COMPARACIÓN DE UN MODELO DE APRENDIZAJE PROFUNDO FRENTE A UN MÉTODO DE RECOMENDACIÓN PARA PREDICCIÓN DE CRÍMENES EN BUCARAMANGA

JUAN DAVID CORCHUELO MORENO

UNIVERSIDAD JORGE TADEO LOZANO
FACULTAD DE CIENCIAS NATURALES E INGENIERÍA
MAESTRÍA EN INGENIERÍA Y ANALITÍCA DE DATOS
BOGOTA D.C.
2022

COMPARACIÓN DE UN MODELO DE APRENDIZAJE PROFUNDO FRENTE A UN MÉTODO DE RECOMENDACIÓN PARA PREDICCIÓN DE CRÍMENES EN BUCARAMANGA

JUAN DAVID CORCHUELO MORENO

Trabajo de grado presentado como requisito para optar al título de
MÁSTER EN INGENIERÍA Y ANALÍTICA DE DATOS

DIRECTOR
OLMER GARCIA BEDOYA, PhD

UNIVERSIDAD JORGE TADEO LOZANO
FACULTAD DE CIENCIAS NATURALES E INGENIERÍA
MAESTRÍA EN INGENIERÍA Y ANALITÍCA DE DATOS
BOGOTÁ D.C.
2022

RESUMEN

Los modelos de predicción del delito son una herramienta útil para construir estrategias de prevención en las ciudades, en Colombia dados los altos índices de criminalidad que se tienen es importante abordar la exploración de soluciones basadas en las últimas tecnologías y enfoques que han dado resultado en otros países para actuar de manera preventiva frente al delito, ya que las estrategias actuales en nuestro país son más reactivas, y además es importante analizar los resultados obtenidos con datos de una ciudad intermedia que tienen un volumen y calidad de datos menor a los que se tienen para ciudades más grandes como Bogotá o Medellín. La finalidad es explorar dos métodos propuestos recientemente para realizar predicción de crimen, usándolos sobre una ciudad de Colombia y evaluar su rendimiento. Los datos usados son las estadísticas criminales para la ciudad de Bucaramanga de los últimos 10 años, y también se toman datos de Google Maps para contextualizar el tipo de sitios presentes en el área donde se realiza la predicción. Para el desarrollo de los modelos se siguió la metodología CRISP-DM. Uno de los modelos propuestos está basado en el uso de redes neuronales profundas que se implementa usando el framework H2O y el otro en el uso de un método de recomendación basado en factorización de matrices. Los dos modelos mostraron un rendimiento similar, teniendo mejor resultado el modelo basado en la técnica de recomendación. Realizar este estudio aporta un conocimiento importante en cuanto al uso de técnicas distintas para la predicción de crimen en el caso de una ciudad colombiana de tamaño intermedio, y además el modelo de red neuronal es un objeto de java reusable para otras ciudades.

ABSTRACT

Crime prediction models are a useful tool to build prevention strategies in cities. In Colombia, given the high crime rates, it is important to explore solutions based on the latest technologies and approaches that have proven successful in other countries to act preventively against crime, since current strategies in our country are more reactive, and it is also important to analyze the results obtained with data from an intermediate city that has a lower volume and quality of data than those available for bigger cities like Bogotá or Medellín. The purpose is to explore two recently proposed methods for crime prediction, using them on a Colombian city and evaluate their performance. The data used are crime statistics for the city of Bucaramanga for the last 10 years, and data from Google Maps is also taken to contextualize the type of sites present in the area where the prediction is made.

The CRISP-DM method was used to develop the models. One of the proposed models is based on the use of deep neural networks implemented using the H2O framework and the other one is based on the use of a recommendation method based on matrix factorization. The two models showed similar performance, with the recommendation technique model having a better result. This study supplies important knowledge about the use of different techniques for crime prediction in the case of a medium-sized Colombian city, and the neural network model is a reusable java object for other cities.

CONTENIDO

LISTA DE TABLAS

LISTA DE GRÁFICAS

1. INTRODUCCIÓN

Para la ciudad de Bucaramanga el año pasado la Defensoría del pueblo emitió una alerta sobre la presencia de crimen organizado y grupos armados relacionados al narcotráfico [8], Si bien para el año 2020 se redujeron las cifras de delitos debido a las restricciones por COVID, al igual que sucedió en otras grandes ciudades[23], para el 2021 los crímenes en la ciudad se han incrementado. Por ejemplo, el número de homicidios llego a 120 superando la cifra del 2020 y 2019 [25].

Con la reactivación económica, la vuelta de la vida nocturna y la presencia del narcotráfico se espera un incremento en algunos delitos en la ciudad. Además, Bucaramanga está clasificada como la quinta ciudad con mayor criminalidad en el país [7]. Por esto es importante además de las medidas tradicionales para combatir el crimen, hacer uso de herramientas tecnológicas que mejoren la toma de decisiones en cuanto a la disposición de los recursos para proteger a los ciudadanos y prevenir el crimen.

Se ha explorado poco el uso de métodos alternativos para realizar la predicción del crimen en las ciudades, y solo en los últimos años se ha iniciado el estudio y puesta a prueba de modelos basados en aprendizaje profundo.

Sin embargo, para algunos de estos modelos de aprendizaje profundo, en ocasiones la cantidad o calidad de los datos no hace posible que se puedan reusar o transferir estos modelos ya construidos y probados con una ciudad, a ciudades diferentes. Por ello es importante explorar alternativas, basadas en otros enfoques donde la cantidad de información no sea un obstáculo y que entreguen resultados iguales o mejores a los que se obtienen actualmente con enfoques tradicionales.

Para Bucaramanga ahora se dispone de varias fuentes de datos, incluida información espaciotemporal. Pero para la ciudad no hay estudios comparativos de modelos de predicción de crimen basados en *deep learning* ni de modelos alternativos.

Por ello se plantea hacer uso de un método no tradicional para este tipo de problema; cambiando el enfoque y usar una técnica usada en sistemas de recomendación para hacer la predicción de crimen; y comparar su precisión frente a un modelo de aprendizaje profundo para el caso de Bucaramanga.

El presente proyecto tiene como finalidad construir dos modelos de predicción del crimen en Bucaramanga y comparar los resultados obtenidos. La diferencia respecto a estudios o comparaciones anteriores, como el estudio hecho en 2021[7] radica en dos puntos principales, los modelos usados no harán uso de metodologías tradicionales para predicción de crimen (*forecasting crime* en inglés), sino que se usará una técnica de Aprendizaje Profundo y una técnica usada en problemas de sistemas de recomendación.

Realizar este estudio aporta un conocimiento importante en cuanto al uso de técnicas distintas para la predicción de crimen en el caso de Bucaramanga, y además realizar la

comparación de dos técnicas no tradicionales para examinar cual entrega la mejor predicción.

El documento está organizado de la siguiente forma. Inicia presentando lo objetivos del trabajo, luego se resume el marco teórico. A continuación, se expone el estado del arte, resumiendo los estudios relacionados y las ciudades donde se han realizado. Seguido se explica la metodología escogida, para en seguida dar el detalle del desarrollo de la propuesta para cada una de las etapas de la metodología CRISDM. Por último, se resumen los resultados y las conclusiones obtenidas a partir de los mismos.

2. OBJETIVOS

2.1 Objetivo General

Definir entre dos métodos no tradicionales para predicción de crimen cual tiene mejor rendimiento para el caso de la ciudad de Bucaramanga.

2.2 Objetivos específicos

- Preprocesar datos estadísticos de crimen en Bucaramanga para que sean usables en el entrenamiento de modelos predictivos.
- Construir un modelo basado en *deep learning* que realice predicción de crímenes de manera espacio temporal para la ciudad de Bucaramanga
- Construir un modelo basado en un método de recomendación que realice predicción de crímenes de manera espacio temporal para la ciudad de Bucaramanga.
- Evaluar mediante diferentes técnicas, el desempeño de los modelos usando datos de Bucaramanga para los crímenes de más impacto, (homicidio y hurto).

3. MARCO TEORICO

Para el proyecto se hará uso de diferentes conceptos y definiciones las cuales se presentan a continuación

SIEDCO
Siglas del Sistema de Información Estadístico, Delincuencial, Contravencional y Operativo de la Policía Nacional. Es el sistema donde se registra la información de delitos, con el detalle de lugar, tipo y fecha de ocurrencia para todo el país.[28]

Crime Forecasting
El *crime forecasting*, en español la previsión o predicción de delitos es definido por Shan N y Bhagat N como el proceso básico que permite predecir la ocurrencia de un delito [30].

Otro termino asociado es "*Spatial Crime Forecasting*", que es una inferencia sobre el crimen tanto en el tiempo como el espacio [18], es decir, no solo se predice la ocurrencia del crimen, sino que además se ubica espacialmente.

Los primeros usos de este proceso se dieron en 1998, cuando el Instituto Nacional de Justicia de EE. UU. (NIJ por sus siglas en inglés) otorgó cinco subvenciones para estudiar la predicción del crimen para uso policial [12].

Así mismo en 1999 la Oficina Central del Reino publicó uno de los primeros pronósticos de delincuencia y una política de reducción de la delincuencia [12].

Todas estas iniciativas y otros estudios sobre la predicción del crimen han tomado, en su mayoría, como base la literatura sobre teoría criminal de Cohen y Felson, quienes afirmaron que los delitos tienden a ocurrir en intervalos de tiempo predecible y se ubican de manera concentrada en ciertos sitios específicos [6].

Deep Learning
El aprendizaje profundo es un subconjunto del *machine learning*, y se define básicamente como una red neuronal con 3 o más capas. El objetivo de esta red es simular el comportamiento del cerebro humano, permitiéndole aprender de grandes cantidades de datos. Una red neuronal con solo una capa puede hacer predicciones aproximadas, con capas ocultas adicionales puede optimizar y perfeccionar las predicciones [14].

El Deep Learning se caracteriza por hacer uso de una arquitectura de redes neuronales con varias capas ocultas. Se ha convertido en un método poderoso para tareas de clasificación de objetos, reconocimiento de voz y reconocimiento de patrones [1].

Durante el proceso de aprendizaje, un algoritmo de retro propagación (backpropagation) ayuda al modelo a afinar los parámetros de una capa a partir de los parámetros calculados de la capa anterior [1].

Dentro del aprendizaje profundo se identifican 4 grandes arquitecturas:[24]

- Unsupervised pretrained networks UPNs
- Convolutional Neural Network CNNs
- Recurrent Neural Network
- Recursive Neural Network

Funciones de activación
Es aquella función que decide si una neurona debe ser activada o no; normalmente luego de pasar los datos por la capa convolucional, se pasan a través de una función de activación El efecto de usar estas funciones es agregar no linealidad a la CNN. Las más usadas son [31]:

La función sigmoidea

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

La función *ReLu*, o de unidad lineal rectificada

$$f(x) = max(0, x)$$

En el framework usado, H2O, existe una variación de esta última que será la usada en el modelo, *Rectifier with Dropout.* Se selecciona ya que mejora la generalización de la red neuronal y previene el *overfitting*[5]. El *dropout* consiste en establecer aleatoriamente una proporción determinada de nodos en 0 durante la etapa de entrenamiento, es decir, que se asigna una probabilidad de no activación a cada neurona, bien sea en la capa de entrada o en las neuronas de las capas ocultas. La tasa de dropout en H2O se puede especificar para la capa de entrada con el argumento: *input_dropout_ratio*, y para las capas ocultas con el argumento *hidden_dropout_ratios,* en esta última por cada capa se puede variar el valor.

Sistemas de recomendación (RS)

Los sistemas de recomendación son técnicas que proveen sugerencias para ítems que son de mayor interés para un usuario en particular. Las sugerencias se relacionan con procesos de toma de decisiones [26].

El termino ítem es el frecuentemente usado para indicar qué es lo que el sistema recomienda a los usuarios. Los RS tratan de predecir cual es el producto o ítem más apropiado para un usuario basado en las preferencias y restricciones del usuario [26]. El desarrollo de estos sistemas se dio al observar que las personas frecuentemente toman decisiones de su rutina diaria siguiendo recomendaciones de otros, por ejemplo, que película ver, que marca comprar de cierto artículo, etc.

La gran diversidad de productos y servicios que se promocionan hoy por internet hace que los usuarios sean incapaces de explorar toda la oferta por si solos, es allí donde los RS han encontrado un crecimiento y mucha importancia.

Collaborative Filtering.

El RS hace la recomendación basada en la elección de ítems que hicieron otros usuarios con gustos similares en el pasado. La similitud del gusto de dos usuarios es calculada con base en la similitud en el historial de calificaciones de los usuarios. Esta es la técnica más usada en RS [26].

Binary Code o Código binario (BC)

En este método se usa una representación de código binario, semejante a la factorización matricial. El objetivo del método es encontrar dos matrices de factores latentes cuya multiplicación se acerque a la tabla de clasificación. La diferencia radica en que la matriz posee solo valores binarios, no valores reales. Trabajos recientes también han propuesto soluciones que combinan código binario con modelos de aprendizaje profundo [42][19].

4. ESTADO DEL ARTE

Los estudios para predicción de crimen que buscan identificar *hotspots* de manera espacio temporal, basados en técnicas estadísticas y algunas de *machine learning* ha sido ampliamente explorado[3] , pero solo el 15% de los estudios han usado técnicas de aprendizaje profundo [3].

