UNIVERSIDAD TECNOLÓGICA INDOAMÉRICA

FACULTAD DE INGENIERÍAS

MAESTRÍA EN BIG DATA Y CIENCIA DE DATOS

TEMA:

PREDICCIÓN DE LA SEVERIDAD DE ACCIDENTES DE TRÁNSITO EN QUITO MEDIANTE INTELIGENCIA ARTIFICIAL: UN ENFOQUE DE CLASIFICACIÓN BINARIA

Trabajo de Titulación previo a la obtención del título de Magíster en Big Data y Ciencia de Datos

Autor(a)

Ing. Jean Kevin Morillo Hernández

Tutor(a)

Ing. Andrés Xavier Rubio Proaño, PhD.

AMBATO – ECUADOR

2025

AUTORIZACIÓN POR PARTE DEL AUTOR PARA LA CONSULTA, REPRODUCCIÓN PARCIAL O TOTAL, Y PUBLICACIÓN ELECTRÓNICA DEL TRABAJO DE TITULACIÓN

Yo, Jean Kevin Morillo Hernández, declaro ser autor del Trabajo de Titulación con el nombre "PREDICCIÓN DE LA SEVERIDAD DE ACCIDENTES DE TRÁNSITO EN QUITO MEDIANTE INTELIGENCIA ARTIFICIAL: UN ENFOQUE DE CLASIFICACIÓN BINARIA", como requisito para optar al grado de Magíster en Big Data y Ciencia de Datos y autorizo al Sistema de Bibliotecas de la Universidad Indoamérica, para que con fines netamente académicos divulgue esta obra a través del Repositorio Digital Institucional (RDI-UTI).

Los usuarios del RDI-UTI podrán consultar el contenido de este trabajo en las redes de información del país y del exterior, con las cuales la Universidad tenga convenios. La Universidad Indoamérica no se hace responsable por el plagio o copia del contenido parcial o total de este trabajo.

Del mismo modo, acepto que los Derechos de Autor, Morales y Patrimoniales, sobre esta obra, serán compartidos entre mi persona y la Universidad Indoamérica, y que no tramitaré la publicación de esta obra en ningún otro medio, sin autorización expresa de la misma. En caso de que exista el potencial de generación de beneficios económicos o patentes, producto de este trabajo, acepto que se deberán firmar convenios específicos adicionales, donde se acuerden los términos de adjudicación de dichos beneficios.

Para constancia de esta autorización, en la ciudad de Ambato a los 16 días del mes de septiembre de 2025, firmo conforme:

Autor: Jean Kevin Morillo Hernández

Firma: …….

Número de Cédula: 2100400791

Dirección: Pichincha, Quito, Calderón, Carapungo.

Correo Electrónico: jeanmorillo1998@gmail.com

Teléfono: 0962576157

APROBACIÓN DEL DIRECTOR

En mi calidad de Director del Trabajo de Titulación "PREDICCIÓN DE LA SEVERIDAD DE ACCIDENTES DE TRÁNSITO EN QUITO MEDIANTE INTELIGENCIA ARTIFICIAL: UN ENFOQUE DE CLASIFICACIÓN BINARIA" presentado por Jean Kevin Morillo Hernández, para optar por el Título de Magíster en Big Data y Ciencia de Datos.

CERTIFICO

Que dicho Trabajo de Titulación ha sido revisado en todas sus partes y considero que reúne los requisitos y méritos suficientes para ser sometido a la presentación pública y evaluación por parte los Examinadores que se designe.

Ambato, 19 de septiembre del 2025

……

Ing. Andrés Xavier Rubio Proaño, PhD.

DIRECTOR

DECLARACIÓN DE AUTENTICIDAD

Quien suscribe, declaro que los contenidos y los resultados obtenidos en el presente Trabajo de Titulación, como requerimiento previo para la obtención del Título de Magíster en Big Data y Ciencia de Datos, son absolutamente originales, auténticos y personales y de exclusiva responsabilidad legal y académica del autor

Ambato, 16 de septiembre del 2025

.....

Jean Kevin Morillo Hernández

2100400791

APROBACIÓN DE EXAMINADORES

El Trabajo de Titulación ha sido revisado, aprobado y autorizada su impresión y empastado, sobre el Tema: PREDICCIÓN DE LA SEVERIDAD DE ACCIDENTES DE TRÁNSITO EN QUITO MEDIANTE INTELIGENCIA ARTIFICIAL: UN ENFOQUE DE CLASIFICACIÓN BINARIA, previo a la obtención del Título de Magíster en Big Data y Ciencia de Datos, reúne los requisitos de fondo y forma para que el estudiante pueda presentarse a la sustentación del Trabajo de Titulación.

Ambato, 16 de septiembre del 2025

Firmado electrónicamente por:
FATIMA ADRIANA AVILES CASTILLO
Validar únicamente con FirmaEC

……..

Ing. Fátima Adriana Avilés Castillo, Mg.

EXAMINADOR

Firmado electrónicamente por:
JANIO LINCON JADAN GUERRERO
Validar únicamente con FirmaEC

……

Ing. Janio Jadan Guerrero, PhD.

EXAMINADOR

DEDICATORIA

A mi gran amiga Fer, por su incondicional apoyo en cada etapa de este camino. Su motivación, paciencia y compañía han sido fundamentales para alcanzar esta meta. A mi madre, por ser mi mayor inspiración y por confiar siempre en mí. Su amor y apoyo inquebrantable han sido la base sobre la que he construido mis logros. Finalmente, a mí mismo por la perseverancia, las largas noches de esfuerzo y la determinación de seguir adelante, a pesar de los desafíos que se presentaron en el camino.

AGRADECIMIENTO

Quiero expresar mi sincero agradecimiento a la Universidad Tecnológica Indoamérica por brindarme la oportunidad de formarme en la Maestría de Big Data y Ciencia de Datos.

Asimismo, agradezco a todos los profesores por su dedicación y por compartir sus valiosos conocimientos, los cuales han sido fundamentales para el desarrollo de esta investigación.

A mi tutor, por su orientación y apoyo constante durante todo el proceso. Su guía fue clave para llevar a cabo este trabajo.

Finalmente, a todos aquellos que, de una u otra forma, contribuyeron a la realización de este proyecto, ya sea con su apoyo, motivación o consejos.

ÍNDICE DE CONTENIDOS

AUTORIZACIÓN POR PARTE DEL AUTOR PARA LA CONSULTA, REPRODUCCIÓN PARCIAL O TOTAL, Y PUBLICACIÓN ELECTRÓNICA DEL TRABAJO DE TITULACIÓN ii
APROBACIÓN DEL DIRECTOR iii
DECLARACIÓN DE AUTENTICIDAD iv
APROBACIÓN DE EXAMINADORES v
DEDICATORIA vi
AGRADECIMIENTO vii
ABSTRACT xv
CAPÍTULO I 1
Introducción 1
Antecedentes 4
Importancia de la seguridad vial 4
Modelos tradicionales en la predicción de accidentes 4
Aplicación de la IA en la prediccion de accidentes 5
Estudios previos 6
Brecha de investigación en Ecuador 8
Justificación 10
Objetivos 12
CAPÍTULO II 13
Área de Estudio 13
Enfoque 15
Enfoque Cuantitativo 15
Enfoque Cualitativo 16
Descripción de la metodología 16
Diseño del trabajo 18
Variable Dependiente 18
Variables Independientes 20
Procedimiento para obtención y análisis de datos 20
Población y muestra 21
Hipótesis 22

Fundamentación 22
Respaldo en la literatura 23
Pertinencia para el contexto ecuatoriano 23
Metodología de contrastación 23
CAPÍTULO III 24
Tratamiento y Detección de Outliers 24
Análisis descriptivo de las variables 25
Variables Temporales 26
Variables categóricas relevantes 29
Ingeniería de Características 31
Transformación de variables temporales 32
Creación de Variables Derivadas 32
Codificación de Variables Categóricas 33
Análisis de Correlación 33
División de Datos 34
Selección de Modelos de Machine Learning 34
XGBoost (Extreme Gradient Boosting) 35
Random Forest 35
LightGBM (Light Gradient Boosting Machine) 36
FFNN (Feed-Forward Neural Network) 37
Métricas e Interpretabilidad 38
Selección de Métricas de Evaluación 38
Importancia de Minimizar los Falsos Negativos 40
Análisis de Interpretabilidad 40
CAPÍTULO IV 41
Interpretación de resultados 41
Desempeño de los Modelos 41
Matrices de Confusión 43
Contraste con investigaciones previas 44
Interpretabilidad 45
Verificación de la hipótesis 47
CAPÍTULO V 49

Conclusiones 49
Recomendaciones 50
Referencias 52
Apéndices 58
Apéndice A - Listado de tipos de siniestros 58
Apéndice B - Distribución de Personas Lesionadas 59
Apéndice C - Matrices de Confusión 60
Apéndice D - Importancia de las características usando SHAP 62
Apéndice E - Código en Python para la verificación de la hipótesis 63

ÍNDICE DE TABLAS

Tabla No. 1: Diccionarios de variables independientes. 20
Tabla No. 2: Modelos en su forma base. 41
Tabla No. 3: Modelos optimizados..... 42
Tabla No. 4: Matriz de confusión. 43

ÍNDICE DE GRÁFICOS

Gráfico No. 1: Vehículos motorizados matriculados. 1
Gráfico No. 2: Personas involucradas en accidentes de tránsito en Ecuador (primer trimestre 2024)..... 2
Gráfico No. 3: Mapa Conceptual..... 13
Gráfico No. 4: Distribución de la severidad de accidentes de tránsito..... 19
Gráfico No. 5: Distribución de variables numéricas seleccionadas. 25
Gráfico No. 6: Distribución temporal de los accidentes de tránsito..... 27
Gráfico No. 7: Distribución de accidentes por hora y severidad..... 28
Gráfico No. 8: Variables categóricas relevantes. 31
Gráfico No. 9: Valores SHAP. 46

ÍNDICE DE IMÁGENES

Imagen No. 1: Área de estudio: Distrito Metropolitano de Quito. 14
Imagen No. 2: Mapa de calor de accidentes de tránsito por severidad. 15

UNIVERSIDAD TECNOLÓGICA INDOAMÉRICA
FACULTAD DE INGENIERÍAS
MAESTRÍA EN BIG DATA Y CIENCIA DE DATOS

TEMA: PREDICCIÓN DE LA SEVERIDAD DE ACCIDENTES DE TRÁNSITO EN QUITO MEDIANTE INTELIGENCIA ARTIFICIAL: UN ENFOQUE DE CLASIFICACIÓN BINARIA.

AUTOR(A): Ing. Jean Kevin Morillo Hernández
TUTOR (A): Ing. Andrés Xavier Rubio Proaño, PhD.

RESUMEN EJECUTIVO

En Ecuador, los accidentes de tránsito representan un desafío creciente para la seguridad vial, agravado por un incremento del 20,10% del parque vehicular en Pichincha durante 2023 y por el hecho de que el 44,38% de los accidentes involucraron víctimas. Esta situación evidencia deficiencias en la planificación y gestión de la movilidad y seguridad vial. El objetivo de esta investigación fue desarrollar modelos de inteligencia artificial para la predicción binaria de la severidad de los accidentes de tránsito en Quito, clasificándolos en "con víctimas" y "sin víctimas". La metodología incluyó la recopilación de datos de la Agencia Nacional de Tránsito entre enero de 2017 y abril de 2024, obteniéndose 35.632 registros tras filtrado geográfico y preprocesamiento mediante codificación cíclica y análisis de correlación. Se evaluaron los algoritmos Random Forest, XGBoost, LightGBM y FFNN en versiones base y optimizadas, priorizando la reducción de falsos negativos. Los resultados mostraron mayor frecuencia de accidentes los sábados y durante horas pico, siendo los atropellos los más propensos a generar víctimas; además, factores como exceso de velocidad y no ceder el paso a peatones resultaron determinantes. El modelo LightGBM optimizado obtuvo el mejor desempeño, con recall del 87% y AUC-ROC de 0,9373, reduciendo los falsos negativos a 648 casos. El análisis con SHAP indicó que variables como "Motocicleta" y "Peatón" aumentan la probabilidad de accidentes con víctimas, mientras que "Conductor Ausente" reduce ese riesgo. En conclusión, los modelos de inteligencia artificial demostraron alta capacidad predictiva, identificando patrones y factores críticos que influyen en la severidad de los accidentes en Quito, apoyando decisiones de gestión vial y priorización de recursos de emergencia.

DESCRIPTORES: (Accidentes de tránsito, Inteligencia artificial, Predicción de severidad, Modelos predictivos)

UNIVERSIDAD TECNOLÓGICA INDOAMÉRICA

FACULTY OF ENGINEERING

MASTER'S IN BIG DATA AND DATA SCIENCE

AUTHOR: MORILLO HERNANDEZ JEAN KEVIN

TUTOR: DR. RUBIO PROAÑO ANDRES XAVIER

ABSTRACT

PREDICTION OF TRAFFIC ACCIDENT SEVERITY IN QUITO USING ARTIFICIAL INTELLIGENCE: A BINARY CLASSIFICATION APPROACH

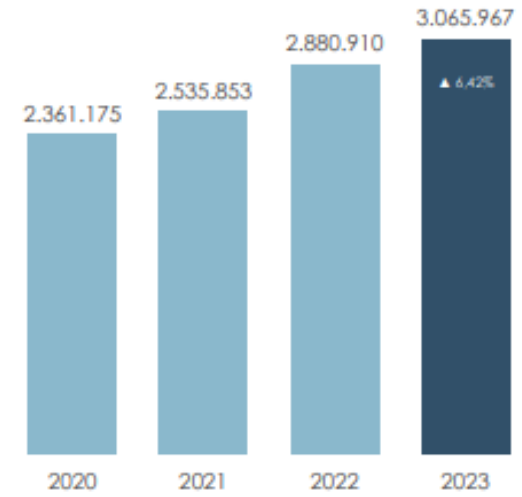
In Ecuador, traffic accidents represent a growing challenge to road safety, exacerbated by a 20.10% increase in the vehicle fleet in Pichincha during 2023 and by the fact that 44.38% of accidents resulted in victims. This situation highlights deficiencies in mobility planning and road safety management. The objective of this research was to develop artificial intelligence models for predicting the severity of traffic accidents in Quito, classifying them as either “with victims” or “without victims”. The methodology involved data collection from the National Transit Agency between January 2017 and April 2024, resulting in 35,632 records after geographic filtering and preprocessing through cyclical encoding and correlation analysis. Random Forest, XGBoost, LightGBM, and FFNN algorithms were evaluated in both baseline and optimized versions, with priority given to reducing false negatives. The results showed a higher frequency of accidents on Saturdays and during peak hours, with pedestrian run-overs being the most likely to result in victims. Additionally, factors such as speeding and failure to yield to pedestrians proved decisive. The optimized LightGBM model achieved the best performance, with a recall of 87% and an AUC-ROC of 0.9373, resulting in a reduction of false negatives to 648 cases. SHAP analysis indicated that variables such as “Motorcycle” and “Pedestrian” increase the probability of victim-related accidents, while “Absent Driver” reduces this risk. In conclusion, the artificial intelligence models demonstrated high predictive capacity, identifying patterns and critical factors that influence accident severity in Quito, supporting road management decisions and the prioritization of emergency resources.

KEYWORDS: Artificial intelligence, predictive models, severity prediction, traffic accidents

CAPÍTULO I
INTRODUCCIÓN

Introducción

La predicción de la severidad de los accidentes de tránsito es un desafío crucial en la gestión de la seguridad vial y la optimización de los recursos de emergencia a nivel mundial. En el contexto ecuatoriano, el crecimiento de la población ha llevado a un aumento en el número de vehículos. Este fenómeno es particularmente notable en la provincia de Pichincha, donde en 2023 se registraron 683.483 vehículos matriculados, lo que representa un incremento del 20,10% en comparación con el año anterior, tal como se muestra en el Gráfico 1 (INEC, 2024). Este crecimiento ha incrementado la incidencia de accidentes, generando no solo costos humanos sino también costos económicos significativos. Según la Organización Mundial de la Salud (2023), los accidentes de tránsito representan una de las principales causas de muerte en el mundo, con aproximadamente 1,4 millones de fallecimientos anuales.



a) Vehículos motorizados matriculados por año

N°	Provincias	2022	2023	Variación
1	Guayas	656.885	715.079	8,86%
2	Pichincha	569.095	683.483	20,10%
3	Manabí	301.986	323.544	7,14%
4	Los Ríos	155.117	156.920	1,16%
5	El Oro	137.687	143.499	4,22%
6	Azuay	185.674	131.284	-29,29%
7	Tungurahua	122.899	130.273	6,00%
8	Santo Domingo	98.783	102.727	3,99%
9	Chimborazo	85.135	90.413	6,20%
10	Cotopaxi	85.682	87.657	2,31%
14	**Resto de Provincias[1]**	**481.967**	**501.088**	**3,97%**
24	**Total Nacional**	**2.880.910**	**3.065.967**	**6,42%**

b) Vehículos motorizados por provincia - 2023

Gráfico No. 1: Vehículos motorizados matriculados.
Elaborado por: INEC (2024).

El Gráfico 2 presenta la distribución de las personas involucradas en accidentes de tránsito en Ecuador durante el primer trimestre de 2024. Tal como se puede apreciar en dicho gráfico, un preocupante 44,38% de los accidentes registrados involucraron víctimas, dato que subraya la necesidad crítica de desarrollar modelos predictivos más

precisos, como los que se exploran en la presente investigación, para anticipar y mitigar la severidad de estos eventos. En contraste, el 35,70% de los involucrados resultaron ilesos y un 19,91% no fueron identificados. Dentro del grupo de víctimas, los pasajeros constituyeron el 43,14%, seguidos de cerca por los conductores (42,15%), los peatones (14,37%) y, en menor medida, los ciclistas (0,34%) (INEC, 2024). Estas cifras no solo evidencian la magnitud del problema de la siniestralidad vial, sino que también refuerzan la importancia de mejorar la gestión de la seguridad en las vías del país.

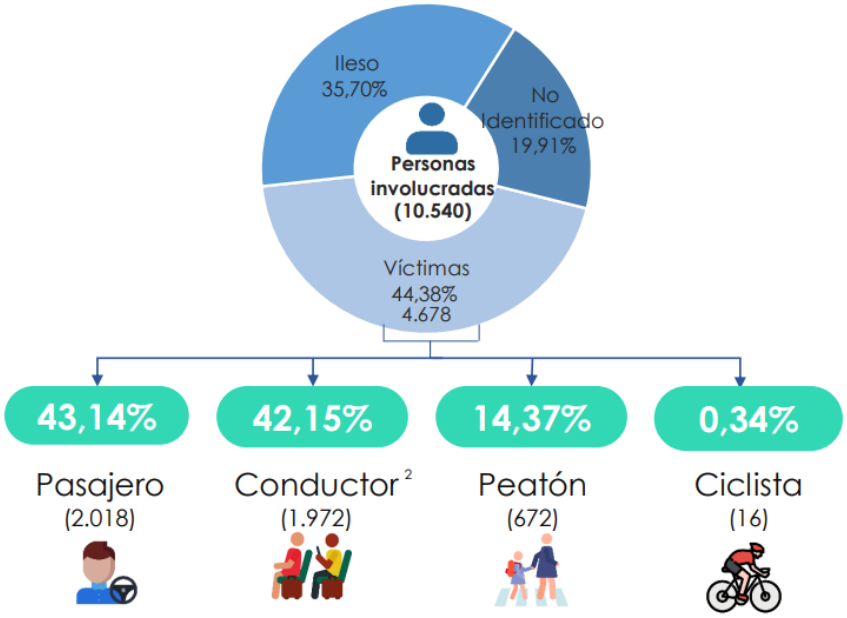


Gráfico No. 2: Personas involucradas en accidentes de tránsito en Ecuador (primer trimestre 2024).
Elaborado por: INEC (2024).

