**Logotipo

Descripción generada automáticamente**

**Análisis predictivo de accidentes de tránsito en Bogotá**

Jose Luis Mena Palomeque

Juan Sebastián Espinosa Uribe

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesor  
Nombres completos, Título académico más alto

Universidad de Antioquia  
Facultad de Ingeniería

Especialización en Analítica y Ciencia de Datos

Medellín, Antioquia, Colombia

2024

|  |  |
| --- | --- |
| **Cita** | (Espinosa Uribe & Mena Palomeque, 2024) |
| **Referencia**  **Estilo APA 7 (2020)** | Espinosa Uribe, J. S., & Mena Palomeque, J. L. (2024). *Análisis predictivo de accidentes de tránsito en Bogotá* Trabajo de grado especialización]. Universidad de Antioquia, Medellín, Colombia. |

**** 

Especialización en Analítica y Ciencia de Datos, CohorteVIII.

Centro de Investigación Ambientales y de Ingeniería (CIA).

|  |  |
| --- | --- |
|  | Diagrama  Descripción generada automáticamente con confianza media |

Centro de Documentación Ingeniería (CENDOI)

**Repositorio Institucional:** http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes.

Decano: Julio Cesar Saldarriaga Molina

Jefe departamento: Diego José Luis Botia Valderrama

El contenido de esta obra corresponde al derecho de expresión de los autores y no compromete el pensamiento institucional de la Universidad de Antioquia ni desata su responsabilidad frente a terceros. Los autores asumen la responsabilidad por los derechos de autor y conexos.

**Tabla de contenido**

[Resumen 6](#_heading=h.3rdcrjn)

[Abstract 7](#_heading=h.26in1rg)

[1. Descripción del problema 8](#_heading=h.35nkun2)

[1.1. Problema de negocio 10](#_heading=h.1ksv4uv)

[1.2. Aproximación desde la analítica de datos 10](#_heading=h.44sinio)

[1.3. Origen de los datos 11](#_heading=h.z337ya)

[1.4. Métricas de desempeño 11](#_heading=h.3j2qqm3)

[2. Objetivos 13](#_heading=h.4i7ojhp)

[2.1. Objetivo general 13](#_heading=h.2xcytpi)

[2.2. Objetivos específicos 13](#_heading=h.3whwml4)

[3. Datos 14](#_heading=h.2bn6wsx)

[3.1. Datos originales 14](#_heading=h.qsh70q)

[3.2. Datasets 14](#_heading=h.3as4poj)

[3.3. Analítica descriptiva 17](#_heading=h.1pxezwc)

[Referencias 25](#_heading=h.1jlao46)

**Lista de figuras**

[**Figura 1** Gráfico QQ Plot (a la izquierda) y gráfico de densidad para la variable "Edad" (a la derecha)](#_heading=h.2r0uhxc) 19

[**Figura 2** Gráfico de frecuencia de personas accidentadas por localidad](#_heading=h.1664s55) 20

[**Figura**](#_heading=h.1664s55) **3** Gráfico de frecuencia de tipo de accidente. 20

[**Figura**](#_heading=h.1664s55) **4** Gráfico de frecuencia por tipo de vehículo accidentado 21

[**Figura**](#_heading=h.1664s55) **5** Gráfico de frecuencia por condición de la persona accidentada. 21

[**Figura**](#_heading=h.1664s55) **6** Gráfico de frecuencia por gravedad del accidente. 22

[**Figura**](#_heading=h.1664s55) **7** Tendencia de personas accidentadas en Bogotá por día. 23

[**Figura**](#_heading=h.1664s55) **8** Gráfica de estacionalidad 24

[**Figura**](#_heading=h.1664s55) **9** Gráfica de volatilidad 24

[**Figura**](#_heading=h.1664s55) **10** Gráfica de correlaciones. Parte superior: función de autocorrelación (ACF). Parte inferior: función de autocorrelación parcial (PACF). 25

[**Figura**](#_heading=h.1664s55) **11** Identificación de Picos y Anomalías. 26

**Siglas, acrónimos y abreviaturas**

**ACF** Autocorrelation Function (Función de Autocorrelación).

**AIC** Akaike Information Criterion (Criterio de Información de Akaike).

**BIC** Bayesian Information Criterion (Criterio de Información Bayesiano).

**GIS** Geographic Information Systems (Sistemas de Información Geográfica).

**KNN** K-Nearest Neighbors (K Vecinos Más Cercanos).

**MAE** Mean Absolute Error (Error Absoluto Medio).

**MB** Mean Bias (Sesgo Medio).

**ML** Machine Learning (Aprendizaje Automático).

**NUMPY** Librería de Python para computación numérica (Numerical Python).

**PACF** Partial Autocorrelation Function (Función de Autocorrelación Parcial).

**PANDAS** Librería de Python para análisis de datos (Python Data Analysis).

**RMSE** Root Mean Square Error (Raíz del Error Cuadrático Medio).

**SARIMA** Seasonal AutoRegressive Integrated Moving Average (Modelo ARIMA estacional).

**SMOTENC** Synthetic Minority Over-sampling Technique for Nominal and Continuous data (Técnica de sobre muestreó sintético para datos nominales y continuos).

**SVM** Support Vector Machines (Máquinas de Soporte Vectorial).

# Resumen

Este proyecto tiene como propósito desarrollar modelos predictivos para estimar la frecuencia y la gravedad de los accidentes de tránsito en Bogotá, empleando datos históricos proporcionados por la Secretaría Distrital de Movilidad. Los datos incluyen información geográfica, demográfica y temporal que, tras un proceso de pre procesamiento y análisis descriptivo, permitieron identificar patrones y anomalías. La estrategia metodológica combinó técnicas de aprendizaje automático y series de tiempo elegidos por su capacidad para manejar problemas de clasificación y predicción. Entre los principales retos se encuentra el manejo de desbalance de la variable de salida (gravedad del accidente) e integración de la frecuencia y gravedad en un indicador, lo cual requirió un enfoque cuidadoso en la limpieza y preparación de los datos. Los resultados obtenidos incluyen un indicador de riesgo por localidades, construido a partir de las predicciones de frecuencia y gravedad, que facilita la priorización de zonas críticas. Este trabajo no solo contribuye a una mejor comprensión de la siniestralidad vial, sino que también genera insumos valiosos para el diseño de estrategias orientadas a la seguridad vial y la optimización de recursos.

*Palabras clave*: Modelos predictivos, Aprendizaje automático, Series de tiempo, Accidentes de tránsito.

Repositorio GitHub:

https://github.com/DCJuanLab/Monograf-a

# Abstract

The purpose of this project is to develop predictive models to estimate the frequency and severity of traffic accidents in Bogotá, using historical data provided by the Secretaría Distrital de Movilidad. The data includes geographic, demographic and temporal information that, after a preprocessing and descriptive analysis process, allowed identifying patterns and anomalies. The methodological strategy combined machine learning techniques and time series chosen for their ability to handle classification and prediction problems. Among the main challenges was the handling of imbalance of the output variable (accident severity) and integration of frequency and severity into an indicator, which required a careful approach to data cleaning and preparation. The results obtained include a risk indicator by locality, constructed from the frequency and severity predictions, which facilitates the prioritization of critical areas. This work not only contributes to a better understanding of road accidents, but also generates valuable inputs for the design of strategies aimed at road safety and the optimization of resources.

