Evaluación del desempeño de diferentes modelos de aprendizaje automático para la predicción de delitos en la ciudad de Medellín

Juan Pablo Areiza Jiménez

Informe de Practica Académica para Optar el Título de Ingeniero de Telecomunicaciones

Modalidad de Práctica Cursada
Proyecto de Investigación

Asesor Interno
Luis Germán García Morales, Ph.D. Ingeniería Electrónica

Universidad de Antioquia
Facultad de Ingeniería
Pregrado en Ingeniería de Telecomunicaciones
Medellín
2024

Cita	Areiza Jiménez [1]
Referencia Estilo IEEE (2020)	[1] J. P. Areiza Jiménez, "Evaluación del desempeño de diferentes modelos de aprendizaje automático para la predicción de delitos en la ciudad de Medellín", Proyecto de Investigación, Pregrado en Ingeniería de Telecomunicaciones, Universidad de Antioquia, Medellín, 2024.

Grupo de Investigación en Sistemas Embebidos e Inteligencia Computacional (SISTEMIC).

UNIVERSIDAD DE ANTIOQUIA
Vicerrectoría de Docencia

Sistema de Bibliotecas

Centro de Documentación de Ingeniería (CENDOI)

Repositorio Institucional: http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes.

Decano/Director: Julio César Saldarriaga Molina.

Jefe departamento: Eduard Emiro Rodríguez Ramírez.

RECONOCIMIENTO

Esta investigación fue apoyada por el Sistema General de Regalías de Colombia, bajo el Proyecto "Administración inteligente de problemas de seguridad ciudadana a través de modelos y herramientas generadas a partir de plataformas para territorios inteligentes apoyadas por estrategias de participación ciudadana en la ciudad de Medellín" (Código BPIN 020000100044).

TABLA DE CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

SIGLAS, ACRONIMOS Y ABREVIATURAS

ML	Machine Learning
AR	Autoregressive
MA	Moving Average
ACF	Autocorrelation Function
PACF	Partial Autocorrelation Function
ARIMA	Autoregressive Integrated Moving Average
SARIMA	Seasonal Autoregressive Integrated Moving Average
ARMAX	Autoregressive Moving Average with Exogenous Variables
VAR	Vector Autoregressive
SVM	Support Vector Machine
SIATA	Sistema de Alerta Temprana de Antioquia
LightGBM	Light Gradient Boosting Machine
MAE	Mean Absolute Error
RMSE	Root Mean Square Error

RESUMEN

El presente trabajo de grado se centra en la aplicación, evaluación y comparación de diversos modelos de aprendizaje automático para predecir la incidencia delictiva en distintas zonas de Medellín. Se analizaron modelos como ARIMA, SARIMA, ARMAX, VAR, Prophet y redes neuronales recurrentes LSTM, utilizando métricas estándar como RMSE y MAE, así como análisis de residuos para evaluar la precisión y capacidad predictiva de cada uno.

Los resultados indican que los modelos autorregresivos con diferenciación de primer orden para gestionar tendencias estacionales, ofrecen resultados prometedores al capturar las características estacionarias de los datos y mejorar la estacionariedad de las series temporales de delitos. Sin embargo, se identificó asimetría positiva en los residuos, lo que sugiere la posibilidad de mejorar la complejidad del modelo para capturar patrones residuales no explicados.

Por otro lado, la inclusión de términos estacionales no condujo a mejoras sustanciales, lo que indica que la estacionalidad podría no estar adecuadamente capturada por enfoques lineales. En contraste, la incorporación de datos exógenos como variables de control y meteorológicas resultó en mejoras significativas en métricas como RMSE y MAE, así como en una menor asimetría en los residuos, lo que refleja una capacidad mejorada del modelo para explicar la variabilidad observada en los datos de delitos.

La implementación de este estudio y los modelos desarrollados para la predicción de delitos tienen como objetivo contribuir al desarrollo de herramientas para la gestión inteligente de la seguridad ciudadana, facilitando la implementación de medidas proactivas que promuevan un entorno más seguro para la comunidad.

Palabras Clave **–** ARIMA, SARIMA, ARMAX, VAR, Prophet, RNN, LSTM, RMSE, MAE, Residuos, Aprendizaje Automático, Variables Meteorológicas, Variables de Control.

1. INTRODUCCIÓN

La mejora de la seguridad ciudadana se erige como un reto fundamental en la constante labor de transformar y optimizar entornos urbanos, impactando directamente en el bienestar de la población. En concordancia con la Agenda 2030 de las Naciones Unidas (ODS-ONU), se destaca la importancia integral de la seguridad ciudadana para salvar la vida, la integridad y los activos personales, buscando reducir significativamente diversas formas de violencia y fortalecer las instituciones nacionales a nivel internacional [1]. Esta perspectiva se alinea con las opiniones de organismos influyentes como la Organización para la Cooperación y el Desarrollo Económicos (OCDE).

La percepción de la seguridad está estrechamente vinculada al riesgo de convertirse en víctima de un delito, intensificándose a medida que crecen las estadísticas delictivas. En el contexto regional de Antioquia, se evidencia una brecha entre los incidentes de seguridad documentados y la cantidad de personas detenidas en relación con estos delitos, con un porcentaje de detención relativamente bajo. Esta disparidad resalta la insuficiencia en la capacidad de respuesta de los sistemas de seguridad urbanos, siendo una de las causas la falta de información oportuna y precisa. La dependencia de informes de víctimas o testigos, sumada a la reserva de algunos ciudadanos a denunciar por desconfianza o temor, contribuye a la complejidad del panorama. Además, los datos existentes de diversas fuentes son analizados manualmente por las agencias de seguridad, prolongando el tiempo necesario para una respuesta rápida.

Esta observación subraya la importancia crucial de la información en la resolución de desafíos de seguridad urbana. Sin embargo, la credibilidad de esta información debe ser garantizada, siendo fundamental para la implementación de estrategias específicas que mejoren la seguridad ciudadana en un territorio determinado. Algunas experiencias exitosas en la región han demostrado que el uso de tecnologías de la información y la comunicación, así como la colaboración entre actores públicos y privados, pueden facilitar la recolección, el procesamiento y el análisis de datos sobre el delito y la violencia [3].

Frente a este escenario complejo, es esencial que las autoridades dispongan de herramientas eficaces y confiables para seguir las tendencias delictivas, prevenir su comportamiento e identificar las áreas de mayor vulnerabilidad. De esta forma, podrán optimizar sus recursos y aplicar estrategias preventivas. En este marco, el presente trabajo de investigación plantea una solución

integral que utiliza datos históricos de delitos de la plataforma MEData, combinándolos con variables meteorológicas, obtenidas de los datos históricos que proporciona el SIATA y variables de control. El objetivo es implementar, evaluar y comparar distintos modelos de aprendizaje automático para la predicción del delito, empleando una o varias métricas de desempeño típicas en problemas de regresión. La incorporación de variables meteorológicas y controles, como la cercanía a fechas de pago, días feriados y fines de semana, se propone como una estrategia para mejorar la capacidad predictiva de los modelos. Estos modelos se fundamentan en técnicas estadísticas y algorítmicas que permiten predecir la cantidad de delitos en una zona específica, a partir de patrones y tendencias detectadas en los datos históricos.

En última instancia, este proyecto busca contribuir al desarrollo de una herramienta para la administración inteligente de la seguridad ciudadana. La implementación de medidas proactivas, respaldadas por la aplicación de técnicas avanzadas de análisis de datos, tiene como objetivo fomentar el establecimiento de territorios inteligentes, promoviendo un entorno más seguro para la comunidad. Esta propuesta se alinea con los principios de la urbanización sostenible, que busca mejorar la calidad de vida de las personas, garantizar el acceso a servicios básicos y fortalecer la cohesión social y la gobernabilidad [4].

Este documento se ha organizado de la siguiente manera: en el capítulo 2 se detallan los objetivos generales y específicos del estudio. El capítulo 3 ofrece un marco teórico, abordando conceptos clave como procesos estocásticos, estacionariedad, autocorrelación y diversos modelos para el pronóstico de series temporales, incluidos ARIMA, SARIMA, ARMAX, VAR, Prophet, regresión logística, LightGBM y redes neuronales. El capítulo 4 revisa trabajos relacionados con la investigación. En el capítulo 5 se describe la metodología empleada, desde la recopilación y comprensión de datos históricos de delitos y meteorológicos, hasta el análisis de características y la aplicación de diferentes modelos de predicción. El capítulo 6 presenta los resultados obtenidos de la aplicación de estos modelos. Finalmente, el capítulo 7 expone las conclusiones del estudio.

2. OBJETIVOS

2.1. Objetivo General

Evaluar el desempeño de diferentes modelos de aprendizaje automático para la predicción de delitos en la ciudad de Medellín, utilizando datos históricos de la plataforma MEData, variables meteorológicas y de control y programación de software mediante Python, con el fin de contribuir a la administración inteligente de problemas de seguridad ciudadana.

2.2. Objetivos Específicos

1. Realizar una revisión de la literatura para la selección de variables de interés, incluyendo factores meteorológicos y variables de control, así como la elección de algoritmos y modelos de aprendizaje automático adecuados, con el fin de analizar su relación y determinar su impacto en la predicción de la cantidad de delitos en zonas específicas de la ciudad de Medellín.
2. Implementar y ajustar diferentes modelos de aprendizaje automático y métodos de forecasting utilizando Python, para la predicción de delitos en la ciudad de Medellín. Se evaluará el desempeño de los modelos mediante métricas típicas para este tipo de escenarios.
3. Comparar el rendimiento de los modelos implementados, identificando aquellos que presentan la mayor eficacia en la predicción de delitos en la ciudad de Medellín.

3. MARCO TEÓRICO

Los desafíos de seguridad ciudadana en la ciudad de Medellín son una problemática que demanda enfoques innovadores respaldados por teorías y modelos que permitan comprender y predecir patrones delictivos. En este contexto, se propone un marco teórico que integra modelos de aprendizaje automático (ML) como herramienta clave para mejorar la capacidad predictiva del delito en Medellín, evaluando también conceptos fundamentales como los procesos estocásticos, la estacionariedad, la autocorrelación y la autocorrelación parcial para una aplicación efectiva de estos modelos.

3.1. Procesos Estocásticos

Un proceso estocástico es una colección de variables aleatorias (X_t) ordenadas en el tiempo. Estos modelos tratan las series temporales como realizaciones de un proceso aleatorio y se describen a través de sus distribuciones de probabilidad conjunta o, de manera práctica, mediante sus momentos (media, varianza, etc.). La media del proceso se denota generalmente como $(\mu_t = E(X_t))$, y la función de autocovarianza mide la dependencia entre las distintas observaciones de la serie [5].

3.2. Estacionariedad

La estacionariedad es una propiedad crucial para los modelos de series temporales. Un proceso estocástico es estacionario en sentido estricto si sus propiedades estadísticas no cambian con el tiempo. En la práctica, se utiliza más comúnmente la estacionariedad en sentido amplio, que requiere que la media del proceso sea constante y que la autocovarianza dependa solo del desfase temporal y no del tiempo en sí. Esto implica que las características del proceso son invariantes en el tiempo, lo cual es esencial para modelar y predecir de manera confiable [5].

3.3. Autocorrelación

La función de autocorrelación mide la correlación entre observaciones de una serie temporal separadas por un cierto desfase temporal (k). Para procesos estacionarios, esta función se denota como (ρ_k) y disminuye a medida que aumenta el desfase. El gráfico de la función de

autocorrelación, conocido como correlograma, es una herramienta útil para identificar patrones y la dependencia temporal en la serie [6].

En aplicaciones prácticas, la función de autocorrelación se estima a partir de las observaciones disponibles y se representa gráficamente para evaluar la estructura temporal de la serie.

3.4. Autocorrelación Parcial

La función de autocorrelación parcial mide la correlación entre dos puntos en una serie temporal, eliminando la influencia de los valores intermedios. Esto es útil para identificar el orden apropiado de un modelo autorregresivo (AR). La función de autocorrelación parcial se estima generalmente a través de regresiones sucesivas donde cada regresión incluye un retardo adicional [6].

El correlograma de autocorrelación parcial es otra herramienta gráfica que ayuda a determinar la estructura de dependencia en una serie temporal y a identificar el orden de los modelos AR.

3.5. Eliminación de la Tendencia

Pocas series temporales en la realidad son estacionarias. Muchas presentan tendencias y variaciones estacionales. La diferenciación es una técnica común para transformar series no estacionarias en estacionarias. Por ejemplo, restando el valor previo de la serie de cada observación (primera diferenciación) se puede eliminar una tendencia lineal. Series que requieren más de una diferenciación se dicen integradas de orden superior ($I(d)$), donde d es el número de diferenciaciones necesarias [6].

3.6. Modelos para el pronóstico de series temporales

Entre los métodos y modelos, se encuentran los modelos de forecasting de series de tiempo, los cuales son técnicas estadísticas que utilizan datos históricos para identificar patrones y tendencias temporales en una variable de interés, proyectando sus valores futuros [5]. Estos modelos son útiles para la predicción de delitos en Medellín, ya que permiten analizar la evolución temporal de la incidencia delictiva y sus factores asociados, así como anticipar posibles escenarios futuros.

3.6.1. ARIMA

Entre los modelos de pronóstico de series de tiempo más utilizados se encuentran los modelos ARIMA (Promedio Móvil Integrado Autorregresivo), el cual es un modelo que combina tres componentes: un componente autorregresivo (AR), que captura la dependencia entre el valor actual y los valores pasados de la serie; un componente de promedio móvil (MA), que captura el efecto de los errores pasados sobre el valor actual, es decir, describe cómo el valor actual puede predecirse a partir de la componente aleatoria actual y de los errores pasados; y un componente de diferenciación (I), que hace que la serie sea estacionaria, es decir, que tenga una media y una varianza constantes a lo largo del tiempo [6]. El modelo ARIMA se representa por tres parámetros: p, d y q.

Los modelos autorregresivos de orden p, denotados como AR(p), describen procesos donde las observaciones actuales son una función lineal de las observaciones pasadas más un término de error aleatorio. El modelo AR(1) más simple se expresa como:

$$X_t = \phi_1 X_{t-1} + a_t$$

Ecuación 1. Modelo autorregresivo de orden 1

Para un modelo AR(p) generalizado:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + a_t$$

Ecuación 2. Modelo autorregresivo de orden p

Un modelo de medias móviles de orden q, denotado como MA(q), describe una serie temporal estacionaria donde el valor actual es una combinación lineal de los términos de error presentes y pasados. Un modelo MA(1) se representa como:

$$X_t = a_t - \nu_1 a_{t-1}$$

Ecuación 3. Modelo de medias móviles de orden 1

Para un modelo MA(q) generalizado:

$$X_t = a_t - \nu_1 a_{t-1} - \nu_2 a_{t-2} - \cdots - \nu_q a_{t-q}$$

Ecuación 4. Modelo de medias móviles de orden q

El modelo ARIMA combina los componentes autorregresivos y de promedio móvil, incluyendo además el componente de diferenciación para manejar la no estacionariedad.

$$\left(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p\right)(1 - B)^d X_t = \left(1 - \nu_1 B - \nu_2 B^2 - \cdots - \nu_q B^q\right) a_t$$

Ecuación 5. Expresión generalizada del modelo ARIMA

Estos modelos son útiles para series temporales con tendencias y estacionalidades, permitiendo un análisis y pronóstico robusto y flexible. Se resalta que la identificación de modelos ARIMA adecuados requiere un análisis detallado de la autocorrelación y la autocorrelación parcial de la serie temporal.

3.6.2. SARIMA

El modelo SARIMA (Seasonal ARIMA) es una extensión del modelo ARIMA que incorpora un componente estacional para capturar la variación periódica de la serie temporal. El modelo SARIMA se caracteriza por seis parámetros: (*p, d, q, P, D*) y (*Q*). Estos parámetros representan, respectivamente, el orden del componente autorregresivo (AR), el grado de diferenciación, el orden del componente de promedio móvil (MA), el orden del componente autorregresivo estacional (AR estacional), el grado de diferenciación estacional y el orden del componente de promedio móvil estacional (MA estacional). Por ejemplo; un modelo SARIMA(1,1,1)(1,1,1,Q) significa que se utiliza: Un término autorregresivo de primer orden (p=1), una diferenciación de primer grado (d=1), un término de promedio móvil de primer orden (q=1), un término autorregresivo estacional de primer orden (P=1), una diferenciación estacional de primer grado (D=1), un término de promedio móvil estacional de primer orden (Q=12) que se refiere a datos mensuales con un patrón estacional anual [7][8].

Este modelo es especialmente útil para series temporales con patrones estacionales claramente definidos, mejorando la precisión de las predicciones al tener en cuenta tanto las tendencias generales como las fluctuaciones estacionales.

Los modelos ARIMA y SARIMA tienen la ventaja de ser flexibles y adaptables a diferentes tipos de series de tiempo, así como de proporcionar intervalos de confianza para las predicciones. Sin embargo, también tienen algunas limitaciones, como la dificultad de elegir los parámetros

óptimos, la sensibilidad a los valores atípicos y la incapacidad de manejar datos no lineales y cambios estructurales en la serie [8].

3.6.3. ARMAX

El modelo ARMAX (Autoregressive Moving Average with Exogenous Variables) combina los componentes AR y MA con variables exógenas (X) que pueden influir en la variable de interés. El modelo ARMAX se representa por la siguiente ecuación:

$$y_t = \beta_0 + \beta_1 x_t + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t$$

Ecuación 6. Expresión generalizada del modelo ARMAX

Donde y_t es la variable de interés en el tiempo t, x_t es una variable exógena en el tiempo t, β_0 y β_1 son los coeficientes de la variable exógena, $\phi_1, \ldots, \phi_p$ son los coeficientes del componente AR, $\theta_1, \ldots, \theta_q$ son los coeficientes del componente MA, y ϵ_t es el término de error en el tiempo t.