Y solo recientemente se ha enfocado el problema para ser resuelto mediante métodos de recomendación. Una primera exploración es la realizada por Zhang Y, Siriaraya P, et al. [41] en 2019, donde toman data disponible para los 2 principales crímenes en la ciudad de San Francisco y modelando el problema como un sistema de recomendación, usando el método de filtro colaborativo (*collaborative filtering*), logran predecir 70% de los robos de manera más eficiente comparado con los métodos tradicionales [41].

Un año más tarde, los mismos autores continuando el trabajo realizando y definiendo unidades de espacio de 200m x 200m logran mejorar el modelo y alcanzan a predecir un 90% de los robos [42]. Adicionalmente en este estudio exponen como principales problemas de los métodos tradicionales la no predicción del tiempo en algunos (al usar ARIMA, por ejemplo), y la dificultad en obtener buenas predicciones cuando se requiere un nivel de granularidad alto. (al usar modelos KDE, *kernel density estimation*) [42].

En cuanto al uso de aprendizaje profundo para resolver este tipo de problemas está el trabajo de Wang B, Yin P, et al donde emplean un método de regresión de aprendizaje profundo, usando datos de la ciudad de Los Angeles referentes a todos los crímenes sin categorizarlos, además de incluir datos meteorológicos [33]. En este estudio adaptan una red llamada ST-ResNet, que mediante CNN realiza pronósticos espaciotemporales para el tráfico [38], dado los buenos resultados que tuvo este modelo prediciendo el tráfico para ciudades como New York y Beijín.

En [20] desarrollan un modelo de aprendizaje profundo para predecir el crimen en Taiwán, basado en la teoría de "*broken window*" en la cual se afirma que las señales visibles de crimen o desorden social crean entornos propicios para el crimen [37]. Los mismos autores desarrollan una propuesta basada en una grilla y en el uso de información geográfica obtenida de Google Maps, para para predecir el robo de vehículos en la ciudad de Taoyuan[21].

También se encuentra el modelo propuesto por [9] llamado "*Spatiotemporal Crime Network"* (STCN) que se basa en el uso de CNN, y puede predecir el riesgo de crímenes para el día siguiente para regiones definidas dentro de la ciudad. Está además el desarrollo de un *framework* llamado *DeepCrime* construido sobre una red neural profunda [13].

Otra exploración de aprendizaje profundo donde se incluyen varias fuentes de datos para hacer la predicción es la documentada por Kang H [17], asume un enfoque de fuentes de datos multimodal, es decir usa además de las estadísticas criminales, datos meteorológicos, demográficos e imágenes de la ciudad de Chicago. Mediante una red neuronal profunda (DNN) fusiona los datos espaciales, temporales y del entorno para hacer la predicción. En sus resultados indica una precisión 18% mejor que la obtenida con un modelo KDE [17].

Otro enfoque que se ha venido dando más frecuentemente al momento de construir modelos de predicción de crimen, es el uso de datos provenientes de la red social Twitter, aplicándolos como una fuente adicional o complementaria, debido a que estos datos por si solos generan formas potenciales de sesgo que son difíciles de ajustar en los modelos de predicción de crimen [36].

Existe por ejemplo el modelo planteado por Chen X, et al [4], en el cual hace un análisis de sentimientos de los tuits y muestra que mejora la capacidad de predicción del modelo de referencia KDE. Igual al caso expuesto por Wang X, et al [34] que al agregar análisis de los tuits a un modelo espaciotemporal generalizado (STGAM por sus siglas en inglés) obtuvo un mejor resultado que en el modelo original para la ciudad de Charlottesville, Virginia.

En el trabajo [32] Twitter fue usado junto con datos de viajes en taxi para predecir crímenes en la ciudad de New York y el modelo mostro un 19% mejores resultados que al usar únicamente estadísticas criminales y variables demográficas.

Como estudios comparativos a nivel internacional Zhang X, et al. realizaron la comparación de 6 métodos de *machine learning*, 2 de ellos correspondientes a métodos de aprendizaje profundo; tomando datos de una ciudad del sureste de la China. Concluyendo que 1 de los modelos de aprendizaje profundo, el modelo LSTM(*Long Short Term Memory*) fue el que mejor rendimiento tuvo frente a los demás [39].

Para Latinoamérica existe el estudio [10], donde se construyó un modelo de predicción de crimen para la ciudad de Buenos Aires, mediante un tipo de red neural llamado perceptrón multicapa.

En el contexto local en este campo existe el estudio comparativo de los métodos tradicionales para predicción de crimen de Barreras, F et al [2], donde se concluyó que el modelo que mostró mejor precisión para Bogotá fue el método KDE.

Mas recientemente se encuentra el estudio de Riascos, A et al [22] en el cual usan una metodología llamada "*Kernel Warping*" que mejora los resultados de usar un modelo KDE estándar, al incluir en el entrenamiento los datos de los homicidios y enriquecerlos con los datos de las peleas callejeras.

Por último, está el estudio del Departamento Nacional de Planeación donde desarrollaron un modelo de *K-Nearest Neighbors* para la predicción de delitos en Bucaramanga [7].

A continuación, un resumen de la literatura revisada y la comparación con el objetivo del proyecto presentado

Estudio	**Método**	**Fuentes**	**Tipo Crimen**	**Ciudad**
[36]	Regresión Lineal Fixed-Effect y Random-Effect	Estadísticas crimen Censo UK 200 millones tweets	9 categorías, incluyendo robo a personas y tiendas	Londres
[9]	CNN	Estadísticas crímenes	-	New York

[10]	*Neural Network: Multilayer perceptron*	Estadísticas crímenes	Asalto, robo y homicidio	Buenos Aires
[13]	*Recurrent Networks*	Datos ubicuos Estadísticas criminales	Todos	-
[20]	*Deep Learning*	-	Crimen de drogas	Taiwán
[21]	*Deep Learning*	Estadísticas criminales oficiales	Robo de vehículos	Taoyuan, Taiwán
[42]	*Hidden Factor as Topics* *Contextual Biased Matrix Factorization* *Collaborative filtering*	Estadísticas criminales 371 mil tweets	Hurto y Asalto	San Francisco
[33]	Regresión de aprendizaje profundo	Estadísticas criminales Datos meteorológicos	Todos	Los Angeles
[17]	Red neuronal profunda	Estadísticas criminales Datos meteorológicos Imágenes de Google Maps	Todos	Chicago
[34]	*Spatio-temporal generalized additive model* (STGAM) y Análisis de texto	Estadísticas criminales Twitter	Todos	Charlottesville, Virginia
[4]	Análisis de sentimientos y KDE	Estadísticas criminales Twitter	Todos	Chicago
[39]	*Deep Learning* - modelo LSTM CNN SVM	Estadísticas criminales	Todos	Anónima, China
[32]	*Gradient boosting machines.* *Neural Networks*	Datos de viajes de taxis Foursquare app Twitter	Crimen violento y crimen a la propiedad	Nueva York
[2]	KDE ARIMA	Estadísticas criminales	8 tipos, están robo, robo a casa, robo a vehículos.	Bogotá

[22]	KDE - *Kernel Warping*	Estadísticas criminales	Homicidios y registro de peleas callejeras	Bogotá
[7]	K-*Nearest Neighbors* SVM	Estadísticas criminales	Homicidios Hurtos personales Lesiones personales	Bucaramanga
Estudio propuesto	*Contextual Biased Matrix Factorization Deep Learning* usando Perceptron multicapa	Estadísticas criminales	Homicidio y Hurto	Bucaramanga

Tabla 2 Revisión Literatura

5. METODOLOGIA

Para el proyecto se hizo uso de la metodología CRISP-DM, la cual es ampliamente usada en proyectos de analítica y de minería de datos.

Esta metodología propone un ciclo de vida de *data mining* constituido por seis fases, en las cuales la secuencia en la que se siguen las mismas no es estricta, permitiendo devolverse a una etapa anterior o realizar ciclos de repetición [15].

Las fases son los siguientes:

- **Entendimiento de negocio**: Se entiende la necesidad a resolver y se plantea el objetivo del proyecto [15] .
- **Entendimiento de los datos**: Se obtienen los datos, se exploran, se describen y se determina la calidad de estos [15].
- **Preparación de datos**: Una de las más importantes que frecuentemente es la que mayor tiempo requiere, en esta etapa se combinan *data sets*, se seleccionan subconjuntos (definiendo data de pruebas), se crean nuevas variables, se resuelven problemas de datos faltantes y se organiza la data para el modelamiento [15].
- **Modelamiento**: Es donde se selecciona la técnica a usar y se construye el modelo. Es una fase iterativa en la cual se van ajustando los parámetros y se va en busca de la afinación del modelo [15].
- **Evaluación**: Fase donde se realizan las mediciones y evaluación de los resultados obtenidos [15].
- **Despliegue**: Se construye un reporte final, y en ambientes empresariales es común que se planee el despliegue y monitoreo de los modelos [16] para uso continuo.

En el siguiente capitulo se desarrolla la metodología partiendo del entendimiento de los datos, teniendo en cuenta que los capítulos anteriores pueden ser interpretados como el entendimiento del negocio. La fase de despliegue se complementa en este documento a través de la sesión de la conclusión, donde se discute como los modelos encontrados pueden ayudar a los diferentes actores de la ciudad.

6. DESARROLLO DE LA PROPUESTA

Siguiendo las fases de la metodología seleccionada, a continuación, se detallan las tareas y resultados obtenidos en cada una de las etapas.

7.2. Entendimiento de los datos

7.2.1. Obtención de los datos

La ocurrencia de crimines en la ciudad de Bucaramanga, se obtuvo a través del portal de datos abiertos del gobierno nacional de Colombia. Los datos son publicados por la secretaria del interior de Bucaramanga, los datos son extraídos de la base de datos SIEDCO[1] de la Policía Nacional de Colombia. La Figura 1 presenta la página web de datos abiertos la cual permite revisar previo a la descarga el origen de la fuente y fecha de actualización. Adicionalmente, datos.gov permite visualizar los datos , exportarlos o adquirirlos a través de un API.

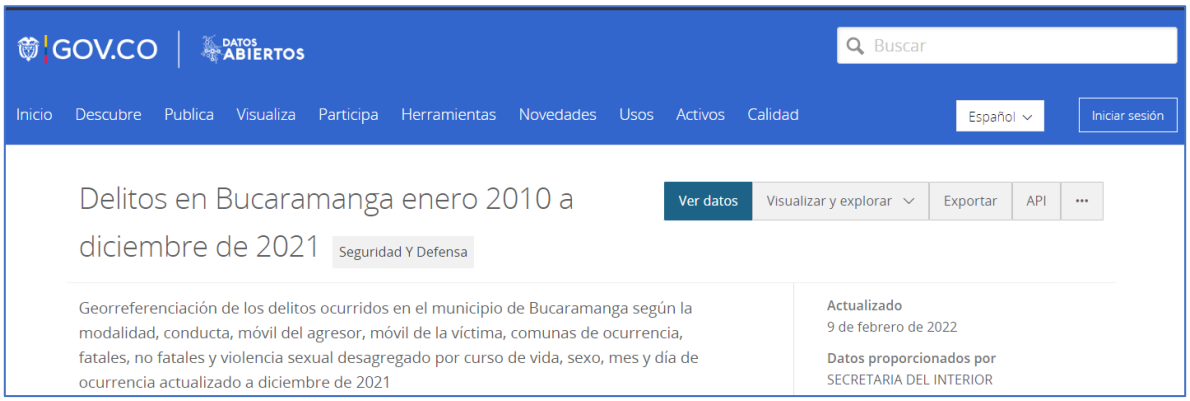


Figura 1. Delitos en Bucaramanga
Fuente: Imagen tomada desde la página web de datos abiertos.

Para los dos modelos se tendrá en cuenta el contexto urbano, tomando este como la presencia de diferentes tipos de sitios en el área, por ejemplo, presencia de estaciones de policía, bancos, hospitales etc. Para obtener este detalle se obtienen los datos de Google

[1] Sistema de Información Estadístico, Delincuencial, Contravencional y Operativo de la Policía

Maps Platform, a través del API Places, mediante el método *NearBy Search*. En la tabla 1, a continuación, el detalle de los parámetros de entrada usados para consumir el API

Parámetro	**Definición**	**Dato enviado**
location	Ubicación a partir de la cual se busca [11]	Se envía latitud y longitud
radius	Distancia en metros dentro de la cual se hace la búsqueda [11]	Se definió en 20000 m
type	Restringe el resultado a los lugares que se especifican [11]	Se definió una lista de 42 tipos de lugares el detalle se indica en la tabla 2.

Tabla 1. Descripción parámetros

Se seleccionaron los sitios tratando de cubrir variedad, es decir, tener en cuenta lugares de tipo económico-comercial, religioso, educativo, de salud y policial. En la tabla 2 se detalla el nombre usado en inglés para realizar la obtención de los lugares a obtener.

atm	florist	liquor_store
bank	funeral_home	local_government_office
cemetery	furniture_store	locksmith
church	gas_station	lodging
clothing_store	gym	meal_delivery
convenience_store	hair_care	meal_takeaway
courthouse	hardware_store	mosque
dentist	home_goods_store	movie_rental
department_store	hospital	movie_theater
doctor	insurance_agency	moving_company
electrician	jewelry_store	museum
electronics_store	laundry	night_club
embassy	lawyer	police
fire_station	library	university

Tabla 2. Lista de sitios a buscar

Para tener mayor precisión del espacio sobre el cual se van a ubicar los datos, se hará uso de un shapefile, el cual tiene la división administrativa de Colombia. El archivo es el que se presenta en la Figura 2 que se obtuvo de la página web *The Humanitarian Data Exchange*[2] que contiene la división administrativa de Colombia a nivel de municipios.