*Keywords*: Predictive models, Machine learning, Time series, Traffic accidents.

GitHub repository:

https://github.com/DCJuanLab/Monograf-a

# Descripción del problema

La siniestralidad vial es un problema crítico en las ciudades (OMS, 2023), particularmente en Bogotá, donde la alta densidad poblacional y el creciente flujo vehicular han incrementado la frecuencia y la gravedad de los accidentes de tránsito. Según los datos de la Secretaría Distrital de Movilidad, desde 2022 se ha observado una tendencia creciente en los siniestros viales. A pesar de este aumento en la cantidad de accidentes graves, Bogotá ha logrado reducir su tasa de mortalidad, alcanzando 7 muertes por cada 100.000 habitantes, una cifra significativamente menor a la media nacional. Además, en comparación con 2022, se ha reportado un incremento del 0.8% en los siniestros graves en 2023 (Secretaría Distrital de Movilidad, 2023). Estos datos resaltan la importancia de entender las dinámicas de los accidentes para optimizar recursos y diseñar estrategias eficaces que mitiguen el impacto de estos eventos en la ciudad, tanto en términos de seguridad vial como en la capacidad de respuesta de los sistemas de salud y emergencias.

Uno de los principales retos es la capacidad de tomar decisiones informadas para abordar la siniestralidad vial en su conjunto. La optimización de los recursos de emergencia, como ambulancias y equipos médicos, es solo una de las áreas que puede beneficiarse de un análisis predictivo de la frecuencia y gravedad de los accidentes (Coffman, 2022; Kang & Cheong, 2023). Sin embargo, este análisis también puede tener un impacto significativo en otras áreas, como el diseño de estrategias de prevención y la asignación de recursos para la mejora de infraestructuras viales. Por ejemplo, la predicción precisa de los patrones de siniestralidad puede ayudar a identificar las zonas de mayor riesgo, lo que permitiría a las autoridades implementar medidas preventivas más efectivas, como la mejora de señales viales, la implementación de normativas de seguridad más estrictas o incluso campañas de sensibilización para la población. Asimismo, contar con información predictiva sobre la gravedad de los accidentes podría contribuir a un enfoque más proactivo en el diseño de políticas públicas orientadas a la seguridad vial, facilitando la toma de decisiones que prioricen las áreas más críticas de manera oportuna y eficiente (Gutierrez Osorio & Pedraza, 2019).

La literatura relacionada con la predicción de la gravedad de los accidentes de tránsito se centra principalmente en el análisis de factores determinantes, como el estado de las personas involucradas, la velocidad de los vehículos implicados y sus características, los factores ambientales. Estos estudios han implementado modelos de aprendizaje automático categóricos, como las máquinas de soporte vectorial (SVM), árboles de decisión y técnicas de aprendizaje profundo; estos últimos han demostrado ser efectivos en la clasificación de la gravedad de los accidentes (Ali, Hussain, & Haque, 2024; Santos, Saias, Quaresma, & Nogueira, 2021; Silva, Andrade, & Ferreira, 2020). Con respecto a la predicción de la frecuencia de accidentes, se han empleado modelos de aprendizaje automático como Random Forest y K-Nearest Neighbors (KNN), además de metodologías que combinan técnicas de aprendizaje automático con herramientas de sistemas de información geográfica (GIS). Esta combinación ha resultado útil para identificar áreas con gran cantidad de accidentes y anticipar zonas de riesgo, como se observa en los trabajos de Khan y Hussain (2024). En Bogotá, se han realizado estudios para determinar los factores de mayor influencia en los accidentes de tránsito mediante minería de datos descriptiva, analizando información geográfica, temporal y la gravedad de los accidentes (Gutierrez Osorio & Pedraza, 2019).

Desde el análisis de datos, este trabajo contribuye al entendimiento de los accidentes de tránsito en Bogotá mediante la predicción de la frecuencia y la gravedad de estos. A través de modelos de aprendizaje automático, se analizaron datos históricos de accidentes, incluyendo información geográfica, demográfica y temporal. Este enfoque permitió identificar patrones y anomalías relevantes, generando información clave para anticipar zonas de riesgo y apoyar la toma de decisiones estratégicas orientadas a la seguridad vial.

La fuente principal de los datos incluye registros de accidentes viales en Bogotá durante 2023

de la secretaría de movilidad, con información detallada por día y por localidad. Estos datos permitieron analizar tanto la frecuencia de eventos como la cantidad de personas afectadas, proporcionando una base para realizar análisis descriptivos, identificar picos de siniestralidad y desarrollar modelos predictivos. A través de la integración de la predicción de frecuencia y gravedad, se construye un índice aproximado de riesgo por localidad que facilite la priorización de esfuerzos en seguridad vial y atención de emergencias.

## Problema de negocio

En las grandes ciudades, los accidentes de tránsito representan un riesgo constante para la seguridad pública y exigen una atención rápida y eficaz por parte de los servicios de emergencia. En Bogotá, durante el año 2023, se reportaron 543 muertes por accidentes de tránsito (Secretaría Distrital de Movilidad, 2023), lo que la posiciona como la ciudad con el mayor número de muertes en Colombia por accidentes de tránsito. Este incremento en la frecuencia y gravedad de los accidentes plantea desafíos significativos para la asignación de recursos críticos y para la implementación de políticas de prevención efectivas. Para abordar estos desafíos, se desarrollaron modelos predictivos que anticipan tanto la frecuencia de accidentes por localidad como la gravedad de los mismos, con el fin de establecer un índice aproximado de riesgo que combine ambas dimensiones y permita identificar el nivel de riesgo de accidente en cada localidad mediante una ponderación matemática. La información generada es de utilidad para abordar problemáticas claves de seguridad vial, optimizar la asignación de recursos de atención de emergencias y guiar el diseño de campañas de concientización vial.

## Aproximación desde la analítica de datos

La modelización de la predicción de los diferentes aspectos de las colisiones mediante machine learning ha evolucionado en las últimas dos décadas. Estos modelos se han convertido en una herramienta ampliamente utilizada en el campo de la accidentalidad vial (Ali, Hussain, & Haque, 2024).

Empleando modelos de clasificación, pudimos estimar las probabilidades de que un individuo sufra lesiones, fallezca o salga ileso en un accidente de tránsito dentro de las localidades de Bogotá. Paralelamente, los modelos de series temporales nos permiten pronosticar el número total de accidentes en una localidad específica, basándonos en su historial.

Para preparar los datos, unificamos las tablas del archivo de accidentalidad en un único conjunto de datos y abordamos los valores faltantes mediante imputación o eliminación y los registros duplicados fueron eliminados.