El modelo ARMAX tiene la ventaja de poder incorporar información adicional que puede afectar a la variable de interés, mejorando el ajuste y la precisión del modelo. Sin embargo, también tiene algunas limitaciones, como la dificultad de interpretar el coeficiente de la variable exógena, que depende de los valores pasados de la variable de interés y del término de error, y la posibilidad de sufrir de multicolinealidad, es decir, de que las variables exógenas estén correlacionadas entre sí o con la variable de interés [9].

3.6.4. VAR

El modelo vectorial autorregresivo (VAR) es una técnica ampliamente utilizada en econometría y análisis de series temporales para entender y predecir la interrelación entre múltiples variables temporales. En un modelo VAR, cada variable del sistema se expresa en función de sus propios valores pasados y de los valores pasados de todas las demás variables en el sistema. El "orden" del modelo se refiere al número de retardos considerados en las ecuaciones [10]. En un modelo VAR(1) con dos variables, las ecuaciones pueden formularse como:

$$y_{1t} = \beta_{10} + \beta_{11} y_{2t} + \beta_{12} y_{1t-1} + \beta_{13} y_{2t-1} + \epsilon_{1t}$$
$$y_{2t} = \beta_{20} + \beta_{21} y_{1t} + \beta_{22} y_{1t-1} + \beta_{23} y_{2t-1} + \epsilon_{2t}$$

Ecuación 7. Expresión modelo VAR(1) con dos variables

Aquí, cada variable depende de su propio valor pasado y del valor pasado de la otra variable, además de un término de error, en esta expresión se asume la existencia de variables exógenas o determinísticas.

Los modelos VAR pueden extenderse a órdenes superiores, donde las variables explicativas incluyen múltiples retardos. Estos modelos también se pueden descomponerse en modelos univariantes para cada variable y de esta manera, se puede mostrar que cada variable sigue una estructura ARMA, donde los términos de error del modelo VAR también presentan una estructura de ruido blanco, es decir, sin autocorrelación [10].

En un modelo VAR, todas las variables se tratan de manera simétrica, siendo explicadas por los valores pasados de todas las demás. Una vez estimado el modelo, se pueden excluir algunas variables explicativas en función de su significancia estadística. No obstante, es esencial considerar la colinealidad entre variables explicativas, ya que puede reducir la precisión de la estimación.

3.6.5. Prophet

El modelo Prophet de Facebook es una herramienta de código abierto diseñada para la previsión de series temporales, desarrollada por el equipo de ingeniería de Facebook. Esta herramienta se destaca por su capacidad para manejar series temporales con características complejas, como estacionalidades múltiples y cambiantes, y la inclusión de eventos externos como días festivos [11]. Prophet descompone la serie temporal en tres componentes principales: tendencia, estacionalidad y días festivos. La componente de tendencia modela los cambios a largo plazo en la serie, que pueden ser lineales o logísticos, adecuados para datos con límites de capacidad. La componente estacional captura variaciones periódicas que ocurren en periodos específicos, como anuales, semanales o diarias. La componente de días festivos permite incorporar el efecto de eventos importantes que pueden afectar significativamente la serie temporal [12].

En términos de implementación, Prophet ha demostrado ser superior a modelos tradicionales como ARIMA, especialmente en la capacidad de capturar estacionalidades complejas y adaptarse a cambios en las tendencias. Estudios comparativos han mostrado que Prophet ofrece menores errores de predicción y mejor ajuste a los datos históricos en comparación con ARIMA y otros modelos estándar de series temporales. Prophet es una herramienta poderosa y versátil para la predicción de series temporales. Su capacidad para manejar estacionalidades complejas, su robustez frente a valores atípicos y su facilidad de uso lo convierten en una opción preferida en

muchos escenarios de previsión [12]. Al considerar la selección y comparación de modelos de predicción delictiva, incluir Prophet puede aportar una perspectiva robusta y precisa, mejorando así la calidad y fiabilidad de las predicciones.

3.7. Regresión logística

La regresión logística, o Logistic Regression, es un algoritmo de clasificación fundamental en el campo del Machine Learning. Su principal aplicación es la predicción de la probabilidad de una variable dependiente categórica, especialmente cuando dicha variable es binaria, es decir, tiene dos posibles estados como 1 - 0, sí - no, abierto - cerrado, entre otros.

En la regresión logística, la variable dependiente es dicotómica, lo que significa que solo hay dos posibles clases o resultados. Este modelo es ampliamente utilizado en situaciones donde se necesita calcular la probabilidad de que ocurra un evento dado. La naturaleza dicotómica facilita la interpretación de resultados en términos probabilísticos, proporcionando una comprensión clara de la relación entre las variables predictoras y la variable objetivo [13].

La regresión logística se basa en la función logística, también conocida como función sigmoide. Esta función es una curva en forma de S que puede asignar cualquier valor real a un rango entre 0 y 1. La función sigmoide se define como:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Ecuación 8. Representación matemática de la función sigmoide

Donde $\sigma(x)$ representa la probabilidad de que la variable dependiente tome el valor de 1 para un conjunto dado de variables independientes. Algunas características clave de la regresión logística incluyen:

- Variable de salida binaria: Destinada para problemas de clasificación binaria.
- Eliminación de ruido: No asume errores en la variable de salida, recomendando la eliminación de valores atípicos en los datos de entrenamiento.
- Distribución gaussiana: Aunque es un algoritmo lineal, una transformación no lineal en la salida puede mejorar la precisión del modelo.

- Eliminar entradas correlacionadas: Similar a la regresión lineal, la presencia de múltiples entradas altamente correlacionadas puede llevar al sobreajuste. Es importante calcular y eliminar las entradas altamente correlacionadas.
- Problemas de convergencia: La estimación de probabilidad puede no converger si hay muchas entradas altamente correlacionadas o si los datos son muy escasos.

3.8.LightGBM

LightGBM, conocido como Light Gradient Boosting Machine, es un marco de trabajo de alto rendimiento desarrollado por Microsoft que se centra en algoritmos de aprendizaje basados en árboles. Destaca por su eficiencia distribuida y optimización, siendo ampliamente utilizado para tareas de clasificación y regresión en ML. Una de sus características distintivas es su capacidad para manejar grandes volúmenes de datos de manera eficiente en memoria, lo que lo convierte en una opción popular entre los practicantes de aprendizaje automático. LightGBM ofrece un análisis de importancia de características, proporcionando dos tipos: la importancia de "Split", que cuenta cuántas veces se utiliza una característica para dividir los datos en todos los árboles del modelo; y la importancia de "Gain", que cuantifica la mejora en la precisión del modelo al utilizar una característica específica para dividir, ofreciendo una visión más detallada y precisa de la importancia de las características según la calidad de las divisiones realizadas [14].

3.9. Redes Neuronales

Surgen como alternativa los modelos de aprendizaje automático que imitan el funcionamiento del cerebro humano, mediante una combinación de ciencia informática y estadística para resolver problemas en el campo de la inteligencia artificial. Una red neuronal consta de varias capas de unidades de procesamiento llamadas neuronas, que reciben, procesan y transmiten información. Las neuronas se conectan entre sí por pesos sinápticos, que se ajustan mediante un proceso de aprendizaje basado en los datos de entrada y salida [15]. Una capa de interés es la capa LSTM (Long Short-Term Memory), esta capa está presente en una red neuronal recurrente (RNN) que se utiliza para procesar datos secuenciales, es decir, datos donde el orden cronológico es importante, como en la predicción de delitos.

Una RNN tiene la capacidad de recordar la información pasada y utilizarla para la predicción actual, mediante un bucle de retroalimentación que conecta la salida con la entrada, como se puede apreciar en la Ilustración 1. Sin embargo, una RNN simple tiene problemas para aprender dependencias a largo plazo, debido al problema del desvanecimiento o la explosión del gradiente. Una LSTM resuelve este problema mediante una estructura de celda de memoria, que consta de tres compuertas: una compuerta de entrada, que decide qué información se almacena en la celda; una compuerta de olvido, que decide qué información se elimina de la celda; y una compuerta de salida, que decide qué información se envía a la capa siguiente [16]. El uso de esta red neuronal, permite aprender secuencias largas y complejas, así como de evitar el sobreajuste y la dependencia de los datos de entrenamiento, lo cual es necesario en la predicción de una serie temporal como lo es la cantidad de delitos.

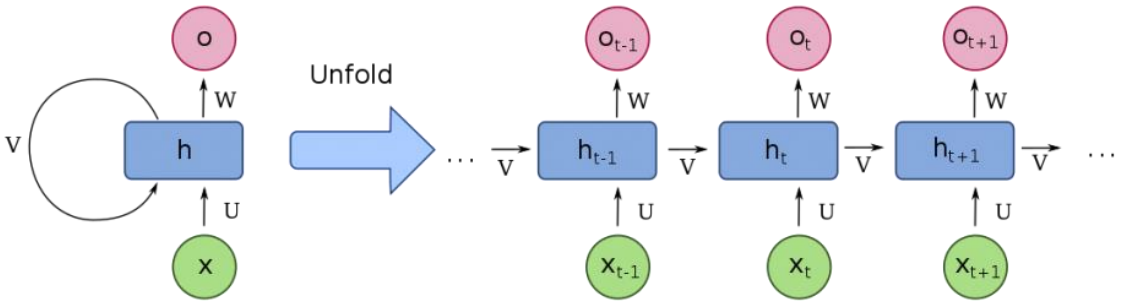


Ilustración 1. Diagrama de una Red Neuronal Recurrente (RNN) mostrando la propagación de estados ocultos a lo largo del tiempo [16]

La selección y comparación de los modelos propuestos se llevará a cabo mediante la aplicación de criterios clave que aborden aspectos fundamentales en el contexto de la predicción delictiva. La precisión, medida por métricas como el Error Cuadrático Medio (RMSE), será un criterio central para evaluar el rendimiento de cada modelo en la predicción de delitos en la ciudad de Medellín. Además, se evaluará la capacidad de generalización, es decir, la habilidad de los modelos para realizar predicciones precisas en nuevos conjuntos de datos. Para complementar esta evaluación, también se realizará un análisis de los residuos de cada predicción, lo que permitirá identificar posibles patrones no capturados por los modelos y garantizar que los errores sean distribuidos de manera aleatoria, asegurando así la robustez y fiabilidad de los modelos seleccionados.

4. TRABAJOS RELACIONADOS

La literatura actual sobre la predicción de delitos utilizando modelos de aprendizaje automático proporciona un marco robusto y variado que es esencial para el desarrollo del presente trabajo en Medellín. A continuación, se presenta un análisis detallado e integrado de los estudios más relevantes en esta área, destacando sus metodologías, resultados y recomendaciones para futuros trabajos.

Un estudio significativo realizado en Medellín evaluó tres modelos de Machine Learning: un clasificador de bosques aleatorios, un modelo de regresión logística y una máquina de vectores de soporte (SVM), para la predicción de crímenes en la ciudad. La metodología implementada siguió el proceso estándar CRISP-DM (Cross Industry Standard Process for Data Mining), una estrategia general utilizada en la industria para la minería de datos. Este estudio utilizó datos históricos sobre incidentes delictivos y la tasa de desempleo, enfocándose en la predicción del hurto a personas en modalidades específicas como atraco, descuido, cosquilleo y raponazo. Los resultados mostraron que el modelo basado en SVM fue el más efectivo, alcanzando un F1-Score del 76,06% y un Recall del 80%, superando tanto a los otros modelos como a un modelo basado en reglas. Este estudio subraya la importancia del registro preciso de la información por parte de las autoridades para desarrollar algoritmos predictivos confiables. Además, se sugirió la inclusión de variables adicionales, como eventos especiales en la ciudad, para mejorar la capacidad predictiva de los modelos futuros [17].

Otro trabajo relevante se llevó a cabo en Bogotá, donde se desarrollaron dos modelos de Machine Learning para analizar los patrones de criminalidad durante el periodo post-pandemia, desde enero 2021 hasta mayo 2023. El primer modelo, un clasificador, permitió identificar la ocurrencia de delitos de alto impacto (con un F1-score entre 0.70 y 0.80), mientras que el segundo modelo, una regresión, predijo la cantidad de estos delitos con un error absoluto medio (MAE) entre 0.2 y 3.13. Los datos utilizados provinieron de la Secretaría Distrital de Seguridad de Bogotá y la Encuesta Multipropósito Bogotá-Cundinamarca, combinando variables categóricas como fecha, año, rango del día, localidad, sexo y arma empleada, con variables numéricas como ingreso per cápita, número de personas, pobreza multidimensional y desempleo. Este enfoque proporciona herramientas basadas en inteligencia artificial para apoyar la toma de decisiones en seguridad

ciudadana, y destaca cómo los datos socioeconómicos pueden ser integrados con datos delictivos para mejorar la precisión de los modelos predictivos [18].

En Bucaramanga, se aplicaron modelos de ML para predecir delitos utilizando procesamiento de señales para grafos y adaptaciones del modelo TF-IDF. Los mejores resultados se obtuvieron utilizando modelos espaciales de grafos semanales, en particular, el modelo de clasificación KNN destacó al lograr un recall del 59% y una precisión superior al 60%. Este estudio resalta las limitaciones y desafíos de aplicar modelos predictivos en ciudades con menor cantidad de datos disponibles en comparación con las ciudades principales. Los autores concluyen que, aunque los modelos de predicción del delito son herramientas útiles para construir estrategias de prevención, es esencial tener en cuenta la calidad y cantidad de los datos disponibles [19].

Además de estos estudios específicos de ciudades colombianas, otros trabajos han explorado diversas técnicas y enfoques para la predicción de delitos en diferentes contextos. Por ejemplo, en un estudio, se utilizaron modelos ARIMA, tradicionales en el análisis de series de tiempo, para predecir incidentes de desorden público y asignar recursos de manera efectiva, demostrando cómo las técnicas de series temporales pueden contribuir significativamente a la planificación de la seguridad [20]. En el estudio "Urban Infrastructure Safety System Based on Mobile Crowdsensing", los autores utilizaron técnicas de aprendizaje automático para modelar el comportamiento de los ladrones en sistemas de transporte público, empleando datos de cámaras y tarjetas de identificación para predecir y prevenir delitos [21].

Una revisión sistemática elaborada por Kounadi evaluó estrategias de análisis espacial del crimen, utilizando enfoques estadísticos y algoritmos de aprendizaje automático como bosques aleatorios, perceptrón multicapa (MLP) y máquinas de vectores de soporte (SVM), para identificar y analizar puntos calientes de criminalidad [22]. Esta revisión destaca cómo los enfoques espaciales y temporales pueden ser combinados para ofrecer predicciones más precisas y contextualizadas, proporcionando una base teórica sólida para la implementación de modelos predictivos en diferentes contextos urbanos.

5. METODOLOGÍA

En esta sección, se describen los métodos y procedimientos utilizados para evaluar el desempeño de diversos modelos de aprendizaje automático en la predicción de delitos en la ciudad de Medellín. La metodología sigue un enfoque estructurado que abarca desde la recopilación y preparación de datos hasta la implementación, ajuste y evaluación de los modelos predictivos.

La literatura revela múltiples técnicas para predecir crímenes en una ciudad, incluyendo modelos autorregresivos y de aprendizaje automático. Estos modelos han sido empleados para estimar el número de crímenes a lo largo del tiempo, proporcionando herramientas valiosas a las autoridades para mejorar la gestión de recursos y la respuesta ante incidentes.

Además, se han explorado inferencias relacionadas con las "zonas calientes" en la ciudad y la predicción del número y la tasa de crímenes. Estas perspectivas buscan ofrecer a las autoridades una mayor comprensión de la distribución de delitos a corto y largo plazo en la región, optimizando así sus estrategias de gestión de recursos y atención oportuna de incidentes.

Es importante destacar que algunos estudios previos no han proporcionado todos los detalles necesarios para reproducir sus resultados. Por lo tanto, este trabajo se enfoca en ofrecer una descripción detallada del proceso metodológico para garantizar la reproducibilidad y validez de los hallazgos.

5.1. Comprensión y Recopilación de Datos

El primer paso consiste en la comprensión y recopilación de datos. Los datos utilizados en este estudio provienen de múltiples fuentes, incluyendo la plataforma MEData, que proporciona datos históricos de delitos en Medellín, y el Sistema de Alerta Temprana de Antioquia (SIATA), que ofrece datos históricos meteorológicos. Además, se han identificado y recopilado variables de control relevantes, tales como la cercanía a fechas de pago, días feriados y fines de semana, las cuales podrían influir en la ocurrencia de delitos.

5.1.1. Datos Históricos de Delitos

La recopilación de datos históricos de delitos se realizó a través de la plataforma MEData. Se extrajeron registros detallados que incluyen la ubicación, la fecha y la hora del incidente. Esta información es crucial para la identificación de patrones temporales y espaciales en la criminalidad.

Todos estos datos históricos de delitos obtenidos a través de la plataforma fueron recopilados en la fase temprana del proyecto "Administración inteligente de problemas de seguridad ciudadana a través de modelos y herramientas generadas a partir de plataformas para territorios inteligentes apoyadas por estrategias de participación ciudadana en la ciudad de Medellín" para su aplicación de zonas seguras, del cual este proyecto de investigación hace parte.

En la Ilustración 2, se presenta gráficamente la solicitud de datos para su análisis. La información se recopiló en el back-end del proyecto de ciudades inteligentes. En el lado del back-end, se almacenan los registros históricos de delitos en una base de datos organizada. El proceso comienza con una solicitud específica para el análisis de datos. Esta solicitud incluye información sobre las zonas para las cuales se requiere el análisis y el back-end responde proporcionando los datos solicitados.

