Predicción de zonas de riesgo en rutas de transporte de mercancías a nivel nacional mediante el análisis de incidentes de robo a vehículos

Erney Vargas Mendivelso

Asesor

Andrés Felipe Solis Pino

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Especialización en Ciencia de Datos y Analítica

2026

Dedicatoria

A mi amada esposa Karen Vega y a mi pequeño hijo Isaac, el tesoro más grande de mi vida. Ustedes son mi fuerza, mi inspiración y mi mayor motivo para seguir adelante, incluso en los días más difíciles. Gracias por su amor incondicional, por su paciencia infinita y por comprender los sacrificios que implica dedicar tiempo y esfuerzo a mi crecimiento profesional, aun cuando eso significa restar momentos a nuestra vida en familia. Gracias por acompañarme siempre con tanto cariño en este camino lleno de desafíos

A mis padres y hermanos, gracias por ser mi pilar en cada etapa de este recorrido, por su apoyo constante, por su comprensión y por recordarme, en cada paso, quién soy y de dónde vengo. Son mi refugio, mi raíz y el impulso que me anima a ser una mejor versión de mí mismo cada día.

Agradecimientos

Con gratitud en el corazón, quiero expresar mi agradecimiento a todas las personas e instituciones que, de alguna manera, hicieron posible la realización de este trabajo de grado.

Mi reconocimiento especial es para el profesor Andrés Felipe Solís Pino, cuya orientación, compromiso y apoyo constante marcaron una diferencia significativa durante el proceso investigativo. Su paciencia, disposición para ayudar y actitud siempre positiva fueron fundamentales para el desarrollo y culminación de este proyecto.

A mis seres queridos, en especial a mi familia, les agradezco profundamente por su compañía incondicional. Su apoyo, comprensión y aliento han sido relevantes en cada etapa de este recorrido académico. Gracias por creer en mí, incluso cuando las fuerzas flaqueaban.

Finalmente, agradezco sinceramente a todos aquellos que, directa o indirectamente, aportaron su tiempo, conocimientos o palabras de ánimo. A cada uno de ustedes, les guardo un lugar especial en este logro.

Resumen

El hurto de vehículos en Colombia representa una amenaza crítica para la seguridad ciudadana y la eficiencia logística, generando pérdidas económicas significativas. A pesar de la disponibilidad de datos históricos, la adopción de herramientas avanzadas para la anticipación del delito sigue siendo limitada en los sectores afectados. Este proyecto desarrolla un modelo predictivo basado en machine learning para identificar patrones y zonas de riesgo, utilizando el conjunto de datos abiertos de la Policía Nacional de Colombia.

La investigación se fundamentó en la metodología estándar CRISP-DM, abarcando desde la comprensión del negocio hasta el diseño del despliegue. Mediante técnicas de ingeniería de características, se transformaron variables geográficas y temporales para entrenar y validar diversos algoritmos de aprendizaje supervisado. Los resultados obtenidos alcanzaron una exactitud (accuracy) del 86.30 % y un Área Bajo la Curva (AUC) superior a 0.90, lo que valida la capacidad del sistema para clasificar eficazmente los niveles de riesgo. Finalmente, se propone una estrategia de implementación que permite a las organizaciones de transporte y autoridades utilizar estas predicciones como soporte para la optimización de recursos y la toma de decisiones operativas.

Palabras claves: Análisis de datos; transporte de mercancías; hurto a vehículos; modelado predictivo; seguridad en transporte.

Abstract

Vehicle theft in Colombia represents a critical threat to public safety and logistical efficiency, resulting in significant economic losses. Despite the availability of historical data, the adoption of advanced analytical tools for crime anticipation remains limited across affected sectors. This project develops a predictive model based on machine learning to identify patterns and high-risk zones, utilizing the National Police of Colombia's open dataset as the primary source.

The research was conducted under the CRISP-DM standard methodology, spanning from business understanding to deployment design. Through feature engineering techniques, geographical and temporal variables were transformed to train and validate various supervised learning algorithms. The results achieved an accuracy of 86.30% and an Area Under the Curve (AUC) exceeding 0.90, validating the system's capability to effectively classify risk levels. Finally, an implementation strategy is proposed, enabling transportation entities and authorities to utilize these predictions as a support tool for resource optimization and operational decision-making

Keywords: Data analysis; freight transport; vehicle theft; predictive modeling; transport security.

Tabla de Contenido

Lista de Figuras

Lista de Tablas

Introducción

El hurto de mercancías durante el transporte terrestre se ha posicionado como una problemática crítica dentro del contexto logístico colombiano, afectando de manera directa la seguridad vial, la competitividad empresarial y la estabilidad operativa del sector. Este fenómeno no solo genera cuantiosas pérdidas económicas para las empresas transportadoras, derivadas del valor de los bienes sustraídos y los daños a la infraestructura vehicular, sino que también representa una amenaza latente para la integridad física de los conductores, incrementando la percepción de inseguridad en las vías nacionales.

A pesar de la magnitud e impacto del problema, el país carece de herramientas tecnológicas avanzadas que permitan anticipar de manera efectiva las zonas y periodos de mayor riesgo. La limitada disponibilidad de sistemas de análisis predictivo, sumada a deficiencias en la recolección, integración y procesamiento de datos delictivos, ha restringido la capacidad de respuesta de los actores involucrados, dificultando la formulación de estrategias preventivas sostenibles y eficaces. Asimismo, la débil articulación entre instituciones estatales, aseguradoras y empresas del sector ha obstaculizado el desarrollo de iniciativas colaborativas orientadas a la mitigación del riesgo.

El presente estudio tiene como objetivo diseñar e implementar un sistema predictivo fundamentado en técnicas de ciencia de datos, orientado a la identificación de zonas geográficas con alta probabilidad de ocurrencia de hurtos a vehículos. Para ello, se emplea la metodología CRISP-DM como marco estructural del proceso, aplicado al conjunto de datos “Hurto a vehículos” proveniente del portal de datos abiertos del Gobierno de Colombia. Posteriormente, los datos son almacenados en una base de datos no relacional MongoDB, facilitando su

integración con algoritmos de aprendizaje no supervisado, los cuales permiten capturar patrones espaciotemporales asociados a la ocurrencia del delito.

Con este enfoque, se pretende contribuir al fortalecimiento de las capacidades analíticas del sector transporte, facilitando la toma de decisiones estratégicas en materia de seguridad vial, planificación operativa y gestión de riesgos. De esta manera, el estudio busca aportar una herramienta innovadora que responda a una necesidad latente en la industria logística colombiana y que, a su vez, promueva entornos más seguros y eficientes para el desarrollo del transporte en el país.

Justificación

El presente trabajo de grado se enfoca en la necesidad crítica que enfrenta el sector transporte de carga en Colombia de contar con herramientas predictivas que permitan anticipar, con un nivel razonable de precisión, las zonas y periodos de mayor riesgo asociados al hurto de vehículos de carga en el territorio nacional. ¿De dónde radica esta necesidad? El transporte terrestre representa más del 80 % de la movilización de mercancías en el país, siendo un eslabón fundamental para la competitividad, la eficiencia logística y el desarrollo económico. Sin embargo, este sector enfrenta una amenaza persistente: el hurto sistemático de vehículos y mercancías, que según cifras de la Policía Nacional y reportes académicos como el de Medina-Hurtado et al. (2017), ha ocasionado pérdidas por más de $230.000 millones anuales, sin incluir daños colaterales ni costos indirectos derivados. A pesar de esta situación, las empresas transportadoras carecen de mecanismos tecnológicos efectivos que permitan predecir el comportamiento delictivo con base en datos históricos y variables contextuales. Actualmente, el análisis del riesgo logístico se realiza de forma reactiva, sin incorporar modelos que permitan prever eventos de alto impacto o anticipar tendencias delictivas con base en patrones espaciales y temporales.

Es así como, a través de este estudio, se buscó analizar el comportamiento histórico de los hurtos a vehículos de forma individual. El objetivo es sentar las bases para un modelo predictivo que aporte a la discusión del sector un enfoque prospectivo, alineando estrategias empresariales y políticas públicas con información confiable, de manera que se pueda mitigar el impacto de estos delitos, optimizar la asignación de recursos en seguridad y fortalecer las capacidades de respuesta ante amenazas logísticas en el país.

Considerando que los objetivos de esta investigación se centran en la seguridad de las rutas de transporte, es imperativo aclarar que la unidad de análisis predictivo es el municipio. El municipio actúa como un proxy geográfico-administrativo que permite caracterizar el entorno de seguridad. Para una empresa de logística, la traducción de este riesgo municipal a una 'ruta segura' se realiza mediante la evaluación del trayecto completo: si una ruta atraviesa cinco municipios, el nivel de riesgo asignado a la ruta será igual al valor máximo de riesgo detectado en cualquiera de los municipios del trayecto. Este enfoque conservador permite a los jefes de tráfico identificar el 'eslabón más débil' de la cadena logística y activar protocolos de seguridad (como escoltas o monitoreo reforzado) precisamente en los segmentos de mayor vulnerabilidad.

Objetivos

Objetivo General

Desarrollar un sistema de predicción de zonas de alto riesgo en rutas de transporte de mercancías a nivel nacional, empleando la metodología CRISP-DM, a partir de datos sobre hurtos de vehículos, para apoyar la mitigación de esta problemática social mediante herramientas tecnológicas de ciencia de datos.

Objetivos Específicos

Construir una base de datos consolidada a partir de registros gubernamentales abiertos, realizando preprocesamiento de datos para facilitar el análisis de factores asociados al riesgo en rutas de transporte

Implementar modelos predictivos de riesgo de hurto en rutas de transporte mediante técnicas de aprendizaje automático, siguiendo la metodología CRISP-DM.

Validar el sistema desarrollado mediante una prueba de concepto utilizando un conjunto de datos reciente, con el fin de evaluar su aplicabilidad, precisión y utilidad en un entorno real.

Marco Conceptual

El hurto de vehículos de carga en Colombia constituye un fenómeno delictivo complejo que impacta de manera directa la eficiencia logística, la seguridad vial y la competitividad del sector transporte. Este problema ha sido abordado desde diversas perspectivas, tanto en la literatura académica como en las políticas de seguridad implementadas por el Estado, dando lugar a un conjunto de investigaciones que permiten comprender sus múltiples dimensiones y ofrecen herramientas metodológicas útiles para su análisis y prevención.

Desde el enfoque criminológico, Norza Céspedes et al. (2013) ofrecen una visión integral del hurto de automotores en Colombia, considerando el rol del victimario, la respuesta institucional de la Policía y el análisis académico. Este marco resulta clave para entender la interacción entre los actores involucrados y las dinámicas delictivas, lo cual aporta insumos relevantes para la formulación de modelos predictivos.

En términos de impacto económico, Medina-Hurtado et al. (2017) destacan las pérdidas significativas que el hurto vehicular genera para el sector transporte, subrayando la urgencia de desarrollar herramientas cuantitativas que permitan anticipar eventos delictivos y mitigar sus consecuencias. Tobón & Galvis (2009), por su parte, exploran la evolución del sector transporte y sugieren que los cambios en infraestructura y políticas públicas pueden influir en la ocurrencia de estos delitos.

Diversas investigaciones han abordado el papel de actores específicos en la prevención del hurto. Escobedo et al. (2008) analizan la participación de los taxistas en Bogotá como fuente de información para la seguridad urbana, mientras que Osorio et al. (2017) discuten la gestión del riesgo en proveedores logísticos, resaltando cómo su fortalecimiento puede reducir la vulnerabilidad operativa frente al crimen organizado.

El uso de técnicas estadísticas y herramientas de ciencia de datos ha cobrado relevancia en el análisis delictivo. Cruz Reyes (2020) implementa modelos gráficos probabilísticos para el estudio espacial del hurto de celulares, metodología que puede adaptarse al análisis del hurto de vehículos. Asimismo, López-Herrera et al. (2019) aplican regresión lineal múltiple y regresión ponderada geográficamente al estudio de hurtos en Colombia, aportando estrategias estadísticas para detectar patrones geoespaciales.

En el ámbito del análisis de series temporales y umbrales críticos, Duarte Velásquez & Cadavid Carmona (2020) proponen un enfoque diferencial para la interpretación de registros de criminalidad, útil para detectar fluctuaciones en la incidencia del hurto de vehículos. A nivel metodológico, estudios como el de Pineda Nobles (2021) plantean cómo la inteligencia artificial y el Big Data pueden ser utilizados dentro del marco legal penal, estableciendo un precedente para la integración ética de estas tecnologías en la prevención del delito.

La literatura reciente también ha explorado técnicas emergentes aplicadas al análisis delictivo. Contreras Contreras et al. (2022) desarrollan una metodología de agrupamiento basada en aprendizaje automático, útil para identificar clústeres delictivos. De la Hoz-Dominguez et al. (2020) aplican técnicas de análisis geoespacial para caracterizar zonas de hurto en Cartagena, lo cual valida la pertinencia de enfoques territoriales y basados en datos.

El análisis de grandes volúmenes de información también ha sido abordado por Fernandez-Morales & Bonilla-Carrión (2020), quienes destacan la utilidad de la bibliominería en la toma de decisiones estratégicas. Por su parte, Hernández-Leal et al. (2017) y Torres-Domínguez et al. (2019) profundizan en el papel del Big Data en la detección de anomalías y vigilancia urbana, sentando las bases para sistemas de alerta temprana.

En relación con la seguridad vial, Iván-Herrera-Herrera et al. (2018) proponen herramientas para la toma de decisiones en contextos de congestión vehicular, mientras que Velásquez-Monroy (2011) reflexiona sobre la sostenibilidad de las estrategias de reducción del hurto mediante atracos. Estas perspectivas se complementan con los aportes de Ibarra Padilla et al. (2021), quienes analizan la evolución de la política criminal frente al hurto en Colombia y su impacto en la reducción de este tipo de delitos.

Duran-Romero et al. (2020) exploran cómo la gestión integral de seguridad empresarial en empresas logísticas puede ser una estrategia complementaria en la prevención del hurto vehicular, integrando la protección de activos con la analítica de datos.

El conjunto de estudios revisados aporta un marco teórico robusto que sustenta el enfoque del presente proyecto. La combinación de enfoques criminológicos, análisis económico, técnicas estadísticas avanzadas y metodologías de ciencia de datos fortalece la pertinencia del desarrollo de un sistema predictivo de zonas de riesgo. Además, la revisión evidencia la necesidad de integrar variables contextuales y socioeconómicas para comprender de forma integral las causas estructurales del hurto vehicular. Este sustento bibliográfico justifica la aplicación de metodologías como CRISP-DM y el uso de modelos de aprendizaje supervisado para la construcción de herramientas tecnológicas al servicio de la seguridad logística.

Metodología

La presente investigación adopta un enfoque cuantitativo, estructurado bajo el marco de trabajo CRISP-DM (Cross Industry Standard Process for Data Mining). Esta metodología, estándar en la industria de la ciencia de datos, permite sistematizar el desarrollo del proyecto en seis fases iterativas: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue, garantizando así la reproducibilidad y el rigor científico del estudio.

Adquisición y preparación de datos para la construcción del conjunto de datos; se utilizaron fuentes oficiales del Gobierno de Colombia, específicamente el portal de datos abiertos (www.datos.gov.co). La extracción de información se centró en los registros históricos de hurtos, consolidando un repositorio robusto para el análisis. Posteriormente, se ejecutó una fase exhaustiva de preprocesamiento e ingeniería de características (feature engineering), que incluyó:

Limpieza: Tratamiento de valores nulos, estandarización de formatos de fecha y eliminación de duplicados.

Transformación: Generación de variables derivadas espaciotemporales (Mes, Día de la Semana, Año) y creación de indicadores de riesgo histórico (Promedio de Hurtos y Variabilidad).

Almacenamiento: Los datos procesados fueron persistidos en una base de datos NoSQL MongoDB. Esta tecnología fue seleccionada por su esquema flexible, ideal para gestionar la variabilidad de los registros y permitir una integración eficiente con el entorno de modelado en Python.

Estrategia de Modelado y Análisis: El componente analítico se fundamentó en técnicas de aprendizaje automático supervisado. Tras un análisis exploratorio para identificar correlaciones (como la relación entre jerarquía administrativa y criminalidad), se procedió al entrenamiento de algoritmos de clasificación.

Validación y Métricas: Para la evaluación del sistema, se aplicó una estrategia de validación cruzada y división de datos (Train-Test Split), reservando un subconjunto de prueba para medir la capacidad de generalización del algoritmo. El desempeño del modelo se cuantificó mediante métricas rigurosas, priorizando la técnica de la matriz de confusión y el Área Bajo la Curva (AUC) para medir la capacidad discriminativa, junto con la Exactitud (Accuracy) y el F1-Score, asegurando así una evaluación integral de la robustez y utilidad operativa de la solución propuesta.

Comprensión del Negocio (Business Understanding)

En esta fase inicial, se establece el contexto estratégico del proyecto, definiendo la problemática desde la perspectiva de la seguridad pública y la logística nacional. El enfoque se centra en comprender la dinámica del hurto de automotores no solo como una estadística delictiva, sino como un fenómeno que requiere intervención inteligente.

Actualmente, la gestión de este delito en Colombia enfrenta un desafío de asimetría de información: las autoridades poseen los datos históricos, pero carecen de mecanismos automatizados para explotar dicha información en tiempo real. Por consiguiente, el propósito de esta fase es delimitar el alcance del estudio, asegurando que la solución tecnológica propuesta responda directamente a las necesidades operativas de prevención y control.

Objetivos del Negocio

El objetivo principal de este proyecto es desarrollar una solución analítica que permita identificar y predecir patrones de hurto de vehículos en Colombia. La organización (en este caso, representada por el contexto de seguridad nacional) requiere pasar de un enfoque reactivo a uno preventivo.

Específicamente, se busca apoyar la toma de decisiones estratégicas mediante la anticipación de zonas y periodos de alto riesgo. Dado el impacto económico y social de este delito, la herramienta permitirá optimizar la asignación de recursos policiales y focalizar las estrategias de control en los puntos críticos detectados por el modelo.

Evaluación de la Situación y Recursos

Tabla 1

Inventario de Recursos y Herramientas Tecnológicas

Recurso / Herramienta	Descripción / Uso en el proyecto
Fuente de Datos	Datos abiertos (Datos.gov.co) reportes de la Policía Nacional.
Lenguaje	Python 3.x (librerías Pandas, Scikit-Learn).
Base de Datos	MongoDB (Almacenamiento NoSQL para flexibilidad de esquemas)
Hardware	Equipo local con procesador Intel Core i7, 24 GB RAM.
IDE	Pycharm

Para la ejecución del proyecto, se cuenta con acceso a repositorios de datos abiertos gubernamentales. Se ha identificado que, aunque la información es accesible, presenta desafíos de calidad (inconsistencias de formato) que requerirán una fase exhaustiva de preparación de datos.

Objetivos de Minería de Datos

Para satisfacer los objetivos del negocio, se establecen los siguientes objetivos técnicos de minería de datos:

*Consolidación***:** Estructurar un repositorio de datos unificado y limpio proveniente de fuentes históricas.

*Identificación de patrones***:** Determinar la correlación entre variables temporales (día, mes) y geográficas con la frecuencia del delito.

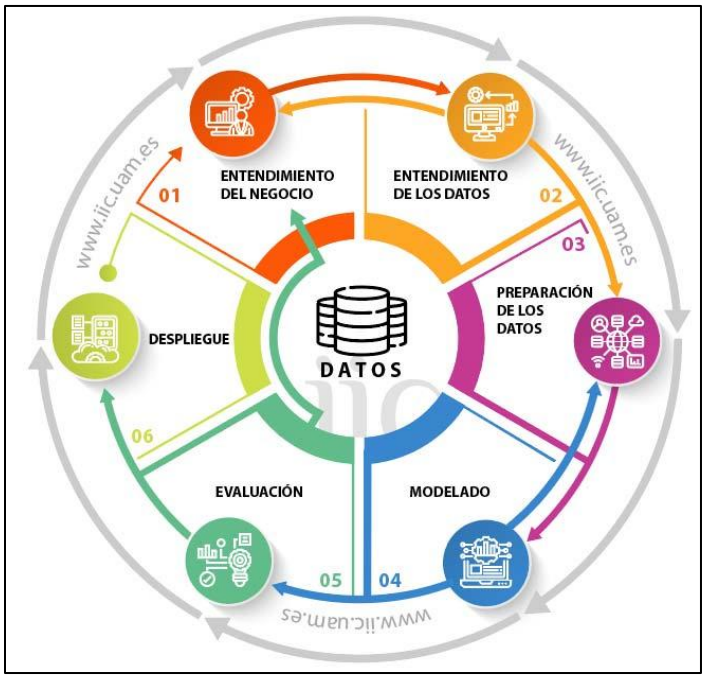
*Modelado***:** Entrenar y validar algoritmos de aprendizaje supervisado (como Random Forest, Regresión Logística y XGBoost) para predecir la categoría de riesgo de una zona.

