Análisis y Predicción de la Siniestralidad Vial en Ibagué Mediante Inteligencia Artificial: Un Enfoque Desde la Ingeniería Industrial

Predicting Road Traffic Accidents in Ibagué with Artificial Intelligence: An Industrial Engineering Perspective

David Trujillo Durán[1]

Fecha de recepción: 04-07-2025 | Fecha de aprobación: 19-09-2025

Resumen

Este estudio aborda la siniestralidad vial en Ibagué, Colombia, mediante el uso de inteligencia artificial para identificar patrones y predecir accidentes. Analizar datos históricos (2020–2023) y desarrollar un modelo predictivo basado en el algoritmo Random Forest, complementado con un dashboard interactivo para visualizar dimensiones temporal, espacial y demográfica. Se procesaron bases de datos de siniestros viales mediante técnicas de análisis exploratorio (EDA) y balanceo de clases Synthetic Minority Over-Sampling Technique (SMOTE) para mitigar sesgos. El modelo alcanzó un F1-score de 0.71, demostrando capacidad moderada para clasificar causas de accidentes. Se identificó que los jóvenes hombres de 18 a 24 años son el grupo más vulnerable, con alta incidencia en motocicletas durante noches de fin de semana, especialmente en la comuna 1. La integración de IA y análisis multidimensional permite priorizar intervenciones en zonas críticas y horarios de riesgo. Estos hallazgos refuerzan la necesidad de políticas de seguridad vial focalizadas, como campañas educativas y optimización de recursos en áreas de alta accidentalidad.

Palabras Clave
Siniestro Vial, Inteligencia Artificial, Predicción, Análisis de Datos, Ingeniería Industrial

Abstract

This study addresses road safety in Ibagué, Colombia, using artificial intelligence to analyze and predict traffic accidents. To identify patterns and develop a predictive model based on the Random Forest algorithm, supported by an interactive dashboard for temporal, spatial, and demographic visualization. Historical accident data (2020–2023) was processed through exploratory data analysis (EDA) and class balancing Synthetic Minority Over-Sampling Technique (SMOTE) to reduce bias. The model achieved

1 Maestro Universidad de Ibagué, Ingeniero Industrial. Contacto: dtrujillo1003@gmail.com

an F1-score of 0.71, showing moderate effectiveness in classifying accident causes. Young men aged 18–24 were identified as the most vulnerable group, with high motorcycle accident rates during weekend nights, particularly in Comuna 1. Integrating AI and multidimensional analysis enables prioritization of high-risk zones and timeframes. These insights highlight the need for targeted road safety policies, such as educational campaigns and resource optimization in critical areas.

Keywords
Accident, Artificial Intelligence, Prediction, Data analysis, Industrial Engineering

La siniestralidad vial representa un desafío crítico para la salud pública, la seguridad y el desarrollo urbano sostenible en Ibagué, Colombia. En 2023, la ciudad reportó aproximadamente 17 muertes por cada 100.000 habitantes debido a incidentes viales, lo que la posiciona entre las más afectadas del país según El Colombiano (2023). La Agencia Nacional de Seguridad Vial (ANSV) (s.f.) también evidenció un índice de fatalidad municipal significativamente alto durante este periodo. En el departamento del Tolima, se registraron 69 muertes de mujeres por siniestros viales en 2022, ubicándolo en el sexto lugar nacional. Además, se identificó que estos hechos constituyen la principal causa de muerte violenta entre las mujeres en la región.

Como plantea Infocop (2024), los siniestros viales afectan no solo la integridad física, mental, social y profesional de las víctimas, sino también la estabilidad económica de sus familias y de las organizaciones en las que participan. Las consecuencias abarcan desde incapacidades y pérdida de movilidad hasta ansiedad, estrés postraumático y fallecimiento. Las empresas también enfrentan costos significativos derivados de estas situaciones, desde aseguradoras hasta entidades empleadoras.

En este contexto, resulta pertinente diferenciar entre los términos *accidente* y *siniestro vial*. Según Educación Bogotá (s. f.):

> es el que permite vincular causas, consecuencias y responsabilidades de la persona en un evento de tránsito. Incluso, la palabra "siniestro" tiene un significado de catástrofe y se asocia con circunstancias dolorosas, como las lesiones o la pérdida de una vida, las cuales se pudieron haber prevenido en el marco de la responsabilidad y la autorregulación. En este sentido, en seguridad vial se opta por siniestro vial y no accidente vial, ya que éste es un suceso imprevisible e inevitable asociado al azar donde se exonera a la persona de toda responsabilidad.

Esto sugiere que los siniestros viales deben ser entendidos como eventos prevenibles, con responsabilidades identificables, y no como simples hechos fortuitos lo cual a título personal cambia completamente la perspectiva desde la que se aborda estos eventos ya que abordarlos como un evento fortuito exime responsabilidades e invisibiliza a las víctimas.

Frente a este contexto social, conceptual y estadístico, esta investigación parte de la hipótesis de que el uso integrado de inteligencia artificial (IA) y análisis multidimensional de datos históricos permite predecir causas comunes de siniestros y priorizar zonas de intervención. Su relevancia radica en ofrecer herramientas aplicables a la gestión pública en movilidad, promoviendo decisiones basadas en evidencia. Además, se alinea con los Objetivos de Desarrollo Sostenible (ODS) 8, 9, 10 y 11, y se inserta en el enfoque sistémico de la ingeniería industrial, orientado a la optimización de procesos urbanos.

El objetivo general es realizar un análisis multidimensional de los siniestros viales ocurridos en Ibagué entre 2020 y 2023, con énfasis en la identificación de zonas críticas, perfiles de actores viales y condiciones contextuales asociadas. Asimismo, se desarrollará un modelo predictivo basado en el algoritmo Random Forest para clasificar causas probables de siniestros, proporcionando insumos para el diseño de políticas públicas orientadas a la reducción de la accidentalidad vial.

En cuanto a la revisión documental el enfoque fue utilizar estudios previos para el análisis de la siniestralidad vial en Colombia e implementaciones de algoritmos de *machine learning* a nivel internacional, identificando avances metodológicos, vacíos operativos y limitaciones técnicas. Por consiguiente, esta exploración resulta necesaria para fundamentar el diseño metodológico adoptado en la presente investigación, que combina análisis exploratorio de datos con modelos de aprendizaje automático aplicados al contexto de una ciudad intermedia como Ibagué. Se examinan estudios con enfoques espaciales, multivariados y computacionales, con especial atención a su aplicabilidad en entornos con restricciones de datos y capacidades institucionales limitadas.

En el ámbito colombiano, diversos estudios han abordado la siniestralidad vial desde enfoques descriptivos, espaciales y temporales. Sin embargo, la mayoría presenta limitaciones en el uso de tecnologías emergentes como la inteligencia artificial. Uno de los trabajos más robustos es el de González (2024), quien desarrolló un análisis espacial y multiescalar de los siniestros viales ocurridos en Bogotá entre 2007 y 2022. El estudio se apoyó en el procesamiento de información proveniente de bases de datos oficiales (como el Observatorio de Movilidad y la Policía Metropolitana) y aplicó técnicas como el análisis de puntos calientes (Getis-Ord Gi*), análisis espacio-temporal, y regresión geográficamente ponderada multiescala (MGWR). Estas herramientas permitieron identificar zonas de concentración significativa de siniestros, así como modelar la relación entre características del entorno urbano (densidad poblacional, infraestructura vial, iluminación, uso del suelo) y la probabilidad de ocurrencia de incidentes.

