HURTO A PERSONAS EN LA CIUDAD DE MEDELLÍN: ANÁLISIS PREDICTIVO DE LA CANTIDAD DE CASOS EN DIFERENTES ZONAS DE LA CIUDAD A PARTIR DE MODELOS DE MACHINE LEARNING IMPLEMENTANDO TÉCNICAS DE MLOPS.

Robbery of Individuals in the City of Medellín: Predictive Analysis of the Number of Cases in Different Areas of the City Using Machine Learning Models Implementing MLOps Techniques.

Jeferson Stiven Arboleda Colorado

Proyecto de grado

Asesor

Juan David Martínez Vargas

UNIVERSIDAD EAFIT

Escuela de Ciencias Aplicadas e Ingeniería

Maestría en Ciencias de los Datos y Analítica

Medellín

2023

Contents

1 Resumen

El hurto a personas en la ciudad de Medellín es un problema que requiere una atención inmediata, ya que hace parte de los hechos de mayor victimización en la ciudad. Para garantizar una respuesta efectiva y eficiente, es crucial contar con un marco estructurado y ordenado para el desarrollo y despliegue de soluciones analíticas, las cuales me permitan tomar decisiones basadas en datos. De esta manera, se pueden obtener soluciones de alta calidad que satisfagan las necesidades y expectativas de los usuarios finales.

En este contexto, los modelos de aprendizaje de máquina (ML) se han convertido en una herramienta ampliamente utilizada en diversas áreas del conocimiento. Estos modelos permiten realizar predicciones basadas en los resultados obtenidos a partir de los datos, lo que facilita la toma de decisiones. Integrar de manera ágil el proceso detrás de los proyectos de aprendizaje de máquina se vuelve crucial para hacerlos más confiables y productivos. Es en este punto donde entra en juego la metodología de Machine Learning Operations (MLOps), la cual permite la actualización y el mantenimiento eficaz de los modelos.

En esta investigación, nos centraremos en abordar el problema del hurto a personas en la ciudad de Medellín mediante la implementación de las etapas del ciclo de vida de un proceso analítico. Comenzaremos desde la recopilación de datos y construcción de la base de datos, hasta el modelado y la puesta en producción. Siguiendo los principios de MLOps, nuestro objetivo es lograr resultados de manera ágil y tomar decisiones de manera eficaz. Además, al utilizar herramientas de inteligencia de negocios como Tableau, obtendremos representaciones gráficas nítidas y dinámicas que permitirán identificar patrones, tendencias y relaciones de manera intuitiva. Esto simplificará la toma de decisiones informadas y agilizará la obtención de valiosas perspectivas, específicamente en relación con el hurto a personas. Permitiendo, encontrar tanto vínculos entre los incidentes y las características urbanas correspondientes, como aquellos espacios de mayor densidad de sucesos en la ciudad.

En particular, utilizaremos modelos de regresión de ML para predecir los casos de hurto a personas en diferentes zonas de la ciudad de Medellín. Esto nos permitirá identificar las áreas con mayor incidencia de hurtos y comprender las características que las definen. Para llevar a cabo esta implementación, haremos uso de Amazon Web Services (AWS) como proveedor de nube, aprovechando los patrones de arquitectura utilizados en proyectos de datos.

Es crucial destacar la importancia de abordar este problema y cómo los modelos de aprendizaje de máquina pueden ayudarnos a encontrar soluciones efectivas. La combinación de un enfoque estructurado y ordenado, junto con el poder predictivo de los modelos de ML, nos permitirá generar resultados significativos para combatir el hurto a personas en Medellín y tomar acciones preventivas y correctivas basadas en la naturaleza de los datos recopilados.

En este sentido, exploraremos cómo, por un lado, al emplear modelos de redes neuronales convolucionales recurrentes, se obtiene un valor de 33.68 para el error cuadrático medio en la predicción de casos semanales para las distintas comunas de la ciudad. Por otro lado, al utilizar bosques aleatorios, se alcanza un valor máximo de 0.84 para el coeficiente de determinación (R^2) en la estimación de los hurtos totales por barrio durante el período analizado (2018-2022). Además, se focalizarán los lugares con mayor incidencia de los hechos y se identificarán aquellas características que podrían propiciarlos, permitiendo así la formulación de recomendaciones tanto cuantitativas como cualitativas en el estudio del fenómeno del hurto a personas en la ciudad de Medellín.

Palabras clave: MlOps, Machine learning, Predicción, Redes neuronales, Hurtos, Medellín.

2 Planteamiento del problema

Los temas concernientes a la seguridad ciudadana juegan un papel primordial en las políticas públicas colombianas y latinoamericanas, ya que representa un reto importante el abordar aquellas dinámicas complejas que deterioran la tranquilidad de la ciudadanía, esto debido a razones como la diversidad de escenarios donde ocurren, sus motivaciones, las estrategias de los agresores y la interacción entre diversos factores socioeconómicos y ambientales (Balbin, [1]). Esto no solo incluye las violencias asociadas a homicidios, narcotráfico y crimen organizado, sino también aquellas que con mayor frecuencia afectan la vida de la ciudadanía, como la violencia intrafamiliar, la extorsión, el hurto, entre otros (Departamento Nacional de Planeación, [2]).

Medellín, como el segundo centro urbano del país, no es ajena a las dinámicas mencionadas. Existe una alta incidencia en muchos de los delitos de alta frecuencia, incluyendo el hurto. Según cifras del Sistema de Información para la Seguridad y Convivencia (SISC) de Medellín, a mediados de marzo de 2023 se habían reportado 7 846 casos de hurto en lo corrido del año, de los cuales el 72% (5 655) correspondían a hurtos a personas. Estos delitos se concentran principalmente en las comunas de la Candelaria, Laureles-Estadio, el Poblado y Belén, un fenómeno territorial que ha sido constante en el tiempo. Surge, por lo tanto, la necesidad de analizar y comprender las dinámicas territoriales que rodean los hechos de hurto a personas en la ciudad de Medellín, para plantear estrategias de protección y prevención de los bienes e integridad de los habitantes de la ciudad.

El objetivo del trabajo de grado es estudiar el fenómeno del hurto a personas en la ciudad de Medellín desde la relación de los hechos con las características urbanas que lo rodean, utilizando el marco del ciclo de vida de un proceso en analítica. Este enfoque permitirá identificar características territoriales que propician la ocurrencia de los hechos y definir estrategias de prevención e intervención para los entes de control pertinentes.

La comprensión ágil del fenómeno del hurto en el contexto de la ciudad es crucial para la toma de decisiones oportunas y efectivas. Por lo tanto, se utilizarán técnicas de vanguardia como las proporcionadas por MLOps para mostrar como implementar, entrenar y desplegar modelos predictivos que ofrezcan información de calidad. Además, se buscará implementar herramientas visuales que permitan una interpretación ágil de la información y que sean útiles para aquellos interesados en el fenómeno en el futuro.

3 Justificación

El hurto a personas es uno de los delitos con mayor índice de victimización en la ciudad de Medellín, por lo que su estudio reviste gran importancia, ya que es un indicador de interés primordial en la toma de decisiones relacionadas con la seguridad ciudadana. En este contexto, es importante realizar, en el marco del ciclo de vida de los proyectos en analítica, un análisis estructurado y eficiente de la información, que permita soluciones enfocadas en los objetivos propuestos. Para ello, se recopilarán, procesarán y analizarán los datos, con el fin de determinar, a través de técnicas de aprendizaje de máquina la relación entre las variables urbanas y el número de casos de hurto en diferentes zonas de la ciudad. La implementación, entrenamiento y despliegue de los modelos se realizarán a través de la metodología ágil y confiable de Machine Learning Operations (MLOps), que es altamente productiva en términos de desarrollo y operaciones en aprendizaje de máquina. Además de esto, como método de comunicación de resultados, se utilizarán herramientas visuales que faciliten los procesos de decisión y la comprensión de los datos. Este enfoque a la vanguardia tecnológica permitirá tomar decisiones focalizadas y oportunas por parte de los entes de control en relación a las dinámicas que rodean el hurto.

4 Objetivos

4.1 Objetivo general

Predecir la tasa de hurtos en diferentes zonas de Medellín a partir de la relación de los hechos y las características urbanas que lo rodean, esto valiéndonos de técnicas de aprendizaje de máquina y su posterior implementación usando MLOps.

4.2 Objetivos específicos

- Construcción de base de datos adecuadas para el estudio del fenómeno.
- Determinar, a partir de la georreferenciación, los lugares de mayor densidad en casos de hurto a personas en la ciudad de Medellín.
- Encontrar relaciones entre el hurto a personas en la ciudad de Medellín y las características urbanas del hecho.
- Implementar, entrenar y desplegar modelos de aprendizaje de máquina para la predicción de hurtos, en un proveedor de nube usando técnicas de MLOps.
- Desplegar un tablero con información resumida de hurtos a personas en la ciudad de Medellín usando una herramienta de inteligencia de negocios.

5 Marco conceptual

En América Latina, el crimen y la violencia son temas de gran relevancia en el contexto social. Según la alcaldía de Medellín [3], el 33% de los casos de criminalidad en el mundo se concentran en esta región, a pesar de que sólo representa el 8% de la población. Esta característica convierte a la seguridad ciudadana en un problema de alta prioridad, no solo para los ciudadanos, sino también para los entes de control cuya misión es velar por la seguridad de la población [3]. Tal como explica Balbín en [1], la victimización y la violencia en estas zonas no se limitan únicamente a homicidios, narcotráfico y crimen organizado, sino que también hay dinámicas nacionales y locales que deterioran la seguridad, donde aparecen amenazas como el delito callejero, los delitos menores y la delincuencia organizada, factores que perturban la tranquilidad de las personas.

En esta línea, el hurto es uno de los delitos callejeros más comunes en Latinoamérica, pasando de ser hechos excepcionales a transformarse en actos cotidianos y de alta frecuencia [1]. Según muestra Jaitman en [4], esto obedece a un problema más endémico de la región, que a lo largo de su historia y desarrollo ha presentado escenarios de desigualdad, generando tensión social e incentivos económicos que son factores importantes para este tipo de hechos victimizantes [5]. Para Colombia, el hurto es uno de los delitos que más afecta los bienes particulares y que tiene mayor incidencia en la percepción de inseguridad ciudadana, según lo evidencia el análisis de crimen en el país durante el periodo 2016-2020 realizado por Padilla, et al.,[6]. En Medellín, según cifras del Sistema de Información para la Seguridad y la Convivencia (SISC), se han presentado 7 845 casos de hurto en lo corrido del año 2023 hasta la segunda semana de marzo, de los cuales aproximadamente el 72 % corresponden a hurtos a personas. Por lo tanto, es indispensable realizar estudios desde diferentes frentes para la intervención y creación de políticas públicas que permitan abordar y prevenir estos hechos victimizantes en la población. De esta manera, se podrá garantizar la seguridad e integridad de las personas y sus bienes.

Diferentes autores, valiéndose de la recopilación de datos por los diferentes entes de control, han buscado aportar en la prevención del crimen a través del análisis de estos a partir de diferentes técnicas de predicción e inferencia, buscando así comprender tanto las relaciones entre las variables que pueden propiciar el delito, como la estructura propia que enmarca las dinámicas

de estos y que llevan así a ocasionar una cantidad determinada de hechos. Es así como Kapoor en [7] usando aprendizaje de máquina y algoritmos de predicción como los árboles de decisión, en los que se busca dividir el espacio de características en zonas y así tener tanto tareas de clasificación como regresión [8], buscó realizar la predicción de casos asociados a diferentes delitos basado en las características del lugar de hecho. De manera similar Wheeler et al. [9], encontraron relaciones asociadas al territorio y los hechos victimizantes que se presentaban en el sector a través del análisis de riesgo usando características propias del lugar e implementando algoritmos como los bosques aleatorios [10] y los kernel de densidad [11]. Autores como Andresen et al. [12] por otro lado se han centrado en el análisis espacial de los hechos, buscando encontrar patrones y concentraciones en territorio que puedan ser abordadas desde los entes de control valiéndose de herramientas geoespaciales que permiten análisis de densidades a nivel de segmentos de vía. En general, múltiples autores como Ingilevich [13], Saraiva [14], Shukla [15], entre otros, han buscado utilizar herramientas estadísticas y de aprendizaje de máquina para abordar y aportar en la comprensión de hechos criminológicos, ya sea desde la predicción de estos, la comprensión de patrones o la distribución de densidades a nivel territorial.

En este sentido, abordar problemáticas de alta importancia social como las asociadas al crimen y, en particular el hurto a personas, usando herramientas cuya base son los datos, su estructura y la dinámica del fenómeno que se refleja en estos, toman relevancia en el contexto actual, donde debido a la vanguardia tecnológica en que nos encontramos, conocer los datos representa una ventaja estratégica, que en aras de la prevención y el cuidado de las personas y sus bienes toman un rol fundamental en el diseño de planes de intervención. Siguiendo esta línea y basados en las etapas del ciclo de vida de un proceso en analítica [16], se recopilaran y procesarán datos de fuentes como el Sistema de Información Estadístico, Delincuencial Contravencional y Operativo de la Policía Nacional (SIEDCO), el SISC, Medata, entre otros, para construir una base que permita realizar un análisis del fenómeno de hurto a personas en la ciudad de Medellín entre el año 2018 y 2022, el cual permitirá identificar los lugares con mayor densidad de hechos y a partir de estos evidenciar relaciones geoespaciales con el fenómeno, similar a lo realizado por Deryol [17] en el 2016. A través del aprendizaje automático y el uso de algoritmos de regresión supervisada, como la regresión lineal [18], los bosques aleatorios [10], xgboost [19] y los basados en grafos [20], se modelará la cantidad de hurtos asociados a diferentes zonas de la ciudad [16]. Estas zonas se definirán previamente según un análisis de la unidad mínima de territorio elegido para la división, como se realizó en Kim [21] y Das [22] para el caso de los grafos.

Debido a la relevancia actual en la toma de decisiones basadas en datos, se ha convertido en una necesidad tener nuestros modelos como un producto disponible que genere conocimiento a medida que la información aumenta y evoluciona en el tiempo, de allí que se hacen necesarias herramientas y técnicas de automatización de software que en unión con el aprendizaje de máquina nos permita construir sistemas complejos para la disponibilización de nuestros productos como un servicio [23]. Es allí cuando aparece lo que se conoce actualmente como machine learning operations (MLOps), que consiste en el procedimiento colaborativo de agilizar el proceso de llevar los modelos de aprendizaje automático a producción, y luego mantenerlos y monitorearlos [24], lo cual autores como Symeonidis et al. [25] y Matsui et al. [26] se han encargado de plantear y definir de manera efectiva. En esta línea, se diseñará, planteará y desplegará una arquitectura en la nube, que pueda alojar un trabajo de características similares al ejecutado, en el cual el resultado sea puesto en producción basándonos en los fundamentos de MLOps y las herramientas disponibles para ejecutarlo, logrando de esta manera tener un producto de ágil implementación y adaptación ante nuevos datos, cuyo análisis y monitoreo dependerá del momento temporal, lo cual hará dichos resultado más fiables al instante de describir los fenómenos, en este caso el hurto a personas y, por ende tendremos una asertividad mayor en la toma de decisiones y la descripción de las dinámicas asociadas a dicha manifestación del crimen.

La toma de decisiones ágil y cercana al tiempo real requiere herramientas que ofrezcan una visualización clara y concisa de los datos dentro de una organización. En este sentido, los tableros de analítica desempeñan un papel fundamental al permitir a las organizaciones identificar

tendencias y patrones en los datos. Esto, a su vez, posibilita predecir futuros resultados y tomar medidas preventivas para mitigar riesgos potenciales. Con este objetivo en mente, se busca disponibilizar un tablero que resuma las principales características del hurto a personas en la ciudad de Medellín, brindando utilidad a todos aquellos interesados en comprender estas dinámicas.

Lo mencionado en párrafos anteriores se basa en lo definido en [16] y [27], donde el ciclo de vida de un proyecto en analítica busca establecer objetivos claros, recopilar los datos necesarios, procesarlos, generar modelos que aporten nuevo conocimiento, evaluar su implementación y finalmente disponibilizarlos como un servicio. Por lo tanto, se considera que este proyecto tiene fundamentos sólidos y está a la vanguardia de los proyectos actuales a nivel de la industria.

6 Metodología

En cuanto a la metodología de trabajo, en línea con lo mencionado anteriormente sobre el ciclo de vida de un proceso en analítica, se hará uso de la denominada Cross-Industry Standard Process for Data Mining (CRISP-DM), la cual según Wirth [28] posee seis fases que definen los lineamientos y proporcionan una descripción detallada de los pasos a seguir en un proyecto estándar de análisis de datos. Estas fases son: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue.

Para nuestro caso, la comprensión del negocio se centrará inicialmente en el entendimiento de las dinámicas asociadas al fenómeno de hurto a personas en la ciudad de Medellín en los últimos años, lo cual guiará nuestro análisis tanto desde una mirada cualitativa como cuantitativa, permitiéndonos plantear hipótesis y planes de trabajo que nos lleven a responder preguntas de interés para la prevención y cuidado de la ciudadanía en este ámbito, tales como los lugares de mayor incidencia del hecho en la ciudad, comportamiento a lo largo del tiempo y la predicción de la tasa de estos a partir de las características urbanas del entorno.

Buscando los objetivos propuestas de encontrar relaciones entre el hurto a personas y las características urbanas de los hechos, se usarán las fuentes de registros de casos dispuesta por el SIEDCO y el SISC en el cual se tabulan datos asociados a las características de las acciones victimizantes que son denunciadas a la policía a lo largo del año. Seguido a esto, usando tanto bases de datos dispuesta por los entes gubernamentales de la ciudad (Medata), como aquellas obtenidas de fuentes como el api de google maps, en los cuales se registran espacios de relevancia como los bares, estaciones de transporte, parques, entre otros, se harán cruces geoespaciales con los registros de los hechos para construir una colección de datos que reúna las características necesarias inicialmente para nuestros análisis. De esta manera tendremos una preparación inicial de nuestra información, la cual, con una posterior limpieza de esta y un análisis de características definirán un punto de partida para comprender los datos asociados al fenómeno en cuestión y así realizar una distinción minuciosa que permita, valiéndonos de estadísticas descriptivas, generar hipótesis iniciales acerca de las relaciones del hurto a personas y el espacio que rodea los hechos en la ciudad de Medellín.

Valiéndonos de frameworks como Scikit-learn [29] y XGBoost [19], así como de nuestro conocimiento en estadística y técnicas de aprendizaje de máquina, modelaremos nuestros datos utilizando diferentes algoritmos de regresión. En el caso de la predicción de la tasa de hechos, comenzaremos con algoritmos clásicos como la regresión lineal [18], y posteriormente utilizaremos algunos de ensamble, como los bosques aleatorios [10], XGBoost y métodos de grafos [20]. Estos algoritmos nos permitirán hacer predicciones y evaluar su ajuste en el momento de generalizar, definiendo así un modelo para nuestro trabajo a partir del que genere el mejor rendimiento. De forma paralela, construiremos un tablero analítico en el que podamos visualizar las principales características asociadas con la dinámica de hurto en Medellín.

Finalmente, se plantea una implementación que busca ofrecer nuestro trabajo como un servicio, el cual sea fácilmente manejable ante las variaciones propias de la dinámica del fenómeno

en el tiempo y, por ende, de los datos asociados. Para lograr esto, se utilizarán técnicas de MLOps y herramientas propias de un proveedor de nube como AWS para diseñar e implementar una arquitectura que permita tener un servicio disponible de manera constante, generando conocimiento y de fácil acceso para el usuario final.

7 Recopilación de los datos

Con el fin de alcanzar nuestras metas, fue necesario recopilar la información necesaria para construir una base de datos que estuviera alineada con la finalidad del trabajo de grado, es decir, datos asociados con los hurtos a personas y las características del territorio, donde esta última presentara una componente espacial que permitiera vincularla con el lugar de los hechos delictivos estudiados. Para esto se definieron dos estrategias:

- Rastreo de bases de datos en la web
- Utilización de la API de google maps

Entre ambas, se logró recopilar un número significativo de características, las cuales se describirán a grandes rasgos a continuación, al igual que el proceso de obtención de las mismas.

7.1 Rastreo de bases de datos en la web

Los hechos delictivos que son puestos en conocimiento de la Policía nacional de una manera formal, son recopilados y registrados en bases de datos que estos administran y de las cuales otras organizaciones se alimentan para nutrir sus diferentes análisis; una de estas es el Sistema de Información para la Seguridad y Convivencia (SISC), el cual se encarga de hacerle seguimiento al comportamiento espaciotemporal de los principales indicadores de seguridad y convivencia en la ciudad de Medellín, entre los cuales encontramos el hurto a personas.

Los datos estandarizados y anónimizados por el SISC son públicados y actualizados constantemente en la pagina web de medata[1] de forma tal que toda persona interesada en estos temas tenga acceso a la información, de esta manera se obtuvo una sábana de datos con el total de Hurtos a personas reportados en la ciudad de Medellín entre el $2003 - 2022$ [30], aunque como se verá más adelante, solo se tomaron los registros entre el $2018 - 2022$, pues es en este periodo que se tiene una similitud en la cantidad de casos entre cada año. Dichos datos tabulares consistían de alrededor de 295405 registros y 36 características, de estas últimas, algunas relacionadas con la georreferenciación y categorizacion del hecho, los bienes hurtados en este, entre otras.

Seguidamente, se llevó a cabo una búsqueda de datos geoespaciales relacionados con la ciudad de Medellín, con el fin de obtener información sobre los objetos geográficos en esta, donde tuvieramos referencias como los límites en los mapas, nombres de barrios, comunas y carreteras dentro de la ciudad. En particular, se buscaron datos que representaran las diferentes subdivisiones de la ciudad de Medellín a través de polígonos (comunas, barrios y manzanas), ya que esto nos permitiría establecer una relación directa entre los delitos estudiados y el espacio territorial en el que fueron cometidos. De esta manera, se obtuvieron los datos que se detallan en la tabla 1:

[1] Portal de datos de la Alcaldía de Medellín.

Tabla 1: Fuentes geoespaciales

Fuente geoespacial	Descripción
Límite barrio vereda catastral	Esta capa representa el límite de los Barrios en la ciudad
Límite catastral de comunas y corregimientos	Esta capa representa el límite de las comunas y corregimientos en la ciudad
Manzana catastral	Esta capa contiene los contornos de las manzanas en la ciudad

Finalmente se hizo una búsqueda minuciosa de data que tuviera información asociada con características propias de la ciudad de Medellín y su configuración, es decir, elementos como iglesias, centros de espectáculo, árboles, instituciones educativas, entre otros, los cuales en conjunto aportan a la construcción del entorno propio de la ciudad, y del cual nos valímos para encontrar sus relaciones con los hechos delictivos asociados al hurto a personas. Se encontraron alrededor de 27 fuentes que nos entregaban información asociada con algunos de los elementos listados a continuación

- Acopio de taxis
- Árboles zona urbana
- Atractivos turísticos
- Camaras ARS - SIMM
- Camaras CCTV Movilidad -SIMM
- Camaras Fotodeteccion - FDT - SIMM
- Ciclorrutas
- Cruces semafóricos
- Ecoparques
- Estaciones (Transporte masivo)
- Cantidad establecimientos - hacienda
- Paradas de Transporte Público Colectivo
- Bienes de Interes Cultural BIC
- Quebradas
- Senalizacion
- Sitios de aprovechamiento de residuos sólidos
- Usos del Predio
- Zonas verdes
- Usos Rurales
- Rutas unificadas de Transporte Público Colectivo
- Instituciones Educativas
- Escenarios deportivos

Una información más detallada acerca de las fuentes mencionadas anteriormente puede ser encontrado en la tabla cuyo link se adjunta en los anexos. Es de anotar que cada una de las fuentes encontradas, se filtraba inicialmente basandose en si tenía o no su información de georreferencia, pues como se ha mencionado, esta es importante para su asociación con el espacio.

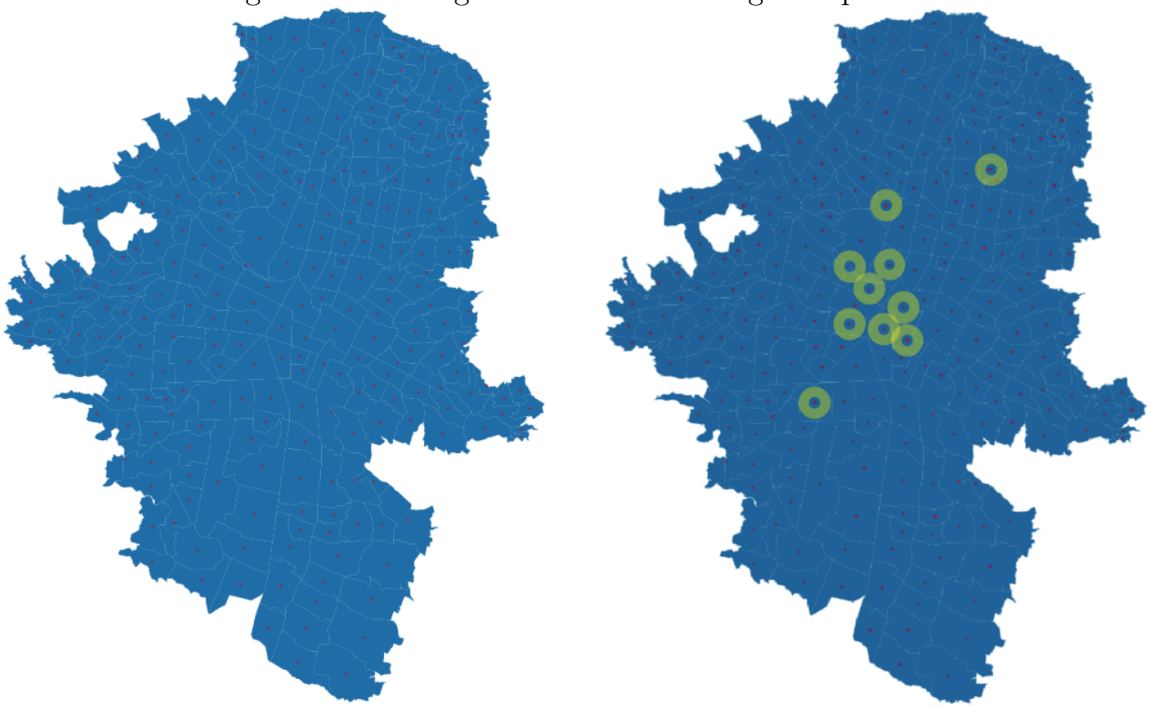
7.2 Utilización de la API de google maps

Si bien la información recopilada a través de la búsqueda en la web nos proporcionó un punto de partida importante, hubo características territoriales que no pudimos obtener mediante fuentes públicas. Por esta razón, buscamos una alternativa para aliviar esta situación. Optamos por utilizar el API de Google Maps, el cual nos permite obtener información sobre las zonas geográficas cercanas a un punto específico mediante su método nearbysearch (Google Maps API, [31]), teniendo ciertas restricciones radiales. Sin embargo, esto presentó un reto inicial, por lo que detallaremos de manera general el proceso que llevamos a cabo.

El método de búsqueda "nearby search" posee, entre muchos de sus parámetros, el radio y el tipo de lugar a buscar. El radio tiene un límite de 50 km, mientras que el segundo se restringe a 97 categorías que engloban los diferentes lugares registrados en Maps, donde un lugar físico puede tener múltiples categorías. Además, el número máximo de elementos o lugares que se pueden obtener desde la api por token es de 60, y se paginan en conjuntos de 20. En otras palabras, puedo realizar tres peticiones con un mismo token, después de lo cual debo generar uno nuevo. Los 60 elementos mostrados son aquellos cuya cercanía al punto origen definido es mayor. Si hay más de 60 elementos dentro del límite radial, no se devolverán.