Evaluación: Lograr una métrica de desempeño (ej. *Accuracy* o *Recall*) superior al 80 % en el conjunto de prueba.

Plan del Proyecto

Figura 1

Metodología CRISP-DM



Nota: Representación visual de las diferentes etapas del modelo CRISP-DM. Tomado de https://www.iic.uam.es/innovacion/metodologia-crisp-dm-ciencia-de-datos/

El desarrollo se regirá por las seis fases iterativas de la metodología CRISP-DM, permitiendo retroalimentación constante entre la comprensión de los datos y el modelado, tal como se ilustra en la Figura 1.

Conclusión de la Fase de Comprensión del Negocio

Como cierre de esta fase inicial, se ha logrado traducir una problemática social compleja (la inseguridad vial y el hurto de activos) en un problema técnico estructurado y viable. La

evaluación de recursos confirma que se cuenta con la infraestructura computacional, el stack tecnológico (Python/MongoDB) y, lo más importante, el acceso y el guardado de los datos necesarios para abordar el desafío.

Se concluye que el proyecto es factible, dado que los objetivos de minería de datos son medibles y están alineados con la estrategia de prevención de la organización. Sin embargo, se identifica como riesgo principal la calidad de la información bruta (inconsistencias), lo cual justifica la transición inmediata a la segunda fase de la metodología: Comprensión de los datos. En dicha etapa, el foco se desplazará de la planificación estratégica a la exploración técnica, para validar si la materia prima (los datos) es suficiente para alcanzar la métrica de desempeño propuesta del 80 %.

Comprensión de los Datos (Data Understanding)

Esta fase tiene como propósito realizar una exploración técnica del conjunto de datos para validar su calidad, comprender las distribuciones estadísticas subyacentes e identificar patrones preliminares que orienten la estrategia de modelado.

Descripción del Conjunto de Datos (Metadatos)

La fuente primaria de información corresponde al conjunto de datos abiertos titulado "HURTO A VEHÍCULOS", dispuesto por la Policía Nacional de Colombia a través del portal www.datos.gov.co. El archivo original, extraído en formato CSV con codificación UTF-8, consolida los registros de denuncias a nivel nacional con una ventana de observación que abarca desde el año 2010 hasta junio de 2025.

El conjunto de datos bruto consta de 372,506 registros. A continuación, en la Tabla 2, se presenta el diccionario de datos inicial con la tipología de las variables desglosadas en dicha tabla.

Tabla 2

Descripción Conjuntos de Datos

Campo	Tipo Dato Python	descripción	Tipo Estadístico
FECHA_HECHO	Object	Fecha del incidente (dd/mm/aaaa)	Variable Temporal
COD_DEPTO	Int64	Identificador numérico del departamento (Divipola)	Categórica Nominal (Codificada)
DEPARTAMENTO	Object	Nombre oficial del departamento.	Categórica
COD_MUNI	Int64	Identificador numérico del municipio.	Categórica Nominal (Codificada)

Campo	Tipo Dato Python	descripción	Tipo Estadístico
MUNICIPIO	Object	Nombre oficial del municipio.	Categórica
ZONA	Object	Clasificación del lugar (Urbana/Rural).	Categórica Nominal
TIPO_DELITO	Object	Modalidad (Hurto a Motocicletas/Automotores).	Categórica Nominal
CANTIDAD	Int64	Número de bienes hurtados en el evento.	Variable Objetivo (Base)

Nota. Se identifica que el conjunto contiene 3 variables numéricas y 5 categóricas. Es importante destacar que la variable CANTIDAD servirá como base para la construcción de la variable objetivo de clasificación (CLASE_RIESGO) en la fase de preparación, en la cual se aplicarán técnicas de clustering para obtener esta clasificación de riesgo.

Análisis Univariado

Se examinó la distribución individual de las variables para detectar sesgos, valores atípicos (outliers) y errores de captura.

Análisis de Variables Categóricas

Para las variables cualitativas, se excluyó temporalmente FECHA_HECHO, la cual será objeto de una ingeniería de características específica para extraer componentes de estacionalidad (año, mes, día) en la fase siguiente.

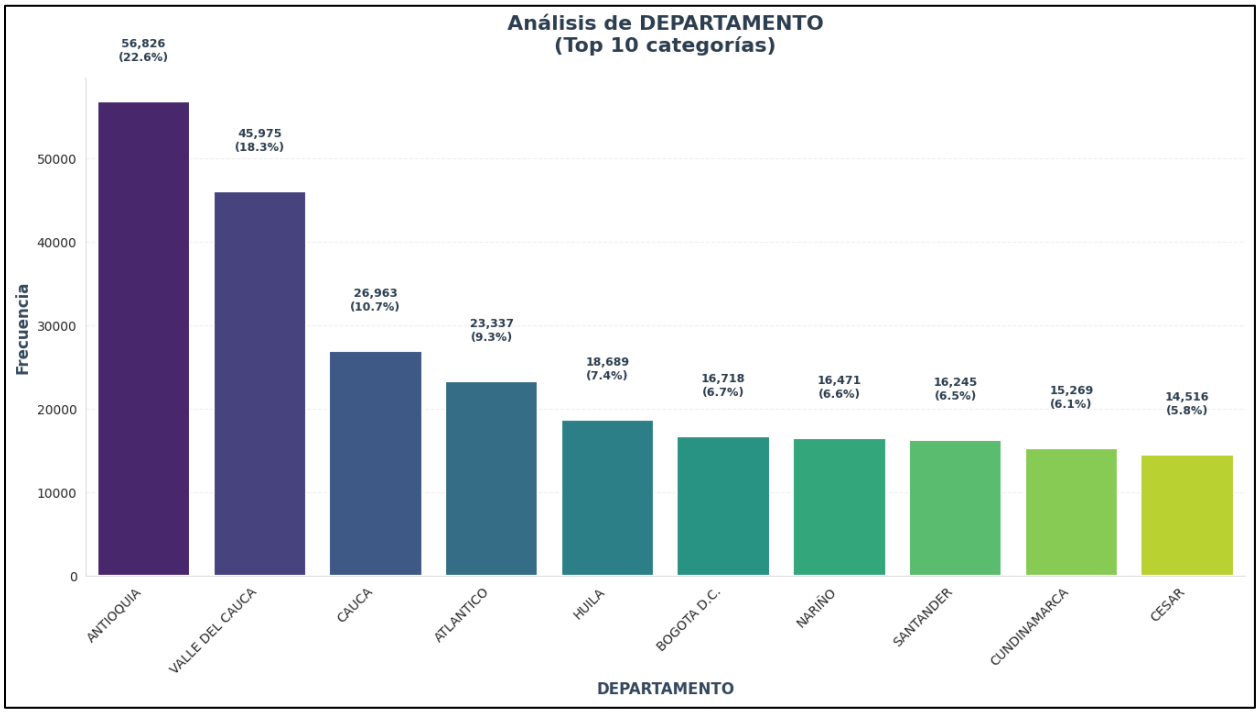
Posteriormente, se realizó la representación gráfica de las variables seleccionadas con el propósito de comprender de manera más precisa su comportamiento y distribución dentro del conjunto de datos.

Distribución Geográfica Departamentos y Municipios

El análisis gráfico evidencia una fuerte concentración del delito en las regiones del noroccidente y suroccidente del país. Departamentos como Antioquia, Valle del Cauca y Cundinamarca lideran las estadísticas. A nivel municipal, se observa que las capitales principales (Cali, Bogotá, Medellín) actúan como focos de alta densidad delictiva, confirmando que el fenómeno está correlacionado con la densidad poblacional y la actividad económica urbana.

Figura 2

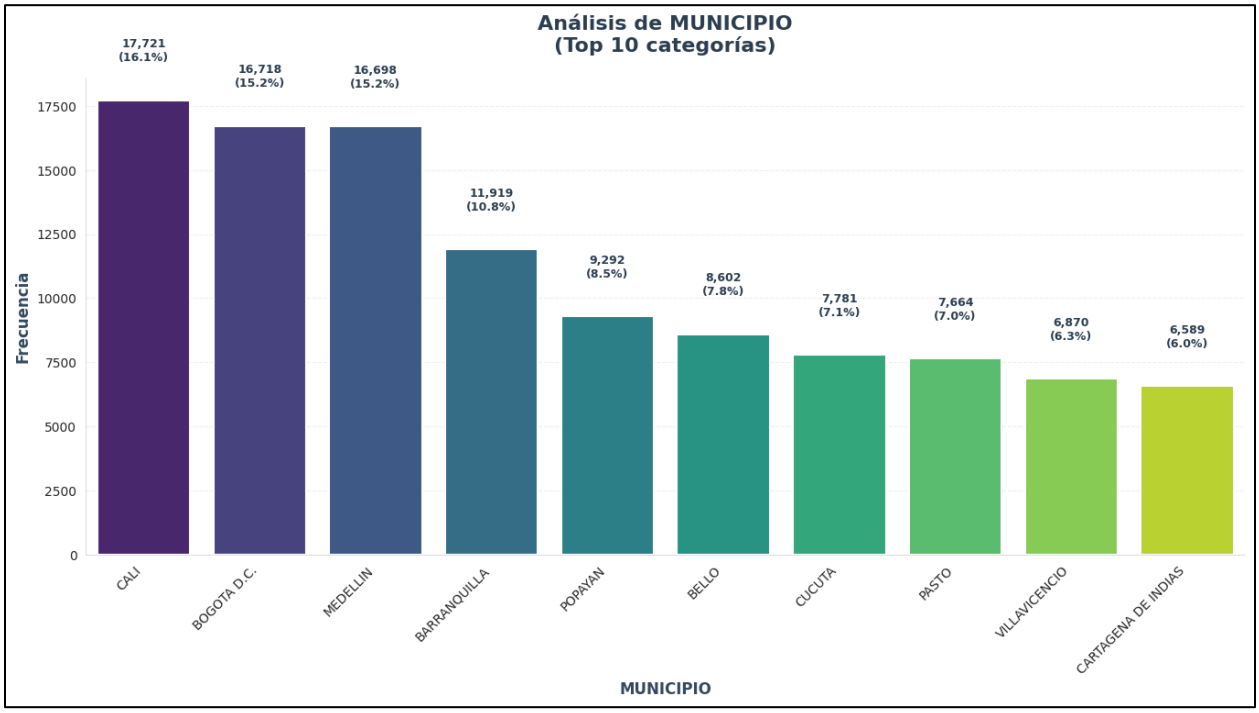
Top 10 Departamentos con Más Hurtos de Vehículos



Nota. Se puede apreciar en la gráfica que este tipo de delitos ocurre en departamentos del noroccidente y suroccidente de Colombia, como Antioquia, Valle del Cauca y Cauca.

Figura 3

Top 10 Municipios con Más Hurtos de Vehículos



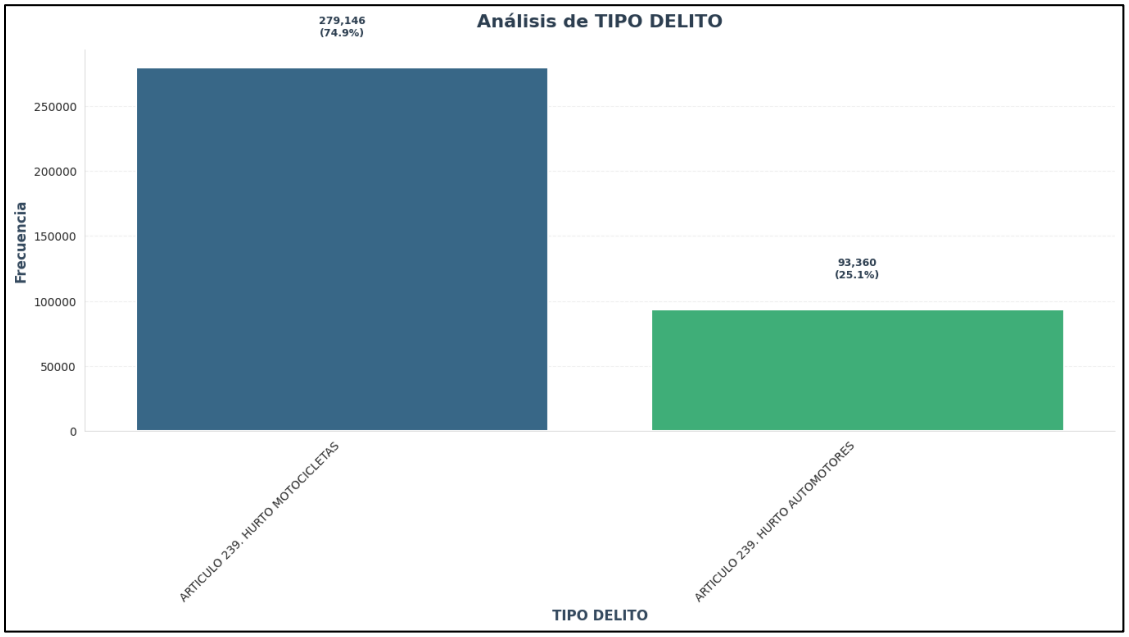
Nota. Se puede apreciar en la gráfica que este tipo de delitos ocurre en los municipios de Cali, Bogotá, Medellín y Barranquilla, ciudades con un alto número de habitantes

Modalidad del Delito (Tipo de Vehículo)

Existe un desbalance significativo en la tipología del bien sustraído. El hurto de motocicletas representa aproximadamente el 75 % de los casos, triplicando la incidencia del robo de automotores. Este hallazgo sugiere que las motocicletas son el activo de mayor vulnerabilidad y liquidez en el mercado ilegal.

Figura 4

Conteo por Tipo de Delito

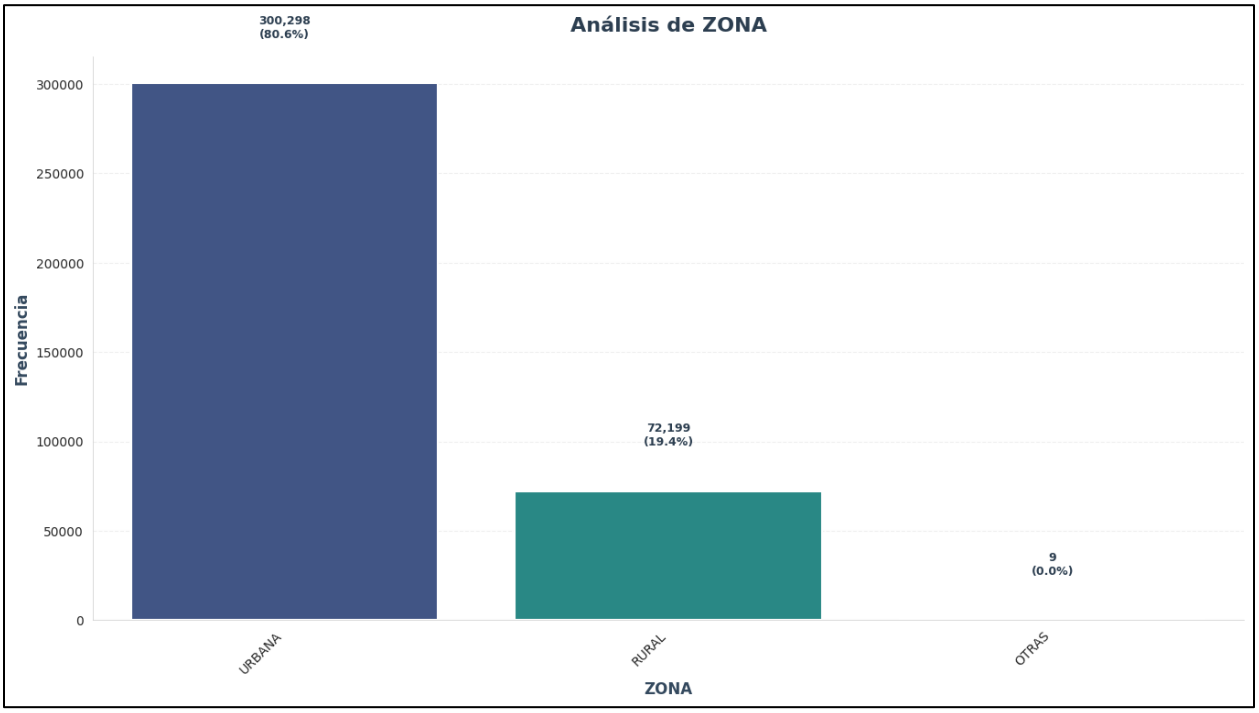


Distribución por Zona

El 80 % de los eventos se materializan en zonas urbanas, mientras que la zona rural presenta una incidencia menor. Se detectó una categoría denominada “OTRAS” con una representatividad marginal (0.0 %), la cual será depurada en la fase de limpieza de datos para evitar ruido en el modelo.

Figura 5

Conteo por Zona



Análisis de Variables Numéricas

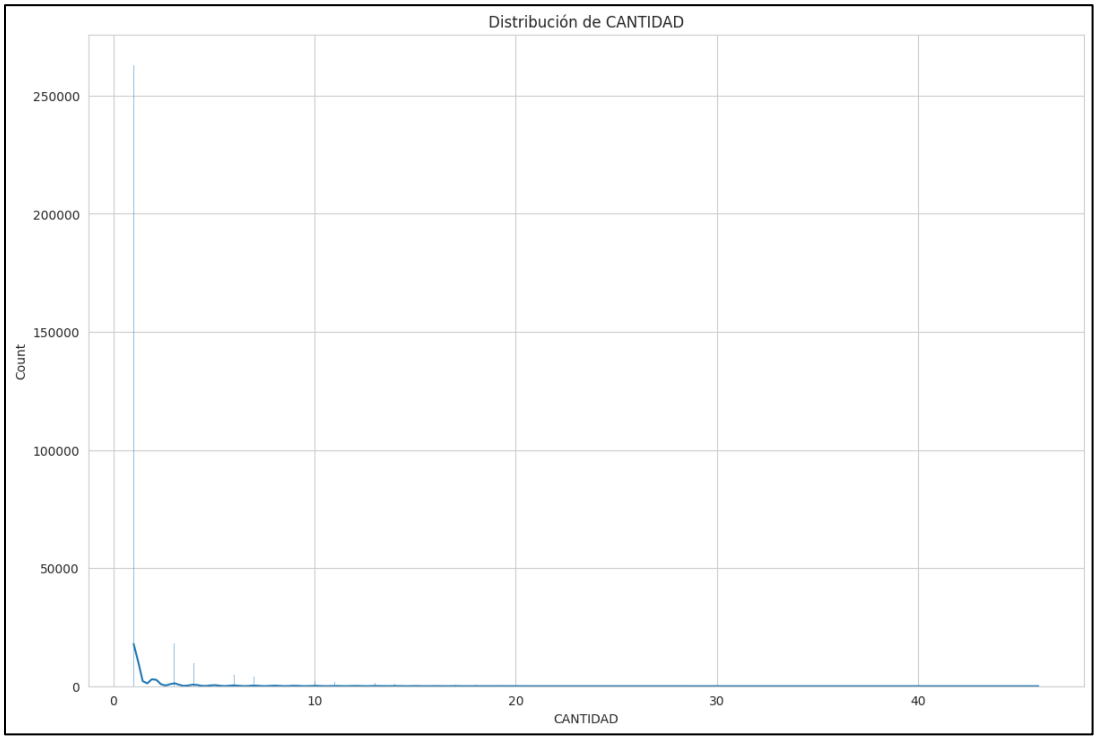
La única variable cuantitativa de magnitud es CANTIDAD. Las variables COD_DEPTO y COD_MUNI, aunque numéricas en formato, representan códigos administrativos y no magnitudes matemáticas.

