Análisis predictivo de accidentes viales en Bogotá: un enfoque basado en ciencia de datos

Alexander Guerrero Caro

Asesor

Sebastián Vélez Jaramillo

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Especialización en Ciencia de Datos y Analítica

2025

Resumen

La ciudad de Bogotá D.C. enfrenta un elevado índice de accidentalidad vial, en el que se ven involucrados diversos actores o causas como automóviles, motociclistas, ciclistas, peatones, servicio público, infraestructura, mal estado de los vehículos, e incluso eventos atípicos como incendios. Este proyecto de grado aplica técnicas de ciencia de datos para analizar, crear, entrenar y seleccionar cuál es modelo óptimo para describir accidentes en Bogotá D.C., con un enfoque particular en los actores más recurrentes. Los hallazgos buscan aportar insumos valiosos para la toma de decisiones en políticas públicas orientadas a la reducción de los accidentes y la mejora de la seguridad vial.

Palabras clave: Datos, análisis de datos, accidentalidad, modelos predictivos, toma de decisiones.

Abstract

The city of Bogotá D.C. faces a high rate of road accidents, involving various actors or causes, such as automobiles, motorcyclists, cyclists, pedestrians, public services, infrastructure, poor vehicle condition, and even atypical events like fires. This thesis applies data science techniques to analyze, create, train, and select the optimal model for predicting accidents in Bogotá D.C., with a particular focus on the most common actors. The findings seek to provide valuable input for decision-making in public policies aimed at reducing accidents and improving road safety.

Keywords: Data, Data Analysis, Accident Rate, Predictive Models, Decision Making.

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Introducción

La ciudad de Bogotá D.C. enfrenta una problemática creciente en términos de accidentalidad vial, que involucra principalmente a peatones, motociclistas, ciclistas y conductores de vehículos particulares. Estos siniestros no solo representan un reto en materia de seguridad vial, sino que también generan consecuencias significativas a nivel social y económico, afectando en mayor proporción a personas jóvenes en edad productiva.

A pesar de los esfuerzos institucionales, normativas y campañas de prevención promovidas por las entidades reguladoras de la movilidad, los niveles de accidentalidad continúan siendo alarmantes. Una de las posibles limitaciones radica en la falta de herramientas analíticas que permitan anticiparse a estos eventos y orientar la toma de decisiones basada en datos. En este contexto, la ciencia de datos adquiere gran relevancia como alternativa para comprender los factores asociados a los accidentes de tránsito y describir su ocurrencia bajo los parámetros del modelo seleccionado.

El presente proyecto de grado tiene como objetivo principal crear, entrenar y seleccionar un modelo de regresión que permita determinar cuáles son los factores determinantes en la problemática de accidentalidad en la ciudad de Bogotá D.C., utilizando distintas variables y técnicas propias de la ciencia de datos. Para ello, se implementan tres enfoques de regresión: regresión lineal simple, regresión lineal múltiple y regresión no lineal, evaluando su desempeño para identificar el modelo más adecuado.

Descripción del Problema

La accidentalidad vial representa una problemática de alcance global que afecta a millones de personas cada año. En diversas partes del mundo, esta situación ha sido abordada mediante estrategias apoyadas en la tecnología, el análisis de datos y la formulación de políticas públicas fundamentadas en evidencia. Aun así, los accidentes de tránsito siguen siendo una de las principales causas de lesiones, discapacidad e incluso muerte, especialmente entre personas en edad productiva.

Colombia no es ajena a esta realidad. En el contexto nacional, la accidentalidad vial constituye un problema persistente en materia de seguridad y bienestar social, con impactos profundos en la salud pública, la economía y la calidad de vida de los ciudadanos. Particularmente en la ciudad de Bogotá D.C., los siniestros viales que involucran a actores como peatones, motociclistas, ciclistas y conductores de vehículos particulares generan anualmente un número considerable de víctimas. Muchas de estas personas sufren secuelas físicas y cognitivas que afectan gravemente su vida personal, familiar y laboral.

A pesar de los esfuerzos institucionales, como la creación de normas, campañas de concientización y mejoras en la infraestructura, los niveles de accidentalidad se mantienen altos. Una de las principales limitaciones radica en la posible falta de herramientas que permitan anticiparse a los accidentes con base en datos históricos y condiciones contextuales. En este sentido, la ciencia de datos emerge como una alternativa para comprender los factores que influyen en la ocurrencia de estos eventos y construir modelos predictivos que sirvan de insumo para la toma de decisiones informadas.

El presente proyecto parte de esta necesidad, buscando aplicar técnicas de análisis y modelado predictivo para identificar variables, relaciones y comportamiento que influyen en la

accidentalidad vial en Bogotá D.C. y, con ello, contribuir en la toma de decisiones informadas y en la formulación de estrategias orientadas a la reducción del riesgo y la protección de los actores más vulnerables.

Justificación

La alta accidentalidad vial en la ciudad de Bogotá D.C., especialmente entre actores como automovilistas, motociclistas, ciclistas y peatones, representa un serio problema de seguridad y salud pública, que conlleva elevados costos sociales, económicos y humanos. Si bien existen políticas públicas, campañas de prevención y mejoras en la infraestructura vial, los esfuerzos realizados hasta el momento no han logrado una disminución significativa en la frecuencia ni en la gravedad de los siniestros.

Una de las posibles falencias en la gestión de esta problemática es la falta de análisis profundo y sistemático de los datos disponibles, que permita comprender los factores contextuales, temporales y espaciales que influyen en la ocurrencia de los accidentes. En este contexto, la ciencia de datos se posiciona como una herramienta clave para transformar la información en conocimiento útil, mediante la identificación de patrones, puntos críticos y variables determinantes en la accidentalidad vial.

Este proyecto adquiere relevancia al aplicar modelos de regresión que permiten determinar la correlación e importancia de las variables que utiliza la Secretaria de Movilidad de Bogotá D.C., en el registro, descripción y caracterización de un Siniestro Vial. En donde los resultados obtenidos pueden ser utilizados por entidades gubernamentales, organismos de control y actores sociales para priorizar intervenciones, asignar recursos de manera más eficiente y diseñar estrategias focalizadas según el tipo de actor vial y la localización del riesgo.

En consecuencia, esta propuesta no solo tiene un aporte académico y técnico, sino también un valor práctico y social, al contribuir con evidencia empírica a la formulación de políticas orientadas a la prevención de accidentes y a la mejora integral de la seguridad vial en Bogotá D.C.

Objetivos

Objetivo General

Analizar la correlación y relevancia entre variables asociadas a la accidentalidad vial en Bogotá D.C., mediante técnicas de análisis exploratorio de datos y la aplicación y evaluación de modelos de regresión, con el fin de identificar los factores más influyentes en la ocurrencia de siniestros viales.