Teniendo esto claro se definió una estrategia en la cual, valiéndonos de la división espacial de la ciudad de medellín por barrios, se encontraron los centroides de estos polígonos y seguido a esto se generaron peticiones por cada uno de estos a la API de google maps como se ve en la figura 1.

Figura 1: Estrategia de rastreo en Google Maps API



Título de la figura.

En la anterior figura se evidencia que al generar peticiones basandonos en los centroides, obteniamos zonas que podían solaparse debido a que lugares cercanos a un centroide podían también serlo a uno vecino, como se evidenica en la representación de la derecha, donde las circunferencias amarillas representan este hecho. Para alivianar esto, con ayuda de los valores coordenados de los lugares encontrados, al finalizar nuestras ejecuciones se hizo un eliminado de elementos duplicados. Con esta estrategia se superó la restricción del número de elementos regresados en cada petición.

En cuanto a las categorías se eligieron 24 para ser rastreadas, esto con el fin de optimizar costos y no solapar la información obtenida a través de bases de datos públicas. Estas categorías fueron:

- Farmacias
- Tiendas de muebles
- Estaciones de policía
- Estanciones de suministro de gasolina
- Cajeros automáticos
- Gimnasios

- Bancos
- Restaurantes
- Bares
- Parqueaderos
- Joyerías
- Tiendas de calzado
- Mall comerciales
- Cafés
- Tiendas de licores
- Tiendas
- Supermercados
- Casinos
- Hostales
- Iglesias
- Tiendas de ropa
- Centros nocturnos
- Centros médicos

Iterando sobre los centroides y las categorías, se tabuló la información relacionada con los lugares de la ciudad de Medellín. Esta información incluía la latitud y longitud como aspectos georreferenciadores, el nombre del lugar, la dirección y la categoría. Este conjunto de datos resultó ser de gran utilidad para definir la cantidad de elementos de cierta categoría asociados con los polígonos que dividen la ciudad. Como se mostrará más adelante, esta información fue fundamental para nuestro análisis.

Las implementaciones asociadas con la obtención de la data que se acaba de describir, puede encontrarse en el repositorio asociado con el trabajo de grado, el cual se referencia y aborda más adelante.

8 Descripción del proceso realizado

Una vez realizado el proceso de recopilación de datos descrito en la sección anterior, se inició con un proceso de tratamiento de los datos asociados a las diferentes fuentes, donde se enmarca la exploración de estas, su procesamiento y su posterior unión valiéndonos de la georreferenciación en cada una.

Luego de finalizado este proceso se procedió a modelar de acuerdo a los objetivos planteados, por un lado a partir de modelos asociados a grafos y por el otro con modelos de regresión. A continuación se entrará en detalle en cada una de estas fases.

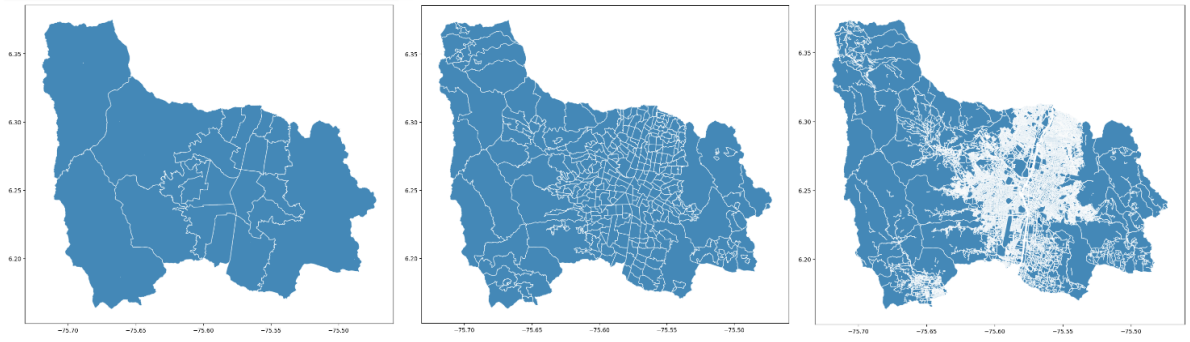
8.1 Análisis y exploración de las diferentes bases de datos

8.1.1 Shapefiles.

Como se mencionó anteriormente, en nuestra búsqueda de información geospacial asociada con la ciudad de Medellín se encontraron tres shapefiles[2], los cuales a través de polígonos dividían el municipio, en manzanas, barrios y comunas, como se observa en la figura 2.

[2]Formato sencillo y no topológico que se utiliza para almacenar la ubicación geométrica y la información de atributos de las entidades geográficas [32]

Figura 2: División geospacial Medellín. Comunas, barrios y manzanas



Cada uno de estos shapefiles viene acompañado de características del territorio definido por los poligonos, como su nombre, ínidice, área apróximada, id, entre otras. Para cada uno de estos se encontraron las dimensiones en la tabla 2.

Tabla 2: Dimensiones bases de datos geoespaciales

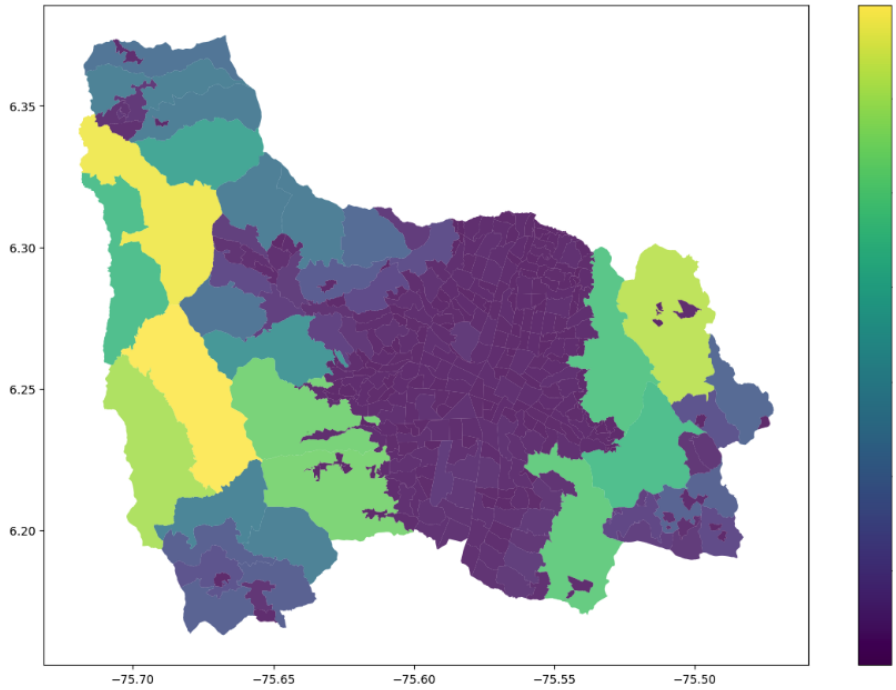
Fuente geoespacial	**columnas**	**filas**
Límite barrio vereda catastral	15	349
Límite catastral de comunas y corregimientos	8	21
Manzana catastral	7	10800

Donde el número de filas corresponde a el número de divisiones que se tienen, es decir, para el caso de nuestros polígonos asociados a las comunas encontramos 21 registros, los cuales corresponden, como es sábido, con el número de comunas y corregimientos que posee el municipio, donde estos últimos son 5 y se asocian con la zona rural cercana de Medellín. Caso similar sucede con las demás subdivisiones.

Una de las tareas iniciales consistió en elegir la subdivisión adecuada para la ciudad. Por un lado, no podía ser demasiado refinada, ya que al realizar conteos de características por zona, se obtendrían matrices altamente dispersas y quizás las características no tendrían mucha relevancia en dichas áreas. Por otro lado, las zonas no podían ser muy amplias, ya que se definiría una cantidad baja de muestras, lo que no sería muy eficiente para utilizar un modelo y un análisis riguroso. Por esta razón, era necesario buscar un punto intermedio que cumpliera con nuestro objetivo.

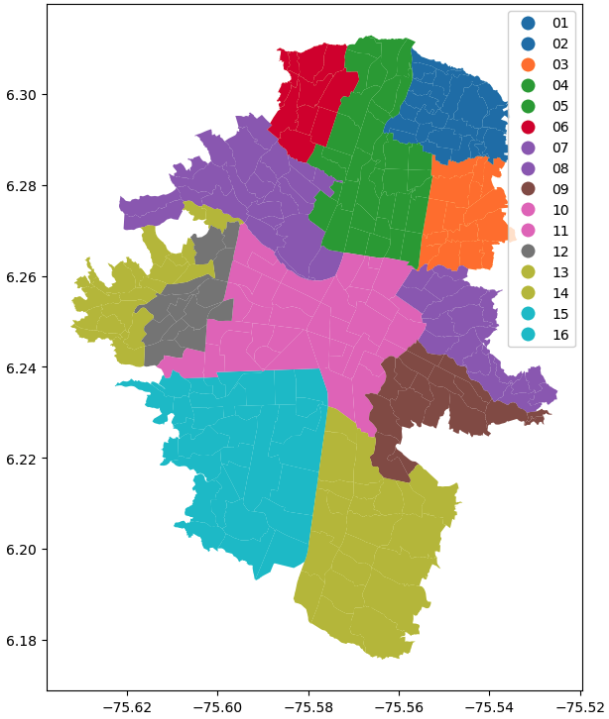
Adicionalmente, al conocer el municipio de Medellín, podemos identificar que en los corregimientos, debido a su alta ruralidad, muchas de las características analizadas en el casco urbano de la ciudad no estarían presentes o lo estarían en menor cantidad. Además, como se verá más adelante, la cantidad de hurtos en estas zonas es muy baja en comparación con el resto de la ciudad debido a características como la densidad de población, las dinámicas propias de estos lugares y la cantidad de zonas dedicadas a la agricultura. Esto hace que no sean un escenario propicio para los encuentros delictivos en la misma cantidad que en las zonas urbanas. Además, la diferencia en el área que cubren estos territorios podría no representar una equiparación con respecto a las posibles subdivisiones de la zona urbana de la ciudad, lo que llevaría a tener desbalances intrínsecos al momento de analizar densidades de hurto. Por estas razones, centraremos nuestro análisis solo en la zona urbana de Medellín, y los corregimientos no serán considerados en el estudio. Tomaremos una división por barrios, siendo este un punto intermedio entre los shapefiles encontrados. Como se puede ver en la figura 3, a nivel de áreas cubiertas por estas divisiones, encontramos que la zona urbana (color morado) se equipara entre sí, por lo que, al menos en relación de áreas, se da peso a la decisión tomada.

Figura 3: División geospacial por barrios de Medellín. En color el área cubierta



Dada la decisión tomada, en nuestro análisis se considerará la incidencia de las 16 comunas de la parte urbana de la ciudad de Medellín, las cuales abarcan 265 barrios. A modo de referencia, se puede ver su división a continuación en la figura 4.

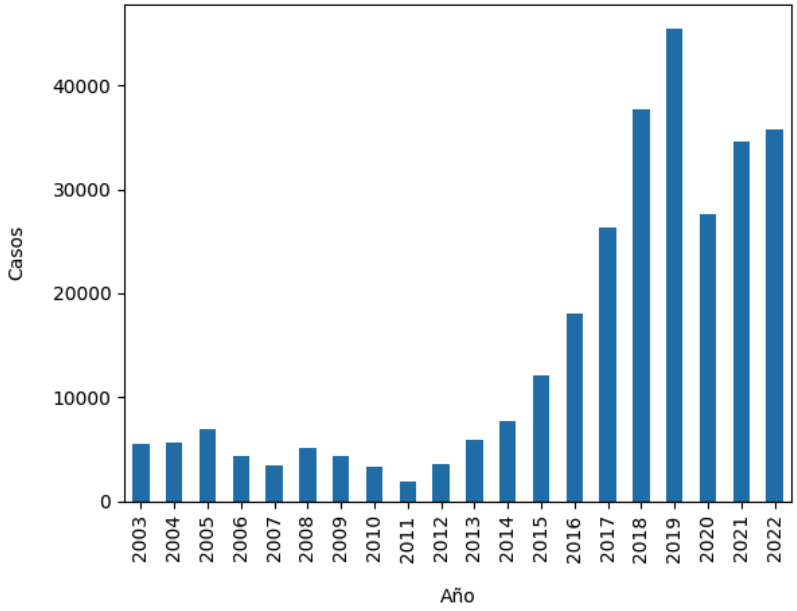
Figura 4: División geospacial por barrios y comunas de Medellín.



8.1.2 Base asociada al hurto a personas.

Nuestra base principal contiene los registros asociados a la denuncia de hurtos a personas en la ciudad de Medellín durante el periodo 2003-2022, como se aprecia en la figura 5. Los últimos años son aquellos con mayor cantidad de registros debido a la sistematización y control de los datos asociados a este delito. En total, la base de datos cuenta con 295 405 registros y 36 características, como se mencionó anteriormente.

Figura 5: Cantidad histórica de hurtos por año en la base de datos



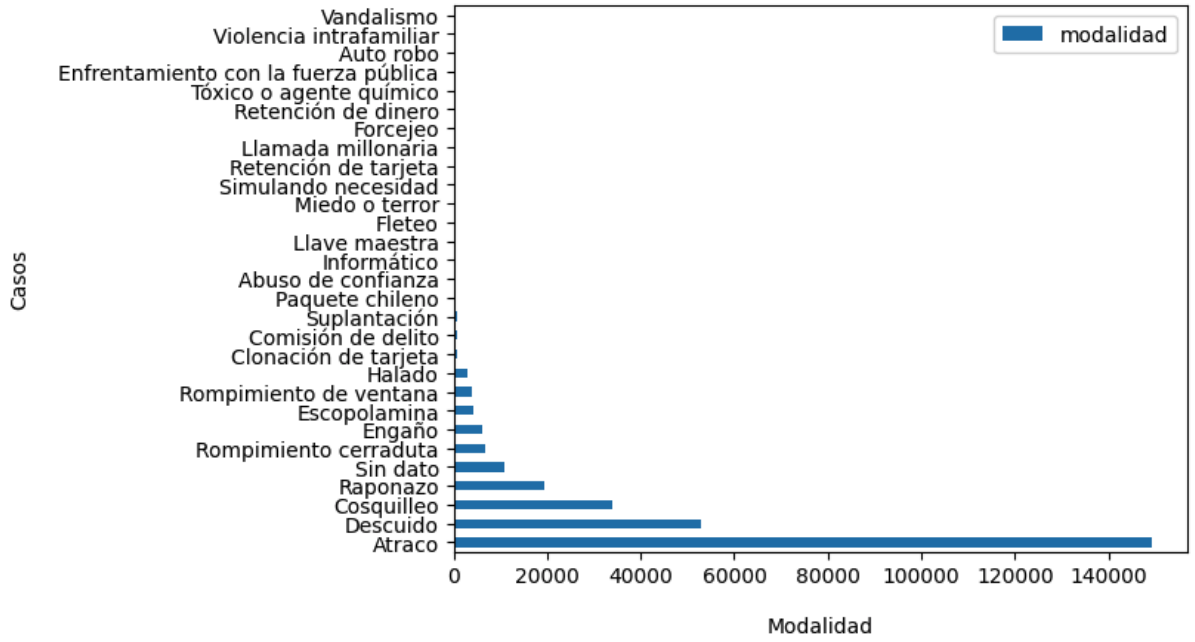
En cuanto a las características y su esquema tenemos la situación siguiente

- fecha_hecho (datetime)
- cantidad (float)
- latitud(float)
- longitud (float)
- sexo (string)
- edad (int)
- estado_civil(string)
- grupo_actor (string)
- actividad_delictiva (string)
- parentesco (string)
- ocupacion(string)
- discapacidad (string)
- grupo_especial(string)
- medio_transporte(string)
- nivel_academico (string)
- testigo(string)
- conducta (string)
- modalidad(string)
- caracterizacion(string)
- conducta_especial (string)
- arma_medio(string)
- articulo_penal(string)
- categoria_penal (string)
- nombre_barrio (string)
- codigo_barrio (string)
- codigo_comuna (string)
- lugar(string)
- sede_receptora(string)
- bien (string)
- categoria_bien(string)
- grupo_bien(string)
- modelo (int)
- color(string)
- permiso(string)
- unidad_medida (string)
- fecha_ingestion (string)

Donde se puede observar que muchos de estos atributos hacen referencia a las características propias del hecho, como la modalidad, la caracterización, el medio o bien hurtado, y muchas

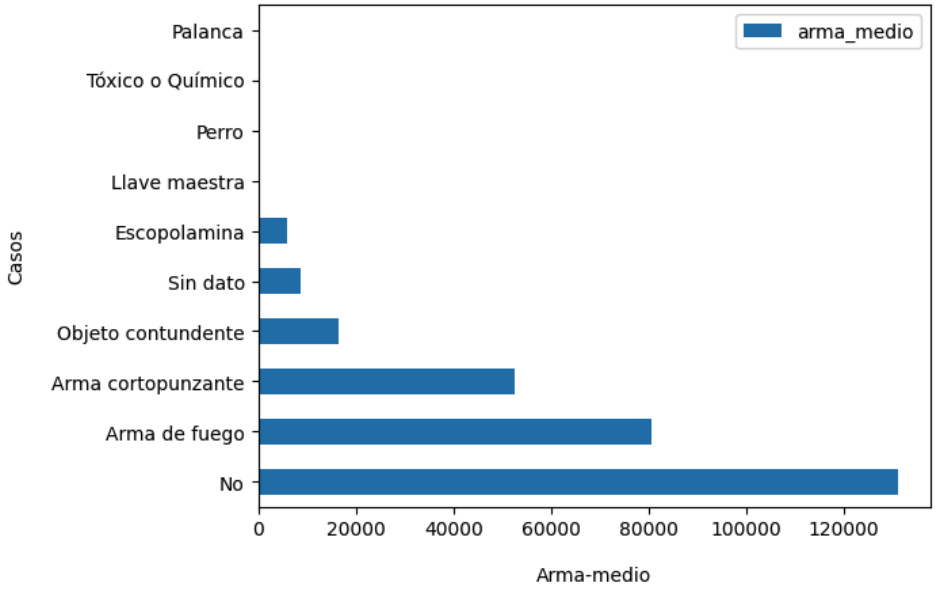
otras que en gran medida se centran en la víctima del vejamen, aludiendo a su sexo, edad, estado civil, entre otras. Si bien nuestro análisis se centrará en el lugar del hecho y sus características espaciales, valores como la modalidad y el arma empleada en el hurto nos pueden proporcionar información relevante sobre la dinámica del fenómeno, lo que nos permitirá entenderlo a nivel de la ciudad. Como veremos más adelante, esto se tuvo en cuenta al crear nuestro tablero. Gráficamente, los hurtos de acuerdo a su modalidad se distribuyen de la siguiente manera:

Figura 6: Distribución modalidad de hurto a personas en Medellín 2018-2022



Vemos que la mayoría de estos se encuentran en la categoría de atraco, descuido, cosquilleo y raponazo, lo cual, dado el caso de que un ente de control guíe sus esfuerzos a combatir el hurto a personas en la ciudad, podría centrarse inicialmente en aquellos bajo estas categorías. Con este panorama, se espera que al preguntarnos por el arma usada durante los hechos prevalezca aquellas que propicien los hurtos en modalidad de atracos(aquellos con violencia o intimidación), como armas de fuego o cortopunzantes. Al fijarnos en la distribución de esta característica encontramos lo representado en el gráfico 7.

Figura 7: Distribución arma usada en los hurtos a personas en Medellín 2018-2022



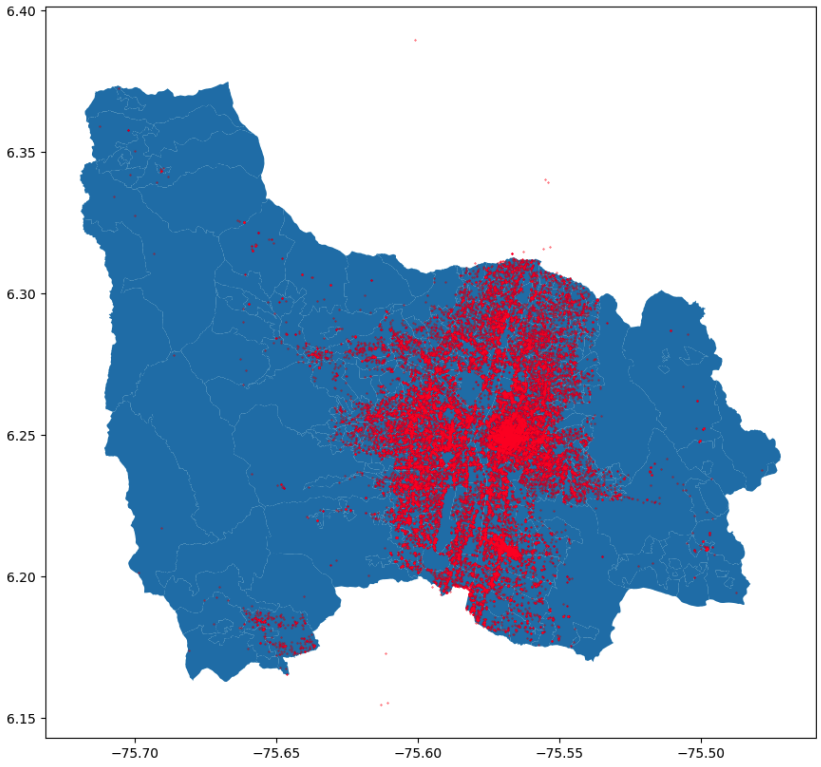
Lo anterior está en concordancia con lo evidenciado en la figura (6) pues un gran número de hurtos se han dado en modalidades que no presentan contacto físico directo entre víctima y agresor, como el cosquilleo o el descuido, por lo que no deberían tener involucradas armas, caso contrario a los atracos que pueden tener de por medio armas, lo que hace que también encontremos un gran número de casos con la presencia de estas.

Como se ha mencionado, nuestro análisis se centrará en el periodo de tiempo comprendido entre el 2018 y el 2022, donde se tiene una cierta madurez en el tratamiento de los datos relacionados a los hurtos, al igual que se presentan comportamientos similares en sus dinámicas, por lo que cuatro años definenen una ventana temporal aceptable, en la cual tendremos un total de 181204 casos registrados; es de anotar que esta cifra puede ser menor a la real, pues se tiene un fenómeno de desconfianza generalizado hacia las entidades protectoras por parte de la ciudadanía, por lo que las personas no acuden a denunciar, lo cual genera un subregistro en la información.

Nuestro análisis de manera principal se centrará en las características del hurto y su relación con el territorio, es por esto que al hacer una depuración de características nos quedamos con aquellas que aporten en este fin, quedando así con fecha_hecho, latitud, longitud, modalidad, arma_medio, nombre_barrio, codigo_barrio, codigo_comuna. Estas columnas tienen la particularidad de no tener valores nulos y además alinearse con nuestros objetivos, pues tienen la fecha y la georreferenciación de los hechos, lo que en principio nos permitiría sacar un conteo por polígonos en la ciudad de Medellín. De igual manera las demás columnas conservadas buscan responder preguntas que de manera posterior se verá son resueltas a través de una herramienta visual, presente en nuestros objetivos.

Al analizar las coordenadas asociadas con los hechos, se encontró que algunos registros en un primer momento se encontraban mal georreferenciados, pues su ubicación se daba lejos de la ciudad, es por eso que se hizo una depuración inicial valíendonos de las coordenas de la región en que, de forma aproximada, se encuentra el municipio de Medellín. Un primer acercamiento a la visual de los hurtos en el espacio de la ciudad se encuentra en la figura 8.

Figura 8: Distribución espacial de hurtos a persona en Medellín 2018-2022



Si bien la imagen no es del todo clara, nos da una idea de dónde encontraremos los lugares con mayor densidad de casos en la ciudad. De forma rápida, podemos evidenciar que cerca de la comuna 10 se presenta un gran número de casos, lo cual es de esperarse ya que esta comuna alberga el centro de la ciudad, lugar por el que, según cifras de Medellín Cómo Vamos [33], se mueven alrededor de 1.2 millones de personas diariamente, lo que lo convierte en un lugar donde el factor de oportunidad es mayor y, por ende, se propicia un alto número de hurtos. Además, esta imagen nos permite validar visualmente lo mencionado anteriormente, donde se afirmó que la densidad de hurtos en los corregimientos es mucho menor que en la parte urbana de la ciudad.

Es importante tener en cuenta que la georreferenciación puede no ser completamente precisa debido a las dinámicas del proceso de denuncia y registro de los casos de hurto. En muchos casos en los que no hay contacto directo, las víctimas proporcionan una estimación del lugar en el que pudieron haber sido afectadas, lo que significa que la ubicación registrada puede hacer referencia al lugar donde la víctima se dio cuenta del hurto o a un lugar cercano en espacio a ese momento. Sin embargo, este problema se ha reducido en gran medida gracias a la elección de la división de la ciudad por barrios en el análisis, ya que proporciona un amplio intervalo de precisión espacial.

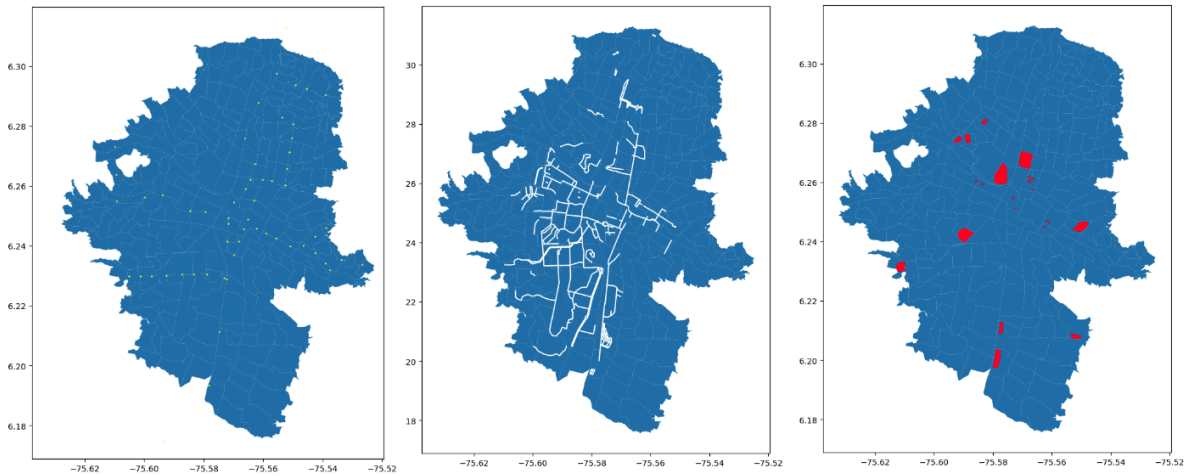
8.1.3 Bases de datos relacionadas con características territoriales.

Como se mencionó en la sección de recopilación de datos, se realizó una búsqueda de información asociada con características territoriales de la ciudad, en total fueron alrededor de 30 bases de datos, a las cuales se realizó un análisis inicial para validar el número de registros asociados, la calidad de su georreferenciación y su pertinencia en el análisis a realizar. Al referirnos a la calidad de la georreferenciación, nos aseguramos de que los registros estén dentro de los límites de la ciudad, sin embargo, no se tuvo en cuenta su precisión real dentro de estos límites.

En todas estas sábanas de datos encontramos coincidencia en su ubicación con la ciudad de Medellín, validación realizada usando las características geospaciales que poseían cada una, por lo que inicialmente cumplen con las condiciones necesarias para su uso. De estas, alrededor del

70% poseen su característica geospacial de tipo puntual, el 23.3% de tipo lineal y los restantes de carácter poligonal. En la siguiente figura se pueden ver sus diferencias.

