MODELO DE CLASIFICACIÓN Y PREDICCIÓN DE DELITOS DE ALTO IMPACTO EN LA CIUDAD DE BOGOTÁ

YVONNE MARITZA SUÁREZ RUIZ
ANDRES FELIPE BEDOYA CRUZ

FACULTAD DE INGENIERÍA

UNIVERSIDAD DE LA SABANA
MAESTRÍA EN ANALÍTICA APLICADA

2023

PÁGINA DE ACEPTACIÓN

Juan Pablo Mojica Macias
Jurado 1

Marco Tulio Teran De La Hoz
Jurado 2

Juan Manuel Aranda Lopez King
Tutor

Chía, 31 de Julio del 2023

DEDICATORIA

A nuestros padres y familias, por ser nuestra mayor motivación y apoyo incondicional durante toda nuestra vida. Sus consejos, valores y principios nos han permitido crecer como personas y profesionales.

AGRADECIMIENTOS

Queremos expresar nuestro más sincero agradecimiento a todas aquellas personas que hicieron posible la realización de este trabajo de grado.

En primer lugar, a nuestros padres, Eliana Cruz y Rubiel Bedoya; y a Jorge Suárez Burbano y Amparo Ruiz Doncel, por brindarnos su respaldo constante, sus sabios consejos y motivación durante toda nuestra formación académica y profesional.

Un agradecimiento especial merece el esposo de Yvonne, Juan Francisco Pedraza Vélez, por su paciencia y dedicación al cuidado de su pequeño hijo recién nacido, mientras ella culminaba este proyecto. Yvonne también agradece profundamente a sus padres, por el invaluable apoyo y compañía durante el proceso de gestación y crianza de su hijo, lo cual le permitió avanzar en el desarrollo de la tesis.

Asimismo, expresamos nuestro agradecimiento a Juan Manuel Aranda Lopez King, director de esta tesis, por compartir sus conocimientos y orientarnos acertadamente en la investigación.

Agradecemos también a Felix Vivian Mohr por su acompañamiento técnico en el desarrollo de los modelos durante la tesis.

Finalmente, a la Universidad La Sabana y a todos los profesores del programa de Maestría en Analítica Aplicada, por brindarnos las herramientas y formación para crecer profesionalmente.

TABLA DE CONTENIDO

ILUSTRACIONES

TABLAS

ANEXOS

ACRÓNIMOS

DANE: Departamento Administrativo Nacional de Estadística
ETL: Extract, Transform, Load
OAIEE: Oficina de Análisis de Información y Estudios Estratégicos
SDP: Secretaría Distrital de Planeación
SEMMA: Sample, Explore, Modify, Model, y Assess (Evaluación)
SMOTE: Synthetic Minority Oversampling Technique
UPZ: Unidad de Planeamiento Zonal

PALABRAS CLAVE

Análisis descriptivo
Análisis predictivo
Clasificación
Delitos de alto impacto
Machine Learning
Naive AutoML
Predicción
Preprocesamiento de datos

RESUMEN GRÁFICO

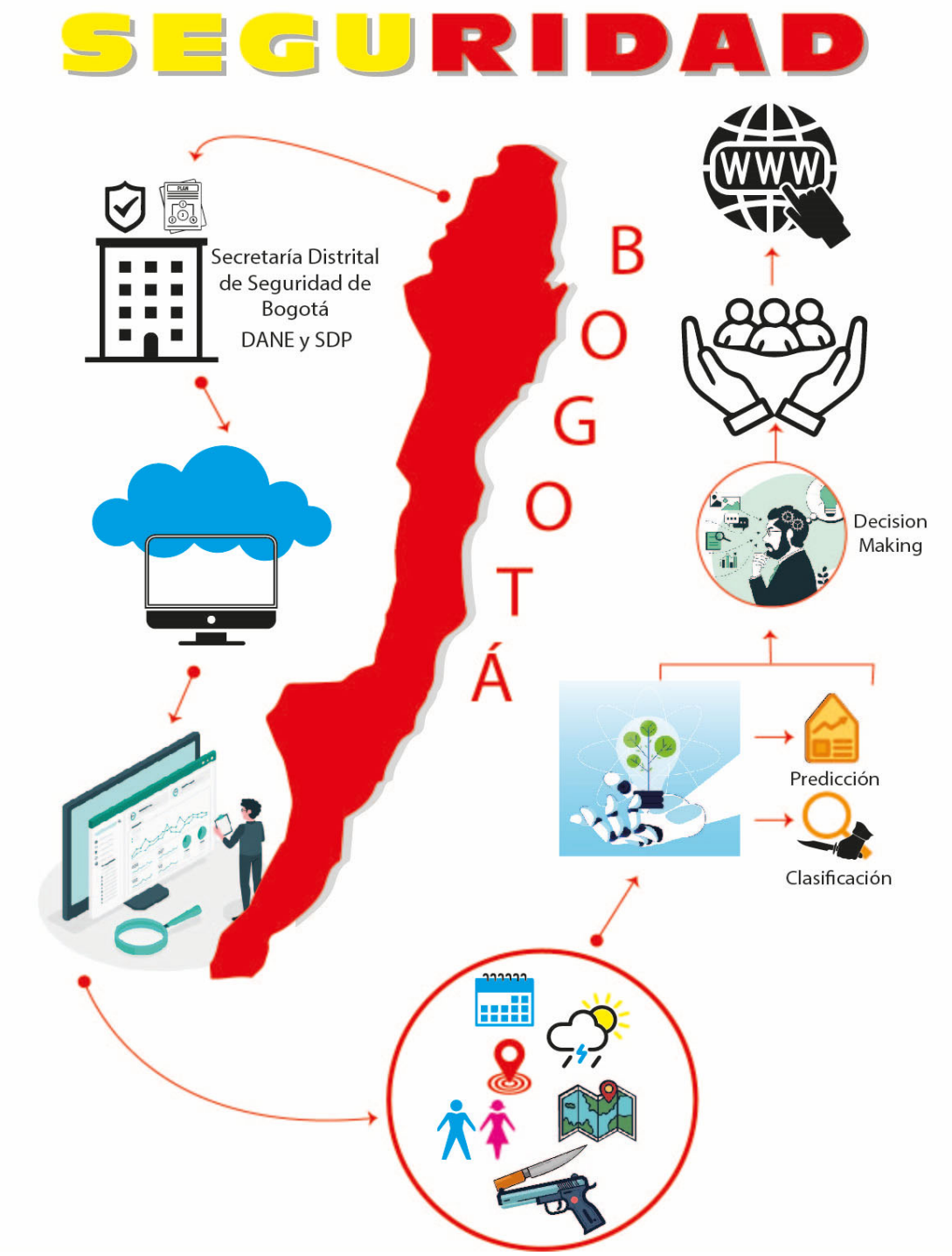

Ilustración 1 - Resumen Gráfico. Elaboración propia a partir de recursos de Freepik.com

RESUMEN

La seguridad ciudadana es esencial para el bienestar de los habitantes. Sin embargo, el aumento de las tasas de criminalidad ha generado la necesidad de proveer a las autoridades y la ciudadanía herramientas efectivas para la toma de decisiones, con esto la implementación de modelos de Machine Learning toma mayor relevancia para apoyar estos esfuerzos.

Con el objetivo de contribuir a la solución del problema de inseguridad, se desarrollaron dos modelos de Machine Learning enfocados en analizar los patrones de criminalidad en Bogotá durante el periodo post-pandemia de enero 2021 a mayo 2023. El primero es un modelo de clasificación que permite identificar la ocurrencia de delitos de alto impacto (si o no), obteniendo un F1-score entre 0.70 y 0.80. El segundo es un modelo de regresión para predecir la cantidad de estos delitos, con un error absoluto medio (MAE) entre 0.2 y 3.13. Ambos modelos brindan resultados prometedores para apoyar la toma de decisiones informadas en seguridad ciudadana en la ciudad.

Para el desarrollo del modelo se utilizaron 2 bases de datos públicas, una de ellas de la Secretaría Distrital de Seguridad de Bogotá, la cual contiene los reportes mensuales de delitos en el periodo mencionado, si bien se cuenta con datos históricos desde 2010 para no afectar el análisis de la situación actual, estos no se tendrán en cuenta debido a que pueden ser menos relevantes para las condiciones actuales y futuras, ya que la pandemia ha modificado los patrones delictivos de forma significativa y posiblemente duradera, y la otra corresponde a la Encuesta Multipropósito Bogotá – Cundinamarca (EM) realizada en 2021 y publicada en 2022, dicha encuesta se realizó debido a un convenio entre el Departamento Administrativo Nacional de Estadística (DANE) y la Secretaría Distrital de Planeación (SDP), los cuales buscaban obtener información socioeconómica y de

entorno urbano de los habitantes de Bogotá. Al combinar las 2 bases de datos manejamos variables categóricas como fecha, año, rango del día, localidad, upz, sexo, arma empleada, y variables numéricas como ingreso per cápita, número de personas, pobreza multidimensional, desempleado, personas por hogar entre otras.

El aporte de este trabajo será brindar una solución basada en inteligencia artificial que apoye la labor de las entidades a cargo de la seguridad y la toma de decisiones informadas para proteger a los ciudadanos, así como también a los mismos ciudadanos cuando acceden a la herramienta, básicamente el entregable es un enlace de acceso y al momento de ingresar tienen la opción de utilizar listas desplegables para seleccionar y de esta manera tendrán la información del tipo y cantidad del delito.

INTRODUCCIÓN

La seguridad ciudadana es un aspecto fundamental para garantizar el bienestar y la calidad de vida de los habitantes de una ciudad. El aumento en un 26.3% de los índices de hurto a personas entre 2021 a 2022 según el boletín de indicadores de Seguridad y Convivencia emitido por la Oficina de Análisis de Información y Estudios Estratégicos (OAIEE, 2022). Este incremento sostenido de la criminalidad ha generado gran preocupación e incertidumbre entre los ciudadanos, quienes se sienten vulnerables e inseguros en el espacio público y en sus propios barrios. Asimismo, la victimización y el miedo al delito han obligado a muchos a alterar sus rutinas y a limitar su libertad de movimiento en la ciudad.

Ante esta situación crítica, es imperativo que las autoridades cuenten con herramientas confiables y efectivas para monitorear las tendencias delictivas,

anticipar su evolución e identificar las zonas de mayor riesgo. Solo así podrán focalizar adecuadamente sus recursos y desplegar estrategias preventivas.

El propósito de este trabajo es proponer un modelo de aprendizaje automático de detección y un modelo de predicción de delitos considerando en la ciudad de Bogotá.

El modelo utilizará 2 bases de datos:
Una proporcionada por la Secretaría Distrital de Seguridad, Convivencia y Justicia de Bogotá, que contiene los hechos delictivos denunciados por las víctimas de delitos de alto impacto con datos categóricos. Estos hechos delictivos están agrupados por temporalidad (rango del día, día, mes y año), ubicación del delito (upz, localidad) y característica del hecho (sexo de la víctima, arma empleada y tipo de delito). La base de datos se actualiza de manera mensual, lo que garantiza que el modelo tenga acceso a la información más reciente. La base de datos se encuentra de manera pública en la página oficial de la Secretaría de Seguridad, Convivencia y Justicia de Bogotá (Secretaría Distrital de Seguridad, 2023) y para el desarrollo del proyecto se tomó el periodo de Enero de 2021 a Mayo de 2023, contamos con datos desde 2010 pero para no afectar el análisis de la situación actual, estos no se tendrán en cuenta debido a que pueden ser menos relevantes para las condiciones actuales y futuras, ya que la pandemia ha modificado los patrones delictivos de forma significativa y posiblemente duradera. Además, agrupamos los datos en 3 grupos principales: Delitos contra la propiedad, Delitos contra las personas y Delitos de tránsito.

La otra corresponde a la Encuesta Multipropósito realizada por el convenio entre el Departamento Administrativo Nacional de Estadística (DANE) y la Secretaría Distrital de Planeación (SDP) con periodicidad de publicación de 3 años, su última publicación fue en Julio de 2022 sobre la encuesta con datos recopilados entre 15

de Abril de 2021 a Noviembre 30 de 2021, esta base de datos se encuentra de manera pública en la página del DANE (DANE, 2021). Esta proporciona información socioeconómica y de entorno urbano de los habitantes de Bogotá con datos numéricos debido a las variables manejadas como ingreso per cápita, número de personas, pobreza multidimensional, desempleado, personas por hogar entre otras. Cabe mencionar que las preguntas de la encuesta anterior realizadas en 2017 no son las mismas realizadas en 2021, entonces solamente se tomaron en cuenta las del último periodo.

ESTADO DEL ARTE

Se buscaron artículos indexados y no indexados relacionados con las temáticas: criminalidad en ciudades, predicción del crimen, clasificación del crimen, redes neuronales y Machine Learning. Se encontró que la mayoría de los artículos eran recientes, debido a que estos fueron publicados entre los años 2017 y 2023.

Diversos estudios han cuestionado el uso de regresiones lineales, métodos tradicionales para predecir el crimen en ciertas áreas, debido a los supuestos no cumplidos y a los posibles resultados erróneos (Alves et al., 2018; He & Zheng, 2021). Como alternativa, se han propuesto métodos de aprendizaje automatizado para identificar variables explicativas del crimen y realizar predicciones (Alves et al., 2018; He & Zheng, 2021; Wheeler & Steenbeek, 2021a). Otros enfoques incluyen el análisis espacial (ToppiReddy et al., 2018) y el uso de indicadores de riesgo (Lisowska-Kierepka, 2021) para identificar zonas con altas tasas de criminalidad.

En cuanto a las características que mejor describen el crimen, se ha encontrado que los indicadores urbanos, como el desempleo, el analfabetismo y la población masculina, son relevantes (Alves et al., 2018). Además, la cercanía a parques y escuelas primarias, así como la reducción de carreteras internas y callejones sin salida, se asocian con menores tasas de criminalidad (He & Zheng, 2021). Por último, se identificó que las zonas con mayor probabilidad de delincuencia se

encuentran en el centro de la ciudad y en las áreas adyacentes (Lisowska-Kierepka, 2021)

Nos enfocamos en gran medida en la revisión sistemática de literatura realizada por (Jenga et al., 2023a), quienes llevaron a cabo un extensivo proceso de recopilación y síntesis de la investigación previa sobre predicción de delitos mediante aprendizaje automático. Partiendo de una lista inicial de 353 publicaciones identificadas, los autores aplicaron criterios de inclusión y exclusión rigurosos, culminando en una selección final de 68 artículos relevantes publicados entre 2010 y 2022. Su estudio proporciona información sobre los objetivos de investigación, fuentes y conjuntos de datos, las variables independientes, algoritmos de Machine Learning, métricas de evaluación aplicadas y desafíos comunes en este campo. Con base en esta investigación, nos centramos en la selección de aquellos artículos que mejor se alinean con nuestros objetivos de investigación y que utilizan bases de datos similares a las que tenemos disponibles. Nos centramos en los artículos que utilizan datos históricos relacionados con características espaciales, temporales y de ocurrencia de delitos, y que tienen como objetivo principal predecir la ocurrencia de un delito o la cantidad de estos, bajo esto para este proyecto nos centramos en 8 artículos indexados y 2 no indexados, de 18 artículos publicados entre 2017 a 2023, los cuales trabajaron modelos de Machine Learning aplicados fueron LSTM (Redes Neuronales Recurrentes con Memoria a Largo Plazo), redes neuronales que combinan capas convolucionales CLSTM, random forest, SVM, regresión logística, ensemble y XGBoost.

En la Tabla 1 se encuentran detallados los artículos tanto indexados como no indexados, siendo los 8 primeros indexados y los 2 siguientes no indexados.

Tabla 1 - Artículos analizados

Título / Journal	***Base de datos utilizada / Variables***	***Metodología***	***Resultados***
Predictability Comparison of Three Kinds of Robbery Crime Events Using LSTM / Association for Computing Machinery New York, NY, United States **2019**	Se utilizaron datos de crímenes diarios en las ciudades de Atlanta (2009-2017) y Baltimore (2011-2016). Los datos incluyen tipo de crimen, fecha, latitud y longitud de ocurrencia. / Se compararon 3 tipos de crímenes de robo: comercial, residencial y peatonal.	**Modelo de Predicción** El artículo propone utilizar el modelo de Redes Neuronales Recurrentes con Memoria a Largo Plazo (LSTM, por sus siglas en inglés) para predecir diferentes tipos de robos en distintas escalas espacio-temporales. Se realiza un análisis comparativo de la predictibilidad de tres tipos de robos (robos comerciales, robos en áreas residenciales y robos a peatones) en dos ciudades, Atlanta y Baltimore, utilizando datos de ocurrencia diaria de crímenes. Se evaluó la predicción en diferentes escalas espacio-temporales: tamaño de celda (resolución espacial) y longitud del historial de entrada (dependencia temporal). Como métrica se usó el coeficiente de correlación R entre los valores predichos y reales (Mei & Li, 2019).	* Los tres tipos de robos tienen diferentes niveles de predictibilidad en cada ciudad y en diferentes escalas espacio-temporales. En Atlanta, los robos comerciales son más predecibles que los robos en áreas residenciales y los robos a peatones. En Baltimore, los robos en áreas residenciales son los más predecibles. * Los robos en áreas comerciales y residenciales tienden a tener una distribución más regular en una escala espacial más grande, mientras que los robos a peatones tienen una distribución más aleatoria en todas las escalas espaciales. * La duración óptima del tiempo varía para cada tipo de robo y refleja diferentes ciclos periódicos o dependencias temporales. * En Atlanta, el robo comercial tuvo mejor predicción (R=0.75) que residencial (R=0.6) y peatonal (R=0.5). * En Baltimore, el robo residencial (R=0.9) y comercial (R=0.88) tuvieron mejor predicción que el peatonal (R=0.65).

Título / Journal	*Base de datos utilizada / Variables*	*Metodología*	*Resultados*
Risk Prediction of Theft Crimes in Urban Communities: An integrated Model of LSTM and ST-GCN / IEEE **2020**	Se utilizaron datos de crímenes en 77 comunidades de Chicago entre 2015-2020, incluyendo fecha, ubicación (latitud, longitud), tipo de crimen y factores climáticos. / * Variables utilizadas: Cantidad de crimenes diarios, día de la semana o fin de semana, día festivo, temperatura * Se enfocaron en predecir robos en cada comunidad.	**Modelo de Predicción** * Modelo integrado de LSTM & ST-GCN: - Módulo de extracción espacio temporal es combinación de GCN and ST-ResNet (ST-GCN) para extraer la transición de crímenes en el espacio a lo largo del tiempo (Spatial-Temporal Graph Convolutional Networks) - Módulo de extracción de caracteristica temporal: lo realizan con Long Short-Term Memory network (LSTM) para calcular el número de delitos y Recurrent Red neuronal (RNN) - algoritmo de tiempo (BPTT) y el parámetro de Adam algoritmo de optimización Método propuesto: Combinando la característica temporal dentro de una comunidad y las características espacio temporales entre comunidades, es posible capturar la ocurrencia del crimen y predecir efectivamente la tendencia del crimen y luego predecir el número de delitos de robo en las comunidades al día siguiente (Han et al., 2020).	* Predijeron el número de robos por día en cada comunidad. * El modelo tuvo mejor desempeño en comunidades con mayor número promedio de robos (>2 robos/día), con MAPE 0.4 y RMSE 1 crimen. En comunidades con menos robos (<2/día) el desempeño fue inferior, con MAPE 0.6 y RMSE 1 crimen. * Comparado con otros modelos (Ridge, Random Forest, LSTM simple) el modelo integrado propuesto obtuvo mejor desempeño. * La mejor predicción del modelo es en los días de semana más que en los días festivos y los fines de semana debido a que el ritmo de vida es irregular y la afluencia de turistas durante las vacaciones * Efectividad para captar la ocurrencia del crimen y predecir la tendencia en el crimen, y luego predecir el número de delitos de robo en las comunidades el próximo día. * Modelo se ve severamente afectado por variables sociales * El modelo tuvo un desempeño razonablemente bueno, con errores promedio entre 0.4% y 0.6% en el conteo de robos diario por comunidad. Y un RMSE promedio de 1 crimen al día.

Título / Journal	*Base de datos utilizada / Variables*	*Metodología*	*Resultados*
Spatio-Temporal Prediction of Baltimore Crime Events Using CLSTM Neural Networks / IEEE **2020**	* Se utilizaron datos de crímenes ocurridos en la ciudad de Baltimore, EE. UU. entre enero de 2016 y diciembre de 2018. Los datos incluyen información espacio-temporal (fecha, hora, latitud, longitud) así como tipo de crimen / Variables espaciales y temporales. * Se seleccionaron dos tipos de crímenes: robos callejeros (street robbery) y hurtos (larceny), por ser los más frecuentes en los datos.	***Modelo de Predicción*** * Manejaron redes neuronales CLSTM, que combina capas convolucionales y capas LSTM para extraer características espaciales y temporales de los datos de entrada. La red CLSTM se entrena utilizando el conjunto de datos históricos y se ajusta para capturar patrones y relaciones complejas entre las variables espaciales y temporales. * La metodología consistió en dividir el área de la ciudad en una grilla de 8x8 o 16x16 celdas. En cada celda se contabiliza la cantidad de crímenes por día. Secuencias de matrices de crimen por día se usan como entrada a una red neuronal Convolutional LSTM (CLSTM) para predecir la ocurrencia de al menos 1 crimen en la celda para el día siguiente. Se evaluaron 3 escenarios: Matrices de días previos para predecir próximo día Matrices de eventos agregados en d días previos para predecir próximos d días (8x8 celdas) Matrices de eventos agregados en d días previos para predecir próximos d días (16x16 celdas) (Esquivel et al., 2020)	*El objetivo principal del artículo es predecir la probabilidad de ocurrencia de nuevos eventos delictivos en ubicaciones y momentos específicos en Baltimore, aplicaron una red neuronal CLSTM a datos espacio-temporales de crímenes en Baltimore, logrando buenos resultados de predicción en algunos escenarios. * Las métricas de evaluación fueron precisión, AUC-ROC y AUC-PR. * Los mejores resultados se obtuvieron para hurtos con **grilla 8x8**, usando 7 días previos agregados. Se obtuvo **precisión de 86%**, AUC-ROC de 80% y AUC-PR 93%. Para robos, en el mismo escenario (8x8, 7 días), precision 83%, AUC-ROC 81% y AUC-PR 77%.

