejemplos de estos programas en el mundo de la Defensa son el Cyber Targeted-Attack Analyzer y el Cyber-Insider Threat (CINDER), ambos de DARPA.

En la actualidad las organizaciones deben hacer frente a nuevos riesgos originados principalmente por dos desafíos:

1. Disolución de los límites de las redes: Las aplicaciones y datos corporativos son cada vez más accesibles mediante servicios en la nube y dispositivos móviles, rompiendo los últimos límites de la red corporativa e introduciendo nuevos riesgos en la información y vectores de amenaza.

2. Adversarios más sofisticados: Los ciberatacantes se han vuelto más adeptos a realizar ataques complejos y muy específicos que evitan las defensas tradicionales. A menudo, los ataques o fraudes no son detectados hasta que el daño se ha realizado.

Es por esto, que sea hacen necesario soluciones más ágiles basadas en evaluaciones dinámicas de riesgo. El análisis de grandes volúmenes de datos y operaciones de seguridad en tiempo real son esenciales para proporcionar una seguridad significativa.

Análisis de vídeo en tiempo real y recuperación rápida en librerías de vídeo.

Con el crecimiento de los sistemas de video vigilancia instalados en las ciudades se recogen una cantidad enorme de datos de vídeo. Esta cantidad es especialmente masiva en labores de inteligencia, vigilancia, adquisición de objetivos y reconocimiento. Debido a este gran y rápido crecimiento surgen las siguientes necesidades/aplicaciones en las que el Big Data puede ser de gran utilidad:

• Análisis de vídeo en tiempo real.

• Búsqueda y recuperación rápida en librerías de vídeo.

De esta manera es posible para los analistas establecer alertas asociadas a diferentes actividades y sucesos de interés mientras estos están ocurriendo (por ejemplo, alertas del tipo “una persona acaba de entrar en el edificio”). O incluso de manera predictiva en base a patrones de hechos ya conocidos, que permitan adelantarse a los acontecimientos en un tiempo suficiente para poder reaccionar frente a las amenazas.

En el caso de la segunda aplicación, se persigue el desarrollo de agentes inteligentes que permitan encontrar contenidos de vídeo de interés en librerías con en miles de horas de grabaciones de vídeo. Hasta ahora, la mayoría de las búsquedas en librerías de video requieren una gran cantidad de intervención

humana. El objetivo por tanto, es detectar de manera rápida y precisa en estas grandes librearías de video actividades potencialmente sospechosas, o trabajar en la detección de objetos (por ejemplo vehículos: acelerando, girando, parando, adelantando, explotando o en llamas, etc.), o también identificar comportamientos extraños de una persona (excavando, dejando abandonado un objeto, etc.), o en interacciones hombre-hombre (siguiendo, reuniéndose, moviéndose conjuntamente, intercambiando objetos, etc.) que puedan resultar de interés para una investigación.

Ejemplos de programas en esta línea son el VIRAT (Video and Image Retrieval and Analysis Tool) de DARPA y el Intelligent Video Analytics de IBM.

Inteligencia visual en máquinas.

El objetivo o reto en este punto es que el sistema de visión artificial sea verdaderamente "inteligente", siendo capaz no sólo de reconocer los objetos, sino también de identificar y razonar sobre las acciones y relaciones que se establezcan entre dichos objetos. De esta manera, un sistema de estas características sería capaz no sólo de identificar los objetos, sino de describir lo que está sucediendo en la escena, y procesar alertas en caso de ser necesario.

El desafío es también que las tecnologías de inteligencia visual permitan construir cámaras inteligentes capaces de responder a preguntas del tipo "¿Qué acaba de pasar en la escena?". En esta evolución, las tecnologías de Big Data juegan un papel determinante.

Entre las iniciativas destacadas en este campo se puede mencionar el programa PERSEAS (Persistent Stare Exploitation and Analysis System) y el Mind's Eye, ambos de la Agencia DARPA.

Análisis de texto como apoyo a la toma de decisiones en tiempo real en entornos intensivos de datos.

En un gran número de operaciones militares y operaciones de seguridad, los datos en formato de texto proporcionan información esencial sobre actitudes, sucesos o relaciones entre los distintos actores implicados en la operación.

En las últimas décadas, el acelerado ritmo de avance de las tecnologías de las telecomunicaciones (internet, telefonía móvil, etc.) ha hecho posible que se

disponga de cantidades muy grandes de textos en una gran variedad de formatos. La aplicación de las técnicas de análisis textual a este gran volumen de datos ofrece nuevas posibilidades para la extracción de las informaciones relevantes para la toma de decisiones. Las técnicas de Big Data ayudan a que este análisis textual se realice en tiempo real y se obtengan resultados exhaustivos y precisos, apoyando eficazmente a la toma de decisión en las ventanas de tiempo marcadas por las misiones.

Son importantes aspectos como la representación eficiente de la información, la capacidad de alcanzar altos ratios de compresión de la información sin pérdida de fidelidad o el comportamiento anticipatorio de los recursos computacionales.

Ejemplos de estos programas son el Data to Decisions (D2D), promovido por la OSD (Office of the Secretary of Defense) o el programa XDATA de la agencia DARPA.

Traducción automática a gran escala.

A pesar de que en años recientes se produjeron notables avances en la recuperación de textos y tratamiento de la información (un ejemplo son los motores de búsqueda inteligentes y adaptativos), el progreso en la “digestión” o la comprensión de la información no avanza con el mismo ritmo.

Este hecho se complica al considerar que muchas de las áreas donde se realizan las operaciones son muy ricas en idiomas y dialectos, o que la globalización requiera una concepción de la seguridad no circunscrita a las fronteras, y por lo tanto con gran diversidad lingüística.

Por estas razones, las capacidades de traducción automática a gran escala (en número de idiomas y en volumen de texto/audio procesable) se convierten en fundamentales para las necesidades operativas de muchas misiones de defensa y de seguridad. Esto es especialmente crítico en el caso de las misiones internacionales, donde el lenguaje proporciona muchos de los elementos clave para entender la cultura, el carácter, las opiniones y apreciaciones de los naturales del país donde se desarrolla la operación.

En este sentido, el desarrollo de algoritmos computacionalmente eficientes para el procesamiento (ontologías y minería de datos), la extracción de información a

gran escala, el reconocimiento de voz y el resumen automático, representan una oportunidad para la aplicación de Big Data a las necesidades de traducción en defensa y seguridad.

Además de todas las aplicaciones descritas anteriormente, existen muchas otras en las que el Big Data puede ayudar, como pueden ser la predicción de eventos, el control y comportamientos de multitudes, la preparación de seguridad en eventos singulares (deportivos, políticos, etc.), la identificación de anomalías, patrones y comportamientos en grandes volúmenes de datos, planteamiento táctico de misiones, etc.

Predicción de eventos.

Los analistas de los servicios de inteligencia necesitan de la capacidad de monitorizar y analizar rápidamente la información de eventos formada por grandes volúmenes de datos de texto no estructurados con el fin de lograr y mantener un entendimiento de los acontecimientos y poder formular sobre ellos estimaciones y predicciones a futuro. La cantidad de datos de texto no estructurados disponibles va mucho más allá de lo que se puede leer y procesar en un tiempo determinado. Es aquí donde el Big Data puede contribuir y facilitar el manejo y análisis de toda esta cantidad enorme de datos.

Uno de los grandes retos tecnológicos en este campo es avanzar en la extracción de eventos con sus atributos de modalidad, polaridad, especificidad y tiempo. La modalidad de un evento indica si se trata de un evento real o no. Ejemplos de modalidad de un evento son asertiva “la bomba explotó el domingo”, creencia “se cree que será condenado”, hipotética “si fuese arrestado, se le acusaría de asesinato”, etc. La polaridad de un evento indica si el evento ocurrió o no.

Para lograr estos retos se investigan nuevos métodos innovadores que puedan apoyar la predicción de eventos, entre ellos destacan:

- Mediante análisis textual, la extracción, caracterización y monitorización continua de redes sociales y grupos de interés. Extracción de las tendencias temporales, es decir, la frecuencia de los contactos entre los nodos o grupos, tiempo con el contacto, contactos periódicos, ruta de retraso en la difusión de la información, etc.

- Detección de la presencia de nodos puente que permiten descubrir subredes ocultas, y determinar el flujo de recursos (información, dinero, influencia) dentro de la red social.

Áreas de aplicación relacionadas	En ámbito militar en humint/operaciones en entornos urbanos. Control y comportamientos de multitudes. Seguridad ciudadana. Preparación de seguridad de eventos singulares (deportivos, políticos, etc.)

Factores impulsores y limitadores en la aplicación de Big Data en Seguridad y Defensa.

Con objeto de poder valorar la evolución del Big Data en general, y sobre todo en lo referente a su aplicación a Seguridad y Defensa, se ha realizado un análisis de cuáles son los factores más influyentes en el desarrollo de estas aplicaciones de Big Data. Estos factores responden a varios aspectos como pueden ser las características técnicas de Big Data, el entorno de seguridad y defensa, el contexto social-económico, etc. Se han dividido entre **factores impulsores del desarrollo**, es decir que demandan o favorecen el desarrollo y aplicación de Big Data, y **factores limitadores**, que, por el contrario, frenan o presentan desafíos destacables para el desarrollo y aplicación de Big Data.

Imagen 3. Factores a favor y en contra y su importancia para la aplicación de Big Data en Seguridad y Defensa.