Figura 9: Diferencias entre tipos de objetos geométicos. (a) puntual, (b) lineal, (c) polígono



La figura 9 ejemplifica las diferencias entre objetos geométricos encontrados en el rastreo de bases, de forma partícular se presentan los datos espaciales para las estaciones del metro de la ciudad, ciclorrutas y universidades, de izquierda a derecha respectivamente.

Es de anotar que cada una los datos recopilados por medio de la api de google maps, tienen una georreferenciación de forma puntual y que cada uno de estos presenta dicho dato, pues fue a partir de este, como se explicó anteriormente, que se filtraron los diferentes registros.

8.2 Procesamiento de los datos

Una vez explorados los datos, se procedió a realizar un procesamiento sencillo en el que se enfocó en contar los casos de acuerdo con la subdivisión espacial de la ciudad, es decir, los barrios, por lo que el procesamiento no se llevó a gran profundidad.

Inicialmente, se validó el objeto geométrico que definía la georreferenciación de cada sábana de datos y se aseguró de que cada registro tuviera dicho valor; se descartaron aquellas muestras sin componente espacial. Luego, con la ayuda de herramientas como Geopandas, se leyeron los dataframes y se garantizó la lectura correcta de la componente espacial. En algunos casos, estas componentes se encontraban aisladas en características, es decir, la latitud y longitud definían de manera individual atributos del dataframe y no un objeto geométrico, por lo que se procedió a fusionarlos de manera correcta.

Una vez que las sábanas fueron leídas como un geodataframe, el siguiente paso fue validar el sistema de referencia de coordenadas que se utilizaba, ya que era necesario unificar dicha característica para poder realizar cruces a nivel espacial dentro de las diferentes fuentes. Para esto se eligió usar el EPSG:4326 como sistema de georreferencia, el cual utiliza la latitud y longitud para definir ubicaciones en la superficie terrestre; también es conocido como WGS 84, que significa Sistema Geodésico Mundial de 1984 [34]. La elección de este sistema se fundamentó en dos argumentos: en primer lugar, porque es ampliamente utilizado en sistemas de información geográfica (GIS) y aplicaciones de cartografía, lo que lo hace una buena elección al trasversalizar diferentes aplicaciones en el área; en segundo lugar, se decidió utilizar este sistema porque la fuente de origen de los datos asociados a hurtos a personas lo utiliza en la georreferenciación de los hechos, por lo que se consideró importante utilizar dicha característica de nuestra fuente de datos principal.

Para la sábana de datos obtenida con la API de google maps, simplemente se debió unificar sus componentes de latitud y longitud, para luego leerla usando un geodataframe.

Con cada una de nuestra sábanas de datos con sus componentes geograficas definidas y en un sistema de referencia estandar, estás, se encontraban listas para continuar con el proceso de unificación. De forma posterior veremos como el proceso que se acaba de detallar puede ser logrado utilizando la arquitectura de nuestro datalake de la figura (33) y el procesamiento descrito en la sección con el mismo nombre, donde los jobs de glue juegan un papel fundamental en el paso entre zonas del datalake y las trasnformaciones necesarias.

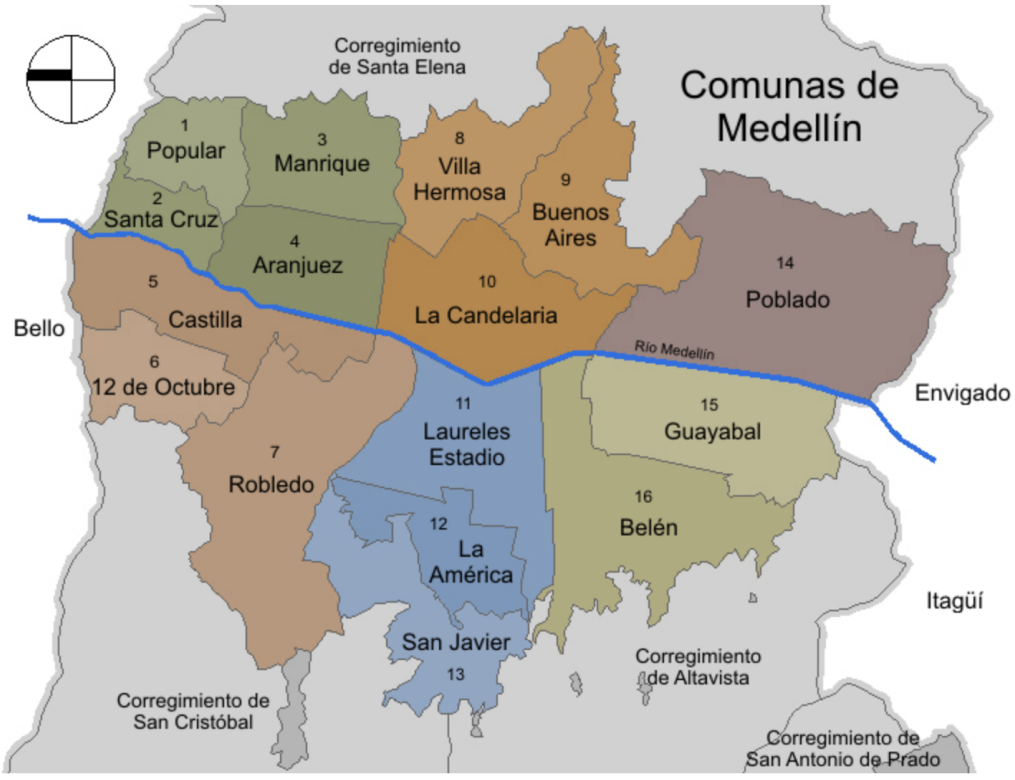
De forma más detallada este proceso puede consultarse en el notebook del repositorio llamado *analisis de otras bases*.

8.3 Creación de bases de datos

Teniendo listas cada una de las sábanas de datos asociadas a las fuentes, el siguiente paso era buscar la manera de cruzarlas entre sí de forma tal que se cumpliera el primer objetivo específico propuesto en este trabajo, la construcción de una base de datos adecuada para el estudio del fenómeno.

Para esto, inicialmente se trabajó con la sábana de registros de hurtos, en la que se identificó la necesidad de tener una clave dentro de las características trabajadas que permitiera diferenciar de manera inequívoca cada uno de los barrios en la ciudad. Después de un análisis de la unicidad en las diferentes características, se encontró que la combinación del índice asociado a la comuna por un lado y al del barrio por el otro, permitían definir una clave única para cada una de las zonas de nuestra subdivisión espacial del municipio. La lógica detrás de esto radica en el hecho de que las comunas son nombradas no solo por un nombre, sino también por un número que la representa (ver figura 10)[35], lo cual se traslada a la nomenclatura de los barrios, en los cuales tenemos una numeración dentro de la comuna, además del título de conocimiento común.

Figura 10: Comunas del municipio de Medellín



Luego de tener definida la llave, se realizó un agrupamiento por cada una de estas y posteriormente

un conteo, fue así que se obtuvo el número de casos de hurtos por barrio en el periodo estudiado, es decir, entre el 2018 y el 2022. De esta manera, para el caso de la comuna $1 - Popular$, se obtuvo los valores en la tabla 3.

Tabla 3: Conteo de casos entre el 2018 y 2022 para la comuna 01-Popular

Barrio	**Cantidad**
01_01	360
01_02	89
01_03	294
01_04	148
01_05	148
01_06	138
01_07	86
01_08	35
01_09	15
01_10	34
01_11	73

Se creó un identificador similar para el shapefile usado, lo que permitió cruzar los resultados anteriores con sus respectivos barrios. De esta manera, se obtuvo una nueva característica asociada con el número de casos por división geoespacial, lo cual se puede apreciar en la figura 11.

Figura 11: Resultado parcial del cruce entre el shapefile de los barrios y el conteo de hurtos

	OBJECTID	CODIGO	COMUNA	BARRIO	NOMBRE_BAR	SECTOR	INDICADOR_	NOMBRE_COM	SHAPE__Are	SHAPE__Len	geometry	key	casos
0	6754	1305	13	05	Metropolitano	4	U	SAN JAVIER	101866.863770	1537.700241	POLYGON ((-75.60795 6.26491, -75.60781 6.26482...	13_05	72
1	6755	0701	07	01	Universidad Nacional	2	U	ROBLEDO	490437.147949	2838.152470	POLYGON ((-75.57688 6.26663, -75.57672 6.26666...	07_01	342
2	6756	0510	05	10	Tricentenario	2	U	CASTILLA	421343.229492	2922.008690	POLYGON ((-75.56621 6.29586, -75.56619 6.29586...	05_10	661
3	6757	1511	15	11	La Colina	6	U	GUAYABAL	689537.432617	4825.014239	POLYGON ((-75.58774 6.20276, -75.58819 6.20163...	15_11	587
4	6758	1113	11	13	El Estadio	4	U	LAURELES	365142.314453	2636.136878	POLYGON ((-75.59328 6.26189, -75.59324 6.26188...	11_13	1862

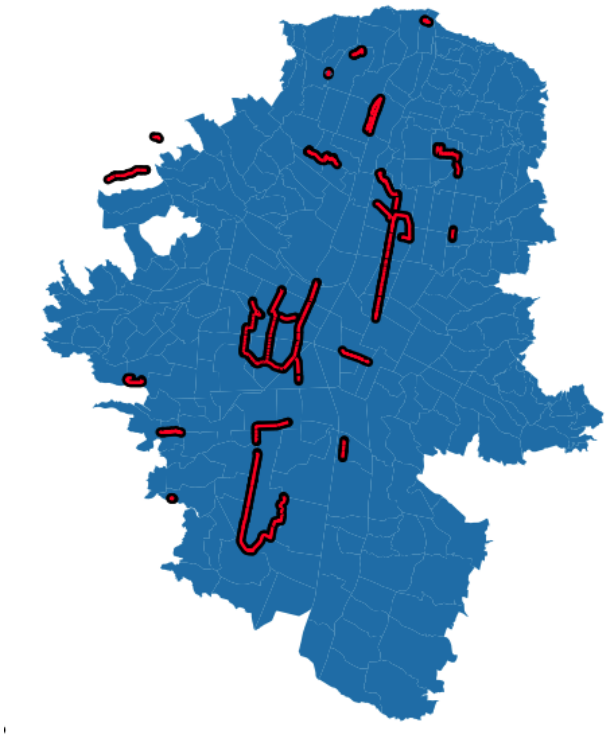
El procedimiento, descrito previamente, realizado con la fuente de casos de hurtos y el shapefile fue definido y posteriormente llevado a cabo con cada una de las bases que contenían registros georreferenciados de forma puntual. Para lograr esto, se creó una función denominada **countPunts**, la cual se encuentra en el notebook *Análisis de otras bases*, permitiendo llevar a cabo un conteo de características por barrio y agregarlas como nuevas características en el shapefile. Al finalizar este proceso se obtuvieron alrededor de 27 nuevas características en el geodataframe asociado a la subdivisión por barrios.

Se descartaron las fuentes cuya georreferenciación se daba a través de polígonos, tanto porque en general tenían pocos registros, lo que resultaba en una falta de relevancia en la característica asociada, como porque no eran pertinentes en nuestro análisis. Un ejemplo de esto fue la fuente de los usos rurales, cuyas características residían principalmente en los corregimientos, los cuales no fueron considerados para el análisis del hurto en este caso.

En las fuentes que tenían georreferenciación de tipo linestring, como en el caso de las ciclorrutas, se empleó una estrategia diferente. Se buscó determinar cuántos de esos objetos pasaban como máximo a 100 metros de cada barrio. En otras palabras, en el caso de las ciclorrutas, nos preguntamos cuántas ciclorrutas pasan por cada barrio o están alejadas de estos como máximo a 100 metros. La decisión de permitir que estuvieran fuera del barrio, pero cercanos en distancia, se debió a considerar que las dinámicas que se dan en lugares como vías principales, rutas de autobuses o rutas de ciclistas, pueden extenderse a su alrededor. Por tanto,

no se considera una idea ilógica. Para esto, se definieron buffers de 100 metros alrededor de los objetos de este tipo, tal y como se puede apreciar en la figura 12.

Figura 12: Ciclorutas mapeadas cada una con un buffer de 100 metros



En esta imagen se aprecia en color rojo las ciclorrutas y en negro un poco más amplio, los buffer asociados, donde recordemos, el buffer es una zona o área alrededor de un elemento geográfico que se crea dibujando una distancia alrededor de este elemento, lo que genera una nueva área geográfica [36].

Después de obtener la estructura en la figura 12, se procedió a verificar cómo se superponían los buffers entre los diferentes barrios utilizando la función *overlay* de Geopandas. Una vez identificado esto, se realizó una agrupación y un conteo por barrio de los linestrings que cumplían la condición impuesta. De esta manera, se agregaron nuevas características al geodataframe de los barrios relacionadas con el número de rutas de autobuses, ciclorrutas y ríos que pasaban por ellos o estaban cerca a una distancia de 100 metros.

Realizado el proceso definido en esta sección, se logra cumplir con el primer objetivo propuesto en nuestro trabajo de grado, el cual consiste en obtener la base de datos que nos permita realizar el análisis del hurto a personas en la ciudad de Medellín, en la cual, tenemos relación entre los barrios, sus características y el número de hurtos en el periodo estudiado. El análisis de este fenómeno se describirá en los apartados subsiguientes. De igual manera, en este punto, siguiendo el ciclo de vida de nuestro proyecto, hemos completado la recopilación y procesamiento de los datos.

Para un análisis más detallado de lo referido en esta sección remitirse a los notebooks *analisis de otras bases y cruce objetos lineales*.

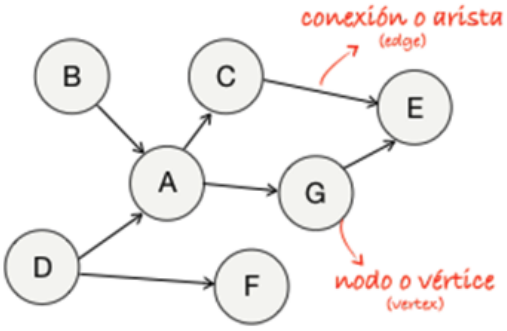
8.4 Modelado

El uso de modelos predictivos es de gran importancia en los contextos actuales, en los cuales los datos han adquirido un papel primordial en la sociedad. Estos modelos nos permiten hacer

predicciones basadas en datos históricos y otros factores relevantes asociados con las dinámicas estudiadas. Pueden resultar de gran ayuda en la toma de decisiones, en la optimización de procesos y en la precisión de las acciones a tomar, ya que nos permiten detectar patrones y tendencias ocultas que de otra forma podrían pasar desapercibidas.

Dentro de los modelos mencionados, encontramos los modelos asociados con grafos [20]. Estos utilizan un enfoque para modelar relaciones complejas entre objetos mediante el uso de grafos. Recordemos que un grafo es una estructura de datos que consta de nodos (también llamados vértices) y aristas (también llamadas bordes), donde los nodos representan objetos y las aristas representan las relaciones entre ellos, de forma más clara se puede observar esto en la figura 13.

Figura 13: Estructura de los grafos [37]

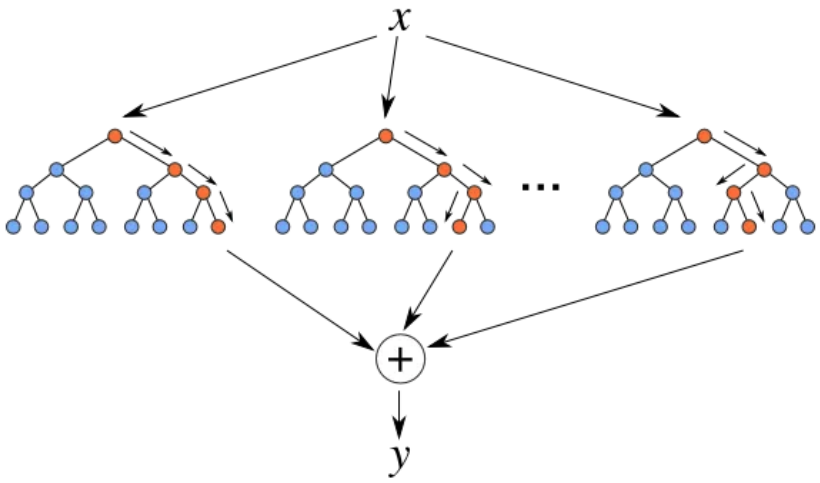


Los modelos de grafos se utilizan comúnmente para resolver problemas en los que la relación entre los datos es compleja y no se puede modelar fácilmente utilizando técnicas convencionales, como las redes neuronales simples o los modelos lineales. Estos modelos se pueden utilizar para resolver problemas de clasificación, regresión y agrupamiento, y se han aplicado en una amplia variedad de campos, tales como la recomendación de productos, el análisis de redes sociales y la biología molecular.

Dentro de los tipos de modelos de grafos utilizados en el aprendizaje automático se encuentra aquel que usa redes neuronales convolucionales y recurrentes, este extiende estas últimas, las cuales se utilizan de forma habitual para el procesamiento de imágenes y series de tiempo.

Otro modelo o algoritmo utilizado en el aprendizaje de máquina es el denominado *Random Forest* (bosque aleatorio en español). Este es un algoritmo del tipo ensemble que combina múltiples árboles de decisión en un solo modelo. Durante el proceso de entrenamiento, el algoritmo crea un conjunto de árboles de decisión, donde cada árbol se entrena en una submuestra aleatoria de los datos y en un conjunto aleatorio de características del conjunto de datos original. Luego, el algoritmo combina las predicciones de todos los árboles individuales para producir una predicción final. Esto puede ser apreciado en la figura 14.

Figura 14: Estructura de los bosques aleatorios [38]



El algoritmo *Random Forest* es útil en una amplia variedad de problemas de aprendizaje automático, como la clasificación, la regresión y la detección de anomalías. Este es de gran uso pues es de alta interpretabilidad, escalabailidad y robustez.

En esta línea se buscó implementar y entrenar modelos de aprendizaje de máquina, los cuales posteriormente sirvieron como puente para el despliegue en la nube utilizando técnicas de MlOps, logrando así cumplir uno de los objetivos establecidos al inicio. En particular, se eligieron los modelos mencionados al principio de esta sección, después de experimentar con algunos, como la regresión lineal y Xgboost, y encontrar que estos tenían el mejor rendimiento. El proceso detallado se explicará a continuación.

8.4.1 Modelo de grafos usando redes de Convolución Recurrentes

Buscando hacer uso de los modelos de grafos mencionados, se planteó un análisis temporal de los sucesos en la ciudad de Medellín durante el periodo de 2018 a 2022. Es decir, se estudiaron y analizaron los cambios y tendencias de los datos a lo largo del tiempo para cada uno de los barrios de la ciudad, utilizando ventanas de 4 semanas como división de nuestra serie temporal.

Para esto se planteó la estructura del grafo teniendo como nodos los diferentes barrios de Medellín. Se definió que solo estarían interconectados a través de aristas (figura 13) aquellos que fueran vecinos entre sí. De esta manera, se buscaba fundamentar y reflejar la idea de que las dinámicas asociadas a dicho fenómeno en barrios vecinos tienen comportamientos similares.

Inicialmente, para el planteamiento mencionado, fue necesario construir la matriz de adyacencia, la cual es una matriz cuadrada que representa todos los nodos (vértices) del grafo y las aristas (enlaces) que los conectan. En una matriz de adyacencia, las filas y columnas representan los nodos del grafo y cada elemento de la matriz representa si hay una arista entre los nodos correspondientes [39]. Es decir, representaremos las diferentes conexiones entre los barrios vecinos entre sí a través de una matriz.

Para la construcción de la matriz de adyacencia se utilizaron los métodos touches y overlap asociados con los geodataframes. Estos permiten encontrar tanto los objetos geométricos que se tocan como los que se solapan. Una vez encontrados, se validan cuáles son los vecinos asociados, en este caso con un polígono, lo cual es el primer paso para obtener la matriz de adyacencia. Una vez obtenido lo anterior, se procede a construir la matriz, que toma los diferentes barrios tanto en las filas como en las columnas y asocia un 1 en las intersecciones donde se encuentran vecinos.

Figura 15: Matriz de adyacencia para los barrios

	13_05	07_01	05_10	15_11	11_13	03_07	05_08	14_15	03_01	08_04	...	08_13	02_10	07_08	11_08	13_14	03_11	05_07	13_07	06_02	08_05
13_05	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	1	0	0
07_01	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
05_10	0	0	0	0	0	0	1	0	0	0	...	0	1	0	0	0	0	0	0	0	0
15_11	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
11_13	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
03_11	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
05_07	0	0	0	0	0	0	1	0	0	0	...	0	1	0	0	0	0	0	0	0	0
13_07	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
06_02	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
08_05	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

265 rows × 265 columns

Para nuestro caso, como se ha mencionado anteriormente, tenemos 265 barrios al interior de la zona urbana del municipio, por lo que nuestra matriz de adyacencia es una simétrica de tamaño 265×265 con valores de ceros y unos como se logra apreciar en la figura 15. Donde recordemos que el índice es una combinación del indicador de la comuna y el barrio.

Además, con el objetivo de validar la evolución de los hechos en el tiempo en ventanas de 4 semanas, se construyó un nuevo dataframe con una segregación de casos por semanas. Para lograr esto, se extrajo la información de la fecha de los hechos separando año y semana. Luego se creó un nuevo índice concatenando ambos valores, lo que permitió una unicidad en el índice y un posterior agrupamiento y conteo de los casos por barrio y semana entre 2018 y 2021. De esta manera, se obtuvo información detallada del número de casos por semana en cada barrio.

Para más detalle de la construcción de la matriz de adyacencia y el dataframe de conteos por semana consultar el notebook *matrix_adyacencia_count_semanas*.

Una vez calculada la matriz de adyacencia, el siguiente paso era crear los objetos que nos permitieran realizar un análisis a través de redes neuronales de convolución recurrentes, donde estas fueron elegidas debido a su capacidad para modelar secuencias de datos ya que sus conexiones recurrentes les permiten recordar información de las entradas anteriores y usarla para procesar las entradas futuras. Esto las hace muy efectivas en el modelado de secuencias de datos, donde las entradas están relacionadas en el tiempo [40].

Para tal fin se usó las implementaciones dadas por PyTorch Geometric, la cual es una biblioteca de Python que proporciona herramientas para el procesamiento de grafos y el aprendizaje profundo en estos utilizando PyTorch. Está diseñada específicamente para trabajar con datos de grafos y ofrece una amplia gama de operaciones y modelos para procesar, analizar y modelar estos [41].

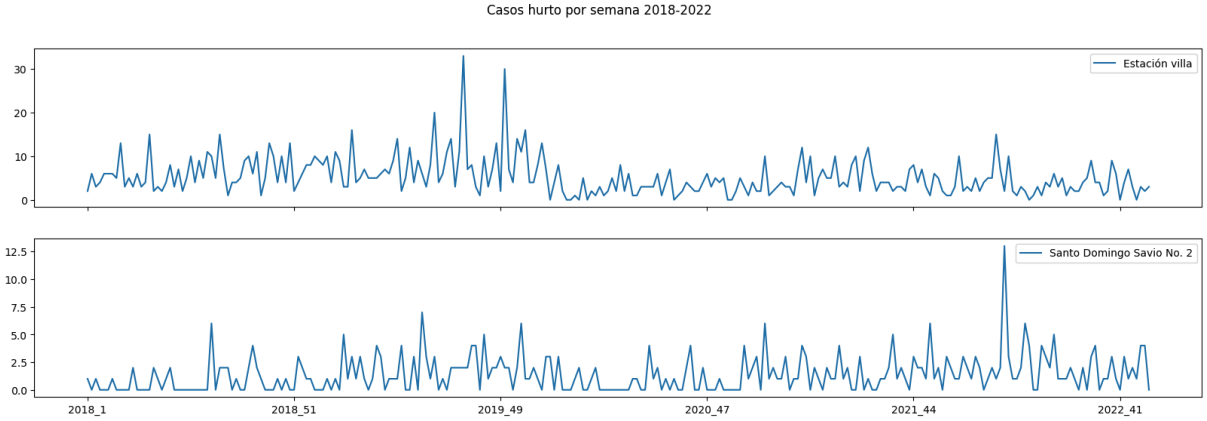
De esta manera, el primer paso fue crear un índice de aristas (edge index), que es una estructura de datos utilizada en el procesamiento de grafos para representar las aristas de estos. Se utiliza dos vectores de índices que representan los nodos (vértices) que están conectados por las aristas. Para crear este índice, partimos de nuestra matriz de adyacencia en la figura 15 y nos valimos de diferentes métodos tanto de torch como numpy. El resultado fue un arreglo de 2×1448 que representa las conexiones entre los nodos; en la figura 16 se aprecia parte de este.

Figura 16: Índice de aristas

```
edge_index

array([[  0,   0,   0, ..., 264, 264, 264],
       [  1,   2,   3, ..., 236, 238, 244]])
```

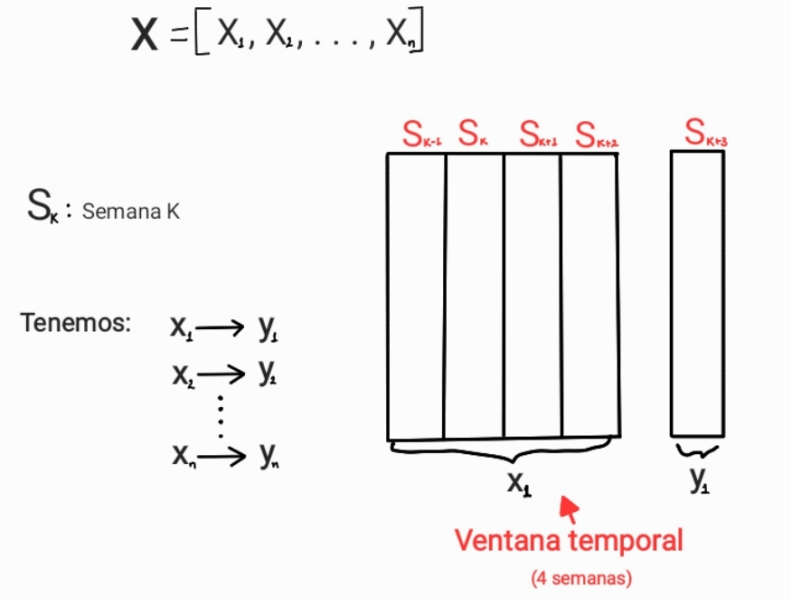
Según la cantidad de casos de hurto entre el 2018 y el 2022 para los barrios Estación villa y Santo Domingo Savio No. 2, tenemos que estos presentan un comportamiento oscilante en el tiempo con unos pocos picos que sobresalen del resto en dicho periodo; esto puede verse en la figura 17, a continuación.

Figura 17: Serie temporal asociada a los barrios Estación villa y Santo Domingo Savio No. 2 durante el periodo 2018-2022



Como se mencionó previamente, se decidió utilizar una ventana temporal en el análisis de los datos para tener en cuenta la dependencia del tiempo en la predicción del modelo. En este caso, se empleó una ventana de tamaño de 4 semanas, lo que requirió la creación de dos objetos, X y Y. El objeto X contenía las matrices con el conteo de hechos por cada 4 semanas, donde las filas representaban los barrios y las columnas las semanas. Por otro lado, el objeto Y contenía los valores correspondientes a la semana siguiente para cada elemento de X. Se puede observar en la figura 18 que dadas cuatro semanas $S_{k-1}, ..., S_{k+2}$, estas representan una matriz en X, y la semana S_{k+3} es su correspondiente vector en Y.