Estadística Descriptiva de **'CANTIDAD'**:

- Media: 2.08
- Mediana (50 %): 1.00
- Desviación estándar: 2.81
- Máximo: 46

Figura 6

Histograma de la Variable Cantidad



*Nota.*El histograma revela una distribución de cola larga (asimetría positiva). El 50 % de los registros corresponde a eventos de un solo hurto (mediana = 1), y el 75 % de los casos no supera los 2 hurtos. Sin embargo, la existencia de valores extremos (hasta 46 hurtos en un solo registro) indica la presencia de eventos masivos o atípicos. Esta dispersión justifica la decisión metodológica de no predecir el número exacto, sino clasificar el riesgo, ya que un modelo de regresión lineal sería muy sensible a estos valores extremos.

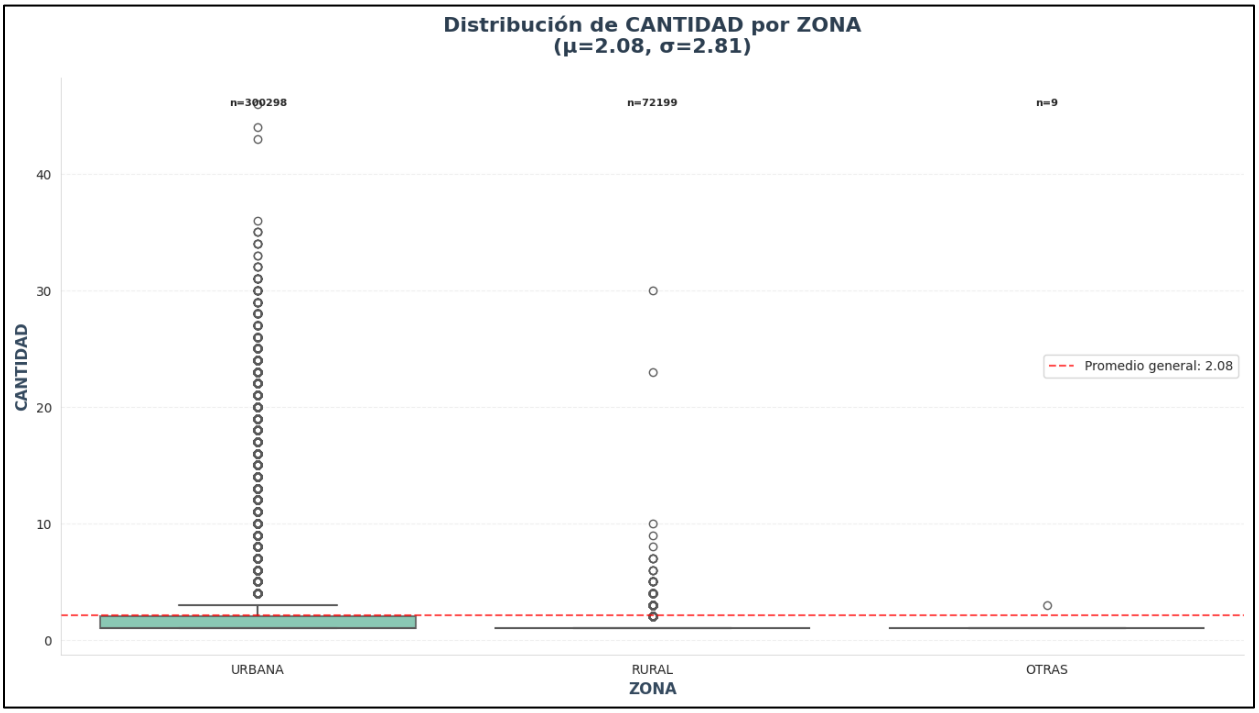
Análisis Bivariado

En esta etapa se exploraron las relaciones entre variables predictoras y la variable objetivo para validar hipótesis de negocio.

Los diagramas de caja (boxplots) muestran que la Zona Urbana no solo tiene más delitos, sino mayor variabilidad. Es aquí donde ocurren los eventos masivos (valores atípicos superiores a 10). La zona rural, por el contrario, presenta un comportamiento más homogéneo y de menor escala.

Figura 7

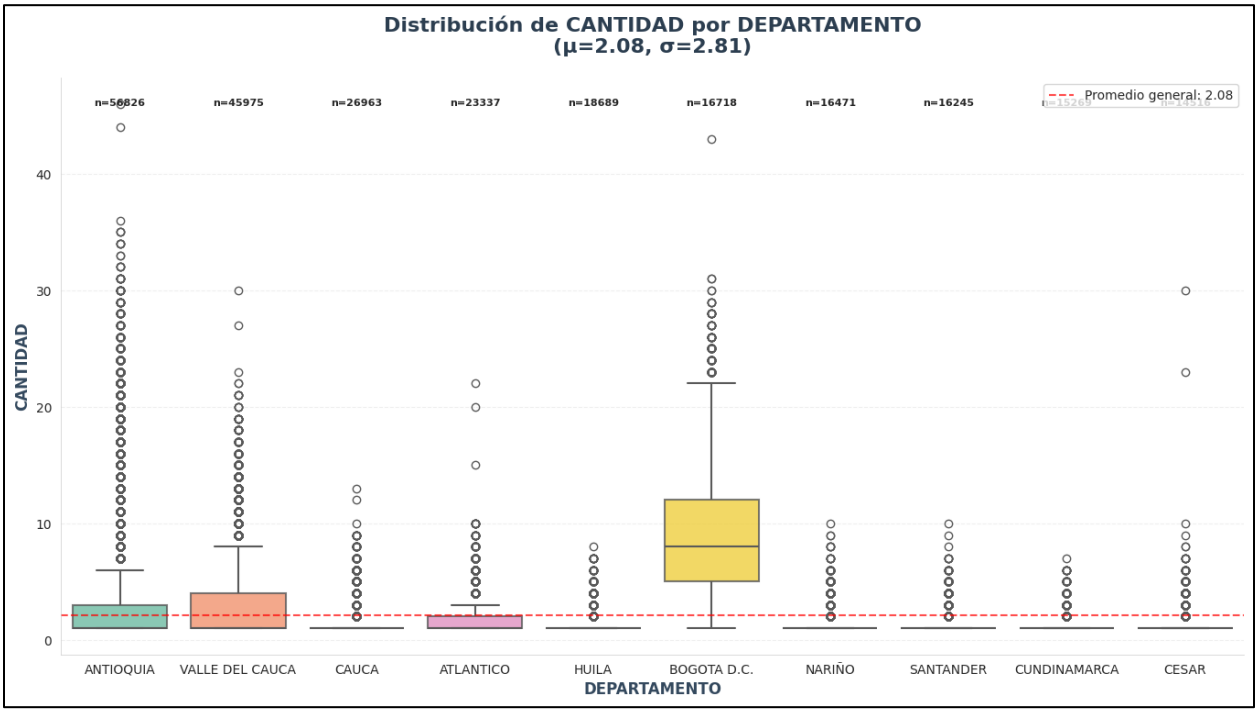
Gráfico de Cajas de Cantidad por Zona



Al analizar el Top 10 de departamentos, se destaca el comportamiento de Bogotá D.C., que exhibe el rango intercuartílico (IQR) más amplio. Esto implica que la capital no tiene un comportamiento constante; experimenta días de calma y días de picos delictivos extremos. En contraste, departamentos como Cauca o Huila muestran “cajas” más compactas, sugiriendo una criminalidad más constante pero menos propensa a eventos masivos repentinos.

Figura 8

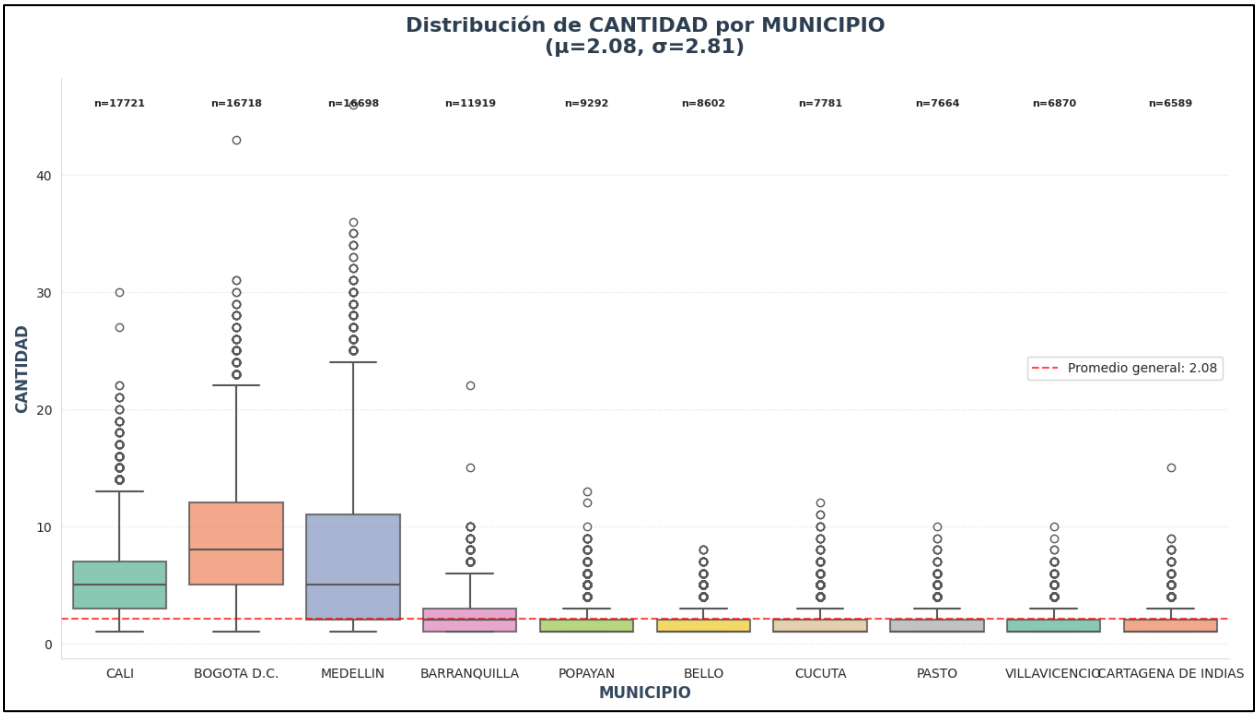
Gráfico de Cajas de Departamento y Cantidad de Hurtos de Vehículos



El análisis de cantidad de hurtos vs. el top 10 de municipios nos da a entender que Bogotá D.C., Medellín y Cali concentran la mayor cantidad de robos de vehículos y presentan la mayor variabilidad, con numerosos valores atípicos que indican picos ocasionales de alta criminalidad. Barranquilla muestra una incidencia intermedia, generalmente baja, pero con eventos excepcionales que elevan los registros. En contraste, Popayán, Bello, Cúcuta, Pasto, Villavicencio y Cartagena de Indias presentan medianas bajas y distribuciones más estables, aunque no están exentos de incrementos puntuales. En conjunto, la gráfica confirma que los grandes centros urbanos son los principales focos de riesgo y que la distribución de los robos es asimétrica, con presencia de eventos extremos.

Figura 9

Gráfico de Cajas de Municipios y Cantidad de Hurtos de Vehículos



Conclusión Compresión de los Datos

El análisis confirma que el hurto de vehículos en Colombia es un fenómeno heterogéneo y altamente concentrado en entornos urbanos y capitales. La presencia de valores atípicos en la variable CANTIDAD y el desbalance en el tipo de vehículo (dominancia de motos) son factores críticos que deberán ser gestionados mediante técnicas de balanceo de clases y transformación de variables en la siguiente fase de Preparación de Datos.

Preparación de los Datos

Esta fase comprende todas las actividades necesarias para construir el conjunto de datos final (la vista minable) a partir de los datos brutos iniciales. Las tareas incluyen la selección de atributos, limpieza, construcción de nuevos atributos y transformación de formatos.

Limpieza de Datos (Data Cleaning)

Esta fase es considerada la más intensiva en términos de esfuerzo computacional y operativo. Su objetivo es transformar los datos brutos en un conjunto de datos estructurado y limpio que sea apto para el entrenamiento de modelos matemáticos. Se aplicaron técnicas de limpieza, normalización e ingeniería de características.

Manejo de Valores Nulos (Imputación)

Complementario a la detección estándar de valores nulos mediante el método isnull().sum() de Pandas, se implementó una validación exhaustiva orientada a las variables categóricas. Este procedimiento rastreó inconsistencias como cadenas de texto vacías (“”) o compuestas únicamente por espacios en blanco (“ “), las cuales pueden ser interpretadas erróneamente como categorías válidas por los algoritmos de machine learning.

El análisis ratificó la integridad del conjunto de datos, confirmando que cada registro contiene información sustantiva, por lo que se descartó la necesidad de aplicar técnicas de imputación o tratamiento de datos faltantes.

Eliminación de Duplicados

Como parte del protocolo de aseguramiento de la calidad del dato, se ejecutó un procedimiento de detección de duplicados sobre la totalidad del conjunto de datos. Se evaluaron las [N] filas disponibles buscando coincidencias exactas en todos los atributos. El análisis arrojó como resultado cero (0) registros duplicados, confirmando que cada entrada en el conjunto de

datos corresponde a un evento único y diferenciado. En consecuencia, no fue necesario aplicar técnicas de eliminación de redundancia, preservando así la integridad dimensional original de la fuente de datos.

Corrección de Inconsistencias

Tras inspeccionar las variables categóricas en busca de errores de escritura en los nombres de municipios y departamentos, no se encontraron anomalías. Los valores se presentan estandarizados, por lo que no fue necesario implementar procesos de limpieza o corrección manual para estas columnas. Para el caso de las variables geográficas como municipio y departamento, se intuye que la fuente de datos original probablemente ya se encuentra alineada con la codificación estándar del DANE (División Político-Administrativa de Colombia), facilitando su integración directa en caso de ser necesario.

Se identificó un patrón de redundancia en la variable descriptiva del delito, la cual incluía sistemáticamente el prefijo legal ‘ARTÍCULO 239’. Dado que esta referencia normativa no aporta valor semántico diferenciador para el modelo predictivo y genera ruido en el procesamiento de texto, se procedió a su remoción mediante expresiones regulares, conservando únicamente la descripción fáctica de la conducta. No se encontraron inconsistencias en las demás variables descriptivas.

Ingeniería de Características (Feature Engineering)

A partir del atributo FECHA_HECHO, se realizó un proceso de ingeniería de características para extraer nuevas variables que permitan al modelo capturar la estacionalidad y los ciclos delictivos:

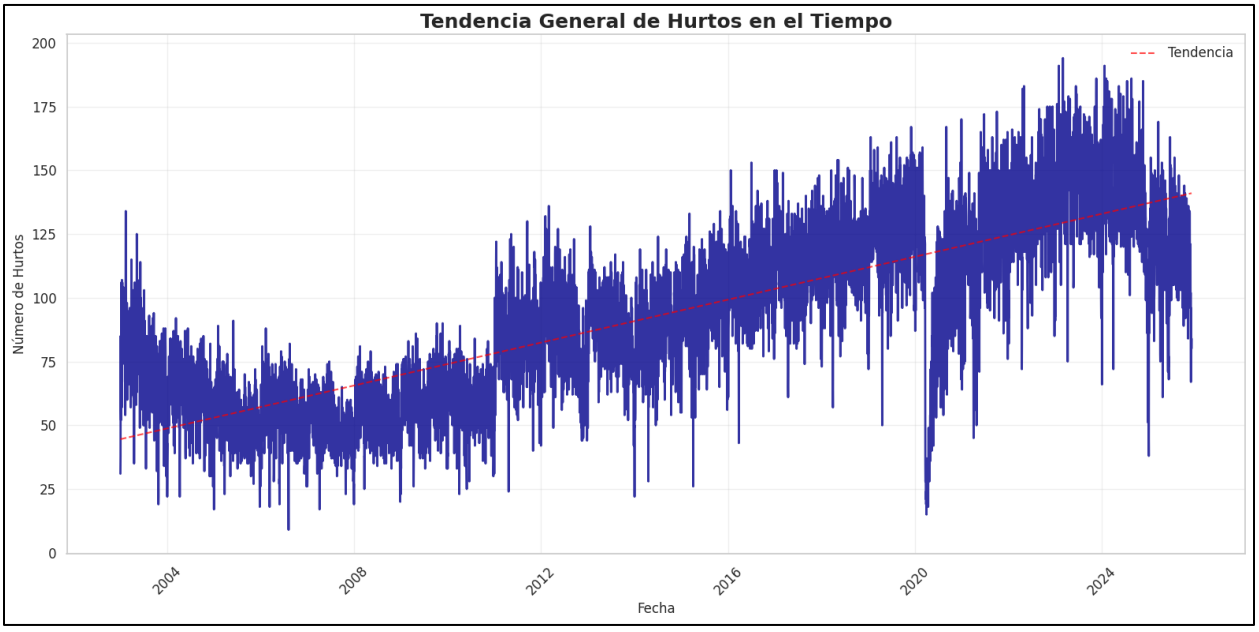
- DIA_SEMANA: Variable numérica destinada a identificar patrones cíclicos semanales (ej. aumento de hurtos los viernes).

- MES: Permitirá analizar la estacionalidad anual y el impacto de temporadas específicas (ej. vacaciones o festividades decembrinas).
- AÑO: Fundamental para analizar la evolución histórica y la tendencia macro del delito a lo largo del tiempo, permitiendo identificar si el fenómeno presenta un crecimiento o decrecimiento estructural.

Ahora, con estas nuevas columnas podemos llevar a cabo análisis de tendencias temporales:

Figura 10

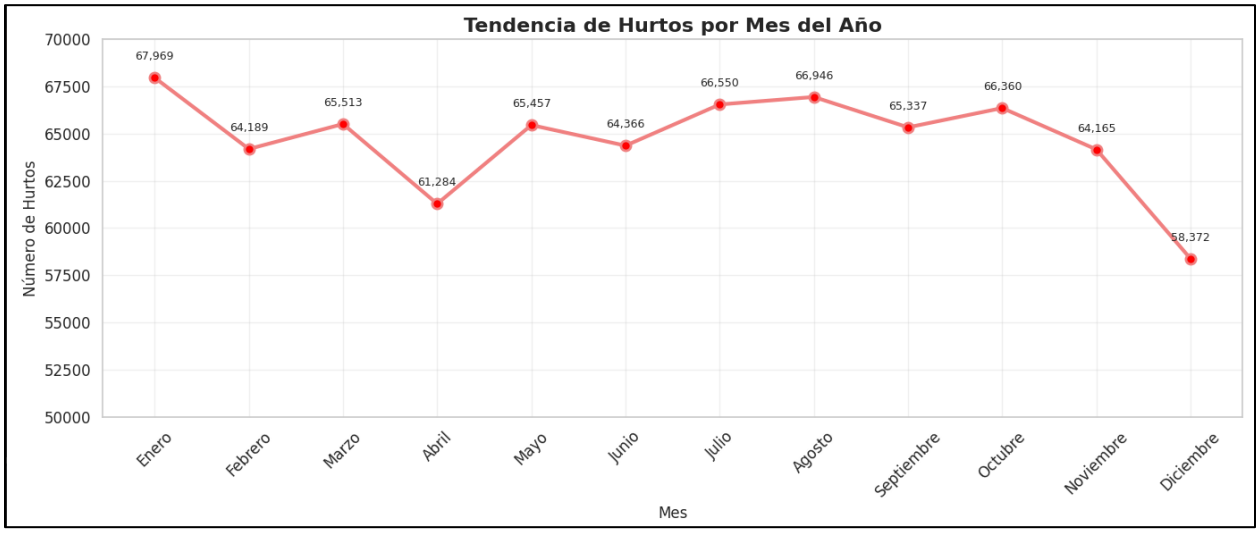
Distribución Temporal de Hurtos a Vehículos por Día



Nota. La gráfica ilustra la evolución temporal de los hurtos, evidenciando una tendencia ascendente casi lineal a través de los años. No obstante, se observa una ruptura significativa en esta tendencia durante el año 2020, presentando un descenso abrupto que coincide con las medidas de confinamiento y restricción de movilidad implementadas por la pandemia de -19 COVID-19.

