Aplicación de técnicas de *machine learning* (aprendizaje automático) para la detección y prevención de hurtos a personas en Bucaramanga

Integrantes del Equipo:

MELISSA GARCIA HERRERA

VICTOR ANDRES NIÑO VARGAS

DANIEL ENRIQUE GRANADOS IGLESIAS

DAVID HERNANDO HENAO MARULANDA

Universidad EAN

Especialización en Machine Learning

Seminario de Investigación

Grupo MLRU1A - M5V - Virtual – 2024

Profesora. Marie José Chery Leal

Diciembre de 2024

Resumen

Este trabajo analiza los patrones de criminalidad en Bucaramanga entre 2016 y 2023, utilizando técnicas de análisis de datos y *Machine Learning* para identificar zonas de alto riesgo y predecir delitos futuros. A través de análisis descriptivo, georreferenciación y modelos de clustering, se identificaron áreas con alta incidencia delictiva, especialmente hurtos a personas. Se desarrolló un visualizador interactivo en Power BI para facilitar a las entidades locales en la toma de decisiones y diseño de estrategias de prevención, destacando la importancia de intervenciones basadas en patrones temporales, geoespaciales y demográficos.

Palabras claves: seguridad ciudadana, hurtos a personas, detección, machine learning.

Problema de Investigación

La seguridad ciudadana es una preocupación constante para los gobiernos, ya que afecta profundamente la vida cotidiana de las personas y es un factor crítico para el crecimiento y desarrollo económico de cualquier país (BID, 2024). En este contexto, es importante destacar que muchos Estados consideran la seguridad ciudadana como una prioridad, ya que solo en un entorno pacífico se pueden establecer las condiciones sociales, económicas y políticas adecuadas para el progreso y desarrollo de una nación (Gonzales Rodriguez & Barbarán Mozo, 2021). Un entorno seguro no solo mejora la calidad de vida, sino que también atrae inversiones, fomenta el comercio y promueve el turismo, todos ellos elementos esenciales para el desarrollo económico. (IMF, 2023).

En Colombia, la seguridad ciudadana ha sido históricamente una de las principales preocupaciones. Según el Índice Mundial de Crimen Organizado del 2023, *"*Colombia ocupa el segundo lugar a nivel global en criminalidad*"* (Índice global de crimen organizado, 2023). Además, la situación de criminalidad ha mostrado una tendencia preocupante, con un aumento constante en delitos como el hurto y la extorsión. De acuerdo con el Centro de Estudios de Justicia (CEJ), cada día más de mil personas en Colombia son víctimas de estos delitos (CEJ, 2023), lo que evidencia un deterioro en la seguridad ciudadana y un creciente desafío para las autoridades.

Esta realidad subraya la urgente necesidad de fortalecer las políticas de seguridad y de buscar soluciones innovadoras que puedan revertir esta tendencia y restaurar la confianza pública. "Reducir el nivel de delincuencia en América Latina al nivel del promedio mundial incrementaría el crecimiento económico anual de la región en 0,5 puntos porcentuales, lo que equivale a aproximadamente un tercio del crecimiento registrado en América Latina entre 2017 y 2019" (FMI, 2023, párrafo 5).

En Bucaramanga, el informe de Calidad de Vida elaborado por el programa Bucaramanga Metropolitana Cómo Vamos (BMCV) en 2023 señala que, en 2022, 11.326 personas reportaron ser víctimas de hurto en el área metropolitana. Además, se evidenció un incremento del 38.85% en los hurtos a personas entre enero y junio de 2023 en comparación con el mismo período de 2022. Más de la mitad de los casos de hurto a personas en 2022 representaron el 68% de estos hechos en el área metropolitana, lo que resalta la necesidad de mejorar la seguridad en la región (Bucaramanga Metropolitana como vamos, 2023).

En este contexto, el uso de técnicas avanzadas como el *Machine Learning* se presenta como una opción prometedora para mejorar la detección y prevención de actividades delictivas. Estas técnicas ya han demostrado ser efectivas en diversos campos, como en finanzas, un ejemplo es que se utilizan para la evaluación de escenarios donde valuan clientes con perfiles riesgosos e inclusive determinar operaciones que signifiquen fraudes (Alvarado Zabala, Martillo Alchundia, & Guzman Seraquive, 2022); en salud, mejorando el diagnóstico temprano de enfermedades como el cáncer (Musheer Aziz, Yaqoob, & Kumar verma, 2023); y en marketing y ventas, facilitan la personalización de ofertas y la predicción de comportamientos de consumo (Herhausen, Bernritter, Ngai, Kumar, & Delen, 2024). En todos estos ámbitos, el Machine Learning se ha consolidado como una herramienta valiosa para enfrentar desafíos complejos.

Mediante la recolección y análisis de datos vinculados a incidentes delictivos, el Machine Learning puede asistir a las agencias de seguridad y a la sociedad civil en la previsión y disuasión de actos delictivo. La implementación de estas tecnologías podría ser un factor decisivo en la reducción de la criminalidad, contribuyendo así al desarrollo económico y social de la región (Karabo, Cagatay, & Gorkem, 2023).

Por esta razón, la presente investigación se centra en la siguiente pregunta: ¿Cómo pueden las técnicas de *machine learning* mejorar la detección y prevención de hurtos a personas en Bucaramanga? Este trabajo busca abordar esta problemática mediante la aplicación de técnicas de *machine learning* para la detección y prevención de hurtos a personas, con el fin de ofrecer una herramienta predictiva que permita a las autoridades locales anticipar la ocurrencia de estos delitos y tomar decisiones informadas para reducir su impacto.

Objetivos

Objetivo General

Aplicar técnicas de *machine learning* para desarrollar una herramienta de visualización que permita la prevención de hurtos a personas en la ciudad de Bucaramanga.

Objetivos Específicos

- Identificar por medio de *machine learning* las zonas en Bucaramanga con características similares en cuanto a la ocurrencia de hurtos a personas, utilizando variables socioeconómicas, demográficas y de criminalidad.
- Desarrollar un modelo de series de tiempo para predecir la frecuencia y probabilidad de hurto en Bucaramanga en el año 2025.
- Desarrollar un visualizador que identifique las zonas de mayor riesgo de hurtos en la ciudad de Bucaramanga para facilitar a las autoridades las intervenciones preventivas en las áreas más vulnerables.

Justificación

El aumento en la tasa de criminalidad es un problema significativo en muchas ciudades alrededor del mundo. Según el Índice Mundial de Crimen Organizado 2023, elaborado por Global Initiative, Colombia ocupa el segundo lugar en el ranking mundial de países con mayor criminalidad (Global Initiative Against, 2023). Este índice, que analiza la criminalidad y la resiliencia frente al crimen organizado en los 193 países miembros de la ONU, revela que el 83% de la población mundial vive en países con altos niveles de criminalidad, un incremento respecto al 79% registrado en 2021. A nivel global, Colombia se encuentra entre los países con mayor criminalidad junto con Birmania, México y Paraguay, y lidera la lista en América Latina (Mark Shaw, 2023).

El hurto a personas (como el robo de carteras o teléfonos móviles) sigue siendo una de las formas más comunes de crimen a nivel mundial. Aunque el porcentaje exacto de hurtos varía según la región, en muchos países representa una parte significativa de los delitos totales. En general, representan entre el 30% y el 50% de los delitos denunciados en varias regiones del mundo. Si bien no se puede obtener un número exacto a nivel global, varios estudios e informes, como el Índice Mundial de Crimen Organizado 2023, muestran que los delitos como el robo y hurto siguen siendo prevalentes en todo el mundo, especialmente en áreas urbanas con alta densidad de población y una presencia limitada de fuerzas de seguridad (GIJN Staff, 2023).

Bucaramanga, una de las ciudades más importantes de Colombia por su desarrollo económico y calidad de vida, enfrenta un aumento en la percepción de inseguridad entre sus habitantes. Aunque los robos han disminuido un 31% en comparación con años anteriores y el esclarecimiento de los delitos ha aumentado en un 54%, persiste un sentimiento de temor en la población. Según la Policía Metropolitana de Bucaramanga, los métodos de robo más comunes incluyen el cosquilleo, atraco a mano armada, suplantación de identidad, uso de escopolamina y el halado de vehículos, los cuales afectan principalmente a barrios como Centro, Cabecera del Llano y San Francisco. A pesar de las mejoras en las estadísticas, un reciente sondeo reveló que el 82% de los bumangueses se sienten inseguros en las calles, lo que pone de manifiesto una desconexión entre las cifras oficiales y la percepción cotidiana de los ciudadanos (Euclides Kilo Ardila, 2024).

La detección y prevención temprana de actividades delictivas es crucial para garantizar la seguridad pública y proteger a los ciudadanos. Sin embargo, las técnicas tradicionales de vigilancia y análisis de datos pueden ser insuficientes para identificar patrones complejos de comportamiento criminal que podrían prevenir futuros delitos.

El uso de Machine Learning (ML) en la seguridad pública ofrece una nueva dimensión para abordar este problema. Las técnicas de ML pueden analizar grandes volúmenes de datos de diversas fuentes, incluyendo cámaras de vigilancia, redes sociales, y registros policiales, para identificar patrones de comportamiento que podrían indicar una actividad delictiva inminente. El objetivo de esta investigación es evaluar la eficiencia del uso de ML en la detección de actividades delictivas buscando la reducción significativa de las tasas de criminalidad, al permitir una respuesta más rápida y eficaz por parte de las fuerzas de seguridad (Ordóñez, H., Cobos, C., & Bucheli, V. 2020).

El enfoque basado en Machine Learning es adecuado debido a su capacidad para aprender y adaptarse a nuevos patrones, mejorando continuamente la precisión de las predicciones.

Además, las técnicas de ML pueden ser integradas con sistemas de vigilancia existentes, lo que facilita su implementación. El uso de modelos predictivos avanzados, como redes neuronales y algoritmos de clasificación, permitirá detectar anomalías y comportamientos sospechosos con alta precisión (Ordóñez, H., Cobos, C., & Bucheli, V. 2020).

Este proyecto busca beneficiar la seguridad pública reduciendo la tasa de hurtos a personas en la ciudad de Bucaramanga. Además, al prevenir estos delitos antes de que ocurran, se reduce la carga sobre los sistemas judiciales y se promueve un entorno más seguro para el desarrollo social y económico. El impacto positivo de este proyecto podría servir como modelo para otras ciudades con problemas similares. También se alinea con los esfuerzos locales para mejorar la seguridad ciudadana mediante el uso de tecnología de punta (Vanguardia, 2024). Además, está en consonancia con las políticas de innovación tecnológica en la administración pública, que buscan utilizar la inteligencia artificial para resolver problemas complejos en la sociedad (G. A. Vergel-Clavijo y A. M. Guerrero-Bayona, 2023).

Esta es una iniciativa innovadora y necesaria que tiene el potencial de transformar la manera en que se maneja la seguridad pública en las ciudades. La implementación de este proyecto por parte de las autorizades competentes abordaría un problema urgente y, de llevarse a cabo, podría establecer un nuevo estándar en la prevención del delito mediante el uso de tecnologías avanzadas.

Marco Teórico

En la actualidad el aumento de las actividades delictivas representa un desafío significativo para las autoridades y la sociedad en general, particularmente en países como Colombia, donde la criminalidad sigue siendo un problema persistente. La identificación de patrones delictivos y la predicción de eventos criminales han sido objetivos clave en la lucha contra el crimen, pero las metodologías tradicionales han demostrado ser insuficientes para enfrentar la complejidad y dinámica de los delitos modernos.

En este contexto, el uso de técnicas avanzadas como el *machine learning* ha emergido como una solución prometedora para mejorar la detección y prevención de actividades delictivas. Estas técnicas permiten el análisis de grandes volúmenes de datos y la identificación de patrones complejos que serían difíciles de discernir mediante métodos convencionales. Sin embargo, la implementación de *machine learning* en el ámbito de la seguridad plantea desafíos tanto técnicos como éticos (Ordóñez et al, 2020).

El presente marco teórico busca proporcionar una base sólida para la investigación. En primer lugar, se realiza una revisión exhaustiva de la literatura existente, identificando estudios previos que han aplicado técnicas de *machine learning* en la detección y prevención de delitos, así como las limitaciones y desafíos que estos estudios han revelado. Posteriormente, se exploran las teorías criminológicas y conceptos clave que subyacen a la predicción del crimen, integrándolos con los principios fundamentales del *machine learning*. Finalmente, se describen los modelos y enfoques metodológicos que se consideran más adecuados para abordar el problema de investigación, sentando las bases para la implementación práctica en el contexto colombiano y en la ciudad de Bucaramanga.

Este marco teórico, por lo tanto, no solo sirve como fundamento conceptual para la investigación, sino que también orienta el desarrollo de un modelo aplicable a la realidad delictiva de Colombia y Bucaramanga, con el objetivo de contribuir al mejoramiento de la seguridad pública y la reducción de la criminalidad en esta región.

1. Estado del arte.

Son varias las investigaciones y estudios que se han realizado en torno a la gestión de la criminalidad con soporte del *Machine Learnig* como elemento habilitador. En ese apartado se analizan algunas de las investigaciones más representativas, como un primer acercamiento al estado del arte de la propuesta de valor.

Un primer acercamiento es el de Gelvez, J., Nieto, M., & Rocha, C. (2022), quienes presentan un modelo de *Machine Learning* para predecir el crimen en Bucaramanga, una ciudad intermedia en Colombia, la cual es el foco de la presente investigación. Se emplearon técnicas de procesamiento de señales en grafos y una adaptación del modelo TF-IDF (frecuencia de término – frecuencia inversa de documento) para *text mining* (minería de texto). Los resultados mostraron que los modelos espaciales de grafos con periodicidad semanal ofrecen las mejores predicciones. El modelo de KNN (K vecinas más cercanas o *K-nearest neighbors*) de clasificación fue el más efectivo, con un 59% de *recall* (sencibilidad) y más del 60% de exactitud. Aunque los modelos de predicción son útiles para estrategias de prevención en grandes ciudades, enfrentan restricciones en ciudades intermedias con datos limitados.

Muñoz, V. (2021), evalúa tres modelos de clasificación que son *Random Forest* (bosque aleatorio), regresión logística y SVM (Máquina de Vectores de Soporte), para predecir el crimen en Medellín, enfocándose en diferentes tipos de hurto a personas (atraco, descuido, cosquilleo y raponazo). La investigación sigue el proceso CRISP-DM (Proceso Estándar Inter-Indsutrias para Minería de Datos), que incluye la recopilación y análisis de datos históricos y tasas de desempleo, y compara el rendimiento de estos modelos con un enfoque basado en reglas, utilizando métricas como exactitud, sensibilidad, precisión y valor F1. Entre las fuentes de información utilizadas por Muñoz se encuentran:

- Base de datos de crímenes y accidentes en la página web MEDATA de Medellín.
- Base de datos de las comunas y barrios de Medellín, en el portal GEOMEDELLIN.
- Base de datos de las variables meteorológicas del SIATA (Sistema de Alertas Temprana de Medellín y el Valle de Aburra).
- Base de datos con las variables demográficas.

Ordóñez, H., Cobos, C., & Bucheli, V. (2020) presentan un modelo de *Machine Learning* basado en Máquinas de Soporte Vectorial para regresión (SVR) para predecir las tendencias de hurto en Colombia y sus tres principales ciudades. El modelo fue validado utilizando un extenso *dataset* de la Fiscalía Nacional de Colombia, que abarca más de 2,6 millones de registros de delitos desde 1960 hasta 2019. Comparado con un modelo de regresión lineal estándar y un modelo SVR sin ajuste, los resultados del modelo ajustado son prometedores y podrían ser útiles para la toma de decisiones en la lucha contra el hurto.

Otro de los estudios más relevantes, es el de Ardila (2023), quien estudia la relación entre el crimen y los factores socioeconómicos en la ciudad de Medellín. Su estudio se centra en analizar

los patrones espaciales de delitos en Medellín, Colombia, utilizando modelos de *Machine Learning* para predecir la probabilidad de ocurrencia de diversos crímenes anualmente, considerando datos históricos y sociodemográficos. Se emplearon modelos como Mínimos Cuadrados Ordinarios, *Random Forest* y *Extreme Gradient Boosting* (refuerzo de gradientes extremo), los cuales mostraron un buen desempeño en términos de precisión. Se observó que variables socioeconómicas, como la proporción de jóvenes, el desempleo, la pobreza multidimensional y las condiciones de vivienda, tienen un alto poder predictivo tanto a nivel de barrios como de grillas.

López, G. y Manosalvas, P. (2023) estudian la relación entre variables socioeconómicas y la ocurrencia de delitos mediante un modelo de aprendizaje de *random forest*, cuyo objetivo es comprender los patrones, las tendencias y los factores determinantes que fomentan la criminalidad. El conjunto de datos se obtuvo a través de organismos como el Consejo de la Judicatura y el Instituto Nacional de Estadística y Censos (INCE), e incluía variables como la tasa de desempleo, la población y la pobreza multidimensional. Mediante el modelo, se evidenció que las variables con mayor importancia para determinar la criminalidad son el cantón (municipio) y la provincia (departamento).

Salazar Isairias, S. (2024) aborda el desarrollo de un modelo de *machine learning*, específicamente un modelo de redes neuronales basadas en grafos, para la predicción de la criminalidad en Bogotá D.C. El modelo se basa en el concepto de *crime forecasting* (pronóstico del crimen) para su funcionamiento. El modelo es entrenado con variables como la fecha, localidad, barrio, tipo de crimen (homicidios, hurto a personas y hurto de celulares), y la cantidad de crímenes. Los grafos generados corresponden a las localidades, ya que no es posible seleccionar los barrios debido a los cambios que estos han sufrido durante el periodo de los datos utilizados, que corresponde al periodo entre 2016 y 2019. Los datos fueron tomados en conjunto del portal queremosdatos.co y de portal de reportes de crímenes de la Policía Nacional.

El *Machine Learning* y la modelación estadística avanzada también tienen participación en los estudios psicológicos y sociológicos relacionados con la criminología. En su estudio, Nedelec & Di Mienzo (2023) exponen que la criminología evolutiva integra conceptos de psicología evolutiva para mejorar la comprensión del crimen y el comportamiento antisocial, cuestionando las teorías criminológicas tradicionales. Un ejemplo reciente es la teoría de la taxonomía evolutiva, que se basa en la teoría de la historia de vida de Moffitt. Aunque esta teoría ha sido útil, aún no se ha evaluado exhaustivamente cómo las medidas de la teoría de la historia de vida predicen la clasificación en grupos de delincuentes según la taxonomía de Moffitt. Este estudio utilizó datos del Estudio Longitudinal Nacional de la Salud de la Adolescencia a la Adultez (n = 12,012) y, haciendo uso de Regresión Lineal Múltiple, se halló que las medidas relacionadas con el esfuerzo somático y el entorno de desarrollo eran predictivas para la clasificación en grupos de delincuentes, mientras que las relacionadas con el esfuerzo reproductivo no lo eran. Los resultados tienen implicaciones tanto para la criminología evolutiva como para la criminología tradicional.

Por su parte, Watts (2023) afirma que las técnicas de aprendizaje automático tienen un gran potencial en psiquiatría forense para prever resultados individuales de los pacientes, aunque el campo aún está en desarrollo. Su capítulo ofrece una visión general de los modelos predictivos para evaluar la violencia y el comportamiento delictivo, así como la simulación de enfermedades mentales graves tras un arresto para evitar penas de prisión. También se exploran diversas

técnicas de aprendizaje automático y diseños experimentales aplicables a problemas persistentes en el área. El objetivo es proporcionar recomendaciones metodológicas para avanzar hacia una ciencia forense más precisa.

