Diseño y validación de un modelo predictivo para la estimación de la movilidad urbana en la localidad de Kennedy, Bogotá

Andrés Mauricio Martínez Cardona

Juan Diego Diaz Bermeo

Universidad Ean

Facultad de Ingeniería

Ingeniería de sistemas

Bogotá, Colombia

2026

Diseño y validación de un modelo predictivo para la estimación de la movilidad urbana en la localidad de Kennedy, Bogotá

Andres Mauricio Martinez Cardona

Juan Diego Diaz Bermeo

Trabajo de grado presentado como requisito para optar al título de:

Ingeniería de Sistemas

Director (a):

Lina María Chacon Rivera

Universidad Ean

Facultad de Ingeniería

Ingeniería de sistemas

Bogotá, Colombia

2026

Nota de aceptación:

Firma del jurado

Firma del jurado

Firma del director del trabajo de grado

Bogotá, 02/06/2026

Tabla de contenido

Resumen **6**

Abstract **7**

Introducción **8**

Marco teórico **11**

Smart Cities y movilidad inteligente 11

Machine Learning aplicado a la predicción de tráfico 12

Big Data y análisis espacial en movilidad 12

APIs geográficas y datos en tiempo real 13

Crowdsourcing ciudadano y sistemas colaborativos 14

Seguridad urbana y análisis de criminalidad geográfica 14

Estado del arte **16**

Estrategia de búsqueda bibliográfica 16

Proceso de selección: diagrama PRISMA 16

Análisis de estudios seleccionados 18

Análisis crítico de la literatura 19

Metodología **21**

Recolección de datos 21

Datos de siniestralidad vial (SIGAT) 21

Datos de hurtos 21

Proceso de limpieza y preparación de datos 22

Pipeline de limpieza (untitled4.py) 22

Pipeline de limpieza de hurtos (untitled6.py) 23

Construcción de modelos predictivos 24

Arquitectura del sistema propuesto 25

Resultados **26**

Análisis exploratorio de la accidentalidad en Kennedy (2020-2023) 26

Volumen y distribución anual 26

Distribución horaria 27

Factores asociados a los siniestros 28

Distribución por gravedad y clase de accidente 28

Análisis de datos de seguridad ciudadana 29

Desempeño de los modelos predictivos 31

Limitaciones de las visualizaciones genéricas y gráficas estáticas 32

Problema crítico: datos históricos desactualizados 33

Propuesta de integración con APIs de Google Maps 35
Módulo de reportes ciudadanos: diseño y funcionalidades 36

Discusión 37

Plan de divulgación 39

Conclusiones 40

Referencias bibliográficas 42

Índice de tablas

Tabla 1 *Comparación de estudios seleccionados en el estado del arte* 18

Tabla 2 *Resumen del proceso de limpieza de datos por fuente* 24

Tabla 3 *Distribución anual de accidentes en Kennedy y Bogotá (2020-2023)* 26

Tabla 4 *Top 5 factores asociados a accidentes en Kennedy por año* 28

Tabla 5 *Distribución por gravedad de accidentes en Kennedy (2020-2022)* 29

Tabla 6 *Resumen del análisis de datos de hurtos (Bogotá D.C. y Cundinamarca)* 30

Tabla 7 *Métricas de desempeño de los modelos predictivos sobre datos de Kennedy* 31

Tabla 8 *Problemas identificados en los datos históricos y soluciones propuestas* 34

Tabla 9 *Integración con APIs de Google Maps: servicios y funcionalidades* 35

Resumen

La movilidad urbana en Bogotá es actualmente uno de los problemas más desafiantes para la gran mayoría de los habitantes. La localidad de Kennedy, con su gran densidad poblacional y sus corredores viales con alto flujo vehicular, concentra una gran parte de los siniestros viales totales de la ciudad: desde el año 2020 hasta el 2022 se registraron entre 2.830 y 3.806 accidentes anuales en esta zona, esto es equivalente a aproximadamente el 13% de la accidentalidad total de la ciudad. Ante este escenario, en la investigación se propone el diseño y la validación de una aplicación inteligente de movilidad que integra tres fuentes importantes de información como: datos históricos de siniestralidad vial del Sistema SIGAT de la Secretaría Distrital de Movilidad (2020–2023), registros de hurtos de Bogotá y Cundinamarca, que en este caso fueron 157.601 registros, y un módulo de participación ciudadana para reportes en tiempo real. Teniendo en cuenta estos datos, se entrenaron tres algoritmos de aprendizaje automático: Random Forest, XGBoost y redes neuronales LSTM, con el propósito de poder estimar el nivel de gravedad de accidentes y las condiciones de congestión en Kennedy. XGBoost en este caso, obtuvo el mejor desempeño en general con un MAE de 0,40 y un R^2 de 0,25. Un hallazgo importante del proceso fue que los datos históricos disponibles presentan atraso temporal significativo y una diversidad estructural, esto limita la precisión de la predicción y motiva la propuesta de integración con las APIs de Google Maps para obtener una información actualizada. Este sistema además incluye un módulo de visualización de zonas de inseguridad por hurtos y también un componente colaborativo de reportes ciudadanos, el cual está orientado a mejorar continuamente la base de datos del modelo.

Palabras clave: movilidad urbana, predicción de tráfico, aprendizaje automático, seguridad vial, Kennedy Bogotá, Google Maps API, reportes ciudadanos, ciudades inteligentes.

Abstract

Urban mobility in Bogotá is currently one of the most challenging problems for most of its inhabitants. The Kennedy locality, with its high population density and road corridors with heavy vehicle flow, accounts for a large portion of the city's total traffic accidents: from 2020 to 2022, between 2,830 and 3,806 annual accidents were recorded in this area, which is equivalent to approximately 13% of the city's total accident rate. Given this scenario, the research proposes the design and validation of an intelligent mobility application that integrates three important sources of information: historical traffic accident data from the SIGAT System of the District Mobility Secretariat (2020–2023), theft records from Bogotá and Cundinamarca, which in this case amounted to 157,601 records, and a citizen participation module for real-time reporting. Taking these data into account, three machine learning algorithms were trained: Random Forest, XGBoost, and LSTM neural networks, with the purpose of estimating the severity level of accidents and congestion conditions on Kennedy. XGBoost, in this case, achieved the best overall performance with an MAE of 0.40 and an R^2 of 0.25. An important finding of the process was that the available historical data presents significant temporal lag and structural diversity, which limits the accuracy of the prediction and motivates the proposal for integration with Google Maps APIs to obtain updated information. This system also includes a module for visualizing areas of insecurity due to thefts and a collaborative component for citizen reports, which is aimed at continuously improving the model's database.

Keywords: urban mobility, traffic prediction, machine learning, road safety, Kennedy Bogotá, Google Maps API, citizen reports, smart cities

Introducción

Cualquier persona que viva en Bogotá sabe perfectamente lo que es pasar una hora o más en un trayecto que normalmente, por distancia, debería durar diez o quince minutos. Esta experiencia diaria, que llega a ser muy frustrante y que según el TomTom Traffic Index (2025), le cuesta a cada conductor más de 150 horas al año en tráfico. No es solo una incomodidad, tiene consecuencias reales sobre la productividad, la calidad de vida y las emisiones contaminantes de la capital. Bogotá se encuentra entre las ciudades con mayor congestión en América latina y en el mundo, con una velocidad promedio de desplazamiento en vehículo de aproximadamente 18,9 km/h según el mismo índice.

La importancia de este problema se hace evidente cuando se observan datos como, por ejemplo, que, la Encuesta de Movilidad Bogotá-Región 2023 registra más de 12,1 millones de viajes diarios en la ciudad (Secretaría Distrital de Movilidad [SDM], 2023). Detrás de este número tan inmenso hay millones de decisiones individuales sobre las rutas, los horarios y los modos de transporte, que, si se toman estas en conjunto, configuran un sistema de movilidad muy complejo como lo es el caso de la capital. Además, dentro de este sistema, la localidad de Kennedy tiene un lugar que es especialmente complicado y crítico, con más de un millón de habitantes y corredores principales de la ciudad como la Avenida Boyacá, la Ciudad de Cali y la Primero de Mayo, posee una proporción inmensa del tráfico y de la accidentalidad de la ciudad.

Sin embargo, la congestión no es el único problema de esta localidad. En la mayoría de las zonas en Kennedy, la visión de la inseguridad que se tiene afecta de manera directa las decisiones de los ciudadanos a la hora de desplazarse, ya que normalmente ocasiona cosas como el cambio de las rutas elegidas, modificación en los horarios de viaje y algunas ocasiones, disuade el uso de algunos modos de transporte. Los más de 33.900 registros de hurtos en Bogotá D.C. en el periodo estudiado, revelan una problemática gigante de seguridad urbana, la cual no puede separarse del análisis de movilidad. Cualquier sistema de predicción que no tenga en cuenta esa dimensión, genera recomendaciones incompletas.