Figura 18: Representación de la ventana temporal usada



De esta manera, obtuvimos las dimensiones $X_{254\times265\times4}$ y $Y_{254\times265\times1}$, para nuestras cantidades tensoriales asociadas a los comportamientos de los hechos por semanas para cada uno de los barrios en el periodo estudiado. Con dichos arreglos se buscará hacer una predicción de cada

uno de los elementos en Y a partir de sus correspondientes en X. Es importante destacar que, en nuestro análisis, se encontró un intervalo sin datos entre la semana 48 y la 52 del año 2022, por lo que se decidió eliminar la última semana del año anterior y usar datos hasta la semana 48 del mismo, esto explica el por qué de la dimensión temporal tener un valor de 254.

En este punto se construyó una estructura que representara una secuencia de grafos estáticos que evolucionan en el tiempo, para esto se usó StaticGraphTemporalSignal de la librería PyTorch Geometric. Estos son estructuras tensoriales con una de las componentes temporales, es decir, diferentes instantes en el tiempo, donde cada entrada en dicha dimensión representa un grafo estático en un momento específico, para nuestro caso representa una fotografía en 4 semanas de los hechos de hurtos asociados a cada barrio de la ciudad de Medellín (figura 18). Esta construcción se eligió ya qu es útil para aplicaciones en las que se desea modelar cómo un grafo cambia a lo largo del tiempo, como en el análisis de redes sociales o en la dinámica de sistemas complejos.

Es de anotar que en la estructura mencionada en el parrafo anterior se usaron pesos iguales para las diferentes aristas. Podría considerarse en trabajos futuros variar dicho valor en relación con alguna característica asociada a las dinámicas del hurto.

Una vez se propició el escenario a nivel de objetos, se buscó construir un regresor usando redes neuronales de Convolución Recurrentes, la cual me permitiera predecir el número de casos semanales de hurtos. De forma previa se hizo una división en 80% para datos de entrenamiento y el restante para datos de testeo.

Se definió una arquitectura de red neuronal de Convolución Recurrente para resolver esta tarea figura 19, basados en lo planteado en la documentación de pytorch para una tarea similar [41].

Figura 19: Red implementada

```python
import torch
import torch.nn.functional as F
from torch_geometric_temporal.nn.recurrent import DCRNN

class RecurrentGCN(torch.nn.Module):
    def __init__(self, node_features):
        super(RecurrentGCN, self).__init__()
        self.recurrent = DCRNN(node_features, 32, 1)
        self.linear = torch.nn.Linear(32, 1)

    def forward(self, x, edge_index, edge_weight):
        h = self.recurrent(x, edge_index, edge_weight)
        h = F.relu(h)
        h = self.linear(h)
        return h
```

La arquitectura de la red neuronal consta de una capa DCRNN (Red Neuronal Convolucional Recurrente Dinámica) y una capa feedforward. El modelo DCRNN combina dos técnicas de aprendizaje profundo: redes neuronales convolucionales (CNN) y redes neuronales recurrentes (RNN). La red neuronal convolucional se utiliza para aprender características espaciales de los datos, mientras que la red neuronal recurrente se encarga de aprender las relaciones temporales y dinámicas entre estas características a lo largo del tiempo. Finalmente, la capa feedforward transforma estas características en una representación de mayor nivel y más abstracta, que se utiliza en la regresión.

Nuesto modelo, al tomar ventanas de cuatro semanas, se definió un número de nodos igual a ese valor en el momento de instanciar la red neuronal. El entrenamiento se dio a través de 200

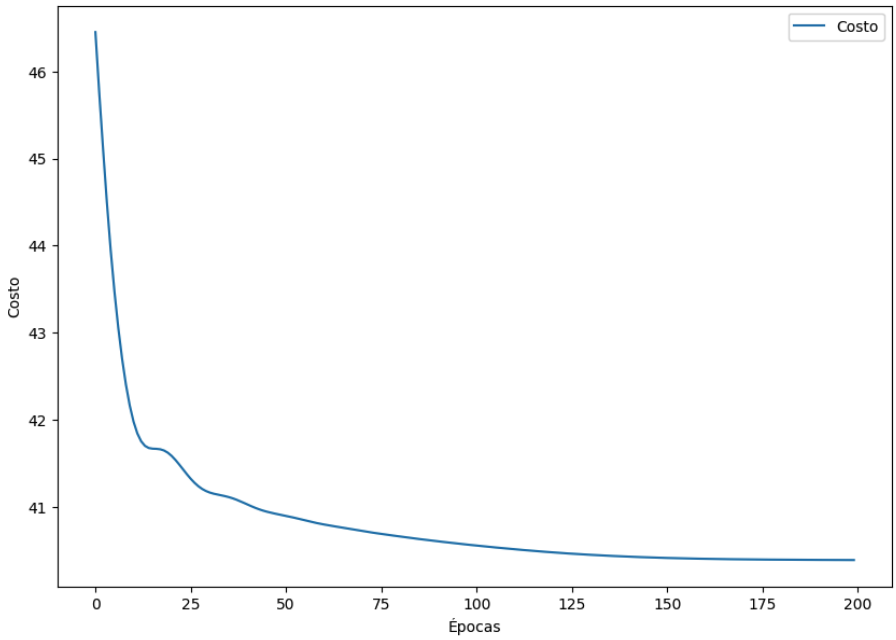
épocas y se uso un optimizador Adam [42] con una tasa de aprendizaje de 0.01. Esto puede ser apreciado en la figura 20.

Figura 20: Optimizador Adam

```
from tqdm import tqdm

model = RecurrentGCN(node_features = 4)

optimizer = torch.optim.Adam(model.parameters(), lr=0.01)

model.train()

for epoch in tqdm(range(200)):
    cost = 0
    for time, snapshot in enumerate(train_dataset):
        y_hat = model(snapshot.x, snapshot.edge_index, snapshot.edge_attr)
        cost = cost + torch.mean((y_hat-snapshot.y)**2)
    cost = cost / (time+1)
    cost.backward()
    optimizer.step()
    optimizer.zero_grad()
```

Bajo estas condiciones, al revisar el cambio obtenido en el costo durante el entrenamiento, se observa que efectivamente, en cada iteración, se logra una disminución de esta magnitud. Esto es lo que se esperaría en un modelo con una optimización adecuada de sus parámetros, ya que implica un mayor ajuste a los datos. Esta afirmación se encuentra respaldada en la figura 21.

Figura 21: Relación entre el costo del modelo y las épocas en el momento del entrenamiento



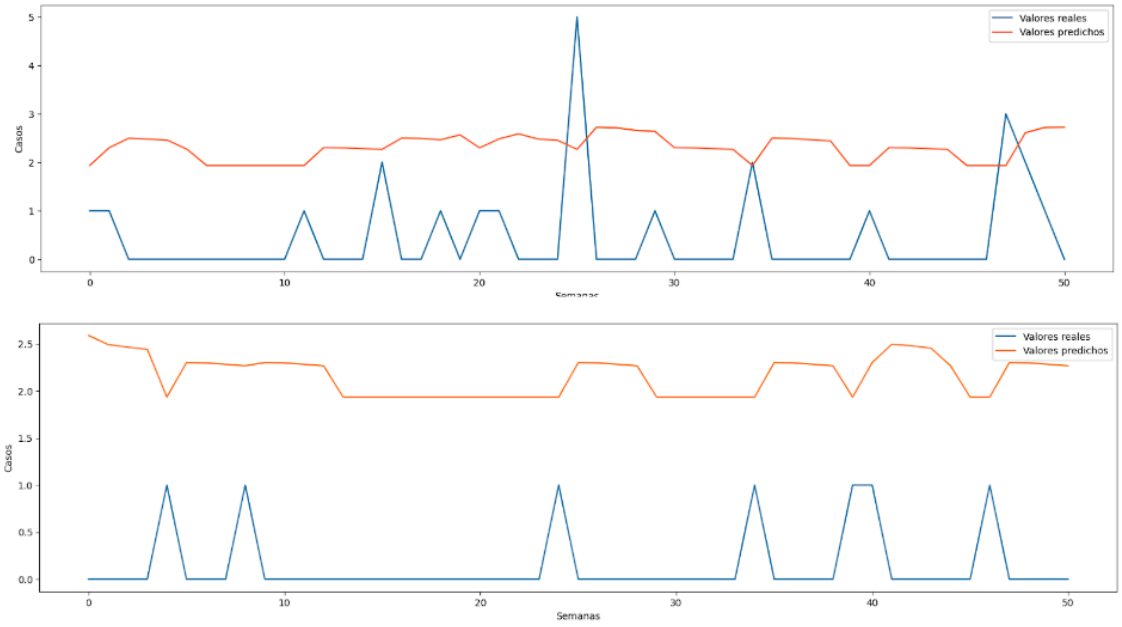
Luego de implementar el proceso de entrenamiento, al validar la eficacia de nuestro modelo en los datos de testeo se obtuvo el resultado en la tabla 4 para el Mean squared error.

Tabla 4: Resultado MSE para el modelo de grafos

Métrica	**resultado**
MSE	33.68

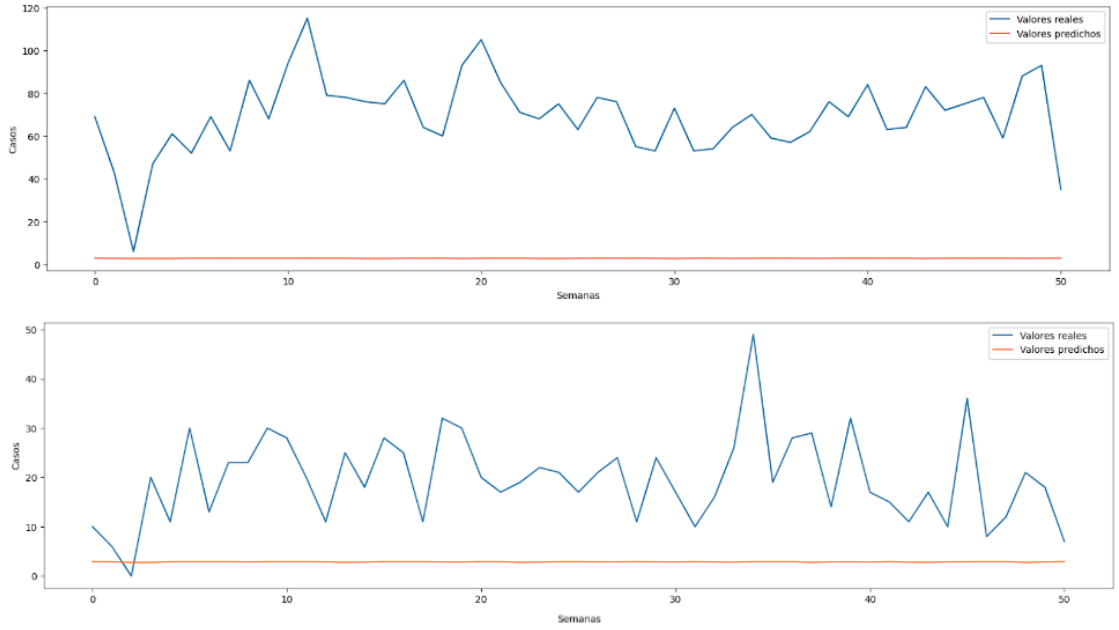
Este resultado no es muy diciente, por lo que se procede a realizar una análisis visual. Para algunos de los resultados de las predicciones por barrio encontramos que para aquellos en los que se tenía un número bajo de casos por semana, si bien la escala está en el orden de los valores reales, su comportamiento se aleja de lo encontrado en el fenómeno a lo largo del tiempo. Puede ser observado como algunos de los patrones en el comportamiento pueden predecirse, como se observa en la figura inferior en 22, donde en varios de los tramos los valles y comportamientos crecientes en los hechos fueron predichos, al igual que sus comportamientos cercanamente oscilantes. Sin embargo, ambos gráficos distan entre si, sin siquiera solaparse, por lo que podríamos decir que se debe considerar una alternativa diferente para la predicción usando el arreglo mencionado previamente.

Figura 22: Comparación valores predichos vs reales para el barrio La Rosa y la Cruz



Por otro lado, en barrios dónde el número de casos sobresale, encontramos que el modelo no logra hacer una correcta predicción de los hechos, inclusive fallando en la escala de estos. Esto se debe en parte al comportamiento átipico y acumulado de casos en estas zonas, con una alto desbalance al compararse con otros barrios en el municipio de Medellín. Este hecho se puede ver en la figura 23, donde tenemos el resultado de la predicción para los barrios la Candelaria y el Poblado, algunos de los de mayor densidad de casos de manera histórica.

Figura 23: Comparación valores predichos vs reales para el barrio La Candelaria y el Poblado



Como podemos observar, el modelo implementado tiene limitaciones al realizar predicciones basadas en el número de casos de hurtos a personas por semana en cada barrio de la ciudad. Por lo tanto, existen oportunidades de mejora y optimización futura para lograr resultados más precisos y cercanos a la realidad. En este sentido, es importante considerar también las características estáticas asociadas a las diferentes regiones (barrios) de la ciudad. Los resultados anteriores han servido de motivación para realizar un análisis predictivo desde diferentes perspectivas, que se detallan en la siguiente sección.

8.4.2 Modelo de bosque aleatorio

Conocer cómo las características territoriales influyen en la cantidad de hurtos en la ciudad es una inquietud que genera bastante interés en diferentes estudios de estas dinámicas delictivas en sus diversos contextos. En esta línea, el uso de modelos que permitan validar cuáles características influyen en la variable a predecir juega un papel importante en el análisis. Por esta razón, se eligió el modelo de Bosques Aleatorios, definido en Chaya ([38]), para predecir el número de casos de hurto en los diferentes barrios durante el intervalo de tiempo de 2018 a 2022, utilizando las características territoriales recopiladas y mencionadas en secciones anteriores. Este enfoque permitirá definir, basados en las características de un sector, si estas propician en mayor o menor medida la cantidad de hurtos en dicho lugar; similar a lo realizado por Ho en [10]. Es de anotar que en la fase de experimentación se validó el rendimiento de algoritmos de regresión como la regresión lineal y la regresión con xgboost, pero fue el modelo de Bosques Aleatorios el que presentó mejor rendimiento, de allí que sea el que se exponga a continuación.

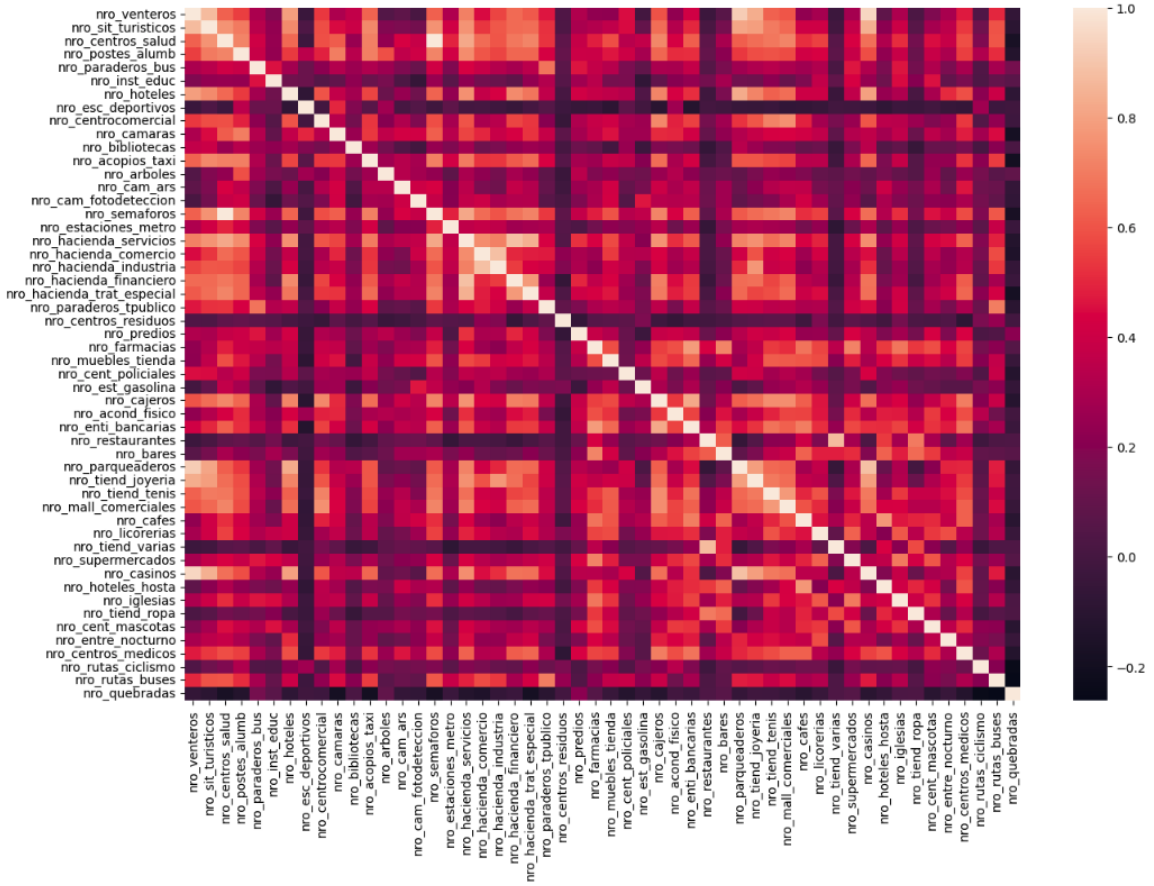
En un primer momento se tomaron como variables predictoras la totalidad de las características territoriales recopiladas y el número de casos de hurtos por barrio como variable a predecir. Seguido a esto, utilizando un proceso de optimización de parámetros usando la función Grid SearchCV de Scikit-Learn encontramos que para este conjunto de datos y el modelo de bosques aleatorios, los parámetros que mejor dan un ajuste a los datos son los mostrados en la tabla 5.

Tabla 5: Resultados iniciales al optimizar parámetros en el modelo de Bosque aleatorio

max_depth	max_features	n_estimators	R^2
7	auto	500	0.84

Usando dichos valores, encontramos un $R^2 = 0.84$, lo cual es una buen resultado para un regresión.

Figura 24: Gráfico de correlaciones entre las diferentes variables empleadas



Buscando una mayor precisión y una forma de depurar características, se graficaron las correlaciones entre las variables predictoras. Estas correlaciones se muestran en la figura 24. Guiados por los colores, podemos observar que existen variables que se correlacionan entre sí, por lo que fue necesario identificarlas y depurarlas, ya que la multicolinealidad puede causar problemas, adicional a esto tener variables innecesarias en nuestros datos puede generar costos añadidos en almacenamiento y procesamiento, especialmente si consideramos una tendencia a gran escala en la cantidad de los datos, por lo que optimizar dichos valores, generan ventajas a nivel de ingeniería de datos a grandes volumenes, de allí que en este caso se optara por dicho análisis, aún estando en una baja escala. Para esto, validamos cuales variables poseen, con otras características, una correlación mayor a 0.9 y las eliminamos en nuestro análisis. En este caso, aquellas que presentaron tal situación fueron:

- nro_semaforos
- nro_parqueaderos
- nro_casinos

Además de esto, se agregó una variable dummy como característica, resultado de tomar valores aleatorios de una distribución normal. El motivo de esta variable es que aquellas características cuya importancia en el modelo esté por debajo de esta, indican que no aportan lo suficiente en el modelo de predicción y pueden ser descartadas.

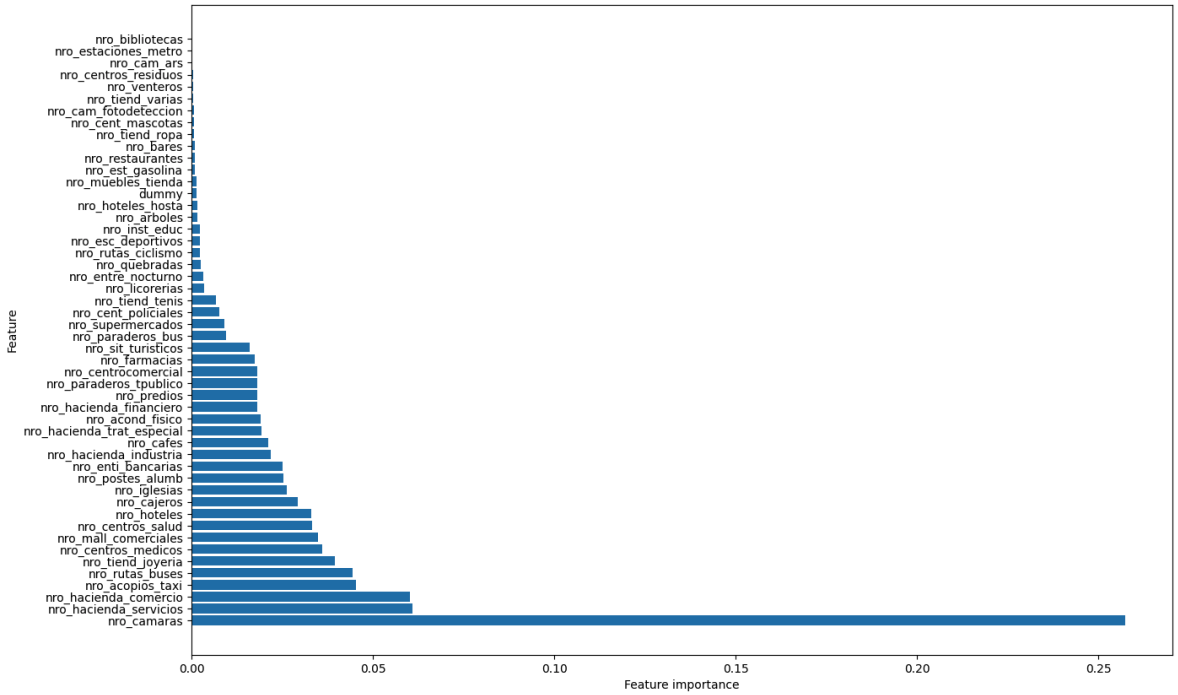
Al aplicar de nuevo el proceso de optimización de parámetros, descrito anteriormente, obtenemos como valores óptimos y de R^2 , lo que se muestra en la tabla 6.

Tabla 6: Resultados al optimizar parámetros en el modelo de Bosque aleatorio luego de depurar características

max_depth	max_features	n_estimators	R^2
None	auto	100	0.83

De nuevo obtenemos un buen resultado para nuestra regresión. Al validar la importancia de cada característica en el modelo encontramos lo representado en la figura 25.

Figura 25: Importancia de las características en el modelo luego de depuración



Allí se evidencia, al comparar con las demás, que por un lado características como el número de cámaras en un barrio son de gran aporte en el modelo, y por otro lado, el número de estaciones de gasolina no lo es tanto. Por debajo de nuestra variable dummy se ubicaron 13 características, las cuales fueron descartadas como parte de la depuración del modelo. Entre ellas encontramos las siguientes:

- nro_muebles_tienda
- nro_cent_mascotas
- nro_cam_ars
- nro_est_gasolina
- nro_cam_fotodeteccion
- nro_restaurantes
- nro_tiend_varias
- nro_estaciones_metro
- nro_bares
- nro_venteros
- nro_tiend_ropa
- nro_centros_residuos
- nro_bibliotecas

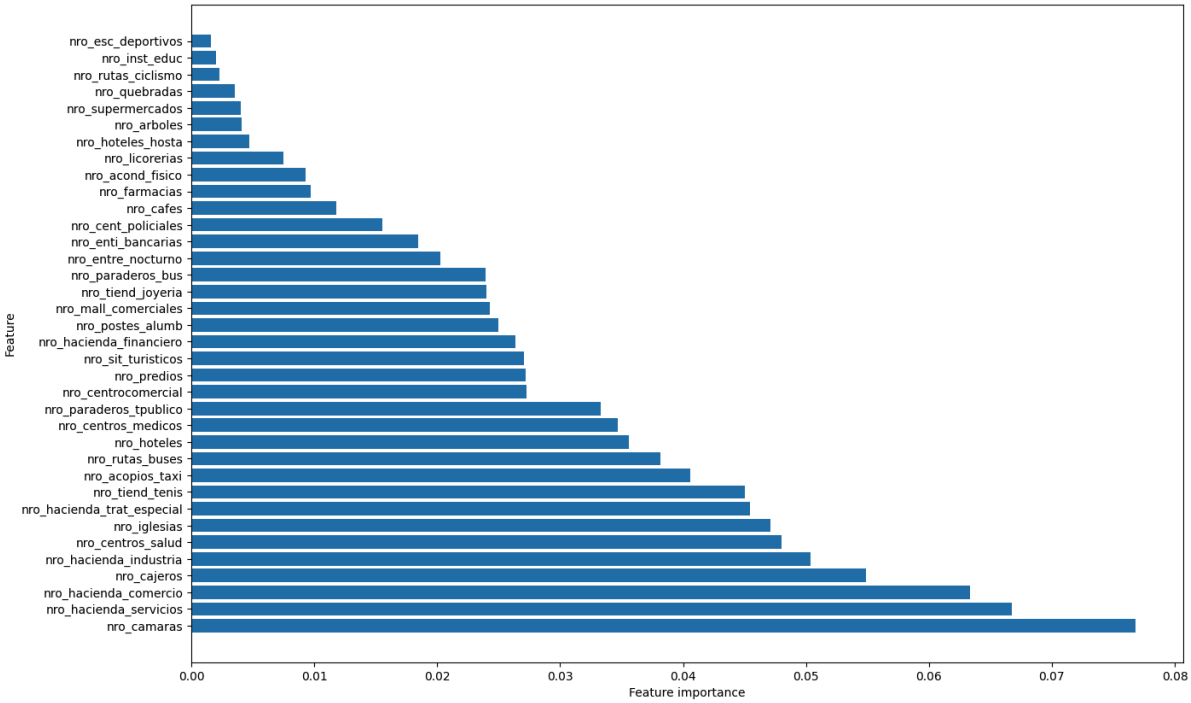
Al realizar de nuevo el proceso de optimización de parámetros y posteriormente el de entrenamiento, se obtienen los siguientes resultados en la tabla 7

Tabla 7: Resultados al optimizar parámetros en el modelo de Bosque aleatorio luego de depurar características usando la variable dummy

max_depth	max_features	n_estimators	R^2
None	sqrt	500	0.84

De nuevo, al guiarnos por el R^2, podemos afirmar que estamos teniendo un buen desempeño de nuestro modelo.

