



Modelo Predictivo de Hurto de Motocicletas en las Capitales Departamentales de Colombia

Nicolas Pernet Pérez

Juan Sebastián Gutiérrez Cudris

David Bernardo Vallejo Villegas

Universidad Ean

Facultad de Ingeniería

Programa de Ingeniería de Sistemas

Bogotá, Colombia

26/02/2026

Modelo Predictivo de Hurto de Motocicletas en las Capitales Departamentales de Colombia

Nicolas Pernet Pérez

Juan Sebastián Gutiérrez Cudris

David Bernardo Vallejo Villegas

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

Ingeniero de Sistemas

Directora:

Lina María Chacón Rivera

Universidad Ean

Facultad de Ingeniería

Programa de Ingeniería de Sistemas

Bogotá, Colombia

26/02/2026

Tabla de Contenido

Resumen.....	6
Abstract.....	7
1. Introducción.....	8
1.1 Planteamiento del problema.....	8
1.2 Objetivo general y objetivos específicos.....	9
1.3 Pregunta de investigación e hipótesis.....	9
2. Marco Teórico.....	11
2.1 Teorías criminológicas sobre la distribución espaciotemporal del delito...11	
2.2 Modelado predictivo aplicado a la seguridad urbana.....	12
2.3 Contexto del hurto de motocicletas en Colombia.....	14
3. Estado del Arte.....	16
3.1 Metodología de la revisión sistemática (PRISMA).....	16
3.2 Bases de datos consultadas y ecuación de búsqueda.....	16
3.3 Criterios de inclusión y exclusión.....	17
3.4 Diagrama PRISMA.....	18
3.5 Análisis crítico y comparativo de los estudios seleccionados.....	19
4. Metodología.....	26
4.1 Enfoque y alcance de la investigación.....	26

4.2 Fuente de datos y población de estudio.....	26
4.3 Fases del desarrollo metodológico.....	27
Fase 1. Consolidación de datos.....	27
Fase 2. Análisis exploratorio de datos (EDA).....	27
Fase 3. Entrenamiento de modelos predictivos.....	28
Fase 4. Evaluación de modelos predictivos.....	28
Fase 5. Desarrollo de la aplicación web.....	29
4.4 Instrumentos.....	29
4.5 Técnicas para el análisis de la información.....	30
5. Resultados.....	31
5.1 Consolidación del dataset.....	31
5.2 Análisis exploratorio de datos.....	31
5.3 Comparación de modelos predictivos.....	32
5.4 Clasificación por niveles de riesgo.....	32
5.5 Análisis complementario.....	33
5.6 Análisis de resultados y cumplimiento de objetivos.....	34
6. Propuesta de Solución.....	35
6.1 Descripción de la solución (MotoSegura).....	35
6.2 Cumplimiento de requerimientos.....	35
6.3 Desempeño y sostenibilidad.....	35
6.4 Costos.....	36

7. Análisis de Requerimientos de Ingeniería.....37

7.1 Intención del producto.....37

7.2 Verificación de parámetros de diseño.....37

7.3 Estimación de características y especificaciones.....39

7.4 Alcance de la solución de ingeniería.....39

7.5 Comparación con productos y estándares similares.....40

7.6 Análisis de costos.....40

7.7 Análisis de sostenibilidad.....42

7.8 Marco legal y regulatorio.....43

8. Discusión.....45

9. Plan de Divulgación.....46

10. Conclusiones.....47

Referencias.....48

Índice de Tablas

Tabla 1. Criterios de inclusión y exclusión aplicados en la revisión sistemática

Tabla 2. Comparación de estudios seleccionados en la revisión sistemática

Tabla 3. Comparación de métricas de error entre modelos predictivos

Tabla 4. Verificación de parámetros de diseño del producto

Índice de Figuras

Figura 1. Diagrama de flujo PRISMA del proceso de selección de estudios (2018–2025)

Resumen

El hurto de motocicletas en Colombia constituye un problema creciente de seguridad urbana con implicaciones socioeconómicas significativas que afecta tanto a la percepción de seguridad ciudadana como a sectores productivos dependientes de este medio de transporte. A pesar de la disponibilidad de estadísticas oficiales publicadas por la Policía Nacional, el uso de estos datos se ha limitado principalmente al análisis retrospectivo, sin que existan herramientas que permitan anticipar el riesgo de hurto en ciudades y periodos específicos. El presente estudio tiene como objetivo desarrollar un modelo predictivo que, a partir de registros oficiales de hurto de motocicletas correspondientes al periodo 2010-2025, permita estimar el nivel de riesgo en las 32 capitales departamentales de Colombia para periodos y ciudades definidos por el usuario. La investigación adopta un enfoque cuantitativo, de alcance descriptivo-correlacional, con un diseño no experimental longitudinal. Se emplea un conjunto de datos de más de 400.000 registros de hurto de motocicletas, sobre los cuales se realiza un proceso de limpieza, transformación y extracción de características temporales. Como técnica de modelado se evalúan modelos de series temporales (Prophet) y algoritmos de aprendizaje automático (XGBoost), comparando su capacidad predictiva mediante métricas como MAE, RMSE y la precisión en la clasificación por niveles de riesgo. Los resultados se materializan en una aplicación web interactiva que presenta un mapa de Colombia con indicadores visuales de riesgo (verde, amarillo y rojo) para cada ciudad y periodo consultado. Se espera que esta herramienta apoye la toma de decisiones de motociclistas, viajeros, empresas del sector logístico y entidades encargadas de la seguridad pública.

Palabras clave: hurto de motocicletas, modelo predictivo, series temporales, seguridad urbana, Colombia, aprendizaje automático, análisis espaciotemporal.

Abstract

Motorcycle theft in Colombia represents a growing urban security problem with significant socioeconomic implications affecting both public safety perception and productive sectors that rely on this mode of transportation. Despite the availability of official statistics published by the National Police, the use of these data has been primarily limited to retrospective analysis, with no existing tools to anticipate theft risk in specific cities and time periods. This study aims to develop a predictive model that, based on official motorcycle theft records from 2010 to 2024, estimates the risk level across the 32 departmental capitals of Colombia for user-defined periods and cities. The research adopts a quantitative approach with a descriptive-correlational scope and a non-experimental longitudinal design. A dataset of 472,760 motorcycle theft records is used, undergoing cleaning, transformation, and temporal feature extraction processes. Time series models (Prophet) and machine learning algorithms (XGBoost) are evaluated as modeling techniques, comparing their predictive capacity through metrics such as MAE, RMSE, and risk-level classification accuracy. Results are delivered through an interactive web application, displaying a map of Colombia with visual risk indicators (green, yellow, and red) for each city and queried period. This tool is expected to support decision-making for motorcyclists, travelers, logistics companies, and public security agencies.

Keywords: *motorcycle theft, predictive model, time series, urban security, Colombia, machine learning, spatiotemporal analysis*

Introducción

El hurto de motocicletas constituye uno de los delitos patrimoniales de mayor frecuencia en Colombia y representa un problema de seguridad urbana con repercusiones económicas, sociales y de movilidad que trascienden el ámbito individual de las víctimas. Según datos de la Policía Nacional de Colombia, entre 2010 y 2025 se registraron más de 400.000 denuncias de hurto de motocicletas en el territorio nacional, con una tendencia creciente que pasó de aproximadamente 13.000 casos anuales en 2010 a cerca de 40.000 en 2023. Este fenómeno se concentra de manera diferenciada en las capitales departamentales, siendo Medellín, Bogotá, Cali, Popayán y Barranquilla las ciudades que reportan los volúmenes más elevados (Policía Nacional de Colombia, 2024).

Desde la perspectiva criminológica, la concentración espaciotemporal de los delitos ha sido explicada por marcos teóricos consolidados como la Teoría de las Actividades Rutinarias (Cohen & Felson, 1979), que postula la convergencia de un agresor motivado, un objetivo atractivo y la ausencia de un guardián capaz como condición necesaria para la ocurrencia del delito. De manera complementaria, la Teoría del Patrón Delictivo (P. J. Brantingham & P. L. Brantingham, 1993) describe cómo los delincuentes operan dentro de sus nodos de actividad y rutas habituales, lo que genera patrones espaciales identificables y, por tanto, predecibles. Estos enfoques, aplicados al contexto colombiano, sugieren que el hurto de motocicletas no es un fenómeno aleatorio, sino que sigue dinámicas temporales y geográficas susceptibles de ser modeladas mediante técnicas cuantitativas.