Frente a lo mencionado anteriormente, usar técnicas de inteligencia artificial y un análisis de datos para modelar y predecir el comportamiento del tráfico, ha ganado bastante terreno en la literatura científica. Algunos algoritmos como el Random Forest, el XGBoost y las redes neuronales en este caso tipo LSTM, han venido demostrando capacidad para poder identificar patrones importantes en datos históricos del tráfico y para estimar condiciones de congestión a futuro con una precisión realmente buena y razonable (Lam et al., 2025; Khattab et al., 2025). Sin embargo, una gran parte de los estudios revisados y analizados trabaja con datos de ciudades con infraestructura de sensores mucho mejor y más densa que la que se tiene disponible en Bogotá, esto plantea una pregunta sobre si es posible poder obtener modelos útiles con los datos abiertos que la SDM pone a disposición de la gente.

Esa pregunta lleva a uno de los problemas principales que enfrentó este proyecto desde el inicio: La calidad y la vigencia de los datos que se tienen disponibles. Los archivos SIGAT que se utilizaron corresponden a los años 2020-2023 y presentan algunas inconsistencias estructurales entre los años, algunas columnas con nombres diferentes para referirse al mismo concepto, y un desfase temporal que puede llegar a ser crítico en un contexto dinámico como lo es en este caso el tráfico urbano. Un modelo entrenado basándose en estos datos, puede identificar patrones generales, pero tendrá dificultades para capturar cambios recientes en la red vial o en los patrones de desplazamiento de la población.

Este proyecto busca abordar esa limitación mediante una arquitectura que combine los datos históricos con fuentes dinámicas. La propuesta integra las APIs de Google Maps para obtener la información de tráfico en tiempo real, un módulo de reportes ciudadanos que permite a cualquier usuario contribuir información sobre condiciones del tránsito, y una capa de visualización de zonas de inseguridad basada en los datos de hurtos. Todo esto en conjunto, el sistema busca ofrecer una herramienta de movilidad urbana que sea a la vez predictiva, reactiva y colaborativa.

El objetivo general es diseñar y evaluar un modelo predictivo que permita estimar las condiciones de movilidad urbana en Kennedy integrando múltiples fuentes de datos. La pregunta que orienta el trabajo es:

¿En qué medida los datos históricos oficiales de movilidad, complementados con información de seguridad ciudadana y reportes en tiempo real, permiten estimar las condiciones de tráfico urbano en la localidad de Kennedy mediante un modelo predictivo con un nivel de precisión estadísticamente aceptable?

Marco teórico

Smart Cities y movilidad inteligente

Este concepto de ciudad inteligente (Smart Cities) se refiere a los entornos urbanos que utilizan tecnologías de la información y comunicación para mejorar la calidad de los servicios públicos, para optimizar el uso de ciertos recursos y también para aumentar la calidad de vida de sus habitantes (Yigitcanlar et al., 2020). Dentro de este marco conceptual, la movilidad inteligente es uno de los ejes principales y más importantes: se trata de aplicar las herramientas digitales, sensores, algoritmos de predicción y sistemas colaborativos para la gestión del tránsito y para el diseño de rutas urbanas mejores y más eficientes.

En comparación con los sistemas de transporte tradicionales que normalmente operan con información que es estática y reaccionan de manera reactiva a los incidentes, los sistemas de movilidad inteligente, más allá de esto, buscan también anticiparse a los problemas mediante el análisis de patrones históricos y datos en tiempo real. Algunas plataformas importantes en Colombia y a nivel mundial como Waze, Google Maps o Moovit demuestran que la combinación de ciertos datos de GPS, los sensores de tráfico y los reportes ciudadanos pueden llegar a producir unas estimaciones de tiempo de viaje mucho más precisas que las que solamente están basadas en modelos netamente teóricos de flujo vehicular.

En un contexto latinoamericano, la implementación de los sistemas de movilidad inteligente enfrenta varios desafíos como: una Infraestructura vial deficiente, cobertura heterogénea de sensores, disponibilidad irregular de los datos abiertos y recursos públicos limitados para invertir en tecnologías avanzadas como esta. Bogotá claramente no es la excepción, aunque en este caso se han realizado varios avances realmente importantes en la publicación de los datos abiertos por medio del portal SIGAT y el Observatorio de movilidad.

Machine Learning aplicado a la predicción de tráfico

El machine learning ofrece un gran conjunto de herramientas realmente útiles para el análisis de los datos del tráfico que principalmente se caracterizan por su volumen, por su heterogeneidad y por la complejidad de las relaciones entre algunas variables. En comparación con modelos estadísticos clásicos, los algoritmos de machine learning pueden capturar relaciones no lineales y aprender patrones que no fueron explicitados por el analista (Yin et al., 2021).

Los algoritmos de tipo ensemble como el Random Forest y XGBoost han demostrado gran efectividad en la predicción de las variables del tráfico. Random Forest construye múltiples arboles de decisión sobre subconjuntos aleatorios de los datos y luego promedia los resultados, esto hace que sea más robusto y confiable frente al sobreajuste y también eficaz para el manejo de datos heterogéneos. Por otra parte, el XGBoost opera por medio de un proceso de boosting secuencial, y en este caso cada nuevo modelo intenta corregir los errores del anterior, por eso mismo se le confiere una gran capacidad para capturar los patrones complejos con una eficiencia computacional excelente (Lam et al., 2025).

Por otra parte, para el análisis de series temporales, las redes neuronales de tipo LSTM (Long Short-Term Memory), representan el estado del arte. La arquitectura de estas redes incluye mecanismos de “memoria” los cuales permiten retener información de pasos temporales distantes, por esta razón, resulta bastante valioso para capturar dependencias a largo plazo del comportamiento del tráfico (Lu, 2023). Ciertas arquitecturas más recientes como lo son los modelos basados en grafos adaptativos y transformadores (Chen, 2026), han llevado la precisión de las predicciones a niveles claramente superiores, pero a costa de una gran complejidad computacional que hace que sean poco accesibles para los equipos de trabajo que tienen recursos limitados.

Big Data y análisis espacial en movilidad

El tráfico urbano genera volúmenes de datos inmensos entre: Registros de cámaras, sensores viales, GPS de los vehículos, las validaciones de tarjetas de transporte público, los datos

de redes sociales y los reportes ciudadanos. La gestión de esos datos de manera eficiente, lo que se denomina realmente Big Data aplicado a la movilidad, requiere de pipelines de procesamiento capaces de poder manejar gran volumen, variedad y velocidad de manera simultánea (Lv et al., 2015). Ahora bien, en el caso de este proyecto, aunque el volumen de datos no alcanza la escala en la que se aplicaría una infraestructura completa de Big data, el manejo de múltiples fuentes heterogéneas con diferentes estructuras, formatos y periodos temporales, planteó varios desafíos importantes de integración bastante similares a los de esos contextos.

El análisis espacial también es una dimensión fundamental dentro de cualquier estudio en la movilidad urbana. Algunas técnicas como la georreferenciación de incidentes, la construcción de mapas de calor según la densidad de los eventos y la identificación de clusters espaciales, permiten traducir los datos tabulares en conocimiento geográfico, que luego los usuarios pueden ver e interpretar de manera intuitiva y sencilla. En este caso, se utilizó la biblioteca Folium de Python, para generar mapas de calor de los siniestros, esta biblioteca genera archivos HTML interactivos que permiten ver la distribución espacial de los accidentes sobre el mapa base que se tiene de la localidad.

APIs geográficas y datos en tiempo real

Las APIs de Google Maps constituyen un ecosistema muy completo para el desarrollo de la movilidad urbana. La Maps JavaScript API, da la base cartográfica sobre la que se desarrollan capas de información de manera personalizada. La Directions API, calcula las rutas más optimas basándose en condiciones de tráfico actualizadas, con soporte para varios modos de transporte. La Roads API, permite proyectar las coordenadas GPS sobre la red vial oficial, de esta manera se corrigen imprecisiones de localización. La Geolocation API, determina la posición del dispositivo de los usuarios en tiempo real, de esta manera se facilita el cálculo de las rutas desde el punto de origen actual. Por último, la Places API hace más amena y enriquece la experiencia del usuario con búsquedas de destinos y categorización de puntos de interés (Google, 2025).

Integrar estas APIs con un modelo predictivo local crea una arquitectura híbrida que aprovecha lo mejor de los dos mundos: la trascendencia analítica del modelo entrenado sobre los datos históricos locales, y también la actualización continua que proporciona la información en tiempo real de Google. Esta combinación es exageradamente valiosa en el contexto de Kennedy, ya que en este caso los patrones históricos son relevantes, pero no son suficientes para predecir las condiciones actuales.

