Analítica de datos para hurtos a personas en la ciudad de Medellín a través de modelos de Machine Learning y Deep Learning

Jhonatan Camilo Arévalo Álvarez
Mario Andrés Fernández García

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Tutor
Javier Fernando Botia Valderrama, Doctor (PhD)

Universidad de Antioquia
Facultad de Ingeniería
Especialización en Analítica y Ciencia de Datos
Medellín, Antioquia, Colombia
2022

Cita	(Arévalo Álvarez & Fernández García, 2022)
Referencia **Estilo APA 7 (2020)**	Arévalo Álvarez, J., & Fernández García, M. A. (2018). *Analítica de datos Analítica para hurtos a personas en la ciudad de Medellín a través de modelos de Machine Learning y Deep Learning.* [Trabajo de grado especialización]. Universidad de Antioquia, Medellín, Colombia.

Especialización en Analítica y Ciencia de Datos, Cohorte III.

Grupo de Investigación Seleccione grupo de investigación UdeA (A-Z).

Seleccione centro de investigación UdeA (A-Z).

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes.
Decano/Director: Jesús Francisco Vargas Bonilla.
Jefe departamento: Diego José Luis Botia Valderrama.

TABLA DE CONTENIDOS

1. RESUMEN EJECUTIVO

El incremento acelerado de los hurtos a personas en la ciudad de Medellín y en Colombia ocasionado por diversos fenómenos sociales, económicos, migratorios entre otros generan impactos negativos en la sociedad. Con el propósito de abordar esta problemática, desde un aporte significativo del análisis de datos, se identifican las áreas de mayor ocurrencia, a partir de la información contenida en la web pública de datos abiertos metadata, cuya muestra es de más de 227 mil datos para un horizonte de 18 años. Esta monografía busca aportar un desarrollo en analítica de datos para que las autoridades y ciudadanos puedan tomar mejores decisiones y acciones competentes en la lucha contra el hurto.

Para ello, se desarrolla un análisis basado en el reporte histórico de hurtos a personas contemplado en el Art. 239 del Código Penal Colombiano, en la ciudad de Medellín, extraídas del Sistema de Información para la Seguridad y la Convivencia (SISC), cuyas cifras documentan los reportes desde el año 2003 hasta el año 2020, teniendo en cuenta la ubicación del suceso, objeto robado, localidad o comuna, bien hurtado, tipo de arma, género de la víctima, y fecha.

Con base en el análisis de datos, se establece que históricamente en la Comuna 10 es donde se registra un mayor porcentaje de hurtos a personas, seguido de la Comuna 11; en ambas localidades, los teléfonos móviles son los objetos más hurtados y los hombres las principales víctimas de tales delitos. Por este motivo, se recomienda a las autoridades centrar su atención en estas zonas e implementar nuevos Comandos de Atención Inmediata, cámaras de videovigilancia y rotación más frecuente del personal policial, y campañas de prevención ciudadana, teniendo en cuenta que no se debe dar lugar al desplazamiento de los delincuentes a nuevas zonas por presentar poca atención a estas.

Para esta monografía, se realizó un comparativo entre dos modelos analíticos de inteligencia artificial, específicamente con el modelo supervisado LGBM se estimaron las probabilidades de ser hurtado teniendo en cuenta un vector de etiquetas y determinadas características; como resultado general el modelo tuvo una precisión del 99%, F1, Recall y AUC del 99%. Por su parte la red neuronal convolucional sequential obtuvo un desempeño bajo dado que no alcanzó a superar el 46% de la precisión.

2. DESCRIPCIÓN DEL PROBLEMA

La concepción del crimen como una de las principales amenazas a los cimientos de la sociedad se sostiene tanto en la percepción ciudadana como los hechos tangibles. La reducción de las necesidades básicas insatisfechas y el aumento de la velocidad en el flujo de la información han generado un sentido de urgencia y de vigilancia constante que impulsa a la ciudadanía a evaluar con preocupación cualquier hecho delictivo (Cadena & Letelier, 2018).

Es así, como los debates en la opinión pública alrededor de la violencia urbana, y más particularmente sobre los atracos, han tomado fuerza en las últimas décadas a razón de la urbanización de países de pasado agrícola, el fin de las guerras civiles y la transformación casi orgánica de la delincuencia organizada o espontánea. En el caso de Colombia, por ejemplo, el aumento de los atracos fue considerado como un efecto esperado de los Acuerdos de la Habana y el Acuerdo de Santa Fe de Ralito debido a la desarticulación de la infraestructura criminal vinculada al conflicto con las FARC y los paramilitares que conllevó a la desocupación masiva de excombatientes, milicianos y operadores (Aguirre, 2017).

Sumado a ello, el desplazamiento forzado en Colombia, especialmente a mediados de los años 90’s y principios de la década del 2000 hizo que las personas desplazadas se dirigieran a las principales ciudades del país en condiciones de pobreza y vulnerabilidad, convirtiéndolas en un blanco fácil para el crimen organizado y la delincuencia común; como consecuencia se presentaron aumentos en diferentes delitos entre ellos el hurto y los homicidios; adicionalmente el efecto de la migración ha podido también tener algún impacto en la delincuencia común.

Por otro lado, el hurto a personas se ha convertido en un fenómeno socioeconómico que impacta la calidad de vida de los ciudadanos y de los diversos actores económicos; llegando de tal forma a todos los hemisferios de la sociedad. Este delito no tiene en cuenta el sexo, la edad, el estado civil, tampoco el estrato y lugar. De acuerdo con lo mencionado, se hace imperativo analizarlo con mucha atención y en la medida de lo posible intentar reducirlo a partir de iniciativas gubernamentales y privadas, utilizando como una de las herramientas la tecnología disponible. Por

ejemplo, el uso de cámaras de vigilancia tradicionales, cámaras de reconocimiento facial, drones y analítica de datos, este último es el que se abordará en este estudio.

Dado que, el enfoque diferencial que se puede obtener, teniendo en cuenta la diversidad de variables que convergen en estos actos delictivos, pueden ser oportunamente analizadas para lograr acciones que nos acerquen a ciudades más seguras, teniendo en cuenta no sólo la tipificación de la acción delincuencial, sino también las características del evento, de la víctima, del contexto y de la intención del delincuente.

Cabe resaltar, que la investigación y análisis del comportamiento delictivo, aporta significativamente en la formulación de nuevas estrategias de aplicación de la ley para la prevención y el control del delito. Es así, como algunos gobiernos han implementado proyectos en un contexto de ciudades inteligentes, suministrando algunas bases de datos relacionadas con diversos fenómenos sociales que ocurren dentro de su territorio, incluidos los asociados con actos delictivos. Por su parte, Medellín, Colombia, cuenta con diversas bases de datos públicas donde es posible obtener información sobre educación, salud, movilidad, infraestructura, y seguridad.

Asimismo, muchos investigadores del campo analítico, han realizado aportes significativos al análisis del comportamiento delictivo en Colombia, tal es el caso de la Corporación Excelencia en la Justicia —CEJ— quienes presentaron el "Reloj de la Criminalidad", que establece las horas en las que más se cometieron los delitos de mayor impacto en Colombia para el año 2021 (con corte al 31 de agosto). En este estudio se realizó un análisis metodológico y estadístico de las cifras extraídas del Sistema de Información Estadístico, Delincuencial, Contravencional y Operativo (SIEDCO) de la Policía Nacional.