Tras seleccionar las variables más relevantes y aplicar técnicas como *one-hot encoding* y *label encoding*, obtenemos un conjunto de datos inicial, listo para ser procesado por el modelo. En caso de desbalanceo en las clases, utilizamos submuestreo o la técnica SMOTENC para equilibrar los datos.

Para el análisis de serie de tiempo, el dataset fue preparado para tener el conteo de accidentes ocurridos por día en Bogotá y a partir de ello estudiar su tendencia, autocorrelación, estacionalidad, variabilidad, volatilidad, picos y posibles anomalías.

El objetivo final es construir conjuntos de datos sólidos para entrenar, evaluar y ajustar los modelos para hacer una predicción precisa.

## Origen de los datos

Los datos provienen del Observatorio de Movilidad de Bogotá y contienen registros de las características de los siniestros viales ocurridos en las 20 localidades de la ciudad para el año 2023.

## Métricas de desempeño

**Métricas para evaluar la predicción de la gravedad de los accidentes**

Precisión: Mide la proporción de predicciones correctas sobre el total de predicciones. Es útil para evaluar la capacidad del modelo e identificar correctamente las zonas con mayor riesgo de accidentes. Se espera un valor cercano al 70% en esta métrica.

Recall: Mide la proporción de casos positivos (accidentes) que el modelo identificó correctamente. Es importante para asegurar que no se estén pasando por alto zonas con alto riesgo. Se espera un valor cercano al 70% en esta métrica.

F1-score: Es la media armónica de precisión y recall, proporcionando un equilibrio entre ambas métricas. Se espera un valor cercano al 70% en esta métrica.

Matriz de confusión: Permite visualizar la distribución de las predicciones correctas e incorrectas, lo que dará una idea más detallada del desempeño del modelo.

AUC-ROC: Mide la capacidad del modelo para distinguir entre clases (en este caso, zonas con y sin accidentes), independientemente del umbral de clasificación. Se espera un valor cercano al 0.7 en esta métrica.

Los valores aceptables para estas métricas fueron elegidos basándose en los resultados obtenidos en la comparación de modelos hecha por (Ahmed, Hossain, Ray, Bhuiyan, & Sabuj, 2023; Silva, Andrade, & Ferreira, 2020) .

**Métricas para Evaluar la predicción de la Frecuencia de Accidentes**

**Criterio de información de Akaike (AIC):** Analiza la calidad del modelo penalizando su complejidad mediante el número de parámetros utilizados, según la literatura, valores bajos indican un modelo más eficiente. En el estudio de Parvareh, y otros. (2018), el mejor modelo SARIMA (1,1,1) (0,0,1) [12] para predecir la frecuencia de accidente que involucran peatones obtuvo un **AIC = 684.98**, mientras que para motociclistas obtuvo un AIC **= 753.13**.

**Criterio de información bayesiano (BIC):** tiene la misma finalidad que el AIC, pero Introduce una penalización más estricta, tomando en cuenta el tamaño de la muestra. Esto favorece los modelos más simples y robustos al igual que el AIC se recomiendan valores bajos, En el trabajo de Parvareh, y otros. (2018), el modelo para la frecuencia de accidentes relacionados con peatones el modelo SARIMA (1,0,2) (1,0,1) [12] tuvo un **BIC = 694.03** y para motociclistas un **BIC = 766.79**.

**Error cuadrático medio (RMSE):** Mide la desviación estándar de los errores, evaluando la magnitud promedio del error al cuadrarlo. Es útil para interpretar cuánto se desvían las predicciones del modelo con respecto a los valores reales. Aunque no se especifican valores concretos en el estudio de Ye, Hu, & Wang. (2019), se buscó que el RMSE fuera lo más cercano posible a cero para asegurar alta precisión.

**Error absoluto medio (EMA):** Calcula la media de los errores absolutos entre las predicciones y los valores reales, ofreciendo una medida robusta frente a valores extremos. Similar al RMSE, se busca que el EMA sea mínimo para garantizar una alta precisión en la predicción.

**Métricas de Negocio**

**Reducción en el número de accidentes:** Comparación del número de accidentes en las zonas donde se implementaron acciones basadas en las predicciones del modelo versus zonas donde no se implementaron.

**Reducción en la gravedad de los accidentes:** Analizar si los accidentes en las zonas tratadas son menos graves (por ejemplo, menos heridos o fallecidos).

**Optimización de recursos:** Evaluar si la asignación de recursos (ambulancias, agentes de tránsito) es más eficiente gracias a las predicciones del modelo.

**Satisfacción de las partes interesadas:** Obtener feedback de las autoridades de tránsito, los ciudadanos y otras partes relevantes sobre la utilidad de las predicciones y las acciones implementadas.

# Objetivos

## Objetivo general

Analizar y predecir la frecuencia y la gravedad de los accidentes de tránsito en Bogotá mediante el uso de modelos de Machine Learning, para generar un indicador de riesgo por localidad que combine ambas predicciones y proporcione una herramienta útil para abordar problemáticas relacionadas con la accidentalidad, como la prevención de accidentes, campañas públicas de seguridad vial y la planificación de recursos de atención de emergencias.

## Objetivos específicos

1) Analizar los patrones de frecuencia y gravedad de los accidentes de tránsito en las localidades de Bogotá utilizando datos históricos.

2) Predecir la gravedad de los accidentes de tránsito utilizando modelos de clasificación, ajustando hiper parámetros y evaluando el desempeño con métricas como F1-Score, Recall y matriz de confusión.

3) Predecir la frecuencia de los accidentes en las localidades de Bogotá utilizando modelos de series temporales.

4) Construir un índice de riesgo por localidad, que combine las predicciones de la gravedad y frecuencia de los accidentes utilizando una ponderación aritmética.

# Datos

## Datos originales

Se obtuvieron los datos de accidentalidad vial de la Secretaría Distrital de Movilidad de Bogotá. Este conjunto de datos es de acceso público y puede consultarse oprimiendo clic en el [enlace](https://datos.movilidadbogota.gov.co/documents/7b8d619e7fa84f0a8a451c6e618a6b69/about).

El conjunto de datos original consta de un documento de Excel con tamaño de 14 MB, en el cual se registraron las características de 14.106 siniestros viales ocurridos en las 20 localidades de la ciudad de Bogotá para el año 2023.

Este documento está ordenado en tres tablas.

* **Siniestros:** Contiene la información general de cada accidente como código de identificación, ubicación y fecha.
* **Vehículos:** Incluye detalles sobre los vehículos involucrados en los accidentes.
* **Actor vial:** Recoge información sobre las personas implicadas en los accidentes, incluyendo sus roles y condiciones.

Los campos seleccionados para el estudio, así como los campos originales del set de datos se pueden consultar en el [Diccionario de Datos](https://github.com/DCJuanLab/Monograf-a/blob/main/Diccionario%20de%20Datos.xlsx).

## Datasets

A partir de los datos originales y utilizando las librerías de Pandas y Numpy, se construyeron conjuntos de datos específicos para el modelo de machine learning.