Objetivos Específicos

Recopilar y consolidar datos sobre accidentalidad vial en Bogotá D.C., a partir de diversas fuentes disponibles.

Procesar y transformar los datos recopilados que permitan una mejor comprensión de los factores asociados a la accidentalidad.

Evaluar y comparar el desempeño de tres modelos de regresión mediante métricas estadísticas (como R^2, RMSE y MAE) con el fin de identificar el modelo que mejor describa el comportamiento de los datos.

Marco Teórico

La accidentalidad vial constituye un fenómeno social complejo, influenciado por múltiples factores humanos, técnicos, ambientales y estructurales. A nivel global, la Organización Mundial de la Salud (OMS) ha identificado los accidentes de tránsito como una de las principales causas de muerte, especialmente entre personas jóvenes en edad productiva. En países como Colombia, y particularmente en ciudades densamente pobladas como Bogotá D.C., esta problemática tiene efectos significativos en la salud pública, la movilidad y la economía.

Figura 1

Total Accidentes Para el Año 2023

Figura 2

Total de Accidentes por Mes

Figura 3

Accidentes por Nivel de Gravedad

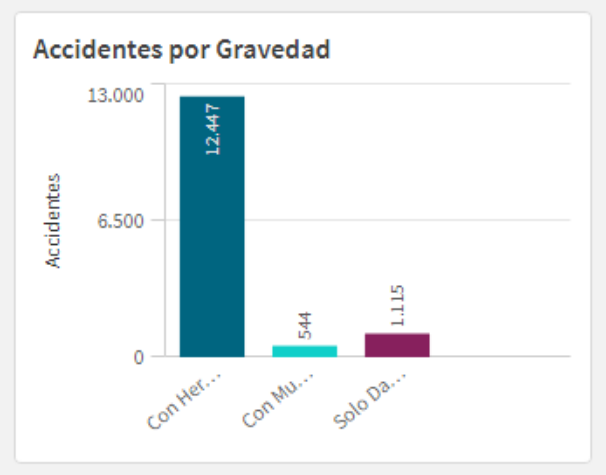
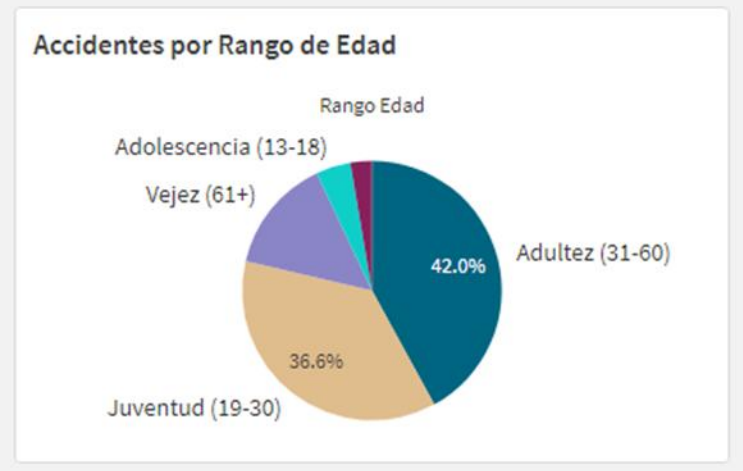


Figura 4

Accidentes Por Rango de Edad

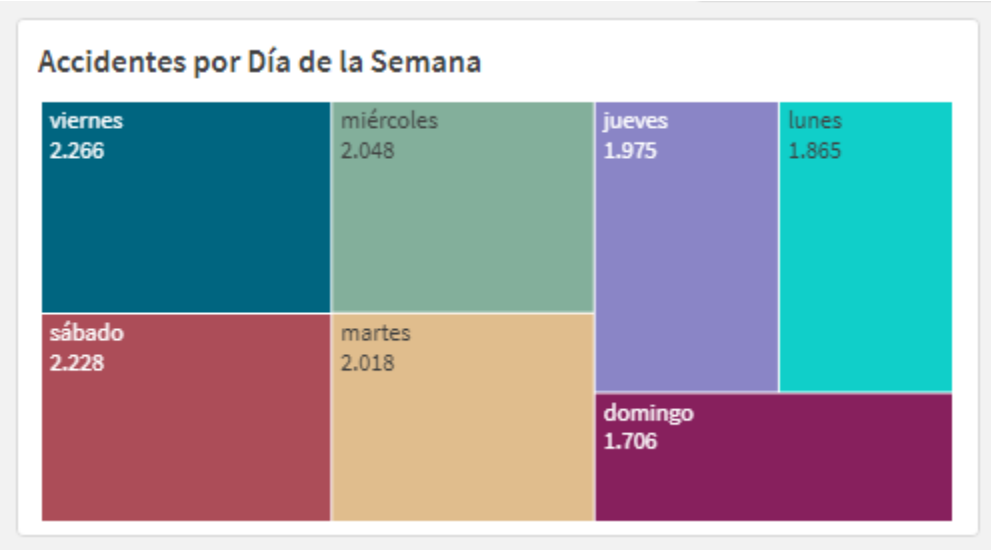


Asimismo, publicaciones de la (OMS) informan que cada año se cuentan con 1.9 millones de muertes producto de accidentes de tránsito, lo que equivale a un individuo cada dos minutos, es decir, 3200 fallecimientos diarios. En este contexto, la ciencia de datos permite

abordar el problema desde una perspectiva cuantitativa, aplicando técnicas de análisis exploratorio, minería de datos y modelado predictivo, lo que facilita la identificación de patrones de comportamiento y la estimación del riesgo en diferentes escenarios urbanos.

Figura 5

Accidentes por Día de la Semana



Entre los métodos más utilizados para determinar la correlación entre variables continuas se encuentran los modelos de regresión, que permiten establecer relaciones entre una variable dependiente como la gravedad y una o más variables independientes como tipo de actor vial, día de la semana, hora, entre otras). En este proyecto se implementan tres tipos de regresión, pero antes de describirlos, nos gustaría abordar un poco de contexto sobre modelos de regresión.

¿Qué es un Modelo de Regresión?

Es un modelo matemático que busca determinar la relación entre una variable dependiente (Y), con respecto a otras variables independientes (X).

Con el análisis de regresión, es posible modelar la relación entre las variables elegidas,

así como predecir valores basándose en el modelo.

El uso de un modelo de regresión puede ayudar en la solución de los siguientes tipos de problemas:

1. Determinar qué variables independientes están más relacionadas con la variable dependiente.
2. Comprender una relación existente entre las variables dependientes e independientes.
3. Predecir valores de la variable dependiente.

Para lograr abordar nuestros objetivos planteados, se han seleccionado tres modelos de regresión, donde cada uno cuenta con características y enfoques que, en base a su resultado permiten una medición y comparación integral de desempeño, los modelos que serán trabajados son los siguientes:

Regresión Lineal Simple

Es una regresión lineal con una variable independiente, y una variable dependiente, en donde se tiene como característica que la variable dependiente es continua.