Tradicionalmente, el análisis de accidentes de tránsito se ha basado en modelos estadísticos clásicos, como la regresión logística y los modelos de series temporales. Sin embargo, estos enfoques presentan limitaciones en la captura de interacciones complejas entre las variables involucradas. Dado el aumento en la complejidad de los datos disponibles y la necesidad de modelos más precisos para abordar la problemática de la siniestralidad vial, el uso de técnicas de Inteligencia Artificial (IA) se ha convertido en una alternativa prometedora para mejorar la capacidad predictiva de los modelos de severidad de accidentes (Baykal et al., 2023; Kumar Mohanta et al., 2022).

El presente estudio propone la implementación de modelos de AI para la predicción binaria de la severidad de los accidentes de tránsito en el cantón Quito, diferenciando

entre accidentes "Con Víctimas" y "Sin Víctimas". Se utilizarán diversas técnicas, incluyendo Random Forest, XGBoost, LightGBM y redes neuronales, con un énfasis en la explicabilidad de los modelos mediante la herramienta SHAP (Shapley Additive Explanations). La investigación busca mejorar la gestión de la seguridad vial mediante modelos de IA, optimizando la respuesta de emergencias y priorizando incidentes según su gravedad estimada.

Para abordar este problema, el resto de este documento comenzará proporcionando el contexto del problema de investigación, destacando la importancia de la seguridad vial y los enfoques previos utilizados para predecir la severidad de los accidentes. A continuación, se presentará la justificación del estudio, resaltando la necesidad de un modelo basado en inteligencia artificial para optimizar la gestión de accidentes en Quito. Luego, se describirán los objetivos generales y específicos de la investigación. Posteriormente, se explorará el marco teórico, donde se discutirán los conceptos clave relacionados con la predicción de accidentes y las técnicas de AI utilizadas. La metodología detallará el área de estudio, los datos analizados y los procedimientos de preprocesamiento aplicados. En la sección de resultados, se evaluará el desempeño de los modelos propuestos, analizando métricas clave para determinar su efectividad. Finalmente, el documento concluirá con un análisis de los hallazgos obtenidos y recomendaciones para mejorar la gestión de la seguridad vial en la ciudad. Este estudio contribuye significativamente a la seguridad vial en Quito mediante la aplicación innovadora de Inteligencia Artificial para la prevención efectiva de accidentes de tránsito.

Antecedentes

Importancia de la seguridad vial

La mejora de la seguridad vial y la reducción de pérdidas económicas y humanas derivadas de los accidentes de tránsito son temas de gran relevancia a nivel global. Estos siniestros representan el 3% del PIB en países de ingresos medios (WHO, 2022). Este enfoque ha impulsado una considerable cantidad de investigaciones centradas en el análisis y predicción de la gravedad de los accidentes de tránsito.

En el cantón Quito, la gestión de accidentes de tránsito sigue un protocolo estandarizado que, aunque busca eficiencia, presenta fallas sistémicas como procedimientos manuales lentos, falta de priorización según severidad y datos inconsistentes, lo que genera respuestas ineficientes y demoras críticas (Quito Informa, 2023). Un proceso basado en inteligencia artificial podría optimizar este proceso mediante modelos predictivos que clasifiquen rápidamente los accidentes según su severidad, diferenciando entre aquellos con víctimas y sin víctimas, utilizando información básica como ubicación, causa probable, categoría de edad, etc. Esto permitiría activar respuestas adaptativas, priorizar recursos de manera eficiente y reducir los tiempos de atención, transformando el sistema de emergencias en un modelo más ágil y basado en evidencia.

Modelos tradicionales en la predicción de accidentes

Tradicionalmente, se han utilizado técnicas estadísticas como la regresión y el análisis de series temporales para identificar factores clave, tales como las condiciones climáticas, el tráfico y la calidad de las carreteras (Al-Masaeid & Khaled, 2023). Los modelos de regresión convencional han sido uno de los primeros enfoques en el desarrollo de Modelos de Predicción de Accidentes (APMs), caracterizados por su alta flexibilidad de aplicación en diversos tipos de vías, desde autopistas urbanas hasta intersecciones (Al-Masaeid et al., 2020; Khasawneh et al., 2018).

Sin embargo, estos modelos de regresión presentan limitaciones significativas al no poder explicar todas las variabilidades en los datos de accidentes de tránsito, especialmente cuando se trata de información temporal. Esto se debe principalmente a

que la naturaleza de los datos de accidentes, valores no negativos, aleatorios y enteros discretos, contradice los supuestos básicos de los modelos de regresión tradicionales.

Como respuesta a estas deficiencias, los investigadores adoptaron técnicas de modelado de series temporales, particularmente los modelos ARIMA (autoregresivos integrados de media móvil), que ofrecen ventajas al trabajar con datos ordenados cronológicamente (Avuglah et al., 2014; Hassouna et al., 2020). Estos modelos permiten capturar patrones estacionales y tendencias a largo plazo, proporcionando mejoras significativas en la predicción de accidentes bajo determinadas condiciones.

Aunque incluso con la incorporación de estos avances en los modelos estadísticos tradicionales, persisten importantes desafíos metodológicos. Tanto los modelos de regresión como los de series temporales tienen dificultades para capturar adecuadamente las complejas interacciones no lineales entre múltiples variables que influyen simultáneamente en la severidad de los accidentes. Además, estos métodos convencionales muestran limitaciones al procesar los grandes volúmenes de datos heterogéneos disponibles actualmente en el ámbito de la seguridad vial.

Es precisamente este conjunto de limitaciones lo que ha impulsado a los investigadores hacia la exploración de enfoques alternativos más sofisticados, particularmente aquellos basados en inteligencia artificial, que ofrecen capacidades superiores para el procesamiento de grandes volúmenes de datos y la detección de patrones complejos no lineales en la información disponible.

Aplicación de la IA en la prediccion de accidentes

La Inteligencia Artificial se ha convertido en una de las tecnologías más transformadoras de la era moderna. Definida como la capacidad de las máquinas para realizar tareas que tradicionalmente requerían inteligencia humana, como el aprendizaje, la toma de decisiones y la resolución de problemas, la IA está impactando una amplia gama de sectores que van desde campos como la medicina hasta la seguridad vial (Akinade, Adepoju, Ige, & Afolabi, 2024). En el contexto de la gestión de accidentes de tránsito, la IA ofrece una capacidad única para analizar grandes volúmenes de datos, identificar patrones y hacer predicciones precisas en tiempo real, lo que puede revolucionar la manera en que las autoridades pertinentes gestionan los incidentes viales.

Los avances en Inteligencia Artificial y Machine Learning han revolucionado la manera en que se abordan estos estudios, permitiendo una comprensión más profunda y detallada del fenómeno. Modelos avanzados de ML, como Random Forest, Support Vector Machines y redes neuronales profundas, han demostrado una mayor capacidad para identificar patrones complejos en los datos de accidentes, lo que se traduce en una mayor precisión en la predicción de la severidad de los accidentes en comparación con los métodos tradicionales (Infante et al., 2023). Estos modelos no solo mejoran la capacidad de predicción, sino que también facilitan la identificación de las variables más influyentes, proporcionando información valiosa para la toma de decisiones en políticas de seguridad vial.

Adicionalmente, la implementación de técnicas de Inteligencia Artificial Explicable (XAI) ha sido un avance significativo, ya que permite interpretar y comprender mejor los factores que influyen en las predicciones realizadas por los modelos de ML. Este enfoque facilita la toma de decisiones informadas, lo cual es crucial para implementar medidas efectivas que mejoren la seguridad vial.

Estudios previos

En este contexto, Adefabi et al. (2023) realizaron un estudio en el que utilizaron el algoritmo de aprendizaje automático Random Forest para predecir la severidad de los accidentes de tránsito en una gran área metropolitana. El modelo fue entrenado con un conjunto de datos de registros de accidentes y evaluado mediante varias métricas. Los resultados demostraron que el modelo Random Forest alcanzó una precisión superior al 80%. Además, identificaron las variables más importantes en la predicción, entre las que se incluyen la velocidad del viento, la presión, la humedad, la visibilidad, las condiciones claras y la cobertura de nubes. El modelo ajustado mostró un Área Bajo la Curva (AUC) del 80%, un recall del 79.2%, una precisión del 97.1% y una puntuación F1 de 87.3%. Estos resultados refuerzan la viabilidad y confiabilidad de modelos de machine learning para predecir la severidad de los accidentes.

De manera similar, Ahmed et al. (2023) demostraron que el modelo Random Forest supera a otros modelos como Decision Jungle, AdaBoost y XGBoost en la predicción de la severidad de accidentes, con una precisión hasta un 15.84% mayor. Los resultados del modelo fueron 81,45% de exactitud, 81,68% de precisión, 81,42% de recall y 81,04% de

puntuación F1. Su estudio en Nueva Zelanda destacó que factores como la categoría de la carretera, el número de vehículos y la edad del conductor influyen en la gravedad del accidente.

Jianjun Yang, Siyuan Han y Yimeng Chen (2023) desarrollaron un modelo Random Forest que integró factores como la ubicación del accidente, la forma del accidente, la calidad de la carretera y la velocidad. Su investigación concluyó que el patrón de colisión es el factor más influyente en la severidad del accidente, seguido de la estructura del vehículo y la calidad de las carreteras. Estos hallazgos subrayan la importancia de mejorar la infraestructura vial y fortalecer la capacitación en conducción segura como medidas clave para reducir los accidentes.

Otro estudio relevante fue el realizado por Juan Li et al. (2023), quienes analizaron la severidad de los accidentes en autopistas montañosas. Su investigación identificó que factores como la intensidad de la lluvia, el tipo de colisión, el número de vehículos involucrados y el tipo de sección de carretera son determinantes en la gravedad de los accidentes. Además, descubrieron que, aunque los accidentes son más comunes en días secos, la tasa de accidentes en días lluviosos es aproximadamente tres veces mayor, lo que resalta la importancia de implementar medidas preventivas en condiciones climáticas adversas. Además, su análisis mostró que las colisiones traseras son frecuentes y que los accidentes con múltiples vehículos, aunque menos comunes, tienen una mayor probabilidad de causar lesiones graves.

En un contexto más amplio, Baykal et al. (2023) llevaron a cabo un estudio en 49 estados de Estados Unidos, donde compararon varios algoritmos de ML, incluyendo Random Forest, SVM, Gradient Boosting (GB) y Multi-Layer Perceptron (MLP), entre otros, para estimar la gravedad de los accidentes. Sus resultados demostraron que el modelo Random Forest fue el más efectivo, con una precisión de 81.6%, superando a los demás modelos en términos de rendimiento de predicción. De manera similar, Obasi y Benson (2023) realizaron un estudio en el Reino Unido, donde también encontraron que el modelo Random Forest ofreció el mejor rendimiento en la predicción de la severidad de los accidentes. Este estudio identificó que los factores más influyentes en la predicción fueron la capacidad del motor, la edad del vehículo, la marca del vehículo, la maniobra del vehículo, la edad del conductor, el horario y la clase de la carretera, subrayando la importancia de estos elementos en la seguridad vial.

En el contexto de la interpretabilidad de los modelos, la metodología SHAP (Shapley Additive Explanations), utilizada en estudios recientes como el de Zihang, Zhang y Das (2023), ha demostrado ser una herramienta poderosa para interpretar los resultados de los modelos de ML. En su estudio sobre la severidad de los accidentes en carreteras rurales de Texas, encontraron que XGBoost era el más efectivo en un conjunto de datos desbalanceado. Aplicando SHAP, lograron identificar que los factores relacionados con las condiciones meteorológicas tenían una contribución significativa en la severidad de los accidentes, mientras que la distribución de la velocidad tenía un impacto más fuerte en los accidentes graves.

Brecha de investigación en Ecuador

Sin embargo, a pesar de la abundancia de estudios sobre este tema en diversas regiones del mundo, existe una notable escasez de investigaciones en América Latina, y especialmente en Ecuador. Esta falta de estudios resalta la necesidad de desarrollar modelos de predicción adaptados al contexto local, donde factores específicos pueden influir en la severidad de los accidentes de manera distinta a otras regiones.

La investigación más similar a la presente se encuentra en la tesis de Argüello & Alcívar (2023), realizada en la Universidad ESPOL, que utilizó modelos de ML para analizar accidentes de tránsito en Guayaquil durante 2021 y 2022. Aunque el modelo SVM obtuvo la mayor precisión (95.22%), los autores optaron por Random Forest debido a su capacidad para manejar un gran número de predictores y la colinealidad entre variables. Este modelo alcanzó un performance del 94.87% y una tasa de error OOB del 3.19%, destacando su fiabilidad en la clasificación de la severidad de los accidentes. No obsante, la tesis se centra principalmente en la métrica de precisión, mientras que, en este contexto, es crucial minimizar los falsos negativos. Estos ocurren cuando un accidente con víctimas es clasificado erróneamente como sin víctimas, lo que podría afectar negativamente la asignación de recursos de emergencia y la gestión de la seguridad vial.

Los hallazgos mencionados anteriormente destacan la importancia de desarrollar modelos basados en inteligencia artificial para predecir la severidad de los accidentes de tránsito en Ecuador. Estos modelos permiten una gestión más eficiente de los accidentes, lo que contribuiría significativamente a mejorar la seguridad vial en el país. Además, subrayan la necesidad de continuar ampliando la investigación en este ámbito, con el fin

de optimizar las respuestas y los recursos disponibles para mitigar los efectos de los siniestros.

A pesar del avance global en la aplicación de Inteligencia Artificial para predecir la severidad de accidentes, Ecuador carece de estudios que implementen modelos avanzados con capacidad explicativa. Esta investigación aborda dicha brecha mediante algoritmos de Machine Learning optimizados para Quito y herramientas de explicabilidad como SHAP, permitiendo identificar los factores locales que influyen en la severidad de los siniestros. Los resultados sentarán las bases para mejorar la gestión de emergencias viales, desarrollar políticas de seguridad basadas en evidencia y estimular futuras investigaciones en este ámbito crítico para la ciudad.

Justificación

Este estudio tiene como objetivo desarrollar un modelo predictivo basado en inteligencia artificial para clasificar accidentes de tránsito según su severidad, utilizando información básica como características geográficas, variables temporales y causa probable. La importancia de este enfoque radica en que atiende un problema crítico para el cantón de Quito. La necesidad de contar con herramientas que permitan anticipar de forma precisa el nivel de gravedad de un accidente en el mismo momento en que ocurre. La relevancia del estudio se justifica en tanto que la severidad de los siniestros tiene consecuencias directas en la pérdida de vidas humanas y en el incremento de los costos sociales y económicos. Por lo tanto, disponer de un modelo predictivo que permita disminuir los tiempos de respuesta y optimizar el uso de los recursos representa un componente estratégico en la planificación y gestión de la seguridad vial en Quito.

El impacto de esta investigación se evidencia en el potencial de transformar los procesos de atención y gestión de accidentes. La implementación de un modelo de predicción de severidad no solo contribuiría a disminuir la mortalidad y las lesiones, sino que también mejoraría la eficiencia del sistema de emergencias al priorizar casos críticos en tiempo real. En el caso de Quito, ciudad con alta densidad vehicular y un parque automotor en crecimiento, este tipo de herramienta puede marcar una diferencia significativa en la capacidad de respuesta institucional. A nivel más amplio, el impacto se reflejaría en la posibilidad de aplicar políticas públicas más informadas y focalizadas, ya que los resultados del modelo servirían como evidencia empírica para orientar intervenciones preventivas y correctivas.

Más allá de la predicción inmediata, este estudio tiene una utilidad estratégica en la generación de conocimiento sobre los factores que determinan la severidad de los accidentes. Comprender cómo influyen variables como el tipo de siniestro, la hora del día o la causa probable permite no solo mejorar las respuestas inmediatas, sino también diseñar estrategias preventivas de largo plazo. En este sentido, la utilidad del modelo se extiende a ámbitos como la planificación urbana y la educación vial, al proporcionar información precisa que respalde la toma de decisiones. A nivel técnico, el empleo de algoritmos avanzados como Random Forest, XGBoost y LightGBM permite detectar patrones complejos que los métodos tradicionales difícilmente logran identificar, lo que

convierte a este estudio en una herramienta de apoyo tanto para la operación como para la formulación de políticas viales.

Los principales beneficiarios de esta investigación son los organismos encargados de la atención de emergencias, quienes podrán contar con un recurso que priorice los casos más graves y mejore la asignación de recursos. Los servicios de salud, los cuerpos de socorro y las autoridades de tránsito obtendrán ventajas directas en la eficiencia de sus operaciones. Al mismo tiempo, los ciudadanos se beneficiarán indirectamente de un sistema de respuesta más rápido y eficaz, que incremente las posibilidades de supervivencia y reduzca las secuelas de los accidentes graves. A nivel comunitario, los beneficios alcanzan a la sociedad en su conjunto, al disminuir los costos asociados a los siniestros como gastos médicos, daños materiales y pérdidas de productividad, y al fomentar un entorno urbano más seguro y resiliente.

La factibilidad del estudio se sustenta en la disponibilidad de datos oficiales de la Agencia Nacional de Tránsito, que proporcionan un volumen significativo de información histórica para el análisis. Estos datos, complementados con herramientas de software de libre acceso como Python y librerías especializadas en Machine Learning e interpretabilidad (SHAP), permiten llevar a cabo el procesamiento y modelado sin incurrir en altos costos. Además, investigaciones previas como las de Baykal et al. (2023) o Ahmed et al. (2023) demuestran que algoritmos como Random Forest o XGBoost alcanzan precisiones superiores al 80 % en la predicción de severidad, lo que respalda la viabilidad metodológica de esta propuesta. En consecuencia, la factibilidad técnica, académica y práctica está asegurada, y sus resultados pueden aplicarse no solo en Quito, sino también replicarse en otras ciudades del país que enfrentan problemáticas similares en la gestión de la movilidad y la seguridad vial.

Objetivos

Este capítulo presenta los objetivos que guiarán el desarrollo de modelos de Inteligencia Artificial para predecir la severidad de accidentes de tránsito en Quito. La investigación sigue un enfoque estructurado desde el análisis de datos hasta la interpretación de resultados, con el propósito de mejorar la gestión de la seguridad vial mediante predicciones precisas.

Objetivo General: Desarrollar y evaluar modelos de aprendizaje automático para la predicción binaria de la severidad de accidentes de tránsito en el cantón Quito.

Objetivos Específicos:

- Realizar un análisis exploratorio de los datos de la ANT para identificar patrones y tendencias clave que influyan en la severidad de los accidentes.
- Desarrollar y entrenar diversos modelos de inteligencia artificial para la predicción de la severidad de los accidentes de tránsito, utilizando técnicas avanzadas de inteligencia artificial.
- Comparar el desempeño de los modelos de aprendizaje automático desarrollados para la predicción binaria de la severidad de accidentes de tránsito en el cantón Quito, utilizando métricas de rendimiento adecuadas.
- Utilizar SHAP como técnica de Inteligencia Artificial para interpretar los modelos predictivos, identificando y analizando las variables más relevantes que influyen en la severidad de los accidentes de tránsito.

CAPÍTULO II
METODOLOGÍA

Este capítulo presenta el marco metodológico que guía la investigación, estructurado en siete apartados principales. En primer lugar, se delimita el área de estudio para contextualizar el alcance geográfico y temporal del análisis. A continuación, se describe el enfoque adoptado, el diseño del trabajo y el procedimiento para la obtención y el análisis de los datos. Finalmente, se formulan las hipótesis de contraste que orientan la validación estadística del modelo propuesto. En el Gráfico 3, se puede observar el mapa conceptual de la investigación, que ilustra de manera visual los componentes y flujos del proceso.

