Machine Translated by Google

Análisis de datos de lesiones por accidentes de tráfico

Mohamed K. Nour1
Facultad de Informática y Sistemas de información
Universidad Umm Al-Qura

Atif Naseer2
Unidad de Ciencia y Tecnología
Universidad Umm Al-Qura

Basem Alkazemi3
Facultad de Informática y Sistemas de información
Universidad Umm Al-Qura

Muhammad Abid Jamil4
Facultad de Informática y Sistemas de información
Universidad Umm Al-Qura

Resumen: Los investigadores en seguridad vial que trabajan con datos de accidentes de tráfico han logrado resultados satisfactorios en el análisis de accidentes de tráfico mediante la aplicación de técnicas de análisis de datos. Sin embargo, se ha avanzado poco en la predicción de lesiones en carretera. Este artículo aplica métodos avanzados de análisis de datos para predecir la gravedad de las lesiones y evalúa su rendimiento. El estudio utiliza técnicas de modelado predictivo para identificar los riesgos y los factores clave que contribuyen a la gravedad de los accidentes. El estudio utiliza datos públicos del Departamento de Transporte del Reino Unido, que abarcan el período comprendido entre 2005 y 2019. El artículo presenta un enfoque lo suficientemente general como para aplicarse a diferentes conjuntos de datos de otros países. Los resultados indican que las técnicas basadas en árboles, como XGBoost, superan a las basadas en regresión, como las redes neuronales artificiales (RNA). Además del artículo, se identifican relaciones interesantes y se reconocen problemas relacionados con la calidad de los datos.

Palabras clave: Análisis de accidentes de tráfico (RTA); minería de datos; aprendizaje automático; XGBOOST

I. INTRODUCCIÓN

Un accidente de tránsito es un evento inesperado que ocurre de manera involuntaria en la carretera y que involucra un vehículo u otros usuarios de la vía y causa víctimas o pérdidas de propiedad.
Más del 90% de las muertes en las carreteras del mundo ocurren en países de ingresos bajos y medios, que representan solo el 48% de los vehículos registrados a nivel mundial [1]. La pérdida financiera, que asciende a aproximadamente US$518 mil millones, supera la asistencia para el desarrollo asignada a estos países. Si bien los países desarrollados y ricos mantienen tasas de mortalidad por accidentes de tránsito estables o en descenso gracias a esfuerzos coordinados de corrección de diversos sectores, los países en desarrollo aún pierden entre el 1% y el 3% de su producto nacional bruto (PNB) debido a la endémica cantidad de víctimas de accidentes de tránsito. La Organización Mundial de la Salud (OMS) teme que, a menos que se tomen medidas inmediatas, los accidentes de tránsito se conviertan en la quinta causa principal de muerte para 2030, lo que resultará en un estimado de 2,4 millones de muertes al año [1].

Por lo tanto, las medidas para reducir los accidentes basadas en una comprensión profunda de las causas subyacentes son de gran interés para los investigadores. El siglo XXI ha presenciado un rápido crecimiento de la motorización vial debido al rápido aumento de la población, la urbanización masiva y el aumento de la movilidad de la sociedad moderna; los riesgos de muerte por accidente de tráfico (RTF) también pueden aumentar y los RTA también pueden asumirse como una "epidemia moderna". Este documento presenta un marco analítico para predecir la gravedad de los accidentes de tráfico [1]. Las investigaciones anteriores sobre el análisis de accidentes de tráfico se habían basado principalmente en métodos estadísticos como la regresión lineal y de Poisson. Este documento presenta un marco analítico para predecir la gravedad de los accidentes de tráfico. En particular, el documento aborda cuestiones relacionadas con el preprocesamiento y la preparación de datos, como la agregación de datos, la transformación, la ingeniería de características y los datos desequilibrados. En Además, el artículo busca aplicar modelos de aprendizaje automático para permitir predicciones más precisas. Por lo tanto, compara el rendimiento de varios algoritmos de aprendizaje automático en la predicción de la gravedad de las lesiones por accidente. En particular, el artículo aplica regresión logística, máquinas de vectores de soporte, árboles de decisión, bosques aleatorios XGBoost y modelos de redes neuronales artificiales. El resto de este artículo está organizado de la siguiente manera: la Sección II presenta algunos trabajos previos. La Sección III muestra la metodología utilizada en este trabajo, la Sección IV describe la gestión de datos y los patrones de los datos de accidentes de tráfico. La Sección V muestra los resultados y el análisis de todos los enfoques utilizados en este trabajo. La Sección VI presenta las conclusiones y trabajos futuros.

II. REVISIÓN DE LITERATURA

Mehdizadeh et al. [2] presentaron una revisión exhaustiva de los métodos de análisis de datos en seguridad vial. Los modelos analíticos se pueden agrupar en dos categorías: modelos predictivos o explicativos que buscan comprender y cuantificar el riesgo de colisión, y (b) técnicas de optimización que se centran en minimizar el riesgo de colisión mediante la selección de rutas y la programación de descansos. Su trabajo presentó fuentes de datos públicas y técnicas de análisis descriptivo (resumen de datos, visualización y reducción de dimensiones) que pueden utilizarse para lograr rutas más seguras y proporcionar código que facilite la recopilación y exploración de datos por parte de profesionales e investigadores. El artículo también revisó los modelos estadísticos y de aprendizaje automático utilizados para la modelización del riesgo de colisión.
Hu et al. [3] categorizaron los modelos analíticos prescriptivos y de optimización que se enfocan en minimizar el riesgo de accidentes. Ziakopou-los et al. [4] revisaron críticamente la literatura existente sobre diferentes enfoques espaciales que incluyen la dimensión del espacio en sus diversos aspectos en sus análisis para la seguridad vial. Moosavi et al. [5] identificaron debilidades en la investigación de accidentes de tránsito que incluyen: conjuntos de datos a pequeña escala, dependencia de un conjunto extenso de datos y no ser aplicable para fines en tiempo real. El trabajo propuso una técnica de recopilación de datos con un modelo de red neuronal profunda llamado Deep Accident Predic-tion (DAP); Los resultados mostraron mejoras significativas para predecir eventos de accidentes poco frecuentes. Zagorodnikh et al. [6] desarrollaron un sistema de información que muestra la concentración de accidentes en un mapa electrónico del terreno de modo automático para la RTA rusa para ayudar a simplificar el análisis de la RTA.

Kononen et al. [7] analizaron la gravedad de los accidentes ocurridos en Estados Unidos utilizando un modelo de regresión logística.
Reportaron un rendimiento del 40% y el 98% en sensibilidad y especificidad, respectivamente. Además, identificaron que los predictores más importantes del nivel de lesión son: cambio en la velocidad, uso del cinturón de seguridad y dirección del impacto.

Machine Translated by Google

Las redes neuronales artificiales (RNA) son una de las herramientas de minería de datos y técnicas no paramétricas con las que los investigadores han analizado la gravedad de los accidentes y las lesiones entre los involucrados en dichos accidentes. Delen et al. [8] aplicaron una RNA para modelar las relaciones entre los niveles de gravedad de las lesiones y los factores relacionados con los accidentes. Utilizaron datos de accidentes de EE. UU. con 16 atributos. El trabajo identificó cuatro factores que influyen.
el nivel de lesión: cinturón de seguridad, uso de alcohol o drogas, edad, sexo y vehículo.