Este modelo de regresión lineal simple ayuda en la ejecución de predicciones y a permite la compresión existente entre una variable dependiente y una variable independiente.

Regresión Lineal Múltiple

Este modelo permite generar un modelo lineal en el que el valor de la variable dependiente (Y), se logra determinar a partir de un conjunto de variables independientes a las cuales son llamadas predictoras (X1, X2, X3…Xn).

Esta regresión puede ser considerada una extensión de la regresión lineal simple, con la diferencia de que este permite emplearse para predecir el valor de la variable dependiente o para

evaluar la influencia que tienen los predictores sobre ella.

Regresión no Lineal

Este modelo de regresión no lineal es una forma de análisis de regresión en la que los datos observacionales se modelizan mediante una función que es una combinación no lineal de los parámetros del modelo.

Este desempeña un papel vital en la interpretación de los datos, ayudando a comprender las complejas relaciones existentes entre las variables.

Este modelo funciona según el principio de minimizar la suma de los cuadrados de los residuos, es decir, la diferencia entre los valores observados y los predichos, lo cual comúnmente es conocido como término de error.

Para evaluar el desempeño de estos modelos, se utilizan métricas estadísticas como el coeficiente de determinación (R^2), el error absoluto medio (MAE), el error cuadrático medio (MSE) y la raíz del error cuadrático medio (RMSE). Estas métricas permiten cuantificar la precisión del modelo y comparar su capacidad predictiva, seleccionando así el modelo más adecuado para el problema planteado.

Para el desarrollo de los objetivos planteados en este proyecto de grado, la selección de los anteriores modelos de regresión fue vital, dado que la combinación de estos tres modelos permite una evaluación entre la simplicidad, capacidad predictiva e interpretativa, como también permite comparar cuál se ajusta mejor a los datos disponibles.

Métricas de Desempeño

Las métricas de desempeño son herramientas cuantitativas estándar que permite evaluar la calidad de las predicciones de un modelo, midiendo la diferencia entre los valores predichos y los valores reales.

Estas métricas serán cruciales para evaluar los distintos modelos trabajados, por este motivo se tomarán en consideración las siguientes:

R2 (Coeficiente de Determinación)

Esta métrica es el coeficiente de determinación, que nos indica que tanta variación tiene la variable dependiente que se puede predecir desde la variable independiente.

Cuando el coeficiente de determinación (R2) es utilizada, todas las variables independientes que se estén utilizando dentro del modelo, contribuyen a su valor.

Para está métrica el mejor valor posible obtenido es 1, y el peor valor obtenido es 0. Aunque un punto crítico de está métrica, es que está asume que todas las variables dentro del modelo ayudan a explicar su variación en la predicción, y esta afirmación no siempre se cumple.

MSE (Mean Squared Error)

Esta métrica es muy útil, para poder determinar qué tan cerca es la línea de ajuste de nuestra regresión a las observaciones. Como también en caso contrario, se logra evitar que un error con valor positivo anule a un valor negativo, dando la ilusión de una mejora del modelo.

MAE (Mean Absolute Error)

La métrica de error absoluto medio es una medida de la diferencia entre dos valores, lo que nos indica que tan diferente es el valor predicho y el valor real y observado.

Dado que para el desarrollo de este proyecto de grado nos interesa conocer el comportamiento del error de todas las observaciones y no solamente de una, está métrica nos permitirá obtener el valor promedio de los valores absolutos de la diferencia.

RMSE (Raíz del MSE)

Sabiendo que en la métrica MSE, obtenemos un resultado en unidades cuadradas, esta métrica nos permite una interpretación más fácil, sacando la raíz cuadrada del resultado

obteniendo el resultado en unidades enteras.

En conjunto, este marco teórico sustenta la aplicación de métodos exploratorios y algoritmos predictivos como una vía eficiente para comprender y mitigar la problemática de los accidentes viales en entornos urbanos complejos como Bogotá D.C.

Metodología

Este proyecto adopta un enfoque cuantitativo y exploratorio, centrado en el análisis de relaciones entre variables asociadas a la accidentalidad vial en Bogotá D.C. A través de técnicas de análisis exploratorio de datos (EDA), estadísticas descriptivas y análisis de correlación, se busca identificar patrones, asociaciones significativas y variables con mayor influencia en los siniestros viales.

Recolección y Selección de Datos

En esta etapa, se realizó la investigación de diferentes fuentes de información, en donde se creó un banco de opciones, posteriormente, se eliminaron las fuentes que no tenían información importante para el entrenamiento de los modelos, después, se seleccionó la data más idónea para el proyecto de grado que cual proviene del portal oficial de datos abiertos de la Secretaría de Distrital de Movilidad de Bogotá D.C.

Fuente de Información

La fuente de información utilizada en el presente proyecto utiliza los datos abiertos proporcionados por la Secretaría Distrital de Movilidad de Bogotá D.C., disponibles en el portal oficial de datos abiertos de la ciudad (datos.movilidadbogota.gov.co). Este conjunto de datos contiene registros detallados de siniestros viales, incluyendo variables como fecha, hora, tipo de actor vial involucrado, ubicación geográfica (localidad, barrio, intersección), número de víctimas, y condiciones del entorno.

Esta información es clave para desarrollar modelos predictivos que reflejen la realidad urbana de Bogotá y permitan tomar decisiones fundamentadas en evidencia.

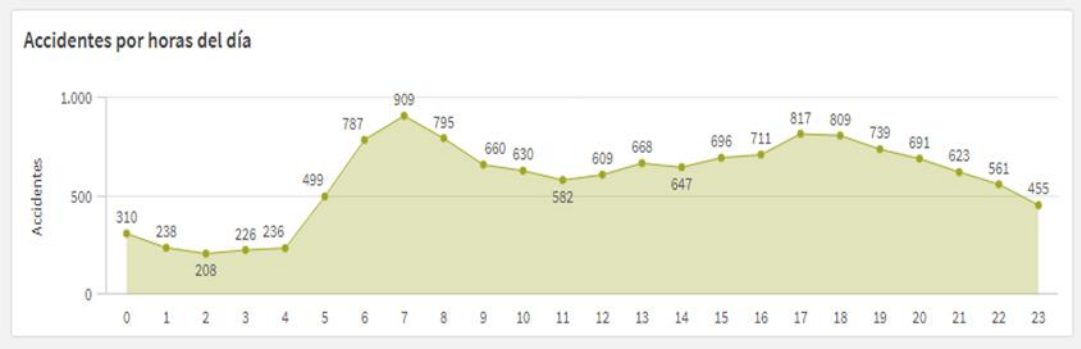
Adicionalmente, se hace uso de herramientas de análisis exploratorio de datos (EDA) para entender la distribución, correlaciones y características generales del conjunto de datos. La

visualización de datos, mediante gráficos y mapas, también cumple un rol fundamental en la identificación de zonas críticas de accidentalidad y en la comunicación efectiva de los hallazgos.