A modo de resumen y de visión global, en la Imagen 3 se presenta el conjunto total de estos factores, quedando en la parte superior los favorables en la aplicación de Big Data, y en la parte inferior, aquellos que suponen un desafío o una limitación en la aplicación de Big Data en seguridad y defensa (y posiblemente en también en otros ámbitos).

Adicionalmente, se ha estimado la importancia de cada uno de estos factores, tanto para el favorecimiento como para la limitación de Big Data, reflejando esta importancia con el tamaño de cada una de las flechas.

Tanto en el ámbito de la seguridad como en el ámbito de la defensa la aplicación del Big Data está muy enfocada a la **prevención**, pues esta requiere tomar decisiones en ventanas de tiempo muy definidas y con un alto nivel de síntesis de la inmensidad de datos y factores involucrados.

2.1.- La Criminología Computacional

Una de las áreas donde los conceptos de Big Data encuentran una aplicación directa es la Criminología Computacional, donde la capacidad de analizar grandes volúmenes de datos relacionados con actividades criminales multiplica las posibilidades de neutralización de las amenazas relacionadas con dichas actividades. En este contexto, Big Data estaría relacionado con técnicas como el minado de datos criminales, el análisis de agrupaciones y el aprendizaje de reglas de asociación para la predicción del crimen, análisis de redes criminales, análisis de textos multilingües, análisis de opiniones y sentimientos, etc.

Programas de investigación como el COPLINK o el Dark Web Research de la Universidad de Arizona ofrecen un excelente ejemplo del potencial de estas tecnologías. COPLINK, desarrollado inicialmente con fondos de la National Science Foundation y el Departamento de Justicia de Estados Unidos, es un sistema de compartición de información y de minado de datos criminales utilizado por más de 4500 departamentos de policía en los Estados Unidos y por 25 países OTAN. El sistema COPLINK fue adquirido por IBM en 2011. Dark Web, financiado por la National Science Foundation y el Departamento de Defensa, ha generado una de las mayores bases de datos existentes para la investigación del terrorismo, con cerca de 20 terabytes de información sobre sitios web y contenidos de redes sociales relacionados con terrorismo.

Otro punto de vista de la criminología computacional donde Big Data es de aplicación es en el seguimiento de actividades sospechosas en diferentes redes (tanto internet como de una organización). Siendo de interés, identificar qué información es recogida como soporte para la actividad criminal.

3.- Estructura para el análisis del Big Data

En esta sección, proponemos un flujo genérico para análisis de Big Data, basado en el libro Big Data Analytics: A hands-On Approach (A. Bahga y V. Madisetti, 2019), donde los autores detallan los pasos involucrados en la implementación de una aplicación de análisis típica y las opciones disponibles en cada paso. La Imagen 4 muestra el flujo de análisis con varios pasos, seleccionar las opciones

para cada paso en el flujo de análisis puede ayudar a determinar las herramientas y los marcos adecuados para realizar los análisis.

Recopilación de datos: La recopilación de datos es el primer paso para cualquier aplicación de análisis. Antes de que los datos puedan ser analizados, los datos deben recopilarse e ingerirse en una gran pila de datos. La elección de herramientas y marcos para la recopilación de datos depende de la fuente de datos y el tipo de datos que se están recogiendo. Para la recopilación de datos, se pueden utilizar varios tipos de conectores como marcos de mensajería de publicación-suscripción, colas de mensajería, conectores fuente-receptor, conectores de base de datos y conectores personalizados.

Preparación de los datos: Los datos a menudo pueden contener problemas que deben resolverse antes de que los datos puedan procesarse. Ejemplos de problemas en los datos pueden ser registros corruptos, valores faltantes, duplicados, abreviaturas inconsistentes, unidades inconsistentes, errores tipográficos, ortografía incorrecta o formato incorrecto, etc. La preparación de datos implica varias tareas, como la limpieza de datos, la manipulación de datos, la desduplicación, la normalización, el muestreo y filtrado. Por ejemplo, cuando recopilamos registros como archivos de texto sin procesar de diferentes fuentes, puede haber inconsistencias en los separadores de campos utilizados en diferentes archivos: algún archivo puede estar usando coma como separador de campo, otros pueden estar usando tabulador como separador, etc. Es necesario resolver estas inconsistencias analizando los datos sin procesar de diferentes fuentes y transformándolo en un formato consistente. La normalización es necesaria cuando los datos de diferentes fuentes usan diferentes unidades o escalas o tiene diferentes abreviaturas para la misma cosa. Por ejemplo, los datos meteorológicos informados por algunas estaciones pueden contener la temperatura en la escala Celsius, mientras que los datos de otras estaciones pueden usar la escala Fahrenheit. El filtrado y el muestreo pueden ser útiles cuando queremos procesar solo los datos que cumplen ciertas reglas. El filtrado también puede ser útil para rechazar malos registros con valores incorrectos o fuera de rango.

Tipos de Análisis: El siguiente paso en el flujo de análisis es determinar el tipo de análisis para el problema concreto sobre el que deseamos realizar el estudio. En la Imagen 4 aparece una clasificación de los tipos de análisis y algoritmos o métodos populares para cada tipo de análisis. Por ejemplo, los modelos de clustering se encargan de la segmentación de conjuntos de datos en varios grupos o clusters. Los modelos de clasificación desarrollan procesos para clasificar eficientemente. Los modelos de regresión buscan determinar la relación entre una variable dependiente, con respecto a otras variables, y en base a estas relaciones realizar predicciones. Otro de los tipos de análisis es el análisis de series temporales, que analiza secuencia de datos recopilados durante un intervalo de tiempo. Algunos de estos tipos de análisis y algoritmos o modelos concretos los analizaremos en las próximas secciones.

Modos de Análisis: Con los tipos de análisis seleccionados, el siguiente paso es determinar el modo de análisis. Puede ser por lotes, en tiempo real o interactivo. La elección del modo depende de los requisitos de la solicitud. Si la aplicación exige que se actualicen los resultados después de breves intervalos de tiempo (por ejemplo, cada pocos segundos), se elige el modo de análisis en tiempo real. Sin embargo, si la aplicación solo requiere que los resultados se generen y actualicen en escalas de tiempo (digamos diarias o mensuales), entonces se puede usar el modo por lotes. Si la aplicación en cambio tiene flexibilidad para consultar datos, entonces el modo interactivo es útil. Una vez que se tiene la elección del tipo de análisis y el modo de análisis, se puede determinar el patrón de procesamiento de datos que se puede utilizar. Por ejemplo, para estadísticas básicas como el tipo de análisis y el modo de análisis por lotes, MapReduce puede ser una buena opción. Mientras que para el análisis de regresión como o el análisis en tiempo real (predicción de valores en tiempo real), el Stream Processing es una buena opción. La elección del tipo de análisis, el modo de análisis y el patrón de procesamiento de datos puede ayudar a preseleccionar las herramientas y los marcos adecuados para el análisis de los datos.

Visualización: Es importante, una vez realizado el análisis, visualizar los datos. Las visualizaciones pueden ser estáticas, dinámicas o interactivas. Las visualizaciones estáticas son aquellas que se utilizan cuando tiene los resultados

del análisis están almacenados en una base de datos de servicio y simplemente se desea mostrar los resultados. Sin embargo, si la aplicación exige que los resultados estén actualizados regularmente, entonces se utilizan las visualizaciones dinámicas (con widgets en vivo, gráficos o calibres). Si la aplicación requiere que el usuario indique valores o datos, entonces se usan visualizaciones interactivas.

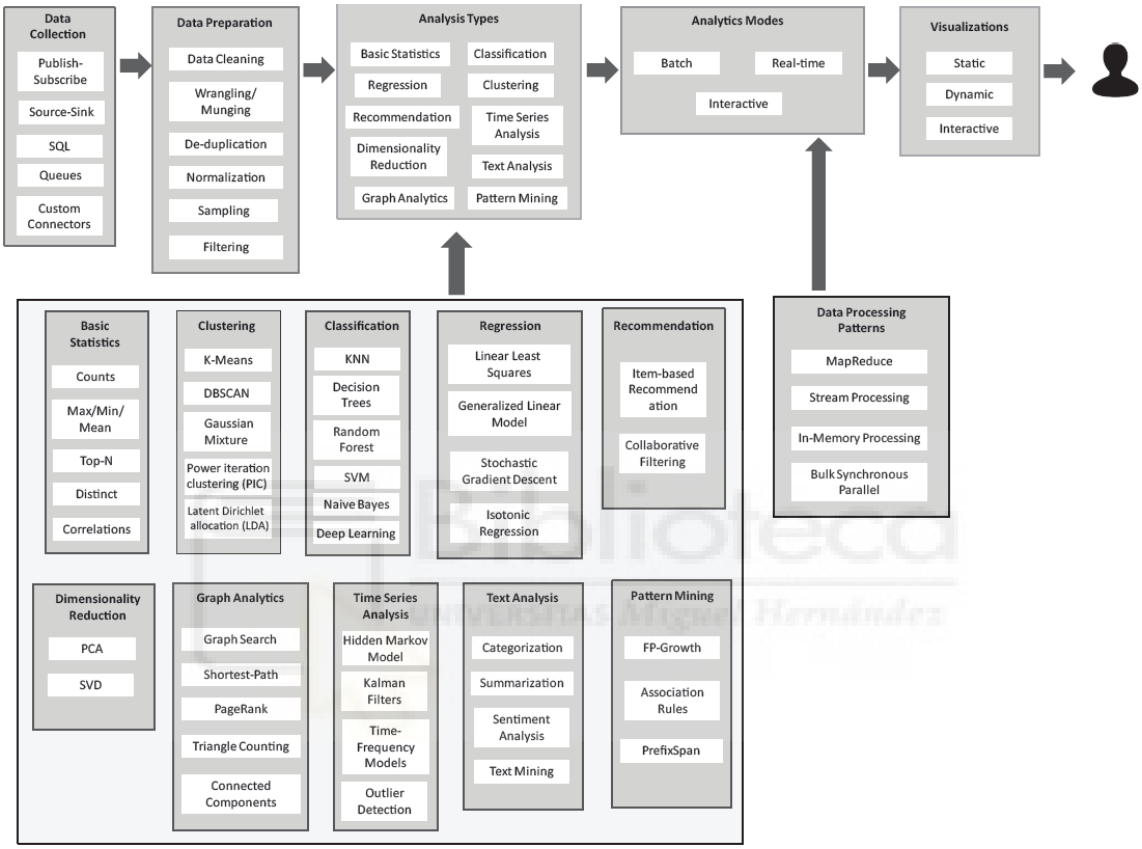


Imagen 4. Flujo de análisis de Big Data. Big Data Analytics: A Hands-On Approach.