Naseer et al. [9] presentan un método de análisis de accidentes de tráfico basado en aprendizaje profundo. Destacaron técnicas de aprendizaje profundo para construir modelos de predicción y clasificación a partir de datos de accidentes de tráfico.

Sharma et al. [10] aplicaron máquinas de vectores de soporte con diferentes funciones de kernel gaussianas para accidentes con el fin de extraer características importantes relacionadas con la ocurrencia de accidentes. El artículo comparó las redes neuronales con las máquinas de vectores de soporte. El artículo indicó que las máquinas de vectores de soporte (SVM) son superiores en precisión. Sin embargo, el método de las SVM presenta las mismas desventajas que las RNA en la predicción de la gravedad de los accidentes de tráfico, como se mencionó anteriormente.

Meng et al. [11] utilizaron XGBoost para predecir accidentes utilizando datos de accidentes de tráfico de múltiples fuentes. Utilizaron datos históricos, junto con datos meteorológicos y de tráfico. Schlogl et al. [12] realizaron múltiples experimentos para demostrar que XGBoost ofrece un mejor rendimiento que varios algoritmos de aprendizaje automático.

Ma et al. [13] propusieron el marco basado en XGBoost, que analiza la relación entre la colisión, el tiempo, los factores ambientales y espaciales y la tasa de mortalidad. Los resultados muestran que el método propuesto ofrece el mejor rendimiento de modelado en comparación con otros algoritmos de aprendizaje automático. El artículo identificó ocho factores que influyen en la mortalidad por accidentes de tráfico.

Cuenca et al. [14] compararon el rendimiento de Naive Bayes, Deep Learning y Gradient Boosting para predecir la gravedad de las lesiones en accidentes de tráfico en España. Su trabajo reveló que Deep Learning supera a otros métodos.

III. METODOLOGÍA

La metodología adoptada en este trabajo se muestra en la Figura 1.
El primer paso del proceso de análisis de datos es la recopilación de datos, considerada fundamental para el éxito de un proyecto de análisis de datos. Existen numerosas fuentes de datos, como sensores, datos visuales a través de cámaras, IoT y dispositivos móviles, que capturan datos en diferentes formatos y deben almacenarse en tiempo real o sin conexión. Además, se recopilan datos de diferentes autoridades relacionados con el volumen de tráfico, detalles de accidentes e información demográfica. El almacenamiento puede realizarse en servidores locales o en la nube. La clave de la pirámide de gestión de datos es el preprocesamiento. Los datos adquiridos de las ubicaciones de almacenamiento no se pueden utilizar tal como están. Requieren preprocesamiento antes de realizar cualquier análisis. Los datos adquiridos pueden incluir información faltante que debe rectificarse, así como mucha información que debe eliminarse debido a la duplicación. El preprocesamiento puede implicar la transformación de datos, ya que facilita la normalización, la selección de atributos, la discretización y la generación de jerarquías. La reducción de datos puede ser necesaria a gran escala, ya que el análisis de una gran cantidad de datos es más difícil. La reducción de datos aumenta la eficiencia del almacenamiento y...

El costo del análisis. El análisis de datos utiliza múltiples algoritmos de aprendizaje automático para obtener información valiosa. El análisis de datos es crucial para cualquier organización, ya que proporciona información detallada sobre los datos y es útil para la toma de decisiones y la realización de predicciones sobre el negocio. Los datos pueden presentarse de diversas formas según el tipo de dato utilizado. Pueden mostrarse en tablas, gráficos o diagramas organizados. La presentación de datos es fundamental para los usuarios empresariales, ya que proporciona los resultados del análisis en un formato visual.

Una de las tareas más importantes del análisis y modelado de riesgos viales es predecir la gravedad de los accidentes. Este artículo analiza la construcción de un modelo predictivo para la gravedad de los accidentes e investiga el proceso de construcción de un modelo de clasificación para predecir la gravedad de los accidentes, en particular el estudio:

- Se presenta el marco de gestión de datos. A continuación, se explica cómo se prepararon los datos antes del modelado, lo que incluye el preprocesamiento y la depuración de datos. Esta sección se presenta en la sección "Marco de Gestión de Datos". • Se identifican las deficiencias en las técnicas de modelado

predictivo de las ATR. Esta sección ofrece una breve introducción a cada técnica utilizada en este documento y presenta trabajos previos en la predicción de accidentes de tráfico, junto con los requisitos de datos y los resultados de rendimiento registrados.

Este tema se presenta en la sección Análisis de Datos • Creación de

modelos de predicción. Esta sección comienza con la descripción de las métricas de rendimiento y, a continuación, presenta los datos utilizados. A continuación, se comparan los clasificadores, en particular: regresión logística, máquinas de vectores de soporte, redes neuronales, árboles de decisión, bosques aleatorios y el árbol de refuerzo de gradiente extremo (XGboost). En esta sección, se investiga la distribución desequilibrada de clases para determinar su impacto en la gravedad de las lesiones durante los accidentes. Esto se presenta en la sección de resultados.

IV. GESTIÓN DE DATOS

A. Recopilación de datos

Los datos comprenden datos disponibles públicamente del gobierno del Reino Unido que abarcan el período de 2005 a 2019 [15]. Aunque el departamento de transporte del Reino Unido proporciona datos desde 1979, se informó que los datos recopilados a partir de 2005 son más precisos y contienen menos datos faltantes. Los registros muestran información sobre colisiones de tráfico que implican lesiones personales ocurridas en vías públicas que han sido reportadas a la policía. Los datos son recopilados por las autoridades en la escena de un accidente o, en algunos casos, reportados por un miembro del público en una estación de policía, luego procesados y transmitidos a las autoridades. Los datos incluyen 2 millones de colisiones únicas, con coordenadas espaciales x, y disponibles. Los datos relacionados con el flujo de tráfico y la información sobre todas las carreteras de la red del Reino Unido y las autoridades locales también están disponibles por separado. El conjunto de datos contiene una sola entrada para cada accidente con 33 atributos (características).

Los atributos se pueden agrupar por geografía, centrados en accidentes, clima y tiempo. Los datos relacionados con los vehículos implicados en los accidentes se almacenan en un archivo independiente con 16 atributos. Los datos sobre las víctimas se almacenan en un archivo de 23 campos. La relación entre estos tres archivos es de uno a muchos; es decir, una fila de accidente puede contener varias filas de víctimas y varias filas de vehículos, con el índice de accidente como campo de enlace.