[2] https://data.humdata.org/dataset/cod-ab-col

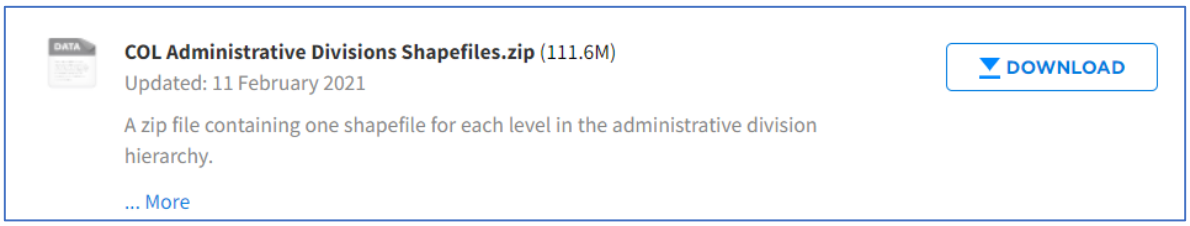


Figura 2. Imagen sitio web para descarga de shapefile
Fuente: https://data.humdata.org/dataset/cod-ab-col

7.2.2. Descripción de los datos

Los datos del SIEDCO que contiene un histórico de los crimines ocurridos desde el año 2010 al 2020 en la ciudad de Bucaramanga, tiene para cada hecho los campos descritos en la Tabla 3:

Año	Año desde 2010 a 2021
Armas_medios	descripción del arma según listado suministrado por la policía nacional
Barrios_hecho	Son los barrios que se generó el delito y están en el municipio de Bucaramanga
Mes	Nombre del mes completo cuando se registró el delito
Dia	día de la ocurrencia del delito
Conducta	es la descripción de la conducta
Descripcion_conducta	es el número del artículo del delito más la descripción de la conducta según la policía nacional
Dia_semana	Nombre en letras del día de la semana
Latitud	es la latitud del punto georreferenciado
Longitud	es la longitud del punto georreferenciado
Zona	son las zonas que se clasifica el barrio que son: OTROS, RURAL, URBANA
Edad	Edad de la víctima en años
Estado civil persona	son los estados civiles de la víctima que reporta la policía nacional y son: CASADO, DIVORCIADO, NO REPORTA, SEPARADO, SOLTERO, UNION LIBRE, VIUDO
Genero	son el género de la víctima y son: MASCULINO, FEMENINO, NO REPORTA

Tabla 3. Descripción campos del dataset [29]

Para los datos obtenidos de la API la respuesta viene en formato JSON, con varios detalles. Sin embargo, para el alcance del estudio solo se capturan de la respuesta los campos indicados en la tabla 4:

Campo	**Descripción**
place_id	Id interno de Google Maps
geometry.location.lat	Valor de latitud del lugar

geometry.location.lng	Valor de longitud del lugar

Tabla 4. Descripción campos obtenidos por API

4.2.3 Exploración de los datos

En la Figura 3 observa que el delito de mayor ocurrencia es el Hurto a personas, seguido de las lesiones personales. Le siguen otros tipos de hurto. La violencia intrafamiliar y los delitos sexuales no se tienen en cuenta en este estudio ya que para estos no está registrada la ubicación exacta por confidencialidad.

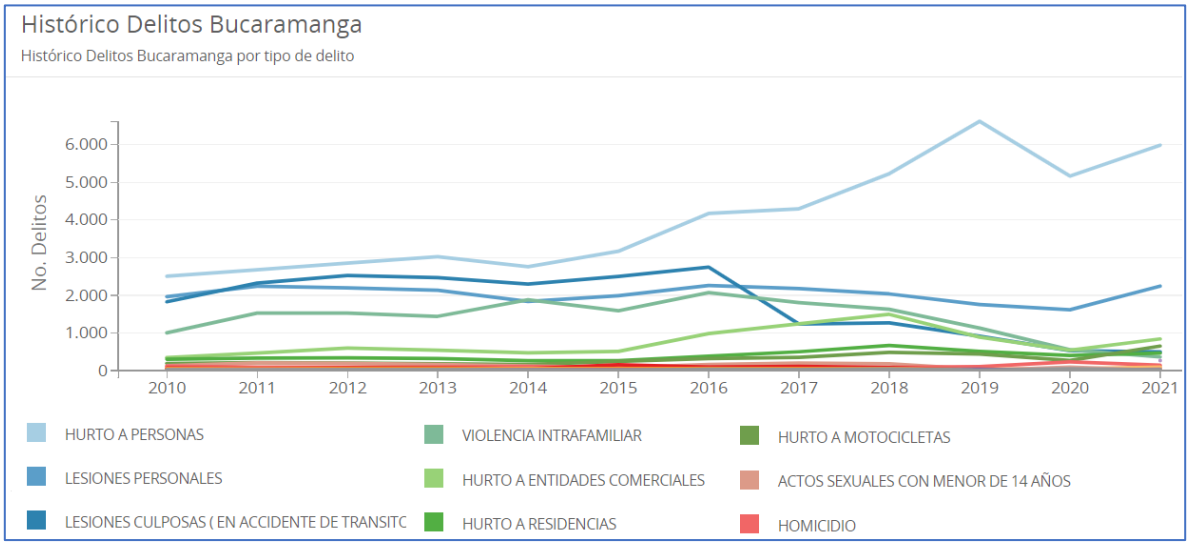


Figura 3. Histórico Delitos en Bucaramanga.
Fuente: Imagen elaborada en la página web de datos abiertos.

Se tendrán en cuenta los delitos de hurto a personas y los homicidios por el alto impacto social que tienen.

Se explora la calidad del conjunto de datos para los datos de Longitud y Latitud, ya que serán los datos principales usados posteriormente en los modelos. Además del dato, o datos que determinan la fecha.

En la Figura 4 se ilustra para la latitud la cantidad de valores existentes por año.

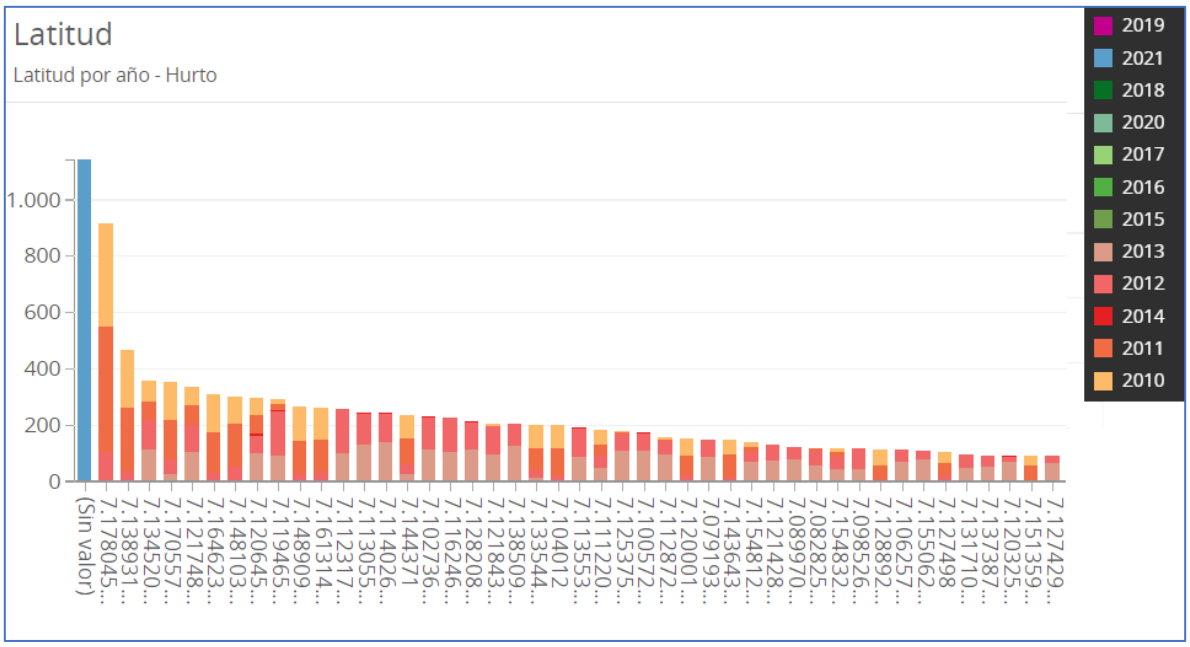


Figura 4. Latitud por año - hurto.
Fuente: Imagen elaborada en la página web de datos abiertos.

Se observa que para el año 2021 hay una gran cantidad de registros sin valor para los delitos de hurto considerados en el estudio. De igual manera se observa en la Figura 5 con los datos de Homicidios

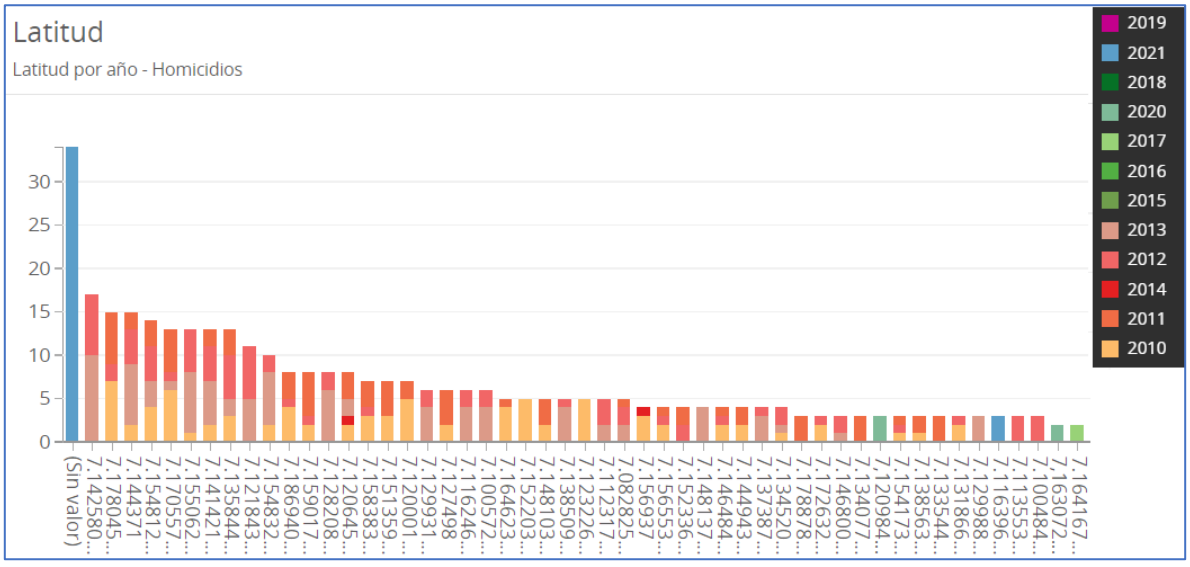


Figura 5. Latitud por año - Homicidios
Fuente: Imagen elaborada en la página web de datos abiertos.

Se realiza la misma exploración para el dato de longitud.

Como puede verse en las Figuras 6 y 7 tanto para hurto como homicidio se obtiene el mismo comportamiento, gran cantidad de registros sin valor.

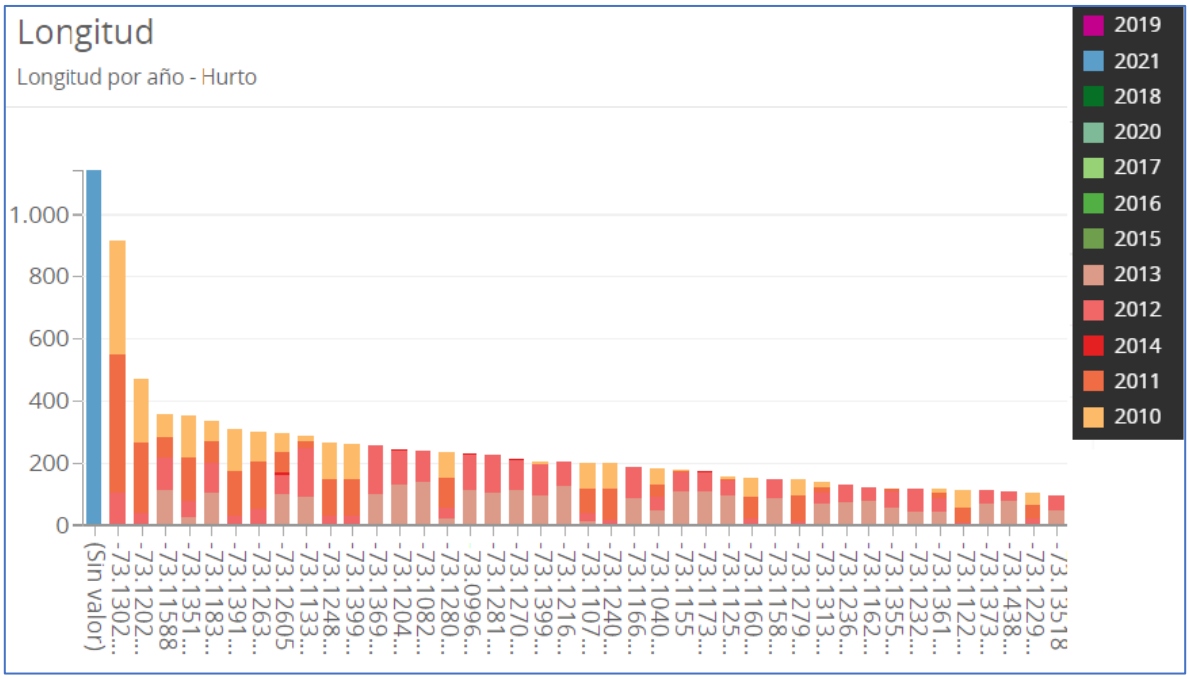


Figura 6. Longitud por año - Hurto
Fuente: Imagen elaborada en la página web de datos abiertos.

