Algoritmo para la predicción del delito al patrimonio económico en la ciudad de Bucaramanga mediante el uso de machine learning.

Diego Julian Plaza Quintero y Santiago Tarazona Rios

Trabajo de investigación para optar al título de Ingeniero Electrónico.

Director

Jeison Arley Castillo Bohorquez

Ingeniero Elecrtónico

Codirector

Jaime Guillermo Barrero Perez

Magíster en Ingeniería Electrónica

Universidad Industrial de Santander

Facultad de Ingenierías Fisicomecánicas

Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones

Ingenieria Electrónica Bucaramanga

2023

Tabla de Contenido

Lista de Figuras

Lista de Tablas

Lista de Apéndices

Apéndice A. Códigos de programación implementados.

Resumen

Título: Algoritmo para la predicción del delito al patrimonio económico en la ciudad de Bucaramanga mediante el uso de machine learning. [*]

Autores: Diego Julian Plaza Quintero, Santiago Tarazona Rios. [**]

Palabras Clave: Capas Neuronales, Aprendizaje de maquina, Delito, Relu, Sigmoid.

Descripción: Este proyecto se centra en la aplicación de técnicas de aprendizaje de maquina para predecir delitos relacionados con el patrimonio económico en la ciudad de Bucaramanga, con el objetivo de mejorar la seguridad de la población. El enfoque de este trabajo aborda la variedad de datos que pueden influir en el rendimiento del modelo. Para lograr una mayor eficacia en la predicción, se dividió el tiempo en ventanas de 15 días, generando así un total de 24 a 25 ventanas por año. Cada ventana se compone de vectores de datos que incluyen información de latitud, longitud y un conjunto de datos preseleccionados, con la variación de 3 a 5 ventanas temporales que contribuyen a formar el vector correspondiente a la ventana siguiente. La arquitectura de la red neuronal con mejores resultados que fue el modelo de capas sencillas, consta de una capa de entrada con 512 neuronas y función de activación Relu, seguida de una capa de dropout con una tasa de 0.2 y una capa de salida con función sigmoid. El estudio evalúa el rendimiento de la predicción en diferentes escenarios de entrenamiento, considerando cambios en el tamaño de la entrada, el tamaño del vector de datos, la salida de la red y el tipo de datos resultante, incluyendo valores normalizados y valores clasificados en 5 clases por los autores vistos en la tabla 7. Los resultados revelan un modelo de capas neuronales con un error promedio de 1,5254 y un error cuadrático medio de 0,0024 para los datos clasificados por los autores.

---

[*] Trabajo de Grado.

[**] Facultad de Ingeniería Fisicomecánicas. Escuela de Ingenierías Eléctrica, Electrónica y Telecomunicaciones. Director: Jeison Arley Castillo Bohorquez. Codirector: Jaime Guillermo Barrero Perez

Abstract

Title: Algorithm for the prediction of economic crime in the city of Bucaramanga through the use of machine learning. [*]

Authors: Diego Julian Plaza Quintero, Santiago Tarazona Rios [**]

Keywords: Neural Layers, Machine Learning, Crime, Relu, Sigmoid.

Description: This project focuses on the application of machine learning techniques to predict crimes related to economic assets in the city of Bucaramanga, with the objective of improving the safety of the population. The focus of this work addresses the variety of data that can influence the performance of the model. To achieve greater efficiency in prediction, the time was divided into 15-day windows, thus generating a total of 24 to 25 windows per year. Each window is composed of data vectors that include latitude, longitude and a pre-selected data set, with the variation of 3 to 5 time windows contributing to form the vector corresponding to the next window. The best performing neural network architecture which was the single layer model, consists of an input layer with 512 neurons and Relu activation function, followed by a dropout layer with a rate of 0.2 and an output layer with sigmoid function. The study evaluates the prediction performance in different training scenarios, considering changes in input size, data vector size, network output and resulting data type, including normalized values and values classified into 5 classes by the authors table 7 . The results reveal a neural layer model with an average error of 1.5254 and a mean square error of 0.0024 for the author-classified data.

---

[*] Undergrad Work

[**] Faculty of Engineering Physicomechanics. School of Electrical, Electronic and Telecommunications Engineering. Advisor: Jeison Arley Castillo Bohorquez; Co-Advisor: Jaime Guillermo Barrero Perez

Introducción

La predicción de delitos en una ciudad es de vital importancia para la seguridad publica. Esto permite a las autoridades locales anticiparse a los delitos y tomar medidas preventivas para reducir la incidencia de delitos. Esto también ayuda a las autoridades a identificar tareas problemáticas y a desarrollar estrategias para abordar los problemas de seguridad. La predicción de delitos también puede ayudar a las autoridades a identificar patrones de delincuencia y a desarrollar estrategias para prevenir futuros delitos. Pero para esto, primero definiremos la palabra "delito" la cual se define según la RAE (Real Academia Española) como la acción u omisión voluntario o imprudente penada por la ley del (2023). Identificar y realizar un planteamiento cuantitativo de estos delitos que suceden en la ciudad de Bucaramanga es necesario para nuevas políticas de seguridad, ya que con un problema bien identificado mejor sera la toma de decisiones de los recursos y en la administración de estos. Nuevas herramientas surgen con el avance de la tecnología y los modelos predictivos son cada vez más usados para temas de seguridad ciudadana.

Para la convivencia de una población se debe garantizar 4 factores, entre esos esta la seguridad que es "la acción de garantizar la protección de los derechos y libertades constitucionales y legales de las personas en el territorio nacional" (Ley 1801 de 2016, art.6º) PIS (2020). El incumplimiento de esta acción genera problemáticas como el desplazamiento de personas y/o empresas,

hurto a personas y/o viviendas llegando a escalar en la afectación de derechos fundamentales como la integridad física y la vida. De acuerdo con datos proporcionados por la Policia Nacional PIS (2020), algunos delitos han disminuido desde el 2017 al 2019, delitos como el acceso carnal violento y la violencia intrafamiliar se vio disminuido en este tiempo, sin embargo, otros delitos como el hurto a personas o el homicidio aumentaron considerablemente en esta ventanas de tiempo; el homicidio aumento un 24 % y el hurto a personas un 47 %. En cifras dadas por el DANE en el 2018 el hurto a personas produjo más de treinta mil víctimas en Bucaramanga, mucho más que cualquier otro delito estudiado por la misma entidad. Dicho todo esto nos permite inferir que es necesario dar atención a los delitos que ocurren en la ciudad, sin dejar de lado los que se vieron disminuidos.

Un caso actual de estudio es la ciudad de Bucaramanga donde diversos factores como la emergencia sanitaria, el fenómeno migratorio, la inflación, la movilidad, entre otros, han desarrollado en la ciudad un comportamiento atípico del delito que afecta notablemente a la seguridad de sus ciudadanos. Esta es una problemática que aqueja a todos sus habitantes y hace necesario la creación de nuevas y mejores políticas de seguridad que atiendan a todas las necesidades de su población. Para esto, es necesario contar con información confiable y actualizada para evaluar el progreso y los resultados de las diferentes políticas publicas PIS (2020).

La importancia de resolver este problema nace de la falta de información de datos reales, confiables y actualizados que permitan tomar decisiones más certeras a la hora de administrar los recursos de seguridad. Los principales beneficiados serian la comunidad Bumanguesa en general, con un estudio profundo sobre la seguridad y como esta puede ser considerablemente mejorada.

Con una mayor fuente de información las políticas de seguridad se podrán aplicar de manera correcta a zonas en las que son más necesarias, los organismos de seguridad se podrán centrar en puntos estratégicos para combatir los delitos de manera más directa centrando el foco en lo que podría ser el verdadero problema.

Este estudio no solo dejara un precedente en la ciudad, también sera el comienzo de más y mejores soluciones a las problemáticas de seguridad que ocurren en la ciudad de Bucaramanga. Mediante el PISCC (Plan Integral de Seguridad y Convivencia Ciudadana Para una Bucaramanga Segura) PIS (2020) se plantea la creación de nuevas políticas de seguridad basadas en datos confiables y correctamente analizados, con esto se espera un mejor diseño e implementación de políticas de seguridad para una ciudad más segura. Para esto es necesario el fortalecimiento del Observatorio de seguridad con el que se busca mejorar en 3 diferentes frentes, el primero consta de la contratación de diversos servidores públicos con diferentes habilidades en el manejo y análisis de datos cuantitativos y cualitativos, el segundo frente trata de la inversión de hardware y software para que sean manejados por los analistas anteriormente mencionados. Para el último frente, se busca la unificación de una base de datos de todos los procesos dentro de la secretaría del Interior y será de vital importancia en la toma de decisiones y en la coordinación de los organismos de seguridad de la ciudad.

En conclusión, en la ciudad de Bucaramanga se cuenta con una organización temprana y una hoja de ruta establecida para la unificación de los datos que se recolectan por medio de las diferentes instituciones que normalmente son las encargadas de recaudar este tipo de informaciones.

Es decir, se cuenta con la información, sin embargo esta se encuentra dispersa lo que dificulta la posibilidad de dar un análisis con mayor profundidad. Sumado a esto no permite dar un seguimiento claro y conciso sobre las políticas y las decisiones que se toman por parte de los organismos de seguridad, basadas en decisiones informadas.

1. Objetivos

1.1. Objetivo General

Implementar un algoritmo que permita predecir de forma espacio temporal, los delitos en contra del patrimonio económico cometidos en el área urbana del municipio de Bucaramanga - Santander, mediante el uso de machine learning.

1.2. Objetivos Específicos

- Revisar el estado del arte relacionado con algoritmos de machine learning para la predicción espacio temporal del delito.
- Preparar la base de datos, mediante el uso de ingeniería de datos para garantizar un buen proceso de entrenamiento.
- Implementar un algoritmo con machine learning para la predicción de delitos con base en lo descubierto en el estado del arte.
- Evaluar el desempeño del algoritmo en comparación con los modelos propuestos por la alcaldía de Bucaramanga.

2. Marco de referencia

En varios países se han implementado modelos similares de machine learning para la predicción del delito, en cada caso depende de la finalidad que se le dará, la forma en que se realice el tratamiento de los datos y la ventana temporal escogida. En Boston, se realizó la predicción del delito visualizándolo mediante “puntos calientes” donde la frecuencia de ocurrencia del crimen es alta y viceversa.Sus mayores falencias estaban en que sus datos no estaban correlacionados entre sí y que los algoritmos de aprendizaje automático supervisado no funcionan bien en este tipo de datos. Como posible mejora proponen el estudio de diferentes modelos de predicción a los que usaron y el uso de un mayor número de registros por años. En Colombia se realizó el estudio para definir cuál es la mejor ventana de tiempo para la predicción de un delito teniendo un sistema de localidades dentro del territorio de Bogotá, observando como la ventana de tiempo apropiada es de 14 díasSharma *et al.* (2021). En Bangladesh se realizó un estudio de 3 diferentes modelos de regresión con machine learning para predecir las tendencias y los patrones de la delincuencia en la ciudad Biswas y Basak (2019). Realizaron la predicción para solo 1 año, utilizaron modelos como regresión lineal y la regresión de árboles aleatorios, sin embargo, no realizaron una división espacial de la ciudad y la ventana temporal para la predicción fue de 1 año. Como conclusión final, se percibió que este tipo de acercamientos no son suficientes para la toma de decisiones.

En el desarrollo de la investigación se recurrió a múltiples fuentes para la definición de aspectos clave, entre los que destaca la determinación de la ventana temporal y la selección de modelos de machine learning. La elección de la ventana temporal se basó en la revisión de diversas literaturas. En ellas, se identificó que para este tipo de problemáticas, la ventana temporal más recurrente varía entre 7 y 15 días. Esta variación se debe, en gran medida, al volumen de datos disponibles en las bases de datos consultadas. Sin embargo, para determinar la duración óptima de la ventana temporal en su estudio, se guió por el trabajo de Zhuang et al., citado en la referencia cri (2017). En este documento se adopta una ventana de 2 semanas. Esta elección se traduce en aproximadamente 24-25 ventanas temporales por año, lo que se consideró adecuado y pertinente para la investigación en cuestión. En la elección de modelos de machine learning para la predicción del delito, se basó los diferentes documentos en el estado del arte, seleccionando aquellos que se adaptan mejor al tipo de datos en cuestión. Tras una exhaustiva revisión bibliográfica, se determinó que los modelos más prevalentes para este tipo de datasets incluyen árboles de decisión, capas densas con variadas arquitecturas y capas convolucionales. Estos modelos han demostrado ser efectivos en múltiples investigaciones previas, evidenciando su capacidad para gestionar la complejidad y las particularidades de los datos relacionados con el delito. La elección de estos modelos también consideró su flexibilidad y adaptabilidad a diferentes escenarios y variaciones en los datos.

