.UBAeconómicas posgrado

ENAP Escuela de Negocios y Administración Pública

Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Negocios y Administración Pública

CARRERA DE ESPECIALIZACIÓN EN MÉTODOS CUANTITATIVOS PARA LA GESTIÓN Y ANÁLISIS DE DATOS EN ORGANIZACIONES

TRABAJO FINAL INTEGRADOR

Predicción de accidentes de tránsito en la Ciudad de Buenos Aires

Aplicación de métodos predictivos para la identificación de zonas de riesgo de accidentes de tránsito en la Ciudad de Buenos Aires.

AUTOR: SERGIO EZEQUIEL CHESKO SORROZA

TUTOR: PABLO NICOLAS CAVIEZEL

NOVIEMBRE 2024

Resumen

Los accidentes de tránsito son una problemática significativa en la Ciudad de Buenos Aires, influenciada por factores como la infraestructura vial, el comportamiento de los conductores y las características urbanas. Este trabajo analiza patrones espacio-temporales del tránsito y variables urbanas para desarrollar modelos predictivos capaces de clasificar zonas de riesgo.

Basándose en las primeras etapas de la metodología CRISP-DM, el estudio integra datos geoespaciales y temporales para anticipar la probabilidad de siniestralidad en distintas áreas de la ciudad. La fase de comprensión del negocio define el problema de la siniestralidad vial. En la fase de comprensión de los datos, se recopila información sobre patrones de tráfico, condiciones ambientales y características urbanas, creando una base sólida para el análisis predictivo.

El proceso de preparación de los datos incluye limpieza, transformación y normalización, lo que asegura que los datos estén listos para el modelado. En esta etapa, se implementan y evalúan modelos predictivos, como Random Forest y Gradient Boosting, para estimar la probabilidad de siniestralidad. La evaluación de los modelos se realiza a través de matrices de confusión, que permiten medir la precisión en la clasificación de zonas de alto, medio y bajo riesgo.

Los resultados obtenidos constituyen una herramienta predictiva importante para optimizar la asignación de recursos en seguridad vial y guiar políticas públicas orientadas a reducir la siniestralidad urbana en la Ciudad de Buenos Aires. Este modelo no solo mejora la toma de decisiones, sino que también contribuye a una planificación urbana más eficiente y proactiva, con un enfoque en la prevención de accidentes.

Palabras clave

Siniestralidad Vial, Análisis Espacio-Temporal, Predicción de Accidentes, Modelos predictivos, Zonas de Riesgo.

Índice

Introducción 4
1. Revisión del conocimiento sobre factores de siniestralidad vial 5
1.1. Comprensión de la influencia de las variables en la siniestralidad 6
1.2. Análisis de los factores de riesgo en los accidentes de tránsito 6
1.3. Modelos predictivos para la predicción de siniestralidad vial 8
1.4. Contexto Nacional 9
1.5. Modelos Predictivos para la Siniestralidad Vial en Buenos Aires 11
2. Metodología del análisis predictivo para la seguridad vial 11
2.1. Recopilación de datos 12
2.2. Procesamiento de datos 13
2.3. Modelos predictivos y evaluación de resultados 14
3. Desarrollo e implementación del modelo predictivo 16
3.1. Captura, Transformación de Datos y Análisis Exploratorio 16
3.1.1. Captura y Transformación de Datos 16
3.1.2. Análisis Exploratorio de los Datos de Accidentes 18
3.1.3. Combinación de Accidentes con Celdas y Generación de Variables Lag. 21
3.1.4. Discretización de la Variable Objetivo 22
3.1.5. Limpieza y Filtrado de Celdas sin Accidentes 23
3.1.6. Implementación del Modelo Predictivo 23
3.1.7. Modelo Random Forest 24
3.2. Evaluación de los Modelos 24
3.2.1. Modelo Base de Regresión Logística 25
3.2.2. Modelo Random Forest 26
3.2.3. Modelo Gradient Boosting 27
3.2.4. Enfoque de Ventana Rolling 28
3.2.5. Balanceo de Clases 29
3.3. Selección de los Modelos: Matriz de Confusión e Interpretación. 31
3.3.1. Selección del Modelo. 31
3.3.2. Matriz de Confusión Ponderada: Evaluación con Costos Diferenciados. 32
3.3.3. Interpretación del Modelo. 33
3.3.4. Estabilidad del modelo 35
Conclusiones 36
Referencias bibliográficas 38

Introducción

En la Ciudad de Buenos Aires, los accidentes de tránsito representan un problema crítico para la seguridad vial, resultado de la interacción entre diversos factores, como la infraestructura vial, el comportamiento de los conductores y las condiciones ambientales. Las dinámicas urbanas cambiantes requieren una comprensión más profunda de cómo estos factores se combinan para generar siniestralidad.

En línea con los objetivos del Plan de Seguridad Vial 2020-2023 de la Ciudad de Buenos Aires, que busca reducir en un 50% las víctimas fatales para 2030, surge la necesidad de adoptar enfoques innovadores que complementen las medidas tradicionales. Este plan enfatiza la importancia de infraestructuras seguras, controles efectivos y educación vial, pero también reconoce la velocidad y precisión en la toma de decisiones como factores clave para alcanzar los objetivos. En este sentido, el uso modelos predictivos puede ser crucial para identificar y mitigar zonas de alto riesgo, permitiendo optimizar la asignación de recursos en seguridad vial.

Los enfoques convencionales de gestión del tránsito, basados en la regulación y el control del flujo vehicular, han mostrado limitaciones en la reducción efectiva de los accidentes. Estas limitaciones evidencian la necesidad de adoptar herramientas analíticas avanzadas, capaces de detectar patrones críticos en la ocurrencia de siniestros. Este estudio busca llenar un vacío en la comprensión de la siniestralidad urbana, explorando cómo los patrones espacio-temporales del tránsito y las características viales de la ciudad afectan tanto la frecuencia como la severidad de los accidentes.

En este contexto, surge la pregunta central de este trabajo: ¿Cómo afectan los patrones de tráfico y las características urbanas a la frecuencia y gravedad de los accidentes de tránsito en la Ciudad de Buenos Aires, y cómo puede utilizarse esta información para predecir zonas con mayores probabilidades de accidentes?

El objetivo general de este estudio es desarrollar un sistema predictivo capaz de anticipar zonas y momentos de mayor riesgo de siniestralidad en Buenos Aires, mediante el análisis de patrones de tránsito y características urbanas en sus dimensiones espaciales y temporales. Entre los objetivos específicos se encuentran comprender la influencia de variables relacionadas con el tránsito y las características urbanas en la frecuencia y gravedad de los accidentes, identificar patrones espacio-temporales críticos en la ocurrencia de siniestros, y proveer herramientas analíticas que apoyen la toma de decisiones estratégicas en políticas de tránsito y optimicen la asignación de recursos destinados a la seguridad vial.

Se espera que los hallazgos de este trabajo contribuyan al diseño de políticas públicas más efectivas, orientadas a la prevención de accidentes y la mejora de la seguridad vial en la Ciudad de Buenos Aires. Para abordar esta problemática, el estudio se estructura siguiendo las primeras etapas de la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), que guía el análisis en cinco etapas clave: comprensión del negocio y los datos, preparación de los datos, modelado, evaluación e implementación.

Con este enfoque, se busca no solo comprender las dinámicas de la siniestralidad urbana, sino también ofrecer una herramienta práctica para anticipar riesgos, facilitando una planificación urbana más eficiente y proactiva.

1. Revisión del conocimiento sobre factores de siniestralidad vial

Los accidentes viales en las zonas urbanas representan un desafío significativo para la seguridad pública y la gestión del tráfico. Comprender cuáles son los factores que contribuyen a los accidentes de tránsito es esencial para el diseño de estrategias efectivas de prevención. Por ello, la capacidad de predecir el riesgo de accidentes se vuelve fundamental, ya que permite no solo prevenir la ocurrencia de siniestros, sino también mitigar los daños que estos causan. En este contexto, el estado actual del conocimiento sobre siniestralidad vial se abordará a través de cuatro enfoques, fundamentados en la bibliografía existente sobre la materia.

El primero se centrará en la comprensión de la influencia de las variables en la siniestralidad, analizando cómo la interacción entre factores como la infraestructura, el comportamiento de los conductores y las condiciones ambientales afecta la frecuencia y gravedad de los accidentes. El segundo explorará el análisis de los factores de riesgo, identificando los elementos que incrementan la probabilidad de siniestros, como el comportamiento del conductor y la calidad de la infraestructura vial. En el tercer enfoque se discutirán los modelos predictivos utilizados para anticipar zonas de alto riesgo de accidentes, presentando ejemplos de estudios que aplican técnicas avanzadas para mejorar la seguridad vial. Finalmente, el cuarto enfoque proporcionará un contexto nacional, revisando informes estadísticos que detallan la siniestralidad vial en Argentina y en la Ciudad Autónoma de Buenos Aires, y resaltando la importancia de estos datos para la formulación de políticas públicas.

1.1. Comprensión de la influencia de las variables en la siniestralidad

La siniestralidad vial es un problema complejo que requiere un análisis detallado para identificar las diversas variables que influyen en la ocurrencia de accidentes. Comprender estas variables es esencial no solo para identificar patrones, sino también para desarrollar intervenciones efectivas que mejoren la seguridad en las vías urbanas.

Los accidentes de tránsito son el resultado de la interacción de múltiples factores, que van desde las características de la infraestructura vial y el comportamiento de los conductores hasta las condiciones ambientales. Comprender cómo estos factores afectan la frecuencia y gravedad de los accidentes de tránsito es crucial para desarrollar políticas de seguridad vial efectivas. Como señala Evans (1991), la seguridad vial depende de una combinación de elementos que incluyen tanto los factores humanos como las características del entorno vial, y su análisis es fundamental para entender cómo las variables individuales y ambientales contribuyen a la ocurrencia de accidentes. Esta comprensión ofrece una base sólida para la implementación de estrategias de prevención más efectivas.

Además, estos factores no actúan de forma aislada, sino que se influyen mutuamente, lo que dificulta el análisis de los accidentes de tránsito. Por ejemplo, la combinación de una vía mal señalizada y la conducción a alta velocidad podría aumentar significativamente el riesgo de siniestros. Por ello, es fundamental que los análisis de accidentes de tránsito en zonas urbanas incluyan la evaluación de múltiples variables mediante técnicas adecuadas, que permiten no solo identificar correlaciones, sino también comprender mejor la interacción entre los factores que influyen en la siniestralidad vial.

1.2. Análisis de los factores de riesgo en los accidentes de tránsito

El análisis de los factores de riesgo en los accidentes de tránsito es fundamental para desarrollar estrategias efectivas de prevención y seguridad vial. Al identificar los elementos que aumentan la probabilidad de sufrir un siniestro, se pueden implementar medidas que reduzcan estos riesgos y mejoren la seguridad en las vías.

Ramírez y Valencia (2021) analizaron datos de accidentes de tránsito en Bogotá entre 2014 y 2016, utilizando variables espacio-temporales como clima, eventos estacionales y características de la infraestructura vial. Sus resultados destacan los principales factores que aumentan el riesgo de accidentes y muertes, identificando zonas críticas que requieren mayor atención y desarrollando una herramienta predictiva para pronosticar accidentes futuros basada en estas variables.

Entre los principales factores de riesgo se destacan, por un lado, los factores relacionados con el conductor. El comportamiento del conductor es uno de los determinantes más importantes de los accidentes de tránsito, ya que conductas como el exceso de velocidad, la conducción bajo los efectos del alcohol o drogas, el uso del teléfono móvil y la falta de atención son causas frecuentes de siniestros viales. Por otro lado, los factores relacionados con la infraestructura vial, su calidad y diseño también influyen en la ocurrencia de accidentes. La mala señalización, iluminación deficiente, mal estado de las vías y la ausencia de semaforización o sendas peatonales pueden incrementar el riesgo de siniestros.