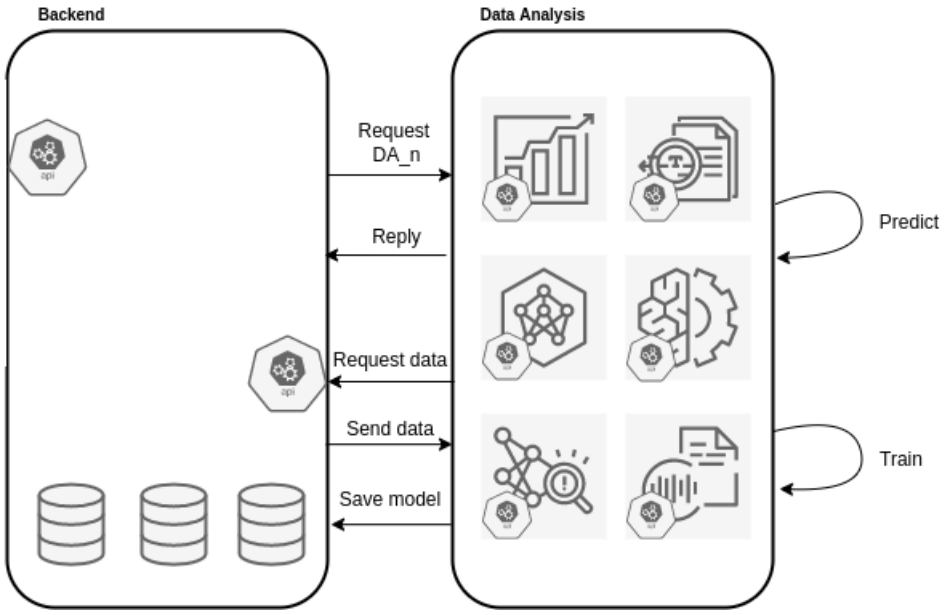


Ilustración 2. Solicitud de datos a Back-end para entrenamiento y predicciones de los modelos de ML

Para estructurar estos datos de manera efectiva, en el proyecto de investigación mencionado se optó por dividir el territorio de Medellín en hexágonos, como se presenta en la Ilustración 3, cada uno de estos hexágonos posee un identificador único, un enfoque que facilita el análisis espacial y la predicción de delitos. La decisión de dividir el territorio por hexágonos se basó en varias consideraciones metodológicas y prácticas. La utilización de una división hexagonal, en lugar de una cuadrícula tradicional, ofrece varias ventajas significativas en el análisis espacial. Los

hexágonos son polígonos regulares que cubren el plano de manera uniforme sin superposición ni espacios vacíos. Esto permite una cobertura consistente del área de estudio, asegurando que cada parte del territorio se considere por igual, lo cual es beneficioso para análisis que dependen de la proximidad espacial, como la predicción de delitos, ya que facilita la identificación de patrones y tendencias.

Los hexágonos pueden adaptarse mejor a la geometría irregular de los patrones urbanos y naturales, proporcionando una herramienta más flexible para el análisis de áreas con formas no rectangulares. Esto es especialmente relevante en una ciudad como Medellín, que tiene una topografía compleja y una estructura urbana densa.

Ilustración 3. División Hexagonal del Territorio

La aplicación de zonas seguras, utilizando esta estructura hexagonal, recopiló y agrupó datos de delitos para cada hexágono. Este proceso implicó un compilado exhaustivo de todos los delitos que ocurrieron dentro de cada área hexagonal desde el año 2012 hasta la fecha. Esta agrupación de datos facilita el análisis de la distribución espacial de los delitos y permite una evaluación más granular de las áreas de alta incidencia criminal, conocidas como "zonas calientes". La recopilación de datos incluyó múltiples aspectos detallados como la ubicación precisa de cada incidente, mapeada a las coordenadas geográficas correspondientes dentro del hexágono, y las fechas exactas de los incidentes para permitir análisis temporales detallados.

Esta información detallada y estructurada es fundamental para desarrollar modelos predictivos precisos. Al realizar esta división del territorio, se facilita la identificación de patrones espaciales y temporales en la ocurrencia de delitos. Esta metodología no solo mejora la precisión

de las predicciones, sino que también proporciona una base sólida para la implementación de estrategias de prevención del delito basadas en datos.

Cada hexágono presenta una serie temporal que refleja la cantidad diaria de delitos, proporcionando una perspectiva detallada y continua de la actividad delictiva en diferentes áreas de la ciudad. Los hexágonos se dividieron en cuartiles con base en la cantidad total de delitos registrados con el fin de obtener una clasificación más precisa de las áreas según sus niveles de criminalidad, lo que facilita el análisis comparativo entre zonas de alta y baja incidencia delictiva. Además, esta segmentación ayuda a adaptar los modelos predictivos a las características específicas de cada grupo de hexágonos, mejorando la precisión y la relevancia de las predicciones.

5.1.2. Datos Históricos Meteorológicos

Los datos meteorológicos históricos fueron obtenidos del SIATA, que proporciona información sobre variables como la temperatura y la precipitación, indicada por dos pluviómetros. Estas variables se integraron al conjunto de datos para analizar su posible impacto en la ocurrencia de delitos. Además, el SIATA ofrece registros de todas las estaciones meteorológicas activas, incluyendo sus coordenadas geográficas. Utilizando esta información, se asociaron las estaciones meteorológicas con cada hexágono del territorio basándose en su proximidad, con el fin de enriquecer las series temporales. La ilustración 4 muestra una visualización general del territorio dividido en hexágonos junto con la ubicación de cada estación. Los datos históricos de estas estaciones fueron consultados a través del portal web del SIATA y recopilados considerando la fecha de activación de cada una de ellas.

5.1.3. Variables de Control

Se identificaron variables de control adicionales que podrían tener un impacto significativo en la predicción de delitos. Además de factores como las condiciones climáticas y la incidencia delictiva histórica, se incorporaron elementos como la proximidad a fechas de pago, días feriados y fines de semana. La integración de estas variables en los modelos analíticos permite realizar ajustes que capturan variaciones contextuales específicas y ciclos económicos relevantes para la población. Esto enriquece la capacidad predictiva de los modelos al considerar no solo los patrones históricos de delitos, sino también los factores temporales y sociales que potencialmente influyen en la incidencia delictiva.

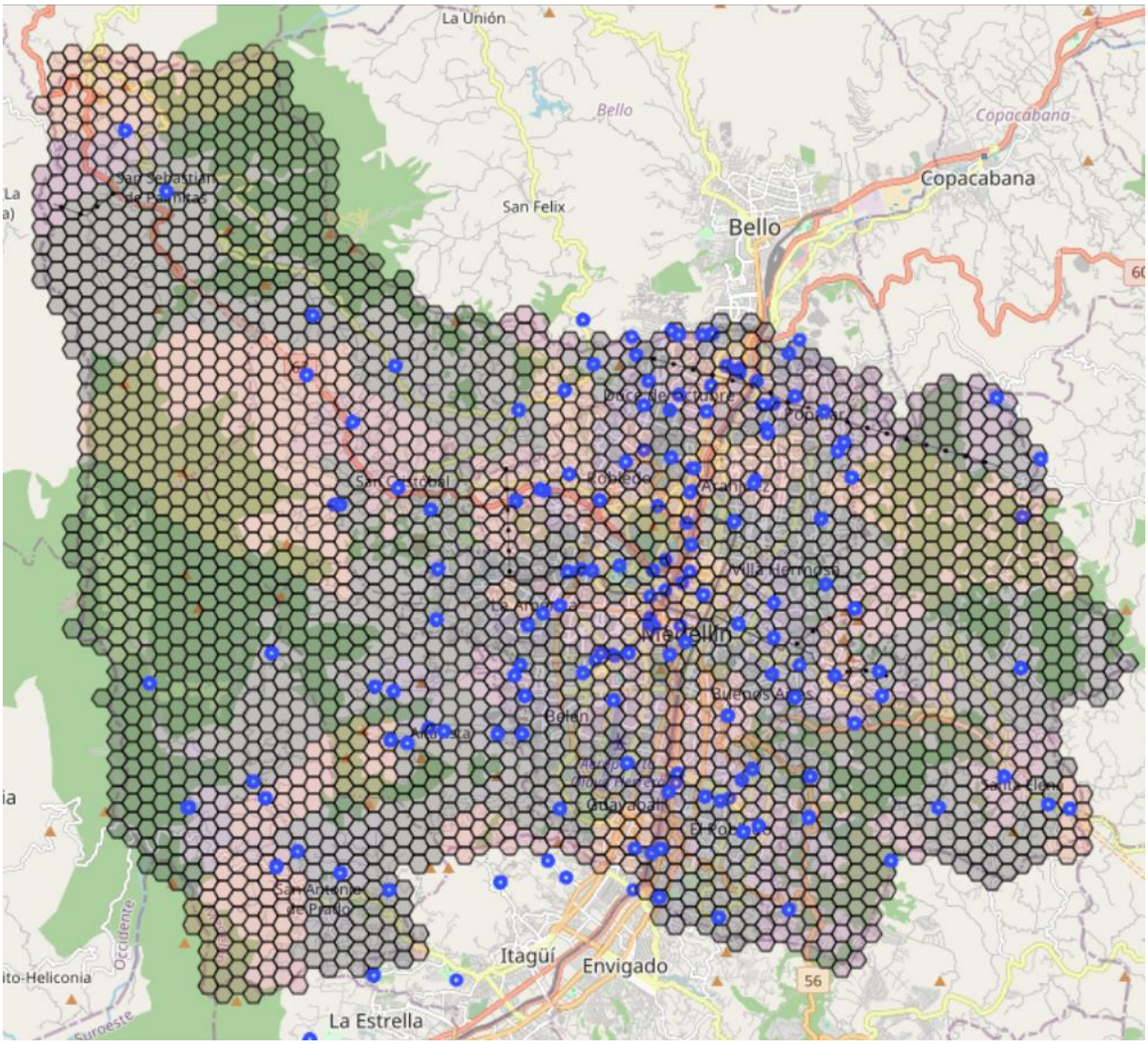


Ilustración 4. División hexagonal del territorio e identificación de estaciones meteorológicas

5.2. Configuración del sistema y librerías usadas

Para la realización de los experimentos se utilizó Python 3.10.12.

5.2.1 Características del sistema de computo

Los experimentos se llevaron a cabo en una CPU de Google Colab, utilizando el backend de Google Compute Engine que emplea Python 3. Las especificaciones del sistema son las siguientes:

- Sistema Operativo: Linux
- Versión del Kernel: 6.1.85+
- Versión del Sistema: #1 SMP PREEMPT_DYNAMIC Thu Jun 27 21:05:47 UTC 2024
- Procesador: x86_64
- Memoria RAM Total: 12.67 GB
- Espacio en Disco Total: 107.72 GB

5.2.2 Librerías utilizadas

Para el desarrollo y la ejecución de los experimentos, se emplearon las siguientes librerías y herramientas:

Regresión Logística:

- `sklearn.linear_model`: Librerías para implementar la regresión logística.
- `sklearn.metrics`: Proporciona funciones para evaluar modelos como un reporte de clasificación, precisión y matrices de confusión .
- `sklearn.model_selection`: Facilita la división del conjunto de datos en entrenamiento y prueba.
- `sklearn.preprocessing import MinMaxScaler`: Para la implementación de un escalador que normaliza los datos para mejorar el rendimiento del modelo.

LightGBM:

- `lightgbm`: Biblioteca de gradiente boosting para tareas de clasificación y regresión.

Modelos ARIMA:

- `statsmodels.tsa.arima.model`: Proporciona el modelo ARIMA para análisis de series temporales.

- `statsmodels.tsa.statespace.sarimax`: Extensión de ARIMA que soporta estacionalidad y regresores exógenos.
- `statsmodels.tsa.stattools`: Funciones estadísticas para análisis de series temporales, incluye ACF, PACF y el test adfuller.

VAR:

- `statsmodels.tsa.vector_ar.var_model`: Modelo VAR para series temporales multivariadas.

Prophet:

- `sktime.forecasting.fbprophet`: Herramienta de predicción de series temporales desarrollada por Facebook.
- `sktime.forecasting.base`: Utilizado para definir horizontes de predicción en modelos de series temporales.

Redes Neuronales:

- ` keras.models`: Permite la creación de modelos de redes neuronales.
- `from keras.layers import LSTM, GRU, Dense, Dropout`: Capas comunes para construir redes neuronales recurrentes y profundas.
- `from keras.callbacks import EarlyStopping`: Detiene el entrenamiento cuando no hay mejoras en el rendimiento.
- `from keras.optimizers import Adam`: Optimizador basado en el descenso de gradiente.
- `from keras.regularizers import l1_l2`: Regularización para prevenir sobreajuste.

Manipulación y Análisis de Datos:

- `pandas`: Biblioteca para manipulación y análisis de datos estructurados.
- `numpy `: Biblioteca para cálculo numérico y manipulación de matrices.
- `datetime`: Biblioteca estándar de Python para manejar fechas y horas.

Visualización de Datos:

- `matplotlib `: Herramienta para la generación de gráficos estáticos.
- `seaborn`: Biblioteca basada en matplotlib para visualización de datos estadísticos.
- `folium`: Utilizada para la visualización geoespacial y la generación de mapas interactivos.

Solicitudes HTTP:

- `requests`: Biblioteca para hacer solicitudes HTTP y trabajar con APIs.

Estas herramientas y librerías fueron esenciales para la implementación de los diferentes modelos y análisis realizados en el proyecto. La combinación de estas tecnologías permitió una exploración exhaustiva y la obtención de resultados precisos y reproducibles.

5.3. Análisis de Características

Antes de la implementación de los modelos, se llevó a cabo un análisis de características para comprender mejor la naturaleza de los datos y las relaciones entre las diferentes variables. Para el análisis de las características se utilizó un hexágono ubicado en el sector de La Candelaria, el cual corresponde al cuartil 4, que contiene a los hexágonos con mayor cantidad de delitos.

La primera etapa del análisis consistió en transformar el problema de predicción de delitos en una tarea de clasificación, considerando las tasas altas y bajas de delitos, se determinó que el 39.94% corresponde a una tasa alta y el 60.06% a una tasa baja. Esta clasificación se realizó promediando la cantidad de delitos, con el fin de crear un modelo inicial. Este enfoque permitió evaluar las características que contribuyen significativamente a la toma de decisiones y determinar si las variables adicionadas a la serie temporal aportan información valiosa o simplemente introducen ruido al problema.

La clasificación se establece asignando la etiqueta "1" a la tasa alta y "0" a la tasa baja. Se implementaron modelos de regresión logística y LightGBM para realizar la clasificación. Sin embargo, el modelo de regresión logística mostró dificultades al predecir la clase "1". Ante esta limitación, se decidió mejorar el modelo de regresión logística mediante la incorporación de regularizaciones L1 (Lasso) y L2 (Ridge) y abordar el desequilibrio en las clases utilizando un balanceador.

La introducción de LightGBM como modelo para la clasificación agrega una dimensión interesante a la evaluación. La ventaja significativa de LightGBM radica en su capacidad para generar métricas como la importancia de características, estas métricas proporcionan una visión detallada sobre qué variables están influyendo más en la toma de decisiones del modelo, lo cual es esencial para evaluar la contribución de las variables introducidas en el problema de predicción de la cantidad de delitos. Para el modelo de clasificación inicial se compilaron los resultados en la Tabla 1.

Modelo	Accuracy Score	Precision (Clase 0)	Precision (Clase 1)	Recall (Clase 0)	Recall (Clase 1)	F1-score (Clase 0)	F1-score (Clase 1)
Regresión Logística	0.6399	0.64	0.67	1.00	0.01	0.78	0.02
Regresión Logística Balanceada	0.5590	0.70	0.42	0.55	0.58	0.61	0.49
LightGBM	0.6443	0.65	0.59	0.98	0.05	0.78	0.10

Tabla 1. Métricas obtenidas en problema de clasificación inicial

El modelo de regresión logística predijo correctamente 739 casos de la clase 0 y 4 casos de la clase 1, pero se confundió en 416 casos de la clase 0 con la clase 1, y en 2 casos de la clase 1 con la clase 0, como se presenta en la Ilustración 5. La escasa precisión en la clase 1 se debe a la ausencia de verdaderos positivos, lo que muestra la importancia de tener en cuenta el contexto y la proporción entre las clases al analizar la matriz de confusión.

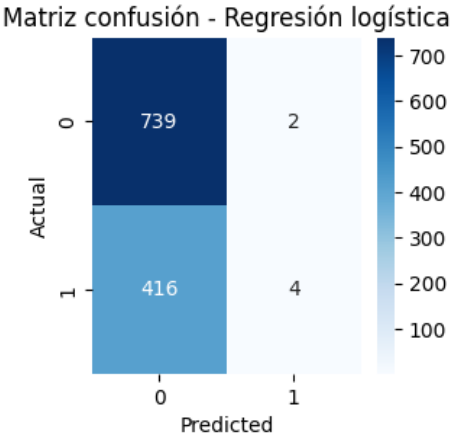


Ilustración 5. Matriz de confusión para regresión logística

Como existe una diferencia muy grande entre el número de elementos de cada clase, esto puede estar afectando al rendimiento del modelo de clasificación, ya que puede sesgar hacia la clase mayoritaria y no aprender bien la clase minoritaria. Después de aplicar el balanceo y probar el modelo, se puede observar en la Ilustración 6 que el modelo predijo correctamente 407 casos de la clase 0 y 242 casos de la clase 1, pero se ha confundido en 334 casos de la clase 0 con la clase 1, y en 178 casos de la clase 1 con la clase 0.

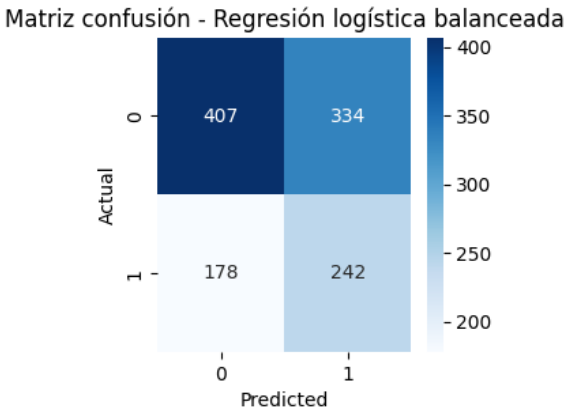


Ilustración 6. Matriz de confusión para regresión logística con balanceo

Estos resultados indican que el balanceo de clases ha mejorado la capacidad del modelo para predecir la clase minoritaria, pero ha empeorado la capacidad para predecir la clase mayoritaria. Esto indica que se requiere un punto óptimo entre el balanceo y el desbalanceo de clases que maximice el rendimiento del modelo.