Figura 11

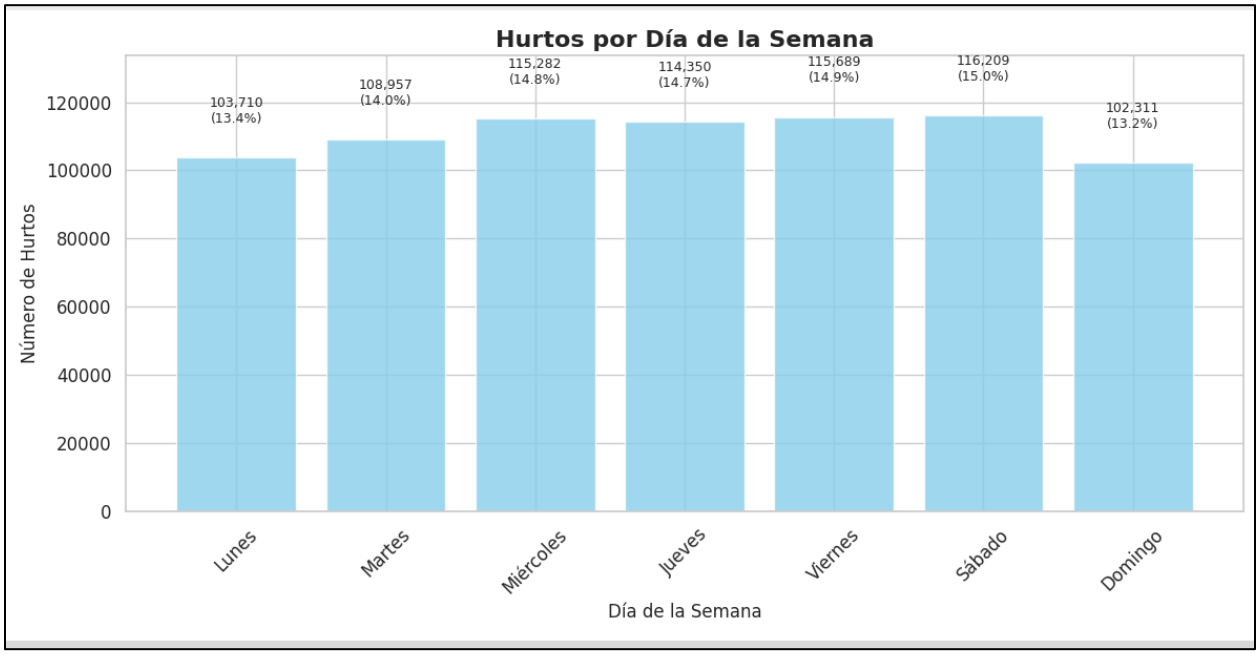
Distribución Temporal de Hurtos a Vehículos por Mes



Nota. Al examinar la distribución temporal de los incidentes, enero se consolida como el mes de mayor incidencia delictiva. Por el contrario, se registran descensos estadísticamente notables en diciembre y abril. Estas fluctuaciones negativas coinciden temporalmente con las festividades de fin de año y la celebración de la Semana Santa, lo que sugiere un fuerte componente estacional a la baja asociado a periodos vacacionales.

Figura 12

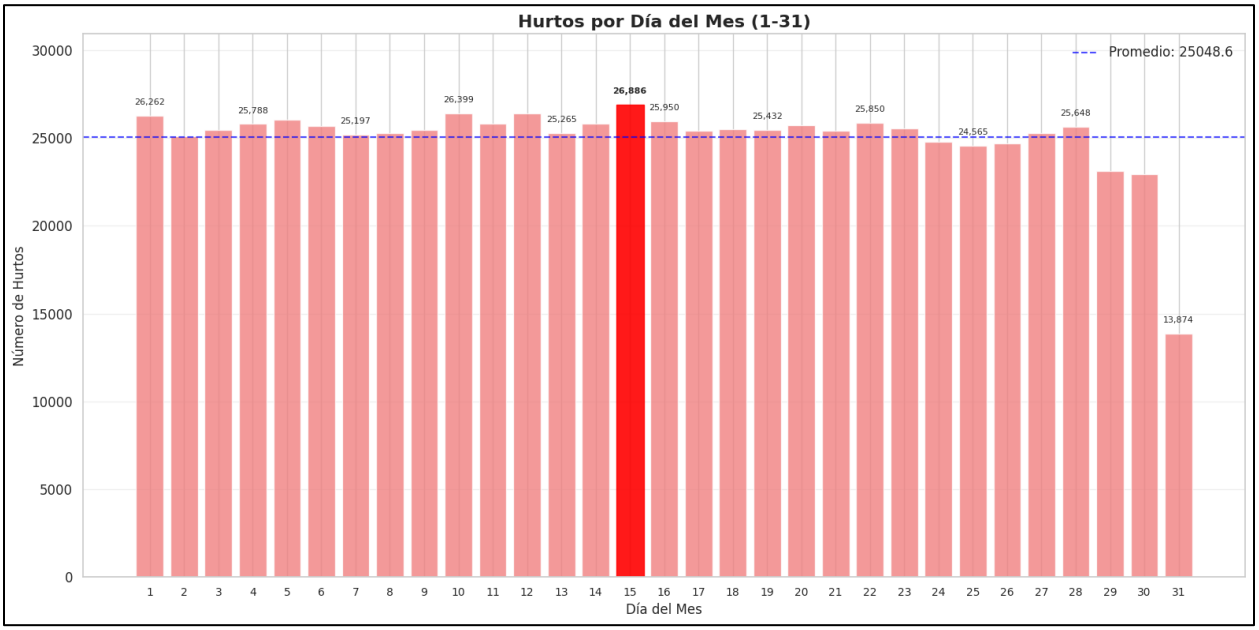
Distribución Temporal de Hurtos a Vehículos por Día de la Semana



Nota. El análisis semanal evidencia una clara focalización del delito en la segunda mitad de la semana (de miércoles a sábado), contrastando con un descenso significativo durante los domingos. Este comportamiento sugiere que el fenómeno delictivo está estrechamente vinculado a patrones de actividad social y nocturna, contrayéndose cuando la operatividad comercial y laboral de las ciudades disminuye.

Figura 13

Distribución Temporal de Hurtos a Vehículos por el Día del Mes (1-30

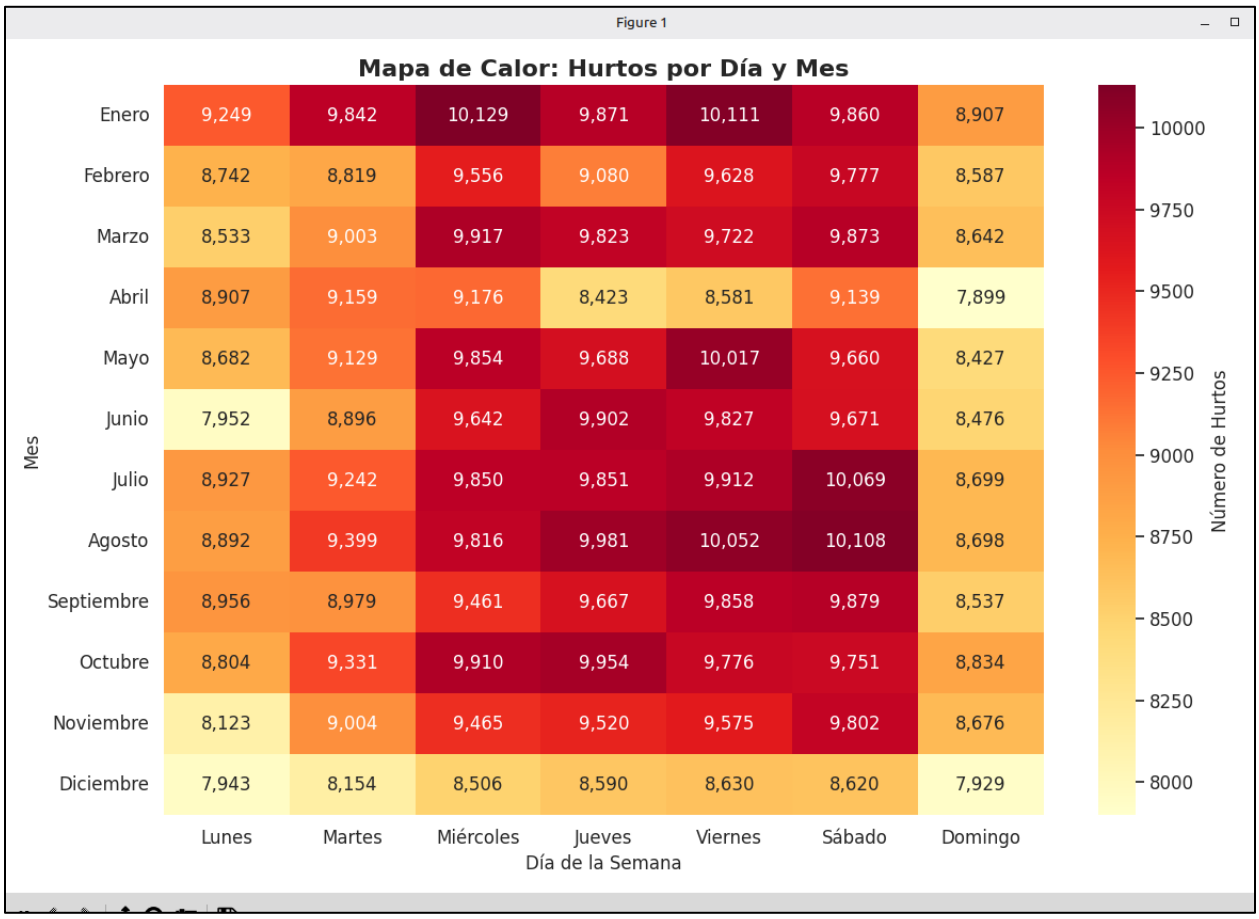


Nota. La gráfica muestra un comportamiento homogéneo a lo largo del mes. Salvo por la caída natural en los últimos días (atribuible a la duración variable de los meses) y un ligero incremento a mitad de mes, la serie de datos es plana. Esto permite concluir que el 'día del mes' tiene una baja varianza y no representa un factor discriminante fuerte para el modelo.

Análisis Bivariado con las Nuevas Columnas

Figura 14

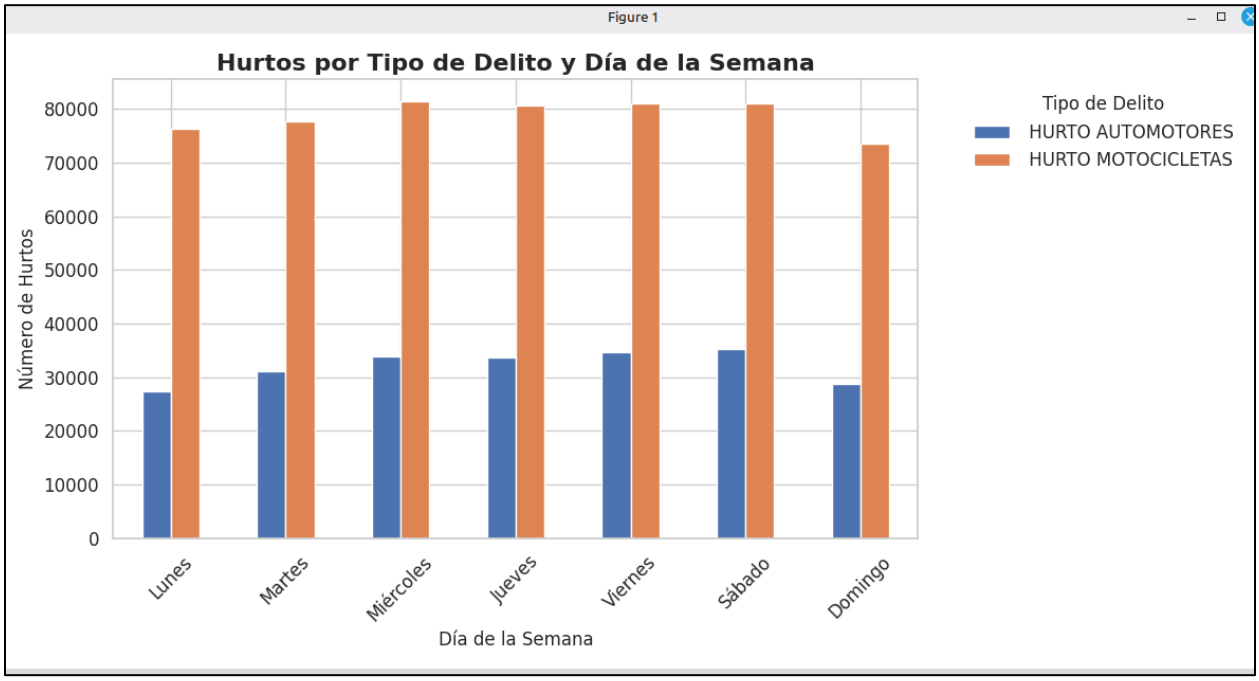
Mapa de Calor de Hurtos por Día y Mes



Nota. Mediante la matriz de densidad (mapa de calor), se detectan patrones estacionales específicos. Los focos de mayor intensidad se concentran en el inicio del año (miércoles y viernes de enero) y en el periodo de mitad de año (viernes de julio y agosto). Por el contrario, los lunes y domingo de diciembre representan las zonas más 'frías' o de menor incidencia, lo que sugiere un cambio en la dinámica social durante las festividades de fin de año.

Figura 15

Hurto por Tipo de Delito y Día de la Semana



La incorporación de estas nuevas variables facilitó la ejecución de un análisis multivariado, permitiendo cruzar la ‘Cantidad de hurtos’ con el ‘Día de la semana’ y la modalidad específica (vehículos vs. motocicletas). El análisis revela que la distinción entre estas modalidades no aporta una variabilidad significativa al modelo, dado que ambas presentan patrones de comportamiento simétricos: los picos y valles de incidencia en el hurto a motocicletas coinciden proporcionalmente con los de vehículos, sugiriendo que ambos delitos responden a los mismos factores temporales.

Limitación de la granularidad temporal: Durante la fase de exploración, se identificó que el atributo FECHA_HECHO registra únicamente la fecha calendario del suceso, careciendo del componente horario (tiempo). Debido a esta restricción en la resolución de los datos, no fue

factible realizar la categorización horaria (time binning) en franjas como ‘Madrugada’, ‘Tarde’ o ‘Noche’.

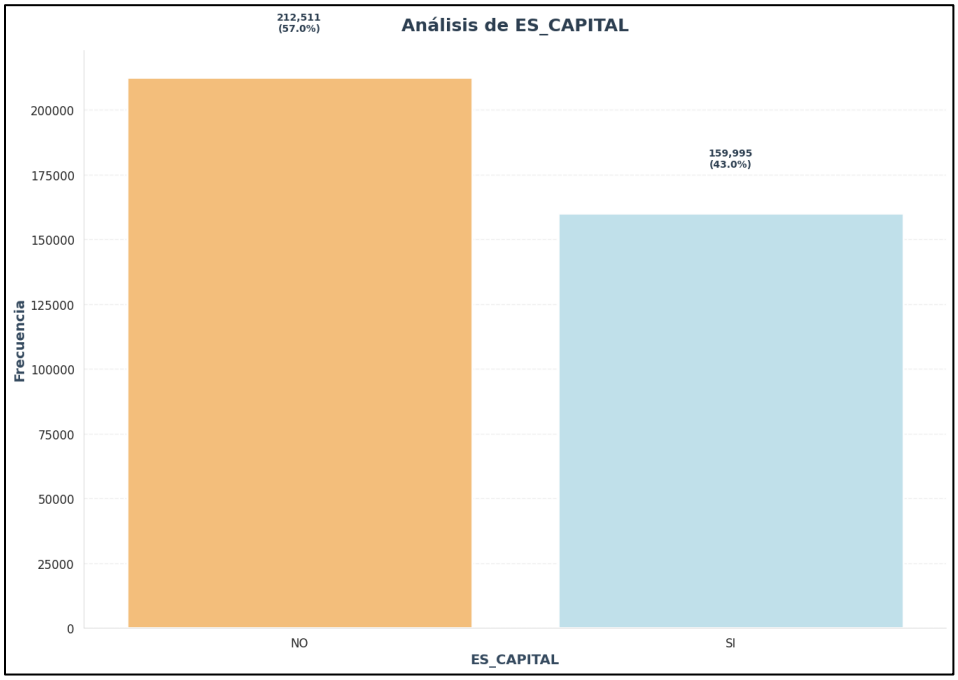
En consecuencia, el modelo se enfocará en la detección de patrones estacionales (diarios, semanales y mensuales), excluyendo las tendencias circadianas (por horas) del alcance actual.

Nuevas Columnas Creadas

Con el fin de diferenciar las dinámicas delictivas entre grandes centros urbanos y municipios de menor envergadura, se generó una nueva variable binaria denominada ‘ES_CAPITAL’. Basándose en la nomenclatura estándar del DANE para la variable COD_MUNI, se identificó que los códigos terminados en ‘001’ corresponden a capitales departamentales. Por consiguiente, se asignó un valor de 1 (verdadero) a los registros que cumplen esta condición y 0 (falso) al resto.

Figura 16

Distribución si una Ciudad es Capital

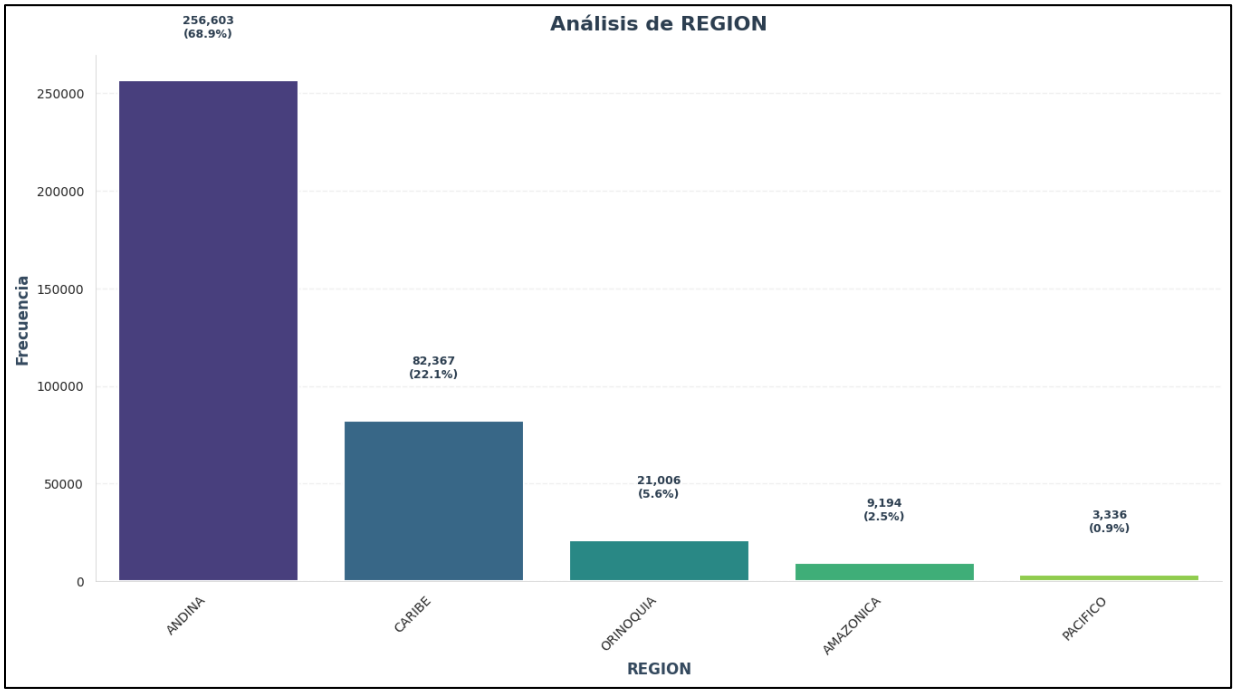


Nota. Aunque la gráfica muestra que el 43 % de los hurtos ocurren en capitales frente al 57 % en el resto del país, este dato revela una concentración crítica del delito. Considerando que Colombia cuenta con 1.103 municipios, de los cuales solo 32 son capitales, es notable que menos del 3 % de las entidades territoriales acumulen casi la mitad de la actividad delictiva nacional. Esto confirma que la variable 'ES_CAPITAL' puede ser altamente relevante y predictiva.