Título / Journal	*Base de datos utilizada / Variables*	*Metodología*	*Resultados*
Machine learning in crime prediction / Journal of Ambient Intelligence and Humanized Computing **2023**	*Buscaron en 5 bases de datos científicas: ACM Digital Library, IEEE Xplore, Springer Link, Science Direct y Scopus. En total se encontraron 353 publicaciones, de las cuales 68 cumplieron con los criterios de inclusión después de aplicar criterios de exclusión y evaluación de calidad. * El 80% de los estudios utilizaron conjuntos de datos públicos, principalmente de registros policiales/gubernamentales. Algunos utilizaron datos de Twitter, Facebook, censos, clima, etc. / Las variables o características (features) utilizadas en los estudios fueron agrupadas en: parámetros del crimen, parámetros de fecha/tiempo, parámetros del delincuente, parámetros de la víctima y parámetros de localización.	***Metodología de Revisión Sistemática de la Literatura (Systematic Literature Review - SLR)*** *La investigación consistió en una revisión sistemática de literatura sobre el uso del aprendizaje automático (Machine Learning) en la predicción de crímenes entre 2010 y 2022. * Los algoritmos más utilizados fueron redes neuronales, Random Forest y KNN. La mayoría de los estudios utilizaron aprendizaje supervisado. *Las métricas de evaluación más comunes fueron precisión, área bajo la curva ROC y precisión. Precisión (Precision): relación de verdaderos positivos sobre verdaderos positivos + falsos positivos. Área bajo la curva ROC (AUC): capacidad del clasificador para diferenciar entre clases. Precisión (Recall): relación de verdaderos positivos sobre verdaderos positivos + falsos negativos. Exactitud (Accuracy): proporción de predicciones correctas sobre el total. Error cuadrático medio (RMSE): medida del error entre los valores predichos y reales. Error absoluto medio (MAE): medida del error absoluto entre predicción y real. Puntaje F1: media armónica de precisión y sensibilidad (Jenga et al., 2023b).	*El propósito de este artículo es proporcionar una mejor comprensión de algoritmos de ML utilizados en la predicción de delitos y sus análisis. Literatura analizada desde enero 2010 hasta agosto 2022 *Entre los principales desafíos identificados están la recolección, almacenamiento, preprocesamiento de datos y problemas de rendimiento de los modelos. *El documento concluye que el aprendizaje automático puede ayudar a autoridades a mejorar la seguridad ciudadana, pero se requiere más investigación en áreas como aprendizaje no supervisado y explicabilidad de los modelos.

Título / Journal	***Base de datos utilizada / Variables***	***Metodología***	***Resultados***
Applied Machine Learning in Social Sciences: Neural Networks and Crime Prediction / MDPI **2021**	Base de datos del Ministerio de Justicia y Seguridad de la Ciudad de Buenos Aires, del periodo del 2016 al 2019 / El modelo tiene en cuenta varias características predictivas como año, mes, día, zona horaria, comuna (15) y el tipo de delito (hurto, robo,lesiones y homicidio) X (variables independientes): “ año ”, “ mes ”, “ día ”, “ franja_horaria ”, “ comuna ”, “ tipo_delito ”. Y (variable dependiente): “ cantidad_registrada ”	***Modelo de Predicción*** Uso de redes neuronales en la predicción del crimen en la ciudad de Buenos Aires, generando predicción de delitos basados en las 15 comunas; aplicaron Kmeans, método del codo y redes neuronales. Los clusters fueron manejados por día y franja horaria. 361.184 records con train 80% y test 20%	*El mejor modelo fue una red neuronal de 3 capas, con 32 neurones en la capa de entrada, 8 en la oculta y 1 en la de salida * MAE (error absoluto medio) : **0.3317 en train y 0.4095 en test**, predice con un error de menos de **0.5 casos de diferencia**, por ejemplo si en realidad se registraron 3 casos, la predicción arrojaría valores entre 2.6 y 3.6 lo que en términos de la posible modos de uso del modelo es apropiado * Para el modelo se recomienda usar año, mes, día, zona horaria, comuna y tipo de delito * Una vez que la red neuronal ha sido entrenada, puede ser utilizada para predecir futuros delitos dada una nueva serie de datos de entrada.
Predicting Crime and Other Uses of Neural Networks in Police Decision Making / Frontiers in Psychology **2021**	Se utilizaron datos de crímenes en la ciudad de Detroit entre 2016-2020 / Fecha, hora, código de crimen, ubicación (latitud/longitud).	***Modelo de Predicción*** Se desarrollaron dos modelos de redes neuronales (RN): RN para predecir el tipo de crimen dado la ubicación y hora. RN para predecir la zona (código postal) dado el tipo de crimen y hora. Las RN fueron entrenadas con backpropagation. Se probaron arquitecturas con 1 y 2 capas ocultas (Forradellas et al., 2021).	* El primer modelo predijo el tipo correcto de 27 categorías de crimen el 16.4% de veces sobre ~263k datos de validación. * Agrupando los crímenes **en 7 categorías, la precisión aumentó a 27.1%.** * El segundo modelo predijo la zona correcta (de 30 zonas) el 8.2% de veces, y dentro de zonas cercanas el 31.2% de veces. * Obtuvieron **precisiones de 16-31%** en condiciones limitadas de información.

Título / Journal	*Base de datos utilizada / Variables*	*Metodología*	*Resultados*
Crime event prediction with dynamic features / EPJ Data Science a Springer Open Journal **2018**	* Datos de delitos en Brisbane, Australia y Nueva York, EEUU entre 2012 y 2013. * Datos demográficos de ambas ciudades * Datos de Foursquare con información de lugares (POIs) y check-ins de usuarios en ambas ciudades./ * Variables históricas: densidad de delitos en los últimos 7 y 30 días. * Variables geográficas: densidad y distribución de categorías de POIs, diversidad regional. * Variables demográficas: ratios de género, ingresos, edad, hogares alquilados, etc. * Variables dinámicas: ratio de visitantes, popularidad regional, entropía de visitantes, homogeneidad de visitantes, frecuencia de observación y recuento de visitantes.	***Modelo de Predicción*** * Se dividió cada día en 8 intervalos de 3 horas y se predijo si ocurriría cada tipo de delito en cada región censal en el siguiente intervalo. * Se probaron modelos de Machine Learning como random forest, redes neuronales, SVM, regresión logística y ensemble. * Se estimaron las variables dinámicas faltantes usando factorización matricial. Se evaluó el desempeño con y sin variables dinámicas para cada modelo Se balancearon los datos de entrenamiento para manejar la escasez de datos de delitos (clase minoritaria), mediante **undersampling** aleatorio de ejemplos de la clase mayoritaria (no delito) (Rumi et al., 2018).	* El desempeño en términos de AUC mejoró al incluir variables dinámicas, especialmente para delitos como robo, entrada ilegal, delitos de drogas y fraude. * El modelo ensemble, que combinaba SVM, árboles de decisión, bosques aleatorios y análisis discriminante lineal, obtuvo el mejor desempeño. * Las variables dinámicas estimadas mantuvieron correlación con los delitos, lo que permite usarlas a pesar de la escasez de datos. * El modelo Random Forest demostró ser el más robusto y constante en la mejora de precisión al incorporar las variables dinámicas derivadas de los datos de Foursquare, reportando: AUC (Area Under ROC Curve): Se observan aumentos de hasta 16% para asalto en Brisbane y 4% para robo en Nueva York al añadir las variables dinámicas. Precisión (Accuracy): Se muestran los valores de precisión antes y después de incluir variables dinámicas. Por ejemplo, para robo en Brisbane entre las 21-24h, la precisión aumenta de 92.88% a **97.96%** al añadir las variables. F1-score: Para robo en Nueva York entre 15-18h, el F1-score crece de 85.18% a **86.46%** al incluir las variables dinámicas.

Título / Journal	*Base de datos utilizada / Variables*	*Metodología*	*Resultados*
Research on prediction of multi-class theft crimes by an optimized decomposition and fusion method based on XGBoost / ELSEVIER **2022**	Datos de casos de robos registrados en la "ciudad de H" durante el año 2019, proporcionados por la Oficina de Seguridad Pública. Contiene 1521 casos con 19 variables como lugar, hora, objeto robado y categoría del caso / * Lugar comercial (clasificado en 19 categorías) * Hora del crimen (5 categorías) * Condición climática (4 categorías) * Objeto robado (11 categorías) * Categoría del caso (4 categorías: robos de vehículos, allanamiento, hurtos y otros)	***Modelo de Clasificación*** Se utilizó XGBoost como modelo base. Se probaron dos métodos de descomposición y fusión: OVR-XGBoost y **OVO-XGBoost.** Se dividió el conjunto de datos en entrenamiento (70%) y prueba (30%). Se utilizó **SMOTENN** para balancear los datos de entrenamiento. Se entrenaron los modelos y evaluaron con métricas como precisión, recall y F1-score (Yan et al., 2022).	* El modelo OVO-XGBoost tuvo la mejor precisión promedio **(81.24%)**, superando al modelo base XGBoost (59.51%). * OVO-XGBoost también mejoró el recall promedio (macroR **78.08%** vs 61.40% en XGBoost). * Para la categoría con menos muestras (t1), OVO-XGBoost tuvo una precisión de **52.7%** vs 24.55% de XGBoost.

Título / Journal	*Base de datos utilizada / Variables*	*Metodología*	*Resultados*
Modelo para la Caracterizac ión del Delito en la Ciudad de Bogotá, Aplicando Técnicas de Minería de Datos Espaciales. / Repositorio Institucional - Universidad Distrital Francisco Jose de Caldas **2017**	* Delitos contra el patrimonio ocurridos en la localidad de Chapinero en Bogotá entre el 1 de enero de 2015 y el 26 de septiembre de 2015 *De un total de 27.606 registros, se toma como muestra 2.101 registros/ *Infraestructura de seguridad: comandos, estaciones de policía, CAIs, cuadrantes *Tipo de delito: cosquilleo, atraco, raponazo, engaño, factor de oportunidad *Armas: arma blanca, arma de fuego, sin uso de armas, cortopunzante *Fecha del delito: semana, mes, día, hora, rango horario *Identificación del delito: modalidad, arma utilizada * Ubicación (coordenadas geográficas)	Modelo de Agrupamiento La metodología consistió en una combinación de recolección de datos espaciales, diseño de base de datos, aplicación de algoritmos de minería de datos espaciales y análisis de resultados. Plantea el diseño de un modelo de caracterización del delito contra el patrimonio en Bogotá, aplicando técnicas de agrupamiento de minería de datos K-means y DBscan (Peña Suarez, 2017).	* Se identificaron 3 focos delictivos aplicando los algoritmos K-means y DBSCAN en la localidad de Chapinero: Calles 43-45, Carrera 6-13 (DBSCAN), Calles 43-45, Av. Caracas-Carrera 13 (DBSCAN), Calles 52-55, Carrera 9-Av. Caracas (K-means) * La modalidad de delito más frecuente fue el atraco. * El arma más utilizada fueron las armas blancas. * Los rangos horarios con más delitos fueron entre las 10am-3pm y 4pm-12am * Los días con más delitos son los martes, jueves y viernes. * Los meses con más delitos fueron marzo y mayo. * Los algoritmos de agrupamiento permitieron caracterizar el comportamiento delictivo en Chapinero. * La mayoría de delitos ocurren en zonas de alta afluencia y poca presencia policial. * Las estaciones de policía están alejadas más de 400 metros de los focos delictivos. * El modelo permite identificar zonas de riesgo con características similares para focalizar la vigilancia pero no especifica resultado de métricas

Título / Journal	*Base de datos utilizada / Variables*	*Metodología*	*Resultados*
Predicción del delito en Colombia: experiencia en ciudades intermedias / Dirección de Estudios Económicos **2021**	Datos del Sistema de Información Estadístico, Delincuencial, Contravencional y Operativo de la Policía Nacional (SIEDCO) / * Tipo de delito (homicidio, violencia intrafamiliar, hurtos a personas y lesiones personales) * Departamento, municipio * Fecha (mes, día, año) * Edad y género de la víctima * Zona urbana o rural * Hora del día * Código DIVIPOLA * Latitud y longitud	Modelo de Predicción Relaciones espaciales de grafos por semanas. Análisis descriptivo de los datos a nivel nacional y para Bucaramanga Selección de Bucaramanga para el estudio piloto por ser una de las ciudades con mayor criminalidad y tener información georreferenciada disponible Desarrollo de un modelo de predicción utilizando procesamiento de señales para grafos y una adaptación del modelo TF-IDF El modelo predecía la probabilidad de ocurrencia de un delito por sección DANE y semana Se probó con modelos de aprendizaje supervisado KNN y SVM (Gélvez-Ferreira et al., 2021)	* Estudio realizó una predicción del delito en Bucaramanga, Colombia * Encontraron que cada hecho puede clasificarse en cuatro categorías: homicidio, violencia intrafamiliar, hurtos y lesiones personales. * Cada categoría tiene un comportamiento distinto tanto en el tiempo como el espacio y su frecuencia (Gélvez-Ferreira et al., 2021). * El modelo resultante fue el KNN con frecuencia semanal y obtuvo un nivel de precisión en la predicción entre el 50% y 60%(Gélvez-Ferreira et al., 2021). * Concluyen que la predicción en centros urbanos pequeños puede ser ineficiente por la cantidad de hechos registrados. Además, los delitos en Bucaramanga responden a una estructura más no a un componente aleatorio (Gélvez-Ferreira et al., 2021).

Con el objetivo de realizar un análisis comparativo entre los trabajos del estado del arte y el presente estudio, se seleccionaron 7 de los 8 artículos indexados revisados previamente. El artículo descartado correspondía a una revisión sistemática de literatura sobre el uso de Machine Learning en predicción delictiva, por lo que no fue incluido al no representar un desarrollo específico de modelos. Los 7 artículos restantes fueron analizados en detalle para evidenciar los aportes del presente trabajo. En la Tabla 2 se presenta una comparación cualitativa entre cada uno de los artículos analizados.

Tabla 2 - Comparación cualitativa de trabajos relacionados

Título / Journal	***Comparación cualitativa con el presente trabajo***
Risk Prediction of Theft Crimes in Urban Communities: An integrated Model of LSTM and ST-GCN 2020	El presente trabajo modela múltiples categorías de delitos en Bogotá incorporando variables espaciotemporales y socioeconómicas relevantes para la ciudad. A diferencia de este artículo que se enfoca sólo en robos en comunidades de Chicago, el valor agregado de este trabajo es modelar diversos tipos de delitos específicos en la capital de Colombia, capturando así dinámicas delictivos locales vinculadas a factores sociales y económicos particulares del contexto de Bogotá.
Spatio-Temporal Prediction of Baltimore Crime Events Using CLSTM Neural Networks 2020	A diferencia de este artículo que se centra únicamente en robos y hurtos en Baltimore, este trabajo modela varios delitos de alto impacto en la ciudad de Bogotá, incorporando también variables socioeconómicas de contexto como ingresos, empleo, pobreza, entre otros. El valor agregado es entender las dinámicas delictivas en Bogotá en función de factores espaciales, temporales y sociales específicos.
Applied Machine Learning in Social Sciences: Neural Networks and Crime Prediction 2021	A diferencia de este artículo que aborda un único tipo de delito en Buenos Aires, este trabajo modela múltiples categorías delictivas relevantes para la ciudad de Bogotá, permitiendo entender variaciones y particularidades entre distintos tipos de delitos. El valor agregado es generar conocimiento específico sobre las dinámicas delictivas en la capital de Colombia.
Predicting Crime and Other Uses of Neural Networks in Police Decision Making 2021	Si bien este artículo predice tipo y ubicación de delitos en Detroit mediante redes neuronales, el presente trabajo modela múltiples delitos relevantes para Bogotá, incluyendo variables socioeconómicas de contexto e implementando técnicas de vanguardia como AutoML que explotan la gran cantidad de datos y variables disponibles. El valor agregado es generar conocimiento específico al contexto de Bogotá para el diseño de políticas de seguridad ciudadana.
Crime event prediction with dynamic features 2018	A diferencia de este artículo que aborda 6 categorías de delitos en ciudades de Estados Unidos, este trabajo modela múltiples tipos de delitos relevantes para la realidad específica de Bogotá, incorporando variables espaciales, temporales y socioeconómicas del contexto de la ciudad. El valor agregado es entender las dinámicas delictivas en la capital colombiana considerando factores sociales, económicos y espaciales particulares

Research on prediction of multi-class theft crimes by an optimized decomposition and fusion method based on XGBoost 2022	Mientras que este artículo se enfoca en predecir única y exclusivamente robos en una ciudad de China, el presente trabajo modela múltiples categorías de delitos relevantes al contexto de Bogotá, incorporando variables espaciales, temporales y socioeconómicas de la ciudad. El valor agregado es entender varios tipos de delitos asociados a factores particulares del contexto de Bogotá.

Este trabajo realiza valiosos aportes al estado del arte existente sobre la modelación en términos de clasificación y predicción de delitos, los cuales se resumen a continuación:

- A diferencia de estudios previos centrados en una o pocas categorías delictivas, este trabajo aborda un amplio conjunto de delitos de alto impacto en Bogotá, incorporando variables espaciales, temporales y socioeconómicas relevantes para caracterizar cada tipo de actividad criminal.

- El enfoque específico en la ciudad de Bogotá permite modelar dinámicas delictivas propias de la capital colombiana, a diferencia de investigaciones previas en otras ciudades o países.

- La implementación de una técnica de agrupamiento semántico de delitos representa una innovación valiosa en la preparación de datos para modelado predictivo. A diferencia de enfoques tradicionales que tratan cada delito de forma aislada, este trabajo consolidó los distintos tipos de delitos en tres categorías principales según su naturaleza.
- Se implementan y comparan rigurosamente diversas técnicas de vanguardia para clasificación (Naive Bayes, Árboles de Decisión, Análisis Discriminante Lineal, Regresión Logística, Bosques Aleatorios, KNeighbors) y predicción (Regresión Lineal, Bosques Aleatorios, KNeighbors, Árboles de Decisión).

Adicionalmente, se aplica la herramienta Naive AutoML para optimizar los modelos predictivos.

- A diferencia de estudios previos que desarrollan un solo tipo de modelo, este trabajo construye modelos integrales de clasificación y predicción, explotando al máximo la información disponible.

- Los modelos incorporan efectivamente variables categóricas y numéricas, tanto espaciales y temporales como socioeconómicas.

- Los modelos permiten predecir volúmenes delictivos por tipo específico, generando información de utilidad práctica para la prevención focalizada en Bogotá.

- El uso de datos públicos promueve la transparencia, el acceso a información valiosa para la ciudadanía y la toma de decisiones informada en materia de seguridad.

PREGUNTA DE INVESTIGACIÓN

¿Cuáles son los modelos de Machine Learning que permiten obtener la mayor precisión en la clasificación de tipos de delitos y el menor error en la predicción de cantidades de delitos en Bogotá en el periodo 2021-2023?

OBJETIVOS

GENERAL

Desarrollar modelos de machine learning para la clasificación de tipo de delitos y la predicción de cantidades de delitos en Bogotá entre 2021 y 2023, mediante la preparación y procesamiento de datos, la evaluación de algoritmos y selección de los más precisos.

ESPECÍFICOS

- Diseñar la estructura de procesos requerida que incluya las etapas de consolidación de bases de datos, preprocesamiento, entrenamiento y validación de modelos, para el análisis predictivo de delitos en la ciudad de Bogotá.

- Entrenar, evaluar y comparar al menos 3 algoritmos de clasificación y 3 de regresión, seleccionando en cada caso aquel que presente mayor precisión, para obtener los modelos óptimos de predicción.

- Implementar una interfaz sencilla para permitir a los usuarios acceder a las predicciones de los modelos.

MARCO CONCEPTUAL

Para recolectar y procesar información sobre las zonas de riesgo de criminalidad se debe contar con la información más reciente y que provenga de grandes y variadas bases de datos. A esto se le conoce como Big Data que se caracteriza por el volumen, la variedad y velocidad de los datos. Esta información se analiza y sirve de apoyo para la toma de decisiones (Larson & Chang, 2016). La analítica convierte

rápidamente estos datos en información para su análisis. Este análisis puede ser descriptivo y/o predictivo (Larson & Chang, 2016).

El análisis descriptivo examina los datos para resumir sus principales características, como tendencias centrales y dispersión, se pueden utilizar herramientas de visualización de datos, que es una forma de comunicación visual de los datos de forma clara y efectiva mediante gráficos y diagramas estadísticos (ToppiReddy et al., 2018).

Además del análisis descriptivo, el machine learning se ha establecido como una herramienta esencial en el campo de la analítica de datos. Generalmente, el machine learning comprende algoritmos y técnicas estadísticas que permiten a los sistemas aprender y mejorar con la experiencia, sin ser explícitamente programados. Estos algoritmos pueden categorizarse en aprendizaje supervisado, donde los algoritmos aprenden de datos etiquetados para predecir resultados para datos desconocidos, y aprendizaje no supervisado, donde los algoritmos descubren patrones y relaciones en los datos por sí mismos (Jagadish et al., 2014).

El análisis predictivo utiliza datos históricos para hacer inferencias sobre eventos futuros por medio de modelos estadísticos o algoritmos de machine learning. Adicionalmente, los algoritmos de machine learning permiten clasificar el tipo de delito ocurrido en una zona y momento específicos, así como predecir la probabilidad de riesgo de criminalidad en una ubicación determinada.

En este contexto, es importante mencionar el papel esencial del preprocesamiento de datos antes de aplicar cualquier algoritmo de machine learning. Este paso puede incluir la limpieza de datos, la eliminación de valores atípicos, la normalización y la ingeniería de características. Estos pasos ayudan a mejorar la calidad de los datos y a optimizar su estructura para una mejor performance de los algoritmos.

El análisis delictivo enfocado en patrones y tendencias requiere de técnicas analíticas avanzadas. Los algoritmos de machine learning son cada vez más utilizados para modelar fenómenos complejos como el crimen (Mohler et al., 2011; Perry et al., 2013). Desde la criminología ambiental, teorías como la de las actividades rutinarias (Cohen & Felson, 1979), la de las ventanas rotas (Kelling & Wilson, 1982) y la ecología del crimen (Brantingham & Brantingham, 1995) sustentan el análisis de los patrones espacio-temporales de criminalidad.

Entre los algoritmos de machine learning más utilizados en el análisis predictivo del crimen se encuentran las redes neuronales artificiales, Random Forest, Support Vector Machines y regresión logística. Cada uno de estos algoritmos ha demostrado ser eficaz en distintos aspectos del análisis delictivo, y su elección depende en gran medida del tipo de problema y los datos disponibles.

Un modelo predictivo preciso permite focalizar intervenciones preventivas y el patrullaje policial, maximizando su efectividad. Pero es importante mencionar que, además de la precisión, otros factores como el sobreajuste (overfitting), el subajuste (underfitting) y la validación cruzada son cruciales para la evaluación y optimización de los modelos de machine learning.