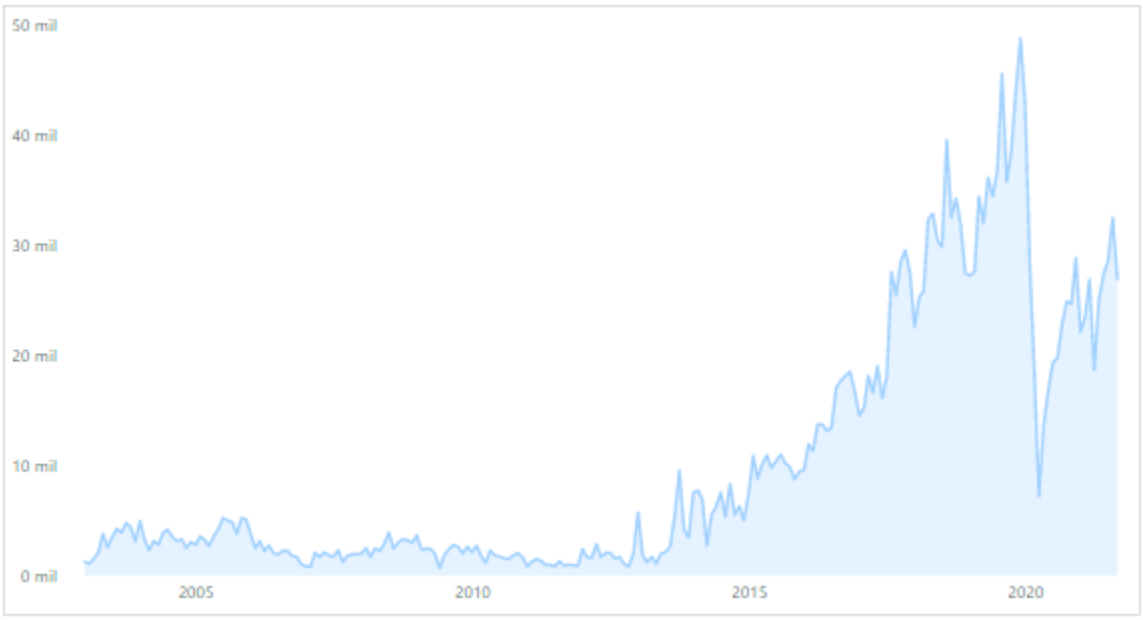
Así las cosas, se realiza un estudio acerca de los hurtos a personas en la ciudad de Medellín en un horizonte de 18 años (2003-2021), partiendo del análisis de las zonas que tienen mayor ocurrencia de hurtos. Se implementa un modelo de inteligencia artificial de aprendizaje profundo o Deep Learning utilizando redes neuronales convolucionales que permitiría estimar las probabilidades por medio de imágenes satelitales identificando las zonas con mayor y menor ocurrencia de hurtos a personas, este análisis también se realiza con un modelo supervisado multiclase LGBM el cual

estima probabilidades de caer en un determinado vector de etiquetas cuya interpretación se asocia con la probabilidad de que se presente un hurto para cada muestra o registro de la base de datos.

2.1 PROBLEMA DE NEGOCIO

El delito de hurtos a personas en los últimos años se ha incrementado en todas sus modalidades. Tal es así, que después del año 2015 los hurtos a personas sobrepasan los 12 mil, lo que representa un incremento del 62% con respecto al año anterior. La tendencia al alza continúa destacando que los años 2018 y 2019 han sido los de mayor cantidad de robos con más de 37 mil y 45 mil respectivamente. En periodo de pandemia (2020 y 2021), los hurtos han disminuido con respecto a los años mencionados, sin embargo, se han mantenido por encima de los 22 mil una cifra que aún es muy alta teniendo en cuenta que es superior a los hurtos presentados durante el 2016 y 2017.

Gráfica 1. Incremento de los hurtos a personas en Medellín



Fuente Elaboración propia

Particularmente en Medellín, ciudad que históricamente ha padecido los azotes de la violencia y la delincuencia común, el hurto es uno de los flagelos sociales que representan un problema para las autoridades locales y para su población en general, dado que muchas veces el acto delictivo en mención, resulta ir acompañado de uso de la violencia y porte de armas por parte del victimario, lo que representa un riesgo real para la salud, integridad y para la vida de los ciudadanos. Para

mayor precisión, previamente se ha demostrado que en la ciudad los delitos contra el patrimonio económico, especialmente el atraco en vía pública, constituyen el mayor porcentaje de casos que aporta al nivel de victimización (“Informe calidad de vida de Medellín, 2018.”).

Si bien, se han realizado aportes significativos por parte de las autoridades para minimizar y/o mitigar estos actos delictivos, entre los que se encuentran el uso de dispositivos tecnológicos y tecnologías de vigilancia ciudadana, se hace necesario hacer un uso significativo de los antecedentes y datos históricos, que permitan caracterizar y estimar los eventos de hurto, dado el aumento tan pronunciado de esta problemática. Por lo cual, se hace necesario aportar un desarrollo analítico basado en datos para que las autoridades y ciudadanos puedan tomar mejores decisiones en la lucha contra el hurto. En un marco en el que las ciudades inteligentes cada vez tienen mayores resultados dado el gran volumen de información con el que cuentan y las herramientas tecnológicas que facilitan el procesamiento, modelamiento y visualización de la información.

Cabe destacar que el origen de los datos es un factor importante, y a la vez una limitante, ya que en este estudio se pretende suministrar una herramienta viable para la toma de decisiones en torno al control de hurtos en la ciudad de Medellín, por lo cual el hallazgo y tratamiento de los mismos, es determinante para dar fiabilidad al trabajo mismo, y que las acciones que se desprenden a partir de este trabajo, sean significativas y generen un impacto positivo en la seguridad de los ciudadanos.

2.2 APROXIMACIÓN DESDE LA ANALÍTICA DE DATOS

Para llegar a la estimación final del modelo de redes neuronales convolucionales y del multiclase LGBM fue necesario en primer lugar utilizar un modelo no supervisado de tipo K Means que permite encontrar el vector de etiquetas o número de clústeres a partir de un análisis de validación interna. La calidad del agrupamiento de datos se incrementó al usar una red neuronal de Auto Encoder o Auto – Codificador el cual se utilizó para realizar una reducción de dimensionalidad de los datos. Luego, Luego, a través de los datos de baja dimensionalidad, se aplicó el algoritmo K-Means para obtener un nuevo vector de clases. El (LGBM) es un modelo supervisado de Machine Learning que se utilizó para clasificar las muestras de la base de datos en 3 vectores de etiquetas, estimando la probabilidad de clasificarse en cada una de las clases. Los valores estimados desde

el modelo se interpretan como la probabilidad de que se presente un hurto de acuerdo con las diferentes características de entrada del modelo.

Como tal el modelo de redes neuronales convolucionales de arquitectura Secuencial es un modelo de Deep Learning que busca a través de imágenes satelitales de tipo Openstreetmap (OSM) reconocer las zonas con mayor frecuencia de hurtos, estimando un valor numérico interpretado como una probabilidad, que a su vez se asocia a un registro de la base de datos que contiene el nombre del barrio y la comuna, de tal manera que se obtiene la probabilidad de ser hurtado en determinada zona. El análisis se puede extender a los horarios, lugares como parques, centros comerciales, etc. con mayor probabilidad de hurtos en algunos sectores de la ciudad de Medellín.