Entre los hallazgos más relevantes se encuentra la identificación sistemática de clústeres persistentes de alta siniestralidad en corredores viales primarios y áreas con alta densidad de usos mixtos, así como la concentración temporal de los eventos en franjas horarias específicas. La desagregación por UPZ permitió generar mapas de riesgo que sustentan una priorización territorial de intervenciones públicas, como el mejoramiento de señalización, control de velocidad y rediseño urbano.

No obstante, este estudio fue desarrollado en un contexto metropolitano con un sistema robusto de información georreferenciada y una estructura institucional avanzada para el monitoreo del tránsito, lo cual favorece la aplicación de modelos estadísticos multivariados de alta resolución. En contraste, ciudades intermedias como Ibagué enfrentan limitaciones en la calidad, estandarización y cobertura de los registros, lo que restringe la aplicabilidad directa de metodologías como MGWR. Por esta razón, el presente estudio adopta un enfoque complementario, basado en algoritmos de aprendizaje automático como Random Forest y técnicas de balanceo de clases (SMOTE), que resultan más tolerantes a datos incompletos o desbalanceados y ofrecen capacidades predictivas útiles para priorizar zonas de intervención.

Además, mientras que el estudio de González (2022) enfatiza el análisis espacial estático con herramientas de SIG, esta investigación añade un componente interactivo mediante la implementación de *dashboards* dinámicos que permiten a los tomadores de decisiones explorar visualmente los datos, filtrar por dimensiones clave (actor, vehículo, tiempo, comuna) y utilizar los resultados como insumo directo para el diseño de estrategias preventivas. En conjunto, ambos enfoques son complementarios y refuerzan la necesidad de adaptar las metodologías al contexto de disponibilidad de datos, escala urbana y capacidades institucionales.

De manera complementaria, Ardila (2017) desarrolló un análisis espacial y temporal de la accidentalidad y mortalidad asociada al uso de motocicletas en el Área Metropolitana de Bucaramanga. Su estudio adoptó un enfoque mixto que combinó análisis estadístico, revisión documental y georreferenciación de casos con el fin de identificar los factores que explican la alta incidencia de siniestros entre motociclistas. La investigación utilizó datos de fuentes institucionales como Medicina Legal, el Instituto Nacional de Salud (INS) y la Dirección de Tránsito, abarcando variables sociodemográficas, técnicas, ambientales y comportamentales. El modelo teórico incorporado se basó en el enfoque de causalidad estructural, que considera tanto causas directas (conducción imprudente, exceso de velocidad, desobediencia de normas) como causas indirectas (déficit en infraestructura vial, débil control institucional, falta de formación vial).

En términos de propuestas, Ardila (2017) planteó una batería de acciones orientadas a mitigar el riesgo: campañas educativas para motociclistas, incremento de la vigilancia en puntos críticos, mejora de la infraestructura vial en zonas priorizadas, fortalecimiento de la capacidad institucional para el monitoreo de la movilidad, e integración de la gestión del riesgo vial en los planes de desarrollo municipal y metropolitano. No obstante, aunque el estudio generó insumos relevantes para la formulación de políticas públicas, no incorporó herramientas de análisis predictivo ni automatizado, lo que reduce su aplicabilidad en contextos donde se requiere proyección de escenarios futuros, estimación de riesgo en tiempo real o priorización dinámica de recursos. Tampoco se consideraron mecanismos de visualización interactiva que facilitaran la consulta de resultados por parte de actores no técnicos, lo cual puede limitar su impacto en procesos de toma de decisiones distribuidos o multisectoriales.

El contraste con el presente estudio es evidente en términos metodológicos y operativos. Mientras Ardila (2017) se centró en un análisis retrospectivo con fuerte anclaje institucional y énfasis cualitativo, esta investigación implementa un enfoque computacional automatizado basado en técnicas de aprendizaje automático (Random Forest), balanceo de clases (SMOTE), validación cruzada y visualización interactiva de datos con herramientas como Dash, Plotly y Folium. La segmentación por hipótesis causales, la modelación de perfiles de riesgo y la implementación de un *dashboard* web abierto constituyen avances sustantivos que permiten transformar los datos en decisiones operativas, especialmente en contextos de ciudades intermedias como Ibagué.

Finalmente, resulta pertinente resaltar que la investigación de Ardila (2017) proporciona un marco conceptual y estratégico valioso para entender los factores estructurales de la siniestralidad motociclista, pero requiere ser complementada con enfoques algorítmicos que permitan generar alertas tempranas, predicciones territorializadas y herramientas visuales adaptables al usuario. La combinación de ambos enfoques estructural y computacional representa una oportunidad concreta para avanzar hacia sistemas inteligentes de gestión del riesgo vial.

A nivel internacional, la aplicación de técnicas de *machine learning* ha mostrado resultados prometedores en el ámbito de la predicción de siniestros. Arakelyan (2023) desarrolló un modelo predictivo para estimar la probabilidad de ocurrencia de accidentes automovilísticos en Armenia, utilizando una base de datos del sector asegurador. El estudio empleó algoritmos como Regresión Logística, Random Forest, XGBoost y Redes Neuronales Artificiales (ANN), y evaluó su desempeño sobre una variable binaria (accidente/no accidente), en un conjunto de datos con un desbalance significativo (aproximadamente 6% de registros positivos frente a 94% negativos). En términos de rendimiento, los modelos basados en XGBoost y ANN lograron los mejores resultados, alcanzando un *F1-score* cercano a 0,80 y una métrica *Cohen's Kappa* por encima de 0,74, lo que indica una alta capacidad del modelo para discriminar eventos en escenarios con clases desproporcionadas.

El conjunto de variables utilizadas en el estudio incluyó factores como la edad del conductor, historial de reclamos, monto asegurado, tiempo de exposición y tipo de vehículo, todos ellos comparables conceptualmente con las dimensiones exploradas en el presente trabajo, como edad, género, localización y contexto del siniestro. Esta coincidencia confirma la relevancia operativa de integrar atributos sociodemográficos y de comportamiento histórico en modelos predictivos de riesgo vial. Asimismo, el uso de técnicas como *undersampling* y *oversampling* en Arakelyan (2023) se alinea con la aplicación de SMOTE en esta investigación, reafirmando la necesidad de estrategias de balanceo para evitar sesgos hacia las clases mayoritarias en contextos de siniestralidad vial.

No obstante, el enfoque de Arakelyan (2023) se limita a contextos privados del sector asegurador, centrado en la optimización del riesgo financiero y la segmentación de pólizas. No considera la dimensión territorial ni variables espaciales como latitud, longitud o comunas, lo que restringe su aplicabilidad directa a políticas públicas urbanas o sistemas de alerta georreferenciados. Tampoco se incorpora una interfaz de visualización para la interpretación institucional o ciudadana de los

resultados, lo que representa una barrera para su adopción por parte de gestores públicos sin formación técnica especializada. En contraste, el presente estudio amplía el alcance metodológico al incorporar variables espaciales y temporales, desarrollar una plataforma visual interactiva y vincular los resultados a escenarios reales de intervención territorial.