Sin embargo, a pesar de que la Policía Nacional publica estadísticas desagregadas por municipio, fecha, modalidad de hurto y otras variables, el uso de estos datos se ha concentrado mayoritariamente en reportes retrospectivos de carácter descriptivo. No se han identificado en la

literatura colombiana herramientas públicas que transformen dicha información en predicciones operativas orientadas a anticipar el riesgo de hurto en ciudades y periodos específicos. Esta brecha entre la disponibilidad de datos y su aprovechamiento predictivo constituye el problema central que motiva la presente investigación.

En este contexto, el objetivo general del estudio es desarrollar un modelo predictivo basado en registros oficiales de hurto de motocicletas (periodo 2010-2025) que permita estimar el nivel de riesgo en las 32 capitales departamentales de Colombia para periodos definidos por el usuario, materializando los resultados en una aplicación web interactiva con visualización geoespacial. Para alcanzar este objetivo, se plantean los siguientes objetivos específicos: (a) consolidar los datos oficiales de hurto de motocicletas publicados por la Policía Nacional, filtrando las 32 capitales departamentales y estructurando series temporales continuas por ciudad; (b) realizar un análisis exploratorio que identifique patrones temporales (mensuales, semanales, estacionales) y diferencias entre ciudades; (c) entrenar modelos predictivos de series temporales y aprendizaje automático para estimar la frecuencia de hurtos por ciudad y periodo; (d) evaluar el rendimiento de los modelos entrenados mediante métricas de error y precisión en la clasificación por niveles de riesgo; y (e) desarrollar una aplicación web que permita al usuario ingresar las ciudades y fechas de un viaje y visualice el nivel de riesgo en un mapa interactivo de Colombia.

La pregunta que orienta esta investigación se formula de la siguiente manera: ¿es posible desarrollar un modelo predictivo, basado en datos históricos oficiales de hurto de motocicletas en Colombia, que estime de manera confiable el nivel de riesgo para las capitales departamentales en un periodo futuro definido por el usuario?

A partir de los patrones identificados en el análisis exploratorio de los datos, se formulan dos hipótesis de trabajo. La primera hipótesis (H1) plantea que los modelos de series temporales y aprendizaje automático, alimentados con variables temporales derivadas de los registros históricos (mes, día de la semana, festividad, tendencia anual), producen predicciones de frecuencia de hurto con un error medio absoluto (MAE) significativamente inferior al de un modelo ingenuo basado en el promedio histórico de la ciudad. La segunda hipótesis (H2) sostiene que existe una diferencia estadísticamente significativa en la distribución temporal de los hurtos de motocicletas entre las capitales departamentales, lo que justifica el uso de modelos diferenciados por ciudad en lugar de un modelo único nacional.

Marco Teórico

El presente marco teórico establece los fundamentos conceptuales y teóricos que sustentan el desarrollo de un modelo predictivo de hurto de motocicletas en las capitales departamentales de Colombia. Para ello, se articulan tres ejes conceptuales: las teorías criminológicas que explican la distribución espaciotemporal del delito, los fundamentos del modelado predictivo aplicado a la seguridad urbana, y el contexto específico del hurto de motocicletas en Colombia.

Teorías Criminológicas sobre la Distribución Espaciotemporal del Delito

La comprensión de los patrones delictivos en entornos urbanos se ha beneficiado de marcos teóricos que abandonan las explicaciones centradas exclusivamente en el perfil del delincuente para incorporar las condiciones ambientales y situacionales que facilitan la ocurrencia del crimen. La Teoría de las Actividades Rutinarias, propuesta por Cohen y Felson (1979), constituye uno de los pilares de la criminología ambiental. Esta teoría postula que un delito ocurre cuando convergen en el espacio y el tiempo tres elementos: un agresor motivado, un objetivo adecuado (en este caso, la motocicleta) y la ausencia de un guardián capaz. Aplicada al contexto del hurto de motocicletas, esta teoría permite explicar por qué ciertos horarios, días de la semana y zonas urbanas presentan mayores frecuencias de hurto: la confluencia de motocicletas estacionadas sin vigilancia, en horarios de alta actividad comercial, crea las condiciones situacionales propicias.

De manera complementaria, la Teoría del Patrón Delictivo (P. J. Brantingham & P. L. Brantingham, 1993) sostiene que los delincuentes no seleccionan sus objetivos de forma aleatoria, sino que operan dentro de un espacio de actividad definido por sus nodos (residencia, trabajo, ocio), rutas habituales y las áreas de conciencia que se generan alrededor de estos. Esta

perspectiva implica que los hurtos de motocicletas se concentran en corredores urbanos predecibles, lo que fundamenta la viabilidad de un enfoque predictivo basado en patrones históricos. A nivel departamental, la teoría se extiende al reconocer que cada capital presenta una estructura socioespacial particular (densidad poblacional, distribución del comercio, infraestructura vial) que configura patrones delictivos diferenciados.

Un tercer aporte relevante proviene de la Teoría de la Desorganización Social (Shaw & McKay, 1942), que vincula las tasas de criminalidad con condiciones estructurales del territorio como la pobreza, la heterogeneidad étnica y la movilidad residencial. Aunque esta teoría fue desarrollada originalmente para el contexto norteamericano, su aplicación al caso colombiano resulta pertinente dada la heterogeneidad socioeconómica entre departamentos. Ciudades con mayores índices de informalidad laboral, menor presencia institucional o mayor desigualdad de ingresos podrían presentar condiciones que facilitan la ocurrencia del hurto de motocicletas, lo cual constituye una variable contextual relevante para la interpretación de los patrones observados.

Modelado Predictivo Aplicado a la Seguridad Urbana

El uso de técnicas de modelado predictivo para anticipar la ocurrencia de delitos ha ganado relevancia en la última década, impulsado por la disponibilidad de datos abiertos y los avances en aprendizaje automático. En el ámbito de la predicción criminal, se distinguen dos grandes familias de métodos: los modelos de series temporales y los algoritmos de aprendizaje supervisado.

Los modelos de series temporales, como ARIMA y Prophet (Taylor & Letham, 2018), son especialmente adecuados cuando el objetivo es capturar tendencias, estacionalidad y efectos cíclicos en datos secuenciales. Prophet, desarrollado por Meta, presenta ventajas para este tipo de

estudios al manejar de manera nativa la estacionalidad múltiple (semanal, mensual, anual), incorporar el efecto de festividades y ser robusto frente a datos faltantes y valores atípicos. En el dominio específico del hurto de motocicletas, Azhari y Utomo (2018) aplicaron el modelo ARIMAX con función de transferencia para predecir la tasa de hurtos en función del parque vehicular, obteniendo valores de MAPE de 32,30 y RMSE de 6,68. De manera análoga, Pranata et al. (2020) compararon modelos AR, MA y ARIMA para pronosticar el hurto de motocicletas en Lampung, Indonesia, identificando que el modelo MA(1) ofrece el mejor ajuste con un RMSE de 6,56.

Por otra parte, los algoritmos de aprendizaje supervisado como XGBoost (Chen & Guestrin, 2016) y LightGBM (Ke et al., 2017) permiten integrar múltiples variables predictoras (mes, día de la semana, festividad, tendencia, ciudad) en un marco de clasificación o regresión. Estos modelos han demostrado un rendimiento competitivo en tareas de predicción criminal, particularmente cuando los datos presentan heterogeneidad territorial y desbalance de clases. Chen y Guestrin (2016) mostraron que XGBoost supera a otros algoritmos en competencias de aprendizaje automático, mientras que Ke et al. (2017) reportaron que LightGBM alcanza precisiones comparables con tiempos de entrenamiento significativamente menores.

Las redes neuronales recurrentes, en particular las redes LSTM (Long Short-Term Memory), propuestas por Hochreiter y Schmidhuber (1997), constituyen una alternativa de mayor complejidad para el modelado de secuencias temporales. Estas arquitecturas son capaces de capturar dependencias de largo plazo y patrones no lineales, lo que las hace potencialmente aplicables a series delictivas con comportamientos complejos. Li y Law (2024) ofrecen una revisión exhaustiva de los modelos de aprendizaje profundo aplicados a la predicción de series temporales, incluyendo arquitecturas CNN, RNN, LSTM, GRU y Transformer, y concluyen que

el modelo SCINet alcanza el mejor desempeño en el conjunto de datos ETTh1. Sin embargo, el mayor requerimiento computacional y la necesidad de grandes volúmenes de datos por serie individual hacen que la aplicación de estas arquitecturas sea más adecuada para ciudades con registros abundantes.

Contexto del Hurto de Motocicletas en Colombia

Colombia ocupa una posición destacada en el mercado motociclistico de América Latina, con un parque automotor que superó los 10 millones de motocicletas registradas según datos del Registro Único Nacional de Tránsito (RUNT). Este crecimiento sostenido ha sido impulsado por factores como la accesibilidad económica de las motocicletas frente al automóvil, las deficiencias del transporte público en ciudades intermedias y el uso extendido como herramienta de trabajo en sectores de domicilios, mensajería y comercio minorista.