2. Una aproximación a la criminalidad

La Real Academia de la Lengua Española (RAE) define a la delincuencia como "el conjunto de delitos, ya en general o ya referidos a un país, época o especialidad en ellos" (Real Academia de la Lengua Española, s.f., párr. 1). Enfocando específicamente en el hurto como el foco de análisis del presente documento, este es definido como "el delito consistente en tomar con ánimo de lucro cosas muebles ajenas contra la voluntad de su dueño, sin que concurran las circunstancias que caracterizan el delito de robo" (Real Academia de la Lengua Española, s.f., párr. 1). González (2019) lo define como el apoderamiento de bienes muebles ajenos con ánimo de lucro, sin emplear violencia ni intimidación. Esta conceptualización, presente en la mayoría de los códigos penales contemporáneos, protege no solo la propiedad, sino también la posesión. Los autores del hurto son aquellos que sustraen bienes sin la voluntad de su dueño, y este delito puede variar en gravedad según el valor de lo hurtado.

Históricamente, el hurto tiene sus raíces en el Derecho romano, donde las XII Tablas sentaron las bases legales y se diferenciaron infracciones como el hurto y el robo. En el ámbito del Derecho Penal español, desde el Fuero Juzgo hasta el Código de 1882, las sanciones fueron variando, incluyendo castigos pecuniarios y severos, como la pena de muerte para delitos de mayor gravedad. Esta evolución refleja la complejidad del fenómeno del hurto y su tratamiento a lo largo del tiempo (González, 2019).

La criminología, como disciplina científica, busca comprender las causas, patrones y consecuencias del comportamiento delictivo dentro de la sociedad (Del Olmo, 1999). A lo largo del tiempo, diversos enfoques teóricos han intentado explicar por qué ocurren los delitos y cómo pueden ser prevenidos o controlados. Estas teorías no solo proporcionan una comprensión profunda de los factores que influyen en la criminalidad, sino que también ofrecen un marco conceptual esencial para la aplicación de técnicas modernas en la predicción y prevención del crimen.

Entre las teorías más influyentes, Buil (2016) menciona la Escuela Clásica, Cartográfica, Positiva, de Chicago, la teoría de la asociación diferencial, de la anomia, del control, las corrientes críticas, las teorías de la oportunidad y la teoría de las ventanas rotas. La teoría de la asociación diferencial, de Sutherland, sostiene que el delito es aprendido a través de la interacción social, especialmente en grupos íntimos, donde se adquieren tanto las técnicas delictivas como las justificaciones antisociales (Matsueda, 2006; Sutherland et al., 1992).

La teoría de la anomia sugiere que cambios macrosociales pueden debilitar las normas sociales, lo que incrementa la criminalidad (Serrano-Maíllo, 2004). Por su parte, Hirschi (1969) argumenta en las teorías del control que vínculos sociales fuertes previenen la conducta delictiva, como apego, compromiso, implicación y creencias prosociales (Cid & Larrauri, 2001).

Las teorías de la oportunidad, desarrolladas entre los años 60 y 70, destacan que el entorno físico influye en la manifestación de comportamientos delictivos. Felson y Clarke (1998) proponen que las oportunidades delictivas son específicas, concentradas en tiempo y espacio, y pueden ser reducidas para prevenir delitos. Finalmente, la teoría de las ventanas rotas de

Wilson y Kelling (1982) postula que el desorden urbano fomenta la criminalidad al deteriorar el control comunitario.

Conocer los principios básicos de la criminología, al igual que las principales teorías criminológicas es esencial para entender cómo piensa y actúa el ser humano en contextos delictivos, lo que permite diseñar soluciones a estas problemáticas y, en particular, con el apoyo de *machine learning*, eje que constituye esta investigación permite alinear estas soluciones con las motivaciones y dinámicas del comportamiento criminal.

3. *Machine Learning*: Fundamentación teórica

Según Fernandes (2018), el *Machine Learning* es una rama de la inteligencia artificial que se enfoca en desarrollar algoritmos y modelos que permiten a las computadoras aprender y tomar decisiones a partir de datos, sin ser programadas explícitamente para cada tarea. Sirve para automatizar procesos, hacer predicciones, identificar patrones y tomar decisiones informadas en diversos campos, como la detección de fraudes, el reconocimiento facial, la personalización de contenidos, el marketing personalizado y mucho más. Existen varios tipos de modelos en *Machine Learning*, que se pueden clasificar principalmente en supervisados, no supervisados, semisupervisados y de aprendizaje por refuerzo (Kubat, 2017). A continuación, se presentan algunos de los modelos más reconocidos en la industria.

3.1. Modelos Supervisados

Estos modelos se entrenan con un conjunto de datos etiquetados, donde se conoce la salida deseada para cada entrada (Fernandes, 2018). Dentro de los modelos supervisados, se encuentran dos subcategorías principales: regresión y clasificación.

3.1.1. Regresión

La regresión se utiliza para predecir valores numéricos continuos. Kuhn y Johnson (2013) exponen entre los principales modelos de regresión, los siguientes:

- Regresión Lineal: Predice un valor numérico como una combinación lineal de las características de entrada.
- Regresión Polinómica: Extensión de la regresión lineal que puede capturar relaciones no lineales entre las características.
- Regresión Ridge y Lasso: Versiones regularizadas de la regresión lineal que previenen el sobreajuste.
- Decision Tree Regressor: Un modelo de árbol de decisión que divide los datos en segmentos homogéneos basándose en las características de entrada, haciendo predicciones mediante la media de los valores en las hojas del árbol. Es simple, pero puede ser propenso al sobreajuste.
- Random Forest Regressor: Un modelo de ensamble basado en la técnica de *bagging* (agregación) que combina múltiples árboles de decisión entrenados en diferentes subconjuntos del conjunto de datos. La predicción final es el promedio de las predicciones de todos los árboles, lo que reduce la varianza y mejora la precisión.
- XGBoost Regressor: Un modelo de *boosting* (potenciar) que combina secuencialmente árboles de decisión, donde cada nuevo árbol corrige los errores de los árboles

anteriores. Es conocido por su alta precisión y eficiencia computacional, especialmente en problemas con grandes conjuntos de datos.
- AdaBoost Regressor: Aplica *boosting* a modelos base (generalmente árboles de decisión simples), asignando mayor peso a las instancias mal predichas en cada iteración para mejorar la precisión en tareas de regresión.

Entre las medidas de desempeño para evaluar este tipo de modelos, presentan las siguientes:

- Error Cuadrático Medio (MSE): Promedio de los cuadrados de las diferencias entre los valores predichos y los valores reales.
- Error Absoluto Medio (MAE): Promedio de las diferencias absolutas entre los valores predichos y los valores reales.
- Coeficiente de Determinación (R^2): Proporción de la varianza en la variable dependiente que es predecible a partir de las variables independientes.

3.1.2. Clasificación

La clasificación se utiliza para asignar entradas a una de varias categorías o clases. Entre los principales modelos de clasificación, James et al. (2013) presentan y definen los siguientes:

- Regresión logística: Es un algoritmo de clasificación que estima la probabilidad de que una instancia pertenezca a una clase binaria (por ejemplo, sí/no, 0/1). A diferencia de la regresión lineal, que predice valores continuos, la regresión logística utiliza la función sigmoide para transformar las predicciones en probabilidades, que luego se clasifican en una de las dos clases. Es ampliamente utilizado para tareas de clasificación binaria como la detección de fraudes y la clasificación de correos electrónicos como spam o no spam.
- Naive Bayes: Modelos basados en la teoría de probabilidad que asumen independencia entre las características.
- Máquinas de Vectores de Soporte (SVM): Modelos que encuentran un hiperplano que separa las clases con el mayor margen posible.
- Redes Neuronales: Modelos inspirados en el cerebro humano que pueden aprender representaciones complejas de los datos.
- Decision Tree Classifier: Un modelo de clasificación que divide los datos en segmentos basados en características, asignando una clase a cada segmento. Es fácil de interpretar, pero puede sobre ajustarse si no se regula adecuadamente.
- Random Forest Classifier: Similar al *Random Forest Regressor*, pero utilizado para tareas de clasificación. Combina múltiples árboles de decisión, donde cada árbol vota por una clase, y la clase con más votos se convierte en la predicción final.
- XGBoost Classifier: Un modelo de *boosting* para clasificación que optimiza la precisión al corregir los errores de clasificación de los árboles anteriores. Es extremadamente popular en competiciones de *Machine Learning* debido a su alto rendimiento y capacidad para manejar datos desbalanceados.
- AdaBoost Classifier: Similar al *AdaBoost Regressor*, pero utilizado para tareas de clasificación, ajustando los pesos de las instancias mal clasificadas en cada ronda de entrenamiento.

Entre las principales medidas de desempeño para evaluar estos modelos, estos mismos autores presentan las siguientes:

- Exactitud (*Accuracy*): Proporción de predicciones correctas entre el total de predicciones.
- Precisión (*Precision*): Proporción de verdaderos positivos entre las predicciones positivas.
- *Recall* (Sensibilidad): Proporción de verdaderos positivos entre todos los casos positivos reales.
- Medida F1: Promedio armónico de precisión y *recall*, útil en situaciones de clases desbalanceadas.
- ROC-AUC: Área bajo la curva (ROC), que mide la capacidad del modelo para distinguir entre clases.

3.2. Modelos No Supervisados

James et al. (2013) exponen que los modelos no supervisados trabajan con datos no etiquetados y buscan descubrir patrones o estructuras ocultas en los datos.

3.2.1. *Clustering* (agrupamiento)

El *clustering* agrupa datos en subconjuntos donde los elementos dentro de cada grupo son más similares entre sí que a los de otros grupos (James et al., 2013). Entre los principales algoritmos de *clustering*, tenemos:

- K-Means: Agrupa los datos en k *clusters* (grupos), minimizando la distancia entre los puntos y el centroide del *cluster*.
- DBSCAN: Agrupa los puntos que están densamente conectados y trata los puntos fuera de esos grupos como ruido.
- Agrupamiento Jerárquico: Crea una jerarquía de *clusters* mediante la combinación de los más similares en cada paso.

Algunas de las principales medidas de desempeño presentadas por James et al. (2013), son:

- Silueta: Mide qué tan cerca están los puntos dentro de un *cluster* y qué tan lejos están de los puntos en otros *clusters*.
- Puntuación de Davies-Bouldin: Mide la relación entre la dispersión dentro de los *clusters* y la distancia entre *clusters*.
- Inercia: Suma de las distancias cuadráticas entre los puntos y el centroide del *cluster*.

3.2.2. Reducción de Dimensionalidad

Este enfoque se utiliza para reducir el número de características en un conjunto de datos mientras se conserva la mayor cantidad posible de información (James et al., 2013).

Entre los principales algoritmos de reducción de dimensionalidad, tenemos:

- Análisis de Componentes Principales (PCA): Transforma los datos a un espacio de menor dimensión, manteniendo la varianza máxima.
- t-SNE: Técnica no lineal para reducir la dimensionalidad, enfocada en la preservación de relaciones locales entre puntos.

Para medir el desempeño de este tipo de modelos, James et al. (2013) presentan las siguientes opciones:

- Varianza Explicada: Proporción de la varianza original de los datos que es capturada por los componentes principales.
- Calidad Visual: Evaluación cualitativa de la capacidad de la técnica para separar grupos en visualizaciones bidimensionales.

3.3. Modelos Semisupervisados

Estos modelos combinan una pequeña cantidad de datos etiquetados con una gran cantidad de datos no etiquetados para mejorar la precisión del modelo (Zhou, 2021).

3.4. Modelos de Aprendizaje por Refuerzo

Los modelos de aprendizaje por refuerzo aprenden a tomar decisiones a través de un sistema de recompensas y castigos en un entorno dinámico (Zhou, 2021).

Entre los principales modelos de Aprendizaje por Refuerzo, tenemos el Q-Learning, el cual es un modelo que aprende una política que maximiza la recompensa acumulada en un entorno. Entre las principales medidas de desempeño, tenemos:

- Recompensa Acumulada: Suma de todas las recompensas obtenidas durante un episodio o conjunto de episodios.
- Tiempo de Convergencia: Número de episodios necesarios para que el modelo aprenda una política óptima.
- Este panorama ofrece una visión estructurada de los diferentes tipos de modelos de *Machine Learning*, sus aplicaciones principales y cómo se evalúan sus desempeños.

3.5. Métodos de pronóstico de series de tiempo

Los métodos de series de tiempo se utilizan para analizar datos recolectados en intervalos regulares y hacer predicciones futuras. Se basan en patrones históricos como tendencia, estacionalidad y ruido, aprovechando estas regularidades para prever comportamientos futuros (Nielsen, 2019).

Los componentes de las series de tiempo son:

- **Tendencia (T):** Movimiento general de la serie a largo plazo (crecimiento, decrecimiento o constante).
- **Estacionalidad (S):** Fluctuaciones periódicas que se repiten en intervalos regulares (mensuales, trimestrales, anuales).
- **Ciclo (C):** Variaciones a largo plazo no regulares asociadas a factores económicos o externos.
- **Ruido o error (E):** Componentes aleatorios no explicados por los patrones anteriores.

A continuación, se describen algunos de los principales métodos de series de tiempo utilizados en la industria.

3.5.1. Método de descomposición

La descomposición separa una serie en sus componentes principales (Nielsen, 2019). Según la naturaleza de los datos, se utilizan modelos:

- **Aditivos:** Yt = Tt + St + Et, cuando los componentes tienen magnitudes constantes, es decir, la varianza es aproximadamente igual a lo largo del tiempo.
- **Multiplicativos:** = Tt*St*Et, cuando las fluctuaciones son proporcionales al nivel de la serie.

Esto facilita analizar cada componente por separado y mejorar la precisión en el pronóstico.

3.5.2. Métodos de suavizamiento

A continuación, se presentan los métodos de suavizamiento expuestos por Nielsen (2019):

- **Medias móviles:** Calculan un promedio móvil de los datos para eliminar fluctuaciones aleatorias y resaltar tendencias.
    - Utilizamos Medias Móviles Simples para series estacionarias sin estacionalidad.
    - Utilizamos Medias Móviles Dobles para series con tendencia lineal sin estacionalidad.
- **Suavizamiento exponencial:** Asigna pesos decrecientes a los datos más antiguos para captar tendencias recientes.
    - Se utiliza Suavizamiento Exponencial Simple para series estacionarias sin estacionalidad.
    - Se utiliza Suavizamiento Exponencial Doble para series con tendencia lineal sin estacionalidad.
    - Se utiliza Suavizamiento Exponencial Triple para series con tendencia cuadrática sin estacionalidad.
- **Método de Holt-Winters:** Útil para datos con tendencia y estacionalidad. Incluye coeficientes de suavizamiento para nivel (α, alpha), tendencia (β, beta) y estacionalidad (γ, gamma).
    - Alpha (nivel): Determina cuánto peso se le da al valor más reciente para actualizar el nivel general de la serie. Un valor alto significa que los datos recientes tienen mayor influencia.
    - Beta (tendencia): Controla cómo se ajusta la tendencia en función de los cambios recientes. Valores altos responden rápidamente a cambios, pero pueden introducir ruido.
    - Gamma (estacionalidad): Ajusta la importancia de los patrones estacionales en los datos. Un gamma alto prioriza fluctuaciones recientes en la estacionalidad.

3.5.3. Método Box-Jenkins:

El método Box-Jenkins es un enfoque iterativo utilizado para identificar, estimar y diagnosticar modelos de series de tiempo, especialmente modelos ARIMA (AutoRegressive Integrated Moving Average) (Nielsen, 2019).

- **Identificación del modelo:** La primera etapa consiste en identificar el modelo adecuado para los datos. Esto implica observar la serie de tiempo y luego analizar gráficos de ACF (AutoCorrelation Function) y PACF (Partial AutoCorrelation Function) para determinar los parámetros p, d y q.
    - ACF (AutoCorrelation Function o Función de autocorrelación): La ACF mide la correlación entre los valores de la serie y sus rezagos (valores pasados). Ayuda a identificar la cantidad de autocorrelación a diferentes rezagos. La ACF es útil para identificar el parámetro q (el número de términos de media móvil en un modelo ARIMA).
    - PACF (Partial AutoCorrelation Function o Función de autocorrelación parcial): La PACF mide la correlación entre un valor y sus rezagos, eliminando la influencia de los rezagos intermedios. Esto ayuda a identificar el parámetro p(el número de términos autoregresivos en el modelo ARIMA). Si la PACF cae bruscamente después de un determinado rezago, ese rezago puede indicar el valor de p.
- **Estimación de los parámetros:** Una vez que se han identificado los posibles valores de p, d y q, se ajusta el modelo ARIMA correspondiente utilizando métodos de estimación (generalmente, máxima verosimilitud). Esta etapa se centra en encontrar los valores óptimos de los parámetros del modelo para los cuales el error de predicción es el más pequeño posible.
    - p: Es el número de términos autoregresivos (AR). Se refiere a la cantidad de rezagos anteriores de la serie que se utilizan para predecir el valor actual.
    - d: Es el número de diferenciaciones necesarias para hacer que la serie sea estacionaria. En otras palabras, se refiere a cuántas veces debes restar los valores previos para que la serie no tenga una tendencia.
    - q: Es el número de términos de media móvil (MA). Representa la cantidad de rezagos de los errores pasados que se incorporan al modelo.
- **Validación de supuestos:** Esta etapa implica revisar si los residuos del modelo ajustado (la diferencia entre las predicciones y los valores reales) son aleatorios. Para ello, se analiza si los residuos tienen una distribución normal, si no hay autocorrelación (usando ACF y PACF de los residuos) y si su media es cero. Los residuos deben ser independientes y no mostrar patrones. Si el modelo no cumple con estos criterios, se deben ajustar los parámetros o intentar otros modelos hasta lograr un buen ajuste.

Metodología

1. Enfoque de la investigación

El enfoque de esta investigación fue cuantitativo, ya que se basó en el análisis de datos numéricos provenientes de la base de datos de incidentes delictivos de Bucaramanga, con el fin de desarrollar modelos predictivos utilizando técnicas de Machine Learning. Este enfoque permitió analizar patrones y relaciones entre variables socioeconómicas, demográficas y de criminalidad para identificar zonas de alta incidencia de hurtos y predecir la ocurrencia de estos delitos. La naturaleza cuantitativa del estudio permitió obtener resultados objetivos y medibles, lo cual fue crucial para proporcionar recomendaciones concretas a las autoridades locales.

El alcance de la investigación fue de tipo descriptivo y correlacional. Fue descriptivo porque buscó analizar y caracterizar los patrones de hurtos en Bucaramanga, utilizando datos históricos para identificar las zonas de mayor incidencia delictiva. También fue correlacional, ya que exploró las relaciones entre diversas variables (como características socioeconómicas y demográficas) y la ocurrencia de hurtos, buscando identificar posibles factores que influencian la criminalidad en la ciudad.

El diseño de la investigación fue no experimental, ya que no se manipularon variables, sino que se trabajó con los datos ya existentes sobre hurtos en Bucaramanga. Además, fue un estudio transversal, realizado en un único momento, analizando los datos históricos disponibles hasta octubre de 2023. El estudio tuvo un carácter aplicado, puesto que el objetivo final fue ofrecer recomendaciones prácticas y accionables para que las autoridades locales optimicen la distribución de recursos para la prevención del delito, basadas en los resultados del modelo predictivo desarrollado.