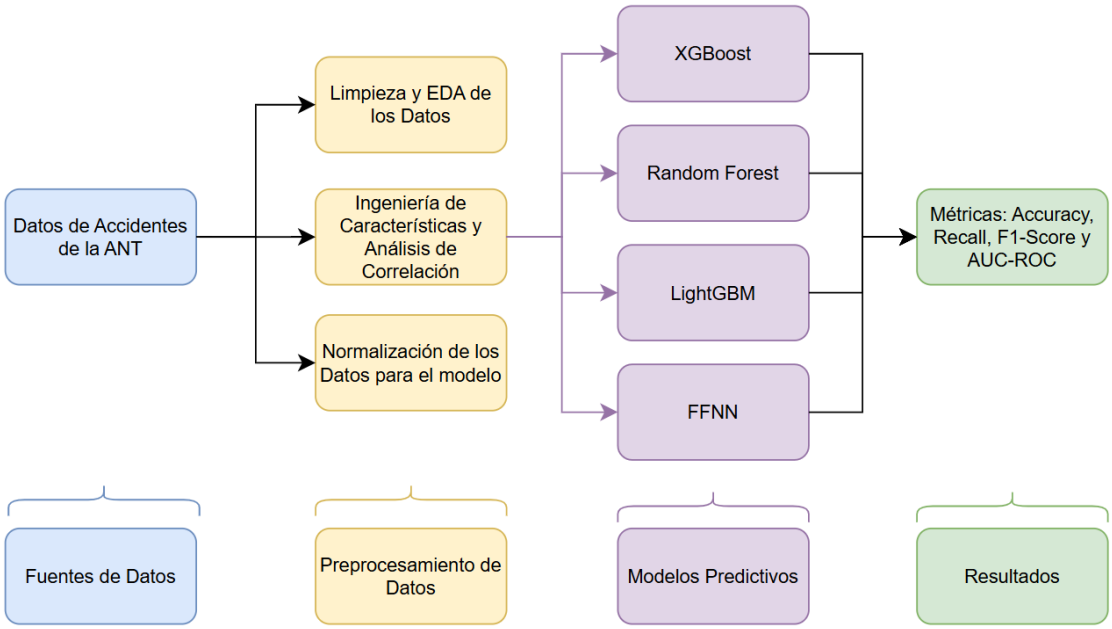


Gráfico No. 3: Mapa Conceptual
Elaborado por: Morillo, Jean (2025).

Área de Estudio

Este estudio se centra en los accidentes de tránsito en el cantón Quito, una de las ciudades más importantes de Ecuador. Quito cuenta con una infraestructura vial diversa, que incluye vías urbanas congestionadas y carreteras rurales, lo que genera desafíos significativos en seguridad vial. Dada su diversidad geográfica y socioeconómica, el análisis de los patrones de accidentes en todo el cantón permitirá identificar factores clave

que influyen en su severidad, proponiendo medidas integrales para mejorar la seguridad vial en áreas urbanas, rurales y periféricas.

El área de estudio se representa en un mapa de calor que muestra la distribución de los accidentes de tránsito en el cantón Quito. En la Imagen 1, las zonas con menor incidencia de accidentes se destacan en azul, mientras que aquellas con mayor concentración de siniestros aparecen en rojo. Se observa que la mayoría de los accidentes ocurren dentro de la parroquia de Quito, lo cual es coherente con la alta densidad vehicular y el flujo de tráfico en esta zona.

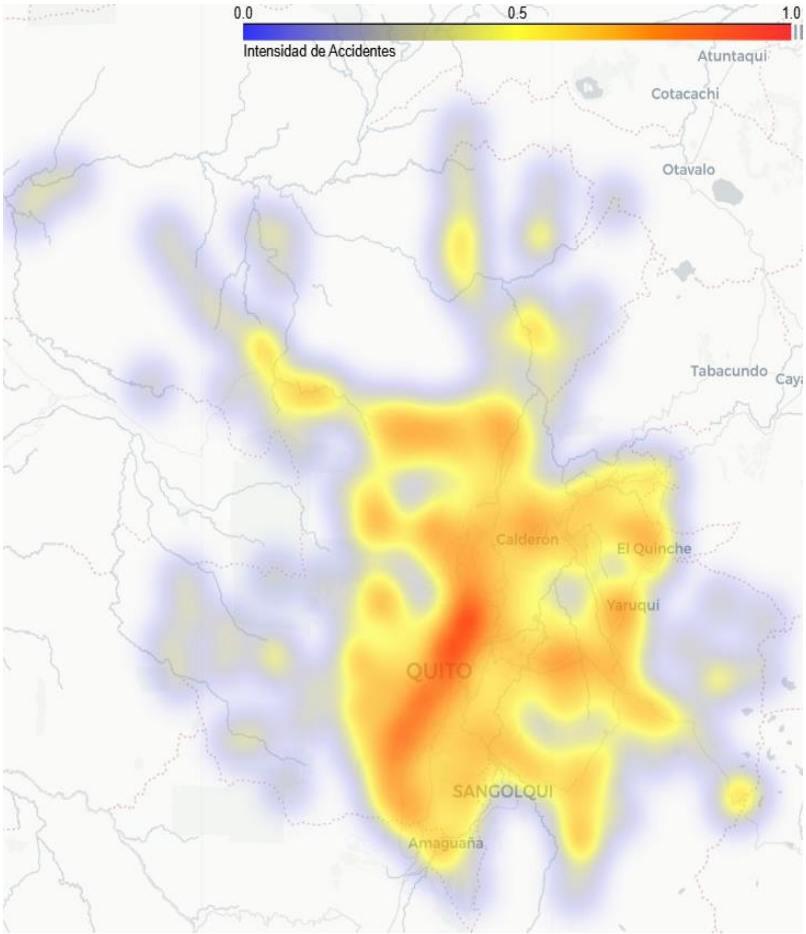
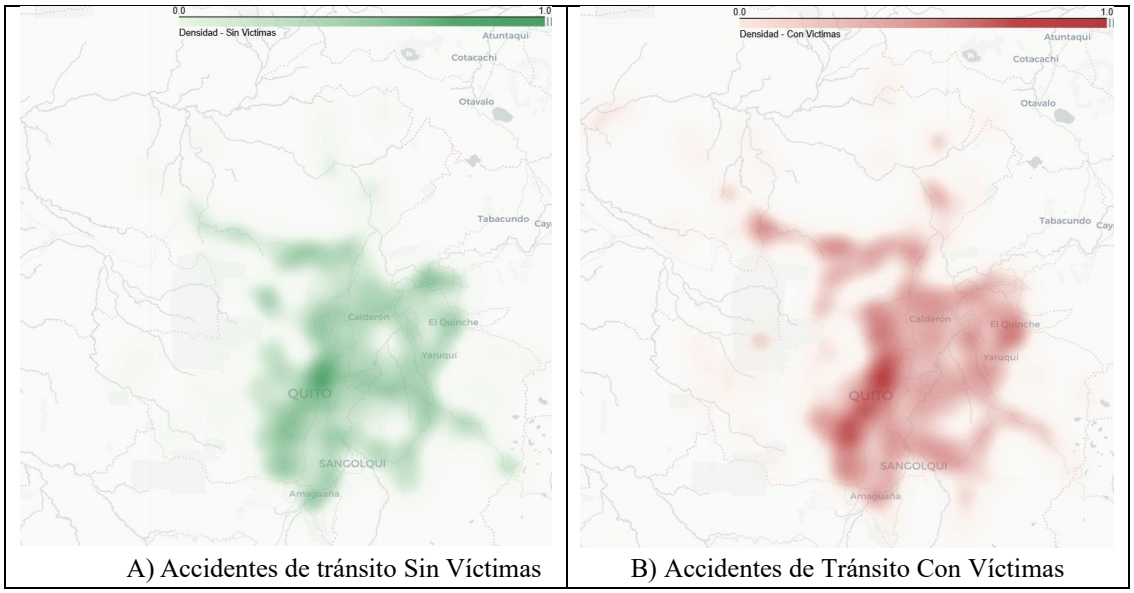


Imagen No. 1: Área de estudio: Distrito Metropolitano de Quito.
Elaborado por: Morillo, Jean (2025).

La Imagen 2 muestra la distribución de los accidentes de tránsito segmentados según su severidad, diferenciando entre accidentes con víctimas y sin víctimas. En este análisis se evidencia que los accidentes con víctimas tienden a concentrarse en el centro de la ciudad, lo que podría estar asociado a factores como la mayor densidad de tráfico, la

velocidad vehicular en ciertas vías y la interacción entre distintos tipos de usuarios de la vía.



A) Accidentes de tránsito Sin Víctimas | B) Accidentes de Tránsito Con Víctimas

Imagen No. 2: Mapa de calor de accidentes de tránsito por severidad.
Elaborado por: Morillo, Jean (2025).

Enfoque

Para alcanzar los objetivos de esta investigación, se adopta una perspectiva metodológica con predominio cuantitativo, complementada con una dimensión interpretativa propia del análisis cualitativo. Esta integración permite medir y validar estadísticamente el desempeño de los modelos predictivos, al mismo tiempo que se generan explicaciones comprensibles sobre los factores que influyen en sus resultados.

Enfoque Cuantitativo

La investigación se enmarca predominantemente en un enfoque cuantitativo, fundamentado en el análisis sistemático de un conjunto de datos estructurados de gran volumen que incluyen variables numéricas y categóricas obtenidas de la Agencia Nacional de Tránsito. Este enfoque se justifica por la naturaleza del objetivo principal: desarrollar y evaluar modelos predictivos mediante algoritmos matemáticos y estadísticos que permitan clasificar la severidad de accidentes de tránsito de manera objetiva y reproducible.

La metodología cuantitativa facilita el análisis de patrones complejos mediante técnicas de machine learning como Random Forest, XGBoost, LightGBM y redes neuronales, permitiendo la validación estadística de hipótesis y la evaluación del

desempeño a través de métricas estandarizadas como Precisión, Recall, F1-Score y AUC-ROC. La robustez del enfoque cuantitativo radica en su capacidad para identificar relaciones causales y predictivas entre variables, proporcionando evidencia empírica medible y generalizable sobre los factores que influyen en la severidad de los accidentes en el contexto específico del cantón Quito.

Enfoque Cualitativo

Complementariamente, la investigación incorpora elementos del enfoque cualitativo a través de la implementación de técnicas de Inteligencia Artificial Explicable (XAI), específicamente mediante el análisis SHAP (Shapley Additive Explanations). Esta dimensión cualitativa se justifica por la necesidad crítica de interpretar y comprender los procesos de toma de decisiones de los modelos de machine learning, transformando resultados numéricos en conocimiento contextualizado y comprensible para los tomadores de decisiones en seguridad vial.

El componente cualitativo permite analizar cómo cada variable contribuye a las predicciones, identificando patrones de comportamiento vial y factores de riesgo que van más allá de la simple cuantificación estadística. Esta perspectiva interpretativa es fundamental para garantizar transparencia y confiabilidad en los modelos predictivos, facilitando su aplicación práctica en la gestión de emergencias y en el diseño de políticas de seguridad vial.

Descripción de la metodología

La investigación se desarrolla bajo un diseño no experimental, transversal y correlacional, con un enfoque explicativo, orientado a identificar y analizar los factores que inciden en la severidad de los accidentes de tránsito en el cantón Quito. Este tipo de diseño resulta adecuado porque permite estudiar los hechos en su contexto natural, sin manipular las variables, y al mismo tiempo establecer relaciones de asociación y causalidad entre ellas, lo que se ajusta al objetivo de construir un modelo predictivo confiable.

Se emplean tres métodos de investigación complementarios. En primer lugar, el método bibliográfico–documental permitió revisar los antecedentes teóricos y empíricos disponibles en la literatura especializada, lo que sirvió para fundamentar la selección de variables y metodologías aplicadas. En segundo lugar, el método de campo se utilizó de

manera indirecta mediante el análisis de los registros históricos proporcionados por la Agencia Nacional de Tránsito (ANT), que constituyen la principal fuente de información empírica de este estudio. Finalmente, se aplicó un método experimental de carácter computacional, orientado a la construcción, entrenamiento y validación de modelos de machine learning, lo que permitió contrastar hipótesis y evaluar el desempeño predictivo bajo diferentes configuraciones algorítmicas.

La metodología integra técnicas de análisis de datos masivos y algoritmos de aprendizaje automático con procedimientos estadísticos de validación, lo que garantiza la solidez y la confiabilidad de los resultados. En términos operativos, el proceso comprendió varias fases interrelacionadas. En primer lugar, se efectuó una revisión documental que permitió identificar factores de riesgo relevantes y seleccionar las variables a analizar. Posteriormente, se procedió a la depuración y preparación de los datos, eliminando registros incompletos o inconsistentes y transformando las variables para hacerlas compatibles con los algoritmos predictivos. En esta misma etapa se realizó la operacionalización de las variables, definiendo como dependiente la severidad del accidente y como independientes un conjunto de factores geográficos, temporales, relacionados con el vehículo y con el conductor.

En la fase de modelado, se construyeron y entrenaron diferentes modelos predictivos empleando algoritmos como Random Forest, XGBoost, LightGBM y redes neuronales, con un proceso de optimización de hiperparámetros para maximizar el desempeño. La evaluación de los modelos se llevó a cabo mediante métricas estandarizadas de clasificación binaria, entre las que destacan recall, precisión, F1-score y AUC-ROC. Para reforzar la validez estadística de los resultados se aplicó el método bootstrap, que permitió estimar intervalos de confianza y realizar contrastes de hipótesis sobre la capacidad predictiva de los modelos. Finalmente, se incorporaron técnicas de Inteligencia Artificial Explicable, específicamente SHAP (Shapley Additive Explanations), con el propósito de interpretar la importancia relativa de las variables y facilitar la comprensión de los resultados para su aplicación en la gestión de la seguridad vial.

En conjunto, esta metodología proporciona un marco integral que combina sustento teórico, evidencia empírica y rigor estadístico, garantizando la pertinencia y confiabilidad de los hallazgos en el contexto de la predicción de la severidad de los accidentes de tránsito en Quito.

Diseño del trabajo

En esta sección se presenta el diseño del trabajo, centrado en la operacionalización de las variables utilizadas para el análisis de la severidad de los accidentes de tránsito en Quito. Se describe cómo se organizan y clasifican las variables dependientes e independientes, así como su definición, tipo de dato y relevancia dentro del estudio. Este capítulo proporciona un marco claro que facilita la comprensión de los datos y sienta las bases para el análisis estadístico y el modelado predictivo que se desarrolla en los capítulos posteriores.

Variable Dependiente

La variable dependiente en este estudio es la severidad del accidente, registrada en el dataset con cuatro categorías: Ileso, Lesionado, Fallecido y No Identificado. Esta variable es de tipo categórica, ya que clasifica los accidentes en función de las consecuencias humanas. Para simplificar el análisis y mejorar la efectividad del modelo, se decidió agrupar las categorías originales en dos grupos binarios: Con Víctimas (Lesionados y Fallecidos) y Sin Víctimas (Ilesos y No Identificados). Esta decisión se basó en varios factores descritos a continuación.

En primer lugar, estudios previos, demostraron que los modelos binarios lograron una mayor precisión, alcanzando un 95% en comparación con el 75% obtenido en clasificación multiclase (Kibria & Matin, 2022). Este resultado sugiere que la clasificación binaria puede ser más adecuada para el tipo de análisis que se busca realizar en este estudio.

Además, en el contexto específico de Quito, la respuesta ante accidentes de tránsito es similar tanto si hay personas lesionadas como fallecidas. Las acciones de las autoridades y servicios de emergencia no varían significativamente dependiendo de si las víctimas están heridas o han perdido la vida, lo que hace que agrupar estas categorías sea relevante para mejorar la calidad del modelo.

Por último, los registros clasificados como "No Identificados" se agruparon en la categoría Sin Víctimas, ya que generalmente se asume que estos accidentes involucran personas que huyen de la escena del accidente, lo cual normalmente deja accidentes sin consecuencias humanas graves, más allá de los daños materiales.

De esta manera, la variable dependiente final se organiza en dos categorías: Con Víctimas (Lesionados y Fallecidos) y Sin Víctimas (Ilesos y No Identificados). La figura

que se presenta a continuación permite apreciar de forma detallada cómo se distribuyen los accidentes según su severidad, mostrando tanto la clasificación original como la agrupación binaria adoptada en este estudio. Esta representación no solo facilita la visualización de la proporción relativa de cada categoría, sino que también sirve como referencia para comprender la estructura de los datos y la lógica detrás de la simplificación a dos grupos, lo cual resulta fundamental para la interpretación y el análisis posterior en el modelado predictivo.

De esta manera, la variable dependiente final se organiza en dos categorías: Con Víctimas (Lesionados y Fallecidos) y Sin Víctimas (Ilesos y No Identificados). El Gráfico 4 permite apreciar de forma detallada cómo se distribuyen los accidentes según su severidad, mostrando tanto la clasificación original como la agrupación binaria adoptada en este estudio. Esta representación no solo facilita la visualización de la proporción relativa de cada categoría, sino que también sirve como referencia para comprender la estructura de los datos y la lógica detrás de la simplificación a dos grupos, lo cual resulta fundamental para la interpretación y el análisis posterior en el modelado predictivo.

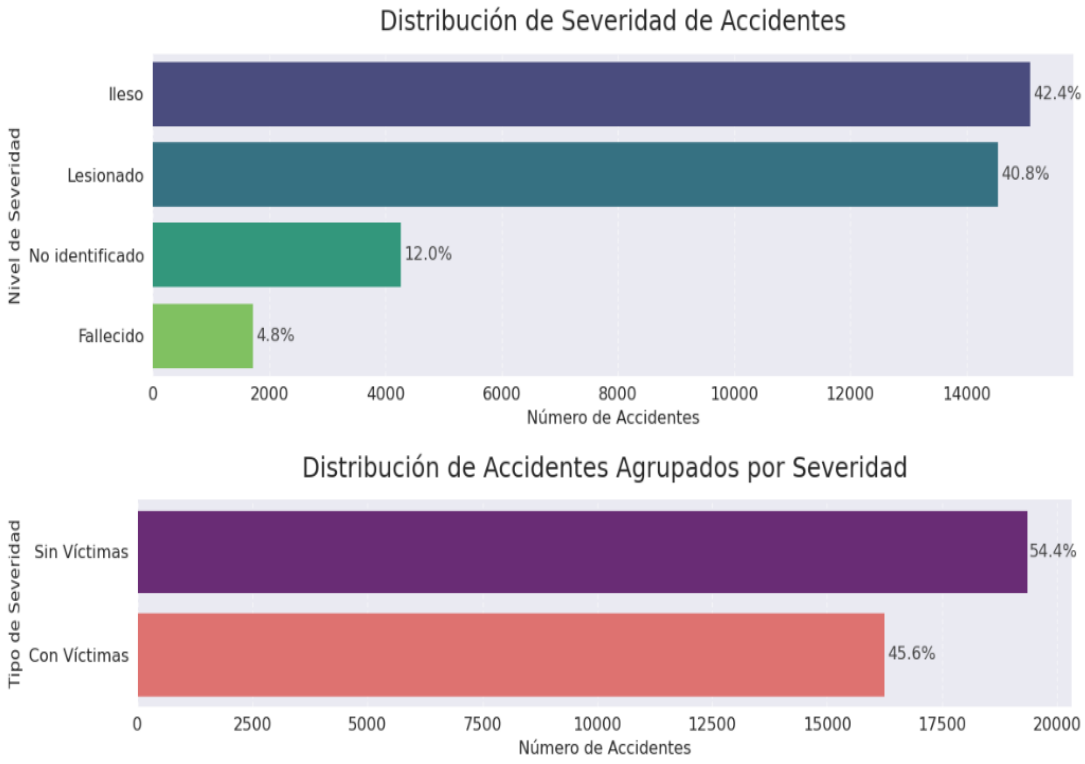


Gráfico No. 4: Distribución de la severidad de accidentes de tránsito.
Elaborado por: Morillo, Jean (2025).