Paralelamente, el hurto de motocicletas ha experimentado un crecimiento sostenido. Los datos de la Policía Nacional muestran que los registros anuales pasaron de 13.354 en 2010 a 41.705 en 2023, lo que representa un incremento superior al 200 % en poco más de una década. Este fenómeno no se distribuye de manera homogénea: capitales como Medellín, Bogotá y Cali concentran más del 35 % del total nacional, mientras que capitales de departamentos con menor población como Mitú, Inírida y Puerto Carreño registran cifras marginales. Adicionalmente, se han identificado patrones temporales relevantes; por ejemplo, en Bogotá los días martes y miércoles presentan frecuencias de hurto significativamente superiores a los fines de semana, y los meses de septiembre y octubre muestran picos estacionales (Policía Nacional de Colombia, 2024).

Estudios previos en el contexto colombiano han abordado el problema desde perspectivas descriptivas y de análisis espacial. Campo Agredo (2017) desarrolló modelos de predicción del

hurto a personas en Bogotá mediante la aplicación de algoritmos de aprendizaje automático sobre datos abiertos. Desde una perspectiva metodológica comparada, Pranata et al. (2020) demostraron la viabilidad del modelado ARIMA para predecir series temporales de hurto de motocicletas en contextos con recursos de datos limitados, e Iqbal et al. (2025) propusieron métodos de clustering espaciotemporal para identificar zonas de alta concentración delictiva. No obstante, ninguno de estos antecedentes ha desarrollado una herramienta predictiva interactiva que permita a un usuario consultar el riesgo estimado para ciudades y periodos específicos en el contexto colombiano, lo cual constituye la brecha que el presente estudio pretende abordar.

Estado del Arte

Metodología de la Revisión Sistemática (PRISMA)

Con el propósito de garantizar un proceso riguroso, transparente y replicable en la identificación y selección de estudios previos, la presente revisión se estructuró de acuerdo con las directrices del protocolo PRISMA (Page et al., 2021). Esta metodología proporciona un marco estandarizado que permite documentar de forma sistemática cada etapa del proceso de búsqueda, cribado y selección de la literatura, asegurando la trazabilidad de las decisiones tomadas y la calidad de los estudios incluidos en la síntesis.

Bases de datos consultadas

La búsqueda se realizó en las siguientes bases de datos académicas: Scopus, Web of Science (WOS), IEEE Xplore y Google Scholar. Se seleccionaron estas fuentes por su cobertura en las áreas de ciencias de la computación, ingeniería, criminología y seguridad urbana. Adicionalmente, se consultaron repositorios institucionales colombianos para identificar trabajos de grado y tesis relacionados.

Ecuación de búsqueda

Se emplearon los siguientes términos y operadores booleanos, adaptados a la sintaxis de cada base de datos:

("motorcycle theft" OR "vehicle theft" OR "hurto de motocicletas") AND ("prediction" OR "forecasting" OR "machine learning" OR "deep learning" OR "time series") AND ("urban" OR "crime" OR "spatiotemporal")

Criterios de inclusión y exclusión

Se establecieron criterios de inclusión y exclusión para garantizar la pertinencia y calidad de los estudios seleccionados. La Tabla 1 presenta de forma detallada cada criterio, su clasificación y la justificación correspondiente.

Tabla 1

Criterios de inclusión y exclusión aplicados en la revisión sistemática

Tipo	Criterio	Justificación
Inclusión	Artículos publicados entre 2018 y 2025	Garantizar la actualidad de los hallazgos y su relevancia para el estado actual de la investigación.
Inclusión	Documentos escritos en inglés o español	Asegurar la comprensión y análisis adecuado por parte de los investigadores.
Inclusión	Estudios que aborden la predicción de delitos patrimoniales (hurto de vehículos, motocicletas o delitos urbanos) mediante técnicas cuantitativas de modelado predictivo	Delimitar la revisión al campo específico de la investigación: predicción criminal con enfoque cuantitativo.
Inclusión	Publicados en revistas con revisión por pares o conferencias indexadas	Garantizar la calidad y el rigor científico de las fuentes seleccionadas.
Exclusión	Estudios puramente descriptivos sin componente predictivo	El proyecto se centra en el desarrollo de modelos predictivos, no en la caracterización descriptiva del fenómeno.
Exclusión	Artículos de opinión, editoriales y revisiones narrativas sin datos	Priorizar estudios con evidencia empírica y metodología replicable.

	empíricos	
Exclusión	Estudios que aborden exclusivamente delitos no patrimoniales (homicidio, lesiones, delitos sexuales)	Mantener la coherencia temática con el objeto de estudio: hurto de vehículos y motocicletas.

Nota. Elaboración propia con base en los lineamientos PRISMA.

Periodo temporal e idiomas

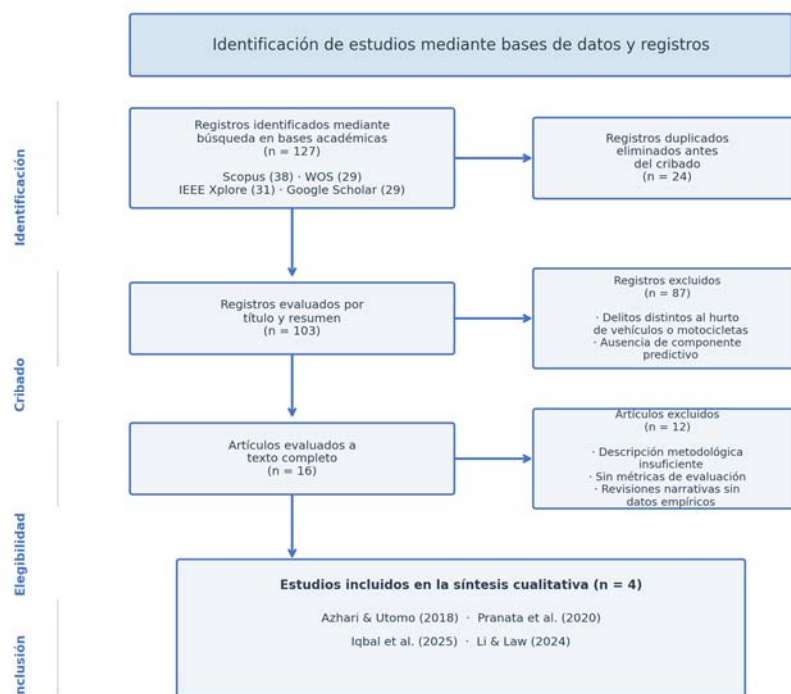
El periodo de búsqueda abarcó publicaciones entre enero de 2018 y febrero de 2026. Se incluyeron documentos en inglés y español.

Diagrama PRISMA

A continuación se presenta el flujo de selección de documentos según la metodología PRISMA. La búsqueda sistemática recuperó un total de 127 registros distribuidos en las cuatro bases de datos consultadas: Scopus (38), Web of Science (29), IEEE Xplore (31) y Google Scholar (29). Tras la eliminación de 24 duplicados, se sometieron a cribado por título y resumen 103 registros, de los cuales 87 fueron excluidos por no cumplir los criterios temáticos (estudio de delitos distintos al hurto de vehículos o motocicletas, ausencia de componente predictivo). Los 16 artículos restantes se evaluaron a texto completo, excluyéndose 12 por carecer de descripción metodológica suficiente, por no reportar métricas de evaluación del modelo, o por tratarse de revisiones narrativas sin datos empíricos. El proceso resultó en la inclusión de 4 estudios primarios que cumplen todos los criterios establecidos.

Figura 1

Diagrama de flujo PRISMA del proceso de selección de estudios (2018–2025)



Nota. Elaboración propia con base en el proceso de revisión sistemática siguiendo los lineamientos PRISMA (Page et al., 2021).

Análisis Crítico y Comparativo de los Estudios Seleccionados

A partir de la revisión realizada, se identifican las siguientes tendencias en la literatura sobre predicción de delitos urbanos con énfasis en el hurto de vehículos y motocicletas.

En cuanto a las tendencias metodológicas, los estudios revisados evidencian una transición progresiva desde modelos estadísticos clásicos hacia enfoques basados en aprendizaje automático y aprendizaje profundo. Azhari y Utomo (2018) y Pranata et al. (2020) ilustran el uso de modelos ARIMA y ARIMAX para la predicción del hurto de motocicletas, destacando su interpretabilidad y la posibilidad de aplicarlos con series temporales relativamente cortas. Por otra parte, Iqbal et al. (2025) demuestran que los algoritmos de clustering espaciotemporal —K-means, DBSCAN y agrupamiento jerárquico— permiten identificar zonas de alta concentración delictiva con precisión significativa, siendo el agrupamiento aglomerativo con 4 clústeres el que mejor separa los hotspots. Li y Law (2024) documentan la superioridad de arquitecturas profundas como SCINet sobre métodos clásicos en benchmarks estándar de series temporales, aunque advierten que su ventaja disminuye con conjuntos de datos pequeños.