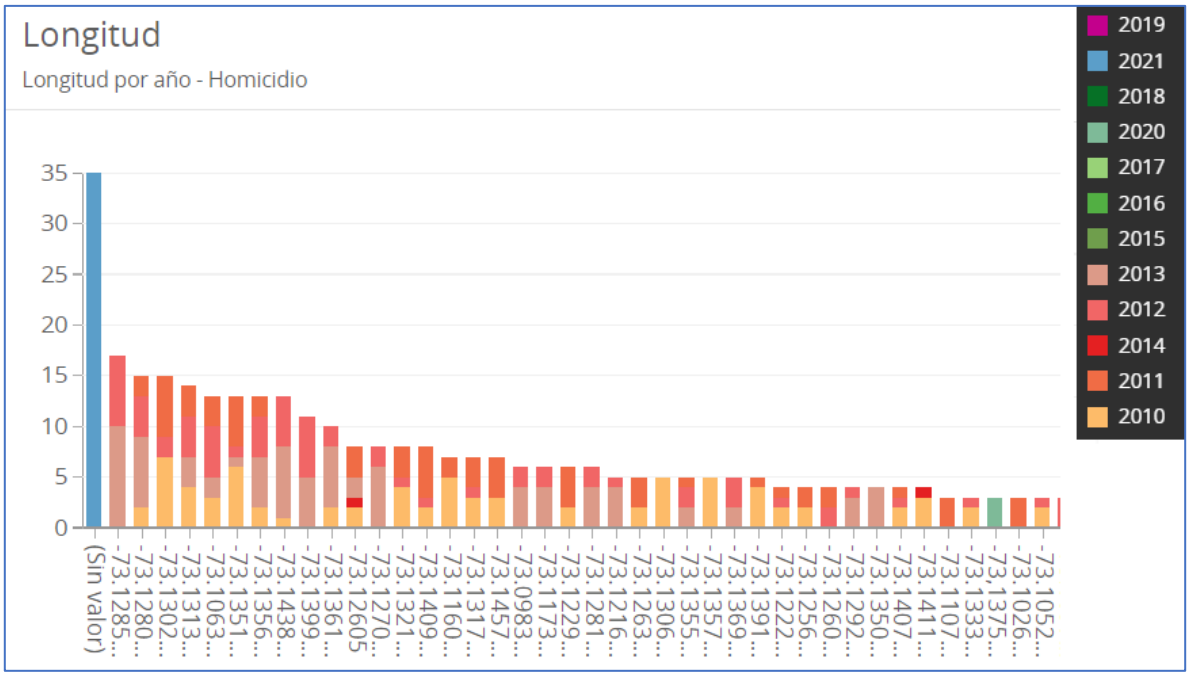


Figura 7. Longitud por año - Homicidios
Fuente: Imagen elaborada en la página web de datos abiertos.

El componente temporal, se encuentra en el conjunto de datos en 3 campos AÑO, MES y DIA.

En la Figura 8 se explora la distribución de estos datos, teniendo en cuenta el resultado de la exploración anterior, se omiten los registros sin valor para la latitud o longitud. Desde 2016 se ve un incremento, y el año con más registros es el 2019.

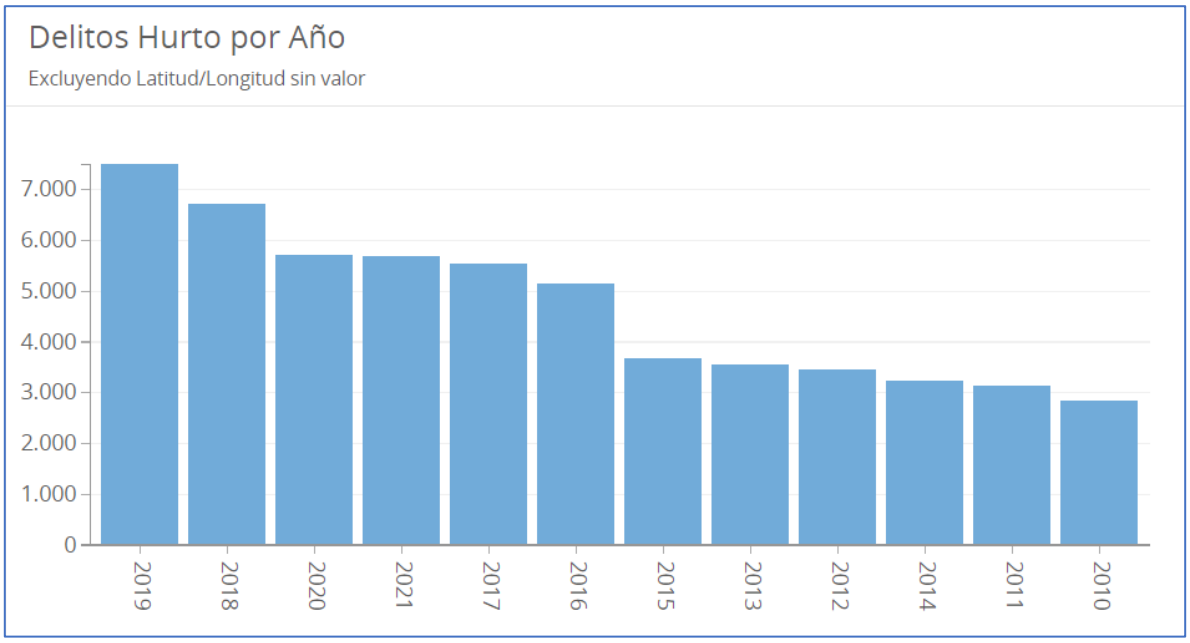


Figura 8. Hurtos por año
Fuente: Imagen elaborada en la página web de datos abiertos.

Para homicidio como se observa en la Figura 9, el año 2020 muestra la mayor cantidad, el resto de los años muestran unas cifras similares.

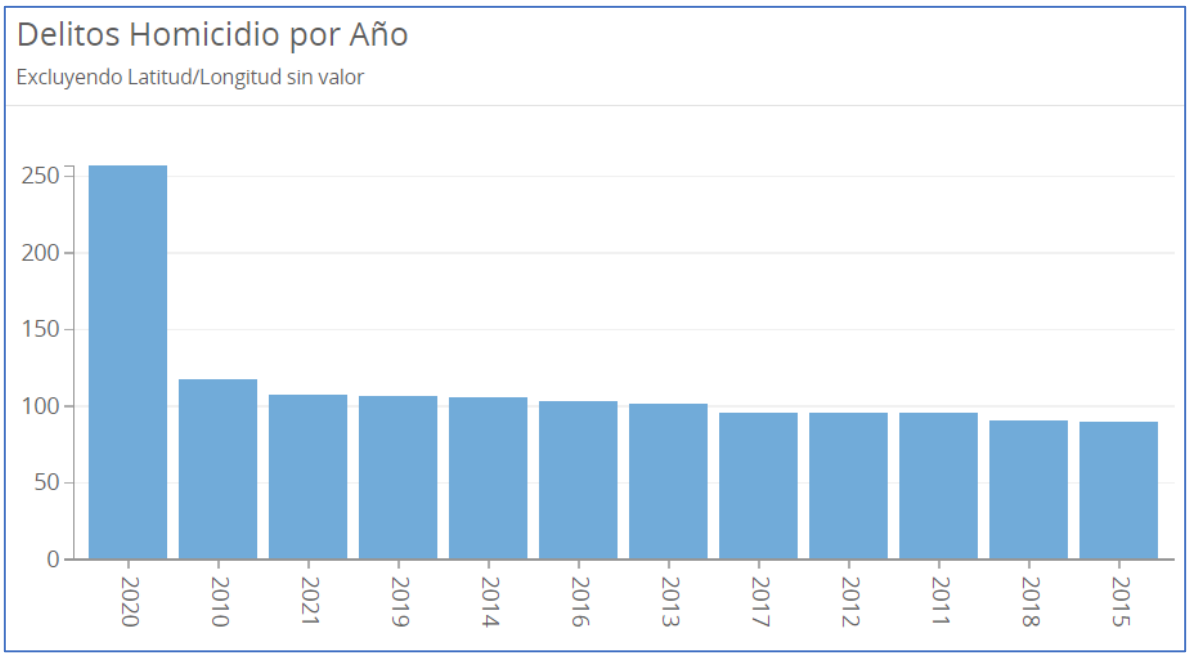


Figura 9. Homicidios por año.
Fuente: Imagen elaborada en la página web de datos abiertos.

Se realiza la exploración por mes

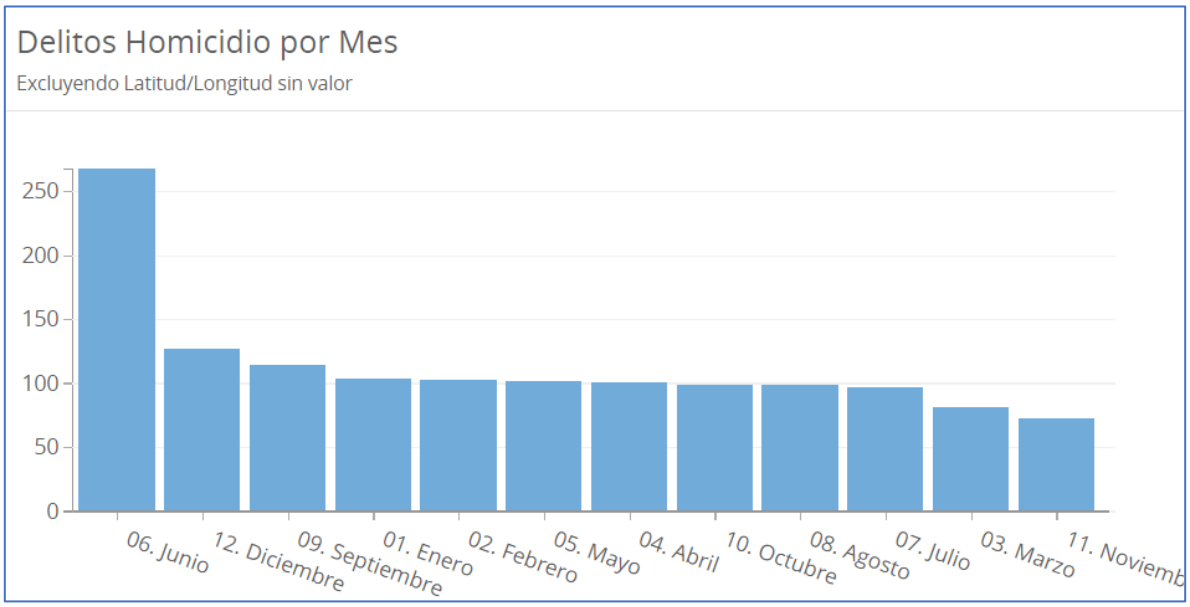


Figura 10. Homicidios por mes.
Fuente: Imagen elaborada en la página web de datos abiertos.

En la Figura 10 se puede observar que, a excepción del mes de junio, los demás meses muestran cifras similares.

En la Figura 11 se ilustra el comportamiento para hurto, se ve que el mes de diciembre muestra una pequeña diferencia a los demás meses.