Variables Independientes

Las variables independientes en la Tabla 1 fueron seleccionadas según su relevancia teórica y su disponibilidad en el conjunto de datos. Estas se clasifican en numéricas y categóricas, y se presentan a continuación en su forma original, sin aplicar transformaciones, agrupaciones ni técnicas de preprocesamiento. Esta presentación permite una comprensión clara de las características brutas del conjunto de datos antes de su utilización en el modelado predictivo.

Tabla No. 1: Diccionarios de variables independientes.

Variable	Tipo de Dato	Descripción
Año	Numérica	Año en el que ocurrió el accidente.
Latitud	Numérica	Coordenada de latitud del lugar del accidente.
Longitud	Numérica	Coordenada de longitud del lugar del accidente.
Parroquia	Categórica	Parroquia en la que ocurrió el accidente.
Zona	Categórica	Zona geográfica donde ocurrió el accidente.
Nombre de la Vía	Categórica	Nombre de la calle o vía donde ocurrió el accidente.
Hora	Numérica	Hora del día en la que ocurrió el accidente
Día de la Semana	Categórica	Día de la semana en el que ocurrió el accidente (ej. Lunes, Martes).
Mes	Categórica	Mes del año en el que ocurrió el accidente (ej. Enero, Febrero).
Feriado	Categórica	Indica si el accidente ocurrió en un día feriado (Sí/No).
Causa Probable	Categórica	Causa probable del accidente (ej. exceso de velocidad, distracción).
Tipo de Accidente	Categórica	Tipo de accidente o siniestro (ej. colisión, atropello).
Tipo de Vehículo	Categórica	Tipo de vehículo involucrado (ej. automóvil, motocicleta).
Servicio del Vehículo	Categórica	Tipo de servicio del vehículo (ej. particular, público).
Número de Vehículos	Numérica	Número total de vehículos involucrados en el accidente.
Edad del Conductor	Numérica	Edad del conductor o persona involucrada en el accidente.
Género del Conductor	Categórica	Género del conductor (ej. Masculino, Femenino).
Uso de Casco	Categórica	Indica si el conductor o persona involucrada usaba casco (Sí/No).
Uso de Cinturón	Categórica	Indica si el conductor usaba cinturón de seguridad (Sí/No).

Elaborado por: Morillo, Jean (2025).

Procedimiento para obtención y análisis de datos

El proceso de obtención y análisis de los datos se llevó a cabo de manera sistemática, siguiendo un flujo de trabajo diseñado para garantizar la calidad, consistencia y relevancia de la información utilizada. Los registros históricos de accidentes de tránsito fueron

descargados desde la página web de la Agencia Nacional de Tránsito (ANT) en formato Excel, abarcando el período comprendido entre enero de 2017 y abril de 2024. Posteriormente, se aplicó un filtro geográfico para centrarse únicamente en los incidentes ocurridos dentro del cantón Quito, descartando registros fuera de esta jurisdicción.

La limpieza y preparación de los datos se realizó utilizando Python en Jupyter Notebooks, lo que permitió integrar el código, los resultados y las visualizaciones en un único entorno reproducible. Durante esta etapa, se eliminaron registros incompletos o inconsistentes, se corrigieron errores de formato y se descartaron variables no relevantes para los objetivos predictivos del estudio. Las variables temporales se transformaron mediante codificación cíclica, y las categóricas se adaptaron a un formato compatible con algoritmos de aprendizaje automático mediante one-hot encoding.

El análisis exploratorio de datos incluyó la generación de gráficos y estadísticos descriptivos para identificar patrones, tendencias y valores atípicos, proporcionando una visión detallada de la distribución de los accidentes según severidad, ubicación, tipo de vehículo y otras características. Esta fase fue fundamental para seleccionar las variables independientes más relevantes y estructurar los conjuntos de datos de entrenamiento y prueba mediante muestreo estratificado, asegurando la preservación de la proporción de casos en cada categoría de severidad.

El modelado predictivo se implementó directamente en Jupyter Notebooks, utilizando algoritmos como Random Forest, XGBoost, LightGBM y redes neuronales feedforward, con optimización de hiperparámetros y evaluación mediante métricas de clasificación binaria (recall, precisión, F1-score y AUC-ROC). Para complementar el análisis, se aplicó SHAP (Shapley Additive Explanations), lo que permitió interpretar el impacto relativo de cada variable en las predicciones. Además, se empleó el método bootstrap para estimar intervalos de confianza y contrastar hipótesis sobre el desempeño de los modelos de manera robusta.

Población y muestra

El conjunto de datos utilizado en esta investigación fue obtenido directamente de la página web de la Agencia Nacional de Tránsito (ANT) de Ecuador, y consiste en un archivo Excel que registra 166.684 accidentes de tránsito ocurridos entre el 1 de enero de 2017 y el 30 de abril de 2024. Inicialmente, el conjunto de datos contenía 56 variables

descriptivas, incluyendo ubicación geográfica, fechas, modalidades de accidentes y consecuencias humanas (lesionados, fallecidos).

Dado que el estudio se centra exclusivamente en el cantón Quito, se aplicó un filtro geográfico para eliminar registros fuera de esta jurisdicción, reduciendo el conjunto de datos a 35.632 registros. Esta selección permite enfocar el análisis en un contexto específico y garantizar que los resultados sean representativos de la realidad local, considerando las particularidades viales, demográficas y de movilidad de la ciudad.

Adicionalmente, se descartaron columnas redundantes o irrelevantes para los objetivos del modelo predictivo, tales como identificadores administrativos (ID de siniestro, provincia, cantón) y variables duplicadas (por ejemplo, días de la semana expresados tanto en texto como en números). Esta depuración asegura que únicamente se consideren variables pertinentes, optimizando el procesamiento y la capacidad predictiva de los modelos de inteligencia artificial.

De esta manera, la población analizada se ajusta tanto a los criterios metodológicos como a los objetivos del estudio, garantizando que los patrones identificados sean aplicables a la gestión de la seguridad vial y a la priorización de recursos de emergencia en Quito.

Hipótesis

El objetivo de esta investigación es evaluar si un modelo de inteligencia artificial entrenado con datos de la Agencia Nacional de Tránsito (ANT) puede predecir de forma binaria la severidad de un accidente de tránsito en el cantón Quito, clasificando cada registro como Con Víctimas o Sin Víctimas, con un nivel de sensibilidad suficiente para apoyar la priorización de recursos de emergencia, por lo tanto:

- Hipótesis alternativa: Los modelos de aprendizaje automático alcanzan un recall poblacional igual o superior a 0,85 en la clasificación binaria de la severidad de los accidentes de tránsito (H_1: Recall $\geq 0,85$).
- Hipótesis nula: Los modelos de aprendizaje automático alcanzan un recall poblacional menor que 0,85 en la clasificación binaria de la severidad de los accidentes de tránsito (H_0: Recall $< 0,85$).

Fundamentación

La hipótesis se sustenta en los siguientes argumentos:

Relevancia de la métrica

En la gestión de emergencias viales, la métrica crítica es el recall (sensibilidad), porque permite minimizar los falsos negativos, es decir, evitar que un accidente con víctimas sea clasificado como "sin víctimas", lo que retrasaría la movilización de recursos de auxilio. La fundamentación completa de este criterio y de la importancia de minimizar los falsos negativos se desarrolla en el siguiente capítulo.

Respaldo en la literatura

Los estudios previos analizados en el capítulo anterior demuestran que los modelos de machine learning pueden alcanzar niveles de recall entre 0,79 y 0,89, valores que respaldan el umbral de 0,85 propuesto.

- Adefabi et al. (2023) obtuvieron un recall del 79,2 % en un área metropolitana.
- Ahmed et al. (2023) reportaron un recall del 81,4 % en Nueva Zelanda.
- Baykal et al. (2023) alcanzaron rangos de 0,84–0,88 en un estudio en 49 estados de EE. UU.
- Wu et al. (2024) lograron 0,86 con LightGBM en la provincia de Jiangsu, China.

Estos resultados confirman que un recall $\geq 0{,}85$ es exigente pero realista para la predicción de la severidad de accidentes de tránsito.

Pertinencia para el contexto ecuatoriano

Tal como se argumentó en la "Brecha de investigación en Ecuador", la ausencia de estudios locales que combinen modelos avanzados y herramientas de interpretabilidad justifica evaluar si es posible replicar dichos niveles de desempeño con datos del cantón Quito, considerando factores propios de esta área.

Metodología de contrastación

El desempeño se evaluará sobre el conjunto de prueba estratificado (30 %), calculando el recall y su intervalo de confianza del 95 % mediante bootstrap de 1000 iteraciones. Se aplicará una prueba unilateral de hipótesis ($p < 0{,}05$) para contrastar H_0: Recall $< 0{,}85$ vs. H_1: Recall $\geq 0{,}85$, garantizando una verificación estadística objetiva y reproducible.

CAPÍTULO III

DESARROLLO DE LA INVESTIGACIÓN

Tratamiento y Detección de Outliers

Durante el análisis de los datos realizado con Python y librerías como Pandas y Seaborn, se identificaron valores atípicos en la variable Lesionados (ver apéndice B). Tras evaluar la distribución, se decidió eliminar los registros con más de 10 personas lesionadas por accidente, lo que representó solo el 0,07% del total. Esta decisión se fundamenta en la distribución empírica de los datos, donde los accidentes con más de 10 lesionados representan solo el 0,07% del total, siendo eventos atípicos que podrían sesgar los modelos predictivos hacia patrones no generalizables. Asimismo, variables como Lesionados y Fallecidos se excluyeron de los predictores del modelo, al ser resultados del accidente y no factores predecibles en tiempo real, reservándose únicamente para análisis descriptivo que se presentará en la siguiente sección del reporte.

El Gráfico 5 muestra la distribución de las principales variables numéricas seleccionadas, considerando únicamente los accidentes con más de 10 lesionados. Se observa que algunas variables, como lesionados, latitud, longitud y edad, aún presentan valores que se encuentran fuera del rango intercuartílico, señalando la presencia de posibles valores atípicos que podrían influir en el análisis posterior. Sin embargo, estos valores no fueron eliminados, ya que representan casos reales dentro del comportamiento esperado de los accidentes de tránsito y no constituyen errores en los datos. La decisión de conservar estos valores se basa en la necesidad de mantener la variabilidad inherente a los datos y evitar la pérdida de información relevante para el análisis y la construcción del modelo predictivo. Además, al haber eliminado únicamente los accidentes con más de 10 lesionados, se garantiza que los modelos no se vean influenciados por eventos extremadamente raros.

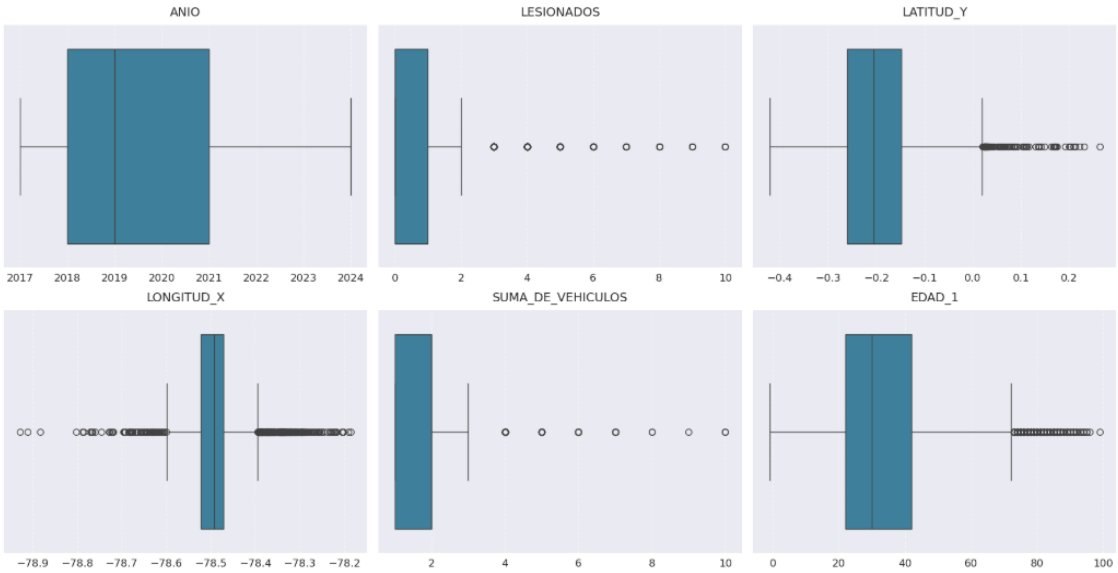


Gráfico No. 5: Distribución de variables numéricas seleccionadas.
Elaborado por: Morillo, Jean (2025).

El tratamiento de outliers aplicado impactó positivamente la calidad del conjunto de datos. La eliminación del 0,07% de registros con valores extremos redujo significativamente el ruido estadístico, mejorando la estabilidad de los modelos predictivos sin comprometer la representatividad de los datos. Este enfoque equilibrado permite conservar la variabilidad natural de los accidentes de tránsito mientras se eliminan casos anómalos que podrían distorsionar los patrones identificados por los algoritmos. Los análisis preliminares mostraron que los modelos entrenados con el conjunto de datos depurado exhibieron menor varianza y mayor capacidad de generalización, aspectos fundamentales para desarrollar predicciones confiables sobre la severidad de futuros accidentes en el contexto específico de Quito.

Análisis descriptivo de las variables

Esta sección presenta un análisis visual y analítico de algunas variables clave del conjunto de datos, con el objetivo de comprender mejor el contexto en el que ocurren los accidentes de tránsito. A través de la exploración de variables temporales y categóricas, se identifican patrones y tendencias que permitirán un mejor entendimiento de los factores asociados a la siniestralidad. Estos hallazgos servirán como base para los análisis posteriores y la construcción de los modelos predictivos.

Variables Temporales

En esta sección, se examinan las variables relacionadas con el año, mes, día y hora. El Gráfico 6 ilustra esta información en tres subgráficos:

A) Accidentes por año

El subgráfico superior muestra la distribución de accidentes de tránsito por año. Se observa una disminución notable en el número de accidentes entre 2020 y 2023. Esta tendencia es atribuible a la pandemia global de COVID-19, que llevó a la implementación de restricciones de movilidad y confinamientos, reduciendo así la circulación vehicular durante este período.

Además, en 2024 se registran menos accidentes, lo cual se debe a que el conjunto de datos solo incluye información hasta el 30 de abril de 2024. Adicionalmente, se observa que la mayoría de los accidentes de tránsito no resultan en víctimas fatales.

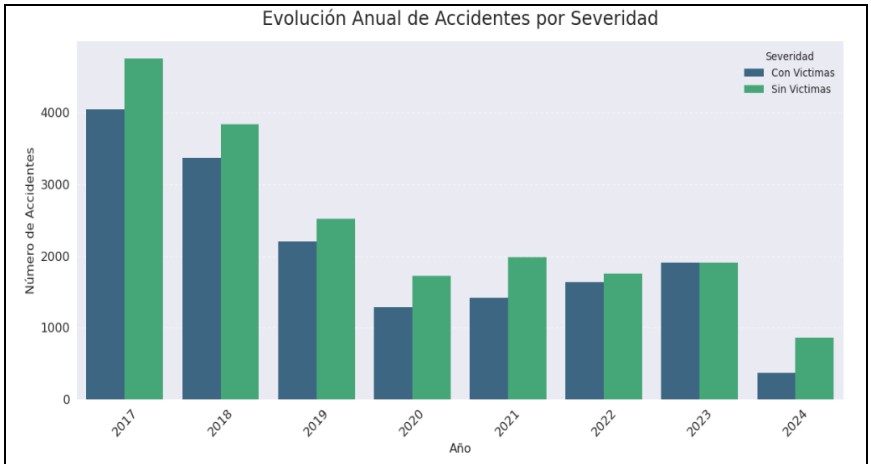
B) Accidentes por mes

El subgráfico central presenta la distribución de accidentes por mes. No se aprecia una variación significativa en el número de accidentes entre los distintos meses, lo que sugiere que la variable "mes" podría no ser relevante en el análisis. Esta hipótesis se evaluará en detalle en la sección de resultados.

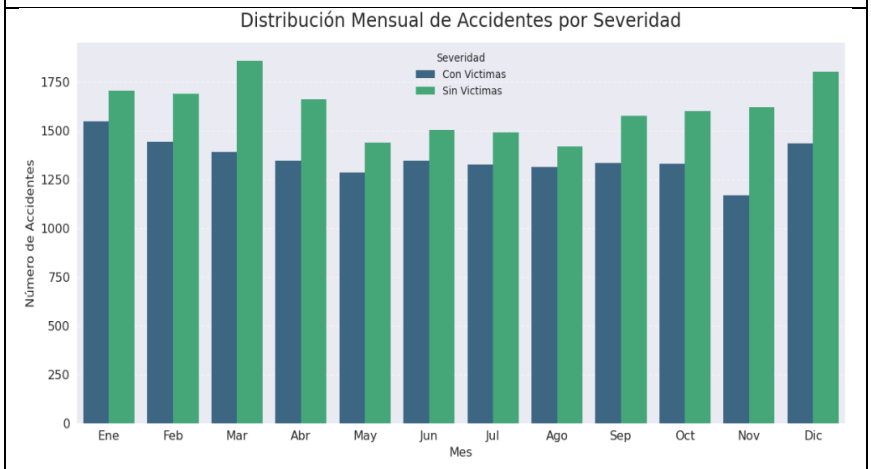
C) Accidentes por día

El subgráfico inferior muestra la distribución de accidentes por día de la semana. Se identifica un patrón cíclico de siete días, donde el número de accidentes aumenta progresivamente a lo largo de la semana, alcanzando su pico los sábados y disminuyendo los lunes. Este comportamiento es consistente con datos reportados por la AMT, donde el sábado es el día con mayor cantidad de accidentes, seguido del viernes y el domingo (Escobar, 2024).

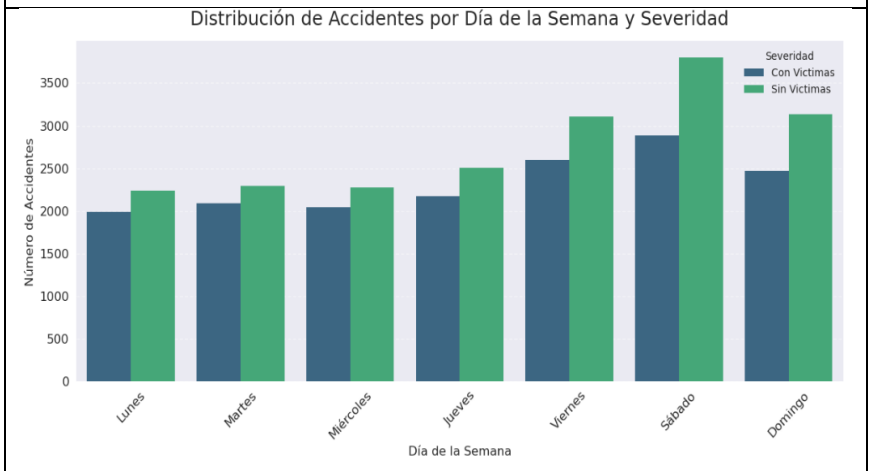
Este análisis preliminar proporciona una visión general de cómo las variables temporales influyen en la ocurrencia de accidentes de tránsito y sienta las bases para análisis más profundos en las secciones posteriores.



A) Accidentes por año



B) Accidentes por mes



C) Accidentes por día

Gráfico No. 6: Distribución temporal de los accidentes de tránsito.
Elaborado por: Morillo, Jean (2025).

La hora del día es otra variable clave en la ocurrencia de accidentes de tránsito. Con mayor circulación vehicular, aumenta la probabilidad de que ocurra un accidente, lo que

se refleja en los datos analizados. Se observa un incremento en los accidentes durante las horas pico, especialmente entre las 6:00 am a 8:00 am y de 5:00 pm a 8:00 pm, lo cual coincide con los picos mostrados en el Gráfico 7.