El primer paso fue la selección de los campos considerados relevantes de cada tabla del archivo para implementar un modelo de clasificación de la gravedad y predicción de la frecuencia de los accidentes.

# Seleccionar campos a utilizar del DataFrame siniestros.  
siniestros = siniestros[['Codigo\_Accidente','Longitud','Latitud','Fecha\_Acc','Hora\_Min\_Acc','Localidad','Clase\_Acc','Gravedad\_Indicador\_Tradicional',  
 'Con\_Bicicleta','Con\_Carga','Con\_Embriaguez','Con\_Huecos','Con\_Menores','Con\_Moto','Con\_Peaton','Con\_Persona\_Mayor','Con\_Velocidad','Con\_Trans\_Public']]  
  
# Seleccionar campos a utilizar del DataFrame vehículos.  
vehiculos = vehiculos[['Codigo\_Accidente','Codigo\_Vehiculo','Clase','Vehiculo\_Viajaba\_Clasificado']]  
  
# Seleccionar campos a utilizar del DataFrame actores.  
actores = actores[['Codigo\_Accidente','Codigo\_Vehiculo','Edad','Sexo','Gravedad\_Indicador\_Tradicional','Muerte\_Posterior','Condicion']]

***Código 1****: Selección de columnas relevantes de las tablas de siniestros, vehículos y actores para el análisis.*

Mediante la función *merge ()* se unieron los campos seleccionados, lo cual nos dejó con un data frame inicial de 33542 filas y 26 columnas.

Posteriormente, los valores faltantes de las variables binarias fueron rellenados con la palabra “No”, ya que estos campos únicamente tenían la palabra “Si” en caso de que algún registro presentara esa característica en particular. Para esto, se utilizó la función *fillna ()* de la librería Pandas, permitiendo que estos campos fueran completados de manera consistente. Fueron eliminados 1158 registros que no contenían información en campos esenciales como ‘sexo, edad y vehículos\_viajaba\_clasificado, utilizando la función *dropna ()* de pandas. Esta decisión se tomó porque se considera que estos campos son críticos para el análisis de la gravedad y frecuencia de los accidentes, los registros con información incompleta pueden generar sesgos o afectar la precisión, la eliminación se consideró adecuada ya que estos registros no representaban una porción significativa de los datos. Al verificar la duplicidad de registros, se encontraron 147 filas duplicadas que se eliminaron mediante la función *drop\_duplicates()* de Pandas.

#eliminacion de registros sin informacion  
df =df.dropna(subset=['Sexo', 'Edad']) # Se eliminan los datos que no tienen sexo o edad  
df = df[df['Vehiculo\_Viajaba\_Clasificado'] !='SIN INFORMACIÓN'] # se elimina los datos que estan sin informacion en el campo'Vehiculo\_Viajaba\_Clasificado'  
  
#Reiniciar el indice  
df.reset\_index(drop=True, inplace=True)  
#Se elimina Gravedad\_Indicador\_Tradicional\_x  
df = df.drop('Gravedad\_Indicador\_Tradicional\_x', axis=1)  
#df = df.drop('Codigo\_Accidente', axis=1)  
df = df.drop('Codigo\_Vehiculo', axis=1)

***Código 2****: Eliminación de registros con valores faltantes y duplicados para garantizar la calidad del conjunto de datos.*

El análisis de valores atípicos se aplicó a la única variable numérica del conjunto de campos seleccionado para predecir la gravedad (edad), utilizando un gráfico de caja y bigotes, el cual permitió identificar edades atípicas en los registros, específicamente aquellas mayores de 77 años. Se detectaron un total de 403 registros con edades fuera del rango esperado, siendo la edad más alta de 112 años. Dado que es posible que existan personas con edades avanzadas involucradas en accidentes, se decidió no eliminar estos registros, ya que podrían representar una parte legítima de la población.

Después de tratar los datos vacíos y eliminar los duplicados, se realizó una agregación de los datos por fecha (Fecha\_Acc) y localidad (Localidad). Este proceso permitió consolidar la cantidad de personas involucradas en accidentes para cada combinación de fecha y localidad. La transformación resultante generó un dataset estructurado para análisis de series de tiempo, ideal para modelar y predecir la frecuencia de accidentes diarios por localidad.

La construcción de los conjuntos de entrenamiento y validación se realizó considerando las diferencias entre los modelos de predicción de gravedad y de frecuencia de accidentes:

* División en conjuntos de entrenamiento y validación (Predicción de la gravedad)

La división de los datos se realizó utilizando la función train\_test\_split de la librería Scikit-learn, separando aleatoriamente el dataset en un 80% para entrenamiento y un 20% para prueba. Antes de esta división, se aplicaron varios pasos clave de procesamiento para garantizar que los datos estuvieran listos para su uso en el entrenamiento y la prueba, evitando que se introdujera sesgos o alteraciones en el modelo. Estos pasos incluyeron la detección de valores atípicos, la codificación de variables categóricas y la estandarización de la variable "Edad".

# Se crean los datasets de entrenamiento y prueba para las variables de entrada y salida

X\_train, X\_test, y\_train, y\_test = train\_test\_split(  
 X,  
 y.values.reshape(-1,1),  
 train\_size = 0.8,  
 random\_state = 123,  
 shuffle = True  
 )

*Código 3: División del conjunto de datos en conjuntos de entrenamiento y prueba utilizando la función train\_ test\_ split de la biblioteca Scikit-learn.*

* División en Conjuntos de Entrenamiento y Validación para la Predicción de la Frecuencia de Accidentes (Series de Tiempo)

Para la predicción de la cantidad de personas accidentadas a lo largo del año, se decidió dividir los datos de la siguiente manera: los primeros nueve meses del año se utilizarán como conjunto de entrenamiento, mientras que los últimos tres meses se reservaron para la validación del modelo. Esta estrategia permite al modelo aprender patrones y tendencias a partir de un conjunto de datos más amplio antes de ser evaluado en un período donde se espera un aumento en la frecuencia de accidentes.