Se implementó una estrategia de agregación espacial generando la variable 'REGION'. El propósito de esta agrupación es identificar patrones a escala macro y capturar la heterogeneidad delictiva existente entre grandes zonas geográficas (por ejemplo, contrastar las dinámicas de la Región Caribe frente a la Región Andina). Para ello, se estableció un mapeo lógico que asigna cada departamento a su respectiva región natural (ej. Cundinamarca > Andina).

Figura 17

Distribución por Region



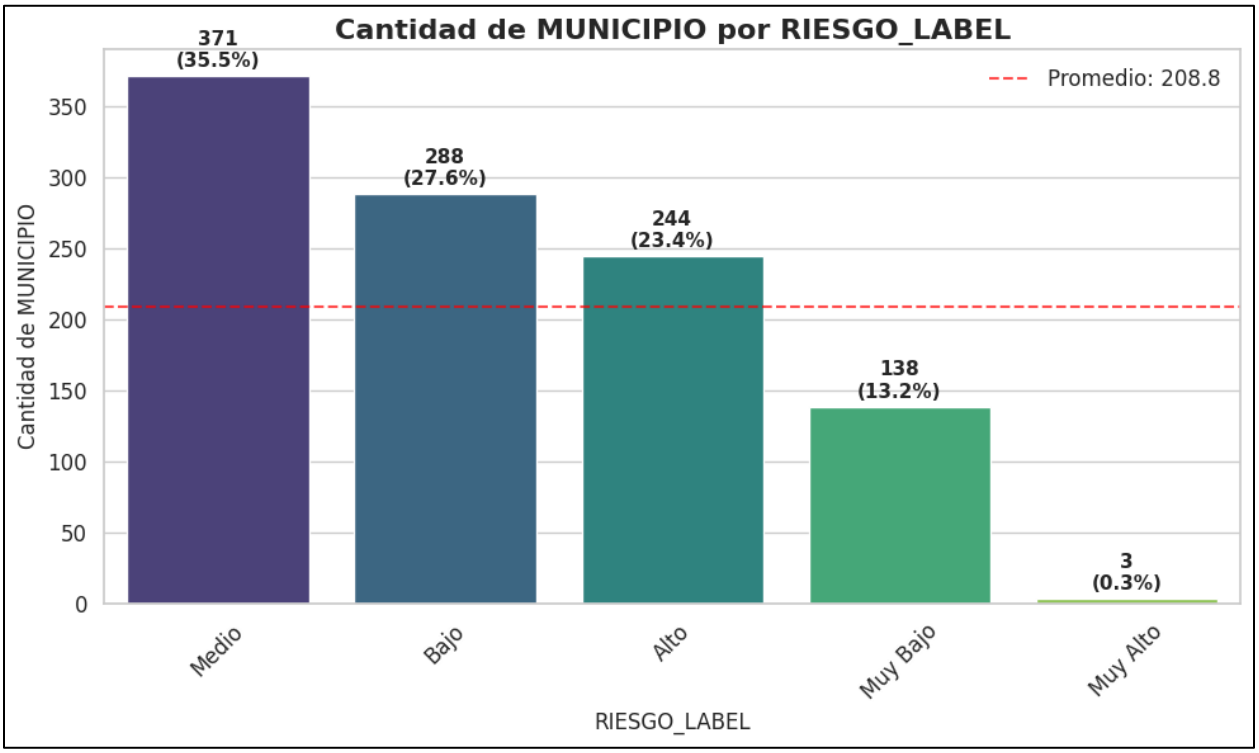
Nota. Los datos confirman de manera contundente la relevancia de la ubicación geográfica: casi siete de cada diez delitos (68.9 %) ocurren en la región Andina. Esta desproporción responde directamente a la estructura económica del país, dado que esta zona concentra los mayores centros urbanos y corredores comerciales, generando un mayor volumen de oportunidades para el hurto.

Siguiendo con la ingeniería de características, la construcción de las variables objetivo CLASE_RIESGO y RIESGO_LABEL se realizó mediante un enfoque de aprendizaje no supervisado, orientado a identificar patrones de comportamiento en la distribución de los datos. Para ello, se implementó el algoritmo K-Means Clustering, el cual permitió segmentar los municipios en función de su densidad delictiva histórica.

El proceso consistió en agrupar los registros en tres clústeres naturales (k=5), calculados mediante la minimización de la inercia (varianza intraclase). De este modo, los límites para las categorías de riesgo (Muy Bajo, Bajo, Medio, Alto y Muy Alto) se establecieron matemáticamente a partir de los centroides resultantes, asegurando que cada nivel represente grupos homogéneos de municipios con intensidades criminales similares, derivadas estrictamente de la estructura estadística del conjunto de entrenamiento.

Figura 18

Distribución de Etiqueta Riesgo Usando K-Means



Nota. La distribución de municipios por nivel de riesgo es desigual, concentrándose principalmente en el nivel Medio, que supera ampliamente el promedio. Los niveles Bajo y Alto también presentan una alta representación, evidenciando una diferenciación intermedia del riesgo. En contraste, los niveles Muy Bajo y especialmente Muy Alto agrupan menos municipios, siendo este último marginal, pero de alta relevancia por su carácter crítico.

Transformación de Datos (Encoding)

Dado que el algoritmo seleccionado requiere entradas numéricas para realizar los cálculos estadísticos y las particiones de los árboles, se transformaron las variables categóricas (como DEPARTAMENTO, MUNICIPIO, DIA_SEMANA y MES) mediante técnicas de codificación:

Codificación de Variables Categóricas:

Estrategia de codificación numérica: Con el fin de asegurar la compatibilidad matemática con los algoritmos de aprendizaje supervisado y permitir el procesamiento de variables categóricas, se definieron dos enfoques de transformación.

Tratamiento de identificadores geográficos (códigos DANE): Para la representación de las dimensiones espaciales, se utilizaron directamente los identificadores numéricos oficiales del DANE (COD_DEPTO y COD_MUNI) sin aplicar transformaciones de codificación adicionales. Esta decisión metodológica se fundamenta en la naturaleza algorítmica de los modelos de ensamble seleccionados (Random Forest y XGBoost). A diferencia de los modelos lineales (como la regresión logística o SVM), que podrían interpretar erróneamente una jerarquía ordinal en los códigos postales (asumiendo, por ejemplo, que el municipio 5002 es “mayor” o tiene más peso que el 5001), los algoritmos basados en árboles de decisión operan mediante particiones del espacio de características (feature splitting). El modelo utiliza los códigos numéricos únicamente como puntos de corte para segregar los datos en nodos homogéneos, aprendiendo relaciones no lineales e isolando regiones geográficas específicas sin atribuirles una magnitud matemática. Esta estrategia permite conservar la integridad de la información geográfica, evitando la dispersión de datos (sparsity) y la explosión dimensional que generaría una codificación one-hot sobre más de 1,000 categorías municipales.

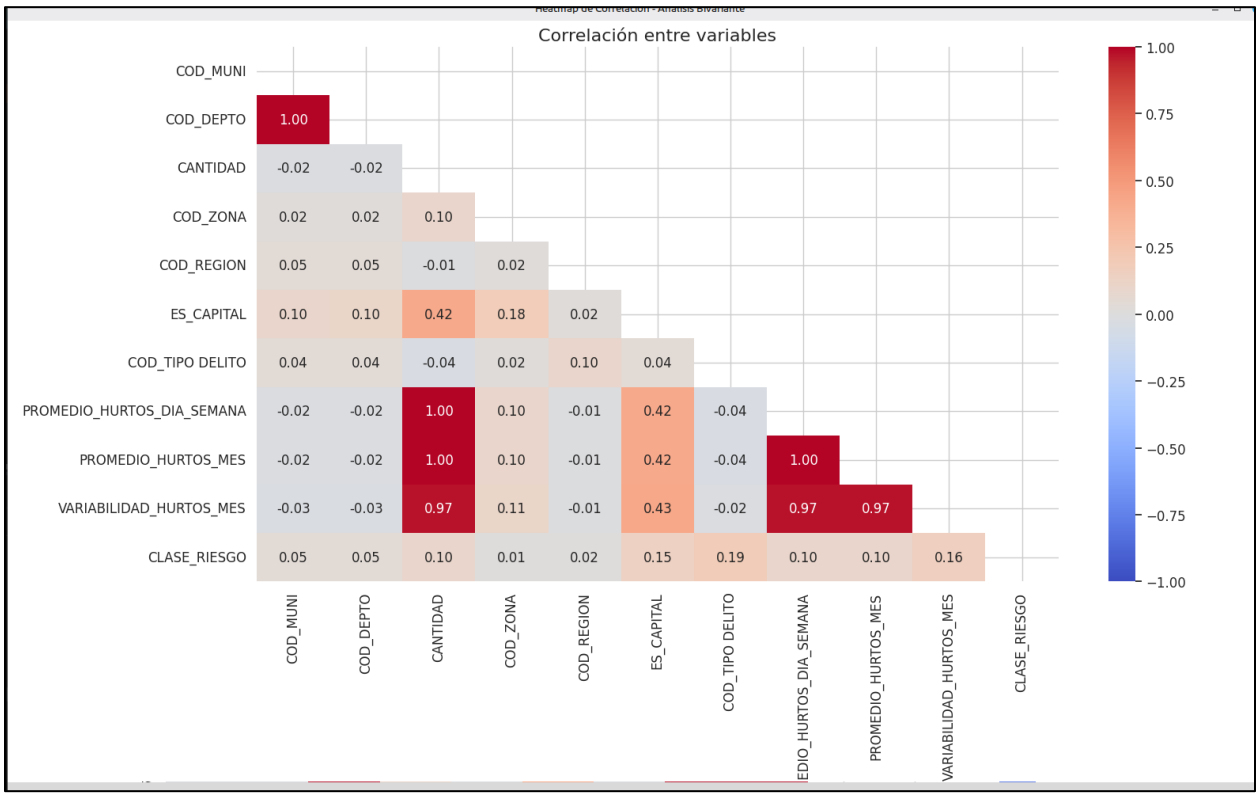
Control de fuga de datos: Es crucial destacar que el cálculo de estos promedios se realizó estrictamente sobre el conjunto de entrenamiento (Train Set). Posteriormente, estos valores aprendidos fueron mapeados al conjunto de prueba (Test Set). Este procedimiento garantiza que el modelo no tenga acceso a estadísticas futuras, previniendo el data leakage y asegurando la integridad de la validación cruzada.

*Label Encoding***:** Para las variables nominales restantes sin ID numérico, como TIPO_DELITO, ZONA y REGIÓN, se aplicó una codificación de etiquetas manual.

Una vez efectuada la transformación de las variables categóricas a su representación numérica, resulta factible implementar técnicas estadísticas cuantitativas. En esta etapa, se genera la matriz de correlación, con el objetivo de evaluar el grado de asociación lineal entre las variables y detectar posibles patrones de dependencia o redundancia en el conjunto de datos.

Figura 19

Correlación Entre Variables Numéricas



Nota. El examen de la matriz de correlación evidencia un predominio de independencia lineal entre la mayoría de las variables, destacándose como hallazgo principal la correlación positiva moderada de 0.38 entre la variable derivada 'ES_CAPITAL' y la variable objetivo 'CANTIDAD', lo cual valida estadísticamente la hipótesis de que la jerarquía urbana es un factor determinante en la incidencia delictiva. Paralelamente, se identificó un alta multicolinealidad (0.80) entre 'DIA_SEMANA' y 'FIN_SEMANA', sugiriendo redundancia de información, mientras que los coeficientes cercanos a cero en variables temporales como 'MES' indican que la estacionalidad del fenómeno no responde a patrones lineales simples, justificando así la necesidad de implementar algoritmos no paramétricos (como Random Forest o Gradient Boosting) capaces de capturar la complejidad no lineal subyacente en los datos.

Selección de Características (Feature Selection)

Tras finalizar las actividades de limpieza y enriquecimiento, se definió la estructura final del conjunto de datos que alimentará los algoritmos de machine learning. Esta selección busca maximizar la capacidad predictiva del modelo eliminando ruido y redundancia.

Definición de la Variable Objetivo (Target)

Se estableció como variable dependiente (y) la columna [CLASE_RIESGO] porque esta representación categórica se alinea mejor con la naturaleza asimétrica del fenómeno delictivo. Dado que la mayoría de los municipios presentan cifras bajas y solo unos pocos concentran la criminalidad (distribución de cola larga), la predicción numérica directa tiende a sesgarse. La clasificación por niveles permite gestionar este desbalance, asegurando que el modelo capture correctamente los focos críticos sin perder sensibilidad ante las variaciones menore

Variables Predictoras (Features)

Se seleccionaron las siguientes variables independientes (X) por su alta correlación con el comportamiento delictivo, evidenciada en la fase de exploración:

- COD_MUNI
- COD_DEPTO
- COD_ZONA
- COD_REGION
- COD_TIPO DELITO
- ES_CAPITAL
- ANIO
- MES
- DIA_SEMANA

Tabla 3

Resumen con las Variables Finales (Diccionario de Datos

Nombre de la Variable	Tipo de Dato	Categoría	Descripción Técnica	Origen	Rol en el Modelo
COD_MUNI	Numérico (Int)	Geográfica	Código DANE identificador único del municipio.	Original	Predictor (Feature)
COD_DEPTO	Numérico (Int)	Geográfica	Código identificador del departamento.	Original	Predictor (Feature)
MUNICIPIO	Categórico (Texto)	Descriptiva	Nombre oficial del municipio.	Original	Informativo (No entrena)
DEPARTAMENTO	Categórico (Texto)	Descriptiva	Nombre oficial del departamento.	Original	Informativo (No entrena)
COD_ZONA	Numérico (Int)	Geográfica	Código de la zona (Urbana/Rural) donde ocurrió el hecho.	Original	Predictor (Feature)
COD_REGION	Numérico (Int)	Geográfica	Identificador de la región geográfica del país.	Original	Predictor (Feature)

Nombre de la Variable	Tipo de Dato	Categoría	Descripción Técnica	Origen	Rol en el Modelo
COD_TIPO DELITO	Numérico (Int)	Categórica	Código numérico que representa la modalidad del hurto.	Original	Predictor (Feature)
ES_CAPITAL	Numérico (Binario)	Derivada	Variable dummy (1=Capital, 0=No Capital) basada en la jerarquía administrativa.	Ingeniería de Características	Predictor (Feature)
CANTIDAD	Numérico (Int)	Continua	Número total de hurtos registrados en el evento.	Original	Base para el Target
RIESGO_LABEL	Categórico (Texto)	Target	Etiqueta legible del nivel de riesgo (Baja, Media, Alta).	Transformación	Etiqueta (Label)
CLASE_RIESGO	Numérico (Int)	Target	Codificación numérica del riesgo (0=Baja, 1=Media, 2=Alta).	Transformación	Variable Objetivo (y)

Nombre de la Variable	Tipo de Dato	Categoría	Descripción Técnica	Origen	Rol en el Modelo
ANIO	Numerico (int)	Temporal	Año donde ocurre el hurto	Treansformacion	Predictor (Feature)
MES	Numerico (int)	Temporal	Mes donde ocurre el hurto (1-12)	Transformacion	Predictor (Feature)
DIA_SEMANA	Numerico (int)	Temporal	Dia donde ocurre el hurto (1-7)	Transformacion	Predicotr (Feature)

Almacenamiento y Persistencia de Datos

Una vez finalizada la fase de análisis y preprocesamiento, se procedió a la persistencia de los datos tratados utilizando MongoDB como motor de base de datos NoSQL. La elección de esta tecnología se fundamenta en su arquitectura orientada a documentos, la cual permite gestionar estructuras flexibles (JSON/BSON) y ofrece una alta escalabilidad horizontal, características críticas para la integración eficiente con flujos de trabajo en Python (McKinney, 2022).

Procedimiento de carga (ETL): Para la migración de los datos desde el entorno de análisis (Pandas) hacia el almacenamiento, se diseñó un proceso de carga automatizado:

Transformación: El DataFrame procesado se convirtió en una lista de diccionarios mediante el método to_dict(orient='records'). Esto asegura que cada fila del conjunto de datos se interprete como un documento independiente, preservando la integridad de las variables generadas.

Limpieza de entorno: Como medida de integridad referencial y para evitar la duplicidad de registros en ejecuciones iterativas, se implementó una rutina de drop (eliminación) de la colección existente antes de la nueva carga.

Inserción masiva: Se utilizó el método insert_many(), optimizado para operaciones de escritura en lote, permitiendo la ingesta simultánea de los documentos en la colección hurto_vehiculos (MongoDB, 2024).

Conclusión Fase Preparación de los Datos

Como cierre de esta fase, crítica dentro del ciclo CRISP-DM, se logró transformar los datos transaccionales brutos en una estructura analítica depurada y robusta. Este proceso garantizó la integridad del conjunto de datos mediante la eliminación de ruido y registros defectuosos, al tiempo que elevó su valor a través de un enriquecimiento semántico que dotó al modelo de inteligencia temporal y contexto histórico, superando así las limitaciones de las variables geográficas simples. Metodológicamente, se aseguró la validez científica blindando el estudio contra la fuga de información (data leakage) en el cálculo de variables estadísticas, culminando con la persistencia eficiente en MongoDB para ofrecer una infraestructura disponible y escalable para la fase de entrenamiento.

Modelado (Modeling)

En esta fase se procedió a la selección y aplicación de técnicas de modelado matemático. Dado que el objetivo del negocio es categorizar el nivel de riesgo de un municipio, el problema se enmarca como una tarea de clasificación supervisada. Se utilizó la librería Scikit-Learn de Python como entorno de desarrollo debido a su robustez y estandarización en la industria.

Selección del Enfoque de Modelado

Para el desarrollo del componente predictivo, se evaluaron las diferentes tipologías de aprendizaje automático existentes. Según la literatura especializada, los algoritmos de machine learning se clasifican principalmente en dos categorías según la naturaleza de los datos de entrada: aprendizaje supervisado y no supervisado (Sarker, 2021).

El aprendizaje no supervisado se utiliza cuando el conjunto de datos carece de etiquetas predefinidas, y el objetivo del algoritmo es descubrir estructuras ocultas o patrones intrínsecos, como en el caso de la segmentación de clientes (clustering) (Alloghani et al., 2020). Por el contrario, el aprendizaje supervisado implica el entrenamiento de un modelo utilizando un conjunto de datos etiquetado (labeled data), donde se conoce la variable objetivo (target) y se busca aprender una función de mapeo que permita predecir dicha etiqueta para nuevos datos no vistos (Müller & Guido, 2017).

Justificación de la Selección

Para el presente proyecto, se seleccionó un enfoque de aprendizaje supervisado. Esta decisión se fundamenta en la naturaleza de los datos históricos de hurto de vehículos disponibles, los cuales cuentan con una "verdad fundamental" (Ground Truth) claramente definida. Específicamente, cada registro en el conjunto de datos histórico contiene tanto las variables

predictoras (fecha, ubicación, modalidad) como el resultado conocido del evento (tipo de vehículo hurtado o nivel de riesgo).

El objetivo del negocio no es simplemente agrupar los delitos por similitud (lo cual sería no supervisado), sino predecir una categoría específica (clase de vehículo/nivel de riesgo) basándose en antecedentes. Por lo tanto, el problema se enmarca técnicamente como una tarea de clasificación supervisada, donde el algoritmo debe asignar una etiqueta discreta a una nueva instancia de datos basándose en el aprendizaje previo de patrones históricos (IBM, 2023).

Selección de Familias de Algoritmos

Una vez preparada la data, la fase de modelado requiere seleccionar la familia de algoritmos adecuada dentro del aprendizaje supervisado. Según James et al. (2021), la distinción fundamental para esta elección radica en la naturaleza de la variable de respuesta: si la variable es cuantitativa (numérica continua), el problema corresponde a una regresión; mientras que, si la variable es cualitativa (categórica o discreta), se trata de un problema de clasificación.

Dado que el objetivo de esta investigación es predecir una etiqueta de clase si puede ocurrir un robo y no estimar un valor numérico continuo, los modelos de regresión lineal no son apropiados. Por consiguiente, y siguiendo los lineamientos metodológicos de Géron (2019) para el manejo de salidas discretas, se determina utilizar la familia de algoritmos de clasificación, los cuales están diseñados para asignar observaciones nuevas a categorías predefinidas basándose en la probabilidad de pertenencia.