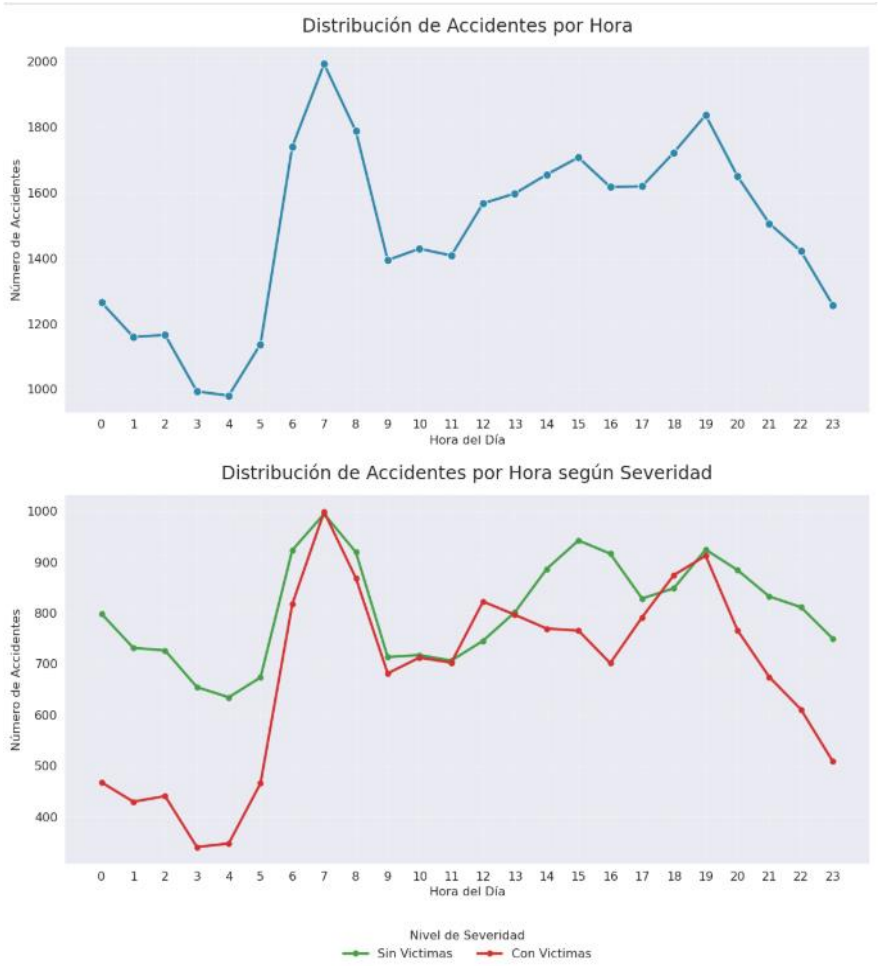


Gráfico No. 7: Distribución de accidentes por hora y severidad.
Elaborado por: Morillo, Jean (2025).

A pesar de la implementación de la normativa "Pico y Placa", que busca reducir la congestión en estos horarios, sigue siendo evidente que los mayores números de accidentes ocurren durante estas franjas. Esto sugiere que, aunque se han tomado medidas para disminuir el tráfico, la alta densidad vehicular en estos momentos sigue siendo un factor determinante en la siniestralidad. Finalmente, se puede observar que los accidentes que ocurren en la noche-madrugada (de 11:00 pm a 5:00 am) son más probables a ser accidentes sin víctimas.

Variables categóricas relevantes

En esta sección, se analizan las variables categóricas más relevantes del conjunto de datos para entender su influencia en los accidentes de tránsito. Se consideran especialmente las categorías relacionadas con los vehículos involucrados, las principales causas probables y los tipos de accidentes. El Gráfico 8 presenta esta información en tres subgráficos, lo que permite observar con claridad cómo cada una de estas variables influye en la ocurrencia de los siniestros.

A) Vehículos involucrados

El análisis de los tipos de vehículos involucrados en accidentes de tránsito revela patrones significativos en la gravedad de los siniestros. Los automóviles encabezan la lista de vehículos más involucrados en accidentes; sin embargo, la mayoría de estos incidentes resultan en accidentes sin víctimas. En contraste, las motocicletas, aunque ocupan el segundo lugar en frecuencia de involucramiento, presentan una mayor probabilidad de resultar en accidentes con víctimas. Esto se debe a que las motocicletas ofrecen menos protección en caso de colisión, lo que incrementa la severidad de las lesiones en los conductores y pasajeros (Santos, Firme, Dias, & Amado, 2023).

Por otro lado, los accidentes que involucran bicicletas son menos frecuentes en comparación con los automóviles y las motocicletas. No obstante, es notable que una proporción significativa de estos incidentes también resulta en víctimas, a pesar de la menor frecuencia de ocurrencia. Esto puede atribuirse a la vulnerabilidad inherente de los ciclistas, quienes carecen de la protección estructural que ofrecen otros vehículos, lo que aumenta el riesgo de lesiones graves en caso de accidente (Liasidis, Benjamin, Jakob, Lewis, & Demetriades, 2023).

B) Principales causas

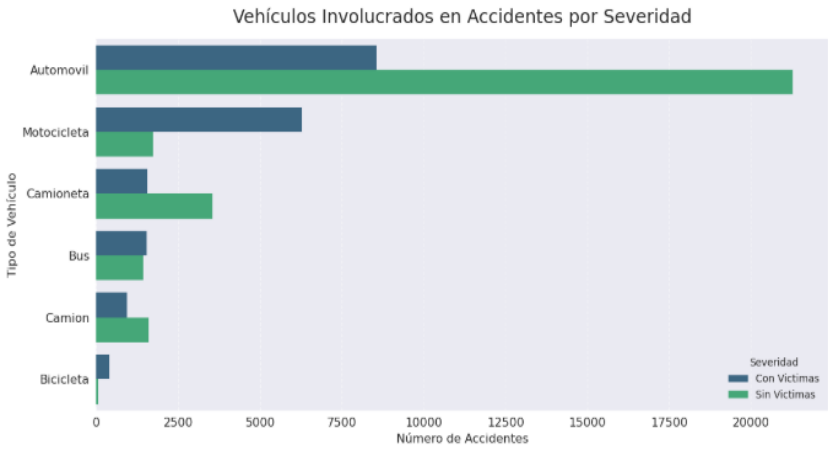
El análisis de las principales causas de accidentes de tránsito confirma que el exceso de velocidad es la más frecuente. Este resultado coincide con múltiples estudios que destacan la velocidad excesiva como un factor crítico en la ocurrencia de accidentes. Por ejemplo, Nasiri et al. (2019) analizaron datos de Sistan, Baluchestán e Irán, concluyendo que la velocidad no solo es un factor determinante en la ocurrencia del accidente, sino que también influye en su gravedad, ya que una mayor velocidad aumenta la probabilidad de víctimas severas.

Además, se encuentran otras causas relevantes, como no respetar las señales de tránsito y la conducción bajo los efectos de alcohol o estupefacientes. Sin embargo, una causa que destaca en la ocurrencia de accidentes con víctimas es la de no ceder el paso a peatones. Este comportamiento debe ser considerado por las autoridades pertinentes para implementar medidas que puedan garantizar la seguridad de los peatones.

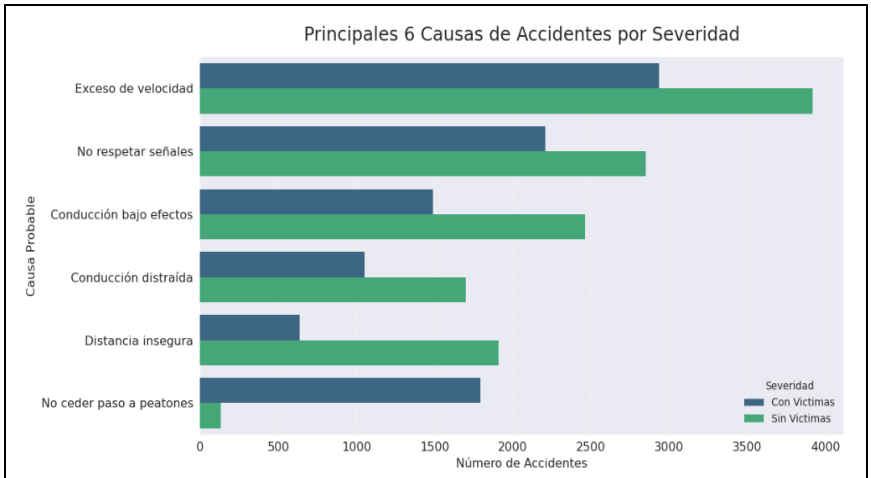
C) Accidentes más comunes

El análisis de los tipos de accidentes de tránsito revela que algunos tipos de siniestros son significativamente más frecuentes y presentan diferencias en la severidad de sus consecuencias. Entre los seis tipos de accidentes más comunes, el choque lateral y los estrellamientos son los de mayor ocurrencia, con un alto número de incidentes tanto con víctimas como sin víctimas (ver apéndice A).

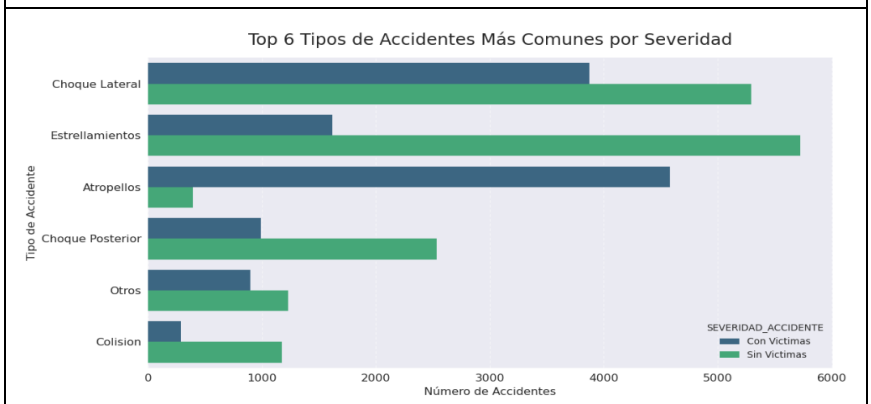
Los atropellos son los accidentes más propensos a resultar en víctimas, lo cual es coherente con su naturaleza, ya que implican el impacto directo de un vehículo contra una persona. A diferencia de otros tipos de accidentes, en los atropellos los peatones no cuentan con ninguna estructura de protección, lo que incrementa significativamente el riesgo de lesiones graves o fatales. La vulnerabilidad de los peatones ante estos incidentes resalta la necesidad de implementar medidas de seguridad vial, como el refuerzo de pasos peatonales, señalización clara y campañas de concienciación para conductores y transeúntes.



A) Vehículos involucrados en accidentes



B) Principales causas de accidentes



C) Tipos de accidents más comunes

Gráfico No. 8: Variables categóricas relevantes.
Elaborado por: Morillo, Jean (2025).

Ingeniería de Características

La ingeniería de características es un proceso esencial en la construcción de modelos de aprendizaje automático, ya que permite transformar los datos brutos en representaciones más informativas y adecuadas para el análisis predictivo (Zheng & Casari, 2018). En este estudio, se aplicaron diversas técnicas de preprocesamiento y codificación de variables para optimizar la predicción de la severidad de accidentes de tránsito en Quito.

Transformación de variables temporales

Las variables Hora, Día de la Semana y Mes fueron transformadas mediante codificación cíclica, utilizando funciones trigonométricas seno y coseno para capturar su naturaleza circular. Esta técnica evita discontinuidades artificiales en los modelos, como la consideración errónea de las 23:59 y 00:01 como valores distantes, y mejora la capacidad predictiva al reflejar adecuadamente patrones temporales repetitivos. Estudios recientes destacan que la codificación cíclica es esencial para modelar variables temporales en análisis predictivos de accidentes, ya que preserva relaciones intrínsecas como la estacionalidad horaria o mensual (Mahajan, Singh, & Bruns, 2021).

Adicionalmente, a partir del análisis exploratorio de datos, se observó un incremento significativo de accidentes en horas pico, es decir, en franjas horarias donde el tráfico vehicular es más denso. Por ello, se generó una variable binaria que identifica si el accidente ocurrió durante estos periodos, lo que permite capturar patrones de riesgo asociados a la congestión vial.

Creación de Variables Derivadas

Vehículo Pesado: A partir de la variable Tipo de Vehículo, se generó la variable binaria Vehículo Pesado, definida como la presencia de buses, camiones, camionetas o furgonetas en el accidente. Esta agrupación se fundamenta en investigaciones actuales que vinculan a los vehículos pesados con un riesgo significativamente mayor de accidentes graves debido a su masa, inercia y menor capacidad de frenado en comparación con vehículos ligeros (WHO, 2022).

Agrupación de Edades: La variable Edad se categorizó en tres grupos: <18, 18-35, 35-65 y >65, con el objetivo de reflejar diferencias en comportamientos y riesgos entre conductores jóvenes, adultos y mayores. Esta segmentación se alinea con enfoques recientes en seguridad vial, donde la edad se analiza como un factor no lineal con impactos diferenciados en la severidad de accidentes (Islam et al., 2022). Además, la edad se separó en grupos porque la finalidad del modelo es predecir la severidad utilizando variables fácilmente visibles en el momento del accidente. Mientras que predecir la edad exacta de una persona es complicado, identificar un grupo de edad es más factible y aporta información útil para el análisis.

Codificación de Variables Categóricas

Todas las variables categóricas fueron transformadas mediante one-hot encoding, convirtiendo cada categoría en una variable binaria independiente. Esta técnica evita asignar un orden arbitrario a categorías nominales (ej. Parroquia o Tipo de Accidente), lo que podría sesgar los modelos predictivos. Investigaciones recientes respaldan su uso en estudios de accidentes de tráfico, ya que facilita la interpretación de coeficientes en modelos de clasificación sin introducir jerarquías ficticias (Brownlee, 2022).

Análisis de Correlación

Para evaluar la correlación entre las variables, se seleccionaron todas las columnas numéricas del conjunto de datos y se calculó su matriz de correlación. Se estableció un umbral de 0.7 para considerar una correlación como alta y determinar posibles redundancias entre las variables. El análisis reveló varios pares de variables altamente correlacionadas, incluyendo:

Tipo de Vehículo: Se encontró una alta correlación entre la columna general "tipo de vehículo" y las variables individuales que representan la presencia de cada tipo de vehículo en el accidente (automóvil, bus, camión, motocicleta, etc.). Debido a que las variables individuales proporcionan información más detallada, como el número total de vehículos involucrados de cada tipo, se decidió eliminar la columna general "tipo de vehículo" para evitar redundancias y mejorar la interpretabilidad del modelo.

Variables de Tiempo: Se encontró una fuerte correlación entre las representaciones normales del tiempo y sus transformaciones trigonométricas:

- Hora vs. Hora seno
- Mes vs. Mes seno
- Día de la semana vs. Día de la semana seno

Como se explicó en la sección anterior, las transformaciones seno ofrecen una mejor representación de los patrones cíclicos del tiempo, por lo que se optó por eliminar las columnas originales (hora, mes y día de la semana) y conservar las variables cíclicas para capturar de mejor manera las tendencias temporales.

División de Datos

Para la división del conjunto de datos en este estudio, se empleó la técnica de muestreo estratificado con el objetivo de preservar la distribución de la variable objetivo en ambas particiones. Se optó por dividir los datos en un 70% para el conjunto de entrenamiento y un 30% para el conjunto de prueba, siguiendo la metodología adoptada en la mayoría de los estudios revisados sobre predicción de la severidad de accidentes de tránsito. Esta proporción permite entrenar el modelo con una cantidad suficiente de datos mientras se reserva una muestra representativa para evaluar su desempeño en datos no vistos, asegurando una estimación confiable de su capacidad predictiva.

Si bien la división 80-20 es también común en problemas de machine learning, la elección de una proporción 70-30 responde a características específicas del conjunto de datos de accidentes de tránsito en el cantón Quito. Primero, el desbalance natural entre accidentes con y sin víctimas requiere un conjunto de prueba más amplio para asegurar una representación adecuada de la clase minoritaria. Segundo, la alta variabilidad en los tipos y circunstancias de accidentes viales hace necesaria una evaluación más robusta, que se beneficia de un mayor volumen de datos de prueba.

Por lo tanto, la división 70-30 ofrece un balance óptimo entre la capacidad del modelo para aprender patrones relevantes y la confiabilidad de las métricas de evaluación, especialmente en términos de sensibilidad y especificidad, cruciales para minimizar los falsos negativos en la clasificación de accidentes graves.

Selección de Modelos de Machine Learning

Para la predicción binaria de la severidad de los accidentes de tránsito, se seleccionaron cuatro modelos de aprendizaje automático con enfoques complementarios: Random Forest, XGBoost, LightGBM y una Red Neuronal Feedforward (FFNN). La elección de estos modelos se fundamenta en un exhaustivo análisis de la literatura, en el cual se identificó que estos algoritmos han reportado consistentemente altos niveles de precisión, recall y AUC-ROC en estudios relacionados con la predicción de eventos viales. Además, presentan una capacidad comprobada para manejar datos tabulares, capturar relaciones no lineales y ofrecer interpretabilidad mediante técnicas como la importancia de características o valores SHAP.

Modelos como Support Vector Machines (SVM) y regresión logística también fueron considerados, ya que han sido tradicionalmente utilizados en problemas de clasificación binaria. Sin embargo, se decidió descartarlos debido a ciertas limitaciones: la regresión logística, aunque útil por su simplicidad e interpretabilidad, tiende a tener un bajo desempeño en presencia de relaciones no lineales complejas. Por su parte, el modelo SVM mostró problemas de escalabilidad y tiempos de entrenamiento elevados al trabajar con conjuntos de datos grandes, como los utilizados en este estudio. Estas limitaciones ya han sido documentadas en trabajos como el de Amini et al. (2022) y Qi et al. (2021), donde los modelos basados en árboles y redes neuronales superaron a SVM y regresión en contextos similares.

XGBoost (Extreme Gradient Boosting)

Este algoritmo de gradient boosting ha demostrado superioridad en la clasificación de severidad de accidentes, particularmente en entornos urbanos con relaciones no lineales entre variables (Ahmed et al., 2023). XGBoost optimiza una función de pérdida mediante el ensamblado secuencial de árboles de decisión, minimizando el error residual en cada iteración. Su fortaleza radica en la regularización incorporada y la gestión eficiente del sobreajuste, lo que mejora la generalización del modelo en tareas de clasificación (ídem).

En el presente estudio, XGBoost será evaluado tanto en su configuración estándar como en una versión optimizada mediante la selección de hiperparámetros. Inicialmente, el modelo se entrenará con sus valores predeterminados para establecer una línea base de rendimiento. Posteriormente, se implementará un ajuste de hiperparámetros utilizando GridSearchCV, explorando combinaciones de n_estimators, learning_rate, max_depth, min_child_weight, subsample y colsample_bytree. Este proceso permitirá identificar la mejor configuración para mejorar la capacidad predictiva del modelo, optimizando su balance entre sesgo y varianza, y evaluando su impacto en la clasificación de la severidad de los accidentes.

Random Forest

Matemáticamente, Random Forest construye múltiples árboles de decisión a partir de diferentes subconjuntos de los datos de entrenamiento. La predicción final se obtiene mediante la agregación de las predicciones individuales de cada árbol, generalmente a través de una votación mayoritaria en problemas de clasificación o promediando las

salidas en problemas de regresión (Breiman, 2001). Esta técnica mejora la precisión y controla el sobreajuste al reducir la varianza de las predicciones.