En síntesis, el trabajo de Arakelyan (2023) valida técnicamente el uso de algoritmos avanzados como Random Forest, XGBoost y ANN en la predicción de eventos infrecuentes con alto costo social, y sirve como referencia metodológica para el diseño de modelos similares en dominios públicos. Su principal aporte radica en la demostración empírica de que, incluso con variables estructuradas de carácter administrativo, es posible alcanzar una precisión aceptable en contextos de alta incertidumbre. Sin embargo, su alcance se ve limitado por la falta de articulación con políticas públicas, visualización geográfica o análisis multidimensional del territorio, aspectos que esta investigación busca integrar en el caso de Ibagué como ciudad intermedia.

A pesar de estos avances, la literatura evidencia vacíos significativos. En primer lugar, existen pocos estudios en Colombia que apliquen inteligencia artificial en contextos locales con datos incompletos o de baja calidad. En segundo lugar, la integración de modelos predictivos con dashboards interactivos sigue siendo escasa, lo que limita la exploración dinámica de los datos por parte de actores institucionales. Por último, persiste una débil conexión entre los hallazgos técnicos y la formulación efectiva de políticas públicas. Esta investigación busca contribuir a cerrar estas brechas mediante el desarrollo de una herramienta predictiva e interactiva adaptada a la ciudad de Ibagué, que no solo permita identificar zonas críticas y causas frecuentes de siniestros, sino que también funcione como insumo estratégico para la gestión pública de la movilidad urbana.

En conjunto, los estudios revisados muestran avances significativos en el análisis de la siniestralidad vial desde perspectivas geoespaciales, estructurales y predictivas. Sin embargo, persisten limitaciones comunes en la articulación entre análisis técnico y herramientas operativas para la gestión pública. La escasa integración de inteligencia artificial en entornos urbanos intermedios, la baja disponibilidad de dashboards interactivos institucionales, y la ausencia de sistemas de predicción adaptados a variables contextuales, justifican la necesidad de un enfoque metodológico híbrido como el desarrollado en este estudio. La combinación de modelos de *machine learning*, balanceo de clases y visualización interactiva representa un aporte concreto para avanzar en sistemas de análisis de siniestralidad más accesibles, aplicables y orientados a la toma de decisiones en ciudades con capacidades limitadas de infraestructura de datos.

En esta investigación se empleó inteligencia artificial (IA) como herramienta complementaria para la estructuración metodológica y el desarrollo computacional. El modelo seleccionado fue Qwen 2.5 Max, accesible a través de la plataforma Qwen Chat, elegido por su capacidad para procesar entradas extensas (hasta 10.000 palabras), realizar razonamiento contextual, mantener coherencia temática entre turnos conversacionales y acceder a información en línea de forma integrada. A la fecha de elaboración de este estudio, se consideraron alternativas como ChatGPT, Claude, Gemini y Copilot.

No obstante, muchas de estas plataformas restringen sus funciones avanzadas a planes de suscripción, lo cual limita su accesibilidad para investigaciones con restricciones presupuestarias. En este sentido, la elección de Qwen se basó en su equilibrio entre capacidad técnica, apertura de funciones, velocidad de respuesta y compatibilidad con flujos de trabajo iterativos en entornos de desarrollo.

Por lo anterior, desde un criterio práctico, se valoró especialmente su desempeño en tareas que requerían generación de código estructurado, depuración de errores, segmentación de funciones, reformulación de scripts y adaptación al entorno Python, sin necesidad de recurrir constantemente a documentación externa. El punto de partida fue la formulación del siguiente *prompt* de arranque:

> Quiero realizar una investigación, el proyecto sería en el municipio de Ibagué, como recursos cuento con una base de datos recopilatoria de los siniestros entre 2020–2022 y 2023, cuento con Python y HTML, dame ideas de qué podríamos hacer.

Este *prompt* inicial no buscaba únicamente generar ideas temáticas, sino también evaluar el alcance funcional de la IA en términos de interpretación de contexto, diseño de flujo metodológico, y sugerencias de implementación computacional viables.

A lo largo del desarrollo, la IA fue utilizada como soporte continuo para acelerar la producción de código en Python y facilitar tareas de análisis estructurado, como la limpieza de datos, construcción de visualizaciones y elaboración del modelo predictivo. El flujo de trabajo se articuló en torno a tres scripts principales, cada uno asociado a una fase distinta del proceso:

1. Preprocesamiento y depuración de las bases de datos (detección y eliminación de registros nulos, tipificación de variables, y estandarización de formatos).
2. Visualización exploratoria con generación de gráficos y publicación de resultados en un entorno web.
3. Modelado predictivo mediante Random Forest con balanceo de clases y validación cruzada.

Cada bloque funcional se diseñó mediante *prompts* específicos que indicaban de manera explícita las entradas esperadas (input), las salidas deseadas (output), y los formatos requeridos para integración en el flujo de trabajo. Un ejemplo recurrente fue: “Necesito el código de una función con las siguientes entradas (edad, género, día) y quiero que me retorne un gráfico de barras con la frecuencia de siniestros.”

Este tipo de estructuras facilitó una interacción precisa y reproducible con el modelo, reduciendo errores de interpretación y favoreciendo la consistencia del código.

Durante la investigación, cuando surgieron errores durante la ejecución, se reutilizó el entorno conversacional de la IA como sistema de depuración, copiando directamente los mensajes arrojados

por la consola para obtener diagnósticos rápidos y soluciones específicas. En casos donde el modelo empezaba a generar respuestas erráticas o inconsistentes, se optó por reiniciar la conversación desde cero, aplicando una estrategia de iteración controlada para evitar acumulación de errores lógicos o alucinaciones. Esta práctica, aunque rudimentaria, permitió mantener trazabilidad sobre los cambios implementados y preservar el control en cada etapa crítica.

En conjunto, esta metodología asistida por IA no solo aceleró el tiempo de desarrollo, sino que permitió mantener una estructura lógica en la producción de código, facilitando la modularidad, reutilización y documentación interna del proceso. Si bien no se delegaron decisiones sustantivas ni análisis estadísticos a la IA, su papel como soporte computacional resultó fundamental para alcanzar un producto replicable, escalable y con valor operativo concreto. Este enfoque evidencia el potencial de los modelos de lenguaje como asistentes técnicos en investigaciones aplicadas, especialmente en contextos de recursos limitados donde se requiere maximizar eficiencia sin comprometer la calidad técnica.

Una vez estructurado el flujo de trabajo inicial con IA, se procedió al análisis exploratorio de datos (EDA), desarrollado a partir de las bases de datos suministradas por los líderes del Observatorio de Movilidad y Transporte de la Universidad de Ibagué (OMTU), Juliana Rojas y Juan Zuluaga quienes amablemente suministraron la información y brindaron su apoyo en el proceso de análisis e interpretación de la información allí contenida. Las bases de datos comprenden registros de siniestralidad vial ocurridos entre los años 2020 y 2023. Estas bases consolidan información procedente de reportes oficiales, integrando variables como tipo de siniestro, actores involucrados, ubicación aproximada que el equipo del observatorio se dio a la tarea de georreferenciar ya que la base original no cuenta con esa información, fecha, hora y características de los vehículos. Exceptuando la base correspondiente a 2021, debido a que, presentaba limitaciones de cobertura y ausencia de coordenadas geográficas, el resto de los años contaban con referenciación espacial aproximada (latitud y longitud), lo que permitió su incorporación en los análisis territoriales. Sin embargo, desde este punto ya se evidencia una oportunidad de mejora a nivel institucional en la medición de datos ya que se debería contar con bases georreferenciadas de siniestralidad al entender el valor que aporta conocer donde ocurren los siniestros viales.