Además, factores urbanos como la densidad de la población, el flujo vehicular intenso y la cercanía a lugares de interés también pueden ser determinantes. Asimismo, los factores temporales, como el horario, el día de la semana, el mes y la estación del año, juegan un papel importante en la siniestralidad vial. Por último, las condiciones climáticas adversas, como la lluvia y la niebla, aumentan el riesgo de accidentes al reducir la visibilidad y el control del vehículo.

De manera similar, un estudio realizado en la ciudad de Zagreb, Croacia, también encontró que factores como la velocidad excesiva, la falta de visibilidad y las condiciones de la infraestructura vial son determinantes clave para la gravedad de los accidentes urbanos. Vorko-Jović et al. (2005) identificaron que los accidentes más fatales ocurren con mayor frecuencia durante la noche, en vías urbanas, y en condiciones de visibilidad reducida, factores que elevan significativamente el riesgo de lesiones graves y muertes. Estos hallazgos son consistentes con los resultados obtenidos en Bogotá, lo que sugiere que, a pesar de las diferencias geográficas, existen patrones comunes que pueden ser aplicados globalmente para diseñar medidas preventivas más efectivas.

En consonancia con lo planteado por Híjar et al. (1999), factores como el consumo de alcohol y las condiciones climáticas adversas son fundamentales en la ocurrencia de accidentes de tránsito. En su estudio, se encontró que las condiciones de visibilidad reducida y el mal tiempo incrementan significativamente el riesgo de accidentes, lo que refuerza la importancia de tener en cuenta estos aspectos en las estrategias de prevención. Además, conductores jóvenes, particularmente aquellos menores de 25 años, presentan una mayor propensión a involucrarse en siniestros, lo que destaca la necesidad de centrar las políticas de seguridad vial en este grupo de riesgo.

En resumen, el análisis de los factores de riesgo en los accidentes de tránsito permite identificar y comprender mejor las diversas variables que contribuyen a la siniestralidad vial,

lo que resulta esencial para el desarrollo de políticas y estrategias de prevención más efectivas.

1.3. Modelos predictivos para la predicción de siniestralidad vial

La siniestralidad vial es un problema complejo que requiere un enfoque proactivo para mejorar la seguridad. En este contexto, el uso de modelos predictivos se ha convertido en una herramienta esencial para anticipar zonas de alto riesgo de accidentes de tránsito. Estos modelos permiten identificar patrones en datos históricos y prever la probabilidad de siniestros en determinadas áreas, facilitando así la implementación de medidas preventivas y de intervención.

Se han realizado varios trabajos que involucran la utilización de modelos para la predicción de accidentes en áreas urbanas. Alhaek et al. (2024) propusieron un enfoque de aprendizaje profundo para predecir la gravedad de los accidentes de tráfico, considerando tanto patrones espaciales como dependencias temporales. Utilizaron una red neuronal para extraer características espaciales de datos de alta dimensión. Los resultados de su estudio, basado en datos de accidentes de tráfico de dos ciudades, demostraron la eficacia de su enfoque para mejorar la seguridad vial.

Song et al. (2020) realizaron un estudio en Carolina del Norte para analizar la gravedad de las lesiones de peatones en colisiones con vehículos, considerando patrones espaciotemporales. Utilizando modelos jerárquicos, el estudio identificó diez patrones espacio temporales de accidentes. El estudio destacó la importancia de considerar los factores humanos, del vehículo, del choque y ambientales en el análisis espacio temporal para proporcionar recomendaciones específicas para áreas propensas a accidentes.

Li et al. (2021) analizaron las infracciones de tráfico en intersecciones. El estudio examinó 24 factores, incluyendo variables relacionadas con el tiempo, espacio, tráfico y clima, y encontró que el mediodía, los fines de semana, los distritos residenciales, las rutas secundarias, el tráfico congestionado y las bajas temperaturas aumentan la probabilidad de infracciones. Utilizando un modelo de RandomForest para la predicción, el estudio demostró que este algoritmo superó a otros modelos, como la regresión logística.

Cui et al. (2024) propusieron un enfoque avanzado para predecir accidentes de tráfico urbanos mediante un marco de aprendizaje espacio temporal. Este enfoque mejora la precisión en la predicción al capturar tanto las dependencias locales como globales entre los datos de accidentes. Utilizando datos de ciudades como Nueva York y Londres, demostraron que su

método supera a los modelos tradicionales en la identificación de patrones espacio temporales complejos, lo que resulta en una mayor precisión en la predicción de accidentes.

Cardona Álvarez (2023) analiza la predicción del riesgo de accidentes de tráfico en Manizales, Colombia, mediante técnicas de aprendizaje automático. El estudio evalúa la eficacia de diferentes modelos en función del contexto y horizonte temporal. Los resultados muestran que la efectividad de los modelos varía según el área y el periodo analizado, sin que haya un algoritmo universalmente superior.

En resumen, la aplicación de modelos predictivos en la siniestralidad vial representa un avance significativo en la comprensión y gestión de los riesgos asociados a los accidentes de tráfico. A través de estos modelos, es posible identificar áreas críticas y patrones de comportamiento que contribuyen a la ocurrencia de siniestros, lo que permite a las autoridades y planificadores implementar medidas de seguridad más efectivas y basadas en datos. La continua investigación en este campo no solo mejora la precisión de las predicciones, sino que también contribuye a la creación de entornos viales más seguros para todos los usuarios.

1.4. Contexto Nacional

La siniestralidad vial es un problema significativo en Argentina, que requiere atención y acciones coordinadas a nivel nacional y local. En este sentido, se han elaborado diversos informes estadísticos que proporcionan una visión detallada sobre la siniestralidad vial en el país y en la Ciudad Autónoma de Buenos Aires. Estos informes, producidos por diferentes entidades gubernamentales, ofrecen datos actualizados sobre accidentes de tráfico, sus causas y características, y son fundamentales para la toma de decisiones y el diseño de políticas públicas en materia de seguridad vial. A continuación, se presentan los informes más relevantes en el ámbito nacional.

El Informe de Siniestralidad Vial Fatal de 2023, elaborado por la Agencia Nacional de Seguridad Vial (ANSV), presenta una visión general de las estadísticas de accidentes de tráfico en Argentina, con un enfoque especial en Buenos Aires. El informe incluye datos sobre el número de siniestros, las causas más comunes y un análisis de las tendencias observadas en el país.

El Informe estadístico sobre las víctimas fatales a causa de siniestralidad vial en la Ciudad Autónoma de Buenos Aires de 2022, publicado por el Gobierno de la Ciudad Autónoma de Buenos Aires, proporciona un análisis detallado de la cantidad y características de los

accidentes de tránsito fatales en la ciudad. El documento abarca datos sobre mortalidad, factores contribuyentes y la distribución espacial de los siniestros.

El Anuario Estadístico de Siniestralidad Fatal de 2021, elaborado por el Ministerio de Transporte de Argentina a través de la Dirección Nacional de Observatorio Vial, reporta estadísticas de accidentes de tránsito resaltando la importancia de contar con datos precisos para la toma de decisiones basadas en evidencia, con el objetivo de reducir la siniestralidad vial en el país.

Además de los informes sobre siniestralidad vial, es fundamental considerar las consecuencias socioeconómicas de estos siniestros, que tienen un impacto significativo en la salud pública y la economía de Argentina. Según el estudio sobre las Consecuencias de la siniestralidad vial en Argentina (Dirección de Investigación Accidentológica del Observatorio Nacional Vial, 2022), los accidentes de tráfico en el país generan pérdidas sustanciales. Estos siniestros no solo afectan la salud de las personas, con altas tasas de mortalidad y morbilidad, sino que también tienen consecuencias económicas severas, representando entre el 1% y el 3% del Producto Bruto Interno (PBI) de los países, según la OMS. En Argentina, los costos sociales de la siniestralidad vial para 2019 fueron estimados en más de 354 mil millones de pesos, considerando tanto los costos directos como indirectos. Esto incluye gastos médicos, daños a la propiedad, pérdida de productividad y los costos humanos relacionados con el dolor y el sufrimiento de las víctimas y sus familias. Estos datos resaltan la importancia de abordar la siniestralidad vial no solo desde una perspectiva de seguridad, sino también como un problema con graves repercusiones económicas, lo que justifica una mayor inversión en políticas de prevención y seguridad vial.

Plan de Seguridad Vial 2020-2023 revela que la siniestralidad vial sigue siendo un grave problema en Argentina. A pesar de los esfuerzos realizados para mejorar la seguridad en las vías, el informe destaca que las cifras de siniestros no han disminuido de manera significativa, lo que mantiene los índices elevados. Los accidentes de tránsito continúan siendo una de las principales causas de muerte en el país, lo que subraya la necesidad urgente de reforzar las políticas de seguridad vial.

Estos informes estadísticos sobre siniestralidad vial proporcionan una base crucial para comprender la magnitud del problema y las dinámicas involucradas. A partir del análisis de estos datos, las autoridades pueden desarrollar e implementar políticas más efectivas que apunten a reducir la incidencia de accidentes y mejorar la seguridad en las vías del país.

1.5. Modelos Predictivos para la Siniestralidad Vial en Buenos Aires

Este trabajo introduce un enfoque innovador en la predicción de la probabilidad de siniestralidad en la Ciudad de Buenos Aires, combinando el análisis de patrones de tránsito y características urbanas en su dimensión espacio-temporal. Aunque existen estudios previos que han abordado el análisis de accidentes de tránsito y su relación con diversas variables, este proyecto busca proporcionar una herramienta útil para la planificación urbana y la gestión del tráfico en la Ciudad de Buenos Aires. Al predecir con mayor precisión las zonas y momentos de mayor riesgo de accidentes, se pueden diseñar estrategias más efectivas y optimizar la asignación de recursos destinados a la seguridad vial.

El uso de estadísticas de accidentes y de tránsito, en combinación con datos detallados sobre la infraestructura urbana de Buenos Aires, permitirá desarrollar modelos predictivos capaces de capturar la complejidad del entorno urbano. La metodología propuesta adopta un enfoque sistemático para la captura y limpieza de datos, seguido de un análisis riguroso mediante técnicas de aprendizaje automático. Esto permitirá identificar patrones críticos en la siniestralidad vial, considerando variables espacio-temporales y urbanas que son fundamentales para mejorar la precisión predictiva y apoyar decisiones en la planificación y gestión del tráfico.

2. Metodología del análisis predictivo para la seguridad vial

En este apartado se describe el enfoque metodológico diseñado para construir un modelo predictivo de siniestralidad vial en la Ciudad de Buenos Aires, con el objetivo de entender y anticipar la ocurrencia de accidentes de tránsito. La metodología de este estudio se basa en el marco CRISP-DM (Cross-Industry Standard Process for Data Mining), que proporciona una estructura estándar para el proceso de minería de datos, asegurando un enfoque sistemático y eficiente (Shearer, 2000).

El proceso comienza con la comprensión del negocio, donde se define el problema de siniestralidad vial y se establecen los objetivos del estudio. Posteriormente, en la comprensión de los datos, se lleva a cabo la recopilación de información relevante sobre patrones de tráfico y características urbanas, así como datos ambientales y temporales.

En la fase de preparación de los datos, se realiza un procesamiento exhaustivo que incluye la limpieza, transformación y normalización de los datos, asegurando que estén listos para el análisis. Esto permite manejar la complejidad de los patrones de tráfico y las variables involucradas en la siniestralidad.

La fase de modelado implica la aplicación de diferentes modelos predictivos, como regresiones logísticas y Random Forest, para entender la relación entre las variables y predecir la siniestralidad vial. Se evaluarán estos modelos utilizando métricas específicas, garantizando que sean precisos y robustos.