Respecto a los marcos teóricos, la mayoría de los estudios se apoyan en la criminología ambiental, particularmente en la Teoría de las Actividades Rutinarias y la Teoría del Patrón Delictivo, para justificar la viabilidad de los enfoques predictivos. Se observa, sin embargo, un vacío en la integración de estas teorías con los resultados empíricos de los modelos, lo que sugiere una oportunidad para fortalecer la interpretabilidad de las predicciones.

En términos de resultados, los estudios reportan de manera consistente que la dimensión temporal (hora del día, día de la semana, mes) y la concentración espacial constituyen los predictores más relevantes. Algunos trabajos han encontrado que hasta el 30 % de los delitos se concentran en apenas el 10 % de las zonas urbanas, y que la incorporación de periodicidad semanal y estacional mejora significativamente el rendimiento predictivo.

Los principales vacíos de conocimiento identificados incluyen: la escasez de estudios enfocados específicamente en el hurto de motocicletas (la mayoría aborda el hurto vehicular en

general), la falta de aplicaciones interactivas orientadas al usuario final, y la limitada investigación en el contexto latinoamericano, particularmente colombiano.

Tabla 2

Comparación de estudios seleccionados en la revisión sistemática

Autor (es) / Año	Título del estudio	Objetivo principal	Metodología empleada	Variables analizadas	Principales hallazgos	Limitaciones reportadas	Relación con el proyecto actual
Azhar y Utomo (2018)	Prediction in the crime motorcycles of theft using ARIMAX-TFM with single input	Predecir la tasa de hurto de motocicletas incorporando como variable externa el número total de motocicletas registradas	ARIMAX con función de transferencia de entrada única (ARIMAX-TFM); evaluación mediante MAPE y RMSE	Registros históricos de hurto de motocicletas; número de motocicletas registradas como variable exógena	Mejor modelo ARIMAX con MAPE = 32,30 y RMSE = 6,68. La incorporación de la variable exógena mejora la precisión respecto a ARIMA	Contexto geográfico único (Indonesia); período de análisis corto; no incorpora variables socioeconómicas adicionales	Precedente directo en el uso de ARIMAX para el hurto de motocicletas; justifica el enfoque de series temporales adoptado

					puro		en este estudio
Pran ata et al. (2020)	Prediksi pencurian sepeda motor menggun akan model time series (Studi Kasus: Polres Kotabum i Lampung Utara)	Comparar modelos AR, MA y ARIMA para pronosticar el número mensual de hurtos de motocicleta s y proyectar los 6 meses siguientes	Modelos AR(1), AR(3), MA(1), ARIMA(1, 1,1) y ARIMA(3, 1,1); selección por criterio AIC y menor RMSE	Registros mensuale s de hurto de motocicle tas (2015- 2019), Polres Kotabumi , Lampung Utara, Indonesia	Modelo MA(1) con mejor ajuste: RMSE = 6,56 y AIC = 394,82. Pronóstic o para los 6 meses siguientes con intervalos de confianza del 95%	Conjunto de datos reducido (60 observacion es); ámbito geográfico local; no incorpora variables explicativas externas	Valida la aplicabili dad del modelado ARIMA a series de hurto de motocicle tas con datos limitados; metodolo gía replicada en el presente estudio para las capitales colombia nas

Iqbal et al. (2025)	A 2D clustering based hotspot identification approach for spatio-temporal crime prediction	Identificar zonas de alta concentración delictiva (hotspots) mediante clustering 2D espacio-temporal para mejorar la predicción de crimen	K-means, DBSCAN y agrupamiento jerárquico aglomerativo; validación con coeficiente de Silhouette e índice de Dunn	529.531 registros de crimen en Chicago (2001-2024); coordenadas geoespaciales; tipo de delito; marca temporal	Agrupamiento aglomerativo con 4 clústeres obtiene el mejor desempeño (coeficiente de Silhouette = 0,71); identifica hotspots persistentes con alta precisión predictiva	Requiere alta densidad de registros; no evaluado en contextos latinoamericanos; validado únicamente sobre datos de Chicago	Fundamenta el uso de análisis espacial para complementar los modelos temporales; relevante para la extensión futura del modelo propuesto hacia análisis intramunicipal
Li y Law	Deep learning	Revisar y comparar	Revisión sistemática	Conjuntos de datos	SCINet obtiene el	Los benchmarks	Fundamenta la

(2024)	models for time series forecasting: A review	sistemática mente arquitectura s de aprendizaje profundo (CNN, RNN, GNN, Transformer) para la predicción de series temporales	a con benchmarks empíricos sobre conjuntos de datos estándar (ETT, tráfico, clima, electricidad)	de series temporales s multivariadas: ETT, Solar Energy, Electricity, Traffic (PeMS), Weather, COVID-19, MIMIC	mejor desempeño en el conjunto ETTh1; los modelos Transformer destacan en horizontes largos; el rendimiento es sensible al dominio de aplicación	no incluyen series delictivas; resultados dependientes de hiperparámetros; requiere alta capacidad computacional	selección de arquitecturas LSTM como opción avanzada para el modelo predictivo; justifica el uso de Prophet sobre modelos Transformer dados los volúmenes de datos disponibles
--------	--	---	--	--	--	--	--

Nota. Los estudios fueron seleccionados a partir de la revisión sistemática descrita en la sección anterior, aplicando los criterios de inclusión y exclusión definidos en la Tabla 1. La información fue extraída directamente de los artículos originales consultados en Scopus, Web of Science, IEEE Xplore y Google Scholar.

Metodología

Enfoque y Alcance de la Investigación

La presente investigación adopta un enfoque cuantitativo, en tanto se fundamenta en la recolección y el análisis de datos numéricos provenientes de registros oficiales de hurto de motocicletas para identificar patrones, tendencias y relaciones entre variables que permitan desarrollar un modelo predictivo (Hernández-Sampieri & Mendoza, 2018). El alcance del estudio es descriptivo-correlacional: descriptivo porque caracteriza el comportamiento temporal y espacial del hurto de motocicletas en las capitales departamentales de Colombia, y correlacional porque busca identificar y cuantificar las relaciones entre variables temporales (mes, día de la semana, festividades, tendencia anual) y la frecuencia de hurtos, con el fin de generar predicciones fundamentadas.

El diseño de investigación es no experimental y longitudinal. Es no experimental porque no se manipulan variables; se trabaja con datos históricos tal como fueron registrados por la Policía Nacional. Es longitudinal porque se analiza la evolución del fenómeno a lo largo de un periodo de 16 años (2010-2025), lo que permite identificar tendencias de largo plazo y ciclos estacionales recurrentes.

Fuente de Datos y Población de Estudio

La fuente primaria de datos es el reporte oficial de hurto por modalidades de la Policía Nacional de Colombia, disponible como datos abiertos. El conjunto de datos contiene 624.698 registros que abarcan hurtos de motocicletas y automotores entre 2010 y 2025, con las siguientes variables: departamento, municipio, código DANE, armas o medios empleados, fecha del hecho, género y grupo etario de la víctima, tipo de hurto y cantidad.

La población de estudio comprende las 32 capitales departamentales de Colombia. Se seleccionan exclusivamente las capitales por tres razones: concentran la mayor proporción de registros de hurto del país, son los destinos más frecuentes para viajeros en motocicleta, y disponen de volúmenes de datos suficientes para el entrenamiento de modelos estadísticos confiables. Del total de registros, se filtran únicamente aquellos correspondientes al tipo de hurto "Artículo 239. Hurto Motocicletas", lo que arroja un subconjunto de 472.760 registros. De estos, 30 capitales cuentan con más de 200 registros en el periodo analizado, mientras que Mitú (33 registros) e Inírida (48 registros) presentan volúmenes insuficientes para el modelado individual, por lo que se les asignará un nivel de riesgo fijo basado en su promedio histórico.

Fases del Desarrollo Metodológico

El desarrollo del estudio se estructura en cinco fases secuenciales, cada una alineada con un objetivo específico.

Fase 1. Consolidación de datos

En esta fase se importa el conjunto de datos en formato Excel (.xlsx) utilizando la librería pandas de Python en el entorno Google Colab. Se aplican los siguientes procedimientos de limpieza: filtrado de registros correspondientes al hurto de motocicletas, selección de las 32 capitales departamentales mediante el identificador "(CT)" en el campo municipio, conversión de la variable de fecha al formato datetime, tratamiento de valores nulos y fechas fuera de rango, y agregación de los datos a nivel diario por ciudad (sumando la cantidad de hurtos por día). Los días sin registros se completan con valor cero para generar series temporales continuas.