Este conjunto de datos contiene registros históricos detallados de siniestros viales, incluyendo variables como Fecha y hora del accidente, variable de tipo date, donde se registra la fecha exacta del accidente y una hora muy cercana al accidente. Esta variable es funcional, porque los modelos pueden llegar a ser entrenados en fechas especiales, para ver su comportamiento y validar si estos eventos causan algún efecto en la accidentalidad.

Figura 6

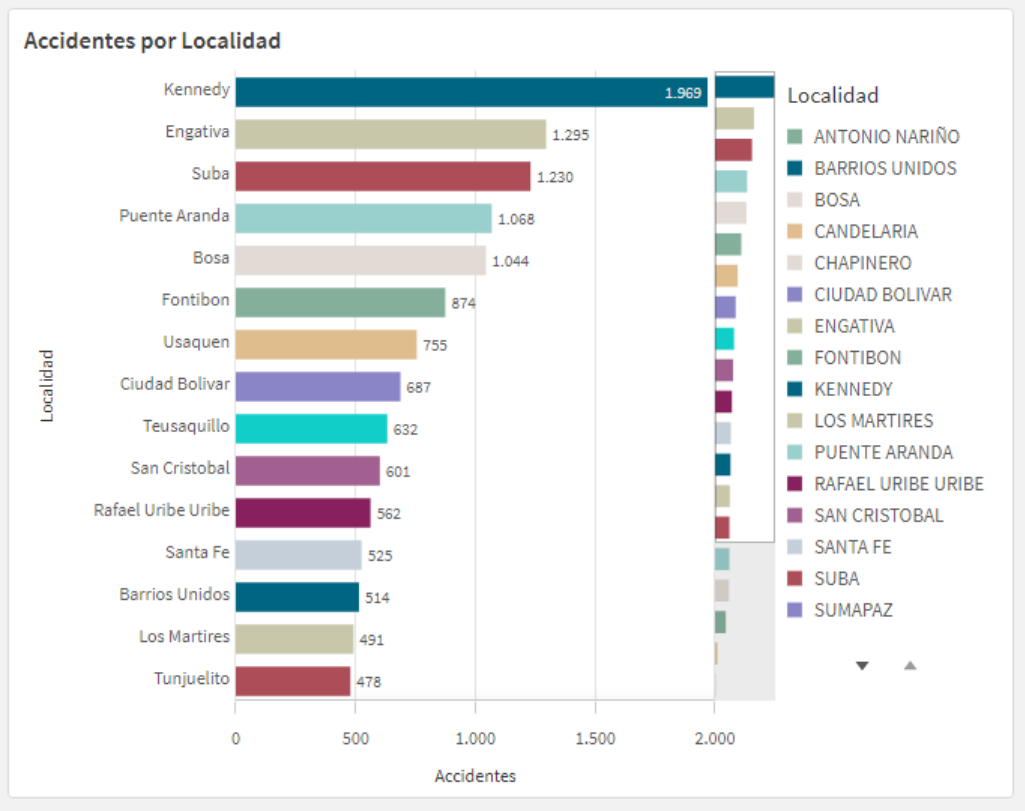
Accidentes por Horas del Día



Localidad: Variable de tipo cadena (String), en donde se evidencia en que zona geoespacial de Bogotá D.C., ocurrió el accidente. Esta variable además de ser valiosa para el entrenamiento de los modelos es una variable vital para la toma de decisiones y asignación de recurso humano que ayude en la gestión de disminución de siniestros viales.

Figura 7

Accidentes por Localidad



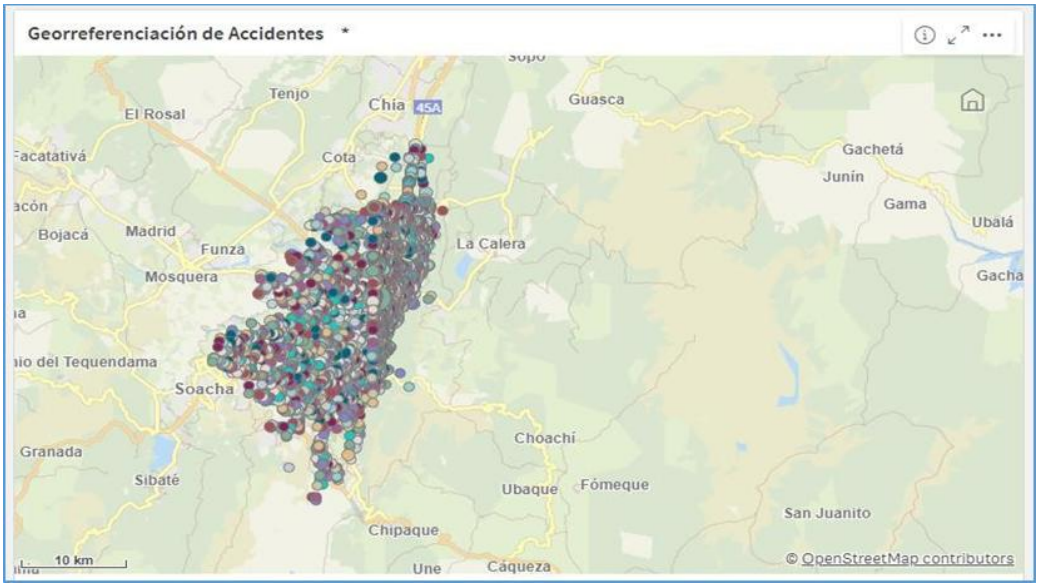
Georreferenciación: Variable que será compuesta de las variables Latitud y Longitud, la cual permite una demarcación más precisa de la zona del siniestro vial.

Figura 8

Georreferenciación de los Accidentes de Tránsito

Figura 9

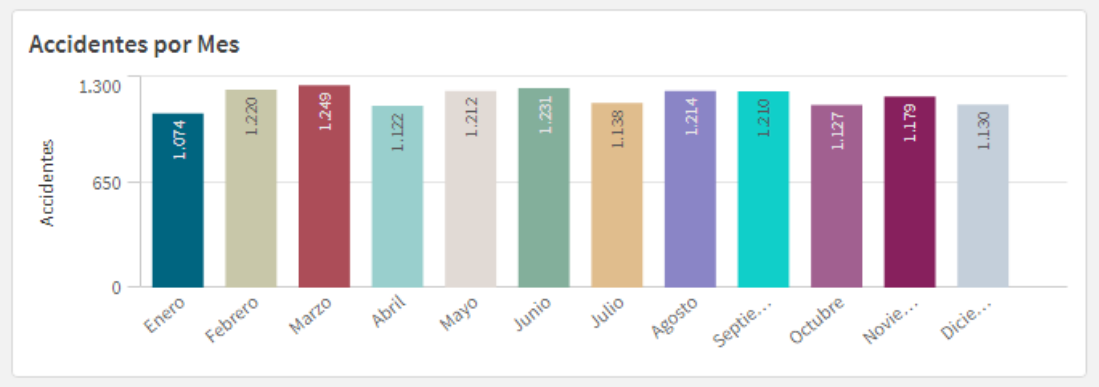
Accidentes Geo Ubicados - Bogotá, D.C.