Este algoritmo de aprendizaje automático ha sido ampliamente utilizado en la predicción de la severidad de accidentes de tráfico debido a su capacidad para manejar datos complejos y reducir el sobreajuste mediante la combinación de múltiples árboles de decisión. Por ejemplo, en un estudio realizado por Domingo Gesteiro (2018), se empleó Random Forest para predecir la severidad de accidentes con víctimas, logrando resultados significativos en la clasificación de la gravedad de los siniestros con una precisión del 95%.

El modelo Random Forest será analizado en dos configuraciones: una versión con hiperparámetros predeterminados y otra optimizada mediante la búsqueda sistemática de los parámetros más adecuados. Para la optimización, se empleará GridSearchCV, ajustando valores clave como n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features y bootstrap. Este proceso permitirá evaluar el efecto de la selección de hiperparámetros en la estabilidad y precisión del modelo, asegurando una mejora en la capacidad de generalización y reduciendo potenciales problemas de sobreajuste en la predicción de la severidad de los accidentes.

LightGBM (Light Gradient Boosting Machine)

Este algoritmo de aprendizaje automático está basado en técnicas de gradient boosting, diseñado para ser altamente eficiente y escalable en el manejo de grandes volúmenes de datos. Su arquitectura se basa en el crecimiento de árboles por hojas (leaf-wise) y en el uso de histogramas para acelerar el entrenamiento, lo que le permite manejar interacciones complejas entre variables de manera efectiva (Kun, Haocheng, & Xiao, 2022).

LightGBM ha demostrado ser una herramienta efectiva en la predicción de la severidad de accidentes de tráfico. Un estudio realizado en el Reino Unido, empleó este algoritmo para analizar registros de accidentes de 2017. Los resultados mostraron un rendimiento superior en comparación con otros modelos, como Random Forest, destacándose por su alta precisión en la clasificación de la severidad de los siniestros, la cual alcanzó un 92% de precisión (ídem).

Este modelo será sometido a una evaluación comparativa entre su versión estándar y una versión optimizada. En la optimización, se exploraron diferentes combinaciones de

n_estimators, learning_rate, max_depth, num_leaves, min_child_samples, subsample y colsample_bytree, con el objetivo de mejorar su eficiencia y capacidad predictiva. Dado que LightGBM utiliza un crecimiento basado en hojas y técnicas avanzadas de histogramas, su desempeño puede verse significativamente afectado por la correcta elección de hiperparámetros. Este análisis permitió determinar el impacto de la optimización en la precisión del modelo y su adaptabilidad a los datos de accidentes de tránsito.

FFNN (Feed-Forward Neural Network)

Este modelo de aprendizaje profundo basado en redes neuronales artificiales, donde la información fluye en una única dirección, desde la capa de entrada hasta la capa de salida, sin ciclos ni retroalimentación. Su arquitectura se compone de múltiples capas interconectadas, incluyendo una capa oculta con funciones de activación no lineales, lo que permite capturar relaciones complejas en los datos.

Este enfoque ha sido aplicado con éxito en la predicción de la severidad de accidentes de tráfico. Un estudio realizado en la Arabia Saudita utilizó un modelo FFNN entrenado con datos de siniestros ocurridos entre 2017 y 2019 en carreteras rurales, optimizando su desempeño mediante el algoritmo de backpropagation y una función de activación logística. Los resultados mostraron que el modelo era capaz de predecir con buena precisión la severidad de los accidentes, consiguiendo una precisión del 77.5%, superando métodos estadísticos tradicionales y destacando la importancia de variables como el volumen de tráfico, velocidad promedio y condiciones climáticas en la severidad de los siniestros (Jamal & Waleed, 2020).

El modelo de red neuronal Feed Forward (FFNN) será evaluado en dos configuraciones: una versión base y una versión optimizada. La versión optimizada incorpora mejoras clave en la arquitectura y el proceso de entrenamiento para aumentar su capacidad de aprendizaje y generalización. Se emplea una red más profunda con un mayor número de capas y neuronas, junto con técnicas avanzadas de regularización como Dropout y Batch Normalization. Además, se implementa Early Stopping para mitigar el sobreajuste.

En conclusión, para este estudio se evaluarán los modelos mencionados anteriormente en su configuración estándar y también en una versión optimizada mediante la búsqueda de los mejores hiperparámetros. En su forma natural, cada modelo será entrenado con sus

configuraciones predeterminadas, mientras que en la versión optimizada se aplicará la técnica de ajuste de hiperparámetros, en este caso, GridSearchCV, para mejorar su desempeño. Por lo tanto, en el caso de XGBoost, Random Forest y LightGBM, se una versión ajustada que optimiza hiperparámetros clave como n_estimators, learning_rate y max_depth, entre otros. Esta estrategia permitirá evaluar el impacto del ajuste de parámetros en la precisión del modelo y su capacidad de generalización.

Métricas e Interpretabilidad

La evaluación del desempeño de los modelos mencionados anteriormente es un aspecto fundamental para garantizar su aplicabilidad en el ámbito de la seguridad vial y toma de decisiones. En este estudio, se emplearán diversas métricas de clasificación para medir la efectividad de los modelos en la tarea de identificar correctamente los accidentes de tránsito con y sin víctimas. Además, se realizará un análisis de interpretabilidad utilizando la técnica SHAP descrita anteriormente.

Selección de Métricas de Evaluación

En el contexto de la seguridad vial, la correcta identificación de accidentes con víctimas resulta crítica, ya que una predicción errónea puede tener consecuencias humanas irreversibles. Por ello, la selección de métricas no debe basarse únicamente en la precisión global del modelo, sino en su capacidad para minimizar los errores más costosos: los falsos negativos (FN). Es decir, aquellos casos en los que un accidente con víctimas es clasificado erróneamente como sin víctimas. Este tipo de error puede retrasar o impedir el envío oportuno de ambulancias, bomberos o personal de emergencia, aumentando el riesgo de muertes o lesiones graves.

Por esta razón, en este estudio se priorizan métricas que penalicen con mayor severidad la presencia de falsos negativos y permitan evaluar el comportamiento del modelo en un escenario desbalanceado, donde los accidentes con víctimas pueden representar una minoría de los casos. Las métricas seleccionadas son las siguientes:

- Precisión (Precision): Evalúa la proporción de verdaderos positivos entre todas las predicciones positivas. Es útil para saber cuántas de las predicciones que el modelo clasificó como “con víctimas” realmente lo son. Si bien es importante reducir las falsas alarmas (falsos positivos), esta métrica no refleja adecuadamente los falsos negativos, por lo que se considera complementaria en este estudio.

- Recall (Sensibilidad o Tasa de Verdaderos Positivos): Mide la capacidad del modelo para identificar correctamente todos los casos positivos reales, es decir, accidentes con víctimas. Esta métrica es la más relevante en este estudio, ya que penaliza directamente los falsos negativos. Maximizar el recall significa reducir la posibilidad de que un accidente grave no sea detectado, lo cual es esencial desde una perspectiva de seguridad pública.
- F1-Score: Es la media armónica entre precisión y recall, y se utiliza cuando se busca un equilibrio entre ambos. En escenarios desbalanceados como el presente, el F1-Score permite obtener una métrica global sin que el modelo se sesgue hacia la clase mayoritaria.
- Área Bajo la Curva ROC (AUC-ROC): Evalúa la capacidad general del modelo para discriminar entre clases. Es particularmente útil cuando se comparan múltiples modelos, ya que resume la relación entre la tasa de verdaderos positivos (TPR) y la tasa de falsos positivos (FPR) en diferentes umbrales de decisión.
- Matriz de Confusión: Proporciona un desglose detallado de los aciertos y errores del modelo, permitiendo observar cuántos casos fueron clasificados correctamente o incorrectamente en cada clase. Es una herramienta diagnóstica esencial para identificar el tipo de errores cometidos.

La elección de estas métricas se justifica además por su uso estandarizado en la literatura sobre predicción de severidad de accidentes. Estudios como el de Amini et al. (2022) ha utilizado precisamente estas métricas para evaluar el desempeño de modelos de clasificación binaria aplicados a datos de tránsito. En dichos trabajos, se destaca que el recall y el AUC-ROC son indicadores especialmente importantes cuando el costo de un falso negativo es elevado, como en casos de riesgo humano.

En resumen, las métricas seleccionadas no solo permiten una evaluación robusta y multifacética del desempeño de los modelos, sino que además están alineadas con el objetivo principal del estudio: maximizar la detección oportuna de accidentes con víctimas para reducir el riesgo humano asociado. En consecuencia, el recall se prioriza como la métrica clave, sin dejar de considerar otras métricas complementarias para obtener una visión integral de la eficacia del modelo.

Importancia de Minimizar los Falsos Negativos

En el contexto de la predicción de la severidad de accidentes de tránsito, un falso negativo ocurre cuando un accidente con víctimas es clasificado erróneamente como sin víctimas. Como se mencionó anteriormente, esto puede tener consecuencias graves, ya que subestimar la severidad de un accidente podría afectar la asignación de recursos de emergencia, la toma de decisiones de políticas de seguridad vial y la gestión del tráfico.

Según Sokolova y Lapalme (2009), en problemas donde la detección de la clase positiva (en este caso, accidentes con víctimas) es crítica, la métrica de recall debe ser priorizada sobre otras métricas como precisión o exactitud general. Además, estudios en inteligencia artificial aplicada a la salud y seguridad indican que, en escenarios donde una clasificación incorrecta puede generar riesgos, los modelos deben ser diseñados para minimizar los FN (Sanskriti, 2022). La ecuación para el cálculo de recall es la siguiente:

$$Recall = \frac{TP}{TP + FN}$$

Donde:

- TP: son los verdaderos positivos (accidentes con víctimas correctamente identificadas).
- FN: son los falsos negativos (accidentes con víctimas clasificados incorrectamente como sin víctimas).

Análisis de Interpretabilidad

Con el objetivo de garantizar que los modelos sean no solo precisos, sino también comprensibles y confiables, se emplearán técnicas de interpretabilidad basadas en Explainable AI (XAI). En particular, se utilizará SHAP (Shapley Additive Explanations), una herramienta avanzada que permite analizar la contribución de cada variable en la predicción del modelo (Li, Guo, Li, Liu, & Wang, 2023). Esta técnica proporcionará información detallada sobre los factores que influyen en la clasificación de un accidente, ya sea como con víctimas o sin víctimas, permitiendo comprender cómo cada característica impacta en la decisión del modelo.

CAPÍTULO IV
RESULTADOS Y DISCUSIÓN

Interpretación de resultados

En esta sección se presentan los resultados obtenidos en la evaluación de los modelos de machine learning empleados para la predicción de la severidad de los accidentes de tránsito en Quito. Se analizan las métricas de desempeño de cada modelo y sus respectivas matrices de confusión con el objetivo de determinar el modelo más adecuado para minimizar la incorrecta clasificación de accidentes con víctimas (False Negatives - FN), dado el impacto significativo de estos errores en términos de seguridad vial y gestión del tráfico.

Desempeño de los Modelos

Se evaluaron un total de 8 modelos, descritos a detalle en la sección anterior: XGBoost, Random Forest, LightGBM y una Red Neuronal Feedforward (FFNN), tanto en su configuración estándar como en su versión optimizada con hiperparámetros ajustados.

Estos modelos se evaluaron utilizando métricas de desempeño clave, tales como precisión, recall, F1-score y la puntuación ROC-AUC. Estas métricas nos proporcionan una visión detallada sobre la capacidad de cada modelo para diferenciar entre accidentes con y sin víctimas.

Tabla No. 2: Modelos en su forma base.

Modelo	Precision	Recall	F1-Score	AUC-ROC
XGBoost	0,87	0,86	0,86	0,9326
Random Forest	0,87	0,86	0,86	0,9314
LightGBM	0,87	0,87	0,87	0,9360
FFNN	0,82	0,74	0,78	0,8840

Elaborado por: Morillo, Jean (2025).

La Tabla 2 presenta las métricas de desempeño de los modelos en su configuración inicial. No obstante, con el objetivo de mejorar la capacidad predictiva en la clasificación de la severidad de los accidentes de tránsito, se llevó a cabo un proceso de optimización basado en el ajuste de hiperparámetros y la implementación de técnicas para mejorar la generalización, como se detalló en el capítulo previo. Estas estrategias permitieron

incrementar la precisión y sensibilidad de los modelos, facilitando una identificación más efectiva de la severidad de accidentes de tránsito. Los resultados obtenidos tras la optimización evidenciaron un rendimiento superior en comparación con las versiones originales, los cuales se presentan en la Tabla 3.

Tabla No. 3: Modelos optimizados.

Modelo	Precision	Recall	F1-Score	AUC-ROC
XGBoost Optimizado	0,88	0,86	0,86	0,9357
Random Forest Optimizado	0,88	0,86	0,86	0,9341
LightGBM Optimizado	0,88	0,87	0,88	0,9373
FFNN Optimizado	0,86	0,85	0,85	0,9071

Elaborado por: Morillo, Jean (2025).

Los modelos optimizados presentaron mejoras significativas en sus métricas de clasificación, lo que sugiere que la optimización de hiperparámetros contribuyó a un mejor desempeño en la identificación de accidentes con víctimas.

LightGBM Optimizado no solo obtuvo la mejor puntuación AUC-ROC (0.9373), sino que también presentó el recall más alto (0.87) entre todos los modelos evaluados. Esto indica que el modelo tiene una gran capacidad para identificar correctamente los accidentes con víctimas, minimizando los falsos negativos. En un contexto donde la prioridad es detectar con precisión los siniestros más graves, este alto recall es un factor clave, ya que sugiere que LightGBM Optimizado es el modelo más eficaz para capturar los casos de mayor riesgo sin sacrificar demasiado la precisión.

XGBoost Optimizado y Random Forest Optimizado mostraron métricas muy similares, con valores de precisión y recall de 0.88 y 0.86 respectivamente. Estos resultados sugieren que ambos modelos son confiables para la tarea de clasificación, ofreciendo una buena capacidad predictiva y reduciendo el número de falsas alarmas y omisiones en la detección de accidentes con víctimas.

Por otra parte, la Red Neuronal Feed Forward (FFNN) Optimizada, aunque mejoró significativamente en comparación con su versión sin optimizar, sigue por debajo de los modelos de boosting en términos de AUC-ROC (0.9071) y recall (0.85). Si bien este modelo puede ser útil en combinación con otros enfoques, su menor desempeño en recall indica que aún tiene dificultades para identificar correctamente todos los accidentes con víctimas, lo que podría afectar su aplicabilidad en escenarios donde la prioridad es minimizar los falsos negativos.

Matrices de Confusión

El análisis de la matriz de confusión de los modelos optimizados permite evaluar con mayor detalle su capacidad para predecir correctamente los accidentes con víctimas (True Positives) y minimizar los casos en los que estos son clasificados erróneamente como accidentes sin víctimas (False Negatives). Dado que en este contexto la correcta identificación de los accidentes con víctimas es prioritaria, se otorga mayor importancia a aquellos modelos que minimizan los FN, ya que un error en esta predicción podría tener consecuencias significativas (ver apéndice C).

El análisis de la matriz de confusión para los modelos optimizados se muestra en la Tabla 4 y revela diferencias clave en su capacidad para clasificar correctamente los accidentes con y sin víctimas. LightGBM Optimizado nuevamente destaca como el modelo con mejor desempeño, logrando el menor número de falsos negativos (FN = 648) y el mayor número de verdaderos positivos (TP = 4231). Esto confirma su capacidad superior para identificar accidentes con víctimas, lo que es fundamental en un sistema donde la prioridad es minimizar los falsos negativos.

Tabla No. 4: Matriz de confusión.

Modelo	TN	FP	FN	TP
XGBoost Optimizado	5157	647	685	4194
Random Forest Optimizado	5130	674	660	4219
LightGBM Optimizado	5119	685	648	4231
FFNN Optimizado	5093	711	748	4131

Elaborado por: Morillo, Jean (2025).

XGBoost Optimizado y Random Forest Optimizado muestran un desempeño similar, con 685 y 660 falsos negativos, respectivamente. Aunque ambos modelos presentan una reducción considerable de FN en comparación con sus versiones sin optimizar, aún están por debajo de LightGBM en términos de capacidad de detección de accidentes con víctimas. Sin embargo, XGBoost Optimizado logra un menor número de falsos positivos (FP = 647), lo que indica que genera menos falsas alarmas en comparación con los demás modelos.

Por otro lado, FFNN Optimizado, aunque mejoró significativamente respecto a su versión original, sigue presentando la mayor cantidad de falsos negativos (FN = 748) y falsos positivos (FP = 711). Esto sugiere que, si bien ha mejorado en precisión y capacidad de generalización, todavía tiene dificultades para identificar correctamente todos los

casos de accidentes con víctimas. Esta deficiencia podría afectar su aplicabilidad en un sistema donde es crucial reducir al mínimo los casos en los que se subestime la gravedad del accidente.

Finalmente, LightGBM Optimizado se confirma como el modelo más efectivo para la tarea, minimizando los falsos negativos sin incrementar significativamente los falsos positivos. XGBoost Optimizado y Random Forest Optimizado siguen siendo opciones viables con un buen equilibrio entre FP y FN, mientras que FFNN Optimizado, aunque mejorado, aún necesita ajustes para alcanzar el nivel de desempeño de los modelos basados en árboles de decisión.

En conclusión, el modelo LightGBM Optimizado se posiciona como la mejor opción para la predicción de la severidad de los accidentes de tránsito debido a su capacidad para minimizar los falsos negativos y maximizar la detección de accidentes con víctimas. Su alto AUC-ROC y recall indican que es especialmente efectivo en la identificación de los casos de mayor riesgo, lo que resulta crucial en aplicaciones donde la prioridad es garantizar una respuesta oportuna en la logística de emergencia. Su equilibrio entre precisión y recall lo convierte en una herramienta valiosa para la toma de decisiones en seguridad vial, reduciendo el riesgo de subestimar la gravedad de los accidentes. Estos resultados están en línea con los obtenidos por Kun, Haocheng, & Xiao (2022), quienes, en un estudio similar realizado en el Reino Unido, también encontraron que LightGBM presentó un rendimiento superior en comparación con otros modelos. Aunque XGBoost y Random Forest también presentan un desempeño sólido, LightGBM Optimizado se destaca como el modelo más confiable para esta tarea.

Contraste con investigaciones previas

Los resultados de esta investigación muestran que los modelos de Machine Learning basados en boosting y árboles de decisión alcanzaron un desempeño sobresaliente en la predicción de la severidad de los accidentes de tránsito en Quito. Entre ellos, el LightGBM Optimizado destacó como el mejor clasificador, alcanzando un recall de 0,87 y un AUC-ROC de 0,9373, lo que confirma su capacidad para minimizar falsos negativos y maximizar la detección de accidentes con víctimas.

Este hallazgo coincide con lo reportado por Kun, Haocheng & Xiao (2022) en el Reino Unido, donde LightGBM superó a otros algoritmos en precisión y capacidad de

generalización. Asimismo, el buen desempeño observado en XGBoost (AUC-ROC = 0,9357) se alinea con lo encontrado por Ahmed et al. (2023), quienes destacan que este modelo es especialmente efectivo en entornos urbanos con alta complejidad en las variables. De manera similar, el Random Forest mostró métricas robustas (AUC-ROC = 0,9341), en concordancia con lo documentado por Domingo Gesteiro (2018), donde alcanzó un 95% de precisión en la predicción de la severidad de siniestros.

En general, los resultados obtenidos son consistentes con la literatura internacional, que indica que los modelos de boosting (XGBoost, LightGBM) y de ensamble como Random Forest tienden a superar a otros enfoques más simples en la predicción de accidentes de tránsito (Amini et al., 2022). Estos modelos ofrecen no solo un alto nivel de precisión, sino también una reducción significativa de errores críticos como los falsos negativos, lo que refuerza su utilidad en la práctica para apoyar la gestión de emergencias.