2.1. Problemática y relevancia

En la actualidad no se cuenta con un mecanismo automatizado capaz de predecir los delitos contra el patrimonio económico ya sea mediante el uso de machine learning, programación o mé-

todos estadísticos que permita realizar este tipo de inferencias. Es por esto que se ha identificado la necesidad de diseñar e implementar una versión de una herramienta capaz de predecir este tipo de delitos en la ciudad de Bucaramanga, a la cual se restringirá únicamente este proyecto.

2.1.1. Contextualización del problema

El crimen contra el patrimonio económico ha sido una preocupación constante en muchas ciudades alrededor del mundo, y Bucaramanga no es la excepción. Esta ciudad, reconocida por su dinamismo económico y cultural, ha enfrentado en las últimas décadas retos significativos en cuanto a la seguridad y protección de bienes y activos de sus ciudadanos.

Segun "Human security handbook" hecho por The United Nations Trust Fund For Human Security para la convivencia de una población la seguridad es fundamental dado que es "la acción de garantizar la protección de los derechos y libertades constitucionales y legales de las personas en el territorio nacional" (Ley 1801 de 2016, art.6º) PIS (2020). El incumplimiento de esta acción genera problemáticas como el desplazamiento de personas y/o empresas, hurto a personas y/o viviendas llegando a escalar en la afectación de derechos fundamentales como la integridad física y la vida. De acuerdo con datos proporcionados por la Policía Nacional, algunos delitos han disminuido desde el 2017 al 2019, delitos como el acceso carnal violento y la violencia intrafamiliar se vio disminuido en este tiempo, sin embargo, otros delitos como el hurto a personas o el homicidio aumentaron considerablemente en esta ventanas de tiempo; el homicidio aumentó un 24 % y el hurto a personas un 47 %. En cifras dadas por el DANE en el 2018 el hurto a personas produjo

más de treinta mil víctimas en Bucaramanga, mucho más que cualquier otro delito estudiado por la misma entidad. Dicho todo esto nos permite inferir que es necesario dar atención a los delitos que ocurren en la ciudad, sin dejar de lado los que se vieron disminuidos.

La complejidad de la situación no solo radica en la ocurrencia de estos crímenes, sino también en la dificultad para su predicción. Diversos factores como la desigualdad económica, el crecimiento urbano descontrolado y la falta de oportunidades laborales son algunos de los catalizadores que, influyen directamente en la comisión de delitos patrimoniales en zonas urbanas.

En varios países se han implementado modelos similares de machine learning para la predicción del delito, en cada caso depende de la finalidad que se le dará, la forma en que se realice el tratamiento de los datos y la ventana temporal escogida. En Boston, se realizó la predicción del delito visualizándolo mediante “puntos calientes” donde la frecuencia de ocurrencia del crimen es alta y viceversa.Sus mayores falencias estaban en que sus datos no estaban correlacionados entre sí y que los algoritmos de aprendizaje automático supervisado no funcionan bien en este tipo de datos. Como posible mejora proponen el estudio de diferentes modelos de predicción a los que usaron y el uso de un mayor número de registros por años. En Colombia se realizó el estudio para definir cuál es la mejor ventana de tiempo para la predicción de un delito teniendo un sistema de localidades dentro del territorio de Bogotá, observando como la ventana de tiempo apropiada es de 7 días Daniel (2021). Por otro lado en Medellín se desarrollaron 3 modelos para la predicción de zonas calientes los cuales se evaluaron por exactitud,exactitud, la precisión, la exhaustividad y el valor F1. Los resultados de la regresión logística es 73.53 % de exactitud, exhaustividad de

66,27 %, un valor F1 de 73,01 % y una precisión de 81,38 % ; Para el modelo de árboles de decisión se tiene una exactitud de 76,25 %,una exhaustividad de 71,87 %, un valor F1 de 75,13 % y una precisión de 78,75 %. El que presentó mejores características fue el modelo de máquinas de soporte vectorial que obtuvo una exactitud de 76.06 %, exhaustividad de 80,57 %, un valor F1 de 76,06 % y una precisión de 72,11 % Daniel (2021).

En Bangladesh se realizó un estudio de 3 diferentes modelos de regresión con machine learning para predecir las tendencias y los patrones de la delincuencia en la ciudad Biswas y Basak (2019). Realizaron la predicción para solo 1 año, utilizaron modelos como regresión lineal y la regresión de árboles aleatorios, sin embargo, no realizaron una división espacial de la ciudad y la ventana temporal para la predicción fue de 1 año. Como conclusión final, se percibió que este tipo de acercamientos no son suficientes para la toma de decisiones.

2.2. Algoritmos actuales de detección

En la actualidad existen diferentes maneras de abordar este problema mediante el uso de machine learning, para brindar una solución se debe considerar que existen numerosos algoritmos robustos de machine learning. Estos algoritmos se dividen en tres grupos principales el no supervisado (unsupervised learning), que se comprende como un proceso de aprendizajes a través de tareas que se le asignan a la máquina en cuestión; el aprendizaje supervisado (supervised learning), que se entiende como el aprendizaje con una base de datos donde se conoce una respectiva estructura del dato a identificar, y el aprendizaje por refuerzo (Reinforcement learning), un aprendizaje

que consiste en estímulos que se aplican después de una prueba de el resultado, llegando así a un desarrollo casi perfecto.

Los algoritmos encontrados que se usan para resolver un problema de forecasting son los árboles de decisión (decision tree), bosque aleatorio (random forest), soporte de máquinas vectoriales y la regresión lineal aunque esta última no es una opción viable ya que se considera la media variable sensible a los valores atípicos Sharma *et al.* (2021).

En la ciudad de Boston se realizó un estudio de predicción del crimen donde se optaron por 2 opciones de las cuales estas cuenta con dos alternativas. Una opción es un modelo de árboles de decisión con una precisión de 53 % bajo el criterio de gini, gini es un parámetro predeterminado en el algoritmo del árbol de decisión que se usa para medir la homogeneidad de un nodo, y su alternativa, son árboles de decisión junto PCA (Principal Component Analysis) con una precisión de 53 % bajo criterio de gini ; La segunda opción es un modelo de bosque aleatorio con una precisión de 52 % y su alternativa de un modelo de bosque aleatorio junto PCA con una precisión de 60 % mostrando una mejora entre los dos modelos Ordóñez *et al.* (2020).

En el problema propuesto se investigó sobre un proyecto realizado en Colombia para la predicción de la tendencia del hurto, en el cual la Fiscalía suministró un dataset de 2,662,402 registros tomados desde el año 1960 hasta el 2019, para el cual después de realizar una limpieza de filtrado de la base de datos usada se obtuvo un total de 350,320 registros. Se propuso la creación de un modelo de machine learning utilizando máquinas vectoriales orientado a problemas de regresión

con la métrica de epsilon igual a 0.1 con el cual se puede predecir el aumento o disminución de actos delictivos, obteniendo un buen comportamiento en relación con los datos proporcionados, llegando a una correlación creciente que oscila entre el 87 % y 100 % Reyes *et al.* (2020).

En el proceso de abordar el desafío presentado, una de las principales dificultades a las que se enfrenta, radica en la obtención de un conjunto de datos robusto y confiable. La calidad del dataset es esencial, ya que de ella depende la precisión y generalización de cualquier modelo de aprendizaje automático. A este reto se suma el proceso de evaluación de los resultados generados por el modelo. Es imperativo establecer métricas adecuadas que se alineen con los objetivos del estudio. La complejidad radica en que, a menudo, el error no debe cuantificarse simplemente basándose en el número absoluto de predicciones erróneas. En lugar de ello, es más apropiado considerar la discrepancia entre los valores reales y los predichos por el modelo. Esta diferencia, conocida técnicamente como residuo, es crucial para entender la efectividad y precisión del modelo. En este contexto, el error no debe interpretarse simplemente como la cantidad de predicciones incorrectas, sino que debe analizarse en términos de la magnitud de la desviación entre el valor esperado y el estimado, lo cual ofrece una perspectiva más matizada sobre el desempeño del sistema en cuestión. Es fundamental adoptar un enfoque riguroso en estas áreas para garantizar que las conclusiones derivadas del estudio sean válidas y aplicables en situaciones prácticas.

3. Conjuntos de datos

3.1. Descripción de los datos

Para el desarrollo del proyecto se creó una base de datos a partir de los datos proporcionados por el Sistema de Información Estadístico, Delincuencial, Contravencional y Operativo (SIEDCO). Esta base de datos es proporcionada a través del tablero digital de observación del delito de la alcaldía de Bucaramanga y es de uso libre. Después de un análisis preliminar, se limitó a sólo los datos del área metropolitana de la ciudad de Bucaramanga, en los cuales se encontró un total de 135.076 registros con las siguientes etiquetas:

Tabla 1
Registros con las etiquetas y tipo de datos.

DATO	TIPO DE DATO
ORDEN	INT
ARMAS_MEDIOS	STRING
BARIOS_HECHO	STRING
LATITUD	STRING
LONGITUD	STRING
ZONA	STRING
NOM_COMUNA	INT
AÑO	INT
MES	INT
DIA	INT
DIA_SEMANA	INT
DESCRIPCION_CONDUCTA	STRING
CONDUCTA	STRING
CLASIFICACIONES DELITO	STRING
EDAD	INT
CURSO_DE_VIDA	STRING
ESTADO_CIVIL_PERSONA	STRING
GÉNERO	STRING
MOVIL_AGRESOR	STRING
MOVIL_VICTIMA	STRING

3.2. Metodología de filtrado

La fase inicial del proyecto implicó la recopilación de datos existentes desde el año 2016 hasta el 2021, proporcionados por la base de datos, lo que resultó en un conjunto de 73.302 registros disponibles. Durante el análisis de estos registros, se identificaron errores en el llenado de la base de datos, que incluyen:

- Nan (Not a Number)
- Guiones mal posicionados
- Ambigüedades en puntos y comas para decimales
- Latitud y longitud intercambiadas
- Latitud y longitud fuera de los límites del polígono de Bucaramanga

A través de programación en Python se implementa un primer filtro el cual elimina y corrige todas las inconsistencias presentadas anteriormente y cuyo algoritmo está contenido en el código del apéndice A. Como se puede apreciar en la figura 1 , en el recuadro amarillo, uno de los errores más comunes fue el intercambio de latitudes por longitudes.

Para esto se implementa el siguiente código de la figura 2, el cual busca los valores de latitud y longitud que están intercalados en el dataframe, evaluando y corrigiendo los datos dejando