2. Etapas

Etapas	Alcance	Actividad
Etapa 1: Recolección de información	Garantizar la obtención de un conjunto de datos relevante, confiable y suficiente para abordar los objetivos del análisis.	Identificación y selección de fuentes de datos relevantes.
Etapa 2: Exploración de Datos	Preparar y explorar los datos para identificar tendencias iniciales.	Análisis exploratorio de los datos: desarrollo de los análisis descriptivos para identificar patrones y correlaciones iniciales.
		Georreferenciación de los datos de criminalidad para mapear la incidencia delictiva en Bucaramanga.
		Selección de variables: Obtención del conjunto de datos limpio y organizado con las variables seleccionadas para el análisis de *Machine Learning*.
Etapa 3: Desarrollo y Aplicación de Modelos de *Machine Learning*	Identificar por medio de *machine learning* las zonas en Bucaramanga con características similares en cuanto a la ocurrencia de hurtos a personas, utilizando variables socioeconómicas, demográficas y de criminalidad	Aplicación de *machine learning* para agrupar zonas con características similares en cuanto a criminalidad y variables socioeconómicas.
		Evaluación de la efectividad del clustering y ajustar parámetros para mejorar la precisión.
		Organización de los datos históricos para aplicar un

	Desarrollar un modelo de series de tiempo para predecir la frecuencia y probabilidad de hurto en Bucaramanga en el año 2025.	modelo predictivo de series de tiempo.
		Entrenamiento del modelo con los datos históricos de hurtos y validar su precisión con datos recientes.
		Eso del modelo de series de tiempo para predecir la frecuencia de hurtos en Bucaramanga durante el año 2025.
Etapa 4: Desarrollo de visualizador de zonas en alto riesgo	Desarrollar un visualizador que identifique las zonas de mayor riesgo de hurtos en la ciudad de Bucaramanga, con el objetivo de generar un mapa de riesgo que facilite a las autoridades focalizar intervenciones preventivas en las áreas más vulnerables.	Desarrollo de visualizador con zonas de mayor riesgo
		Análisis de resultados

Tabla 1

Etapas de desarrollo

Nota: Elaboración propia

2.1. Etapa 1: Recolección de información

En esta etapa, se realizó un proceso sistemático para identificar y recopilar datos relevantes, confiables y adecuados al problema de estudio. Se exploraron diversas fuentes, incluyendo plataformas de datos abiertos como la de Datos Abiertos del Gobierno de Colombia, los registros estadísticos del DANE y bases de datos específicas de la Policía Nacional relacionadas con hurtos. Estas fuentes fueron analizadas para seleccionar los conjuntos de datos más pertinentes, asegurando su alineación con los objetivos del estudio y el problema planteado.

2.2. Etapa 2: Exploración de Datos

En la segunda etapa, se alcanzó el objetivo de preparar y explorar los datos para identificar tendencias iniciales, lo que permitió una comprensión más profunda del comportamiento delictivo en Bucaramanga y estableció una base sólida para los análisis futuros.

- **Análisis exploratorio de los datos:** Se realizó un análisis descriptivo para identificar patrones y correlaciones iniciales en los datos de criminalidad. A través de estadísticas básicas como frecuencias, promedios y distribuciones, se detectaron las concentraciones de delitos en ciertos barrios y comunas, así como su variación en función de factores como el horario, el día de la semana y el tipo de delito. Este análisis permitió revelar posibles concentraciones delictivas en horarios nocturnos o en barrios específicos, que se utilizaron como insumos para las etapas posteriores.

- **Georreferenciación de los datos:** Con el uso de coordenadas de latitud y longitud, se mapearon los delitos ocurridos en Bucaramanga mediante técnicas de georreferenciación. Esto facilitó la creación de visualizaciones, como mapas de calor, que mostraron las zonas con mayor concentración de criminalidad. Los resultados iniciales identificaron los puntos críticos en barrios específicos donde el hurto y otros delitos violentos tienden a concentrarse, lo que ayudó a determinar áreas prioritarias para la intervención.

- **Selección de variables:** Se procedió a limpiar exhaustivamente los datos y seleccionar las variables clave para el análisis de Machine Learning. Entre ellas estuvieron las variables demográficas como la edad y el género de las víctimas, características del delito como la modalidad, el tipo de arma utilizada, el móvil del agresor, así como las variables temporales y geográficas. El conjunto de datos quedó organizado y listo para ser utilizado en las siguientes etapas, optimizando así la calidad y relevancia de la información para los futuros modelos predictivos.

Con estas actividades completadas, se logró preparar los datos adecuadamente, cumpliendo con el objetivo propuesto y estableciendo un punto de partida sólido para los análisis avanzados que siguieron en las siguientes fases del proyecto.

2.3. Etapa 3: Desarrollo y Aplicación de Modelos de *Machine Learning*

En la etapa 3, se alcanzó el objetivo de identificar las zonas en Bucaramanga con características similares en cuanto a la ocurrencia de hurtos a personas, mediante el uso de técnicas de *machine learning* que consideraron variables socioeconómicas, demográficas y de criminalidad.

- **Aplicación de *machine learning* para agrupar zonas**: Se implementaron algoritmos de *clustering*, como *K-means* o DBSCAN (Agrupamiento Espacial Basado en Densidad de Aplicaciones con Ruido), para agrupar zonas con características similares en cuanto a criminalidad y variables socioeconómicas. Estas técnicas permitieron identificar patrones ocultos en los datos, donde los barrios y comunas con perfiles delictivos similares se agruparon en clústeres. Estas agrupaciones ayudaron a segmentar Bucaramanga en áreas con niveles de riesgo comparables, facilitando una mejor comprensión de las dinámicas delictivas a nivel local. Además, se utilizaron variables como la frecuencia de hurtos, el tipo de arma, el perfil demográfico de las víctimas y las condiciones socioeconómicas de las zonas para construir estos modelos.

- **Evaluación y ajuste de parámetros**: Una vez obtenidos los resultados del clustering, se evaluó la efectividad de los modelos mediante métricas como la inercia o la silueta. Este análisis ayudó a determinar la coherencia interna de los clústeres y la capacidad del modelo para distinguir correctamente entre zonas con diferentes perfiles delictivos. En función de los resultados, se ajustaron los parámetros de los modelos, como el número de clústeres o el peso de ciertas variables, para mejorar la precisión y fiabilidad de las predicciones. Esta etapa también incluyó la validación cruzada para asegurar que el modelo generalizara adecuadamente en nuevos datos.

En esta etapa también se desarrolló el objetivo de crear un modelo de series de tiempo para predecir la frecuencia y probabilidad de hurto en Bucaramanga en el año 2025.

- **Organización de datos históricos para series de tiempo**: Se recopilaron y organizaron los datos históricos de criminalidad, asegurando que estuvieran preparados para ser utilizados en un modelo predictivo de series de tiempo. Esto implicó la estructuración en intervalos de tiempo regulares, como días, semanas o meses, considerando las variaciones temporales en la frecuencia de los hurtos.

- **Entrenamiento y validación del modelo de series de tiempo**: Se entrenó el modelo predictivo utilizando los datos históricos de hurtos en Bucaramanga. Se emplearon técnicas para identificar patrones en los datos, y se validó la precisión del modelo utilizando los datos más recientes disponibles, ajustando los parámetros del modelo para mejorar sus predicciones.

- **Predicción de la frecuencia de hurtos en 2025**: Con el modelo entrenado y validado, se utilizaron las predicciones para estimar la frecuencia de hurtos en Bucaramanga durante el año 2025. Esto proporcionó una proyección de los picos de criminalidad y permitió planificar estrategias preventivas y la asignación de recursos policiales de manera más eficiente.

Con la culminación de estas actividades, se identificó las zonas con características delictivas similares en Bucaramanga, lo que proporcionará una herramienta clave para la toma de decisiones en materia de seguridad y la asignación eficiente de recursos. No solo se identificó las zonas de riesgo mediante *clustering*, sino también se generó predicciones sobre la frecuencia de hurto para 2025, cumpliendo con los objetivos clave del proyecto.

Las siguientes son las variables usadas para el cumplimiento de los objetivos en la etapa.

Objetivo	Variables y tipos de variable	Definición operacional	Definición conceptual
Identificar por medio de *machine learning* las zonas en Bucaramanga con	BARRIOS_HECHO (Independiente)	Descripción categórica (nombre del barrio, texto)	Barrio donde ocurre el delito según la Policía.
	LOCALIDAD (Independiente)	Descripción categórica (nombre de la localidad, texto)	Comuna donde ocurre el delito según la Policía.
	NUM_COM (Independiente)	Número de comuna (número entero)	Número de la comuna según planeación.

características similares en cuanto a la ocurrencia de hurtos a personas, utilizando variables socioeconómicas, demográficas y de criminalidad.	NOM_COM (Independiente)	Descripción categórica (nombre de la comuna, texto)	Nombre de la comuna según planeación.
	ARMAS_MEDIOS (Independiente)	Descripción categórica (tipo de arma/medio empleado, texto)	Descripción del arma usada en los delitos según la Policía.
	SEXO (Independiente)	Categórica (Masculino, Femenino, Otro)	Sexo de la víctima del delito según la Policía.
	EDAD (Independiente)	Años (número entero)	Edad de la víctima del delito según la Policía.
	CURSO_VIDA (Independiente)	Descripción categórica (etapas de la vida, texto)	Clasificación por rango de edades de la víctima.
	CLASE_SITIO (Independiente)	Descripción categórica (tipo de lugar, texto)	Tipo de sitio en el cual ocurre el delito según la Policía.
	DESCRIPCION_CONDUCTA (Dependiente)	Descripción categórica (texto)	Descripción completa con número de artículo y descripción de la conducta según la Policía.
Desarrollar un modelo de series de tiempo para predecir la frecuencia y probabilidad de hurto en Bucaramanga en el año 2025.	FECHA_HECHO (Independiente)	Fecha (día/mes/año)	Fecha cuando se generan los hechos del delito según la Policía.
	HORA_HECHO (Independiente)	Hora exacta (formato HH o categórica por rango horario)	Hora cuando se generan los hechos del delito según la Policía.
	AÑO_NUM (Independiente)	Año (número entero)	Año en el que ocurre el delito según la Policía.
	MES_NUM (Independiente)	Número del mes (número entero de 1 a 12)	Mes en el que ocurre el delito según la Policía.
	DIA_NUM (Independiente)	Número del día (número entero de 1 a 31)	Día en el que ocurre el delito según la Policía.

	RANGO_HORARIO (Independiente)	Categórica (intervalo de horas, texto o número entero)	Clasificación por rango de horas en las que sucede el delito.
	DELITO_SOLO (Dependiente)	Descripción categórica (tipo de delito, texto)	Descripción del delito únicamente.
	CANTIDAD_UNICA (Dependiente)	Cantidad (número entero, total de delitos registrados)	Cantidad representativa de afectados.

Tabla 2

Varibles para desarrollo de modelos

Nota: Elaboración propia

2.4. Etapa 4: Desarrollo de visualizador de zonas en alto riesgo

En esta etapa se desarrolló un visualizador interactivo para identificar las zonas con mayor riesgo de hurtos en Bucaramanga, basado en el análisis de datos históricos y técnicas de visualización geoespacial. Además, se evaluaron los resultados obtenidos para garantizar que el visualizador cumpliera con los objetivos de facilitar la focalización de intervenciones preventivas y optimizar la asignación de recursos en las áreas más vulnerables.

- **Desarrollo de visualizador con zonas de mayor riesgo:** Para desarrollar el visualizador de zonas en alto riesgo, se consolidaron los datos históricos de hurtos en Bucaramanga, incluyendo variables como ubicación geográfica, frecuencia de ocurrencia y desagregación temporal. Se generaron mapas de calor utilizando herramientas como Folium, una biblioteca de Python para la creación de mapas interactivos (Folium, s.f.), y Kepler.gl, una plataforma de visualización geoespacial basada en la web desarrollada por Uber (Kepler.gl, s.f.). Estas herramientas permitieron resaltar las áreas de mayor incidencia delictiva mediante gradientes de color, sin embargo, este desarrollo no se pudo integrar al resultado final por dificultades de compatibilidad. Además, se aplicaron algoritmos de clusterización, como K-Means, para identificar patrones de riesgo y agrupar zonas con características similares. El visualizador, implementado con Power BI, una herramienta de inteligencia empresarial desarrollada por Microsoft, que permite conectar, visualizar y analizar datos de manera interactiva (Power BI, s.f.). El visualizador incluye filtros interactivos que permite exploración de los datos históricos.

- **Análisis de resultados**: Se llevó a cabo un análisis exhaustivo de los resultados generados por los modelos de *clustering* y series de tiempo. Esto incluyó la identificación de patrones clave en la ocurrencia de hurtos, la comparación de las predicciones con los datos reales recientes y la evaluación de la precisión del modelo predictivo para el año 2025. Este análisis proporcionó una visión clara de las zonas con mayor riesgo de hurtos.

Con esta etapa finalizada, se generaron recomendaciones prácticas y basadas en datos, proporcionando a las autoridades las herramientas necesarias para mejorar la seguridad pública y reducir la incidencia de hurtos en Bucaramanga.

Las siguientes son las variables usadas para el cumplimiento de los objetivos en la etapa.

Objetivos	Variables y tipos de variable	Definición operacional	Definición conceptual
Desarrollar un visualizador que identifique las zonas de mayor riesgo de hurtos en la ciudad de Bucaramanga, con el objetivo de generar un mapa de riesgo que facilite a las autoridades focalizar intervenciones preventivas en las áreas más vulnerables.	MOVIL_VICTIMA (Independiente)	Descripción categórica (texto)	Móvil en el que se desplazaba la víctima en el momento del delito según la Policía.
	MOVIL_AGRESOR (Independiente)	Descripción categórica (texto)	Móvil en el que se desplazaba el agresor en el momento del delito según la Policía.
	CLASE_SITIO (Independiente)	Descripción categórica (tipo de lugar, texto)	Tipo de sitio en el cual ocurre el delito según la Policía.
	ARMAS_MEDIOS (Independiente)	Descripción categórica (tipo de arma/medio empleado, texto)	Descripción del arma usada en los delitos según la Policía.
	TIPOLOGÍA (Dependiente)	Descripción categórica (tipo de delito, texto)	Clasificación del delito según el orden jurídico al que pertenece el mismo.
	DESCRIPCION_CONDUCTA (Dependiente)	Descripción categórica (texto)	Descripción completa con número de artículo y descripción de la conducta según la Policía.

Tabla 3

Varibles para desarrollo de visualizador

Nota: Elaboración propia

Hipótesis

H1. Las zonas de Bucaramanga con mayor incidencia de hurtos a personas comparten características demográficas.

H2. La frecuencia de hurtos a personas en Bucaramanga sigue un patrón cíclico influenciado por factores temporales como la hora del día, el día de la semana y el mes del año.

Análisis de resultados y conclusiones

1. Etapa 1: Recolección de información

La base de datos de Información delictiva de la ciudad de Bucaramanga está enfocada en el ámbito de Seguridad y Defensa, proporcionando un registro detallado de los delitos ocurridos en la ciudad. Esta información cubre las denuncias realizadas por las víctimas el período comprendido entre enero de 2016 y octubre de 2023, permitiendo un análisis exhaustivo y longitudinal de la actividad delictiva en Bucaramanga. Los datos en cuestión fueron recuperados de la plataforma de Datos Abiertos de Colombia, la cual facilita el acceso a este tipo de información pública para el desarrollo de estudios de seguridad y prevención del delito (Datos Abiertos Colombia, 2023)

Los datos están organizados por modalidad y conducta delictivas, lo que facilita la identificación de patrones y tendencias en los tipos de delitos cometidos. También se incluye una segmentación geográfica basada en los barrios y comunas de ocurrencia, proporcionando una visión clara de las zonas con mayor incidencia delictiva.

La fuente de datos detalla las armas y medios empleados en la comisión de los delitos, junto con el móvil del agresor y de la víctima, lo que permite un análisis profundo de las dinámicas delictivas. También se recoge el curso de vida de las víctimas, es decir, su edad y etapa de desarrollo, así como su género, brindando una perspectiva sociodemográfica sobre las personas involucradas.

Un componente crucial de la base de datos es la desagregación temporal de los delitos, que registra la fecha y hora de ocurrencia por mes, día y hora, lo que permite identificar patrones temporales en la criminalidad, como los momentos del día o las épocas del año con mayor incidencia de delitos.

Los datos son suministrados por la Policía Metropolitana de Bucaramanga, en coordinación con la Secretaría del Interior de la Alcaldía de Bucaramanga, que es la entidad responsable de la gestión de la seguridad y el orden público en el municipio.

Información adicional de la información:

- **Departamento:** Santander
- **Municipio:** Bucaramanga
- **Nombre de la Entidad:** Alcaldía de Bucaramanga
- **Orden:** Territorial
- **Sector:** Defensa
- **Área o Dependencia:** Secretaría del Interior
- **Propietario del Conjunto de Datos:** Alcaldía de Bucaramanga

2. Etapa 2: Exploración de Datos

En esta etapa se analizaron los datos de georreferenciación de los delitos ocurridos en el municipio de Bucaramanga según la modalidad, conducta, móvil del agresor, móvil de la víctima, comunas de ocurrencia, fatales, no fatales y violencia sexual desagregado por curso de vida, sexo, mes y día de ocurrencia, esto para el periodo comprendido entre 2016 y octubre de 2023.

La Figura 1 presenta el comportamiento del número de delitos en la ciudad de Bucaramanga a lo largo del periodo 2016-2023:

Figura 1

Comportamiento del número de delitos en la ciudad de Bucaramanga

Nota: Elaboración propia

Se puede observar una fuerte tendencia al alza en el número de delitos a lo largo de los años, con una disminución fuerte en 2020, lo cual se debe al inicio del SARS-CoV-2. Por otra parte, dado que en 2023 no se tienen las cifras del año completo, se presenta un decrecimiento en la gráfica.

En la Figura 2 se muestra la distribución del número de delitos para cada una de las comunas de la ciudad de Bucaramanga, ocurridos entre 2016 y octubre de 2023:

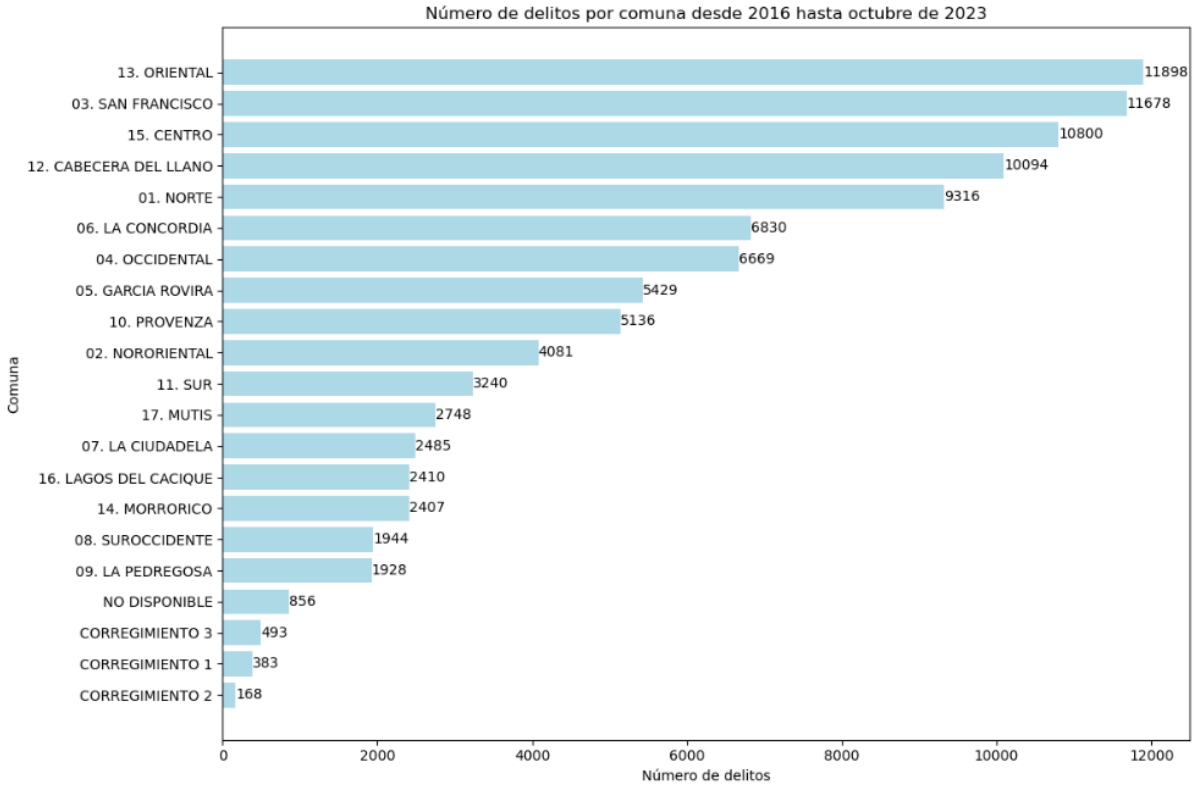

Figura 2

Distribución del número de delitos por comuna

Nota: Elaboración propia

Se observa que, en el periodo antes mencionado, la comuna con mayor cantidad de delitos presentados es ORIENTAL, con 11.898, seguida de las comunas SAN FRANCISCO y CENTRO con 11.678 y 10.800, respectivamente.