2.3 ORIGEN DE LOS DATOS

A partir de la información recolectada por el proyecto municipal Sistema de Información para la Seguridad y la Convivencia (SISC) de la ciudad de Medellín, se obtiene la base de datos de los hurtos desde el año 2003 hasta el año 2021. Estos datos se extrajeron de la página web pública de Datos Abiertos metadata y cuenta con más de 227 mil registros. Adicionalmente, se cuenta con más de 226 mil imágenes satélites de los puntos de ocurrencia de robo con una distancia a la redonda de 10 metros de donde se efectuó el hurto.

Para descargar las imágenes se realizó un script desarrollado en Python, el cual toma como insumo los datos de la latitud y longitud correspondiente a registros en donde ocurrieron los robos, también fue necesario conseguir los archivos en formato.shp (Shape Files) correspondientes a los polígonos de Medellín, los cuales se descargaron de la siguiente página del geoportal del DANE.[1]
El script con los datos de las coordenadas crea un polígono a partir del punto y del parámetro del área de estudio (Medellín), luego dibuja ese punto en un mapa especializado, le elimina los ejes para que no se vean en la imagen final, se adiciona un mapa base a la representación espacial (OSM), luego con un ciclo se le da un nombre a cada imagen geo-referenciada y las va guardando en una ubicación predeterminada.

[1] https://n9.cl/p78tq

2.4 MÉTRICAS DE DESEMPEÑO

Para determinar el número óptimo de clúster del K-Means se utilizaron medidas para evaluar la calidad de agrupamiento mediante el puntaje de las métricas de la inercia, Davies Bouldin, y de la silueta. Para evaluar el desempeño del modelo de redes neuronales convolucionales y del modelo multiclase LGBM se tuvo en cuenta la precisión, la precisión en la clasificación (acurracy), la precisión balanceada (balance acurracy) F1, Memorización (Recall), el área bajo de la curva (AUC) y la característica Operativa del Receptor (curva ROC), además de las respectivas probabilidades. Se espera obtener una curva ROC cuyo valor del AUC sea mayor igual al 80% y para las demás métricas un porcentaje superior al 80%. Con esto se espera que el modelo tenga un buen ajuste con un mínimo aceptable de precisión. Para la red neuronal se escoge la exactitud y el F1 como las medidas principales desempeño, estos valores deben estar también por encima del 80%.

3. DATOS

3.1 DATOS ORIGINALES

La base de datos para este estudio es de dominio público y se encuentra alojado en la siguiente dirección http://medata.gov.co/dataset/hurto-persona. Esta base de datos contiene información detallada de los hurtos cometidos a personas. La información contenida tiene una cobertura temporal desde enero de 2003 hasta septiembre de 2021 y contiene 248.024 filas y 36 columnas. Para efectos de esta investigación, sólo se tuvieron en cuenta 19 variables con 227.370 registros o muestras, el tamaño es de aproximadamente 35 MB. En la tabla 1 se presenta una breve descripción de la información más relevante.

Para acceder a los datos originales basta con ingresar al sitio web o al hosting de Firebase donde se almacenó en la nube[2] el dataset con el fin de poder cargarlo al notebook de Google Colab de manera más fácil y rápida, lo anterior por el tamaño de la base de datos. Por otro lado, las 20 mil imágenes de muestra de Openstreetmap son obtenidas en formato png con resolución de150 x 150

[2] https://tuwebnetco.web.app/data/Hurtos_Personas_Medellin.xlsx

y en conjunto tienen un tamaño de 11 GB. La descarga masiva de las imágenes se realiza a nivel local y posteriormente se suben a la nube de Google Drive todas las imágenes y de ahí al notebook.

Tabla 1. Descripciones variables del conjunto de datos

Nombre	Tipo	Descripción
SEXO	string	sexo de la víctima, o cuando lo que se mide es la comisión de un delito del presunto indiciado
EDAD	integer	edad de la víctima, o cuando lo que se mide es la comisión de un delito, será del presunto indiciado
ESTADO_CIVIL	string	estado civil de la víctima, o cuando lo que se mide es la comisión de un delito, será del presunto indiciado
MEDIO_TRANSPORTE	string	medio transporte donde se movilizaba la víctima o el presunto indiciado según el caso
NOMBRE_BARRIO	string	nombre del barrio donde ocurrieron los hechos
CODIGO_COMUNA	string	código de la comuna donde ocurrieron los hechos
LUGAR	string	lugar donde ocurrieron los hechos. El lugar es una tipificación del urbanismo más cercano al hecho
MODALIDAD	string	es la forma como se materializa el hecho
ARMA_MEDIO	string	es el arma, medio o mecanismo con el que se comete el hecho
BIEN	string	nombre del bien
CATEGORIA_BIEN	string	agrupación de bienes, nivel intermedio de agregación
LONGITUD	number	longitud geográfica sistema de coordenadas wgs84
LATITUD	number	latitud geográfica sistema de coordenadas wgs84
CANTIDAD	number	cantidad que debe ser entendida en el contexto de la unidad de medida

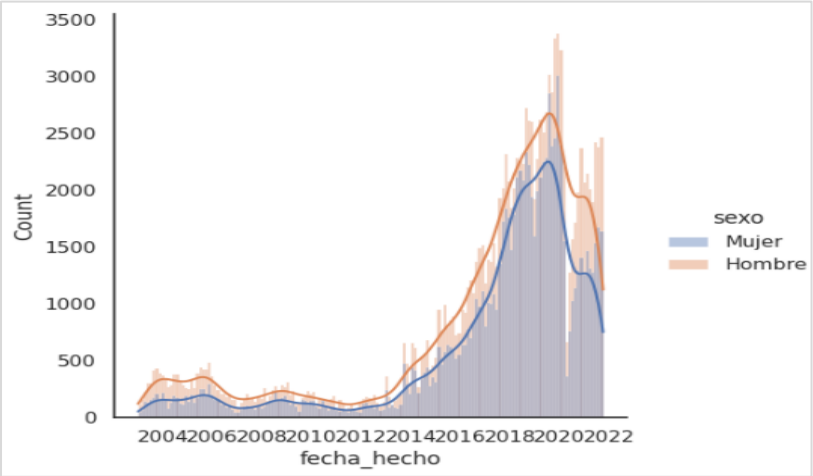
3.2 DATASETS

Los conjuntos de datos para los modelos de Machine Learning y Deep Learning se obtuvieron realizando una partición del 70% para los datos de entrenamiento y 30% para datos de validación. Cabe resaltar que para los datos del entrenamiento en el modelo de (LGBM), se realizó el procedimiento de validación cruzada con el enfoque de *k* -fold, aquí el conjunto de entrenamiento se divide en *k* conjuntos más pequeños, de manera que para este caso se utilizaron 5 pliegues. Para la red neuronal se escogió un conjunto de validación del 20% del conjunto total de los datos de entrenamiento.

3.3 DESCRIPTIVA

Se realiza un análisis de datos a nivel general usando cierto tipo de gráficos que nos permitan entender de mejor manera como es el comportamiento de las variables a través del tiempo. De acuerdo con los datos en un periodo de 18 años se encontró que, después del 2012 los hurtos han aumentado considerablemente. En efecto, durante el año 2019 se presentaron más de 43 mil hurtos a personas, posicionándolo como el año de mayor cantidad. El año 2020, a pesar de ser un año con pandemia, fue el tercero donde más se presentó este delito. El fenómeno de hurtos sigue una distribución con sesgo hacia la derecha para ambos sexos con una mayor concentración en los últimos años (grafica 2). Sin embargo, son los hombres a quienes más roban especialmente en la modalidad de atraco, esta modalidad representa el 50% de los robos.