ARMAS MEDIOS	BARRIOS HECHO	LATITUD	LONGITUD	ZONA	NOM COMUNA	AÑO	MES	DIA	DIA SEMANA	DESCRIPCION CONDUCTA
ARMA BLANCA / CORTOPUNZANTE	BOLARQUI	7.1095641729	-73.1139885602	URBANA	12. Cabecera del Llano	2020	11. Noviembre	29	07. Domingo	ARTÍCULO 239. HURTO A PERSONAS
ARMA BLANCA / CORTOPUNZANTE	PORTON DEL TEJAR	7.1014178413	-73.1033155741	URBANA	16. Lagos del Cacique	2020	11. Noviembre	29	07. Domingo	ARTÍCULO 239. HURTO A PERSONAS
ARMA DE FUEGO	URB. MONTE REDONDO	7.0934214471	-73.134076406	URBANA	17. Mutis	2020	11. Noviembre	29	07. Domingo	ARTÍCULO 239. HURTO A PERSONAS
SIN EMPLEO DE ARMAS	CENTRO	7.1215293293	-73.1259339038	URBANA	15. Centro	2020	11. Noviembre	29	07. Domingo	ARTÍCULO 239. HURTO A PERSONAS
SIN EMPLEO DE ARMAS	NUEVO SOTOMAYOR	7.1168572036	-73.1139472884	URBANA	12. Cabecera del Llano	2020	11. Noviembre	29	07. Domingo	ARTÍCULO 239. HURTO A PERSONAS
PALANCAS	QUINTAS DEL CACIQUE	7.1008333875	-73.1010263845	URBANA	16. Lagos del Cacique	2020	11. Noviembre	29	07. Domingo	ARTÍCULO 239. HURTO A RESIDENCIAS
SIN EMPLEO DE ARMAS	SAN GERARDO	7.1035470779	-73.1196123924	URBANA	08. Sur Occidente	2020	11. Noviembre	29	07. Domingo	ARTÍCULO 239. HURTO A RESIDENCIAS
SIN EMPLEO DE ARMAS	NUEVA GRANADA	7.0979676471	-73.113998237	URBANA	09. La Pedregosa	2020	11. Noviembre	29	07. Domingo	ARTÍCULO 239. HURTO A RESIDENCIAS
ARMA BLANCA / CORTOPUNZANTE	SAN FRANCISCO	7.133538947	-73.1257484542	URBANA	03. San Francisco	2020	11. Noviembre	30	01. Lunes	ARTÍCULO 239. HURTO A PERSONAS
SIN EMPLEO DE ARMAS	ANTONIA SANTOS	7.1263915456	-73.1206472279	URBANA	13. Oriental	2020	11. Noviembre	30	01. Lunes	ARTÍCULO 239. HURTO A PERSONAS
SIN EMPLEO DE ARMAS	MEJORAS PUBLICAS	7.122418966	-73.1130526369	URBANA	13. Oriental	2020	11. Noviembre	30	01. Lunes	ARTÍCULO 239. HURTO A PERSONAS
ARMA BLANCA / CORTOPUNZANTE	JARDINES DE COAVICONSA	7.0856666937	-73.1243549209	URBANA	11. Sur	2020	11. Noviembre	30	01. Lunes	ARTÍCULO 111. LESIONES PERSONALES
ARMA DE FUEGO	LA VICTORIA	7.0978433356	-73.1191366143	URBANA	06. La Concordia	2020	11. Noviembre	30	01. Lunes	ARTÍCULO 111. LESIONES PERSONALES
ARMA BLANCA / CORTOPUNZANTE	SAN FRANCISCO	7.133538947	-73.1257484542	URBANA	03. San Francisco	2020	11. Noviembre	30	01. Lunes	ARTÍCULO 239. HURTO A PERSONAS
SIN EMPLEO DE ARMAS	ANTONIA SANTOS	7.1263915456	-73.1206472279	URBANA	13. Oriental	2020	11. Noviembre	30	01. Lunes	ARTÍCULO 239. HURTO A PERSONAS
SIN EMPLEO DE ARMAS	MEJORAS PUBLICAS	7.122418966	-73.1130526369	URBANA	13. Oriental	2020	11. Noviembre	30	01. Lunes	ARTÍCULO 239. HURTO A PERSONAS
SIN EMPLEO DE ARMAS	GIRARDOT	7.1224986192	-73.1347855288	URBANA	04. Occidental	2020	11. Noviembre	30	01. Lunes	ARTÍCULO 239. HURTO A RESIDENCIAS
ARMA BLANCA / CORTOPUNZANTE	GIRARDOT	-73,1342680301	7,124289317	URBANA	04. Occidental	2020	12. Diciembre	1	02. Martes	ARTÍCULO 239. HURTO A PERSONAS
SIN EMPLEO DE ARMAS	SAN FRANCISCO	-73,1277098195	7,1348339071	URBANA	03. San Francisco	2020	12. Diciembre	1	02. Martes	ARTÍCULO 239. HURTO A PERSONAS
SIN EMPLEO DE ARMAS	BALCONES DEL SUR	-73,1217085423	7,0792874118	URBANA	11. Sur	2020	12. Diciembre	1	02. Martes	ARTÍCULO 239. HURTO A PERSONAS
ARMA DE FUEGO	VILLA CANDADO	-73,122685627	7,083038818	URBANA	11. Sur	2020	12. Diciembre	1	02. Martes	ARTÍCULO 239. HURTO A PERSONAS
SIN EMPLEO DE ARMAS	PROVENZA	-73,1167192585	7,0862769857	URBANA	10. Provenza	2020	12. Diciembre	1	02. Martes	ARTÍCULO 239. HURTO A PERSONAS
SIN EMPLEO DE ARMAS	MEJORAS PUBLICAS	-73,1143725079	7,1224205647	URBANA	13. Oriental	2020	12. Diciembre	1	02. Martes	ARTÍCULO 239. HURTO A PERSONAS

Figura 1. Errores de latitud y longitud en la base de datos.

así corregido los datos con un total de 64.602 registros listos para la importación y su filtrado dependiendo de los días y tipo de delito que se necesite para el ingreso a los modelos.

La base de datos contiene 4 tipologías de delitos las cuales son: delitos contra la libertad, integridad y formación sexuales, delitos contra el patrimonio económico, delitos contra la libertad, integridad y formación sexuales y por último delitos contra la vida y la integridad personal; y el filtro comprende todo delito que esté en la tipología de delito contra el patrimonio económico dentro el cual se encuentra : hurto a personas, hurto a residencias, hurto a motocicletas, hurto a automotores, hurto a entidades comerciales y extorsión. La cantidad de delitos se recopila en la tabla 2.

Con esto se hace el segundo filtro por tipologías. Se retiró el Hurto a automotores y Extor-

```
df3.reset_index(drop = True, inplace=True)
df3["LATITUD"]=df3["LATITUD"].replace(",",".",regex=True)
df3["LONGITUD"]=df3["LONGITUD"].replace(",",".",regex=True)
df3['LONGITUD']=df3['LONGITUD'].astype(str)
df3['LATITUD']=df3['LATITUD'].astype(str)

df3["LONGITUD"][33872]="-"+df3["LONGITUD"][33872]
i=0
temp=0
for i in df3.index:
    if df3.at[i,"LATITUD"][1]==".": #df3.at[i,"LONGITUD"].__contains__('.')
        if df3.at[i,"LATITUD"][0]!="-":
            df3['LATITUD'][i]=float(df3['LATITUD'][i])
    else:
        if df3.at[i,"LATITUD"][0]=="-":
                temp=df3['LATITUD'][i]
                df3['LATITUD'][i]=df3['LONGITUD'][i]
                df3['LONGITUD'][i]=temp
                df3['LATITUD'][i]=float(df3['LATITUD'][i])
        else:
            df3["LATITUD"][i]=df3["LATITUD"][i][:1]+"."+df3["LATITUD"][i][1:]
            df3['LATITUD'][i]=float(df3['LATITUD'][i])

    if df3.at[i,"LONGITUD"][3]==".":
        df3['LONGITUD'][i]=float(df3['LONGITUD'][i])
    else:
        df3["LONGITUD"][i]=df3["LONGITUD"][i][:3]+"."+df3["LONGITUD"][i][3:]
        df3['LONGITUD'][i]=float(df3['LONGITUD'][i])
```

Figura 2. Código para corregir los errores de latitud y longitud en la base de datos.

Tabla 2
Cantidad de cada delito de la base de datos.

Delito contra el patrimonio económico	Cantidad de delito
Hurto a personas	28.335
Hurto a residencias	2.914
Hurto a motocicletas	2.505
Hurto a automotores	106
Hurto a entidades comerciales	6.338
Extorsión	358

sión por su baja incidencia; dando así un total de 40.556 registros con los cuales se comenzó con el pre-procesamiento de los datos para el entrenamiento de los modelos.

4. Modelos para la predicción de delitos

Para el desarrollo del proyecto se hizo uso de 5 modelos diferentes en los cuales se encuentran:

4.1. Modelo capas densas sencillo

En el presente modelo, se propone una arquitectura compuesta por cuatro capas distintas, estructuradas de la siguiente manera: Inicialmente, se encuentra la capa de entrada que sirve como interfaz para los datos iniciales que se desean procesar. A continuación, se halla una capa densamente conectada, comúnmente denominada “fully connected”, que cuenta con 512 neuronas. Esta capa emplea la función de activación Rectified Linear Unit (ReLU), la cual es ampliamente reconocida en el campo de las redes neuronales debido a su eficiencia y sencillez. La función ReLU opera tomando la entrada y devuelve el valor entrante si es positivo, mientras que devuelve cero para valores negativos. Esto se traduce matemáticamente como $f(x) = max(0,x)$, lo que implica que la salida variará entre 0 y un valor positivo infinito dependiendo de la operación lineal que involucra los valores entrantes, así como los pesos y bias correspondientes.

Posterior a esta capa densa, se implementa una capa de dropout, técnica ampliamente utilizada para mitigar el sobreajuste en modelos de aprendizaje profundo. En este modelo en particular,

se establece un ratio de 0.2, lo que significa que durante el proceso de entrenamiento, el 20 % de las neuronas se desactivan aleatoriamente en cada iteración, promoviendo así una distribución más robusta y generalizada de las características aprendidas.

Finalmente, la arquitectura culmina con la capa de salida, que, según se especifica, puede tener un rango de neuronas que varía entre 1 y 5. Esto sugiere que el modelo podría estar diseñado para clasificación en múltiples categorías, aunque se requeriría más detalle para determinar la naturaleza exacta de las clases o el tipo de problema que se intenta abordar.

En el marco del experimento, el algoritmo de aprendizaje automático fue entrenado mediante el empleo del método '.fit', una función provista por la biblioteca Scikit-learn. Este proceso de optimización iterativa se ejecutó a lo largo de un total de 25 épocas. A fin de mejorar la eficiencia y la robustez del modelo en fase de entrenamiento, se implementaron dos mecanismos de retroalimentación denominados "callbacks".

El primer callback incorporado fue "EarlyStopping", configurado con un margen de tolerancia de cinco iteraciones. Este mecanismo monitoreaba la métrica "val_loss" (error de validación) con el propósito de interrumpir el ciclo de entrenamiento si no se observan avances significativos en la minimización de dicha métrica. Esto contribuye a evitar el sobreajuste y a optimizar el uso de recursos computacionales.

El segundo callback utilizado fue "ReduceLROnPlateau", que estaba diseñado para adaptar dinámicamente la tasa de aprendizaje en caso de que el descenso en el error de validación se

estancara. Este callback se ajustó con un factor de reducción de 0.1 en la tasa de aprendizaje y se programó para tener una paciencia de 10 iteraciones antes de activar su funcionalidad. Al igual que el primero, también focalizó su atención en la métrica “val_loss”.

4.2. Modelo capas densas

El presente modelo se tomo de un trabajo el cual se baso en la detección automática de primeros arribos usando redes neuronales artificiales en trazas terrestres reales del catálogo sísmico colombianoJosé (2023),en la cual se ha implementado una arquitectura compuesta por cinco estratos distintos, diseñados de manera secuencial. Inicialmente, se encuentra una capa densamente conectada que aloja un total de 680 neuronas. Esta capa adopta una función de activación conocida como Rectified Linear Unit"(ReLU), la cual es ampliamente empleada debido a su capacidad de añadir no linealidad al modelo sin requerir de cálculos computacionales excesivos. Además, el proceso de inicialización de los pesos de esta capa se lleva a cabo mediante la estrategia "Glorot Uniform". Esta técnica, también conocida como inicialización Xavier, se basa en establecer valores dentro de un rango predeterminado que depende de la cantidad de unidades de entrada y salida en el tensor de peso, optimizando así la propagación de gradientes en las etapas iniciales del entrenamiento.

Posterior a esta capa densa, se ha integrado una capa "dropout"que funciona como una técnica de regularización. Esta capa actúa deshabilitando aleatoriamente una fracción de 0.1 de las entradas, reduciendo de esta forma la posibilidad de sobre-ajuste al inhibir la dependencia excesiva

en determinadas neuronas durante el proceso de entrenamiento.

Finalmente, para la salida del modelo, se implementa una capa densa cuyo número de neuronas oscila entre 1 y 5, dependiendo de la estructura intrínseca de los datos a procesar. Esta capa de salida emplea la función de activación "sigmoid", que permite obtener respuestas en un rango que se extiende entre -1 y 1, siendo adecuada para tareas de clasificación binaria y multi-clase cuando se adecúa el número de neuronas en la capa de salida.

El algoritmo de aprendizaje automático fue entrenado mediante el empleo del método ".fit" Este proceso de optimización iterativa se ejecutó a lo largo de un total de 25 épocas. A fin de mejorar la eficiencia y la robustez del modelo en fase de entrenamiento, se implementaron dos mecanismos de retroalimentación denominados "callbacks" los cuales fueron "EarlyStopping" y "ReduceLROnPlateau" .

4.3. Modelo capas densas siames

En el diseño propuesto se introduce una arquitectura neural de capas densas ramificadas, caracterizada por la operación en paralelo de dos secuencias de capas densas. La primera rama cuenta con una capa densa compuesta por 128 neuronas, mientras que la segunda se estructura con 64 neuronas. Ambas ramas implementan la función de activación Rectified Linear Unit (ReLU), ampliamente reconocida por sus propiedades de regularización y eficiencia en la optimización del proceso de entrenamiento.

Posterior a estas ramificaciones, las salidas de ambas secuencias convergen hacia una capa de concatenación. Está, a su vez, se conecta a una sucesión de tres capas densas con 256, 64 y 32 neuronas, respectivamente. Nuevamente, se opta por la función de activación ReLU debido a su efectividad en la propagación de gradientes y minimización de la saturación.

El algoritmo de aprendizaje automático fue entrenado mediante el empleo del método ".fit" Este proceso de optimización iterativa se ejecutó a lo largo de un total de 25 épocas. A fin de mejorar la eficiencia y la robustez del modelo en fase de entrenamiento, se implementaron dos mecanismos de retroalimentación denominados "callbacks"los cuales fueron "EarlyStoppingz "ReduceLROnPlateau".

4.4. Modelo de árboles de decisión regresiva

El modelo propuesto se caracteriza por una configuración basada en tres parámetros esenciales. Uno de los parámetros cruciales es el criterio de evaluación, que se ha seleccionado en función del error absoluto medio (EAM). Dicho error representa la desviación promedio entre los valores pronosticados y los valores reales observados en el conjunto de datos. Adicionalmente, para evitar la complejidad excesiva del modelo y prevenir posibles situaciones de sobreajuste, se ha establecido una restricción en la profundidad del árbol de decisión, limitándose a un máximo de cinco niveles. Esta limitación garantiza un equilibrio entre la precisión del modelo y su capacidad generalizadora al enfrentarse a nuevos datos no vistos anteriormente.