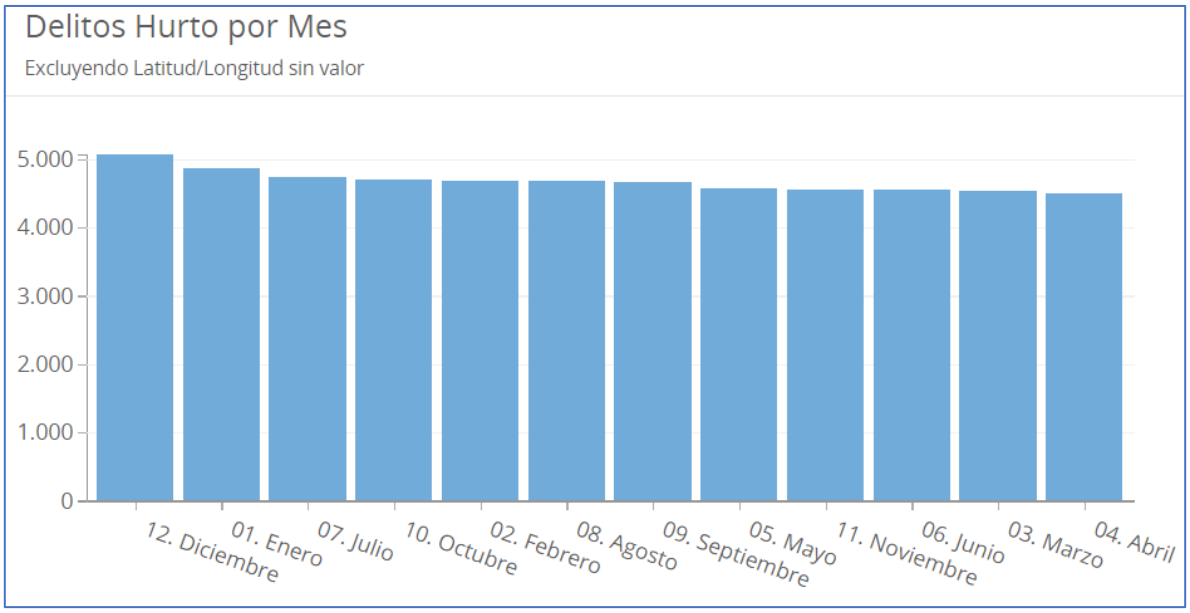


Figura 11. Hurtos por mes
Fuente: Imagen elaborada en la página web de datos abiertos.

Por último, se hace la exploración por día del mes.

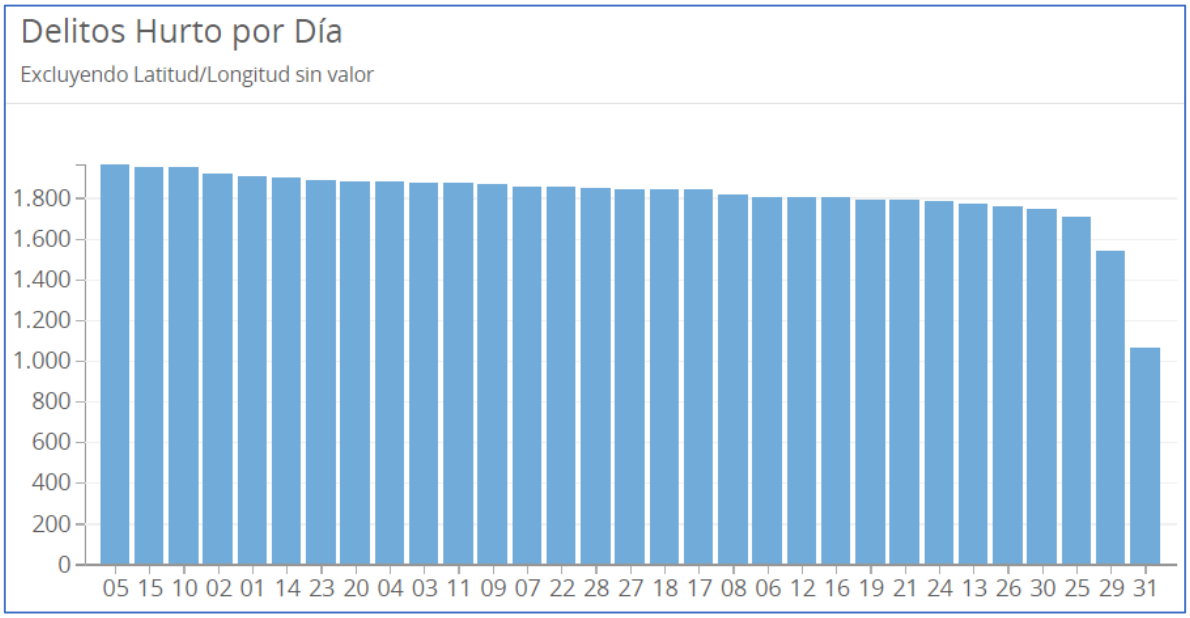


Figura 12. Hurtos por día
Fuente: Imagen elaborada en la página web de datos abiertos.

En la Figura 12 se observa que el día 29 muestra una diferencia respecto a los demás. Como es de esperar el día 31 al no estar presente en todos los meses tiene menos registros.

Para homicidio se ilustra el comportamiento en la Figura 13, el primer día del mes muestra una diferencia, con más de 20 registros que los demás días.

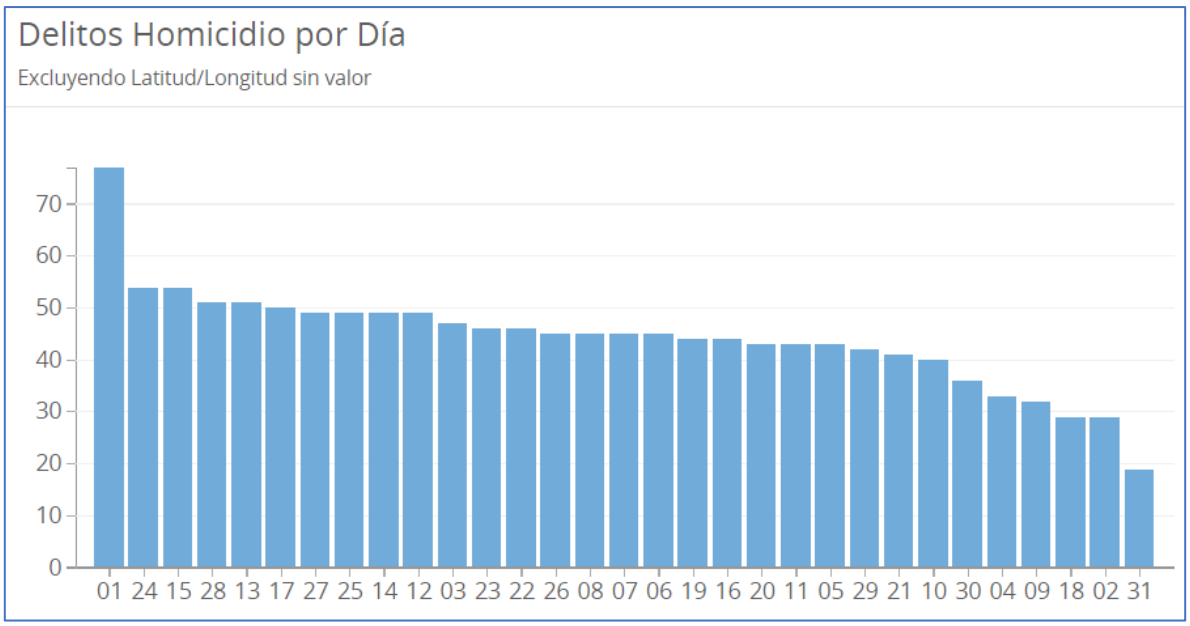


Figura 13. Homicidios por día
Fuente: Imagen elaborada en la página web de datos abiertos

Como resultado de esta exploración se concluye que en el componente espacial para el año 2021 la calidad de los datos no es buena ya que se omitió la información de las coordenadas geográficas, en aproximadamente 16% de los registros.

Para el componente temporal, la calidad es óptima ya que no se encontraron valores fuera de rango, o registros sin valor.

7.3. Preparación de los datos

Para el componente espacial, a pesar de que los campos latitud y longitud tengan un valor, no implica que ese valor sea completamente admitido en el sistema de georreferencia, y aunque el valor sea válido se debe verificar que se encuentra dentro del territorio objeto del estudio, es decir, Bucaramanga.

Se lleva a cabo una actividad cargando los datos en una BD Oracle. Posteriormente se eliminan los que no tienen datos espaciales. Y como se observa en la Figura 14 se valida que no existan valores nulos.

```
--Verificar si hay campos de longitud nulos
select count(*) from ADMIN.DELITOS_BUCARAMANGA where longitud is null or  latitud is null
```

count(*)
0

Figura 14. Verificación de coordenadas en la BD
Fuente: Imagen tomada de entorno Oracle Database

Una vez verificado se agrega un campo de tipo *sdo_geometry* generado a partir de los campos latitud y longitud. Como se muestra en la Figura 15:

```
--GENERAR CAMPO DE UBICACION
alter table ADMIN.DELITOS_BUCARAMANGA add UBICACION sdo_geometry;
update ADMIN.DELITOS_BUCARAMANGA
set UBICACION = sdo_geometry(2001,4326,sdo_point_type(LONGITUD_CORREGIDA,LATITUD_CORREGIDA,null),null,null);
```

Figura 15. Adición de campo de tipo espacial
Fuente: Imagen tomada de entorno Oracle Database

Luego haciendo uso de la herramienta *Oracle Spatial Studio* se visualizan los datos. En las primeras visualizaciones se encontraron puntos fuera de Colombia, los cuales se corrigieron ya que las coordenadas estaban, al contrario. Al final se obtiene la visualización de la Figura 16, se obtienen puntos fuera del perímetro y varios que se encuentran en los límites; realizar una validación visual y manual de estos lleva mucho tiempo y está sujeto a errores. Por lo cual se continua el filtrado de los datos usando R, haciendo un filtro por la latitud y longitud máximas para la ciudad de Bucaramanga.

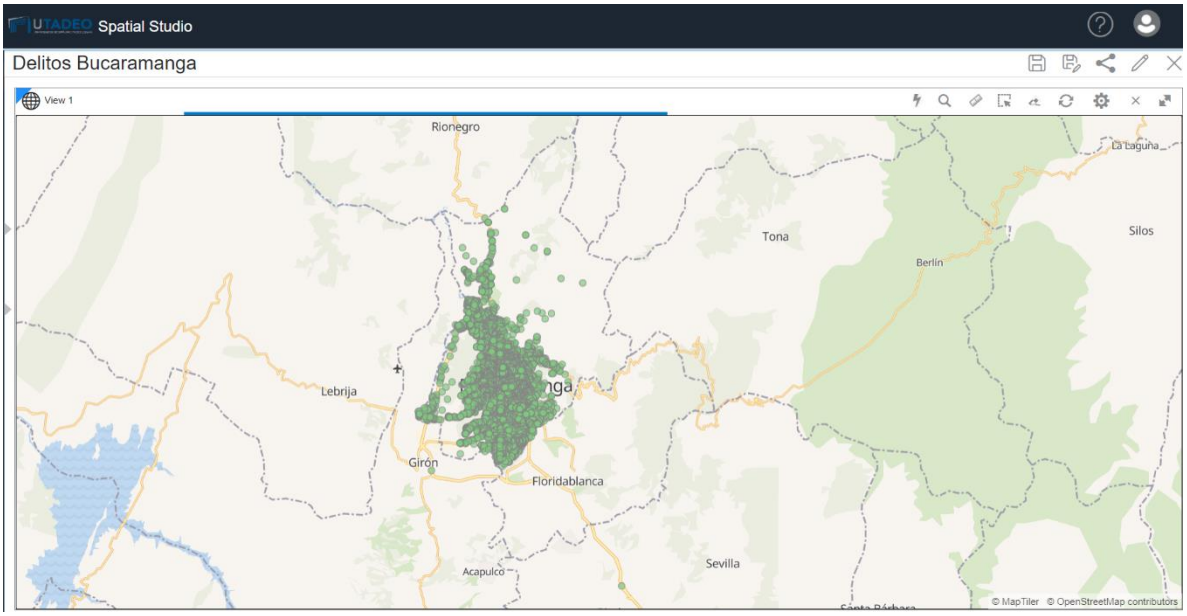


Figura 16. Visualización geográfica de delitos
Fuente: Imagen elaborada en Oracle Spatial Studio.

Dado que se usará un modelo basado en una grilla, a los datos obtenidos de Google Maps se les hace un procesamiento para asignarle la latitud y longitud centroide de los polígonos que crean la grilla, de tal forma que sean fáciles de cruzar en la etapa de modelamiento.

Para el componente de tiempo se agrega un dato nuevo concatenando los campos de año y mes.

Por último, se filtran los crímenes que se tienen dentro del alcance y que están tipificados en el conjunto de datos como HOMICIDIO, FEMINICIDIO, HURTO A ENTIDADES COMERCIALES y HURTO A PERSONAS.

7.4. Modelamiento

7.4.1. Modelo Deep Learning

Para el modelo de *Deep Learning* se seleccionó el enfoque descrito en el trabajo de Lin, Y at al. [21] en el cual se usa una grilla para dividir el espacio donde se hará la predicción de crimen.

En R mediante el uso de librerías para proceso de datos geoespaciales se carga el *shapefile*. Luego se define una grilla de 100x100 que es la que se usaran para ubicar los datos y asignarles un ID de ubicación. Al final se cruza la grilla con el mapa para obtener la división, como se observa en la Figura 17 a continuación:

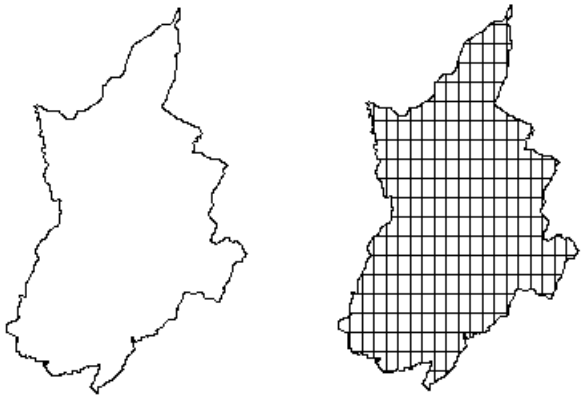


Figura 17. Visualización mapa y grilla superpuesta.
Fuente: Imagen elaborada en R.