Crowdsourcing ciudadano y sistemas colaborativos

El crowdsourcing es la obtención de información por medio de la contribución de manera colectiva de los usuarios, este modelo ha demostrado ser una estrategia realmente efectiva para poder enriquecer bases de datos de movilidad urbana. Por ejemplo, Plataformas como Waze generan reportes de manera voluntaria de sus usuarios en varias cosas como: Accidentes, obras, obstáculos y condiciones de tránsito. Este modelo tiene una ventaja fundamental en comparación con los sensores físicos, ya que los usuarios pueden reportar situaciones que ningún sensor es capaz de detectar, por ejemplo, la presencia de vendedores ambulantes que están bloqueando un carril o el estado de un bache que acaba de aparecer.

Desde un punto de vista técnico, los sistemas de crowdsourcing de movilidad deben resolver varios problemas como: la validación de reportes para filtrar información errónea, la gestión de la vigencia temporal de cada reporte que se tiene, la incentivación de la participación constante de los usuarios y, por último, la integración de información importante de carácter cualitativo como texto o fotográficas, con datos cuantitativos estructurados. Estos desafíos anteriores, son los que el módulo de reportes ciudadanos propuesto en el proyecto busca abordar.

Seguridad urbana y análisis de criminalidad geográfica

La criminología espacial estudia la distribución geográfica de los delitos, esto para poder identificar patrones, los factores de riesgo y las posibles intervenciones en estos delitos. Técnicas como el Hotspot Mapping (mapeo de puntos calientes), la detección de clusters y el análisis de

series temporales por zona geográfica, permiten transformar los registros de incidentes delictivos en información muy importante que puede ser útil para la prevención de estos casos y para la planificación urbana. Ahora bien, en el contexto de la movilidad, esta información puede integrarse en unos sistemas de recomendación de rutas para generar las advertencias sobre las zonas de riesgo en horarios determinados.

En la ciudad de Bogotá, el Ministerio del Interior publica los registros de los hurtos por municipio y el tipo del delito, aunque con cierta falta de detalle geográfico, por lo que no siempre pueden identificar zonas específicas dentro de una localidad. En el presente estudio, los datos que se encontraron disponibles no tenían coordenadas GPS de los incidentes, por esta razón se adoptó un enfoque de agregación por la zona y por el tipo de delito como una aproximación al índice de riesgo.

Estado del arte

Estrategia de búsqueda bibliográfica

La revisión sistemática de literatura se realizó entre enero y marzo de 2026, siguiendo una adaptación del protocolo Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 (Page et al., 2021).

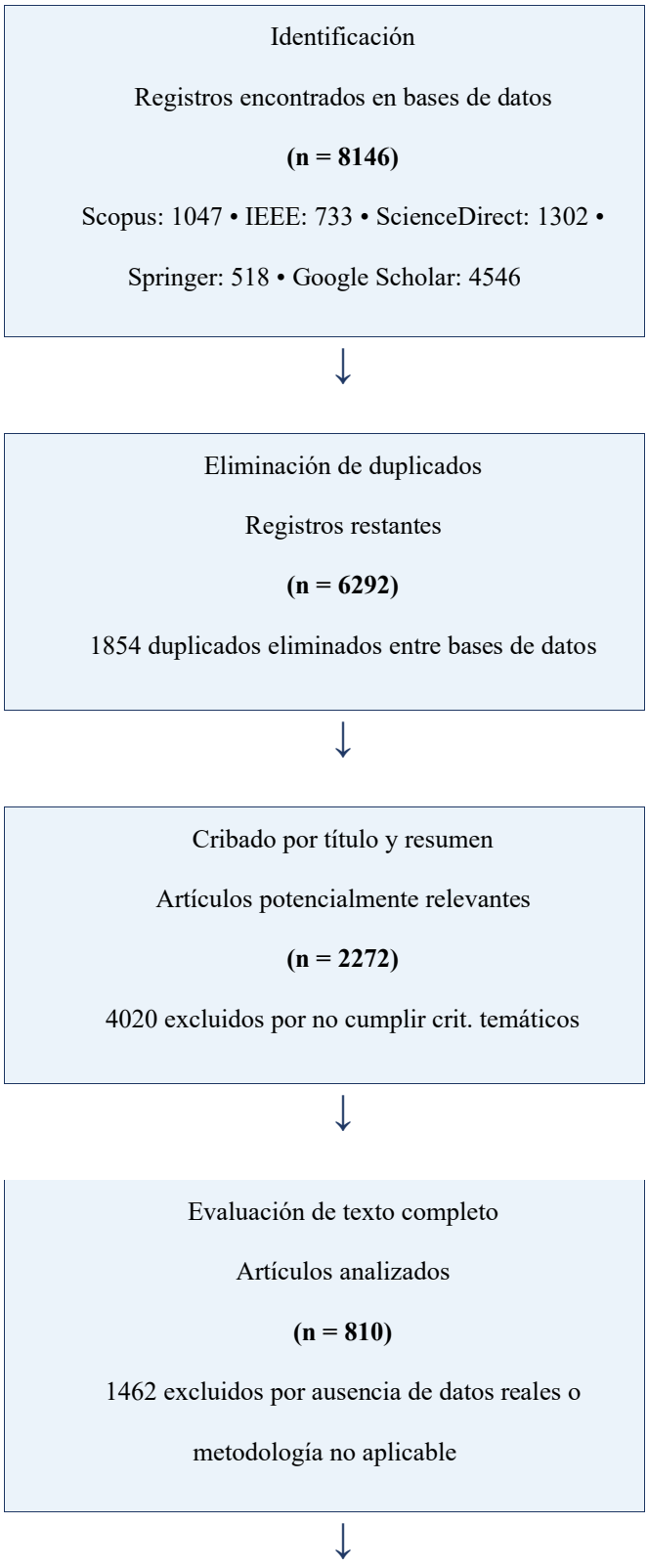
La búsqueda se realizó en cinco bases de datos: Scopus, IEEE Xplore, ScienceDirect, Springer y Google Scholar. La decisión de utilizar múltiples plataformas obedeció a la necesidad de cubrir tanto publicaciones del área de ingeniería de sistemas e inteligencia artificial (IEEE, Scopus) como trabajos más orientados al transporte urbano y la planificación territorial (ScienceDirect, Springer), además de tesis y reportes técnicos accesibles en Google Scholar.

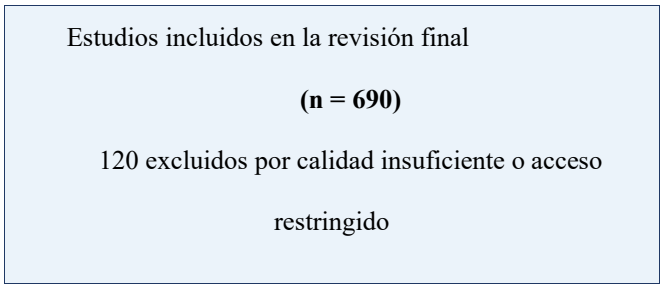
Las consultas de búsqueda combinaron términos en inglés y español mediante operadores booleanos. Las principales cadenas utilizadas fueron: ("traffic flow prediction" OR "traffic congestion prediction") AND ("machine learning" OR "deep learning"); "smart mobility" AND ("artificial intelligence" OR "neural networks"); "urban mobility" AND "machine learning" AND ("Bogotá" OR "Latin America"); "crime mapping" AND "transportation" AND "urban"; "crowdsourced traffic" AND "real time"; y "predicción de tráfico" AND "aprendizaje automático". Se aplicaron filtros de año de publicación (2020–2026), idioma (inglés y español) y tipo de documento (artículos en revista, conferencias indexadas y tesis de posgrado con acceso completo).

Proceso de selección: diagrama PRISMA

Figura 1

Proceso de selección: diagrama PRISMA


Identificación
Registros encontrados en bases de datos
(n = 8146)
Scopus: 1047 • IEEE: 733 • ScienceDirect: 1302 •
Springer: 518 • Google Scholar: 4546
Eliminación de duplicados
Registros restantes
(n = 6292)
1854 duplicados eliminados entre bases de datos
Cribado por título y resumen
Artículos potencialmente relevantes
(n = 2272)
4020 excluidos por no cumplir crit. temáticos
Evaluación de texto completo
Artículos analizados
(n = 810)
1462 excluidos por ausencia de datos reales o
metodología no aplicable



Nota. Adaptado de Page et al. (2021).