En el marco del experimento, el algoritmo de aprendizaje automático fue entrenado me-

diante el empleo del método ".fit", con el 80 % total de los datos, haciendo el test con el 20 % restante.

4.5. Modelo de capas convolucionales unidimensional

En el desarrollo del modelo propuesto, se implementó inicialmente una capa convolucional unidimensional (Conv1D). Esta capa se caracterizó por contener 32 filtros, cada uno de tamaño 3x1, optimizados para extraer características específicas del conjunto de datos unidimensional en estudio. Tras el proceso de convolución, se integró una capa de aplanamiento (Flatten), cuyo propósito es transformar la estructura multidimensional resultante en un vector unidimensional, facilitando así su integración con capas subsecuentes. Posteriormente, se diseñó una estructura secuencial de tres capas densamente conectadas (Dense layers), de las cuales las primeras dos poseen 128 y 32 neuronas respectivamente. Ambas utilizan la función de activación Rectified Linear Unit (ReLU), favoreciendo la no linealidad y mejorando la capacidad del modelo para aprender patrones complejos del conjunto de datos. Finalmente, la capa de salida, también de tipo densa, varía su cantidad de neuronas entre 1 y 5, adaptándose a la estructura específica de los datos a ser evaluados. Esta capa utiliza la función de activación sigmoide, idónea para tareas de clasificación binaria o multiclase, dependiendo de la configuración adoptada.

El aprendizaje se hizo mediante el empleo del método ".fit", una función provista por la biblioteca Scikit-learn. Este proceso de optimización iterativa se ejecutó a lo largo de un total de 25 épocas.

Para el manejo del conjunto de datos de todos los modelos anteriormente mencionados, se segmentan en dos subconjuntos distintos para facilitar tanto el entrenamiento como la validación del modelo. Específicamente, el 80 % del total de datos se destinó para el entrenamiento, mientras que el 20 % restante se reservó para las operaciones de validación. Esta división se efectuó con un tamaño de lote compuesto por 32 muestras, optimizando así la eficacia del proceso de aprendizaje y permitiendo una estimación más precisa del gradiente durante la optimización de los parámetros del modelo.

4.6. Evaluación y métricas de desempeño

En la etapa evaluativa del presente estudio, los modelos propuestos se sometieron a un minucioso análisis bajo dos métricas específicas: el Valor Medio Absoluto del Error (MAE, por sus siglas en inglés) y el Error Cuadrático Medio (MSE, por sus siglas en inglés). El MAE es un parámetro que cuantifica la media de las diferencias absolutas entre las predicciones y las observaciones reales, proporcionando una medida directa y comprensible de la magnitud del error de predicción. El MSE, por otro lado, mide el promedio de los cuadrados de los errores, penalizando de manera más severa los errores grandes, lo que lo hace particularmente útil en contextos donde estos pueden ser especialmente perjudiciales.

Estas métricas fueron escogidas con sumo cuidado, dada la naturaleza intrínseca de los datos manejados en este estudio, los cuales se caracterizan por valores no clasificables sino comparables. Esta particularidad realza la complejidad en la determinación de una precisión acertada,

ya que cualquier divergencia con respecto al valor real se categoriza de inmediato como un error, sin tomar en cuenta la magnitud de la discrepancia entre el valor predicho y el valor real perseguido. Este escenario subraya la importancia de implementar métricas evaluativas que proporcionen una vista más holística y matizada del rendimiento del modelo. Asimismo, permiten una interpretación más precisa y ajustada de los resultados obtenidos, facilitando la implementación de mejoras sustanciales y la optimización continua del modelo en cuestión, asegurando su robustez y confiabilidad en aplicaciones prácticas y reales.

5. Procesamiento de datos

5.1. Preparación de los datos

Para llevar a cabo el entrenamiento de los modelos, es imperativo someter los datos recopilados de la base de datos a un proceso de filtrado, con el objetivo de extraer únicamente la información relevante, que en este caso corresponde a los delitos que serán utilizados en las predicciones. Este proceso requiere, en primer lugar, definir qué categorías de datos serán empleadas en la construcción de los vectores de entrenamiento. Es necesario realizar una división tanto espacial como temporal, permitiendo así la generación de datos adecuados para la alimentación de los modelos de predicción. Finalmente, se procede a la creación de un conjuntos de datos y a la agrupación de vectores únicos, lo cual facilita el análisis comparativo y la evaluación del rendimiento de los modelos con el objetivo de determinar si se logra una mejora en los resultados.

5.1.1. Definiciones

Se especificaron las variables de “fecha inicio”, “fecha fin” y una “fecha variable”. Esta última actúa como una referencia para ajustar la “fecha inicio” hacia el siguiente intervalo, permitiendo así registrar la cantidad de delitos en cada segmento de la grilla durante ese período específico. Con las variables y parámetros del conjunto $S = \{s1, s2, ...\}$, que abarca todas las di-

visiones espacio-temporales, es posible identificar las características correspondientes para cada sector espacial:

$$L^s = \{x_s, y_s\}$$

$$M_t^s = \{m_{t-3}^s, m_{t-2}^s, m_{t-1}^s, \ldots, m_{t-i}^s\}$$

$$P_s = \{P_s\}$$

$$A_s = \{A_s\}$$

Siendo m_{t-3}, m_{t-2}, m_{t-1} el histórico del número de delitos en cada ventana temporal el cual se normaliza con el valor máximo de delitos posible en una casilla correspondiendo a 43, el periodo se normalizo entre la cantidad de ventanas que hay posible en un año siendo esta de 24 a 25 y la variable del año normalizado para un espacio de 10 años. Siendo m_t la semana que se está evaluando o la matriz siguiente a los valores ingresados en el vector.

5.1.2. Selección de datos

En la selección de los datos se tuvieron en cuenta 3 factores:

1. En la construcción del conjunto de datos definitivo, se requiere realizar una agrupación de los incidentes que comparten similitudes en el espacio temporal. Sin embargo, a medida que se

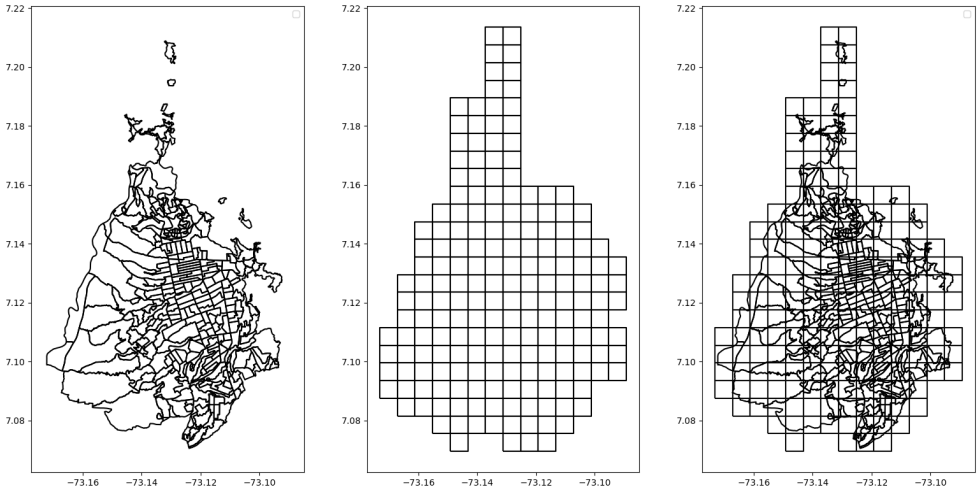
introducen nuevas características en esta división, se observa una disminución en la cantidad de incidentes disponibles para su análisis. En consecuencia, se toma la determinación de utilizar exclusivamente los atributos asociados a la fecha y a las coordenadas geográficas. Se opta por descartar los campos que contienen información sobre las características de los demás delitos.

2. Se decide trabajar con un periodo de tiempo entre enero de 2016 y octubre 4 del 2021, un periodo de aproximadamente 70 meses.

3. Se opta por excluir del conjunto de datos final los delitos de extorsión y hurto de automóviles debido a su baja frecuencia, lo que resulta en una disminución en el número total de datos de 66,350 a 40,556.

5.2. Etapa de división espacial

A partir de los datos recolectados se ve la necesidad de generar una división espacial para la asignación de todos los crímenes. Para el caso que nos compete, en la ciudad de Bucaramanga esta distribución se puede hacer mediante las divisiones administrativas de la ciudad como son los barrios, pero uno de los problemas que se presenta es la diversidad de áreas que contiene cada sector. Por tanto para darle un nivel de regularización a los datos se realiza una división por malla espacial para asegurar densidades comparables del crimen. La malla espacial es la mostrada en la figura 3(b).



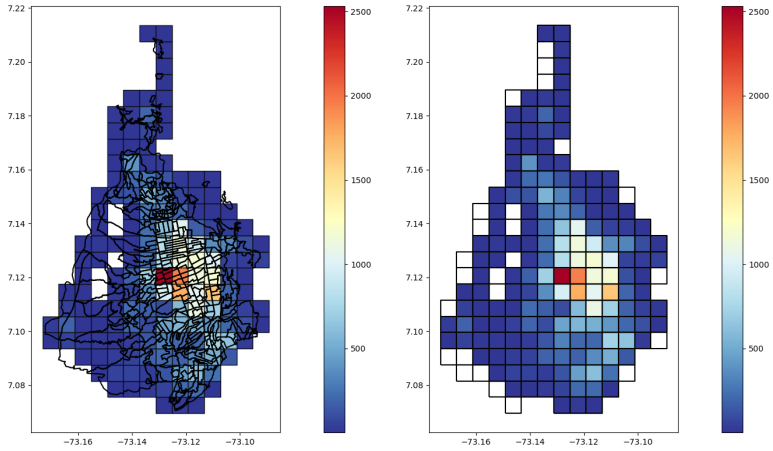
(a) Mapa Bucaramanga (b) Grilla de Bucaramanga (c) Grilla + mapa de Bucaramanga

Figura 3. Escala espacial para la ciudad de Bucaramanga.

En específico, se ha tomado como punto de partida la superficie promedio de los barrios de Bucaramanga, que se cifra en 0.4436 km². Como resultado de este cálculo, se ha determinado que las celdas de la malla tienen un lado de 666 metros, lo que permite un agrupamiento satisfactorio del crimen en áreas adecuadas.

El área total de la ciudad cubierta por cada sección de la malla corresponde a 0.3175 km². La disposición de las celdas en esta malla genera un total de 375 unidades distribuidas en un arreglo de 25 columnas y 15 filas. No obstante, dado que el enfoque del estudio se centra en el área urbana de la ciudad, se ha procedido a una selección rigurosa de las celdas que se hallan dentro de esta zona. Como resultado de este proceso, se ha obtenido un conjunto de 192 celdas de interés, que servirán como base fundamental en el desarrollo de los experimentos subsiguientes.

Una vez efectuada la asignación de incidentes a las respectivas celdas, se procede a realizar una detenida observación de los resultados, los cuales se exhiben de manera detallada en la figura 4(b). En dicha representación gráfica, se constata que únicamente un conjunto de 167 celdas, lo que constituye un porcentaje del 86.97 % respecto al total de la cuadrícula, alberga registros válidos durante el periodo comprendido entre el primero de enero de 2016 y el diez de octubre de 2021. Conviene destacar que, en contraposición, 25 celdas se excluyen de este análisis, ya que no han registrado incidentes en el contexto de la grilla en cuestión. Este análisis revela patrones significativos en la distribución de datos, ofreciendo una visión precisa de la dinámica espacial de los incidentes y sus relaciones con las celdas de la cuadrícula.



(a) Mapa Bucaramanga (b) Grilla de Bucaramanga

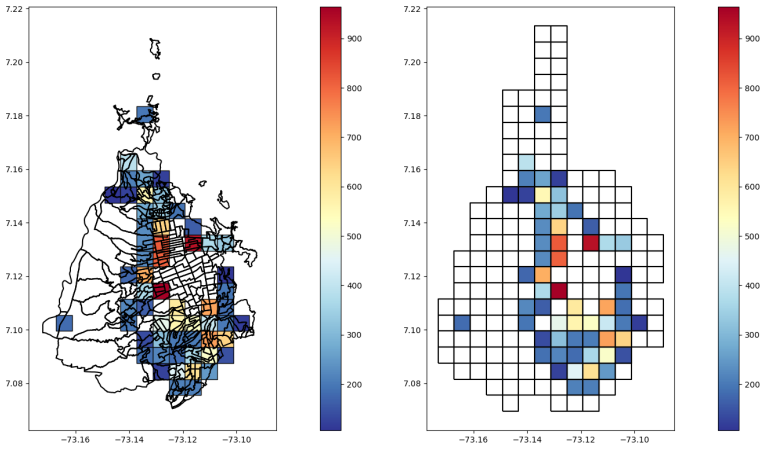
Figura 4. Distribución de hurtos en la ciudad de Bucaramanga entre el 2016 y 2021.