Luego se agrega a cada uno de los cuadros de la grilla, la frecuencia de los crímenes que ocurrieron dentro de esa área. Además, se agrega la información obtenida de Google Maps sobre los sitios que se ubican en esa área, que como se mencionó anteriormente tiene las coordenadas propias y coordenadas del centroide en el que se encuentra. Todo este cruce se realiza con base a las coordenadas geográficas, ya que la grilla se maneja como un polígono georreferenciado. Luego para optimizar el trabajo computacional y evitar

desbalances se eliminan de la grilla los espacios que quedaron vacíos,[21] que seguramente son regiones no pobladas o montañosas.

Se obtiene la visualización de la Figura 18, donde se observa que la grilla toma una forma similar a la visualización de los datos inicial.

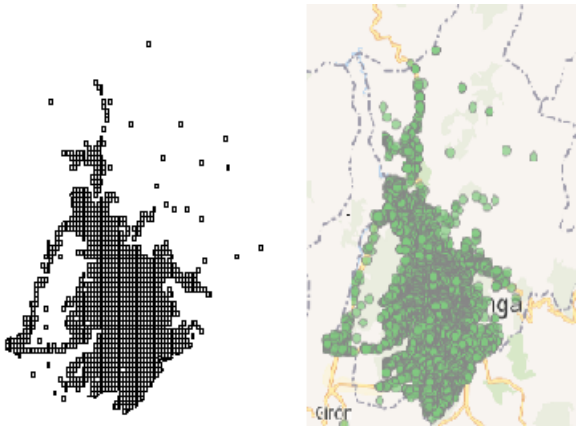


Figura 18. Visualización grilla vs carga inicial
Fuente: Elaboración propia.

Por cada cuadro en la grilla se tienen en este punto las variables de tiempo (en meses), la cantidad de crímenes y los datos de los sitios que existen en el área. En el modelo que se está siguiendo, agregan otras variables para tener en cuenta el cambio en la dinámica de ocurrencia de los crímenes[21]. Las primeras asociaran el número de crímenes que se presentaron en la ubicación, en meses anteriores, para tener variables del comportamiento histórico. Las otras variables nombradas como *near_x* tendrán en cuenta los crímenes en los cuadros vecinos en diferentes tiempos (se suman la cantidad de crímenes de los vecinos para los meses 1,3,6,9 y 12 anteriores) como se observa en la Figura 19, ya que se sabe que los criminales tienden a desplazarse a ubicaciones cercanas según la teoría de "*Broken Window"* [37].

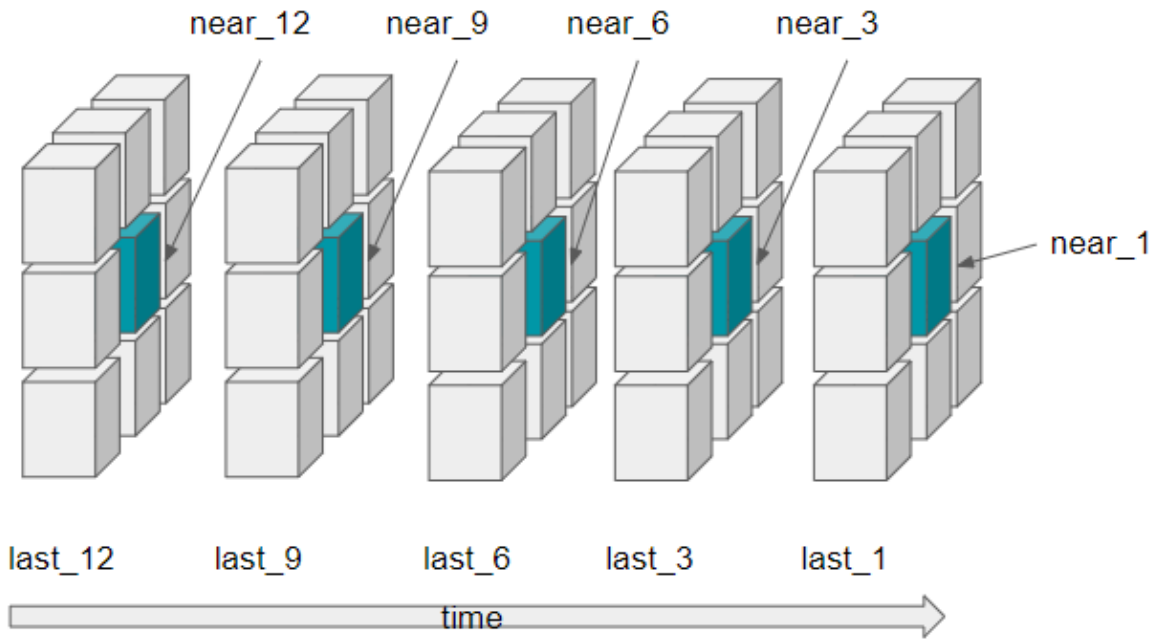


Figura 19. Detalle de las variables basadas en el tiempo y vecinos que se agregan a cada espacio de la grilla
Fuente: Elaboración propia

Con los datos preparados, se procede a dividirlos en datos de entrenamiento, y en datos de prueba. Dado que las variables tienen en cuenta los meses anteriores, se hace la división teniendo en cuenta ese orden, además de lo observado en la exploración de datos. Se toma como data de entrenamiento los meses desde el 2011 al 2018. Se toma desde el 2011 para poder calcular las variables de crímenes del año anterior tomando los datos del año 2010. Y se dejan para validación los primeros 8 meses del 2019.

Con los datos definidos se crea un modelo de *Deep Learning* en la plataforma H2O, la arquitectura de esta red neuronal es *Feed Forward*, supervisada, o en otras palabras un perceptrón multicapa. Donde se configura como variable a predecir la variable *hotspot*, la cual es una variable binaria que indica si se presentará un crimen en esa ubicación.

Como variables independientes se tienen las definidas en modelamiento descrito anteriormente, en total son 54, en la Figura 20 se listan la manera en la que quedaron definidas para el modelo.

```
> dependent
[1] "hotspot"
> independent
 [1] "last_1"              "last_3"              "last_6"
 [4] "last_9"              "last_12"             "before_12"
 [7] "near_1"              "near_3"              "near_6"
[10] "near_9"              "near_12"             "near_b12"
[13] "atm"                 "bank"                "cemetery"
[16] "church"              "clothing_store"      "convenience_store"
[19] "courthouse"          "dentist"             "department_store"
[22] "doctor"              "electrician"         "electronics_store"
[25] "embassy"             "fire_station"        "florist"
[28] "funeral_home"        "furniture_store"     "gas_station"
[31] "gym"                 "hair_care"           "hardware_store"
[34] "home_goods_store"    "hospital"            "insurance_agency"
[37] "jewelry_store"       "laundry"             "lawyer"
[40] "library"             "liquor_store"        "local_government_office"
[43] "locksmith"           "lodging"             "meal_delivery"
[46] "meal_takeaway"       "mosque"              "movie_rental"
[49] "movie_theater"       "moving_company"      "museum"
[52] "night_club"          "police"              "university"
```

Figura 20. Detalle de las variables que se agregan a cada espacio de la grilla
Fuente: Elaboración propia obtenida en R.

La arquitectura de la red neuronal se muestra en la Figura 21, se usan 7 capas ocultas, con 100 nodos cada una, como input están las 54 variables indicadas anteriormente y la salida esperada es el valor de la variable *hotspot*.

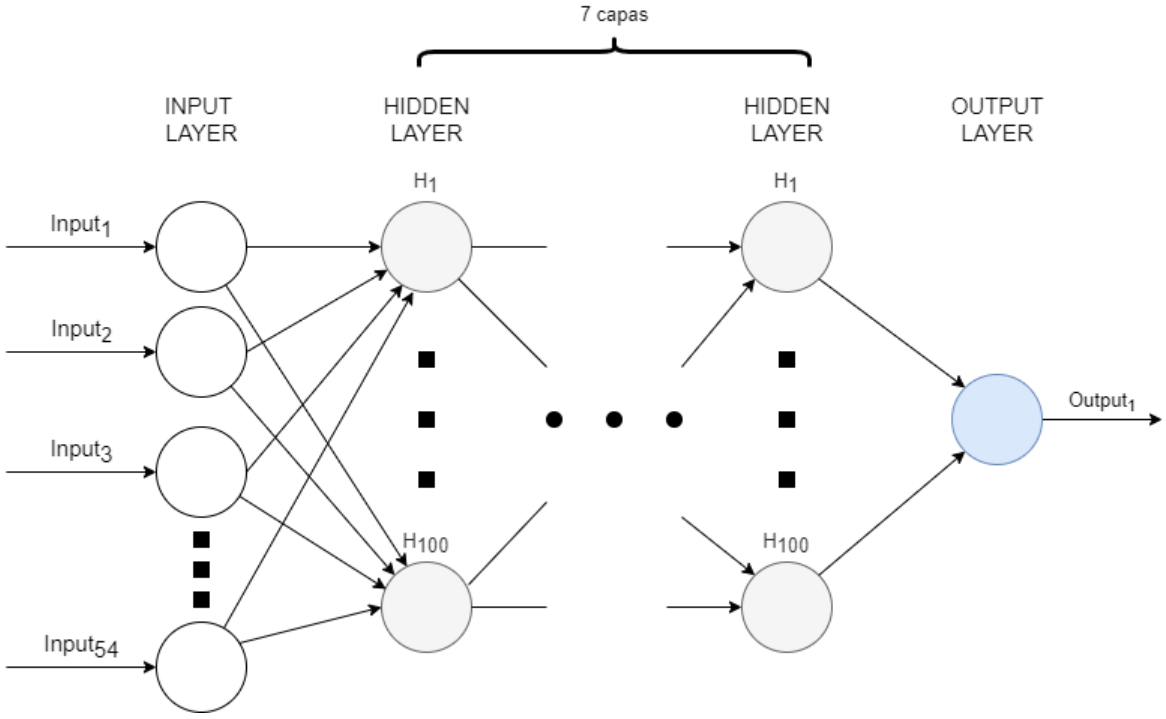


Figura 21. Arquitectura de la red neuronal.
Fuente: Elaboración propia

Como función de activación se selecciona *RectifierWithDropout* ya que mejora el rendimiento de grandes redes neuronales[35] además que en H20 permite parametrizar dos valores de esta función permitiendo experimentar el comportamiento al variar estos parámetros.

Para el parámetro *input_dropout_ratio,* en el cual se define la probabilidad de que una neurona suprima su activación en la capa de entrada, se usó el valor sugerido de 0.2. Para el parámetro *hidden_dropout_ratios*, en el cual se define la probabilidad de que una neurona suprima su activación en las capas ocultas, se usó el mismo valor 0.2 para todas las capas ocultas. Esta combinación de valores fue la que mejor rendimiento mostró.

Por último, se determinaron 50 *epochs* después de realizar pruebas con 20 y 40.

7.4.2. Modelo con método de Sistema de Recomendación

Para el modelo que usa una técnica de sistemas de recomendación, se tomó la propuesta de Zhang, Y at al[42] en la cual se plantea el problema de predicción de crimen como un problema de recomendación, y por ende se soluciona con una de las técnicas usadas para ese tipo de problemas.

Como se observa en la figura en el problema clásico de recomendación, se busca conocer la calificación que obtendría una película por parte de un usuario, con base en las calificaciones ya obtenidas dadas por otros usuarios. En el planteamiento de Zhang, Y at al[42] se homologa el problema, pero para calcular el número de crímenes que se presentan en un espacio de tiempo para un lugar específico, basado en los valores ya conocidos.

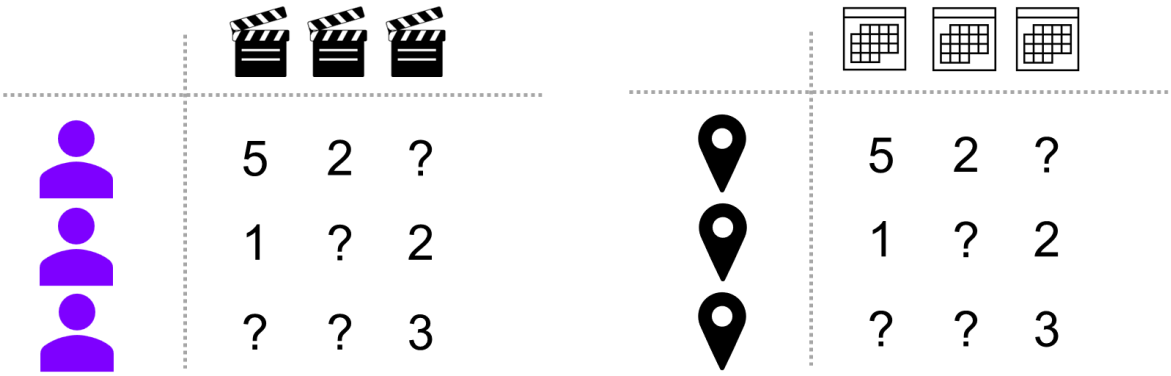


Figura 22 Problema de recomendación clasico y propuesta parapredicción de crimen
Fuente: Adaptado de https://docplayer.net/228062535-Time-and-location-recommendation-for-crime-prevention-yihong-zhang-panote-siriaraya-yukiko-kawai-adam-jatowt.html

La primera definición importante es identificar los dos factores que siempre se tienen en un problema de recomendación, los usuarios y los ítems. Los autores de la propuesta consideran el tiempo y la ubicación como los factores homologables [42] y adicionalmente argumentan que como en los problemas de recomendación por lo general el número de

usuarios es mayor que el número de ítems, para el problema de predicción de crimen se puede tomar la ubicación como el usuario, y el tiempo como el ítem [42]. Esto coincide con el modelamiento que se realizó con los datos de Bucaramanga, ya que se tienen más de 5200 ubicaciones mientras que el tiempo tomado en meses presenta una cantidad menor; por esto se usará la misma definición propuesta.