Además, el análisis de interpretabilidad con valores SHAP corroboró hallazgos de investigaciones previas sobre los factores de mayor influencia en la severidad de accidentes. Variables como motocicleta y peatón se identificaron como altamente asociadas con accidentes con víctimas, lo que coincide con lo reportado por Santos et al. (2023) y Liasidis et al. (2023) acerca de la vulnerabilidad de estos actores viales. Asimismo, el atropello emergió como uno de los tipos de siniestro más críticos, en línea con la literatura que resalta su fuerte relación con lesiones graves o fatales. Por otro lado, el exceso de velocidad fue ratificado como un factor determinante, reforzando lo señalado por Nasiri et al. (2019).

En conjunto, los resultados de este estudio no solo validan lo reportado en contextos internacionales, sino que también aportan evidencia empírica para el caso de Quito, confirmando que los modelos de boosting constituyen herramientas efectivas para la predicción de la severidad de accidentes de tránsito y para apoyar la toma de decisiones orientadas a la seguridad vial.

Interpretabilidad

El análisis de valores SHAP permite evaluar la contribución de las variables más importantes en la predicción de la severidad de los accidentes de tránsito. Basada en la interpretación del modelo LightGBM optimizado, esta técnica de Explainable AI (XAI) proporciona una comprensión detallada del impacto de cada variable en la salida del

modelo. Al analizar cómo y por qué se toman determinadas decisiones en la predicción, SHAP facilita una interpretación más transparente y confiable del comportamiento del modelo.

En el Gráfico 9, el eje vertical muestra las siete características más influyentes en la clasificación de los accidentes en las categorías de "Sin Víctimas" y "Con Víctimas". Estas variables han sido seleccionadas según su impacto acumulado en la predicción, lo que significa que su presencia en el conjunto de datos tiene una influencia significativa en la determinación del resultado. Cada punto en la figura representa un caso individual del conjunto de prueba, donde la posición horizontal refleja el valor SHAP asociado a cada observación. El eje horizontal indica la magnitud y dirección del impacto de cada variable en la predicción. Los valores positivos sugieren que la variable aumenta la probabilidad de que el accidente sea clasificado como "Con Víctimas", mientras que los valores negativos indican que la variable está más asociada con accidentes clasificados como "Sin Víctimas".

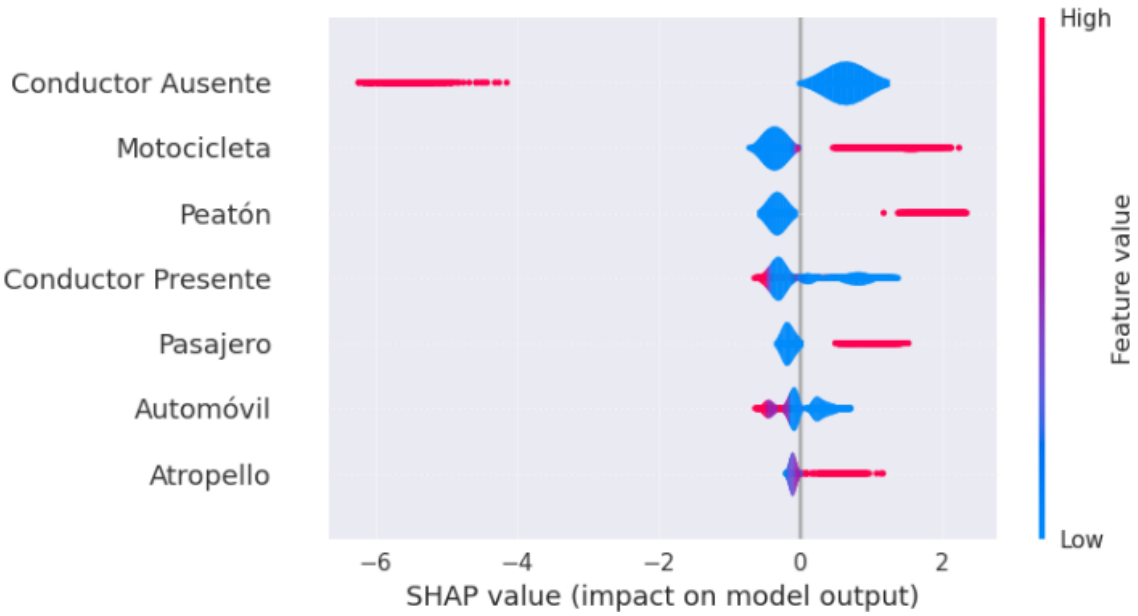


Gráfico No. 9: Valores SHAP.
Elaborado por: Morillo, Jean (2025).

El color de los puntos representa el valor de la característica en cada observación. En la escala de colores, los tonos rojos indican valores altos de la variable, mientras que los tonos azules representan valores bajos. Esto permite observar patrones importantes en los datos. Por ejemplo, en la figura se aprecia que la variable "Conductor Ausente" tiene una fuerte relación negativa con la clasificación de accidentes con víctimas, lo que indica que,

cuando un conductor está ausente, es más probable que el accidente no tenga consecuencias fatales. Por otro lado, variables como "Motocicleta" y "Peatón" muestran un impacto positivo en la predicción de accidentes con víctimas, lo que es consistente con la literatura previa sobre seguridad vial, donde estos actores viales presentan mayor vulnerabilidad ante un siniestro.

El gráfico también destaca la importancia del tipo de siniestro en la severidad del accidente. En particular, la variable "Atropello" aparece como una de las más relevantes, lo que indica que este tipo de accidente tiende a estar fuertemente asociado con la presencia de víctimas. Esto tiene sentido desde una perspectiva de seguridad vial, ya que los atropellos suelen involucrar a peatones, quienes son especialmente vulnerables en un evento de tránsito.

En conclusión, la técnica SHAP proporciona una visión intuitiva y basada en datos sobre los factores clave que influyen en la predicción de la severidad de los accidentes de tránsito. Al utilizar valores SHAP, se logra una mayor transparencia en la interpretación del modelo de Machine Learning, lo que permite no solo evaluar su rendimiento, sino también entender mejor los factores que contribuyen a la gravedad de los accidentes en Quito. Esta información es fundamental para el desarrollo de políticas públicas enfocadas en la reducción de la siniestralidad vial y la implementación de medidas preventivas dirigidas a los actores más vulnerables del tránsito (ver apéndice D).

Verificación de la hipótesis

La hipótesis planteada en el capítulo 2 se verificó mediante un procedimiento de bootstrap no paramétrico, técnica apropiada para estimar la variabilidad del recall sin asumir una distribución específica. Se generaron 1.000 réplicas a partir del conjunto de prueba estratificado, calculando en cada una el recall del modelo LightGBM Optimizado. A partir de estas réplicas se obtuvo la media, la desviación estándar, el intervalo de confianza del 95 % y el p-valor empírico correspondiente a la prueba unilateral H_1: Recall $\geq 0{,}85$ vs H_0: Recall $< 0{,}85$.

Los resultados muestran un recall promedio de $0{,}8673 \pm 0{,}0049$, con un intervalo de confianza del 95 % de [0,8571, 0,8770]. Además, ninguna de las 1.000 réplicas presentó valores inferiores al umbral de 0,85, lo que corresponde a un p-valor empírico $< 0{,}001$. Esto significa que, bajo la hipótesis nula, la probabilidad de observar un desempeño como

el alcanzado es extremadamente baja. (El código empleado para este análisis se presenta en el apéndice E)

Con base en estos resultados, el límite inferior del intervalo de confianza es mayor que 0,85, por lo que se rechaza la hipótesis nula y se acepta la alternativa. En consecuencia, se concluye que el recall poblacional del modelo es al menos 0,85, confirmando la validez de la hipótesis y respaldando el uso del modelo LightGBM Optimizado como herramienta predictiva confiable en el contexto de la seguridad vial en el cantón Quito.

CAPÍTULO V
CONCLUSIONES Y RECOMENDACIONES

Conclusiones

Este estudio tuvo como objetivo predecir la severidad de accidentes de tránsito en el cantón Quito, utilizando técnicas de Inteligencia Artificial. Para ello, se desarrollaron cuatro modelos principales (XGBoost, Random Forest, LightGBM y FFNN), evaluados en su configuración base y optimizada. Finalmente, se aplicó la técnica de interpretabilidad SHAP para comprender la influencia de cada variable en la predicción final.

El análisis exploratorio de datos permitió identificar patrones relevantes en la ocurrencia y severidad de los accidentes de tránsito en Quito. Se observó que los siniestros son más frecuentes los sábados y en horarios de alta congestión vehicular (6:00–8:00 y 17:00–20:00), y que los atropellos presentan una alta probabilidad de generar víctimas. Asimismo, factores como el exceso de velocidad y la falta de ceder el paso a peatones se consolidan como determinantes en la gravedad de los accidentes.

En cuanto a los modelos predictivos evaluados, LightGBM optimizado se consolidó como el de mejor desempeño, alcanzando un recall del 87 % y un AUC-ROC de 0,9373. Estos resultados reflejan una alta capacidad para identificar accidentes con víctimas y, en particular, para minimizar falsos negativos, aspecto crucial en contextos donde no detectar un caso grave puede tener consecuencias críticas.

El uso de técnicas de interpretabilidad mediante SHAP permitió, además, identificar las variables con mayor influencia en la predicción. Destacan factores como la presencia de motocicletas y peatones, así como el tipo de siniestro “atropello”, que aumentan la probabilidad de accidentes con víctimas. En contraste, situaciones de “conductor ausente” reducen dicha probabilidad, lo cual aporta evidencia empírica sobre la relevancia diferencial de las variables incluidas en el modelo.

Los hallazgos obtenidos confirman que la Inteligencia Artificial constituye una herramienta eficaz para apoyar la gestión de la seguridad vial, al ofrecer modelos capaces de anticipar la severidad de los siniestros. Su implementación práctica podría contribuir a mejorar la asignación de recursos de emergencia y optimizar los tiempos de respuesta

ante accidentes graves, fortaleciendo así las capacidades de gestión de las instituciones responsables.

Finalmente, este estudio demuestra que la Inteligencia Artificial constituye una herramienta eficaz para apoyar la gestión de la seguridad vial, al ofrecer modelos capaces de anticipar la severidad de los siniestros y mejorar la asignación de recursos de emergencia, optimizando así los tiempos de respuesta ante accidentes graves. No obstante, se reconoce que el fenómeno de la accidentalidad vial mantiene una complejidad significativa. Factores externos como las condiciones climáticas, el estado de las vías o la señalización no fueron considerados en este estudio, pero su inclusión en futuras investigaciones podría mejorar aún más la capacidad predictiva y proporcionar una comprensión más integral del contexto de Quito.

Recomendaciones

A pesar de que los resultados obtenidos en este estudio son prometedores, la falta de información sobre ciertos factores clave limita el rendimiento del modelo. A continuación, se presentan algunas recomendaciones para mejorar la calidad de las predicciones y su aplicación en escenarios reales:

- Incorporación de datos sobre condiciones climáticas y estado de las carreteras: Factores como la lluvia, niebla, visibilidad y el estado del pavimento pueden influir directamente en la severidad de los accidentes. La integración de estos datos en tiempo real mediante sensores o registros históricos podría mejorar la capacidad predictiva del modelo.
- Inclusión de información sobre la velocidad de los vehículos: La velocidad al momento del impacto es un factor crítico en la gravedad de los accidentes. Se recomienda la recopilación de datos de radares de tráfico o sensores de velocidad en la vía para evaluar su impacto en la predicción.
- Desarrollo de una aplicación basada en IA: Se propone la implementación de una aplicación móvil o web que utilice los modelos desarrollados para proporcionar predicciones en tiempo real. Esta herramienta podría ser utilizada por organismos de tránsito y servicios de emergencia para priorizar la atención de accidentes según su gravedad estimada.

En conclusión, el presente estudio resalta el potencial de la Inteligencia Artificial para mejorar la respuesta ante accidentes de tránsito en Quito. Las recomendaciones propuestas buscan no solo perfeccionar los modelos desarrollados, sino también facilitar su implementación práctica y maximizar su impacto en la reducción de víctimas por siniestros viales. La integración de información más detallada y el desarrollo de herramientas tecnológicas específicas representan los próximos pasos lógicos para transformar los hallazgos de esta investigación en soluciones concretas que contribuyan significativamente a la seguridad vial en el cantón Quito.

Referencias

Adefabi, A., Olisah, S., Obunadike, C., Oyetubo, O., Taiwo, E., & Tella, E. (2023). Predicting Accident Severity: An Analysis Of Factors Affecting Accident Severity Using Random Forest Model. *International Journal on Cybernetics & Informatics, 12*(6). doi:10.5121/ijci.2023.120609

Ahmed, S., Hossain, A., Ray, S. K., Mafijul, B. I., & Sabuj, S. R. (2023). A study on road accident prediction and contributing factors using explainable machine learning models: analysis and performance. *Transportation Research Interdisciplinary Perspectives, 19*. doi:https://doi.org/10.1016/j.trip.2023.100814

Akinade, A. O., Adepoju, P. A., Ige, A. B., & Afolabi, A. I. (2024). Artificial Intelligence in Traffic Management: A Review of Smart Solutions and Urban Impact. *Iconic Research And Engineering Journals, 7*(12), 511-522.

Al-Masaeid, H. R., & Khaled , F. (2020). *Regression-based traffic accident prediction models: applicability across different road types.* Jordan Journal of Civil Engineering. doi:https://doi.org/10.14525/JJCE.v17i1.04

Al-Masaeid, H., & Khaled, F. (2023). Performance of Traffic Accidents' Prediction Models. *Jordan Journal of Civil Engineering, 17*(1).

Amini, M., Bagheri, A., & Delen, D. (2022). Discovering injury severity risk factors in automobile crashes: A hybrid explainable AI framework for decision support. *Reliability Engineering & System Safety, 226*.

Argüello, D., & Alcívar, E. (2023). *Análisis de accidentes de tránsito en el cantón Guayaquil usando Machine Learning.* ESPOL.FCNM.

Baykal, T., Ergezer, F., Eriskin, E., & Terzi, S. (2023). Accident Severity Prediction in Big Data Using Auto-Machine Learning. *Scientia Iranica*, 1026 - 3098. Obtenido de https://scientiairanica.sharif.edu/article_23141_094e079a349ba28650650cc0c3118155.pdf

Benson, C., & Obasi, I. (2023). Evaluating the effectiveness of machine learning techniques in forecasting the severity of traffic accidents. *Heliyon*. doi:https://doi.org/10.1016/j.heliyon.2023.e18812

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5-32. doi:10.1023/A:1010933404324

Brownlee, J. (2022). *Data Preparation for Machine Learning: Transform, Encode, and Scale.* Machine Learning Mastery. Obtenido de Machine Learning Mastery

Chakradhara, P., Mishra, A. K., & Nawab, A. K. (2023). Predicting and explaining severity of road accident using artificial intelligence techniques, SHAP and feature analysis. *International Journal of Crashworthiness, 28*, 186-201. doi:10.1080/13588265.2022.2074643

Domingo, A. R. (2018). *"Predicción de la severidad de accidentes de tráfico con víctimas mediante Random Forest.*

Escobar, J. C. (19 de 09 de 2024). *Ecuador Chequea*. Obtenido de La vía con mayor cantidad de siniestros es la Simón Bolívar: https://ecuadorchequea.com/la-via-con-mayor-cantidad-de-siniestros-es-la-simon-bolivar/

Hassouna, F. M., & Al-Sahili, K. (2020). *Analysis and modeling of road crash trends in Palestine* (Vol. 45). Arabian Journal for Science and Engineering. doi:https://doi.org/10.1007/s13369-020-04728-6

Infante, P., Jacinto, G., Santos, D., Pedro, N., Anabela, A., Quaresma, P., . . . Manuel, P. (2023). Prediction of Road Traffic Accidents on a Road in Portugal: A Multidisciplinary Approach Using Artificial Intelligence, Statistics, and Geographic Information Systems. *Information*. doi:doi:10.3390/info14040238

Instituto Nacional de Estadística y Censos. (2024). *Siniestros de Tránsito - I Trimestre, 2024. Agencia Nacional de Tránsito (ANT).*

Jamal, A., & Waleed, U. (2020). Exploring the Injury Severity Risk Factors in Fatal Crashes with Neural Network. *International Journal of Environmental Research and Public Health, 17*. doi:10.3390/ijerph17207466

Jiang, T., Gradus, J. L., & Rosellini, A. J. (2020). Supervised Machine Learning: A Brief Primer. *Behavior Therapy, 51*(5), 675-687. doi:https://doi.org/10.1016/j.beth.2020.05.002.

Khasawneh, M. A., Al-Omari, A. A., & Ganam, B. (2018). *Forecasting traffic accidents in Jordan using regression techniques* (Vol. 12). Jordan Journal of Civil Engineering. doi:https://doi.org/10.21203/rs.3.rs-4187484/v1

Kibria , H. B., & Matin, A. (2022). The severity prediction of the binary and multi-class cardiovascular disease − A machine learning-based fusion approach. *Computational Biology and Chemistry, 98*, 107672. doi:https://doi.org/10.1016/j.compbiolchem.2022.107672

Kumar Mohanta, B., Debasish , J., Niva , M., Somula , R., Bharat S., R., Hasmat , M., . . . Smriti , S. (2022). Machine learning based accident prediction in secure IoT enable transportation system. *Journal of Intelligent & Fuzzy Systems: Application in Engineering and Technology, 42*(2). doi:10.3233/JIFS-189743

Kun, L., Haocheng, X., & Xiao, L. (2022). Analysis and visualization of accidents severity based on LightGBM-TPE. *Chaos, Solitons & Fractals, 157*. doi:https://doi.org/10.1016/j.chaos.2022.111987

Li, J., Guo, F., Zhou, Y., Yang, W., & Ni, D. (2023). Predicting the severity of traffic accidents on mountain freeways with dynamic traffic and weather data. *Transportation Safety and Environment*. doi:10.1093/tse/tdad001

Li, J., Guo, Y., Li, L., Liu, X., & Wang, R. (2023). Using LightGBM with SHAP for predicting and analyzing traffic accidents severity. En *7th International Conference on Transportation Information and Safety (ICTIS)* (págs. 2150-2155). doi:10.1109/ICTIS60134.2023.10243855

Li, Y., Zhang, H., & Liu, Y. (2023). Cyclical Encoding of Temporal Features for Accident Severity Prediction. *Journal of Transportation Safety & Security, 15*(2), 145-160. doi:10.1080/19439962.2023.1023456

Liasidis, P., Benjamin, E., Jakob, D., Lewis, M., & Demetriades, D. (2023). Injury patterns and outcomes in motorcycle passengers. *European Journal of Trauma and Emergency Surgery, 49*(6), 2447-2457. doi:10.1007/s00068-023-02296-8

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems, 30*.

Mahajan, T., Singh, G., & Bruns, G. (2021). An Experimental Assessment of Treatments for Cyclical Data. *Computer Science Conference for CSU Undergraduates.*

Mohammad Tamim , K. (2024). Robust spatiotemporal crash risk prediction with gated recurrent convolution network and interpretable insights from SHapley additive explanations. *Engineering Applications of Artificial Intelligence, 127*(Part B). Obtenido de https://doi.org/10.1016/j.engappai.2023.107379.

Nasiri , N., Nazari , P., Kamali , A., Sharifi , A., & Sharifi , H. (2019). Factors contributing to fatal road traffic accidents in the South of Kerman during the period from 2013 to 2017, Iran. *Journal of Occupational Health and Epidemiology, 8*, 6-11. Obtenido de http://johe.rums.ac.ir/article-1-325-en.html

Organization, W. H. (2022). *Global Status Report on Road Safety 2022.* World Health Organization.