## Analítica descriptiva

**Análisis Descriptivo de Variables Categóricas y Numéricas**

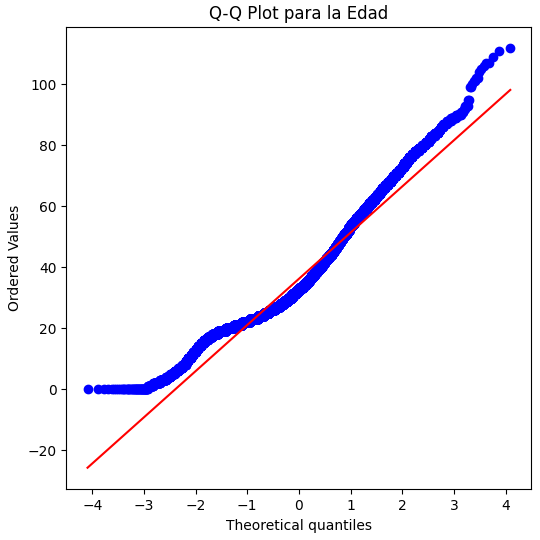
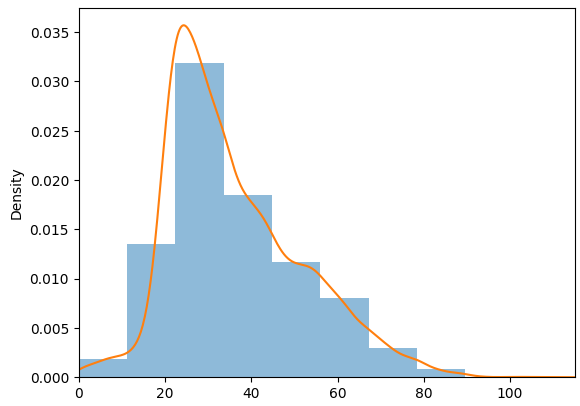
En esta sección, se decidió separar las variables categóricas de las numéricas para facilitar el análisis descriptivo. La mayoría de las columnas en el DataFrame son variables categóricas, mientras que la única variable numérica es 'Edad'. Esta separación permite aplicar técnicas de análisis específicas para cada tipo de variable y optimizar el estudio de las distribuciones, frecuencias y otras características relevantes de los datos.

Para las variables categóricas, se calcularon frecuencias y proporciones que permitieron identificar los valores predominantes y detectar posibles patrones. En el caso de la variable 'Edad', se analizó su distribución utilizando estadísticas descriptivas como la media, mediana, desviación estándar, rangos para entender su variabilidad dentro del dataset.

Análisis de la Distribución de la Variable Edad

Para evaluar la normalidad de la variable "Edad", se utilizaron tanto el gráfico Q-Q plot como pruebas estadísticas de Shapiro-Wilk y Kolmogórov-Smirnov. El Q-Q plot muestra un desvío significativo de los puntos en los extremos respecto a la línea de comportamiento normal (roja), la cual representa el comportamiento esperado si los datos siguieran una distribución normal. Este desvío indica que la distribución de "Edad" no es normal. Además, el histograma de frecuencias respalda esta observación, mostrando una distribución sesgada a la derecha, lo que sugiere una acumulación de valores en edades menores y una disminución en los valores de edades más altas.

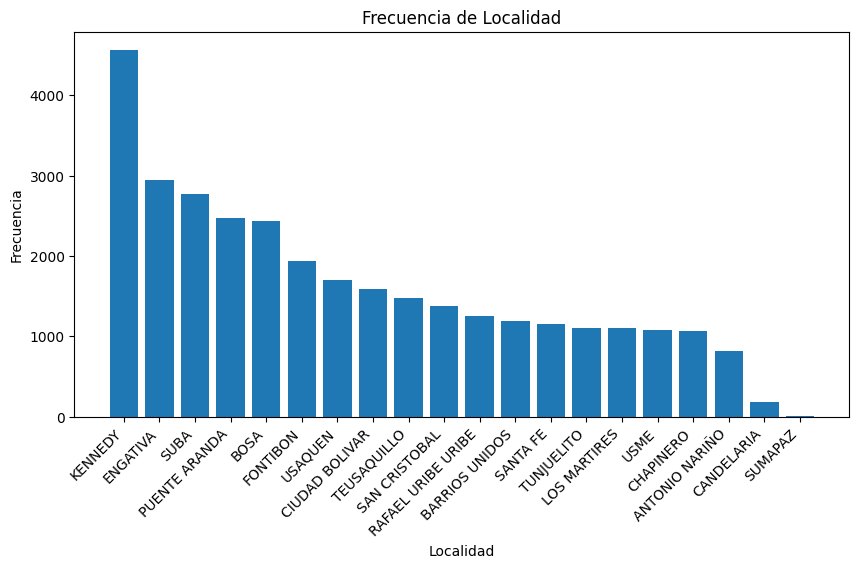
Las pruebas estadísticas confirman esta observación, ya que los valores p de Shapiro-Wilk y Kolmogorov-Smirnov son menores a 0.05, el nivel de significancia comúnmente aceptado. Esto significa que podemos rechazar la hipótesis nula de normalidad con un 95% de confianza, concluyendo que la variable "Edad" no sigue una distribución normal.

***Figura 1.*** *Gráfico QQ Plot (a la izquierda) y gráfico de densidad para la variable "Edad" (a la derecha)*

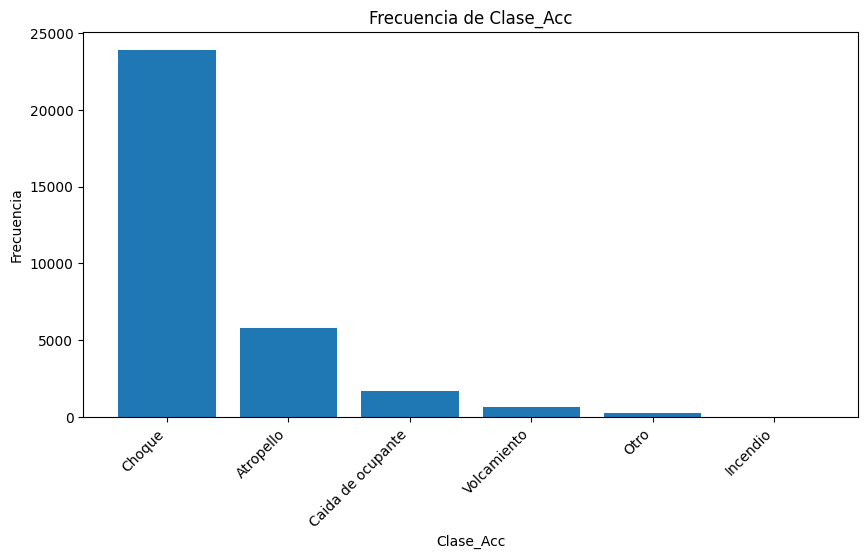
**Variables categóricas:**

* **Localidades**: El gráfico de barras de localidades muestra que Kennedy es la zona con la mayor cantidad de accidentes de tránsito, mientras que Sumapaz presenta la menor cantidad. Además, las cinco localidades con más personas involucradas en accidentes son Kennedy, Engativá, Suba, Puente Aranda y Bosa, las cuales concentran el 47% del total de incidentes registrados.