LightGBM tiene un rendimiento similar al de la regresión logística, sobresaliendo en la predicción de la clase mayoritaria (tasa baja de delitos). Sin embargo, tiene problemas para identificar la clase minoritaria (tasa alta de delitos), aunque los resultados son un poco mejores. La importancia de estos resultados radica en el análisis de relevancia de las características usadas en el problema de clasificación, como se muestra en la Ilustración 7 y la Ilustración 8.

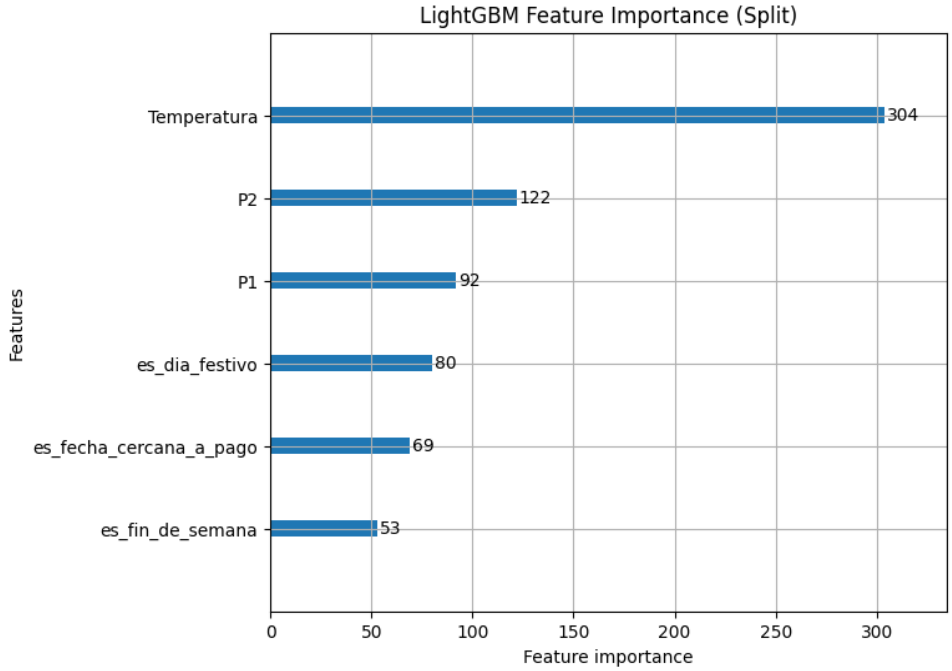


Ilustración 7. Importancia de características con el enfoque Split

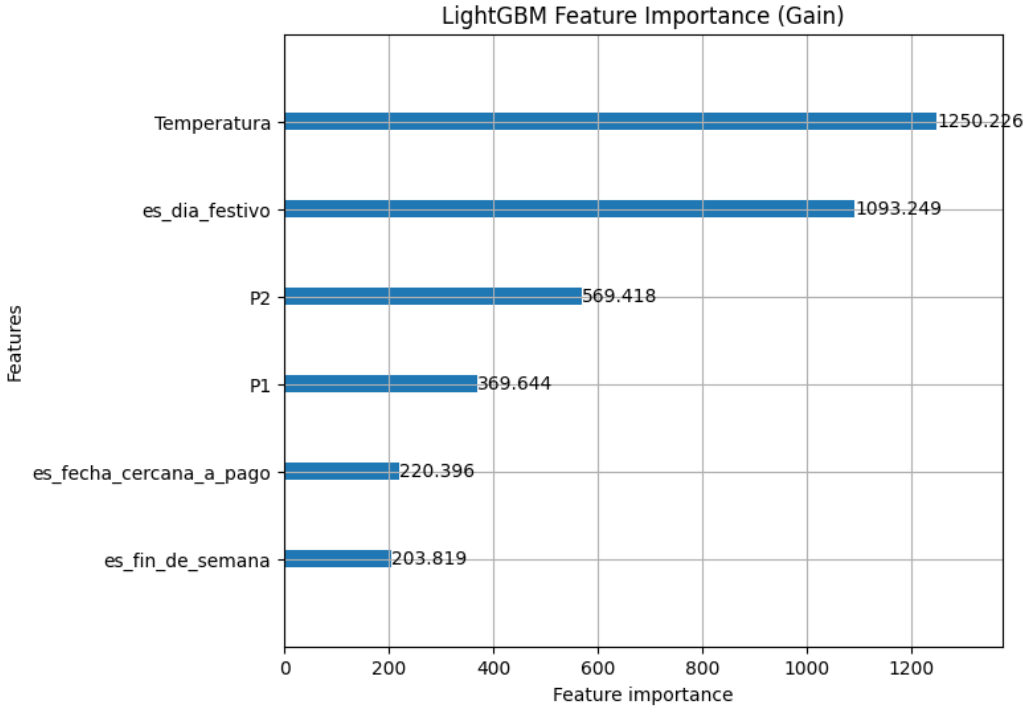


Ilustración 8. Importancia de características con el enfoque Gain

Al evaluar la importancia de las características con las métricas de Split y Gain del modelo LightGBM, se observa que la característica más relevante en el modelo es la temperatura. Este hallazgo resalta la influencia significativa de la temperatura en la predicción de la tasa de delitos, apoyado visualmente en la Ilustración 9 para explorar la relación entre el número de delitos y la temperatura en grados Celsius (°C).

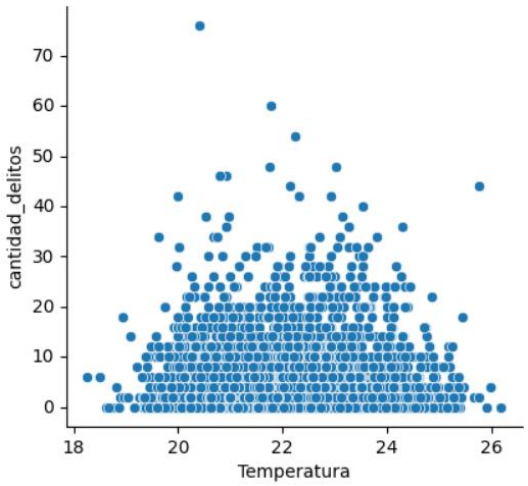


Ilustración 9. Relación temperatura y cantidad de delitos

Además, la característica relacionada con la precipitación, presentada por los pluviómetros P1 y P2, también desempeñan un papel crucial en el modelo, indicando que la cantidad de lluvia está correlacionada con la tasa de delitos. En la Ilustración 10, se presenta la relación entre la cantidad de delitos y la precipitación en milímetros (mm):

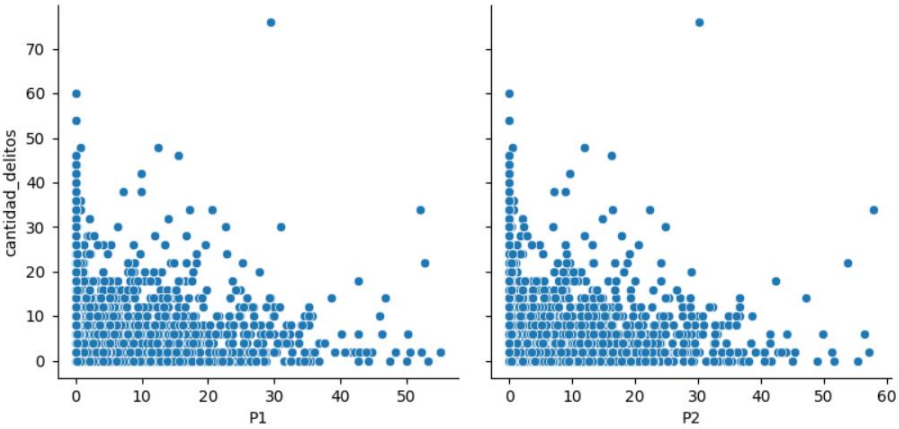


Ilustración 10. Relación precipitaciones y cantidad de delitos

Las variables que indican si la fecha corresponde a un fin de semana, día festivo o fecha cercana a pago tienen importancias relativamente menores en el modelo. Esto sugiere que, en términos de predicción de la tasa de delitos, la distinción entre días hábiles y fines de semana, así como los días festivos, tiene una influencia más tenue en comparación con las variables meteorológicas; sin embargo, se logró observar que existe un impacto de los eventos especiales y ciclos económicos en la dinámica delictiva.

Estos resultados ofrecen una visión integral de la capacidad predictiva de los modelos, destacando la relevancia de las variables meteorológicas en la predicción de la tasa de delitos, así como la importancia de considerar el equilibrio entre las clases al interpretar los resultados de la clasificación.

5.4. Análisis de estacionariedad

Para poder aplicar correctamente los modelos de forecasting y ML, es fundamental asegurar la estacionariedad de los datos, donde el tiempo, la varianza y la covarianza exhiben independencia temporal. Con este fin, se llevó a cabo una prueba de Dickey-Fuller en múltiples hexágonos distribuidos en diferentes cuartiles. Los resultados presentados en la Tabla 2, revelaron un valor de p estadísticamente significativo ($p < 0,05$), indicando una fuerte evidencia en contra de la hipótesis nula de no estacionariedad. Esto sugiere que el orden óptimo de diferenciación es cero y confirma que la serie temporal es estacionaria, como se evidencia también por el hecho de que el estadístico de prueba es menor que los valores críticos. Este hallazgo es consistente a través de todos los hexágonos analizados, lo que valida la estacionariedad de los datos y permite proceder con seguridad al entrenamiento de los modelos de predicción.

Hexágono por cuartil	Test Statistic	p-value	Critical Value (1%)	Critical Value (5%)	Critical Value (10%)
Altavista (Q1)	-46.292548	0.000000	-3.433405	-2.862890	-2.567488
San Cristóbal (Q2)	-61.337260	0.000000	-3.432103	-2.862315	-2.567182
Laureles (Q3)	-10.568317	1.464246 e-28	-3.432020	-2.862278	-2.567163
La Candelaria (Q4)	-4.126592	0.000875	-3.432054	-2.862293	-2.567171

Tabla 2. Resultados del test Dickey-Fuller para diferentes hexágonos

5.5. Aplicación de Modelos

Para la aplicación de los modelos de machine learning y forecasting, se seleccionaron los hexágonos presentados en la Tabla 2. Estos hexágonos fueron escogidos de diferentes cuartiles según la cantidad de delitos históricos registrados en cada área. La Ilustración 11 muestra la ubicación exacta de cada uno de estos hexágonos, junto con la estación meteorológica más cercana. Esta información es crucial para contextualizar los datos y garantizar la precisión de los modelos al considerar tanto la distribución geográfica de los delitos como las condiciones meteorológicas locales.

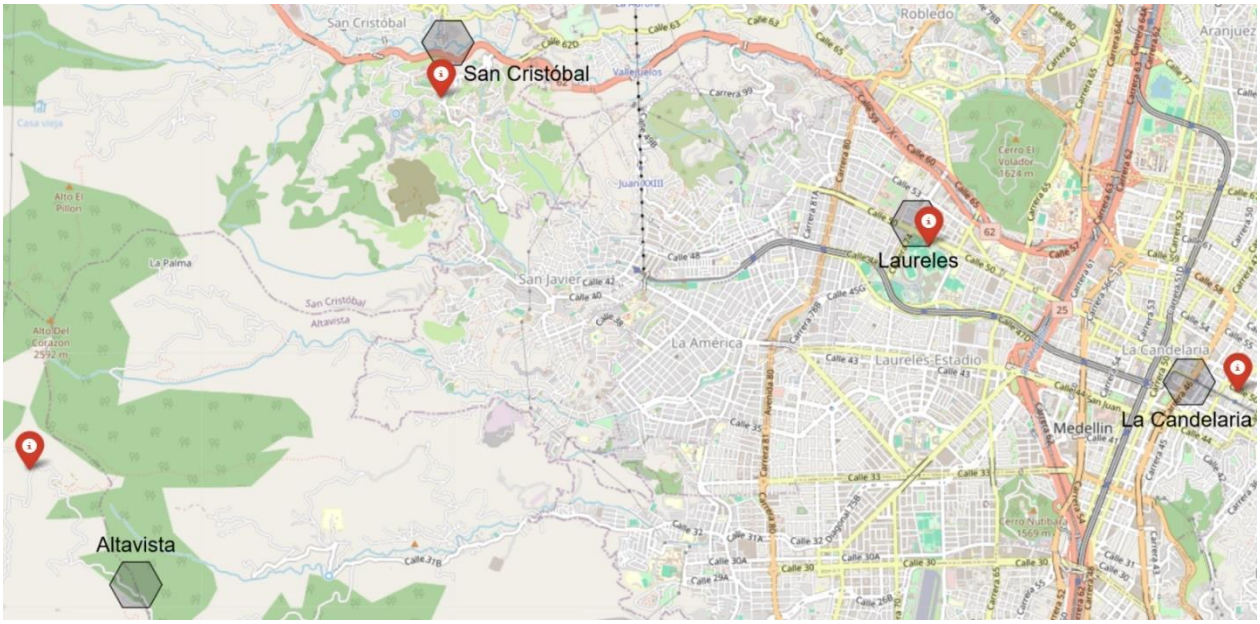


Ilustración 11. Hexágonos seleccionados para análisis

En las ilustraciones 12, 13, 14 y 15, se presentan gráficamente las series temporales que muestran la cantidad de delitos históricos diarios en cada hexágono. Es importante destacar que no todos los hexágonos disponen de la misma cantidad de datos, lo cual puede influir significativamente en el entrenamiento de los modelos.

Cuando se tienen menos datos disponibles para ciertos hexágonos, los modelos pueden enfrentar dificultades para captar patrones consistentes y generar predicciones precisas. La insuficiencia de datos puede llevar a una mayor variabilidad en los resultados y a un riesgo incrementado de sobreajuste, donde el modelo se adapta demasiado a las peculiaridades del conjunto de entrenamiento en lugar de generalizar bien a nuevos datos. Por lo tanto, la cantidad

desigual de datos entre los hexágonos debe considerarse cuidadosamente durante el proceso de modelado. También se resalta que los hexágonos del cuartil 1 poseen una cantidad histórica de delitos muy baja, entre 0 y 1. A pesar de esta baja incidencia, se realizará el análisis para tener una comprensión más completa y representativa de la distribución delictiva en toda el área de estudio. Ignorar estos datos podría sesgar los resultados y las conclusiones, limitando la efectividad de cualquier estrategia de intervención.