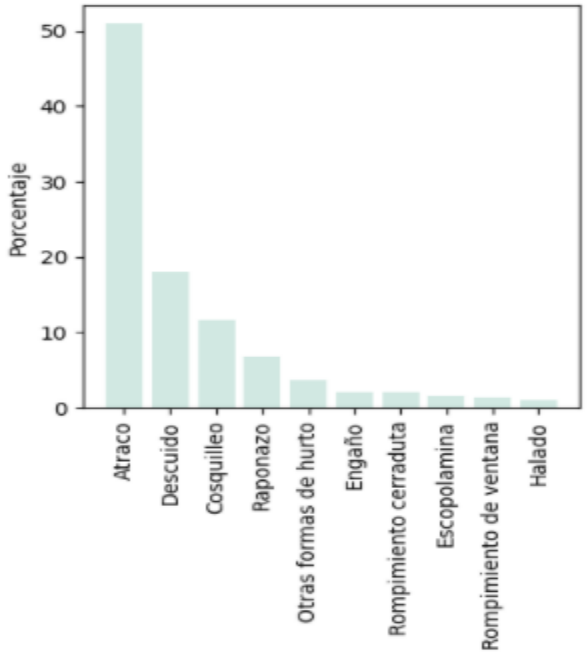
Grafica 2. Distribución de los hurtos a personas



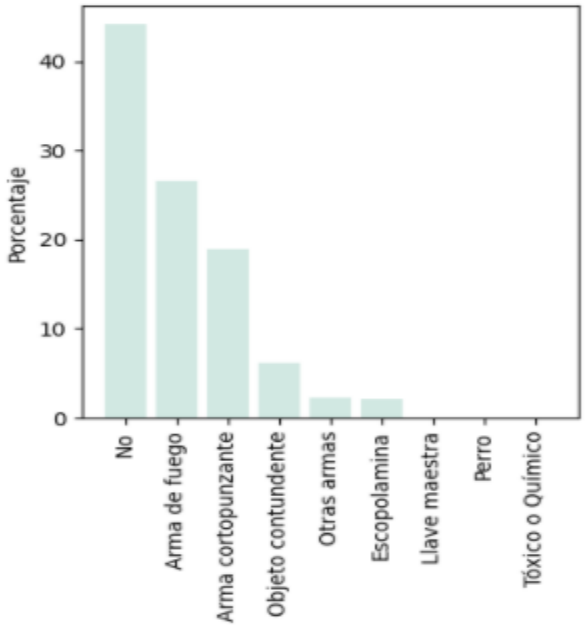
Fuente: elaboración propia

Los elementos tecnológicos son los que más roban estos representan algo más del 40%, dentro de esta categoría son los celulares los que más frecuentemente suelen hurtarse, de allí que el descuido y el cosquilleo representan casi el 30% dentro de las modalidades de hurto, esto está muy relacionado con el tipo de arma usada durante el robo, dado que más del 50% de los hurtos se realizan sin utilizar armas. (graficas 3, 4, y 5).

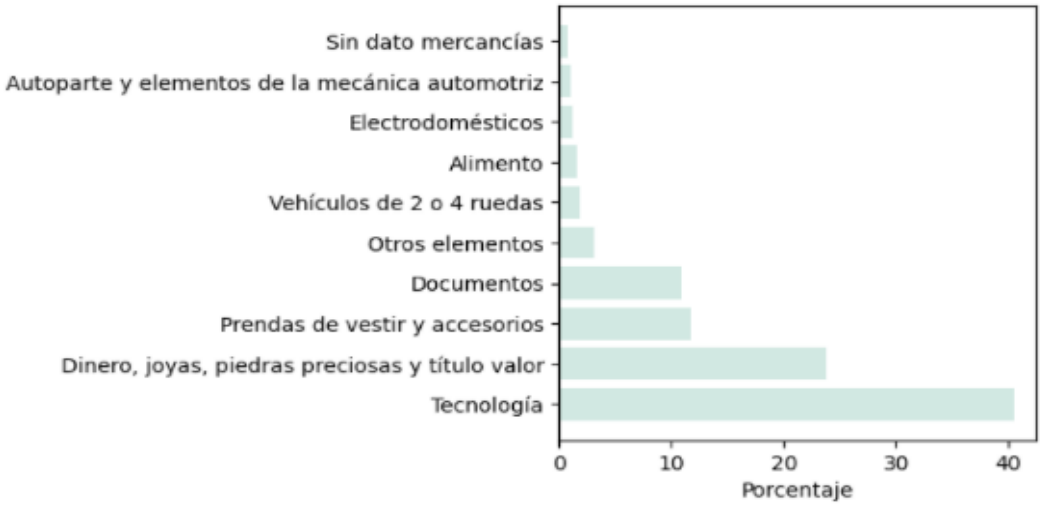
Grafica 3. Participación por Modalidad de hurto