Mes: Esta variable es de tipo fecha y es obtenida de la fecha, indicando el mes exacto donde ocurrió uno o varios accidentes. Esta variable es importante para el entrenamiento y predicción de datos del modelo, porque podemos llegar a un nivel de detalle donde se podría indicar algo como "Para el mes 04, 05 y 06, se ha obtenido un promedio de accidentes "X", lo que indica que para el mes 07 podría ocasionarse "Y" número de accidentes."

Figura 10

Frecuencia de Accidentes por Mes



Procesamiento de Datos

En esta etapa del proyecto se realiza una limpieza y transformación del conjunto de datos seleccionado, lo que puede incluir:

- Eliminación o imputación de registros incompletos o inconsistencias.
- Conversión de variables categóricas a variables numéricas.
- Generación de nuevas variables a partir del campo fecha (día de la semana, hora, mes y número de mes).
- Detección y tratamiento de valores atípicos.
- Normalización o escalamiento de variables cuando es necesario para los modelos.

Así mismo, se realiza un análisis exploratorio de datos (EDA) con el fin de visualizar patrones, correlaciones entre variables y tendencias temporales o geográficas que permitan la predicción de accidentes.

Construcción y Entrenamiento de Modelos

Se construyen y entrenan los modelos de regresión seleccionados para este proyecto de grado en base a las variables seleccionadas mediante el EDA y el resultado obtenido en una matriz de correlación, teniendo el siguiente comportamiento.

- Regresión Lineal: Modelo que será entrenado con una sola variable.
- Regresión Lineal Múltiple: Modelo que será entrenado con múltiples variables predictoras.
- Regresión no Lineal: Posible modelos polinomiales o transformaciones no lineales que serán entrenados, si la relación entre las variables lo requiere.

Evaluación y Comparación de Modelos

Dado nuestro objetivo principal, en donde se tiene la creación, entrenamiento, mejora y entrenamiento de modelos de regresión, seleccionando al más idóneo.

Esta selección no deberá de ser al azar, o el que más rápido reacciones a los datos entregados, por contrario se hará uso de las métricas estándar destinadas para este tipo de modelos, de las cuales haremos uso.

Coeficiente de Determinación (R2)

Esta métrica indica que proporción de la variable dependiente puede ser explicada por el modelo. Su valor oscila entre 0 y 1, donde los valores cercanos a 1 indican un mejor ajuste.

Criterio: Se dará prioridad a modelos con un R2 más alto, pero no será la única métrica decisiva, esto teniendo en cuenta que un R2 alto puede llegar a ser engañoso si los errores (MSE

y RMSE) son altos.

Error Absoluto Medio (MAE)

Mide el promedio de los errores absolutos entre los valores reales y los valores predichos. Es fácil de interpretar y no penaliza fuertemente los errores grandes. Criterio: Se buscará un MAE lo más bajo posible, lo que indica que en promedio las predicciones del modelo se alejan poco de los valores reales.

Error Cuadrático Medio (MSE)

Está métrica calcula el promedio de los errores al cuadrado. Penalizando más fuertemente los errores grandes, lo cual lo hace sensible a valores atípicos.

Criterio: Se evaluará que el MSE sea bajo, y que no exista un aumento considerable frente al MAE, lo que podría indicar la presencia de predicciones con errores extremos.

Raíz del Error Cuadrático Medio (RMSE)

Esta métrica literalmente es tomada como la raíz cuadrada del MSE. Donde tiene las mismas unidades que la variable a predecir, lo que facilita su interpretación directa, además también penaliza más los errores grandes.

Criterio: Se considerará deseable un RMSE bajo, y será una métrica clave para comparar modelos entre sí, especialmente en términos de su aplicabilidad real.

Criterio Global de Selección del Modelo Óptimo

El modelo seleccionado, y quien será el encargado de futuras predicciones de la accidentalidad en Bogotá D.C., será aquel que presente un equilibrio adecuado entre las cuatro métricas, teniendo en consideración la siguiente configuración.

- Alto R2: Pero sin que MSE y RMSE se eleven.
- MAE, MSE y RMSE bajos y consistentes: Que estas métricas tengan diferencias

pequeñas entre ellas, lo cual nos entregaría un modelo estable.

En caso de que un modelo tenga un R2 levemente más bajo, pero conserve mejores métricas de error (especialmente en RMSE), podría ser ese el seleccionado como óptimo, dado que su error promedio real es menor y es más confiable en la predicción.

Herramientas Utilizadas

En la fase de desarrollo(código) del proyecto, se hará uso del entorno notebooks de Python, haciendo uso de la plataforma de Google de Google Colab, la cual permitirá un trabajo asincrónico entre nosotros (estudiantes especialización), además que nos facilita y permite el uso de las siguientes librerías:

- Pandas y Numpy: Estas librerías permitirán la manipulación y análisis de datos.
- Matplotlib y Seaborn: Librerías que permitirán hacer visualizaciones de gráficos para comparar comportamientos.
- Scikit-learn: Esta librería permitirá el uso de modelos predictivos y gestionar las métricas de medición.

Limpieza y Transformación de la Data

- Transformación de la variable mesAcc (1 al 12): 1= enero, 2=febrero, 3=marzo, 4=abril, 5=mayo, 6=junio, 7=julio, 8=agosto, 9=septiembre, 10=octubre, 11=noviembre, 12=diciembre.
- Transformación de la variable Localidad: 1=Usaquén, 2=Chapinero, 3=Santa Fe, 4=San Cristóbal, 5=Usme, 6=Tunjuelito, 7=Bosa, 8=Kennedy, 9=Fontibón, 10=Engativá, 11=Suba, 12=Barrios Unidos, 13=Teusaquillo, 14=Los Mártires, 15=Antonio Nariño, 16=Puente Aranda, 17=La Candelaria, 18=Rafael Uribe Uribe, 19=Ciudad Bolívar, 20=Sumapaz.
- Transformación de la variable Gravedad_Indicador_Tradicional: 1= Solo Daños,

2= Con Heridos, 3=Con Muertos

- Transformación de la variable Clase_Acc: 1=Choque, 2=Atropello, 3=Volcamiento, 4=Caída de ocupante, 5=Incendio, 6=Otro
- Transformación de la variable Dia_Semana_Acc: 1=lunes, 2=martes, 3=miércoles, 4=jueves, 5=viernes, 6=sábado, 7=domingo.
- Eliminación de variables: Formulario, Longitud, Latitud, Dirección. Estas variables fueron eliminadas por considerar que no representaban valores significativos en la descripción de la accidentalidad y en su combinación con otras variables no generaría cálculos o métricas para el análisis de los datos.
- Imputación de la variable: Elemento_Choque, con la categoria "NO REGISTRA", cuando en la columna Clase_Acc se igual a Choque y Elemento_Choque sea vacío o nulo.
- Binarización de columnas: Las columnas cuyo nombre inicie con "Con", 1=SI y 0=No.