Fase 2. Análisis exploratorio de datos (EDA)

Se realiza un análisis exploratorio exhaustivo que incluye: estadísticas descriptivas por ciudad (media, mediana, desviación estándar, percentiles), gráficos de tendencia temporal por ciudad y año, mapas de calor (heatmaps) que crucen mes del año con día de la semana para identificar patrones estacionales, diagramas de caja (boxplots) comparativos entre ciudades, y análisis de correlación entre las variables temporales construidas (mes, día de la semana, semana del año, indicador de festividad, indicador de fin de semana) y la frecuencia de hurtos. Para la identificación de festividades colombianas se utiliza la librería holidays de Python. Las visualizaciones se generan con las librerías matplotlib y seaborn.

Fase 3. Entrenamiento de modelos predictivos

Se evalúan dos aproximaciones de modelado complementarias. La primera emplea Prophet (Taylor & Letham, 2018), un modelo de series temporales desarrollado por Meta que descompone la serie en tendencia, estacionalidad y efecto de festividades. Se entrena un modelo Prophet independiente para cada capital con datos suficientes, incorporando las festividades colombianas como regresores adicionales. La segunda aproximación utiliza XGBoost (Chen & Guestrin, 2016), un algoritmo de gradient boosting que recibe como variables de entrada las características temporales extraídas (mes, día de la semana, semana del año, festividad, tendencia, ciudad codificada) y como variable objetivo la frecuencia diaria de hurtos.

Fase 4. Evaluación de modelos predictivos

Los datos se dividen temporalmente en un conjunto de entrenamiento (2010-2022) y un conjunto de validación (2023-2024), manteniendo el orden cronológico para respetar la naturaleza temporal del problema. Se evalúa el rendimiento de ambos modelos mediante las métricas de error medio absoluto (MAE), raíz del error cuadrático medio (RMSE) y,

adicionalmente, la precisión en la clasificación por niveles de riesgo, que constituye la salida principal de la aplicación.

La conversión de la predicción numérica a categorías de riesgo se realiza mediante umbrales basados en percentiles de la distribución histórica de cada ciudad: verde (percentil 0-33, riesgo bajo), amarillo (percentil 33-66, riesgo moderado) y rojo (percentil 66-100, riesgo alto). Esta estrategia permite que los umbrales se adapten a la realidad de cada ciudad, evitando comparaciones absolutas entre capitales con volúmenes de hurto muy diferentes.

Fase 5. Desarrollo de la aplicación web

La materialización de los resultados se realiza mediante una aplicación web interactiva desarrollada con Streamlit (framework de Python para aplicaciones de datos) y Folium (librería de Python para visualización de mapas basada en Leaflet.js). La interfaz permite al usuario seleccionar las capitales por las que viajará y las fechas de estancia en cada una; el sistema ejecuta la predicción del modelo entrenado para el periodo seleccionado, calcula el nivel de riesgo y lo visualiza en un mapa de Colombia donde cada ciudad se colorea según la escala definida (verde, amarillo, rojo). La aplicación se despliega en Streamlit Cloud, que permite su acceso público sin costo.

Instrumentos

Los instrumentos utilizados en la investigación se detallan a continuación. Como fuente de datos se emplea el reporte oficial de la Policía Nacional de Colombia, un instrumento institucional validado por la entidad como fuente primaria de estadísticas delictivas del país. El entorno de procesamiento es Google Colab (Python 3), y las librerías principales son: pandas y NumPy para manipulación de datos, matplotlib y seaborn para visualización, Prophet para modelado de series temporales, XGBoost y scikit-learn para aprendizaje automático y evaluación

de modelos, holidays para la identificación de festividades colombianas, Streamlit para el desarrollo de la interfaz web, y Folium para la visualización geoespacial. No se requiere validación de instrumentos de medición adicionales dado que la fuente de datos es un registro administrativo oficial de acceso público.

Técnicas para el Análisis de la Información

El análisis de la información se estructura en tres niveles. En el nivel descriptivo, se calculan medidas de tendencia central y dispersión, frecuencias absolutas y relativas, y se generan visualizaciones exploratorias para caracterizar el comportamiento del hurto de motocicletas por ciudad y periodo. En el nivel correlacional, se aplican pruebas estadísticas para evaluar la significancia de las diferencias temporales y geográficas observadas, incluyendo pruebas de Kruskal-Wallis para comparar distribuciones entre ciudades y análisis de autocorrelación para identificar ciclos temporales. En el nivel predictivo, se emplean los modelos Prophet y XGBoost descritos previamente, evaluados mediante MAE, RMSE y matrices de confusión para la clasificación de riesgo. Todo el procesamiento se realiza en Python mediante las librerías mencionadas.

Resultados

En esta sección se presentan los resultados obtenidos en cada fase del desarrollo metodológico, incluyendo la consolidación de datos, el análisis exploratorio, el entrenamiento y evaluación de modelos predictivos, y la clasificación de riesgo.

Consolidación del Dataset

Del total de 624.698 registros del reporte de la Policía Nacional, se filtraron los correspondientes al hurto de motocicletas en las 32 capitales departamentales, obteniendo un subconjunto de 294.161 registros para el periodo 2010-2025. Las capitales concentran el 62 % del total nacional de hurtos de motocicletas. Se estandarizaron las categorías de armas o medios empleados, consolidando cinco variantes de arma blanca en una sola categoría. Los 679 registros con género nulo y los 190 con grupo etario nulo se completaron con la categoría "No reporta". Se generaron series temporales diarias continuas por ciudad, completando con valor cero los días sin registros de hurto.

Análisis Exploratorio de Datos

El análisis exploratorio reveló patrones temporales consistentes. A nivel semanal, los días martes y miércoles presentan las frecuencias más altas de hurto, mientras que los domingos registran los valores más bajos. A nivel mensual, septiembre y octubre muestran picos estacionales, en tanto que diciembre presenta el mínimo anual. Estos patrones son coherentes con la Teoría de las Actividades Rutinarias: entre semana, la mayor circulación de motocicletas en zonas comerciales y la menor vigilancia generan condiciones propicias para el hurto.

La prueba de Kruskal-Wallis, aplicada para evaluar la Hipótesis H2, arrojó un estadístico H de 87.999,16 con un valor p prácticamente igual a cero ($p < 0,001$), lo que permite rechazar la hipótesis nula de igualdad de distribuciones. Este resultado confirma que la distribución temporal

de hurtos difiere significativamente entre las capitales departamentales, justificando el uso de modelos diferenciados por ciudad.

Comparación de Modelos Predictivos

Se entrenaron tres modelos con datos del periodo 2010-2022 y se validaron con el periodo 2023-2025.

Tabla 3

Comparación de métricas de error entre modelos predictivos

El modelo Baseline (promedio histórico por ciudad, mes y día de la semana) obtuvo un MAE de 1,0706 y un RMSE de 2,1973. Prophet, con modelos independientes por ciudad, alcanzó un MAE de 1,0697 y un RMSE de 1,3803, mejorando sustancialmente el RMSE (-37 %) respecto al Baseline. XGBoost, como modelo global con variables temporales, obtuvo el mejor MAE (0,9508) y un RMSE de 1,7319, superando al Baseline en un 11 %. El análisis de importancia de variables de XGBoost reveló que la ciudad codificada es la variable más determinante (59,9 %), seguida del indicador de fin de semana (16,7 %), el indicador de festividad (8,8 %), el año (6,9 %) y el día de la semana (5,0 %).

Estos resultados confirman la Hipótesis H1: el modelo XGBoost (MAE = 0,9508) produce predicciones con un error medio absoluto significativamente inferior al del modelo Baseline (MAE = 1,0706).

Clasificación por Niveles de Riesgo

La conversión de predicciones numéricas a categorías de riesgo se realizó mediante umbrales basados en terciles de la distribución mensual histórica de cada ciudad: verde (percentil 0-33, riesgo bajo), amarillo (percentil 33-66, riesgo moderado) y rojo (percentil 66-100, riesgo alto). Se adoptaron tres categorías en lugar de las cuatro planteadas inicialmente, dado que la

evaluación preliminar con cuatro categorías y umbrales diarios produjo un accuracy del 25,4 %, equivalente al azar. El cambio a tres categorías con umbrales mensuales incrementó el accuracy global a 42,6 % con un F1-Score ponderado de 0,3761.