Grafica 4. Participación por tipo de arma



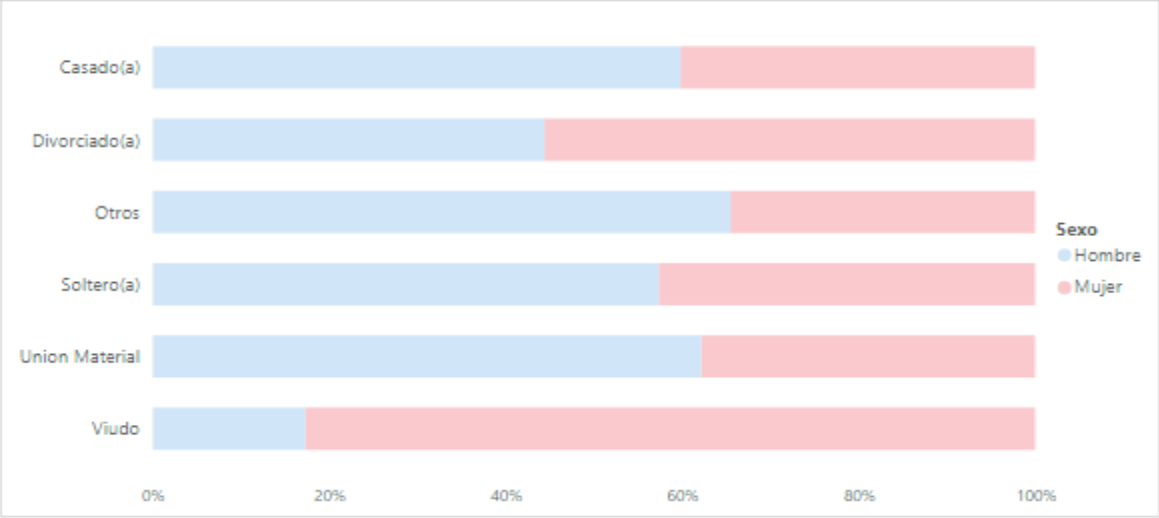
Grafica 5. Participación por Categoría



Fuente: elaboración propia

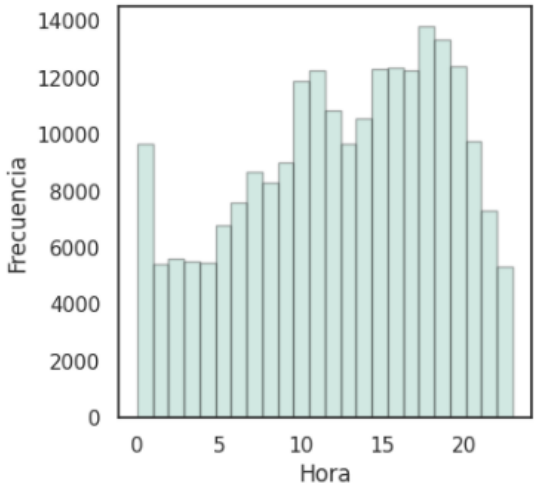
En cuanto a los hurtos por estado civil se encuentran que, las personas solteras son las que más sufren por este delito, particularmente los hombres con más de 27 mil hurtos representan más del 55% para esta categoría. Un comportamiento muy similar se observa para los hombres casados y los de unión marital. No obstante, el comportamiento cambia para las categorías viudo(a) y divorciado(a) puesto que aquí las mujeres son a las que más roban, esto puede estar asociado a que las mujeres de estas categorías son mujeres mayores lo cual las hace un poco más vulnerables que las mujeres solteras más jóvenes.

Grafica 6. Demografía de los hurtos

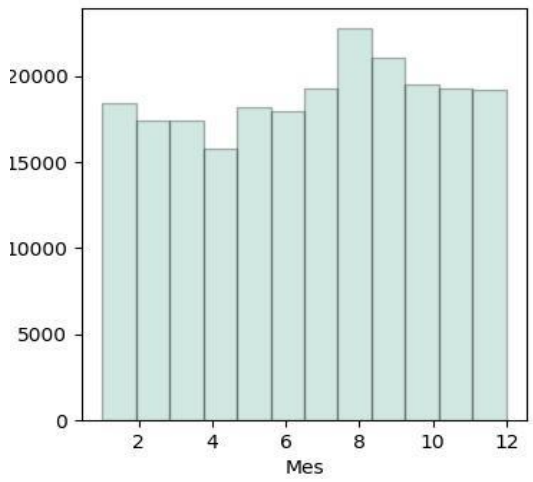


Fuente: elaboración propia

Grafica 7. Frecuencia de hurtos por hora



Grafica 8. Frecuencia hurtos por mes

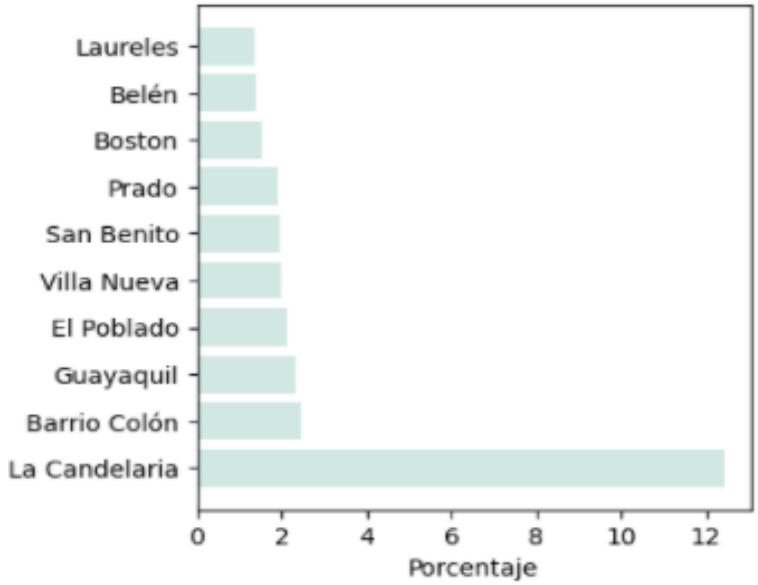


Fuente: elaboración propia

Analizando los histogramas de la hora del día y del mes, no se encuentra un sesgo en particular. Sí se encuentra un patrón y es que entre las 6 pm y 8 pm son las horas con que mayor frecuencia se llevan a cabo los robos (casi 40 mil) y que después de mitad de año suelen haber más hurtos, especialmente en el mes de agosto y septiembre (más de 20 mil).

Respecto a los barrios se tiene que La Candelaria es el barrio donde más hurtos se cometen (12%), algo que llama mucho la atención es que los barrios El Poblado y Laureles también aparecen como uno de los barrios con esta característica. Se desconoce puntualmente porque estas zonas se ven afectadas por este fenómeno social, sin embargo, a priori tal vez una de las causas puede estar relacionada con la densidad poblacional ya que son sectores muy turísticos y comerciales.