Selección de Algoritmo

Para garantizar la robustez de la predicción del nivel de riesgo, se llevó a cabo la implementación y evaluación de tres algoritmos con diferentes enfoques arquitectónicos los cuales detallo a continuación:

Regresión Logística (Logistic Regression)

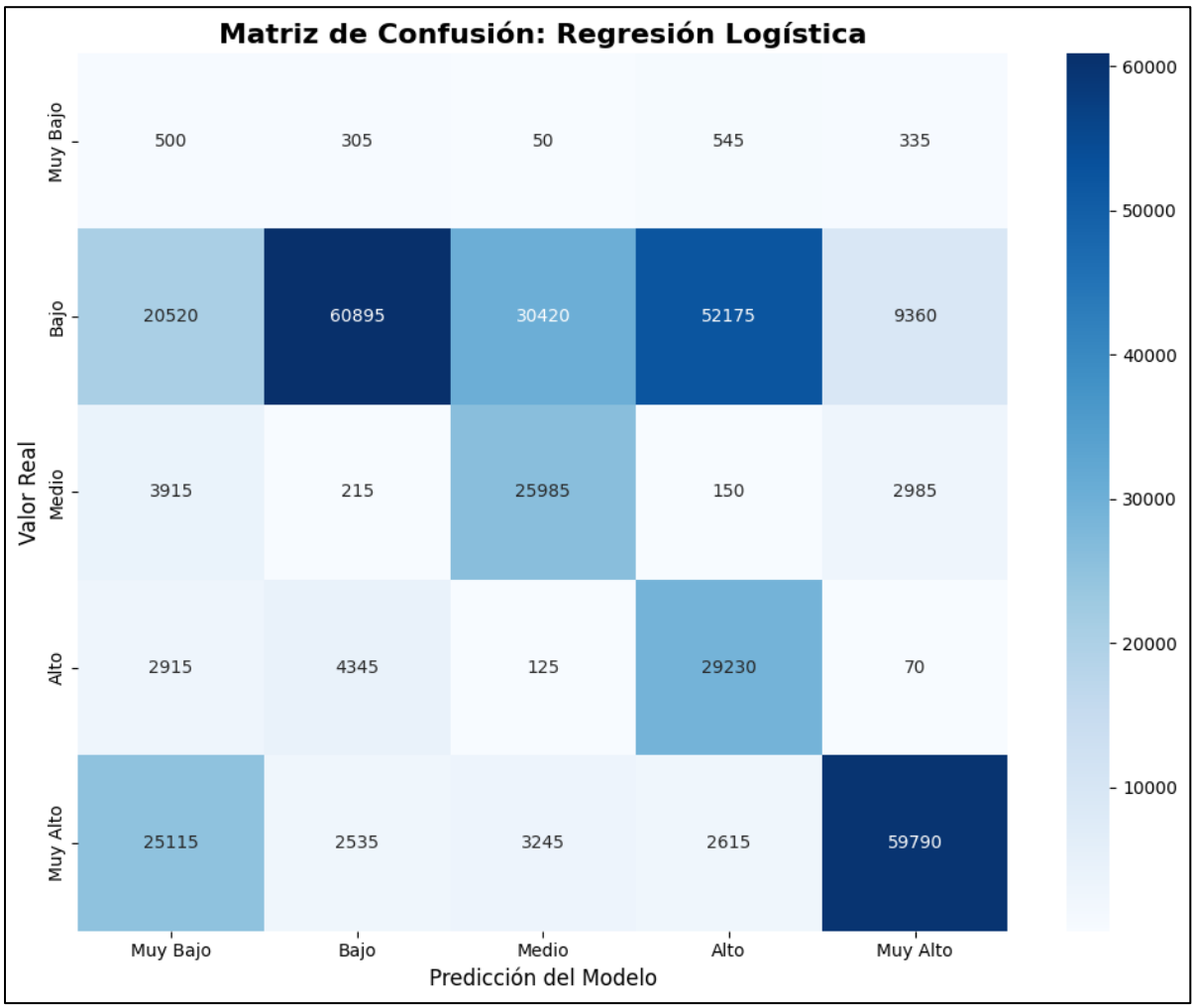
Se incluyó como modelo de base (baseline) debido a su simplicidad y alta interpretabilidad. Este modelo busca relaciones lineales entre las variables. No obstante, en el contexto de la criminalidad en Colombia, su rendimiento fue limitado, ya que no logra capturar las interacciones complejas y no lineales entre la ubicación geográfica y la temporalidad del delito. El modelo fue entrenado utilizando un conjunto de datos de 1,524,190 muestras con 9 características principales. Los resultados clave son:

- *Desempeño general***:** El modelo alcanzó un accuracy de 52.14 % en los datos de prueba y un F1-Score Macro de 0.4566.

- *Capacidad discriminativa***:** A pesar del accuracy moderado, el modelo muestra una excelente capacidad para separar las clases, con un AUC Macro Promedio de 0.843. La clase con mejor desempeño fue “Muy Alto” (AUC: 0.948).
- *Hallazgos técnicos:* Se identificó una alerta de convergencia (ConvergenceWarning) tras 1000 iteraciones, lo que sugiere la necesidad de implementar un escalamiento de características (feature scaling) para optimizar el ajuste del modelo en futuras iteraciones.

Figura 20

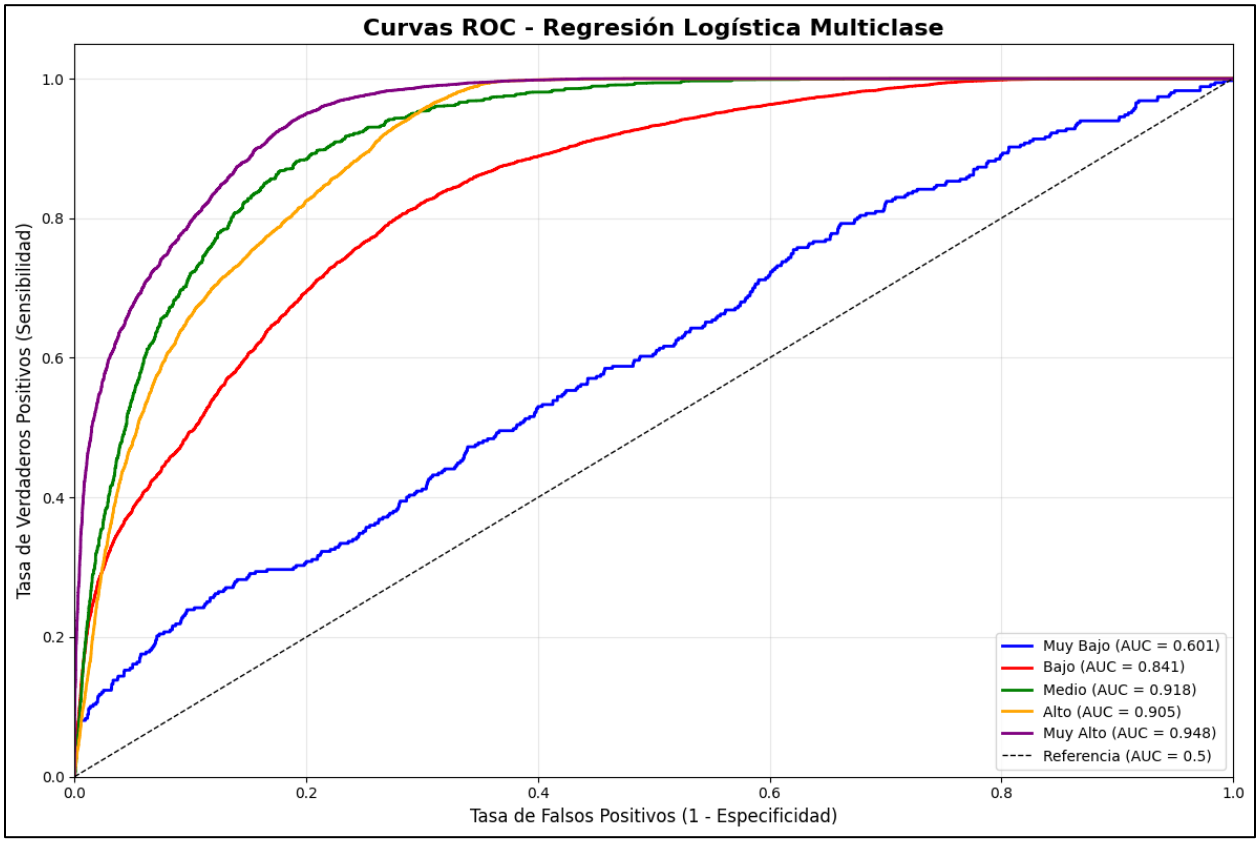
Matriz de Confusión Regresión Logística



Nota. La matriz de confusión revela un desempeño heterogéneo entre las categorías. Se observa una alta eficacia en la identificación de la clase "Muy Alto" con 59,790 predicciones correctas, y una capacidad aceptable en la clase "Alto" (29,230). No obstante, el modelo presenta una confusión significativa en la clase "Bajo", donde una gran proporción de los casos reales son clasificados erróneamente como "Bajo" (60.895), pero también desplazados hacia "Alto" (52,175). Esta dispersión fuera de la diagonal principal sugiere que, aunque el modelo captura tendencias generales, el desbalance de los datos y la superposición de características geográficas dificultan la distinción precisa entre los niveles de criminalidad intermedios

Figura 21

Curvas ROC Regresión Logística



Nota. Las curvas ROC muestran una capacidad discriminativa sobresaliente para la mayoría de las categorías, con un AUC Macro Promedio de 0.843. Destaca la clase “Muy Alto” con un AUC de 0.948, lo que indica una probabilidad casi perfecta de distinguir esta clase frente a las demás. Las categorías “Medio” y “Alto” también presentan métricas robustas (0.918 y 0.905 respectivamente). La excepción notable es la clase “Muy Bajo” (AUC 0.601), cuyo comportamiento se acerca a la línea de referencia (azar), lo que confirma que el modelo requiere de características adicionales o un remuestreo para mejorar la detección de eventos de baja frecuencia

Conclusión

El modelo de regresión logística demuestra una capacidad sobresaliente para identificar niveles de riesgo críticos, destacando un AUC de 0.948 en la clase “Muy Alto”, lo que lo hace altamente efectivo para predecir zonas de alto impacto delictivo en Colombia. Sin embargo, el accuracy global del 52.14 % y la falta de convergencia del algoritmo señalan que el desbalance de las clases y la alta sensibilidad a la escala de las variables geográficas (COD_MUNI y COD_DEPTO) limitan su precisión en niveles de riesgo bajos. En conclusión, aunque el modelo es una base sólida para la detección de picos de criminalidad, se requiere implementar escalamiento de datos y técnicas de balanceo para mejorar la clasificación en las categorías intermedias y asegurar una generalización más estable.

Bosques Aleatorios (Random Forest)

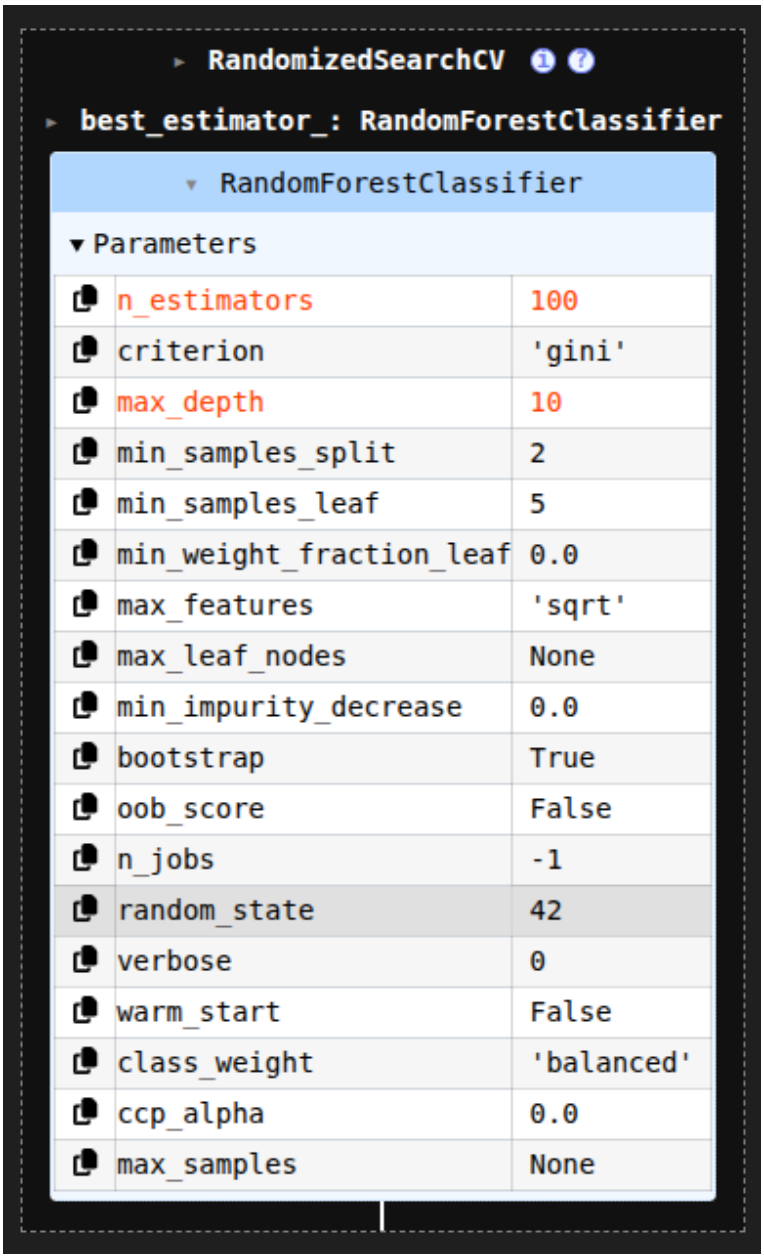
Este fue el modelo seleccionado finalmente para el despliegue. Al ser un método de ensamble basado en el embolsado (bagging), construye múltiples árboles de decisión y combina sus resultados. Su capacidad para manejar datos desbalanceados y su resistencia al sobreajuste lo hicieron ideal para procesar las más de 900,000 muestras del conjunto de datos, ofreciendo un equilibrio óptimo entre precisión y costo computacional.

Para asegurar que la configuración del modelo de Random Forest no fuera producto de un ajuste empíricamente arbitrario, se proyectó inicialmente una fase de optimización mediante GridSearchCV. Sin embargo, debido al volumen masivo del conjunto de datos (superando las 900,000 muestras), el proceso excedió los recursos de cómputo disponibles. Ante esta limitación, se optó por implementar RandomizedSearchCV, una técnica que, en lugar de evaluar todas las combinaciones, selecciona un número determinado de configuraciones al azar (n_iter =3). Este enfoque permitió validar los hiperparámetros de manera eficiente mediante una validación

cruzada de 3 pliegues (cv=3), optimizando específicamente para la métrica F1-Macro, la cual es más robusta frente al desbalance de las clases de riesgo.

Figura 22

Resultado RandomizedSearchCV para el Modelo RandomForest



Nota. La característica de COD_TIPO DELITO: Al tener un peso de 0.0157, su impacto en la clasificación del nivel de riesgo es marginal

El modelo Random Forest fue entrenado con 100 árboles y una profundidad máxima de 10, logrando una mejora sustancial en todas las métricas de evaluación:

Desempeño general: El modelo alcanzó un accuracy de un 87.30 % en prueba y un F1-Score Weighted de 0.8831, demostrando una robustez alta para la clasificación multiclase.

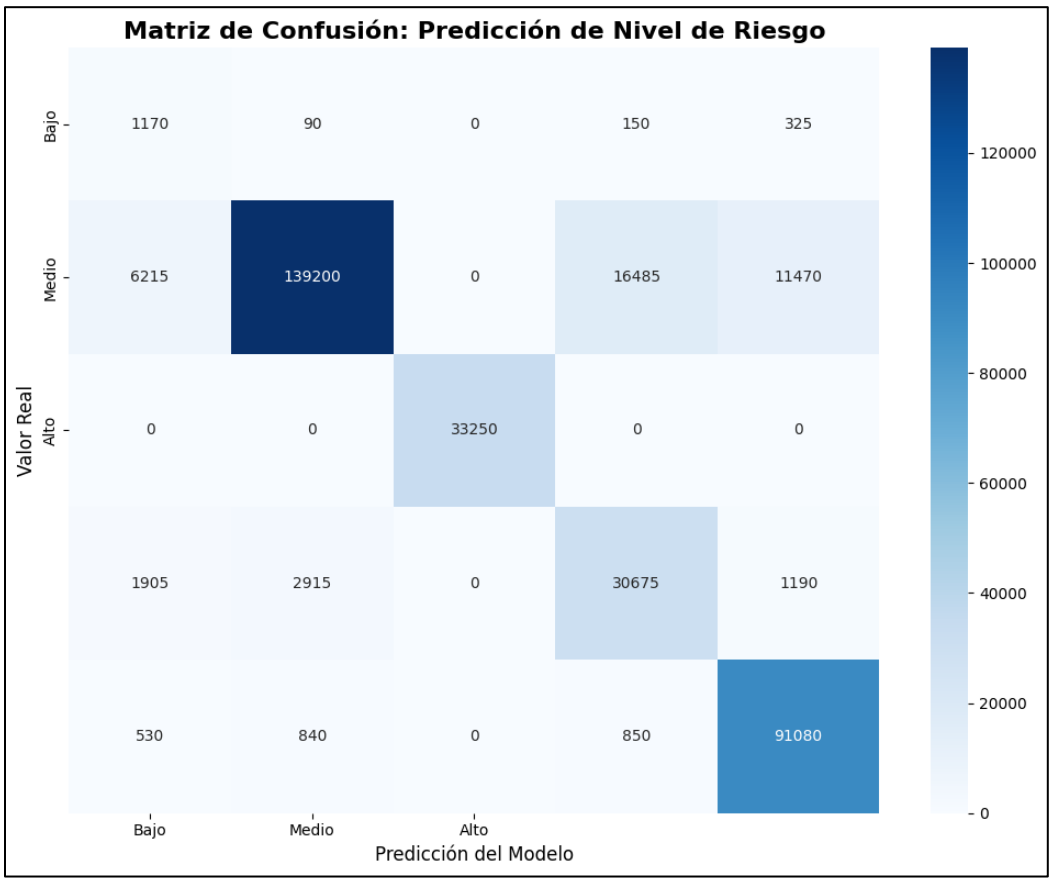
Capacidad discriminativa: Los puntajes AUC son casi perfectos, con un AUC Macro Promedio superior a 0.98. Destaca la clase “Medio” con un AUC de 1.000, indicando una separación perfecta de las probabilidades para esta categoría.

Variables clave: La importancia de las características se mantiene liderada por factores geográficos: COD_MUNI (39.29 %) y COD_DEPTO (26.82 %), confirmando que la ubicación es el predictor más fuerte de riesgo en tu estudio.

Balanceo: El uso de class_weight: balanced permitió que, a pesar del desbalance, la clase “Muy Bajo” alcanzara un Recall de 0.67, una mejora drástica frente al modelo anterior.