El accuracy global debe interpretarse considerando la heterogeneidad del conjunto de ciudades. Las ciudades con mayores volúmenes de datos presentan accuracy significativamente superiores: Bogotá D.C. alcanza un 94,1 %, Barranquilla un 91,2 %, Medellín un 82,4 %, Bucaramanga un 76,5 % y Neiva un 73,5 %. En contraste, ciudades con registros muy escasos presentan accuracy inferiores al 30 %, lo que reduce el promedio global. Cinco de las 32 ciudades superan el 70 % de accuracy.

En cuanto al desempeño por categoría, el modelo clasifica con mayor precisión el riesgo alto (rojo: precisión 0,52, recall 0,67, $F1 = 0,59$) que el riesgo bajo (verde: precisión 0,32, recall 0,04, $F1 = 0,07$). El modelo es más confiable detectando periodos de alto riesgo, lo cual resulta preferible desde la perspectiva de protección al usuario.

Análisis Complementario: Configuración de Ciudades con Datos Suficientes

De manera complementaria, se evaluó una segunda configuración (Opción C) que restringe el análisis a las 24 ciudades con más de 1.000 hurtos registrados en el periodo, excluyendo ocho capitales con datos insuficientes (San Andrés, Manizales, San José del Guaviare, Leticia, Tunja, Puerto Carreño, Inírida y Mitú). Bajo esta configuración, el accuracy global asciende a 46,9 % con un F1-Score ponderado de 0,4543, superando al modelo Baseline en 20,5 puntos porcentuales. Esto representa una mejora respecto a la configuración de 32 ciudades (42,6 %), atribuible a la exclusión de ciudades cuyos volúmenes de datos no permiten una clasificación confiable.

Esta comparación entre configuraciones confirma la robustez del modelo: al concentrar la evaluación en las ciudades con datos suficientes, el desempeño mejora de manera consistente, mientras que las cifras de las ciudades principales (Bogotá 94,1 %, Barranquilla 91,2 %, Medellín 82,4 %) se mantienen estables en ambas configuraciones. Se adopta la configuración de 32 ciudades como principal, por su cobertura nacional completa y su coherencia con el objetivo del proyecto.

Análisis de Resultados

Cumplimiento de Objetivos

Objetivo (a): Consolidar los datos oficiales. Se consolidaron 294.161 registros de hurto de motocicletas en las 32 capitales, estructurando series temporales diarias continuas por ciudad con procedimientos de limpieza, estandarización y tratamiento de nulos. Objetivo cumplido integralmente.

Objetivo (b): Análisis exploratorio. Se identificaron patrones de estacionalidad semanal y mensual. La prueba de Kruskal-Wallis confirmó diferencias significativas entre ciudades ($H = 87.999$, $p < 0,001$). Objetivo cumplido integralmente.

Objetivo (c): Entrenar modelos predictivos. Se entrenaron tres modelos: Baseline, Prophet y XGBoost. XGBoost obtuvo el mejor rendimiento con MAE de 0,9508. Objetivo cumplido integralmente.

Objetivo (d): Evaluar el rendimiento. Se evaluaron los modelos con MAE, RMSE, matrices de confusión, accuracy y F1-Score. Se confirmaron las Hipótesis H1 y H2. Objetivo cumplido integralmente.

Objetivo (e): Desarrollar la aplicación web. Se desarrolló MotoSegura con Streamlit y Folium, permitiendo visualizar el nivel de riesgo en un mapa interactivo de Colombia con tres niveles de color. Objetivo cumplido integralmente.

Propuesta de Solución

Descripción de la Solución

MotoSegura es una aplicación web interactiva donde un usuario que planea un viaje en motocicleta selecciona las ciudades y fechas de estancia. El sistema ejecuta internamente el modelo XGBoost, calcula la frecuencia esperada de hurtos y clasifica el riesgo en tres niveles (verde, amarillo, rojo) comparando contra los umbrales históricos de cada ciudad. Los resultados se visualizan en un mapa de Colombia con popups informativos y la ruta del viaje dibujada entre las ciudades seleccionadas.

Cumplimiento de Requerimientos

Los requerimientos técnicos establecían un modelo con MAE inferior a 1,0 y clasificación en niveles de riesgo. XGBoost obtuvo un MAE de 0,9508, cumpliendo el umbral. La clasificación alcanza accuracy superiores al 70 % en las cinco ciudades principales, garantizando la utilidad operativa de la herramienta.

Desempeño y Sostenibilidad

La aplicación responde en menos de 5 segundos, es compatible con cualquier navegador y está disponible permanentemente a través de Streamlit Cloud. El costo operativo es nulo: hosting gratuito, modelo de menos de 3 MB, sin servidores dedicados. La actualización se realiza reejecutando el notebook cuando la Policía publique nuevos datos.

Costos

El desarrollo se realizó íntegramente con herramientas gratuitas y de código abierto: Google Colab, Python con librerías open-source, datos abiertos de la Policía Nacional y hosting en Streamlit Cloud. El costo total de desarrollo y operación es de \$0 COP.

Análisis de Requerimientos de Ingeniería

El análisis de requerimientos de ingeniería tiene como objetivo establecer las especificaciones de diseño y el alcance de la solución, con el fin de evitar cambios costosos en etapas tardías del proyecto. Este análisis se estructura en cuatro componentes: la intención del producto, la verificación de los parámetros de diseño, la estimación de características y especificaciones, y el alcance de la solución de ingeniería.

Intención del Producto

Descripción del producto. MotoSegura es un sistema de software predictivo que estima el nivel de riesgo de hurto de motocicletas en las 32 capitales departamentales de Colombia, presentado a través de una aplicación web con visualización geoespacial.

Objetivos del producto. Transformar los datos abiertos de criminalidad en información accionable que permita a los usuarios anticipar el riesgo y tomar decisiones informadas, reduciendo la probabilidad de pérdida patrimonial.

Usuarios finales. Motociclistas y viajeros que se desplazan entre ciudades, empresas del sector logístico y de domicilios que gestionan flotas, compañías aseguradoras que tarifican pólizas, y entidades de seguridad pública.

Verificación de Parámetros de Diseño

Dado que MotoSegura es un producto de software, los parámetros de diseño tradicionalmente físicos se reinterpretan en términos de calidad de software, siguiendo el estándar ISO 25010. La Tabla 4 presenta la verificación de los parámetros de diseño aplicables al producto.

Tabla 4

Verificación de parámetros de diseño del producto

Parámetro	Especificación del producto	Norma
Tolerancia	Error en calidad de datos < 5%	ISO 25010
Dependencias	Python, XGBoost, Matplotlib, Joblib, Holidays	Doc. técnica
Capacidad	Datasets de hasta millones de registros	Google Colab
Desempeño	Precisión > 80%, respuesta < 2s	ISO 25010
Validación	Validación cruzada, train/test, accuracy/recall	ISO 25010
Seguridad	Control de errores, respaldo, protección de datos	ISO 27001
Eficiencia	Optimización de código y recursos	Green Software
Fiabilidad	Estabilidad sin caídas ni sobrecarga	ISO 25010
Compatibilidad	Integración con CSV, APIs, Power BI	ISO 25010
Trazabilidad	Versionado Git, registro de datasets/modelos	ISO 9001

En cuanto a la tolerancia, al tratarse de un sistema de software no aplican tolerancias dimensionales físicas; el concepto se reinterpreta como tolerancia en la calidad de los datos, con un error permitido inferior al 5 % (ISO 25010). Las especificaciones de materiales se sustituyen por las dependencias de software: versiones específicas de Python y las librerías XGBoost, Matplotlib, Joblib y Holidays. La capacidad máxima de procesamiento depende de los recursos de Google Colab (RAM y CPU), permitiendo el manejo de datasets de hasta millones de registros. Los parámetros de desempeño establecidos son una precisión del modelo superior al 80 % en las ciudades principales, un tiempo de respuesta inferior a 2 segundos y la estabilidad del sistema.

Los métodos de validación recomendados incluyen la validación cruzada, la división train/test y métricas como accuracy, precision y recall. Los criterios de seguridad comprenden el control de errores, el manejo de excepciones, el respaldo de modelos en formato joblib y la protección de datos conforme a ISO 27001. La eficiencia se mide como eficiencia computacional mediante la optimización del código y el uso eficiente de recursos, siguiendo las buenas prácticas de la Green Software Foundation. La fiabilidad del sistema (equivalente digital a los límites de vibración o ruido) se garantiza mediante la estabilidad sin caídas ni sobrecargas (ISO 25010). La

compatibilidad e interoperabilidad se logra mediante la integración con sistemas de datos (CSV, APIs) y dashboards como Power BI. Finalmente, la trazabilidad se asegura mediante la documentación del código, el versionado con Git y el registro de datasets y modelos (ISO 9001, ISO 25010).