Quito Informa. (18 de Mayo de 2023). Obtenido de Conoce el procedimiento que utilizan los agentes de tránsito en un siniestro: https://www.quitoinforma.gob.ec/2023/05/18/conoce-el-procedimiento-que-utilizan-los-agentes-de-transito-en-un-siniestro/

Sanskriti , S. (2022). Emphasis on the Minimization of False Negatives or False Positives in Binary Classification. *arXiv preprint*. Obtenido de https://arxiv.org/abs/2204.02526

Santos, K., Dias, J. P., & Amado, C. (2022). A literature review of machine learning algorithms for crash injury severity prediction. *Journal of Safety Research, 80*, 254-269. doi:https://doi.org/10.1016/j.jsr.2021.12.007.

Santos, K., Firme, B., Dias, J. P., & Amado, C. (2023). Analysis of Motorcycle Accident Injury Severity and Performance Comparison of Machine Learning Algorithms. *Transportation Research Record, 2678*(1), 736-748. doi:10.1177/03611981231172507

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427-437. doi:https://doi.org/10.1016/j.ipm.2009.03.002.

Wang, Y., & Zhang, W. (2017). Analysis of Roadway and Environmental Factors Affecting Traffic Crash Severities. *Transportation Research Procedia, 25*, 2124-2130. doi:10.1016/j.trpro.2017.05.407

World Health Organization. (2023). *Global status report on road safety 2023.* Ginebra.

Yan, M., & Shen, Y. (2022). Traffic Accident Severity Prediction Based on Random Forest. *Sustainability, 14*, 1729. doi:https://www.mdpi.com/2071-1050/14/3/1729

Yang, J., Han , S., & Chen , Y. (2023). Prediction of Traffic Accident Severity Based on Random Forest. *Journal of Advanced Transportation, 2023*(1). doi:https://doi.org/10.1155/2023/7641472

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing,, 415*, 295-316. doi:https://www.sciencedirect.com/science/article/pii/S0925231220311693

Zeng, Q., & Huang, H. (2014). A stable and optimized neural network model for crash injury severity prediction. *Accident Analysis & Prevention, 73*, 351-358. doi:https://doi.org/10.1016/j.aap.2014.09.006.

Zheng, A., & Casari, A. (2018). *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists.* O'Reilly Media, Inc.

Zihang , W., Zhang , Y., & Das, S. (2023). Applying Explainable Machine Learning Techniques in Daily Crash Occurrence and Severity Modeling for Rural

Interstates. *Transportation Research Record, 2677*(5), 611-628. doi:10.1177/03611981221134629

Apéndices

Apéndice A - Listado de tipos de siniestros

Listado de tipos de siniestro.

Tipo de Siniestro	Definición	Causas Frecuentes
Choque lateral	Impacto en el costado de un vehículo	No respetar señales de tránsito, invasión de carril
Choque frontal	Impacto frontal entre dos vehículos	Adelantamientos imprudentes, invasión de carril
Colisión por alcance	Impacto en la parte posterior de un vehículo	Distancia de seguridad insuficiente, distracción
Estrellamiento	Impacto contra objeto fijo	Pérdida de control, distracción, fatiga
Atropello	Impacto de vehículo contra peatón	Imprudencia peatonal o del conductor, exceso velocidad
Arrollamiento	Vehículo pasa con sus ruedas sobre la víctima	Imprudencia, falta de visibilidad
Volcamiento	Vehículo gira sobre su eje	Exceso de velocidad, maniobras bruscas

Elaborado por: Morillo, Jean (2025).

Apéndice B - Distribución de Personas Lesionadas

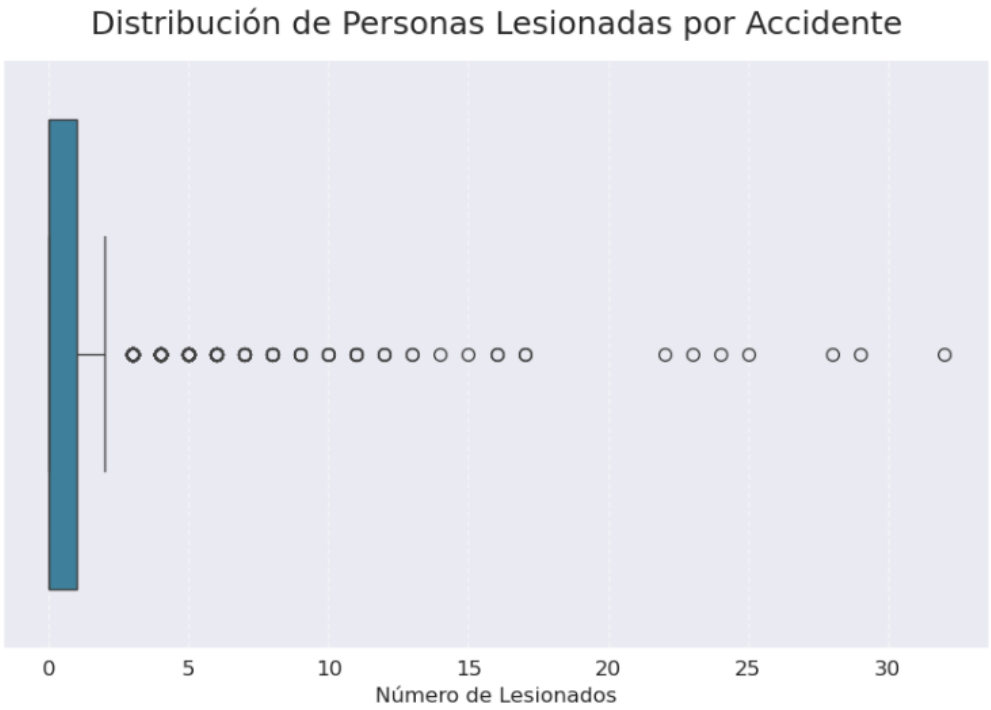
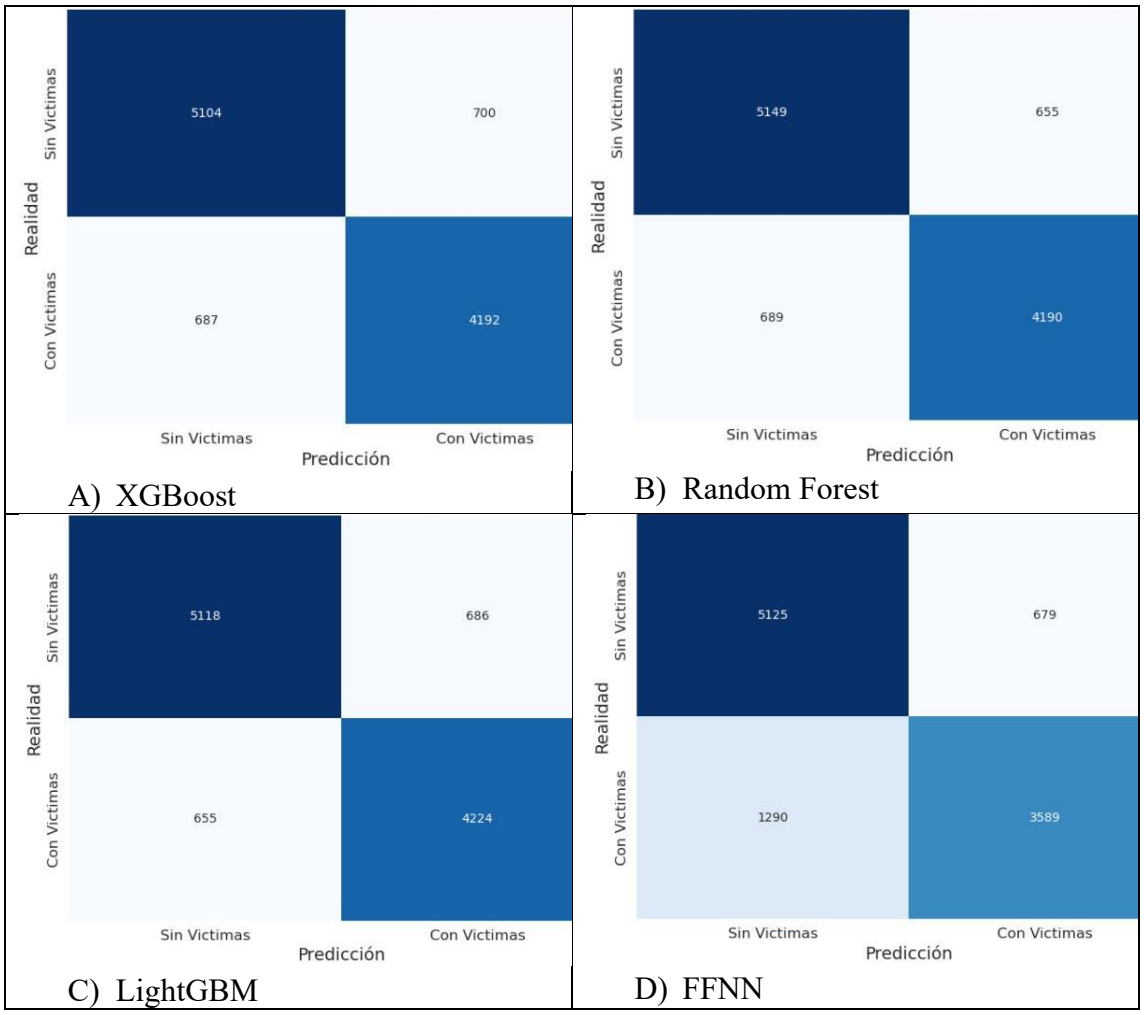
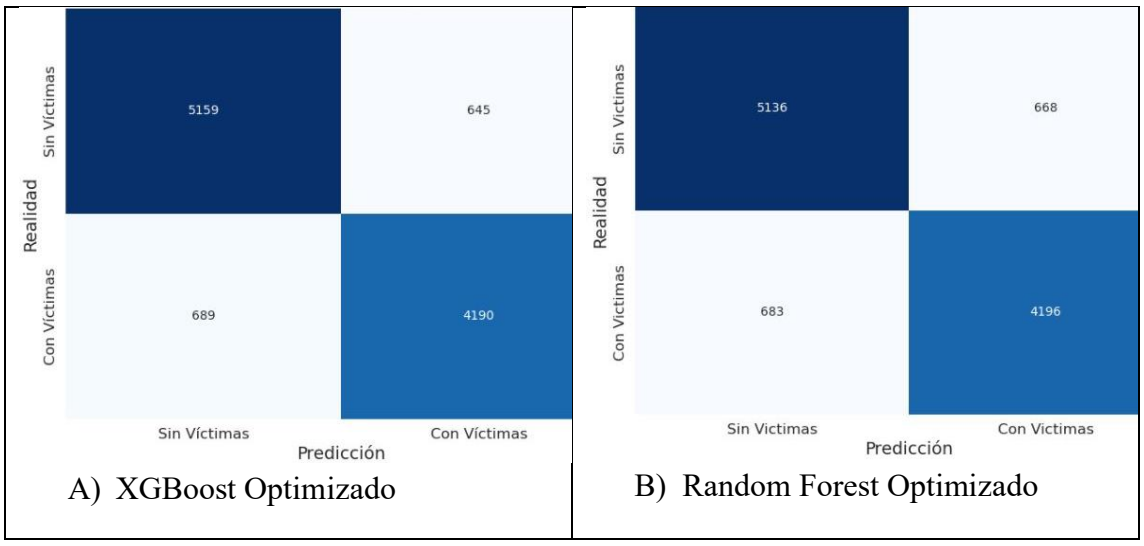


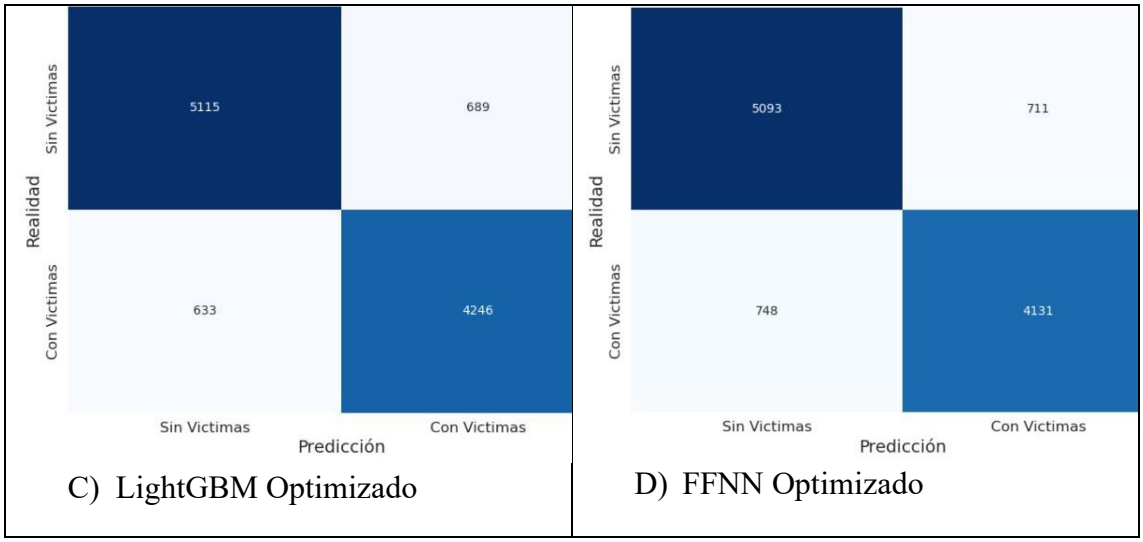
Diagrama de caja y bigotes de la variable Lesionados
Elaborado por: Morillo, Jean (2025).

Apéndice C - Matrices de Confusión



Matrices de confusión para modelos base.
Elaborado por: Morillo, Jean (2025).



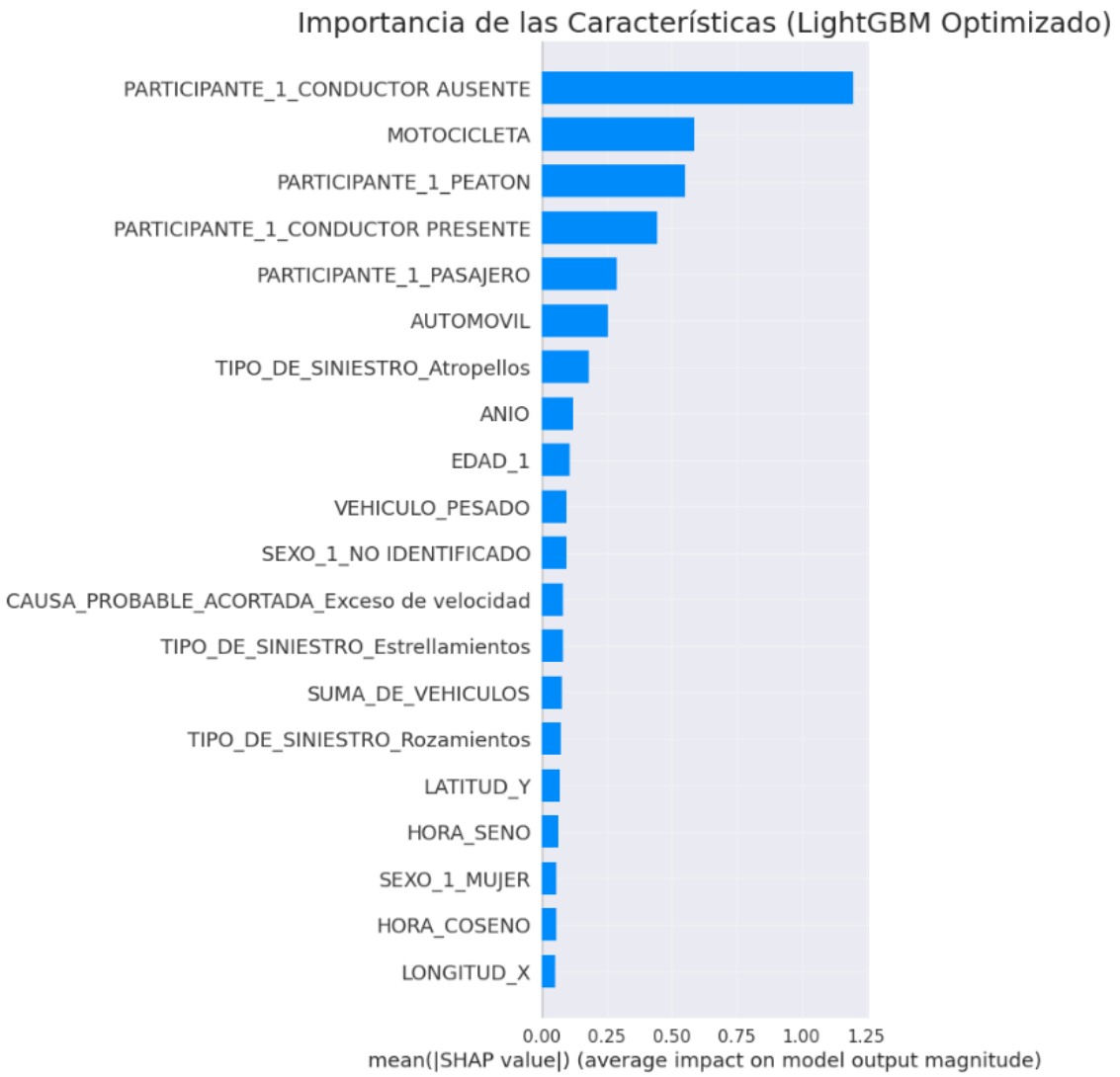


C) LightGBM Optimizado

D) FFNN Optimizado

Matrices de confusión para modelos optimizados.
Elaborado por: Morillo, Jean (2025).

Apéndice D - Importancia de las características usando SHAP



Impacto de las variables en el modelo.
Elaborado por: Morillo, Jean (2025).

Apéndice E - Código en Python para la verificación de la hipótesis

```
# ----------------------------
# Bootstrap para IC de Recall
# ----------------------------
import numpy as np
from sklearn.metrics import recall_score

# Parámetros bootstrap
n_iterations = 1000
rng = np.random.default_rng(42)  # reproducibilidad
n_test = len(y_test)

recalls = np.empty(n_iterations)

# Conertir a arrays
y_true_arr = np.asarray(y_test)
y_pred_arr = np.asarray(y_pred_lgb_opt)

for i in range(n_iterations):
    # muestreo con reemplazo: índices aleatorios del conjunto de prueba
    sample_idx = rng.integers(0, n_test, n_test)
    recalls[i] = recall_score(y_true_arr[sample_idx],
y_pred_arr[sample_idx])

# Estadísticos bootstrap
recall_mean = recalls.mean()
recall_std = recalls.std(ddof=1)
ic_low, ic_high = np.percentile(recalls, [2.5, 97.5])

# Prueba unilaterial H0: recall < 0.85  vs  H1: recall >= 0.85
threshold = 0.85
p_value_one_sided = (recalls < threshold).mean()  # proporción de recalls
< 0.85

# Resultados
print(f"\nBootstrap (n={n_iterations}) - Recall promedio:
{recall_mean:.4f} ± {recall_std:.4f}")
print(f"IC 95% (percentiles): [{ic_low:.4f}, {ic_high:.4f}]")
print(f"Proporción de muestras bootstrap con recall < {threshold}:
{p_value_one_sided:.4f}")

# Interpretación simple automática (umbral alfa = 0.05)
alpha = 0.05
if ic_low >= threshold:
    print(f"INTERPRETACIÓN: El límite inferior del IC 95% (={ic_low:.4f})
>= {threshold}. "
```

```
          "Se puede concluir que el recall poblacional es al menos 0.85
(rechazamos H0).")
elif p_value_one_sided < alpha:
    print(f"INTERPRETACIÓN: p-valor unilaterl = {p_value_one_sided:.4f} <
{alpha}. "
          "Evidencia en favor de H1 (recall >= 0.85).")
else:
    print(f"INTERPRETACIÓN: No hay evidencia suficiente para aceptar H1
al nivel alfa={alpha}. "
          "No se rechaza H0 (recall poblacional < 0.85).")
```