El uso de información geográfica producto de la minuciosa labor del equipo del OMTU, aportó para la agregación por comunas y la generación de mapas de calor, clústeres y visualizaciones comparativas por zonas de la ciudad. Esta condición permitió realizar segmentaciones espaciales de la siniestralidad, identificar áreas con alta recurrencia de incidentes y establecer correlaciones entre la ocurrencia de siniestros y su localización urbana. Adicionalmente, la integración de la dimensión temporal (año, mes, día de la semana y hora) posibilitó el análisis multivariable de patrones espacio-temporales, facilitando la caracterización de franjas críticas según tipo de vehículo, actor vial o severidad del evento.

En esta etapa se aplicaron procesos sistemáticos de depuración, transformación y normalización de los datos. Esto incluyó la eliminación de registros incompletos, homogenización de formatos, recodificación de variables categóricas y ajuste de escalas temporales. A partir de esta base depurada, se organizaron los datos en estructuras específicas para cada una de las dimensiones analíticas definidas en el estudio: actores viales, demografía, características del siniestro, variables temporales, ubicación espacial y tipo de vehículo.

El proceso de limpieza incluyó la eliminación de registros incompletos, columnas sin información significativa y datos inconsistentes que comprometían la calidad analítica. Esta etapa fue fundamental para garantizar que los modelos y visualizaciones posteriores se basaran en información confiable y estructurada. La base consolidada original contaba con 110 columnas, muchas de las cuales eran redundantes, presentaban nomenclaturas irregulares o contenían vacíos sistemáticos. Tras la depuración, se conservaron 4.544 registros válidos, representativos del fenómeno de siniestralidad vial en Ibagué entre 2020 y 2023.

Una vez estabilizada la estructura del conjunto de datos, se definieron cinco dimensiones clave de análisis, orientadas a descomponer el fenómeno en componentes interpretables:

- Actores viales: permitiendo diferenciar entre conductores, peatones, acompañantes u otros tipos de participantes, considerando variables como edad y género.
- Demografía: enfocada exclusivamente en la caracterización sociodemográfica de los involucrados, facilitando análisis por grupos etarios, género y su distribución relativa.
- Características del siniestro: abarcando el tipo de evento (choque, volcamiento, atropello), su gravedad (con heridos o fallecidos) y la participación relativa de cada tipo de actor.
- Dimensión espacial y temporal: con base en la ubicación geográfica (latitud, longitud, comuna) y aspectos cronológicos (día de la semana, hora, mes, año) que permiten construir perfiles espacio-temporales del riesgo vial.
- Tipo de vehículo: clasificando las unidades involucradas en motocicletas, vehículos particulares, transporte público, bicicletas, entre otros.

Para cada dimensión, se diseñaron scripts personalizados en Python que automatizaron el procesamiento, transformación y representación gráfica de los datos. Estas herramientas facilitaron el desarrollo modular de los análisis y permitieron replicar la lógica exploratoria para distintas combinaciones de variables.

Inicialmente se consideró la exportación de resultados en formato estático (imágenes, PDF), pero dicha opción limitaba la capacidad de exploración cruzada de los datos. En respuesta, y aprovechando la infraestructura digital del OMTU, se optó por implementar un dashboard interactivo en línea que permite a los usuarios explorar los resultados.

Para este entorno interactivo se integraron herramientas de desarrollo específicas:

- Pandas y NumPy: para el tratamiento, filtrado y agregación eficiente de grandes volúmenes de datos.
- Plotly: para la generación de gráficos dinámicos (barras, líneas, histogramas) con capacidad de interacción directa.
- Folium: para la representación de los datos georreferenciados en mapas por comunas, permitiendo identificar zonas críticas.
- Dash: como marco principal para construir la arquitectura del dashboard, integrando gráficos, menús desplegables, selectores de tiempo y filtros dinámicos.
- Render: utilizado para el despliegue y publicación web del producto final, asegurando su disponibilidad pública desde el portal del OMTU.

La visualización interactiva, disponible en línea, refuerza el valor operativo del análisis exploratorio al facilitar una interpretación directa y segmentada de los datos. Este componente no solo convierte las bases de datos en información legible, sino que habilita la toma de decisiones basada en evidencia, sin requerir formación técnica avanzada por parte del usuario final. La arquitectura modular del sistema permite, además, su futura ampliación con nuevos años, variables o capas de análisis, manteniendo la lógica de navegación ya construida.

Este enfoque modular permitió una lectura más precisa de los datos y facilitó la generación de visualizaciones dirigidas, adaptadas a cada eje de análisis. En paralelo, se diseñaron scripts reutilizables en Python para automatizar la producción de gráficos y facilitar su posterior integración en un entorno web interactivo. Esta combinación de depuración estructurada, segmentación temática y visualización multicanal constituyó la base operativa sobre la cual se construyó el dashboard exploratorio final.

Para el desarrollo del modelo predictivo, se definió como variable objetivo la columna denominada Hipótesis, correspondiente al juicio preliminar emitido por el agente de tránsito sobre la causa probable del siniestro vial. Esta variable representaba una entrada textual no estandarizada, por lo que su uso requería un tratamiento previo de limpieza, normalización y clasificación.

Durante la revisión inicial se identificaron inconsistencias ortográficas, uso de sinónimos no uniformes, abreviaciones irregulares y presencia de múltiples criterios superpuestos en una misma entrada. Además, algunos registros estaban vacíos o eran irrelevantes para el modelo (por ejemplo: “se desconoce” o “en investigación”). Ante esto, se procedió a una consolidación semántica, que consistió en agrupar las hipótesis dispersas en un conjunto de 10 categorías unificadas, construidas con base en criterios conceptuales, frecuencia de aparición y claridad operacional:

1. Incumplimiento de prelaciones
2. Falta de atención a la vía y actores viales
3. No mantener distancia de seguridad

4. Desobediencia de señales y normas de tránsito
5. Uso incorrecto de carriles
6. Falta de precaución
7. Exceso de velocidad
8. Impericia en el manejo
9. Conductas de riesgo en maniobras específicas
10. Embriaguez

Este proceso de normalización categórica fue crítico para garantizar que el modelo pudiera operar sobre clases discretas, homogéneas y con suficiente volumen de datos por grupo. Cabe aclarar que la construcción de estas categorías fue de tipo inductivo, basada tanto en la recurrencia empírica como en el criterio técnico.