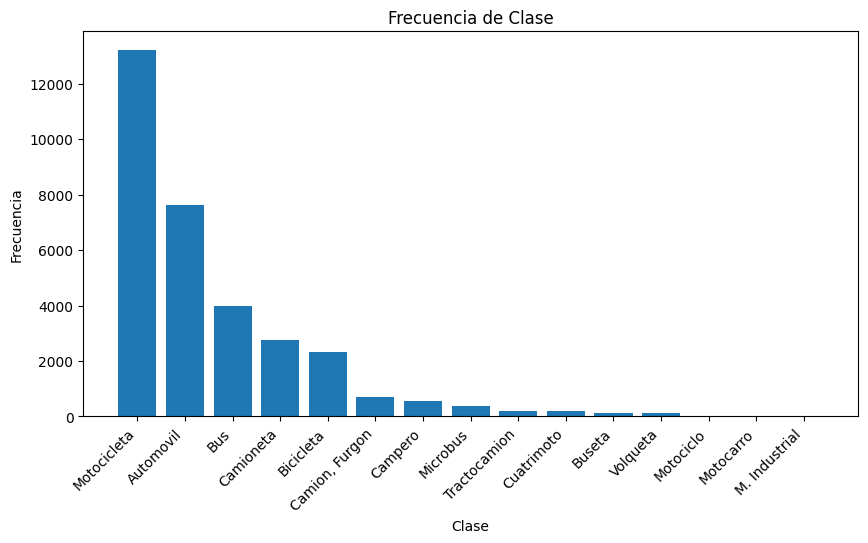
***Figura 2:*** *Gráfico de frecuencia de personas accidentadas por localidad*

* **Tipo de Accidente (Clase\_Acc)**: En el gráfico relacionado con la Clase de Accidente, se observa una clara predominancia de choques, que representan el 74% de los incidentes. Los atropellos ocupan el segundo lugar, con un 18% de los casos.

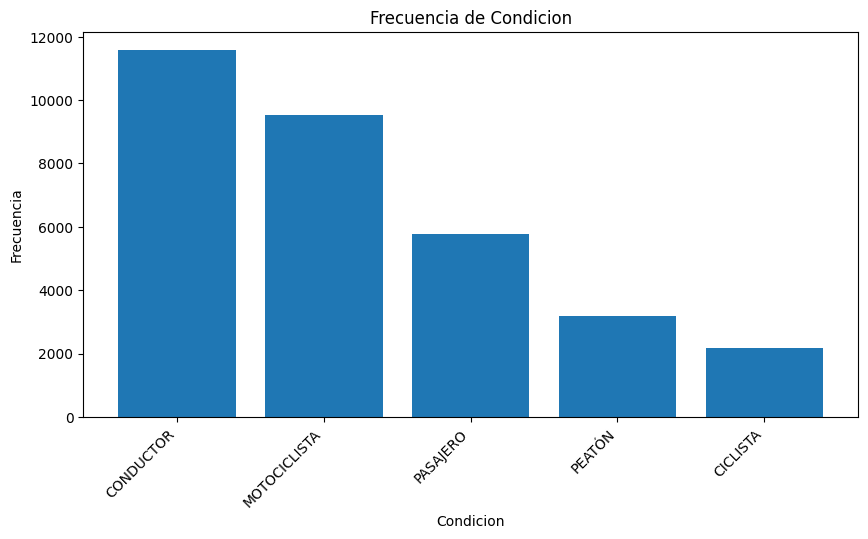


***Figura 3****: Gráfico de frecuencia de tipo de accidente.*

* **Involucrados por tipo de vehículo y condición**: Los gráficos sobre la frecuencia de vehículos clasificados y la condición de los accidentes revelan un patrón destacado: la mayoría de las personas involucradas en accidentes fueron motociclistas.

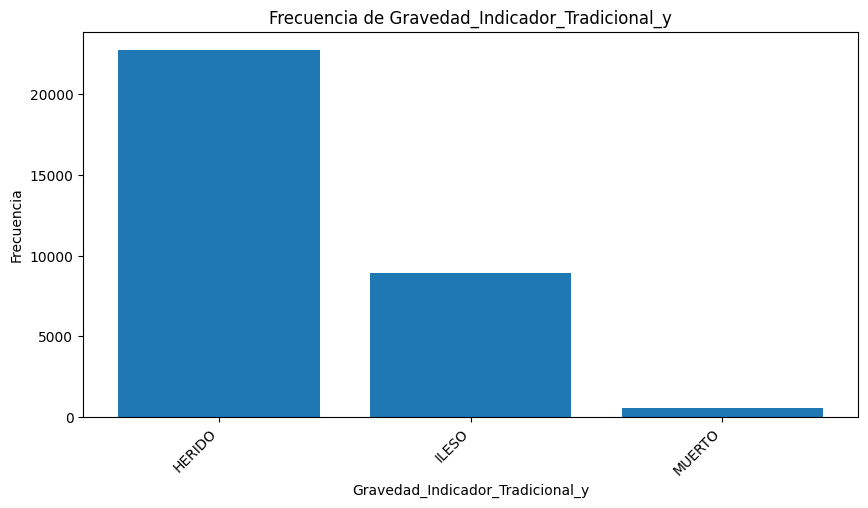


***Figura 4:*** *Gráfico de frecuencia por tipo de vehículo accidentado*



***Figura 5****: Gráfico de frecuencia por condición de la persona accidentada.*

* **Gravedad de los Accidentes**: En cuanto a la gravedad de los accidentes, se observa que el 71% de las personas resultaron heridas, el 28% fueron ilesas y solo un 1% sufrió fallecimientos. Esta distribución muestra que la variable de gravedad está desbalanceada, con una gran mayoría de accidentes que no resultan en víctimas fatales.