El segundo punto es la escogencia del método de resolución, dentro del *collaborative filtering*, existen dos enfoques, uno basado en métodos sobre los vecinos donde se procesan las relaciones entre usuarios o entre ítems, y otro donde se procesan las relaciones entre usuarios e ítems a la vez, llamado '*Latent factor method'*. En este ultimo los modelos se basan en la factorización de matrices[27].

Para la predicción de crimen se escoge este enfoque ya que precisamente se quiere procesar la relación entre los dos factores definidos, la ubicación y el tiempo. Además, en el estudio propuesto argumentan que la factorización de matrices ideal para el problema es la factorización de matrices con sesgos, o BMF, por sus siglas en ingles *biased matrix factorization* ya que en la realidad algunas áreas presentan mayor ocurrencia de crímenes que otras, y ese sesgo o *bias* puede ser modelado con este método[42].

Teniendo

U = usuarios

I = ítems

R = rating o calificación que dio el usuario al ítem

F = características latentes

Lo que se quiere con la factorización de matrices es encontrar dos matrices, P de tamaño $|U| \times K$ Y Q de tamaño $|D| \times K$ tal que

$$P \times \mathrm{Q}^{\mathrm{T}} \approx R \quad [42]$$

Luego se plantea, la siguiente ecuación para la predicción de cada par usuario-ítem

$$\hat{\mathrm{r}}_{ij} = p_i^T q_j = \sum_{k=1}^{K} p_{ik}\, q_{kj} \quad [42]$$

Al agregar los sesgos o *bias* a la predicción la ecuación queda de la siguiente forma

$$\hat{\mathrm{r}}_{ij} = bu_i + bd_j \sum_{k=1}^{K} p_{ik}\, q_{kj} \quad [42]$$

Quedan 3 componentes, el sesgo del usuario, el sesgo del ítem y la relación usuario-ítem. El modelo se entrena para reducir el error cuadrado.

Por último, en el método propuesto, llamado ***contextually-biased matrix factorization*** (CBMF)[42] se agregan variables para la información de contexto

$$\hat{r}_{ij} = cbl_i + cbt_j + bu_i + bd_j \sum_{k=1}^{K} p_{ik}\, q_{kj} [42]$$

Donde cbl_i y cbt_j representan la información de contexto de la ubicación i y el tiempo j respectivamente; y se calculan mediante una regresión lineal

$$cbl_i = \Theta X_i$$

Donde X_i son las características de contexto usadas en este estudio, es decir, los sitios presentes en el área.

Para este modelo la data de entrenamiento y de prueba se definió de la misma forma que en el modelo de Deep learning, para entrenamiento los meses del 2010 al 2018. Y se deja para validación los meses del 2019.

Se uso la implementación original en R publicada por el profesor Yiong Zhang [40].

Dado que el método original retorna una probabilidad continua, se agrega un paso para convertir el resultado a binario y poder comparar el resultado con la red neuronal que hace clasificación binomial. Se toman las probabilidades con valor igual o mayor a 1 como 1, y las que sean menores se toman como cero o no ocurrencia de crimen.

7.5. Evaluación

7.5.1. Resultados

El modelo de *Deep Learning* obtuvo un AUC de 0.8078 y alcanzo un RMSE de 0.3511

Se hicieron pruebas con los meses de prueba y se obtuvieron las siguientes medidas de rendimiento

PPV	TPR	NPV	TNR	FPR	FNR	Accuracy	F1
0.533181	0.620795	0.899787	0.805344	0.100213	0.466819	0.794909	0.590116
0.526185	0.655488	0.879877	0.818529	0.120123	0.473815	0.773091	0.579515
0.432927	0.710438	0.90487	0.741187	0.09513	0.567073	0.734545	0.536213
0.128641	0.779412	0.984424	0.725325	0.015576	0.871359	0.709091	0.209486
0.597059	0.616477	0.856039	0.86608	0.143961	0.402941	0.785455	0.595336
0.580925	0.649123	0.862974	0.859632	0.137026	0.419075	0.789818	0.60573
0.555035	0.618919	0.859705	0.810945	0.140295	0.444965	0.776	0.597911
0.472477	0.687919	0.902023	0.786444	0.097977	0.527523	0.763636	0.557823

Tabla 5. Métricas de rendimiento modelo deep learning

Para el modelo con la factorización de matrices, se obtuvo un AUC de 0.6304 y un RMSE de 0.71, haciendo la ejecución con 108 iteraciones.

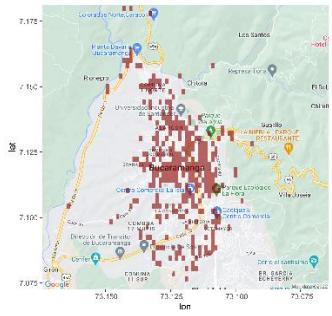
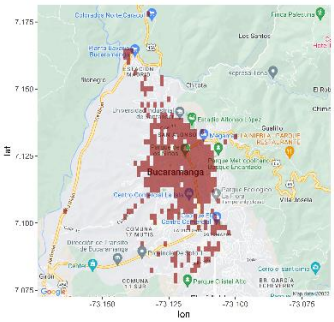
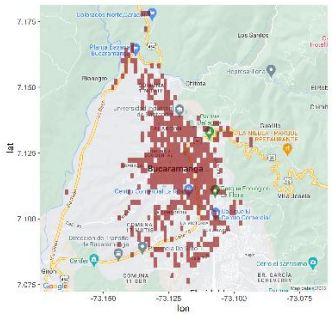
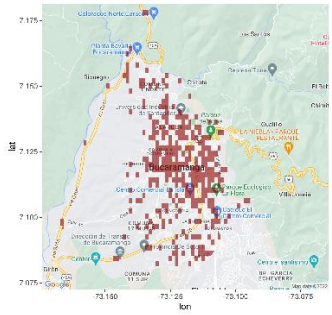
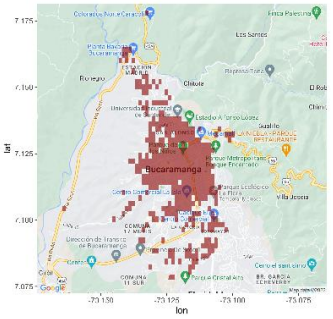
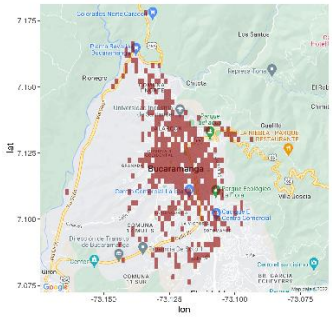
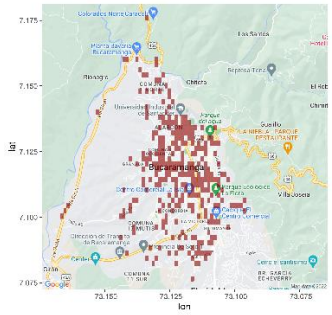
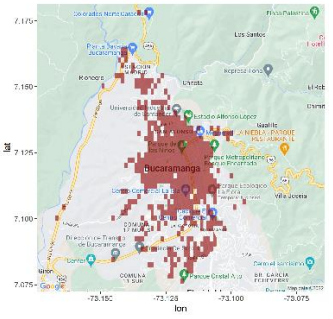
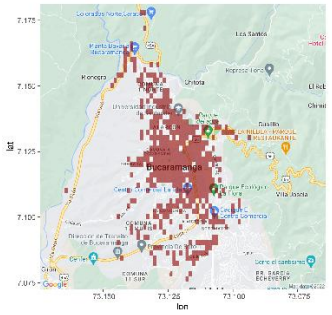
PPV	TPR	NPV	TNR	FPR	FNR	Accuracy	F1
0.473592	0.840625	0.848665	0.488889	0.151335	0.526408	0.61326	0.605856
0.450704	0.797508	0.807122	0.465753	0.192878	0.549296	0.583425	0.575928

0.40493	0.784983	0.813056	0.447712	0.186944	0.59507	0.556906	0.534262
0.110915	0.954545	0.991098	0.398093	0.008902	0.889085	0.438674	0.198738
0.498239	0.822674	0.818991	0.491979	0.181009	0.501761	0.61768	0.620614
0.487676	0.809942	0.807122	0.483126	0.192878	0.512324	0.60663	0.608791
0.491197	0.766484	0.747774	0.465804	0.252226	0.508803	0.58674	0.598712
0.419014	0.801347	0.824926	0.457237	0.175074	0.580986	0.570166	0.550289

Tabla 6. Métricas de rendimiento de CBMF

7.5.2. Comparación

A continuación, se realiza la comparación grafica de los resultados entre los dos modelos, para los 3 primeros meses del 2019.

MES	Original	Deep Learning	CBMF
01			
02			
03			

Para hacer la comparación de los modelos se tomó como modelo de referencia, el modelo de clasificación del framework H2O *randomforest,* el cual genera un bosque de árboles de clasificación. Se usaron los parámetros por default que usa el framework, 50 como el número de árboles y una profundidad máxima de 20 por árbol.

En el modelo se da igual importancia a la precisión y a la exhaustividad, por lo cual la medida final que se tendrá en cuenta es el puntaje F1.

A continuación, la tabla resumen de la comparación con el promedio para los ocho meses de prueba.

	Random Forest			**Deep Learning**			**CBMF**		
Mes	**Recall**	**Accuracy**	**F1**	**Recall**	**Accuracy**	**F1**	**Recall**	**Accuracy**	**F1**
201901	0.798165	0.067623	0.345865	0.620795	0.794909	0.590116	0.840625	0.61326	0.605856
201902	0.759146	0.081612	0.388206	0.655488	0.773091	0.579515	0.797508	0.583425	0.575928
201903	0.774411	0.069937	0.448441	0.710438	0.734545	0.536213	0.784983	0.556906	0.534262
201904	0.838235	0.01087	0.842975	0.779412	0.709091	0.209486	0.954545	0.438674	0.198738
201905	0.741477	0.090278	0.288828	0.616477	0.785455	0.595336	0.822674	0.61768	0.620614
201906	0.783626	0.07906	0.389522	0.649123	0.789818	0.60573	0.809942	0.60663	0.608791
201907	0.740541	0.098664	0.318408	0.618919	0.776	0.597911	0.766484	0.58674	0.598712
201908	0.758389	0.073394	0.426396	0.687919	0.763636	0.557823	0.801347	0.570166	0.550289
Promedio	0.774249	0.07143	0.43108	0.667321	0.765818	0.534016	0.822263	0.571685	**0.536649**

Tabla 7. Comparación de modelos

7. CONCLUSIONES

En la comparación realizada, aunque el *accuracy* del modelo de Deep learning es mejor, siempre es importante revisar otras medidas según el objetivo que se busca con el desarrollo del modelo, para este caso al medir mediante el puntaje F1, la propuesta de usar una factorización de matrices ajustada para la predicción mostró un rendimiento levemente mejor.

Los modelos mejoraron la predicción al incluir variables explicitas para indicar los crímenes en el pasado, pero esto conlleva tener que reentrenar el modelo cada vez que se quiere hacer una predicción para el siguiente mes.

El enfoque de usar una grilla en lugar de usar la división propia de la ciudad, en barrios o comunas, muestra varias ventajas, ya que se divide en espacios de igual tamaño el espacio sobre el cual se hace la predicción, evitando sesgos. Adicionalmente también permite construir modelos más dinámicos, ya que el tamaño de la grilla se puede ajustar según la precisión que se busque y el rango del área sobre el cual hacer la predicción.

La propuesta de usar un método de recomendación para la predicción de crimen muestra buenos resultados, a pesar de que en este estudio no se contaba con una variable de contexto cambiante en el tiempo, como la usada en el estudio de referencia, consistente en datos de Twitter[42].

El método de recomendación no requirió tanta capacidad computacional como el modelo de *Deep Learning*, lo que facilita su reentrenamiento con nuevos datos en menor tiempo.

Para una ciudad de tamaño medio como la que se seleccionó en este estudio el rendimiento de los dos enfoques es similar. Como trabajo futuro se podrían usar los datos de ciudades más grandes, tanto en área como en ocurrencia de crímenes.

El despliegue y uso de estos modelos implica que las instituciones definan procesos articulados para la recolección y actualización de los datos, de tal forma que los modelos tengan mayor precisión. Presentando así una oportunidad para que las entidades públicas en coordinación con la Policía Nacional creen una cultura de manejo y uso adecuado de los datos.