Además, la interpretación y explicabilidad del modelo son aspectos esenciales para su aplicación práctica, especialmente cuando los resultados pueden influir en las políticas públicas y las vidas de las personas. Los modelos de machine learning también deben ser optimizados para equilibrar la precisión con la interpretabilidad de los patrones detectados.

Algunos algoritmos con buenos rendimientos son Random Forest (Bosque Aleatorio) es un algoritmo de aprendizaje supervisado que funciona construyendo

múltiples árboles de decisión durante el entrenamiento y luego promediando sus predicciones al hacer inferencias; calibra el sesgo-varianza para obtener un modelo conjunto robusto con alta precisión predictiva (Breiman, 2001). Un ejemplo de este fue aplicado en el estudio de Wheeler y Steenbeek (2021), el cual buscó mapear zonas de riesgo delictivo, logrando buenos resultados de predicción (Wheeler & Steenbeek, 2021b).

Naive Bayes Multinomial es un clasificador probabilístico que modela las variables predictoras como distribuciones multinomiales. Este modelo funciona bien en datos de texto y categóricos donde las frecuencias son importantes (McCallum & Nigam, 1998). Un ejemplo de este se encuentra en el trabajo de (Walczak, 2021), Naive Bayes fue utilizado para predecir el tipo de crimen dado ciertas características, alcanzando un desempeño razonable.

El Análisis Discriminante Lineal (LDA) encuentra combinaciones lineales de características que separan clases diferentes. Modela la distribución condicional de cada clase como una gaussiana (Balakrishnama & Ganapathiraju, 1998). Un ejemplo de este lo empleo (Liu et al., 2012) en combinación con otras técnicas para clasificar eventos delictivos, obteniendo buena capacidad de discriminación entre clases.

Pero se deben considerar desafíos éticos como evitar la perpetuación de sesgos y el uso discriminatorio de los resultados. Los sesgos en los datos de entrenamiento pueden influir en los modelos resultantes y, por lo tanto, en las predicciones. Es necesario implementar técnicas para minimizar estos sesgos y garantizar que los modelos de machine learning sean justos y éticos (Lum & Isaac, 2016).

MARCO METODOLÓGICO

Este trabajo emplea la metodología de minería de datos SEMMA (Sample o Muestreo, Explore o Exploración, Modify o Modificación, Model o Modelado y Assess o Evaluación) (Azevedo & Santos, 2008). Esta metodología fue seleccionada por las siguientes razones:

- Es una metodología sencilla de implementar y se ajusta al objetivo de este proyecto.

- Se cuenta con bases de datos públicas, lo cual facilita el acceso a los datos.

- No requiere un análisis interdisciplinario para la creación de los modelos ni para la interpretación de resultados.

En la Ilustración 2 se presenta el proceso aplicado para la creación del modelo de clasificación y predicción, siguiendo la metodología SEMMA.

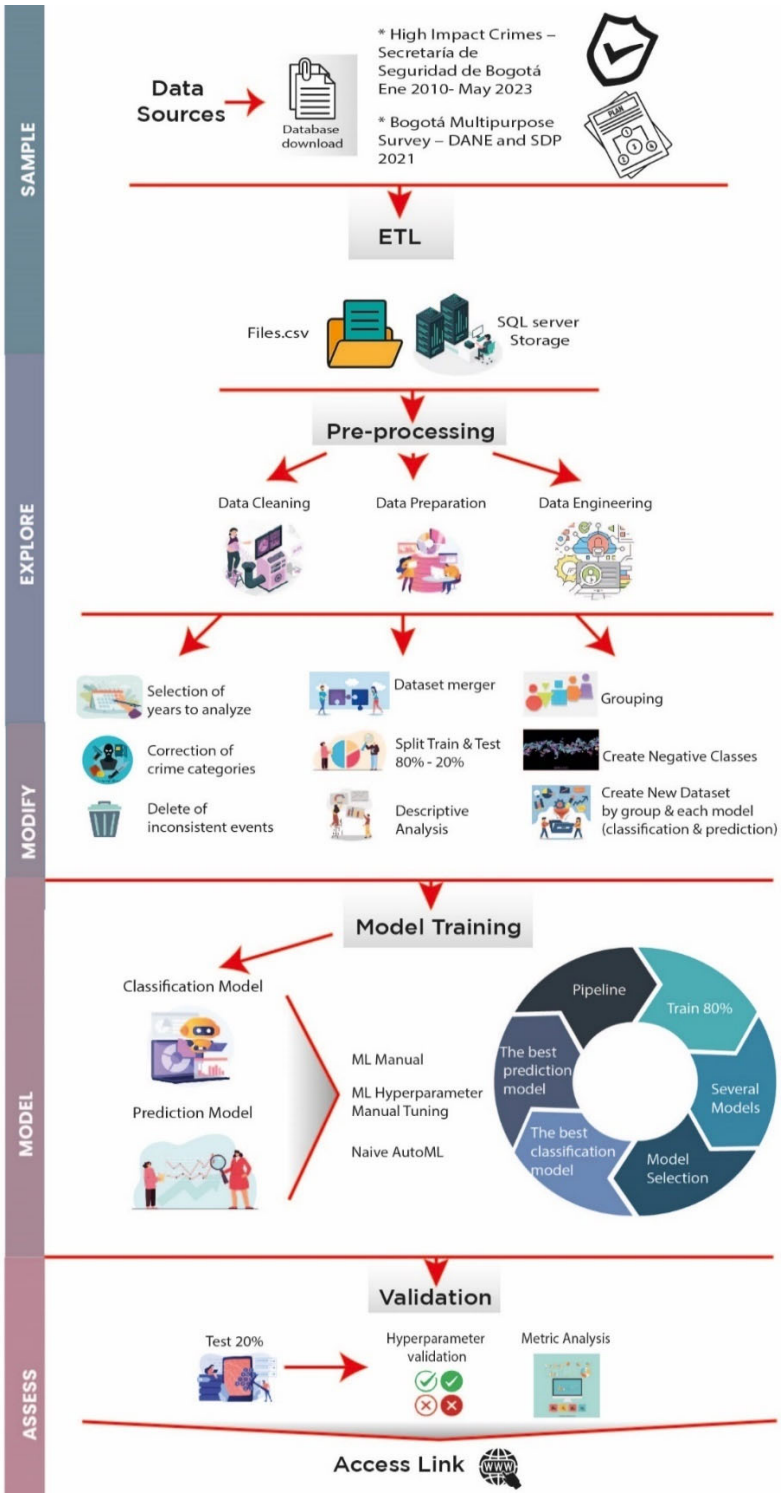

Ilustración 2 - Proceso del proyecto realizado. Elaboración propia a partir de recursos de Freepik.com e imagen de google slides.

SAMPLE / MUESTREO

Este módulo está compuesto por Data Sources y ETL (Extracción, transformación y cargue). A continuación, se presentan los detalles:

Identificación y selección de datos.

Se realizó un proceso de verificación de bases de datos públicas que registran los delitos en la ciudad de Bogotá. Para construir modelos robustos de clasificación y predicción, se decidió utilizar bases de datos con variables tanto numéricas como categóricas. En esta búsqueda se identificaron 2 bases de datos públicas:

- La primera base de datos corresponde a los delitos de alto impacto que contiene los hechos denunciados por víctimas de conductas que violan sin justa causa un bien jurídico en la ciudad de Bogotá(Secretaría Distrital de Seguridad, 2023). Estos hechos se denominan delitos de alto impacto según la Secretaría Distrital de Seguridad, Convivencia y Justicia de Bogotá. Los hechos delictivos se agrupan por temporalidad (rango del día, día, mes y año), ubicación (UPZ, localidad) y características (sexo de la víctima, arma empleada, tipo de delito). La base de datos se actualiza mensualmente y contiene 2,252,056 registros entre enero de 2010 y mayo de 2023. Los datos están disponibles para acceso público en el sitio web de la Secretaría de Seguridad, Convivencia y Justicia de Bogotá.

  En la Tabla 3 se describen las variables contenidas en la base de datos, las cuales aportan las variables categóricas.

Tabla 3 - Variables Base de datos Delitos de Alto Impacto

Variables	Tipo de dato	Posibles valores	Descripción
Fecha	Date	01/01/2021 - 21/05/2023	Fecha en la que ocurrió el delito
Anio	Int	2021-2023	Año en el que ocurrió el delito
nro_del_mes	Int	1-12	Número correspondiente al mes en el que ocurrió el delito.
Mes	Varchar	Ene, Feb, Mar, etc	Mes en el que ocurrió el delito.
nombre_dia	Varchar	Lunes, Martes, etc	Día de la semana en el que ocurrió el delito.
rango_del_dia	Varchar	Madrugada, Mañana, Tarde, Noche	Rango del día (por ejemplo, mañana, tarde, noche) en el que ocurrió el delito.
Localidad	Varchar	1- 20 ej: 11 – Suba	Localidad en la que ocurrió el delito.
Upz	Varchar	Nombres de UPZ	Zona de planeamiento (Unidad de Planeamiento Zonal) en la que ocurrió el delito
Sexo	Varchar	Femenino, Masculino	Género de la víctima del delito
Delito	Varchar	Nombre de delitos	Tipo de delito que se cometió
Modalidad	Varchar	Nombre de modalidades	Modalidad en la que se cometió el delito
arma_empleada	Varchar	Nombre de armas	Tipo de arma empleada en el delito
numero_hechos	Int	Cantidad	Número de incidentes reportados en esa entrada de datos
arma_empleada_nueva	Varchar	Con arma, Sin arma	Nueva variable creada que clasifica "SIN EMPLEO DE ARMAS" como "Sin empleo de armas" y el resto de opciones como "Con empleo de armas"

- La segunda base de datos corresponde a la Encuesta Multipropósito (EM), una investigación estadística periódica realizada por el Departamento Administrativo Nacional de Estadística (DANE) en convenio con la Secretaría

Distrital de Planeación (SDP) de Bogotá (DANE, 2021). Esta encuesta busca recopilar información socioeconómica de los hogares de Bogotá y municipios cercanos, sirviendo de insumo para formular políticas públicas focalizadas en mejorar la calidad de vida (DANE, 2021). La EM incrementa su representatividad geográfica en cada versión: inició en 2011 con 19 localidades de Bogotá, luego agregó cabeceras municipales de Cundinamarca, después aumentó a nivel de UPZ y localidades rurales (DANE, 2021). La edición más reciente realizada en 2021 cubrió 80 UPZ urbanas, 15 agrupaciones, 7 localidades rurales de Bogotá y 21 municipios urbanos y 7 rurales de Cundinamarca. Esta encuesta se actualiza cada tres años y la edición de 2021 contiene 96 registros recopilados entre el 15 de abril y el 30 de noviembre de 2021. Los microdatos están disponibles en el sitio web de la Secretaría Distrital de Planeación de Bogotá.

En la Tabla 4 se describen a continuación las variables contenidas en la base de datos, las cuales aportan las variables numéricas.

Tabla 4 - Variables Base de datos Encuesta Multipropósito

Variables	Tipo de dato	Posibles valores	Descripción
cod_upz	int	Códigos únicos de UPZ	Código único para identificar cada UPZ
Upz	varchar	Nombres de UPZ	Nombre de la Unidad de Planeamiento Zonal
cod_loc	int	Códigos únicos de Localidad	Código único para identificar cada Localidad
Localidad	varchar	Nombres de Localidad	Nombre de la Localidad
ingreso_total	float	Valores numéricos	Suma de los ingresos de todas las personas en el hogar
ingreso_per_capita	float	Valores numéricos	Ingreso per cápita del hogar
num_personas_hogar	int	Valores numéricos	Número de personas que viven en el hogar

Variables	Tipo de dato	Posibles valores	Descripción
n_hogares	int	Valores numéricos	Número total de hogares en la UPZ
indice_pobreza_multidimensional	float	Cantidad de hogares	Hogares en situación de pobreza multidimensional
pobre_monetario	int	Cantidad de hogares	Hogares en situación de pobreza monetaria
pobre_extremo_monetario	int	Cantidad de hogares	Hogares en situación de pobreza monetaria extrema
poblacion_edad_trabajar	int	Valores numéricos	Población en edad de trabajar
poblacion_fuerza_trabajo	int	Valores numéricos	Población en la fuerza laboral
poblacion_ocupada	int	Valores numéricos	Población ocupada (trabajando)
poblacion_desocupada	int	Valores numéricos	Población desempleada
ocupados_informales	int	Valores numéricos	Trabajadores del sector informal
N	int	Valores numéricos	Cantidad de individuos en la UPZ
iluminacion_via_noche	int	Cantidad de personas	Personas con percepción de mala iluminación nocturna
cerca_fabricas_industrias	int	Cantidad de hogares	Hogares cerca de fábricas e industrias
cerca_bares_discotecas	int	Cantidad de hogares	Hogares cerca de bares y discotecas
cerca_expendios_droga	int	Cantidad de hogares	Hogares cerca de expendios de droga
cerca_lotes_oscuros_peligrosos	int	Cantidad de hogares	Hogares cerca de lotes oscuros o peligrosos
cerca_canos_aguas_residuales	int	Cantidad de hogares	Hogares cerca de caños de aguas residuales
problema_entorno_inseguridad	int	Cantidad de personas	Personas que perciben inseguridad en el entorno
problema_entorno_invasion	int	Cantidad de personas	Personas que perciben invasiones en el entorno
victima_atracos_robos	int	Cantidad de personas	Personas victimizadas por atracos o robos

Variables	Tipo de dato	Posibles valores	Descripción
victima_homicidios_asesinatos	int	Cantidad de personas	Personas victimizadas por homicidio o asesinato
victima_persecucion_amenazas	int	Cantidad de personas	Personas victimizadas por persecución o amenazas
victima_extorsion_chantaje	int	Cantidad de personas	Personas victimizadas por extorsión o chantaje
victima_acoso	int	Cantidad de personas	Personas victimizadas por acoso

Se utilizaron dos bases de datos con el fin de enriquecer el análisis al incorporar información socioeconómica del contexto donde ocurrieron los delitos. La primera base contiene los registros de delitos denunciados, lo que permite analizar su distribución temporal y espacial, así como sus características. La segunda base proviene de una encuesta multipropósito que recopila indicadores socioeconómicos a nivel de UPZ y localidades. De esta manera, al cruzar ambas fuentes de datos es posible caracterizar no sólo los patrones de los delitos, sino también las condiciones sociales y económicas de los entornos en los que se producen. El uso complementario de las dos bases otorga mayor profundidad al análisis, al vincular las dinámicas delictivas con aspectos contextuales como ingresos, empleo, pobreza y percepción de seguridad de los habitantes. Así, se pueden explorar posibles relaciones entre estas variables socioeconómicas y las tendencias observadas en los delitos, generando conocimiento más integral sobre este fenómeno.

ETL - Extracción, transformación y cargue.

Se procedió a descargar la base de datos de delitos de la Secretaría de Seguridad de Bogotá por Unidad de Planeamiento Zonal (UPZ). Fue un proceso laborioso ya que los datos no estaban consolidados, sino únicamente disponibles por mes y año. Se realizó la descarga y consolidación para cada mes y año desde enero 2010 hasta

mayo 2023, generando una base de datos unificada para el período de 01 de enero de 2010 a 31 de mayo de 2023.

Posteriormente se descargó la Encuesta Multipropósito 2021, incluyendo la base de datos anonimizada con los resultados de la encuesta y la base de variables e indicadores adicionales (déficit habitacional, pobreza, empleo, etc.). Se realizó una preparación y selección de las preguntas y variables de interés para el modelo Anexo 1 asegurando que tuvieran la UPZ como llave para unirlas con la base de delitos.

Como se mencionó anteriormente, para la sección de MUESTREO concluimos lo siguiente:

- La elección de integrar dos bases de datos, una de delitos y otra socioeconómica, se justifica por cuanto permite vincular las tendencias delictivas con factores contextuales como ingresos, empleo y percepción de seguridad. De esta manera, se otorga mayor profundidad al análisis al explorar posibles relaciones entre variables socioeconómicas y la evolución de los delitos.
- Enfocarse en el periodo 2021-2023 evita sesgos derivados de cambios en la recolección de datos en años previos. Además, se descartan los años de pandemia con dinámicas atípicas en materia de delitos. Centrarse en el periodo más reciente y consistente maximiza la relevancia de los hallazgos para el contexto actual.
- La consolidación de las bases de datos unificadas en un servidor SQL facilita el acceso y preservación de los datos, evitando tener que reprocesar y unir repetidamente las fuentes originales.

EXPLORE / EXPLORACIÓN

En este módulo se realizó Data Cleaning en la base de delitos y preparación de datos en ambas bases, para asegurar la calidad de estos. A continuación, se describen los pasos:

Data cleaning.

Para limpiar la base de delitos se realizaron las siguientes tareas:

- Corrección de categorías de delitos que tenían abreviaturas o descripciones incompletas. Se unificaron categorías como Lesiones en AT y Homicidios en AT.
- Eliminación de 254,212 eventos inconsistentes que especificaban "sin localización" en UPZ, "-" en sexo o "sin UPZ".
- Verificación y confirmación de que no había datos faltantes en la base.
- Selección del periodo de tiempo relevante, descartando años previos a 2021 que contenían menos información y los años de pandemia para evitar sesgos. Con dicha selección, quitando los años de 2010 a 2021 se eliminaron 1,329,887 registros, generando como base final para análisis 667,957 registros.

Se decidió enfocarse en el periodo de enero 2021 a mayo 2023, que representa la mayor proporción de los datos, como se muestra en el siguiente gráfico:

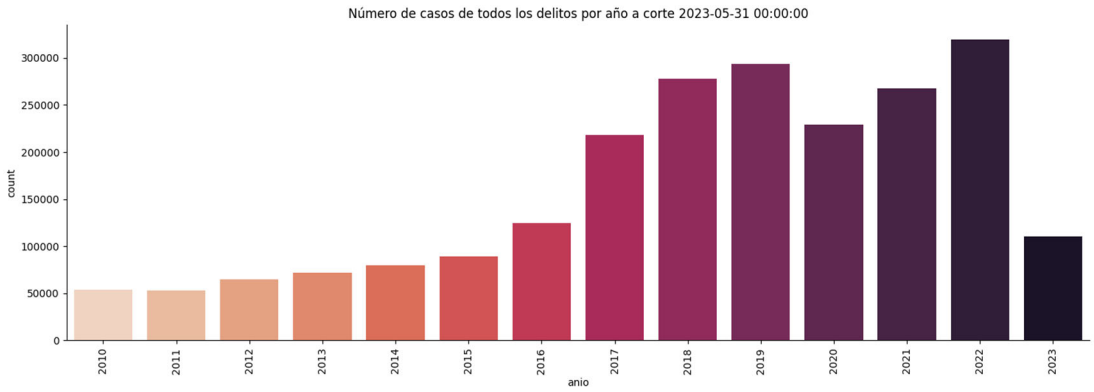

Ilustración 3 - Número de casos de todos los delitos por año

La Ilustración 3 muestra la evolución anual de eventos delictivos en Bogotá. Se observa una tendencia creciente en los años previos, que es interrumpida por una disminución significativa en 2020 producto de las restricciones de movilidad impuestas durante la pandemia de COVID-19. Sin embargo, esta caída fue temporal y los niveles delictivos retornaron una trayectoria al alza en 2021 y 2022. Esto refleja el impacto coyuntural de la pandemia en la actividad delictiva, pero evidencia que la problemática subyacente de inseguridad en Bogotá persiste y sigue manifestándose en aumentos sostenidos de los delitos años tras año.

Data preparation.

En esta parte se consolido la base final. El proceso se detalla a continuación:

- Se unificaron las dos bases de datos generando la base final para el análisis, la cual cuenta con 667,957 instancias y 45 variables. Esta base de datos unificada se alojó en un servidor Azure SQL para facilitar el acceso. Inicialmente se tenía una base de delitos con 2,252,056 instancias y 16 variables, y una base de la encuesta con 96 instancias y 29 variables. Luego del preprocesamiento y unión se obtuvo la base integrada detallada en la Tabla 5.

Tabla 5 - Cantidad de variables e instancias

Base de datos		Delitos de Alto Impacto	Encuesta Multipropósito
Inicial	Q Variables	16	29
	Q Instancias	2,252,056	96
Preprocesada	Q Variables	16	29
	Q Instancias	1,997,844	95
Base de datos final	Q Variables	45	
	Q Instancias	667,957	

- Se realizó una separación aleatoria de los datos en conjuntos de entrenamiento (80%, 534,365 instancias) y prueba (20%, 133,592 instancias). Se guardaron en archivos CSV independientes para garantizar el uso de solo el entrenamiento en la creación de los modelos y la evaluación imparcial sobre los datos nuevos de prueba.

Análisis Descriptivo.

Realizamos análisis por medio de gráficas de serie de tiempo, mapas de calor, tendencias, verificando diferentes variables para identificar patrones. A continuación, detalles del proceso realizado:

La Ilustración 4 muestra la serie de tiempo del total de eventos delictivos por día entre enero 2021 y mayo 2023 en Bogotá, mientras que en la Ilustración 5 el eje ‘x’ representa el tiempo en escala mensual para visualizar la tendencia, mientras que el eje ‘y’ indica la cantidad de eventos delictivos agregados mensualmente. Se puede observar que hay una tendencia creciente en el número de delitos durante el periodo analizado, con picos pronunciados en los meses de mayo, junio y julio de cada año. Los meses con menos eventos delictivos son los de fin e inicio de año. En conclusión, los gráficos permiten analizar visualmente el comportamiento diario

y mensual del total de delitos en la ciudad durante la etapa post-pandemia, lo cual es útil para identificar meses críticos y tendencias para la toma de decisiones.

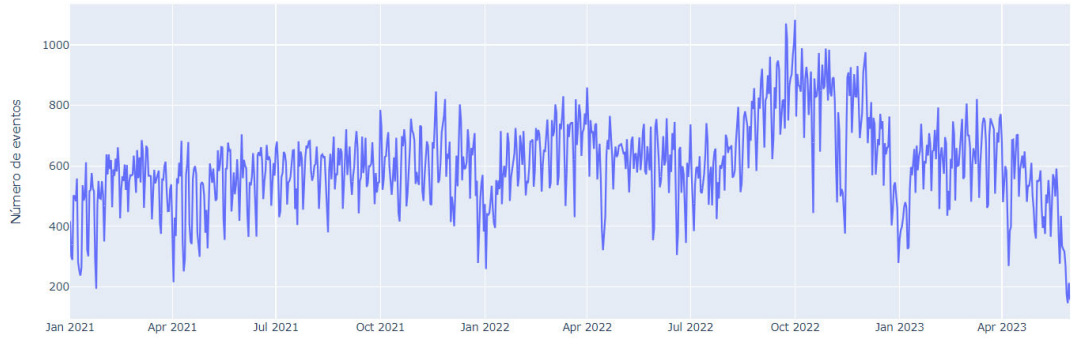

Ilustración 4 - Serie de tiempo de todos los eventos delictivos diarios (Eje x: Días)

Ilustración 5 - Serie de tiempo de todos los eventos delictivos mensuales (Eje x: Meses)

La Ilustración 6 evidencia el desequilibrio presente en la distribución de las clases de nuestra base de datos. Específicamente, se observa una alta concentración en las 4 categorías de delito más frecuentes, que en conjunto representan el 83.25% de los casos. Las clases restantes, a pesar de su diversidad, corresponden únicamente al 16.75% de los eventos registrados. Esta disparidad entre las clases mayoritarias y minoritarias es un reto importante para considerar en el modelado.