Finalmente, en la etapa de evaluación, se validarán los modelos mediante datos de prueba, y se analizarán los resultados para extraer conclusiones y recomendaciones. Este enfoque metódico permite no solo predecir la siniestralidad, sino también ofrecer información valiosa para mejorar la seguridad vial y la planificación urbana en la Ciudad de Buenos Aires.

2.1. Recopilación de datos

Para alcanzar los objetivos de este estudio, se ha diseñado una metodología que abarca la recopilación, el procesamiento y la transformación de datos provenientes de múltiples fuentes. A diferencia de un enfoque que parte de un conjunto de datos ya armado, este trabajo se basará en la construcción de un dataset consolidado que integre información de diferentes fuentes, haciendo fundamental no solo la recopilación, sino también la transformación y estandarización de los datos para su correcta aplicación en el modelo predictivo.

La recopilación de datos se divide en varias etapas de acuerdo con las diversas fuentes de información. Primero, se utilizarán registros de accidentes de tránsito obtenidos de fuentes oficiales de la Ciudad de Buenos Aires. Estos datos incluyen información clave como la fecha, la hora y la ubicación geográfica de cada accidente, lo cual permite identificar patrones espaciales y temporales en la ocurrencia de siniestros. La ubicación y el horario son particularmente relevantes, ya que permiten observar tendencias en zonas y horarios específicos.

En segundo lugar, los datos sobre infraestructura vial y características urbanas se extraerán de repositorios públicos. Estos datos incluyen aspectos como el ancho de las calles, el número de carriles, la presencia de semáforos, cruces peatonales y otros elementos que podrían influir en el riesgo de accidentes. La infraestructura vial puede determinar zonas de mayor riesgo, por ejemplo, en áreas con intersecciones complicadas o vías rápidas. Estas variables urbanas aportan una dimensión adicional al análisis, ya que permiten considerar el entorno físico en el que ocurren los accidentes y cómo ciertas características pueden contribuir a la siniestralidad.

Por último, los datos meteorológicos se obtendrán de fuentes meteorológicas públicas e incluirán información sobre variables como la precipitación, temperatura y visibilidad. Las condiciones climáticas tienen un impacto directo en la seguridad vial, ya que factores como la lluvia, la niebla o las bajas temperaturas pueden reducir la visibilidad, afectar la adherencia de

los vehículos a la calzada y, en general, incrementar el riesgo de accidentes. Estas variables aportan una dimensión ambiental, esencial para capturar la influencia de las condiciones del clima en la ocurrencia de siniestros.

Esta recopilación de datos, segmentada y justificada, permitirá construir un conjunto de datos integral, adecuado para el desarrollo y la evaluación de un modelo predictivo que tome en cuenta tanto los patrones de tráfico y las características del entorno como las condiciones climáticas en la Ciudad de Buenos Aires.

2.2. Procesamiento de datos

El procesamiento de datos en este estudio se lleva a cabo en dos etapas principales: la preparación de datos y la ingeniería de características, cada una con un rol específico en la creación de un conjunto adecuado para el análisis predictivo.

En la preparación de datos, se realiza una limpieza, estandarización y normalización exhaustiva del conjunto de datos para asegurar su calidad y cohesión. Este proceso incluye la eliminación de registros incompletos o inconsistentes, lo cual es esencial para garantizar la integridad de la información. Además, se estandarizan los nombres de las variables y se normalizan los formatos de datos provenientes de distintas fuentes, facilitando así la unificación y reduciendo el riesgo de problemas durante el modelado. Este paso asegura que el conjunto final esté libre de errores estructurales y listo para el análisis.

Posteriormente, la ingeniería de características se enfoca en enriquecer el conjunto de datos para mejorar la precisión del modelo predictivo. Uno de los principales desafíos en este proyecto es la transformación de los datos recopilados en un formato que permita la generación de nuevos atributos significativos. Para ello, se construye una grilla espacial que subdivide la ciudad en celdas, lo cual permite observar patrones localizados de siniestralidad. Esta estructura facilita el análisis de patrones espaciales específicos, como la frecuencia de accidentes en zonas determinadas, y permite calcular variables asociadas a cada área, como características del entorno y condiciones temporales.

Además de las variables de tráfico y ambientales, se incorporan variables exógenas como la presencia de escuelas, zonas comerciales, hospitales y otros puntos de interés que pueden influir en el riesgo de accidentes. Por ejemplo, la proximidad a escuelas o áreas comerciales suele estar asociada con un incremento en la actividad peatonal y vehicular, lo cual podría afectar la probabilidad de siniestros. Estas variables se integran como datos adicionales en cada celda de la grilla para evaluar su efecto en la siniestralidad.

En términos de procesamiento temporal, los datos se estructuran mensualmente. Para cada celda de la grilla, se organizan y agregan datos correspondientes a cada mes dentro de un período determinado. Este enfoque permite observar tendencias a largo plazo, como fluctuaciones estacionales o cambios progresivos en la siniestralidad en diferentes áreas de la ciudad. Esta estructura mensual facilita la identificación de patrones recurrentes y ayuda a analizar cómo los factores temporales interactúan con los demás atributos en la ocurrencia de accidentes.

Todo este procesamiento detallado asegura que los datos estén limpios, organizados y preparados para su integración en el modelo predictivo, optimizando así su precisión y utilidad para anticipar la ocurrencia de accidentes en la Ciudad de Buenos Aires.

2.3. Modelos predictivos y evaluación de resultados

Respecto al análisis de datos, se desarrollarán diferentes modelos predictivos para estimar el riesgo de siniestros en función de las variables analizadas. En particular, se utilizará el algoritmo Random Forest, el cual es adecuado para este tipo de predicción debido a su capacidad para manejar grandes volúmenes de datos y su robustez frente al sobreajuste (Liaw & Wiener, 2002). Además, se evaluará la efectividad de este modelo en comparación con otros métodos, como regresiones logísticas, para determinar cuál proporciona mejores resultados en la predicción de accidentes.

La validación de los modelos se llevará a cabo utilizando datos de prueba, lo que permitirá asegurar la precisión de las estimaciones. Para evaluar la efectividad de los modelos, se emplearán métricas específicas como la precisión y el recall, que medirán la sensibilidad y especificidad de las predicciones (Saito & Rehmsmeier, 2015). Esto proporcionará una visión clara de cómo cada modelo se desempeña en diferentes aspectos de la predicción de siniestralidad.

Además, se realizará un análisis de la importancia de las variables, especialmente en el caso del modelo Random Forest, para identificar qué factores tienen un mayor peso en la predicción de accidentes. Esta información no solo ayudará a entender mejor los determinantes de la siniestralidad vial, sino que también podrá guiar futuras investigaciones y estrategias de intervención.

Finalmente, se evaluarán los resultados obtenidos, lo cual es una fase crítica del estudio, ya que determinará la efectividad de los modelos desarrollados en predecir la siniestralidad vial en la ciudad. Los hallazgos de esta evaluación permitirán mejorar los modelos y proporcionar recomendaciones basadas en los resultados. Las conclusiones del estudio se centrarán en la

efectividad de los modelos predictivos aplicados y las oportunidades futuras para la investigación en la predicción de accidentes de tránsito.

3. Desarrollo e implementación del modelo predictivo

En este apartado se detalla el proceso llevado a cabo para desarrollar un modelo predictivo que permita anticipar la siniestralidad vial en la Ciudad de Buenos Aires. Este proceso incluye desde la recopilación y preparación de los datos, pasando por su transformación y análisis exploratorio, hasta el modelado, evaluación y explicación de los resultados obtenidos. La implementación sigue un enfoque estructurado que garantiza la coherencia y efectividad en cada etapa del desarrollo, asegurando la aplicabilidad del modelo en contextos reales.

3.1. Captura, Transformación de Datos y Análisis Exploratorio

El proceso de captura, transformación y análisis exploratorio de datos se orientó a construir un conjunto de datos consolidado que permita la predicción de la siniestralidad vial en la Ciudad de Buenos Aires. Este proceso incluyó la segmentación espacial de la ciudad en sectores homogéneos para asociar datos relevantes como infraestructura vial, puntos de interés, variables temporales y meteorológicas. Se integraron diferentes fuentes de información, incluyendo aspectos urbanos, climáticos y de infraestructura, para enriquecer el análisis.

El análisis exploratorio de los datos de accidentes permitió identificar patrones temporales y espaciales, lo que facilitó la creación de variables clave como el número de accidentes por celda y mes. A su vez, se generaron variables lag para capturar el impacto de accidentes pasados sobre la siniestralidad futura. Este enfoque integrador garantizó que los datos estuvieran listos para la posterior modelización predictiva.

A continuación, se detallan los pasos específicos en el proceso de captura y transformación de los datos, el análisis exploratorio de accidentes y su combinación con la base de celdas.

3.1.1. Captura y Transformación de Datos

El primer paso consistió en la creación de una grilla espacial que subdividiera la ciudad en sectores homogéneos. Esta segmentación fue desarrollada específicamente para este trabajo y permite asociar cada dato con un área determinada, facilitando el análisis de patrones localizados. Se dividió la ciudad en una grilla de 1190 celdas, donde cada celda representa una unidad espacial a la que se asocian las distintas variables recopiladas, constituyendo la base para integrar datos de diversas fuentes con un enfoque geográfico.

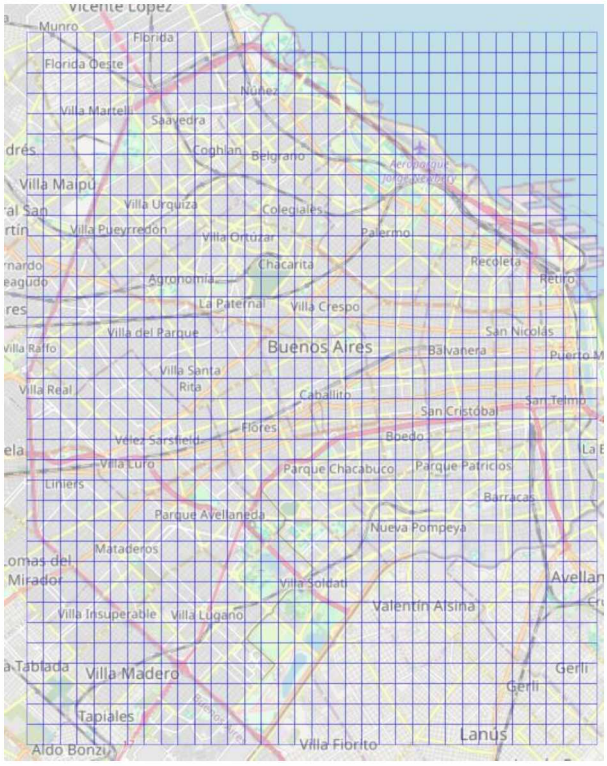


Figura 1: Grilla espacial (Elaboración Propia)

Una vez definida la grilla, se incorporó información sobre la infraestructura vial de la ciudad proveniente de OSMnx, una herramienta que permite descargar, modelar y analizar redes urbanas basadas en datos de OpenStreetMap. Entre las variables consideradas se incluyeron la cantidad de intersecciones dentro de cada celda, el número de semáforos, la presencia de paradas de buses, los límites de velocidad superiores y la existencia de calles de un solo sentido. Estas características permiten evaluar cómo las condiciones estructurales de cada área pueden influir en el riesgo de accidentes.

Adicionalmente, se integraron datos relacionados con puntos de interés que impactan la dinámica vial, obtenidos también de OSMnx. Entre estos se encuentran la cantidad de hospitales, escuelas, universidades, comisarías, estaciones de bomberos, bibliotecas, así como zonas de ocio como bares, restaurantes y clubes nocturnos. Estas variables enriquecen el análisis al incorporar aspectos sociales y funcionales del entorno urbano que pueden influir en el comportamiento vehicular y peatonal.