Figura 26: Importancia de las características en el modelo luego de depuración con variable dummy

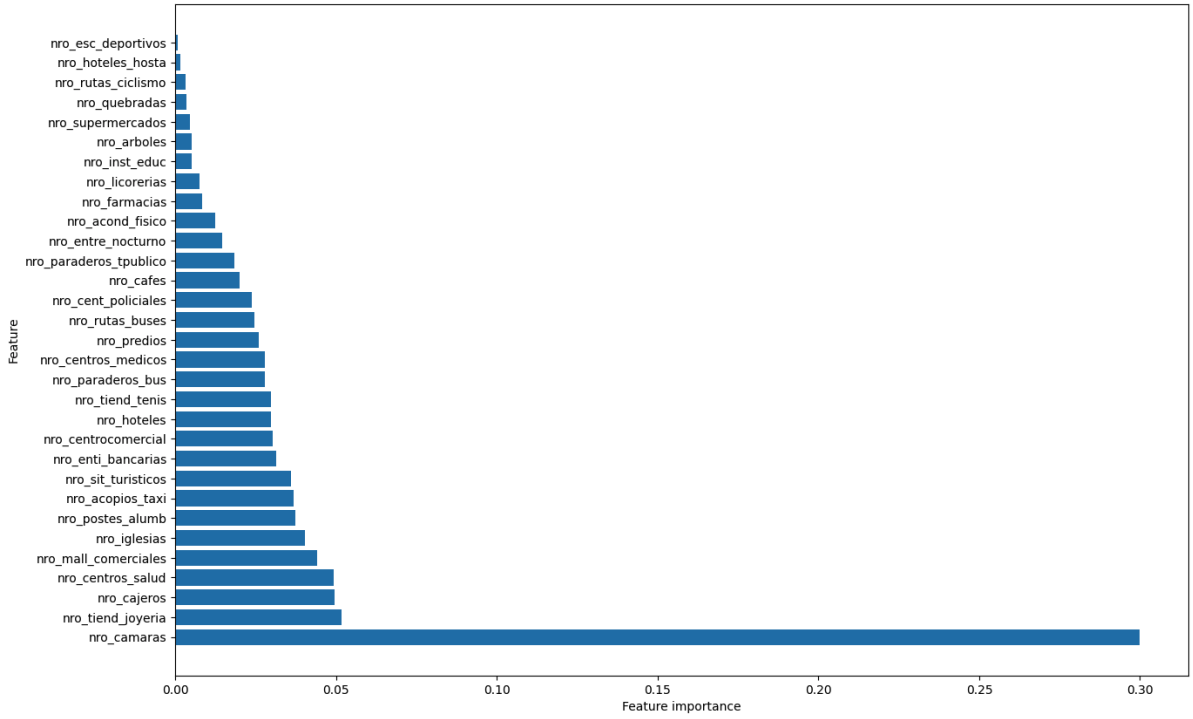


Como se puede observar en la figura 26, las características tienen una buena distribución en su aporte a la relevancia dentro del modelo, aunque sus valores estén en una escala baja. Por lo tanto, no se puede afirmar que una característica tenga una predominancia alta sobre las demás. Se puede evidenciar cómo las características provenientes de la fuente de hacienda, la cuales se relacionan con los registros de industria y comercio, tienen alta relevancia en el análisis. Sin embargo, a nivel conceptual presentan cierta imprecisión, ya que no se tiene certeza de qué comercios están involucrados en estas fuentes. Por lo tanto, eliminarlas es una buena idea para evitar duplicidades en los elementos territoriales. Al realizar el análisis sin estas componentes, se encontraron los resultados en la tabla 8.

Tabla 8: Resultados al optimizar parámetros en el modelo de Bosque aleatorio luego de depurar características de hacienda

max_depth	max_features	n_estimators	R^2
None	auto	500	0.82

Figura 27: Importancia de las características en el modelo luego de depuración variables de hacienda



Si bien en este caso, con respecto al experimento anterior, tenemos un valor menor para el R^2, sin embargo este aún tienen valores considerables, por lo que no se descarta la efectividad del modelo.

Podemos observar, en la figura 27, cómo las características aumentan en importancia con respecto a los casos anteriores, y variables como el número de cámaras cobran mayor relevancia. Una buena idea es validar cuál es el comportamiento de aquellas de mayor relevancia con respecto al número de casos y así determinar de qué manera influyen en estos según nuestro modelo.

Figura 28: Correlación entre las características de mayor importancia en el modelo y los casos de hurtos

nro_camaras	0.555323
nro_tiend_joyeria	0.831591
nro_cajeros	0.764569
nro_centros_salud	0.832970
nro_mall_comerciales	0.736732
nro_iglesias	0.478785
nro_postes_alumb	0.727106
nro_acopios_taxi	0.754874

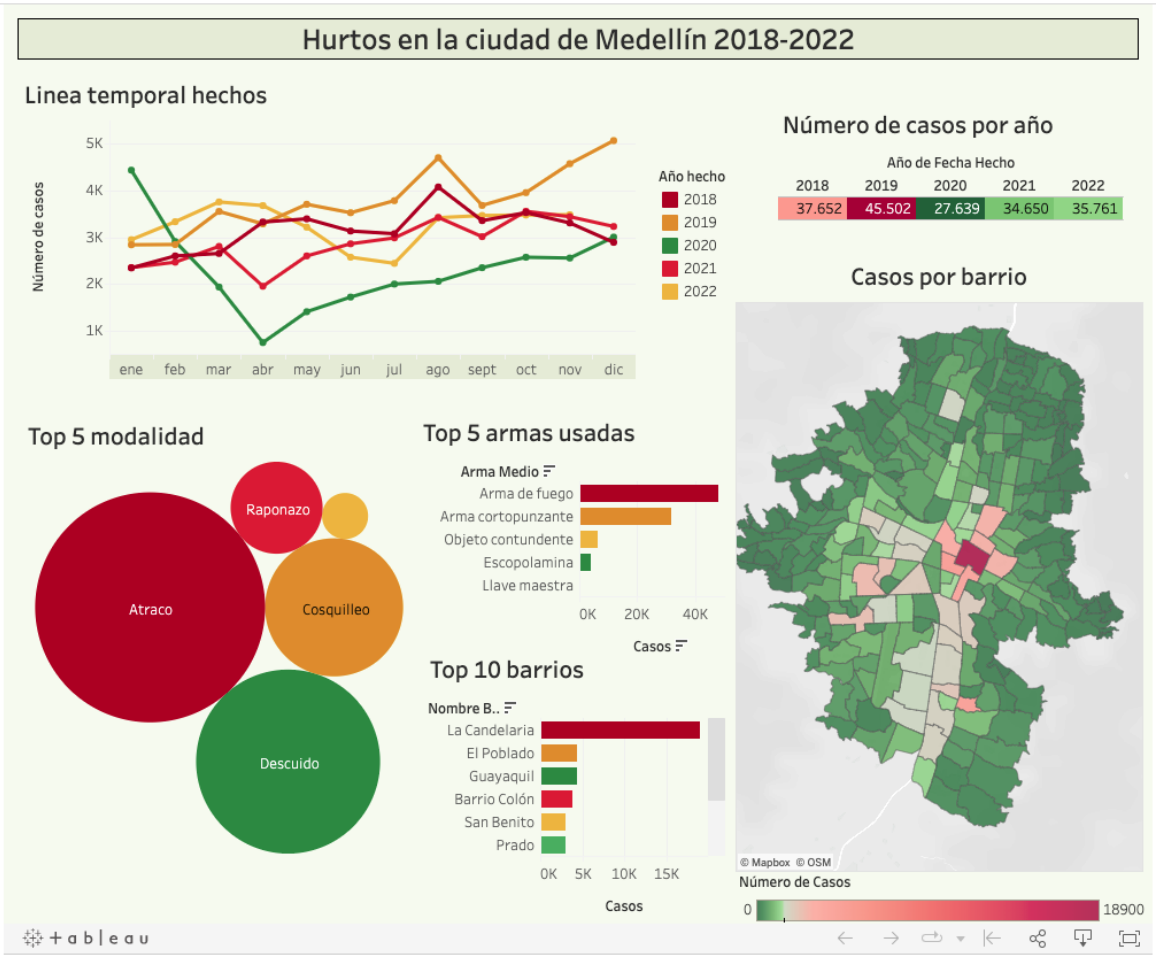
En la imagen anterior, figura 28, se puede observar la correlación de las características más importantes con respecto al número de casos, donde estas tiene un valor positivo, por lo que se espera que a medida que aumenten dichas características, los casos de hurtos también lo harán. Esto tiene sentido en el hecho de que estas características, en general, se relacionan con lugares de convergencia de personas, lo que aumenta el factor de oportunidad y propicia dichos delitos. Llama la atención el hecho de que el número de cámaras tenga una correlación positiva con el número de casos, siendo la característica dominante en la importancia dentro del modelo de bosques aleatorios. Esto implica que el número de cámaras no está cumpliendo su papel disuasivo ante el delito. Abordaremos este tema en la sección de resultados y conclusiones.

9 Tablero de visualización del hurto

Los tableros de visualización son importantes, pues proporcionan una vista rápida y completa de los datos clave en una organización o proyecto. Al reunir datos de múltiples fuentes y presentarlos en una única visualización, los tableros de información permiten a los usuarios obtener una comprensión clara de la situación actual, las tendencias y las áreas que requieren atención. A partir de estos, podemos tomar decisiones de manera ágil, analizar tendencias y medir rendimientos, lo cual nos ahorrará tiempo y esfuerzo.

Por las razones expresadas, buscando cumplir uno de los objetivos propuestos con este trabajo de grado, se construyó un tablero de visualización del hurto en la ciudad de Medellín a través de la utilización de la herramienta de uso común llamada tablau, software de análisis y visualización de datos que permite a los usuarios conectarse a diversas fuentes de datos, crear visualizaciones interactivas, tableros de información y reportes dinámicos. En la figura 29 se puede ver una imagen del tablero construido, el link a este puede encontrarse en los anexos de este escrito.

Figura 29: Vista general tablero de visualización del hurto

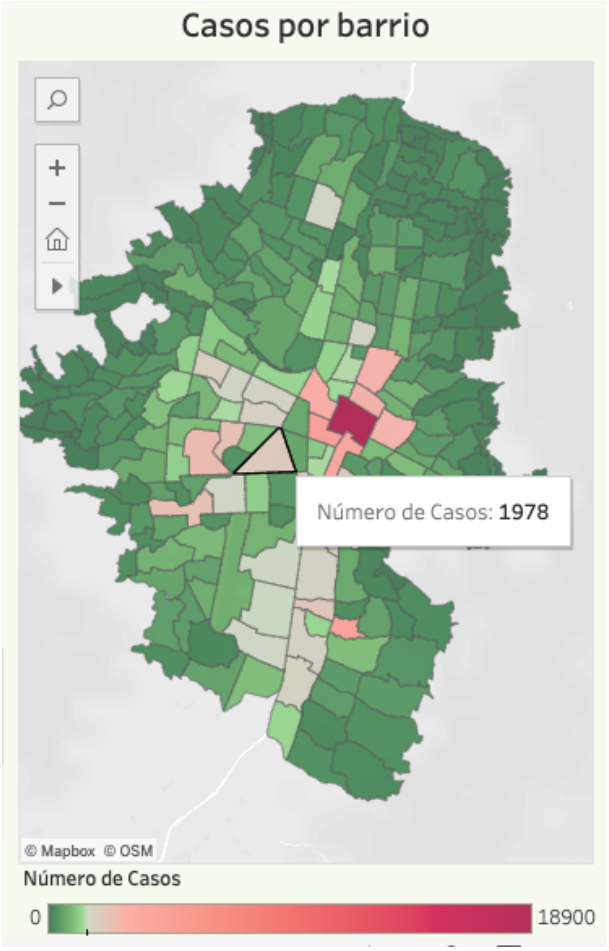


El objetivo de este tablero fue tener una herramienta que permitiera de manera ágil evidenciar la dinámica del hurto a personas en la ciudad de Medellín, es por esta razón que se buscó reflejar algunas de las características primordiales en esta tarea. Es así como inicialmente se definió poner las diferentes series temporales diferenciadas por año, como se observa en la parte superior izquierda de la figura 29, esto nos permite comparar la situación actual del hurto a personas con respecto a años anteriores, con lo que podemos generar hipótesis acerca de situaciones que se presenten en un momento definido del año, además de observar la evolución de esta variable

a lo largo del tiempo y detectar patrones, tendencias, estacionalidad, ciclos y anomalías en la frecuencia del hurto a personas.

Un dato importante para la toma de decisiones y la evaluación de políticas en un período de tiempo es aquel asociado con el acumulado de hurtos en lo que va del año y su diferencia con años anteriores. Es por esta razón que, en el diseño del tablero, se optó por incluir este dato en la parte superior derecha. Además, el número de casos de hurto segregado por barrios y comunas permite a las autoridades competentes definir dónde enfocar sus acciones y determinar si las dinámicas de un entorno propician los hechos delictivos. En línea con esto y buscando cumplir nuestro objetivo de definir los lugares con mayor densidad de hurtos en la ciudad, se decidió incluir un mapa interactivo del municipio a la derecha del tablero. Al hacer clic en cada uno de los barrios de la zona urbana de Medellín, se evidencia el número de casos que se han registrado allí durante el período estudiado. De manera directa puede observarse en la figura 30.

Figura 30: Mapa interactivo de casos



La pertinencia de este mapa es de carácter alto, pues no solo nos da información explícita sobre los casos, lo cual es importante en el análisis, sino que además, valiéndose de los colores, nos permite evidenciar la situación de la zona sin necesidad de tener una interacción directa con él. En la misma dirección, tenemos un gráfico de ranking, el cual nos muestra el top 10 de barrios con el mayor número de casos de hurto a personas reportados en la ciudad, lo que nos da una idea de la focalización del delito en Medellín.

Como se mencionó en la sección de análisis y exploración de las diferentes bases, características como la modalidad del hurto y las armas utilizadas en este brindan información valiosa acerca de la dinámica del fenómeno de hurto a personas en el municipio. Es por esta razón que se incluyeron dos gráficos que proporcionan información relacionada con estas dos características:

un ranking del top 5 de armas utilizadas y otro del top 5 de modalidades en los casos de hurto.

Figura 31: Datalake

Ambos gráficos mostrados en la figura 31 son de fácil lectura y cómodos a la vista, lo que lo hace relevantes y de fácil acceso ante la toma rápida de decisiones.

De esta manera, podemos afirmar que nuestro tablero de visualización de las dinámicas del hurto en Medellín posee las características más relevantes asociadas con el fenómeno, lo que lo convierte en una herramienta de gran utilidad para la toma de decisiones basadas en datos. Esta justificación respalda el uso de este tipo de herramientas en el análisis y la exploración de datos, lo que nos permite obtener una comprensión clara y profunda de la situación actual y las tendencias del fenómeno del hurto en la ciudad.

Los colores utilizados en esta implementación no fueron elegidos al azar, se buscó que las características con el mayor número de hechos asociados tuvieran un color que se asociara con el peligro, como el rojo, ya que esto representa una situación peligrosa para la población de la ciudad. Por esta razón, en los rankings y conteos, cuanto mayor sea el número de casos, más se acerca nuestra paleta al color rojo. Como base del tablero, se eligió un color en el espectro de los verdes, ya que se puede asociar con Medellín y su alta vegetación. Por lo tanto, en nuestra paleta de colores, cuanto más oscuro sea el verde, menos peligroso será, como un símbolo implícito de tranquilidad.

Tanto el fondo del tablero y del título, poseen un color suave y delicado en línea con nuestra paleta, esto para que dichos elementos no se roben la atención del usuario y puedan tenerla concentrada en la información de mayor relevancia.

De acuerdo a lo mencionado en esta sección y partiendo del objetivo del tablero de análisis y toma de decisiones por parte de los entes de control, se considera que cumple con las buenas prácticas del storytelling. Simplifica la información, utiliza gráficos acordes a lo que se busca transmitir, proporciona un contexto del fenómeno y utiliza una paleta de colores coherente con los objetivos y la dinámica reflejada. Por lo tanto, se considera que se ha cumplido con el cometido propuesto al inicio del trabajo de grado.

10 Arquitectura en la nube

En la actualidad, disponemos de varias herramientas que nos permiten diseñar e implementar rápidamente ecosistemas de almacenamiento y procesamiento de datos a gran escala, basados principalmente en sistemas de computación distribuida en la nube. Haciendo uso de estos

servicios, se buscó diseñar e implementar una arquitectura enfocada en datos que permitiera desarrollar el ciclo de vida de un proyecto en analítica, siguiendo los principios de MLOps [43]. Esto nos permitiría tener una referencia y un punto de partida para futuros trabajos enfocados en llevar modelos a producción. Para lograr este objetivo, se utilizaron diferentes servicios de proveedores de nube, especialmente Amazon Web Services (AWS), así como algunas herramientas de código libre como Apache Airflow y Great Expectation.

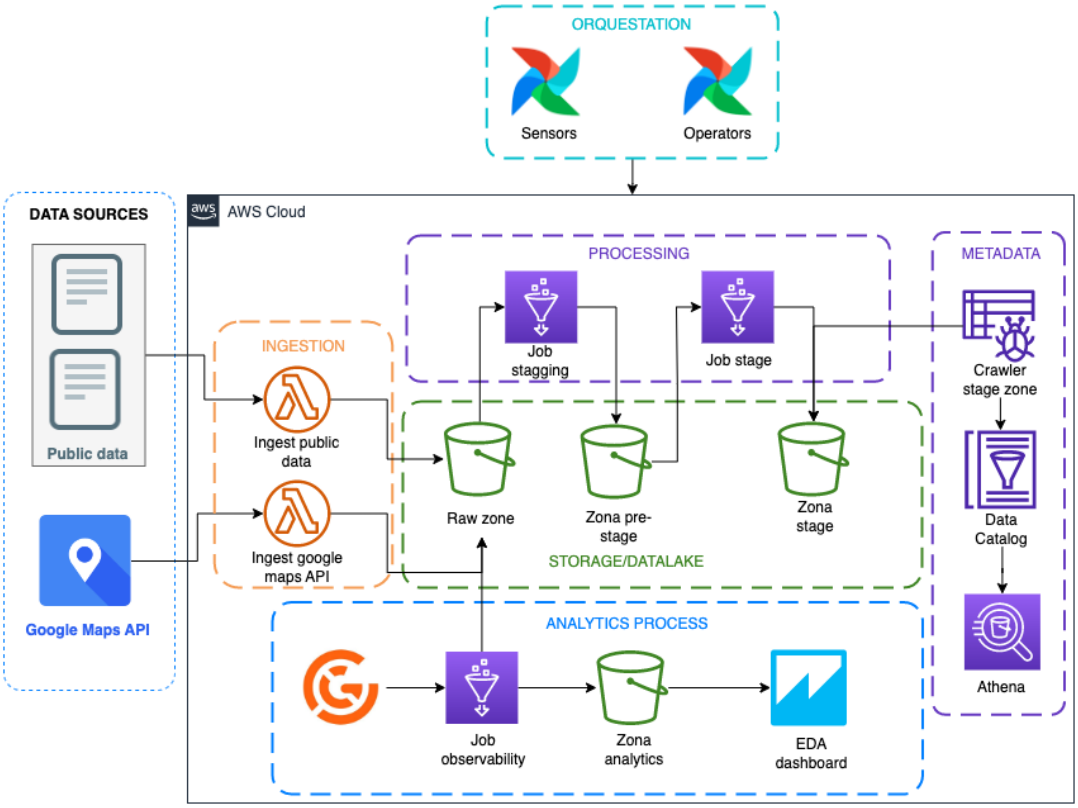
Nuestra arquitectura, la cual se muestra en la figura 32 puede entenderse incialmente en dos grandes fases, una inicial asociada a la ingesta y disponibilización de la data proveniente de todas nuestras fuentes en un datalake y una segunda fase encargada del procesamiento de esta data desde una mirada analítica, manteniendo lineamientos de acuerdo a las principios definidos por MLOPS [43].

A continuación se entrará en detalle en cada una de estas fases.

10.1 Datalake

En la primera fase de nuestro proceso buscamos implementar lo que comúnmente se conoce como un datalake[3]. Un datalake es un repositorio de datos que almacena tanto datos estructurados como semiestructurados y no estructurados. Utilizamos el datalake para almacenar los datos de origen, así como todos los datos resultantes de procesar y transformar la información original. De esta manera, podemos rastrear el ciclo de vida del dato y garantizar un buen funcionamiento y seguimiento del mismo.

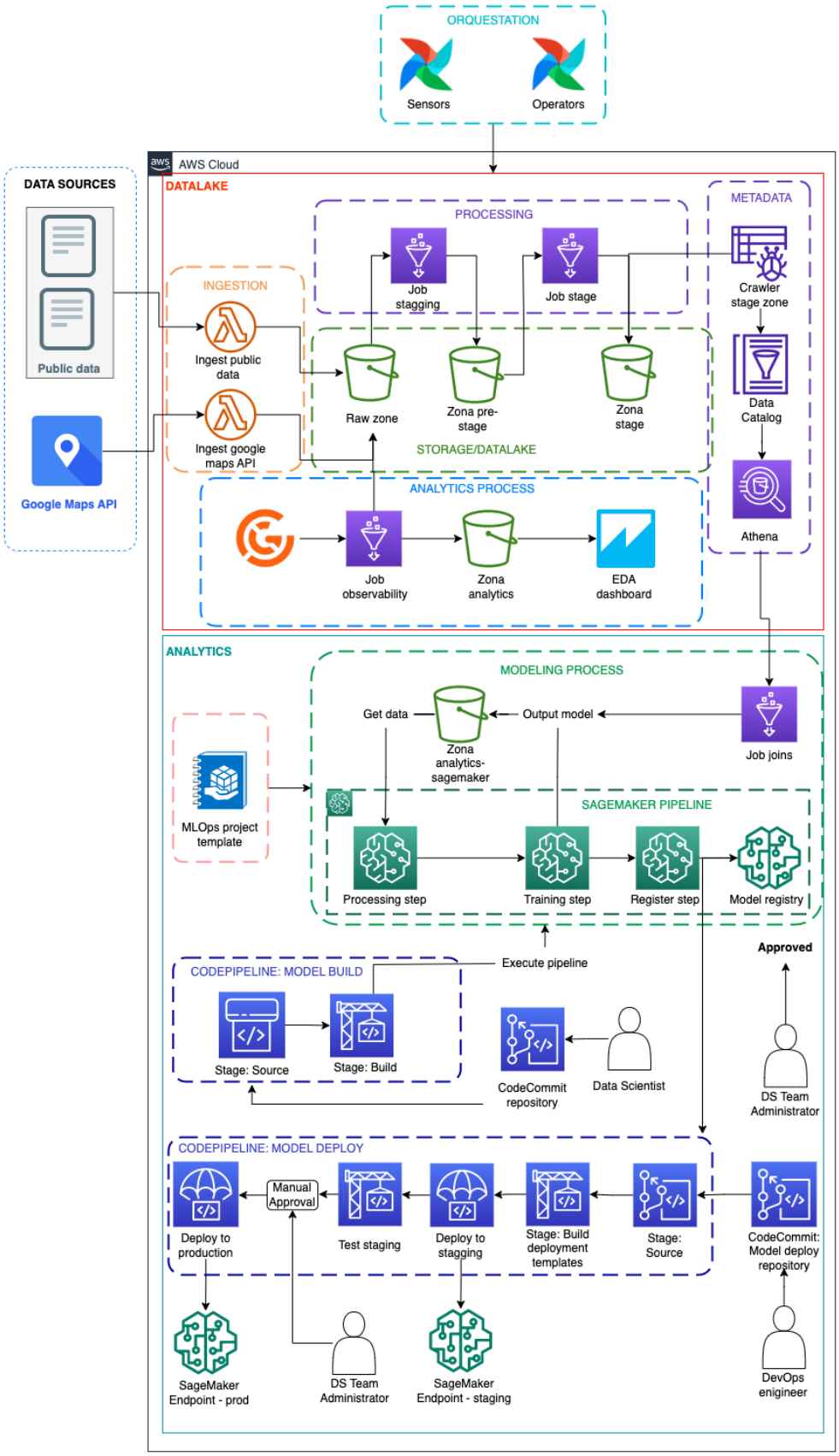
Figura 33: Arquitectura del datalake



Nuestro datalake, como se ve en la figura 33 y 34, fue planteado usando cuatro zonas al interior de este definidas de la siguiente manera:

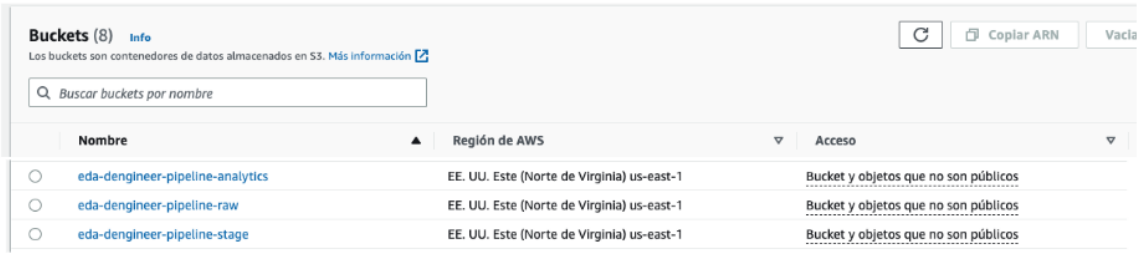
[3]Para obtener un contexto más detallado, remítase a [44]

Figura 32: Arquitectura implementada en el proyecto



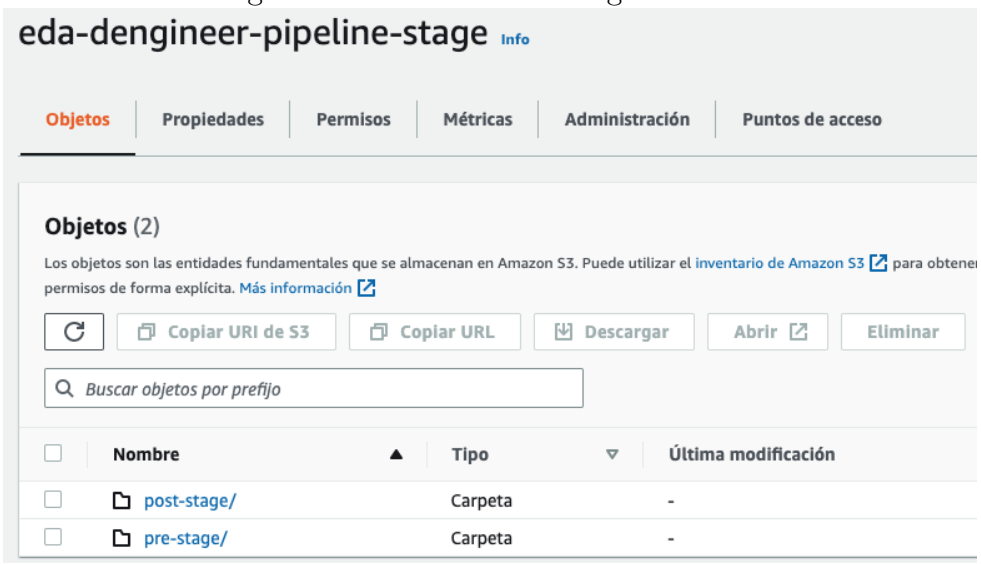
- **Zona Raw:** En esta zona se disponibilizaron los datos de manera cruda, es decir, tal cual se tomaban del origen.
- **Zona pre-stage:** La zona pre-stage fue destinada para poner los datos luego de un ligero preprocesamiento, es decir, datos provenientes de raw luego de una limpieza y estandarización inicial.
- **Zona stage:** En esta zona se disponibilizaron los datos procesados y con un formato adecuado para el procesamiento por terceros, el cual fuera óptimo en miras de la computación distribuída.
- **Zona analytics:** Esta zona fue destinada para la disponibilización de la información asociada con la observabilidad de los datos procesados en nuestro datalake, al igual que los modelos de sagemaker.

Figura 34: Zonas datalake s3



En la imagen anterior se puede ver la implementación de las zonas usadas en el servicio de s3, el cual se usó para nuestra construcción del datalake, donde para el nombramiento de los servicios se usó la sintaxis,*eda-dengineer-pipeline-*{*zone name*}, esto asociado con la exploración y análisis de la data en ingeniería de datos y los pipelines asociados. El bucket asociado con la zona stage (*eda-dengineer-pipeline-stage*)fue a su vez dividido en dos, de acuerdo a lo planteado; esto puede verse a continuación, en la figura 35.

Figura 35: División zona stage datalake



La data fue ingestada usando dos lambdas, una para aquella proveniente de los datos públicos y otra para el consumo de la API de google maps. La lambda asociada con la data de carácter

público (Ingest public data), tiene como objetivo la carga de los datos desde un repositorio local a la zona cruda del data lake. Donde el acumulado de sábanas de datos se encuentra en un repositorio centralizado pues son el resultado del rastreo de información en la web de carácter manual[4]. Para la segunda lambda (Ingest google maps API), tenemos que esta se encarga de generar peticiones a la API de google maps, recopilar los resultados y alojarlos en la zona cruda de manera tabular; el código de ambos desarrollos puede encontrarse en el repositorio asociado, en la carpeta de orquestación y api_google respectivamente.