Para entender la teoría relacionada con el Big Data, es conveniente dominar algunas definiciones. Durante la fase de procesamiento, para generar conocimiento a partir de los datos, se utilizan modelos y algoritmos, un algoritmo es un conjunto de instrucciones o reglas definidas y no ambiguas, ordenadas y finitas que permite solucionar un problema, realizar un cómputo, procesar datos y llevar a cabo otras tareas similares.

Además, para el desarrollo de la mayoría de las técnicas que se emplea en el procesamiento de los datos se utiliza la Estadística, la Matemática y la Informática, mediante la Minería de datos, Machine Learning (aprendizaje

automático), la Computación Paralela, La Inteligencia Artificial (IA), el Deep Learning, etc.

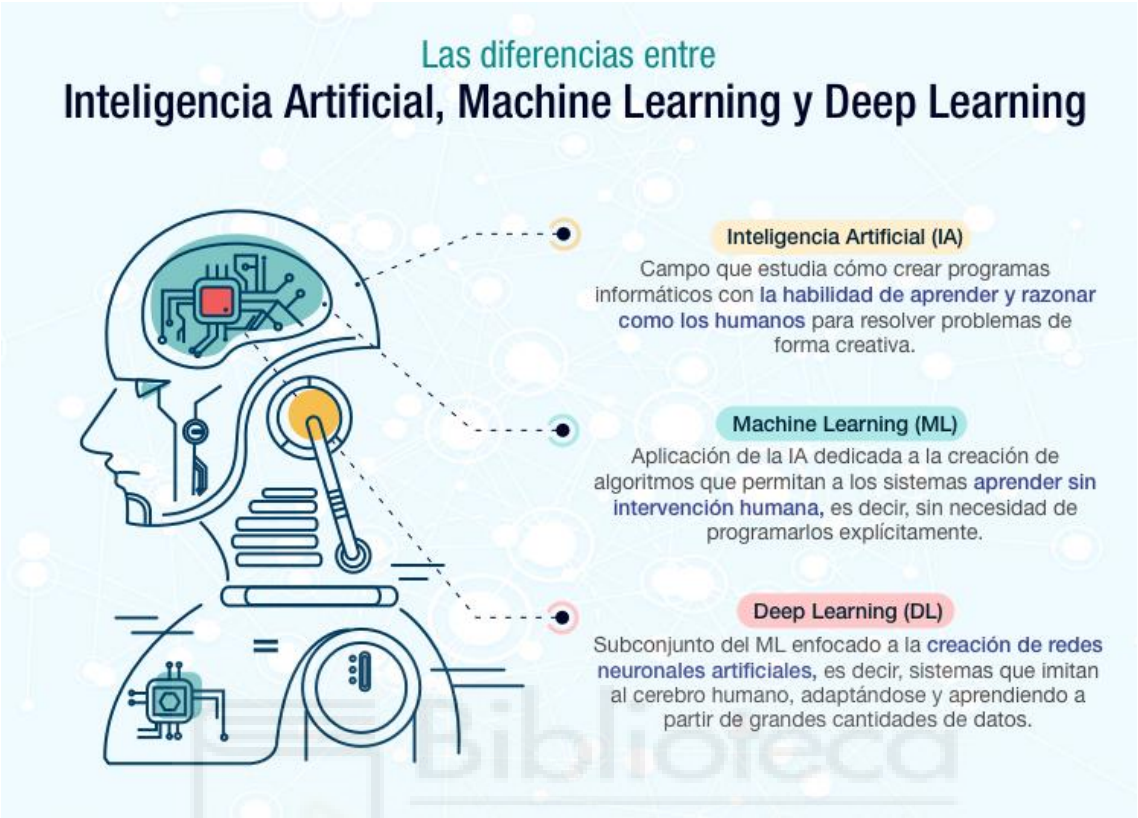


Imagen 5. Diferencias entre Inteligencia Artificial, Machine Learning y Deep Learning. Fuente: Qubole.com

El Machine learning

Se trata de una rama de la inteligencia artificial (IA) y la informática que se centra en el uso de datos y algoritmos para imitar la forma en la que aprenden los seres humanos, con una mejora gradual de su precisión. El primero en acuñar el término fue Arthur Samuel, empleado de IBM.

Machine learning es un componente importante del creciente campo de la ciencia de datos. Mediante el uso de métodos estadísticos, los algoritmos se entrenan para hacer clasificaciones o predicciones, y descubrir información clave dentro de los proyectos de minería de datos.

Los algoritmos más utilizados en Machine Learning son los basados en aprendizaje supervisados, capaces de automatizar un proceso tras haber aprendido a partir de un conjunto de ejemplos conocidos. El usuario facilita a un

algoritmo un ejemplo con pares de entrada/salida para que el algoritmo aprenda la relación de manera que sea capaz de facilitar una salida a partir de una entrada no vista antes y sin la supervisión de ningún ser humano. En este tipo de aprendizaje supervisado, se entrena al algoritmo a partir de un conjunto de datos etiquetados con la respuesta correcta, cuanto mayor es el conjunto de datos más y mejor aprenderá el algoritmo. Una vez finalizado el entrenamiento, se le introducen nuevos datos sin las respuestas correctas y el algoritmo de aprendizaje predecirá el resultado basándose en la experiencia adquirida.

El otro modelo de aprendizaje es el conocido como aprendizaje no supervisado, donde sólo se proporcionan los datos de entrada al modelo. En este caso, el algoritmo es entrenado usado un conjunto de datos que no tienen ninguna etiqueta, es decir, que no se le dice al algoritmo lo que representan los datos. La idea es que el algoritmo aprenda a descubrir por sí solo patrones que ayuden a entender el conjunto de datos.

Los clasificadores de machine learning se dividen en tres categorías principales.

Machine learning supervisado

El aprendizaje supervisado, también conocido como machine learning supervisado, se define por utilizar conjuntos de datos etiquetados para entrenar los algoritmos. Los datos etiquetados son aquellos que tienen una etiqueta o palabra clave que identifica sus características, propiedades o clasificaciones, valores conocidos. A medida que se introducen datos de entrada en el modelo, adapta sus pesos hasta que el modelo se haya ajustado correctamente. El aprendizaje supervisado permite a las organizaciones resolver una amplia variedad de problemas del mundo real a escala como, por ejemplo, la clasificación de spam en una carpeta distinta de la bandeja de entrada. Algunos métodos utilizados en el aprendizaje supervisado son las redes neuronales, Naïve Bayes, la regresión lineal, la regresión logística, el bosque aleatorio, la máquina de vectores de soporte (SVM), etc.

Machine learning no supervisado

El aprendizaje no supervisado, también conocido como machine learning no supervisado, utiliza algoritmos de machine learning para analizar y agrupar

conjuntos de datos sin etiquetar. Estos algoritmos descubren agrupaciones de datos o patrones ocultos sin necesidad de ninguna intervención humana. Su capacidad de descubrir similitudes y diferencias en la información lo convierten en la solución ideal para el análisis de datos exploratorios, las estrategias de venta cruzada, la segmentación de clientes, y el reconocimiento de imágenes y patrones.

Aprendizaje semisupervisado

El aprendizaje semisupervisado ofrece un punto intermedio entre el aprendizaje supervisado y no supervisado. Durante el entrenamiento, utiliza un conjunto de datos etiquetados más pequeño para guiar la clasificación y la extracción de características de un conjunto de datos sin etiquetar de mayor tamaño. El aprendizaje semisupervisado puede resolver el problema de no tener suficientes datos etiquetados (o no poder permitirse etiquetar suficientes datos) para entrenar un algoritmo de aprendizaje supervisado.

El procesamiento aplicando inteligencia, nos devuelve información útil para entre otras cosas, establecer perfiles, generar conocimiento, crear modelos de comportamiento, predecir situaciones futuras.

Algunos usos para el machine learning son, el reconocimiento de voz, chatbots en servicios de atención al cliente, reconocimiento de objetos en imágenes y videos digitales, motores de recomendación, etc.