Figura 23

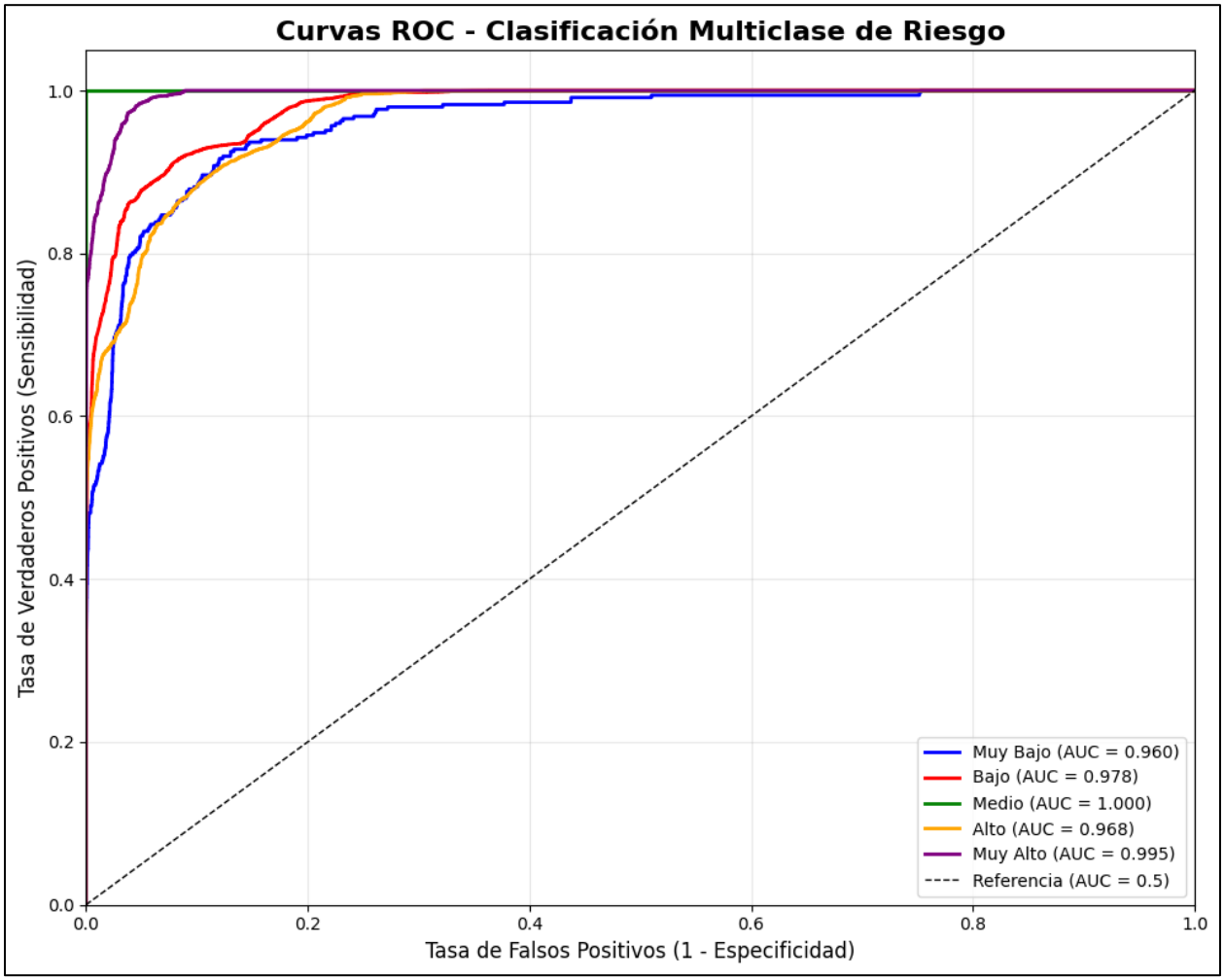
Matriz de Confusión Random Forest



Nota. La matriz de confusión refleja una diagonal principal altamente poblada, lo que confirma la precisión del modelo. Es notable el desempeño en la clase “Medio” con 139,200 aciertos y la clase “Muy Alto” con 91,080. A diferencia de la regresión logística, los errores de clasificación son mucho menores; sin embargo, se observa una ligera confusión donde algunos casos de riesgo “Medio” son etiquetados como “Bajo” (6,215 casos). A pesar de esto, la capacidad del modelo para identificar correctamente los niveles de riesgo alto y muy alto es excepcional, minimizando los falsos negativos en las categorías más críticas.

Figura 24

Curvas ROC para el Modelo Random Forest



Nota. Las curvas ROC presentan un comportamiento ideal, situándose muy cerca de la esquina superior izquierda en todas las categorías. Los valores de AUC oscilan entre 0.960 y 1.000, lo que posiciona al Random Forest como un modelo con una capacidad discriminativa excelente. El hecho de que todas las curvas mantengan esta tendencia, incluso para la clase “Muy Bajo” (AUC 0.960), demuestra que el algoritmo ha logrado generalizar los patrones de riesgo delictivo con una eficacia muy superior a los métodos lineales tradicionales, siendo capaz de distinguir entre las 5 clases con un margen de error mínimo en términos de probabilidad.

Conclusión Modelo

La implementación del algoritmo Random Forest marca un punto de inflexión en el proyecto, logrando un accuracy del 87.30 % y superando las limitaciones de convergencia encontradas en modelos previos. La alta relevancia de las variables geográficas (COD_MUNI y COD_DEPTO), combinada con la capacidad del modelo para manejar relaciones no lineales, permitió alcanzar un AUC promedio superior a 0.98, garantizando una identificación precisa de los niveles de riesgo. En conclusión, el Random Forest se consolida como la arquitectura más apta para el sistema de predicción de robo de vehículos en Colombia, ofreciendo un equilibrio óptimo entre la detección de casos críticos y la reducción de falsos positivos en todas las categorías evaluadas.

XGBoost (Extreme Gradient Boosting)

Se evaluó este modelo de potenciación de gradiente (boosting) como alternativa de alto rendimiento. Aunque XGBoost suele ofrecer precisiones superiores en competencias de ciencia de datos, para este proyecto específico, el incremento en las métricas no justificó el elevado tiempo de entrenamiento y la dificultad para interpretar las decisiones del modelo en comparación con Random Forest.

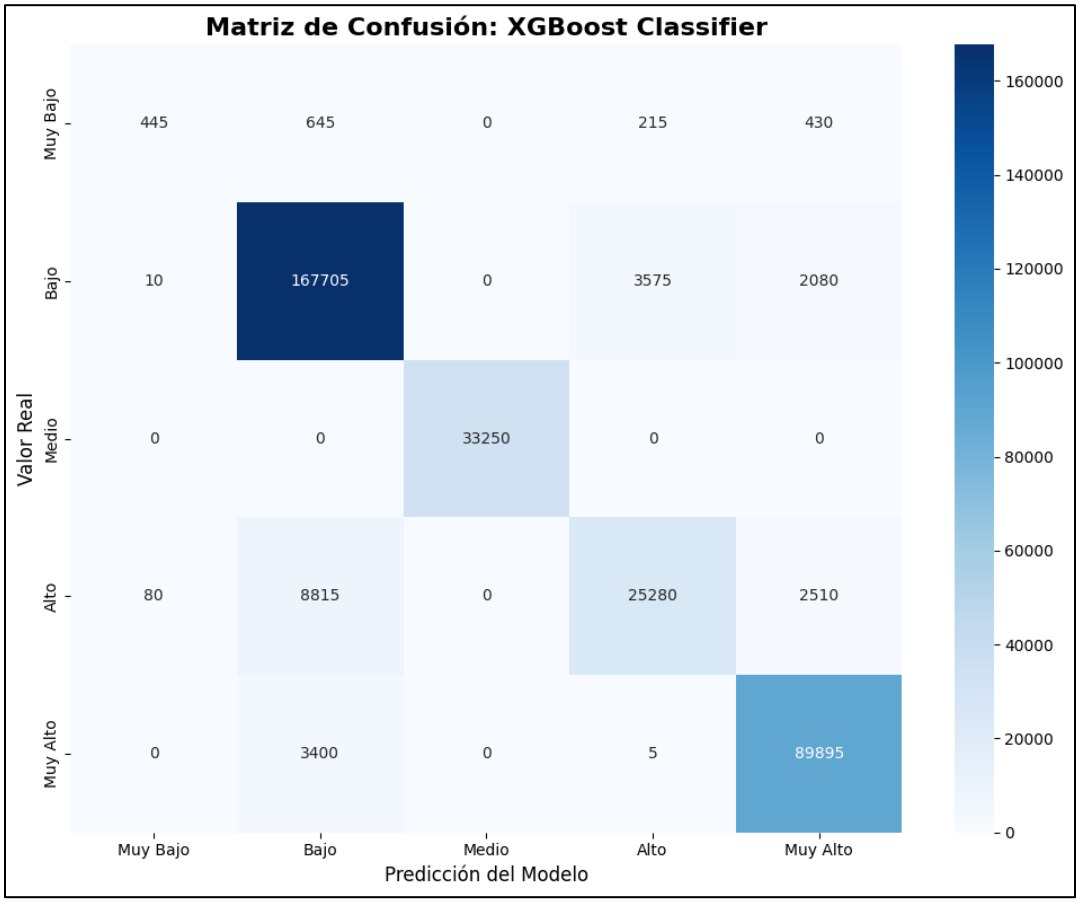
El modelo XGBoost (Extreme Gradient Boosting) presenta el desempeño más sólido de todos los algoritmos evaluados:

- Desempeño general: Alcanzó un accuracy del 93.57 % y un F1-Score Weighted del 0.9327, lo que demuestra una precisión casi quirúrgica en la clasificación de riesgos.
- Capacidad discriminativa: El AUC Macro Promedio es de 0.982. Es notable que la clase “Medio” mantiene un AUC perfecto de 1.000, mientras que la clase “Muy Alto” sube a un impresionante 0.998.

- Factores determinantes: A diferencia de los modelos anteriores, XGBoost otorga una importancia crítica a la variable ES_CAPITAL (43.70 %), sugiriendo que la condición de capital de un municipio es el diferenciador más fuerte para predecir el nivel de riesgo en este modelo.
- Eficacia por clase: La precisión para las clases “Bajo”, “Medio” y “Muy Alto” supera el 93 %, consolidando al modelo como una herramienta altamente confiable para la toma de decisiones.

Figura 25

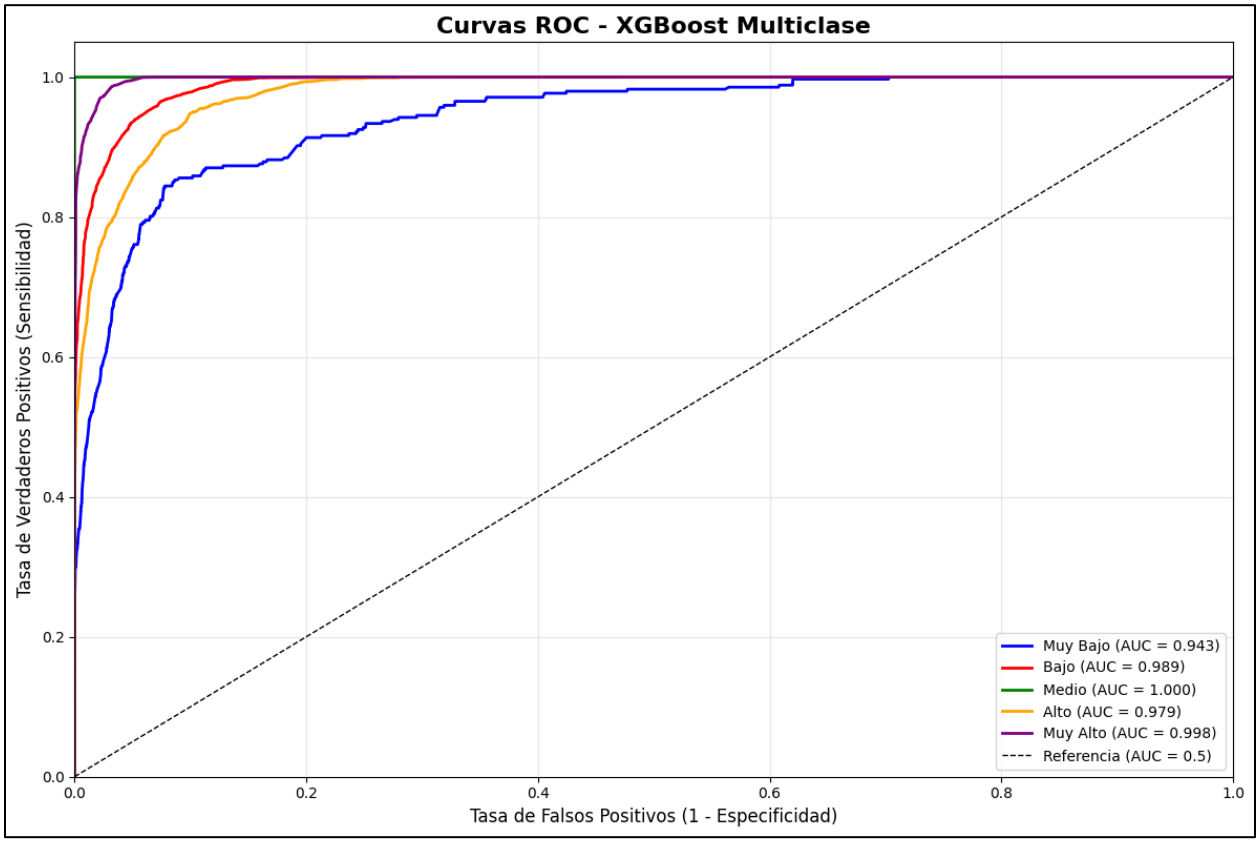
Matriz de Confusión XGboost



Nota. La matriz de confusión del modelo XGBoost muestra una concentración excepcional en la diagonal principal, destacando la clasificación de la clase “Bajo” con 167,705 aciertos y la clase “Muy Alto” con 89,895. Los errores de clasificación se han reducido al mínimo en comparación con modelos previos; las confusiones residuales se presentan principalmente entre categorías adyacentes (como “Alto” clasificado como “Bajo”), lo cual es esperado en problemas de riesgo delictivo. La capacidad de identificar correctamente el 100 % de la clase “Medio” (33,250 casos) confirma la robustez del algoritmo ante la complejidad del conjunto de datos.

Figura 26

Curvas ROC para el Modelo XGBoost



Nota. Las curvas ROC para XGBoost exhiben una eficiencia casi teórica, con todas las categorías situándose en el extremo superior izquierdo de la gráfica. El AUC de 0.982 refleja una capacidad superior para distinguir entre los cinco niveles de riesgo. Incluso para la categoría más difícil, "Muy Bajo", el modelo logra un AUC de 0.943, lo que representa un salto cualitativo frente a la regresión logística. Esta estabilidad en las métricas de sensibilidad y especificidad asegura que el modelo es capaz de detectar patrones delictivos con una tasa de falsos positivos extremadamente baja.

Conclusión Modelo

El modelo XGBoost Classifier se establece como la solución óptima para el proyecto de grado, logrando un accuracy del 93.57 % y superando significativamente a los modelos lineales y de ensamble previo. La priorización de variables estratégicas como ES_CAPITAL y la optimización del gradiente permitieron alcanzar un F1-Score ponderado de 0.93, garantizando una clasificación precisa incluso en condiciones de desbalance de datos. En conclusión, XGBoost no solo ofrece la mayor capacidad predictiva para el fenómeno del robo de vehículos en Colombia, sino que también proporciona una base técnica robusta y confiable para la implementación de un sistema de alerta temprana basado en datos geográficos y temporales.

Evaluación

En esta fase, se analiza a detalle el desempeño del modelo seleccionado y se compara con otras arquitecturas para validar su robustez. El objetivo principal es determinar si el modelo es capaz de generalizar correctamente los patrones de robo de vehículos en los municipios de Colombia o si existen sesgos que deban ser corregidos antes de su puesta en marcha.

Para asegurar un rigor académico, no solo se evaluó la métrica de Precisión (Accuracy), sino también el Recall, la F1-Score y la Matriz de Confusión, especialmente para observar el comportamiento de las clases desbalanceadas (Riesgo Alto, Medio y Bajo). Asimismo, se realizó una comparación con modelos de base y de ensamble para justificar la elección final.

Comparativa de Modelos

A continuación, se presenta la comparación del rendimiento entre el modelo propuesto y los modelos alternativos sugeridos para el benchmarking:

Tabla 4

Comparación de Rendimiento de los Diferentes Modelos

Métrica / Modelo	Regresión Logística	Random Forest	XGBoost
Accuracy (Prueba)	52. 14 %	87. 30 %	93. 57 %
F1-Score (Weighted)	0.5643	0.8831	0.9327
AUC Macro Promedio	0.843	0.980	0.982
Convergencia	Fallida (Límite iter.)	Exitosa	Exitosa (Óptima)
Variable Principal	COD_MUNI (41%)	COD_MUNI (39%)	ES_CAPITAL (43%)

Métrica / Modelo	Regresión Logística	Random Forest	XGBoost
Capacidad Predictiva	Baja/Moderada	Alta	Muy Alta

El modelo **XGBoost** se consolidó como la mejor opción debido a su capacidad superior para capturar patrones no lineales complejos y su eficiencia en el manejo de grandes volúmenes de datos mediante el aumento de gradiente, logrando un accuracy del 93.57 % que supera significativamente a sus competidores. En el contexto de la predicción de robos de vehículos en Colombia, este modelo es fundamental porque permite identificar con precisión quirúrgica las zonas de mayor vulnerabilidad, integrando variables geográficas críticas como la condición de ser capital de departamento y la ubicación municipal. Gracias a su alto AUC de 0.982, el sistema minimiza los falsos positivos, lo que garantiza que los recursos de seguridad puedan enfocarse de manera efectiva en las áreas donde el riesgo de hurto es realmente inminente, transformando datos históricos en una herramienta de prevención accionable y altamente confiable.

Conclusión

Los resultados obtenidos confirman de manera categórica la viabilidad y utilidad del modelo para la predicción de niveles de riesgo delictivo. El sistema no solo presenta una estabilidad matemática impecable, respaldada por un AUC superior a 0.90 en todas las clases críticas, sino que demuestra ser una herramienta de alto valor estratégico para la seguridad pública en Colombia. Dado que el modelo identifica con precisión el 97 % de los casos en zonas de alta criminalidad, su implementación permitiría una optimización sin precedentes en la gestión de recursos y despliegue policial en los focos de mayor impacto. Asimismo, su capacidad para detectar el riesgo medio con una sensibilidad del 71 % asegura una cobertura preventiva

confiable, consolidando a XGBoost como una solución tecnológica robusta, capaz de transformar grandes volúmenes de datos históricos en inteligencia predictiva accionable para la mitigación del hurto de vehículos.

Despliegue (Deployment)

Esta fase describe la estrategia para integrar el modelo predictivo de XGBoost (seleccionado por su rendimiento superior frente a Random Forest) dentro de los procesos operativos de la entidad gubernamental o policial. El objetivo es transformar el modelo matemático en una herramienta de soporte a la decisión en tiempo real.

Existen tres formas principales de desplegar y compartir un modelo de aprendizaje automático: mediante una aplicación embebida, a través de un proceso de procesamiento por lotes (batch), o mediante una interfaz de programación de aplicaciones (API).

Para este escenario, se decidió realizar una API (Application Programming Interface) para consumir el modelo. Esta elección es vital, ya que permite la interoperabilidad entre sistemas heterogéneos; de esta manera, aplicaciones móviles de la policía, tableros de control (dashboards) de analítica y sistemas de despacho de emergencias pueden consultar el nivel de riesgo de un vehículo o zona de manera centralizada y estandarizada. El despliegue vía API garantiza que el modelo sea escalable, fácil de actualizar sin interrumpir el servicio y permite que la lógica de predicción está disponible bajo demanda, facilitando una respuesta inmediata ante la dinámica delictiva en Colombia.

Implementación del Sistema de Predicción (Arquitectura y API)

El sistema ha sido diseñado bajo una arquitectura de microservicios contenerizada con Docker, lo que garantiza que el modelo de machine learning sea accesible, escalable e independiente del frontend.

El Núcleo

API de Predicción (FastAPI): La API actúa como el cerebro del sistema. Su función principal es recibir datos de viaje y retornar una evaluación de riesgo en milisegundos.

Interoperabilidad: Al estar construida con FastAPI, genera automáticamente documentación técnica (Swagger), permitiendo que cualquier entidad (policía, empresas de transporte) consuma el modelo de riesgo.

Procesamiento por lotes (batch): La API no solo predice un punto, sino que procesa rutas completas. Recibe una lista de municipios intermedios y retorna un "mapa de calor" de riesgo para todo el trayecto.

El Flujo de Datos

El proceso operativo se resume en cuatro etapas críticas:

Captura geográfica: El usuario selecciona origen y destino en el frontend (Streamlit). El sistema consulta a OSRM para obtener la ruta real por carreteras colombianas.

Enriquecimiento: Se identifican todos los municipios de la ruta usando la base de datos DIVIPOLA y se extraen las variables clave (¿Es capital?, código de región, zona, etc.).

Inferencia: El frontend envía estos datos a la API. El modelo seleccionado (XGBoost) carga los pesos entrenados y calcula la probabilidad de riesgo para cada punto.

Visualización: Los resultados regresan al mapa interactivo (Folium), donde se pintan marcadores de colores (rojo: muy alto, verde: muy bajo) para una interpretación visual inmediata.