En la sección de procesamiento, contamos con dos jobs de Glue, los cuales son responsables de trasladar los archivos desde las zonas del datalake. El primer job, denominado "pre-stage", se encarga de trasladar los datos desde la zona cruda a la zona de preparación ("pre-stage"), normalizar las columnas, limpiar los caracteres no deseados y eliminar las columnas innecesarias. El segundo job, llamado "stage", se encarga de mover los datos de la zona de preparación a la zona de almacenamiento final ("stage"), convirtiendo los archivos a formato Parquet. El formato Parquet es una opción de almacenamiento columnar optimizada para el procesamiento distribuido.

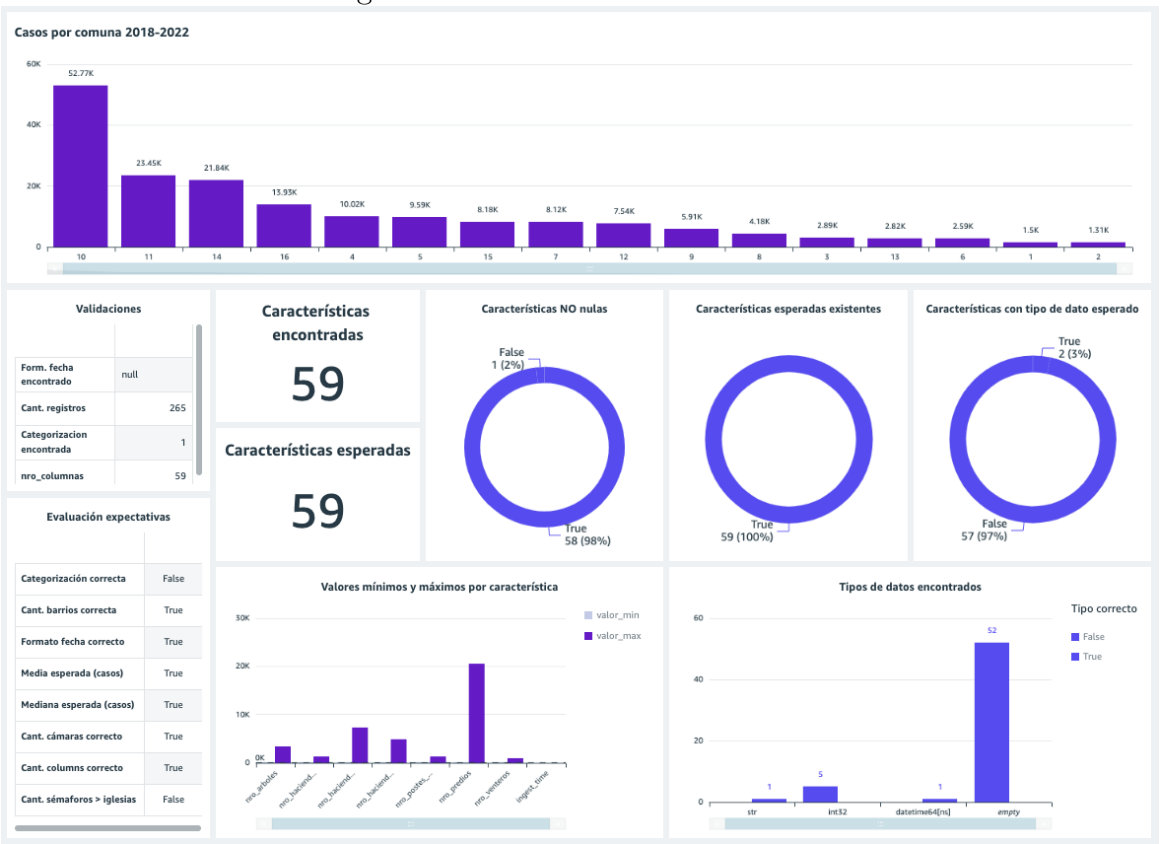
Figura 36: Jobs usados en el datalake

Una vez que todos los archivos se encuentran en la zona stage, se procede a activar los crawlers, los cuales nos permitirán alojar y poblar las tablas asociadas con cada archivo en el data catalog de Glue. De esta manera, podemos consultar y operar con la información de una manera más eficiente y rápida. Uno de los beneficios de este procedimiento es que la información puede ser consultada utilizando el servicio de Athena, además de tener un repositorio de metadatos asociado con nuestra información.

De forma paralela al procesamiento dentro del datalake y las zonas mencionadas anteriormente, los archivos que llegan a la zona raw son sometidos a un proceso de inspección de acuerdo a las expectativas definidas usando Great Expectations. Esto se realiza a través de un job de observabilidad, como se muestra en la figura 36, que luego de evaluar las expectativas, deposita los resultados en la zona analítica del datalake para que puedan ser consumidos por un tablero de validaciones. El tablero de validaciones contiene información asociada al tipo de dato de las columnas, el conteo de casos por comuna, entre otras. De esta manera, se pueden evidenciar de manera ágil anomalías en nuestros datos y tomar decisiones que permitan la correcta ingesta y posterior procesamiento de la información. A continuación, en la figura 37, se muestra una imagen del mencionado tablero.

[4]Para mayor detalle, revisar tabla anexa asociada con las fuentes

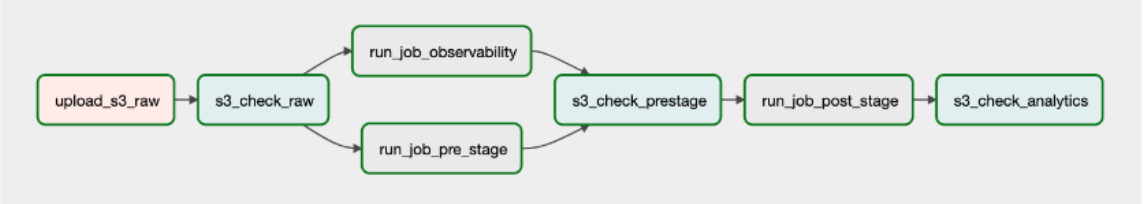
Figura 37: Tablero de observabilidad



Al tener un buen sistema de observabilidad, podemos detectar y solucionar rápidamente problemas, asegurarnos de que los datos se estén procesando correctamente y de que se estén cumpliendo los objetivos del proyecto. Como se puede observar en la imagen anterior, la construcción de este sistema tuvo en cuenta el número de características que se tenían en el conjunto de datos, el tipo de dato que se estaba procesando, se validó la nulidad de columnas y algunas restricciones que podrían tener los datos. Cada una de estas características del tablero entrega un alto valor en el proceso, las cuales al ser refinadas, permitirán tener una herramienta enfocada en garantizar la calidad, confiabilidad y eficiencia del proyecto de datos.

La orquestación durante esta parte del proceso fue realizada usando Apache airflow como se evidencia en la figura 33. Para esto se construyó un DAG (Directed Acyclic Graph) con la forma siguiente:

Figura 38: DAG orquestación datalake



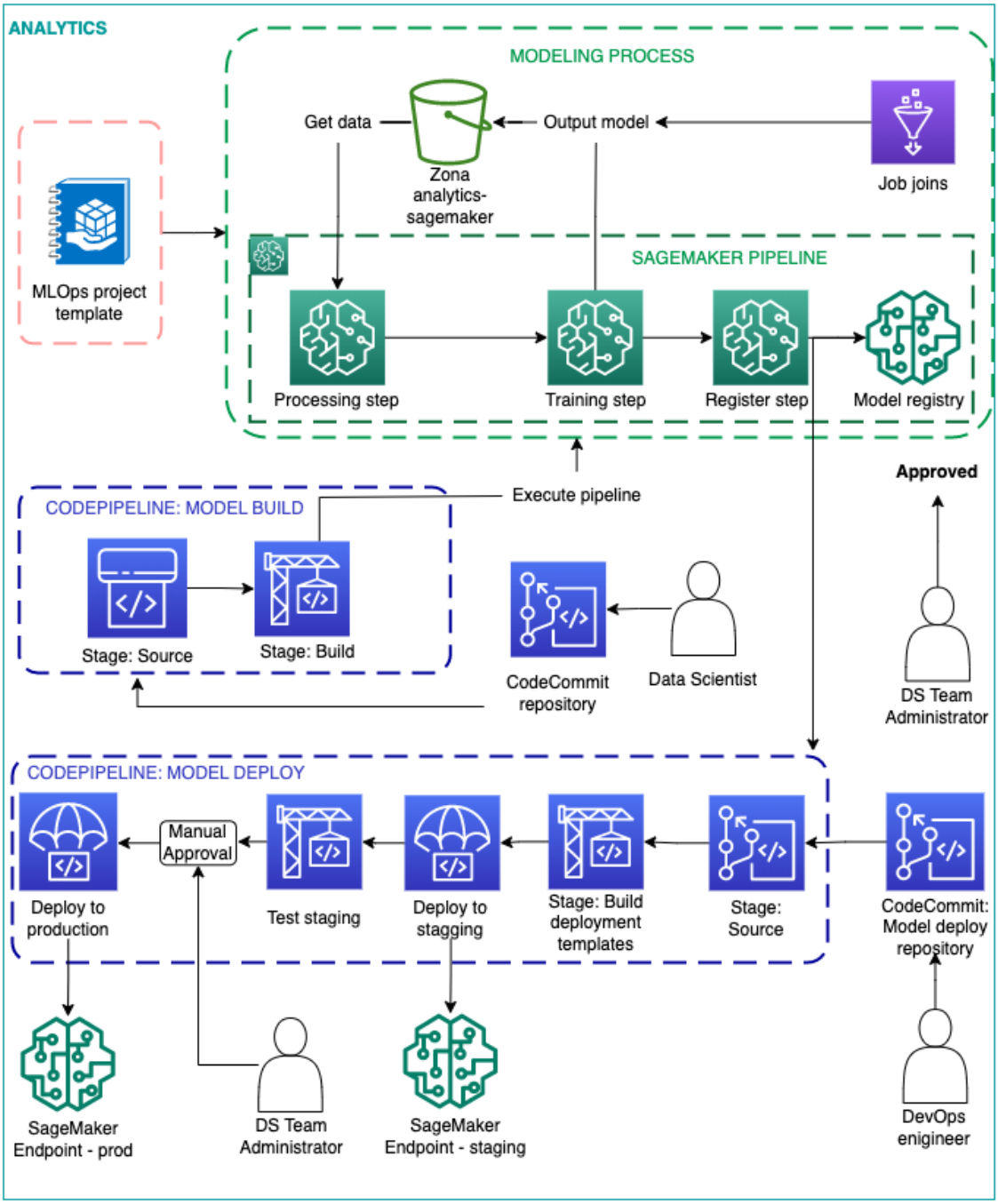
Esta herramienta permite tener un monitoreo de las ejecuciones del flujo de trabajo, por lo que es de gran utilidad ante el control y frecuencia de las ejecuciones, al igual que del orden en los pasos dentro de estas. En la figura 38 se ve que inicialmente se tiene una tarea que carga los datos, luego se comprueba que estos hayan llegado a la zona cruda del datalake, seguido a esto se ejecuta tanto el job de observabilidad como el de pre-stage, se valida que los resultados de

cada uno se alojen en el lugar correspondiente y finalmente se ejecuta el job de post-stage. De esta manera a grandes razgos se tiene una descripción de la primer etapa del proceso descrito en la arquitectura. Para los crawlers se realizan ejecuciones sincrónicas, lo cual permite optimizar los costos asociados a esto.

10.2 Analytics

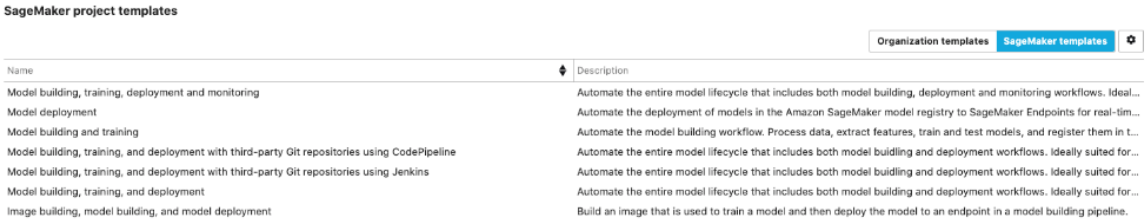
La segunda gran fase del proceso se dedicó a la componente analítica de este, donde una vez se tuvo una sábana de datos proveniente del cruce geospacial entre fuentes, se implementó el flujo de procesamiento, entrenamiento y registro de modelos usando sagemaker, como podemos ver a la figura 39.

Figura 39: Procesamiento analítico



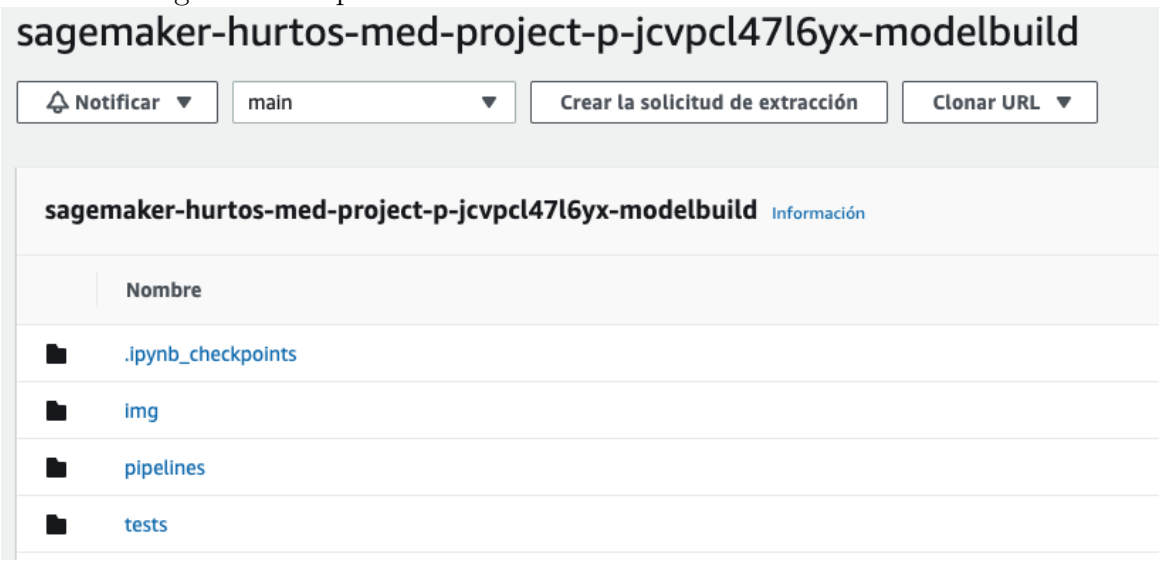
Para la construcción y despliegue de esta arquitectura se hizo uso de dos de los templates definidos por AWS para dicho fin (ver figura 40), acoplándolos de manera adecuada a nuestras necesidades. En particular, se utilizaron el *Model Building and Training* y el *Model Deployment*.

Figura 40: Templates de depliegue sagemaker



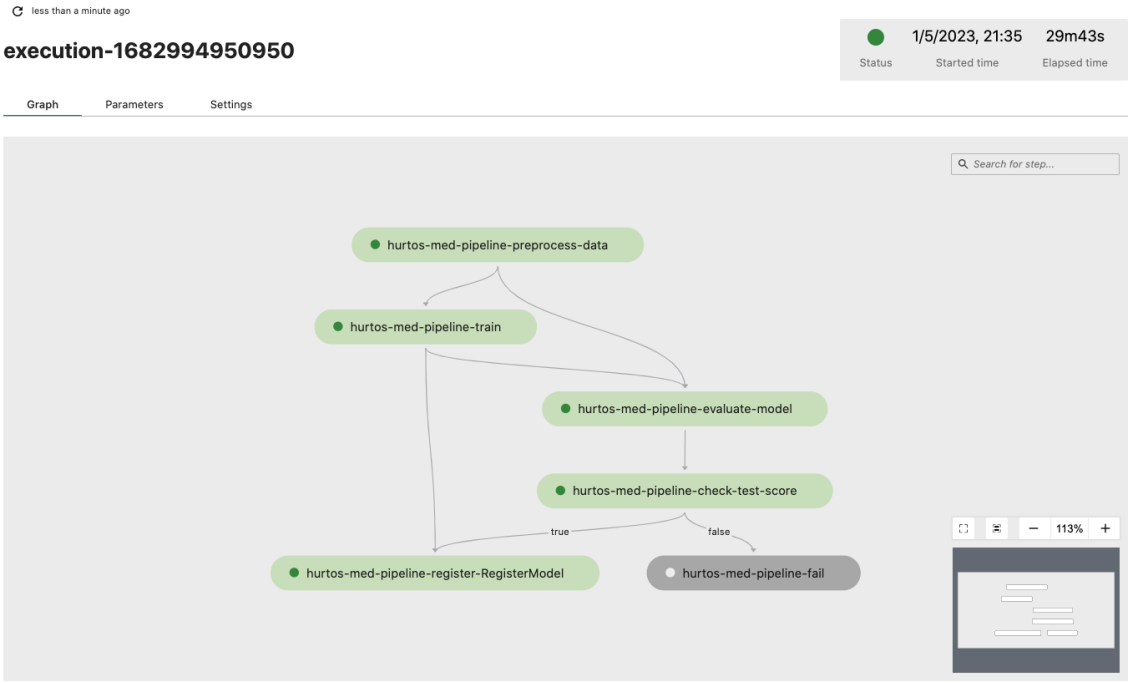
Inicialmente, se utilizó el template *Model Building and Training*, el cual permitió crear un repositorio en CodeCommit y un pipeline con dos etapas. En la Figura 39, en la sección *Model Build* y *Sagemaker pipeline*, podemos ver estas dos etapas. Con la primera se descargaba el código del repositorio en CodeCommit y, en la segunda, se creaba y ejecutaba el pipeline de SageMaker. El repositorio utilizado en esta parte se muestra en la figura 41.

Figura 41: Repositorio asociado al model build en codecommit



En el repositorio mencionado se agregaron los archivos en formato Python que definían el pipeline, el preprocesamiento y la evaluación del modelo. En el archivo asociado a nuestro pipeline, se incluyen no solo la instanciación de parámetros necesarios, como las rutas de S3, las librerías y las instancias a utilizar, sino también los punteros a los archivos de preprocesamiento, la definición del modelo y sus hiperparámetros para el entrenamiento y, la etapa de evaluación, al igual que las diferentes acciones a tomar según los resultados de la ejecución. Donde en caso de éxito, se registra el modelo. A continuación podemos ver gráficamente el resultado de una ejecución del pipeline de sagemaker para este proyecto (figura 42).

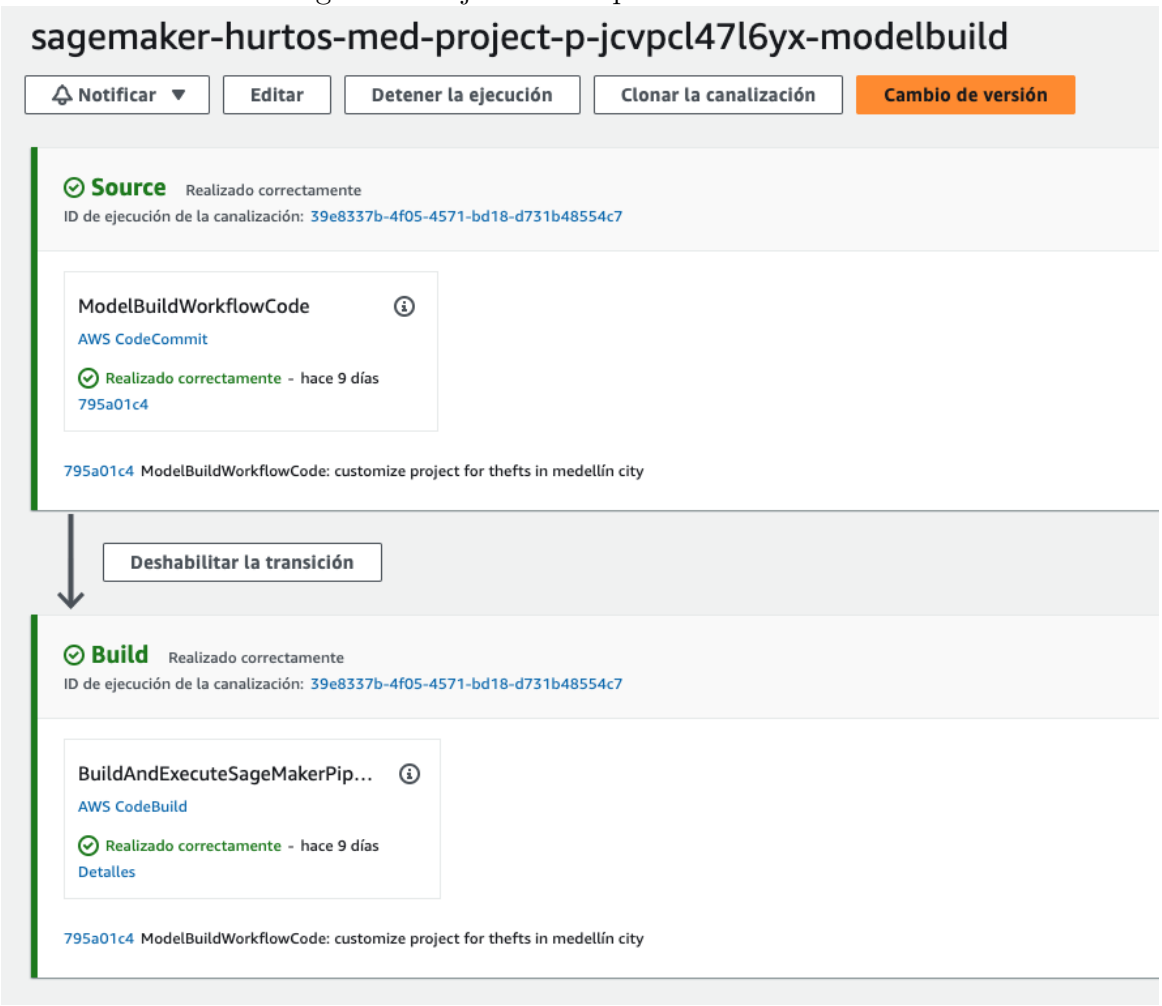
Figura 42: Ejecución Pipeline Sagemaker



En la ejecución anterior, se pueden observar en verde los resultados exitosos de cada paso en el pipeline ejecutado. Se inicia con el preprocesamiento de los datos, seguido por el entrenamiento y una evaluación del modelo. Se realiza una prueba para determinar si el resultado del modelo, en cuanto a la métrica definida, es mejor que el modelo previamente registrado. En caso afirmativo, se registra uno nuevo.

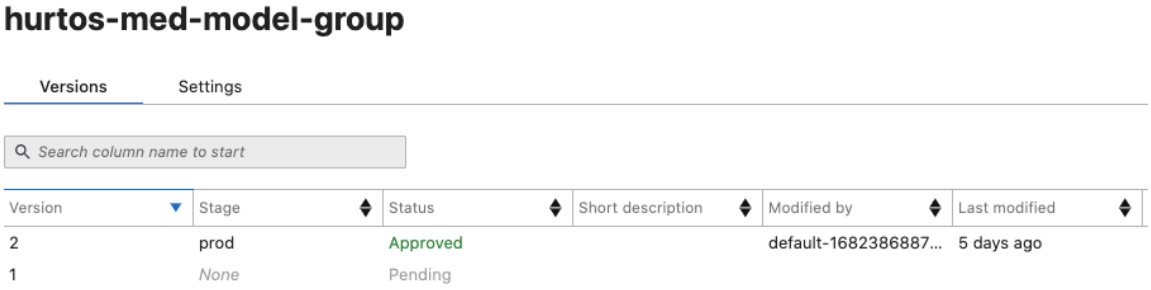
Además de lo anterior, en el repositorio se encuentran archivos como el codebuild-buildspec, que permiten iniciar ejecuciones del pipeline de Sagemaker a través de integración continua, lo cual es fundamental en nuestro planteamiento utilizando MLOps. Básicamente, este repositorio define la sección Sagemaker pipeline que se muestra en la figura 39. Para obtener más detalles, se pueden consultar los archivos en el repositorio del trabajo de grado. En la siguiente imagen (figura 43) se puede observar una ejecución exitosa del pipeline Model Build, el cual basado en los cambios realizados en nuestro repositorio (figura 41) ejecuta el pipeline de sagemaker.

Figura 43: Ejecución Pipeline Model Build



Como se mencionó previamente, una vez se realizaban cambios en nuestro repositorio asociado con el modelo en Sagemaker, se ejecutaba el pipeline y en caso de registrar un nuevo modelo con mejor rendimiento, este debía ser autorizado y aprobado por el científico de datos encargado, tal como se muestra en la figura 39. La decisión es tomada con base en su criterio y conocimiento, al igual que de las métricas obtenidas. En la siguiente imagen, figura 44, podemos ver cómo se expresa esto desde el studio de Sagemaker.

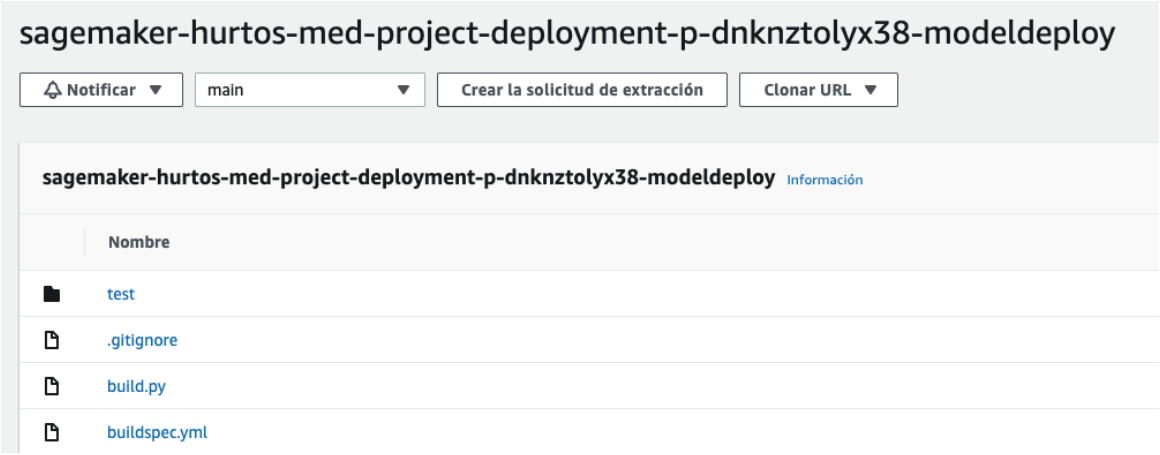
Figura 44: Aprovación del modelo registrado



Desplegar un modelo en producción es importante, pues permite que este entre en funcionamiento y comience a generar valor real para la empresa u organización, validando su eficacia en un entorno real. Por esta razón, el siguiente paso en nuestro planteamiento fue poner el modelo

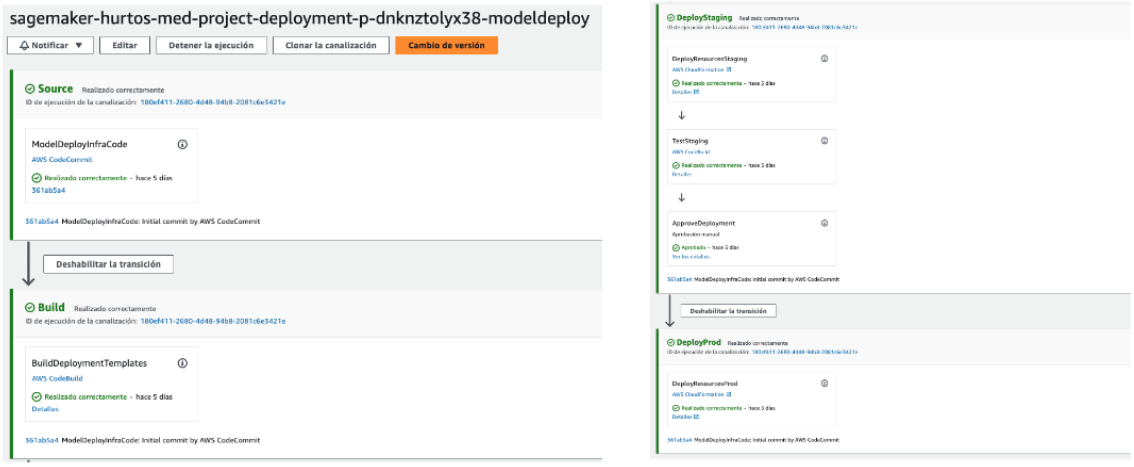
a disposición. Para lograr esto, como ya se mencionó, se utilizó el template *Model Deployment* de SageMaker, el cual se integró con la implementación previamente descrita. Dicho template proporciona, por un lado, un repositorio (figura 45) con el código necesario para encontrar nuestra versión más reciente aprobada del modelo y luego implementarla en el endpoint al detectar un cambio. Por otro lado, define las plantillas de CloudFormation que me permitirán integrar la infraestructura de mis endpoints como código.