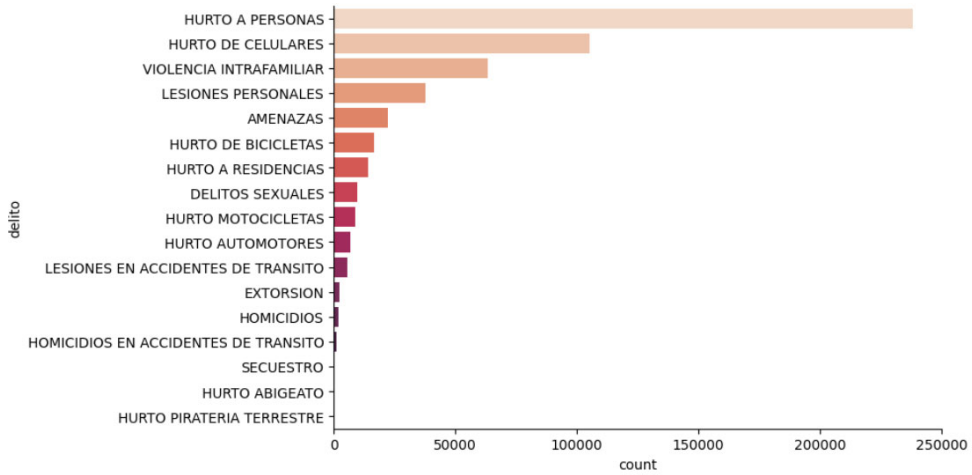

Ilustración 6 - Número de eventos por Delito

Bajo lo anterior y para un mejor análisis visual, decidimos crear agrupamientos correspondientes a los delitos más representativos para evidenciar su comportamiento en una serie de tiempo (Ilustración 7). Se puede evidenciar que el hurto a personas y el hurto a celulares presentan un comportamiento similar entre sí, distinto al de los demás delitos.

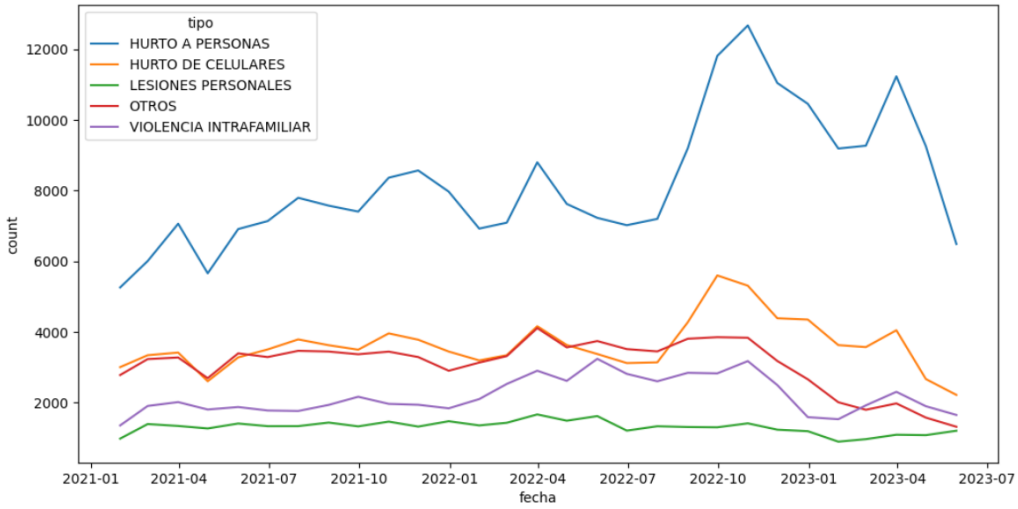

Ilustración 7 - Series de tiempo mensual por principales delitos

Posteriormente, iniciamos la verificación del comportamiento por variables, para identificar patrones y relaciones con mapas de calor (Ilustración 8). Se puede observar que la mayoría de los eventos delictivos, para todos los tipos, fueron sin empleo de armas.

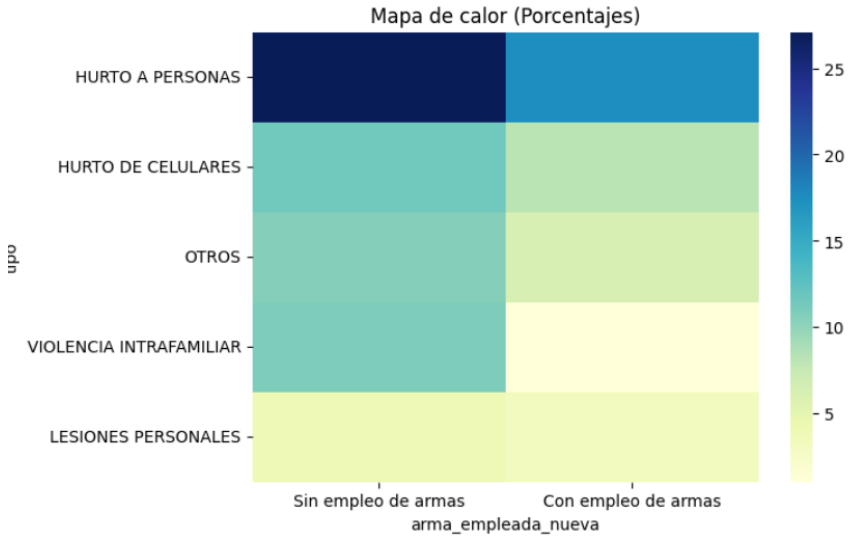

Ilustración 8 - Principales delitos y Empleo de arma

Realizamos un análisis entre variables de tiempo, en este caso día de la semana y rango del día, sobre cada uno de los grupos previamente mencionados Ilustración 9. El hurto a personas y el hurto a celulares presentan un comportamiento similar, los eventos delictivos tienden a ocurrir entre semana en la mañana y los fines de semana en la noche y madrugada. La violencia intrafamiliar tiende a ocurrir

cualquier día en la madrugada. Las lesiones personales ocurren con mayor frecuencia en la noche del sábado y madrugada del domingo.

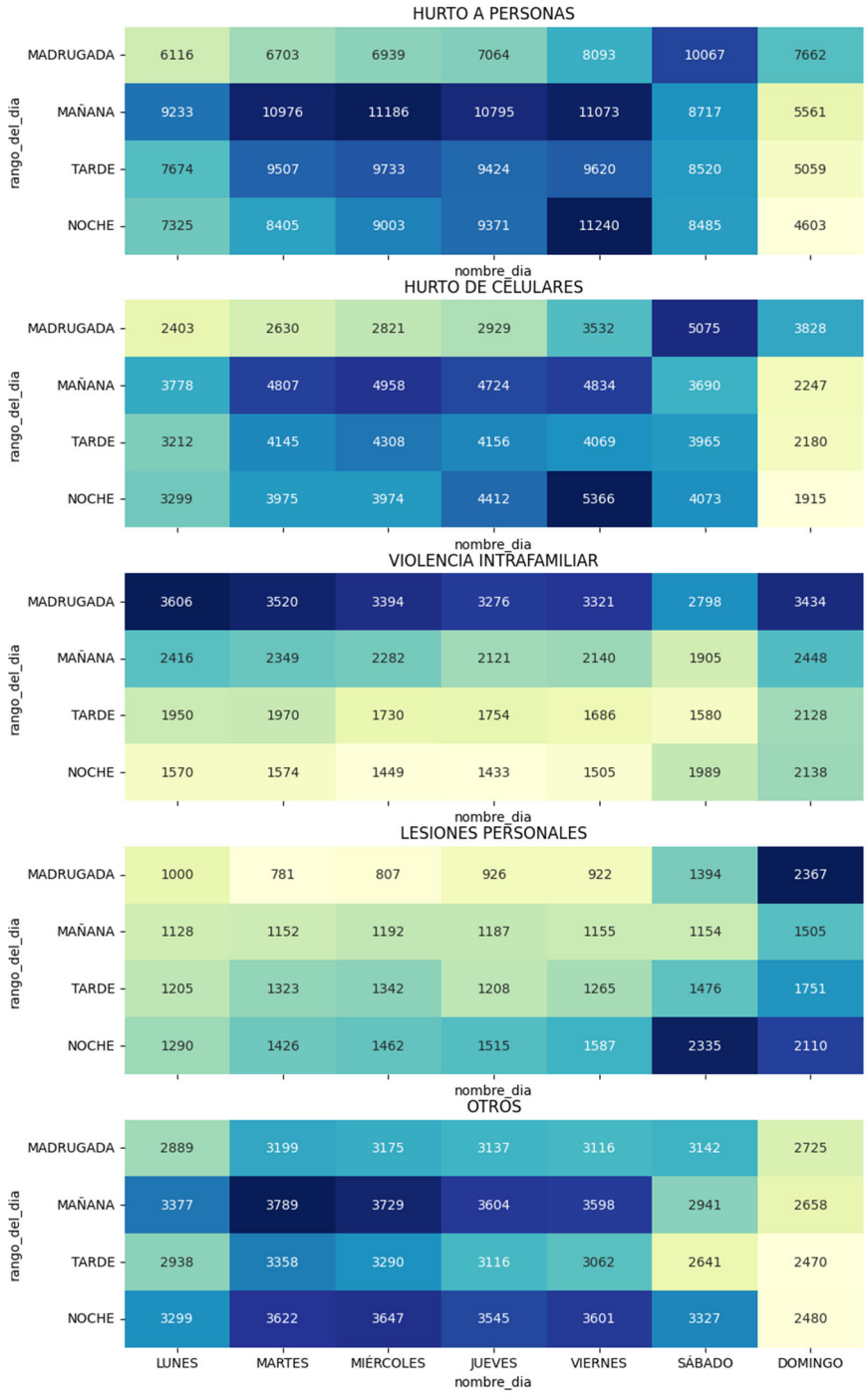

Ilustración 9 - Rango del día y día de la semana por principales delitos

Cuando relacionamos rango del día y empleo de arma, en la mayoría de los delitos se tiende a emplear arma en la noche (Ilustración 10).

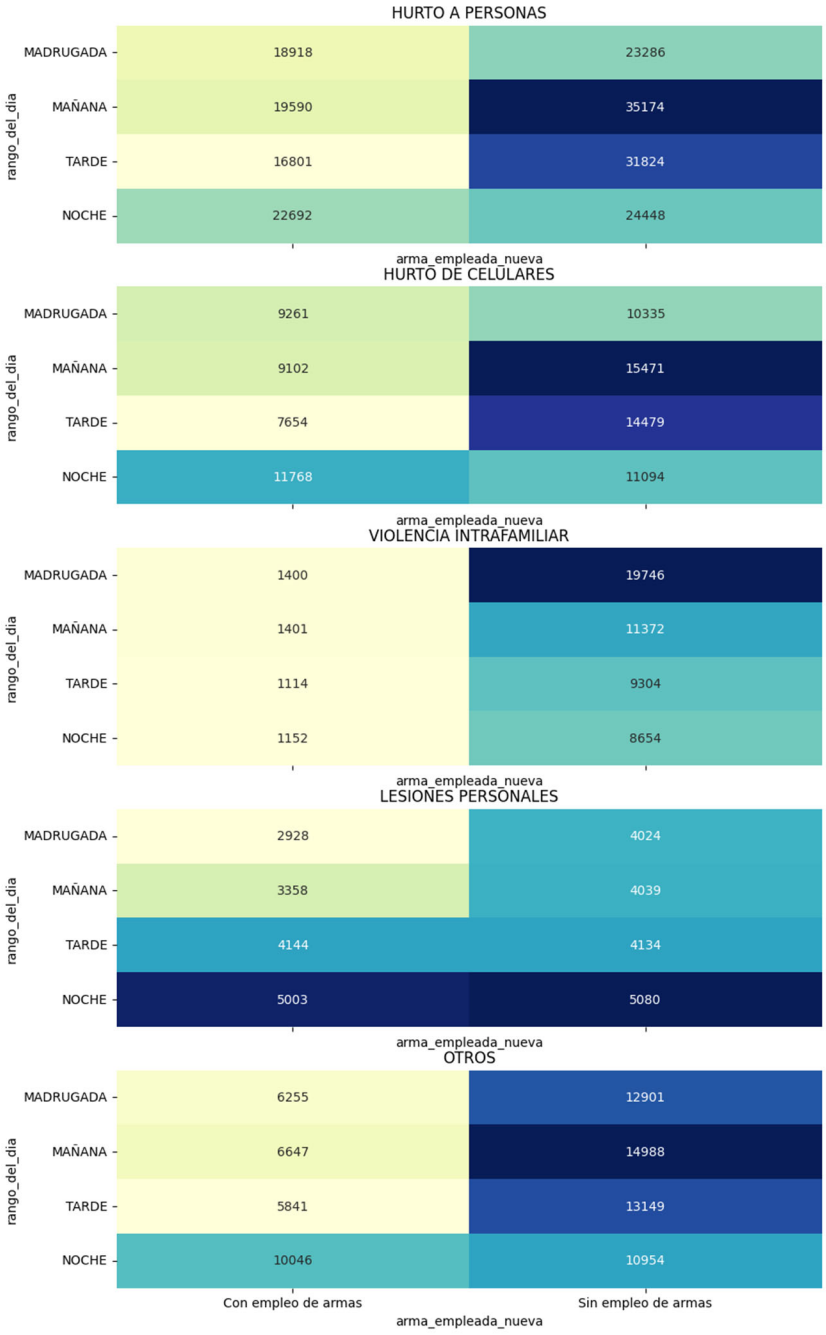

Ilustración 10 - Rango del día y empleo de arma por principales delitos

Como se menciona en las líneas anteriores concluimos para la parte de EXPLORE / EXPLORACIÓN:

- El proceso de limpieza y preparación de datos buscó eliminar registros inconsistentes e imputar valores faltantes para garantizar integridad y calidad de los datos analizados.
- El análisis visual de series de tiempo, mapas de calor y cruces entre variables busca identificar de forma eficiente patrones y relaciones en los datos antes de aplicar técnicas de modelado más complejas.
- La separación de los conjuntos de entrenamiento y prueba previene sesgos y permite evaluar la capacidad de generalización de los modelos sobre datos totalmente nuevos.

MODIFY / MODIFICACIÓN

En esta etapa de preprocesamiento se realizaron diversas transformaciones y preparaciones de los datos originales con el objetivo de obtener una representación adecuada para el posterior modelado.

Agrupamiento.

Inicialmente, contamos con 17 categorías diferentes de delitos. Para simplificar el análisis, se optó por reducir la granularidad agrupándolos en un conjunto menor de categorías. Posteriormente, cada nueva categoría resultante se modeló de forma independiente como un problema de clasificación binaria. El objetivo fue consolidar las 17 categorías en un conjunto reducido de grupos coherentes. Cada grupo se modeló de manera separada para facilitar el proceso de clasificación binaria. Para lograr este agrupamiento se exploraron diferentes enfoques:

Agrupamiento mediante K-means.

Este método realiza una clusterización de los datos en k grupos mediante un proceso iterativo que busca minimizar las distancias intra-cluster. Se probó el

algoritmo variando k entre 2 y 5 clusters. El número óptimo de grupos se determinó aplicando el método del codo sobre los datos transformados, el cual identifica un punto de inflexión en la suma de cuadrados dentro de cluster. Como se observa en la Ilustración 11 - Método del codo para la selección de número de clusters, el valor resultante fue k=2.

Para entrenar el modelo, se utilizaron como variables de entrada las categóricas binarizadas y las numéricas escaladas a distribución normal estándar. Se entrenó K-means con k=2. Al explorar las etiquetas de clase dentro de cada cluster, se encontró que en ambos grupos el delito más frecuente era el hurto a personas, representando más del 50% de los casos. Los demás delitos quedaron mezclados indiferenciadamente entre los clusters. En la Ilustración 12 se muestra la distribución de frecuencias para k=2, evidenciando la predominancia de hurtos a personas en ambos grupos.

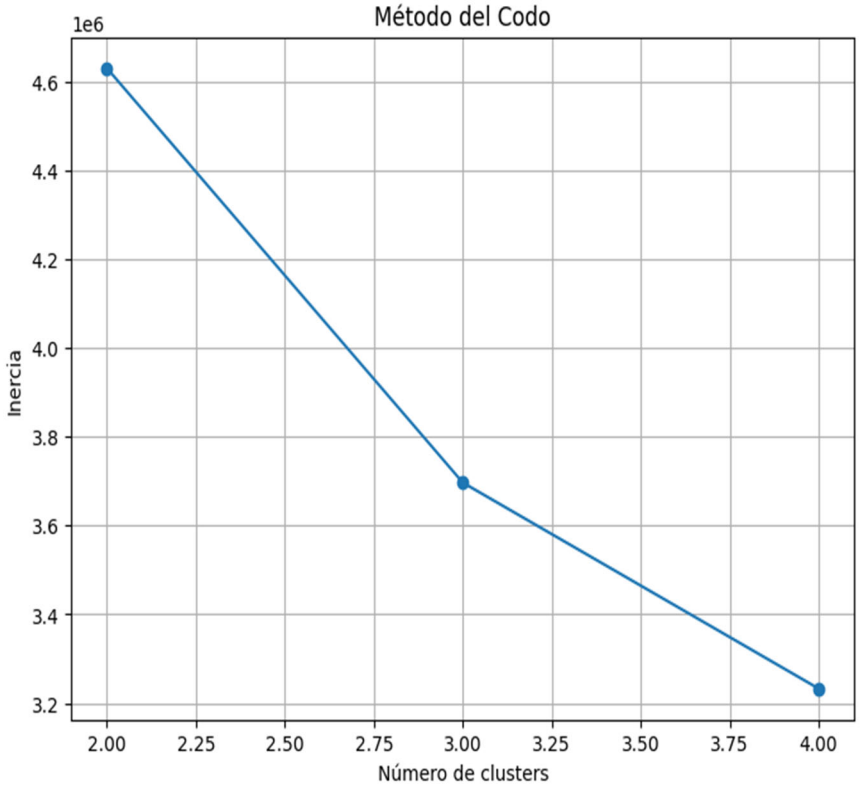

Ilustración 11 - Método del codo para la selección de número de clusters

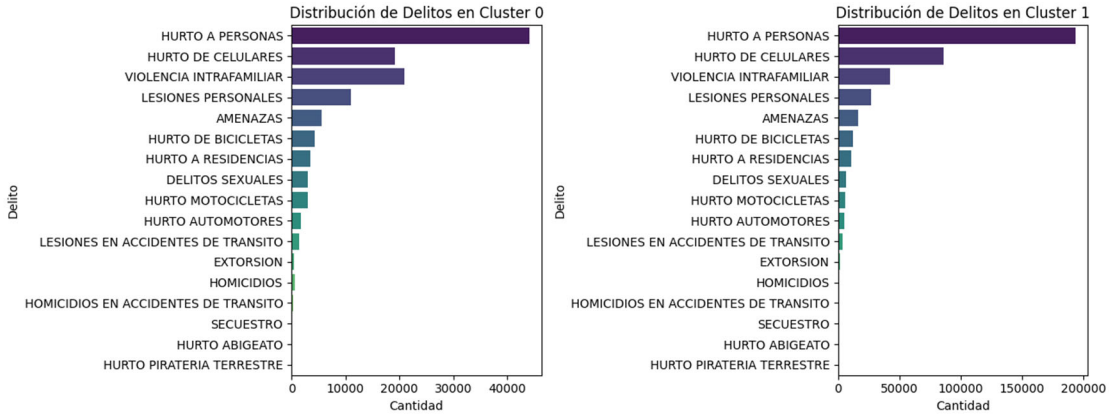

Ilustración 12 - Distribución de frecuencias en clusters por K-means

Adicionalmente, al analizar otras variables como tipo de arma, rango del día, día de la semana y mes, se obtuvo un perfil casi idéntico en los dos clusters como se evidencia en la Tabla 6 .

Tabla 6 - Características de clusters por K-means

Cluster	Delito más común	Arma más común	Rango del día más común	Día más común	Mes más común
0	Hurto a personas	Sin empleo de armas	Mañana	Viernes	Mar
1	Hurto a personas	Sin empleo de armas	Mañana	Viernes	Mar

Por lo tanto, el agrupamiento por K-means no logró distinguir patrones diferenciables entre los grupos resultantes, viéndose dominado por la clase mayoritaria de hurtos a personas.

Agrupamiento por frecuencias.

En este enfoque se consolidaron los delitos menos frecuentes y se dejaron las categorías mayoritarias en grupos separados. Siguiendo este criterio, se obtuvieron 4 grupos: hurtos a personas, hurtos de celulares, violencia intrafamiliar y un grupo de “otros delitos” (Ilustración 13). Sin embargo, el análisis preliminar de las series de tiempo ya había revelado un comportamiento muy similar entre hurtos a personas

y celulares. Por lo tanto, mantener estas categorías por separado no resultaría en grupos claramente distinguibles para la posterior clasificación binaria.

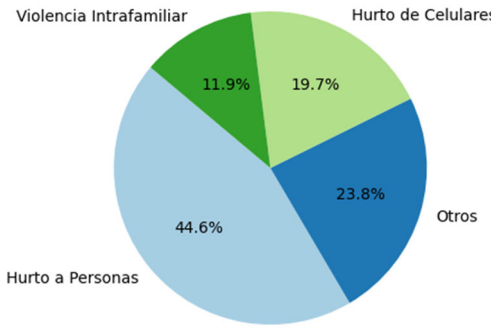

Ilustración 13 - Agrupamiento por frecuencia de delitos

Agrupamiento semántico.

Este método agrupó los delitos según su afinidad semántica y naturaleza en tres categorías principales: delitos contra la propiedad, delitos contra las personas y delitos de tránsito descritas en Tabla 7.