Este estudio es un punto de partida para continuar con pruebas piloto y evaluaciones experimentales cuyos resultados y datos permitan el mejoramiento de las políticas públicas encaminadas a la seguridad ciudadana. Generando un impacto en la percepción de seguridad y la tranquilidad de los ciudadanos. Y a nivel de las instituciones obteniendo beneficios al poder priorizar los esfuerzos y recursos de una manera más acorde al comportamiento y evolución que presenta el crimen.

Por último, los modelos desarrollados pueden extenderse para ser aplicados en otras ciudades del país con una volumetría de datos similar a los que se obtuvieron para Bucaramanga, teniendo en cuenta el tipo de crimen y el tamaño del área sobre la cual se desee realizar la predicción.

REFERENCIAS BIBLIOGRAFICAS

[1] Tariq M. Arif. 2020. *Introduction to Deep Learning for Engineers: Using Python and Google Cloud Platform*. Morgan & Claypool Publishers. DOI:https://doi.org/10.2200/S01029ED1V01Y202007MEC028

[2] Francisco Barreras, Alvaro Riascos, and Monica Ribero. 2017. Una Comparación De Diferentes Modelos Para La Predicción Del Crimen En Bogotá (A Comparison of Different Crime Prediction Models for Bogotá). *SSRN Electronic Journal* (2017). DOI:https://doi.org/10.2139/ssrn.2940343

[3] Umair Muneer Butt, Sukumar Letchmunan, Fadratul Hafinaz Hassan, Mubashir Ali, Anees Baqir, and Hafiz Husnain Raza Sherazi. 2020. Spatiooral crime hotspot detection and prediction: A systematic literature review. *IEEE Access 8*. DOI:https://doi.org/10.1109/ACCESS.2020.3022808

[4] Xinyu Chen, Youngwoon Cho, and Suk Young Jang. 2015. Crime prediction using Twitter sentiment and weather. In *2015 Systems and Information Engineering Design Symposium, SIEDS 2015*. DOI:https://doi.org/10.1109/SIEDS.2015.7117012

[5] Angelo Ciaramella, Antonino Staiano, Maria Raposo, Paulo Ribeiro, and Susana Sério (Eds.). 2020. *Computational Intelligence Methods for Bioinformatics and Biostatistics*. Springer International Publishing.

[6] Lawrence E. Cohen and Marcus Felson. 1979. Social Change and Crime Rate Trends: A Routine Activity Approach. *American Sociological Review* 44, 4 (1979). DOI:https://doi.org/10.2307/2094589

[7] Juan David GÉLVEZ-FERREIRA Pablo MONTENEGRO HELFER María Paula NIETO RODRÍGUEZ Carlos Andrés ROCHA RUIZ, Juan David GÉLVEZ-FERREIRA jgelvez, dnpgovco Pablo MONTENEGRO HELFER pabmontenegro, dnpgovco María Paula NIETO RODRÍGUEZ manieto, dnpgovco Carlos Andrés ROCHA RUIZ crocha, and dnpgovco Resumen. *PREDICCIÓN DEL DELITO EN COLOMBIA: EXPERIENCIA EN CIUDADES INTERMEDIAS*. Retrieved from https://www.dnp.gov.co/estudios-y-publicaciones/estudios-economicos/Paginas/archivos-de-economia.aspxhttp://www.dotec-colombia.org/index.php/series/118-departamento-nacional-de-planeacion/archivos-de-economia

[8] Defensoria del Pueblo. Defensoría emite alerta temprana para Bucaramanga ante riesgo de acciones de grupos armados ilegales y delincuenciales. *Comunicado 235 de 2021*. Retrieved February 23, 2022 from https://www.defensoria.gov.co/es/nube/comunicados/10638/Defensor%C3%ADa-emite-alerta-temprana-para-Bucaramanga-ante-riesgo-de-acciones-de-grupos-armados-ilegales-y-delincuenciales-alerta-temprana-Defensor%C3%ADa-Bucaramanga.htm

[9] Lian Duan, Tao Hu, En Cheng, Jianfeng Zhu, and Chao Gao. 2017. Deep Convolutional Neural Networks for Spatiotemporal Crime Prediction. *Proceedings of the International Conference on Information and Knowledge Engineering (IKE)* (2017).

[10] Ricardo Francisco Reier Forradellas, Sergio Luis Náñez Alonso, Marcela Laura Rodriguez, and Javier Jorge-Vazquez. 2021. Applied machine learning in social sciences: Neural networks and crime prediction. *Social Sciences* 10, 1 (2021). DOI:https://doi.org/10.3390/socsci10010004

[11] Google. 2022. Google Maps Platform Documentación. Retrieved May 20, 2022 from https://developers.google.com/maps/documentation/places/web-service/search-nearby

[12] Wilpen Gorr and Richard Harries. 2003. Introduction to crime forecasting. *International Journal of Forecasting* 19, 4 (2003). DOI:https://doi.org/10.1016/S0169-2070(03)00089-X

[13] Chao Huang, Junbo Zhang, Yu Zheng, and Nitesh v. Chawla. 2018. DeepCrime: Attentive hierarchical recurrent networks for crime prediction. In *International Conference on Information and Knowledge Management, Proceedings*. DOI:https://doi.org/10.1145/3269206.3271793

[14] IBM Cloud Education. 2020. Deep Learning. Retrieved November 14, 2021 from https://www.ibm.com/cloud/learn/deep-learning

[15] IBM Corporation. CRISP-DM Help Overview. Retrieved November 14, 2021 from https://www.ibm.com/docs/en/spss-modeler/SaaS?topic=dm-crisp-help-overview

[16] IBM Corporation. Deployment Overview. Retrieved November 14, 2021 from https://www.ibm.com/docs/en/spss-modeler/SaaS?topic=deployment-overview#crisp_deployment_phase

[17] Hyeon Woo Kang and Hang Bong Kang. 2017. Prediction of crime occurrence from multimodal data using deep learning. *PLoS ONE* 12, 4 (2017). DOI:https://doi.org/10.1371/journal.pone.0176244

[18] Ourania Kounadi, Alina Ristea, Adelson Araujo, and Michael Leitner. 2020. A systematic review on spatial crime forecasting. *Crime Science 9*. DOI:https://doi.org/10.1186/s40163-020-00116-7

[19] Yang Li, Suhang Wang, Quan Pan, Haiyun Peng, Tao Yang, and Erik Cambria. 2019. Learning binary codes with neural collaborative filtering for efficient recommendation systems. *Knowledge-Based Systems* 172, (2019). DOI:https://doi.org/10.1016/j.knosys.2019.02.012

[20] Ying Lung Lin, Tenge Yang Chen, and Liang Chih Yu. 2017. Using Machine Learning to Assist Crime Prevention. In *Proceedings - 2017 6th IIAI International Congress on Advanced Applied Informatics, IIAI-AAI 2017*. DOI:https://doi.org/10.1109/IIAI-AAI.2017.46

[21] Ying Lung Lin, Meng Feng Yen, and Liang Chih Yu. 2018. Grid-based crime prediction using geographical features. *ISPRS International Journal of Geo-Information* 7, 8 (August 2018). DOI:https://doi.org/10.3390/ijgi7080298

[22] Juan S Moreno Pabón, Mateo Dulce Rubio, Yor Castaño, Alvaro J Riascos, and Paula Rodríguez Díaz. 2020. *A Manifold Learning Data Enrichment Methodology for Homicide Prediction*.

[23] Amy E. Nivette, Renee Zahnow, Raul Aguilar, Andri Ahven, Shai Amram, Barak Ariel, María José Arosemena Burbano, Roberta Astolfi, Dirk Baier, Hyung Min Bark, Joris E.H. Beijers, Marcelo Bergman, Gregory Breetzke, I. Alberto Concha-Eastman, Sophie Curtis-Ham, Ryan Davenport, Carlos Díaz, Diego Fleitas, Manne Gerell, Kwang Ho Jang, Juha Kääriäinen, Tapio Lappi-Seppälä, Woon Sik Lim, Rosa Loureiro Revilla, Lorraine Mazerolle, Gorazd Meško, Noemí Pereda, Maria F.T. Peres, Rubén Poblete-Cazenave, Simon Rose, Robert Svensson, Nico Trajtenberg, Tanja van der Lippe, Joran Veldkamp, Carlos J.Vilalta Perdomo, and Manuel P. Eisner. 2021. A global analysis of the impact of COVID-19 stay-at-home restrictions on crime. *Nature Human Behaviour* 5, 7 (2021). DOI:https://doi.org/10.1038/s41562-021-01139-z

[24] Josh Patterson and Adam Gibson. 2019. *Deep learning. A Practitioner's Approach*.

[25] José Luis Pineda Arenas. 2022. Homicidios: "esclarecimiento de casos en Bucaramanga es del 51%". Retrieved February 23, 2022 from https://www.vanguardia.com/area-metropolitana/bucaramanga/homicidios-esclarecimiento-de-casos-en-bucaramanga-es-del-51-YF4689265#:~:text=M%C3%A1s%20de%20120%20homicidios%20registrados,una%20decena%20de%20homicidios%20adicionales.

[26] Francesco Ricci, Lior Rokach, and Bracha Shapira. 2015. *Recommender Systems Handbook*. DOI:https://doi.org/10.1007/978-1-4899-7637-6

[27] Geoffrey I Sammut Claude and Webb (Ed.). 2010. Latent Factor Models and Matrix Factorizations. In *Encyclopedia of Machine Learning*. Springer US, Boston, MA, 571. DOI:https://doi.org/10.1007/978-0-387-30164-8_887

[28] Secretaría de Seguridad Convivencia y Justica de Bogotá. SIEDCO. Retrieved November 14, 2021 from https://scj.gov.co/es/oficina-oaiee/bi/seguridad_convivencia/siedco

[29] BUCARAMANGA SECRETARIA DEL INTERIOR. 2022. Delitos en Bucaramanga enero 2010 a diciembre de 2021. Retrieved February 23, 2022 from https://www.datos.gov.co/Seguridad-y-Defensa/Delitos-en-Bucaramanga-enero-2010-a-diciembre-de-2/75fz-q98y

[30] Neil Shah, Nandish Bhagat, and Manan Shah. 2021. Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention. *Visual Computing for Industry, Biomedicine, and Art 4*. DOI:https://doi.org/10.1186/s42492-021-00075-z

[31] Thomas Wood. Convolutional Neural Network. Retrieved October 16, 2021 from https://deepai.org/machine-learning-glossary-and-terms/convolutional-neural-network

[32] Lara Vomfell, Wolfgang Karl Härdle, and Stefan Lessmann. 2018. Improving crime count forecasts using Twitter and taxi data. *Decision Support Systems* 113, (September 2018), 73–85. DOI:https://doi.org/10.1016/j.dss.2018.07.003

[33] Bao Wang, Penghang Yin, Andrea Louise Bertozzi, P. Jeffrey Brantingham, Stanley Joel Osher, and Jack Xin. 2019. Deep Learning for Real-Time Crime Forecasting and Its Ternarization. *Chinese Annals of Mathematics. Series B* 40, 6 (2019). DOI:https://doi.org/10.1007/s11401-019-0168-y

[34] Xiaofeng Wang, Donald E. Brown, and Matthew S. Gerber. 2012. Spatio-temporal modeling of criminal incidents using geographic, demographic, and twitter-derived information. In *ISI 2012 - 2012 IEEE International Conference on Intelligence and Security Informatics: Cyberspace, Border, and Immigration Securities*. DOI:https://doi.org/10.1109/ISI.2012.6284088

[35] Nathan Watt and Mathys C. du Plessis. 2018. Dropout algorithms for recurrent neural networks. *SAICSIT '18: Proceedings of the Annual Conference of the South African Institute of Computer Scientists and Information Technologists* (2018), 72–78. DOI:https://doi.org/https://doi.org/10.1145/3278681.3278691

[36] Matthew L. Williams, Pete Burnap, and Luke Sloan. 2017. Crime sensing with big data: The affordances and limitations of using open-source communications to estimate crime patterns. *British Journal of Criminology* 57, 2 (2017). DOI:https://doi.org/10.1093/bjc/azw031

[37] James Wilson and George Kelling. 1982. Broken Window. *Atlantic monthly* (March 1982).

[38] Junbo Zhang, Yu Zheng, and Dekang Qi. 2017. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *31st AAAI Conference on Artificial Intelligence, AAAI 2017*.

[39] Xu Zhang, Lin Liu, Luzi Xiao, and Jiakai Ji. 2020. Comparison of machine learning algorithms for predicting crime hotspots. *IEEE Access* 8, (2020). DOI:https://doi.org/10.1109/ACCESS.2020.3028420

[40] Zhang. Y. 2020. CBMF. Retrieved May 12, 2022 from https://www.ringspool.com/yihongzhang/material/kbs2020.zip

[41] Yihong Zhang, Panote Siriaraya, Yukiko Kawai, and Adam Jatowt. 2019. Time and location recommendation for crime prevention. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. DOI:https://doi.org/10.1007/978-3-030-19274-7_4

[42] Yihong Zhang, Panote Siriaraya, Yukiko Kawai, and Adam Jatowt. 2020. Predicting time and location of future crimes with recommendation methods. *Knowledge-Based Systems* 210, (2020). DOI:https://doi.org/10.1016/j.knosys.2020.106503