Para incorporar una perspectiva temporal, los datos fueron mensualizados. Esto implicó generar observaciones correspondientes a cada celda y mes dentro del período de análisis, permitiendo capturar la evolución de las variables a lo largo del tiempo y facilitando el análisis de tendencias temporales. Además, se integraron variables meteorológicas como la precipitación acumulada, la velocidad del viento, la temperatura promedio y mínima, y las condiciones de visibilidad. Estas variables se obtuvieron a través de Meteostat, una

plataforma que proporciona datos climáticos históricos y actuales a nivel global, basada en información de estaciones meteorológicas oficiales. Estos datos son esenciales para evaluar el impacto de las condiciones climáticas en la siniestralidad.

3.1.2. Análisis Exploratorio de los Datos de Accidentes

De manera paralela al armado de la base principal, se trabajó con un conjunto de datos específico de accidentes de tránsito ocurridos en la ciudad. Estos datos, correspondientes al período comprendido entre los años 2019 y 2023, fueron obtenidos de un repositorio público de accidentes de tránsito, disponible en la web de la Ciudad Autónoma de Buenos Aires (CABA) como parte de los datos abiertos proporcionados por el gobierno. La información proviene del Observatorio de Movilidad y Seguridad Vial, que recopila y publica estadísticas sobre los siniestros viales ocurridos en la ciudad.

El Observatorio de Movilidad y Seguridad Vial se basa en datos proporcionados por la Policía de la Ciudad y otras fuentes oficiales, registrando una amplia gama de accidentes de tránsito que afectan la movilidad y seguridad en Buenos Aires. Esta información permite tener una visión integral de la siniestralidad vial en la ciudad y es de gran utilidad para la toma de decisiones en políticas públicas.

A través de su metodología, el Observatorio garantiza que los datos sean lo más completos posible, validándolos con diversas fuentes como SAME, AUSA, hospitales, y otros organismos. Esta información es publicada y está disponible para consulta pública, promoviendo una mayor transparencia y acceso a datos cruciales para la gestión de la seguridad vial.

El conjunto de datos incluye información detallada sobre cada accidente, como el ID del hecho, la cantidad de víctimas (N_VICTIMAS), la fecha y hora del evento (FECHA, HORA), y variables espaciales como la calle y la altura (CALLE, ALTURA), así como la ubicación geográfica especificada mediante latitud y longitud (LATITUD, LONGITUD). También incorpora datos relevantes como la comuna en la que ocurrió el accidente (COMUNA), la gravedad del incidente (GRAVEDAD), y detalles sobre los participantes y víctimas involucradas.

Este análisis exploratorio inicial incluyó la distribución temporal de los accidentes por mes y año, la identificación de concentraciones espaciales, el análisis de horarios pico y la evaluación de factores recurrentes. Este trabajo permitió identificar patrones preliminares y posibles relaciones con las variables incorporadas en las celdas del modelo principal.

Se llevó a cabo un análisis detallado de la evolución diaria y mensual de la cantidad de accidentes, observándose cierta estacionalidad en los datos. En particular, se identificó una disminución en la cantidad de siniestros durante los meses de diciembre, enero y febrero, lo que sugiere una posible relación con la reducción del tráfico en la ciudad debido al período vacacional.

Asimismo, el análisis reveló la presencia de valores atípicos entre marzo de 2020 y mayo de 2021, período en el que las restricciones de movilidad impuestas por la pandemia de COVID-19 impactaron significativamente en la circulación vehicular. Durante esos meses, se registró una reducción drástica en la cantidad de accidentes, seguida de un aumento progresivo a medida que se flexibilizaban las restricciones y la actividad urbana retomaba su ritmo habitual.

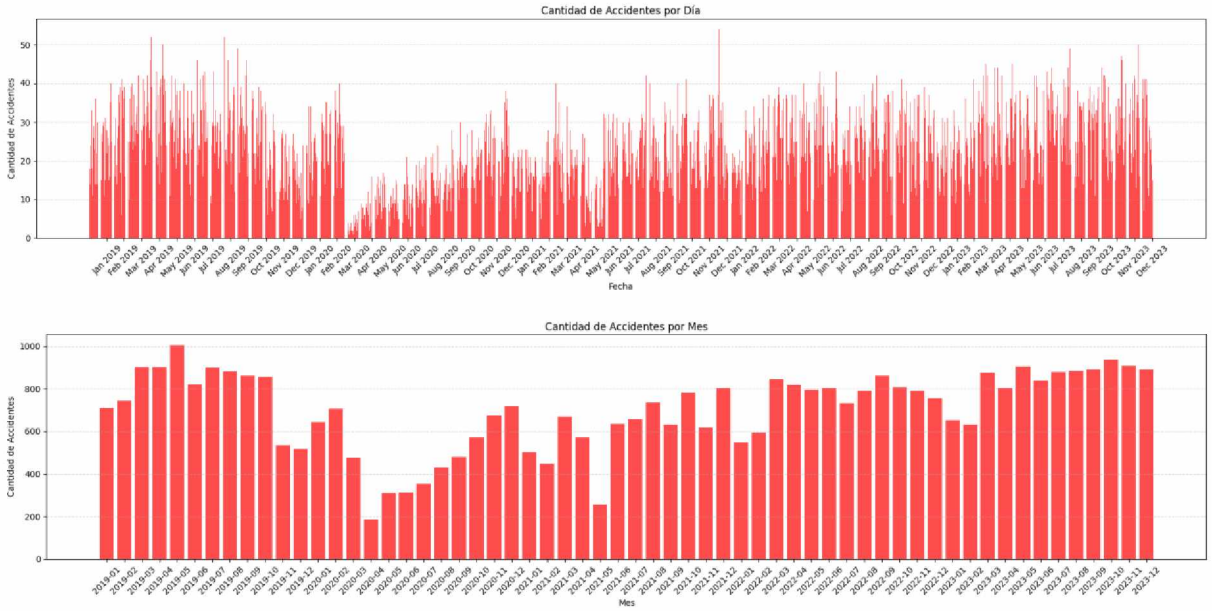


Figura 2: Distribución Temporal de accidentes (Elaboración Propia)

El análisis de la variación horaria de los accidentes permitió identificar patrones significativos en la distribución temporal de los siniestros. Se observó un claro pico en el promedio de accidentes durante las horas de la tarde, coincidiendo con el incremento del tráfico vehicular en ese período. En contraste, se identificó un valle en la madrugada, momento en el que la circulación es considerablemente menor.

Asimismo, al comparar la cantidad de accidentes entre días laborales y fines de semana, se detectó una disminución en la frecuencia de siniestros durante los sábados y domingos en relación con los días hábiles. Este hallazgo podría explicarse por el menor flujo vehicular durante los fines de semana, lo que reduce las probabilidades de accidentes, ya que al haber

menos vehículos en circulación, disminuye la exposición al riesgo. Sin embargo, un hallazgo relevante es el aumento en la cantidad de accidentes durante la madrugada de los fines de semana, lo que podría estar asociado a factores como el ocio nocturno y el consumo de alcohol en estos días.

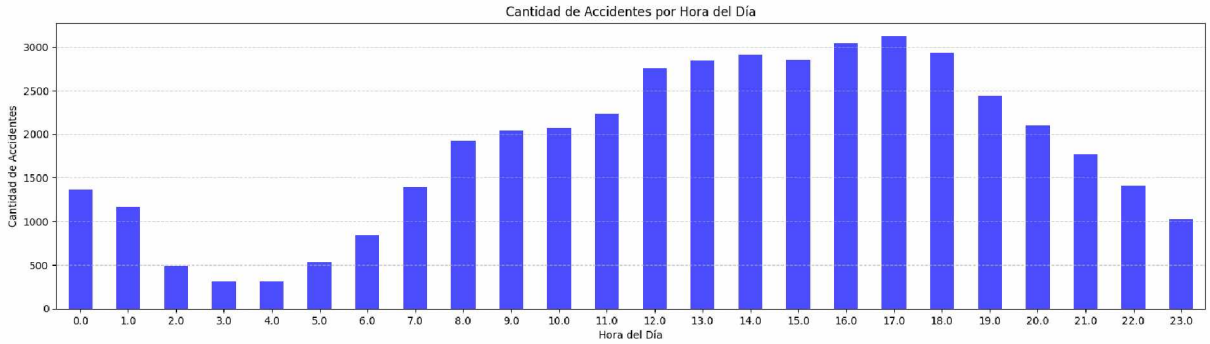


Figura 3: Distribución Horaria de accidentes (Elaboración Propia)

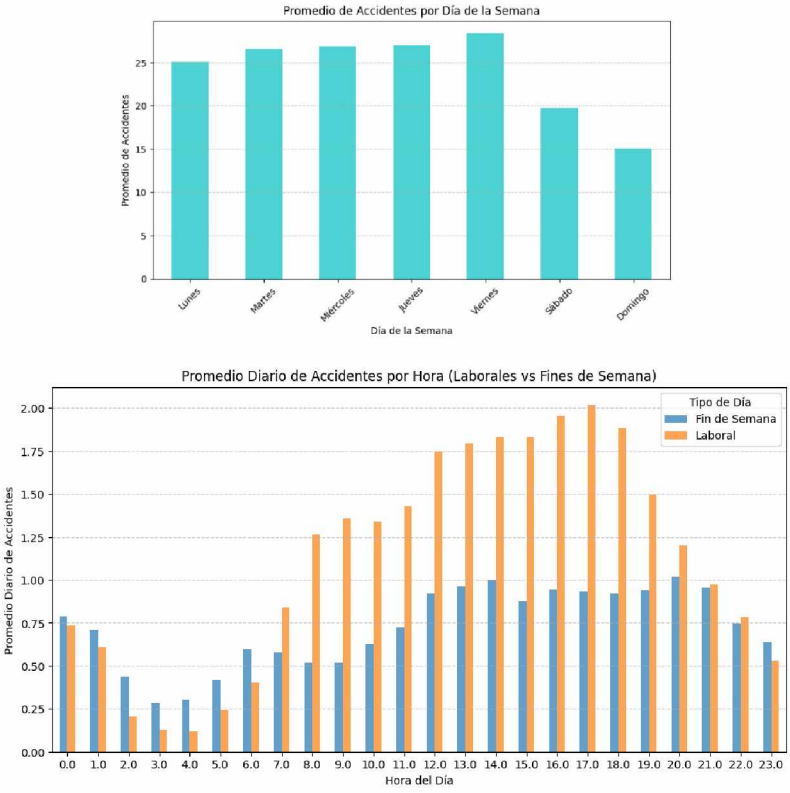


Figura 4: Distribución Temporal de accidentes (Elaboración Propia)

Para complementar el análisis espacial de los accidentes, se elaboró un mapa de calor que permite visualizar las zonas con mayor concentración de siniestros viales. Los resultados muestran una clara intensificación del calor en el área central de la ciudad y a lo largo de las principales avenidas, donde el flujo vehicular es más elevado. Además, se identificaron focos de alta concentración en los barrios de Belgrano y Flores, así como en los principales accesos

a la ciudad, donde convergen múltiples autopistas y avenidas de gran tránsito. Esta distribución sugiere que la densidad del tráfico y la infraestructura vial juegan un papel clave en la ocurrencia de los accidentes.