Por otra parte, se consideró la estructura relacional de los datos, donde un mismo siniestro podía registrar múltiples involucrados (actores viales) asociados a una misma hipótesis. Esta característica generaba el riesgo de duplicidad de registros al momento de entrenar el modelo, lo cual podía distorsionar el balance de clases o sobre representar ciertos perfiles. Para evitar este sesgo, se decidió reducir la unidad de análisis a un actor vial por siniestro, conservando únicamente el primer registro según el orden cronológico de aparición en la base.

En esta selección se priorizó la conservación de atributos como edad y género, que resultaban esenciales para la construcción de perfiles sociodemográficos y además evitaban un número muy grande de datos vacíos. Esta depuración permitió obtener un subconjunto final de 1.621 registros válidos, distribuidos de forma suficientemente representativa entre las categorías de hipótesis.

Este conjunto final sirvió como base estructurada para el entrenamiento del modelo predictivo, garantizando tanto la calidad del etiquetado como la unicidad de los casos, condiciones fundamentales para evitar sobreajuste y asegurar interpretabilidad en los resultados del modelo.

El algoritmo de clasificación seleccionado para el modelo predictivo fue Random Forest, una técnica ampliamente utilizada en problemas de clasificación multiclase por su capacidad para manejar relaciones no lineales, interacción entre variables y conjuntos de datos con ruido o valores atípicos. Este modelo opera mediante la construcción de múltiples árboles de decisión, entrenados sobre subconjuntos aleatorios del conjunto de datos original, lo cual permite reducir el sobreajuste y aumentar la generalización. Cada árbol genera una predicción independiente, y la decisión final se obtiene mediante un proceso de votación mayoritaria entre todos los árboles del bosque. En el presente caso, el objetivo fue predecir la hipótesis más probable asociada a un siniestro vial, a partir de atributos geoespaciales (latitud, longitud), temporales (día de la semana) y sociodemográficos (edad, género, comuna).

Cada árbol de decisión sigue una estructura jerárquica donde, en cada nodo, se selecciona la variable y el umbral de división que maximizan la ganancia de información o reducen la impureza. A través de este proceso recursivo, los árboles agrupan observaciones similares en función de los valores de sus atributos hasta llegar a una clasificación final, en este caso, una de las 10 hipótesis causales previamente normalizadas.

Para evaluar el rendimiento del modelo, se utilizaron métricas estándar de clasificación: precisión, *recall* y F1-score. La precisión indica el porcentaje de predicciones positivas correctas frente al total de predicciones positivas realizadas. El *recall*, en cambio, mide la capacidad del modelo para identificar correctamente los casos positivos reales. Finalmente, el F1-score representa la media armónica entre precisión y *recall* (Kundu, 2022), siendo especialmente útil en contextos donde las clases están desbalanceadas. Esta métrica, al ser menos sensible al desequilibrio que la precisión simple, permite una evaluación más robusta de modelos en escenarios reales de siniestralidad.

Dada la alta disparidad en la frecuencia de aparición de las hipótesis, se implementó la técnica de balanceo SMOTE (*Synthetic Minority Over-sampling Technique*). Esta técnica genera artificialmente nuevos registros para la clase minoritaria mediante interpolación entre observaciones cercanas, en lugar de simplemente duplicar registros existentes. El objetivo es reducir el sesgo del modelo hacia las clases mayoritarias, facilitando un entrenamiento más equilibrado y justo. SMOTE es particularmente efectivo cuando se desea evitar la pérdida de información asociada a técnicas de *undersampling*.

El algoritmo se apoya en el uso del algoritmo k-vecinos más cercanos (k-NN), el cual permite identificar las observaciones más próximas (según una métrica de distancia) para la generación de ejemplos sintéticos. El valor de k especifica cuántos vecinos se consideran en este proceso, lo cual influye directamente en la diversidad y representatividad de los nuevos registros generados (IBM, 2025). Este enfoque mejora la representación de las clases minoritarias sin alterar la distribución estadística global del conjunto de datos.

En cuanto al conjunto de variables utilizadas para el entrenamiento, se incluyeron: Latitud, Longitud, Edad, Género, Día y Comuna, seleccionadas por su relevancia tanto estadística como operacional, y su disponibilidad en los registros procesados. Estas variables aportan información suficiente para capturar patrones de riesgo asociados al contexto físico, temporal y social del siniestro.

Para evitar sobreajuste y asegurar una evaluación más precisa, se empleó la técnica de validación cruzada K-Folds con 10 particiones, mediante la cual el conjunto de datos se divide aleatoriamente en 10 subconjuntos del mismo tamaño. En cada iteración, 9 subconjuntos se utilizan para el entrenamiento y 1 para la validación, rotando sucesivamente hasta que cada subconjunto haya sido utilizado como conjunto de prueba. Este proceso, implementado con la librería Scikit-Learn de Python, maximiza el uso del conjunto de datos disponible, que por su volumen moderado no permitía una separación clásica de entrenamiento/prueba sin pérdida significativa de información.

Finalmente, los resultados se visualizaron mediante una matriz de confusión, que permite identificar cuántas veces el modelo acertó o erró en la predicción de cada clase. Se definió además un umbral de riesgo de 0.9, utilizado para filtrar solo aquellas predicciones cuya probabilidad asignada por el modelo superara dicho valor, garantizando que las salidas finales fueran altamente confiables y aptas para uso operativo.

En conjunto, estos procedimientos metodológicos permitieron construir una arquitectura robusta de análisis predictivo, capaz de operar con volúmenes intermedios de datos, lidiar con desequilibrios estructurales entre clases, y ofrecer resultados interpretables y que son coherentes con la realidad de la ciudad.

Desde una perspectiva personal, esta metodología responde a la necesidad de compatibilizar capacidades técnicas limitadas con una alta exigencia de resultados estructurados, reproducibles y útiles. No se partió de un entorno ideal, sino de un contexto realista en términos de disponibilidad de recursos, tiempo y acceso a herramientas avanzadas. Por ello, la elección de componentes como Python, Dash y la IA conversacional Qwen no obedece únicamente a su condición de libre acceso, sino a su capacidad comprobada de articular un flujo de trabajo controlado, flexible y escalable, sin renunciar a estándares de calidad técnica.

Frente a otras alternativas cerradas, opacas o excesivamente dependientes de licencias, esta arquitectura ofreció autonomía operativa, trazabilidad del código y posibilidad de adaptación modular. Cada componente del sistema fue elegido no por moda o disponibilidad, sino por su alineación funcional con los objetivos analíticos y su compatibilidad con entornos institucionales de bajo presupuesto. Esto permitió una integración racional de tecnologías, donde cada herramienta cumple una función específica en el ciclo de procesamiento, visualización o predicción.

El uso de IA como asistente funcional se justificó no como una delegación de criterio, sino como un acelerador técnico supervisado. En este proyecto, la IA se utilizó con la premisa de ser puntual, dirigida y bajo control constante. Las solicitudes se diseñaron con *prompts* estructurados, las respuestas se validaron línea por línea y cualquier desviación se corregía mediante reinicio o reformulación. Nunca se delegó el diseño analítico ni se aceptaron respuestas automáticas sin revisión crítica. Esta elección metodológica refleja un principio rector: la herramienta debe estar subordinada al diseño lógico del sistema, no al revés.