Tabla 3
Estadísticas descriptivas de la distribución de las celdas.

mean	std	min	25 %	50 %	75 %	95 %	max
242	408	1	8	56	238	1117	2531

En la tabla 3 se muestran las estadísticas descriptivas de las celdas basadas en la cantidad de

delitos. Se evidencia que se debe realizar el enfoque en celdas que tienen una cantidad significativa de delitos, esas que potencialmente podrían identificarse como zonas de alta incidencia, evitando valores atípicos. Al analizar solo las celdas que tienen más de 60 y menos de 1200 delitos (que caen entre los percentiles 50 y 95), la figura 5 da una representación más clara de la distribución de incidentes, con un promedio de 420 y una desviación estándar de 336.



(a) Mapa Bucaramanga (b) Grilla de Bucaramanga

Figura 5. Distribución de hurtos en la ciudad de Bucaramanga entre el 2016 y 2021.

Tabla 4
Estadísticas descriptivas de la distribución de las celdas.

mean	std	min	25 %	50 %	75 %	95 %	max
420	336	66	183	256	607	1127	1185

La distribución de delitos entre las celdas evidencia un sesgo hacia la derecha, lo que denota una presencia marcada de celdas con mínimas incidencias de delito. Sin embargo, existen ciertas celdas que registran cantidades significativamente altas de delitos, impactando la media y emergiendo como potenciales zonas de alto riesgo o “puntos calientes”. La notable desviación estándar revela una heterogeneidad pronunciada en la frecuencia delictiva entre estas celdas. Al restringir

el análisis a celdas que presentan incidentes ni muy bajos ni excesivamente altos, nos enfocamos en una franja intermedia. Esta selección nos brinda una visión más clara de zonas con incidencias moderadas a altas, que, sin intervención, podrían evolucionar en futuros focos de alta criminalidad.

5.3. Etapa de división temporal

Con la delimitación de estas zonificaciones espaciales debidamente definidas, la siguiente etapa crucial radica en el establecimiento de una segmentación temporal precisa, un factor fundamental que allana el camino hacia la generación de las características esenciales para el conjunto de datos final, como se ilustra en la tabla 5, la elección de este parámetro representa un acto de equilibrio entre la disponibilidad de datos necesarios para el proceso de entrenamiento y la profundidad de la información contenida en cada intervalo espacio-temporal.

Para este propósito, se ha llevado a cabo una comparación considerando intervalos temporales de 1 día, 8 días, 15 días y 29 días. El razonamiento detrás de la elección de medidas en días, en lugar de semanas, se basa en la practicidad inherente al proceso de adquisición y cálculo de datos en el contexto de implementación del código. Esta selección cuidadosamente ponderada permitirá un análisis más efectivo y una gestión eficiente de los recursos disponibles durante el desarrollo del modelo de predicción.

Para intervalos de tiempo extremadamente cortos, como un día, se aprecia una densidad significativamente elevada de celdas utilizadas para la modelización de incidentes delictivos. Sin embargo, en cada subdivisión temporal, los recuentos son notoriamente bajos, generalmente osci-

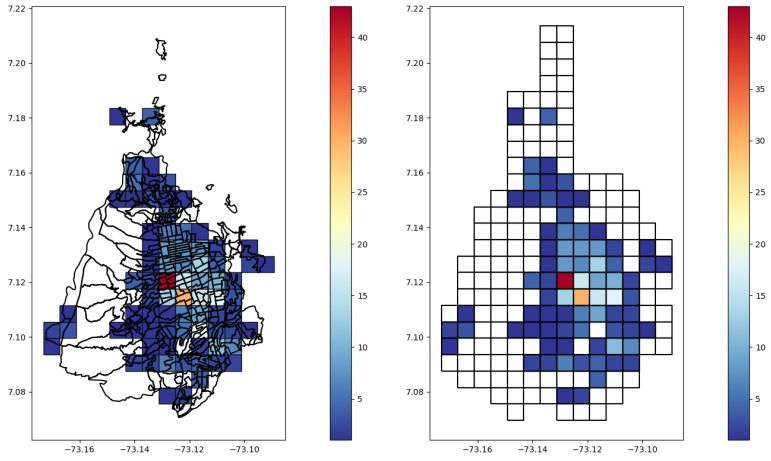
Tabla 5
Comparación de diferentes valores para la ventana temporal.

Variable	1 Día	7 Días	15 Días	29 días
Ventanas Temporales	2103	301	141	73
Total de Divisiones	18924	2709	1269	657
Promedio de incidentes	1	2	2	4
Percentil 95 %	2	8	16	25
Máximo de Incidentes	34	23	43	78
Celdas con Incidentes por Ventana Temporal	9	28	28	72
% Celdas con Incidentes por Ventana Temporal	4.68 %	14.5 %	14.5 %	37.5 %

lando entre 0 y 1 incidente. En contraste, cuando se consideran ventanas temporales más amplias, como la de 29 días, se percibe una notable mejora en la calidad de los datos. En estas condiciones, la media de incidentes por subdivisión espacio-temporal asciende a 4, y el percentil 95 alcanza un valor de 25. No obstante, esta ganancia en calidad se traduce en una disminución sustancial en la cantidad de datos disponibles para la fase de entrenamiento, especialmente cuando se compara con las subdivisiones temporales más breves.

En este proyecto de investigación, se ha tomado la decisión de emplear un valor óptimo basado en fundamentos teóricos y en la revisión de la literatura estudiada. Se ha seleccionado una ventana temporal de 15 días, ya que presenta un porcentaje satisfactorio de muestras disponibles para el entrenamiento de los modelos, en comparación con las de 7 días y 29 días. A pesar de que el promedio de incidentes por subdivisión es de 2, el percentil 95 alcanza un valor de 16, y un valor de delitos máximo de 43 como se muestra en la figura 6. Se asume que las áreas de alta incidencia delictiva se identificarán a partir de este percentil, lo que implica que en cada período quincenal se

estarían manejando, al menos, 16 incidentes por grupo de patrullaje.



(a) Mapa Bucaramanga (b) Grilla de Bucaramanga

Figura 6. Distribución de hurtos en la ciudad de Bucaramanga para la ventana temporal de 15 días en su valor máximo.

Esta elección conlleva la generación de un total de 141 ventanas temporales, distribuidas en 192 celdas. La base de datos resultante consta de 1269 registros que se emplearán en la fase de entrenamiento de los modelos. El promedio de incidentes en todas las celdas se sitúa en torno a los 2 incidentes, con una desviación estándar de 1.5 crímenes. Se registran un máximo de 43 crímenes en una sola subdivisión espacio-temporal. Por último, en cada ventana temporal, se observa un promedio de 28 celdas que contienen incidentes, con una desviación de 16 celdas.

5.4. Selección de los datos

Una vez que se han establecido los modelos, las divisiones temporales y espaciales, así como el conjunto de datos destinado para llevar a cabo la predicción de delitos, se procedió a realizar una segmentación en ocho categorías distintas. Estas categorías se han organizado en dos

grupos principales, los cuales se diferencian por la cantidad de elementos que integran cada vector. Estos valores oscilan desde 3, que representa el número de ventanas temporales contenidas en cada vector, hasta 5, que denota el número de ventanas consideradas para cada vector.

Los resultados proporcionados muestran el rendimiento de diferentes modelos en términos de error absoluto y error cuadrático medio. Cada modelo ha sido evaluado utilizando el mismo conjunto de datos, dependiendo del número de ventanas temporales, compuesto por vectores de la forma:

$$X = (x_s, y_s, m^s_{t-3}, m^s_{t-2}, m^s_{t-1}, P_s, A_s) \tag{1}$$

Dónde:

- x_s, y_s Rrepresentan las coordenadas de longitud y latitud.
- $m^s_{t-3}, m^s_{t-2}, m^s_{t-1}$ son las tasas de delitos en las tres últimas ventanas temporales.
- P_s denota el periodo dentro de cada año.
- A_s es el año correspondiente.

Luego de generar los vectores de entrenamiento, se procede a construir un único vector que albergará los valores de los delitos de la ventana temporal siguiente. Este vector es esencial para el

proceso de entrenamiento de los modelos y se compone de la siguiente manera:

$$Y = (m_t^s) \tag{2}$$

5.4.1. Set de datos creados

Para la ejecución de este proyecto, se procedió a crear ocho set de datos distintas que servirían de conjunto de entrenamiento para los modelos. El propósito detrás de esta diversificación radica en evaluar el desempeño y comportamiento de cada modelo en la tarea de predicción de delitos. El tamaño de los vectores en estas bases de datos varía en función de la cantidad de ventanas temporales asignadas, es decir, tres o cinco. Para tres ventanas temporales, los vectores constan de siete columnas, mientras que para cinco, se extienden a nueve columnas. Entre estas bases de datos, dos de ellas contienen valores nulos, mientras que en otras se han eliminado estos valores. La mayoría de los datos ha sido normalizada hasta 43, que representa el máximo valor de delitos en todas las ventanas temporales. No obstante, para algunas categorías, se optó por una normalización hasta 20 con el fin de explorar los resultados que se obtendrían bajo estas condiciones. Todos estos datos se condensan en la tabla 6.

5.4.2. Vectores únicos agrupados

Adicionalmente, como parte del proceso, se generan vectores únicos definidos por la variable "Y". Estos vectores se organizan en cuatro grupos distintos, lo que permite evaluar el desem-

Tabla 6
Base de datos creadas para el entrenamiento de los modelos

Set de datos usado	**Significado**
all_vectors_3	7 columnas normalizadas a 43 sin ceros
all_vectors_3z	7 columnas normalizadas a 43 con ceros
all_vectors_5	9 columnas normalizadas a 43 sin ceros
all_vectors_5z	9 columnas normalizadas a 43 con ceros
all_vectors_N20_3	7 columnas normalizadas a 20
all_vectors_N20_3a_1	7 columnas normalizadas a 20 de 0 a 1
all_vectors_N20_5	9 columnas normalizadas a 20
all_vectors_N20_5a_1	7 columnas normalizadas a 20 de 0 a 1

peño de cada modelo con respecto al conjunto de datos creado. Estos vectores únicos se utilizan en el entrenamiento de los modelos con el objetivo de mejorar la precisión en las predicciones, intentando así superar los intentos anteriores. La agrupación de estos vectores se encuentra detallada en la tabla 7.

Tabla 7
Clases de los vectores únicos agrupados

Cantidad de delitos	**Tipo de clase**
0 a 1	Clase 0, Muy bajo
2 a 5	Clase 1, Bajo
6 a 10	Clase 2, Medio
11 a 15	Clase 3, Alto
16 ó más	Clase 4, Muy alto

Los resultados revelan el rendimiento predictivo de cada modelo en relación con este conjunto de datos específico. Para decidir si este conjunto de datos es adecuado para la predicción o si se debe evaluar de acuerdo con cada modelo, es importante considerar varios factores:

1. Dependiendo de la naturaleza del problema de predicción de delitos y las características

de los modelos, es posible que algunos modelos sean más adecuados para manejar ciertas características de los datos que otros. Es fundamental considerar si las características de los datos se alinean con los supuestos de cada modelo.

2. Cada modelo debe evaluarse por separado utilizando métricas de rendimiento específicas. Esto permitirá comprender cómo se desempeña cada modelo en el contexto de los datos proporcionados.

3. La comparación de los modelos entre sí es esencial para determinar cuál es el más adecuado para su aplicación específica. Puede hacerlo utilizando métricas como el error absoluto y el error cuadrático medio, pero también debe considerar otros factores, como el costo de implementación y la interpretabilidad del modelo.

6. Experimentación

6.1. Primera fase experimental

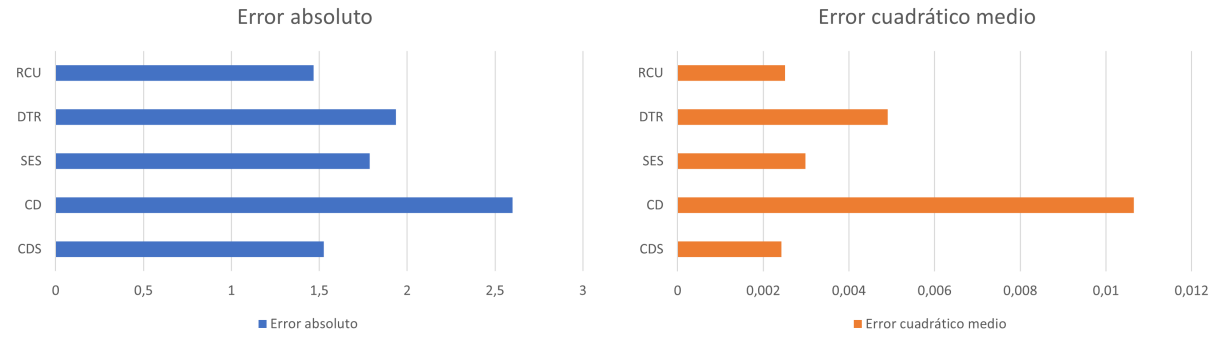
El enfoque adoptado en la primera fase experimental habilitó un análisis meticuloso y comparativo de la evolución temporal de los patrones delictivos. Los resultados obtenidos de este conjunto de datos se exponen en la tabla 8. En ella, se congregan y sintetizan las métricas esenciales utilizadas para evaluar el desempeño de los modelos de predicción de crímenes. Para la elaboración de estos modelos, se emplearon vectores de datos normalizados al número máximo de delitos, eliminando los valores de baja relevancia. Este proceso de normalización y filtrado permitió incrementar la precisión y confiabilidad de las predicciones de los modelos, generando así resultados más robustos y confiables en la predicción de delitos.