Concentrándose en un año particular, por ejemplo, 2022 (último año de análisis completo del *dataset* explorado), es posible tomar como referencia el Plan de Desarrollo de Bucaramanga 2020-2023, en el cual se presenta la población de cada una de las comunas de Bucaramanga a cierre 2020 como se muestra en la Tabla 4. La Figura 3 muestra el número de delitos por año en base a la población del plan de desarrollo.

Comuna	Población
Comuna 1. Norte	61.583
Comuna 2. Nororiental	39.781
Comuna 3. San Francisco	50.712
Comuna 4. Occidental	43.365
Comuna 5. García Rovira	47.845
Comuna 6. La Concordia	31.956
Comuna 7. Ciudadela	32.852
Comuna 8. Sur Occidente	20.906
Comuna 9. La Pedregosa	18.815
Comuna 10. Provenza	36.675
Comuna 11. Sur	32.315

Comuna 12. Cabecera del Llano	37.930
Comuna 13. Oriental	59.373
Comuna 14. Morrorico	26.043
Comuna 15. Centro	9.796
Comuna 16. Lagos del Cacique	17.024
Comuna 17. Mutis	30.344

Tabla 4

Población por comuna en la ciudad de Bucaramanga

Nota: Tomado de Alcaldía de Bucaramanga, 2022

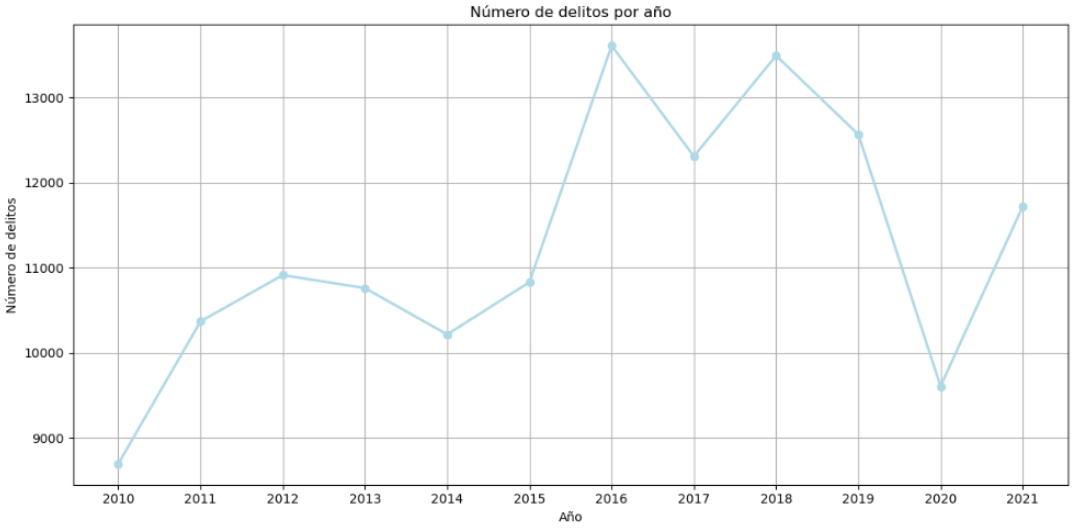

Figura 3

Número de delitos por año con relación a la población de las comunas

Nota: Elaboración propia

Dicho lo anterior, para 2022 se puede visualizar para cada una de las comunas, el número de delitos ocurridos por cada 1.000 habitantes según se muestra en la Figura 4.

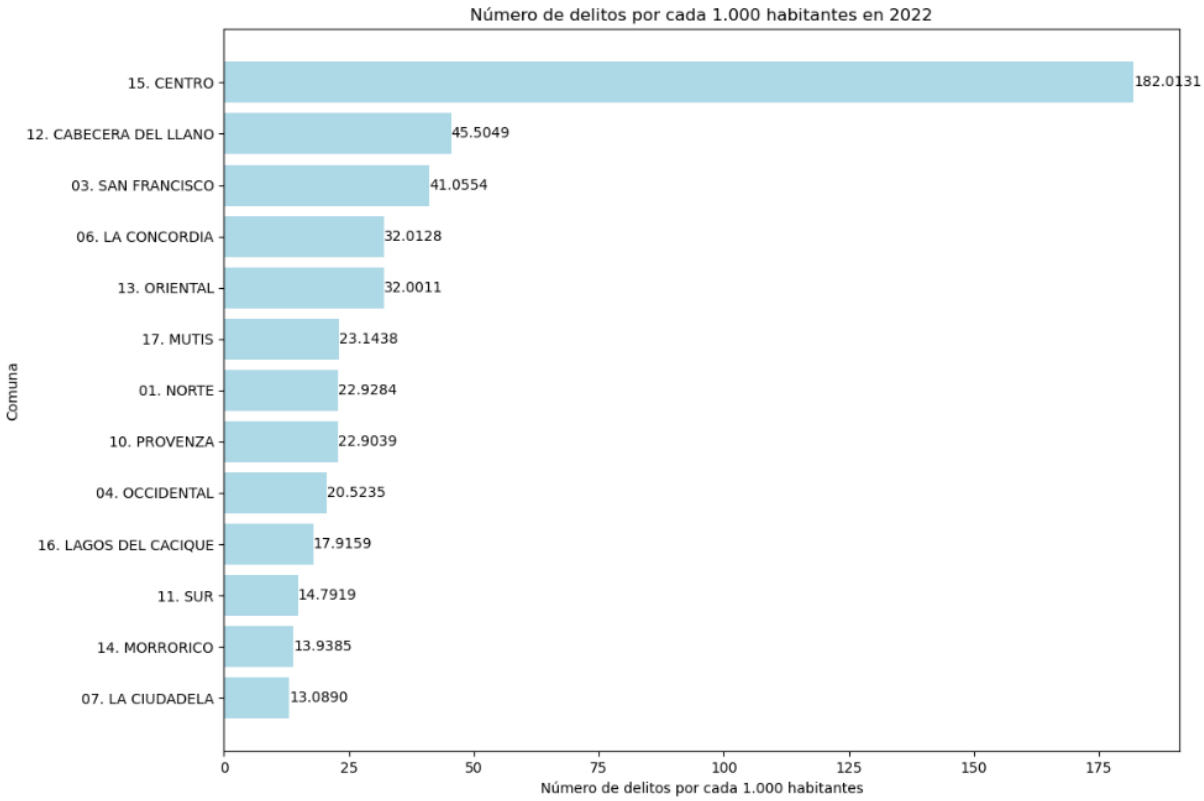

Figura 4

Número de delitos por comuna por cada 1000 habitantes en 2022

Nota: Elaboración propia

Es evidente que, con gran diferencia, la zona centro resulta ser la más peligrosa cuando se utiliza como métrica el número de delitos por cada 1.000 habitantes. Esto se justifica considerando que, al tratarse de una zona principalmente comercial, el número de eventos delincuenciales es elevado, mientras que la cantidad de residentes en dicha comuna es considerablemente baja. Al analizar comunas predominantemente residenciales, se identifica que la comuna Cabecera del Llano es la que presenta el mayor nivel de peligrosidad.

Adicionalmente, la Figura 5 ilustra la clara sectorización de los robos en la ciudad de Bucaramanga, destacando en rojo las zonas más propensas a estos delitos. Esta visualización respalda los datos previamente presentados sobre las comunas con mayor cantidad de casos.

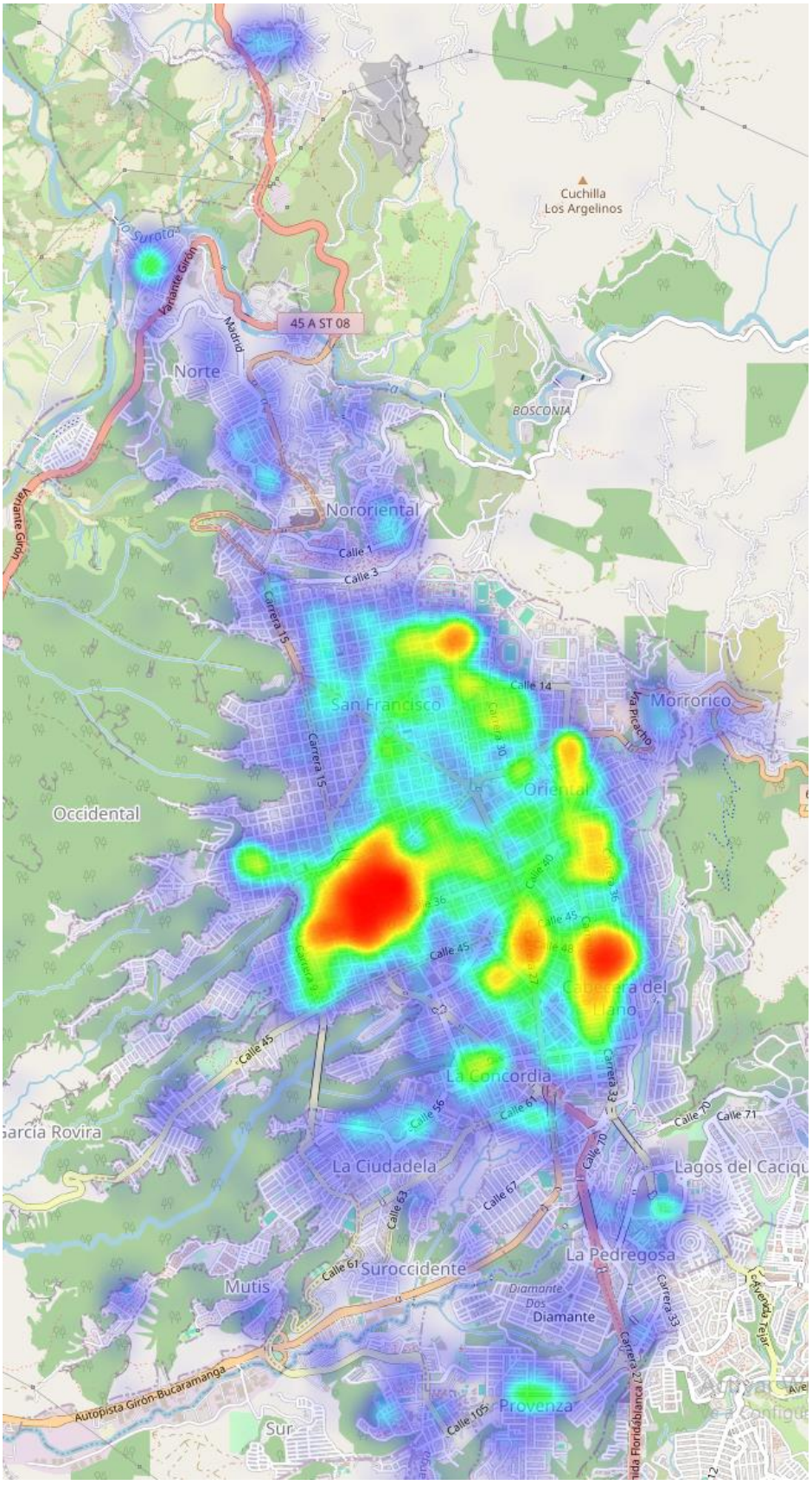

Figura 5

Mapa de calor de delitos en 2022

Nota: Elaboración propia

Algo que también es importante conocer es la frecuencia de los delitos por tipo de arma o medio utilizado para cometer el acto. En el caso de Bucaramanga, para el periodo 2016-2023 se tiene presentado en la Figura 6.

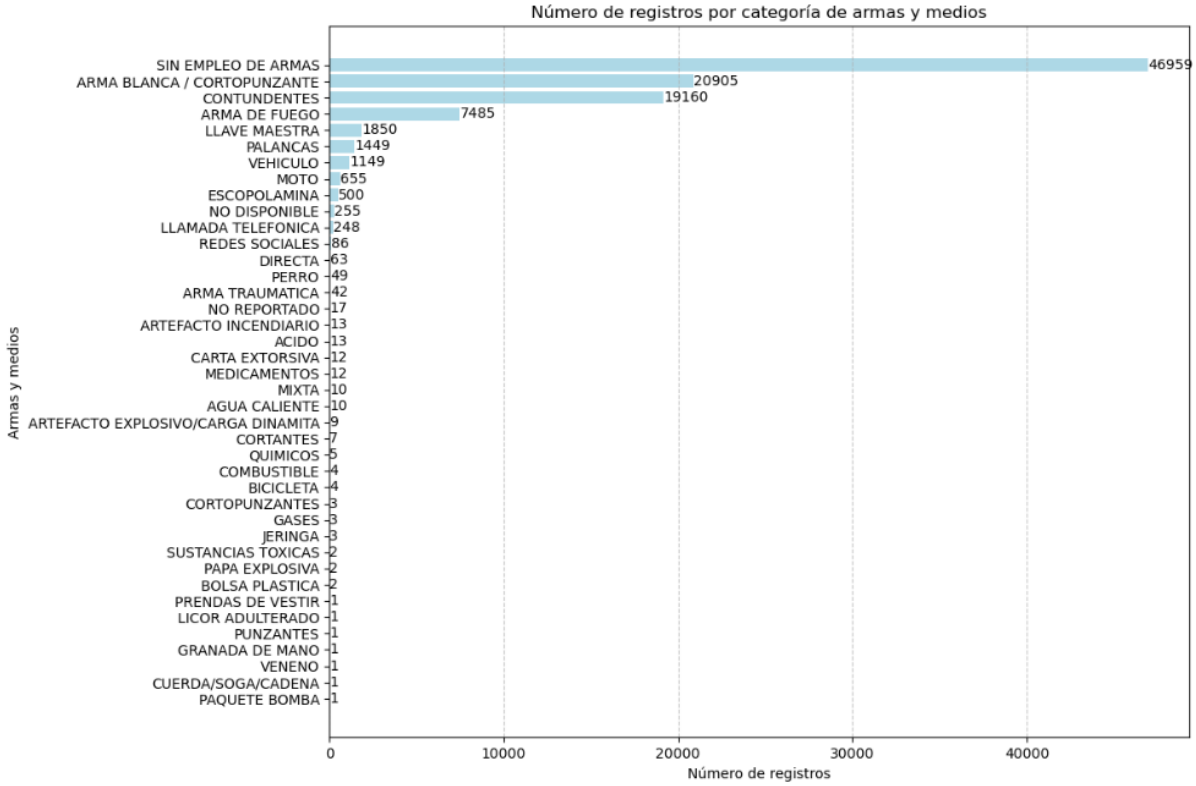

Figura 6

Número de registros por categoría de arma y medios

Nota: Elaboración propia

En este caso, el 86% de los delitos (87.024) se realiza sin empleo de armas, con armas blancas y cortopunzantes, o con armas contundentes.

En cuanto al día de la semana como lo muestra la Figura 7, es lógico pensar que el sábado, al ser el día más comercial y económicamente activo, tendrá el mayor número de delitos.

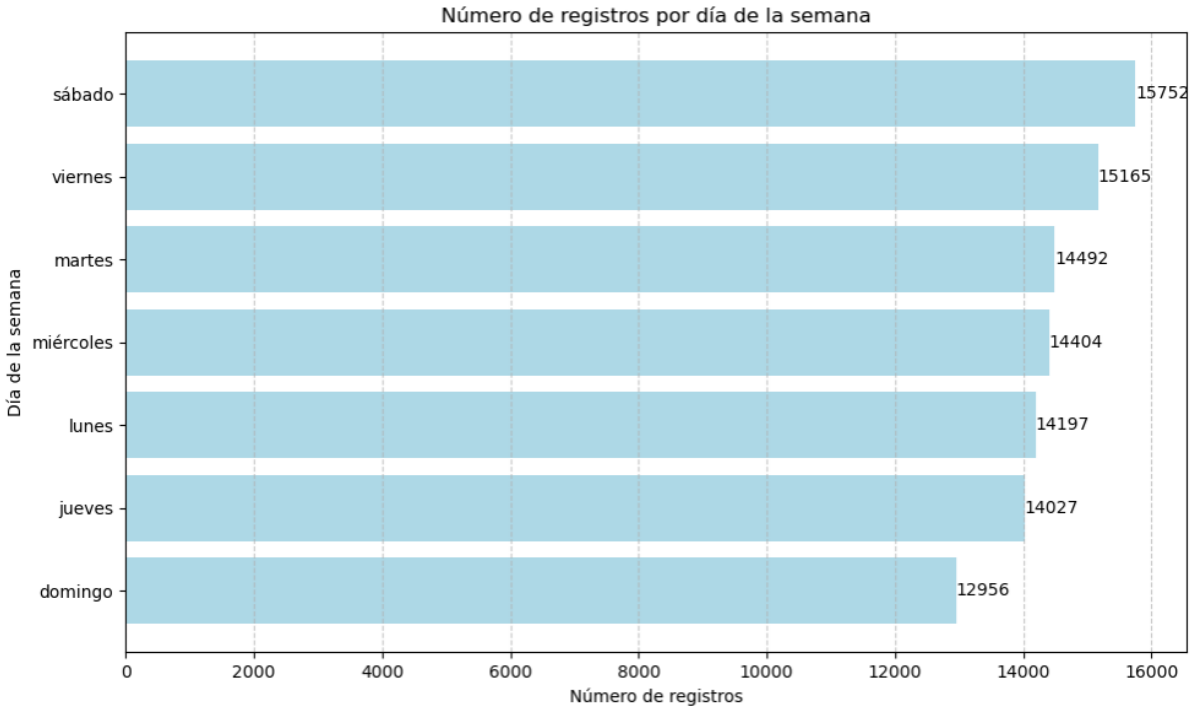

Figura 7

Número de registros por día de la semana

Nota: Elaboración propia

Se observa que, si bien el número de delitos se presenta en mayor medida durante este día, la diferencia no es muy alta con respecto a las demás clases. De hecho, pareciera que el número de delitos se distribuye de manera balanceada entre todos los días de la semana, por lo que no es posible concluir algo específico para esta variable.

Por otra parte, en cuanto al tipo de delito o conducta, se tiene la siguiente distribución mostrada en la Figura 8.