Grafica 9. Participación por barrios

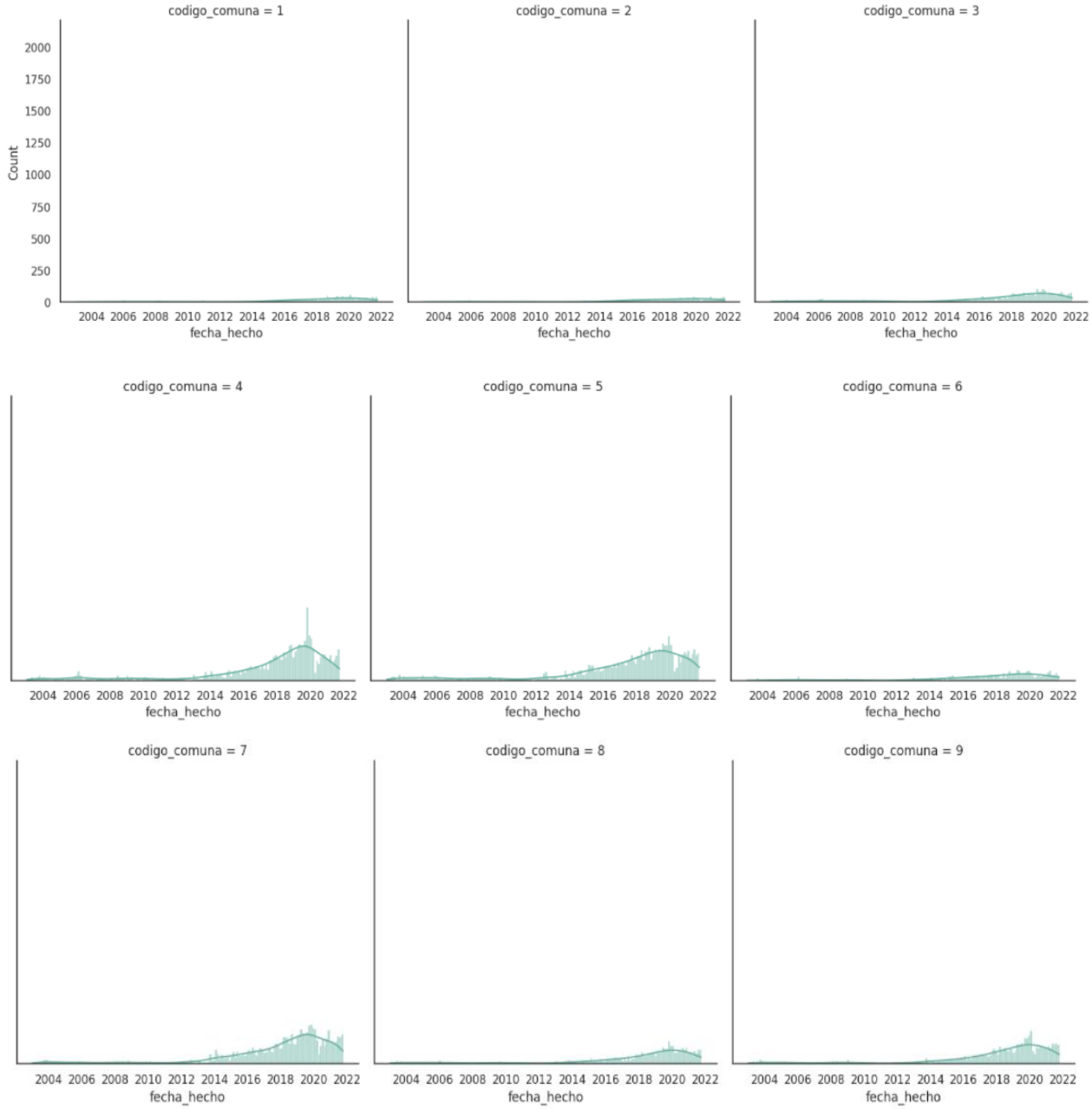


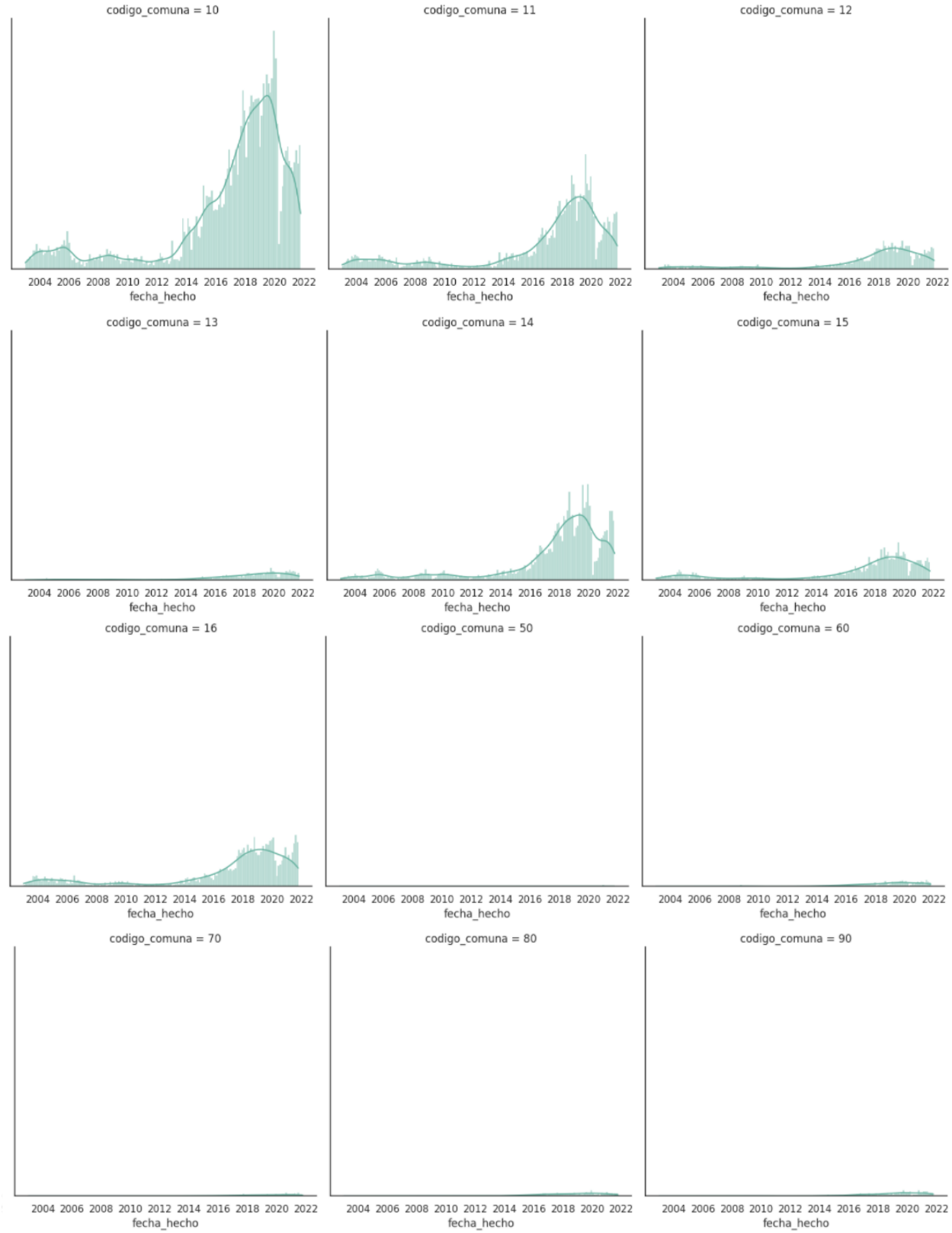
Fuente: elaboración propia

La grafica 10 representa la distribución de los hurtos por comuna, se observa entonces que en las comunas 1, 2, 3 no se han incrementado los robos a personas tal parece que estas comunas pueden ser un poco más seguras. Una distribución muy similar se presenta en las comunas del área rural de Medellín 60, 70, 80 y 90 ya que la curva es casi plana no se refleja como tal una distribución, lo que lleva a pensar que al no haber una densidad población alta los hurtos son poco frecuentes. Caso contrario, se presenta en las comunas 10, 11 y 14 donde la distribución está sesgada hacia la derecha, es decir que los hurtos en estás comunas se han ido incrementado con el paso de los años,

particularmente se ven muy concentrados después del año 2015. Las demás comunas también presentan una tendencia muy similar, aunque la distribución es mucho más plana.