En cuanto al enfoque general, se priorizó un modelo reproducible, documentado y adaptable, antes que uno hipercomplejo, opaco o dependiente de estructuras externas no replicables. Este enfoque no pretende explotar al máximo la sofisticación algorítmica, sino garantizar que el modelo pueda ser comprendido, ajustado y aplicado a nivel general sin requerir capacidades técnicas avanzadas. Esto tiene implicaciones directas en la escalabilidad, la transferencia metodológica y la sostenibilidad operativa en el tiempo.

Adicionalmente, se tomó la decisión consciente de evitar la dependencia de librerías experimentales o entornos de alto mantenimiento, lo que refuerza la viabilidad de actualizar o extender el modelo mediante cambios incrementales, sin necesidad de rediseños completos. Este criterio de diseño es especialmente relevante en contextos públicos, donde la rotación de personal, la obsolescencia tecnológica o los cambios en prioridades institucionales son frecuentes.

En resumen, se construyó un modelo técnicamente sólido, operacionalmente viable y estratégicamente útil para el contexto local, con un diseño orientado a decisiones prácticas, adaptable a futuras extensiones y compatible con procesos institucionales reales.

Como producto del análisis exploratorio, se diseñó un *dashboard* interactivo disponible públicamente en https://resumen-siniestros-ibague.onrender.com. Allí se visualizan los principales patrones espacio-temporales, demográficos y modales identificados. Adicionalmente, en el repositorio https://github.com/Dtrujillo-d/Dashboard.git se encuentra el código para acceder al *dashboard* así como la base de datos implementada y demás elementos para replicar o mejorar el cuadro de control construido. A continuación, se sintetizan los hallazgos más relevantes según cada dimensión analizada.

Actores Viales

- 2.578 de los registros corresponde a conductores.
- Del total de conductores involucrados en siniestros, el 85,5% son hombres.
- La edad promedio de los actores viales varía según su rol, pero se concentra entre los 30 y 34 años.
- Los conductores con mayor participación en siniestros se ubican entre los 21 y 24 años.

Demografía

- El grupo etario con mayor frecuencia de siniestros corresponde a personas entre 21 y 23 años (842 registros).
- En ese rango, el 78% de los involucrados son hombres y el 22% mujeres
- La relación general entre hombres y mujeres involucrados en siniestros es de 4 a 1.

Temporalidad

- El mes con mayor número de siniestros fue marzo.
- Los viernes y sábados concentran la mayoría de los siniestros (707 y 711 casos, respectivamente).
- La hora con mayor número de siniestros fue las 12:00 AM.
- Entre semana se identificaron picos adicionales a las 16:00 y 18:00 horas.
- Los fines de semana, la siniestralidad incrementa significativamente entre las 23:00 y las 5:00 horas.

Características del siniestro

- El tipo de siniestro más común es el choque, que representa el 82,3% de los registros.
- Marzo fue el mes con mayor número de heridos (299 casos).
- Junio presentó el mayor número de fallecimientos (18 casos).
- El 61,1% de los casos, los implicados fueron conductores.
- La proporción entre heridos y fallecidos es de 32 a 1.

Vehículos Involucrados

- Las motocicletas fueron los vehículos más involucrados en siniestros.
- Febrero fue el mes con mayor siniestralidad en motocicletas (326 casos).
- Marzo registró el mayor número total de siniestros vehiculares (353).
- Los vehículos particulares lideran la participación en siniestros.
- Se reportaron 420 buses de servicio público implicados en siniestros viales.

Modelo Predictivo

La variable objetivo que fue seleccionada fue la columna Hipótesis, categorizada en diez causas probables tras un proceso de depuración y normalización de texto. La distribución inicial de clases evidenció un fuerte desbalance entre las categorías, con diferencias de hasta 1.300 registros entre la clase más y menos frecuente. Por esta razón, se implementó la técnica SMOTE para equilibrar las clases. La categoría con menor representación (“Impericia en el manejo”) fue eliminada para evitar distorsión en la predicción.

El modelo de clasificación empleado fue Random Forest, con un esquema de validación cruzada mediante K-Folds (k=10). Las variables predictoras utilizadas fueron: Latitud, Longitud, Edad, Género, Día y Comuna. El rendimiento promedio del modelo fue el siguiente:

- F1-score promedio: 0,71
- Precisión (precisión): 0,70
- Recall (exhaustividad): 0,72

Estos valores reflejan un desempeño equilibrado del modelo, con una capacidad aceptable para predecir correctamente las causas probables de siniestros viales a partir de condiciones espacio-temporales y sociodemográficas.

La matriz de confusión resultante evidenció una adecuada correspondencia entre las clases reales y las predichas, con un margen de error aceptable en la mayoría de las categorías. Adicionalmente, se construyó un gráfico de importancia de variables, donde se identificó que la Edad, seguida por la Latitud y Longitud, fueron los factores más influyentes en la predicción de las hipótesis.

Por último, se estableció un umbral de riesgo de 0,9, lo que permitió priorizar únicamente aquellas predicciones con alta certeza, mejorando así la confiabilidad del modelo para su uso como herramienta de apoyo en la toma de decisiones públicas.

A partir del análisis exploratorio y predictivo realizado en esta investigación, se derivan las siguientes conclusiones clave para la comprensión y gestión de la siniestralidad vial en la ciudad de Ibagué:

- Riesgo elevado en jóvenes conductores: El grupo etario entre 18 y 24 años concentra el mayor número de siniestros viales registrados en la base de datos. Esta concentración no solo se refleja en volumen absoluto, sino también en la proporción relativa dentro de los segmentos poblacionales analizados. La reiterada implicación de este grupo sugiere una exposición significativa al riesgo vial, probablemente asociada a factores como baja experiencia al volante, sobreconfianza, conductas impulsivas y limitado reconocimiento de situaciones de peligro. Este hallazgo refuerza la necesidad de establecer programas de intervención dirigidos específicamente a esta población, incluyendo formación obligatoria en gestión del riesgo vial, simuladores de conducción, campañas de sensibilización con enfoque conductual, y restricciones progresivas de licencia según nivel de experiencia. La estrategia debe ir más allá del enfoque informativo y priorizar acciones que modifiquen la percepción del riesgo y los hábitos concretos de conducción.

- Mayor siniestralidad en horarios nocturnos de fin de semana: Se identificó un aumento marcado de incidentes entre viernes y domingo en las franjas comprendidas entre las 23:00 y las 05:00 horas. Este patrón se desvía significativamente de la distribución esperada bajo condiciones de tráfico promedio y coincide con periodos de alta movilidad asociada al ocio nocturno. La coincidencia temporal sugiere la influencia de factores como consumo de alcohol, fatiga, condiciones de baja visibilidad o tránsito reducido, lo que incrementa las velocidades medias y reduce la percepción de riesgo. Estos hallazgos justifican la implementación de estrategias específicas para estas horas críticas, como refuerzo de los operativos de control de alcoholemia, mejoramiento del alumbrado público, campañas específicas de prevención nocturna y diseño de esquemas de transporte alternativo para el retorno seguro durante la madrugada.