Deep Learning

“El deep learning es un tipo de machine learning que entrena a una computadora para que realice tareas como las hacemos los seres humanos, como el reconocimiento del habla, la identificación de imágenes o hacer predicciones. En lugar de organizar datos para que se ejecuten a través de ecuaciones predefinidas, el deep learning configura parámetros básicos acerca de los datos y entrena a la computadora para que aprenda por cuenta propia reconociendo patrones mediante el uso de muchas capas de procesamiento.” (sas.com/es)

Reconocimiento del habla: Como ejemplo Google Now o Siri de Apple.

Reconocimiento de imágenes: Podría tener uso en el campo de la seguridad al revisar cámaras de vigilancia tratando de identificar a sospechosos. O en el campo de la conducción autónoma para detectar obstáculos o personas.

Procesamiento del lenguaje: Puede ser útil para descubrir patrones en diferentes tipos de textos.

Sistemas de recomendación: Plataformas como Amazon o Netflix entre otras tienen sistemas de recomendación basados en las compras y elecciones de sus usuarios, de igual manera puede utilizarse de forma más compleja para averiguar las preferencias musicales o para vestir de los usuarios.

3.1 Los árboles de clasificación.

Dentro de las técnicas de clasificación supervisada (Machine Learning supervisado) encontramos los **Árboles de decisión**, cuyo objetivo es predecir una variable respuesta en función de otras variables. Para construirlos, lo primero será determinar un conjunto de reglas sucesivas que progresivamente van particionando el conjunto original de datos, de esta forma ayudan a tomar decisiones de clasificación sobre nuevos datos que no pertenecen al conjunto original. Son de uso muy habitual puesto que dan muy buenos resultados, son fáciles de interpretar y sirven para variables cuantitativas y cualitativas.

Los árboles se pueden clasificar en dos tipos:

- Árboles de regresión en los que la variable respuesta es cuantitativa
- Árboles de clasificación en lo que la variable respuesta es cualitativa. En este trabajo utilizaremos este tipo de árboles.

Los elementos principales de un árbol de decisión son:

La Regla de división. Es decir, qué preguntas vamos a realizar y en qué momento. La idea es ir dividiendo los nodos para disminuir la impureza del nodo padre, un nodo será puro cuando contenga observaciones de una única clase o construir reglas de división que minimicen la tasa de clasificación errónea.

El Criterio de parada. Determinará la profundidad del árbol, es decir cuándo dejo de realizar particiones. Cuanto más profundizamos en las ramificaciones, mayor será la pureza de los nodos finales, no obstante, conforme ramificamos

vamos disminuyendo el número de observaciones de cada grupo y por tanto corremos el riesgo de tener un sobreajuste. Para evitar esto, se suele establecer como criterio de parada que el número de observaciones de los grupos esté por encima de un mínimo, por ejemplo 10.

La Asignación de clase. Cuando finalmente el árbol de decisión se encuentra formado con los grupos finales, se asignará a cada grupo la clase mayoritaria entre sus componentes.

Finalmente, para evaluar la precisión de los árboles de decisión podemos hacer uso de las **matrices de confusión**, mediante las cuales podemos relacionar las predicciones que estamos llevando a cabo, con los valores reales. Es decir, básicamente nos informará sobre cuantas veces hemos acertado con nuestro algoritmo de predicción.

4.- Sesgos en el uso del Big Data para la predicción de la delincuencia.

El sesgo de aprendizaje automático, a veces también llamado sesgo de algoritmo o sesgo de IA, es un fenómeno que ocurre cuando un algoritmo produce resultados que son sistemáticamente perjudiciales debido a suposiciones erróneas en el proceso de aprendizaje automático o porque existe sesgo en los datos utilizados.

El aprendizaje automático depende de la calidad, la objetividad y el tamaño de los datos de entrenamiento utilizados para enseñarlo. Los datos defectuosos, deficientes o incompletos darán como resultado predicciones inexactas, lo que refleja la expresión garbage in, garbage out " entra basura, sale basura" utilizada en informática para transmitir el concepto de que la calidad de la salida está determinada por la calidad de la entrada, si los datos originales son erróneos, incluso el programa informático más sofisticado producirá resultados erróneos.

El sesgo en el Machine Learning se deriva generalmente de problemas introducidos por las personas que diseñan y/o entrenan los sistemas de aprendizaje automático. Estas personas podrían crear algoritmos que reflejen

sesgos cognitivos no deseados o prejuicios de la vida real. También puede ocurrir que estas personas introduzcan sesgos porque usan conjuntos de datos incompletos, defectuosos o perjudiciales para entrenar y/o validar los sistemas de aprendizaje automático.

Por ejemplo, en 2016 se descubrió que algunos de los algoritmos de LinkedIn tenían un sesgo de género que, por ejemplo, recomendaba empleos mejor remunerados a hombres en vez de a mujeres. Esta casuística se veía reforzada por el hecho de que, en la sociedad actual, los puestos de elevada remuneración están predominantemente ocupados por hombres. De igual manera, cuando buscabas a un usuario femenino en esa red, era habitual que el motor de búsqueda te sugiriera un nombre masculino similar.

Otro ejemplo, en 2015 un desarrollador de software advirtió que el servicio de reconocimiento facial de Google había etiquetado las fotos de él con un amigo de color como “gorilas”. Google entonó el mea culpa y declaró que estaba “trabajando en soluciones a largo plazo”. Más de dos años después, uno de esos arreglos corresponde a borrar los términos relativos a gorilas y algunos otros primates del léxico del servicio; una torpe solución a todas luces, lo cual ilustra las dificultades a las que se enfrentan las compañías de tecnología cuando buscan ofrecer servicios de calidad fundamentados en aprendizaje automático.

En la Imagen 6 observamos como durante un reconocimiento automático, la máquina etiqueta por error a un hombre como mujer al comprobar que se encuentra cocinado solo.

Imagen 6. La máquina falla al etiquetar el sexo. (T. Simonite, 2017). wired.com

Tipos de sesgo de aprendizaje automático

Sesgo del algoritmo. Ocurre cuando hay un problema dentro del algoritmo que realiza los cálculos que impulsan los cálculos de aprendizaje automático.

Sesgo de la muestra. Sucede cuando hay un problema con los datos utilizados para entrenar el modelo de aprendizaje automático. En este tipo de sesgo, los datos utilizados no son lo suficientemente grandes o representativos para enseñar el sistema. Por ejemplo, el uso de datos de entrenamiento donde sólo se caracterice a mujeres profesoras, entrenará al sistema para concluir que todas las profesoras son mujeres.

Sesgo de prejuicio. En este caso, los datos utilizados para entrenar el sistema reflejan prejuicios, estereotipos y/o suposiciones sociales defectuosas existentes, lo que introduce esos mismos sesgos del mundo real en el propio aprendizaje automático. Por ejemplo, el uso de datos sobre profesionales médicos que incluyan solo enfermeras y médicos, perpetuaría un estereotipo de género en el mundo real sobre los trabajadores de la salud en el sistema informático.

Sesgo de medición. Como sugiere el nombre, este sesgo surge debido a problemas subyacentes con la precisión de los datos y cómo se midieron o evaluaron. El uso de imágenes de trabajadores felices para entrenar un sistema destinado a evaluar el entorno laboral podría estar sesgado si los trabajadores que aparecen en las imágenes supieran que se les mide la felicidad; un sistema que se está entrenando para evaluar con precisión el peso estará sesgado si los pesos contenidos en los datos de entrenamiento se redondearon constantemente.

Sesgo de exclusión. Esto sucede cuando un apartado importante de los datos que se utilizan no es considerado, algo que puede suceder si los programadores no reconocen dichos datos como importantes.

Los factores más influyentes en el desempeño de un algoritmo de Machine Learning, además del propio algoritmo y su implementación, son la calidad y cuantía de los datos utilizados. Un modelo será tan bueno como los datos de los que aprende, y esto se convierte en algo imprescindible para mantener la

integridad de las decisiones tomadas por las representaciones aprendidas por los algoritmos. Entrenar un modelo con datos sesgados en una determinada dirección puede afectar de manera determinante al desempeño de la herramienta, condenando los resultados obtenidos por el modelo. Este sesgo, diferente al sesgo estadístico o muestral, no es más que la proyección de los prejuicios (inconscientes o no) de los desarrolladores en los algoritmos o la falta de rigor en la correcta recogida de datos para el entrenamiento de los mismos.

4.1.- Análisis de un caso real "Machine Bias"

Para ver un caso real del sesgo que puede producirse en el aprendizaje automático (machine learning), voy a comentar el artículo *Machine Bias; there's software used across the country to predict future criminals. And it's biased against blacks*. de Julia Angwin, Left Larson, Surya Maatu y Lauren Kirchner, publicado en la web propublica.org en mayo de 2016.

El artículo trata sobre el sesgo que se produce en las "evaluaciones de riesgo" que se realiza a las personas detenidas por la comisión de delitos, en algunos departamentos de justicia de los Estados Unidos de América. Los datos que se introducen a la máquina se extraen de un test de 137 preguntas, algunas de ellas contestadas por los detenidos y otras extraídas de sus antecedentes penales. Tras el análisis, un software predice que posibilidad hay de que los detenidos vuelvan a delinquir.

La empresa propietaria de este software encargado de las predicciones es Northpointe, fundada en 1989 por Tim Brennan, que entonces era profesor de estadística en la Universidad de Colorado, y Dave Wells, que dirigía un programa correccional en Traverse City, Michigan. Brennan y Wells llamaron a su producto Perfiles de Manejo de Delincuentes Correccionales para Sanciones Alternativas, o COMPAS, por sus siglas en inglés. En 2011, Brenan y Wells vendieron Northpointe al conglomerado Constellation Software.