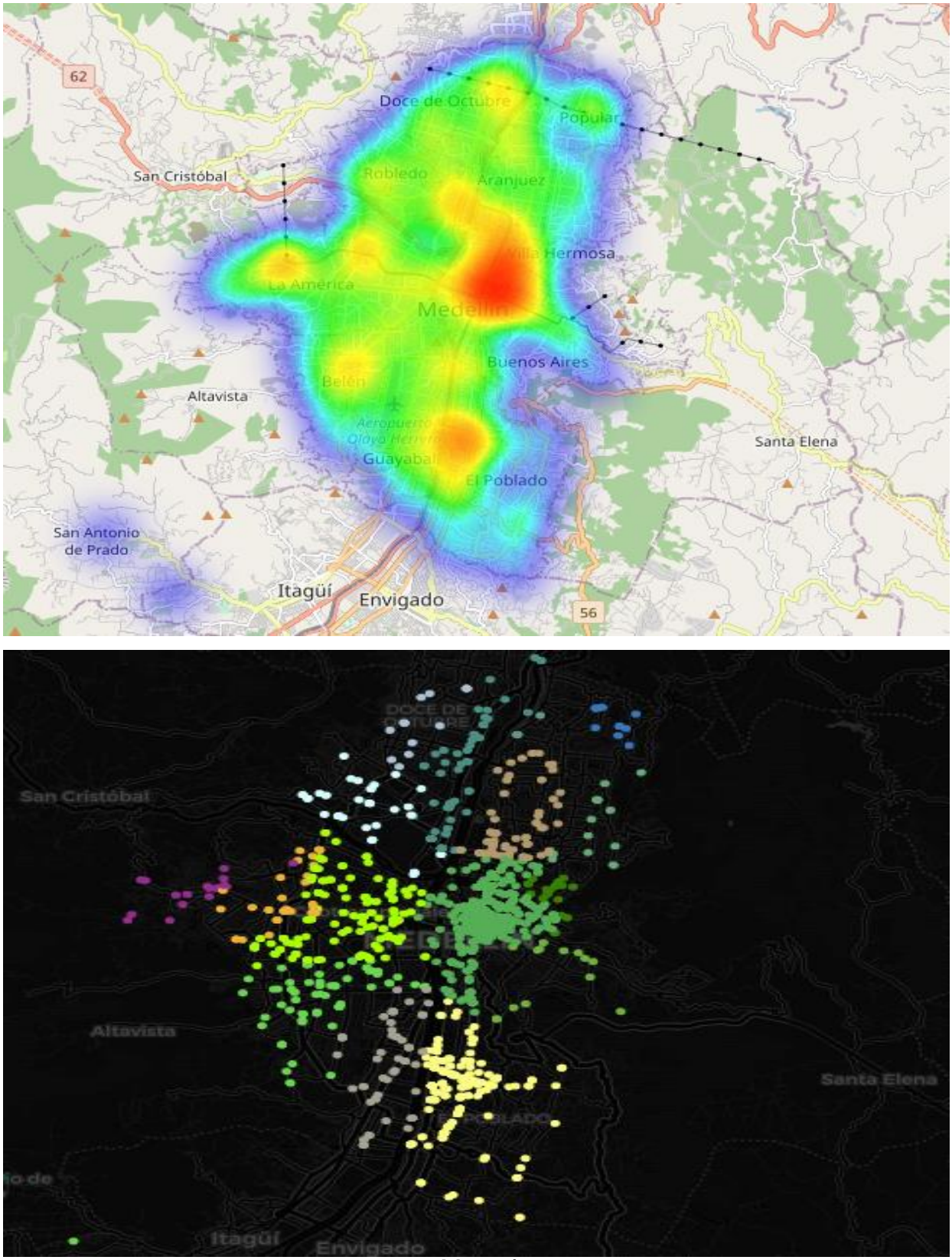
Grafica 10. Distribución de hurtos por comuna





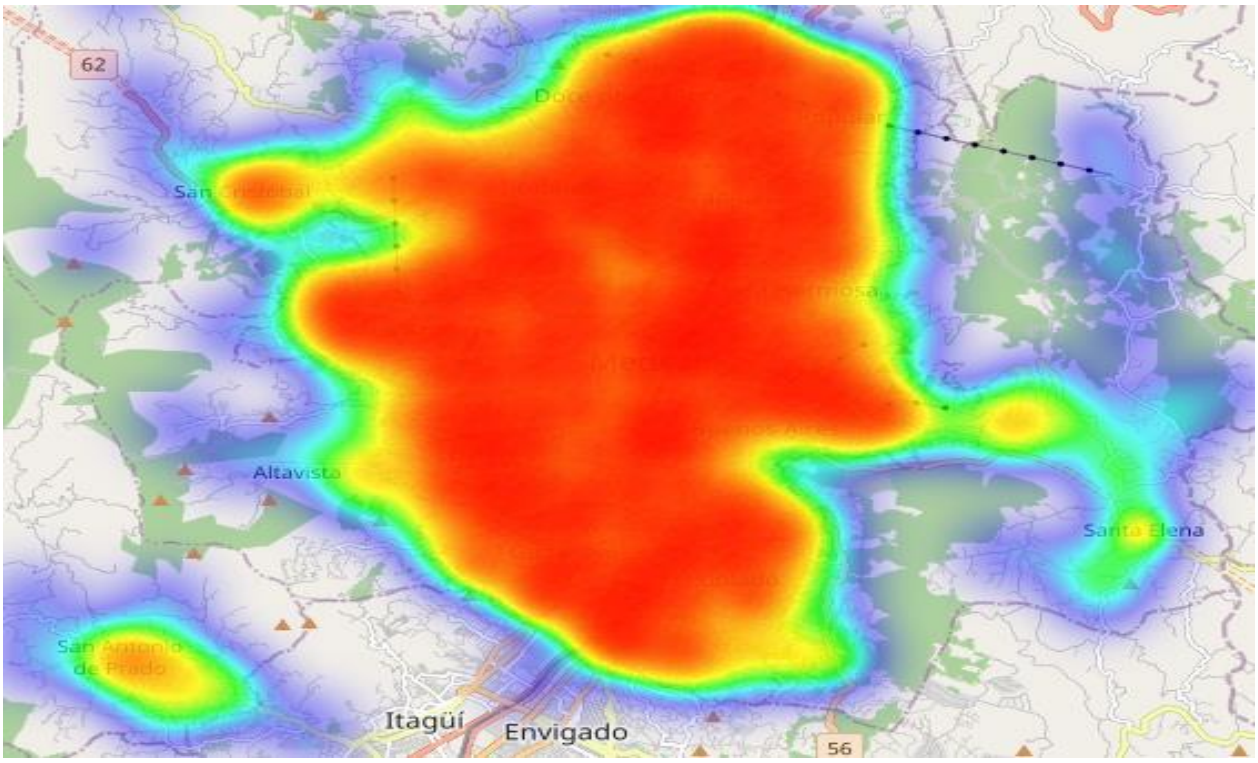
Fuente: elaboración propia

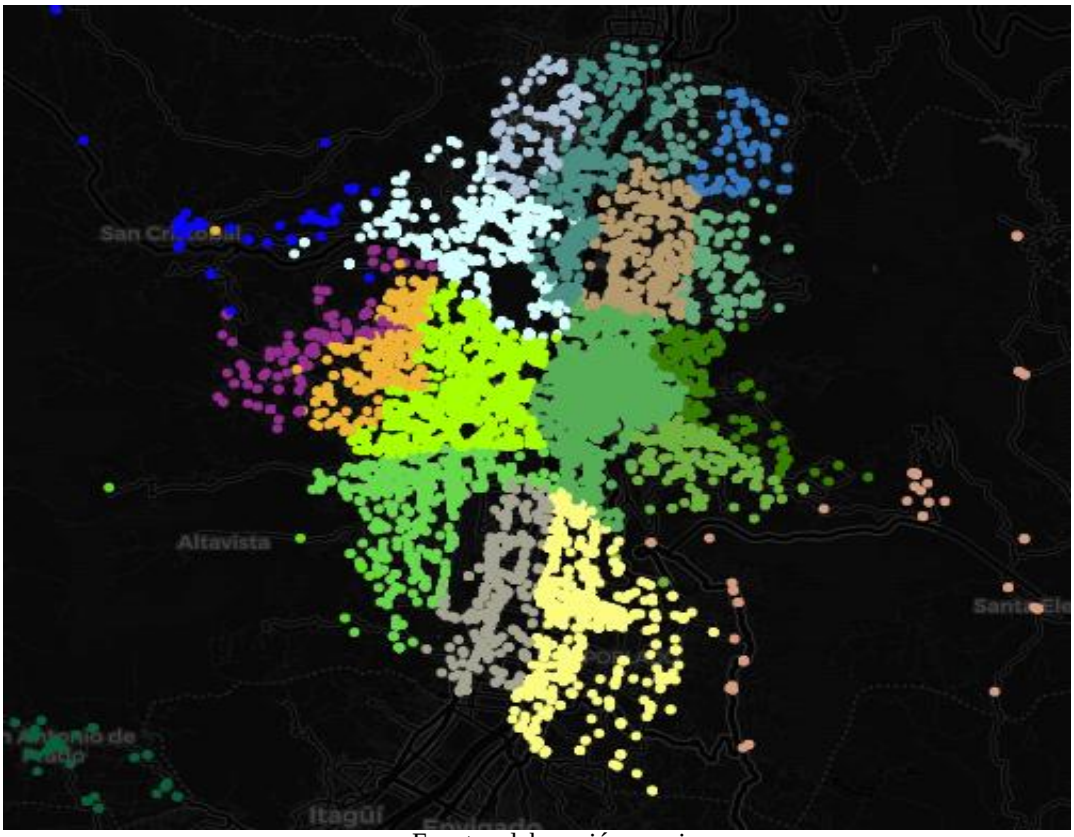
Grafica 11. Concentración de los hurtos en la ciudad de Medellín (2011)