Machine Translated by Google

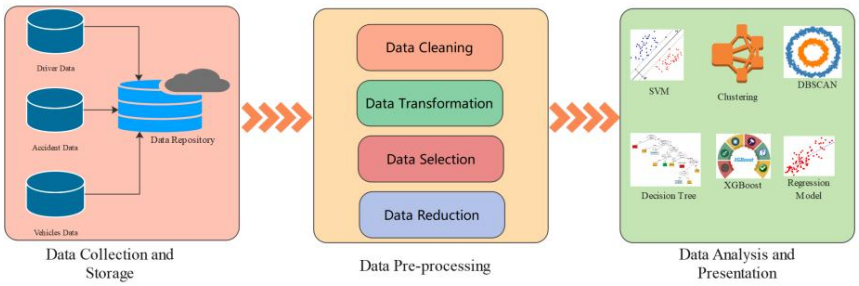

Figura 1. Metodología propuesta

La gravedad de cualquier accidente es el factor más importante para analizar el patrón de lesiones. Según el WSDOT [16], el nivel de gravedad en los accidentes se mide mediante la escala KABCO, que utiliza los parámetros: mortal (K), lesión incapacitante (A), lesión no incapacitante (B), lesión leve (C) y solo daños materiales (PDO u O).

En este trabajo, dividimos los accidentes en tres categorías, es decir, fatales (donde la muerte ocurre dentro de los 30 días del accidente), lesiones graves (donde la persona requiere tratamiento hospitalario) y lesiones leves (donde la persona no requiere ningún tratamiento médico).

En este proyecto, los datos se almacenan en una base de datos relacional. Sin embargo, en trabajos futuros, los datos relacionales se desnormalizarán primero y luego se transformarán en registros clave-valor de Hadoop.

B. Preprocesamiento de datos

Se aplicaron diferentes métodos de preprocesamiento y limpieza de datos. El preprocesamiento de datos implica diversas tareas y técnicas, como la gestión de valores faltantes, la detección de valores atípicos o anomalías y la selección de características. Una etapa clave del análisis de datos es la selección de los datos. Estos deben ser de buena calidad y estar limpios.

Las consideraciones sobre la calidad de los datos incluyen la precisión, la integridad y la consistencia [17]. Además, el volumen de datos también es importante. Los datos deben ser lo suficientemente grandes como para ser útiles en el modelado predictivo. Deben dividirse en subconjuntos de entrenamiento, prueba y validación para evaluar el modelo. Se aplicaron los siguientes pasos de preprocesamiento de datos para prepararlos para el análisis y los algoritmos de aprendizaje automático:

La mayoría de los métodos de aprendizaje automático requieren que los datos estén en formato binario o numérico. Sin embargo, en la vida real, las fuentes de datos incluyen atributos de categoría, como el tipo de vía y la categoría de siniestros. Todos los atributos de categoría se convierten a valores numéricos.

Los valores numéricos difieren en rangos. Para evitar sesgos hacia valores numéricos grandes, todos los valores numéricos se normalizarán a valores entre 0 y 1.

- Se eliminarán los registros con valores faltantes.
- El campo de fecha creará varios campos relacionados como mes, año y semana.
- Determinar atributos de menor calidad. Se eliminarán los atributos con más del 70 % de valores faltantes.
- Calcule la correlación entre el nivel de gravedad y todos los demás atributos. Se eliminarán tanto los atributos con valores de correlación altos como los de correlación muy bajos.
- Cree campos relacionados con el este y el norte para crear zonas de accidentes en lugar de ubicaciones específicas. El valor umbral para el nivel de zona es de 1 km² .

Uno de los problemas que enfrenta la creación de modelos analíticos para la gravedad de los accidentes es el desequilibrio de datos [18], donde la ocurrencia de un evento mortal es poco frecuente o rara en comparación con los accidentes sin lesiones o con lesiones leves. Debido al desequilibrio extremo de los datos de accidentes, la mayoría de los algoritmos no generan buenos modelos predictivos y, con un rendimiento deficiente, probablemente clasificarán erróneamente los accidentes mortales, ya que no son frecuentes en el conjunto de datos [19]. Para conjuntos de datos desequilibrados, como los de accidentes de tráfico, las técnicas de muestreo pueden ayudar a mejorar la precisión del clasificador [19]. A continuación se analizarán dos técnicas de muestreo: el submuestreo y el sobremuestreo.

El submuestreo se utiliza para ajustar la distribución de clases de un conjunto de datos a favor de la clase minoritaria. Con el submuestreo, la clase mayoritaria se reduce o se submuestrea [17] y se eliminan aleatoriamente datos de la clase mayoritaria hasta que ambas clases coincidan. El sobremuestreo es una técnica utilizada en minería de datos para ajustar la distribución de clases de un conjunto de datos a favor de la clase mayoritaria [18]. Por otro lado, el sobremuestreo aumenta o se sobremuestrea la clase minoritaria hasta que su tamaño se iguala al de la clase mayoritaria. Sin embargo, estas técnicas requieren habilidades especializadas y puede llevar un tiempo considerable identificar la mejor muestra.

En este estudio, se aplica la tarea de selección de características al conjunto de datos. El conjunto de datos se preprocesa según lo especificado en las Tablas I, II y III. La Tabla I muestra las características relativas a los accidentes, la Tabla II muestra todas las características de los vehículos y la Tabla III destaca las características relativas a las víctimas y su tipo. Las tablas también muestran el preprocesamiento de las características, de modo que algunas se excluyen de la lista y otras se ajustan con la escala.

Machine Translated by Google

TABLA I. CARACTERÍSTICAS DE LOS ACCIDENTES

Nombre de la variable	Tipo	Preprocesamiento
Índice de accidentes	Campo de enlace	EXCLUIR (único en accidentes)
Cuerpo de policía	Número del 1 al 98	0 Londres 1 de lo contrario
Gravedad del accidente	1 Fatal 2 Serio 3 Ligero	EXCLUIR
Número de vehículos	ESCALA numérica	
Número de víctimas	ESCALA numérica	
Fecha (DD/MM/AAAA)	FECHA	Dividido en polilla, Año y Semana, fin de semana Día de la semana
Día de la semana	ESCALA DEL 1 AL 7	
Hora (HH:MM)	TIEMPO	Dividido en horas punta y horas no pico
Ubicación Este OSGR (Nulo si no se conoce)	Numérico	Eliminar las dos últimas dists y escala
Ubicación Norte OSGR (Nulo si no se sabe)	Numérico	Eliminar las dos últimas dists y escala
Longitud (Nulo si no se conoce)	INCLUIR numérico	
Latitud (Nulo si no se conoce)	INCLUIR numérico	
Autoridad local (Distrito)	1 a 941	excluir
Autoridad local (Autoridad de Carreteras - Código ONS)	208 artículos	excluir
1.ª clase de carretera	1 a 6	0 autopista, 1 de lo contrario
1er número de carretera	Exclusión numérica	
Tipo de carretera	Autopista 1 a 12 1	
Límite de velocidad	ESCALA numérica	
Detalle de la unión	0 A 9	0 sin unión, 1 de lo contrario
Control de uniones	0 A 4	0 sin control de unión, 1 de lo contrario
2da clase de carretera	0 A 6	0 autopista, 1 de lo contrario
2º Número de Carretera	Numérico	excluir
Paso de peatones- Control humano	0 A 2	0 no cruzar peatones, 1 de lo contrario
Paso de peatones- Instalaciones físicas	0 A 8	0 no cruzar peatones, 1 de lo contrario
Condiciones de luz	1 A 7	0 luz del día 1 en caso contrario
Condiciones meteorológicas	1 a 9	0 buenas condiciones, 1 de lo contrario
Superficie de la carretera Condiciones	1 a 7	0 seco, 1 en caso contrario
Condiciones especiales en el sitio	0 a 7	0 sin condiciones especiales, 1 de lo contrario
Peligros en la calzada	0 a 7 1	0 sin peligro, 1 en caso contrario
Área urbana o rural	a 3	0 urbano, 1 en otro caso
¿Lo hizo un oficial de policía? Asistir a la escena de Accidente	1 A 3	0 asisten, 1 de lo contrario