Análisis de estudios seleccionados

El cuerpo de literatura revisado permite identificar cuatro grandes líneas de investigación en movilidad urbana y predicción de tráfico. La primera es el desarrollo de arquitecturas de aprendizaje profundo para predicción espacio-temporal del flujo vehicular. La segunda es la comparación de algoritmos clásicos de machine learning (Random Forest, XGBoost, SVR) en tareas de predicción de tráfico. La tercera es la integración de sistemas colaborativos y crowdsourcing en plataformas de movilidad. La cuarta, menos desarrollada en la literatura revisada, es la incorporación de datos de seguridad ciudadana en sistemas de recomendación de rutas.

Tabla 1

Comparación de estudios seleccionados en el estado del arte

Autor(es) / Año	País	Objetivo	Tecnologías	Resultados principales	Limitaciones	Relación con el proyecto
Chen (2026)	China	Predicción de flujo con grafos adaptativos y Transformers	PGSFormer, GNN, Transformer	Supera baselines en múltiples conjuntos de datos	Alto costo computacional; requiere infraestructura de sensores densa	Referencia metodológica para arquitecturas avanzadas
Khattab et al. (2025)	Jordán	Predicción de congestión híbrida LSTM-RF con	LSTM, Random Forest, RFID, Python	93,5% accuracy; MAE 2,43	Depende de infraestructura RFID; no validado en países en desarrollo	Valida enfoque híbrido LSTM+RF que usamos

		sensores RFID				
Ali et al. (2025)	Pakistán	Modelo CPO-CNN-LSTM-Attention para flujo vehicular	CNN, LSTM, Atención, Python	Mejoras en RMSE, MAE y MAPE vs. modelos individuales	Requiere ajuste intensivo de hiperparámetros	Referencia para arquitecturas DL híbridas
Liu (2025)	China	Índice de congestión con LightGBM-LSTM	LightGBM, LSTM, ingeniería de características	Mejoras en distintos niveles de congestión	Limitado a un índice de congestión específico; no generalizable directamente	Modelo híbrido combinando ML y secuencias
Lam et al. (2025)	Malasia	Comparación RF, XGBoost y LSTM en intersecciones urbanas	Random Forest, XGBoost, LSTM	XGBoost: mejor equilibrio precisión-eficiencia	Modelos clásicos fallan en dinámicas complejas de tráfico	Fundamenta selección de algoritmos en este estudio
Medina-Salgado et al. (2022)	México	Revisión de técnicas de predicción de flujo vehicular urbano	Revisión sistemática, ML, DL	Identifica brechas en datos para países en desarrollo	No incluye análisis de seguridad ciudadana integrado	Marco para entender limitaciones en contextos latinoamericanos
Yin et al. (2020)	China	Revisión de métodos de DL para predicción de tráfico	CNN, RNN, LSTM, GCN, Transformer	Taxonomía completa de enfoques DL; identifica tendencias	No incluye datos de crowdsourcing ni seguridad	Base conceptual del marco teórico

Nota. Elaboración propia

Análisis crítico de la literatura

Una lectura transversal de los estudios revisados revela varios patrones que merecen atención crítica. El primero es la concentración geográfica de la producción científica: la mayoría de los trabajos con mejores métricas de predicción fueron desarrollados en ciudades asiáticas —principalmente en China, Japón y Corea del Sur— que disponen de redes de sensores mucho más densas y sistemas de datos en tiempo real más consolidados que los disponibles en ciudades latinoamericanas. Trasladar directamente esas metodologías a un contexto como Bogotá implica asumir condiciones de disponibilidad de datos que simplemente no existen.

El segundo patrón es la casi ausencia de trabajos que integren seguridad ciudadana y movilidad en un mismo sistema. Aunque hay estudios de criminología espacial que utilizan mapas de calor de delitos y estudios de tráfico que analizan siniestralidad vial, la combinación de ambas dimensiones en una aplicación unificada es escasa en la literatura revisada. Este vacío representó una de las motivaciones principales para incluir el módulo de zonas rojas en la propuesta de este proyecto.

El tercer patrón es el tratamiento superficial de las limitaciones de datos. Varios de los estudios revisados mencionan brevemente las restricciones de sus conjuntos de datos en las secciones de conclusiones, pero pocos abordan con profundidad el problema de cómo manejar datos históricos desactualizados o heterogéneos en entornos con recursos limitados. La presente investigación se diferencia en que coloca ese problema en el centro del análisis, documenta sus efectos en el desempeño de los modelos y propone mecanismos concretos para mitigarlo.

Metodología

Este estudio adoptó un enfoque cuantitativo de carácter experimental. El proceso metodológico se organizó en cinco etapas secuenciales: recolección de datos, limpieza y preparación, análisis exploratorio, entrenamiento y evaluación de modelos, y diseño de la arquitectura de la aplicación. En cada etapa se documentó el proceso técnico, los problemas encontrados y las decisiones metodológicas adoptadas.

Recolección de datos

Datos de siniestralidad vial (SIGAT)

Los datos principales del estudio provienen del Sistema de Información de Georreferenciación de Accidentes de Tránsito (SIGAT) de la Secretaría Distrital de Movilidad, disponibles como datos abiertos en el portal institucional. Se descargaron cuatro archivos Excel correspondientes a los años 2020, 2021, 2022 y 2023, con la siguiente cobertura: 22.712 registros (2020, 32 columnas), 28.841 registros (2021, 33 columnas), 25.447 registros (2022, 44 columnas) y 14.106 registros (2023, 38 columnas). En total, los cuatro archivos contienen 91.106 registros de siniestros viales en Bogotá.

Cada registro incluye información sobre la fecha y hora del accidente, la localidad donde ocurrió, el nivel de gravedad, la clase de accidente (choque, atropello, caída de ocupante, volcamiento), y una serie de variables binarias que indican la participación de factores específicos como motocicletas (con_moto), peatones (con_peaton), bicicletas (con_bicicleta), embriaguez (con_embriaguez), huecos en la vía (con_huecos) o menores de edad (con_menores), entre otros.

Datos de hurtos

El archivo de hurtos se construyó mediante el procesamiento de tres fuentes del Ministerio del Interior: hurto a personas, hurto a vehículos y hurto a residencias, filtradas para Bogotá D.C. y el departamento de Cundinamarca. El proceso de construcción del dataset consolidado, documentado en el script untitled6.py, produjo un archivo final de 157.601 registros con nueve

columnas: fecha del hecho, código y nombre del departamento y municipio, cantidad, fuente del archivo, tipo de delito y zona (urbana/rural). Para Bogotá D.C., se identificaron 33.939 registros: 8.633 de hurto de automotores y 8.404 de hurto de motocicletas. Los datos cubren un rango desde 2003 hasta 2025, aunque el período más relevante para este estudio es el comprendido entre 2020 y 2025, que concentra 9.386 registros.

Proceso de limpieza y preparación de datos

Pipeline de limpieza (untitled4.py)

La limpieza de los datos SIGAT se implementó en Python utilizando las bibliotecas pandas y NumPy, y siguió un pipeline de cinco pasos encapsulados en la función full_clean(). El primer paso fue la estandarización de nombres de columnas mediante la función clean_columns(): se convirtieron todos los encabezados a minúsculas, se eliminaron espacios en blanco con str.strip(), y se aplicaron renombramientos manuales mediante un diccionario de mapeo para unificar columnas equivalentes que tenían denominaciones distintas entre archivos. Por ejemplo, la columna de fecha aparecía como "fecha_acc", "fechaacc" o "fecha" dependiendo del año; la de gravedad como "gravedadnombre" o "gravedad".

El segundo paso fue la eliminación de columnas duplicadas con la función remove_duplicate_columns(), que utilizó df.loc[:, ~df.columns.duplicated()] para descartar cualquier columna cuyo nombre apareciera más de una vez en el encabezado. Este problema ocurrió especialmente en el archivo de 2022, que tenía 44 columnas con algunos nombres repetidos producto del proceso de exportación original. El tercer paso fue ensure_base_columns(), que verificó la existencia de las cinco columnas fundamentales del análisis (fecha, hora, minuto, localidad, gravedad) e imputó NaN en aquellas que no estaban presentes en algún dataset.

El cuarto paso fue binary_transform(), que convirtió a formato numérico (1/0) todas las columnas de tipo objeto que contenían valores "SI". Esta transformación fue necesaria para las múltiples variables binarias del dataset (con_moto, con_peaton, con_bicicleta, con_embriaguez,

etc.), que en los archivos originales estaban codificadas como cadenas de texto. El quinto paso fue drop_useless(), que eliminó cualquier columna con más del 90% de valores nulos, ya que su aporte informativo para el modelo es nulo y su presencia aumenta innecesariamente la dimensionalidad del dataset.