Tabla 8
Resultados Fase 1

Modelos	**Error absoluto**	**Error cuadrático medio**
Capas densas sencilla (CPS)	1,5254	0,0024
Capas densas (CD)	2,6003	0,0106
Siames (SES)	1,7870	0,0029
DecisionTreeRegressor (DTR)	1,9363	0,0049
Red Convolutiva Unidimensional (RCU)	1,4673	0,0025

Los resultados revelaron que el modelo de *Capas Densas Sencillas* superó a los demás modelos, demostrando un error absoluto notablemente bajo y un error cuadrático medio mínimo. Esto

sugiere que las predicciones de este modelo están más cerca de los valores reales en comparación con los otros modelos evaluados. Esto se puede ver en los gráficos de la figura 7 dónde se tiene los gráficos de barras del error absoluto 7(a) y del error cuadrático medio 7(b).



(a) Error absoluto (b) Error cuadrático medio

Figura 7. Gráficos de barras para para el error absoluto y el error cuadrático medio de la primera fase experimental.

El modelo de *Red Convolutiva Unidimensional* también mostró un rendimiento sólido, con un error absoluto bajo y un error cuadrático medio reducido, lo que indica una buena precisión en las predicciones.

En el contexto analizado, es esencial destacar que los modelos de *Capas Densas*, *Siames* y el modelo *Decision Tree Regressor* manifestaron un desempeño subóptimo en términos de exactitud predictiva. Estos modelos exhibieron valores elevados tanto en el error absoluto medio como en el error cuadrático medio, evidenciando así una capacidad reducida para ofrecer predicciones precisas y fiables en el ámbito de la proyección de delitos. Dicha limitación los posiciona como alternativas menos idóneas para ser implementadas en aplicaciones de pronóstico de criminalidad que exijan un alto grado de precisión y confiabilidad, particularmente cuando se manejan conjuntos

de datos con características análogas a las del conjunto evaluado en este estudio. Esta observación enfatiza la necesidad de explorar y evaluar otros métodos y técnicas de aprendizaje automático más avanzados y robustos, que puedan superar las deficiencias observadas y proporcionar estimaciones más certeras y confiables para la predicción de delitos en contextos similares.

6.2. Segunda fase experimental

En esta fase experimental, se procede a llevar a cabo un análisis de los resultados obtenidos a través de la implementación de diversos modelos de predicción de crímenes, utilizando conjuntos de datos que abarcan tres ventanas temporales consecutivas incluyendo los valores ceros en los vectores de entrenamiento. Los resultados se encuentran resumidos en la tabla 9.

Tabla 9
Resultados Fase 2

Modelos	**Error absoluto**	**Error cuadrático medio**
Capas densas sencilla	0,4960	0,0002
Capas densas	0,2435	0,0004
Siames	0,3059	0,0002
DecisionTreeRegressor	0,4514	0,0008
Red Convolutiva Unidimensional	0,4444	0,0007

De manera destacable, se observa que el error absoluto para todos los modelos evaluados es notablemente bajo, lo que sugiere una alta precisión en las predicciones realizadas. Este hecho es especialmente evidente en el caso del modelo de *Capas Densas*, que exhibe el error absoluto más bajo de 0,2435. Este valor representa la magnitud promedio de las diferencias entre las predicciones y los valores reales, lo que indica una gran capacidad predictiva.

Además, el error cuadrático medio para todos los modelos también es muy reducido, con valores que oscilan entre 0,0002 y 0,0008. El error cuadrático medio cuantifica el promedio de los cuadrados de las diferencias entre las predicciones y los valores reales, y su baja magnitud en todos los casos refuerza la idea de la alta precisión de los modelos.

Un aspecto importante a considerar es que estos resultados tan alentadores pueden atribuirse, en parte, a la naturaleza de los datos utilizados. Es crucial señalar que el conjunto de datos utilizado en este experimento contenía una gran cantidad de ceros, lo que en términos de la predicción de crímenes es de gran relevancia, ya que, a la hora de predecir se obtienen valores de cero. La presencia de ceros en los datos tiende a reducir el error, ya que predicciones cercanas a cero son inherentemente precisas en este contexto.

6.3. Tercera fase experimental

En esta etapa experimental, se centra en evaluar el rendimiento de diversos modelos de predicción de delitos, estos resultados se pueden observar en la tabla 10. Los datos de entrenamiento abarcaron un período de 5 ventanas temporales, sin embargo, es importante destacar que, los buenos resultados obtenidos en esta tabla fueron relativamente bajos debido a la presencia significativa de ceros en los datos utilizados para entrenar los modelos.

Uno de los hallazgos más destacados es la presencia de errores relativamente bajos en todos los modelos evaluados. El error absoluto se mantuvo en un rango bastante estrecho, con valores que oscilan entre 0,2433 y 0,8804. En particular, el modelo de *Capas Densas* logró el error absoluto

Tabla 10
Resultados Fase 3

Modelos	**Error absoluto**	**Error cuadrático medio**
Capas densas sencilla	0,4157	0,0002
Capas densas	0,2433	0,0004
Siames	0,2882	0,0002
DecisionTreeRegressor	0,4484	0,0008
Red Convolutiva Unidimensional	0,8804	0,0038

más bajo, registrando un valor de 0,2433, lo que sugiere que sus predicciones están más cerca de los valores reales en promedio.

El error cuadrático medio, que representa el promedio de los cuadrados de las diferencias entre las predicciones y los valores reales, también refleja resultados notables en términos de precisión. Esta vez, el modelo *Siames* presenta el error cuadrático medio más bajo de 0,0002, indicando que las diferencias entre las predicciones y los valores reales son cuadrados pequeños en promedio.

Es importante contextualizar estos resultados, la presencia de valores extremadamente bajos en el error absoluto y el error cuadrático medio puede atribuirse, en parte, a la naturaleza de los datos de entrenamiento. Estos datos parecen contener una gran cantidad de ceros, lo que puede influir en la disminución de los valores de error. En otras palabras, la existencia de una proporción significativa de ceros en los datos puede generar predicciones más cercanas a cero, lo que reduce el error absoluto y el error cuadrático medio.

6.4. Cuarta fase experimental

En esta etapa de experimentación, se realiza un análisis exhaustivo de los resultados generados al emplear múltiples modelos de predicción de crímenes. Estos modelos se entrenaron utilizando conjuntos de datos que abarcan cinco ventanas temporales consecutivas, sin embargo, no se incluyeron valores ceros en los vectores de entrenamiento. Los hallazgos de este experimento se presentan de manera resumida en la tabla 11.

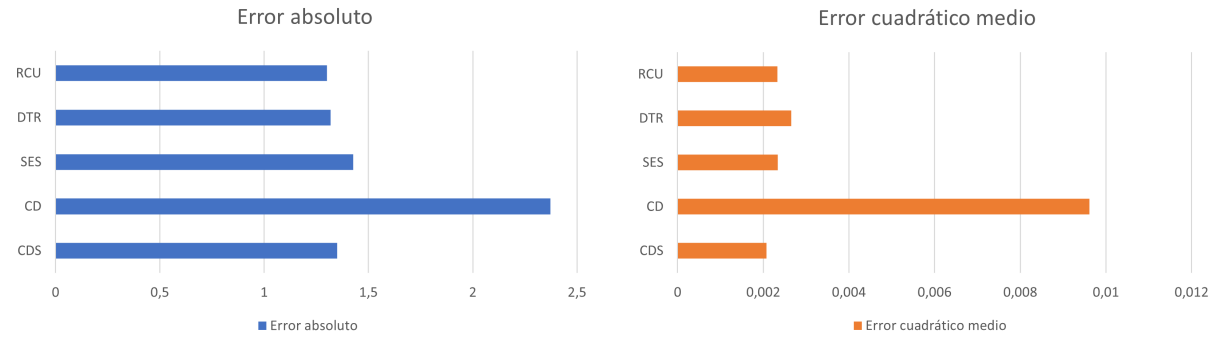
Tabla 11
Resultados Fase 4

Modelos	**Error absoluto**	**Error cuadrático medio**
Capas densas sencilla	1,3500	0,0020
Capas densas	2,3727	0,0096
Siames	1,4270	0,0023
DecisionTreeRegressor	1,3179	0,0026
Red Convolutiva Unidimensional	1,3007	0,0023

Se observa que el error absoluto para todos los modelos es relativamente bajo, lo que indica una capacidad aceptable para realizar predicciones precisas. El modelo de *Capas Densas Sencillas* muestra el error absoluto más bajo con un valor de 1,3500, lo que sugiere que sus predicciones tienden a estar más cerca de los valores reales en promedio.

Por otro lado, el error cuadrático medio también es una métrica relevante para evaluar la precisión de los modelos. Todos los modelos presentan un error cuadrático medio en un rango similar, que oscila entre 0,002073936 y 0,009614019. El error cuadrático medio cuantifica el promedio de los cuadrados de las diferencias entre las predicciones y los valores reales, y la baja magnitud

de estos valores en todos los casos indica una capacidad adecuada de los modelos para predecir crímenes.



(a) Error absoluto (b) Error cuadrático medio

Figura 8. Gráficos de barras para para el error absoluto y el error cuadrático medio para la cuarta fase experimental.

Aunque los resultados son alentadores, es importante destacar que la adición de más ventanas temporales parece tener un efecto en el aumento del error absoluto y cuadrático medio en comparación con los experimentos anteriores, que consideraban conjuntos de tres ventanas temporales. Esto sugiere que, si bien la inclusión de más información temporal puede mejorar la capacidad de predicción en cierta medida, también puede aumentar la complejidad del problema y, en consecuencia, el error.

6.5. Quinta fase experimental

En esta etapa experimental, se empleó la base de datos “all_vectors_N20_5”. En esta base, los vectores unitarios fueron normalizados hasta alcanzar un valor máximo de 20. Esta elección de normalización se realizó con el propósito de investigar las variaciones en las predicciones de

los modelos y determinar si se podían obtener resultados mejorados. Los resultados de este experimento se encuentran detallados en la tabla 12.

Tabla 12
Resultados Fase 5

Modelos	**Error absoluto**	**Error cuadrático medio**
Capas densas sencilla	1,3080	0,0089
Capas densas	1,4360	0,0106
Siames	1,4800	0,0107
DecisionTreeRegressor	1,3192	0,0121
Red Convolutiva Unidimensional	1,3520	0,0107

Los valores de error absoluto y el error cuadrático medio se calcularon para evaluar el desempeño de varios modelos. Como se puede observar, el modelo de *Capas Densas Sencillas* muestra el error absoluto más bajo de todos los modelos evaluados, lo que sugiere una mayor precisión en sus predicciones. Sin embargo, es importante destacar que, dado el cambio en la escala de los datos, el error absoluto y el error cuadrático medio son más altos en comparación con experimentos anteriores.

Los modelos *Capas densas*, *Siames*, *Decision Tree Regressor* y *Red Convolutiva Unidimensional* también se sometieron a estas pruebas. A pesar de mostrar errores absolutos y error cuadrático medio ligeramente más altos que el modelo de *Capas Densas Sencillas*, estos modelos siguen siendo viables para su implementación en la predicción de delitos.

6.6. Sexta fase experimental

En esta fase experimental, se aborda el desafío de normalizar los datos de entrenamiento del vector único hasta un valor máximo de 20, con cinco ventanas temporales al interior de sus datos. Este cambio en la escala de los datos tenía como objetivo explorar cómo afectaría la predicción de delitos en varios modelos. Los resultados de estos experimentos se registraron en la tabla 13.

Tabla 13
Resultados Fase 6

Modelos	**Error absoluto**	**Error cuadrático medio**
Capas densas sencilla	1,3133	0,0089
Capas densas	1,3671	0,0099
Siames	1,3769	0,0094
DecisionTreeRegressor	3,2887	0,0089
Red Convolutiva Unidimensional	2,7134	0,0547

Como se observa, se realizaron pruebas en múltiples modelos. El modelo de *Capas Densas Sencillas* presenta el error absoluto más bajo, lo que sugiere una mayor precisión en sus predicciones en comparación con otros modelos. No obstante, es crucial notar que, debido a la normalización a 20, los valores de error absoluto y error cuadrático medio son más altos en general en comparación con experimentos anteriores con escalas de datos diferentes.