TABLA II. CARACTERÍSTICAS DEL VEHÍCULO

Nombre de la variable	Tipo	Preprocesamiento
Índice de accidentes	Campo de enlace	EXCLUIR (accidente único, uno o más vehículos)
Referencia del vehículo	Campo de enlace	EXCLUIR (vehículo único y una o más víctimas)
Tipo de vehículo	1 A 113 0 coche 1 en caso contrario	
Remolque y articulación	1 A 5	0 sin remolque, 1 en caso contrario
Maniobra del vehículo	1 A 18	0 revirtiendo,1 en caso contrario
Ubicación del vehículo- Carril restringido	0 a 10	0 en la vía principal c, 1 de lo contrario
Ubicación del cruce	0 A 8	0 sin unión, 1 en caso contrario
Derrape y Vuelco	0 A 5	0 sin derrape, 1 en caso contrario
Golpear objeto en Calzada	0 A 12	0 sin objeto, 1 en caso contrario
Salida del vehículo Calzada	0 A 8	0 no se va, 1 de lo contrario
Golpear objeto Primer punto	0 A 11	0 no objeto fuera de c camino, 1 de lo contrario
de impacto de la calzada	0 A 4	0 sin impacto, 1 en caso contrario
¿Se dejó el vehículo? Conducción manual	1 A 2	0 mano derecha, 1 en caso contrario
Propósito del viaje del conductor	1 A 6	0 trabajo, 1 de lo contrario
Sexo del conductor	1 A 3	0 hombres, 1 en otro caso
Banda de edad del conductor	1 a 11	
Capacidad del motor	Numérico	
Propulsión de vehículos Código	1 a 10	0 Gasolina, 1 en caso contrario
Edad del vehículo (fabricar)	Numérico	
Decil IMD del controlador	0 a 10	0 privados 1 de lo contrario
Conductor a casa Tipo de área	1 A 3	0 privados 1 de lo contrario

Machine Translated by Google

TABLA III. CARACTERÍSTICAS DE LAS VÍCTIMAS

Nombre de la variable	Tipo	Preprocesamiento
Índice de accidentes	Campo de enlace	EXCLUIR (accidente único, una o más víctimas)
Referencia del vehículo	Campo de enlace	EXCLUIR (único en la tabla de bajas)
Referencia de accidentes	Campo de enlace	EXCLUIR (único en el vehículo y uno o más en caso de accidente)
Clase de siniestro	1 A 3 0 conductor, 1 en caso contrario	
Sexo de la víctima	1 a 2 0 hombres, 1 en caso contrario	
Franja de edad de la víctima	ESCALA numérica	
Gravedad de la víctima	1 A 3	VALOR OBJETIVO (0 fatal 1 en caso contrario)
Ubicación peatonal	1 A 10 0 cruce,1 en caso contrario	
Movimiento de peatones	1 A 9 0 cruce,1 en caso contrario	
Pasajero de coche	0 A 2 0 pasajeros, 1 en caso contrario	
Autobús o autocar Pasajero	0 A 4	0 pasajeros de autobús o autocar, 1 de lo contrario
Camino peatonal Mantenimiento Trabajador (Desde 2011)	0 A 2	0 trabajador de la carretera, 1 de lo contrario
Tipo de siniestro	0 A 113 0 peatón 1 en caso contrario	
Decil IMD de accidentes	0 A 10 0 privados 1 en caso contrario	
Hogar de Urgencias Tipo de área	1 A 3 0 urbano 1 en caso contrario	

C. Análisis de datos

Los métodos para la predicción de accidentes de tráfico pueden ser muy diversos: clasificados en tres categorías, a saber, modelos estadísticos, enfoques de aprendizaje automático y análisis, y modelos basados en simulación. métodos [17]. En esta investigación nos centraremos en la máquina enfoques de aprendizaje.

El aprendizaje automático es un concepto amplio, que incluye la supervisión Aprendizaje y técnicas no supervisadas. Aprendizaje supervisado Las técnicas incluyen: redes neuronales artificiales y sus variantes (aprendizaje profundo, mapas autoorganizados), máquinas de vectores de soporte (SVM), árboles de decisión e inferencia bayesiana.
El aprendizaje incluye: reglas de asociación y técnicas de agrupamiento.

El aprendizaje no supervisado implica la búsqueda de patrones o agrupaciones previamente desconocidos. Estas técnicas suelen funcionar. Sin una variable objetivo previa. Reglas de agrupamiento y asociación. entran en este grupo de técnicas. El aprendizaje supervisado, en el Por otro lado, implica clasificación, predicción y estimación. técnicas que contienen una variable objetivo. La clasificación es una técnica de aprendizaje automático que asigna una clase a una instancia, es decir, asignar automáticamente un accidente de tráfico a un predefinido Clase de gravedad. La predicción es similar a la clasificación, pero implican asignar un valor continuo a una instancia.

Los métodos de aprendizaje supervisado suelen utilizar dos conjuntos: un entrenamiento y conjunto de prueba. Los datos de entrenamiento se utilizan para aprender el modelo. y requiere un grupo primario de accidentes de tráfico etiquetados. Prueba El conjunto se utiliza para medir la eficiencia del modelo aprendido e incluye casos de accidentes de tráfico etiquetados, que no participar en el aprendizaje de clasificadores.