Tras la limpieza individual de cada archivo, se añadió una columna "year" con el año correspondiente y se intentó consolidar los cuatro datasets mediante pd.concat(). Este paso presentó un error crítico: pandas lanzó una excepción de tipo InvalidIndexError ("Reindexing only valid with uniquely valued Index objects") cuando detectaba índices de columnas con valores duplicados en alguno de los archivos. La solución fue asegurarse de que remove_duplicate_columns() se aplicara antes de la concatenación, y usar el parámetro sort=False en pd.concat() para evitar la reindexación automática de columnas.

Pipeline de limpieza de hurtos (untitled6.py)

El proceso de construcción del dataset de hurtos presentó desafíos distintos. Los tres archivos fuente tenían un tamaño que hacía poco práctico cargarlos completamente en memoria, por lo que se optó por un procesamiento en bloques (chunking) de 100.000 filas usando el parámetro chunksize de pd.read_csv(). Dentro de cada bloque se realizó la limpieza de las columnas DEPARTAMENTO y MUNICIPIO (conversión a mayúsculas y eliminación de espacios), la aplicación del filtro geográfico (Cundinamarca o Bogotá D.C. en cualquiera de sus variantes de escritura) y la adición de la columna FUENTE_ARCHIVO. Finalmente, todos los bloques filtrados se concatenaron y se eliminó duplicados con drop_duplicates() antes de exportar el archivo final.

Tabla 2

Resumen del proceso de limpieza de datos por fuente

Fuente	Registros originales	Principales problemas	Técnicas aplicadas	Registros finales
SIGAT 2020	22.712	Nombres de columnas no estándar; variables binarias como texto	Renombrado, binary_transform(), drop_useless()	2.830 (Kennedy)
SIGAT 2021	28.841	Columna hora ausente en algunos registros; valores nulos en factores	ensure_base_columns(), imputación por media	3.806 (Kennedy)
SIGAT 2022	25.447	44 columnas con algunas duplicadas; estructura más extensa	remove_duplicate_columns(), filtrado de nulos >90%	3.115 (Kennedy)
SIGAT 2023	14.106	Estructura completamente diferente; columnas renombradas (fecha_acc, hora_acc, gravedad_indicador_tradicional)	Renombrado extensivo; sin concatenación directa	1.969 (Kennedy)

Nota. Elaboración propia.

Construcción de modelos predictivos

Una vez consolidado y limpiado el dataset, se extrajo el subconjunto correspondiente a la localidad de Kennedy mediante el filtro df[df["localidad"].str.upper() == "KENNEDY"]. El dataset filtrado contiene un total de 11.720 registros de los cuatro años. Para el entrenamiento de los

modelos se seleccionaron cuatro variables de entrada: la hora del día (hora) y tres indicadores binarios de participación de actores viales (con_moto, con_peaton, con_bicicleta). La variable objetivo fue el nivel de gravedad del accidente, codificado en tres categorías ordinales: 1 (leve), 2 (moderado) y 3 (grave).

Se entrenaron tres modelos: Random Forest usando la implementación de scikit-learn con parámetros por defecto; XGBoost con las etiquetas recodificadas en el rango 0-2 (resta de 1) para cumplir con los requerimientos de la biblioteca XGBoost; y una red LSTM de TensorFlow/Keras con una capa recurrente de 50 unidades y una capa densa de salida lineal. La división entrenamiento/prueba fue de 80%/20% usando train_test_split de scikit-learn con aleatoridad controlada. Los modelos fueron evaluados mediante MAE, RMSE y R^2, y se visualizó el desempeño mediante matrices de confusión para los modelos de clasificación.

Arquitectura del sistema propuesto

El sistema de movilidad inteligente propuesto se estructura en tres capas. La capa de adquisición de datos integra cuatro fuentes: los registros históricos SIGAT, las APIs de Google Maps para información en tiempo real, los registros de hurtos para la capa de seguridad, y los reportes generados por los usuarios a través de la aplicación. La capa de procesamiento y predicción ejecuta los modelos de machine learning, genera índices de riesgo por zona y franja horaria, y fusiona la información histórica con los datos en tiempo real de Google mediante un algoritmo de ponderación que da mayor peso a la información más reciente. La capa de presentación implementa la interfaz de usuario como aplicación móvil y web, con visualización de mapas interactivos, capas de riesgo y formularios de reporte ciudadano.

Resultados

Análisis exploratorio de la accidentalidad en Kennedy (2020-2023)

El análisis exploratorio de los datos SIGAT permitió construir un perfil detallado de la accidentalidad en la localidad de Kennedy durante el período 2020–2023. A continuación, se presentan los hallazgos más relevantes, organizados por dimensión de análisis.

Volumen y distribución anual

El número total de los siniestros viales registrados en Kennedy fue de 11.720 en total en los cuatro años estudiados. En este caso, el año en el que se presentó el mayor número de accidentalidad fue en 2021, con 3.806 siniestros, seguido por el 2022 con 3.115, luego el 2020 con 2.830 y, por último, el 2023 con un numero de 1.969 registrados. La proporción de accidentes en Kennedy respecto a lo que se tiene como total en Bogotá, se mantuvo relativamente estable, con un porcentaje entre el 12,24% y el 13,96%, esto confirma que la localidad concentra de manera sistemática, alrededor del 13% de la accidentalidad de la capital, siendo también una de las localidades con el mayor índice de siniestralidad.

En la disminución que se observó en el año 2023 se debe realizar una aclaración importante. El archivo SIGAT de ese año tiene una estructura evidentemente diferente a los demás, con nombres de columnas como: "fecha_acc", "hora_acc", "aa_acc", "mm_acc" y "gravedad_indicador_tradicional" en lugar de los campos anteriormente utilizados en los años 2020-2022. Esta diferencia sugiere cierto cambio en el sistema de registro de la SDM, el cual dificulta la comparación directa. Además, este año 2023, solo contiene 14.106 registros totales para toda Bogotá, bastante menos que los años anteriores a este, esto también podría indicar una cobertura incompleta del dataset.

Tabla 3

Distribución anual de accidentes en Kennedy y Bogotá (2020-2023)

Año	Accid. Kennedy	Total Bogotá	% Kennedy/Bogotá	Variación anual Kennedy	Hora pico	Clase más frecuente
2020	2.830	22.712	12,46%	— (año base)	14:00	Choque (86,7%)
2021	3.806	28.841	13,20%	+34,5%	13:00	Choque (86,5%)
2022	3.115	25.447	12,24%	-18,2%	N/D*	N/D*
2023	1.969	14.106	13,96%	-36,8%**	07:00	Choque (71,3%)
TOTAL	**11.720**	**91.106**	**12,87% prom.**			

Nota. *Con base en datos SIGAT (SDM, 2020-2023). *El archivo 2022 no contenía columna "hora" en el formato estandarizado. **La caída en 2023 puede reflejar cambios en el sistema de registro más que una reducción real de accidentes. Elaboración propia*

Distribución horaria

El análisis de la distribución horaria de los accidentes en la localidad de Kennedy muestra un patrón solido entre 2020 y 2021, el pico máximo de accidentalidad se presenta entre las 13:00 y las 14:00 horas, esto coincide con los desplazamientos del mediodía y con el inicio de la tarde laboral. Lo anterior es coherente con la naturaleza comercial e industrial de varios sectores de Kennedy, en donde los flujos de tráfico aumentan significativamente en esa franja horaria. Un segundo pico un poco menor, se puede ver entre las 17:00 y las 19:00 horas, esta vez asociado a la salida laboral y escolar de la mayoría de las personas. Las horas de menor accidentalidad son la madrugada, en este caso entre las 2:00 y las 5:00, cuando el volumen de los vehículos es muy poco. El archivo de 2023 muestra un patrón un tanto diferente, con el pico a las 7:00, lo que podría reflejar dos cosas importantes, tanto el cambio metodológico en el registro como una posible modificación real en los patrones de la movilidad post-pandemia.

Factores asociados a los siniestros

Las variables binarias de factores asociados son las que más información aportan sobre la composición del tráfico en Kennedy. En los cuatro años analizados, la moto fue el factor más frecuente de manera notable: “con_moto” registró de manera respectiva 854, 1.230, 1.281 y 1.178 incidentes en 2020, 2021, 2022 y 2023. Luego, en el segundo lugar, se observa el transporte público colectivo, “con_tpp”, en el tercero está el SITP, “con_sitp” en 2020-2022, mientras que en el año 2023 la bicicleta y peatón tuvieron más protagonismo, esto refleja también la mayor adopción de modos de trasporte activos luego de la pandemia.