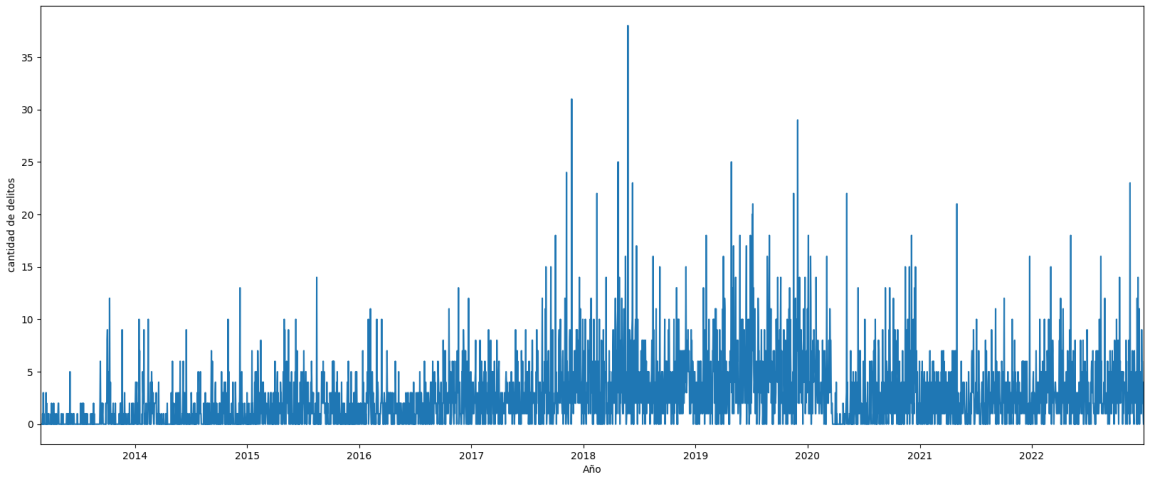


Ilustración 12. Serie temporal hexágono La Candelaria Q4

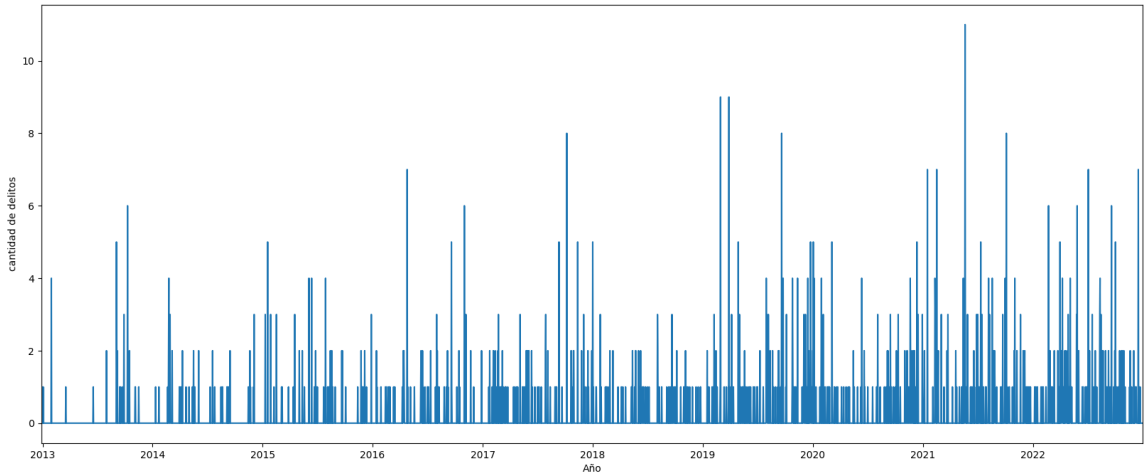


Ilustración 13. Serie temporal hexágono Laureles Q3

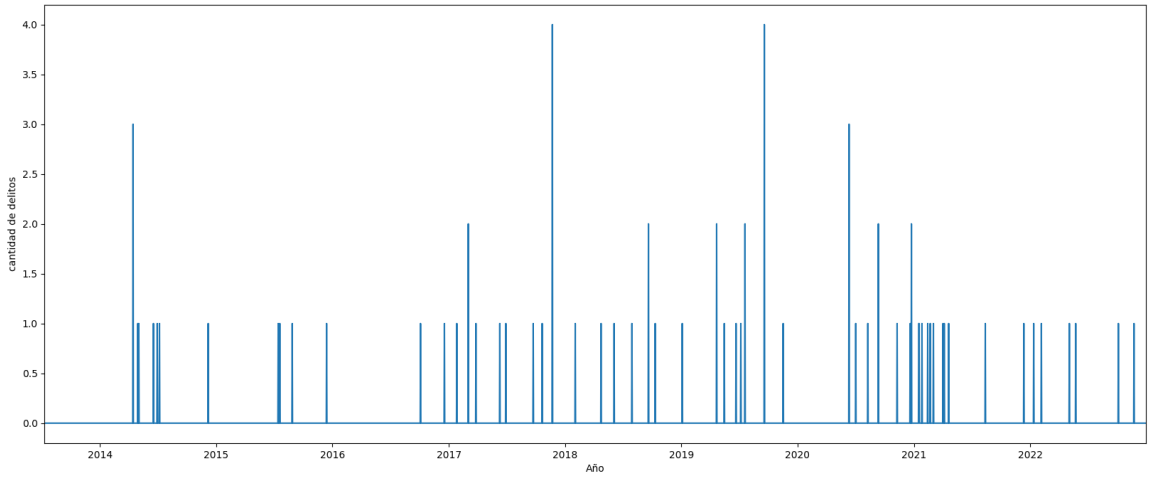


Ilustración 14. Serie temporal hexágono San Cristóbal Q2

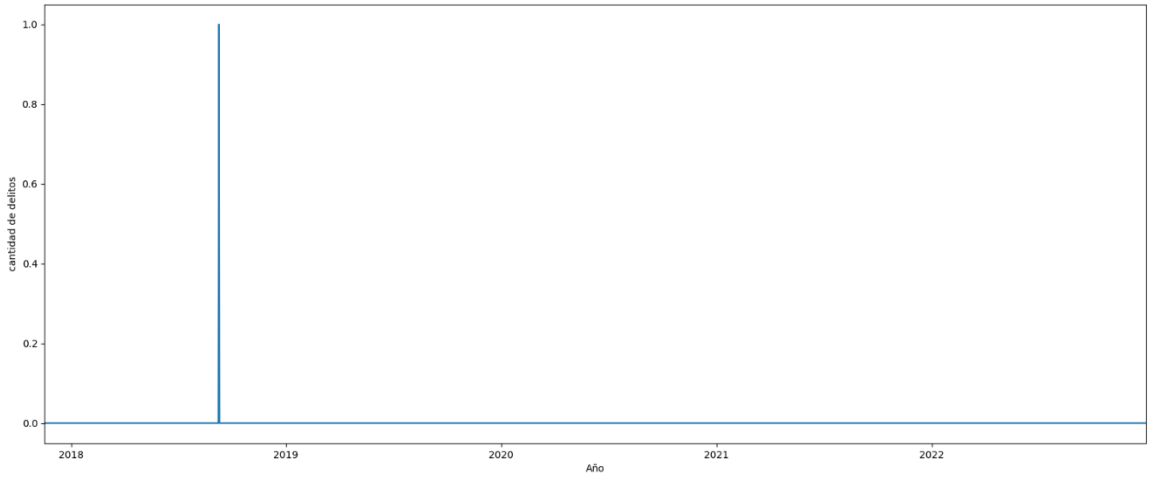


Ilustración 15. Serie temporal hexágono Altavista Q1

5.5.1. Aplicación modelo ARIMA

Para comenzar la evaluación de diferentes modelos, se aplica ARIMA a los conjuntos de datos. Para seleccionar los mejores parámetros p, d y q para el modelo ARIMA, se presenta en la Ilustración 16 la Autocorrelación (ACF) y Autocorrelación Parcial (PACF), observándose que en ambas gráficas hay una caída significativa después del primer rezago, además las series son estacionarias, lo que proporciona un modelo ARIMA(1, 0, 1) como punto de partida. Por lo tanto, la predicción del valor observado de la serie temporal se basa en un valor pasado, donde no se

requirieron niveles de diferenciación para hacer que la serie fuera estacionaria, y el error observado (residual) se basa en un error pasado igual a uno.

En el caso de todos los hexágonos evaluados la forma de la ACF y PACF siguen esta misma forma, por lo que se considera viable la aplicación de un modelo ARIMA de este orden para todos los hexágonos.

Es necesario llevar a cabo una validación del modelo ARIMA(1,0,1) comparándolo con otros modelos de otro orden, para este análisis comparativo se utilizan métricas de desempeño como RMSE (Root Mean Square Error) y MAE (Mean Absolute Error). Además, se realiza la generación de gráficos residuales. Estas visualizaciones son fundamentales para comprender la idoneidad del modelo y proporcionan información valiosa sobre su capacidad para capturar la estructura subyacente de los datos.

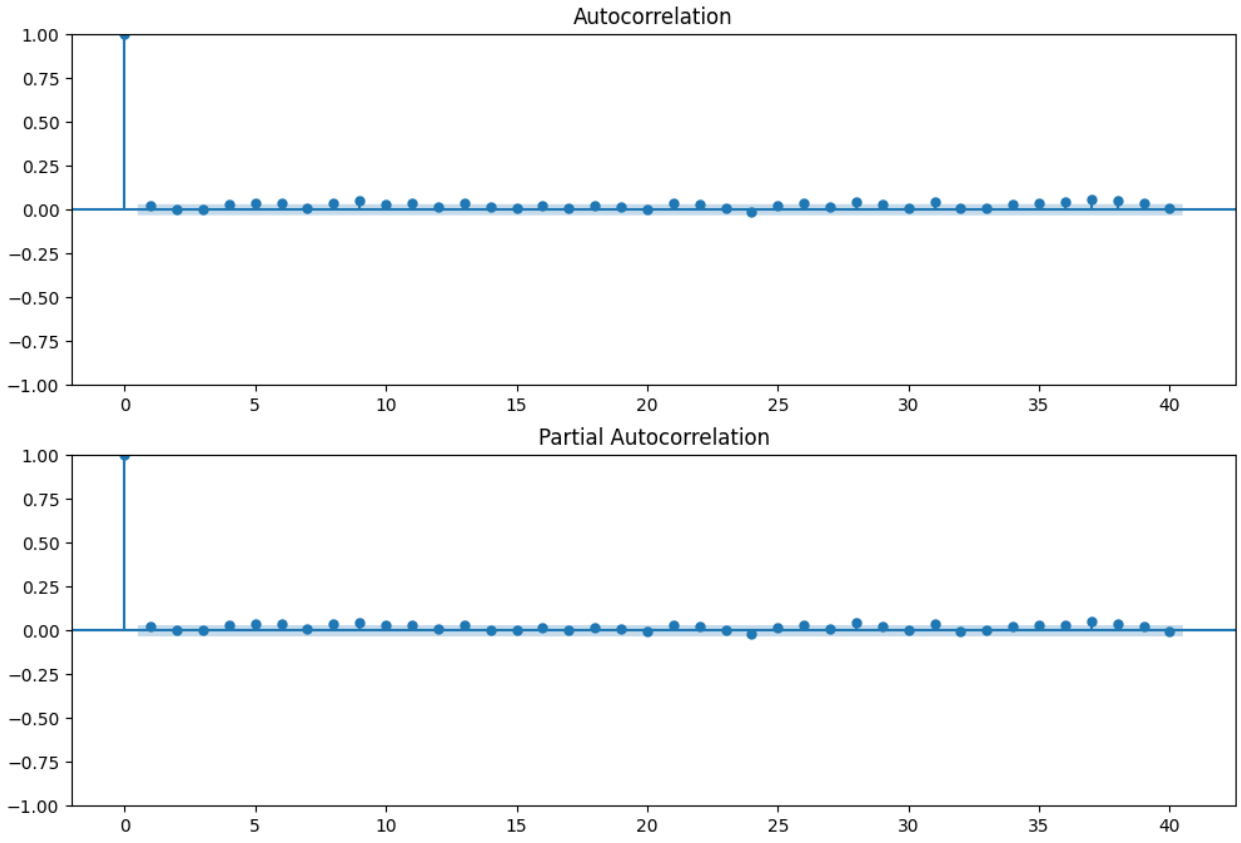


Ilustración 16. Gráficos ACF y PACF para ayudar a determinar el orden del modelo ARIMA

5.5.2. Aplicación modelo SARIMA

Una vez seleccionado el orden del modelo ARIMA, se procedió a mejorar su capacidad predictiva mediante la implementación de SARIMA, que incorpora componentes estacionales adicionales para abordar patrones recurrentes en la serie temporal. Como punto de partida, se utilizó la configuración orden estacional (1, 0, 1, 12), que especifica la presencia de una estacionalidad anual.

5.5.3. Aplicación modelo ARMAX

Hasta este momento, el forecasting se ha centrado exclusivamente en la cantidad de delitos. Sin embargo, se ampliará el enfoque al modelo ARMAX, incorporando variables externas como los datos meteorológicos y las variables de control. Este enfoque permitirá una modelización más completa al considerar factores adicionales que pueden tener impacto en la incidencia de delitos, mejorando así la precisión y capacidad predictiva del modelo.

5.5.4. Aplicación de Otros Modelos de Forecasting

Una vez establecida la estacionariedad de las series y aplicados los modelos ARIMA, se procede a utilizar otro modelo enfocado en la autorregresión, como es el caso del VAR. Adicionalmente, se emplea el modelo Prophet, ampliamente utilizado para la predicción de series temporales debido a su facilidad de uso. Prophet ofrece la ventaja de capturar automáticamente tendencias, estacionalidades y efectos de días festivos en los datos. Para la implementación de estos dos modelos no se requieren consideraciones adicionales, ya que su uso se basa en la simplicidad de su aplicación y en la capacidad de adaptarse de manera eficiente a las características inherentes de las series temporales analizadas.

5.5.5. RNN LSTM

Para abordar la predicción de delitos en series temporales, se aplicó una Red Neuronal Recurrente (RNN) con capas LSTM (Long Short-Term Memory) y capas de dropout. Las RNN son un tipo de red neuronal artificial especialmente diseñadas para procesar datos de series temporales e información secuencial. A diferencia de las redes neuronales tradicionales, las RNN poseen elementos de memoria que les permiten evaluar la entrada actual utilizando información histórica, lo que las hace particularmente adecuadas para esta tarea.

Las capas LSTM se eligieron por su capacidad de retener información relevante durante largos períodos y manejar el problema del desvanecimiento del gradiente, que es común en las RNN tradicionales. Estas capas son efectivas para capturar patrones temporales complejos y dinámicos en los datos. Las capas de dropout, por otro lado, se incorporaron para prevenir el sobreajuste, proporcionando una regularización efectiva al apagar aleatoriamente neuronas durante el entrenamiento y mejorando así la generalización del modelo.

Se evaluaron diferentes hiperparámetros para optimizar la predicción del modelo, ajustando aspectos como el número de unidades LSTM, la tasa de dropout, el tamaño del lote (batch size), y el número de épocas (epochs). Además, se implementó el parámetro Early Stopping, que detiene el entrenamiento si el rendimiento en el conjunto de validación no mejora después de 10 épocas, restaurando los mejores pesos obtenidos durante el entrenamiento. Esto asegura que el modelo no se sobreajuste y se mantenga eficiente y preciso.

Para garantizar la robustez del modelo, los pesos del entrenamiento se guardaron para cada cuartil, evaluando la posibilidad de utilizar estos modelos ya entrenados en otros hexágonos que correspondan al mismo cuartil. Esta estrategia permite aprovechar el conocimiento adquirido y mejorar la eficiencia del proceso de modelado, facilitando la transferencia de aprendizaje entre hexágonos con características delictivas similares.

6. RESULTADOS

En esta sección se presentan los resultados obtenidos a partir de la aplicación de los diferentes modelos de aprendizaje automático para la predicción de delitos. Los modelos fueron evaluados utilizando métricas específicas como RMSE y MAE, adicionalmente se realiza un análisis de residuos y se analizó la influencia de variables meteorológicas y temporales.

6.1. Resultados modelo ARIMA

A continuación, se presentan los resultados obtenidos del modelo ARIMA aplicado para predecir la cantidad de delitos en los diferentes hexágonos. Se evaluaron varios órdenes de ARIMA, inicialmente, se exploraron diversos órdenes de ARIMA, partiendo del ARIMA(1,0,1). Aunque las métricas como el MAE y RMSE no mostraron variaciones significativas entre los diferentes órdenes, el análisis de residuos reveló que el modelo ARIMA(1,1,0) presentaba los mejores resultados en términos de ajuste y capacidad predictiva para los datos. Este orden se adaptó satisfactoriamente a las características estacionales y tendenciales de las series temporales para todos los hexágonos evaluados, como se presenta en las ilustraciones 17, 19, 21 y 23.

Aunque las series temporales muestran componentes estacionales, se optó por aplicar una diferenciación de primer orden (d=1) en ARIMA(1,1,0). Esta decisión fue crucial para capturar adecuadamente la escala de los datos y mejorar la estacionariedad de las series, facilitando así la aplicación efectiva del modelo ARIMA en la predicción de la incidencia delictiva. La diferenciación permitió eliminar la tendencia no estacionaria presente en los datos originales, asegurando que el modelo capturara de manera precisa las fluctuaciones estacionales y la variabilidad a lo largo del tiempo.

Se observó que las métricas como RMSE y MAE varían significativamente dependiendo del cuartil de la distribución de la cantidad de delitos. En los cuartiles con mayor número de delitos diarios, estas métricas tienden a ser mayores debido a la escala absoluta de los valores pronosticados y observados. Esta variabilidad subraya la importancia de considerar la distribución y escala de los datos al interpretar y comparar el desempeño del modelo ARIMA en diferentes contextos espaciales y temporales.

Se realizó la prueba de Ljung-Box para evaluar la presencia de autocorrelación significativa en los residuos del modelo ARIMA(1,1,0). En todos los casos, los resultados indicaron la presencia

de correlación significativa en los residuos, lo cual sugiere que el modelo podría beneficiarse de una mayor complejidad o de la incorporación de variables adicionales para capturar patrones residuales no explicados por el modelo actual. No obstante, los histogramas de residuos presentados en las figuras 18, 20, 22 y 24 muestran una distribución normal. Esto es crucial porque una distribución normal de los residuos es un requisito fundamental para que el modelo ARIMA sea válido y confiable. Indica que los residuos se distribuyen aleatoriamente alrededor de cero, lo cual es consistente con la suposición de que no quedan patrones significativos sin capturar por el modelo, validando así la adecuación del modelo en términos de ajuste y predicción.

Adicionalmente, se calculó la medida de asimetría en la distribución de los residuos del modelo ARIMA(1,1,0), encontrando en todos los casos una asimetría positiva (hacia la derecha). Esta observación sugiere que los residuos tienden a tener colas más largas en el lado derecho de la distribución, indicando que hay más valores extremadamente positivos que negativos en las discrepancias entre los valores pronosticados y observados.

A pesar de los buenos resultados en las métricas de desempeño generales, es fundamental considerar estos aspectos al comparar el modelo ARIMA con otras alternativas como ARIMA y ARMAX. La elección del modelo más apropiado dependerá de la capacidad para manejar la estacionalidad, la complejidad de los datos y la necesidad de mejorar la precisión predictiva mediante la incorporación de variables externas relevantes.