Según Brenan y Wells, su herramienta "evalúa no solo el riesgo, sino también casi dos docenas de las llamada*s necesidades criminogénicas* que se relacionan con las principales teorías de la criminalidad, incluida la *personalidad criminal,* el *aislamiento social,* el *abuso de sustancias* y la *residencia/estabilidad.* Los acusados se clasifican en riesgo bajo, medio o alto en cada categoría." Machine

Bias; there's software used across the country to predict future criminals. And it's biased against blacks. (2016) Propublica.org

Tras la realización de un amplio análisis, los redactores del artículo determinaron que hay un porcentaje de fallo a tener en cuenta, y que en general, los algoritmos suelen cometer el error de dar una puntuación de alto riesgo de reincidencia a personas afroamericanas y a la inversa con personas blancas. A modo de ejemplo, a lo largo del artículo se expone varios casos en los que personas afroamericanas, a las que el software etiquetó como de alto riesgo, no volvieron a cometer delitos en los dos años siguientes y, por el contrario, personas blancas a las que etiquetó de bajo riesgo, sí lo hicieron.

Como muestra de algunos de los errores cometidos por COMPAS, podemos citar el caso de Brisha Borden y su amiga Sade Jones.

Una tarde de 2014 cuando Brisha tenía 18 años y mientras se dirigía a la escuela a recoger a su hermanastra, ella y su amiga decidieron coger y conducir una bicicleta y un patinete que se encontraban en la calle, pero tras unos metros se dieron cuenta que pertenecían a un niño de 6 años, una mujer les llamó la atención y Brisha y su amiga las dejaron y se marcharon, a pesar de ello, un vecino que presenció los hechos las denunció a la policía y Brisha y su amiga fueron acusadas de robo y hurto menor.

De otro lado, el artículo analiza el caso de Prater, un criminal experimentado condenado en varias ocasiones por robo y tentativa de robo a mano armada, por los que cumplió cinco años de prisión. Brisha también tenía antecedentes, pero por delitos menores cuando era menor de edad.

Brisha y Prater (Imagen 7) fueron encarcelados. La sorpresa llegó cuando los algoritmos de COMPAS emitieron una puntuación que predecía cual era la probabilidad de que cada uno volviera a cometer un delito en el futuro. Brisha, que es afroamericana, fue calificada como de alto riesgo, mientras que Prater que es blanco, fue calificado de bajo riesgo.

Dos años después, podemos afirmar que el algoritmo de la máquina lo entendió al revés. Brisha Borden no volvió a ser acusada de ningún delito, mientras que

Prater se encuentra cumpliendo una condena de ocho años de prisión por entrar en un almacén y robar productos electrónicos valorados en miles de dólares.

Imagen 7. Calificaciones de riesgo de Vernon Prater y Brisha Borden. Propublia.org

Estas son las conocidas como **evaluaciones de riesgo**, y son cada vez más comunes en los tribunales de los EE.UU. Estas puntuaciones se utilizan entre otros, para determinar quién puede ser puesto en libertad, o para establecer la fianza.

Para comprender hasta qué punto estas puntuaciones pueden determinar el futuro de una persona, en el caso de Sade Jones, la amiga de Brisa, quien nunca antes había sido arrestada, fue clasificada como riesgo medio y aunque los cargos se redujeron, todavía hoy tiene problemas para encontrar trabajo, *"fui a McDonald's y a una tienda de dólar, y todos me dijeron que no debido a mis antecedentes".*

Los autores del artículo analizaron las puntuaciones de riesgo asignados de más de 7000 personas arrestadas en el condado de Broward, Florida en 2013 y 2014 y comprobaron cuántos fueron acusados de nuevos delitos durante los dos años siguientes. La puntuación resultó ser poco fiable para predecir delitos violentos.

Sólo el 20% de los delincuentes puntuados como de alto riesgo de cometer delitos violentos acabaron por cometerlos.

También detectaron disparidades raciales significativas, al pronosticar quién volvería a delinquir, el algoritmo cometió errores tanto para personas blancas como negras, pero de manera muy diferente. El algoritmo señaló de manera errónea a los acusados negros como futuros delincuentes el doble de veces que a los acusados blancos (Imagen 8). En cambio, los acusados blancos fueron etiquetados como de bajo riesgo con mayor frecuencia que los acusados negros (Imagen 9).

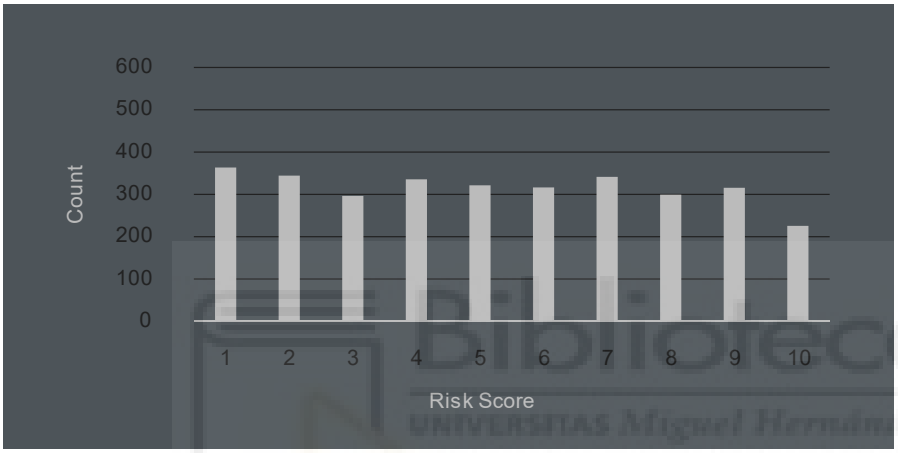


Imagen 8. Puntuaciones de riesgo de los acusados afroamericanos “análisis de datos del condado de Broward, Fla”. Propublia.org.

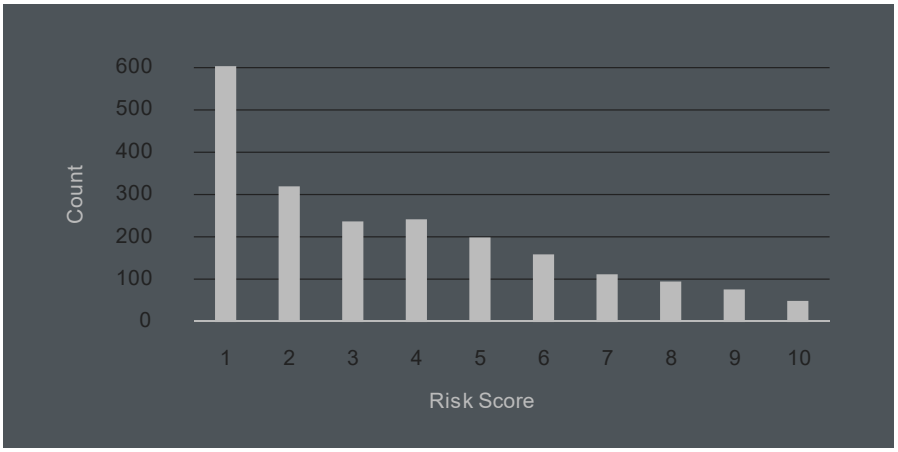


Imagen 9. Puntuaciones de riesgo de los acusados blancos “análisis de datos del condado de Broward, Fla”. Propublia.org

Como se puede ver en los gráficos anteriores, las puntuaciones de riesgo obtenidas por los acusados blancos se encuentran sesgadas hacia las categorías de menor riesgo, no así las obtenidas por los acusados negros.

Otro de los casos expuestos en el artículo analizado es el de James Rivelli de Hollywood, Florida, varón blanco, quien a sus 54 años fue arrestado por robar siete cajas de blanqueante dental en una farmacia. Rivelli tenía antecedentes penales por agresión con agravantes, robos múltiples y tráfico de drogas, aún así el algoritmo lo calificó con un 3 sobre 10 (bajo riesgo), posiblemente porque en su evaluación no se tuvo en cuenta los cinco años que pasó en una prisión en otro estado. El propio Rivelli se sorprendió de su baja puntuación "*Me sorprende que sea tan baja. Pasé cinco años en una prisión estatal en Massachusetts*". Menos de un año más tarde, volvió a ser acusado, esta vez de dos delitos graves por robar herramientas por un valor aproximado de 1000 dólares.

En un estudio de validación publicado en 2009, Brennan afirmó que es difícil construir una puntuación que no incluya elementos que puedan correlacionarse con la raza, como la pobreza, el desempleo y la marginación social. "*Si se omiten en su evaluación de riesgos, la precisión disminuye*".

El análisis de ProPublica reveló que las personas afroamericanas tienen casi el doble de probabilidades que los blancos de ser etiquetados como de mayor riesgo, pero a la hora de la verdad no reinciden. Sin embargo, en el caso de los blancos los algoritmos cometen el error opuesto, son mucho más propensos a ser etiquetados como de bajo riesgo, pero continúan cometiendo delitos.

La predicción falla de manera diferente para los acusados negros

	BLANCO	AFROAMERICANO
Etiquetado de mayor riesgo, pero no volvió a delinquir	23,5%	44,9%
Etiquetado de menor riesgo, pero volvió a delinquir	47,7%	28,0%

Imagen 10. Análisis de datos del condado de Broward, Fla.

5.- Predicciones de reincidencia en crimen. Estudio práctico en R.