Otros modelos, como *Capas densas*, *Siames*, *Decision Tree Regressor* y *Red Convolutiva Unidimensional*, aunque exhiben errores absolutos y error cuadrático medio ligeramente más altos en comparación con el modelo de *Capas Densas Sencillas*, siguen siendo opciones viables para la predicción de delitos en este contexto específico.

6.7. Séptima fase experimental

En esta fase experimental, se aborda el desafío de normalizar los datos de entrenamiento del vector único hasta un valor máximo de 20, pero con tres ventanas temporales dentro de sus datos. Este cambio en la escala de los datos tenía como objetivo explorar cómo afectaría la predicción de delitos en varios modelos. Los resultados de estos experimentos se registraron en la tabla 14.

Tabla 14
Resultados Fase 7

Modelos	Error absoluto	Error cuadrático medio
Capas densas sencilla	1,5033	0,0109
Capas densas	1,4780	0,0111
Siames	1,4270	0,0023
DecisionTreeRegressor	3,3902	0,0134
Red Convolutiva Unidimensional	3,0150	0,0648

Los resultados revelan una variación significativa en el rendimiento de los modelos con respecto a la normalización de los datos hasta 20. El modelo de *Capas Densas Sencillas* muestra un error absoluto y un error cuadrático medio relativamente bajos, lo que indica un rendimiento más preciso en comparación con otros modelos. Sin embargo, es esencial destacar que, debido a la normalización de datos, todos los valores de error absoluto y error cuadrático medio son más altos en general en comparación con experimentos anteriores con escalas de datos diferentes.

Otros modelos, como *Capas densas*, *Siames*, *Decision Tree Regressor* y *Red Convolutiva Unidimensional*, si bien estos modelos presentan errores absolutos y error cuadrático medio ligeramente más altos en comparación con el modelo de *Capas Densas Sencillas*, siguen siendo

consideraciones viables para la predicción de delitos en este contexto particular.

6.8. Octava fase experimental

En esta fase experimental, se aborda la tarea de normalizar los datos de entrenamiento del vector único a 20 pero esta vez en un rango limitado de 0 a 1. Esta normalización fue aplicada en el contexto de datos que representan 3 ventanas temporales. El objetivo principal era explorar cómo esta restricción de la escala de datos afectaría el rendimiento de diversos modelos de predicción de delitos. Los resultados de estos experimentos se registraron en la tabla 15.

Tabla 15
Resultados Fase 8

Modelos	Error absoluto	Error cuadrático medio
Capas densas sencilla	1,4906	0,0105
Capas densas	1,5977	0,0116
Siames	1,4706	0,0111
Decision Tree Regressor	1,3913	0,0118
Red Convolutiva Unidimensional	1,4250	0,0105

Los resultados revelan que, bajo la restricción de la normalización en el rango 0-1 y con datos que representan un período de 3 ventanas temporales, los modelos de predicción de delitos experimentaron una variación en su rendimiento.

El modelo de *Capas Densas Sencillas* muestra el error absoluto más bajo y el error cuadrático medio más bajo en comparación con otros modelos, lo que sugiere un rendimiento más preciso. Esto indica que este modelo es particularmente efectivo cuando los datos se encuentran en el rango limitado de 0 a 1.

Otros modelos, como *Capas densas*, *Siames*, *Decision Tree Regressor* y *Red Convolutiva Unidimensional*, también fueron evaluados en este contexto. Aunque algunos de estos modelos presentan errores absolutos y error cuadrático medio ligeramente más altos que el modelo de *Capas Densas Sencillas*, siguen siendo opciones viables para la predicción de delitos en este escenario de normalización de datos.

Estos resultados ofrecen una perspectiva valiosa sobre cómo la normalización de datos puede influir en el rendimiento de los modelos de predicción de delitos y proporcionan información útil para futuras investigaciones y aplicaciones prácticas en el campo de la seguridad y la prevención del delito.

6.9. Novena fase experimental

En esta fase experimental, se enfoca en evaluar el rendimiento de un modelo específico, la *Red Convolutiva Unidimensional*, bajo diversas condiciones de entrada de datos. El objetivo era comprender cómo la eliminación de ciertas columnas de datos y la normalización de los valores afectaría la capacidad de predicción del modelo en la tarea de pronosticar delitos. Además, se consideraron dos longitudes diferentes de ventanas temporales: 3 semanas (all_vectors_3 y variantes) y 5 semanas (all_vectors_5 y variantes). Los resultados se presentan en la tabla 16.

Se observa que los resultados varían significativamente según las condiciones de entrada. Es importante destacar que algunos conjuntos de datos, especialmente aquellos con la etiqueta "z" al final (como all_vectors_5z y all_vectors_3z), obtuvieron resultados relativamente buenos en

Tabla 16
Resultados Fase 9

Modelos	**Error absoluto**	**Error cuadrático medio**
all_vectors_3	1,2823	0,0094
all_vectors_3z	1,4211	0,0100
all_vectors_5	1,2469	0,0086
all_vectors_5z	1,4942	0,0124
all_vectors_N20_3	0,4555	0,0006
all_vectors_N20_3a1	1,3406	0,0024
all_vectors_N20_5	0,4618	0,0007
all_vectors_N20_5a1	1,5021	0,0026

términos de error absoluto y error cuadrático medio. Esto se debe a que estos conjuntos de datos contienen una cantidad sustancial de ceros, lo que permite al modelo realizar predicciones precisas al aprender patrones de no ocurrencia de delitos.

Por otro lado, los conjuntos de datos sin la etiqueta “z” (all_vectors_5 y all_vectors_3) también mostraron buen rendimiento, aunque ligeramente inferior en comparación con sus contrapartes “z”. Esto sugiere que incluso cuando se eliminan algunos datos, el modelo sigue siendo capaz de hacer predicciones razonablemente precisas.

Los conjuntos de datos etiquetados como “N20” implican normalización de los valores en un rango de 0 a 20, lo que resultó en un error absoluto y un error cuadrático medio notoriamente bajos. Esto indica que, al normalizar los datos en un rango restringido, el modelo pudo hacer predicciones mucho más precisas.

En contraste, los conjuntos de datos etiquetados como “N20a1” aplicaron una normaliza-

ción similar, pero permitiendo valores más altos, lo que resultó en un aumento significativo en el error absoluto y el error cuadrático medio. Esto sugiere que la normalización en un rango más amplio puede llevar a una pérdida de precisión en las predicciones.

Este experimento demostró la influencia significativa de los datos de entrada en el rendimiento del modelo de predicción de delitos. La presencia de ceros en los datos, así como la normalización de valores, tuvo un impacto sustancial en la capacidad del modelo para realizar predicciones precisas. Estos hallazgos proporcionan información valiosa sobre la configuración óptima de datos para futuros desarrollos en la predicción de delitos utilizando modelos de aprendizaje automático.

6.10. Décima fase experimental

En esta fase de experimentación, se lleva a cabo una evaluación exhaustiva del desempeño del modelo *Capas Densas Sencillas* en la tarea de predicción de delitos. Los resultados de este análisis se encuentran detallados en la tabla 17. Un aspecto destacado de este experimento radica en la variación de los conjuntos de datos utilizados para el entrenamiento del modelo, cada uno de los cuales presenta su propia característica de normalización y la inclusión de valores ceros. Es importante mencionar que se ha modificado el vector unitario, asignando valores en un rango del 0 al 4.

El modelo entrenado con datos normalizados en un rango de 0 a 20 (_n_N20_x) mostró un rendimiento considerablemente mejor en comparación con los datos normalizados en un rango de

Tabla 17
Resultados Fase 10

Dato	Error absoluto	Error cuadrático medio
all_vectors_n_N20_5a1	1,3627	3,2495
all_vectors_n_N20_3a1	0,5709	0,9333
all_vectors_n_N20_5	0,9140	1,9848
all_vectors_n_N20_3	0,6971	1,3344
all_vectors_n_5z	1,5954	3,2536
all_vectors_n_5	1,3545	3,4097
all_vectors_n_3z	1,0983	2,7893
all_vectors_n_3	0,5438	0,8173

43 (_n_xz). Esto se evidencia en el error absoluto y el error cuadrático medio, que son más bajos en los conjuntos de datos "_n_N20_x". Esto sugiere que la normalización en un rango más estrecho puede mejorar la precisión del modelo.

La presencia de ceros en los datos (_n_xz) parece afectar negativamente el rendimiento del modelo. Los conjuntos de datos con ceros tuvieron errores absolutos y error cuadrático medio más altos en comparación con sus contrapartes sin ceros (_n_N20_x).

Entre los conjuntos de datos normalizados en un rango de 0 a 20, aquellos con menos ventanas temporales (_n_N20_3a1) mostraron un rendimiento superior en términos de error absoluto y error cuadrático medio en comparación con los conjuntos de datos con más ventanas temporales (_n_N20_5a1).

Este experimento demuestra que la elección de datos de entrenamiento puede influir en el rendimiento del modelo de *Capas Densas*. La normalización y la presencia de ceros son factores

clave que deben considerarse al configurar conjuntos de datos para la predicción.

6.11. Undécima fase experimental

En esta fase experimental, se llevaron a cabo pruebas con dos modelos de predicción de crímenes: el modelo *Siames* y el modelo de *Capas Densas*. Lo sorprendente de este experimento es que ambos modelos arrojaron resultados idénticos en términos de error absoluto y error cuadrático medio, la cual se evidencia en la tabla 18, a pesar de ser entrenados con diferentes conjuntos de datos como se ha hecho con el resto de modelos, los resultados fueron iguales. Esta casualidad merece un análisis detenido.

Tabla 18
Resultados Fase 11

Dato	**Error absoluto**	**Error cuadrático medio**
all_vectors_n_N20_5a1	0,6171	1,1384
all_vectors_n_N20_3a1	0,6798	1,2576
all_vectors_n_N20_5	0,6171	1,1384
all_vectors_n_N20_3	0,6798	1,2576
all_vectors_n_5z	0,0629	0,0769
all_vectors_n_5	0,6171	1,1384
all_vectors_n_3z	0,0627	0,0769
all_vectors_n_3	0,6798	1,2576

Los resultados, como se puede observar en la tabla, muestran que en todos los conjuntos de datos, ya sea con vectores unitarios normalizados a 20 o a 43, con o sin valores cero, los errores absoluto y error cuadrático medio son consistentemente iguales para ambos modelos. Esto sugiere que estos modelos tienen una tendencia constante en la precisión de sus predicciones, indepen-

dientemente de las variaciones en los datos de entrada.

Una posible razón para esta coincidencia podría ser la similitud inherente en la estructura de ambos modelos. Ambos modelos se basan en redes neuronales profundas y podrían compartir arquitecturas, hiper parámetros o configuraciones similares que los lleven a producir resultados casi idénticos. También podría estar relacionado con la naturaleza de los datos de entrada y cómo estos datos se dividen en clases según la gravedad del delito.

Sin embargo, se requiere un análisis más profundo y pruebas adicionales para comprender completamente por qué estos dos modelos arrojaron resultados consistentemente iguales. Esta coincidencia plantea preguntas interesantes sobre la interacción entre la arquitectura del modelo y la naturaleza de los datos en la tarea de predicción de crímenes, lo que podría ser un tema de investigación futuro.

6.12. Duodécima fase experimental

En esta fase experimental, el enfoque se centra en el modelo *Siames*, que fue sometido a una serie de pruebas utilizando diversas bases de datos. Cabe destacar que los vectores unitarios se organizaron en categorías según sus valores, dependiendo de la gravedad de 0 a 4. Estos resultados se pueden observar en la tabla 19.

Los resultados revelan que, en general, el modelo *Siames* tuvo un desempeño notablemente constante en términos de error absoluto y error cuadrático medio. Los valores de error absoluto

Tabla 19
Resultados Fase 12

Dato	Error absoluto	Error cuadrático medio
all_vectors_n_N20_5a1	0,1181	0,0433
all_vectors_n_N20_3a1	0,1294	0,0479
all_vectors_n_N20_5	0,1186	0,0444
all_vectors_n_N20_3	0,1300	0,9844
all_vectors_n_5z	0,8274	0,0769
all_vectors_n_5	0,0551	0,0096
all_vectors_n_3z	0,2435	0,8370
all_vectors_n_3	2,6003	19,6888

oscilaron entre 0,0551 y 2,6003, mientras que los valores de error cuadrático medio variaron desde 0,0096 hasta 19,6888.

En particular, los conjuntos de datos “all_vectors_n_5” y “all_vectors_n_N20_5a1” destacan por su bajo error absoluto y error cuadrático medio, lo que sugiere que el modelo *Siames* fue especialmente eficaz al trabajar con estos datos. Por otro lado, el conjunto “all_vectors_n_3” mostró los resultados menos favorables, con un error absoluto de 2,6003 y un error cuadrático medio de 19,6888.

Estas variaciones en el rendimiento pueden atribuirse a las diferencias en la naturaleza de los datos de entrada y la estructura de los vectores unitarios. La agrupación de los vectores unitarios en clases según la gravedad del delito parece haber influido en el rendimiento del modelo en cada conjunto de datos.