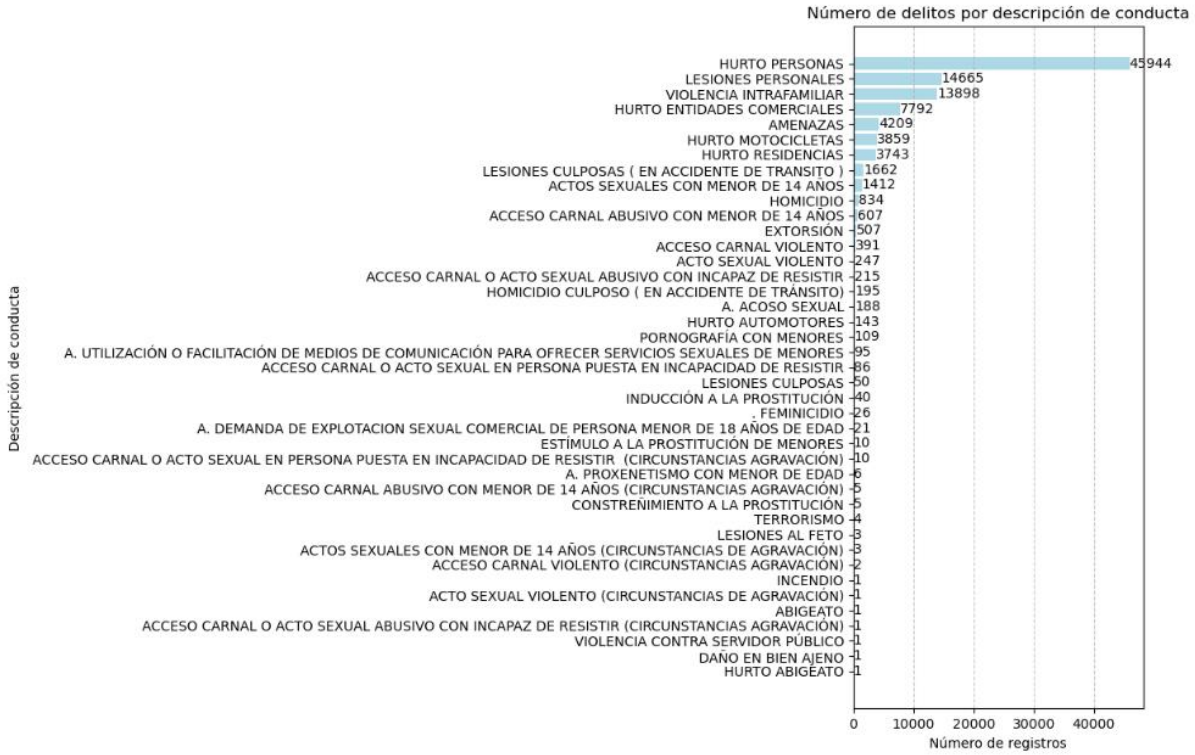

Figura 8

Tipo de delito

Nota: Elaboración propia

Como era de esperar, el hurto a personas ocupa el primer puesto en cuanto a frecuencia de ocurrencia, representando casi el 46% de los delitos.

Algo que también es interesante analizar, es la relación que existe entre el tipo de arma o medio utilizado y el tipo de delito cometido, con el género de la víctima, la cual es otra de las variables que se presentan en el *dataset*.

La Figura 9 se presenta el porcentaje de delitos cuya víctima es mujer, para cada uno de los tipos de arma que se presentan

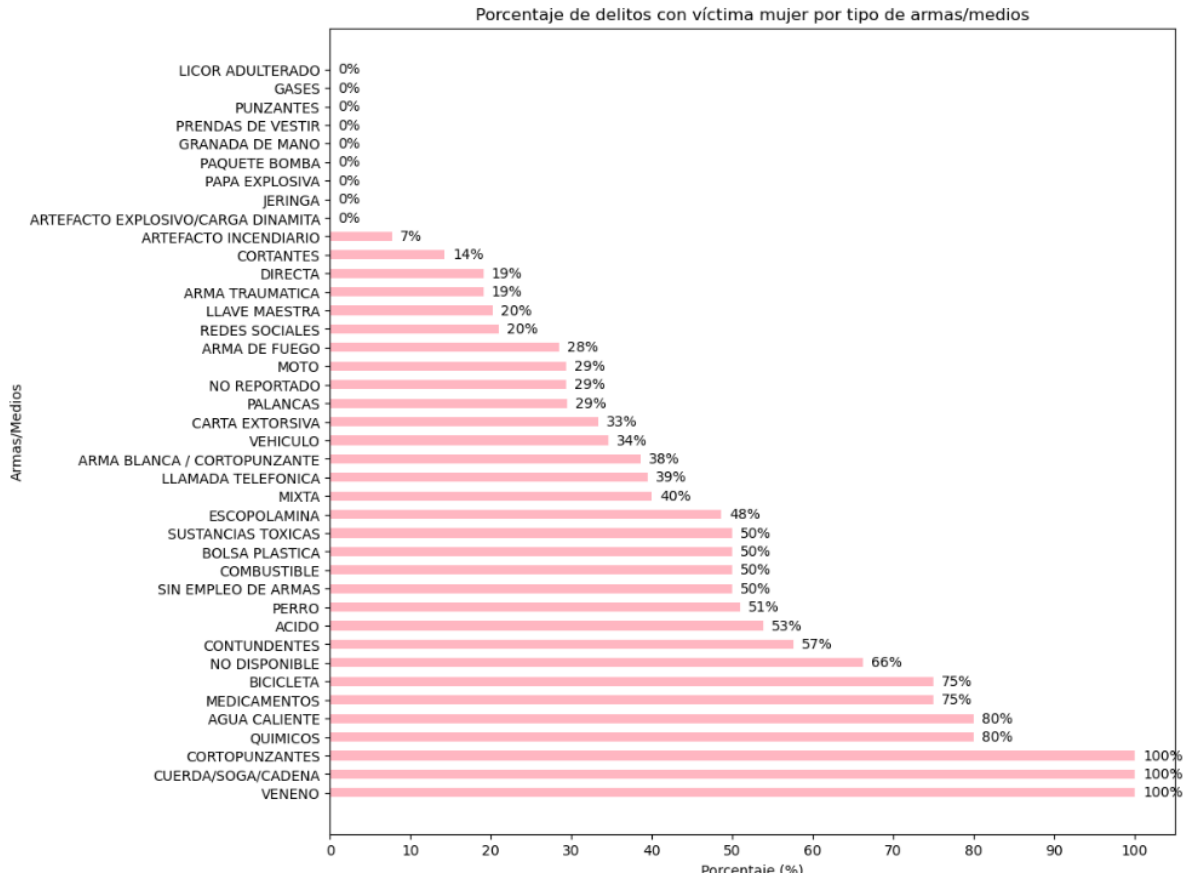

Figura 9

Porcentaje de delitos con mujeres como víctimas

Nota: Elaboración propia

Es notable que armas como veneno, cuerdas, cortopunzantes, químicos y agua caliente, son las más utilizadas en delitos de feminicidio. En particular y adhiriéndose a lo hallado en el *dataset*, el veneno, las cuerdas o sogas y las armas cortopunzantes siempre son utilizadas para atentar contra mujeres.

Por otra parte, se visualiza en la Figura 10 el porcentaje de delitos cuya víctima es mujer, para cada uno de los tipos de delito presentados en el *dataset*.

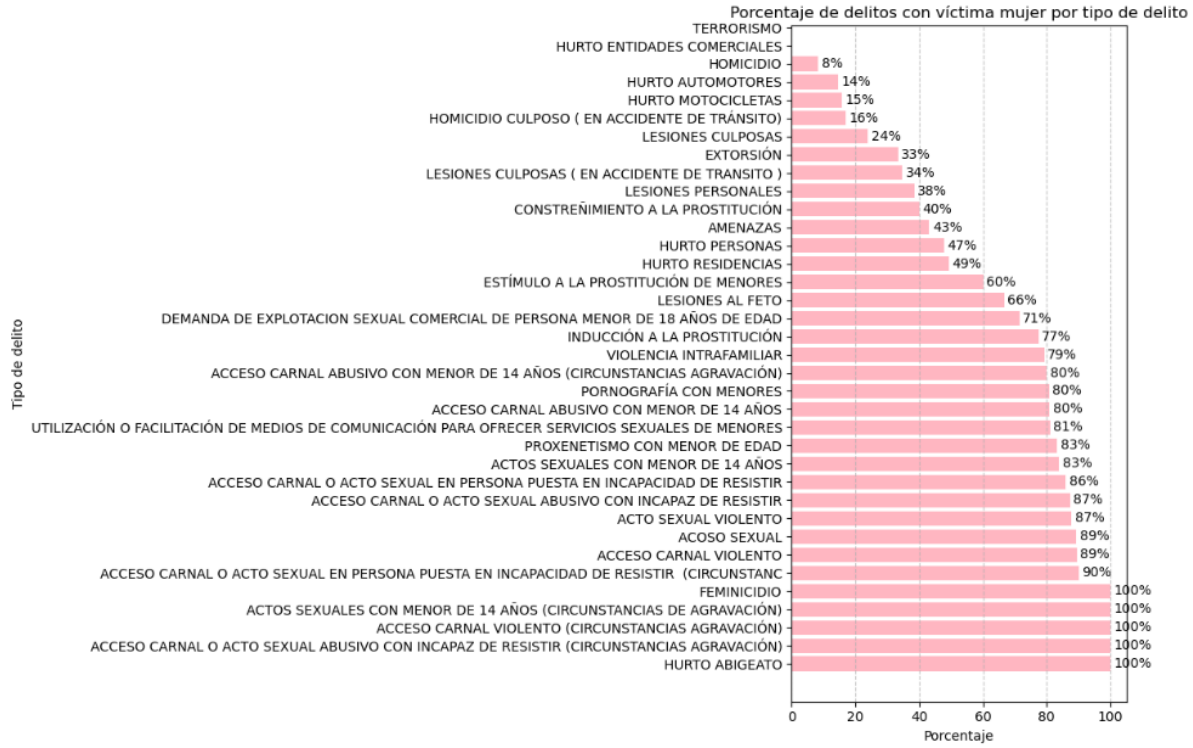

Figura 10

Porcentaje de delitos con mujeres como víctimas

Nota: Elaboración propia

Los actos sexuales violentos, accesos carnales violentos y lesiones personales, se presentan en la mayoría de los casos en contra de las mujeres.

Finalmente, es conveniente saber si la tipología de los delitos varía mucho en función de la comunidad o localidad que se analice. En este caso, se visualiza el top 5 de tipos de delito con más frecuencia, para las localidades con mayor número de delitos por cada 1.000 habitantes en la Figura 11.

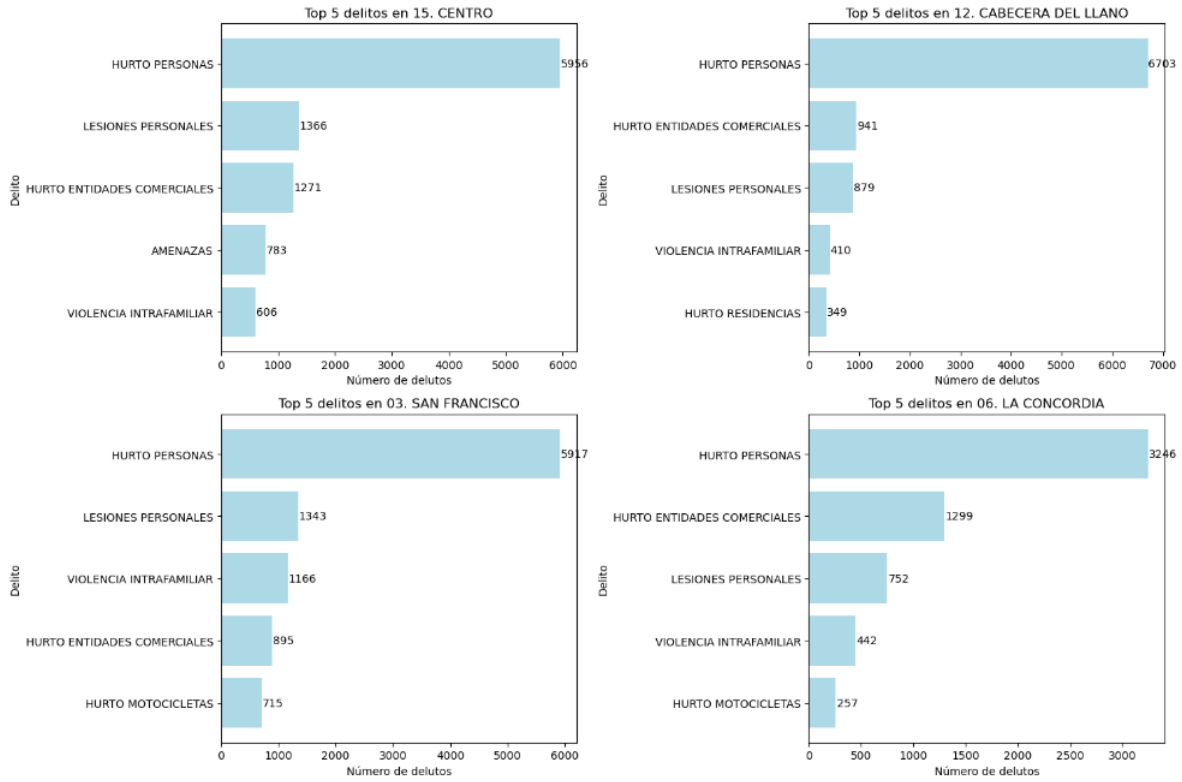

Figura 11

Top 5 delitos en localidades con mayor número de delitos

Nota: Elaboración propia

El análisis descriptivo realizado en esta etapa permitió identificar patrones y correlaciones clave en los datos de criminalidad. Al aplicar el análisis se detectaron concentraciones significativas de delitos en barrios y comunas específicas, y se observaron variaciones en la ocurrencia de los delitos según factores temporales como el horario y el día de la semana. Este análisis inicial reveló que ciertos barrios y horarios, presentaban una mayor incidencia de hurtos y otros delitos violentos. Esta información fue crucial para identificar zonas prioritarias que requieren intervención.

La aplicación de técnicas de georreferenciación utilizando las coordenadas de latitud y longitud permitió mapear los delitos ocurridos en Bucaramanga, creando visualizaciones como mapas de calor. Estos mapas ayudaron a identificar de manera clara las zonas con mayor concentración de criminalidad. Los puntos críticos en barrios específicos, donde el hurto y otros delitos violentos tendían a concentrarse, fueron fácilmente identificables, lo que proporcionó un primer paso para determinar las áreas prioritarias para intervenciones de seguridad.

Se procedió a limpiar y seleccionar las variables clave para el análisis de *Machine Learning*. Se incluyeron variables demográficas (edad y género de las víctimas), características del delito (modalidad, tipo de arma, móvil del agresor) y variables temporales y geográficas. Esta selección y limpieza exhaustiva de los datos aseguraron que solo la información relevante y de calidad se utilizara en las fases posteriores del análisis, optimizando la capacidad predictiva de los modelos a desarrollar.

Para el análisis de agrupamiento (clustering), se emplearon variables como latitud, longitud, rango horario y número de comuna, con el objetivo de identificar patrones espaciales y temporales en la incidencia delictiva. Para el modelo de series de tiempo, únicamente se utilizó la variable 'FECHA_HECHO', correspondiente a la fecha en que ocurrió cada evento.

3. Etapa 3: Desarrollo y Aplicación de Modelos de *Machine Learning*

3.1. Modelo de *clustering*

Este modelo utiliza técnicas de *Machine Learning* y visualización geoespacial para identificar y representar zonas de Bucaramanga con altos niveles de criminalidad. A continuación, se explica el flujo del análisis paso a paso:

Importación de Datos: Se importa un archivo CSV que contiene datos de criminalidad en Bucaramanga, con variables como LATITUD, LONGITUD, tipo de delito, fecha, y hora.

Filtrado de Datos: Se realiza un filtrado de los datos para enfocar el análisis solo en incidentes de "hurto a personas". También se filtran los registros sin ubicación específica o fuera del área geográfica de Bucaramanga.

Transformación de Datos: Normalización Temporal: La variable de tiempo "RANGO_HORARIO_ORDEN" se agrupa en rangos (mañana, tarde, noche, etc.).

Asignación de Comunas Normalizadas: Se crea una nueva columna "NUM_COM_NORM" que clasifica las comunas en categorías específicas usando un diccionario de mapeo.

Preparación para Clustering: Se seleccionan las variables clave para el análisis de agrupación, como LATITUD, LONGITUD, rango horario y comuna. Estas variables se normalizan mediante StandardScaler para mejorar la eficacia del algoritmo DBSCAN en el análisis espacial.

Algoritmo DBSCAN para Agrupación: Se emplea DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*), ideal para detección de patrones en datos geográficos. Los puntos son agrupados en "*clusters*" que representan áreas con patrones de hurto similares; los puntos sin suficiente densidad se consideran ruido.

Visualización Geoespacial: Usando la librería Folium, se crea un mapa de Bucaramanga donde los *clusters* son representados en colores únicos. Los puntos de ruido se muestran en gris.

Cada punto incluye un punto de georreferencia con información sobre el *cluster* y el barrio para facilitar la interpretación geográfica. La Figura 12 representa la primera clasificación del modelo de *clustering* con las características ya mencionadas.

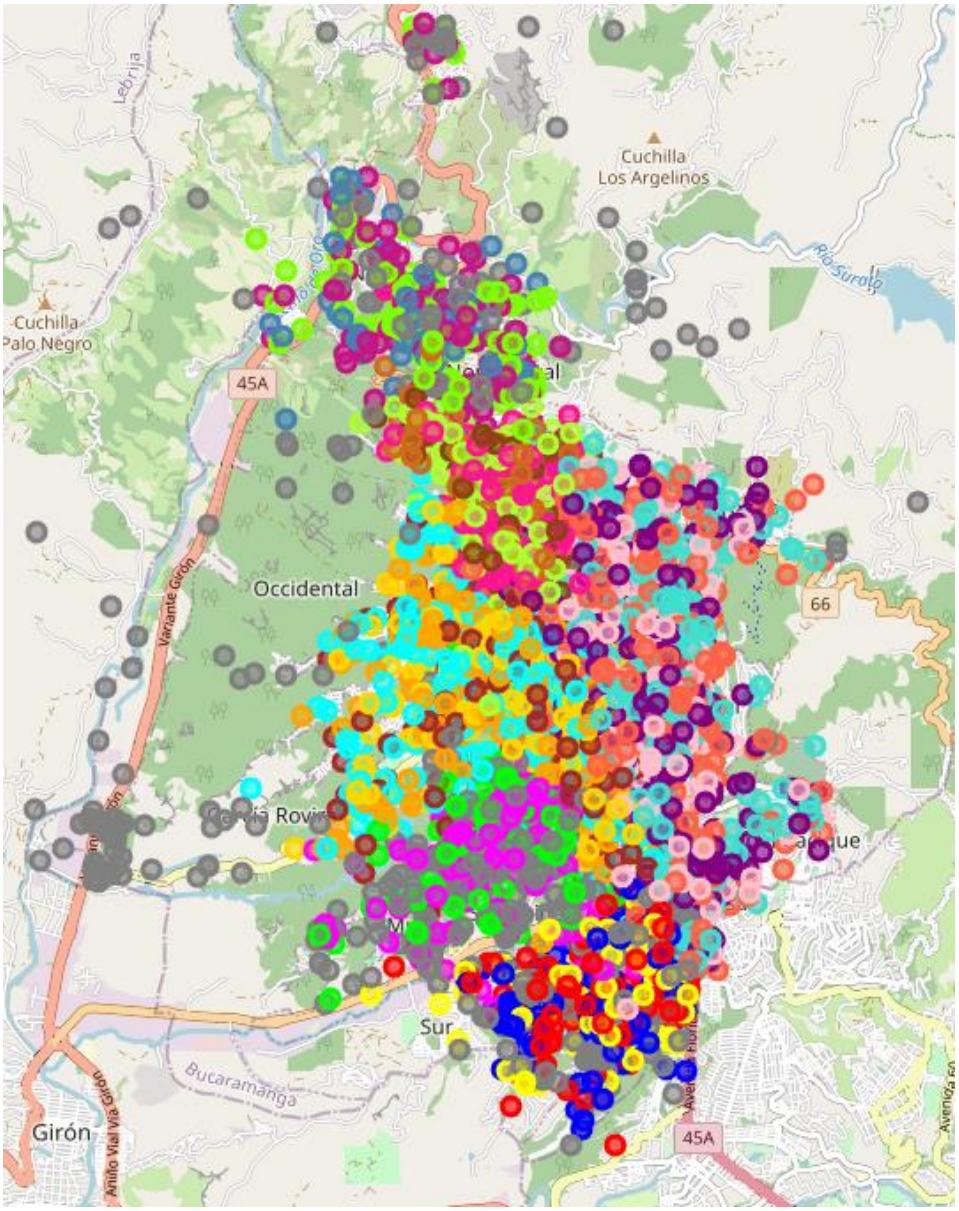

Figura 12

Primera ejecución modelo clustering

Nota: Elaboración propia

Agrupación Secundaria (Clusters de Segundo Nivel):

Se redefine el *clustering* usando una categorización más amplia, asignando ciertos grupos de *clusters* originales a categorías secundarias.