Basándome en el artículo de estudio *Machine Bias,* voy a realizar un estudio práctico utilizando el lenguaje de programación R en el entorno de RStudio, para lo cual emplearé datos reales de delincuentes examinados en Florida (EE.UU.) durante los años 2013-14. Los datos se encuentran alojados en la web *rdocumentation.org*. La base de datos cuenta con un total de 5855 observaciones con información sobre las siguientes variables:

age: variable continua que contiene la edad (en años) de la persona.

juv_fel_count: variable continua que contiene el número de delitos cometidos por menores.

decile_score: variable continua, contiene el decil de la puntuación en COMPAS.

juv_misd_count: variable continua que contiene el número de faltas juveniles.

juv_other_count: variable continua que contiene el número de condenas juveniles previas que no se consideran delitos graves ni delitos menores.

v_decile_score: variable continua que contiene el decil previsto de la puntuación COMPAS.

priors_count: variable continua que contiene el número de delitos anteriores cometidos.

Sex: factor con niveles "Female" y "Male".

two_year_recid: factor con dos niveles "Yes" y "No"(si la persona ha reincidido dentro de los dos años).

race: factor que codifica la raza de la persona;

c_jail_in: variable numérica que contiene la fecha en que la persona ingresó a la cárcel (normalizada entre 0 y 1).

c_jail_out: variable numérica que contiene la fecha en que la persona salió de la cárcel (normalizada entre 0 y 1).

c_offense_date: una variable numérica que contiene la fecha en que se cometió el delito.

screening_date: variable numérica que contiene la fecha en que la persona fue tamizada (normalizada entre 0 y 1).

in_custody: variable numérica que contiene la fecha en que la persona fue internada (normalizada entre 0 y 1).

out_custody: variable numérica que contiene la fecha en que la persona fue liberada (normalizada entre 0 y 1).

Análisis de los datos.

De entrada, comenzaré realizando un breve análisis estadístico descriptivo de los datos en RStudio.

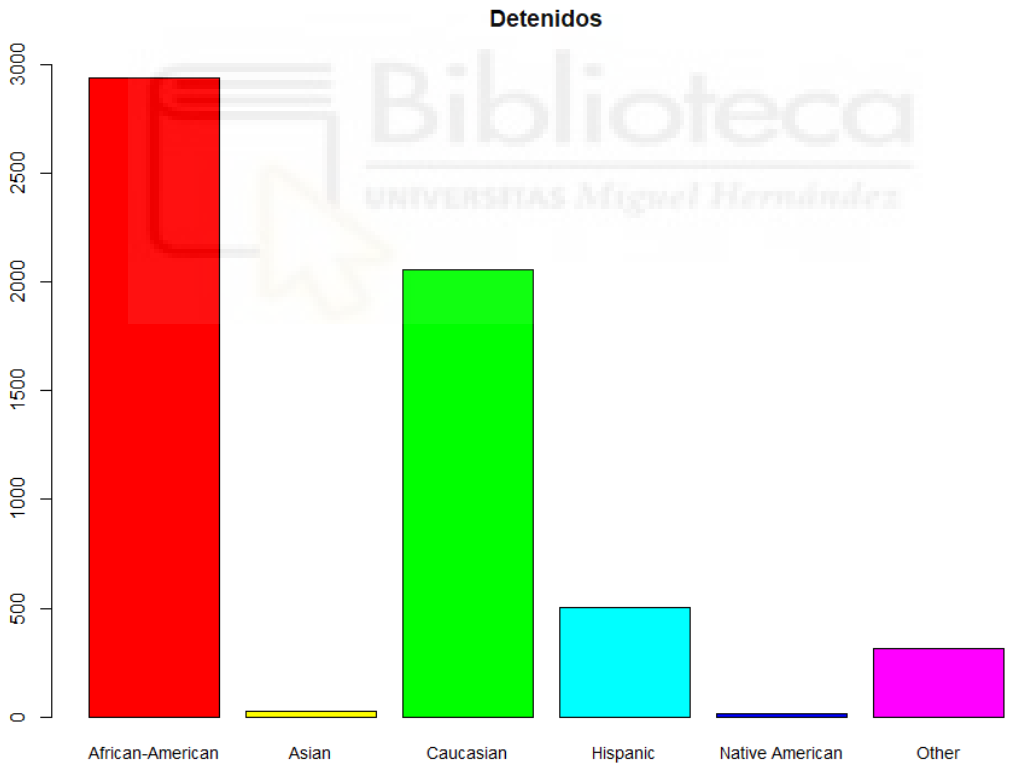


Imagen 11. Detenidos por razas en el estado de Florida entre 2013 y 2014.

En la Imagen 11 podemos observar que, del total de las 5.855 observaciones realizadas, el número de detenidos Afroamericanos es de casi 3.000 sujetos,

notablemente mayor que el del resto de razas, y de casi 1.000 más, si lo comparamos con personas de raza caucásica.

Continuando con el análisis de los datos, en las detenciones por sexo (IMEGEN 12), observamos que la mayoría de los detenidos son hombres, y considerando el gráfico de barras anterior, hombres afroamericanos.

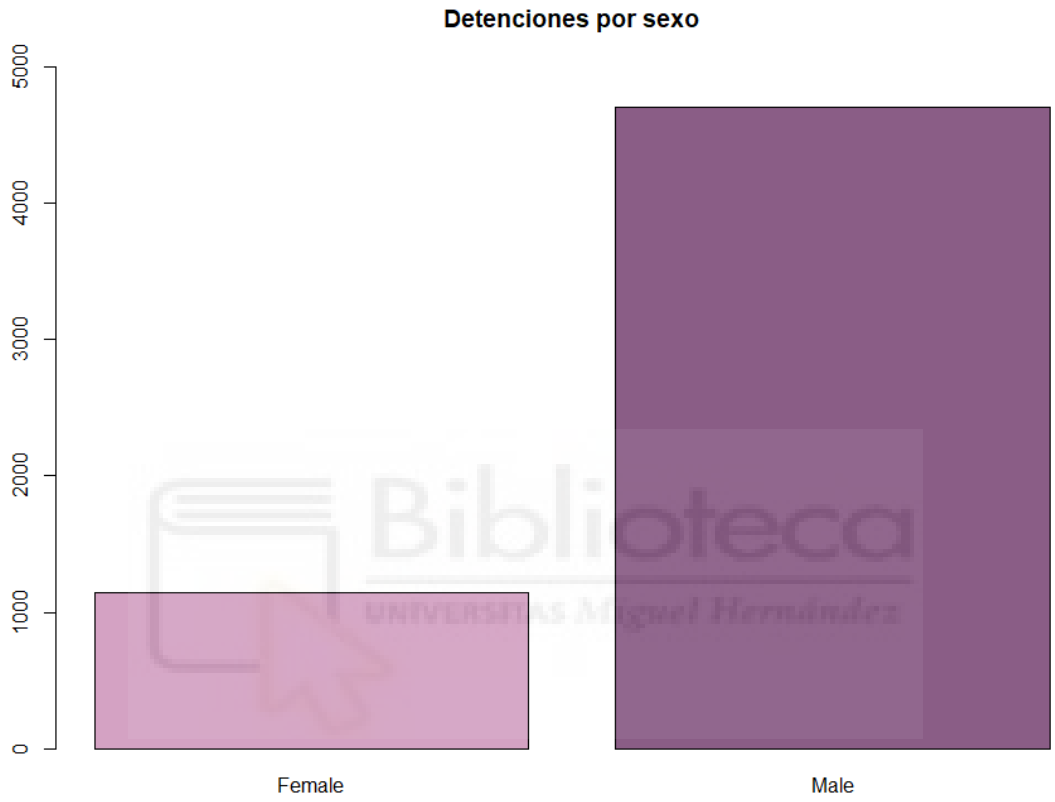


Imagen 12. Detenidos por sexo en el estado de Florida entre 2013 y 2014.

La siguiente imagen (Imagen 13), muestra que la reincidencia real a dos años en base a la raza es algo más del 50% para personas afroamericanas mientras que en general es un poco menor para el resto de razas incluida la blanca.