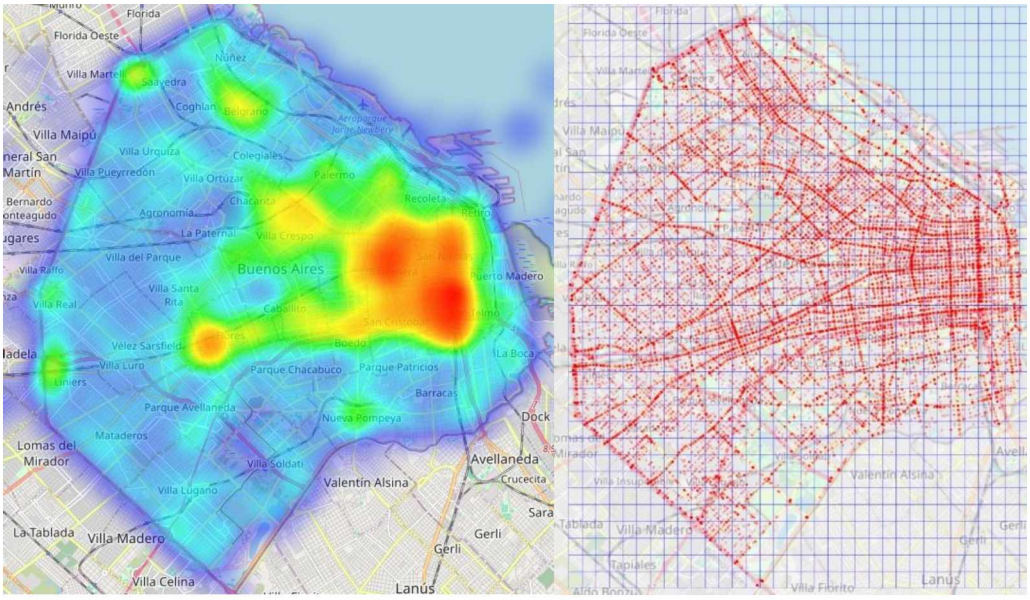


Figura 5: Accidentes Registrados (Elaboración Propia)

Estos hallazgos permitieron una mejor contextualización de los datos y facilitaron la identificación de tendencias relevantes para el estudio de los accidentes de tránsito en la Ciudad Autónoma de Buenos Aires.

3.1.3. Combinación de Accidentes con Celdas y Generación de Variables Lag.

La información sobre accidentes de tránsito se integró con la base de celdas mediante un cruce espacial y temporal. Para cada celda, se asignó el número de accidentes registrados en su área durante cada mes del período analizado (2019-2023), generando una variable clave de accidentes mensuales por celda. Este cruce de datos no solo permitió modelar la siniestralidad en función de las características estructurales, sociales y funcionales de cada celda, sino también identificar tendencias temporales y concentraciones espaciales de los accidentes en la ciudad.

Posteriormente, se generaron variables lag, diseñadas para capturar el efecto de los valores pasados en la predicción de la siniestralidad futura. Estas variables representan el número de accidentes registrados en uno, tres, seis y doce meses anteriores. Su inclusión permitió evaluar si las tendencias pasadas en la siniestralidad afectan los niveles de riesgo en meses posteriores. Este enfoque es particularmente útil para explorar patrones de recurrencia y persistencia en la dinámica de accidentes.

La combinación de datos de accidentes con las celdas y la generación de variables lag enriqueció significativamente la base de datos, permitiendo una representación integral de las condiciones viales y temporales de la ciudad.

3.1.4. Discretización de la Variable Objetivo

Una vez obtenida la información sobre la cantidad de accidentes en cada celda durante el período de análisis, se procedió a discretizar esta variable para transformarla en una variable categórica que permita clasificar los niveles de riesgo de accidentes en la ciudad. La variable de "cantidad de accidentes" se segmentó en tres categorías basadas en umbrales predefinidos, reflejando distintos niveles de riesgo. Estas categorías son: Riesgo Bajo, que se asigna a las celdas con un número de accidentes bajo, según los percentiles de distribución de accidentes; Riesgo Medio, que representa aquellas celdas con un número de accidentes en un rango intermedio; y Riesgo Alto, que se asigna a las celdas con una cantidad elevada de accidentes, que se encuentran en los percentiles más altos de la distribución.

Este proceso de discretización transformó la variable continua de "cantidad de accidentes" en una variable categórica que se utiliza como la variable objetivo del modelo predictivo. El modelo intentará predecir el riesgo de accidente de cada celda, clasificando las celdas en una de las tres categorías mencionadas: bajo, medio o alto.

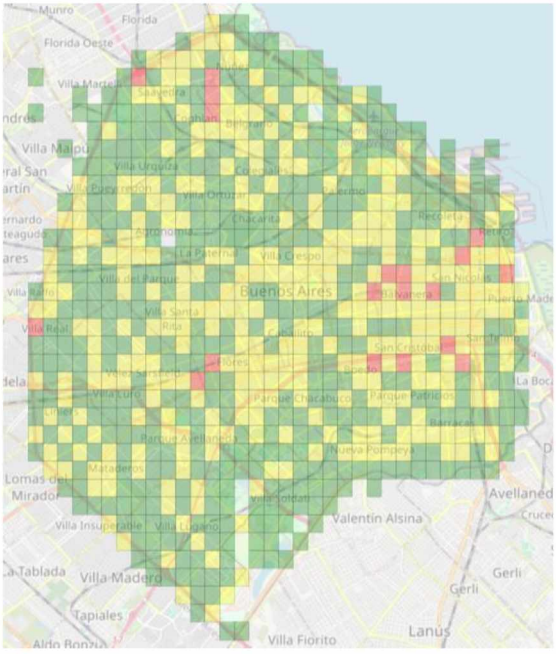


Figura 6: Ejemplo de Grafico de Zonas de Riesgo para un periodo (Elaboración Propia)

La creación de esta variable objetivo fue crucial para el enfoque predictivo, ya que permite aplicar técnicas de clasificación, lo cual facilita la interpretación y utilización de los resultados para la toma de decisiones en la gestión de la siniestralidad vial.

3.1.5. Limpieza y Filtrado de Celdas sin Accidentes

En el proceso de preparación de los datos, se realizó una limpieza adicional de las celdas que nunca tuvieron accidentes registrados durante el período de análisis. Estas celdas, muchas de las cuales se encontraban fuera de los límites urbanos de la Ciudad de Buenos Aires, fueron eliminadas del conjunto de datos.

La presencia de estas celdas sin accidentes en las áreas periféricas y no urbanas generaba un desbalance significativo en la distribución de la variable objetivo, especialmente en la clase de "Riesgo Bajo", al representar una gran cantidad de celdas con valor nulo o cero en cuanto a accidentes. Esta limpieza permitió reducir el sesgo en la clasificación y mejorar la calidad del modelo predictivo, asegurando que las celdas utilizadas fueran representativas de las áreas urbanas y las dinámicas reales de siniestralidad vial en la ciudad.

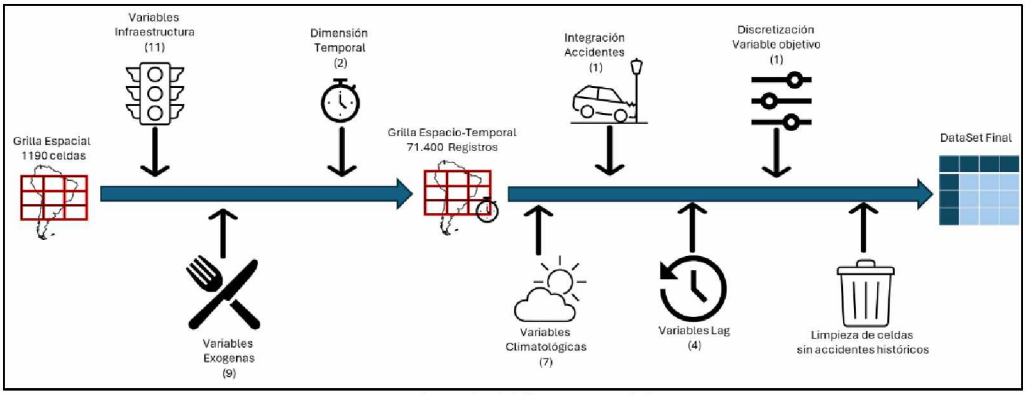


Figura 7: Diagrama Armado del Set Datos (Elaboración Propia)

Este trabajo preparatorio asegura que los datos estén organizados, enriquecidos y listos para su uso en el modelado predictivo, que se abordará en los apartados siguientes.

3.1.6. Implementación del Modelo Predictivo

En este estudio, se implementaron varios modelos de clasificación para estimar el riesgo de accidentes viales, utilizando tres categorías: bajo, medio y alto. Los modelos aplicados fueron el RandomForestClassifier y el GradientBoostingClassifier, ambos adecuados para tareas de clasificación multicategoría.

A continuación, se describe cómo se implementaron y entrenaron estos modelos, así como los enfoques utilizados para evaluar su desempeño.

3.1.7. Modelo Random Forest

El RandomForestClassifier se eligió por su capacidad para manejar grandes volúmenes de datos y por su robustez frente al sobreajuste. Para entrenar este modelo, se utilizó un enfoque de ventana deslizante (sliding window), donde se entrenó el modelo con un período de 12 meses de datos históricos y luego se evaluó su desempeño con el mes siguiente como conjunto de prueba. Este enfoque permite simular un escenario real en el que los modelos se entrenan continuamente con datos recientes, mientras que las predicciones se hacen para períodos futuros.

El modelo de Random Forest fue entrenado y validado de esta manera para predecir el riesgo de accidentes en función de las variables analizadas, clasificando cada celda de la ciudad en una de las tres categorías de riesgo.

Optimización de Parámetros

Para mejorar la precisión del modelo, se realizó una optimización de parámetros utilizando la técnica de búsqueda en malla (GridSearchCV). Este enfoque permite probar diferentes combinaciones de parámetros para encontrar la configuración que maximiza el rendimiento del modelo. Se ajustaron parámetros clave como el número de árboles en el bosque (n_estimators) y la profundidad máxima de los árboles (max_depth). La optimización de estos parámetros garantizó que el modelo estuviera bien ajustado a los datos sin sobre ajustarse, mejorando así su capacidad de generalización y rendimiento en datos no vistos.

3.2. Evaluación de los Modelos

Para la evaluación de los modelos, se utilizó inicialmente un modelo base más simple, con el objetivo de establecer una línea de referencia sobre la cual comparar los modelos más complejos. Este primer modelo base consistió en una regresión logística, que permitió evaluar la capacidad inicial de clasificación de las zonas de riesgo de accidentes viales en las categorías de bajo, medio y alto. A partir de los resultados obtenidos de este modelo base, se avanzó hacia la implementación de modelos más sofisticados, como el Random Forest y el Gradient Boosting, para mejorar la precisión y la capacidad predictiva.

La evaluación se centró en analizar cómo cada modelo clasifica las zonas de riesgo en las tres categorías mencionadas, utilizando métricas clave para su evaluación, así como la construcción de matrices de confusión. También se emplearon enfoques como la metodología

de ventana deslizante y el enfoque de Ventana Rolling para evaluar el impacto de los datos históricos y el comportamiento evolutivo de los riesgos de accidentes en el desempeño de los modelos. A continuación, se detallan los resultados y las comparaciones entre los diferentes modelos implementados.

3.2.1. Modelo Base de Regresión Logística

El modelo base de regresión logística sirvió como un punto de partida para la clasificación de las zonas de riesgo en las categorías de bajo, medio y alto. Este modelo es una técnica de clasificación lineal que asigna probabilidades a cada clase según las características de las zonas de riesgo. Aunque es un modelo simple, proporciona una línea base útil para comparar con modelos más complejos.

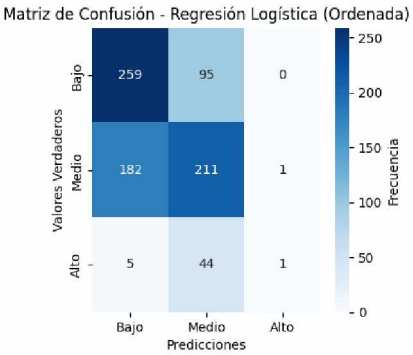


Figura 8: Elaboración Propia

La evaluación del modelo base se realizó mediante la construcción de una matriz de confusión que permitió identificar los aciertos y errores en la clasificación de cada zona. Además, se calcularon métricas clave como la precisión, el recall y el F1-score para evaluar el rendimiento general del modelo.