Esto proporciona un nivel de agrupación más general, útil para visualizar zonas de riesgo elevadas de una manera simplificada.

Exportación y Visualización Final:

Se guarda el mapa en un archivo HTML para su revisión interactiva. Esto permite a los analistas y autoridades explorar áreas específicas y obtener información clave sobre los patrones de hurto, facilitando la focalización de intervenciones preventivas en zonas de alto riesgo.

Este enfoque combina un análisis espacial avanzado con técnicas de visualización geográfica para proporcionar hallazgos útiles sobre la distribución espacial y temporal de los hurtos en Bucaramanga, ayudando a optimizar los recursos de seguridad en las áreas de mayor necesidad como podemos ver en la Figura 13.

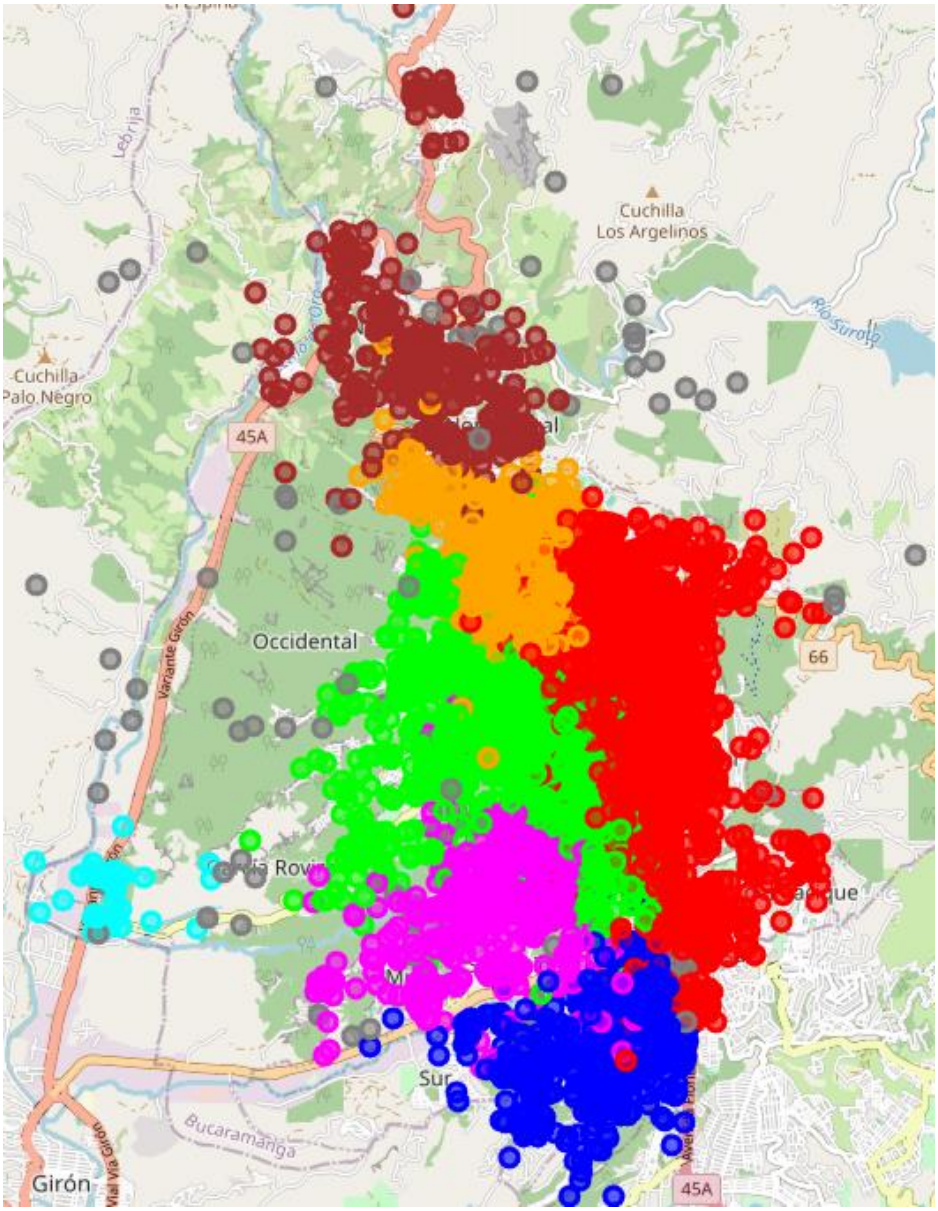

Figura 13

Segunda ejecución modelo clustering agrupada

Nota: Elaboración propia

3.2. Modelo de series de tiempo

Para el modelo de series de tiempo, inicialmente se realizó un análisis exploratorio de los datos desde el punto de vista temporal, a partir del campo 'FECHA_HECHO'. Se graficó en la Figura 14 el número de hurtos a personas, en función de distintos periodos de tiempo como mes, bimestre, trimestre, cuatrimestre y trimestre.

Figura 14

Indicadores hurtos a personas

Nota: Elaboración propia

Debido a que con una frecuencia mensual se tiene un mayor número de datos y, además, se puede percibir patrones que se repiten cada 12 meses, se determina que se debe trabajar con la serie mensualizada, es decir, con un periodo estacional mensual. Dado lo anterior, se tiene la siguiente distribución en la Figura 15 de los delitos de Bucaramanga a lo largo de los meses.

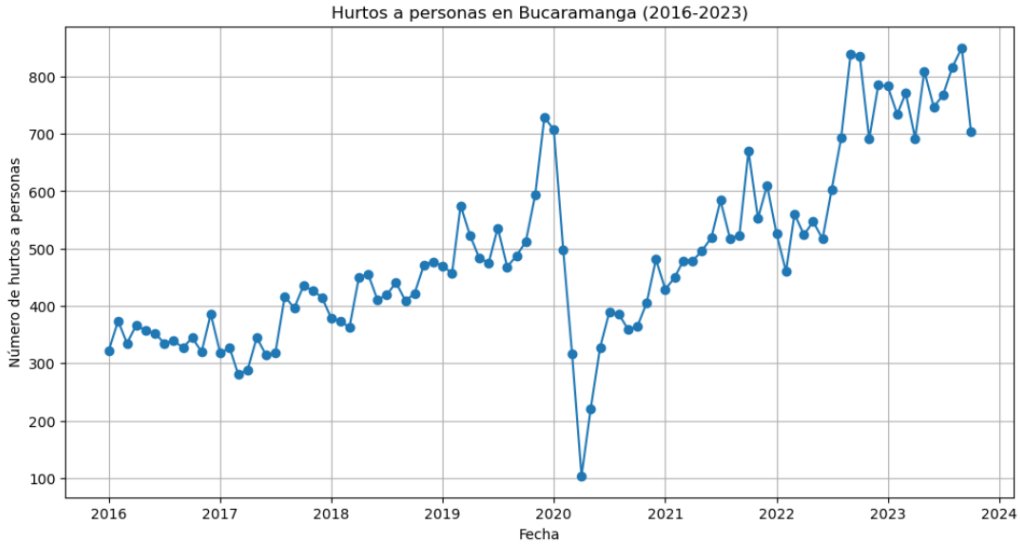

Figura 15

Distribución hurtos por mes

Nota: Elaboración propia

Adicional a lo anterior, se hizo necesario filtrar los registros a fin de excluir el año 2020, ya que debido a la contingencia asociada al SARS-CoV-2, la información es atípica y afectaría el desempeño del modelo. Además de esto, para que el salto desde 2019 hasta 2021 no afecte la continuidad de la serie, se utilizaron índices numéricos para los periodos de tiempo. La serie de tiempo final se presenta en la Figura 16.

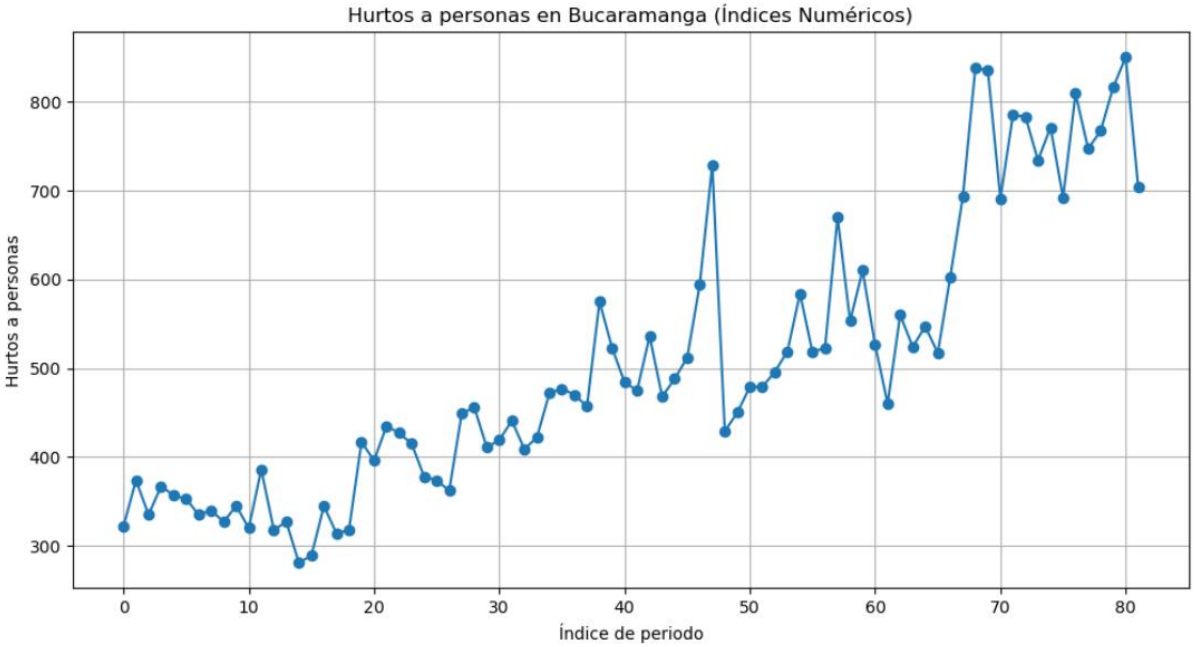

Figura 16

Serie de tiempo final

Nota: Elaboración propia

Análisis de los componentes de la serie de tiempo

Antes de emplear los distintos métodos de modelación de series temporales, realizamos un análisis de las componentes de nuestra serie. Los resultados se pueden observar en la Figura 17.

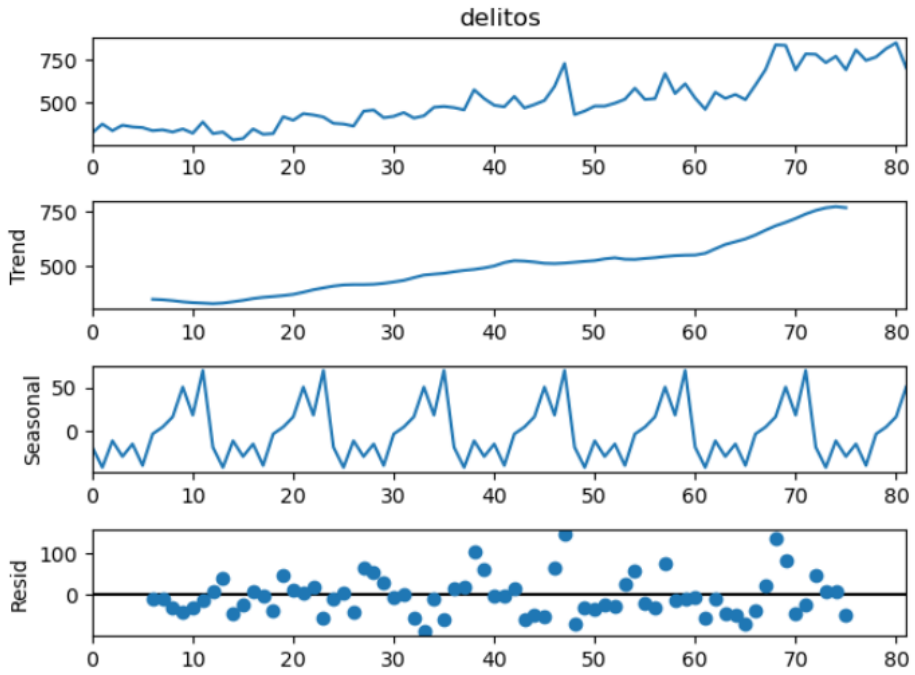

Figura 17

Análisis de componentes

Nota: Elaboración propia

En la primera parte de la Figura 17 tenemos nuestra serie original. En la segunda, tercera y cuarta de la misma Figura 17, se grafican la tendencia, la estacionalidad y los errores residuales, respectivamente.

Como se puede observar, esta es una serie de tiempo no estacionaria con tendencia al alza y con un patrón estacional (es decir, tiene estacionalidad). Aparentemente, parece tener una estacionalidad aditiva, ya que la varianza de los datos parece ser constante a lo largo del tiempo.

En cuanto a la tendencia (o en su defecto, estacionariedad), otra forma de corroborar si la serie es realmente estacionaria o no, es a través de la prueba de Dicky - Fuller. Para esta prueba se validan las siguientes hipótesis:

- H0: La serie temporal no es estacionaria.
- H1: La serie temporal es estacionaria.

Utilizando diferentes niveles de significancia, se obtuvieron los resultados mostrados en la Figura 18.

```
Estadístico ADF: -0.8183611329841576
Valor p: 0.8136690385797227
Valor crítico:
        1%: -3.5159766913976376
        5%: -2.898885703483903
        10%: -2.5866935058484217
La serie no es estacionaria (No podemos rechazar la hipótesis nula).
```

Figura 18

Niveles de significacia

Nota: Elaboración propia

El valor p obtenido, el cual es de 0,81, es superior a cualquiera de los niveles de significancia propuestos, por lo que no se tiene evidencia estadística suficiente para rechazar la hipótesis nula. Esto quiere decir que, con una confianza del 1%, 5% o 10%, la serie temporal es no estacionaria, es decir, tiene tendencia.

Dado que se tiene una serie temporal con tendencia y con estacionalidad, se utilizan dos métodos apropiados para estos contextos:

- Método de Holt-Winters.
- Método de Box-Jenkins.

Método de Holt-Winters.

Este método es perfecto para series de tiempo no estacionarias, con estacionalidad. En este caso, como se tiene una estacionalidad aditiva, se debe aplicar el método de Holt-Winters aditivo.

Se ajusta el modelo en Python indicando tendencia y estacionalidad aditiva, y un periodo estacional de 12.

Se obtuvieron los siguientes parámetros observados en la Figura 19.

```
Parámetros del modelo Holt-Winters:
Coeficiente de suavizado de nivel (alpha): 0.41049641643865764
Coeficiente de suavizado de tendencia (beta): 0.0219724490498169
Coeficiente de suavizado de estacionalidad (gamma): 0.0014503599880213063
Coeficiente de suavizado de tendencia (damping): nan
```

Figura 19

Parámetros modelo Holt-Winters

Nota: Elaboración propia

En el modelo Holt-Winters, el coeficiente de nivel (alpha = 0.4104) indica que el modelo otorga un peso balanceado entre los datos recientes y el historial para actualizar el nivel base. El coeficiente de tendencia (beta = 0.0219) y el coeficiente de estacionalidad (gamma = 0.00145), ambos muy bajos, reflejan que el modelo asume cambios lentos en la tendencia y patrones estacionales, confiando más en los patrones históricos que en fluctuaciones recientes.

En cuanto al desempeño del modelo, se obtuvo la Figura 20 de valores reales vs. Ajustados.

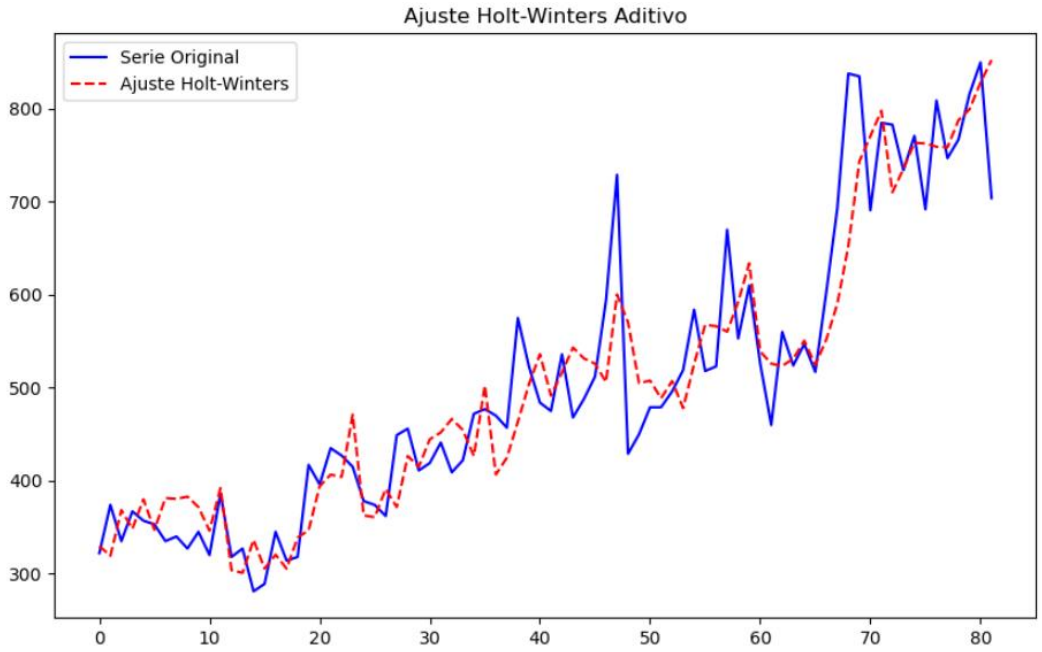

Figura 20

Valores reales vs ajustados Holt Winters

Nota: Elaboración propia

Como se observa en la Figura 20, la serie suavizada (en rojo) se asemeja mucho a la serie real, lo cual se traduce en las siguientes métricas de error:

- MSE: 3.045
- MAPE: 8,25%

Por otra parte, a pesar de que en los métodos de suavizamiento como Holt-Winters no es necesario validar los supuestos de independencia, media cero, homocedasticidad y normalidad de los residuos, se decidió realizar estas pruebas a fin de tener mayor noción del poder predictivo del modelo. El resultado se presenta en la Figura 21.