Este experimento puso de manifiesto la capacidad del modelo *Siames* para adaptarse a di-

ferentes conjuntos de datos y ofrecer resultados coherentes en términos de precisión de predicción. Sin embargo, la influencia de la clasificación de los vectores unitarios en clases también se reflejó en las diferencias de rendimiento observadas.

6.13. Décimo tercera fase experimental

En esta fase experimental, se lleva a cabo el rendimiento del modelo supervisado *Decision Tree Regressor* en la tarea de predicción. La particularidad de este experimento radica en la variación de las bases de datos utilizadas para entrenar el modelo, cada una con su propia configuración de normalización y presencia de ceros, pero agrupando los valores del vector unitario, obteniendo los siguientes resultados que se evidencian en la tabla 20.

Tabla 20
Resultados Fase 13

Dato	**Error absoluto**	**Error cuadrático medio**
all_vectors_n_N20_5a1	6,7725	0,3759
all_vectors_n_N20_3a1	7,7137	0,4331
all_vectors_n_N20_5	6,7725	0,3759
all_vectors_n_N20_3	7,7137	0,4331
all_vectors_n_5z	2,2560	0,1278
all_vectors_n_5	6,7725	0,3759
all_vectors_n_3z	2,3161	0,1302
all_vectors_n_3	7,7137	0,4331

El rendimiento del modelo *Decision Tree Regressor* es más coherente en conjuntos de datos que comparten características similares. Por ejemplo, tanto “all_vectors_N20_5a1” como “all_vectors_N20_5” tienen el mismo rendimiento en términos de error absoluto y error cuadrá-

tico medio, al igual que “all_vectors_N20_3a1” y “all_vectors_n_N20_3”. Esto sugiere que la normalización de datos tiene un impacto limitado en el rendimiento en este contexto.

Los conjuntos de datos “_n_5z” y “_n_3z” se destacan por tener un error absoluto y error cuadrático medio excepcionalmente bajos. Esto se debe a su configuración única de normalización a 43 y la cantidad de ceros en los datos la cual puede afectar la predicción. Estos resultados indican que, en este contexto, la eliminación de ceros y la normalización en un rango amplio pueden contribuir significativamente a la precisión del modelo *Decision Tree Regressor*.

6.14. Décimo cuarta fase experimental

En esta fase experimental,se lleva a cabo una evaluación del rendimiento del modelo de *Red Convolutiva Unidimensional (1D CNN)* en la tarea de predicción. Lo notable de este experimento es que el modelo se entrenó en cada iteración utilizando conjuntos de datos con diferentes configuraciones y agrupaciones de vectores unitarios. Estos resultados se evidencian en la tabla 21. Es importante mencionar que se ha modificado el vector unitario, asignando valores en un rango del 0 al 4.

El rendimiento del modelo *Red Convolutiva Unidimensional (1D CNN)* varía significativamente según el conjunto de datos. Los conjuntos “all_vectors_5z” y “all_vectors_3z”, que contienen datos normalizados a 43 y todos los ceros, destacan por su bajo error absoluto y error cuadrático medio. Esto sugiere que la eliminación de ceros y la normalización en un rango más amplio pueden mejorar la precisión del modelo.

Tabla 21
Resultados Fase 14

Dato	**Error absoluto**	**Error cuadrático medio**
all_vectors_n_N20_5a1	14,1947	1,2973
all_vectors_n_N20_3a1	14,6801	1,2632
all_vectors_n_N20_5	13,6634	1,1979
all_vectors_n_N20_3	14,5290	1,3283
all_vectors_n_5z	10,5738	0,4237
all_vectors_n_5	29,4280	1,3190
all_vectors_n_3z	10,8984	0,4270
all_vectors_n_3	33,3659	1,4959

Por otro lado, los conjuntos “all_vectors_5” y “all_vectors_3” muestran un rendimiento considerablemente peor, con un error absoluto y error cuadrático medio más altos. Estos conjuntos contienen datos normalizados a 43 sin eliminar los ceros. Esto demuestra que la presencia de ceros puede tener un impacto negativo en la precisión del modelo.

En un contexto general, al efectuar una evaluación de este modelo, se observa que sus resultados son los menos favorables en comparación con los experimentos previos. Esta situación podría atribuirse a la naturaleza particular de los datos empleados, especialmente en relación con el vector unitario y las dificultades que podría haber enfrentado en términos de precisión al realizar las predicciones.

7. Conclusiones

Con base a los resultados obtenidos, se logró desarrollar un modelo que exhibe un desempeño óptimo en términos técnicos. No obstante, se observa que los resultados obtenidos no son los idóneos para cumplir con los objetivos establecidos en la fase inicial del proyecto. Esta limitación se atribuye principalmente a la insuficiente cantidad de datos utilizados para el entrenamiento del algoritmo dada la naturaleza sensible de los datos y la imposibilidad de aplicar técnicas de aumento de datos.

Se revela que la disponibilidad de datos históricos sólidos es esencial para construir modelos de aprendizaje automático efectivos. La falta de datos históricos adecuados limita significativamente la capacidad de los modelos para identificar patrones y tendencias relevantes en la ocurrencia de delitos. Por tanto, se tiene la necesidad de mantener bases de datos históricas confiables y actualizadas para respaldar los esfuerzos continuos de predicción de delitos. Uno de los hallazgos más destacados de este proyecto fue la revelación del impacto sustancial que los valores cero en los datos de entrenamiento pueden tener en la precisión de los modelos de predicción de delitos. Este hallazgo resalta la necesidad crítica de comprender y abordar adecuadamente esta paradoja de la .ªusencia de delitos". El análisis reveló que el no tener precaución con estos valores cero pueden generar sesgos significativos en los modelos. Los modelos de aprendizaje automático,

al entrenarse en conjuntos de datos con una presencia desigual de delitos, pueden tender a predecir menos delitos de los que realmente ocurren, lo que podría llevar a una asignación ineficiente de recursos policiales y, en última instancia, a una disminución de la seguridad pública.

Este proyecto enfatiza la necesidad de adaptabilidad y flexibilidad en los modelos de predicción de delitos. La diversidad de conjuntos de datos utilizados, cada uno con sus propias características de normalización y presencia de ceros, subraya la importancia de ajustar los modelos a las particularidades de los datos. La adaptabilidad es esencial para hacer frente a un entorno en constante evolución, donde los datos, las condiciones y el contexto cambian con el tiempo. Los modelos que no pueden adaptarse corren el riesgo de volverse obsoletos y llegar a ofrecer predicciones inexactas. Por lo tanto, la adaptabilidad se convierte en un pilar fundamental en la construcción y el despliegue exitoso de modelos de predicción de delitos efectivos.

Con la información discutida se llega a la conclusión de que modelo de capas densas sencillas se destacó como el mejor en términos de rendimiento predictivo. Este resultado puede explicarse por la simplicidad de los datos utilizados para entrenar los modelos. Se observó que a medida que aumentaba la complejidad de los modelos, tendían a generar más errores en las predicciones. Además, se notó que varios de estos modelos ofrecían resultados cercanos al valor esperado, lo que sugiere que podrían ser adecuados para la tarea de predicción de delitos.

8. Recomendaciones

- Se evidencia la necesidad de un aprendizaje constante y la adaptación continua de los modelos de predicción de delitos en respuesta a un entorno en constante cambio. La dinámica naturaleza de la predicción de delitos exige que los modelos se mantengan actualizados y se ajusten de acuerdo con las tendencias emergentes y las nuevas fuentes de datos. El aprendizaje continuo es crucial para garantizar que los modelos sigan siendo efectivos a medida que evolucionan los patrones delictivos y se introducen nuevas variables en la ecuación. Esto implica una inversión constante en investigación y desarrollo, así como una estrecha colaboración entre expertos en seguridad y científicos de datos. La capacidad de adaptación y aprendizaje continuo se elige como un activo indispensable en la lucha por mejorar la precisión de las predicciones y, en última instancia, contribuir a la prevención y reducción de los delitos.

- Una de las lecciones fundamentales extraídas de este proyecto reside en la diversidad de conjuntos de datos utilizados. Estos conjuntos, cuidadosamente seleccionados, abarcan un amplio espectro de características de normalización y presencia de ceros. La razón detrás de esta elección no fue simplemente académica, sino que subraya la esencia misma de la adaptabilidad requerida en la predicción de delitos. Este enfoque de múltiples conjuntos de datos

condujo a una valiosa conclusión: la adaptabilidad de los modelos es esencial para abordar la heterogeneidad de los datos de delitos. Un modelo que funciona bien en un contexto puede no ser adecuado en otro si no se adapta adecuadamente a las características únicas de los datos.

- Se considera que uno de los principales desafíos que enfrentó en este proyecto fue la calidad y disponibilidad de los datos. Por lo tanto, se enfatiza la importancia de la precisión al definir las fechas con las que se planea trabajar y la creación de su ventana temporal . En este contexto, el proyecto abarcó un período de tiempo que se extiende desde el año 2016 hasta el 2021. Se presume que la disminución en la cantidad de datos en los años 2020-2021 puede estar relacionada con la pandemia global que ocurrió en ese período.

- Se recomienda una ”actualización continua”, esto implica la creación de un sistema dinámico que permita mantener el modelo de predicción de delitos siempre relevante y preciso. Esto puede lograrse mediante la incorporación regular de nuevos datos, la recalibración de los modelos y la reevaluación de las características relevantes a medida que evolucionan los patrones delictivos. Además, se sugiere establecer alertas que indiquen cuando se detecten cambios significativos en las tasas de criminalidad, lo que activaría una revisión inmediata del modelo. Este enfoque adaptativo garantizará que las predicciones sigan siendo útiles y confiables a medida que cambian las circunstancias y permitirá a las autoridades responder eficazmente a las dinámicas cambiantes del crimen.

- Para fortalecer la evaluación del rendimiento del modelo predictivo de delitos, es esencial considerar métodos de validación cruzada robustos. La validación cruzada estratificada, por ejemplo, asegura una distribución proporcional de las categorías de delitos en los conjuntos de entrenamiento y prueba, minimizando así sesgos potenciales en la evaluación. Además, para modelos que se basan en datos de series temporales, es imprescindible emplear la validación cruzada de series temporales, ya que esta considera la correlación temporal de los datos. La aplicación de estas técnicas avanzadas de validación cruzada promete mejorar la precisión en las estimaciones del rendimiento del modelo y, por lo tanto, proporcionar una visión más clara de su aplicabilidad y eficacia en contextos reales.

- Se recomienda ampliar la cantidad de modelos predictivos para el tema tratado en este proyecto, los modelos usados fueron investigados a través del estado del arte, pero se cuenta con otros modelos que puedan mejorar la precisión a la hora de predecir delitos.

Referencias Bibliográficas

(2017). Crime hot spot Forecasting: a recurrent model with spatial and temporal information.

(2020). Plan Integral de Seguridad y Convivencia Ciudadana Para una Bucaramanga Segura 2020-2023.

(2023). delito,RAE.

Biswas, A. A. & Basak, S. (2019). Forecasting the trends and patterns of crime in bangladesh using machine learning model. En *2019 2nd international conference on intelligent communication and computational techniques (ICCT)*, pp. 114–118. IEEE.

Daniel, M. J. V. (2021). Evaluación de modelos de machine learning para la predicción de crímenes en la ciudad de medellín.

José, M. P. J. (2023). Detección automática de primeros arribos usando redes neuronales artificiales en trazas terrestres reales del catálogo sísmico colombiano.

Ordóñez, H., Cobos, C., & Bucheli, V. (2020). Modelo de machine learning para la predicción de las tendencias de hurto en colombia. *Revista Ibérica de Sistemas e Tecnologias de Informação*, (E29):494–506.

Reyes, A. M., Rudas, J., Pulido, C., Victorino, J., Martínez, D., Narváez, L. Á., & Gómez, F. (2020). Characterization of temporal patterns in the occurrence of aggressive behaviors in bogotá (colombia). En *2020 7th International conference on behavioural and social computing (BESC)*, pp. 1–4. IEEE.

Sharma, H. K., Choudhury, T., & Kandwal, A. (2021). Machine learning based analytical approach for geographical analysis and prediction of Boston City crime using geospatial dataset. *GeoJournal*.

Apéndices

A. Códigos de programación implementados

La implementación de lo expuesto durante este trabajo se encuentra alojado en el repositorio de GitLab, en donde se pueden encontrar los códigos de filtrado y conjunto de datos resultantes luego de la eliminación de datos sensibles, junto con una librería desarrollada por los autores junto con la asesoría y orientación del director y codirector del presente proyecto.

El desarrollo de este proyecto fue propuesto usando la versión de Python 3.9.7, bajo un ambiente de Anaconda.

Para acceder al repositorio de este proyecto es necesario la previa autorización por parte de los autores de este proyecto y puede ser clonado posterior a la autorización mediante el `https://gitlab.com/apdd/APDD.git` junto con la herramienta git.