Clase	Precision	Recall	F1-Score	Support
Alto	0.50	0.02	0.04	50
Bajo	0.58	0.73	0.65	354
Medio	0.60	0.54	0.57	394
Accuracy			**0.59**	798
Macro avg	0.56	0.43	0.42	798
Weighted avg	0.59	0.59	0.57	798

Figura 9: Elaboración Propia

El modelo base de regresión logística proporcionó una línea de base con la que comparar otros enfoques más sofisticados, pero sus limitaciones en la clasificación de las zonas de alto riesgo subrayaron la necesidad de técnicas más complejas que pudieran manejar mejor las interacciones no lineales entre las variables.

3.2.2. Modelo Random Forest

Como se mencionó anteriormente, es entrenó un modelo de RandomForestClassifier. Una vez entrenado, se construyó la matriz de confusión para evaluar el desempeño del modelo Random Forest. La matriz de confusión proporcionó información sobre el número de aciertos y errores cometidos al clasificar las celdas en sus respectivas categorías de riesgo (bajo, medio, alto).

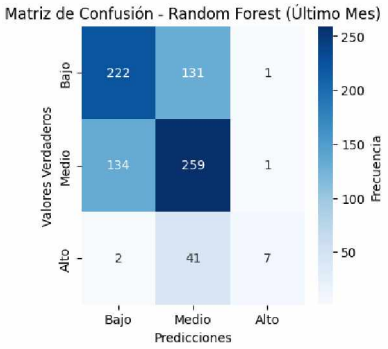


Figura 10: Elaboración Propia

Además, se calcularon métricas clave como la precisión, el recall y el F1-score, lo que permitió evaluar la capacidad del modelo para predecir con precisión el riesgo de accidentes, minimizando los falsos positivos y negativos.

Clase	Precision	Recall	F1-Score	Support
Alto	0.78	0.14	0.24	50
Bajo	0.62	0.63	0.62	354
Medio	0.60	0.66	0.63	394
Accuracy			**0.61**	798
Macro avg	0.67	0.47	0.50	798
Weighted avg	0.62	0.61	0.60	798

Figura 11: Elaboración Propia

El modelo Random Forest ofrece un rendimiento global más robusto y confiable, especialmente en la clasificación de las clases más comunes ("Bajo" y "Medio"), mejorando tanto la precisión como el recall y F1-score en comparación con el modelo base de regresión logística. A pesar de que la clase "Alto" sigue siendo difícil de predecir, este modelo ofrece una mejora significativa y, por lo tanto, se considera una mejor opción para abordar la tarea de clasificación del riesgo de accidentes.

3.2.3. Modelo Gradient Boosting

El GradientBoostingClassifier fue empleado como un modelo alternativo de ensamblaje que construye árboles de decisión de manera secuencial, corrigiendo los errores cometidos por los árboles anteriores. Este enfoque se implementó también utilizando la metodología de ventana deslizante (12 meses de entrenamiento y 1 mes de prueba), lo que permitió comparar su desempeño con el modelo Random Forest en la clasificación de las tres categorías de riesgo de accidentes.

De manera similar al modelo de Random Forest, se construyó la matriz de confusión para evaluar el desempeño del Gradient Boosting.

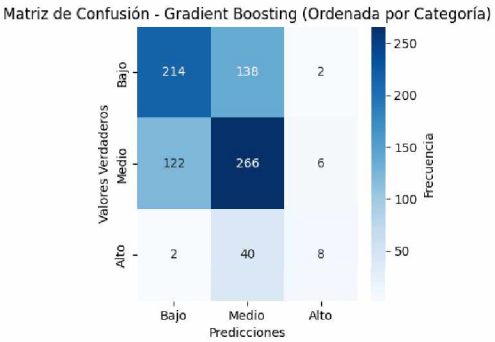


Figura 12: Elaboración propia

Se analizaron las métricas de rendimiento, incluyendo la precisión, el recall y el F1-score, para determinar cuán bien se logró clasificar cada celda de acuerdo con su nivel de riesgo. Este análisis comparativo entre Gradient Boosting y Random Forest ayudó a identificar cuál de los dos modelos ofreció un mejor desempeño en la clasificación de riesgos de accidentes.

Clase	Precision	Recall	F1-Score	Support
Alto	0.50	0.16	0.24	50
Bajo	0.63	0.60	0.62	354
Medio	0.60	0.68	0.63	394
Accuracy	0.61			798
Macro Avg	0.58	0.48	0.50	798
Weighted Avg	0.61	0.61	0.60	798

Figura 13: Elaboración propia

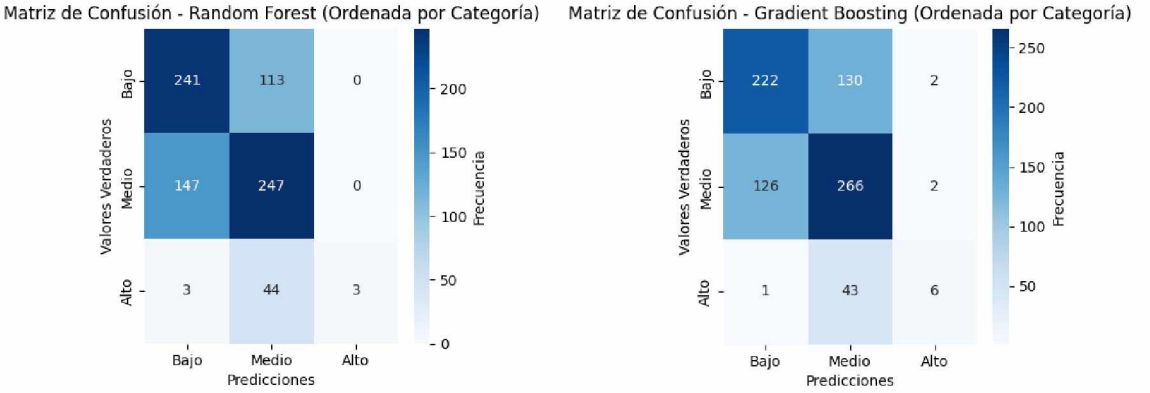
Ambos modelos presentan ventajas y desventajas según la clase evaluada. El modelo de Random Forest muestra un mejor rendimiento en la predicción de la clase "Alto", alcanzando una mayor precisión en comparación con Gradient Boosting. Por otro lado, Gradient Boosting destaca ligeramente en la predicción de la clase "Medio", donde presenta un mayor recall. En términos generales, ambos modelos ofrecen resultados similares en la clasificación de las clases "Bajo" y "Medio". Ninguno de los dos modelos presenta una superioridad clara y definitiva sobre el otro.

3.2.4. Enfoque de Ventana Rolling

Además del enfoque de ventana deslizante, se probó un modelo con la técnica de ventana rolling, en la cual los datos de entrenamiento no se limitan a los 12 meses más recientes, sino que se van acumulando mes a mes, incorporando toda la información disponible hasta el momento. Esta estrategia de entrenamiento acumulativo permite al modelo "recordar" patrones a largo plazo y adaptar sus predicciones con base en toda la historia de los datos, en lugar de limitarse a un período fijo.

Este enfoque fue particularmente útil para evaluar el impacto de los datos más antiguos en las predicciones y para observar cómo los modelos pueden adaptarse a la evolución temporal de los riesgos de accidentes.

Para este enfoque, también se construyeron las matrices de confusión y se analizaron las métricas de desempeño. Comparando los resultados obtenidos con los modelos de ventana deslizante, se evaluó si la inclusión de datos acumulativos en la técnica de ventana rolling mejora la capacidad predictiva de los modelos en términos de clasificación del riesgo de accidentes a lo largo del tiempo.



Clase	Precision	Recall	F1-Score	Support
Alto	1.00	0.06	0.11	50
Bajo	0.62	0.68	0.65	354
Medio	0.61	0.63	0.62	394
Accuracy			**0.62**	798
Macro avg	0.74	0.46	0.46	798
Weighted avg	0.64	0.62	0.60	798

Clase	Precision	Recall	F1-Score	Support
Alto	0.60	0.12	0.20	50
Bajo	0.64	0.63	0.63	354
Medio	0.61	0.68	0.64	394
Accuracy			**0.62**	798
Macro avg	0.61	0.47	0.49	798
Weighted avg	0.62	0.62	0.61	798

Figura 14: Elaboración propia

El desempeño de la ventana rolling muestra ciertas diferencias en comparación con los modelos deslizantes. En el caso de Random Forest rolling, la precisión para la clase "Alto" es muy alta (1.00), pero con un recall extremadamente bajo (0.06), lo que indica que el modelo predice pocos casos de esta clase correctamente. Sin embargo, presenta un recall superior para las clases "Bajo" y "Medio", mejorando ligeramente la clasificación de estas categorías. En cuanto a Gradient Boosting rolling, los resultados son similares a los modelos deslizantes en

términos de precisión y recall, manteniendo un accuracy del 62%. En general, la estrategia rolling no introduce mejoras significativas sobre los modelos deslizantes, pero sí muestra una tendencia a afectar el equilibrio entre precisión y recall en la clase menos representada.

3.2.5. Balanceo de Clases

En el análisis de la predicción de accidentes de tránsito, nos enfrentamos a un desafío importante relacionado con el desbalanceo de clases. En este caso, las clases que intentamos predecir no tienen una distribución equitativa. La mayoría de las celdas presentan una baja cantidad de accidentes, mientras que las celdas con una mayor cantidad de accidentes, que corresponden a la clase Alto, son mucho menos frecuentes.

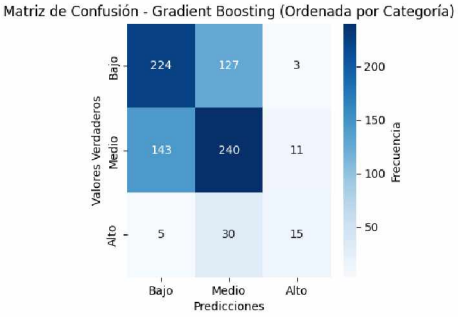
Este desbalance en las clases presenta una dificultad para los modelos de predicción, ya que, sin un manejo adecuado, el modelo puede sesgarse hacia la clase mayoritaria (en este caso, Bajo), lo que significa que podría predecir con mayor facilidad las celdas con menos accidentes, pero fallar al identificar correctamente las celdas con una mayor cantidad de accidentes, que son las de mayor interés para la prevención y la intervención.

Por lo tanto, es fundamental aplicar técnicas de balanceo de clases, las cuales permiten ajustar el modelo para que no solo se enfoque en la clase mayoritaria, sino que también sea capaz de predecir de manera más precisa las celdas con una alta concentración de accidentes. Esto contribuirá a mejorar la efectividad de las predicciones, especialmente en áreas de la ciudad donde se requieren acciones más inmediatas y focalizadas para mitigar los riesgos.

Para abordar el desbalanceo de clases en el conjunto de datos, se aplicó la técnica de SMOTE (Synthetic Minority Over-sampling Technique). Esta metodología permite generar muestras sintéticas de la clase minoritaria, en este caso, las celdas con una alta cantidad de accidentes, para equilibrar la distribución de las clases. Esto provoca que el modelo se vuelve más sensible a las clases menos representadas, mejorando así la capacidad de predicción, especialmente en las áreas con mayor número de accidentes. Este enfoque asegura que el modelo no se sesgue hacia la clase mayoritaria y pueda identificar con mayor precisión las celdas con altos índices de accidentes.