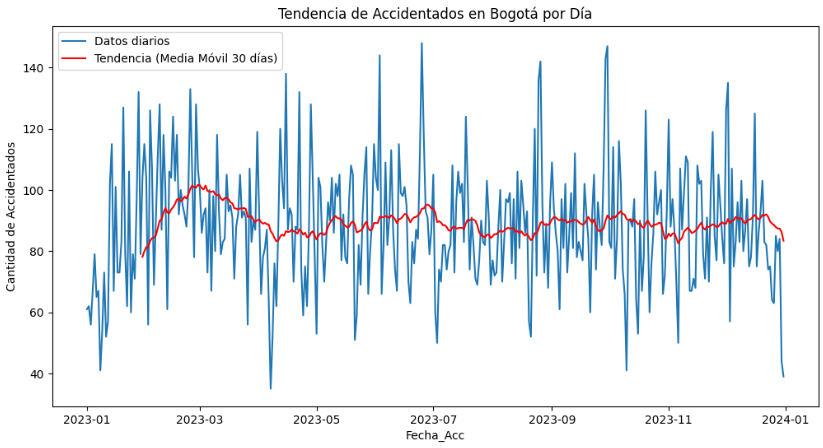


***Figura 6****: Gráfico de frecuencia por gravedad del accidente.*

**Análisis descriptivo de la serie de tiempo de accidentes en Bogotá**

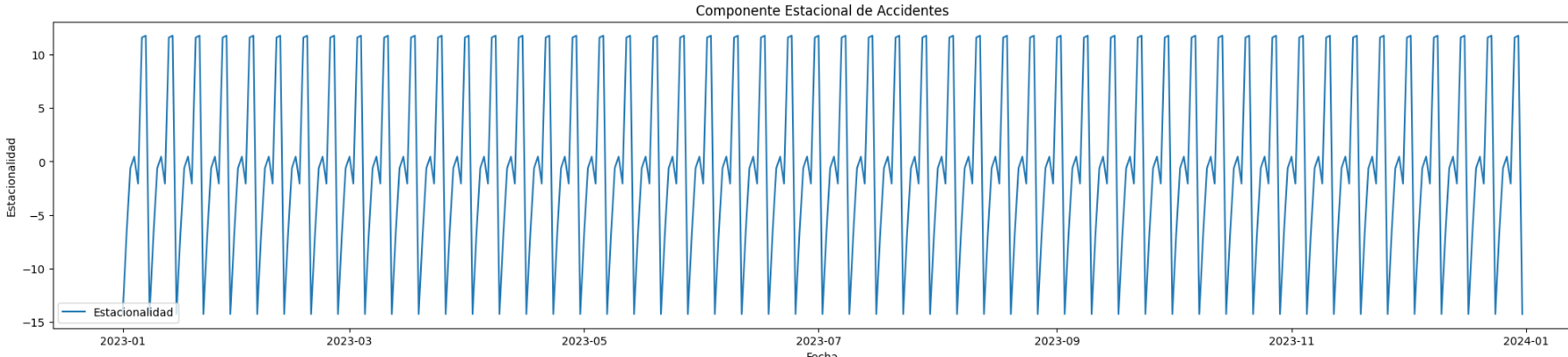
Se llevó a cabo un análisis descriptivo de la serie temporal del número de personas accidentadas en Bogotá durante el año 2023, con el fin de comprender el comportamiento de los accidentes a lo largo del tiempo. En primer lugar, se examina la tendencia, la cual permite determinar si los accidentes aumentaron, disminuyeron o se mantuvieron constantes a lo largo del año. Luego, se analizó la estacionalidad, con el objetivo de identificar si existen patrones repetitivos en intervalos de tiempo similares, como, por ejemplo, entre diferentes semanas. El estudio de la variabilidad y volatilidad, evalúa las fluctuaciones del número de personas accidentadas, observando cómo cambian estos valores a lo largo del tiempo. Posteriormente, el análisis de autocorrelación permitió verificar si los accidentes en fechas específicas podrían estar influenciados por los valores de días anteriores. Finalmente, los picos y anomalías ayudaron a detectar valores atípicos, ya sea por presentar números significativamente más altos de lo habitual (picos) o más bajos (anomalías), destacando que estas anomalías incluyen tanto los valores elevados como los reducidos en comparación con el promedio general.

* **Tendencia:** El análisis de la tendencia en la serie temporal del número de personas accidentadas en Bogotá durante 2023 revela un claro aumento en marzo, seguido de un período de estabilización desde abril hasta diciembre. Este patrón, observado al comparar los datos diarios con la media móvil de 30 días, sugiere que el incremento de marzo pudo haberse producido debido a factores específicos, como el aumento del tráfico o condiciones climáticas adversas. La estabilización posterior sugiere que dicho aumento podría haber sido temporal o que se implementaron medidas para mitigar los accidentes.



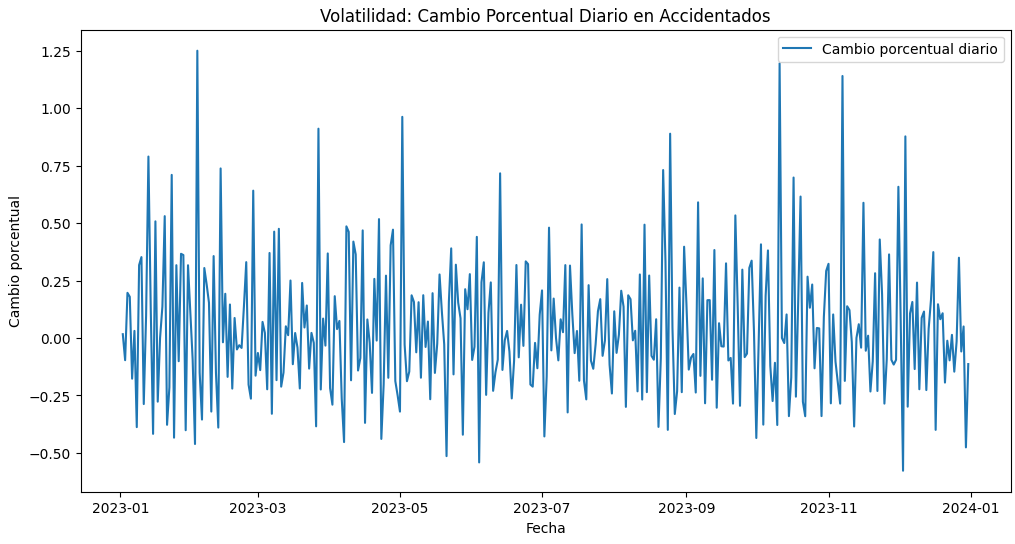
***Figura 7:*** *Tendencia de personas accidentadas en Bogotá por día.*

* **Estacionalidad:** La estacionalidad de los accidentes de tránsito en Bogotá muestra un patrón semanal claro, con el mayor número de accidentes ocurriendo los viernes y sábados (fines de semana), y los domingos registrando los valores más bajos. Este comportamiento sugiere que factores como el aumento del tráfico, el flujo vehicular intenso, eventos sociales y comportamientos típicos de los días de descanso son influyentes en la ocurrencia de accidentes.



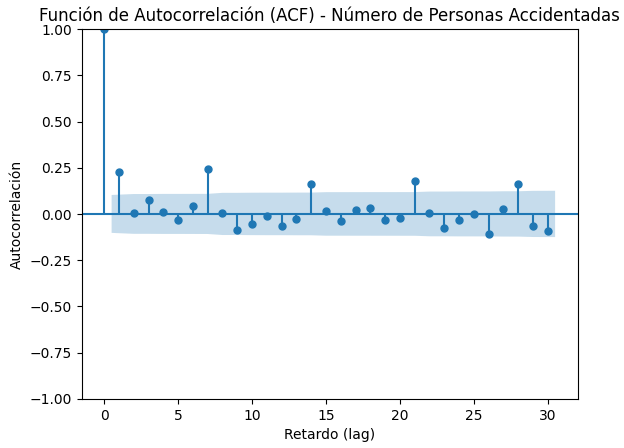
***Figura 8:*** *Gráfica de estacionalidad*

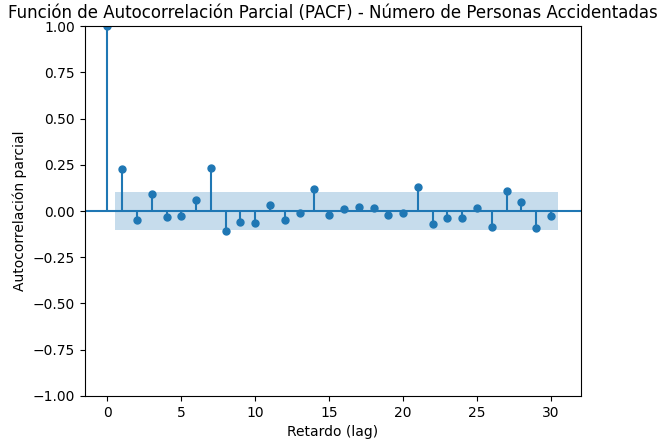
* **Variabilidad y volatilidad:** revela una alta dispersión en el número de personas accidentadas a diario, con una desviación estándar de 19.83 y una varianza de 393.26, lo que indica fluctuaciones significativas entre días. La volatilidad media diaria de 3.9% sugiere un aumento general en los accidentes, pero la alta desviación estándar del cambio porcentual (29.8%) indica que los cambios diarios son erráticos, con aumentos y disminuciones notables, lo que refleja un comportamiento irregular en la serie. temporal.