Estimación de Características y Especificaciones

Desempeño. El modelo XGBoost alcanza un MAE de 0,9508 y una precisión de clasificación superior al 80 % en las ciudades con mayor volumen de datos (Bogotá 94,1 %, Barranquilla 91,2 %, Medellín 82,4 %). El tiempo de respuesta de la aplicación es inferior a 5 segundos por consulta.

Durabilidad y mantenimiento. Al ser un producto de software, su vida útil depende de la actualización periódica del modelo con nuevos datos. El mantenimiento se limita a reejecutar el notebook de entrenamiento cuando la Policía Nacional publique datos actualizados.

Consumo de recursos. El modelo entrenado ocupa menos de 3 MB y opera sobre infraestructura en la nube de bajo consumo energético, alineado con las prácticas de software sostenible.

Alcance de la Solución de Ingeniería

Definición del alcance. El proyecto incluye la consolidación de datos, el entrenamiento y evaluación de modelos predictivos, y el desarrollo de una aplicación web funcional para las 32 capitales departamentales. No incluye predicción a nivel intraurbano (barrios o localidades), predicción de la hora específica del hurto, ni integración con sistemas en tiempo real de la Policía, los cuales se proponen como trabajo futuro.

Cronograma y recursos. El desarrollo se estructuró en cinco fases (consolidación, análisis exploratorio, entrenamiento, evaluación y aplicación web), ejecutadas con recursos de software libre y la participación del equipo de tres desarrolladores.

Comparación con Productos y Estándares Similares

Con el fin de posicionar la solución frente a alternativas existentes, se compararon quince productos y enfoques de predicción criminal y análisis de datos. Entre ellos se incluyen sistemas de predicción policial basados en Big Data e inteligencia artificial avanzada, modelos académicos de aprendizaje automático (Random Forest, SVM, regresión logística), redes neuronales profundas (TensorFlow, Keras), modelos de series temporales (ARIMA), sistemas de información geográfica criminal (ArcGIS), dashboards de inteligencia de negocios (Tableau, Power BI) y el análisis manual policial tradicional. La comparación evidencia que MotoSegura ocupa un nicho diferenciado: combina un modelo XGBoost de alta precisión con una interfaz web accesible y gratuita orientada al usuario final, característica ausente en las soluciones industriales (de alto costo y propietarias) y en los enfoques académicos (sin aplicación práctica para el ciudadano). Esta combinación de precisión, accesibilidad y costo nulo constituye la principal ventaja competitiva del producto.

Análisis de Costos

El análisis de costos de la solución de ingeniería se estructura en costos fijos, costos variables, costo del producto, análisis de viabilidad y modelo de negocio. Dado que MotoSegura es un producto digital basado en datos abiertos y tecnologías de código abierto, su estructura de costos es significativamente más liviana que la de productos físicos.

Costos Fijos

Los costos fijos anuales comprenden la nómina del equipo de desarrollo, el arriendo de espacio de trabajo (\$1.200.000), los servicios de internet (\$1.500.000) y telefonía (\$500.000), la papelería (\$700.000), el outsourcing contable y legal (\$2.000.000) y el dominio con hosting especializado (\$1.000.000). Los gastos fijos operativos suman aproximadamente \$6.900.000 anuales, a los cuales se añade la nómina del equipo.

Costos Variables

Los costos variables están asociados a cada línea de servicio. Para las suscripciones premium, el costo marginal es el hosting y soporte por usuario (aproximadamente \$12.000 anuales por usuario). Para las licencias B2B con aseguradoras y empresas de logística, el costo variable corresponde a la integración técnica y el soporte especializado (\$2.000.000 a \$3.000.000 por contrato). Para los reportes de inteligencia, el costo es el tiempo de elaboración (\$1.000.000 por reporte). El bajo costo variable es característico de los productos digitales escalables.

Costo del Producto y Punto de Equilibrio

La inversión inicial asciende a \$20.819.344, que incluye muebles y enseres (\$900.000), equipo de oficina (\$2.500.000), registro de patentes e intangibles (\$1.500.000) y el capital de trabajo para los primeros meses. El punto de equilibrio se alcanza con aproximadamente 4 unidades equivalentes de servicio, lo que refleja la alta rentabilidad marginal del modelo digital.

Análisis de Viabilidad

La evaluación financiera proyectada a cinco años (2026-2030) arroja indicadores favorables. El Valor Presente Neto (VPN) es de \$295.829.801, lo que indica que el proyecto genera valor por encima de la inversión. La Tasa Interna de Retorno (TIR) es del 113,5 %, muy

superior a la tasa de descuento, y el periodo de recuperación es de 0,33 años (aproximadamente cuatro meses). Estos resultados confirman la viabilidad financiera de la solución.

Modelo de Negocio y Oferta de Valor

El modelo de negocio combina un esquema freemium con ingresos B2B. La aplicación es de acceso gratuito para el usuario final, generando una base de usuarios que sustenta la propuesta de valor. Los ingresos provienen de tres fuentes: suscripciones premium con funcionalidades avanzadas; licencias B2B a compañías aseguradoras que utilizan el modelo de riesgo para tarifar pólizas de motocicletas, y a empresas de logística para la gestión de flotas; y reportes de inteligencia para entidades públicas. Los ingresos proyectados para el primer año ascienden a \$83.280.000, distribuidos entre suscripciones premium (\$17.280.000), licencias a aseguradoras (\$28.000.000), licencias a empresas de logística (\$28.000.000) y reportes de inteligencia (\$10.000.000). La oferta de valor central es la transformación de datos abiertos de criminalidad en información accionable que reduce el riesgo de pérdida patrimonial.

Análisis de Sostenibilidad

El análisis de sostenibilidad se aborda desde tres dimensiones complementarias — ambiental, social y económica— siguiendo el enfoque del Triple Bottom Line, fundamentado en estándares internacionales reconocidos.

Dimensión Ambiental

MotoSegura presenta un impacto ambiental bajo dado su carácter de producto digital. El consumo energético se limita al uso de servidores en la nube, optimizado mediante infraestructura compartida de bajo consumo (ISO 50001). La emisión de CO₂ es mínima al no requerir manufactura ni distribución física (ISO 14064, ISO 14001). El uso de materiales físicos

es prácticamente nulo y los residuos electrónicos son mínimos (ASTM E1961). La eficiencia de recursos se estima en 80 % y el ciclo de vida del producto se evalúa favorablemente (ISO 14040/44). Adicionalmente, al contribuir a la reducción del hurto, el proyecto favorece indirectamente la conservación del parque automotor.

Dimensión Social

El impacto social es alto y positivo. La seguridad del usuario es el eje central de la propuesta (ISO 12100, IEC 62368). La accesibilidad se garantiza mediante una interfaz web gratuita y compatible con cualquier dispositivo (WCAG 2.1). El impacto social se valora en el nivel máximo, dado que la herramienta beneficia a motociclistas, sector conformado mayoritariamente por trabajadores de domicilios, mensajería y comercio (ISO 26000). La formación necesaria para usar la aplicación es mínima y los riesgos de salud ocupacional del desarrollo son bajos (ISO 45001).

Dimensión Económica y de Gobernanza

La sostenibilidad económica se sustenta en la viabilidad financiera (VPN positivo de \$295.829.801 y TIR del 113,5 %) y en un modelo de negocio diversificado. El costo operativo nulo de la infraestructura y el uso de tecnologías de código abierto garantizan la sostenibilidad a largo plazo. En gobernanza, el proyecto se adhiere a la transparencia mediante el uso de datos abiertos y la publicación del código fuente, cumpliendo la normativa colombiana de protección de datos. La generación de empleo calificado constituye un aporte económico adicional.

Marco Legal y Regulatorio

El desarrollo y operación de MotoSegura en Colombia está sujeto a un marco normativo que regula el tratamiento de datos, el uso de información pública y la seguridad informática. A continuación se identifican las normas aplicables y las medidas adoptadas para su cumplimiento.

Protección de Datos Personales (Ley 1581 de 2012 y Decreto 1377 de 2013)

La Ley 1581 de 2012 y su decreto reglamentario 1377 de 2013 regulan el tratamiento de datos personales. Aunque MotoSegura no almacena datos personales de sus usuarios —procesa las consultas en tiempo real sin guardar registros, no requiere registro de cuenta ni recopila información sensible—, el proyecto adopta como medida preventiva la inclusión de un aviso de privacidad y una política de tratamiento conforme a estas normas, garantizando los principios de finalidad, libertad, transparencia y seguridad. En caso de implementar funcionalidades premium con registro, se solicitará autorización expresa del titular y se registrará la base de datos ante el Registro Nacional de Bases de Datos.