Se aplicó tanto al modelo de Random Forest como al Gradient Boosting, y se obtuvieron los siguientes resultados:

Gradient Boosting

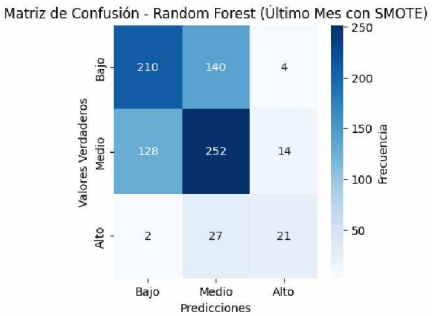


Clase	Precision	Recall	F1-Score	Support
Alto	0.52	0.30	0.38	50
Bajo	0.60	0.63	0.62	354
Medio	0.60	0.61	0.61	394
Accuracy			**0.60**	798
Macro avg	0.57	0.51	0.53	798
Weighted avg	0.60	0.60	0.60	798

Figura 15: Elaboración propia

El modelo de Gradient Boosting con balanceo de clases presenta una mejora notable con respecto a los resultados obtenidos previamente, especialmente en comparación con el modelo de Gradient Boosting sin balanceo de clases. Aunque la precisión y el recall de la clase "Alto" siguen siendo relativamente bajos, se observa un incremento en su recall (30%) en comparación con el modelo original (16%), lo cual es una mejora significativa en la identificación de las celdas con alta cantidad de accidentes. El promedio ponderado de todas las clases se mantiene en 0.60, lo que sugiere que, aunque el modelo ha mejorado en la clasificación de las clases minoritarias, aún existe un espacio para continuar afinando el balance entre precisión y recall para todas las clases.

Random Forest



Clase	Precision	Recall	F1-Score	Support
Alto	0.54	0.42	0.47	50
Bajo	0.62	0.59	0.61	354
Medio	0.60	0.64	0.62	394
Accuracy			**0.61**	798
Macro avg	0.59	0.55	0.57	798
Weighted avg	0.60	0.61	0.60	798

Figura 16: Elaboración propia

El modelo de Random Forest con balanceo de clases muestra una mejora notable respecto a los resultados obtenidos en modelos previos. En particular, el recall para la clase "Alto" ha aumentado significativamente, alcanzando un 42% en comparación con el 14% en el modelo de Random Forest deslizante, lo que indica una mejor capacidad para identificar las celdas con una alta cantidad de accidentes. Aunque el valor de la precisión no es tan alto, el equilibrio entre precisión y recall resulta en un f1-score de 0.47, lo cual es una mejora en comparación con el modelo anterior. El promedio ponderado sigue siendo 0.60, sugiriendo que el modelo logra un buen desempeño general, especialmente al identificar correctamente las clases "Bajo" y "Medio".

3.3. Selección de los Modelos: Matriz de Confusión e Interpretación.

Una vez entrenados y evaluados los modelos, se analizaron las matrices de confusión para comparar la precisión de cada modelo y entender en qué medida estos lograron clasificar correctamente las celdas en sus respectivas categorías de riesgo (bajo, medio y alto). Además, se realizaron análisis sobre las variables más importantes en el modelo, lo que permitió identificar los factores más influyentes en la predicción del riesgo de accidentes. Estos hallazgos proporcionan información valiosa sobre los determinantes de la siniestralidad vial y pueden ser útiles para futuras intervenciones en políticas de seguridad vial.

3.3.1. Selección del Modelo.

Tras analizar los resultados obtenidos de los diferentes modelos aplicados, se concluye que el Random Forest con balanceo de clases es el modelo más equilibrado y eficiente para la clasificación de las tres clases ("Alto", "Bajo" y "Medio"). Este modelo destaca principalmente por su capacidad para manejar el desbalance de clases, logrando un mejor recall para la clase "Alto" en comparación con los otros enfoques. Aunque el accuracy de este modelo es similar al de otros modelos, el macro average F1-score y el weighted average F1-score demuestran un mejor desempeño en comparación con el resto, particularmente en términos de precisión y recall para las clases "Bajo" y "Medio".

En resumen, el Random Forest con balanceo de clases es el modelo que muestra el mejor equilibrio entre las métricas de precision, recall, y f1-score, y maneja de manera más eficiente las desventajas causadas por el desbalance de clases, lo que lo convierte en la opción más robusta y recomendada para este análisis.

3.3.2. Matriz de Confusión Ponderada: Evaluación con Costos Diferenciados.

Finalmente, se implementaron matrices de confusión "ponderadas", que consideran la diferencia en los costos asociados con los distintos tipos de errores en la clasificación de riesgo (bajo, medio y alto). En este contexto, un error grave es clasificar una zona de alto riesgo como de bajo riesgo, ya que puede llevar a la falta de intervención en un área peligrosa. En cambio, errores como clasificar una zona de alto riesgo como medio, o una zona de medio riesgo como bajo, aunque importantes, no tienen el mismo impacto crítico en términos de seguridad.

El costo de error más grave ocurre cuando una zona de bajo riesgo se clasifica como alto riesgo, ya que esto puede generar intervenciones innecesarias o mal dirigidas, lo que podría implicar el uso de recursos en áreas que no lo necesitan. Por otro lado, cuando una zona de bajo riesgo se clasifica como medio riesgo, el costo es menor, aunque aún representa un error en la asignación de recursos.

El costo asociado a la clasificación incorrecta de una zona de medio riesgo, como bajo riesgo, es también importante, aunque no tan crítico como los errores que implican clasificar zonas de alto riesgo. Clasificar una zona de alto riesgo como medio riesgo, aunque es un error, tiene un impacto moderado, ya que aún podría generarse alguna intervención, aunque no la adecuada.

Finalmente, el error más grave se produce cuando una zona de alto riesgo se clasifica como bajo riesgo, ya que esto puede resultar en la falta de acción en áreas que requieren atención urgente para prevenir accidentes, lo que puede tener consecuencias graves.

Al ponderar estos errores según su gravedad, se obtuvo una evaluación más precisa de cómo el modelo influye en la toma de decisiones para la gestión de la siniestralidad vial. Este enfoque permite mejorar la efectividad de las intervenciones preventivas y la asignación de recursos en áreas de mayor riesgo.

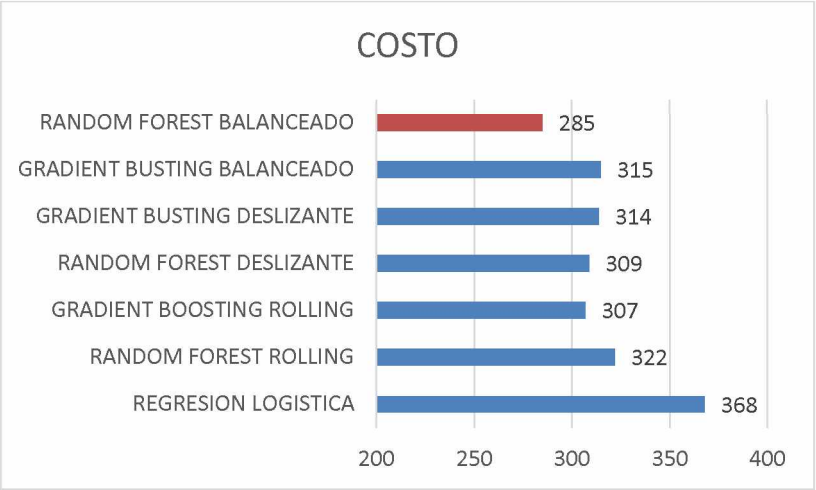


Figura 17: Elaboración propia

De manera consistente con el análisis previo, la evaluación basada en los costos asociados a los errores de clasificación confirma que el modelo Random Forest balanceado es el más efectivo. Con un costo total de 285, se posiciona como la mejor alternativa para minimizar los errores más críticos en la categorización del riesgo vial. Este resultado refuerza su capacidad para optimizar la asignación de recursos y mejorar la gestión de la siniestralidad, permitiendo una toma de decisiones más precisa y eficiente en la prevención de accidentes.

3.3.3. Interpretación del Modelo.

Una parte importante del análisis fue la interpretación de los modelos, particularmente la identificación de las variables más relevantes en la predicción del riesgo de accidentes. El Random Forest es particularmente útil para este tipo de análisis, ya que permite evaluar la importancia de cada variable en función de su contribución al rendimiento del modelo. Al comprender cuáles son los factores más influyentes, se pueden tomar decisiones más informadas sobre las áreas que requieren atención prioritaria.

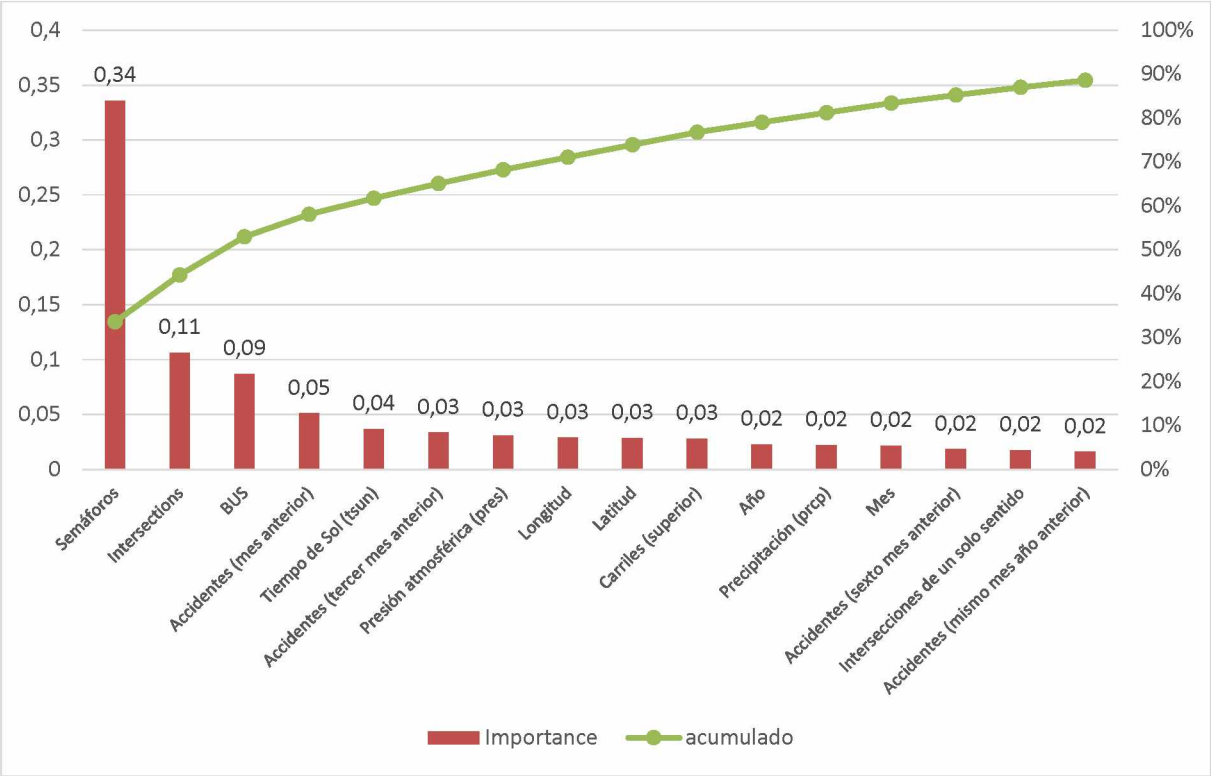


Figura 17: Elaboración propia

El análisis de importancia de las características revela que la variable más influyente en la predicción de la siniestralidad vial es la cantidad de semáforos, lo que indica que su presencia tiene un fuerte impacto en la ocurrencia de accidentes, posiblemente debido a su relación con intersecciones de alto flujo vehicular.

En segundo lugar, la cantidad de intersecciones también juega un papel clave, ya que las zonas con una mayor densidad de cruces suelen presentar un riesgo elevado debido a la complejidad del tráfico y las maniobras de los vehículos.

La presencia de paradas de autobús se destaca como un factor relevante, lo que sugiere que los puntos de ascenso y descenso de pasajeros generan interrupciones en la circulación y posibles conflictos con otros actores viales. Asimismo, los accidentes ocurridos en el mes anterior son un fuerte predictor del riesgo actual, lo que evidencia patrones de siniestralidad persistentes en ciertas áreas.