Tabla 7 - Agrupamiento semántico de delitos

Grupo	Delitos
Delitos contra la propiedad	Hurto a personas, hurto de celulares, hurto de bicicletas, hurto a residencias, hurto motocicletas, hurtos automotores, hurto abigeato, hurto piratería terrestre.
Delitos contra las personas	Violencia intrafamiliar, lesiones personales, amenazas, delitos sexuales, homicidios, extorsión, secuestro.
Delitos de tránsito	Lesiones en accidentes de tránsito, homicidios en accidentes de tránsito

El análisis visual de las distribuciones temporales, por variables categóricas y variables numéricas mostró comportamientos claramente diferenciados entre estos tres grupos consolidados. Por ejemplo, en Ilustración 14 - Rango del día y día de la semana por Grupos delictivos, cuando relacionamos el rango del día y el día de la semana, los tres grupos delictivos presentan comportamientos distintos. Los delitos contra las personas tienden a ocurrir todos los días en la madrugada y los fines de semana en la noche. Mientras que los delitos contra la propiedad muestran un

comportamiento diferente, pues suelen ocurrir principalmente entre semana, entre la mañana y la noche.

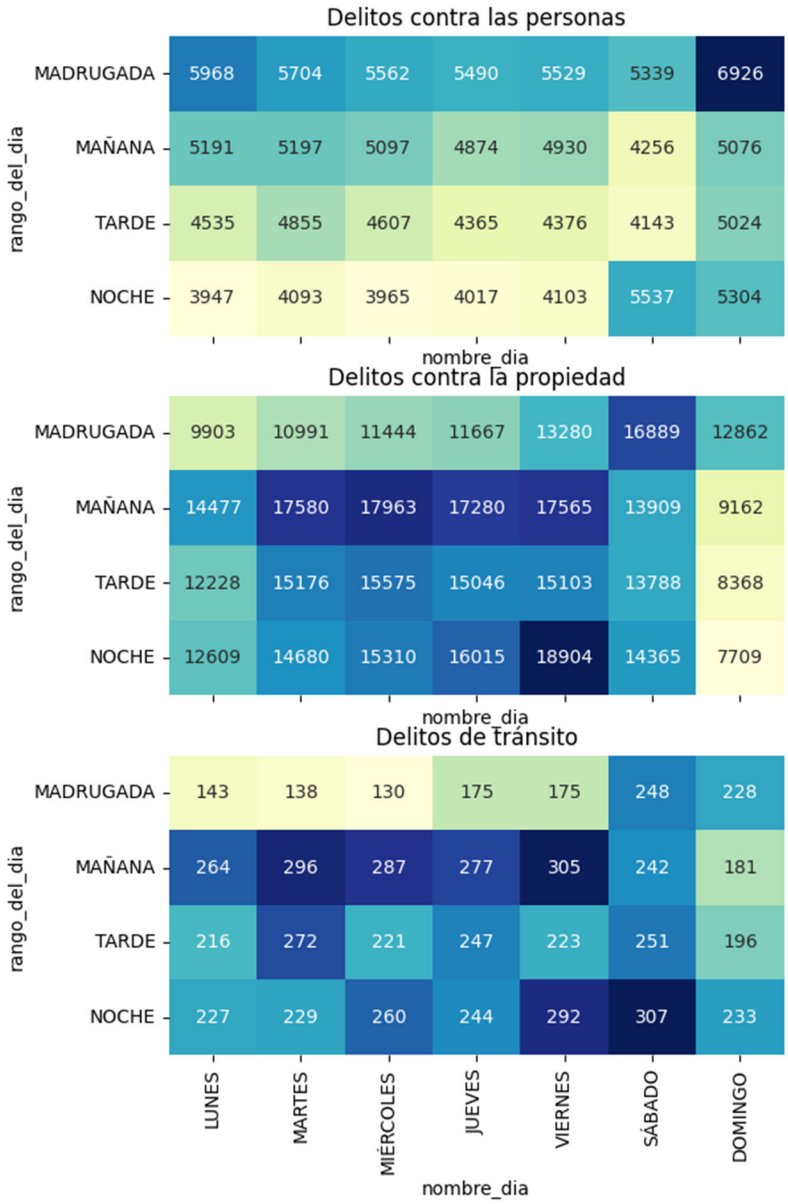

Ilustración 14 - Rango del día y día de la semana por Grupos delictivos

En cuanto al empleo de armas Ilustración 15 - Rango del dia y empleo de armas por grupos delictivos, si bien en los tres grupos predomina la ausencia de estas, al

relacionarlo con el rango del día, el comportamiento varió entre los grupos. Los delitos contra las personas por lo general se cometen sin empleo de armas durante el día. En los delitos contra la propiedad, tiende a haber empleo de armas en la noche, y los delitos de tránsito presentan un comportamiento variable.

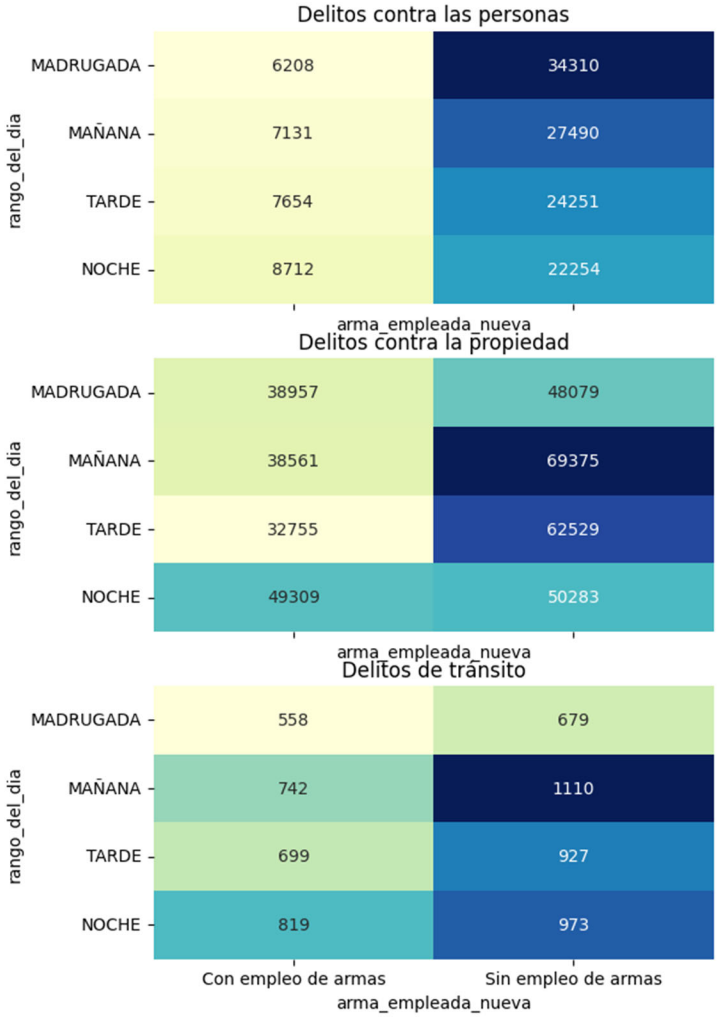

Ilustración 15 - Rango del dia y empleo de armas por grupos delictivos

Si bien la mayor parte de los delitos se concentra en los ingresos más bajos, esto es más evidente en los delitos contra las personas. Mientras que, en los ingresos más altos, los delitos contra la propiedad son los que ocurren con mayor frecuencia como se demuestra en Ilustración 16 - Distribución del Ingreso total por Grupos delictivos.

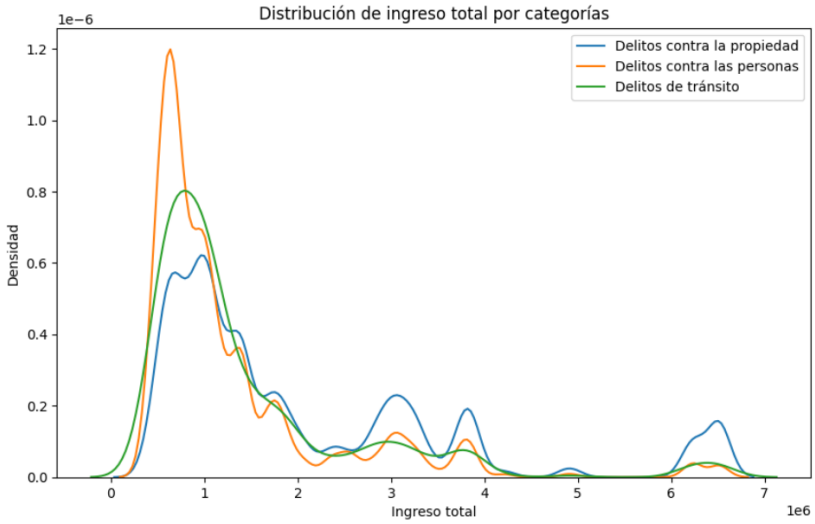

Ilustración 16 - Distribución del Ingreso total por Grupos delictivos

Se decidió utilizar la técnica de agrupamiento semántico para consolidar los 17 delitos iniciales en tres conglomerados principales. Este enfoque permitió agrupar delitos de naturaleza similar en cada conglomerado, los cuales, según el análisis visual de sus distribuciones, mostraban comportamientos diferenciados entre los tres grupos resultantes. Al mismo tiempo, al examinar las dinámicas individuales de los distintos delitos dentro de cada agrupación, se confirmó que compartían patrones temporales y socioeconómicos comunes. Además, mediante el método de la silueta se evaluó si estos 3 grupos estaban bien definidos, dio como resultado 0.004, un valor cercano a 0, esto sugiere que los grupos pueden superponer o que hay puntos que no están claramente asignados a un grupo en particular, sin embargo, no es cercano a -1 lo que indicaría que los grupos están mal definidos.

La distribución de los eventos delictivos se muestra en la Ilustración 17. En la siguiente sección se discutirá sobre la definición de las clases positivas y negativas de cada grupo y de cómo se manejó el desbalance de estas clases.

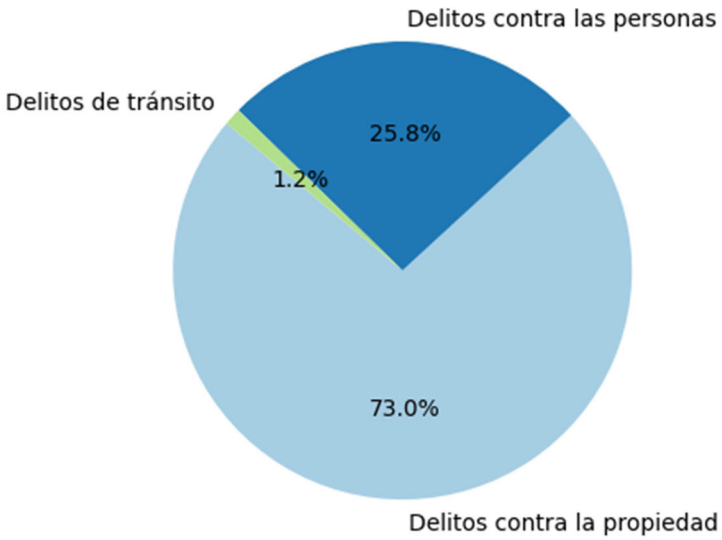

Ilustración 17 - Agrupamiento por similitud semántica

Preparación de los datos para la clasificación.

La clase positiva, al concentrarse en los delitos particulares de cada conglomerado, estaba claramente delineada. En contraste, la clase negativa, al reflejar la ausencia de ese grupo delictivo objetivo, abarcaba múltiples tipos de delitos no directamente relacionados con ese conglomerado.

Para mitigar esta disparidad y mejorar el desempeño del modelo, se evaluaron tres estrategias para conformar las clases negativas:

- **Casos de otros delitos:** Casos reales de otros delitos no asociados al conglomerado modelado. Sin embargo, el modelo podría enfocarse en patrones de esos delitos en lugar de la ausencia del grupo objetivo.
- **Casos sintéticos:** Casos sintéticos generados artificialmente a partir de combinaciones ausentes en los datos reales. Es técnica permite un mejor contraste entre la presencia y la ausencia del conglomerado delictivo. Pero, podría introducir ruido al modelo debido a la generación artificial de datos.
- **Enfoque híbrido**: Combinación de casos reales y sintéticos.

Tras evaluar los pros y contras de los enfoques, se optó por la estrategia híbrida para representar de forma más completa y equilibrada la ausencia de cada conglomerado delictivo modelado. Los delitos contra las personas y de tránsito tienen una clase negativa dominante, mientras que los delitos contra la propiedad tienen una clase positiva dominante.
Para abordar el equilibrio de las clases, se realizó una cuidadosa selección de una muestra representativa de la clase negativa, compuesta por ambas fuentes (casos reales y sintéticos), con el objetivo de igualar el tamaño de la clase positiva. Este enfoque permitió lograr una distribución más equilibrada de las clases y proporcionar al modelo una base de datos más equitativa para el entrenamiento.

Sin embargo, para el caso de los delitos contra la propiedad, no se dispone de suficientes casos negativos para igualar el número de casos positivos. Ya que el número de casos positivos es considerablemente mayor que el número de casos negativos disponibles. Ante esta situación, se tomó la decisión de utilizar todos los casos negativos disponibles, incluso si esto implicaba que el conjunto de datos todavía estaría ligeramente desbalanceado.

La prioridad era mantener una muestra lo más representativa posible de la clase negativa para evitar la pérdida de información relevante en el proceso de modelado. Aunque el conjunto de datos no esté completamente balanceado en esos casos, la utilización de todos los casos negativos disponibles contribuyó a proporcionar al modelo una mayor cantidad de información sobre la ausencia del delito objetivo.
No se utilizaron técnicas de sobre muestreo como SMOTE debido a que se emplearía validación cruzada para evaluar la capacidad de generalización del modelo y evitar el sobre ajuste. Aplicar SMOTE antes de la validación cruzada puede causar filtrado de información (information leakage), al duplicar datos que podrían aparecer tanto en el conjunto de entrenamiento como en el de validación. La solución es aplicar SMOTE dentro de cada iteración de la validación cruzada,

después de dividir los conjuntos de entrenamiento y validación, para evitar la contaminación entre ambos. Sin embargo, este enfoque requiere implementar el proceso completo de validación cruzada de forma manual, ya que librerías como sklearn no tienen incorporada esta funcionalidad.

Preparación de los datos para la predicción.

Para el problema de predicción de eventos delictivos diarios, el objetivo consistió en predecir la cantidad de delitos que ocurrirían cada día para cada uno de los grupos delictivos previamente definidos mediante el agrupamiento semántico.

Con este fin, se agregaron y colapsaron los datos originales calculando el recuento de eventos por cada combinación única de las variables de año, día de la semana, rango del día, sexo, arma empleada y localidad.

Se optó por utilizar la localidad y no la UPZ como nivel de agregación debido a que agrupar a nivel de UPZ resultaba en una distribución altamente concentrada de los recuentos en 1 evento diario. En cambio, la agregación por localidad generaba una distribución de frecuencias más dispersa en la variable objetivo de cantidad de eventos diarios, mejorando así el poder predictivo del modelo.

De esta manera, el procesamiento y transformación inicial de los datos permitió prepararlos adecuadamente para abordar la tarea específica de predicción de volumen diario de delitos para los conglomerados previamente identificados.

Creación bases de datos y almacenamiento.

Se procedió con la creación de bases de datos de entrenamiento (train) independientes para cada modelo, quedando de la siguiente manera:

- Clasificación:
  Delitos contra la propiedad: 38 variables, 582,155 instancias
  Delitos contra las personas: 38 variables, 618,991 instancias
  Delitos de tránsito: 38 variables, 678,883 instancias

- Predicción:
  Delitos contra la propiedad: 39 variables, 162,634 instancias
  Delitos contra las personas: 39 variables, 90,043 instancias
  Delitos de tránsito: 38 variables, 6,167 instancias

Estas seis bases de datos se almacenaron en archivos CSV independientes. Este procedimiento previene la pérdida de datos ante cualquier eventualidad en el entorno de trabajo, evitando la necesidad de re-ejecutar módulos previos. El almacenamiento independiente de las bases de entrenamiento facilita la reproducibilidad del proceso de modelamiento.

Como se menciona en las líneas anteriores concluimos para la parte de MODIFY / MODIFICACIÓN:

- La consolidación de categorías mediante agrupamiento semántico simplifica el problema al enfocarse en conglomerados de delitos relacionados con comportamientos diferenciables.
- La estrategia híbrida para conformar clases negativas equilibra casos reales y sintéticos para representar mejor la ausencia de cada conglomerado delictivo modelado.
- La agregación por localidad en los datos de predicción mejora la distribución de la variable objetivo de eventos diarios respecto a la agregación por UPZ, aumentando poder predictivo.

MODEL / MODELADO

Metodología de modelación y optimización.

En la etapa inicial del desarrollo de los modelos, se realizó una evaluación individual de varios algoritmos. El propósito de este proceso fue comparar diversas métricas de rendimiento y, en última instancia, seleccionar los mejores modelos para los problemas de clasificación y predicción. Posteriormente, a partir de los modelos que ofrecieron un rendimiento superior, se llevó a cabo una optimización de hiperparámetros y preprocesamiento de datos. Este proceso fue esencial para refinar los modelos seleccionados y mejorar su rendimiento.

Finalmente, se implementó una optimización automatizada utilizando la herramienta Naive AutoML (Mohr, 2021). Este enfoque permitió una exploración exhaustiva y eficiente de múltiples combinaciones de modelos, hiperparámetros y técnicas de preprocesamiento. Para el modelado de clasificación, la métrica de evaluación predominante fue el F1-Score. Esta métrica es apropiada en este contexto debido a su capacidad para equilibrar la precisión y el recall, que son métricas ampliamente utilizadas en problemas de clasificación binaria.

En el caso del modelado de regresión, la métrica de evaluación fue el Error Absoluto Medio (MAE). Esta métrica es especialmente adecuada cuando se desea tratar todos los errores de la misma manera, independientemente de su tamaño, lo que es coherente con los objetivos del modelado de predicción en este estudio.

A continuación, se describe en detalle cada una de las tres etapas de esta metodología de modelización y optimización.

Evaluación individual de algoritmos.

La primera etapa involucró una evaluación comparativa de diversos algoritmos de aprendizaje supervisado. Para el problema de clasificación binaria, se seleccionaron seis algoritmos: Naive Bayes (NB), Árboles de Decisión (DT), Análisis Discriminante Lineal (LDA), Regresión Logística (LR), KNeighbors Classifier y Random Forest

Classifier. Los primeros cinco se seleccionaron por su menor costo computacional y Random Forest debido a su prevalencia en el estado del arte.
Para la predicción de eventos diarios, se seleccionaron los siguientes algoritmos: LinearRegression, DecisionTreeRegressor, KNeighborsRegressor y RandomForestRegressor.
En esta fase inicial, se emplearon los hiperparámetros por defecto de cada técnica. Para evaluar la efectividad de cada algoritmo, implementamos dos funciones de evaluación esenciales. La primera, una función de validación cruzada, realizó una validación cruzada de 5 pliegues en los datos de entrenamiento para cada algoritmo. Esta validación proporcionó métricas de rendimiento valiosas: precisión, recall, F1-score y precisión para los problemas de clasificación; y RMSE, MAE y R2 para los problemas de predicción.

La segunda función de evaluación se aplicó al conjunto de prueba. Utilizó el modelo más prometedor derivado de la validación cruzada para proporcionar una evaluación integral de su rendimiento. En el caso de la clasificación, se generó una matriz de confusión y una curva ROC, junto con el correspondiente valor AUC. Para el problema de predicción, se elaboró un gráfico de dispersión que contrastaba las predicciones del modelo contra los valores reales. Esto proporcionó una visualización clara de la efectividad del modelo y su capacidad para generalizar a datos no vistos.

Optimización manual de hiperparámetros y creación de pipelines.
En la segunda etapa, se realizó una optimización de los hiperparámetros del modelo que tuvo el mejor rendimiento en la primera etapa. Paralelamente, se diseñaron pipelines que integraron etapas de preprocesamiento de datos y el modelo de aprendizaje automático.
Las técnicas de preprocesamiento incluyeron la normalización de los datos y la reducción de dimensionalidad a través del Análisis de Componentes Principales

(PCA). El proceso de optimización se llevó a cabo con la ayuda de para encontrar la mejor combinación de hiperparámetros para cada modelo.
Luego se hace una validación cruzada con la combinación de parámetros seleccionados por GridSearchCV y luego se evalúa en el conjunto de prueba

Optimización automatizada con Naive AutoML.
En la etapa final, se utilizó la optimización automatizada para explorar una gran cantidad de combinaciones de modelos, hiperparámetros y técnicas de preprocesamiento. Este proceso se llevó a cabo con la ayuda de una herramienta de Naive AutoML.

La implementación de la optimización automatizada permitió una exploración más amplia y eficiente de combinaciones de modelos y hiperparámetros en comparación con el enfoque manual. Los resultados obtenidos con Naive AutoML se compararon con el mejor pipeline logrado de forma manual para evidenciar las mejoras de rendimiento obtenidas a través de la automatización del proceso.
Debido al costo computacional de aplicar Naive AutoML a toda la base de datos, se realizaron pruebas iniciales sobre muestras aleatorias del 10%. Se ejecutó Naive AutoML en cada muestra, generando distintos pipelines. Al comparar resultados entre muestras no se encontraron diferencias significativas, por lo que se determinó utilizar solo una muestra, evitando mayores tiempos de ejecución.
La metodología final que se siguió con Naive AutoML tanto para los modelos de clasificación como predicción consistió en:

- Muestreo inicial: Se genera una muestra aleatoria estratificada de 10% de los datos disponibles, preservando la distribución de clases.
- Búsqueda de pipelines: Naive AutoML explora combinaciones de modelos, preprocesadores y parámetros sobre la muestra del 10%, buscando rápidamente un pipeline inicial prometedor.

- Validación entrenamiento: El pipeline prometedor se valida sobre los datos de entrenamiento mediante validación cruzada.
- Entrenamiento: El mejor pipeline encontrado se entrena sobre el 80% restante de los datos, permitiendo un mejor ajuste de parámetros.
- Validación prueba: El pipeline entrenado se valida en un 20% de datos reservados exclusivamente para evaluar de forma justa el desempeño final del modelo.
- Análisis: Se realiza un análisis detallado del pipeline seleccionado, su desempeño y posibles mejoras.

En resumen, la metodología de modelización y optimización siguió una secuencia de evaluación inicial de algoritmos, optimización manual de hiperparámetros y creación de pipelines, y finalmente una optimización automatizada con Naive AutoML. Esta secuencia se aplicó tanto al problema de clasificación binaria como a la predicción de eventos diarios.

- La validación cruzada y el conjunto de prueba separado permiten estimar la capacidad de generalización y evitar overfitting durante la evaluación de modelos. En otras palabras, estas técnicas proveen estimaciones confiables del desempeño real de los modelos.
- La optimización manual mediante GridSearch identifica eficientemente mejores configuraciones de hiperparámetros para los modelos iniciales. Es decir, GridSearch explora de manera efectiva distintas combinaciones de parámetros.
- El uso de Naive AutoML permite una exploración más amplia y automatizada de combinaciones de modelos, parámetros y técnicas de procesamiento superando los resultados manuales. En otras palabras, Naive AutoML automatiza y optimiza la búsqueda de mejores pipelines.