Figura 45: Repositorio asociado al model deploy en codecommit



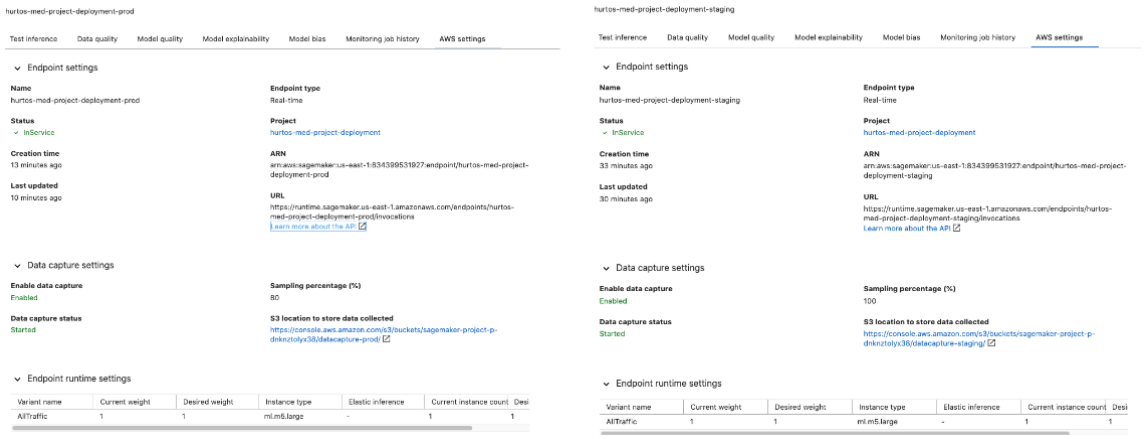
Adicionalmente, se provisiona un pipeline usando CodePipeline, el cual se representa en la figura 39 como Model Deploy. Dentro de este, se encuentra la integración con el repositorio mencionado y la compilación del mismo a través de stacks. Después de este proceso, se despliega el endpoint del ambiente de staging. Para pasar de este ambiente a producción, es necesario contar con una aprobación previa, que por construcción es manual. A continuación, se puede observar el resultado luego de ejecutar el pipeline.

Figura 46: De izquierda a derecha, secuencia de ejecuión del pipeline asociado al deploy



Una vez que el personal a cargo ha aprobado los modelos, se pueden habilitar los endpoints para que se puedan hacer solicitudes de inferencias. Esto nos da la posibilidad de integrar modelos de machine learning en nuestras aplicaciones y sistemas en producción. En la figura 47 podemos ver cómo, en su momento, los endpoints referenciados en la figura 39 estaban disponibles y listos para realizar inferencias.

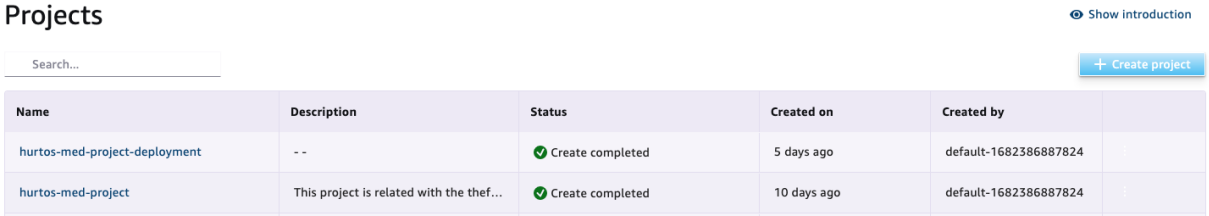
Figura 47: Endpoints disponibles para inferencias con nuestro modelo



En esta parte del flujo del proceso, algunos de los pasos necesitaban aprobación y para lograrlo se usaron reglas de EventBridge. Además, aunque el repositorio asociado con el despliegue a producción está sujeto a cambios, en este caso no fue necesario, ya que al tener un proyecto construido previamente, Sagemaker permite referenciarlo e integrarlo automáticamente.

Es de anotar que los procedimientos descritos anteriormente se enmarcan en los proyectos que se muestran en la figura 48, a continuación.

Figura 48: Proyectos de sagemaker asociados a nuestro trabajo



De esta manera, se describió la parte de nuestra implementación denominada *analítica* en la figura 39. Esta sigue los patrones necesarios y suficientes para, en un primer acercamiento, llevar nuestros modelos de inferencia a un ambiente productivo. Por esta razón, se considera que se ha alcanzado uno de nuestros objetivos, que consiste en el diseño e implementación de una arquitectura para el despliegue de modelos de aprendizaje automático para la predicción de robos en un proveedor de nube usando técnicas de MLops. Para este caso en particular, se implementó un modelo de regresión alineado con lo realizado a lo largo del trabajo. Es importante mencionar que muchos de los recursos mostrados en esta sección fueron removidos debido a que una implementación como la mostrada en la figura 32 conlleva altos costos debido a los múltiples y en algunos casos exigentes servicios que se utilizan, como por ejemplo las instancias utilizadas para la disponibilización de los endpoints (figura 47) que tienen altos costos de servicio.

10.3 Machine learning operations (MLOPs)

Muchas empresas, grupos de investigación y aficionados se han volcado a la experimentación y posterior implementación de modelos utilizando técnicas de aprendizaje automático (machine learning). Es por esta razón que uno de los desafíos actuales en la industria de los datos, y por ende en la obtención de valor de los mismos, es tener metodologías claras que nos permitan no solo tener un control adecuado del ciclo de vida de los datos, sino también una forma ágil de implementar modelos que conviertan nuestra información en activos que reflejen ganancias para nuestras compañías.

La tarea de implementar y operacionalizar nuestros modelos, en este caso enfocados al aprendizaje de máquina (Machine learning), es una tarea que no se considera trivial, por lo que surgen paradigmas como el denominado Machine learning operatios (MLOPs), el cual se encarga de este problema a través de definir conceptos, prácticas, principios y herramientas utilizadas para implementar, escalar y administrar modelos de aprendizaje automático en producción de manera eficiente y confiable [26].

El MLOps como paradigma toma gran valor a medida que el aprendizaje automático se vuelve más central para los negocios y las organizaciones, por lo que los modelos deben ser implementados y actualizados rápidamente para mantenerse al día con los cambios en los datos y las necesidades de los clientes. Esta situación sin técnicas definidas puede representar dificultades de implementación y mantenibilidad, llevando a una deficiencia al momento de escalar y poner en producción los modelos.

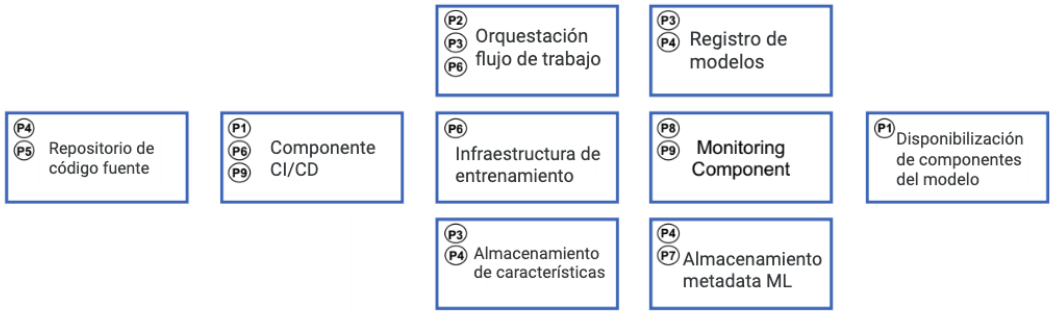
A medida que la adopción del aprendizaje automático se expande en empresas y organizaciones de todo tipo, la necesidad de un enfoque sistemático y estandarizado para la implementación de modelos de aprendizaje automático en producción se ha vuelto cada vez más evidente. Es por esta razón que, a través de los diferentes desarrollos en este proyecto, se buscó mantener el MLOps como enfoque central, lo que resulta en un trabajo a nivel de los realizados en las compañías guiadas por datos, data driven.

Para esto se partió de los principios o buenas prácticas definidas en Matsui et. al [26], los cuales se expresan a continuación y se muestran en la figura 49:

- P1- CI/CD automatización
- P2- Orquestación de flujos de trabajo
- P3- Reproducibilidad
- P4- Versionamiento de la data, el código y los modelos
- P5- Trabajo colaborativo
- P6- Entrenamiento y evaluación continua de los modelos de ML
- P7- Seguimiento de la metadata asociada a los diferentes modelos
- P8- Monitoreo continuo
- P9- Retroalimentación continua

A partir de estos principios se definen los componentes que debe poseer el diseño de un sistema basado MLOps:

Figura 49: Componentes de diseño sistema MLOps [26]



Con base en lo mostrado en la figura 32, se buscó mantener y tener en cuenta la mayoría de estas componentes a lo largo de este trabajo. Esto será evidenciado de manera más notable en los detalles que se presentarán a continuación.

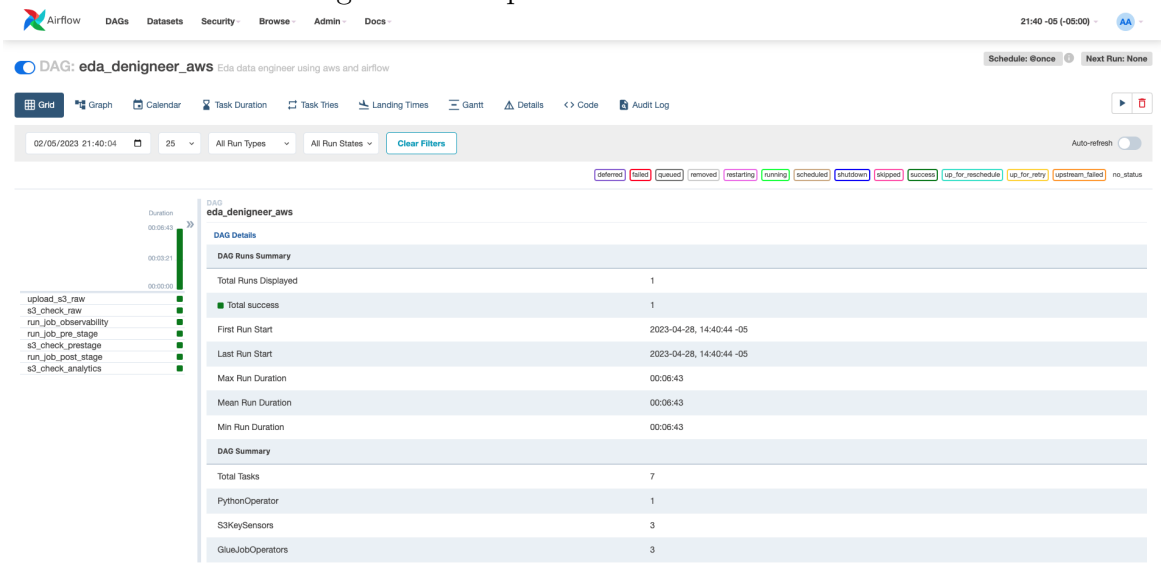
Pensando en el trabajo colaborativo, inicialmente se utilizó herramientas como GitHub y CodeCommit para alojar nuestro código (tabla 9). Esto nos permite realizar un seguimiento de los cambios realizados en él. Además, estos repositorios se pueden integrar con herramientas que nos permiten desplegar rápidamente los servicios necesarios en nuestra implementación y actualizar los cambios realizados en ellos. Entre estas herramientas se encuentran CodeBuild y CodePipeline, los cuales, junto con CodeCommit, nos permiten lograr el objetivo mencionado, como se detalló en la sección de arquitectura. De esta manera, contribuimos tanto al despliegue continuo como al monitoreo de nuestro código.

Tabla 9: Repositorios usados en el proyecto

Nombre repositorio	**Plataforma de versionamiento**
hurto_a_personas_en_medellin	github
sagemaker-hurtos-med-project-deployment-p-dnknztolyx38-modeldeploy	codecommit
sagemaker-hurtos-med-project-p-jcvpcl47l6yx-modelbuild	codecommit

La orquestación es uno de los componentes más importantes en el diseño, ya que nos permite definir cómo, cuándo y con qué frecuencia se ejecutarán las diferentes tareas en nuestro flujo de trabajo. Para ello, se utilizó una herramienta clave como Apache Airflow para la orquestación inicial, como se puede observar en las figuras 33 y 38. Al hacer uso de su estructuración mediante DAGs, logramos orquestar nuestro proceso inicial en el datalake de manera eficiente y efectiva. La implementación con este orquestador fue realizada usando contenedores de docker, de forma partícular docker compose; para más detalles remitirse al repositorio asociado. En la figura 50 podemos ver la interfaz asociada con Apache Airflow.

Figura 50: Orquestación usando Airflow



Por otro lado para el proceso analítico con nuestros datos, se uso el servicio Sagemaker pipelines de AWS como orquestador del proceso de modelado que se evidencia en las figuras

32 y 39, automatizando así tareas como la preparación de datos, la formación de modelos y la implementación en producción.

Para el registro de modelos y su gestión, se utilizó el Model Registry de SageMaker (figura 44). Esta herramienta permite llevar un seguimiento detallado de los diferentes cambios que pueden ocurrir durante el proceso de experimentación, lo que resulta esencial para garantizar la reproducibilidad y trazabilidad de los resultados. Además, el Model Registry facilita el intercambio de modelos entre los miembros del equipo de manera efectiva y asegura que los modelos sean consistentes y estén libres de errores. Al hacer un seguimiento detallado de la evolución del modelo, es posible identificar y resolver problemas de manera eficiente, lo que garantiza una implementación transparente y eficaz.

Como es sabido, la metadata proporciona información adicional sobre los datos y los modelos que se están utilizando, lo que permite a los desarrolladores de estos realizar un seguimiento de los cambios, optimizar el rendimiento, colaborar de manera efectiva y garantizar la seguridad de los datos y de dichos modelos. Es por esto que, además de los servicios de orquestación que ya se han mencionado y que alojan metadata, se han utilizado recursos como los jobs de Glue, que registran las salidas de las diferentes ejecuciones, y el catálogo de Glue, que aloja información sobre los esquemas de las tablas. La combinación de todo este conjunto de herramientas me permite definir métricas que permiten evaluar el rendimiento de mis aplicaciones e implementaciones. La importancia de esta componente se tiene presente en cada una de las fases de nuestro proceso.

En esta misma línea, la implementación de Great Expectations utilizando jobs de Glue, cuyos resultados obtenidos se usaron para su posterior visualización mediante QuickSight, como se muestra en la Figura 37, permite agregar una componente de observabilidad a mi proceso, en este caso, en cuanto a la calidad de la información que se ingresa en el datalake, lo cual es una buena práctica de monitoreo en aplicaciones orientadas a los principios de MLOps.

Finalmente, utilizando AWS SageMaker endpoint, se pudo implementar el servicio de manera que, a partir del modelo con mayor eficiencia, se pueda realizar inferencia para datos a través de solicitudes REST.

En este sentido, es importante destacar que la implementación que se ha llevado a cabo se alinea con las buenas prácticas y los estándares establecidos en el marco de trabajo de MLOps [26], lo que garantiza que el proceso de desarrollo de modelos de aprendizaje automático se realice de manera sistemática y eficiente. Si bien es cierto que no se han cubierto todas las componentes técnicas definidas en la literatura, se ha logrado una implementación robusta y eficaz que cumple con los principales objetivos de MLOps, tales como la automatización de los procesos, la reproducibilidad de los resultados, la gestión del ciclo de vida de los modelos, la colaboración entre equipos y la observabilidad del proceso. En resumen, se ha logrado una implementación de alto nivel, acorde a las necesidades de las empresas que buscan una gestión eficiente de sus proyectos de aprendizaje automático.

11 Resultados

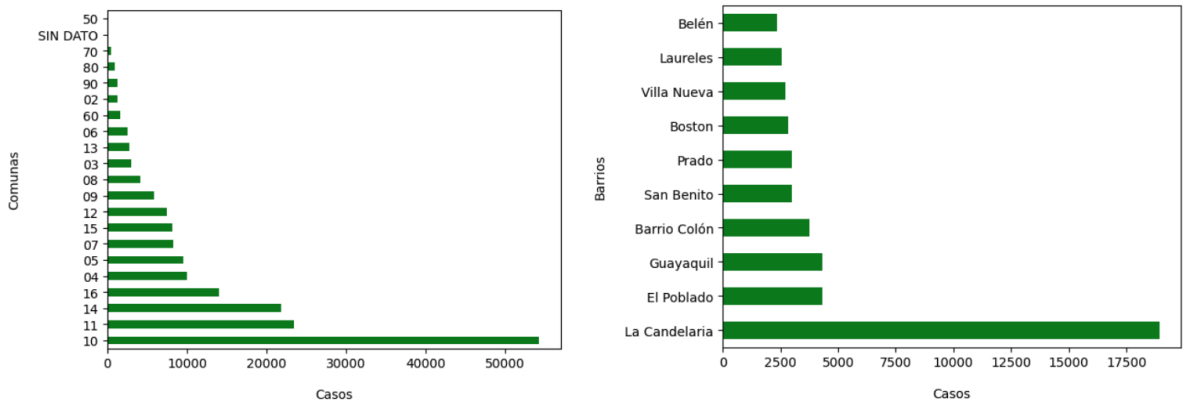
En el presente trabajo, se buscó realizar un análisis del hurto a personas en la ciudad de Medellín. El objetivo general fue cubrir cada una de las etapas dentro del ciclo de vida de un proceso analítico (Han et al., [16]), desde la recopilación de los datos hasta el modelamiento y obtención de resultados de valor. Para lograr esto, se empleó tecnología de vanguardia y servicios avanzados, lo que convierte este proyecto en uno con un enfoque de alto nivel.

Por un lado, desde un enfoque conceptual, este estudio no solo permite responder a la pregunta de cómo se concentra el hurto en la ciudad, sino que también genera hipótesis sobre cómo el territorio y sus características pueden influir en las dinámicas del hurto. De igual manera, desde una perspectiva técnica, nos brinda la oportunidad de descubrir formas de optimizar los procesos utilizando infraestructura y herramientas de última generación, lo cual facilita la toma ágil de decisiones y aumenta la eficacia en su implementación.

Teniendo clara la importancia de comprender y abordar el fenómeno del hurto a personas, específicamente en la ciudad de Medellín, se logró construir una base de datos confiable de información 11. Esta base de datos permite relacionar los casos de hurto con características específicas de los lugares donde se produjeron, lo que nos ha permitido identificar áreas de mayor incidencia, modalidades delictivas y otros factores relevantes. Este logro representa un paso crucial en el desarrollo de este trabajo y se convierte en una herramienta valiosa para la investigación académica y la colaboración interinstitucional. Además, fomenta el intercambio de conocimientos y contribuye al diseño de políticas basadas en evidencia.

En nuestro análisis, encontramos que el hurto a personas en la ciudad de Medellín, durante el periodo estudiado, se ha concentrado de manera predominante en los barrios La Candelaria y El Poblado, pertenecientes a las comunas 10 y 14, respectivamente. Además, se identificaron algunos otros barrios circundantes de La Candelaria, como Guayaquil, el Barrio Colón y San Benito, cómo focos de dicho vejámen (figuras 29 y 51). Como se mencionó anteriormente, esta concentración puede deberse en gran medida al alto número de personas que transitan por estos lugares en diferentes momentos del día, lo cual genera una alta densidad de personas y, por ende, objetivos de valor llamativos para los agresores, dándose así, un factor de oportunidad elevado para la comisión del delito.

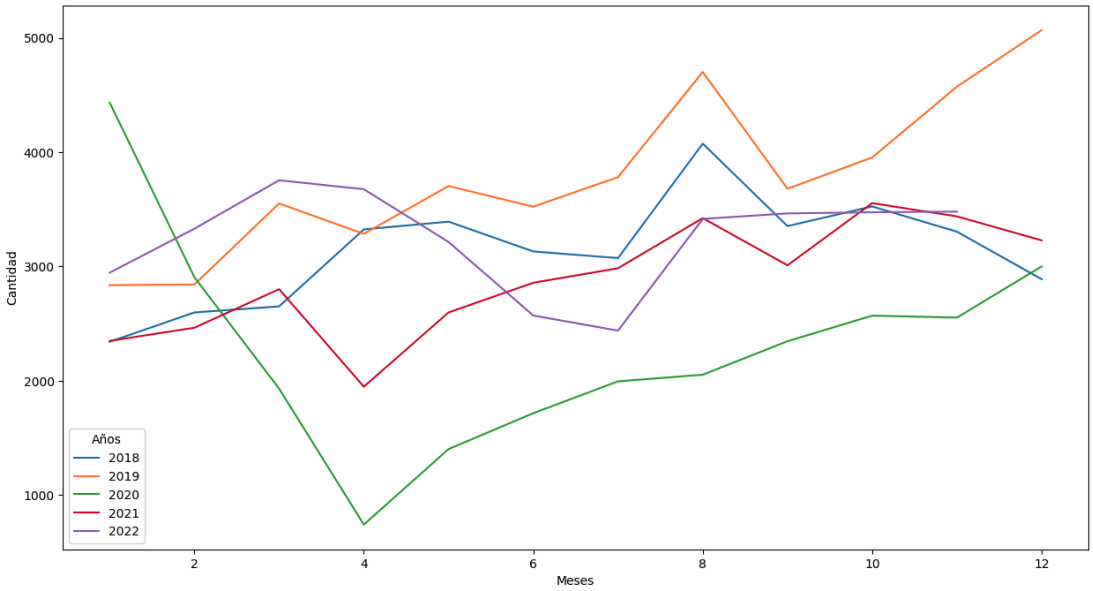
Figura 51: Segregación casos por comuna y barrio. A la derecha el número de casos por comuna, a la izquierda los 10 barrios con mayor número de hechos



Adicionalmente, podemos destacar que dichos lugares, debido a su construcción estructural, pueden representar entornos propicios para una rápida escapatoria de los delincuentes. Esto se debe al alto número de vías de acceso disponibles, lo que les permite perderse fácilmente entre las multitudes y dificultar así su reconocimiento por parte de las autoridades. Esta característica contribuye a la persistencia de los índices de hurto en estas áreas, ya que brinda a los delincuentes una ventaja logística en su actividad delictiva.

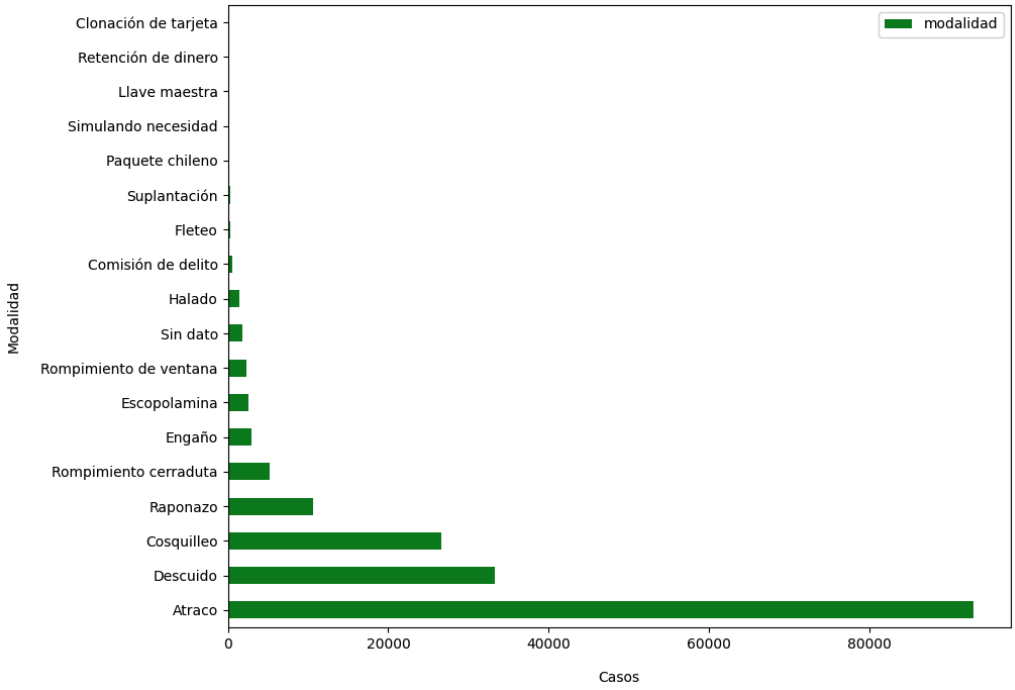
En cuanto al intervalo de tiempo estudiado, encontramos que el hurto a personas presenta un comportamiento similar año tras año, con una variación notable en abril de 2020, coincidiendo con la situación mundial relacionada con la pandemia de COVID-19. En general, podemos observar que la dinámica ha sido bastante consistente a lo largo de los años, con meses como agosto destacándose por presentar picos en la mayoría de ellos (figuras 29 y 52). Esto puede explicarse por la celebración de la Feria de Flores, un evento que atrae a un gran número de personas a la ciudad, lo que incrementa el factor de oportunidad para la comisión del delito. Por lo que podemos afirmar, que no solo las características territoriales afectan la dinámica del hurto a personas, sino además las festividades y eventos sociales que competen a la ciudadanía, como el caso expresado de la feria de flores o navidad, donde vemos ligeros aumentos de casos.

Figura 52: Línea temporal del hurto a personas entre 2018-2022



En el contexto de Medellín, encontramos que los hurtos a personas se caracterizan principalmente por ser realizados mediante el atraco, como se ve en la figura 53. Esta tendencia está en consonancia con el hecho de que en un gran número de casos de hurto se utilizan armas de fuego, armas blancas y objetos contundentes (ver figura 54). Estos delitos son considerados como actos violentos, donde la intimidación a través del uso de armas o la fuerza física es el principal medio utilizado.