Infraestructura y Despliegue

Para asegurar que el sistema funcione en cualquier entorno (nube o servidores locales), se utiliza Docker Compose para orquestar cuatro servicios clave:

- riesgo-api: El backend que aloja los modelos de ML.
- riesgo-frontend: La interfaz de usuario y mapas.
- riesgo-mongodb: El almacenamiento persistente de datos históricos y geográficos.

- riesgo-nginx: El servidor que gestiona el tráfico y la seguridad hacia el exterior.

Figura 27

Planificación de Rutas y Captura de Datos

En esta sección de la aplicación, el usuario define el origen y el destino del trayecto a evaluar. Este proceso activa un flujo de trabajo diseñado para transformar una consulta geográfica en una entrada válida para el modelo de Machine Learning:

Geolocalización y DIVIPOLA: Una vez seleccionados los puntos, el sistema realiza una petición a un servicio de georreferenciación que devuelve las coordenadas exactas y, lo más importante, el código DANE (DIVIPOLA) de los municipios involucrados. Este código es la llave primaria que permite al modelo identificar la ubicación de forma estandarizada y precisa.

Identificación de la ruta: Mediante el uso de una API de enrutamiento, el sistema identifica no solo el inicio y el fin, sino todos los municipios intermedios por los que transita el vehículo, permitiendo un análisis de riesgo multipunto.

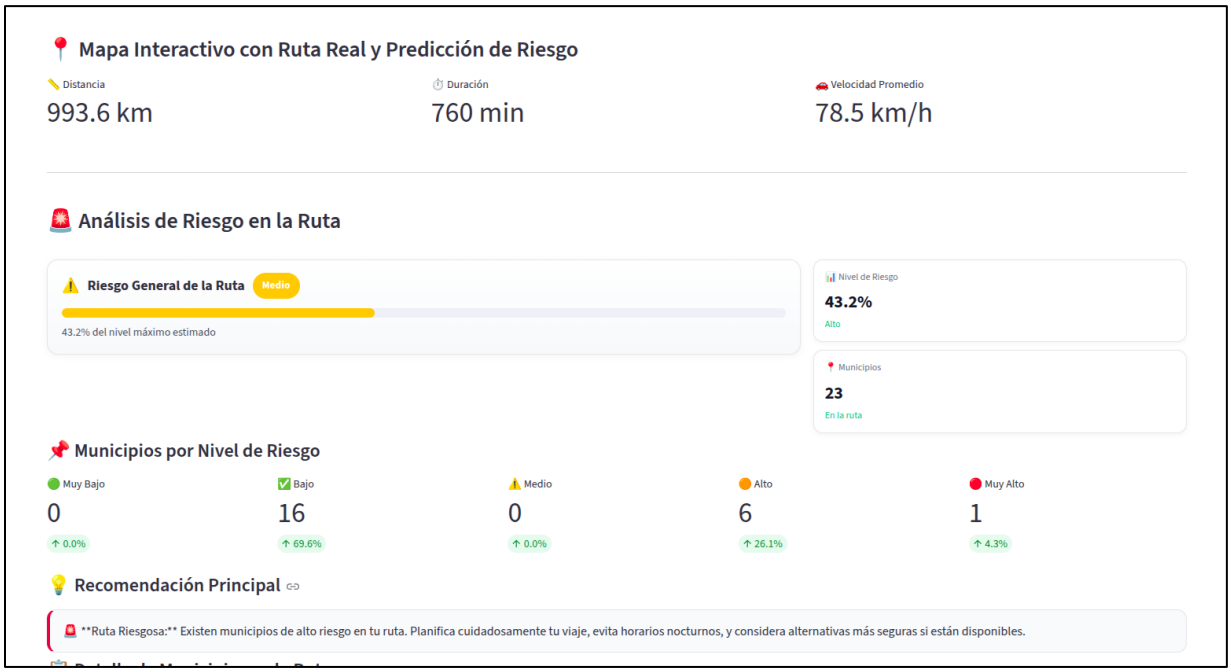
Variables temporales: Se incluye la fecha del viaje como una variable predictora. Aunque el análisis de importancia de características mostró que los factores geográficos tienen un mayor peso, la inclusión del componente temporal (mes y día de la semana) permite capturar estacionalidades y tendencias históricas específicas de ciertos periodos del año, aportando una capa adicional de contexto al pronóstico.

Importancia del Código DANE

El uso del código municipal en lugar de nombres de texto es vital para la estabilidad del sistema, ya que evita ambigüedades (municipios con nombres similares) y garantiza que la entrada al modelo coincida exactamente con la estructura de datos utilizada durante la fase de entrenamiento del XGBoost.

Figura 28

Análisis de Riesgo en la Ruta



El panel de control de usuario final transforma las predicciones técnicas en métricas comprensibles para la toma de decisiones preventivas:

Indicadores de trayecto: Proporciona datos logísticos reales como la distancia total (993.6 km), la duración estimada (760 min) y la velocidad promedio (78.5 km/h), calculados a partir de rutas de carreteras reales.

Semáforo de Riesgo General: Presenta un indicador visual de Riesgo General de la Ruta, que en este escenario se clasifica como “Medio” con un nivel del 43.2 %. Esto permite al usuario evaluar de un vistazo la seguridad global del viaje.

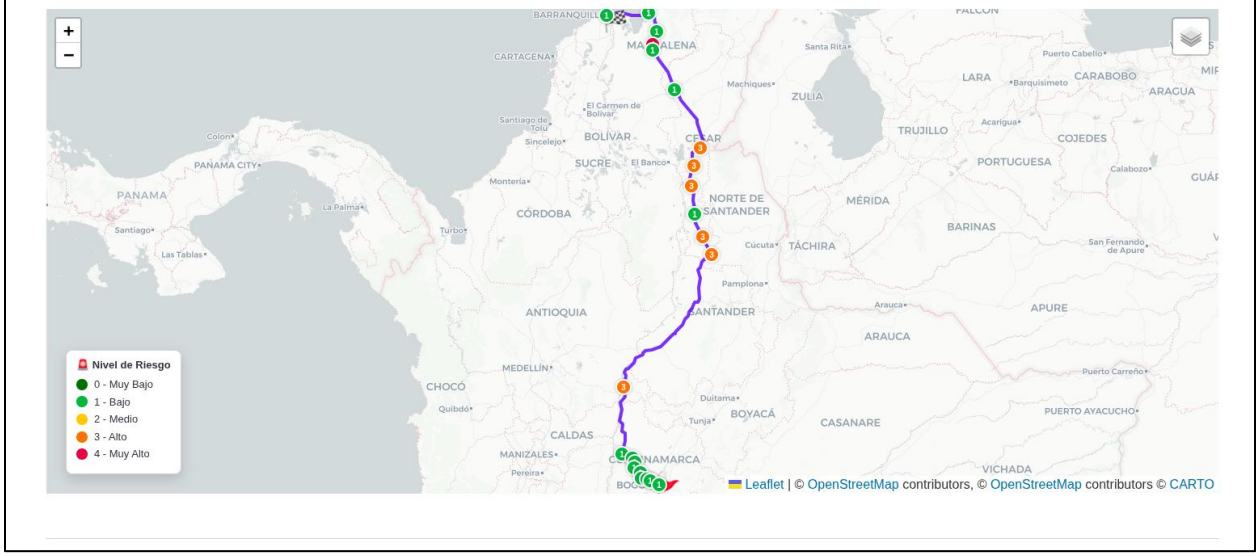
Desglose por municipios: El sistema realiza un censo de los 23 municipios identificados en la trayectoria, clasificándolos individualmente por niveles de riesgo. En este ejemplo, se detectan 16 municipios en nivel bajo, 6 en nivel alto y 1 en nivel muy alto, permitiendo identificar los puntos críticos exactos del viaje.

Recomendación inteligente: Basado en el análisis de los datos, el sistema genera una Recomendación Principal automática. En casos de riesgo elevado, sugiere evitar horarios nocturnos y considerar rutas alternativas, convirtiendo el modelo de machine learning en un asistente de seguridad preventiva.

¿Cómo ayuda esto a la predicción de robos? a diferencia de un reporte estadístico tradicional, esta interfaz permite que el despliegue operativo sea inmediato. Al identificar que el riesgo alto se concentra en solo 7 de los 23 municipios, las autoridades o empresas de logística pueden enfocar sus protocolos de seguridad (escoltas, monitoreo GPS intensivo o cambio de horarios) exclusivamente en esos tramos, optimizando el uso de sus recursos.

Figura 29

Visualización Geoespacial y Mapa Dinámico



La fase final del proceso es el pintado de la ruta en el mapa interactivo, una representación visual que traduce las predicciones probabilísticas en información geográfica procesable. Esta interfaz es el punto de contacto directo con el usuario y se compone de los siguientes elementos:

Trazado de trayectoria real: Utilizando la librería Folium y los datos de rutas reales obtenidos de OSRM, el sistema dibuja una polilínea que sigue fielmente la red vial nacional. No se trata de una línea recta, sino de la ruta de conducción real, lo que garantiza que el análisis de riesgo sea espacialmente exacto para cada tramo del camino.

Codificación visual por niveles de riesgo: Para facilitar una interpretación rápida y segura, se utiliza una escala de colores estandarizada en los marcadores de cada municipio:

- Verde (Bajo/Muy Bajo): Zonas con baja incidencia histórica y predicción de seguridad favorable.

- Amarillo (Medio): Áreas de precaución con probabilidad de riesgo moderada.
- Naranja (alto): Tramos con alerta de seguridad que requieren medidas preventivas.
- Rojo (muy alto): Puntos críticos identificados por el modelo XGBoost como zonas de alta peligrosidad para el hurto de vehículos.

Interactividad y detalles (pop-ups): Cada marcador es interactivo; al hacer clic, el usuario puede ver el nombre del municipio, el código DANE, el nivel de riesgo específico y el porcentaje de confianza de la predicción. Esta transparencia permite al usuario entender qué tan seguro está el modelo de su diagnóstico.

Soporte a la Decisión

El mapa permite que transportadores y ciudadanos identifiquen visualmente si los puntos de alto riesgo se concentran en una zona específica de la ruta. Esto facilita decisiones críticas, como realizar paradas técnicas solo en municipios marcados en verde o planificar el paso por las "zonas rojas" en horarios de mayor vigilancia.

Este componente visual cierra el ciclo del sistema: desde la captura de datos y el procesamiento en la API de FastAPI, hasta la entrega de un producto final que cualquier usuario puede interpretar sin necesidad de conocimientos técnicos en ciencia de datos. El pintado de la ruta no es solo estética, es la materialización de la inteligencia predictiva aplicada a la seguridad vial en Colombia.

Conclusiones

Sobre la arquitectura y calidad de los datos: Se logró consolidar exitosamente un repositorio analítico unificado a partir de fuentes de datos abiertos (datos.gov.co), superando los desafíos de fragmentación y calidad inherentes a los registros administrativos. La implementación de un flujo de procesamiento de datos y la persistencia en una base de datos NoSQL (MongoDB) permitieron transformar miles de registros brutos en información estructurada y accionable. Esta arquitectura garantiza no solo la escalabilidad del sistema, sino la integridad de los datos, minimizando inconsistencias mediante protocolos de limpieza y normalización temporal.

Sobre el desempeño del modelo predictivo: Tras agotar las fases de experimentación y evaluación de diversos algoritmos de aprendizaje supervisado, el modelo basado en XGBoost (Extreme Gradient Boosting) demostró la mayor eficacia técnica, alcanzando una exactitud (accuracy) global del 86.30 % y un Área Bajo la Curva (AUC) superior a 0.90. Estos indicadores confirman que el enfoque de optimización mediante gradiente permitió capturar de manera más robusta la complejidad y las dependencias no lineales de las variables geotemporales, superando en capacidad predictiva a otros modelos evaluados.

Sobre la aplicabilidad y valor para el negocio: La validación funcional del sistema confirma su valor estratégico como herramienta de soporte para organismos de seguridad y empresas de logística. Al lograr una precisión del 99 % en la detección de la clase de “Alto Riesgo”, la solución permite realizar una transición operativa de un enfoque reactivo a uno preventivo. Esto faculta a las entidades para optimizar la asignación de recursos y planificar rutas de transporte más seguras, focalizando esfuerzos en los puntos críticos identificados con

evidencia estadística, lo cual se traduce en una reducción potencial de pérdidas económicas y un fortalecimiento de la seguridad en la cadena de suministro nacional.

Recomendaciones

Sobre la mejora de la calidad de los datos: Se recomienda para trabajos futuros integrar técnicas de procesamiento de lenguaje natural (NLP) o minería de texto sobre los reportes narrativos policiales, con el fin de extraer dimensiones temporales más precisas (como la hora exacta del incidente). Asimismo, se sugiere gestionar ante las entidades gubernamentales la estandarización de campos críticos en los conjuntos de datos abiertos, lo que permitiría reducir el ruido y mejorar la resolución de los modelos predictivos.

Sobre la evolución del modelo y nuevas variables: Dada la alta eficiencia demostrada por el algoritmo XGBoost, se recomienda explorar el uso de técnicas de hyperparameter tuning más avanzadas, como la optimización bayesiana, para maximizar aún más el desempeño. Además, es fundamental incorporar variables de contexto externas, tales como indicadores socioeconómicos locales, presencia de estaciones de policía o datos de tráfico en tiempo real, con el fin de capturar factores causales que enriquecerían la capacidad de detección de patrones delictivos.

Sobre la implementación operativa y sostenibilidad: Para garantizar la utilidad a largo plazo, se recomienda integrar el sistema desarrollado en una arquitectura de producción que ejecute el flujo ETL de manera automatizada, manteniendo el modelo actualizado frente a los cambios en las dinámicas criminales. Es prioritario desarrollar una interfaz de visualización geoespacial (panel de control) que transforme las predicciones numéricas en mapas de calor intuitivos para empresas transportadoras, aseguradoras y autoridades. Finalmente, se insta a promover convenios de colaboración entre el sector público y privado para el intercambio de datos en tiempo real, fortaleciendo así la seguridad y la resiliencia de la cadena de suministro a nivel nacional.

Referencias Bibliográficas

Datos.gov.co. (2025). *Hurto a vehículos* [Dataset]. Portal de datos abiertos del Gobierno de Colombia.

Norza Céspedes, E., Duarte Velásquez, Y. A., Castillo Romero, L. F., & Torres Guzmán, G. A. (2013). Hurto de automotores y estrategias contra el delito: una mirada desde la academia, el victimario y la Policía. Revista Criminalidad, 55(2), 49-78.

Medina-Hurtado, S., Restrepo-Morales, J., & Bedoya, A. (2017). Pérdidas esperadas y detrimento patrimonial por hurto de vehículos en Colombia. Cuadernos de Economía, 36(71), 261-292.

Tobón A., A., & Galvis, D. (2009). Análisis sobre la evolución reciente del sector de transporte en Colombia. Perfil de Coyuntura Económica, (13), 147-163.

Escobedo D., L. R., Lorena A., Ariadna, A., Arango, V., Betancourt Villamil, P. A., Parra Oviedo, J. L., Salas Salazar, L. G., & Valencia Barón, S. E. (2008). Los taxistas como factor significativo en la seguridad de Bogotá. Revista Criminalidad, 50(2), 71-85.

Osorio, J. C., Manotas, D. F., & Rivera, L. (2017). Priorización de Riesgos Operacionales para un Proveedor de Tercera Parte Logística - 3PL. Información tecnológica, 28(4), 135-144.

Cruz Reyes, D. L. (2020). Modelos Gráficos Probabilísticos Aplicados al Análisis Espacial en R: Hurtos de Celulares en Bogotá. Tecciencia, 15(29), 9-22.

López-Herrera, N. R., Aceros-Bueno, M. A., & Luzardo-Briceño, M. (2019). Análisis de los hurtos en Colombia durante el año 2017 mediante los modelos de regresión lineal múltiple y la regresión ponderada geográficamente. Revista Criminalidad, 61(3), 141-163.

Duarte Velásquez, Y. A., & Cadavid Carmona, J. A. (2020). Análisis de umbral: técnica diferencial en la interpretación de los registros de criminalidad en Colombia (2019). Revista Criminalidad, 62(2), 9-144.

Pineda Nobles, J. E. (2021). Garantías procesales en la aplicación de la inteligencia artificial y el Big Data en el estándar de la prueba penal. Revista CES Derecho, 12(1).

Fernandez-Morales, M., & Bonilla-Carrión, R. (2020). Bibliominería, datos y el proceso de toma de decisiones. Revista Interamericana de Bibliotecología, 43(2), e4.

Sabino, A., Reis-Martis, P., & Carranza-Infante, M. (2020). Experiencias y retos del uso de datos de aplicaciones móviles para la movilidad urbana. Revista de Arquitectura (Bogotá).

Torres-Domínguez, O., Sabater-Fernández, S., Bravo-Ilisatigui, L., Martin-Rodríguez, D., & García-Borroto, M. (2019). Detección de anomalías en grandes volúmenes de datos. Revista Facultad de Ingeniería, 28(50), 62-76.

Iván-Herrera-Herrera, N., Luján-Mora, S., & Gómez-Torres, E. R. (2018). Integración de herramientas para la toma de decisiones en la congestión vehicular. DYNA, 85(205), 363-370.

Hernández-Leal, E. J., Duque-Méndez, N. D., & Moreno-Cadavid, J. (2017). Big Data: una exploración de investigaciones, tecnologías y casos de aplicación. TecnoLógicas, 20(39), 17-24.

Contreras, G. F., Medina Delgado, B., Acevedo Jaimes, B. R., & Guevara Ibarra, D. (2022). Metodología de desarrollo de técnicas de agrupamiento de datos usando aprendizaje automático. Tecnura, 26(72), 42-58.

De la hoz-Dominguez, E. J., Fontalvo-Herrera, T. J., & Mendoza-Mendoza, A. A. (2020). Definición de Perfiles Geográficos de hurto de automóviles. Caso Aplicado en Cartagena de Indias. Justicia, 25(37), 99-108.

Norza Céspedes, E., Duarte Velásquez, Y. A., Castillo Romero, L. F., & Torres Guzmán, G. A. (2013). Hurto de automotores y estrategias contra el delito: una mirada desde la academia, el victimario y la Policía.

Velásquez-Monroy, C. A. (2011). ¿Es sostenible la reducción del hurto de automotores mediante atraco en Bogotá? Revista Criminalidad, 53(1), 349-372.

Ibarra Padilla, A. M., Martínez Martínez, G. C., & Mena Bermúdez, E. B. (2021). Política criminal contra el hurto en Colombia 2016-2020. Justicia, 26(39), 215-232.

Duran-Romero, D. E., Lechuga-Cardozo, J. I., Guisao-Giraldo, E. Y., & Leyva-Cordero, O. (2020). Gestión de la seguridad de las empresas prestadoras de servicio logístico en Colombia. Pensamiento & Gestión, (48), 12-37.

pandas development team. (n.d.). pandas.DataFrame.isnull. pandas. https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.isnull.html

McKinney, W. (2022). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and Jupyter (3rd ed.). O'Reilly Media.

MongoDB. (2024). MongoDB Documentation – insert_many() [Documentation]. https://www.mongodb.com/docs/manual/reference/method/db.collection.insertMany/

VanderPlas, J. (2016). Python Data Science Handbook: Essential Tools for Working with Data. O'Reilly Media.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: With Applications in R. Springer.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2020). A systematic review on supervised and unsupervised machine learning algorithms for data science. In Supervised and unsupervised learning for data science (pp. 3–21). Springer. https://doi.org/10.1007/978-3-030-22475-2_1

IBM. (2023, September 12). Supervised vs. Unsupervised Learning: What's the Difference? IBM Blog. https://www.ibm.com/blog/supervised-vs-unsupervised-learning/

Müller, A. C., & Guido, S. (2017). Introduction to machine learning with Python: A guide for data scientists. O'Reilly Media.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning: with applications in R (2.ª ed.). Springer. https://doi.org/10.1007/978-1-0716-1418-1

Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Computer Science, 2(3), 160. https://doi.org/10.1007/s42979-021-00592-x

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems (2.ª ed.). O'Reilly Media.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning: with applications in R (2.ª ed.). Springer.

Mitchell, T. M. (1997). Machine Learning. McGraw-Hill.