Este artículo se centra en la aplicación de métodos de clasificación a Clasificar la gravedad de los accidentes. Se aplicarán cinco técnicas. y comparó: (1) modelos de regresión logística, (2) modelos neuronales profundos redes, (3) máquinas de vectores de soporte, (4) árboles de decisión, (5) aumento de gradiente extremo

1) Regresión logística: Los modelos de regresión se han vuelto un componente integral de cualquier análisis de datos relacionado con la relación entre una variable de respuesta y una o Más variables explicativas. La regresión logística es un método de máxima verosimilitud que se ha utilizado en cientos de estudios. del resultado del accidente. Tradicionalmente, los modelos de regresión estadística Se desarrollan en estudios de seguridad vial para asociar los accidentes frecuencia con las variables más significativas. La logística La regresión es un caso especial del modelo lineal generalizado (GLM), que generaliza la regresión lineal ordinaria mediante permitiendo que el modelo lineal se relacione con una variable de respuesta que sigue a la familia exponencial a través de un enlace apropiado función. La regresión logística puede ser binomial o multinomial.
El modelo de regresión logística binomial tiene la siguiente forma:

$$p(y|x, w) = Ber(y|sigm(w^{T}_{incógnita}))$$

donde w y x son vectores extendidos, es decir,

$$w = (b, w1, w2, ..., wD), x = (1, x1, x2, ..., xD).$$

2) Redes neuronales artificiales: Redes neuronales artificiales (ANN) fue construida para imitar cómo funciona el cerebro humano. Se forma mediante la creación de una red de pequeñas unidades de procesamiento. llamadas neuronas. Cada neurona es muy primitiva, pero la red... Puede lograr tareas complejas como reconocimiento de patrones, imágenes clasificación y detección, procesamiento del lenguaje natural, etc. Matemáticamente, la ANN puede considerarse un tipo de regresión. sistema que predice y estima nuevos valores a partir de datos históricos registros. ANN es capaz de estimar cualquier función no lineal siempre que se suministraran suficientes conjuntos de datos para entrenar la ANN. La arquitectura de ANN se construye con tres capas:

1) Capa de entrada: Esta capa recibe la característica para la modelo.
2) Capa oculta: Esta capa consta de uno o más capas que identifican la profundidad de la RNA. Cada capa está conectado a través de nodos con aristas ponderadas. El El rendimiento del modelo depende en gran medida de la Capas ocultas y su conectividad con la entrada y capas de salida.
3) contenido...

3) Máquinas de vectores de soporte: Máquinas de vectores de soporte (SVM) se han introducido como una máquina nueva y novedosa Técnica de aprendizaje según la teoría del aprendizaje estadístico. Las SVM se utilizan para problemas de clasificación y regresión. La minimización del riesgo estructural (SRM) aplicada por SVM puede ser Superior a la Minimización de Riesgo Empírico (ERM) desde SRM minimiza el error de generalización.

La forma primaria del SVM para la clasificación es:

$$H : y = f(x) = signo(wx + b)$$

Para la regresión, el SVM se representa como:

$$H : y = f(x) = w^{T} x + b$$

4) Árboles de decisión: Los árboles de decisión son poderosos métodos de minería de datos que se pueden utilizar para la clasificación y la predicción. Los árboles de decisión representan reglas que son fáciles de interpretar. Existen múltiples métodos utilizados para crear árboles de decisión para Ejemplo: Dicotomizador iterativo 3 (ID3) y C4.5. Decisión

Machine Translated by Google

Los árboles de decisión son métodos de aprendizaje supervisado. Su mecanismo de funcionamiento consiste en dividir los datos aleatoriamente en conjuntos de entrenamiento y prueba.

Los árboles de decisión presentan una alta varianza debido a que el modelo produce diferentes resultados, como el bagging y el boosting. En las técnicas de bagging, muchos árboles de decisión se construyen en paralelo y forman los aprendices base de la técnica de bagging. Los datos muestreados se utilizan para el entrenamiento de los aprendices.

En las técnicas de boosting, los árboles se construyen secuencialmente con menos divisiones. Estos árboles pequeños, que no son muy profundos, son muy interpretables. Las técnicas de validación como k-fold ayudan a encontrar los parámetros óptimos, lo que a su vez ayuda a determinar la profundidad óptima del árbol. Además, es fundamental detener cuidadosamente los criterios de boosting para evitar el sobreajuste.

5) eXtreme Gradient Boosting (XGBoost): XGBoost, un sistema de aprendizaje automático escalable para potenciar árboles que ha demostrado ser muy popular en competiciones de aprendizaje automático como Kaggle y kdnuggests. La mayoría de los equipos ganadores utilizan o complementan su solución con XGBoost. Este éxito se puede atribuir principalmente a la característica de escalabilidad heredada del algoritmo. Esta escalabilidad se debe al algoritmo de aprendizaje optimizado para trabajar con datos dispersos, al paralelismo y al uso de multiproceso [20]. XGBoost es un algoritmo de boosting que utiliza la técnica de optimización por descenso de gradiente con una función objetivo de aprendizaje regulada. Presenta las siguientes características:

1) Regularización: XGBoost evita el sobreajuste mediante el uso de regularización L1 y L2.
2) Bosquejo de cuantiles ponderado: Encontrar los puntos de división es la tarea principal de la mayoría de los algoritmos de árboles de decisión. Su rendimiento se ve afectado si los datos están ponderados. XGBoost gestiona los datos ponderados mediante un algoritmo distribuido de bosquejo de cuantiles ponderado.
3) Estructura de bloques para aprendizaje paralelo: XGBoost utiliza múltiples núcleos en la CPU mediante una estructura de bloques que forma parte de su diseño. Los datos se ordenan y almacenan en unidades o bloques en memoria, lo que permite su reutilización iterativa. Esto también resulta útil para tareas de búsqueda de divisiones y submuestreo de columnas.
4) Manejo de datos dispersos: Los datos pueden volverse dispersos por diversas razones, como valores faltantes o codificación one-hot. El algoritmo de búsqueda de divisiones XGBoost puede manejar diferentes tipos de patrones de dispersión en los datos.
5) Conocimiento de caché: En XGBoost, se requiere acceso discontinuo a memoria para obtener las estadísticas de gradiente por índice de fila. Por lo tanto, XGBoost se ha diseñado para optimizar el uso del hardware. Esto se logra asignando búferes internos en cada hilo, donde se pueden almacenar las estadísticas de gradiente.
6) Computación fuera del núcleo: esta característica optimiza el espacio de disco disponible y maximiza su uso cuando se manejan conjuntos de datos grandes que no caben en la memoria.

D. Evaluación del modelo

La evaluación es una etapa clave en el análisis de datos que evalúa la capacidad predictiva del modelo e identifica el modelo con mejor rendimiento [17]. Se utilizan varias técnicas para evaluar modelos de clasificación, como la matriz de confusión, la curva operador-receptor (ROC) y el área bajo la curva (AUC). Una matriz de confusión muestra las clasificaciones correctas. Verdaderos positivos (VP) y verdaderos negativos (VN), además de falsos positivos (FP) y falsos negativos (FN) por clasificación incorrecta [21]. La precisión se calcula a partir de la matriz de confusión, que proporciona los valores de precisión (porcentaje de datos correctamente clasificados) y recuperación (porcentaje de datos correctamente etiquetados). Las ecuaciones 1-6 muestran las fórmulas de las matrices de rendimiento.

$$TPR = \frac{TP}{TP + FN} \quad (1)$$

$$FPR = \frac{FP}{FP + TN} \quad (2)$$

$$Precisión = \frac{TP}{TP + FP} \quad (3)$$

$$Precisión = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

$$Recordar = \frac{TP}{TP + FN} \quad (5)$$

$$F - medida = \frac{2 \quad Recordatorio \quad Precisión}{recuperación + precisión} \quad (6)$$

V. RESULTADOS Y ANÁLISIS

Utilizando Python, Jupyter Notebook y las bibliotecas de ciencia de datos Scikit Learn, Pandas y Matplot, hemos desarrollado un flujo de trabajo para procesar el conjunto de datos y generar los modelos de predicción de la gravedad de accidentes correspondientes. Este flujo se compone de varios nodos, a saber:

1) Conjunto de datos: contiene los datos preprocesados para el experimento. 2) Explorar datos: es un nodo opcional para ayudar en la exploración de datos y la visualización de algunas estadísticas sobre los datos antes del modelado.
3) Modelo: contiene los algoritmos que se utilizarán para la generación del modelo.
4) Aplicar: donde se aplica el modelo a los predictores para generar los resultados requeridos. 5) Predictores:
conjunto de datos de muestra para probar la predicción.
6) Predicción: tabla resultante después de aplicar el modelo sobre los predictores.