- El uso de muestras iniciales reduce la carga computacional de Naive AutoML antes de entrenar sobre el conjunto completo de datos. Dicho de otro modo, las muestras disminuyen el costo computacional en la etapa inicial de búsqueda.

ASSESS / EVALUACIÓN

La evaluación rigurosa de los modelos desarrollados constituye una etapa clave dentro de la metodología aplicada. Se utilizaron métricas específicas para cada problema:

Clasificación: Precision, Recall, F1-Score, Matriz de confusión, Curva ROC y AUC. El F1-Score proporciona un balance entre precisión y recall útil en problemas desbalanceados.

Regresión: Error cuadrático medio (MSE), Error absoluto medio (MAE) y coeficiente de determinación R2. El MAE resulta apropiado al ponderar errores grandes y pequeños equitativamente.

Para estimar la capacidad de generalización y evitar overfitting, se emplearon técnicas como validación cruzada sobre el conjunto de entrenamiento y evaluación sobre un conjunto de prueba separado.

Sobre los modelos iniciales más prometedores, se realizó una optimización manual de hiperparámetros mediante GridSearch. Este método evalúa exhaustivamente distintas configuraciones, identificando la de mejor desempeño.

Asimismo, se ensamblaron pipelines que integraban técnicas de preprocesamiento como normalización y PCA con los algoritmos de aprendizaje automático.

Posteriormente, se aplicó optimización automatizada mediante AutoML para acelerar la búsqueda de mejores modelos, superando los resultados del proceso manual. Se utilizaron muestras del 10% de los datos para reducir la carga computacional antes de entrenar sobre el conjunto completo.

En conclusión, la evaluación sistemática, optimización de hiperparámetros y uso de pipelines y AutoML resultaron esenciales para desarrollar modelos precisos y robustos, tanto para la clasificación como la predicción de eventos delictivos.

- Las métricas específicas seleccionadas como F1-Score y MAE son adecuadas para evaluar los problemas de clasificación y regresión respectivamente.
- La optimización de hiperparámetros mediante GridSearch y el uso de pipelines y Naive AutoML buscan mejorar el desempeño y robustez de los modelos desarrollados.
- Las técnicas como validación cruzada y evaluación en conjunto de prueba separado previenen overfitting y proveen estimados confiables del desempeño real.

RESULTADOS

En esta sección, se exponen los resultados alcanzados durante el proceso de análisis y modelado en el contexto del problema de predicción de delitos. Se llevaron a cabo diversas evaluaciones y comparaciones de diferentes algoritmos de clasificación y técnicas de preprocesamiento para enfrentar el desafío de clasificar delitos en diferentes categorías y predecir eventos delictivos diarios. El objetivo principal fue identificar el modelo óptimo que proporcionara un equilibrio entre precisión y recall para las clases de interés en el caso de clasificación, y con el menor error en el caso de predicción.

En primer lugar, se describen los modelos iniciales y se presentan las métricas de desempeño obtenidas para cada uno de ellos. Luego, se analizan detalladamente los resultados de los mejores modelos, considerando las métricas correspondientes, con el fin de identificar el enfoque más prometedor en términos de capacidad predictiva.

A continuación, se explora el proceso de optimización de hiperparámetros realizado para mejorar el rendimiento del modelo seleccionado. Se detalla cómo se llevó a cabo la búsqueda de la mejor combinación de hiperparámetros y se comparan los resultados con los obtenidos mediante la técnica de Naive AutoML.

Finalmente, se presentan tablas de comparación que resumen los resultados finales de los modelos evaluados, centrándose en métricas clave para la toma de decisiones. Se discuten las fortalezas y debilidades de cada enfoque y se proporciona una visión general de los modelos recomendados para abordar el problema de predicción de delitos.

Los hallazgos presentados en esta sección brindan una visión integral del proceso de modelado y orientan hacia los enfoques más adecuados para la clasificación y predicción de delitos. Estos resultados son fundamentales para respaldar la toma de decisiones y la implementación de soluciones efectivas en la prevención y gestión de la delincuencia.

Los resultados de los modelos desarrollados fueron implementados en Google Colab, al cual se puede acceder a través del link descrito en Anexo 2Anexo 2

RESULTADOS DE LA CLASIFICACIÓN BINARIA

Delitos de tránsito

Comparación de modelos Iniciales.

El Accuracy se ubica entre 0.59 y 0.63. Al no haber una clase mayoritaria que beneficie esta métrica, la precisión de clasificación general es menor.

Tabla 8 - Evaluación individual de algoritmos clasificación - delitos de tránsito

Modelo	Accuracy	Precision	Recall	F1	Tiempo (s)
GaussianNB	0.587	0.584	0.605	0.594	0.08
DecisionTreeClassifier	0.584	0.582	0.59	0.586	0.527
LinearDiscriminantAnalysis	0.6	0.596	0.626	0.61	0.893
LogisticRegression	0.6	0.596	0.623	0.609	1.139
KNeighborsClassifier	0.591	0.576	0.693	0.629	1.915
RandomForestClassifier	0.629	0.629	0.635	0.631	8.23

La precisión y recall están muy cercanas, con valores similares alrededor de 0.58 a 0.63. Indica capacidad equilibrada de detectar ambas clases. El F1 score se mantiene moderado, entre 0.59 y 0.63, siguiendo los valores de precisión y recall.

Evaluación de los mejores modelos.

En este caso, al considerar la métrica F1, el mejor modelo es RandomForestClassifier con un F1 de 0.631. Este modelo también tiene la precisión y recall más altos (0.629 y 0.635 respectivamente), lo que indica que el modelo tiene un buen equilibrio en términos de minimizar tanto los falsos positivos (precision) como los falsos negativos (recall). El tiempo de ejecución para este modelo es más alto (8.23s), pero todavía es bastante aceptable comparado con los tiempos de los demás modelos.

El segundo mejor modelo en términos de la métrica F1 es KNeighborsClassifier con un F1 de 0.629, pero con un tiempo de ejecución mucho mayor (1.915s) y una precisión y recall más bajos (0.576 y 0.693, respectivamente).

Entonces, dado este conjunto de datos equilibrado y basándonos en la métrica F1, el modelo óptimo sería RandomForestClassifier. Sin embargo, si el tiempo de ejecución es una preocupación y se prefiere un modelo más rápido con un rendimiento ligeramente peor, el modelo LinearDiscriminantAnalysis o LogisticRegression podrían ser considerados.

Resultados de la optimización de hiperparámetros y preprocesamiento.

Se realizó una optimización de hiperparámetros sobre Random Forest mediante GridSearch, buscando mejorar su capacidad predictiva. Se probaron distintas configuraciones para hiperparámetros clave como n_estimators, max_features, max_depth, min_samples y técnicas de preprocesamiento de variables.

La optimización identificó la mejor combinación, que incluía escalado MinMax de numéricas, PCA a 0.9 componentes, one-hot encoding de categóricas y 100 estimadores de Random Forest. Esta configuración superó el desempeño del modelo inicial sin ajustes. El modelo muestra una precisión y recall aceptables en la validación cruzada, no se desempeña tan bien en el conjunto de prueba, lo que indica que el modelo puede estar sobreajustado a los datos de entrenamiento y no generaliza bien a nuevos datos (Ilustración 18).

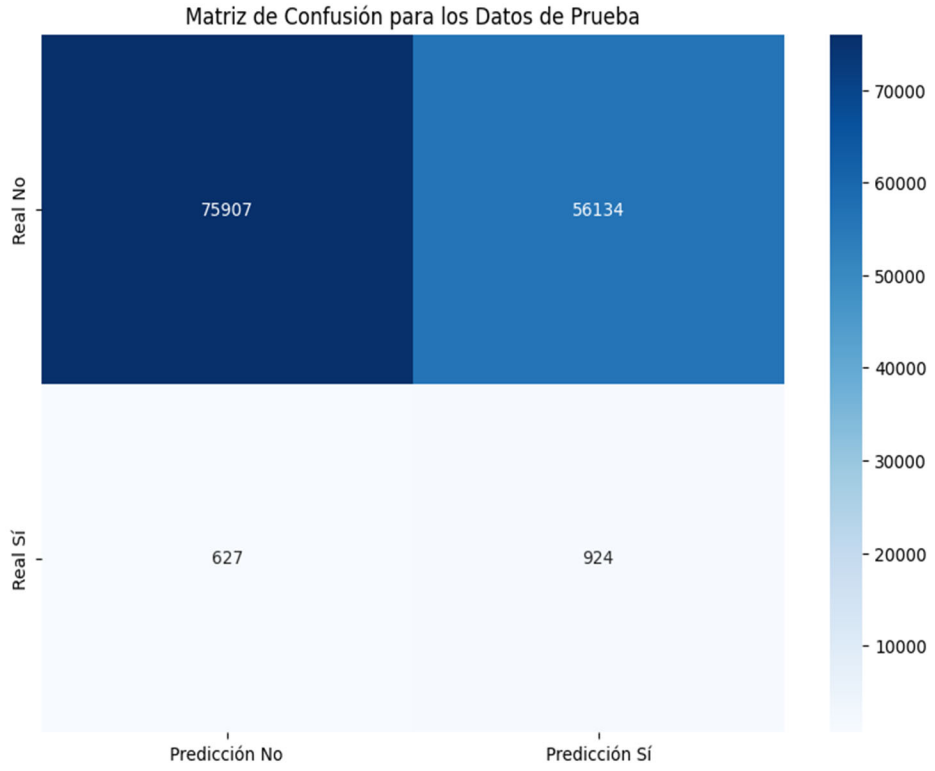

Ilustración 18 - Matriz de confusión Delitos de tránsito (GridSearchCV)

a. Fase de búsqueda (10% de datos).

Tabla 9 - Resultados delitos de tránsito, fase de búsqueda 10% de datos

Modelo	Preprocesador	F1 Score	Precisión	Recall
QuadraticDiscriminantAnalysis	None	0.6234	0.5204	0.7839
KNeighborsClassifier	None	0.5248	0.5248	0.5256
BernoulliNB	None	0.5295	0.5453	0.516

El mejor pipeline fue QuadraticDiscriminantAnalysis sin preprocesamiento, con F1 score de 0.6234.

b. Fase de entrenamiento (80% de datos) y de validación (20% de datos).

Tabla 10 - Resultados delitos de tránsito, fase de entrenamiento y validación

Métrica	Entrenamiento 80%	Validación 20%
Precisión	0.5407	0.98
Recall	0.804	0.3
F1-score	0.646	0.44

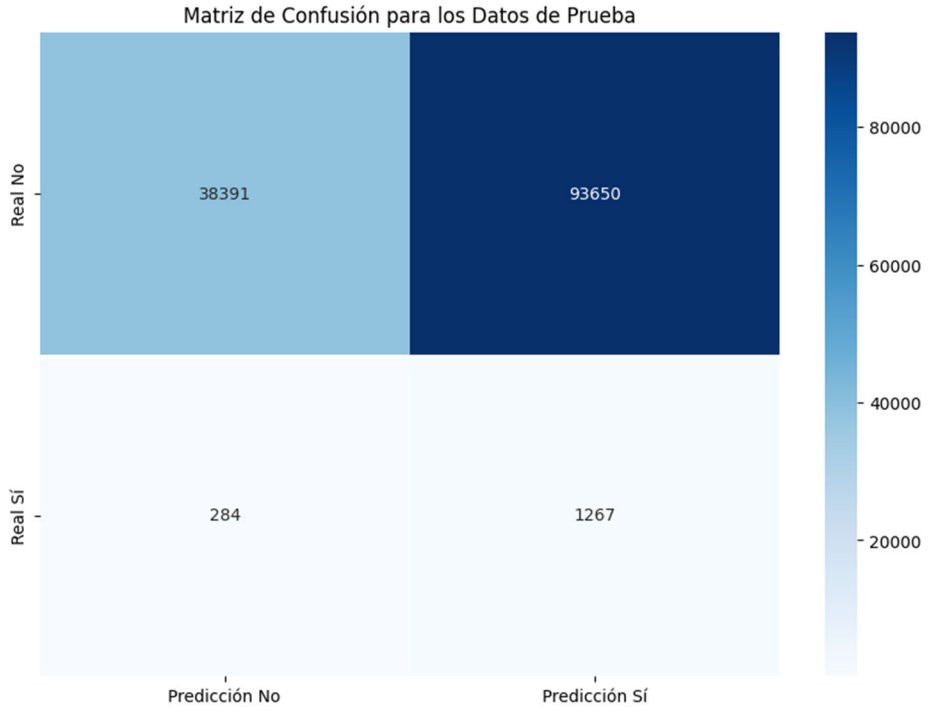

Ilustración 19 – Matriz de confusión Datos de prueba. Naive AutoML Delitos de tránsito

Al comparar la optimización de hiperparámetros con Naive AutoML, Tabla 11 - Delitos de tránsito (Comparación final) y Tabla 12 – Delitos de tránsito (Comparación final por clases), se puede concluir que:

- Rendimiento del modelo en validación cruzada: Según los resultados de la validación cruzada, el método NaiveAutoML supera a la Optimización de Hiperparámetros en términos de la métrica F1. El método NaiveAutoML también tiene un tiempo de ejecución significativamente más corto.
- Rendimiento del modelo en el conjunto de prueba: Sin embargo, en el conjunto de pruebas, el modelo obtenido con Optimización de Hiperparámetros muestra un mejor rendimiento en términos de precisión (accuracy). Sin embargo, el modelo AutoML presenta un recall más alto para la clase 1, lo que significa que puede identificar una mayor proporción de las instancias positivas.

- Precisión y recall de la clase minoritaria: Es notable que ambos modelos tienen una precisión muy baja pero un recall relativamente alto para la clase 1 en el conjunto de prueba. Esto indica que los modelos están identificando correctamente una buena proporción de los casos positivos (alta recall), pero al mismo tiempo están clasificando incorrectamente muchos casos negativos como positivos (baja precisión).
- Balance de las clases: A pesar de tener un recall alto para la clase 1, la precisión muy baja indica que los modelos pueden estar dando muchos falsos positivos. Esto puede ser un problema si las predicciones incorrectas tienen un alto costo.

Tabla 11 - Delitos de tránsito (Comparación final)

Método	Accuracy	Precision	Recall	F1-score	Tiempo (s)
Optimización Hiperparámetros	0.609	0.608	0.614	0.611	458.88
Naive AutoML	0.561	0.541	0.804	0.646	0.83

Tabla 12 – Delitos de tránsito (Comparación final por clases)

Método	Accuracy	Precision (Clase 0)	Recall (Clase 0)	F1-score (Clase 0)	Precision (Clase 1)	Recall (Clase 1)	F1-score (Clase 1)
Optimización Hiperparámetros	0.58	0.99	0.57	0.73	0.02	0.6	0.03
Naive AutoML	0.3	0.99	0.29	0.45	0.01	0.82	0.03

Delitos Contra las personas

Comparación de modelos iniciales.

El Accuracy se ubica entre 0.688 y 0.712 para los distintos modelos. Esto se explica porque al haber la misma proporción de casos positivos y negativos, la precisión de la clasificación mejora. La precisión y recall están equilibrados, ambas con valores

entre 0.67 y 0.72 para todos los algoritmos. Al no haber un sesgo por clase mayoritaria, los modelos logran detectar mejor los verdaderos positivos y tienen menos falsos positivos (ver Tabla 8).

Tabla 13 – Evaluación individual de algoritmos- delitos contra las personas

Modelo	Accuracy	Precision	Recall	F1	Tiempo (s)
GaussianNB	0.682	0.710	0.628	0.665	1.252
DecisionTreeClassifier	0.688	0.701	0.669	0.682	8.191
LinearDiscriminantAnalysis	0.695	0.691	0.726	0.706	7.630
LogisticRegression	0.696	0.694	0.720	0.705	7.442
KNeighborsClassifier	0.689	0.674	0.746	0.707	194.036
RandomForestClassifier	0.712	0.720	0.720	0.716	163.333

Como consecuencia, el F1 score se ubica entre 0.70 y 0.72 para la mayoría de las modelos. Al haber distribución balanceada, la media armónica entre precisión y recall es mayor.

Evaluación de los mejores modelos.

En este caso, dado que las clases están equilibradas, podemos dar más importancia a la métrica de precisión en adición al recall y F1-score. RandomForestClassifier aparece como el mejor modelo basándose en la puntuación F1, con una puntuación de 0.716. Este modelo también tiene la precisión y recall más altos (ambos 0.720), lo que significa que el modelo tiene un buen equilibrio en términos de minimizar tanto los falsos positivos (precision) como los falsos negativos (recall).

Aunque los modelos de LinearDiscriminantAnalysis, LogisticRegression y KNeighborsClassifier tienen puntuaciones F1 cercanas a RandomForestClassifier, este último tiene ligeramente mejores resultados en todas las métricas. Además, aunque RandomForestClassifier tiene un tiempo de ejecución mayor que algunos

otros modelos (163.33s), es significativamente más rápido que KNeighborsClassifier, que es el siguiente mejor modelo en términos de F1-score. Por lo tanto, de acuerdo con estas métricas y considerando tanto el rendimiento del modelo como el tiempo de ejecución, RandomForestClassifier parece ser el modelo óptimo para este conjunto de datos equilibrado.

Resultados de la optimización de hiperparámetros y preprocesamiento.

Se realizó una optimización de hiperparámetros sobre Random Forest mediante GridSearch, buscando mejorar su capacidad predictiva. Se probaron distintas configuraciones para hiperparámetros clave como n_estimators, max_features, max_depth, min_samples y técnicas de preprocesamiento de variables.

Los parámetros óptimos encontrados son los siguientes: 'max_depth': None, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100.

Evaluando este modelo en el conjunto de prueba, obtenemos la matriz de confusión (Ilustración 20).

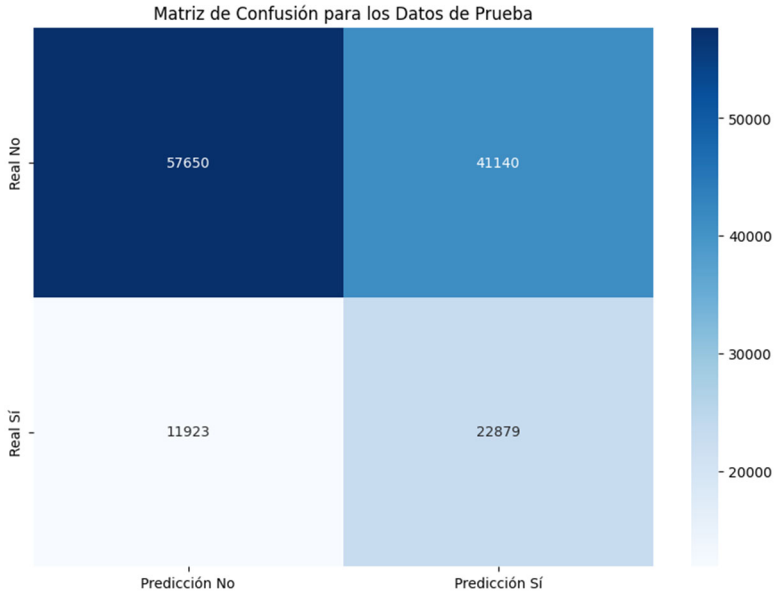

Ilustración 20 – Matriz de confusión para los datos de prueba- Gridsearch. Delitos contra las personas

En términos de rendimiento, el modelo optimizado tiene un buen equilibrio entre precisión (0.696) y Recall (0.716), es decir, el modelo no solo tiene una buena capacidad para predecir correctamente los delitos, sino que también es capaz de capturar una gran proporción de los delitos reales.

A partir de los resultados de prueba, el modelo no mantiene el rendimiento que se observó durante la validación cruzada. La precisión de la clase positiva es baja (0.36) lo que indica que el modelo está teniendo dificultad para predecir correctamente esta clase. Como se muestra en Ilustración 20, se está identificando una buena parte de las instancias positivas, pero a costa de hacer muchas predicciones positivas incorrectas.
El valor AUC-ROC obtenido para el modelo Random Forest en este caso es de 0.67. Ese valor indica que el modelo tiene una capacidad decente para distinguir entre clases positivas y negativas.

Resultados de la automatización con Naive AutoML.
A continuación, presentamos los resultados obtenidos en la fase de búsqueda, entrenamiento y validación.

a. Fase de búsqueda (10% de datos)

El mejor pipeline encontrado en la muestra del 10% fue GradientBoostingClassifier con QuantileTransformer, con F1 score de 0.7166 (ver Tabla 15 Tabla 1).

Tabla 14 - Resultados delitos contra personas, fase de búsqueda 10% de datos

Modelo	Preprocesador	F1 Score	Precisión	Recall
GradientBoostingClassifier	QuantileTransformer	0.7166	0.7238	0.7098
GradientBoostingClassifier	None	0.7163	0.7224	0.7101
RandomForestClassifier	None	0.6979	0.7081	0.6882
KNeighborsClassifier	None	0.6851	0.6787	0.6916

El pipeline con RandomForestClassifier tuvo un buen desempeño con F1 score de 0.6979, solo ligeramente por debajo del mejor modelo Podría ser un buen candidato para explorar en más detalle. Algunos pipelines como SVC y MLPClassifier tuvieron tiempos fuera o excepciones durante la búsqueda, por lo que no se pudieron evaluar completamente.

b. Fase de entrenamiento (80% de datos) y de validación (20% de datos)

Tabla 15 - Resultados delitos contra personas, fase de entrenamiento y validación

Métrica	**Entrenamiento 80%**	**Validación 20%**
Precisión	0.731	0.74
Recall	0.731	0.65
F1-score	0.728	0.67

Según las métricas en la Tabla 15 este modelo tiene un buen equilibrio entre precisión y recall, lo que indica que es bueno para identificar la clase positiva sin incurrir en demasiados falsos positivos o falsos negativos.