Resultados

Una vez finalizado todo el proceso de análisis exploratorio de datos (EDA), y transformación de estos, se prosiguió con la creación, entrenamiento y ajustes de los diferentes modelos planteados en este proyecto de grado, para estos modelos se estableció como variable dependiente "Gravedad_Num", la cual es resultado de la transformación de la data de la variable "Gravedad_Indicador_Tradicional" a valores numéricos, quedando como se mencionó en el punto cinco punto seis (5.6).

Cada modelo fue evaluado bajo los mismos criterios de validación, empleando el mismo conjunto de variables, claro está, cuando aplicaba.

La elección de estos modelos responde al interés de comparar enfoques progresivamente más complejos y observar su capacidad para capturar la relación entre las variables independiente y la gravedad del accidente.

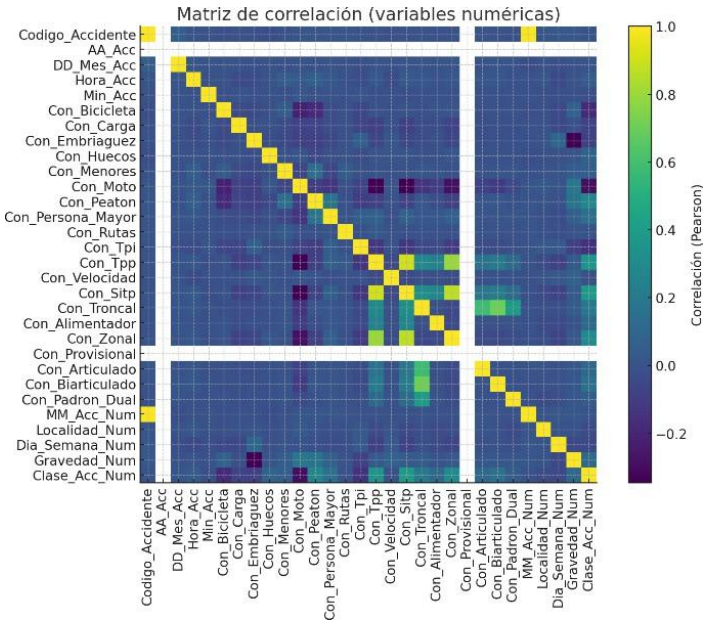
A continuación, se presentarán los resultados obtenidos para cada uno de los modelos, junto con una comparación de sus métricas de desempeño; Teniendo presente el objetivo de determinar cuál ofrece un mejor ajuste y una mayor capacidad predictiva para la problemática expuesta en este proyecto de grado.

Variables Utilizadas

A continuación, se presentan las variables utilizadas y con mayor relevancia dentro de los tres modelos de regresión que fueron evaluados, para lo cual se analizó la matriz de correlación dando como resultado lo referenciado en la Figura 11.

Figura 11

Matriz de Correlación



Panorama de la Correlación

La mayoría de las asociaciones con Gravedad_Num son bajas a moderadas: |r| típicamente < 0.20, salvo Con_Embriaguez (~0.335 en valor absoluto).

Ninguna variable por sí sola explica bien la gravedad → los modelos simples tienen R^2 bajo y los multivariados/no lineales mejoran la predicción, pero no de forma significativa.

Entre predictores no se observan colinealidades extremas generalizadas (pocos pares con |r| > 0.8), pero las interacciones pueden ser relevantes (no las capta Pearson).

Usamos la variable Con_Embriaguez porque (en valor absoluto) es de las más asociadas con la gravedad en el dataset. El signo negativo sólo indica dirección de la relación dadas las escalas y no resta valor como predictor

Con_Embriaguez fue la mejor por correlación absoluta ⇒ modelo simple y también entra en los otros modelos. Múltiple (Top-5): tomó las 5 con mayor |r| (Embriaguez, Moto, Peatón,

Clase_Acc_Num, Tpi) ⇒ $R^2 \sim 0.18$.

Polinómica g2 + Ridge: añadió continuas y dummies con interacciones y cuadráticos ⇒ R^2 CV ~0.193 (mejora sobre simple y múltiple), aunque algunas de esas variables tengan |r| bajo de forma individual.

Regresión Lineal Simple

- Variable dependiente: Gravedad_Num
- Variable independiente: Con_Embriaguez

Figura 12

Aplicación del Modelo de Regresión Lineal Simple

Los valores de las métricas obtenidas en este modelo fueron los siguientes:

Tabla 1

Métricas Obtenidas del Modelo de Regresión Lineal Simple

Métrica	Valor	%
R^2	0.126	12.6
MAE	0.121	12.1
MSE	0.099	9.9
RMSE	0.315	3.15

Este modelo explica aproximadamente el 12.6% de la variación en la gravedad del accidente, lo que lo convierte en el mejor modelo de regresión lineal simple entre las variables disponibles.

Regresión Lineal Múltiple

- Variable dependiente: Gravedad_Num
- Variables independientes sin multicolinealidad: Estas variables fueron utilizadas por las razones expuestas en la explicación realizada en la Tabla 2.