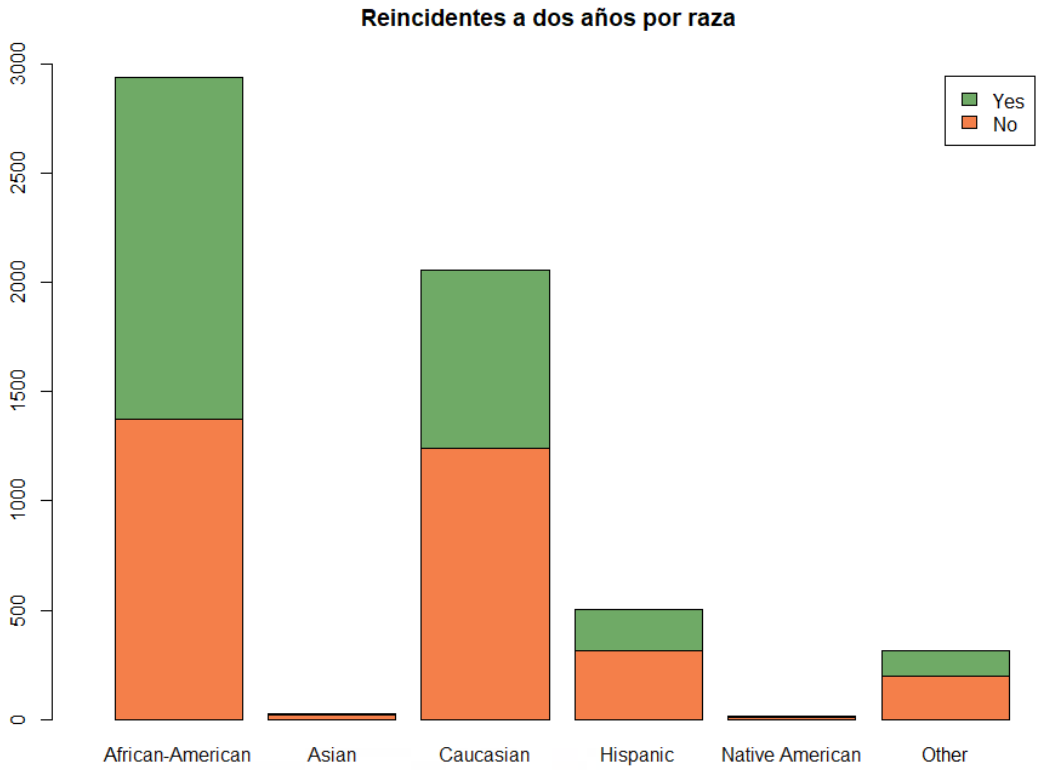
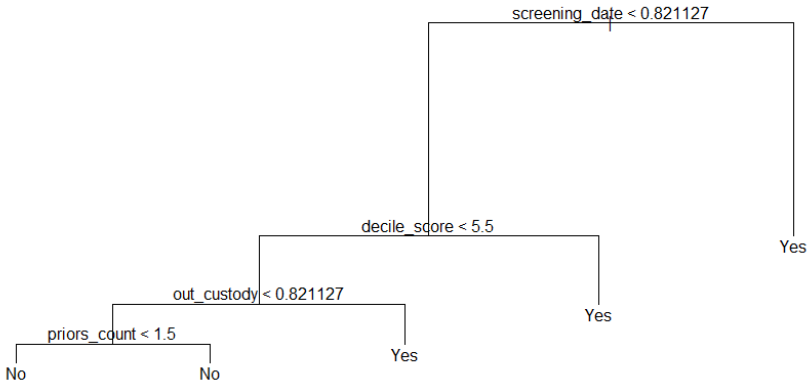


Imagen 13. Reincidencia a 2 años en base a la raza.

Estudio práctico en R sobre la reincidencia a dos años.

Par realizar el estudio utilizaremos la técnica de los árboles de clasificación explicados con anterioridad. Trataremos de predecir la reincidencia a dos años de los sujetos examinados en función de las variables que contiene la base de datos.

En primer lugar, partiremos el conjunto de datos en un conjunto de datos de entrenamiento, y conjunto de datos de prueba. De las 5.855 observaciones que contiene nuestra base de datos, escogeremos al azar 4.000 para el conjunto de datos de entrenamiento y el resto para el conjunto de datos de prueba (1855). Primero, realizamos el árbol de clasificación utilizando todas las variables del conjunto de datos de entrenamiento, es decir, trataremos de predecir la reincidencia a dos años en función del resto de variables. Obtenemos el siguiente árbol de clasificación:



Si analizamos el árbol generado, observamos que en función de la fecha de evaluación (screening_date), la puntuación obtenida en COMPAS (decil_score), la fecha en la que la persona fue liberada (out_score) y el número de delitos cometidos con anterioridad (priors_count) obtenemos la predicción correspondiente a si reincidirá o no. A continuación, añadimos al conjunto de datos de prueba la variable tree.pred con el resultado de la predicción de reincidencia a dos años y generamos la correspondiente **matriz de confusión** empleando la reincidencia real a dos años y la predicha por nuestro algoritmo.

Reincidencia a 2 años	**NO**	**SI**
NO	719	304
SI	152	680

Analizando los datos, podemos determinar que hemos predicho correctamente la NO reincidencia en 719 casos y hemos fallado en 304, y de la misma manera, hemos predicho correctamente que SI reincidirán en 680 casos y hemos fallado en 152.

A continuación, y de un modo más visual, analizamos mediante gráficos de barras los datos del conjunto de prueba (Imagen 14). Como podemos observar, los gráficos muestran que en general la predicción se aproxima mucho a la realidad para la raza caucásica, así como para el resto, pero no es tan buena al predecir la reincidencia de los afroamericanos, para quienes la reincidencia real es menor que la predicha por nuestro algoritmo.

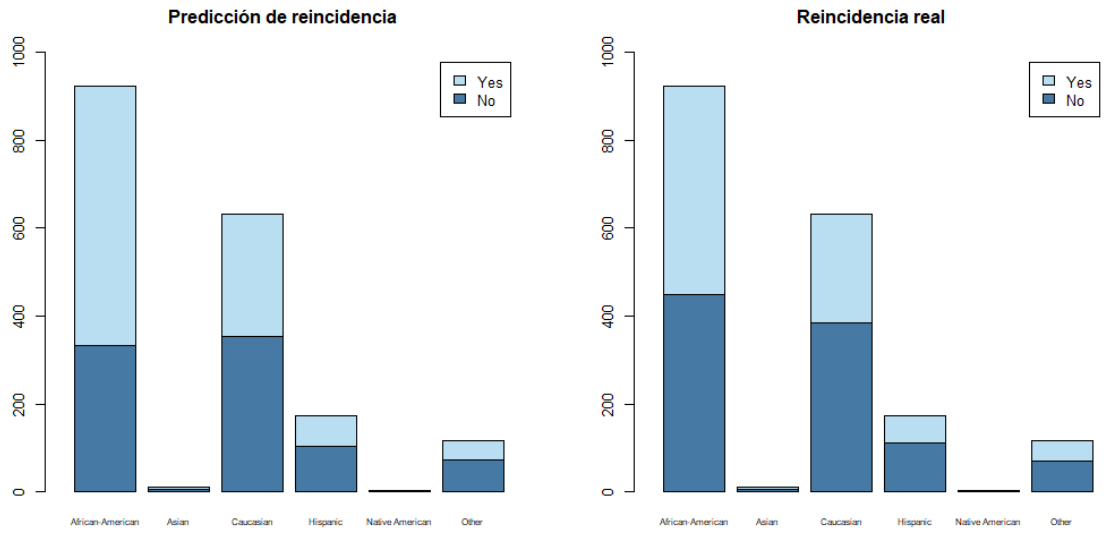
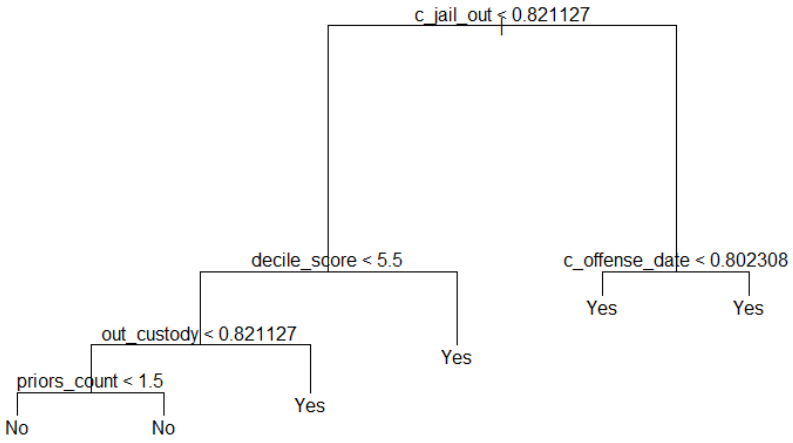


Imagen 14. Predicción de reincidencia en función de todas las variables y reincidencia real (Conjunto de datos de prueba).

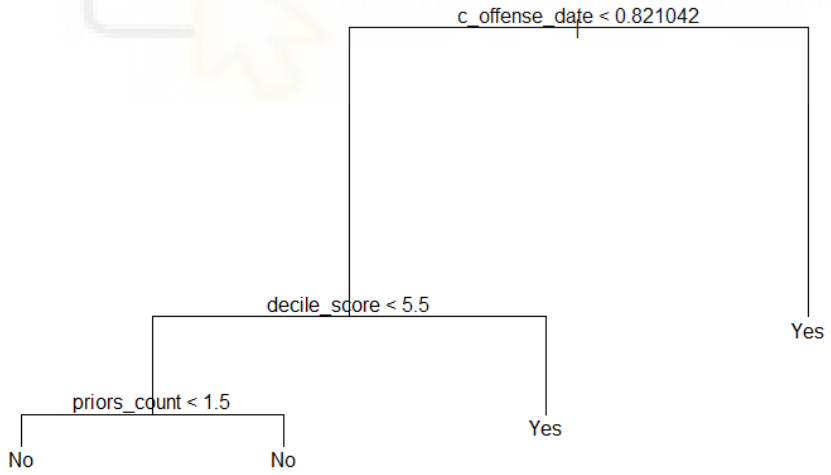
Ahora, tratamos de afinar más nuestro algoritmo eliminando algunas variables del conjunto de datos de entrenamiento. En concreto eliminamos las variables c_jail_in (fecha en la que la persona ingresa en la cárcel), screening_date (fecha en que la persona fue puntuada), y in_custody (fecha en la que la persona fue internada), puesto que, en principio, no aportan valor al algoritmo de predicción. Obtenemos el siguiente árbol de clasificación:



De nuevo, creamos la variable tree.pred en el conjunto de datos de prueba con el valor de la predicción de reincidencia a dos años. Para esclarecer mejor los resultandos obtenemos la matriz de confusión empleando los valores de reincidencia real a dos años y los predichos por nuestro algoritmo. Como podemos observar, los resultados son prácticamente iguales que el caso anterior, es decir, las variables eliminadas no aportaban precisión al algoritmo.