El conjunto de datos que hemos utilizado corresponde a los accidentes de tráfico ocurridos en el Reino Unido entre 2005 y 2019, obtenidos del Departamento de Transporte del Reino Unido. Los datos se presentan en tres archivos diferentes: accidentes, víctimas y vehículos en las Tablas I, II y III. El campo de índice de accidentes une las tres tablas en una relación de uno a muchos, donde un registro de accidente corresponde a uno o más registros de víctimas y vehículos. La tabla de víctimas incluye un campo llamado Referencia de vehículo que vincula un registro de víctima específico con un registro de información del vehículo y del conductor mediante el mismo campo de índice de accidentes. El conjunto de datos original contiene aproximadamente 2 millones de accidentes, 2 millones de víctimas y 5 millones de vehículos. La tabla combinada resultó en 3 millones de registros.

Los datos se analizaron mediante gráficos de barras e histogramas para buscar tendencias y patrones. Se muestran ejemplos de estos gráficos en las figuras 2, 3 y 4.

De los datos se pueden observar las siguientes observaciones:

Machine Translated by Google

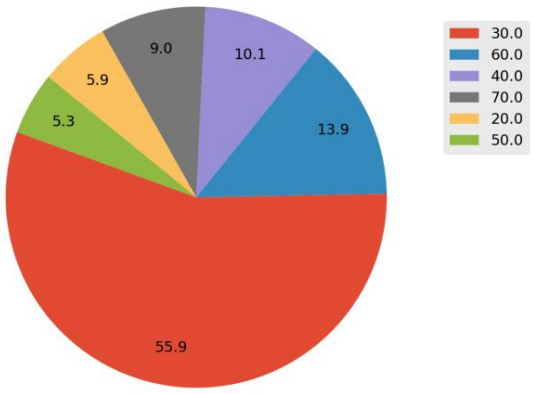

Fig. 2. Accidentes por zona de velocidad

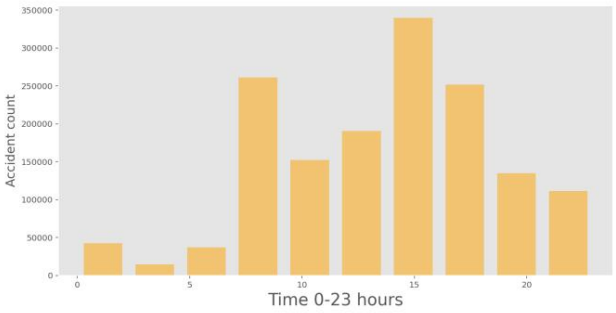

Fig. 3. Hora del día del accidente

1) Gravedad del accidente (más del 90% de registros con gravedad leve).

2) La mayoría de los accidentes involucran menos de cinco automóviles, con una mediana de 2 y una media de 1,8.

3) El número de víctimas oscila entre 1 y 93, con una media de 1,3 y una mediana de 1.

4) Accidentes repartidos a lo largo de la semana con ligero incremento de accidentes los días jueves.

5) Los accidentes se distribuyen a lo largo del día con un ligero aumento en el momento de regreso a la escuela durante los días laborales.

6) Las carreteras de clase 3 de primera clase tienen un número máximo de accidentes y son carreteras de un solo carril y un límite de velocidad de 30 MPH.

7) Los cruces sin control tienen más accidentes que otros tipos de accidentes.

8) La mayoría de los accidentes ocurren en condiciones climáticas agradables y con la superficie de la carretera seca.

Se limpió la fecha de registros incompletos. Todos los registros con celdas vacías o valor -1 se consideraron faltantes y se eliminaron. Se creó un histograma para cada columna y se eliminaron las columnas no extendidas.

1) Pasajero de autobús o autocar 2) Remolque y articulación 3) Ubicación del vehículo: Carril restringido 4) Punto de impacto 5) ¿El vehículo tenía el volante a la izquierda? 6) Clase de carretera, número de carretera 7) Clase de carretera _ _ _ _

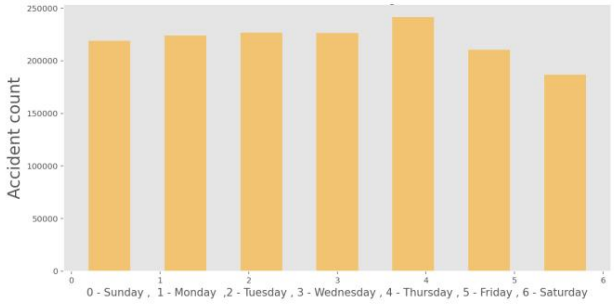

Fig. 4. Día de la semana de accidentes

8) ° Número de Carretera
9) Paso de peatones _
10) Control humano_
11) Condiciones especiales en el sitio

El número resultante de atributos utilizados para la construcción del modelo es de 48 características. Se implementó un preprocesamiento adicional en los atributos de tipo categoría mediante la codificación con valores 0 y 1. Los campos numéricos y ordinales se escalaron para eliminar el sesgo debido a valores elevados. De 3 millones de registros de accidentes, solo 29.000 fueron mortales y 190.000, lesiones graves. Los registros graves y mortales se agrupan en una sola clase, y las lesiones leves en la segunda. Esto reduce el desequilibrio de clases, junto con el método de submuestreo, lo que permite que los modelos obtengan mejores resultados.

Los modelos seleccionados para este experimento son: regresión logística, árboles de decisión, bosque aleatorio, redes neuronales y XGBoost. Se aplicó un ajuste de hiperparámetros a los métodos.
El conjunto de datos se dividió en dos partes: 70% para entrenamiento y 30% como datos de prueba. Las métricas utilizadas para evaluar los algoritmos fueron la precisión equilibrada, que suele emplearse con conjuntos de datos desequilibrados. Los resultados de precisión equilibrada se muestran en la Tabla IV y las curvas ROC en la Figura 5.