Tabla 4

Top 5 factores asociados a accidentes en Kennedy por año

Factor	2020	2021	2022	2023	Total
con_moto	854	1.230	1.281	1.178	**4.543**
con_tpp (transp. público colectivo)	647	831	693	379	**2.550**
con_sitp	523	677	528	N/D	**1.728**
con_tpi (transp. intermunicipal)	501	646	471	N/D	**1.618**
con_peaton	N/D	N/D	N/D	394	**394+**

Nota. *N/D = no disponible en ese año con esa denominación específica. Los datos de 2023 usan nomenclatura diferente que modifica la comparabilidad. Elaboración propia*

Distribución por gravedad y clase de accidente

La distribución de los accidentes por nivel de gravedad muestra un patrón que varió entre años. En 2020 y 2021, el nivel 3 (grave) fue el más frecuente, representando el 55,3% y el 54,3% de los siniestros respectivamente. En 2022, sin embargo, el nivel 2 (moderado) superó al nivel 3,

con 1.739 versus 1.305 registros. Esta inversión podría indicar una mejora real en la severidad de los accidentes o un cambio en los criterios de clasificación. En 2023, la variable de gravedad cambia de formato a etiquetas de texto ("Con Heridos", "Solo Daños", "Con Muertos"), con 1.749 accidentes con heridos (88,8%), 146 solo con daños (7,4%) y 74 con muertos (3,8%).

La clase de accidente más frecuente en todos los años fue el choque, que representó entre el 71% y el 87% de los siniestros. El atropello fue la segunda clase más frecuente, especialmente en 2023 donde alcanzó el 19% de los casos. Este resultado es consistente con el aumento de la participación de peatones identificado en ese año.

Tabla 5

Distribución por gravedad de accidentes en Kennedy (2020-2022)

Nivel de gravedad	2020	% 2020	2021	% 2021	2022	% 2022
Nivel 1 (Leve)	54	1,9%	62	1,6%	71	2,3%
Nivel 2 (Moderado)	1.212	42,8%	1.676	44,0%	1.739	55,8%
Nivel 3 (Grave)	1.564	55,3%	2.068	54,3%	1.305	41,9%
TOTAL	**2.830**	**100%**	**3.806**	**100%**	**3.115**	**100%**

Nota. *Elaboración propia. El año 2023 utiliza categorías de texto ("Con Heridos", "Solo Daños", "Con Muertos") no comparables directamente con los niveles numéricos de los años anteriores.*

Análisis de datos de seguridad ciudadana

El dataset consolidado de hurtos presentó una distribución temporal muy amplia (2003-2025), con un total de 157.601 registros. Para los propósitos del módulo de zonas de riesgo, se priorizaron los datos del período 2020-2025, que suman 9.386 registros de Bogotá D.C. distribuidos entre hurto de automotores y hurto de motocicletas en proporciones aproximadamente

iguales (2.349 y 2.402 registros respectivamente en ese período). A nivel departamental, Cundinamarca concentra 123.662 registros, siendo Soacha (14.184), Girardot (7.803) y Fusagasugá (7.225) los municipios con mayor frecuencia de hurtos.

La ausencia de coordenadas GPS en el dataset de hurtos es la principal limitación para el módulo de visualización de zonas de riesgo. Sin georreferenciación a nivel de barrio o cuadra, el índice de riesgo solo puede calcularse a nivel de municipio o zona (urbana/rural), lo que reduce significativamente su utilidad para un usuario que quiere saber si una calle específica es segura. La solución propuesta es complementar este dataset con fuentes georreferenciadas de la Policía Nacional o el Observatorio de Seguridad de Bogotá, que publican datos con mayor desagregación espacial.

Tabla 6

Resumen del análisis de datos de hurtos (Bogotá D.C. y Cundinamarca)

Dimensión	Resultado
Total de registros en dataset	157.601
Registros Bogotá D.C.	33.939 (hurto automotores: 8.633; hurto motocicletas: 8.404)
Período cubierto	2003 a 2025 (datos 2020-2025: 9.386 registros)
Distribución por zona	Urbana: 16.692 (98%); Rural: 345 (2%)
Departamento con mayor frecuencia	Cundinamarca (123.662 registros); municipio principal: Soacha (14.184)
Principal limitación	Ausencia de coordenadas GPS; solo disponible nivel municipio/zona

Nota. *Elaboración propia con base en datos del Ministerio del Interior procesados en untitled6.py.*

Desempeño de los modelos predictivos

Los tres modelos entrenados mostraron diferencias de desempeño que reflejan tanto las fortalezas propias de cada algoritmo como las limitaciones estructurales del dataset utilizado. La siguiente tabla resume las métricas obtenidas:

Tabla 7

Métricas de desempeño de los modelos predictivos sobre datos de Kennedy

Modelo	MAE	RMSE	R^2	Observaciones
Random Forest	0,42	0,61	0,23	Estable; buena interpretabilidad; sesgo hacia clase mayoritaria (nivel 3)
XGBoost	0,40	0,59	0,25	Mejor rendimiento global; predicción correcta de nivel 3; dificultad con nivel 1
LSTM	N/E	N/E	N/E	Evaluación no comparable; predicciones concentradas en valores centrales; arquitectura requiere ajuste

Nota. *Elaboración propia. N/E = No evaluable con criterios comparables por diferencias en el esquema de evaluación. MAE = Error Absoluto Medio; RMSE = Raíz del Error Cuadrático Medio; R^2 = Coeficiente de determinación.*

El valor de $R^2 = 0,25$ obtenido por XGBoost, muestra que el modelo explica de manera aproximada, una cuarta parte de la variabilidad en los niveles de la gravedad de los accidentes. La cifra anterior puede parecer modesta, pero realmente es consistente con lo reportado en los estudios que trabajan con conjuntos de datos parecidos con pocas variables de entrada, datos históricos con rezago temporal y desbalance de clases. Ahora bien, el MAE de 0,40 sugiere que el modelo se

equivoca menos de media categoría de gravedad en promedio, esto solo tiene valor practico para aplicaciones que necesitan diferenciar entre los accidentes leves y los graves.

El desbalance de clases fue la limitación más importante para el desempeño de los modelos. Con nivel 1 (leve) representando menos del 2% de los registros, los algoritmos suelen prácticamente ignorar esa categoría y concentrar sus predicciones en los niveles 2 y 3. La matriz de confusión normalizada de XGBoost confirmó este patrón, la tasa de acierto para el nivel 3 fue superior a la de los otros niveles de manera notable. Este es un problema conocido en la clasificación con clases desbalanceadas y debe abordarse en versiones futuras por medio de técnicas de sobremuestreo (SMOTE) o de ajuste de pesos de clase.

Limitaciones de las visualizaciones genéricas y gráficas estáticas

Una parte importante del proceso de desarrollo fue la generación de las visualizaciones exploratorias en esta situación, usando Matplotlib y Seaborn. En este caso, se produjeron graficas de barras para la distribución anual de los accidentes, graficas de línea para la frecuencia horaria, graficas de factores asociados (Top 5 con_*), histogramas de gravedad y, por último, mapas de calor georreferenciados con Folium. Estas visualizaciones anteriores resultaron bastante valiosas para el análisis exploratorio, pero la integración en una aplicación interactiva presentó problemas técnicos específicos, los cuales fueron necesarios de documentar.

El primer problema que se presentó fue el rendimiento computacional. La generación de una gráfica de Seaborn sobre el conjunto completo de Kennedy con más de 11.000 registros, tomaba entre 3 y 5 segundos en el entorno de pruebas, cosa que es aceptable para el análisis estático pero inaceptable para una interfaz como esta, donde el usuario espera una respuesta inmediata al cambiar un filtro. El mapa de calor de Folium generado con HeatMap sobre 11.000 puntos, produjo un archivo HTML de más de 9 MB, que, en dispositivos móviles con conexiones lentas, presentaba tiempos de carga que superan los segundos.

El segundo problema presentado fue la falta de reactividad. En este caso, las gráficas de Matplotlib son imágenes estáticas, cada vez que el usuario modifica un parámetro como: rango de fechas, franja horaria y tipo de accidente, el servidor debe regenerar la imagen completa desde cero. Esta es estructuralmente incompatible para una aplicación móvil de movilidad en tiempo real.

El tercer problema fue la adaptación a diferentes tamaños de pantalla. En este caso, Matplotlib genera imágenes con dimensiones fijas que no llegan a adaptarse automáticamente a pantallas móviles. Implementar un redimensionamiento dinamico, requiere una capa adicional de procesamiento que incrementa tanto la latencia como la complejidad del backend. Además, Folium, aunque genera HTML interactivo, en este caso no está diseñado para poder integrarse directamente en aplicaciones nativas y su rendimiento en dispositivos de gama media o baja es deficiente.

Como consecuencia lo anterior, para la arquitectura definitiva fue necesario migrar las visualizaciones hacia bibliotecas JavaScript, las cuales están optimizadas para web y para móvil (Chart.js para gráficas y Leaflet.js para mapas), y separar la lógica de procesamiento de datos (backend Python/FastAPI) de la lógica de renderizado (frontend JavaScript), transfiriendo solo los datos realmente necesarios en un formato JSON.