Reincidencia a 2 años	**NO**	**SI**
NO	719	304
SI	153	679

Continuamos afinando nuestro árbol de clasificación para tratar de aproximar nuestras predicciones a la realidad. Eliminamos ahora las variables c_jail_out (fecha en la que la persona salió de la carcel) y out_custody (fecha en que la persona fue liberada) del conjunto de datos de entrenamiento. Obtenemos el siguiente árbol de clasificación:



Al prescindir de algunas variables, el árbol generado es más sencillo. De nuevo, creamos en el conjunto de datos de prueba la variable tree.pred que contiene el valor de la predicción de reincidencia a dos años y obtenemos la matriz de confusión que nos aclarará mejor el resultado obtenido.

En este caso, aunque mejoramos un poco la predicción de NO reincidencia a 2 años, aumentamos el error al predecir que SÍ reincide:

Reincidencia a 2 años	**NO**	**SI**
NO	806	217
SI	274	558

Al volver a comparar los gráficos de barras (Imagen 15) del conjunto de datos de prueba, entre la reincidencia real y la predicha por nuestro algoritmo, observamos que ahora la predicción de la reincidencia a dos años se aproxima más a la reincidencia real para los afroamericanos y, por el contrario, se aleja un poco para el resto de razas incluida la caucásica.

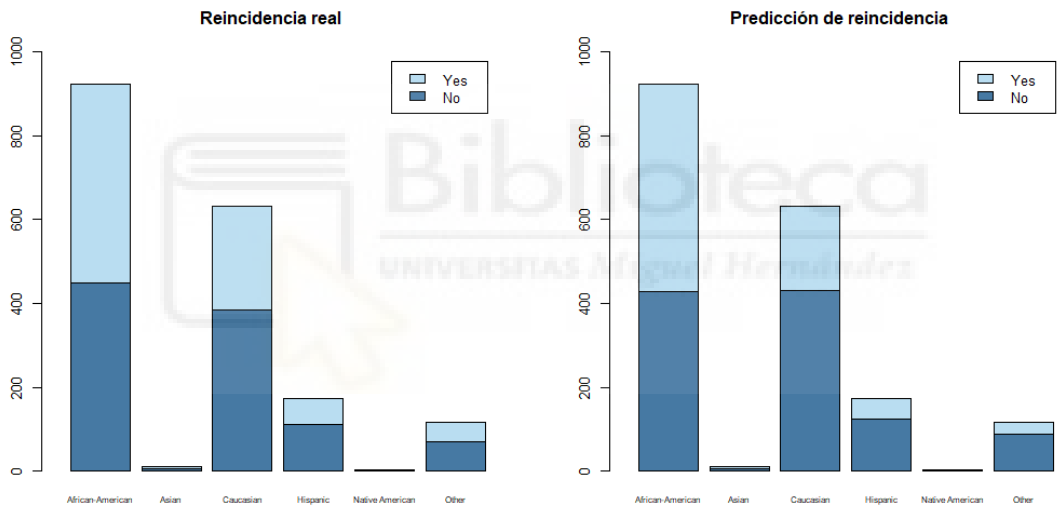


Imagen 15. Reincidencia real a dos años y Predicción de reincidencia. Conjunto de datos de prueba, eliminando las variables c_jail_in, screening_date, in_custody, c_jail_out y out_custody del conjunto de datos de entrenamieto.

6.- Conclusión

Tras analizar la base de datos de detenidos en el estado de Florida entre los años 2013 y 2014, podemos constatar que, el resultado de nuestra predicción será distinto en función de las variables empleadas para el cálculo. Esto no hace más que confirmar, que el resultado de las predicciones varía en función de numerosas variables como son, los datos recopilados, el algoritmo empleado

para la predicción, o incluso la interpretación que realicemos de los resultados. Tal y como explicábamos en el apartado número 4 de este trabajo, cuando hablábamos del sesgo en el uso del Big Data para la predicción, el sesgo de aprendizaje o sesgo de algoritmo ocurre cuando un algoritmo produce resultados que son sistemáticamente perjudiciales debido a suposiciones erróneas en el proceso de aprendizaje automático o porque existe sesgo en los datos utilizados. El aprendizaje automático dependerá de la calidad, la objetividad y el tamaño de los datos de entrenamiento utilizados. Los datos defectuosos, deficientes o incompletos darán como resultado predicciones inexactas.

Por otro lado, el sesgo también puede ser introducido por quienes diseñan y/o entrenan los sistemas de aprendizaje automático, al crear algoritmos que reflejan sesgos cognitivos no deseados o prejuicios de la vida real. No olvidemos, que las máquinas no tienen prejuicios, la Inteligencia Artificial busca la lógica para determinar si algo es verdadero o falso y, por tanto, no pueden tener sesgo por sí misma.

El sesgo procedente de los seres humanos, como es el sesgo de confirmación (cuando aceptamos el resultado en base a una creencia previa), o el sesgo de disponibilidad (cuando ponemos mayor énfasis en la información que es relevante para nosotros mismos), es el responsable de la selección de unas reglas de entrenamiento para la máquina que darán forma a la creación de un modelo de aprendizaje automático defectuosos, que en realidad no es auténtica inteligencia artificial, si no que más bien será un reflejo de sus prejuicios, ocultando sus propias observaciones incompletas o defectuosas, en el interior de una caja negra a la que llamarán máquina.

Finalmente, el artículo “Machine Bias”, comentado en este trabajo, pone de manifiesto este sesgo durante el aprendizaje de la máquina, y aunque según sus autores, para calcular el riesgo de reincidencia de un delincuente no se utilizan directamente las variables de raza y sexo de la persona examinada, si se utilizan otras variables como el lugar de residencia, la escuela donde estudiaron, y muchas otras que en su conjunto terminan por generar sesgo en el aprendizaje de la máquina.

7.- Bibliografía

Schawab, K. (2016) La Cuarta Revolución Industrial.

Real Academia Española. Big Data. En Diccionario panhispánico del español jurídico. Recuperado el 1 de junio de 2022 de https://dpej.rae.es/lema/big-data

Gantz, J., & Reinsel, D. (2012). THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. https://datastorageasean.com/sites/default/files/idc-the-digital-universe-in-2020.pdf

Manyika, J. & Chui, M. & Brown, B. & Bughin, J. & Dobbs, R. & Rowburgh, C. & Hung Byers, A. (2011) Big data: The next frontier for innovation, competition, and productivity [Big data: La próxima frontera para la innovación, competitividad y productividad] Mckinsey Global Insitute. https://www.mckinsey.com/~/media/mckinsey/business%20functions/mckinsey%20digital/our%20insights/big%20data%20the%20next%20frontier%20for%20innovation/mgi_big_data_exec_summary.pdf

Laney, D (2012). Big Data Means Big Business, Gartner, Inc. http://media.ft.com/cms/4b9c7960-2ba1-11e3-bfe2-00144feab7de.pdf

Armenta, A. (2021). Data Lake vs. Big Data for Industrial Applications. https://control.com/technical-articles/data-lake-vs-big-data/

Baha, A., & Madisetti, V. (2019). Big Data Analytics: A Hands-On Approach.

Joyanes Aguilar, L. (2013) Big Data, Análisis de grandes volúmenes de datos en organizaciones.

IT Solutions de BETWEEN (2020). El desafío del almacenamiento de datos en tiempos del big data. https://impulsate.between.tech/almacenamiento-datos-big-data

Proyectos y Desarrollos ISARQ C.A. ¿De qué hablamos cuando hablamos de Big Data? (2018) https://isarq.com/index.php/2018/08/15/de-que-hablamos-big-data/.

Carrillo J. A & Marco de Lucas J. E. & Dueñas J. C. & Cases F. & Cristino F. & Gonzáles G. & Pereda L. F. Big Data en los Entornos de Defensa y Seguridad. Grupo de Trabajo sobre Big Data, comisión de Investigación de Nuevas Tecnologías del Centro Superior de Estudios de la Defensa nacional (CESEDEN). Instituto Español de Estudios Estratégicos (IEEE).

Real Academia Española. Big Data. En Diccionario panhispánico del español jurídico. Recuperado el 1 de junio de 2022 de https://dpej.rae.es/lema/big-data

Wikipedia. La enciclopedia libre. Person of Interest. Recuperado el 1 de junio de 2022 de https://es.wikipedia.org/wiki/Person_of_Interest_(serie_de_televisi%C3%B3n)

Las diferencias entre Inteligencia Artificial, Machine Learning y Deep Learning. Qubole.com

Sas.com/es ¿Qué es el Deep Learning? https://www.sas.com/es_es/insights/analytics/deep-learning.html

Simonite, T. (2017). Machines Learn a Biased View of Women. Wired. https://www.wired.com/story/machines-taught-by-photos-learn-a-sexist-view-of-women/

Angwin, J. & Larson, L. & Maatu, A. & Kirchner, L. (2016) Mahine Bias; There's software used across the country to predict future criminals. And it's biased against blacks. [Sesgo de Máquina; Hay software utilizado en todo el país para predecir futuros delincuentes. Y está sesgado contra los negros.] Propublica.org

Base de datos de delincuentes examinados en Florida (EE.UU.) https://www.rdocumentation.org/packages/fairml/versions/0.6.3/topics/compas