Figura 53: Casos de hurto a personas por modalidad

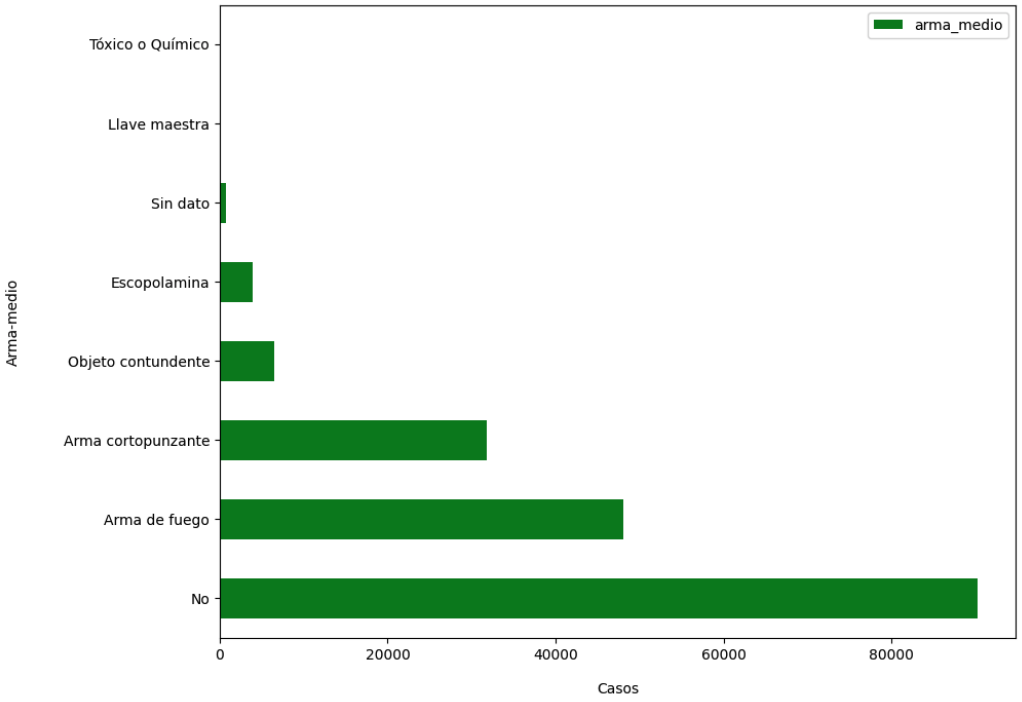


Este hallazgo plantea interrogantes sobre la regulación y el control del uso de armas en la sociedad. Aunque el porte de armas de fuego está regulado en Colombia por la ley 2535 de 1993[45], la cual prohíbe su tenencia para la mayoría de los ciudadanos, resulta preocupante que

este tipo de arma sea el elemento más frecuentemente utilizado en los hurtos, incluso por encima de las armas blancas que son de acceso más fácil.

Esto nos lleva a reflexionar sobre la efectividad de las políticas y medidas implementadas para combatir el uso de armas en la ciudad. Es importante evaluar si se están realizando los esfuerzos suficientes para controlar el acceso y el uso indebido de armas de fuego, así como para promover medidas de prevención y concienciación en la comunidad. La seguridad y protección de los ciudadanos deben ser prioridades, y es fundamental trabajar en conjunto para reducir la incidencia de estos delitos violentos.

Figura 54: Casos de hurto a personas por Arma o medio utilizado



Es curioso el hecho de que entre las principales modalidades del hurto a personas en la ciudad se encuentren el descuido y el cosquilleo, ambas sin un contacto estrecho con la víctima. Esto nos lleva a reflexionar sobre la importancia de buscar estrategias para que la ciudadanía aumente su atención hacia sus pertenencias personales, especialmente en espacios concurridos y de carácter público, aminorando confianzas excesivas que se puedan tener de manera intrínseca en la sociedad.

En gran medida, lo comentado en los párrafos anteriores se ha plasmado en una herramienta visual de fácil acceso, que ha sido desarrollada como parte de este proyecto y que forma parte de los entregables (ver figura 37). Esta herramienta visual representa una validación adicional de la importancia de este tipo de recursos en el análisis de la dinámica del hurto y en la toma de decisiones fundamentadas en datos.

Al analizar nuestro conjunto de datos relacionados con el recuento de casos de hurtos por semana en cada barrio, utilizando redes neuronales y su implementación en grafos, hemos encontrado que la capacidad predictiva del modelo no fue muy alta. Por un lado, esto se debe a la alta dispersión presente en las matrices utilizadas, lo que llevó al modelo a tender a predecir la media de los casos. Por otro lado, la falta de otras variables asociadas a cada espacio territorial impidió obtener un resultado preciso y fiel a la realidad. Como trabajo futuro, se sugiere considerar la inclusión de características estáticas asociadas a cada barrio de la ciudad. De esta manera, no solo se tendrán en cuenta las series de tiempo relacionadas con el número de incidentes, sino también las características territoriales donde ocurrieron.

En cuanto a los modelos utilizados, diferente al de grafos, se realizó un análisis exhaustivo para determinar el mejor enfoque de predicción de casos de hurto a partir de las características espaciales. Inicialmente, se evaluaron regresiones de tipo lineal y regresiones utilizando el algoritmo XGBoost. Sin embargo, se observaron valores bajos para las métricas de evaluación, lo que indicaba que estos modelos no se ajustaban de manera adecuada a nuestros datos.

Fue entonces cuando se decidió emplear el algoritmo de bosques aleatorios, el cual mostró un ajuste significativamente mejor entre la variable a predecir y las variables predictoras. Como se mencionó en la sección de modelado, se obtuvieron valores de R^2 de hasta 0.84 (Tabla 5), lo que evidencia la robustez y versatilidad de dicho algoritmo.

El uso de bosques aleatorios permitió capturar la complejidad y las interacciones no lineales presentes en nuestros datos. Además, su capacidad para manejar grandes conjuntos de variables predictoras y su resistencia al sobreajuste fueron aspectos clave para obtener resultados satisfactorios en nuestro estudio.

Es importante resaltar que en nuestro modelo, las características territoriales que mostraron mayor importancia se relacionan principalmente con lugares que propician la convergencia de personas en los diferentes barrios, como iglesias, tiendas, centros comerciales y lugares turísticos, entre otros (figura 27). Estos hallazgos respaldan lo mencionado anteriormente, donde se argumentaba que la densidad de personas incrementa el factor de oportunidad y, por consiguiente, el riesgo de hurto.

Adicionalmente, al validar la correlación directa entre estas características y el número de hurtos (figura 28), se pudo observar que a medida que aumenta la presencia de dichos lugares en los barrios, también se incrementa la incidencia de casos de hurto. Esta correlación refuerza la relación entre la concentración de personas y la probabilidad de que ocurran hurtos en dichas áreas.

Estos resultados resaltan la importancia de considerar la influencia de las características territoriales en el análisis del hurto a personas. Al identificar y comprender los factores que contribuyen a la concentración del delito, se pueden diseñar estrategias más efectivas para prevenir y mitigar esta problemática. Asimismo, estos hallazgos refuerzan la necesidad de implementar medidas de seguridad en los lugares de alta afluencia de personas, con el fin de reducir la vulnerabilidad y proteger a la comunidad.

Es interesante notar que existe una correlación entre el aumento del número de cámaras en las calles y un incremento en los casos de hurtos, lo cual va en contra de la expectativa de que las cámaras actúen como una herramienta disuasiva del delito (figura 27). Aunque no tenemos un conocimiento preciso sobre el estado de cada una de las cámaras instaladas en el municipio, podría sugerirse que las estrategias de vigilancia y prevención del delito podrían ser optimizadas para cumplir mejor con sus objetivos de salvaguardar la integridad de las personas y sus propiedades.

Por otro lado, es interesante observar que variables relacionadas con escenarios deportivos, hoteles y rutas de ciclismo no tienen un impacto significativo en la capacidad predictiva del modelo. Aunque inicialmente esto podría parecer sorprendente, esta falta de influencia puede explicarse en gran medida por la baja cantidad de estos lugares en la ciudad en comparación con otros espacios de esparcimiento. Esto genera un desequilibrio intrínseco entre algunas características en los datos analizados.

Es importante destacar que la falta de impacto predictivo no significa necesariamente que estos lugares no estén asociados con casos de hurto. Puede deberse simplemente a que la incidencia de hurtos en estas áreas específicas es relativamente baja en comparación con otros lugares más concurridos. Además, existen otros factores que podrían influir en la falta de relación detectada, como la ubicación geográfica de los escenarios deportivos, hoteles y rutas de ciclismo, así como las medidas de seguridad implementadas en estos lugares.

A lo largo del texto, se ha enfatizado la importancia de la parte técnica en los proyectos de analítica de datos y la necesidad de establecer una arquitectura de datos de alto nivel. Por tal razón se diseñó y planteo la arquitectura en la figura 32, la cual proporciona una base sólida

para almacenar, organizar y procesar grandes volúmenes de datos de manera eficiente y efectiva. Esta permite la escalabilidad y la integración de los datos, asegurando su calidad y facilitando el procesamiento de la información. Esto, a su vez, mejora la toma de decisiones, tanto a través de enfoques visuales como mediante el modelado de datos.

Esta arquitectura de alto nivel es fundamental no solo para el proyecto en cuestión, sino también para cualquier proyecto de analítica que busque aprovechar los servicios en la nube. Se ha seguido principios y paradigmas de vanguardia, como MLOps, que resaltan aspectos importantes como el gobierno de los datos y la implementación continua de modelos de aprendizaje automático. Estos aspectos garantizan la confiabilidad y la eficiencia en la ejecución de los proyectos.

El planteamiento de esta arquitectura de datos se considera una referencia para futuros trabajos, no solo en el ámbito de la delincuencia, sino en proyectos analíticos en general. Su adopción permite aprovechar al máximo los avances tecnológicos y garantizar la excelencia en el análisis de datos, impulsando la toma de decisiones basadas en evidencia y maximizando el valor de la información disponible.

12 Conclusiones

Con base en los resultados obtenidos, podemos afirmar que se logró cumplir de manera satisfactoria cada uno de los objetivos planteados al inicio de nuestro proyecto de grado. En términos generales, se logró realizar predicciones sobre los hurtos en Medellín, teniendo en cuenta la relación de los incidentes con las características urbanas circundantes, todo ello enmarcado dentro del paradigma de MLOps.

En particular, se desarrolló una base de datos que se ajusta al objetivo de nuestro trabajo, en la cual se vinculan las características territoriales con los casos de hurto en la ciudad. Esta base de datos puede resultar útil para investigaciones futuras relacionadas con este tema. No obstante, debemos tener en cuenta las posibles imprecisiones derivadas de la falta de actualización de las fuentes y las posibles ambigüedades en estas, por lo que se recomienda realizar mejoras a largo plazo.

Además de esto, se identificaron características fenomenológicas del hurto a personas en Medellín. Por un lado, se evidenció que este delito se caracteriza por su naturaleza violenta, con la presencia de atracos y el uso de armas que representan un gran peligro para la integridad de las personas. Por otro lado, se observó que los delincuentes aprovechan la confianza y distracción de las víctimas para cometer hurtos mediante el cosquilleo y el descuido.

Adicionalmente, se logró focalizar el hurto en la ciudad en aquellos lugares de mayor convergencia de personas. Esto se validó a través del modelado, donde se demostró que los lugares que albergan a un alto número de personas también presentan una mayor incidencia del delito. Esto se debe, en parte, al factor de oportunidad que representa la presencia masiva de personas en un área determinada.

Podemos afirmar que los algoritmos, como los bosques aleatorios, se adaptan de manera efectiva a la complejidad de nuestros datos, lo que resalta su relevancia en proyectos de análisis y toma de decisiones fundamentadas. Esto abre un abanico de oportunidades en el estudio del hurto a personas y nos permite profundizar en su comprensión.

Asimismo, es importante no subestimar la importancia de las variables que no son altamente predictivas en el modelo. Estas variables pueden desempeñar un papel crucial en términos de seguridad y prevención del delito. Es fundamental considerar todas las variables relevantes y adoptar enfoques integrales en la implementación de estrategias de seguridad. Esto implica abordar tanto los lugares con alta incidencia de hurto como aquellos que puedan presentar vulnerabilidades potenciales.

En este trabajo, queda clara la importancia de utilizar tecnologías de vanguardia, ya que nos permiten no solo optimizar nuestros procesos, sino también agilizar el análisis mediante el uso

de herramientas visuales, aliviando la carga cognitiva y facilitando la interpretación de los datos. Queda reafirmada la importancia de utilizar recursos visuales en el análisis de datos y la toma de decisiones. La utilización de herramientas visuales tiene un impacto significativo al comunicar de forma efectiva los hallazgos y resultados obtenidos, lo cual resulta en un enfoque más informado y proactivo en la lucha contra el hurto a personas.

Basados en nuestro análisis, se derivan las siguientes recomendaciones. En primer lugar, es importante promover estrategias de responsabilidad individual y autoprotección para combatir el hurto por cosquilleo y descuido. Esto puede lograrse a través de campañas de concientización y educación dirigidas a la ciudadanía, destacando la importancia de estar atentos a sus pertenencias y evitar confiar en exceso en entornos públicos.

Además, es necesario tener en cuenta que el simple aumento en la cantidad de cámaras de vigilancia no garantiza automáticamente una reducción en la incidencia del hurto. Es fundamental evaluar la calidad y eficacia de la implementación de estas cámaras, así como las estrategias de monitoreo, respuesta y coordinación por parte de las autoridades competentes. Se debe garantizar que las cámaras estén ubicadas estratégicamente en áreas de alto riesgo y que se realice un monitoreo adecuado para detectar y responder de manera oportuna a los incidentes.

Asimismo, se recomienda realizar análisis exhaustivos para identificar posibles brechas en la cobertura de vigilancia. Es necesario determinar si existen áreas con una concentración de casos de hurto a pesar de la presencia de cámaras, lo cual podría indicar la necesidad de ajustar la distribución de las cámaras o implementar medidas adicionales de seguridad en esas zonas.

Este proyecto abre las puertas a múltiples análisis futuros, que van más allá de las características territoriales estudiadas. Podemos explorar la conexión entre el hurto a personas de carácter violento con las lesiones personales y el homicidio en la ciudad, lo cual ampliaría nuestro entendimiento sobre la delincuencia y sus consecuencias. Además, es relevante buscar la implementacion de estrategias de prevención, como las propuestas en los estudios de Marlies et al. [46] y Roach et al. [47], para abordar de manera más efectiva el problema del hurto.

En cuanto al uso de las cámaras de vigilancia de la ciudad, es necesario diseñar y optimizar planes y estrategias de prevención que las involucren, buscando maximizar su efectividad en la disuasión y detección del hurto. Esto implica evaluar su ubicación, calidad de imagen, alcance y coordinación con las autoridades competentes.

En el ámbito técnico, es importante mencionar que queda pendiente mejorar las conexiones a las fuentes de datos. Se puede desarrollar y utilizar scrappers para obtener información más detallada y actualizada. Además, es necesario refinar la arquitectura del sistema, considerando aspectos de seguridad y optimización de recursos. Sería beneficioso implementar métricas de desempeño para evaluar el uso eficiente de los recursos y asegurar un funcionamiento óptimo.

En cuanto al modelado, se sugiere explorar la posibilidad de asignar pesos a las aristas de los grafos utilizados y considerar características adicionales más allá del número de incidentes. Esto podría mejorar la representación de las relaciones entre las variables, ya que la forma en que se abordó inicialmente no permite obtener predicciones precisas. Además, es necesario ajustar los hiperparámetros de los modelos utilizados con el fin de obtener resultados más precisos y sólidos.

En resumen, este proyecto nos brinda oportunidades para futuros análisis, incluyendo la conexión entre distintos tipos de delitos, la implementación de estrategias de prevención basadas en evidencia y la mejora en aspectos técnicos del estudio. Estas recomendaciones permitirán ampliar nuestro conocimiento y contribuir a la seguridad y prevención del hurto a personas en la ciudad.

13 Aspéctos éticos

Los datos que se utilizarán en el desarrollo de este proyecto son tanto de carácter público como privado, y se pueden obtener de fuentes de datos como el SIEDCO, el SISC y Medata, así como de recursos y APIs como Google Maps, aunque el uso de esta última implica un costo. No hay

restricciones asociadas a la privacidad de los datos públicos, mientras que el acceso a los datos privados se ha obtenido de manera legal y se utilizará únicamente con fines académicos. El trabajo de grado propuesto se compartirá para aquellos que lo consideren relevante, sin ninguna intención lucrativa. Al cumplir con los objetivos definidos en este proyecto, se busca desarrollar una herramienta de apoyo para el estudio del fenómeno del hurto a personas en Medellín, que no solo permita comprender las dinámicas del delito, sino que también ayude en la toma de decisiones preventivas en función de los hechos observados.

14 Anexos

14.1 Links a los recursos

A continuación se dejarán los enlaces a los dierentes recursos usados a lo largo de este trabajo.

- **Tabla con información de las bases de datos usadas**
  - `https://docs.google.com/spreadsheets/d/1mEgHiEM45CFV3NQXGm_KYnL32fKhvgH9tEeW8N1Fzv0/edit?usp=sharing`
- **Repositorio público de Github**
  - `https://github.com/jarboledac/hurto_a_personas_en_medellin`
- **Tablero de dinámicas del hurto**
  - `https://public.tableau.com/views/Tablero_16781551878570/Dashboard1?:language=es-ES&:display_count=n&:origin=viz_share_link`
- **Tablero de observabilidad**
  - `https://us-east-1.quicksight.aws.amazon.com/sn/dashboards/1e3d46f5-a566-42d1-aef9-971cd89183c5/views/69b64cf7-4b39-4bd5-87f8-e57d6c6cc5d2`

Referencias

[1] Balbín, D., JaramilloL.UribeC.ArangoM.ArbeláezD.: *La tranquilidad robada. Un análisis institucional del hurto a personas en Medellín.* Alcaldía de Medellín, 2018.

[2] Convivencia, AltaConsejeríaPresidencial para la: *Política Nacional de Seguridad y la convivencia ciudadana.* Departamento Nacional de Planeación, 2011.

[3] Medellín, Alcaldía: *Plan de desarrollo medellín futuro 2020-2023Plan de desarrollo Medellín futuro 2020-2023.* 2020.

[4] Jaitman, Laura: *Los costos del crimen y la violencia en el bienestar en américa latina y el caribeLos costos del crimen y la violencia en el bienestar en América Latina y el Caribe.* Washington DC: Banco Interamericano de Desarrollo, 2015.

[5] Buvinic, Mayra, Andrew Morrison MaríaBeatriz Orlando: *Violencia, crimen y desarrollo social en américa latina y el caribeViolencia, crimen y desarrollo social en América Latina y el Caribe.* Papeles de población, 11(43):167–214, 2005.

[6] Padilla, AdelaidaMaríaIbarra, GloriaCristinaMartínez Martínez EsquidBernardoMena Bermúdez: *Criminal policy against theft in colombia 2016-2020Criminal policy against theft in Colombia 2016-2020.* Justicia, 26(39):237–254, 2021.

[7] Kapoor, Punya: *An analytical approach of crime prediction using machine learningAn Analytical Approach of Crime Prediction Using Machine Learning. Proceedings of Second International Conference in Mechanical and Energy Technology: ICMET 2021, India*, 435–445. Springer, 2022.

[8] Breiman, L, JH Friedman, RA Olshen CJ Stone: *Review of classification and regression treesReview of Classification and Regression Trees.* Biometrics, 40(3):874–874, 1984.

[9] Wheeler, AndrewP Wouter Steenbeek: *Mapping the risk terrain for crime using machine learningMapping the risk terrain for crime using machine learning.* Journal of Quantitative Criminology, 37:445–480, 2021.

[10] Ho, TinKam: *Random decision forestsRandom decision forests. Proceedings of 3rd international conference on document analysis and recognition*, 1, 278–282. IEEE, 1995.

[11] Rosenblatt, Murray: *Remarks on some nonparametric estimates of a density functionRemarks on some nonparametric estimates of a density function.* The annals of mathematical statistics, 832–837, 1956.

[12] Andresen, MartinA, AndreaS Curman ShannonJ Linning: *The trajectories of crime at places: Understanding the patterns of disaggregated crime typesThe trajectories of crime at places: Understanding the patterns of disaggregated crime types.* Journal of quantitative criminology, 33:427–449, 2017.

[13] Ingilevich, Varvara Sergey Ivanov: *Crime rate prediction in the urban environment using social factorsCrime rate prediction in the urban environment using social factors.* Procedia Computer Science, 136:472–478, 2018.

[14] Saraiva, Miguel, Irina Matijošaitienė, Saloni Mishra Ana Amante: *Crime prediction and monitoring in porto, portugal, using machine learning, spatial and text analyticsCrime Prediction and Monitoring in Porto, Portugal, Using Machine Learning, Spatial and Text Analytics.* ISPRS International Journal of Geo-Information, 11(7):400, 2022.

[15] Shukla, Amar, Avita Katal, Saurav Raghuvanshi Shivam Sharma: *Criminal combat: Crime analysis and prediction using machine learningCriminal Combat: Crime Analysis and Prediction Using Machine Learning. 2021 International Conference on Intelligent Technologies (CONIT)*, 1–5. IEEE, 2021.

[16] Han, Jiawei, Jian Pei Hanghang Tong: *Data mining: concepts and techniques*. Morgan kaufmann, 2022.

[17] Deryol, Rustu, Pamela Wilcox, Matthew Logan John Wooldredge: *Crime places in context: An illustration of the multilevel nature of hot spot developmentCrime places in context: An illustration of the multilevel nature of hot spot development*. Journal of Quantitative Criminology, 32:305–325, 2016.

[18] Hastie, Trevor, Robert Tibshirani, JeromeH Friedman JeromeH Friedman: *The elements of statistical learning: data mining, inference, and prediction*, 2. Springer, 2009.

[19] Chen, Tianqi Carlos Guestrin: *Xgboost: A scalable tree boosting systemXgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794, 2016.

[20] Needham, Mark AmyE Hodler: *Graph algorithms: practical examples in Apache Spark and Neo4j*. O'Reilly Media, 2019.

[21] Kim, Dongyoung Sungwon Jung: *Comparison of crime forecasting models based on spatio-temporal data and machine learningComparison of crime forecasting models based on spatio-temporal data and machine learning*. Journal of the Architectural Institute of Korea, 37(1):135–143, 2021.

[22] Das, AsitKumar Priyanka Das: *Graph based ensemble classification for crime report predictionGraph based ensemble classification for crime report prediction*. Applied Soft Computing, 125:109215, 2022.

[23] Burkov, Andriy: *Machine learning engineering*, 1. True Positive Incorporated Montreal, QC, Canada, 2020.

[24] Databircks®: *MlopsMLOps*. `https://www.databricks.com/glossary/mlops#:~:text=What%20is%20MLOps%3F,then%20maintaining%20and%20monitoring%20them`, mes-nomagosto 2022. Accedido en octubre de 2022.

[25] Symeonidis, Georgios, Evangelos Nerantzis, Apostolos Kazakis GeorgeA Papakostas: *Mlops-definitions, tools and challengesMLOps-definitions, tools and challenges. 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*, 0453–0460. IEEE, 2022.

[26] Matsui, BeatrizMA DeniseH Goya: *Mlops: five steps to guide its effective implementationMLOps: five steps to guide its effective implementation. Proceedings of the 1st International Conference on AI Engineering: Software Engineering for AI*, 33–34, 2022.

[27] Shmueli, Galit, PeterC Bruce, Inbal Yahav, NitinR Patel KennethC Lichtendahl Jr: *Data mining for business analytics: concepts, techniques, and applications in R*. John Wiley & Sons, 2017.

[28] Wirth, Rüdiger Jochen Hipp: *Crisp-dm: Towards a standard process model for data miningCRISP-DM: Towards a standard process model for data mining. Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, 1, 29–39. Manchester, 2000.

[29] Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg .: *Scikit-learn: Machine learning in pythonScikit-learn: Machine learning in Python.* the Journal of machine Learning research, 12:2825–2830, 2011.

[30] Medellín, Alcaldía: *MedataMedata.* `https://medata.gov.co`. Accedido entre el 2022 y el 2023.

[31] *Google maps javascript apiGoogle Maps JavaScript API.* `https://developers.google.com/maps/documentation/places/web-service/search-nearby?hl=es-419`. Accedido en el 2022.

[32] Desktop, ArcGIS: *¿qué es un shapefile?¿Qué es un shapefile?* `https://desktop.arcgis.com/es/arcmap/latest/manage-data/shapefiles/what-is-a-shapefile.htm`. Accedido en mayo de 2023.

[33] Telemedellin®: *1 millón 200 mil personas caminan diariamente por las calles del centro de medellín1 millón 200 mil personas caminan diariamente por las calles del Centro de Medellín.* `https://telemedellin.tv/calles-del-centro-de-medellin/172975/`, marzo 2017. Accedido en abril de 2023.

[34] Wikipedia: *Wgs84WGS84.* `https://es.wikipedia.org/wiki/WGS84`, agosto 2022. Accedido en abril de 2023.

[35] Wikipedia: *Archivo:comunas de medellin.pngArchivo:Comunas de Medellin.png.* `https://es.wikipedia.org/wiki/Archivo:Comunas_de_Medellin.png`, junio 2007. Accedido en abril de 2023.

[36] Gis, Wiki: *Buffer (gis)Buffer (GIS).* `http://wiki.gis.com/wiki/index.php/Buffer_(GIS)`, septiembre 2016. Accedido en abril de 2023.

[37] Rojas, Ricardo: *Estructuras de datosEstructuras de datos.* `https://github.com/makeitrealcamp/guides/blob/master/algoritmos/estructuras-de-datos.md`, Agosto 2020. Accedido en abril de 2023.

[38] Chaya: *Random forest regressionRandom Forest Regression.* `https://levelup.gitconnected.com/random-forest-regression-209c0f354c84`, Junio 2020. Accedido en abril de 2023.

[39] Wasserman, Stanley Katherine Faust: *Análisis de redes sociales. Métodos y aplicaciones*, 10. CIS-Centro de Investigaciones Sociológicas, 2013.

[40] Li, Yaguang, Rose Yu, Cyrus Shahabi Yan Liu: *Diffusion convolutional recurrent neural network: Data-driven traffic forecastingDiffusion convolutional recurrent neural network: Data-driven traffic forecasting.* arXiv preprint arXiv:1707.01926, 2017.

[41] Fey, Matthias JanEric Lenssen: *Fast graph representation learning with pytorch geometricFast graph representation learning with PyTorch Geometric.* arXiv preprint arXiv:1903.02428, 2019.

[42] Géron, Aurélien: *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow.* " O'Reilly Media, Inc.", 2022.

[43] Kreuzberger, Dominik, Niklas Kühl Sebastian Hirschl: *Machine learning operations (mlops): Overview, definition, and architectureMachine learning operations (mlops): Overview, definition, and architecture.* IEEE Access, 2023.

[44] Service®, AmazonWeb: *¿qué es un lago de datos?¿Qué es un lago de datos?* https://aws.amazon.com/es/big-data/datalakes-and-analytics/what-is-a-data-lake/. Accedido en abril de 2023.

[45] Colombia, Republica de: *Decreto 2535 de 1993DECRETO 2535 DE 1993.* 1993.

[46] Sas, Marlies, Koen Ponnet, Genserik Reniers Wim Hardyns: *Nudging as a crime prevention strategy: the use of nudges to improve cyclists' locking behavior and reduce the opportunities for bicycle theftNudging as a crime prevention strategy: the use of nudges to improve cyclists' locking behavior and reduce the opportunities for bicycle theft.* Security Journal, 1–23, 2021.

[47] Roach, Jason, Kevin Weir, Paul Phillips, Karen Gaskell Miles Walton: *Nudging down theft from insecure vehicles. a pilot studyNudging down theft from insecure vehicles. A pilot study.* International Journal of Police Science & Management, 19(1):31–38, 2017.

[48] Medellín®, Paola: *Criminalidad y violencia ¿una epidemia en américa latina?Criminalidad y violencia ¿Una epidemia en América Latina?* http://ieu.unal.edu.co/medios/noticias-del-ieu/item/criminalidad-y-violencia-una-epidemia-en-america-latina, marzo 2020. Accedido en octubre de 2022.

[49] Simeone, Osvaldo: *A very brief introduction to machine learning with applications to communication systemsA very brief introduction to machine learning with applications to communication systems.* IEEE Transactions on Cognitive Communications and Networking, 4(4):648–664, 2018.

[50] Ariza, JuanMedina Reka Solymosi: *Crime Mapping and Spatial Data Analysis using R.* CRC Press, 2023.