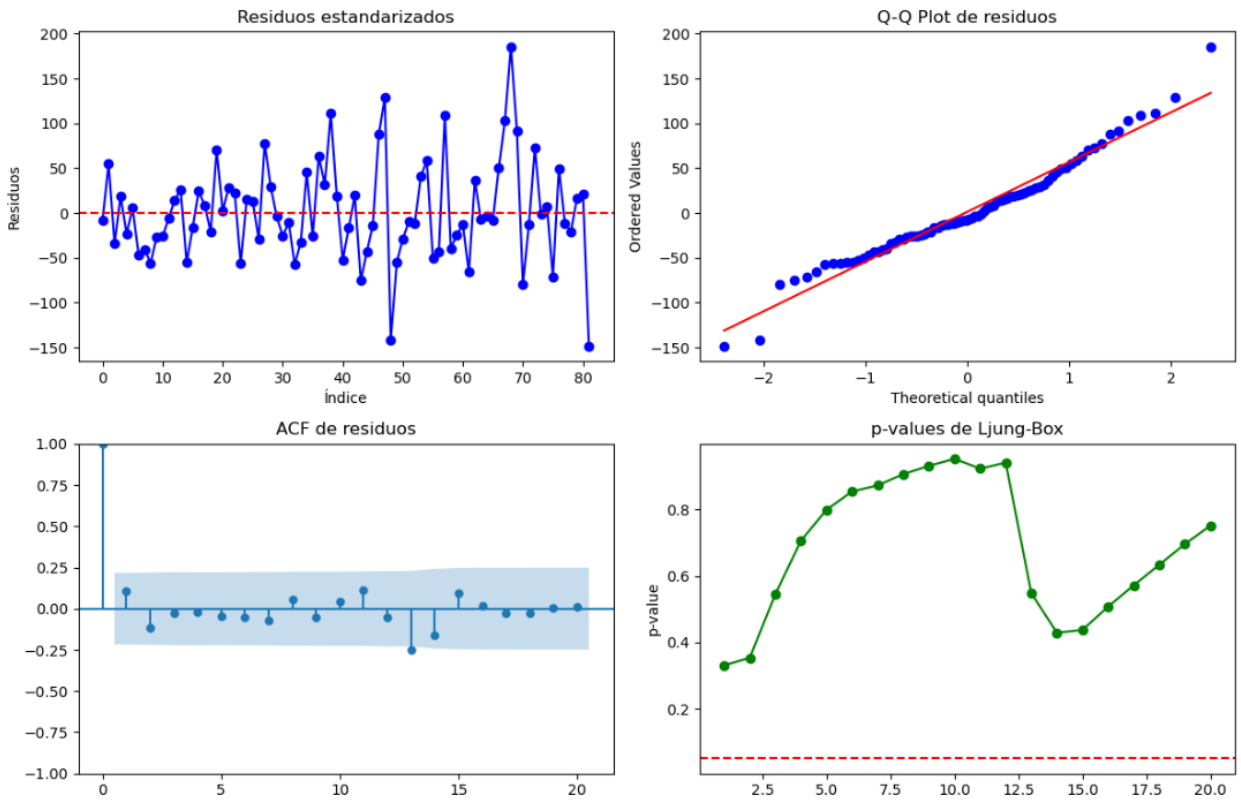

Figura 21

Validación de supuestos Holt Winters

Nota: Elaboración propia

- Con la primera parte de la Figura 21 de residuos estandarizados, podemos corroborar que los residuos tienen media cero.
- El gráfico Q-Q Plot en la Figura 21 nos muestra que los residuos siguen una distribución normal, ya que estos se distribuyen a lo largo de la línea recta de color rojo. Realizando la prueba de Shapiro-Wilk, se obtiene un p-value de 0,0505, lo cual también sugiere la existencia de normalidad con una confianza del 95%.
- La gráfica ACF de los residuos en la Figura 21 nos indica que hay independencia, ya que ningún rezago está por fuera de las bandas de significancia. Esto coincide con la gráfica de p-values de Ljung-Box, en la que todos los puntos están por encima de la línea roja.

Finalmente, se procede a predecir el número de delitos para los siguientes 12 periodos. Dado que los residuos siguen una distribución normal, también se determinan los intervalos de confianza del 95% representados en la Figura 22.

Predicciones para los próximos 12 meses con intervalos de confianza:

	Predicción	Límite Inferior (95%)	Límite Superior (95%)
82	779.361024	671.233507	887.488541
83	838.706784	730.579267	946.834301
84	755.627663	647.500146	863.755180
85	749.400943	641.273426	857.528460
86	776.750152	668.622635	884.877669
87	771.307143	663.179626	879.434660
88	796.142293	688.014776	904.269810
89	773.764883	665.637366	881.892400
90	806.485642	698.358125	914.613159
91	825.864743	717.737226	933.992260
92	846.655382	738.527865	954.782899
93	860.728811	752.601294	968.856328

Figura 22

Intervalos de confianza

Nota: Elaboración propia

Método de Box-Jenkins:

A pesar de haber obtenido un modelo de Holt-Winters bastante bueno, se decidió emplear el método de Box-Jenkins, el cual consiste en ajustar modelos AR (Autorregresivos), MA (de media móvil), o una combinación de estos (ARMA, Modelo autorregresivo de media móvil) y ARIMA (Modelo autorregresivo integrado de media móvil).

Inicialmente, se construyeron las gráficas ACF (función de autocorrelación) y la PACF (función de autocorrelación parcial), obteniendo los resultados que se observan en la Figura 23.

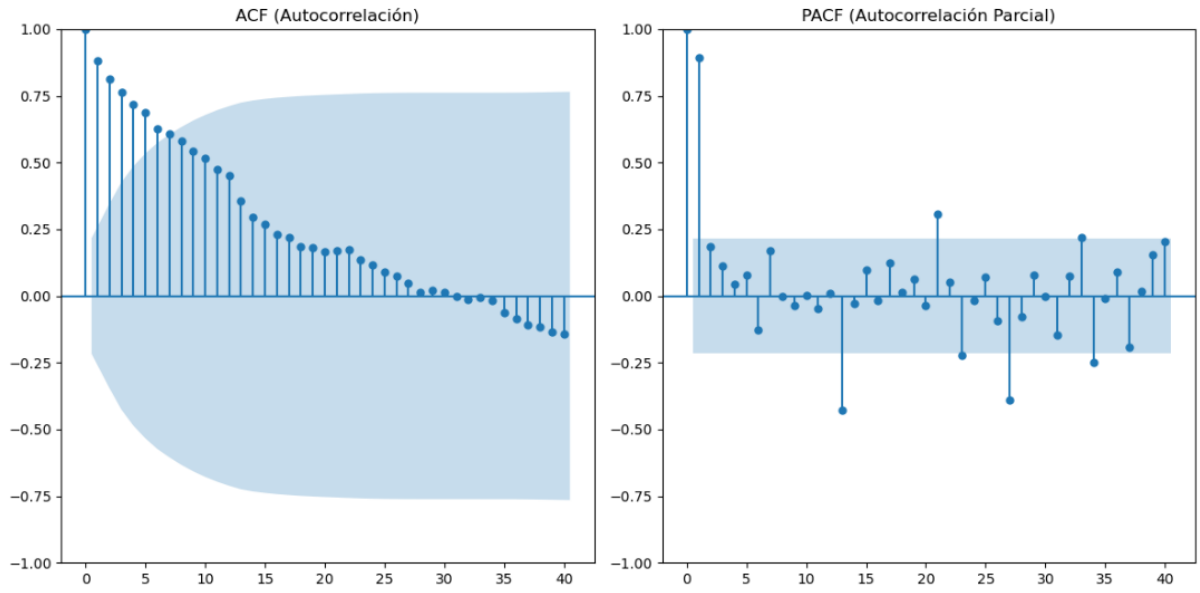

Figura 23

Método Box Jenkins. ACF y PACF

Nota: Elaboración propia

Dado que la serie es no estacionaria, no se puede aplicar Box-Jenkins directamente, lo cual coincide con el hecho de que ambas gráficas, ACF y PACF, no permiten concluir nada sobre el modelo a ajustar.

Por lo anterior, se procedió a diferenciar la serie teniendo en cuenta una tendencia lineal, es decir, utilizando un parámetro d = 1. Esto quiere decir que, en particular, se tendrá un modelo de tipo ARIMA (Modelo autorregresivo integrado de media móvil) con parámetros p = desconocido, d = 1, y q = desconocido. En un modelo ARIMA (p, d, q), p representa el número de términos autorregresivos (dependencia con valores pasados), d es el orden de diferenciación necesario para hacer estacionaria la serie (eliminando tendencias o estacionalidad), y q indica el número de términos de media móvil (dependencia con errores pasados).

Lo siguiente que se realizó fue determinar el valor de los parámetros p y q que minimicen una métrica de error como el MSE. Tal y como se observa en los parámetros del modelo de Holt-Winters, se espera que los valores de p y q sean relativamente altos, teniendo en cuenta que nuestra serie de tiempo tiene una visión de largo plazo, es decir, está muy influenciada por valores pasados.

Luego de diferenciar la serie en d = 1, se encontraron los siguientes valores óptimos de p y q presentados en la Figura 24.

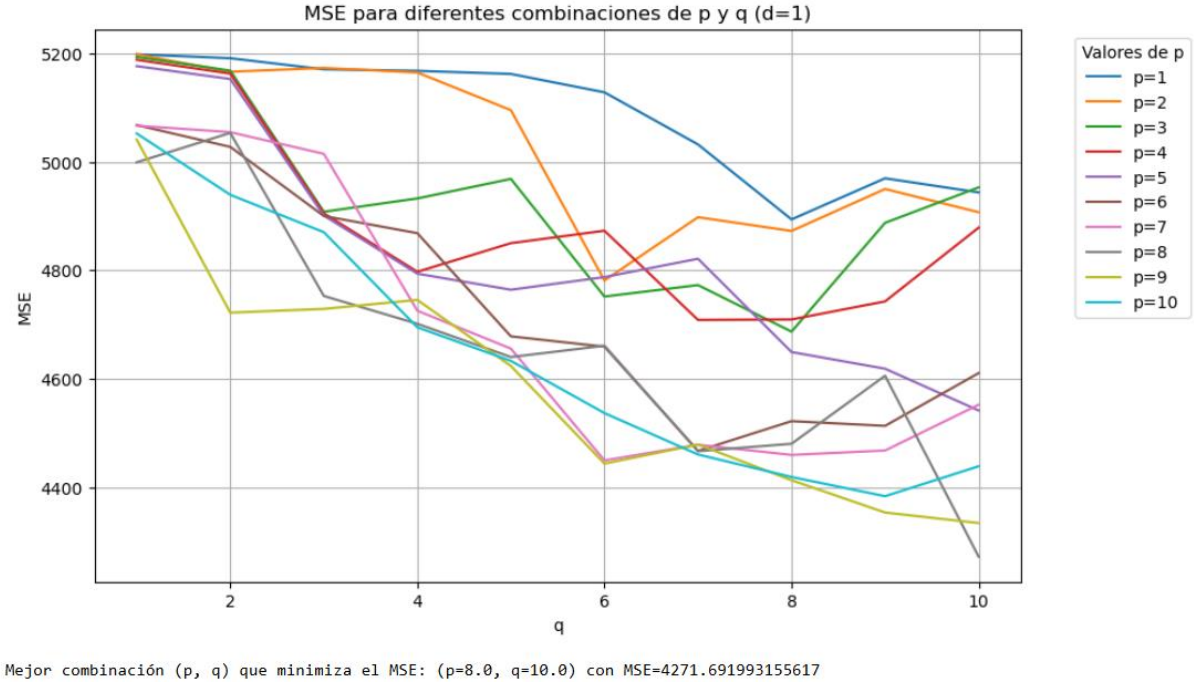

Figura 24

MSE valores óptimos p y q

Nota: Elaboración propia

Con un p = 8 y un q = 10, se tiene un modelo final ARIMA (8,1,10).

En cuanto al error de pronóstico, se obtuvo lo siguiente:

- MSE: 4.271
- MAPE: 8,93%

Ambas métricas son superiores a los resultados obtenidos en Holt-Winters, por lo que se sugiere utilizar Holt-Winters en lugar de Box-Jenkins.

La gráfica de los valores ajustados vs. reales se muestra a en la Figura 25.

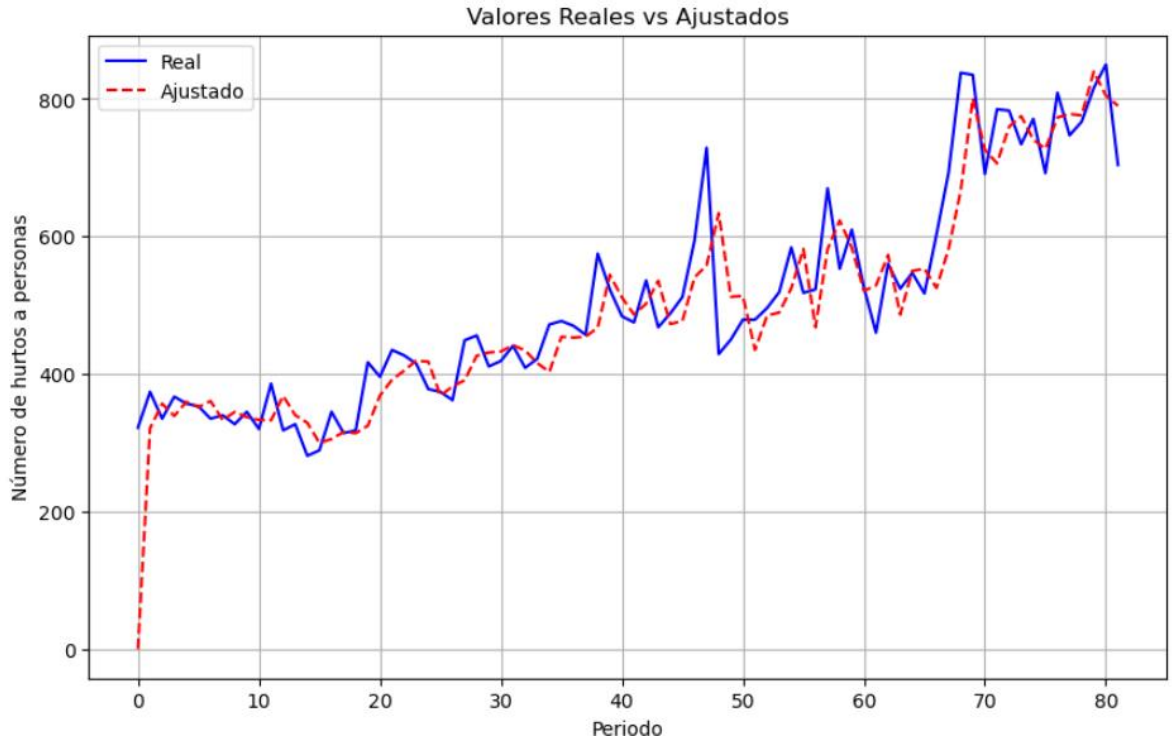

Figura 25

Valores reales vs ajustados Box Jenkins

Nota: Elaboración propia

Finalmente, para el caso de Box-Jenkins sí es estrictamente necesario validar los supuestos. Si bien se prioriza el supuesto de independencia, es necesario que todos los supuestos se cumplan. La Figura 26 presenta la respectiva validación de supuestos.

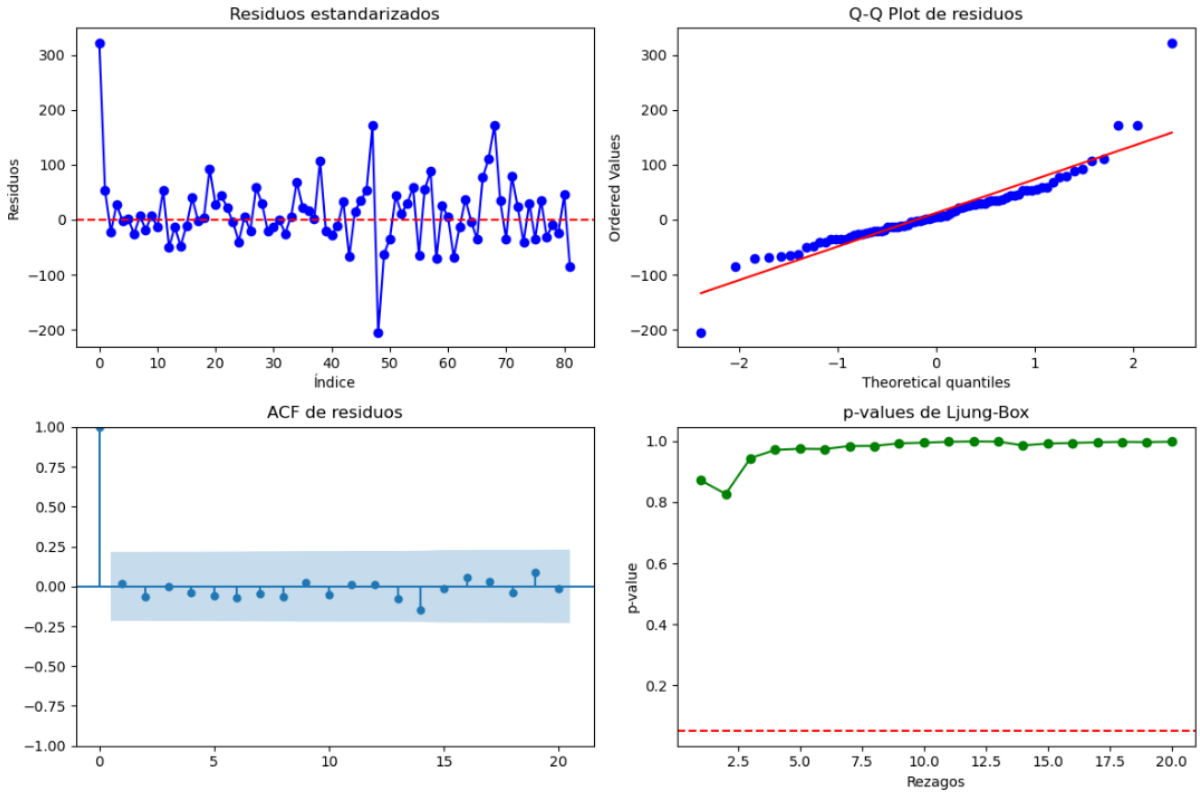

Figura 26

Validación de supuestos Box Jenkins

Nota: Elaboración propia

En este caso, no todos los supuestos se cumplen. En particular, para el supuesto de normalidad y media cero, se observan datos atípicos que perturban la distribución de los residuos. EN el caso de normalidad, luego de realizar la prueba de Shapiro-Wilk se obtuvo un valor p de 1,85x10-6, lo cual indica que se encuentran muy alejados de la distribución normal de los residuos.

En resumen, debido al resultado obtenido para los errores MSE y MAPE, y teniendo en cuenta el cumplimiento de los supuestos de independencia, normalidad, homocedasticidad y media cero, se prefiere utilizar un modelo de Holt-Winters por encima de un modelo ARIMA(8,1,10).