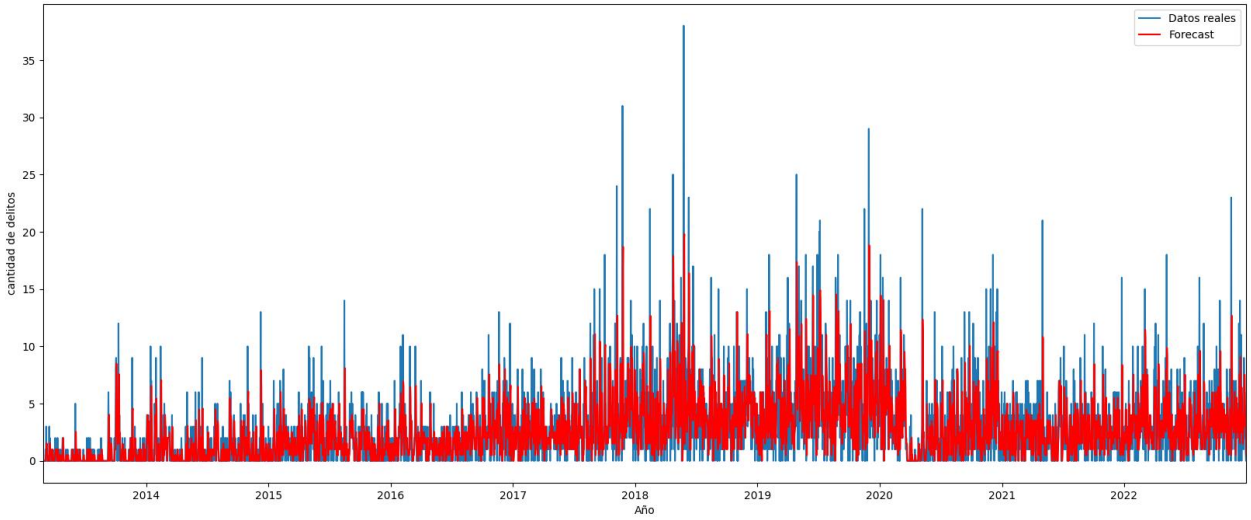


Ilustración 17. Forecasting modelo ARIMA para hexágono de La Candelaria (Q4)

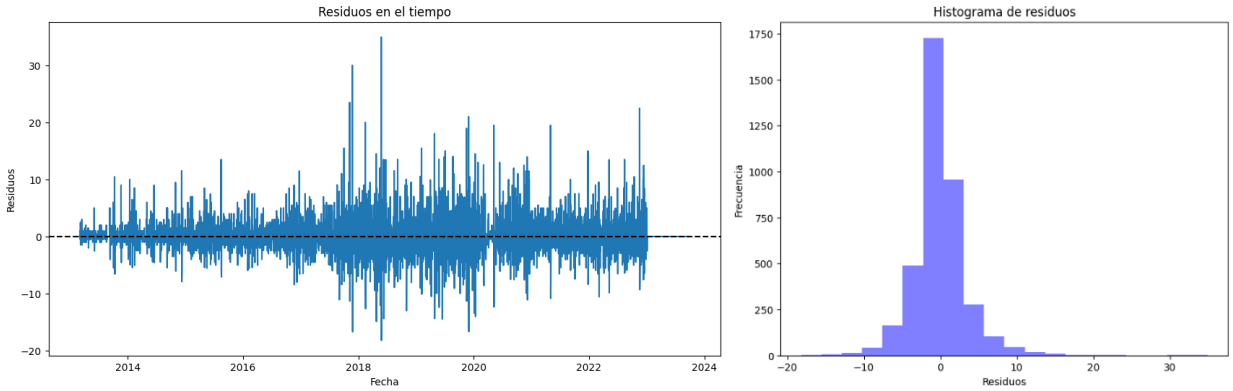


Ilustración 18. Análisis de residuos para hexágono de La Candelaria (Q4)

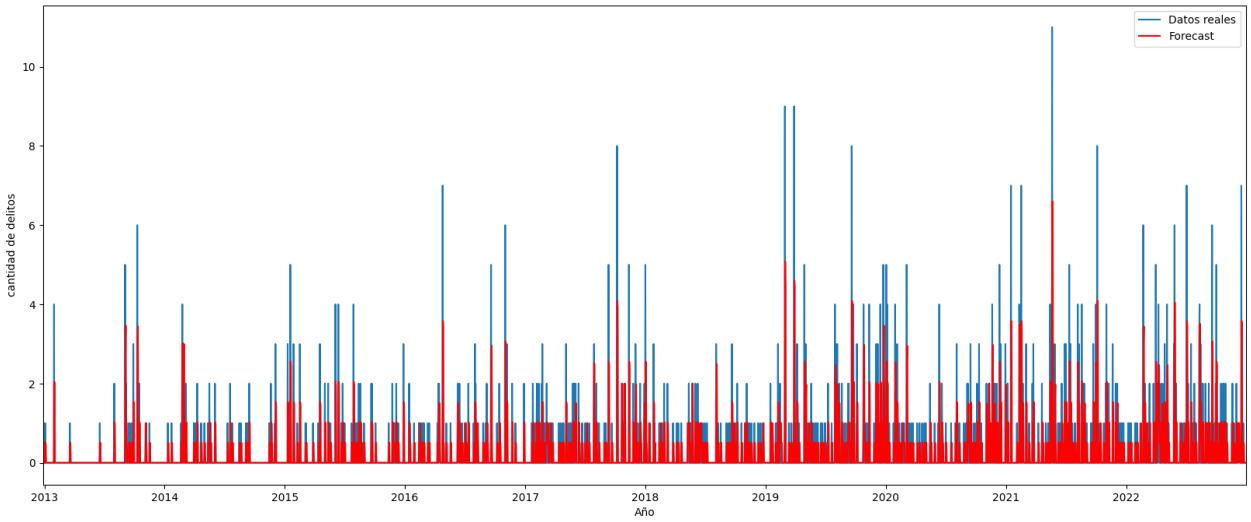


Ilustración 19. Forecasting modelo ARIMA para hexágono de Laureles (Q3)

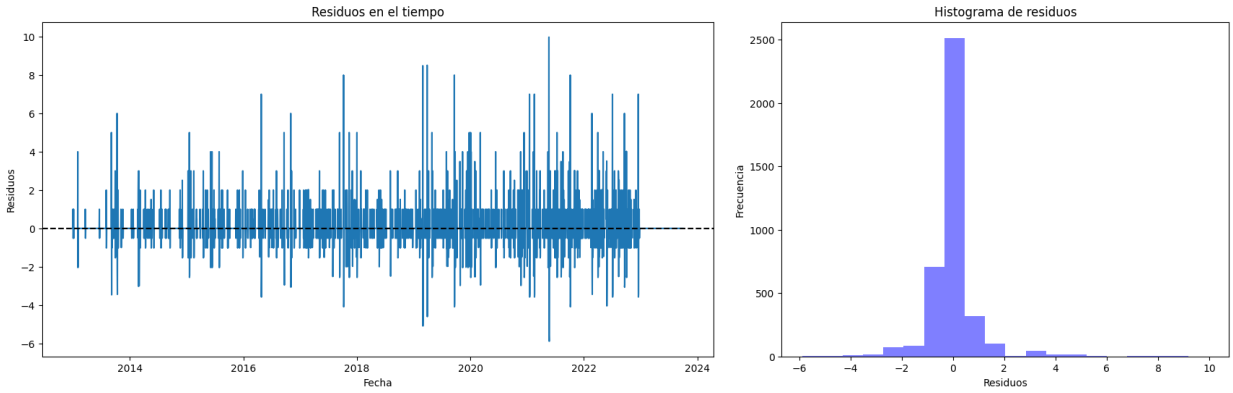


Ilustración 20. Análisis de residuos para hexágono de Laureles (Q3)

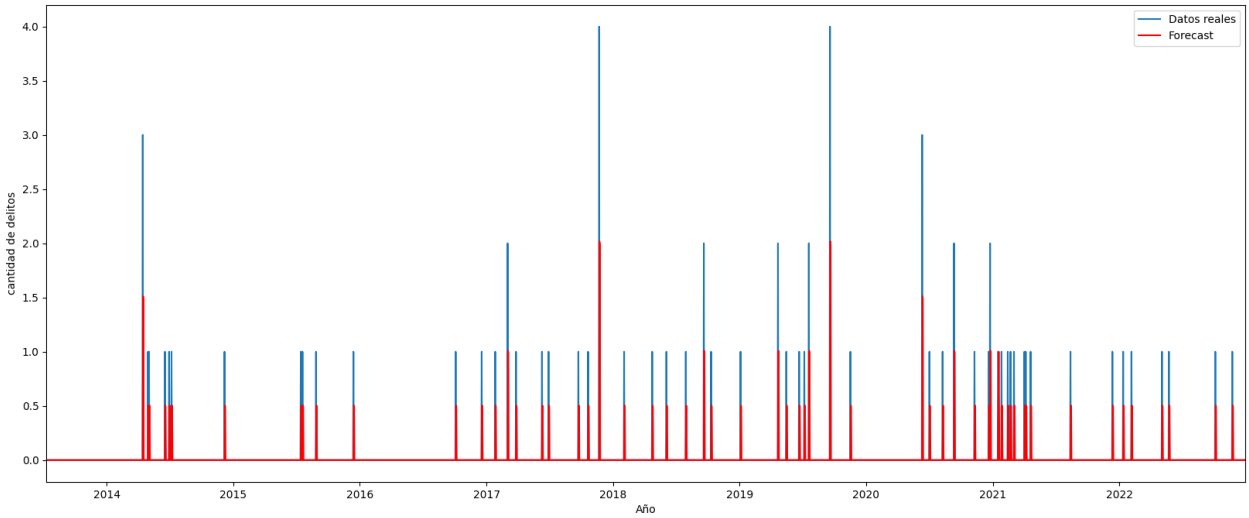


Ilustración 21. Forecasting modelo ARIMA para hexágono de San Cristóbal (Q2)

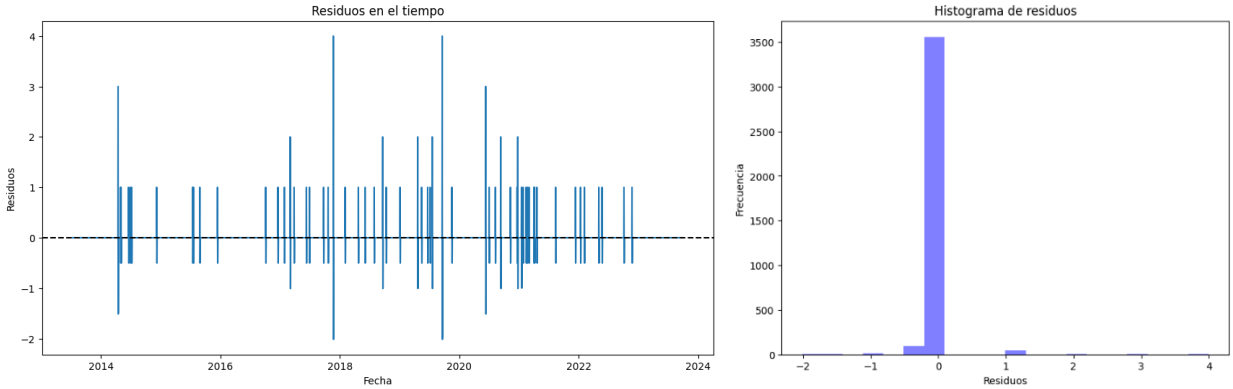


Ilustración 22. Análisis de residuos para hexágono de San Cristóbal (Q2)

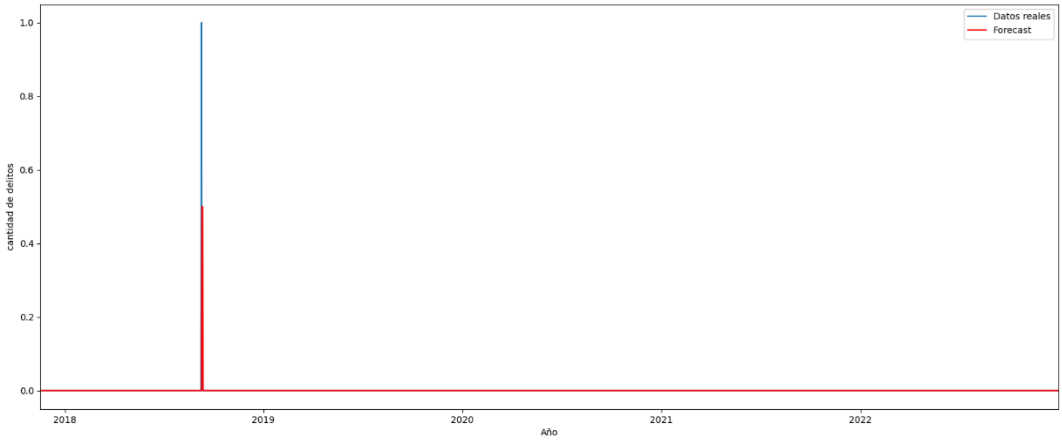


Ilustración 23. Forecasting modelo ARIMA para hexágono de Altavista (Q1)

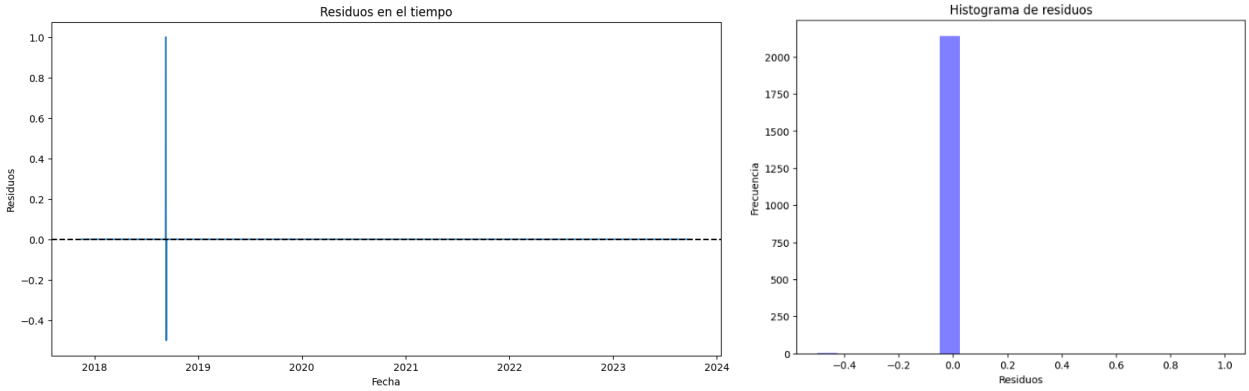


Ilustración 24. Análisis de residuos para hexágono de Altavista (Q1)

6.2. Resultados Modelo SARIMA

Se observó que al agregar un orden estacional al modelo de orden (1,1,0), no se produjeron mejoras significativas en los resultados. Tanto en la prueba de Ljung-Box como en la medida de la distribución de los residuos, los resultados fueron prácticamente idénticos a los obtenidos sin considerar la estacionalidad. Esta falta de cambio puede deberse a varios factores que se explican a continuación:

Primero, aunque la serie temporal de delitos presenta componentes estacionales, estos pueden no ser lo suficientemente fuertes o regulares como para ser capturados de manera efectiva por el término estacional del modelo SARIMA. Segundo, es posible que la variabilidad y los patrones en los datos de delitos sean más complejos y no se ajusten bien a un modelo estacional lineal. Los delitos pueden estar influenciados por una combinación de factores estacionales, tendencias a largo plazo y eventos aleatorios, lo que complica el ajuste de un modelo SARIMA simple.

Estos hallazgos sugieren que, para estas series temporales específicas de delitos, la estacionalidad no es un componente significativo que pueda ser capturado eficazmente mediante la diferenciación estacional en el modelo SARIMA. Por lo tanto, se obtuvieron resultados similares con y sin el término estacional, lo que refuerza la idea de que el modelo ARIMA es igualmente adecuado para esta tarea.

6.3. Resultados Modelo ARMAX

La incorporación de variables exógenas, que incluyen factores de control y datos meteorológicos, resultó en una mejora significativa respecto al modelo ARIMA. Esta inclusión se reflejó en mejoras en todas las métricas de desempeño evaluadas: el RMSE y el MAE mostraron una reducción promedio del 6.08% y 13.84%, respectivamente, mientras que la medida de asimetría de los residuos se redujo en un promedio de 0.83 unidades en comparación con el modelo ARIMA. Esto indica que el modelo ARMAX no solo mejoró la precisión de las predicciones, sino que también logró capturar de manera más efectiva los patrones residuales no explicados por el modelo ARIMA inicial.

La reducción en la medida de asimetría de los residuos es crucial en el análisis de calidad del modelo. En el contexto del histograma de residuos, una menor asimetría indica que los residuos tienden a distribuirse de manera más cercana a una distribución simétrica alrededor de cero. Esto sugiere que el modelo ARMAX está capturando más efectivamente las variaciones sistemáticas en los datos, reduciendo así los sesgos en las predicciones. En consecuencia, una menor asimetría indica una mejor capacidad del modelo para explicar la variabilidad de los datos observados, lo cual es fundamental para la precisión y fiabilidad de las predicciones presentadas en las ilustraciones 25, 26, 27 y 28.

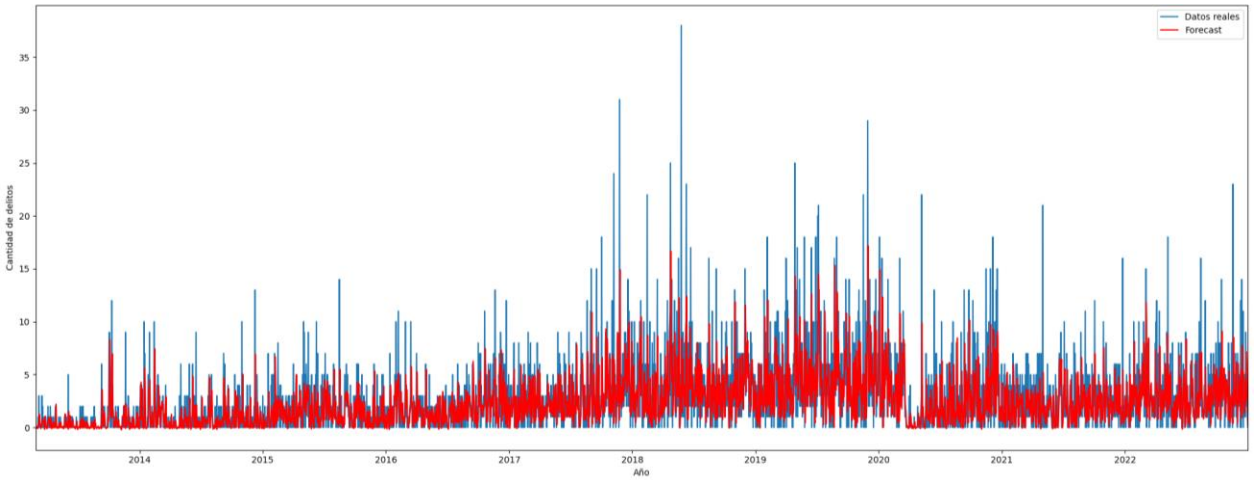


Ilustración 25. Forecasting modelo ARMAX para hexágono de La Candelaria (Q4)