TABLA IV. RESULTADOS DE PRECISIÓN EQUILIBRADA

Método precisión balanceada Regresión logística 66.26 Árboles de decisión 69.42 Máquinas de vectores de soporte 53.22 Redes neuronales 67.23	
Bosque aleatorio	73.82
XGBoost	74.40

XGBoost y Random Forest han demostrado un mejor rendimiento que la regresión logística, las máquinas de vectores de soporte y las redes neuronales. Esto se puede atribuir a la naturaleza de la tarea de modelado y a los datos utilizados. La mayoría de los atributos tienen valores de categoría donde se informa que los métodos basados en árboles de decisión superan a los métodos basados en regresión. Aunque con un gran número de dimensiones, los métodos basados en árboles de decisión tienden a afectar su...
Rendimiento: según estos datos, estos métodos siguen superando a los clasificadores lineales y no lineales. Sin embargo, una desventaja es el rendimiento: a medida que aumenta el tamaño de los datos, el tiempo para obtener los resultados es mayor en comparación con la regresión logística. Además, se necesitaron más investigaciones para comparar el rendimiento con los métodos basados en reglas, que, según se informa, funcionan bien con datos categóricos.

Machine Translated by Google

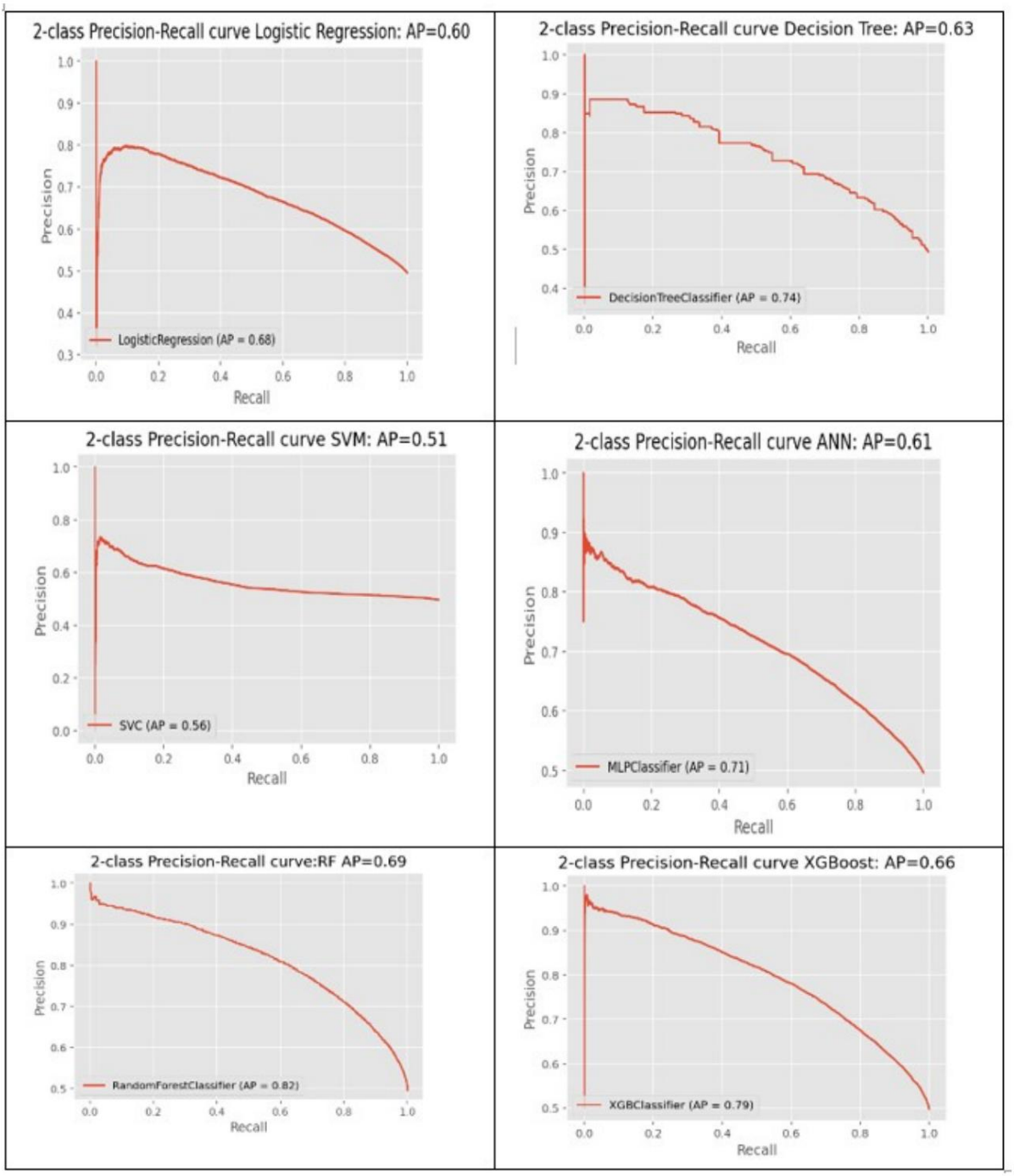

Figura 5. Curvas ROC para LR, SVM, ANN, DT, RF, XGBoost

La figura de importancia de las características se muestra en la Fig. 6. Esta figura muestra los 20 atributos principales que influyen en el nivel de gravedad. El atributo principal es el tipo de víctima, que especifica si se trata de un peatón o un pasajero. Le siguen los atributos de vehículo y área. Si bien el 66 % de los accidentes se produjeron dentro del límite de velocidad de 48 km, la tabla de importancia de las características muestra que el límite de velocidad tiene un menor efecto en el tipo de lesión. Esta información puede ayudar a las autoridades de tráfico a priorizar medidas para reducir los niveles de lesiones.

VI. CONCLUSIÓN

Este artículo presentó un marco de análisis de datos en el que se analizaron los datos de accidentes de tráfico del Reino Unido para establecer un modelo de predicción de la gravedad de las lesiones. El artículo utilizó datos públicos de 2005 a 2019 para construir modelos de predicción del nivel de gravedad de las lesiones. El artículo combinó todos los atributos de tres fuentes de datos para analizar 63 atributos y su relación con la gravedad de los accidentes. El artículo destacó problemas relacionados con la calidad y el desequilibrio de los datos, y aplicó técnicas para abordarlos. El artículo comparó el rendimiento entre

Machine Translated by Google

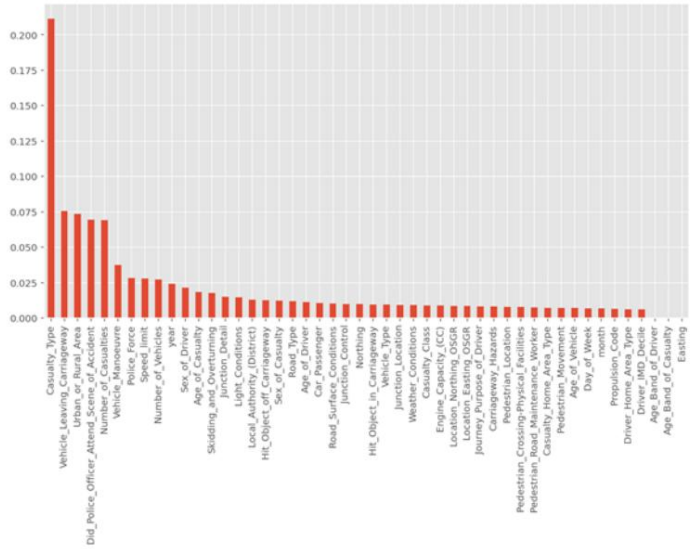

Fig. 6. Importancia de las características

Diferentes técnicas de aprendizaje automático, el algoritmo XGBoost fue Se ha demostrado que supera a otras técnicas con mayor precisión. tasa incluso con datos desequilibrados. Se sugiere seguir trabajando para Utilice bibliotecas de procesamiento paralelo y compare el rendimiento de técnicas basadas en reglas y técnicas basadas en decisiones.