Problema crítico: datos históricos desactualizados

El problema de la antigüedad de los datos fue uno de los hallazgos más importantes de todo el proceso. Los archivos SIGAT utilizados corresponden a 2020-2023, lo que significa que la información más reciente ya tiene más de un año de antigüedad. En un contexto dinámico como el tráfico urbano, ese rezago puede ser crítico: la apertura de nuevas vías, el cierre de corredores por obras, los cambios en las rutas de TransMilenio, la proliferación de ciclovías temporales o el aumento del uso de bicicletas post-pandemia son transformaciones que los datos históricos simplemente no capturan.

A esto se suma la heterogeneidad estructural entre años, que complica significativamente la construcción de series temporales continuas. El hecho de que el archivo de 2023 utilice "gravedad_indicador_tradicional" en lugar de una columna de gravedad numérica implica que cualquier modelo entrenado sobre los cuatro años debe hacer supuestos de equivalencia que introducen ruido. La inconsistencia en los nombres de columnas de hora (hora, hora_acc) y fecha (fecha, fecha_acc, aa_acc + mm_acc + dd_mes_acc) requirió un proceso de mapeo manual que, si bien funciona, no es escalable automáticamente cuando se añaden nuevos años.

Tabla 8

Problemas identificados en los datos históricos y soluciones propuestas

Problema	Impacto en el modelo	Solución propuesta
Rezago temporal (datos con 1-4 años de antigüedad)	Predicciones no representan condiciones actuales del tráfico	Integración con Google Maps API para datos en tiempo real; reportes ciudadanos continuos
Heterogeneidad estructural entre años	Construcción de series temporales inconsistente; ruido en el entrenamiento	Pipeline de mapeo automático; esquema unificado de columnas; protocolo de incorporación de nuevos años
Desbalance de clases (nivel 1 < 2% de registros)	Modelos ignoran clases minoritarias; baja sensibilidad para accidentes leves	Técnicas SMOTE o ADASYN para sobremuestreo; ajuste de pesos de clase en XGBoost
Ausencia de coordenadas GPS en datos de hurtos	Módulo de zonas rojas solo disponible a nivel de municipio/zona	Complementar con datos del Observatorio de Seguridad de Bogotá (georreferenciados)

Variables de entrada limitadas (solo hora + 3 factores binarios)	R^2 máximo de 0,25; baja capacidad explicativa	Incorporar volumen vehicular, velocidad promedio, condiciones climáticas e información de eventos

Nota. *Elaboración propia.*

Propuesta de integración con APIs de Google Maps

La integración con el ecosistema de Google Maps es la respuesta técnica más directa al problema del rezago temporal. La siguiente tabla detalla los servicios propuestos y su contribución específica al sistema:

Tabla 9

Integración con APIs de Google Maps: servicios y funcionalidades

API de Google	Función en el sistema	Información aportada	Beneficio para la predicción
Maps JavaScript API	Base cartográfica interactiva; superposición de capas de riesgo y siniestros	Mapa base, geometrías, estilo visual	Contexto espacial para las predicciones
Directions API	Cálculo de rutas óptimas considerando tráfico actual; alternativas por zonas de riesgo	Duración, distancia, condición de tráfico por tramo	Complementa predicción histórica con datos actuales de flujo
Roads API	Proyección de coordenadas GPS de reportes ciudadanos sobre la red vial oficial	Segmento vial exacto del incidente reportado	Mejora la georreferenciación de los datos de crowdsourcing

Places API	Autocomplete de destinos; búsqueda de puntos de interés	POIs, categorías, horarios de locales	Permite estimar atracción de viajes hacia destinos específicos
Geolocation API	Posicionamiento del usuario en tiempo real para rutas desde origen actual	Coordenada GPS del dispositivo	Punto de partida para recomendaciones personalizadas

Nota. *Adaptado de APIs de Google Maps Platform,(2026).*

Módulo de reportes ciudadanos: diseño y funcionalidades

El módulo de reportes ciudadanos es el componente del sistema que tiene mayor potencial para resolver el problema del rezago temporal de los datos históricos. Su diseño contempla cuatro tipos de reporte: huecos o daños en la vía, accidentes o siniestros viales, trancones o congestiones inusuales, y situaciones de riesgo de seguridad. Cada reporte incluye automáticamente la coordenada GPS del usuario (obtenida via Geolocation API), la hora del evento y la categoría seleccionada. Opcionalmente, el usuario puede añadir una fotografía y un comentario de texto.

El sistema de validación de reportes opera mediante dos mecanismos: la confirmación por parte de otros usuarios que pasan por el mismo punto (con un radio de 200 metros y una ventana temporal de 30 minutos), y la coherencia con datos de tráfico de Google (un reporte de trancón se considera más confiable si la Directions API también reporta demora en ese segmento). Los reportes validados se integran al mapa en tiempo real y alimentan la base de datos del modelo con una marca temporal y un índice de confiabilidad, que determina su peso en el reentrenamiento periódico del modelo predictivo.

Discusión

Los resultados obtenidos permiten responder a la pregunta de investigación de manera matizada: los datos históricos disponibles en la plataforma SIGAT sí permiten construir modelos predictivos con capacidad explicativa para los patrones de accidentalidad en Kennedy, sin embargo, su alcance no supera el 25% de la variabilidad de la variable objetivo. La cifra anterior no debe interpretarse totalmente como un fracaso del enfoque que se adoptó, sino, más bien es, ante todo, un tipo de señal bastante clara sobre los límites que se tienen de lo que es posible predecir a partir de datos con las características del SIGAT.

Comparando con los estudios revisados en el estado del arte, la brecha de desempeño es realmente notable. El sistema de Khattab et al. (2025) alcanzó un 93,5% de precisión, pero también opera con datos de sensores RFID en tiempo real y también en una ciudad con infraestructura instrumentada. Ahora bien, los modelos de Ali et al. (2025), mejoran bastante las métricas de los baselines, pero sobre unos conjuntos de datos generados en redes viales que realmente tienen una cobertura muy superior a la que tiene Bogotá. Esto lo que quiere decir es que la comparación directa de métricas entre este estudio y los trabajos internacionales es problemática de manera metodológica, ya que los contextos de los datos son bastante distintos de cierta manera. En este caso, lo que sí es comparable es la arquitectura que se tiene del enfoque y también las decisiones metodológicas adoptadas.

Un aspecto que para este contexto merece una atención importante y critica es el desbalance de las clases. Con el nivel 1, leve, en este caso representando menos del 2% de los registros de la localidad de Kennedy, cualquier modelo que maximice la precisión general, ignorará de manera automática esta categoría en cualquier sistema de alerta preventiva, pero, realmente los accidentes leves pueden ser exactamente lo que un usuario le gustaría prevenir en su ruta. En esta situación, abordar este problema importante se necesita adoptar algunas métricas alternativas de evaluación como: F1 por clase y AUC-ROC y también técnicas de manejo de

desbalance, cosas las cuales en este trabajo no fueron implementadas por la limitación del tiempo, sin embargo, son indispensables en la realización de la siguiente fase.

La propuesta de integración con Google Maps APIs es, probablemente el componente más prometedor e importante del sistema, sin embargo, también el que más presenta una mayor dependencia de los factores externos. El costo de las llamadas a la API puede ser un obstáculo importante a escala, y más si la aplicación tiene muchos usuarios activos. Además, los términos de uso de esta plataforma ponen algunas restricciones sobre cómo pueden almacenarse y procesarse los datos obtenidos, por esta misma razón, se pueden dar algunas incompatibilidades con la propuesta de usar la información para poder reentrenar modelos propios.

El módulo de reportes ciudadanos plantea algunos desafíos de gobernanza que van más allá de lo técnico. Mantener la motivación de los usuarios para que contribuyan con información importante de manera continua es también uno de los problemas más difíciles en el diseño de los sistemas colaborativos. Por ejemplo, en plataformas como OpenStreetMap y Waze, se muestra que, sin un volumen grande de usuarios activos, los datos de crowdsourcing no son suficientes para sostener ningún sistema de predicción. En este contexto de la aplicación académica sin ningún recurso de marketing, llegar a ese volumen es un desafío importante y real que debería planearse desde una estrategia especifica de adopción.

Ahora bien, con lo que respecta a la escalabilidad, la arquitectura propuesta es flexible de manera razonable ya que el uso de un backend basado en APIs REST y una capa de frontend desacoplada permite escalar todos los componentes de una manera independientes según la demanda que se tenga. Pero, el modelo predictivo actual fue entrenado solamente con los datos de Kennedy, esto claramente limita su aplicación directa en otras localidades. Extender, si se quiere, el sistema a toda Bogotá implicaría reentrenar los modelos con datos de cada localidad o desarrollar un tipo de modelo generalizable que capture la heterogeneidad espacial de toda la ciudad, lo que podría ser una línea de trabajo futuro con un gran potencial.