Acceso a Datos Abiertos (Ley 1712 de 2014)

La Ley 1712 de 2014, de Transparencia y del Derecho de Acceso a la Información Pública, legitima el uso de los datos de hurto publicados por la Policía Nacional. Esta ley establece que la información pública debe estar disponible en formatos abiertos y reutilizables. El proyecto utiliza estos datos respetando su naturaleza pública y citando apropiadamente la fuente.

Delitos Informáticos (Ley 1273 de 2009)

La Ley 1273 de 2009 tipifica los delitos informáticos y protege la información y los datos. El proyecto adopta medidas de seguridad informática para proteger la integridad del

sistema y prevenir accesos no autorizados. El uso de infraestructura en la nube con estándares de seguridad y la ausencia de almacenamiento de datos sensibles reducen la superficie de riesgo.

Habeas Data y Comercio Electrónico (Leyes 1266 de 2008 y 527 de 1999)

La Ley 1266 de 2008 (Habeas Data) complementa el régimen de protección de datos en lo relativo al derecho de los ciudadanos a conocer y rectificar su información. La Ley 527 de 1999 establece el marco para el comercio electrónico y la validez de los mensajes de datos, aplicable a las funcionalidades de suscripción y contratación B2B. El proyecto contempla el cumplimiento de ambas normas en su operación comercial.

Discusión

La confirmación de la Hipótesis H1 es consistente con los hallazgos de Chen y Guestrin (2016), quienes demostraron la superioridad de los algoritmos de gradient boosting en tareas de predicción con datos heterogéneos. El MAE de 0,9508 es inferior al reportado por Azhari y Utomo (2018) con ARIMAX (RMSE = 6,68) y comparable al de Pranata et al. (2020) con MA(1) (RMSE = 6,56), aunque estas comparaciones deben interpretarse con precaución dado que los contextos geográficos difieren.

La confirmación de la Hipótesis H2, con un estadístico H de 87.999 ($p < 0,001$), refuerza lo planteado por la Teoría del Patrón Delictivo (P. J. Brantingham & P. L. Brantingham, 1993): cada capital presenta una estructura socioespacial particular que configura patrones delictivos diferenciados. Este hallazgo justifica que la aplicación utilice umbrales de riesgo relativos a cada ciudad.

La variable más importante para XGBoost fue la ciudad (59,9 %), seguida del fin de semana (16,7 %) y las festividades (8,8 %). Este resultado es coherente con la Teoría de las Actividades Rutinarias (Cohen & Felson, 1979): los fines de semana y festivos, con menor

circulación de motocicletas, presentan menos hurtos. La predominancia de la ciudad sugiere que las condiciones estructurales de cada capital son más determinantes que las variaciones temporales.

Una limitación del estudio es la dependencia exclusiva de variables temporales. El dataset no incluye hora del día, barrio, ni variables socioeconómicas o climáticas, cuya integración podría mejorar la precisión. Adicionalmente, los datos representan denuncias y no el total real de hurtos, aunque el hurto de motocicletas tiene una de las tasas de denuncia más altas dado que las víctimas requieren el denuncia para reclamar el seguro y bloquear el vehículo en el RUNT.

Otra limitación es el accuracy global de la clasificación (42,6 %), influenciado por ciudades con datos insuficientes. Sin embargo, las ciudades con mayor relevancia para los viajeros presentan accuracy superiores al 70 %, validando la utilidad operativa de la herramienta para su público objetivo.

Plan de Divulgación

El plan de divulgación se estructura en tres estrategias. Primero, la aplicación MotoSegura será publicada en Streamlit Cloud con acceso abierto y gratuito, difundida a través de redes sociales, foros de motociclistas colombianos y comunidades de viajeros.

Segundo, el código fuente, los notebooks y los modelos entrenados se publicarán en un repositorio público de GitHub, permitiendo la reproducibilidad del estudio y facilitando que otros investigadores puedan adaptar o extender el modelo.

Tercero, los hallazgos se presentarán como artículo científico conforme a los lineamientos de la Universidad Ean, con la posibilidad de remitirlo a revistas especializadas en seguridad urbana, ciencia de datos o ingeniería de sistemas.

Conclusiones

El presente estudio desarrolló un modelo predictivo de hurto de motocicletas en las 32 capitales departamentales de Colombia, partiendo de 294.161 registros oficiales de la Policía Nacional correspondientes al periodo 2010-2025. Los principales hallazgos son los siguientes.

Se confirmó la Hipótesis H1: el modelo XGBoost, con un MAE de 0,9508, supera significativamente al Baseline (MAE = 1,0706) en un 11 %, demostrando que las técnicas de aprendizaje automático aportan valor predictivo real más allá del promedio histórico.

Se confirmó la Hipótesis H2: la prueba de Kruskal-Wallis ($H = 87.999$, $p < 0,001$) demostró que las 30 ciudades evaluadas presentan distribuciones de hurto estadísticamente diferentes, justificando el enfoque de umbrales relativos por ciudad.

La clasificación de riesgo en tres niveles con umbrales basados en terciles mensuales alcanza un accuracy global del 42,6 %, superando el azar (33,3 %) en 9,3 puntos porcentuales. Las cinco ciudades principales presentan accuracy entre 73,5 % y 94,1 %, validando la utilidad práctica de la herramienta.

El análisis de importancia de variables reveló que la ciudad es el factor más determinante (59,9 %), seguido del fin de semana (16,7 %) y las festividades (8,8 %), hallazgo coherente con las teorías criminológicas de referencia.

La aplicación web MotoSegura materializa los resultados en una herramienta accesible, gratuita y de costo operativo nulo, que convierte estadísticas delictivas en información accionable para el ciudadano. No se identificaron en la literatura colombiana herramientas similares orientadas al usuario final.

Como trabajo futuro se recomienda integrar variables externas (datos climáticos del IDEAM, parque automotor del RUNT, eventos masivos), extender el modelo a localidades

dentro de cada ciudad, implementar actualización automática y desarrollar una versión móvil con notificaciones de riesgo en tiempo real.

Referencias

- Azhari, & Utomo, P. E. P. (2018). Prediction the crime motorcycles of theft using ARIMAX-TFM with single input. En Proceedings of the 3rd International Conference on Informatics and Computing (ICIC 2018) (art. 8780520). IEEE. <https://doi.org/10.1109/IAC.2018.8780520>
- Brantingham, P. J., & Brantingham, P. L. (1993). Environment, routine, and situation: Toward a pattern theory of crime. En R. V. Clarke & M. Felson (Eds.), Routine activity and rational choice (pp. 259–294). Transaction Publishers.
- Campo Agredo, C. A. (2017). Desarrollo de modelos de predicción sobre el hurto a personas en Bogotá a partir del análisis y aplicación de algoritmos en base de datos [Trabajo de grado, Universidad de los Andes]. Repositorio institucional Séneca. <http://hdl.handle.net/1992/50785>
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. <https://doi.org/10.1145/2939672.2939785>
- Cohen, L. E., & Felson, M. (1979). Social change and crime rate trends: A routine activity approach. American Sociological Review, 44(4), 588–608. <https://doi.org/10.2307/2094589>
- Hernández-Sampieri, R., & Mendoza, C. P. (2018). Metodología de la investigación: Las rutas cuantitativa, cualitativa y mixta. McGraw-Hill Education.

- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Iqbal, M. F. B., Ullah, A., Alhomoud, A., Hussain, T., Attar, R. W., Ouyang, J., Alnfai, M. M., & Hatamleh, W. A. (2025). A 2D clustering based hotspot identification approach for spatio-temporal crime prediction. *International Journal of Interactive Multimedia and Artificial Intelligence*, 9(5), 6–16. <https://doi.org/10.9781/ijimai.2024.10.001>
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30, 3146–3154.
- Li, W. X., & Law, K. L. E. (2024). Deep learning models for time series forecasting: A review. *IEEE Access*, 12, 92306–92327. <https://doi.org/10.1109/ACCESS.2024.3422528>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2021 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. <https://doi.org/10.1136/bmj.n71>
- Policía Nacional de Colombia. (2024). Reporte de hurto por modalidades [Conjunto de datos]. Datos abiertos de la Policía Nacional. <https://www.policia.gov.co/>
- Pranata, M., Anggraini, D., Makbuloh, D., & Rinaldi, A. (2020). Prediksi pencurian sepeda motor menggunakan model time series (studi kasus: Polres Kotabumi Lampung Utara). *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, 14(3), 423–432. <https://doi.org/10.30598/barekengvol14iss3pp423-432>

Shaw, C. R., & McKay, H. D. (1942). *Juvenile delinquency and urban areas*. University of Chicago Press.

Taylor, S. J., & Letham, B. (2018). Forecasting at scale. *The American Statistician*, 72(1), 37–45. <https://doi.org/10.1080/00031305.2017.1380080>