RECONOCIMIENTO

Este trabajo está financiado por la subvención número 17-COM-1-01-0007,Decanato de Investigación Científica (DSR) de Umm al Qura Universidad del Reino de Arabia Saudita. Los autores...
Me gustaría expresar mi gratitud por el apoyo y la generosidad. Contribución a la continuación de la investigación en este ámbito

REFERENCIAS

[1] Organización Mundial de la Salud (OMS), Una ruta Salud Técnico Paquete de seguridad, 2017. [En línea]. Disponible: http://iris.paho.org/xmlui/bitstream/handle/123456789/34980/9789275320013-por.pdf?secuencia=1{\&}está permitido=y

[2] A. Mehdizadeh, M. Cai, Q. Hu, MAA Yazdi, N. Mohabbati-Kalejahi, A. Vinel, SE Rigdon, KC Davis y FM Megahed, “Una revisión de Aplicaciones del análisis de datos en la seguridad vial. Parte 1: Descriptivo y modelado predictivo”, Sensors (Suiza), vol. 20, núm. 4, págs. 1–24, 2020.

[3] Q. Hu, M. Cai, N. Mohabbati-Kalejahi, A. Mehdizadeh, MAA Yazdi, A. Vinel, SE Rigdon, KC Davis y FM Megahed, “Una revisión de aplicaciones del análisis de datos en la seguridad vial. Parte 2: Prescriptiva modelado”, Sensors (Suiza), vol. 20, núm. 4, págs. 1–19, 2020.

[4] A. Ziakopoulos y G. Yannis, “Una revisión de los enfoques espaciales en Seguridad vial”, Accid. Anal. Prev., vol. 135, n.º julio, p. 105323, 2020. [En línea]. Disponible en: https://doi.org/10.1016/j.aap.2019.105323

[5] S. Moosavi, MH Samavatian, A. Nandi, S. Parthasarathy y R. Ram-nath, “Descubrimiento de patrones a corto y largo plazo sobre datos geoespaciotemporales a gran escala”, Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Datos mínimos, págs. 2905–2913, 2019.

[6] N. Zagorodnikh, A. Novikov y A. Yastrebkov, “Algoritmo y software para identificar tramos de carretera propensos a accidentes”, Transp. Res. Procedia, vol. 36, págs. 817–825, 2018. [En línea]. Disponible: https://doi.org/10.1016/j.trpro.2018.12.074

[7] DW Kononen, CA Flannagan y SC Wang, “Identificación y validación de un modelo de regresión logística para predecir lesiones graves asociadas con accidentes automovilísticos”, Accid. Anal. Prev., vol. 43, n.º 1, págs. 112-122, 2011. [En línea]. Disponible: http://dx.doi.org/10.1016/j.aap.2010.07.018

[8] D. Delen, R. Sharda y M. Bessonov, “Identificación de predictores significativos de la gravedad de las lesiones en accidentes de tráfico utilizando una serie de redes neuronales", Accid. Anal. Prev., vol. 38, no. 3, págs. 434–444, 2006.

[9] A. Naseer, MK Nour y BY Alkazemi, “Hacia el aprendizaje profundo Análisis de accidentes de tráfico basado en la nube”, en 2020, 10.º Congreso Anual de Computación y Taller y Conferencia de Comunicación (CCWC), 2020, págs. 0817–0820.

[10] B. Sharma, VK Katiyar y K. Kumar, “Predicción de accidentes de tráfico " Modelo que utiliza máquinas de vectores de soporte con núcleo gaussiano A Características de accidentes A Minería de datos Adv. Intel. Sistema. Computación, vol. 437, págs. 1–10, 2016.

[11] H. Meng, X. Wang y X. Wang, “Predicción de accidentes en autopistas basada en sobre big data de tráfico”, ACM Int. Conf. Proceeding Ser., págs. 11–16, 2018.

[12] M. Schlogl, R. St. utz, G. Laaha y M. Melcher, “Una comparación de métodos de aprendizaje estadístico para derivar factores determinantes de accidentes ocurrencia a partir de un conjunto de datos desequilibrados de alta resolución”, Accid. Anal. Prev., vol. 127, n.º enero, págs. 134–149, 2019. [En línea]. Disponible: https://doi.org/10.1016/j.aap.2019.02.008

[13] J. Ma, Y. Ding, JC Cheng, Y. Tan, VJ Gan y J. Zhang, “Analyzing Las principales causas de accidentes de tráfico mortales mediante XGBoost y sistemas basados en la red Análisis: una perspectiva de gestión urbana”, IEEE Access, vol. 7, pp. 148 059–148 072, 2019.

[14] LG Cuenca, E. Puertas, N. Aliane y JF Andrés, “Accidentes de tráfico Clasificación y predicción de la gravedad de las lesiones”, en el 3.er IEEE Int. de 2018. Conf. Intel. Transp. Ing. ICITE 2018, 2018, págs. 52–57.

[15] Departamento de Transporte, “Orientación sobre estadísticas de tráfico por carretera”, págs. 1–13, 2014. [En línea]. Disponible: http://data.dft.gov.uk/gb-traffic-matrix/ metadatos de todos los datos de tráfico.pdf

[16] B. Burdett, “Mejora de la precisión de la evaluación de la gravedad de las lesiones mediante KABCO por parte de las fuerzas del orden”, Univ. Wisconsin-Madison, 2014.

[17] J. Han, M. Kamber y J. Mining Concepts, Pei. (2012) Datos Waltham, Massachusetts. y Técnicas, tercera edición. [En línea]. Disponible en: {http://www.amazon.de/ Conceptos y técnicas de gestión de minería de datos/dp/0123814790/ref= título hrd tmm 0?ie=UTF8\&qid=1366039033\&sr=1-1}

[18] A. Fernández, S. García, M. Galar, RC Prati, B. Krawczyk y F. Herrera, Aprendiendo de conjuntos de datos desequilibrados, 2018.

[19] L. Gautheron, A. Habrard, E. Morvant y M. Sebban, “Aprendizaje métrico a partir de datos desequilibrados”, Proc. - Int. Conf. Herramientas con Artif. Intell. ICTAI, vol. 2019-noviembre, núm. 9, págs. 923–930, 2019.

[20] T. Chen y C. Guestrin, “XGBoost: Un sistema escalable de refuerzo de árboles”, Proc. ACM SIGKDD Int. Conf. Knowl. Descubrimiento. Datos Min., vol. 13-17-Agosto-2016, págs. 785–794, 2016.

[21] P. Hajek, M. Hole na y J. Rauch, “El método GUHA y su significado para la minería de datos”, J. Comput. Syst. Sci., vol. 76, n.º 1, págs. 34–48, 2010.