***Figura 9:*** *Gráfica de volatilidad*

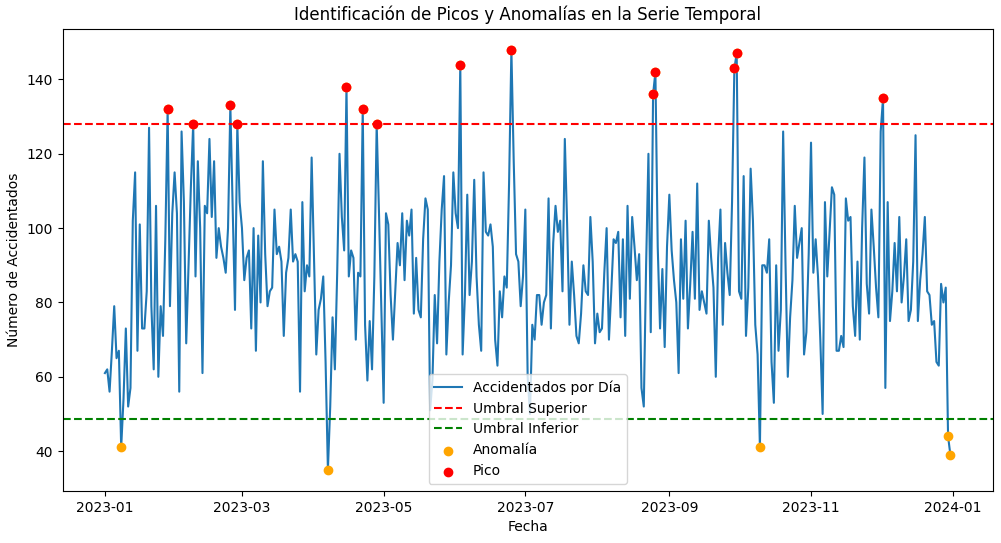
* **Autocorrelación:** muestra que los accidentes en Bogotá presentan cierta relación con los días anteriores. Para un retardo de 1 día, se observa una autocorrelación de 24%, lo que indica que el número de personas accidentadas de un día está un poco relacionado con el del día siguiente. En el retardo de 7 días, la autocorrelación aumenta al 25%, sugiriendo un patrón semanal en los accidentes. También se identifican autocorrelaciones de alrededor del 20% para los retardos de 21 y 28 días, mostrando una posible estacionalidad mensual. Las gráficas de ACF y PACF coinciden en la autocorrelación de 7 días, con el PACF mostrando una relación más directa y menos influenciada por retardos intermedios, lo que refuerza la idea de una estacionalidad semanal.





***Figura 10.*** *Gráfica de correlaciones. Parte superior: función de autocorrelación (ACF). Parte inferior: función de autocorrelación parcial (PACF).*

* **Picos y anomalías**: se encontraron 19 días con valores atípicos en el número de personas accidentadas. De estos, 5 días presentaron cifras inusualmente bajas, destacando el 7 de abril con solo 35 personas accidentadas, mientras que 14 días mostraron picos elevados, siendo el 25 de junio el más alto con 148 personas accidentadas. Estos días anómalos suelen coincidir con fines de semana, festivos y periodos de vacaciones, lo que sugiere que factores como el aumento de la movilidad y eventos especiales pueden influir en estos comportamientos extremos.



***Figura 11:*** *Identificación de Picos y Anomalías.*

# Referencias

Ahmed, S., Hossain, M., Ray, S. K., Bhuiyan, M., & Sabuj, S. (1 de 5 de 2023). A study on road accident prediction and contributing factors using explainable machine learning models: analysis and performance. *Transportation Research Interdisciplinary Perspectives, 19*. doi:10.1016/j.trip.2023.100814

Ali, Y., Hussain, F., & Haque, M. (1 de 1 de 2024). Advances, challenges, and future research needs in machine learning-based crash prediction models: A systematic review. *Accident Analysis and Prevention, 194*. doi:10.1016/j.aap.2023.107378

Coffman, M. (3 de 8 de 2022). *How AI and Machine Learning are Improving Ambulance Response and Dispatch*. Obtenido de https://www.jems.com/administration-and-leadership/communications-dispatch/how-ai-and-machine-learning-are-improving-ambulance-response-and-dispatch/

Gutierrez Osorio, C., & Pedraza, C. A. (3 de 2019). Characterizing road accidents in urban areas of Bogota (Colombia):. doi:10.1109/ITSLATAM.2019.8721334

Kang, S., & Cheong, T. (2023). Optimizing Ambulance Allocation in Dynamic Urban Environments: A Historic Data-Driven Approach. *Applied Sciences (Switzerland)*. doi:10.3390/app132111671

Khan, A., & Hussain, J. (2024). Utilizing GIS and Machine Learning for Traffic Accident Prediction in Urban Environment. *Civil Engineering Journal (Iran), 10*, 1922-1935. doi:10.28991/CEJ-2024-010-06-013

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

Parvareh, M., Karimi, A., Rezaei, S., Woldemichael, A., Nili, S., Nouri, B., & Nasab, N. (2018). Assessment and prediction of road accident injuries trend using time-series models in Kurdistan. *Burns & Trauma*. doi:10.1186/s41038-018-0111-6

Santos, D., Saias, J., Quaresma, P., & Nogueira, V. B. (1 de 12 de 2021). Machine learning approaches to traffic accident analysis and hotspot prediction. *10*. doi:10.3390/computers10120157

Secretaría Distrital de Movilidad. (2023). *Anuario de Siniestralidad Vial de Bogotá 2023.* Bogotá. Obtenido de https://observatorio.movilidadbogota.gov.co/sites/observatorio.movilidadbogota.gov.co/files/2024-09-30/pub/Anuario%20de%20Siniestralidad%20Vial%20de%20Bogot%C3%A1%202023.pdf

Silva, P., Andrade, M., & Ferreira, S. (1 de 12 de 2020). Machine learning applied to road safety modeling: A systematic literature review. *Journal of Traffic and Transportation Engineering (English Edition)*. doi:10.1016/j.jtte.2020.07.004

Ye, Q., Hu, K., & Wang, Y. (28 de 11 de 2019). Application of Time Series Analysis to Traffic. doi:arXiv:1911.12813v1 [stat.AP] 28 Nov 2019