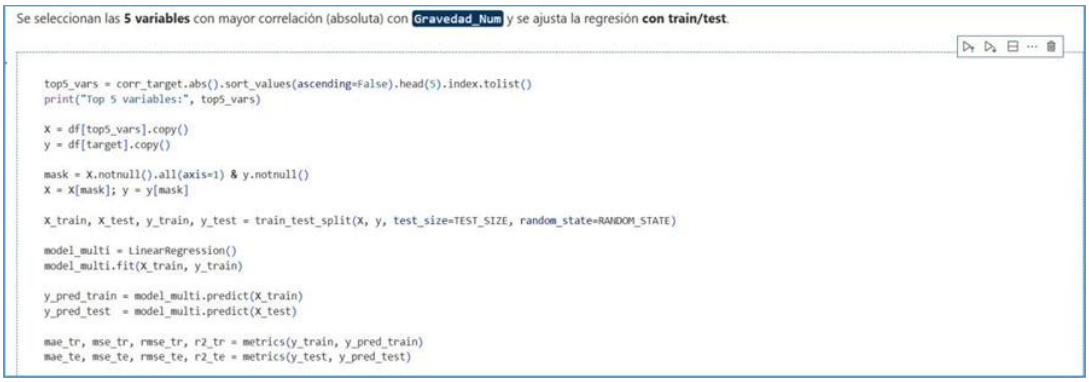
Tabla 2

Variables Independientes Modelo de Regresión Lineal Múltiple

Variable	Explicación
Con_Embriaguez	Tienen fuerte correlación con accidentes graves.
Con_Moto	Actor vulnerable o de alto riesgo.
Con_Peaton	Actor vulnerable o de alto riesgo.
Clase_Acc_Num	Tipo de siniestro (choque, atropello, etc.)
Con_tpi	Vehículo de transporte público individual involucrado

Figura 13

Fragmento del Código Desarrollado en Python



Los valores de las métricas obtenidas en este modelo fueron los que se evidencian en la siguiente tabla:

Tabla 3

Métricas Obtenidas del Modelo de Regresión Lineal Múltiple

Métrica	Valor	%
R^2	0,196	19.6
MAE	0,152	15.2
MSE	0,091	9.1
RMSE	0,302	30.2

El modelo logró un coeficiente de determinación (R^2) del 19.6%, lo que significa que explica cerca del 20% de la variabilidad en la gravedad de los accidentes. Los errores de predicción fueron moderados, así:

- MAE: 0.152.
- MAPE: 9.76%, lo que indica un bajo error relativo medio.

Regresión no Lineal - Polinómica Grado 2

Variable dependiente:

- Gravedad_Num

Variables independientes:

- Con_Embriaguez
- Con_Velocidad
- Con_Moto
- Con_Peaton
- Con_Persona_Mayor
- Clase_Acc_Num
- Hora_Acc
- Dia_Semana_Num
- Localidad_Num

Los valores de las métricas obtenidas en este modelo fueron los siguientes:

Tabla 4

Métricas Obtenidas del Modelo de Regresión No Lineal - Polinómico Grado2

Métrica	Valor	%
R^2	0,220	22
MAE	0,148	14,8
MSE	0,0886	8,86
RMSE	0,298	29,8

La variable Con_Embriaguez es, por mucho, la más influyente individualmente y en

combinación con otras. Combinaciones como Con_Velocidad * Con_Moto y Con_Peaton * Con_Embriaguez también tienen un efecto considerable en la gravedad. Esto sugiere que la interacción entre factores de riesgo (como velocidad y tipo de actor vial) aumenta significativamente la gravedad.

Por qué el R^2 fue bajo (~0.22): Señal débil en las variables (x): las correlaciones individuales con Gravedad_Num son pequeñas (máx|r|≈0.33), así que, aun sumando variables, el poder explicativo es limitado.

Objetivo ordinal y poco variable: Gravedad_Num solo toma {1,2,3} y está sesgada hacia "Solo Daños"; poca variación reduce el R^2.

Predictores en gran parte binarios: con 0/1 la separación posible es acotada; eso pone un "techo" natural al R^2.

Factores omitidos/importantes no modelados: no se dispone de variables como vía, clima, velocidad, localización, etc., que podrían explicar gran parte de la gravedad.

Relaciones complejas: aunque se usó polinómica grado 2, aún quedan no linealidades/ interacciones que el modelo no captura totalmente.

Variables que podrían faltar para unos resultados y análisis más precisos:

Con respecto a las vías y al entorno: Tipo de vía (autopista, arterial, residencial), n.º de carriles, separador, pendiente/curvatura. Intersección (sí/no, rotonda, cruce semaforizado, paso peatonal). Iluminación (día/noche/iluminado), señalización (presencia/estado), estado de la calzada (seco/mojado/bacheado). Límite de velocidad oficial y velocidad medida/estimada.

Con respecto al clima: Precipitación (lluvia/niebla), intensidad y duración previo al siniestro. Visibilidad (metros o categorías), viento.

Factores humanos: Fatiga/somnolencia, distracción (teléfono móvil), exceso de

velocidad, experiencia del conductor.

Confiabilidad del modelo: Capacidad explicativa limitada con $R^2 \approx 0.22$, el modelo solo captura ~22% de la variación de Gravedad_Num. No es malo per se, pero no es suficiente para predicciones finas a nivel caso.

Error absoluto moderado: RMSE≈0.30 (en rango 1–3) implica un error típico de ~0.3 puntos de gravedad. Es decir, puede confundir "Solo daños" (1) con algo cercano a "Heridos" (2) en bastantes casos.

Estabilidad aceptable: la validación cruzada externa mostró desviaciones bajas, lo que sugiere poca varianza (no está sobre ajustado). El modelo es consistente, solo que poco potente con las variables actuales.

Señal débil de las variables: las correlaciones individuales son bajas/moderadas; por eso, aunque el modelo sea estable, su poder predictivo sigue siendo limitado.

Mejoras a realizar para futuros estudios:

- Enriquecer el modelo con las variables antes mencionadas.
- Mejorar la representación de la variable objetivo (Y).
- Utilizar modelos que capturen mejor las no linealidades (Gradient Boosting / XGBoost / LightGBM).
- Calidad de datos y gobernanza de los mismos (completos, consistentes, actualizados, relevantes y uniformes).

Comparación de las Métricas Obtenidas

Con el objetivo de identificar el modelo más adecuado, en función de las variables con correlación más relevantes, se realizó una comparación de las métricas estándar de desempeño, permitiendo contrastar su capacidad explicativa, así como el nivel de error en las predicciones

generadas. A continuación, se presenta el resumen de las métricas obtenidas por cada modelo.

Tabla 5

Comparativa de las Métricas Obtenidas en los Modelos Entrenados

Métrica	Regresión Lineal Simple		Regresión Lineal Múltiple		Regresión No Lineal - Polinómica Grado 2	
	Valor	%	Valor	%	Valor	%
R^2	0.126	12.60%	0.196	19.6	0.220	22
MAE	0.121	12.10%	0.152	15.2	0.148	14.8
MSE	0.099	9.90%	0.091	9.1	0.0886	8.86
RMSE	0.315	31.50%	0.302	30.2	0.298	29.8

Regresión Lineal Simple (Con_Embriaguez) Este modelo fue el más simple de todos, pero mostró una capacidad explicativa limitada ($R^2 = 0.126$), a pesar de esto, logró una baja tasa de error (MAPE ≈ 8.5%), lo que sugiere que la embriaguez es un factor importante en la gravedad del accidente. La variable Hora_Acc, en cambio, no tuvo capacidad predictiva significativa ($R^2 \approx 0$), descartándola como predictor individual.