En cuanto a variables ambientales, factores como la precipitación también presentan una influencia baja a moderada, lo que sugiere que las condiciones climáticas pueden contribuir a la peligrosidad de las vías. Por otro lado, variables como la cantidad de hospitales, universidades y bibliotecas tienen una importancia mínima o nula en el modelo, lo que indica que su impacto en la siniestralidad es irrelevante.

3.3.4. Estabilidad del modelo

Una vez entrenado y evaluado el modelo, es importante analizar su estabilidad y comportamiento a lo largo del tiempo, especialmente en contextos dinámicos como el análisis de riesgos viales. Para evaluar la capacidad del modelo de mantenerse robusto y efectivo frente a cambios en los datos, se implementó una iteración temporal de la predicción utilizando los últimos 12 meses de datos disponibles. Este análisis busca evaluar si el modelo mantiene su precisión y capacidad de generalización cuando se enfrenta a nuevos datos, no utilizados durante el proceso de entrenamiento.

Este enfoque permite observar cómo el modelo maneja la variabilidad de las características a lo largo del tiempo, lo que puede influir en su desempeño predictivo.

La evaluación de la estabilidad tiene como objetivo identificar posibles variaciones en la capacidad del modelo para predecir el riesgo de accidentes. Esta iteración temporal no solo permite validar la robustez del modelo, sino que también proporciona información sobre su capacidad para adaptarse a nuevas condiciones y escenarios, lo que resulta fundamental para su implementación práctica en la gestión de la seguridad vial.

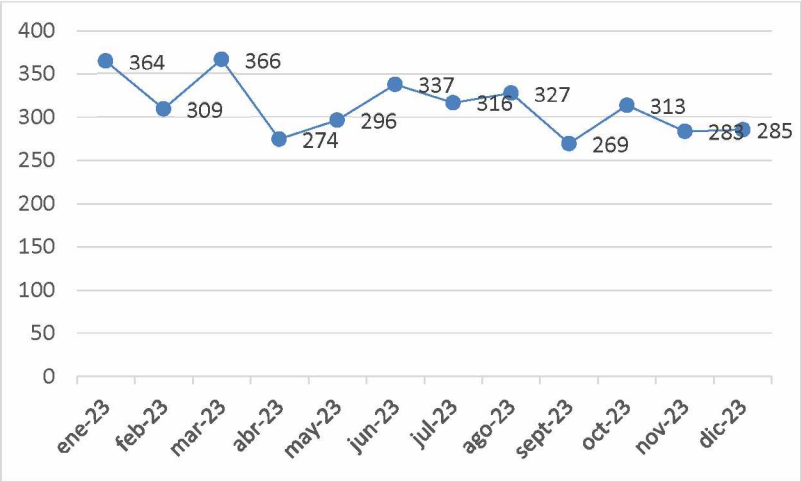


Figura 18: Elaboración propia

La evolución del costo ponderado a lo largo de los 12 meses refleja cómo el modelo ha manejado los errores de clasificación según los diferentes costos asignados en la matriz de confusión. A pesar de algunas fluctuaciones en los primeros meses, la tendencia general muestra una estabilización, indicando que el modelo ha logrado ajustar su desempeño. Los valores más altos en meses específicos pueden indicar errores de clasificación más graves, como la asignación incorrecta de zonas de alto riesgo, mientras que la disminución del costo en meses posteriores refleja una mejora en la precisión del modelo al clasificar correctamente las zonas según su nivel de riesgo. Este análisis resalta la capacidad del modelo para

minimizar los costos asociados a los errores críticos de clasificación, lo cual es fundamental para la correcta asignación de recursos en la gestión de la siniestralidad vial.

Conclusiones

En el presente trabajo se han desarrollado y evaluado modelos de clasificación utilizando algoritmos de Random Forest y Gradient Boosting con el objetivo de predecir el riesgo de accidentes de tránsito en la Ciudad de Buenos Aires, basado en variables urbanas y ambientales, encontrándose en el modelo Random Forest con balanceo de clases el mejor desempeño.

Tras evaluar el modelo con las métricas de desempeño, se obtuvieron resultados mixtos. Las clases 'Bajo' y 'Medio' mostraron un rendimiento aceptable lo que indica que el modelo puede identificar estos casos con una efectividad razonable. El rendimiento de la clase 'Alto' fue menos satisfactorio, lo que sugiere que el modelo tiene dificultades para identificar correctamente los casos de alto riesgo. Aunque la clase "Alto" tiene un bajo recall, es importante señalar que muchos de los errores de clasificación se distribuyen hacia la clase "Medio", lo cual puede no ser tan grave en términos de intervención, ya que los casos de riesgo elevado no se clasifican erróneamente como de bajo riesgo. Este hallazgo resalta la necesidad de ajustar y optimizar el modelo para mejorar su capacidad de detección en situaciones críticas, específicamente en la clasificación de los casos de alto riesgo.

A través de la investigación realizada, se ha evidenciado que, aunque las variables seleccionadas ofrecen información valiosa, el modelo actual presenta un rendimiento que es mejorable.

Este trabajo aporta una base para la comprensión de los factores que influyen en la siniestralidad vial y establece un punto de partida para el desarrollo de modelos más precisos. Además, proporciona una perspectiva de cómo los datos urbanos y ambientales pueden integrarse para predecir el riesgo de accidentes en contextos urbanos complejos. La identificación de las características clave del modelo también abre la puerta a investigaciones adicionales sobre qué variables pueden tener un mayor impacto en la predicción de accidentes.

Los resultados obtenidos, ofrecen una plataforma para futuras investigaciones en el ámbito de la seguridad vial urbana. Los modelos de predicción podrían utilizarse en el diseño de políticas públicas, especialmente para priorizar la implementación de medidas preventivas en áreas de mayor riesgo. Esta investigación tiene el potencial de influir en las decisiones sobre

dónde asignar recursos para la mejora de la infraestructura vial, la regulación del tránsito y la implementación de campañas de concientización.

Entre las líneas de trabajo futuro, destacan la inclusión de variables adicionales, y se recomendaría realizar ajustes en la estrategia de modelado. También es interesante explorar la implementación de técnicas de deep learning y redes neuronales, que podrían capturar patrones más complejos y mejorar la precisión en la predicción de accidentes.

Otra línea de investigación relevante sería la aplicación de modelos de regresión para estimar la cantidad de accidentes en lugar de clasificarlos en categorías de riesgo. Este enfoque permitiría obtener estimaciones más precisas del volumen de accidentes en distintas zonas de la ciudad, lo que facilitaría la toma de decisiones más informadas en términos de políticas públicas y asignación de recursos para la prevención de accidentes.

Referencias bibliográficas

Agencia Nacional de Seguridad Vial (ANSV). (2023). Informe de Siniestralidad Vial Fatal. https://www.argentina.gob.ar/seguridadvial/observatoriovialnacional/estadisticas-observatorio

Alhaek, F., Liang, W., Rajeh, T. M., Javed, M. H., & Li, T. (2024). Learning spatial patterns and temporal dependencies for traffic accident severity prediction: A deep learning approach. Knowledge-Based Systems, Volume 286. https://doi.org/10.1016/j.knosys.2024.111406

Cardona Álvarez, J. (2023). Modelo predictivo de zonas de riesgo espacio temporal de accidentes de tráfico en la ciudad de Manizales. Disponible en: https://repositorio.ucaldas.edu.co/handle/ucaldas/19537

Ciudad Autónoma de Buenos Aires, Secretaría de transporte. Plan de Seguridad Vial de la Ciudad 2020-2023. https://buenosaires.gob.ar/movilidad/plan-de-seguridad-vial/plan-de-seguridad-vial-de-la-ciudad-2020-2023

Cui, P., Yang, X., Abdel-Aty, M., Zhang, J., & Yan, X. (2024). Advancing urban traffic accident forecasting through sparse spatio-temporal dynamic learning. Accident Analysis & Prevention, Volume 200. https://doi.org/10.1016/j.aap.2024.107564

Dirección de Investigación Accidentológica del Observatorio Nacional Vial. (2022). Consecuencias de la siniestralidad vial en Argentina: Actualización de indicadores: Carga Global de Enfermedad y Costos sociales. Año 2019. Noviembre 2022.

Evans, L. (1991). Traffic Safety and the Driver. Van Nostrand Reinhold.

Gobierno de la Ciudad Autónoma de Buenos Aires. (2022). Informe estadístico sobre las víctimas fatales a causa de siniestros viales Ciudad Autónoma de Buenos Aires. Año 2022 Gobierno de la Ciudad. https://buenosaires.gob.ar/sites/default/files/2024-01/Informe_web_Victimas_Fatales_2022_0.pdf

Híjar, M., Carrillo, C., Flores, M., Anaya, R., & Lopez, V. (2000). Risk factors in highway traffic accidents: A case control study. Accident Analysis & Prevention, 32(4), 495-504. https://doi.org/10.1016/S0001-4575(99)00116-5

Li, Y., Li ,M., Yuan J., Lu J., Abdel-Aty, M. (2021). Analysis and prediction of intersection traffic violations using automated enforcement system data. Accident Analysis & Prevention, Volume 162. https://doi.org/10.1016/j.aap.2021.106422

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. R News, 2(3), 18-22.

Ministerio de Transporte Argentina. Dirección Nacional de Observatorio Vial. (2023). Anuario Estadístico de siniestralidad fatal Año 2021. https://www.argentina.gob.ar/sites/default/files/2018/12/anuario_estadistico_2021.pdf

Ramírez, A. F., & Valencia, C. (2021). Spatiotemporal correlation study of traffic accidents with fatalities and injuries in Bogota (Colombia). Accident Analysis & Prevention, Volume 149. https://doi.org/10.1016/j.aap.2020.105848

Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot is More Informative than the ROC Plot. Nature Biotechnology, 37(4), 1-2.

Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Warehousing, 5(4), 13-22.

Song, L., Li, Y., Fan, W. (D.), & Wu, P. (2020). Modeling pedestrian-injury severities in pedestrian-vehicle crashes considering spatiotemporal patterns: Insights from different hierarchical Bayesian random-effects models. Analytic Methods in Accident Research, Volume 28. https://doi.org/10.1016/j.amar.2020.100137

Vorko-Jović, A., Kern, J., & Biloglav, Z. (2006). Risk factors in urban road traffic accidents. Journal of Safety Research, 37(2), 137-145. https://doi.org/10.1016/j.jsr.2005.08.009

Enlace a carpeta de Google Drive:

https://drive.google.com/drive/folders/1QGR0uTJmPuwNST0aIekDr4qX3nnZbMUY?usp=sharing

Reporte del Tutor

El trabajo final presentado por el candidato incluye elementos propios de las técnicas cuantitativas aprendidas en carrera de especialización, junto con la georreferenciación de la información, integrando información geoespacial con información temporal. La motivación de este trabajo se enmarca en los objetivos del Plan de Seguridad Vial 2020-2023 que la Ciudad de Buenos Aires oportunamente definió y que busca reducir en un 50% las víctimas fatales para el año 2030. A partir de allí, hay una clara pregunta de investigación que el candidato buscó responder: "¿Cómo afectan los patrones de tráfico y las características urbanas a la frecuencia y gravedad de los accidentes de tránsito en la Ciudad de Buenos Aires, y cómo puede utilizarse esta información para predecir zonas con mayores probabilidades de accidentes?"
Todas las fases y pasos de la implementación de la metodología son abordadas con lenguaje claro, profesional y técnico, haciendo especial énfasis en la evaluación del modelo a la luz de las conclusiones obtenidas.

Este trabajo refleja una adecuada integración de motivación, metodología y análisis, lo que evidencia un desempeño satisfactorio en el desarrollo del tema.

Propongo entonces, que el mismo sea elevado a la Comisión Académica de la carrera, para su correspondiente consideración.

Pablo Caviezel

pcaviezel@economicas.uba.ar