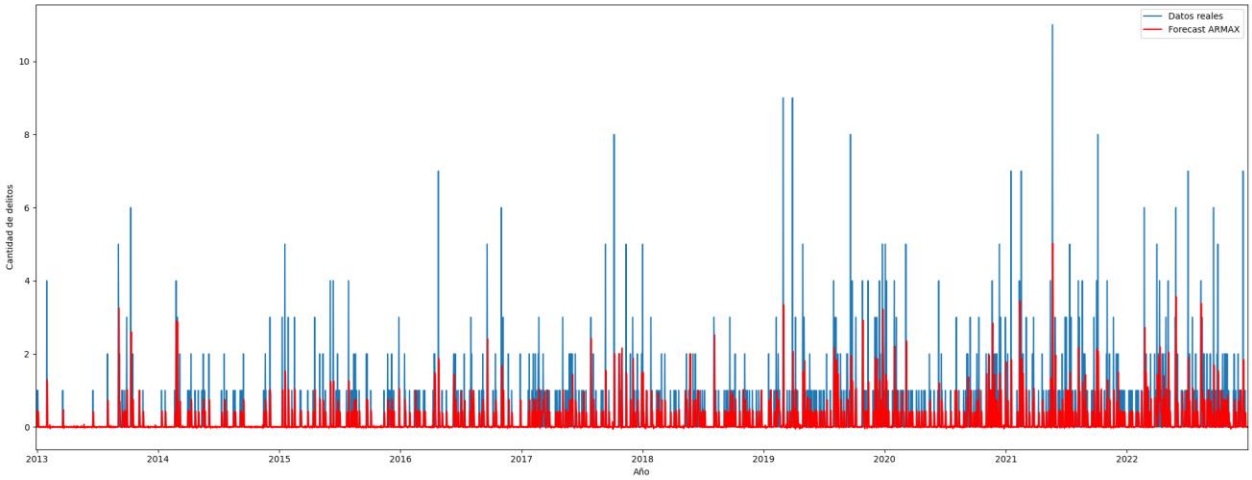


Ilustración 26. Forecasting modelo ARMAX para hexágono del cuartil 3

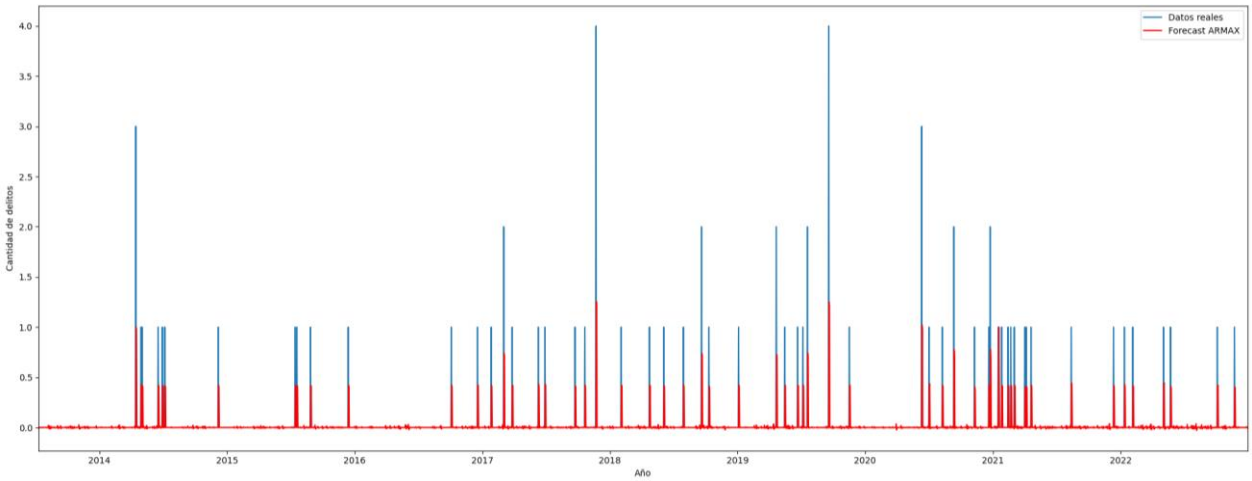


Ilustración 27. Forecasting modelo ARMAX para hexágono del cuartil 2

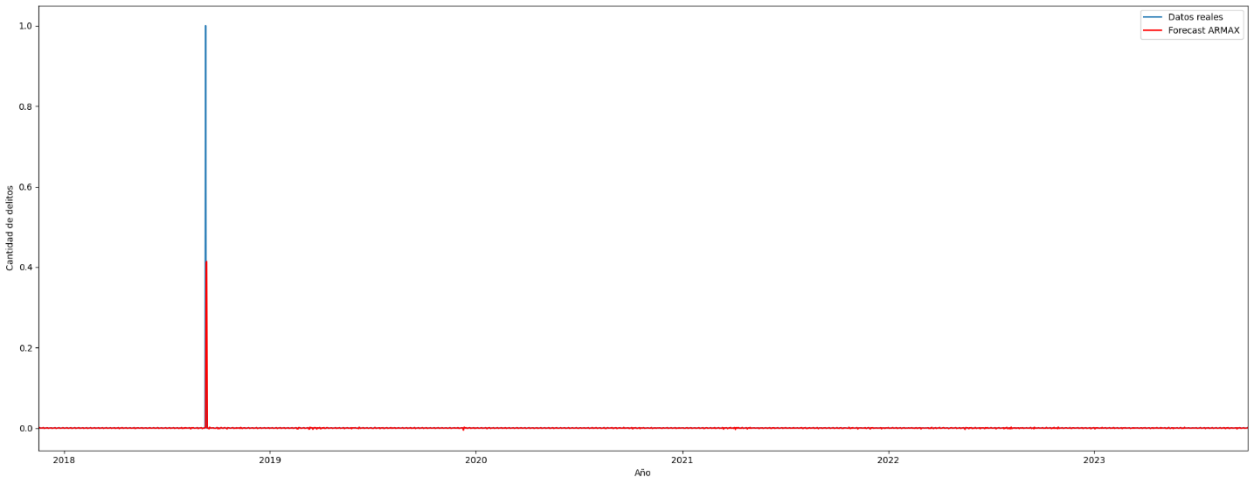


Ilustración 28. Forecasting modelo ARMAX para hexágono del cuartil 1

En la Tabla 3, se presentan las métricas obtenidas para cada uno de los modelos evaluados hasta el momento, destacando la mejora significativa que el modelo ARMAX presenta en comparación con ARIMA y SARIMA. El modelo ARMAX muestra mejoras consistentes en todas las métricas evaluadas, incluyendo RMSE, MAE y la medida de asimetría de los residuos (Skew). Esta mejora indica una mayor precisión en las predicciones y una mejor capacidad para capturar las características de los datos observados, especialmente notable en los cuartiles de mayor actividad delictiva.

Hexágono	Modelo	RMSE	MAE	Skew (medida de asimetría de residuos)
Altavista (Q1)	ARIMA	0.02645	0.00093	18.90
	SARIMA	0.02645	0.00093	18.90
	ARMAX	0.02500	0.00147	18.86
San Cristóbal (Q2)	ARIMA	0.22198	0.03990	5.09
	SARIMA	0.22198	0.04000	5.09
	ARMAX	0.20431	0.03874	3.71
Laureles (Q3)	ARIMA	0.99257	0.44229	2.07
	SARIMA	0.99250	0.44505	2.07
	ARMAX	0.90729	0.39682	1.17
La Candelaria (Q4)	ARIMA	3.51539	2.26703	0.98
	SARIMA	3.51491	2.26888	0.98
	ARMAX	3.35747	2.14896	-0.02

Tabla 3. Resultados modelos ARIMA, SARIMA y ARMAX

6.4. Resultados modelo VAR y Prophet

La aplicación del modelo VAR mostró resultados desfavorables debido a que el algoritmo utilizado para resolver el problema de mínimos cuadrados lineales mediante la descomposición de valores singulares (SVD) no pudo converger. Este problema se atribuyó a la estabilidad numérica de los datos, específicamente a la presencia de variables exógenas con valores muy pequeños, los cuales afectaron negativamente la convergencia del modelo. Estos hallazgos subrayan la importancia de seleccionar modelos que puedan manejar de manera efectiva la estructura particular de los datos utilizados en el análisis predictivo.

Durante la evaluación del modelo Prophet para predecir la cantidad de delitos, se encontraron varios desafíos significativos que afectaron la idoneidad del modelo para los datos específicos. En primer lugar, se observó que las predicciones generadas por Prophet no se

adaptaban correctamente a la escala de nuestros datos reales, como se presenta en la Ilustración 29. Este problema puede atribuirse a la diferencia en la frecuencia de los datos originales y la suposición implícita de Prophet sobre la frecuencia diaria de los datos. Sin embargo, se resalta que, en cuanto a métricas de desempeño, este modelo presentó mejores resultados, por ejemplo, los obtenidos para el hexágono del cuartil 4: RMSE: 2.89567, MAE: 1.95722.

Un hallazgo adicional fue la aparición de predicciones negativas, lo cual es inconsistente con la naturaleza de los datos de cantidad de delitos, que no pueden ser valores negativos. Esta observación indica que las componentes de tendencia o estacionalidad del modelo pueden no estar correctamente especificadas.

Basado en estos problemas observados, se tomó la decisión de descartar el modelo Prophet para este estudio particular. Es fundamental que los modelos de pronóstico proporcionen resultados confiables y útiles para la toma de decisiones, y en este caso, Prophet no cumplió con las expectativas debido a los desafíos mencionados.

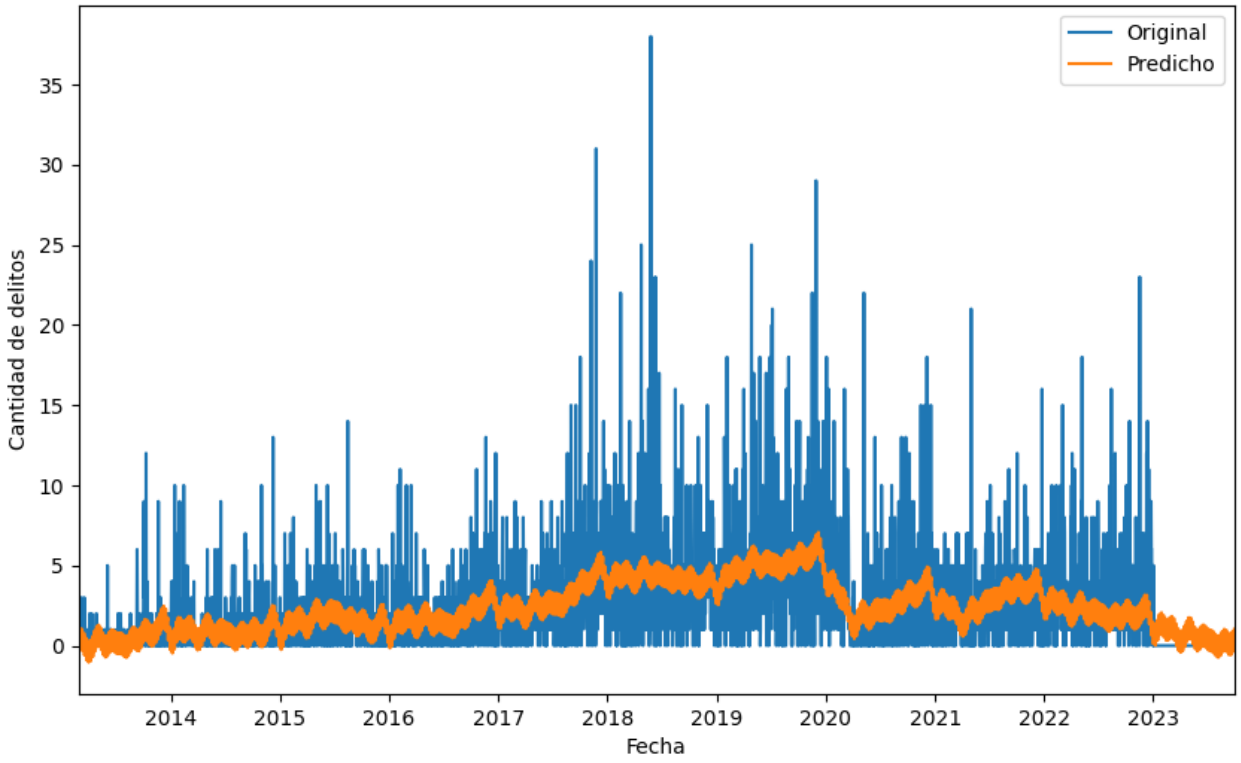


Ilustración 29. Forecasting modelo Prophet para hexágono de La Candelaria (Q4)

6.5. Resultados RNN

La aplicación de la red neuronal recurrente con capas LSTM reveló varios desafíos significativos. A diferencia del modelo ARMAX, que incorporó variables exógenas y mostró adaptabilidad, la red LSTM fue objeto de ajustes exhaustivos en sus hiperparámetros. Se exploraron variaciones en el número y tamaño de las capas dense y LSTM para optimizar la capacidad de la red para capturar relaciones temporales complejas. Además, se experimentó con diferentes configuraciones de dropout para regular el sobreajuste, y se ajustó la cantidad de épocas para equilibrar el rendimiento computacional con la precisión del modelo. A pesar de estos esfuerzos, el modelo LSTM mostró una rápida activación del mecanismo de early stopping en todas las ejecuciones, indicando dificultades persistentes para converger hacia una solución óptima adaptada a la escala y la complejidad del conjunto de datos real.

Los resultados métricos, aunque mejorados en comparación con los otros modelos (RMSE: 2.896 y MAE: 1.957), no son suficientes para considerar que el modelo LSTM se adapta de manera satisfactoria a la estructura de las series temporales, como se evidencia en la Ilustración 30.

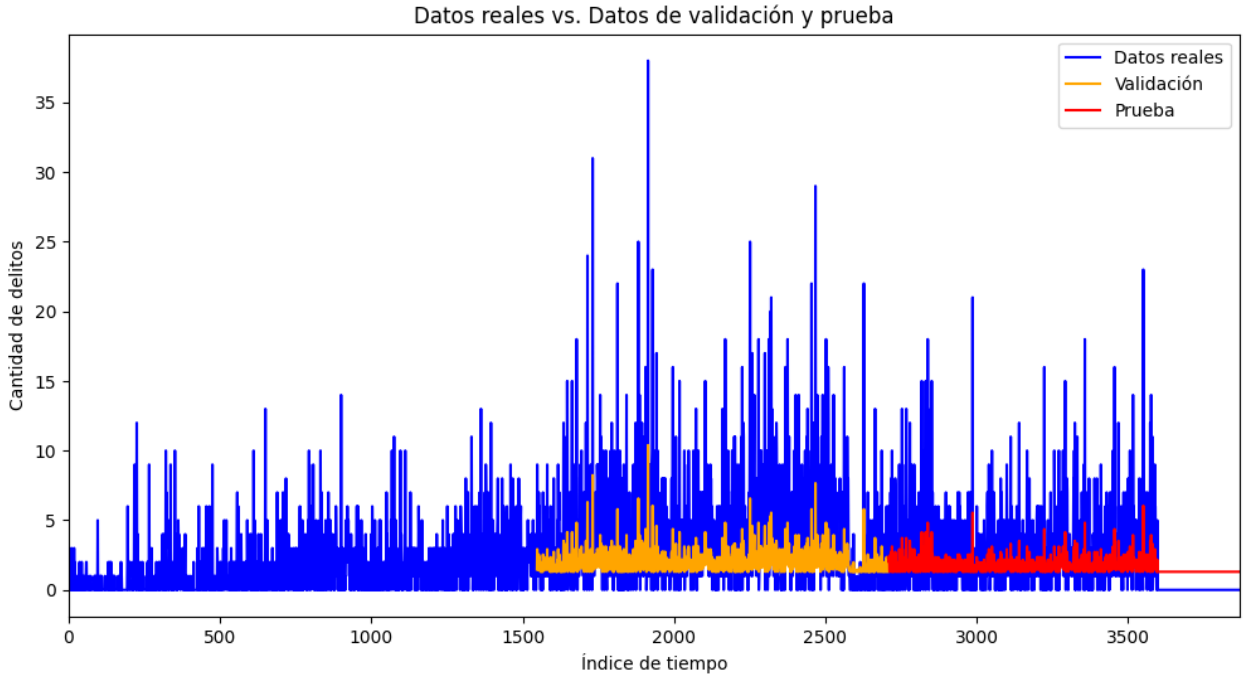


Ilustración 30. Predicción usando una RNN LSTM para hexágono de La Candelaria (Q4)

La decisión de descartar el modelo LSTM se fundamentó en su incapacidad para capturar la complejidad inherente de los datos temporales de delitos, junto con su propensión a activar el

early stopping rápidamente. Esta complejidad adicional y la falta de adaptabilidad sugieren que el enfoque de red neuronal LSTM puede no ser el más adecuado para este problema específico.

En contraste, el modelo ARMAX destacó al integrar de manera efectiva las variables exógenas, mostrando una mejor capacidad para manejar la estructura de los datos temporales y proporcionando resultados más robustos, es por esto que ARMAX emergió como la elección preferida debido a su capacidad para ajustarse adecuadamente a las series temporales de delitos, ofreciendo mejores métricas de rendimiento y una interpretación más clara de las relaciones entre las variables.

En la Tabla 4, se presenta la comparación de las métricas de los diferentes modelos en cuestión de las métricas de desempeño (RMSE y MAE). Se debe recalcar que estas métricas no son los únicos parámetros necesarios para la elección de un modelo en el contexto de la predicción de delitos. Si bien Prophet y RNN LSTM tuvieron las mejores métricas de desempeño, estos modelos no se adaptaron adecuadamente a la escala del conjunto de datos, sugiriendo limitaciones en su aplicabilidad para este contexto específico.

Modelo	**RMSE**	**MAE**
ARIMA	3.51539	2.26703
SARIMA	3.51491	2.26888
ARMAX	3.35747	2.14896
Prophet	2.89567	1.95722
RNN LSTM	2.896	1.957

Tabla 4. Métricas de desempeño de los diferentes modelos evaluados para el hexágono La Candelaria (Q4)

7. CONCLUSIONES

En este trabajo se realizó un análisis de modelos para la predicción de delitos en la ciudad de Medellín, centrándose en diferentes enfoques para el pronóstico de series temporales. Se evaluaron modelos como ARIMA, SARIMA, ARMAX y Prophet para capturar tendencias estacionales en los datos de delitos, utilizando métricas como RMSE y MAE para medir su desempeño. Además, se exploró el potencial de la red neuronal recurrente LSTM, ajustando hiperparámetros para mejorar su adaptación a la complejidad de los datos. A continuación, se presentan las conclusiones derivadas de esta evaluación detallada:

Se encontró que el modelo ARIMA(1,1,0) con diferenciación de primer orden fue efectivo para capturar las tendencias estacionales y mejorar la estacionariedad de las series temporales de delitos de diferentes hexágonos. Aunque este modelo mostró autocorrelación significativa en los residuos, sus métricas de desempeño como RMSE y MAE fueron competitivas, destacando su capacidad para manejar la variabilidad temporal de los datos. Por otro lado, la inclusión de términos estacionales en el modelo SARIMA no proporcionó mejoras sustanciales sobre el ARIMA básico, sugiriendo que la estacionalidad en los datos de delitos podría no ser capturada eficazmente mediante un enfoque lineal.