- Brecha de género en la participación vial: El estudio evidenció una participación masculina desproporcionada en los siniestros viales, con una relación aproximada de 4 a 1 respecto a las mujeres. Esta brecha puede estar relacionada con mayores niveles de exposición, mayor frecuencia de conducción, diferencias en la elección del tipo de vehículo (por ejemplo, motocicleta) y patrones diferenciales de comportamiento vial. Aunque la estadística no permite inferencias causales directas, estos resultados abren la puerta a la formulación de hipótesis operativas y a la necesidad de incorporar un enfoque de género en las políticas de seguridad vial. Esto incluye la revisión de los programas de formación, la identificación de factores de riesgo específicos según género y la promoción de investigaciones complementarias que profundicen en los determinantes sociales y conductuales de esta disparidad.

- Alta incidencia de motocicletas: Las motocicletas representaron el tipo de vehículo con mayor número de siniestros reportados. Esta sobre-representación es consistente con estudios nacionales y evidencia un problema estructural de seguridad en este modo de transporte. La facilidad de acceso a motocicletas, su bajo costo relativo, la alta densidad de uso en ciudades intermedias y la exposición directa del conductor explican en parte esta incidencia. Frente a esto, resulta urgente diseñar e implementar estrategias específicas dirigidas a motociclistas: capacitaciones obligatorias antes de la expedición del pase, fortalecimiento de los controles sobre el uso de casco y elementos de protección, rediseño de la infraestructura para reducir los puntos críticos de fricción con otros vehículos, y políticas de regulación diferenciada por cilindrada o perfil del conductor.

- Concentración espacial de los siniestros: El análisis geográfico reveló una mayor concentración de siniestros en la comuna 1, lo que sugiere una combinación de factores de riesgo urbano que requieren intervención prioritaria. Esta concentración puede estar relacionada con características del entorno como alta densidad poblacional, mezcla de usos del suelo, congestión vial, falta de infraestructura peatonal segura o deterioro en las condiciones de señalización y visibilidad. La identificación de este patrón permite delimitar una zona crítica que debería ser considerada como polígono piloto para la implementación de acciones preventivas integrales. Esto incluye auditorías de seguridad vial, intervenciones de urbanismo táctico y control permanente con monitoreo institucional.

- Cobertura institucional y posibles vacíos en el registro: Finalmente, como se comentó en la metodología, se detectaron indicios de posibles inconsistencias o vacíos en la recolección de información georreferenciada y demográfica en ciertas zonas de la ciudad. Esta situación puede estar vinculada a una distribución desigual del recurso humano (agentes de tránsito), limitaciones logísticas en el levantamiento de información o ausencia de mecanismos estandarizados de reporte. Estos vacíos no solo afectan la representatividad del análisis, sino que pueden conducir a un subregistro sistemático que distorsiona la planeación pública. Se recomienda revisar la cobertura operativa de los agentes de tránsito, evaluar la calidad del proceso de captura de datos en campo, e impulsar el diseño de herramientas móviles unificadas para el reporte de siniestros, con validación automática de datos mínimos obligatorios y georreferenciación en tiempo real.

El modelo predictivo basado en el algoritmo Random Forest arrojó un F1-score promedio de 0,71, lo cual puede considerarse un desempeño aceptable y funcional, especialmente si se contextualiza en un entorno de datos limitados, con calidad heterogénea y un fuerte desbalance de clases. Esta métrica sugiere que el modelo logra una razonable capacidad de generalización y puede ser utilizado como herramienta de apoyo en procesos de clasificación preliminar. El desempeño no solo valida la aplicabilidad del enfoque, sino que también demuestra la pertinencia de utilizar algoritmos de ensamble robustos cuando se trabaja con fenómenos multivariados y distribuciones asimétricas, como ocurre en los registros de siniestralidad vial.

En términos de importancia relativa de las variables predictoras, se destaca que edad, latitud y longitud fueron los atributos con mayor peso dentro de la estructura del modelo. Esto confirma que los factores demográficos y geoespaciales son determinantes para predecir las causas probables de los siniestros, y sugiere que cualquier estrategia de intervención debe tener una base territorial clara, segmentada por grupo etario. El hecho de que el componente espacial tenga un peso comparable al sociodemográfico refuerza la necesidad de trabajar con mapas de riesgo dinámicos y con herramientas de análisis territorial continuo, lo cual abre la posibilidad de integrar capas adicionales como infraestructura, densidad vehicular o presencia de equipamientos urbanos.

La aplicación del método SMOTE (*Synthetic Minority Over-sampling Technique*) fue fundamental para corregir el sesgo del modelo hacia las clases mayoritarias. Este tipo de balanceo es especialmente útil cuando se busca conservar la mayor cantidad de registros posibles, evitando la pérdida de información inherente al undersampling. No obstante, como limitación metodológica, debe señalarse que el uso de muestras sintéticas puede alterar parcialmente la estructura interna de los datos y generar sobreajuste si no se controla adecuadamente. Por este motivo, se sugiere que futuras versiones del modelo incluyan una comparación sistemática con otras técnicas de balanceo.

Adicionalmente, se identificó que la variable objetivo, la columna *Hipótesis*, presenta un desafío metodológico no menor: se trata de una clasificación cualitativa de carácter subjetivo, emitida por agentes de tránsito sin un protocolo estandarizado de codificación. Esto introduce un nivel de variabilidad interobservador que puede afectar la consistencia del modelo. Para reducir este ruido, se sugiere avanzar en un proceso de normalización institucional de esta variable mediante una matriz de codificación técnica, orientada por criterios periciales, que garantice homogeneidad semántica y operativa. Alternativamente, podrían explorarse modelos jerárquicos de predicción que primero clasifiquen el tipo de siniestro y luego infieran la hipótesis probable, disminuyendo el impacto de la ambigüedad inicial.

Por otra parte, la incorporación de un dashboard interactivo construido con Dash, Plotly y Folium representa una mejora sustantiva frente a modelos tradicionales que limitan sus resultados a reportes estáticos o análisis descriptivos. Esta interfaz no solo democratiza el acceso a la información, sino que también habilita la exploración multidimensional de los resultados, filtrando según comuna, franja horaria, tipo de actor vial y tipo de vehículo. Su publicación dentro del portal del OMTU facilita su adopción institucional y convierte el modelo en un instrumento operativo, accesible incluso para usuarios sin conocimientos técnicos en programación o análisis de datos.

En conjunto, el sistema de análisis implementado permite estructurar una lógica operativa para la acción pública en movilidad segura, basada en cinco dimensiones clave:

1. Edad (identificación de poblaciones de riesgo);
2. Tiempo (segmentación por horarios críticos);

3. Geografía (zonificación territorial de alta siniestralidad)
4. Actor (tipificación por tipo de vehículo y perfil del involucrado);
5. Causa probable (clasificación predictiva basada en condiciones observadas).

A partir de los hallazgos y limitaciones del presente estudio, se delinean varias líneas de trabajo que no solo permiten consolidar y perfeccionar el modelo propuesto, sino también escalar su aplicabilidad hacia un sistema integral de apoyo a la toma de decisiones en seguridad vial. Estas líneas se estructuran en cinco dimensiones funcionales: mejora del modelo, expansión de datos, integración tecnológica, evaluación de impacto y replicabilidad territorial.