Regresión Lineal Múltiple (con 9 variables seleccionadas) Este modelo integró factores de riesgo humanos, tipo de accidente, tiempo y localización. Asimismo, logró un R^2 de 0.196, explicando casi el 20% de la variabilidad de la gravedad, fue más robusto que el modelo simple, y conservó buena interpretabilidad, el MAPE (9.76%) indica un error relativo aceptable para fines analíticos.

Regresión No Lineal (Polinómica de Grado 2), Capturó relaciones cuadráticas e interacciones entre variables. Presentó un R^2 de 0.220 y mejoró las métricas de error (MAE =

0.148) frente al modelo múltiple lineal. Las interacciones como Embriaguez * Peatón y Velocidad * Moto demostraron tener un efecto significativo en la gravedad.

Conclusiones

El presente proyecto de grado demostró la aplicabilidad y el valor de la ciencia de datos en el análisis y correlación existente entre las variables asociadas en la accidentalidad vial en la ciudad de Bogotá D.C., lo cual es una problemática persistente que afecta gravemente la seguridad y el bienestar de los ciudadanos. A partir de un enfoque cuantitativo, se logró construir y evaluar modelos de regresión que permiten evidenciar la relevancia en función de variables históricas y contextuales.

Entre los principales hallazgos, se resalta que los modelos de regresión, especialmente la regresión lineal múltiple, lograron un desempeño aceptable en términos de precisión y capacidad predictiva, evidenciado por métricas como el coeficiente de determinación (R^2), el error absoluto medio (MAE) y la raíz del error cuadrático medio (RMSE). Estos resultados reflejan que, aunque el fenómeno de la accidentalidad vial es complejo y multifactorial, existen patrones estadísticos aprovechables para anticipar su ocurrencia bajo ciertas condiciones.

Asimismo, el análisis exploratorio y la visualización de datos permitieron identificar tendencias relevantes, como las localidades con mayor concentración de siniestros, los días y horarios de mayor riesgo, y los actores viales más recurrentemente involucrados. Esta información puede convertirse en un insumo valioso para las autoridades de tránsito, al facilitar la formulación de estrategias preventivas más focalizadas y basadas en evidencia.

Finalmente, se concluye que el uso de modelos de regresión en la gestión de la seguridad vial no solo es viable, si no necesario en contextos urbanos como en la ciudad de Bogotá D.C., donde la alta densidad vehicular y la vulnerabilidad de ciertos actores hacen imperativa una toma de decisiones informadas.

Este proyecto establece una base sólida para futuras investigaciones orientadas a la

incorporación de modelos más sofisticados, así como variables externas relevantes como: condiciones climáticas, estado de la infraestructura vial, deficiencias en la señalización, presencia de conductores bajo el efecto de sustancias psicoactivas o eventos especiales.

Recomendaciones

Incorporar variables de contexto adicionales, como lo podrían ser datos socioeconómicos, densidad vehicular, zonas escolares y reportes de infracciones, con el fin de aumentar la capacidad explicativa del modelo seleccionado.

Implementar el modelo dentro de una plataforma de visualización interactiva (Dashboard) que permita a entidades como la Secretaría de Movilidad de Bogotá identificar en tiempo real zonas de alta probabilidad de siniestros y planificar intervenciones preventivas.

Actualizar y reentrenar periódicamente el modelo con datos nuevos, para asegurar su vigencia y precisión, además de extender su aplicación a otras ciudades o departamentos para evaluar patrones comunes y específicos de accidentalidad vial a nivel nacional.

Promover alianzas interinstitucionales entre entidades públicas y universidades o centros de investigación, para fortalecer la recolección y estandarización de los datos, validar hallazgos y aplicar de forma coordinada las herramientas analíticas desarrolladas.

Dado el avance tecnológico, es vital la actualización de las librerías usadas, o migrarlas de ser necesario, en este caso, se recomienda hacer la migración de la librería de pandas, por la de polars, dado que esta librería está mucho más optimizada y es mucho más veloz.

Bibliografía

A pesar de los notorios progresos, la seguridad vial sigue siendo un problema apremiante para el mundo. (s.f.). (2023). https://www.who.int/es/news/item/13-12-2023-despite-notable-progress-road-safety-remains-urgent-global-issue

Amat, R. (2016). *Introducción a la Regresión Lineal Múltiple*. https://cienciadedatos.net/documentos/25_regresion_lineal_multiple

Análisis de regresión. (s.f.). (2025). https://doc.arcgis.com/es/insights/latest/analyze/regression-analysis.htm

Barrera, S. (2023). *1.9 millones de personas fallecen por accidentes de tránsito: OMS.* https://consultorsalud.com/1-9-millones-pers-accidentes-de-transito-oms/

Base Anuario de Siniestralidad 2023. (n.d.). Gov.co. Retrieved June 4, 2025, from https://datos.movilidadbogota.gov.co/documents/7b8d619e7fa84f0a8a451c6e618a6b69/about

Bock, T. *What is a Correlation Matrix?*. (s.f.). https://www.displayr.com/what-is-a-correlation-matrix/

Cabrera, G., Velásquez, N., Valladares, M. (2009). *Seguridad vial, un desafío de salud pública en la Colombia del siglo XXI.* https://revistas.udea.edu.co/index.php/fnsp/article/view/260/1873

Correlación. (s.f.). https://www.jmp.com/es/statistics-knowledge-portal/what-is-correlation

Introducción a Polars para el Tratamiento de Datos y Comparativa con Pandas. (2024). https://www.vernegroup.com/actualidad/tecnologia/introduccion-polars-tratamiento-datos-comparativa-pandas/

La librería Numpy. (s.f.). (2022). https://aprendeconalf.es/docencia/python/manual/numpy/

Pandas: La biblioteca de Python dedicada a la Data Science. (2022, December 19). DataScientest. https://datascientest.com/es/pandas-python

Qué es la regresión Lasso. (s.f.). https://www.ibm.com/es-es/think/topics/lasso-regression

Regresión no lineal. (s.f.). https://www.studysmarter.es/resumenes/ingenieria/matematicas-de-la-ingenieria/regresion-no-lineal/

Regresión Lineal Simple. (2024). https://www.datacamp.com/es/tutorial/simple-linear-regression

Roldán, P. (2022). *Modelo de regresión*. https://economipedia.com/definiciones/modelo-de-regresion.html

Seguridad vial. (n.d.). Paho.org. Retrieved June 4, 2025, from https://www.paho.org/es/temas/seguridad-vial

Terra, J. (2024, August 14). What is data imputation, and how can you use it to handle missing data? Caltech -. https://pg-p.ctme.caltech.edu/blog/data-science/what-is-data-imputation-for-missing-data

Waskom, M. (2024). *Seaborn: Statistical data visualization*. https://seaborn.pydata.org/