4. **Etapa 4: Desarrollo de visualizador de zonas en alto riesgo.**

En esta etapa, se desarrolló un visualizador interactivo en Power BI como se muestra en la Figura 27. Este proporciona una caracterización detallada de los delitos en Bucaramanga. A continuación, se describen las principales pestañas y sus respectivas funcionalidades:

1. **Caracterización de los Delitos en Bucaramanga**: Esta pestaña ofrece una visión general de los tipos de delitos registrados en la ciudad, destacando las áreas con mayor incidencia y las variaciones a lo largo del tiempo. Los datos son representados a través de gráficos interactivos que permiten identificar rápidamente patrones y tendencias.
2. **Análisis por Localización**: En esta sección, se presenta un análisis geográfico de los hurtos, desglosado por comuna. Se muestran mapas y gráficos que indican la cantidad

de hurtos ocurridos en cada comuna, lo que permite visualizar las zonas más afectadas y facilitar la toma de decisiones en términos de seguridad y prevención.

3. **Caracterización de las Víctimas**: Esta pestaña proporciona información detallada sobre las víctimas de los hurtos, segmentada por género, edad y tipo de movilización (a pie, en moto, en carro, etc.). Los datos permiten identificar patrones relacionados con las características de las víctimas, lo que puede ser útil para diseñar estrategias de prevención específicas según los grupos más afectados.
4. **Predicción para el Próximo Año**: En esta hoja se presenta una proyección de la incidencia de delitos en el siguiente año, basada en los patrones históricos y los análisis realizados. Esta predicción es clave para anticipar las zonas de alto riesgo y planificar medidas de seguridad proactivas.
5. **Análisis de Clusters por Rango Horario y Día de la Semana**: Finalmente, en esta sección se muestran los resultados del análisis de clusters, donde los hurtos se agrupan según el rango horario y el día de la semana. Los clusters ayudan a identificar las franjas horarias y los días con mayor concentración de delitos, lo que permite diseñar estrategias de intervención más efectivas en momentos de mayor vulnerabilidad.

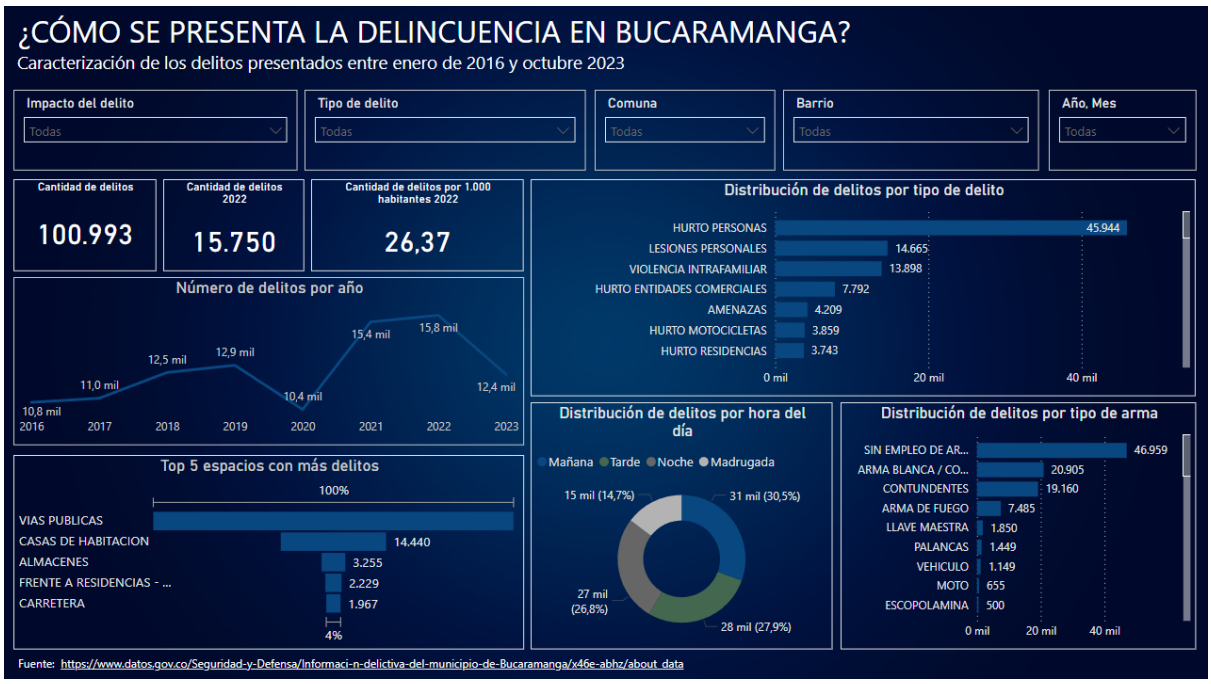

Figura 27

Tablero de visualización

Nota: Elaboración propia

Conclusiones

La primera etapa del proyecto estableció una base sólida al preparar y explorar los datos, identificando patrones iniciales, visualizando áreas de alta criminalidad y seleccionando variables clave para los modelos predictivos. Estos resultados aseguran datos listos para análisis avanzados, como clustering y predicción de criminalidad futura.

Se concluye que el hurto a personas ocupa consistentemente el primer lugar. El hurto a motocicletas prevalece en las comunas San Francisco y La Concordia, mientras que el hurto a entidades comerciales, las lesiones personales y la violencia intrafamiliar se mantienen entre las cinco primeras posiciones del ranking.

El modelo de clustering generó agrupaciones coherentes que reflejan las características clave de los datos originales, como ubicación geográfica y frecuencia temporal. Los patrones observados dentro de cada cluster muestran consistencia y correspondencia con las zonas de alta densidad identificadas mediante mapas de calor y reducción de dimensionalidad. Esto evidencia que el modelo captura eficazmente los patrones subyacentes, siendo útil para identificar áreas de riesgo y analizar tendencias significativas.

Se demostró que los datos se pueden modelar a través de una serie de tiempo cuya tendencia, estacionalidad, periodo estacional y ciclo son componentes claramente definidas, identificables y caracterizables. Además, dado que el ajuste del modelo es bastante aceptable y los residuos cumplen con los supuestos de independencia, normalidad, homocedasticidad y media cero, podemos realizar pronósticos confiables del número de hurtos a personas en la ciudad de Bucaramanga.

El visualizador en Power BI proporciona una herramienta integral para analizar la criminalidad en Bucaramanga desde diferentes perspectivas, permitiendo a las autoridades y a los equipos de seguridad tomar decisiones informadas basadas en datos actualizados y proyecciones futuras. La interactividad de la plataforma facilita la exploración y el análisis dinámico de los datos, mejorando la capacidad para identificar patrones y tomar acciones preventivas.

Referencias

- Alcaldía Bucaramanga (2023). Informe de Calidad de Vida AMB 2023. Bucaramanga.
- Alcaldía de Bucaramanga. (2022). Plan de desarrollo municipal. Alcaldía de Bucaramanga. https://www.bucaramanga.gov.co/wp-content/uploads/2022/02/Primer-Documento-PDM-Final.pdf
- Alvarado Zabala, J., Martillo Alchundia, I., & Guzman Seraquive, G. (2022). Revisión de literatura sobre las técnicas de Machine Learning en la detección de fraudes bancarios. Sapienza.
- Ardila, M. (2023). Crimen y factores económicos en Medellín: un estudio de predicción con Machine Learning. Repositorio Universidad del Rosario. https://repository.urosario.edu.co/server/api/core/bitstreams/3e4db522-c4d6-40cb-9eab-a52dd777874b/content
- BBC. (8 de agosto de 2012). Tecnología para predecir dónde ocurrirá el próximo robo. BBC New Mundo.
- BID. (07 de marzo de 2024). Hoja informativa. Obtenido de Seguridad Ciudadana en América Latina y el Caribe. https://www.iadb.org/es/noticias/seguridad-ciudadana-en-america-latina-y-el-caribe
- BID. (2017). Los costos del crimen y de la violencia: nueva evidencia y hallazgos en América Latina y el Caribe. BID.

- Buil, D. (2016). ¿Qué es la criminología? Una aproximación a su ontología, función y desarrollo. Derecho y Cambio Social, 13(44). https://dialnet.unirioja.es/servlet/articulo?codigo=5456246
- Cid, J. y Larrauri, E. (2001). Teorías criminológicas. Explicación y prevención de la delincuencia. Barcelona: Bosch.
- Corporación Excelencia en la Justicia. (21 de septiembre de 2023). Se agudiza la criminalidad en Colombia. Obtenido de CEJ. https://cej.org.co/destacados-home-page/se-agudiza-la-criminalidad-en-colombia-cada-dia-mas-de-mil-personas-son-victimas-de-hurtos-y-o-extorsion/
- Datos Abiertos Colombia. (2023). Información delictiva del municipio de Bucaramanga [Dataset]. Datos Abiertos Colombia. https://www.datos.gov.co/Seguridad-y-Defensa/Informaci-n-delictiva-del-municipio-de-Bucaramanga/x46e-abhz/about_data
- Del Olmo, R. (1999). América Latina y su criminología. Siglo veintiuno editoriales. https://books.google.com.co/books?id=0fFz0FZQXE8C&lpg=PA9&ots=yrCBkGDxCn&dq=criminolog%C3%ADa&lr&hl=es&pg=PA3#v=onepage&q&f=false
- Felson, M. y Clarke, R.V. (1998). Opportunity makes the thief: Practical theory for crime prevention. Police Research Series, Paper 98, 1-43.
- Fernandes, R., & Antonelli, M. (2018). Machine learning: A practical approach on the statistical learning theory. Springer. https://link-springer-com.ezproxy.uniandes.edu.co/book/10.1007/978-3-319-94989-5
- FMI, F. M. (18 de diciembre de 2023). Latin America Can Boost Economic Growth by Reducing Crime. Obtenido de Latin America Can Boost Economic Growth by Reducing Crime. https://www.imf.org/es/Blogs/Articles/2023/12/18/latin-america-can-boost-economic-growth-by-reducing-crime
- G. A. Vergel-Clavijo y A. M. Guerrero-Bayona, “Ciudad inteligente: mejoramiento de la seguridad ciudadana a través del uso de nuevas tecnologías”, Rev. Ingenio, vol. 20, n°1, pp. 32-39 (2023). https://doi.org/10.22463/2011642X.3510
- García, J., Ordóñez, K., Ruiz, M. (2022). Sistema biométrico de reconocimiento facial mediante redes neuronales artificales para aportar a la seguridad ciudadana. Universidad Nacional Autónoma de Nicaragua, Leon Facultad de ciencias y tecnología. http://riul.unanleon.edu.ni:8080/jspui/bitstream/123456789/9726/1/253079.pdf
- Gelvez, J., Nieto, M. & Rocha, C. (2022). Prediciendo el crimen en ciudades intermedias: un modelo de “machine learning” en Bucaramanga, Colombia. Revista Latinoamericana de Estudios de Seguridad, No. 34. http://scielo.senescyt.gob.ec/pdf/urvio/n34/1390-4299-urvio-34-00082.pdf
- Gelvez, J., Nieto, M., & Rocha, C. (2022). Prediciendo el crimen en ciudades intermedias: un modelo de “machine learning” en Bucaramanga, Colombia. URVIO Revista Latinoamericana de Estudios de Seguridad, (34), 82-98. https://doi.org/10.17141/urvio.34.2022.5395
- GIJN Staff. (2023, September 27). 2023 Global Organized Crime Index – Global Investigative Journalism Network. Global Investigative Journalism Network. https://gijn.org/
- Global Initiative Against Transnational Organized Crime. (2023). Global Organized Crime Index 2023. https://ocindex.net/report/2023/0-3-contents.html

- Gonzales Rodriguez, F., & Barbarán Mozo, H. (2021). La seguridad ciudadana como política gubernamental en América Latina en el últimoquinquenio. Ciencia Latina Revista multidisciplinar.
- González, E. (2019). El delito de hurto y su evolución histórica. Revista Caribeña de Ciencias Sociales. https://www.eumed.net/rev/caribe/2019/03/hurto-evolucion-historica.html
- Herhausen, D., Bernritter, S., Ngai, E., Kumar, A., & Delen, D. (2024). Machine learning in marketing: Recent progress and future research directions. ScienceDirect.
- Hirschi, T. (1969). Causes of delinquency. Berkeley: University of California Press.
- IMF. (20 de diciembre de 2023). América Latina: Reducir la delincuencia para estimular el crecimiento económico. Obtenido de IMF Blog. https://www.imf.org/es/Blogs/Articles/2023/12/18/latin-america-can-boost-economic-growth-by-reducing-crime
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning with applications in R. Springer. https://link-springer-com.ezproxy.uniandes.edu.co/book/10.1007/978-1-4614-7138-7
- Karabo, J., Cagatay, C., & Gorkem, K. (2023). Machine learning in crime prediction. Journal of Ambient Intelligence and Humanized Computing. https://www.researchgate.net/publication/368164162_Machine_learning_in_crime_prediction
- Kubat, M. (2017). An Introduction to Machine Learning. Springer Cham. https://doi-org.ezproxy.uniandes.edu.co/10.1007/978-3-319-63913-0
- Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. Springer. https://link-springer-com.ezproxy.uniandes.edu.co/book/10.1007/978-1-4614-6849-3
- LaRepublica. (25 de octubre de 2023). Colombia está en el segundo lugar dentro de los países con mayores índices de criminalidad. LaRepublica. https://www.larepublica.co/globoeconomia/los-paises-con-mayores-indices-de-criminalidad-3735633#:~:text=Colombia%20sigue%20ocupando%20primeros%20lugares,con%20mayor%20criminalidad%20del%20mundo
- López, G., & Manosalvas, P. (2023). Datos y criminalidad: Machine learning aplicado en modelos predictivos en seguridad. AL DATO OBSERVATORIO. https://al-dato.datalat.org/wp-content/uploads/2024/01/2.-Geovanna-Lopez-Pedro-Manosalva-Criminalidad-Al-Dato.pdf
- Matsueda, R. (2006). Differential social organization, collective action, and crime. Crime, Law and Social Change, 46(1), 3-33. https://doi.org/10.1007/s10611-006-9045-1
- Muñoz, D. (2021). Evaluación de modelos de Machine Learning para la predicción de crímenes en la ciudad de Medellín. Repositorio Universidad Nacional. https://repositorio.unal.edu.co/handle/unal/80976
- Musheer Aziz, R., Yaqoob, A., & Kumar verma, N. (2023). Applications and Techniques of Machine Learning in Cancer. National Library of Medicine.
- Nedelec, J. & Di Rienzo, F. (2023). Predicting Moffitt's Developmental Taxonomy of Antisocial Behavior Using Life History Theory: A Partial Test of the Evolutionary Taxonomy. Evolutionary Psychology, 21(4). https://www-scopus-com.bdbiblioteca.universidadean.edu.co/record/display.uri?eid=2-s2.0-85177067595&origin=resultslist&sort=plf-

f&src=s&sid=881098da453a4af1db3fad0f2217aaef&sot=b&sdt=b&s=TITLE-ABS-KEY%28criminality+prediction%29&sl=31&sessionSearchId=881098da453a4af1db3fad0f2217aaef&relpos=2

- Nielsen, A. (2019). P*ractical Time Series Analysis: Prediction with Statistics and Machine Learning*. O'Reilly Media, Inc. Recuperado de: https://learning-oreilly-com.ezproxy.uniandes.edu.co/library/view/practical-time-series/9781492041641/
- Ordóñez, H., Cobos, C. & Bucheli V. (2020). Modelo de machine learning para la predicción de las tendencias de hurto en Colombia. Revista Ibérica de Sistemas e Tecnologias de Informação. https://media.proquest.com/media/hms/PFT/1/tDYJG?_s=Clf1GMYyFx9XGL0Z9HAW%2BzhFLhA%3D
- Ordóñez, H., Cobos, C., & Bucheli, V. (2020). Modelo de machine learning para la predicción de las tendencias de hurto en Colombia. Revista Ibérica de Sistemas e Tecnologias de Informação, (N.º E29), 494-506. https://www.proquest.com/openview/fb8bfe36673b48be2d035ee8a035c307/1?pq-origsite=gscholar&cbl=1006393
- Real Academia de la Lengua Española (s.f.). Diccionario de la lengua española: Definición delincuencia. Real Academia de la Lengua Española - RAE. https://dle.rae.es/delincuencia
- Real Academia de la Española (2001). Diccionario de la lengua española: Definición delincuencia. Real Academia de la Lengua Española - RAE. https://dle.rae.es/delincuencia
- Salazar Isairias, S. (2024). Predicción geoespacial de crímenes en Bogotá: un enfoque basado en Machine learning para mejorar la seguridad ciudadana. Universidad de los Andes. https://hdl.handle.net/1992/73858
- Serrano-Maíllo, A. (2004). Introducción a la Criminología. 2ª edición. Madrid: Dykinson.
- Sutherland, E., Cressey, D. & Luckenbill, D. (1992). Principles of Criminology. Eleventh edition. Lanham: General Hall.
- Vanguardia. (2024, abril 11). Tecnología de punta, una aliada clave para reforzar la seguridad ciudadana. Vanguardia. https://www.vanguardia.com/mundo/tecnologia/2024/04/11/tecnologia-de-punta-una-aliada-clave-para-reforzar-la-seguridad-ciudadana/
- Vanguardia. (2024, mayo 2). ¿Qué tan seguro se siente en Bucaramanga? Expertos hablan de realidad y de percepción. Vanguardia. https://www.vanguardia.com/area-metropolitana/bucaramanga/2024/05/02/que-tan-seguro-se-siente-en-bucaramanga-expertos-hablan-de-realidad-y-de-percepcion/
- Watts, D. (2023). Digital Mental Health: A Practitioner's Guide (p. 223-235). Springer International Publishing. https://www-scopus-com.bdbiblioteca.universidadean.edu.co/record/display.uri?eid=2-s2.0-85173854121&origin=resultslist&sort=plf-f&src=s&sid=881098da453a4af1db3fad0f2217aaef&sot=b&sdt=b&s=TITLE-ABS-KEY%28criminality+prediction%29&sl=31&sessionSearchId=881098da453a4af1db3fad0f2217aaef&relpos=5
- Wilson, J.Q. y Kelling, G. (1982). Broken Windows: The police and neighbourhood safety. Atlantic Monthly, 249(3), 29-38.
- Zhou, Z.-H. (2021). Machine learning. Springer. https://link-springer-com.ezproxy.uniandes.edu.co/book/10.1007/978-981-15-1967-3

Anexos

Anexo A.

Modelos de *Machine learning*

En este anexo se presenta el repositorio de GitHub con los modelos de *machine learning* utilizados en el proyecto. Los modelos desarrollados están disponibles para su revisión y uso en el siguiente enlace:

Modelos de Machine Learning en GitHub
(https://github.com/victoranv/seminario_investigacion/tree/main).

Anexo B.

Visualizador de zonas en alto riesgo

Este anexo presenta la publicación del visualizador desarrollado en el proyecto, el cual permite identificar las zonas de mayor riesgo en Bucaramanga para intervenciones preventivas. La visualización es una herramienta interactiva que facilita el análisis y la toma de decisiones para mejorar la seguridad en las áreas más vulnerables de la ciudad.

Tablero power BI
(https://app.powerbi.com/view?r=eyJrIjoiNGI4ZTI0NjktNGRlZC00N2FkLTg3M2UtZDEyNWYzNDEwNWU3IiwidCI6Ijg3NDg5NWUwLWQwYjYtNDkzZC05N2YxLWE1MzMxODc0M2I3ZSIsImMiOjR9)