Con base en los hallazgos y limitaciones del presente estudio, se identifican varias líneas estratégicas que pueden fortalecer la continuidad y escalabilidad del análisis, tanto en términos metodológicos como operativos:

- Ampliación del modelo mediante variables contextuales críticas
  Actualmente, el modelo se apoya en variables geoespaciales y demográficas básicas. Sin embargo, existe un margen considerable para enriquecer su capacidad explicativa mediante la incorporación de nuevas variables que capturen mejor el contexto operativo de los siniestros. Entre ellas se destacan:

  1. Condiciones climáticas (lluvia, niebla, temperatura, humedad relativa).
  2. Estado y tipo de vía (pavimento, pendiente, número de carriles, presencia de señalización horizontal o vertical).
  3. Iluminación pública (tipo y cobertura del alumbrado en la zona).
  4. Flujo vehicular estimado por hora, extraído de sensores o estimaciones satelitales.
  5. Eventos extraordinarios (festividades, marchas, conciertos, accidentes previos).

  La inclusión de estas variables no solo incrementaría la precisión del modelo, sino que permitiría generar perfiles de riesgo altamente localizados, útiles para la planificación táctica de intervenciones preventivas.

- Evolución hacia un sistema de predicción en tiempo real
  Una proyección natural de este trabajo es su integración con fuentes de datos en línea y sensores urbanos que operen en tiempo real. Esto habilitaría:
  - Modelos adaptativos capaces de actualizar sus predicciones con base en condiciones dinámicas del entorno.
  - Alertas tempranas para autoridades de tránsito, basadas en umbrales de riesgo derivados del modelo.
  - Integración con plataformas ciudadanas o aplicaciones móviles, permitiendo retroalimentación bidireccional (reportes de eventos, condiciones, incidentes).

Este componente requiere tanto desarrollo tecnológico como protocolos de gobernanza de datos, así como acuerdos interinstitucionales para la interoperabilidad de sistemas.

- Profundización metodológica en modelos geoespaciales y de inteligencia artificial
  El uso de Random Forest demostró ser efectivo para una primera aproximación. No obstante, se recomienda explorar modelos avanzados con mayor sensibilidad espacial y capacidad de modelado no lineal. Además, es pertinente comparar el rendimiento de técnicas de balanceo como SMOTE con métodos alternativos para evitar distorsiones por síntesis artificial.

- Diseño y validación de intervenciones piloto orientadas por evidencia
  La utilidad del modelo debe ser validada operativamente mediante su traducción en acciones piloto territorializadas, especialmente en las zonas y franjas horarias identificadas como críticas. Esto implica:
    - Diseñar intervenciones diferenciadas por comuna, tipo de vehículo y edad de los involucrados.
    - Establecer indicadores de impacto (reducción de siniestros, cambios en tipos de hipótesis, modificación de patrones temporales).
    - Aplicar diseño experimental para medir el efecto real de las políticas basadas en datos.

  Esta línea permitiría cerrar el ciclo de retroalimentación entre análisis predictivo, toma de decisiones y evaluación de resultados, aportando evidencia concreta sobre la eficacia de las acciones.

- Replicabilidad nacional e institucionalización de la herramienta
  El modelo desarrollado para Ibagué es escalable a otras ciudades intermedias de Colombia, bajo un enfoque modular. Para ello se sugiere:
    - Establecer criterios técnicos para la selección de municipios replicables (población, estructura vial, disponibilidad de datos).
    - Elaborar una guía metodológica para la implementación local del modelo.
    - Establecer alianzas entre observatorios de movilidad, gobiernos locales y universidades regionales para garantizar la adopción y sostenibilidad del sistema.

  En paralelo, se recomienda avanzar hacia la institucionalización de un sistema nacional de análisis predictivo vial, que estandarice metodologías, fomente la interoperabilidad de bases de datos, y ofrezca soporte técnico continuo a los entes territoriales.

- Reforzamiento de capacidades institucionales y cultura de datos
  La sostenibilidad del modelo requiere no solo tecnología, sino capacidades institucionales consolidadas en:
    - Recolección, depuración y estandarización de datos
    - Formación técnica en ciencia de datos aplicada a la movilidad.
    - Adopción de protocolos éticos para el uso de inteligencia artificial en decisiones públicas.

Se recomienda promover la creación de células de análisis predictivo en secretarías de movilidad, con perfiles mixtos (ingeniería, estadística, geografía, ciencia de datos), responsables de mantener y escalar los modelos en operación.

En síntesis, el trabajo futuro no debe centrarse únicamente en perfeccionar el modelo desde un punto de vista algorítmico, sino en consolidar un ecosistema completo de datos, capacidades institucionales, herramientas interoperables y cultura organizacional orientada a la acción basada en evidencia. Esta visión permitiría convertir modelos como el desarrollado en este estudio en pilares operativos de una política pública moderna de seguridad vial.

Referencias

Agencia Nacional de Seguridad Vial [ANSV]. (s. f.). Índice de Fatalidad Municipal. Agencia Nacional de Seguridad Vial. Recuperado 23 de abril de 2025, de https://ansv.gov.co/es/observatorio/estad%C3%ADsticas/indice-de-fatalidad-municipal

González, F. H. (2022). *Análisis espacial de siniestros viales en Bogotá, Colombia durante el periodo 2007 – 2022* [Tesis de Maestría en Salud Pública]. Pontificia Universidad Javeriana. https://repository.javeriana.edu.co/items/3bfe2949-59f9-479e-8ce2-460c1ed1d401

Ardila, J. A. (2018). *Análisis espacial y temporal de la accidentalidad y muerte generado por el uso de la motocicleta en el área Metropolitana de Bucaramanga* [Tesis de maestría]. Universidad de Santander, Bucaramanga. https://repositorio.udes.edu.co/entities/publication/ee0eb9e8-cc8f-4906-994a-660732b7291e

Arakelyan, D. (2023). *Machine Learning-Powered Accident Prediction for the Automotive Insurance Industry*. American University of Armenia. https://cse.aua.am/files/2023/12/Dawid_Arakelyan_Capstone.pdf

Educación Bogotá. (s. f.). Protocolo de atención de siniestros viales para establecimientos educativos del Distrito Capital. https://www.educacionbogota.edu.co/portal_institucional/sites/default/files/inline-files/Anexo%2011%20protocolo_atencion_siniestros_viales_Establecimientos_edu.pdf

El Colombiano. (2023, 21 junio). *Las 10 ciudades con más accidentes de tránsito en Colombia, ¿está la suya?* El Colombiano. https://www.elcolombiano.com/colombia/cuales-son-las-diez-ciudades-de-colombia-con-mas-accidentes-de-transito-GM21789609

IBM. (2025, abril 1). *¿Qué es el algoritmo de k vecinos más cercanos (KNN)?* IBM. https://www.ibm.com/mx-es/think/topics/knn

Infocop. (2024, 6 febrero). *Consecuencias psicológicas, físicas y socioeconómicas de los accidentes de tráfico*. Infocop. https://www.infocop.es/consecuencias-psicologicas-fisicas-y-socioeconomicas-de-los-accidentes-de-

trafico/#:~:text=Los%20accidentes%20de%20tr%C3%A1fico%20pueden,depresi%C3%B3n%20y%20trastorno%20de%20ansiedad.

Kundu, R. (2022). *F1 Score in Machine Learning: Intro & Calculation*. *V7*. https://www.v7labs.com/blog/f1-score-guide