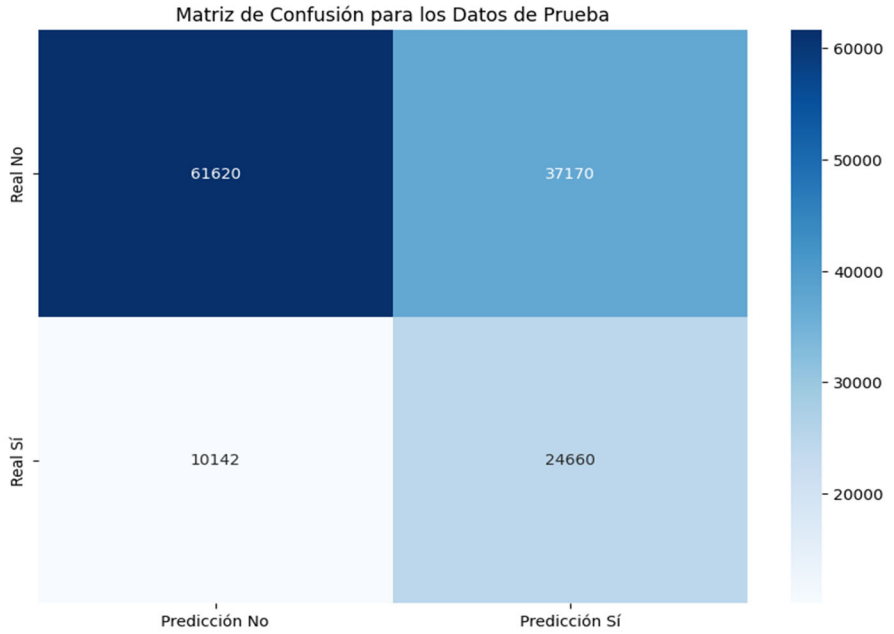

Ilustración 21 – Matriz de confusión para los Datos de prueba. Modelo Naive AutoML Delitos contra las personas

En el conjunto de pruebas, el rendimiento del modelo disminuye un poco en comparación con la validación cruzada. Esto podría indicar que el modelo está ligeramente sobre ajustado a los datos de entrenamiento. Sin embargo, aún supera a otros modelos en términos de rendimiento general (Ilustración 21).

Conclusión.

- **Precisión**: El modelo Gradient Boosting con QuantileTransformer presenta un mejor rendimiento en la precisión en los datos de validación cruzada, así como también en los datos de prueba.
- **Recall**: Ambos modelos tienen valores de recall muy similares tanto en los datos de validación cruzada como en los de prueba. Sin embargo, Gradient Boosting tiene un ligero margen.
- **F1 Score**: De nuevo, el modelo Gradient Boosting con QuantileTransformer tiene un mejor F1 Score en los datos de validación cruzada y en los datos de prueba.
- **Tiempo de entrenamiento**: Aunque Gradient Boosting presenta un rendimiento ligeramente mejor, también es notablemente más lento para entrenar comparado con Random Forest con ajuste de hiperparámetros.

Delitos contra la propiedad

Comparación de modelos iniciales.

Delitos contra la propiedad tiene una distribución de clases 67% positiva y 33% negativa, como se mencionó anteriormente, esta categoría no se balanceo. Sin embargo, el Accuracy es moderado a alto para los distintos modelos, entre 0.645 y 0.720. Se explica por la mayor proporción de casos positivos, que benefician esta métrica.

Tabla 16 - Evaluación individual de algoritmos - delitos contra la propiedad

Modelo	Accuracy	Precision	Recall	F1	Tiempo (s)
GaussianNB	0.646	0.741	0.727	0.734	2.33
DecisionTreeClassifier	0.709	0.769	0.808	0.788	26.56
LinearDiscriminantAnalysis	0.7	0.719	0.906	0.802	14.26
LogisticRegression	0.7	0.718	0.908	0.802	10.37
KNeighborsClassifier	0.709	0.729	0.9	0.806	842.33
RandomForestClassifier	0.72	0.759	0.853	0.803	333.17

La precisión es alta, entre 0.718 y 0.769. Menos falsos positivos al acertar más en la clase mayoritaria real. El recall es muy alto, entre 0.726 y 0.908. Los modelos detectan correctamente la mayoría de los casos positivos. Como consecuencia, el F1 score es alto también, entre 0.733 y 0.808. En general, el desbalance beneficia métricas como recall y F1 score, pero sobreestima la precisión real.

Evaluación de los mejores modelos.

Según la puntuación F1, el RandomForestClassifier parece ser el mejor modelo con una puntuación de 0.803. Esto sugiere que tiene un buen equilibrio entre precisión y sensibilidad. Es decir, tiene capacidad para identificar correctamente los casos positivos y los casos negativos.

Sin embargo, también es importante considerar el tiempo de entrenamiento del modelo. El KNeighborsClassifier, por ejemplo, tiene una puntuación F1 similar a la del RandomForestClassifier, pero tarda mucho más en entrenarse (842.33s vs 333.17s). Por lo que, RandomForestClassifier, según el f1 score y el tiempo de entrenamiento parece ser el mejor modelo.

Resultados de la optimización de hiperparámetros y preprocesamiento

Se realizó una optimización de hiperparámetros sobre Random Forest mediante GridSearch, buscando mejorar su capacidad predictiva. Se probaron distintas configuraciones para hiperparámetros clave como n_estimators, max_features, max_depth, min_samples y técnicas de preprocesamiento de variables. Se utilizó

como preprocesamiento MinMaxScaler y n_estimators igual a 10. El modelo optimizado tiene un F1 score de 0.7615 en la validación cruzada, indica que el modelo tiene un buen equilibrio entre precisión y recall. En el conjunto de prueba, la precisión del modelo para la clase positiva (1) es de 0.79, y el recall es de 0.78, lo que da como resultado un F1 score de 0.78. Aunque estos resultados son más bajos que los obtenidos en la validación cruzada, todavía representan un buen rendimiento El modelo es bueno identificando los delitos, pero tiene dificultades para identificar correctamente cuando un delito no ocurrirá, esto se puede ver claramente en la matriz de confusión ().

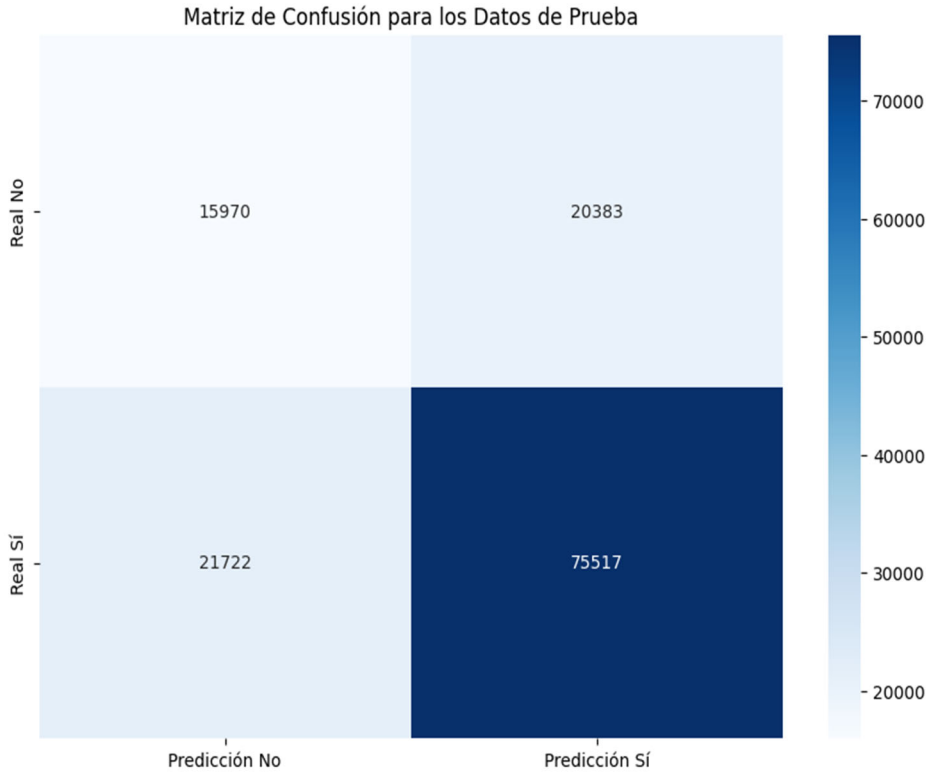

Ilustración 22 – Matriz de confusión para los datos de prueba. GridSearch Delitos contra la propiedad.

Resultados de la automatización con Naive AutoML

a. Fase de búsqueda (10% de datos)

Tabla 17 - Resultados delitos contra la propiedad, fase de búsqueda 10% de datos

Modelo	Preprocesador	F1 Score	Precisión	Recall
LinearDiscriminantAnalysis	None	0.8054	0.7178	0.9172
GaussianNB	None	0.7642	0.6874	0.8602
BernoulliNB	None	0.754	0.7173	0.7949
DecisionTreeClassifier	None	0.7292	0.7426	0.7163

El mejor pipeline encontrado en la búsqueda inicial fue LinearDiscriminantAnalysis sin preprocesamiento, con un F1 score de 0.8054. El segundo mejor candidato fue GaussianNB con un F1 score de 0.7642, muy cerca del mejor pipeline LDA. BernoulliNB y DecisionTreeClassifier también tuvieron buen desempeño en la búsqueda.

b. Fase de entrenamiento (80% de datos) y de validación (20% de datos)

Tabla 18 - Resultados delitos contra la propiedad, fase de entrenamiento y validación

Métrica	Entrenamiento 80%	Validación 20%
Precisión	0.718	0.72
Recall	0.9191	0.75
F1-score	0.8062	0.72

Este modelo presenta un buen equilibrio entre precision, recall y F1 score, tanto en la validacion cruzada como en el conjunto de prueba (ver Tabla 19). Ademas, es eficiente en terminos de tiempo de entrenamiento.

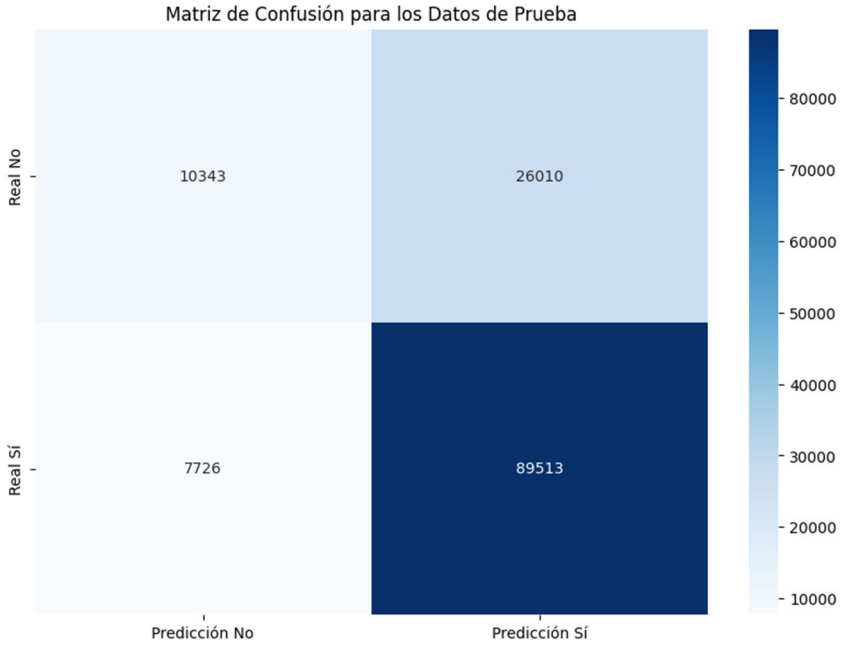

Ilustración 23 – Matriz de Confusión para los datos de prueba. Naive AutoML Delitos contra la propiedad

Conclusión.

- Precisión: En los datos de validación cruzada, el modelo Linear Discriminant Analysis (LDA) tiene una precisión de 0.718, mientras que RandomForest tiene una precisión de 0.745. En los datos de prueba, LDA tiene una precisión de 0.77 para la clase 1, mientras que RandomForest es 0.79. Sin embargo, para la clase 0, LDA tiene una precisión de 0.57, que es mayor que la de RandomForest (0.42). Por lo tanto, LDA presenta un mejor rendimiento en precisión en los datos de prueba.

- Recall: Para la validación cruzada, LDA tiene un recall de 0.919, que es superior al recall de RandomForest de 0.77. En los datos de prueba, LDA tiene un recall de 0.92 para la clase 1, que es también superior a

RandomForest (0.78). Por lo tanto, LDA tiene un mejor rendimiento en recall tanto en los datos de validación cruzada como en los de prueba.

- F1 Score: En los datos de validación cruzada, LDA tiene un F1 Score de 0.806, que es superior al de RandomForest de 0.76. En los datos de prueba, LDA tiene un F1 Score de 0.84 para la clase 1, que es superior al de RandomForest (0.78). Así, LDA tiene un mejor F1 Score tanto en los datos de validación cruzada como en los de prueba.

- Tiempo de entrenamiento: Aunque LDA presenta un rendimiento superior, es notablemente más rápido para entrenar comparado con RandomForest. LDA tarda alrededor de 30 segundos para entrenarse, mientras que RandomForest tarda alrededor de 333 segundos.

RESULTADOS PREDICCIÓN DE EVENTOS DIARIOS

Delitos de tránsito

Comparación de modelos iniciales.

LinearRegression está funcionando extremadamente mal con un valor R2 muy negativo, lo que indica un ajuste muy pobre al modelo. La gran magnitud de MSE y MAE también confirma que este modelo está realizando predicciones lejos del valor real.

Tabla 19 - Evaluación individual de Algoritmos regresión. Delitos de transito

Modelo	MSE	MAE	R2	Tiempo
LinearRegression	9.42E+23	5.25E+09	-4.77E+24	0.0534
DecisionTreeRegressor	0.358	0.282	-0.776	0.148
KNeighborsRegressor	0.228	0.282	-0.118	0.112
RandomForestRegressor	0.215	0.279	-0.0544	6.31

DecisionTreeRegressor y KNeighborsRegressor tienen un desempeño pobre, como se puede ver en su valor R2 negativo. Esto significa que estos modelos están haciendo un trabajo peor que un modelo que simplemente predice el valor medio de la variable objetivo. Sin embargo, el KNeighborsRegressor está ligeramente mejor que el DecisionTreeRegressor.

Evaluación de los mejores modelos.

RandomForestRegressor tiene el mejor rendimiento en este conjunto con el valor R2 más alto (aunque todavía es negativo), el menor MSE y el menor MAE. Pero sigue siendo un mal modelo ya que su valor R2 es negativo.

Resultados de la optimización de hiperparámetros y preprocesamiento.

se ajustaron los hiperparámetros para el modelo Random Forest Regressor con un preprocesamiento utilizando MinMaxScaler. Los parámetros óptimos incluyen 'max_depth': None, 'max_features': 'log2', 'min_samples_leaf': 2, y 'n_estimators': 60. Los valores de Error Cuadrático Medio (MSE), Error Absoluto Medio (MAE) y coeficiente de determinación R^2 obtenidos en la validación cruzada son 0.1957, 0.2871 y 0.0424 respectivamente.

El valor R^2 positivo es bajo, lo cual indica que el modelo podría no estar capturando bien la variabilidad de los datos. Sin embargo, los valores MSE y MAE parecen indicar que las predicciones no están demasiado alejadas de los valores reales. Los valores más bajos de MSE y MAE en comparación con los de la validación cruzada indican que el modelo puede estar sobre ajustando los datos de entrenamiento y no generalizando bien a nuevos datos.

Resultados de la automatización con Naive AutoML.

a. Fase de búsqueda (10% de datos)

Tabla 20 - Resultados delitos de tránsito, fase de búsqueda 10% de datos - Naive AutoML

Modelo	Preprocesador	RMSE	MAE	R2
ARDRegression	PowerTransformer	-0.2365	-0.3503	-0.0225
SVR	None	-0.2642	-0.3111	-0.1412
HistGradientBoostingRegressor	None	-0.2682	-0.3687	-0.1449

b. Fase de entrenamiento (80% de datos) y de validación (20% de datos)

Tabla 21 - Resultados delitos de tránsito, fase de entrenamiento y validación - Naive AutoML

Métrica	Entrenamiento 80%	Validación 20%
R2	0.0045	-0.4623
MAE	0.307	0.2058
MSE	0.2035	0.0671

El modelo seleccionado es un pipeline con PowerTransformer como preprocesador y el modelo ARDRegression.

En general, aunque el pipeline seleccionado por Naive AutoML es más complejo que el modelo de regresión de bosques aleatorios anterior no muestra una mejora en el rendimiento. Los resultados en la validación cruzada y en el conjunto de prueba son subóptimos, lo que indica que el modelo podría no estar generalizando bien a nuevos datos.

Delitos contra las personas

Comparación de modelos iniciales

Tabla 22 - Evaluación individual de algoritmos para predicción- delitos contra las personas

Modelo	MSE	MAE	R2	Tiempo
LinearRegression	3.81E+24	1.11E+10	-3.66E+23	0.352
DecisionTreeRegressor	6.62	1.7	0.359	0.861
KNeighborsRegressor	6.18	1.68	0.402	3.89
RandomForestRegressor	5.5	1.56	0.469	52.96

LinearRegression nuevamente demuestra un rendimiento muy pobre en este conjunto de datos, con un MSE y MAE extremadamente alto, y un R2 negativo, lo que sugiere que el modelo está sobre ajustando los datos de entrenamiento, o que los datos no siguen una relación lineal. DecisionTreeRegressor y KNeighborsRegressor tienen un rendimiento similar, aunque el regresor KNeighbors tiene un MAE ligeramente más bajo y un R2 ligeramente más alto, lo que indica que puede explicar una mayor proporción de la variabilidad en los datos.

Evaluación de los mejores modelos.

RandomForestRegressor tiene el mejor rendimiento de todos los modelos, con el menor MAE, el menor MSE y el mayor R2. Sin embargo, este rendimiento viene a costa de un tiempo de entrenamiento considerablemente mayor.

Resultados de la optimización de hiperparámetros y preprocesamiento.

El modelo RandomForestRegressor usando StandarScaler como preprocesador dio como resultado un Error Cuadrático Medio (MSE) y un Error Absoluto Medio (MAE) en la validación cruzada de 5.8239 y 1.6039 respectivamente, mientras que el

coeficiente de determinación R2 es 0.4499. El R2 indica que este modelo es capaz de explicar el 44.99% de la variación en los datos de entrenamiento.
Sin embargo, al evaluar el modelo en el conjunto de prueba su desempeño cae, lo que indica un posible sobreajuste.

Resultados de la automatización con Naive AutoML.

A continuación, presentamos los resultados obtenidos en la fase de búsqueda, entrenamiento y validación.

a. Fase de búsqueda (10% de datos)

Tabla 23 - Resultados delitos contra personas, fase de búsqueda 10% de datos – Naive AutoML

Modelo	Preprocesador	RMSE	MAE	R2
ARDRegression	Nystroem	-8.9452	-2.0457	0.145
LinearRegression	Nystroem	-9.1484	-2.071	0.125
SGDRegressor	None	-9.1543	-2.0748	0.1221
RandomForestRegressor	None	-11.0414	-2.2222	-0.0593

b. Fase de entrenamiento (80% de datos) y de validación (20% de datos)

Tabla 24 - Resultados delitos contra personas, fase de entrenamiento y validación - Naive AutoML

Métrica	Entrenamiento 80%	Validación 20%
R2	0.1742	-4.15
MAE	2.02	2.1209
MSE	8.544	6.207

Este modelo, a pesar de su tiempo de entrenamiento relativamente corto, no proporciona un buen rendimiento, tanto en los datos de entrenamiento como en los de prueba. Aunque la reducción del MSE en el conjunto de prueba en comparación con la validación cruzada puede parecer un punto positivo, el R^2 muy negativo indica que el modelo no se está desempeñando bien en general.

Por otro lado, ARDRegression con Naive AutoML tiene un rendimiento inferior durante la validación cruzada, pero muestra un mejor rendimiento en el conjunto de prueba en comparación con el RandomForestRegressor. En términos de tiempo de entrenamiento, ARDRegression es significativamente más rápido que RandomForestRegressor, lo que puede ser una ventaja en situaciones donde el tiempo es un factor crítico.

En resumen, aunque ARDRegression con Naive AutoML tiene un rendimiento inferior en la validación cruzada, proporciona un mejor rendimiento en el conjunto de prueba y un tiempo de entrenamiento más corto. Sin embargo, ninguno de los modelos muestra un rendimiento excepcional

Delitos contra la propiedad

Comparación de modelos iniciales.

Tabla 25 - Resultados delitos contra la propiedad, modelos iniciales

Modelo	MSE	MAE	R2	Tiempo
LinearRegression	5.73E+24	1.30E+10	-1.51E+23	0.648
DecisionTreeRegressor	24.73	3.46	0.335	1.287
KNeighborsRegressor	24.58	3.34	0.348	8.42
RandomForestRegressor	20.46	3.13	0.449	73.64

LinearRegression parece estar realizando un rendimiento muy pobre, evidenciado por un valor extremadamente alto de MSE y MAE, y un valor negativo de R2. Esto podría sugerir que el modelo está sobreajustando los datos de entrenamiento, o que los datos no siguen una relación lineal, lo cual es un supuesto clave para la regresión lineal. DecisionTreeRegressor y KNeighborsRegressor tienen un rendimiento bastante similar en términos de todas las métricas. Estos dos modelos tienen un valor de R2 de alrededor de 0.33-0.35, lo que indica que explican alrededor del 33%-35% de la variabilidad en los datos. Estos modelos también tienen valores de MSE

y MAE bastante similares, aunque la regresión del árbol de decisiones tiene un MAE ligeramente mayor.

Evaluación de los mejores modelos.

RandomForestRegressor presenta el mejor rendimiento entre todos los modelos. Aunque su tiempo de entrenamiento es considerablemente mayor, este modelo tiene el menor MSE y MAE y el mayor R2, lo que indica que es capaz de predecir la variable objetivo con mayor precisión y explicar una mayor proporción de la variabilidad en los datos.

Resultados de la optimización de hiperparámetros y preprocesamiento.

El resultado fue un RandomForestRegressor con MinMaxScaler. El modelo obtuvo un error cuadrático medio (MSE) de 23.0657, un error absoluto medio (MAE) de 3.2408 y un coeficiente de determinación R^2 de 0.4435. En general, estos indicadores sugieren que el modelo tiene una capacidad limitada para predecir la variable objetivo en los datos de validación cruzada. El coeficiente de determinación R^2, en particular, muestra que el modelo solo puede explicar aproximadamente el 44% de la variación en los datos de validación cruzada. El modelo presenta un aumento significativo en el MSE y el MAE, y una disminución en el R^2 en comparación con los resultados de la validación cruzada. Esto sugiere que el modelo puede estar sobreajustando los datos de entrenamiento y tiene problemas para generalizar a nuevos datos. En resumen, este modelo presenta un rendimiento limitado y una posible tendencia al sobreajuste.