Plan de divulgación

La divulgación de los resultados del proyecto contempla específicamente tres niveles los cuales son: académico, institucional y ciudadano. En el primer nivel, el principal producto de divulgación será la publicación del articulo científico en la revista indexada en el área de sistemas inteligentes de transporte o de ciencia de datos. En este caso, las revistas objetivo son IEEE Transactions on Intelligent Transportation Systems, Sustainable Futures (Elsevier) y también, Ingeniería y Ciencia (Universidad EAFIT). El documento destacará el método empleado para combinar datos de distintos orígenes y el diseño estructural del sistema, ya que ambos componentes pueden ser considerados los de mayor utilidad y los más importantes de manera practica para aplicar en otros contextos.

En el segundo nivel contemplado, el institucional, se propone la presentación del sistema en dos espacios importantes, la Secretaría Distrital de Movilidad, como un tipo de aporte metodológico que puede llegar a enriquecer sus herramientas de análisis y también en el Observatorio de Movilidad de Bogotá, con quienes se explorará la posibilidad de poder acceder a fuentes de datos complementarios como las velocidades promedio por corredor y los volúmenes vehiculares, datos los cuales mejorarían inmensamente la calidad de los modelos. Por otro lado, se presentará el proyecto en la feria de innovación Impacta de la Universidad EAN para lograr explorar su viabilidad como un emprendimiento tecnológico.

Por último, en el nivel ciudadano, el código fuente del proyecto se publicará en el repositorio GitHub de manera pública con documentación en español, esto con el propósito principal de que otros estudiantes y otros investigadores puedan replicar y extender el trabajo si se desea. Además, se publicarán cosas en LinkedIn y en comunidades de datos como la comunidad de Python Colombia y la comunidad de OpenStreetMap Colombia, apuntando específicamente a audiencias técnicas interesadas en la movilidad urbana en la ciencia de datos aplicada.

Conclusiones

La primera y quizás más importante conclusión de este trabajo es que los datos históricos de siniestralidad de la plataforma SIGAT tienen valor real para comprender los patrones de accidentalidad en Kennedy, pero ese valor tiene un límite claro cuando se busca predecir condiciones actuales. Los modelos entrenados —especialmente XGBoost con $R^2 = 0{,}25$— son capaces de identificar qué variables están asociadas con accidentes más graves (principalmente la hora del día y la presencia de motocicletas), pero su capacidad para predecir la situación específica del tráfico en este momento es limitada. Esto no invalida su uso: los patrones estructurales del tráfico en Kennedy (pico del mediodía, alta participación de motocicletas, concentración de choques) son suficientemente estables como para guiar alertas preventivas generales. Pero para predicciones de alta precisión en tiempo real, se necesitan datos que el SIGAT no puede proporcionar.

La segunda conclusión es que la combinación de datos de seguridad ciudadana con datos de movilidad abre una dimensión de análisis que los sistemas de predicción de tráfico tradicionales no contemplan. Aunque el dataset de hurtos disponible tiene limitaciones de georreferenciación, la lógica del módulo de zonas rojas es sólida y su implementación con fuentes más detalladas (datos del Observatorio de Seguridad de Bogotá) produciría resultados que enriquecerían considerablemente la utilidad de la aplicación.

La tercera conclusión tiene que ver con las lecciones técnicas del proceso. Las visualizaciones genéricas de Matplotlib y Seaborn, aunque útiles para el análisis exploratorio, no son adecuadas para una aplicación de movilidad en tiempo real. La arquitectura que separa el procesamiento de datos (backend Python) del renderizado (frontend JavaScript) es la única viable para los requerimientos de reactividad y rendimiento de la aplicación propuesta. Esta decisión arquitectural, aunque obvia en retrospectiva, requirió haber probado las limitaciones del enfoque de visualización del lado del servidor para poder fundamentarse con evidencia concreta.

El sistema de reportes ciudadanos tiene el potencial de ser el componente con mayor impacto a largo plazo. Si se logra construir una comunidad de usuarios activos en Kennedy, los datos generados por sus reportes pueden superar rápidamente en utilidad a los registros históricos del SIGAT para la predicción en tiempo real. Ese es, probablemente, el camino más viable hacia un sistema de movilidad inteligente que funcione con los recursos disponibles en el contexto bogotano.

Referencias bibliográficas

Topilin, I., Jiang, J., Feofilova, A., & Beskopylny, N. (2025). Traffic flow prediction via a hybrid CPO-CNN-LSTM-attention architecture. Smart Cities, 8(5), 148. https://doi.org/10.3390/smartcities8050148

Ali, R., Ali, A., Naeem, H. M. Y., Asad, M., Alsarhan, T., & Heyat, M. B. B. (2025). A comprehensive survey of deep learning-based traffic flow prediction models for intelligent transportation systems. TACS: Transactions on Advanced Computing and Systems, 1(3), 117–137. https://doi.org/10.62762/TACS.2025.795448

Chen, L. (2026). PGSFormer: Traffic flow prediction based on joint adaptive graph learning. Scientific Reports. https://doi.org/10.1038/s41598-026-35643-x

Google. (2026). Google Maps Platform documentation. Google LLC. https://developers.google.com/maps/documentation

Khattab, O., Saravana Balaji, B., Alghadhoori, F., AlMazyad, F., AlDousari, M., Al-Ameeri, A., & Al-Kadri, M. O. (2025). Real time urban traffic prediction using RFID and a hybrid LSTM random forest model. Scientific Reports, 15, 43722. https://doi.org/10.1038/s41598-025-27485-w

Lam, K. N. (2025). Traffic prediction using LSTM, RF and XGBoost. Applied Sciences, 15(4), 2134. https://doi.org/10.5220/0013515600004619

Liu, S., Gao, J.-J., Hong, Y.-L., & Zhou, J.-C. (2025). Urban road traffic congestion index prediction based on a hybrid LightGBM-LSTM model. Archives of Transport, 76(4), 175–190. https://doi.org/10.61089/aot2025.zf4t0674

Liu, R., & Shin, S. Y. (2025). A review of traffic flow prediction methods in intelligent transportation system construction. Applied Sciences, 15(7), 3866. https://doi.org/10.3390/app15073866

Lu, J. (2023). An efficient and intelligent traffic flow prediction method based on LSTM and variational modal decomposition. Measurement: Sensors, 28, 100843. https://doi.org/10.1016/j.measen.2023.100843

Lv, Y., Duan, Y., Kang, W., Li, Z., & Wang, F. Y. (2014). *Traffic flow prediction with big data: A deep learning approach*. *IEEE Transactions on Intelligent Transportation Systems, 16*(2), 865–873. https://doi.org/10.1109/TITS.2014.2345663

Medina-Salgado, B., Sánchez-Delgadillo, E. R., Morales-Reyes, A., Castelán-Ortega, O., & Cortes-Pena, O. L. (2022). Urban traffic flow prediction techniques: A review. Sustainable Futures, 4, 100084. https://doi.org/10.1016/j.suscom.2022.100739

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71

Secretaría Distrital de Movilidad. (2020–2023). Sistema de Información de Georreferenciación de Accidentes de Tránsito (SIGAT) [Conjunto de datos]. Alcaldía Mayor de Bogotá. https://www.movilidadbogota.gov.co

Secretaría Distrital de Movilidad. (2023). Encuesta de Movilidad Bogotá-Región 2023. Alcaldía Mayor de Bogotá. https://observatorio.movilidadbogota.gov.co/sites/default/files/2025-09/Cartilla%20Encuesta%20de%20Movilidad%202023.pdf

TomTom. (2025). TomTom Traffic Index 2025. TomTom International BV. https://www.tomtom.com/traffic-index/

Yigitcanlar, T., Desouza, K. C., Butler, L., & Roozkhosh, F. (2020). Contributions and risks of artificial intelligence (AI) in building smarter cities: Insights from a systematic review of the literature. Energies, 13(6), 1473. https://doi.org/10.3390/en13061473

Yin, X., Wu, G., Wei, J., Shen, Y., Qi, H., & Yin, B. (2021). Deep learning on traffic prediction: Methods, analysis, and future directions. IEEE Transactions on Intelligent Transportation Systems, 23(6), 4927–4943. https://doi.org/10.1109/TITS.2021.3054840

Anexos

Los siguientes anexos complementan la información metodológica, técnica, financiera y de sostenibilidad del proyecto.

Anexo A. Requerimientos del sistema y matriz de sostenibilidad

Anexo B. Inventario técnico del proyecto y pruebas de validación

Anexo C. Modelo financiero CAPEX/OPEX y cronograma de desarrollo