La red neuronal recurrente LSTM mostró dificultades para adaptarse a la escala y complejidad de los datos de delitos, con un desempeño inferior. Las activaciones frecuentes del mecanismo de early stopping indicaron problemas de convergencia y la necesidad de ajustes adicionales en los hiperparámetros para mejorar su rendimiento.

Además de evaluar los diferentes modelos de aprendizaje automático, este estudio resalta el aporte crucial de la inclusión de variables meteorológicas y de control en la predicción de la cantidad de delitos. La integración de estas variables exógenas en el modelo ARMAX representó un avance significativo, evidenciado por mejoras consistentes en las métricas de desempeño como RMSE y MAE en comparación con los modelos ARIMA y SARIMA.

Las variables meteorológicas, como la temperatura y la precipitación, tienen un impacto directo en el comportamiento humano y, por ende, en los patrones delictivos. La capacidad del modelo ARMAX para capturar estos efectos mejoró la capacidad predictiva al considerar cómo estos factores externos pueden influir en la frecuencia de los delitos en diferentes contextos geográficos y temporales.

Este trabajo subraya la importancia de considerar no solo los datos históricos de delitos, sino también el entorno circundante y las condiciones externas que pueden afectar significativamente las tendencias delictivas. Esta integración amplía la capacidad analítica y predictiva de los modelos de aprendizaje automático, proporcionando herramientas más robustas para la formulación de políticas públicas y estrategias de seguridad.

Departamento de Ingeniería Electrónica y de Telecomunicaciones

UNIVERSIDAD DE ANTIOQUIA

Facultad de Ingeniería

Evaluación del desempeño de diferentes modelos de aprendizaje automático para la predicción de delitos en la ciudad de Medellín

PRACTICANTE: Juan Pablo Areiza Jiménez

ASESOR: Luis Germán García Morales

PROGRAMA: Ingeniería de Telecomunicaciones

Semestre de la práctica: 2024-1

Introducción

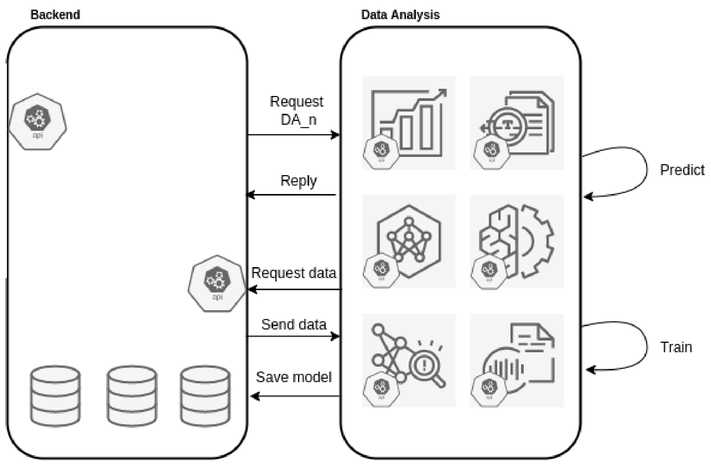
La seguridad ciudadana es crucial para el bienestar urbano y se alinea con los Objetivos de Desarrollo Sostenible de la ONU. En Medellín, la falta de información precisa y oportuna dificulta la respuesta efectiva de las autoridades ante delitos. En este contexto, este proyecto buscó desarrollar una herramienta para la administración inteligente de la seguridad ciudadana, mediante la evaluación de modelos de aprendizaje automático para predecir delitos, utilizando datos históricos de crímenes, variables meteorológicas y de control, con el fin de fomentar un entorno más seguro para la comunidad.

Objetivos

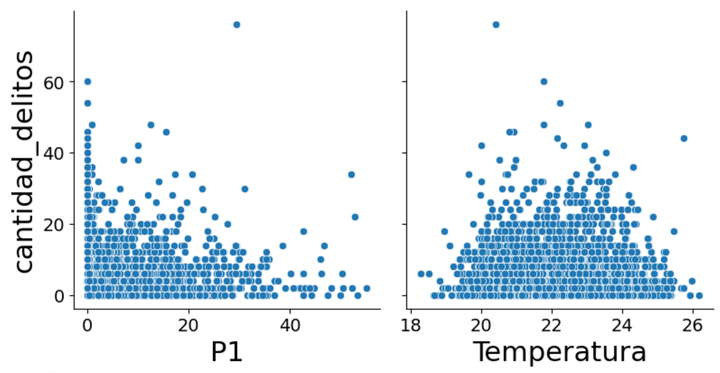
- Revisar la literatura para seleccionar variables y algoritmos adecuados para predecir delitos en Medellín
- Implementar y ajustar modelos de aprendizaje automático en Python para predecir delitos, evaluando su desempeño.
- Comparar los modelos para identificar los más eficaces en la predicción de delitos en Medellín.

Metodología

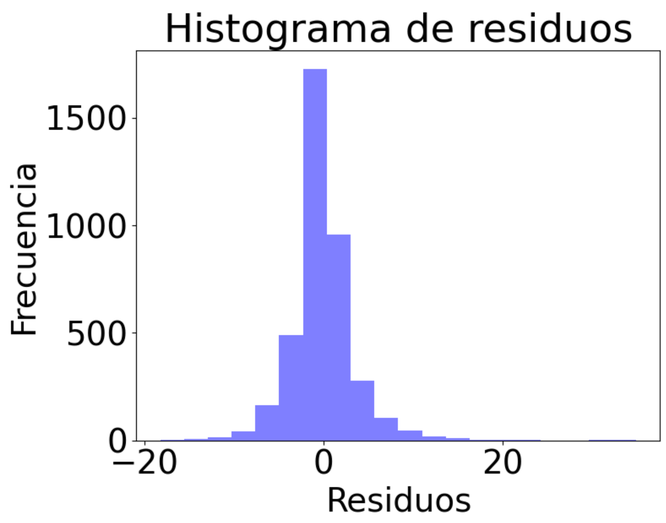
- Comprensión y recopilación de datos.

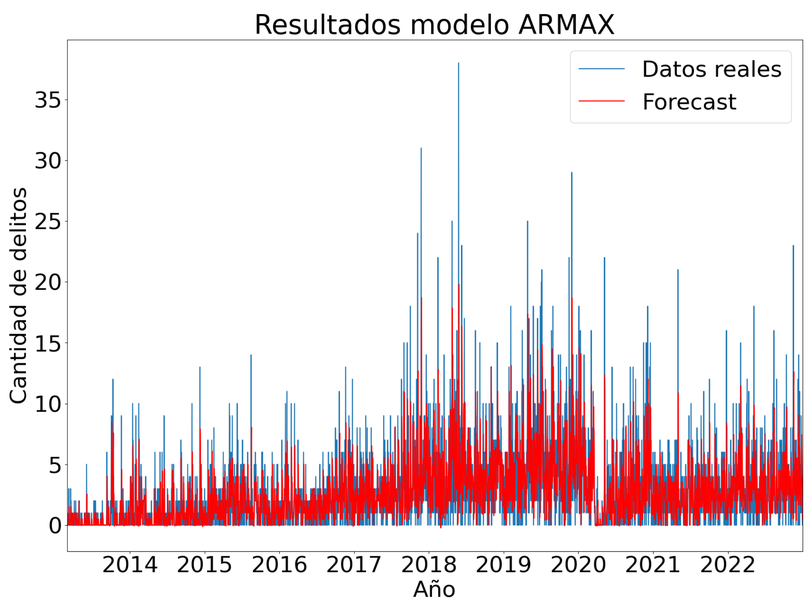


- Configuración del entorno de pruebas
- Análisis de características



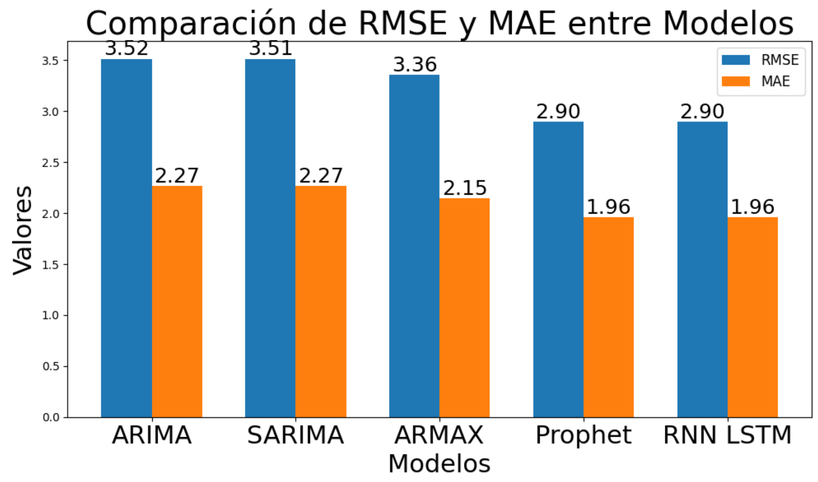
- Análisis de estacionariedad
- Implementación, evaluación y comparación de modelos:
  - ARIMA
  - SARIMA
  - ARMAX
  - VAR
  - Prophet
  - RNNN LSTM





Resultados

Los resultados obtenidos indican que el modelo ARMAX fue el que mejor se adaptó a la escala de los datos de delitos, confirmado mediante análisis de residuos más favorables. Si bien ARMAX no obtuvo las mejores métricas, superó a los modelos ARIMA y SARIMA obteniendo mejoras en todas las métricas de desempeño evaluadas: el RMSE y el MAE mostraron una reducción promedio del 6.08% y 13.84%, respectivamente. Aunque las métricas RMSE y MAE fueron importantes, no fueron lo único a considerar, ya que modelos como Prophet y la RNN no se adaptaban a la escala de los datos. La inclusión de variables exógenas mejoró la capacidad del modelo para capturar la variabilidad observada.



Conclusiones

- El modelo ARMAX se destacó por su capacidad para integrar variables exógenas, mejorando notablemente la precisión predictiva de la incidencia delictiva.
- La selección del modelo adecuado dependió de la capacidad para manejar la estacionariedad, la complejidad de los datos y la inclusión de variables relevantes.
- Aunque las métricas RMSE y MAE fueron importantes, no fueron las únicas consideraciones; el análisis de residuos también fue crucial para confirmar la adaptabilidad del modelo.
- La inclusión de variables meteorológicas y de control mejoró significativamente la capacidad del modelo para explicar la variabilidad en los datos de delitos.

DATOS DE CONTACTO DEL AUTOR:

+57 3046133030

pablo.areiza@udea.edu.co

https://www.linkedin.com/in/juan-pablo-areiza-2240611a5

Este trabajo se realizó en el Grupo de Investigación en Sistemas Embebidos e Inteligencia Computacional (SISTEMIC), con el apoyo del proyecto de regalías código BPIN 020000100044

REFERENCIAS

[1] M. Moran, "Paz y justicia - desarrollo sostenible," Desarrollo Sostenible [en línea]. Consultado: 2024-03-10. Disponible en: https://www.un.org/sustainabledevelopment/es/peace-justice/

[2] "Índice para una Vida Mejor". OECD Better Life Index. [en línea]. Consultado: 2024-03-12. Disponible en: http://www.oecdbetterlifeindex.org/es/topics/safety-es/

[3] "LOS MODELOS DE PREVENCIÓN DEL DELITO: MODELOS DE ORIENTACIÓN 'ETIOLÓGICA' Y MODELOS DE PREVENCIÓN". 1Library.Co - plataforma de documentos compartidos. [en línea]. Consultado: 2024-03-11. Disponible en: https://1library.co/article/modelos-prevención-delito-modelos-orientación-etiológica-modelos-prevención.yr3l7roo

[4] "Abrir espacios para la seguridad ciudadana y el desarrollo humano | Human Development Reports". Home | Human Development Reports. [en línea]. Consultado: 2024-03-08. Disponible en: https://hdr.undp.org/content/abrir-espacios-para-la-seguridad-ciudadana-y-el-desarrollo-humano

[5] G. E. Box, G. M. Jenkins, G. C. Reinsel y G. M. Ljung, "Análisis de series temporales: previsión y control," John Wiley e hijos, 2015.

[6] "Series Temporales, Modelo ARIMA Metodología de Box - Jenkins". Estadística.net. [en línea]. Consultado: 2024-03-09. Disponible en: https://www.estadistica.net/ECONOMETRIA/SERIES-TEMPORALES/modelo-arima.pdf

[7] C. Miranda Chinlli, "Modelización de Series Temporales modelos clásicos y SARIMA", Tesis de maestría, Univ. Granada, Granada, 2021. [en línea]. Consultado: 2024-03-06. Disponible en: https://masteres.ugr.es/estadistica-aplicada/sites/master/moea/public/inline-files/TFM_MIRANDA_CHINLLI_CARLOS.pdf

[8] D. S. R. Perú. "Modelos ARIMA, SARIMA y Método de Selección de Variables LASSO para Series Temporales (Parte 1)". Data Science Research Perú | Substack. [en línea]. Consultado: 2024-03-07. Disponible en: https://datasciencepe.substack.com/p/modelos-arima-sarima-y-metodo-de

[9] "¿Qué es el Modelo ARMAX y la diferencia con el modelo ARIMA?" Todo Econometría y ciencia de datos. [en línea]. Consultado: 2024-03-14. Disponible en: https://todoeconometria.com/armaxvsarima/

[10] A. Novales, "Modelos vectoriales autoregresivos (VAR)", Univ. Complut., 2017. [en línea]. Consultado: 2024-03-15. Disponible en: https://www.ucm.es/data/cont/media/www/pag-41459/VAR.pdf

[11] "Prophet". Facebook. [en línea]. Consultado: 2024-03-16. Disponible en: https://facebook.github.io/prophet/

[12] K. Sharma, R. Bhalla y G. Ganesan, "Time Series Forecasting Using FB-Prophet", 2022. [en línea]. Consultado: 2024-03-18. Disponible en: https://ceur-ws.org/Vol-3445/PAPER_07.pdf

[13] "Regresión Logística - Teoría - Aprende IA". Aprende IA. [en línea]. Consultado: 2024-03-17. Disponible en: https://aprendeia.com/algoritmo-regresion-logistica-machine-learning-teoria/#:~:text=La%20regresión%20logística%20o%20Logistic,,%20abierto%20–%20cerrado,%20etc.

[14] "LightGBM Feature Importance and Visualization - GeeksforGeeks". GeeksforGeeks. [en línea]. Consultado: 2024-03-13. Disponible en: https://www.geeksforgeeks.org/lightgbm-feature-importance-and-visualization/

[15] "RNNs y LSTMs: ¿Qué son las Redes Neuronales Recurrentes?" Remolinator. [en línea]. Consultado: 2024-03-12. Disponible en: https://remolinator.com/rnn-y-lstms/

[16] "¿Qué son los RNN y los LSTM en Deep Learning?" Unite.AI. [en línea]. Consultado: 2024-03-10. Disponible en: https://www.unite.ai/es/¿Qué-son-rnns-y-lstms-en-el-aprendizaje-profundo?/

[17] V. D. Muñoz Jaramillo, "Evaluación de Modelos de Machine Learning para la Predicción de Crímenes en la Ciudad de Medellín", Tesis de maestría, Univ. Nac. Colomb., Medellín, 2021. [en línea]. Consultado: 2024-03-07. Disponible en: https://www.unite.ai/es/¿Qué-son-rnns-y-lstms-en-el-aprendizaje-profundo?/

[18] Y. M. SUÁREZ RUIZ, A. F. BEDOYA CRUZ, "MODELO DE CLASIFICACIÓN Y PREDICCIÓN DE DELITOS DE ALTO IMPACTO EN LA CIUDAD DE BOGOTÁ", Tesis de maestría, UNIV. SABANA, Chía, 2023. [en línea]. Consultado: 2024-03-11. Disponible en: https://intellectum.unisabana.edu.co/bitstream/handle/10818/59141/Tesis_Delitos_Andres_Bedoya_Cruz_Yvonne_Suarez_Ruiz.pdf?sequence=1

[19] J. D. Gelvez Ferreira, M. P. Nieto Rodríguez y C. A. Rocha Ruiz, "Prediciendo el crimen en ciudades intermedias: un modelo de 'machine learning' en Bucaramanga, Colombia1", Sistema de Información Científica Redalyc, Red de Revistas Científicas. [en línea]. Consultado: 2024-03-13. Disponible en: https://www.redalyc.org/journal/5526/552673068006/

[20] A. Araujo, N. Cacho, A. C. Thome, A. Medeiros y J. Borges, "A predictive policing application to support patrol planning in smart cities," in 2017 International Smart Cities Conference (ISC2), 2017, pp. 1–6.

[21] X. Zhao, N. Wang, R. Han, X. Botao, Y. Yu, M. Li y J. Ou, "Urban infrastructure safety system based on mobile crowdsensing," International Journal of Disaster Risk Reduction, vol. 27, 11 2017.

[22] O. Kounadi, A. Ristea, A. Araujo y M. Leitner, "A systematic review on spatial crime forecasting," Crime science, vol. 9, no. 1, pp. 1–22, 2020.

[23] "Medata". [en línea]. Consultado: 2024-03-08. Disponible en: https://medata.gov.co/

[24] "SIATA - Sistema de Alerta Temprana del valle de Aburrá". [en línea]. Consultado: 2024-03-14. Disponible en: https://siata.gov.co/sitio_web/index.php/monitoreo