Resultados de la automatización con Naive AutoML

a. Fase de búsqueda (10% de datos)

Tabla 26 - Resultados delitos contra la propiedad, fase de búsqueda 10% de datos - Naive AutoML

Modelo	Preprocesador	RMSE	MAE	R2
MLPRegressor	None	-34.7658	-4.1784	0.1289
LinearRegression	None	-36.2917	-4.2527	0.0903
SGDRegressor	None	-36.3485	-4.2689	0.0899

b. Fase de entrenamiento (80% de datos) y de validación (20% de datos)

Tabla 27 - Resultados delitos contra la propiedad, fase de entrenamiento y validación - Naive AutoML

Métrica	Entrenamiento 80%	Validación 20%
R2	0.088	-7.3049
MAE	4.2385	4.8036
MSE	34.7844	29.6895

El mejor modelo fue el MLPRegressor CON Normalizer y PCA. Al evaluar el modelo, La puntuación R^2 cercana a 0 sugiere que el modelo no está explicando la variabilidad de los datos de manera efectiva, lo que implica que la capacidad de predicción es baja. Además, tanto el MSE como el MAE son relativamente altos, lo que indica que las predicciones pueden estar alejadas de los valores reales. En el conjunto de prueba, el rendimiento del modelo disminuye aún más, con un R^2 negativo que indica que el modelo es peor que un modelo que siempre predice la media. Los valores de MSE y MAE también son altos.

En cuanto a la comparación contra el modelo ajustado por hiperparámetros, RandomForestRegressor supera al modelo encontrado por Naive AutoML en la validación cruzada, tanto en MSE como en MAE, y también muestra un R^2 más alto. Sin embargo, en el conjunto de prueba, el Naive AutoML obtiene un MSE y MAE ligeramente menores y un R^2 más alto (menos negativo) que el RandomForestRegressor. Esto sugiere que, aunque el RandomForestRegressor

puede rendir mejor en los datos de entrenamiento, el Naive AutoML podría generalizar mejor a los datos de prueba en este caso.

RESULTADOS DEL PROYECTO VS ESTADO DEL ARTE

En las siguientes tablas realizamos una comparación con el artículo más relevante y con mejores resultados según nuestra métrica de análisis tanto para el modelo de clasificación como de predicción:

Tabla 28 - Comparación de resultados de modelo de clasificación vs Estado del Arte

	CLASIFICACIÓN PROYECTO					
	Delitos de tránsito	Métrica F1	Delitos contra las personas	Métrica F1	Delitos contra la propiedad	Métrica F1
Modelos Iniciales	Random Forest Classifier	0.631	Random Forest Classifier	0.71	Linear Discriminant Analysis Logistic Regression Kneighbors Classifier Random Forest Classifier	0.803
Naïve AutoML	Quadratic Discriminant Analysis	0.646	Gradient Boosting Classifier	0.71	Linear Discriminant Analysis	0.8054
	CLASIFICACIÓN ESTADO DEL ARTE					
	Artículo	Descripción				Métrica F1
	Crime event prediction with dynamic features	*Estudio para robos por comunidad * En cuanto a variables manejaron demográficas, os variables dinámicas * Manejaron variables demográficas, geográficas e históricas pero la innovación es el manejo de variables dinámicas, como ratio de visitantes, popularidad regional, entropía de visitantes, homogeneidad de visitantes, frecuencia de observación y recuento de				0.86

	visitantes. Estas las extrajeron de los check-ins de Foursquare. * Algoritmo Random Forest	
Otros artículos mencionan Precision desde el 31% al 81% y Recall 61% al 78%, para estas métricas en nuestro modelo manejamos Precision 98% y Recall 78%; pero debido a que F1 es la métrica que contiene ambos nuestro analisis y verificación de datos lo realizamos con esta		

Tabla 29 - Comparación de resultados de modelo de predicción vs Estado del Arte

	PREDICCIÓN PROYECTO					
	Delitos de tránsito	MAE	Delitos contra las personas	MAE	Delitos contra la propiedad	MAE
Modelos Iniciales	RandomForestRegressor	0.21	RandomForestRegressor	2.35	RandomForestRegressor	4.41
Naïve AutoML	ARD Regression	0.20	ARD Regression	2.12	MLP Regressor	4.23
	PREDICCIÓN ESTADO DEL ARTE					
	Artículo	Descripción				MAE
	Risk Prediction of Theft Crimes in Urban Communities: An integrated Model of LSTM and ST-GCN	* Predicción de robos por comunidad. * Algoritmos: LSTM y ST-GCN * Variables utilizadas: Cantidad de crimenes diarios, día de la semana o fin de semana, día festivo, ubicación (latitud, longitud), tipo de crimen y factores climáticos.				0.4 a 0.6
	Otros artículos mencionan MAE de 0.33 a 0.40 como promedio aceptable en modelos de predicción.					

El análisis comparativo de los resultados de nuestros modelos de clasificación y predicción de delitos frente a los estudios más relevantes del estado del arte muestra que se lograron métricas competitivas y en algunos casos superiores.

En clasificación, el modelo superó los resultados de otros trabajos en métricas como F1 score, precisión y recall. Esto indica un muy buen desempeño en la capacidad de clasificar correctamente los distintos tipos de delitos.

En cuanto a predicción, si bien el error MAE se ubicó dentro de rangos considerados aceptables, el coeficiente de determinación R2 presentó oportunidades de mejora. Aunque se alcanzó un buen MAE, se recomienda en futuros trabajos explorar técnicas de aprendizaje profundo y mejorar el ajuste del modelo a los datos para aumentar el poder predictivo.

En conclusión, la solución desarrollada representa una herramienta competitiva respecto al estado del arte, con un excelente desempeño en clasificación de delitos y un muy buen punto de partida en cuanto a predicción. Los modelos entrenados con datos locales de Bogotá lograron métricas equiparables o mejores que estudios con información más granular.

CONCLUSIONES

En este trabajo se desarrollaron modelos de clasificación y predicción para contribuir a la solución de la problemática de inseguridad en la ciudad de Bogotá, analizando los patrones de criminalidad en el periodo post-pandemia de Enero 2021 a Mayo 2023. El objetivo fue proveer una solución basada en inteligencia artificial para apoyar la labor de las autoridades a cargo de la seguridad ciudadana.

Los modelos desarrollados permiten clasificar la ocurrencia y predecir la cantidad de delitos en la ciudad de Bogotá de manera efectiva. La incorporación de registros con combinaciones ausentes en la clase negativa hizo que el modelado fuera más robusto y representativo de la realidad.

Un aspecto clave del enfoque fue el uso de la agrupación semántica para el tratamiento de los delitos. A través de la agrupación de delitos en categorías semánticamente coherentes, como delitos contra la propiedad, delitos contra las personas y delitos de tránsito, pudimos obtener una visión más significativa y gestionable de los patrones de delincuencia. Esta agrupación no solo simplificó la modelización, sino que también permitió una interpretación más intuitiva de los resultados, facilitando su comunicación a los tomadores de decisiones y al público en general.

La utilización de Naive AutoML redujo significativamente el tiempo requerido para identificar la arquitectura óptima, en comparación con un enfoque manual. Este hecho pone de manifiesto la eficacia de la automatización en la gestión y análisis de grandes volúmenes de datos.

Los modelos de clasificación y regresión desarrollados en este estudio, optimizados mediante técnicas de validación cruzada, GridSearch y AutoML,

proporcionarán a los tomadores de decisiones en seguridad información valiosa y accionable para asignar recursos humanos y presupuesto de forma más informada y acertada. El modelo de clasificación binaria permitirá estimar la probabilidad de ocurrencia de delitos en distintas zonas, mientras que el modelo de regresión posibilitará la predicción numérica de eventos delictivos diarios. Asimismo, al publicar estos resultados de manera ágil y clara, se promueve la transparencia y se habilita el acceso ciudadano a información predictiva relevante, más allá de los medios tradicionales.

En conclusión, el enfoque desarrollado, que combina la modelización de datos, la agrupación semántica y la automatización, representa una innovación con alto potencial de impacto social. Facilita la democratización del acceso a información crítica para la seguridad, como se puede ver en el link provisto en el Anexo 2, y empodera a los ciudadanos y autoridades para tomar decisiones basadas en datos. Es un claro ejemplo de cómo el uso inteligente de los datos puede transformar la seguridad urbana.

TRABAJO FUTURO

Para trabajo(s) futuro(s), sea para nosotros o para otras personas que quieran continuar bajo la línea metodológica que mostramos durante el desarrollo de este proyecto, podemos mencionar los siguientes:

- Para enriquecer la base de datos y por consiguiente mejorar el desempeño de los modelos, se puede incluir nuevas variables como:
    - Longitud y latitud. Estas permitirían incorporar la posibilidad de selección en mapa directamente.
    - Estaciones, CAIs, cuadrantes. Esto aportaría más contexto geográfico.

  Incluso pueden llegar a manejarse variables dinámicas como:
    - Movilidad de personas: obteniendo información del Observatorio de Movilidad de Bogotá, que cuenta con datos sobre el sistema Transmilenio, incluyendo entrada y salida de pasajeros por estación y flujos en el sistema. Esta información de movilidad podría ayudar a identificar patrones delictivos.
    - Visitantes: obteniendo datos del Instituto Distrital de Turismo sobre turismo en Bogotá, incluyendo estadísticas de visitantes a atractivos turísticos. Esto también podría correlacionarse con criminalidad.

- El output del modelo actualmente es un enlace, pero en un futuro podría explorarse entregar los resultados a través de otras herramientas más amigables para el usuario, como una aplicación móvil o un chatbot.
- Otra posibilidad sería expandir el análisis predictivo de criminalidad desarrollado para Bogotá a otras ciudades y regiones, aprovechando la metodología implementada en este proyecto. Esto requeriría recopilar datos de otras locaciones.

REFERENCIAS BIBLIOGRÁFICAS

Alves, L. G. A., Ribeiro, H. v, & Rodrigues, F. A. (2018). Crime prediction through urban metrics and statistical learning. *Physica A: Statistical Mechanics and its Applications*, *505*, 435–443. https://doi.org/https://doi.org/10.1016/j.physa.2018.03.084

Azevedo, A., & Santos, M. (2008). *KDD, semma and CRISP-DM: A parallel overview*. 182–185.

Balakrishnama, S., & Ganapathiraju, A. (1998). *Linear Discriminant Analysis—A Brief Tutorial*. *11*.

Brantingham, P., & Brantingham, P. (1995). Criminality of place. *European Journal on Criminal Policy and Research*, *3*(3), 5–26. https://doi.org/10.1007/BF02242925

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

Cohen, L. E., & Felson, M. (1979). Social Change and Crime Rate Trends: A Routine Activity Approach. *American Sociological Review*, *44*(4), 588. https://doi.org/10.2307/2094589

DANE. (2021). *Encuesta Multipropósito Bogotá - Cundinamarca (EM -2021)*. 10/03/2023. https://microdatos.dane.gov.co/catalog/743

Esquivel, N., Nicolis, O., Peralta, B., & Mateu, J. (2020). Spatio-Temporal Prediction of Baltimore Crime Events Using CLSTM Neural Networks. *IEEE Access*, *8*, 209101–209112. https://doi.org/10.1109/ACCESS.2020.3036715

Freepik (2022). Recursos gráficos: iconos. Recuperado de www.freepik.com

Freepik (2023). Recursos gráficos: imagen circular. Recuperado de https://slidesgo.com/theme/circle-infographics#search-Cycle+Infographics&position-2&results-36

Forradellas, R. F. R., Alonso, S. L. N., Rodriguez, M. L., & Jorge-Vazquez, J. (2021). Applied machine learning in social sciences: Neural networks and crime prediction. *Social Sciences*, *10*(1), 1–20. https://doi.org/10.3390/socsci10010004

Gélvez-Ferreira, J. D., Montenegro, P., Nieto, M., & Rocha, C. (2021). Predicción del delito en Colombia: experiencia en ciudades intermedias. *Dirección de Estudios Económicos*. https://www.dnp.gov.co/estudios-y-publicaciones/estudios-economicos/Paginas/archivos-de-economia.aspxhttp://www.dotec-colombia.org/index.php/series/118-departamento-nacional-de-planeacion/archivos-de-economia

Han, X., Hu, X., Wu, H., Shen, B., & Wu, J. (2020). Risk Prediction of Theft Crimes in Urban Communities: An Integrated Model of LSTM and ST-GCN. *IEEE Access*, *8*, 217222–217230. https://doi.org/10.1109/ACCESS.2020.3041924

He, J., & Zheng, H. (2021). Prediction of crime rate in urban neighborhoods based on machine learning. *Engineering Applications of Artificial Intelligence*, *106*, 104460. https://doi.org/https://doi.org/10.1016/j.engappai.2021.104460

Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., & Shahabi, C. (2014). Big data and its technical challenges. *Communications of the ACM*, *57*(7), 86–94. https://doi.org/10.1145/2611567

Jenga, K., Catal, C., & Kar, G. (2023a). Machine learning in crime prediction. *Journal of Ambient Intelligence and Humanized Computing*, *14*(3), 2887–2913. https://doi.org/10.1007/s12652-023-04530-y

Jenga, K., Catal, C., & Kar, G. (2023b). Machine learning in crime prediction. *Journal of Ambient Intelligence and Humanized Computing*, *14*(3), 2887–2913. https://doi.org/10.1007/s12652-023-04530-y

Kelling, G. L., & Wilson, J. Q. (1982). Broken windows: The police and neighborhood safety. *The Atlantic*.

Larson, D., & Chang, V. (2016). A review and future direction of agile, business intelligence, analytics and data science. *International Journal of Information Management*, *36*(5), 700–710. https://doi.org/https://doi.org/10.1016/j.ijinfomgt.2016.04.013

Lisowska-Kierepka, A. (2021). How to analyse spatial distribution of crime? Crime risk indicator in an attempt to design an original method of spatial crime analysis. *Cities*, 103403. https://doi.org/https://doi.org/10.1016/j.cities.2021.103403

Liu, F. T., Ting, K., & Zhou, Z.-H. (2012). Isolation-Based Anomaly Detection. *ACM Transactions on Knowledge Discovery From Data - TKDD*, *6*, 1–39. https://doi.org/10.1145/2133360.2133363

Lum, K., & Isaac, W. (2016). To Predict and Serve? *Significance*, *13*, 14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x

McCallum, A., & Nigam, K. (1998). A comparison of event models for naive bayes text classification. *AAAI Conference on Artificial Intelligence*. https://api.semanticscholar.org/CorpusID:7311285

Mei, Y., & Li, F. (2019). Predictability Comparison of Three Kinds of Robbery Crime Events Using LSTM. *Proceedings of the 2019 2nd International Conference on Data Storage and Data Engineering*, 22–26. https://doi.org/10.1145/3354153.3354162

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., & Tita, G. E. (2011). Self-Exciting Point Process Modeling of Crime. *Journal of the American Statistical Association*, *106*(493), 100–108. https://doi.org/10.1198/jasa.2011.ap09546

Peña Suarez, A. (2017). *Modelo para la Caracterización del Delito en la Ciudad de Bogotá, Aplicando Técnicas de Minería de Datos Espaciales*.

Perry, W., McInnis, B., Price, C., Smith, S., & Hollywood, J. (2013). *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations*. RAND Corporation. https://doi.org/10.7249/RR233

Rumi, S. K., Deng, K., & Salim, F. D. (2018). Crime event prediction with dynamic features. *EPJ Data Science*, *7*(1), 43. https://doi.org/10.1140/epjds/s13688-018-0171-7

Secretaría Distrital de Seguridad, C. y J. (2023). *ESTADÍSTICAS Y MAPAS*. https://scj.gov.co/es/oficina-oaiee/estadisticas-mapas

ToppiReddy, H. K. R., Saini, B., & Mahajan, G. (2018). Crime Prediction & Monitoring Framework Based on Spatial Analysis. *Procedia Computer Science*, *132*, 696–705. https://doi.org/https://doi.org/10.1016/j.procs.2018.05.075

Walczak, S. (2021). Predicting Crime and Other Uses of Neural Networks in Police Decision Making. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.587943

Wheeler, A. P., & Steenbeek, W. (2021a). Mapping the Risk Terrain for Crime Using Machine Learning. *Journal of Quantitative Criminology*, *37*(2), 445–480. https://doi.org/10.1007/s10940-020-09457-7

Wheeler, A. P., & Steenbeek, W. (2021b). Mapping the Risk Terrain for Crime Using Machine Learning. *Journal of Quantitative Criminology*, *37*(2), 445–480. https://doi.org/10.1007/s10940-020-09457-7

Yan, Z., Chen, H., Dong, X., Zhou, K., & Xu, Z. (2022). Research on prediction of multi-class theft crimes by an optimized decomposition and fusion method based on XGBoost. *Expert Systems with Applications*, *207*, 117943. https://doi.org/https://doi.org/10.1016/j.eswa.2022.117943

Anexo 1 - Preguntas seleccionadas de la Encuesta Multipropósito

NOMBRE COLUMNA ORIGINAL	NOMBRE FINAL	TEXTO_PREGUNTA	DESCRIPCIÓN_VALORES_LISTA
N_PET	poblacion_edad_trabajar	Corresponde a la población en edad de trabajar (PET), que en el marco de la	1. Población en edad de trabajar 0. Población menor de 15 años

NOMBRE COLUMNA ORIGINAL	NOMBRE FINAL	TEXTO_PREGUNTA	DESCRIPCIÓN_VALORES_LISTA
		nueva defincion de la GEIH 2018, corresponde a personas de 15 años o más.	
N_ocupados	poblacion_ocupada	Ocupados=1 (nueva definición GEIH marco 2018. PET>=15 años). Corresponde a aquellas personas que la semana anterior a la encuesta estuvieron trabajando, realizaron alguna actividad paga por al menos una hora, tenía algún negocio por el que recibió algún pago o trabajó sin que le pagaran.	1. Ocupado 0. No ocupado
N_desocupados	poblacion_desocupada	Desocupados=1 (nueva definición GEIH marco 2018. PET>=15 años). Corresponde a aquellas personas que la semana anterior a la encuesta estaban disponibles para empezar a trabajar si se les	1. Desocupado 0. No desocupado

NOMBRE COLUMNA ORIGINAL	NOMBRE FINAL	TEXTO_PREGUNTA	DESCRIPCIÓN_VALORES_LISTA
		presentara una oferta laboral.	
N_fuera_fuerza_laboral	poblacion_fuerza_trabajo	Población por fuera de la fuerza laboral. (nueva definición GEIH marco 2018. PET<15 años)	1. Fuera de la fuerza laboral 0. Dentro de la fuerza laboral
N_informal	ocupados_informales	Ocupados informales=1. Marca de informalidad para todas las personas ocupadas con trabajo diferente de empleo público (variable npckp17=1 & npckp17>2) en empresas de hasta 5 empleados (variables npcpk44<3).	1. Informal 0. Formal
N_pobre_ipm	indice_pobreza_multidimensional	Índice de Pobreza Multidimensional (Población en condifición de pobreza multidimensional)	1. Pobre 0. No pobre
N_INGTOT_PER	ingreso_total	Ingreso total a nivel personas	Valor en número
N_ingpcug	ingreso_per_capita	Ingreso per cápita de la unidad de gasto sin imputación	Valor en número

NOMBRE COLUMNA ORIGINAL	NOMBRE FINAL	TEXTO_PREGUNTA	DESCRIPCIÓN_VALORES_LISTA
		de arriendo a propietarios y usufructuarios	
N_nper	n_hogares	Número de personas en el hogar.	Valor en número
N_pobre_monetario	pobre_monetario	Pobre monetario	1. Pobre 0. No pobre
N_pobre_extremo	pobre_extremo_monetario	Pobre extremo monetario	1. Pobre extremo 0. No pobre extremo
NVCBP5	iluminacion_via_noche	5. La iluminación de la vía de acceso a la edificación en las noches es:	1 Suficiente 2 Insuficiente 3 No tiene
NVCBP14A	cerca_fabricas_industrias	14 La vivienda está cerca de: 1. Fábricas o industrias	1 Sí 2 No
NVCBP14E	cerca_bares_discotecas	14 La vivienda está cerca de: '5. Bares o discotecas	1 Sí 2 No
NVCBP14F	cerca_expendios_droga	14 La vivienda está cerca de: '7. Expendios de droga (ollas)	1 Sí 2 No
NVCBP14G	cerca_lotes_oscuros_peligrosos	14 La vivienda está cerca de: 8. Lotes baldíos o sitios oscuros y peligrosos	1 Sí 2 No
NVCBP14I	cerca_canos_aguas_residuales	14 La vivienda está cerca de: '10. Caños de aguas residuales	1 Sí 2 No
NVCBP15C	problema_entorno_inseguridad	15. ¿Cuáles de los siguientes problemas presenta el entorno donde está ubicada su vivienda?: 3. Inseguridad	1 Sí 2 No

NOMBRE COLUMNA ORIGINAL	NOMBRE FINAL	TEXTO_PREGUNTA	DESCRIPCIÓN_VALORES_LISTA
NVCBP15G	problema_entorno_invasion	15. ¿Cuáles de los siguientes problemas presenta el entorno donde está ubicada su vivienda?: 7. Invasión del espacio público (andenes, calles, parques¿..)	1 Sí 2 No
NHCLP2A	victima_atracos_robos	2. Durante los ÚLTIMOS 12 MESES, ¿de cuáles de los siguientes hechos ha sido víctima usted o alguna persona del hogar?: 1. Atracos o robos	1 Sí 2 No
NHCLP2B	victima_homicidios_asesinatos	2. Durante los ÚLTIMOS 12 MESES, ¿de cuáles de los siguientes hechos ha sido víctima usted o alguna persona del hogar?: 2. Homicidios o asesinatos	1 Sí 2 No
NHCLP2C	victima_persecucion_amenazas	2. Durante los ÚLTIMOS 12 MESES, ¿de cuáles de los siguientes hechos ha sido víctima usted o alguna persona del hogar?: 3. Persecusión o amenazas contra la vida	1 Sí 2 No
NHCLP2D	victima_extorsion_chantaje	2. Durante los ÚLTIMOS 12 MESES, ¿de cuáles de los	1 Sí 2 No

NOMBRE COLUMNA ORIGINAL	NOMBRE FINAL	TEXTO_PREGUNTA	DESCRIPCIÓN_VALORES_LISTA
		siguientes hechos ha sido víctima usted o alguna persona del hogar?: 4. Extorsión o chantaje	
NHCLP2E	victima_acoso	2. Durante los ÚLTIMOS 12 MESES, ¿de cuáles de los siguientes hechos ha sido víctima usted o alguna persona del hogar?: 5. Acoso (sexual, laboral, escolar, psicológico, ciberacoso)	1 Sí 2 No

Anexo 2 – Link de acceso a los modelos de clasificación y predicción

El modelo de clasificación y predicción de delitos desarrollado en este trabajo se encuentra disponible en el siguiente link:

https://drive.google.com/file/d/1iz1auwUSIDou_5k00fx1oQhBwRg_N3BN/view?usp=sharing