Fuente: elaboración propia

Grafica 12. Concentración de los hurtos en la ciudad de Medellín (2019)





Fuente: elaboración propia

A continuación, se presentan diferentes mapas con el objetivo de identificar las zonas y comunas que concentran la mayor y menor cantidad de robos. Para este análisis se toman como referencia el año 2011 a principio de la década pasada que registró el mejor comportamiento en cuanto hurtos a personas con tan solo 1.283 y el 2019 que es el último año antes de pandemia[3] y que coincide como el año donde más hurtos se presentaron 44.209. Ahora bien, de acuerdo con los datos de Proyecciones y retroproyecciones de población municipal para el periodo 1985-2017 y 2018-2035 con base en el CNPV 2018 del Departamento Administrativo Nacional de Estadístico DANE, la población de Medellín para el año 2011 era de 2.213.549 y para el 2019 de 2.483.545 lo que representa un incremento del 10,8%.

En ese orden ideas, con la anterior información y utilizando la respectiva fórmula se calcula el indicador tasa de hurtos. Encontrando que, para el año 2011 por cada 100.000 mil habitantes se presentaban 58 hurtos a personas y para el 2019 se registraron 1.780 hurtos por cada 100.000 mil, o sea que el indicador se incrementó en 2969% en un horizonte de 9 años. En otras palabras, en Medellín durante el 2019 robaron 30,7 veces más que en el año 2011. Estas cifras evidencian la problemática tan aguda que presenta la ciudad en este tipo de delitos.

(1)

$$Tasa\ Hurtos\ = \left(\frac{N°\ hurtos\ a\ personas}{Poblacion}\right) * 100.000$$

En el mapa de puntos se evidencia que todas las comunas de Medellín para el año 2019 han tenido un aumento sustancial en los hurtos a personas respecto al 2011, las comunas del área rural no han sido ajenas a esta tendencia. Por ejemplo, en la comuna 10 se incrementaron los robos en 1.232,89% y en las comunas 14 y 11 el aumento fue de 1.349% y 1635% respectivamente. En la comuna 80 que corresponde a San Antonio del Prado el aumento fue de 6150%

En particular, algunas zonas de la ciudad han visto como este delito se acentúa con el pasar de los años. Con los mapas de calor se hacer un análisis más profundo al identificar zonas de mayor concentración. En efecto, al sur oriente de la ciudad en sectores como El Poblado y los Naranjos que durante el año 2011 no presentaban casi hurtos a personas, en el año 2019 en estas mismas zonas aumentaron significativamente.

[3] Se decide no tomar el año 2020 en este comparativo ya que fue año de pandemia, el cual fue un año atípico y para no entrar en sesgos, se toma el año inmediatamente anterior.

Un comportamiento similar se identifica en la zona sur, que específicamente corresponde a las carreras 43, 44 y 45 atravesadas por la calle 16 sur, en esta zona se encuentran clínicas y algunos edificios. Continuando por la zona de la autopista sur hacia el sector de Aguacatala y por la calle 12 sur se registraban pocos o nada de hurtos, no obstante, en el año 2019 este comportamiento se revirtió, ya que en todo este sector los hurtos se incrementaron de igual manera.

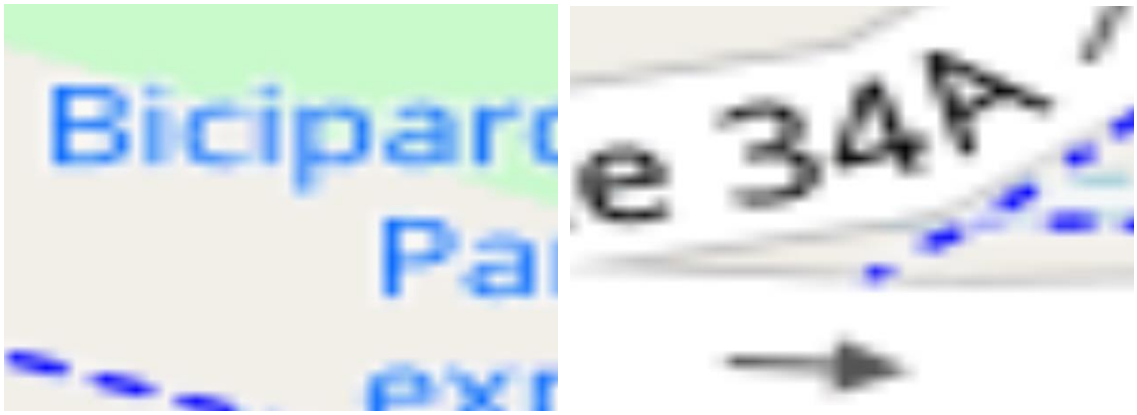
Si bien el sector de la 70 hacia el occidente de la ciudad ya presentaba algunos hurtos en el año 2011, sorprende ver en el mapa de calor como a través de toda la 70 se concentran tantos hurtos, pasó de ser un lugar con algunos robos a ser una zona relativamente insegura para 2019, asimismo sus alrededores donde se ubica el barrio Laureles.

El sector de la candelaria se ha caracterizado por ser unos de los lugares con más frecuencia de robos a personas, en particular si se trata de elementos tecnológicos y dinero. En el 2011 esta zona era de las más inseguras, y para el 2019 el incremento ha sido muy pronunciado convirtiéndola en la zona más insegura de la ciudad, al igual que sus alrededores (estación San Antonio, alrededor del Éxito San Antonio, Universidad ECCI Sede A). De igual manera, toda la avenida que se encuentra en la carrera 46 y que se intercepta de la calle 47 a la 58 también ha aumentado considerablemente los hurtos a personas.

Finalmente, el sector de la terminal de transporte al inicio de la década no registraba casi hurtos, sin embargo, ya para el 2019 estos habían aumentado bastante hacia sus alrededores. Por los lados del Jardín Botánico y de la Estación Universidad, este delito también presenta alzas en la cantidad de los hurtos con respecto al año 2011, cuando se registraban muy pocos.

Las siguientes figuras son ejemplo del tipo de imagen que se consiguió con la descarga masiva y que posteriormente se utilizaran en la red neuronal convolucional. Cada imagen representa un punto (latitud - longitud) en donde sucedió un robo, se toma como referencia un área de 10 metros a la redonda.

Imagen 1. Imágenes satelitales en openstreetmap

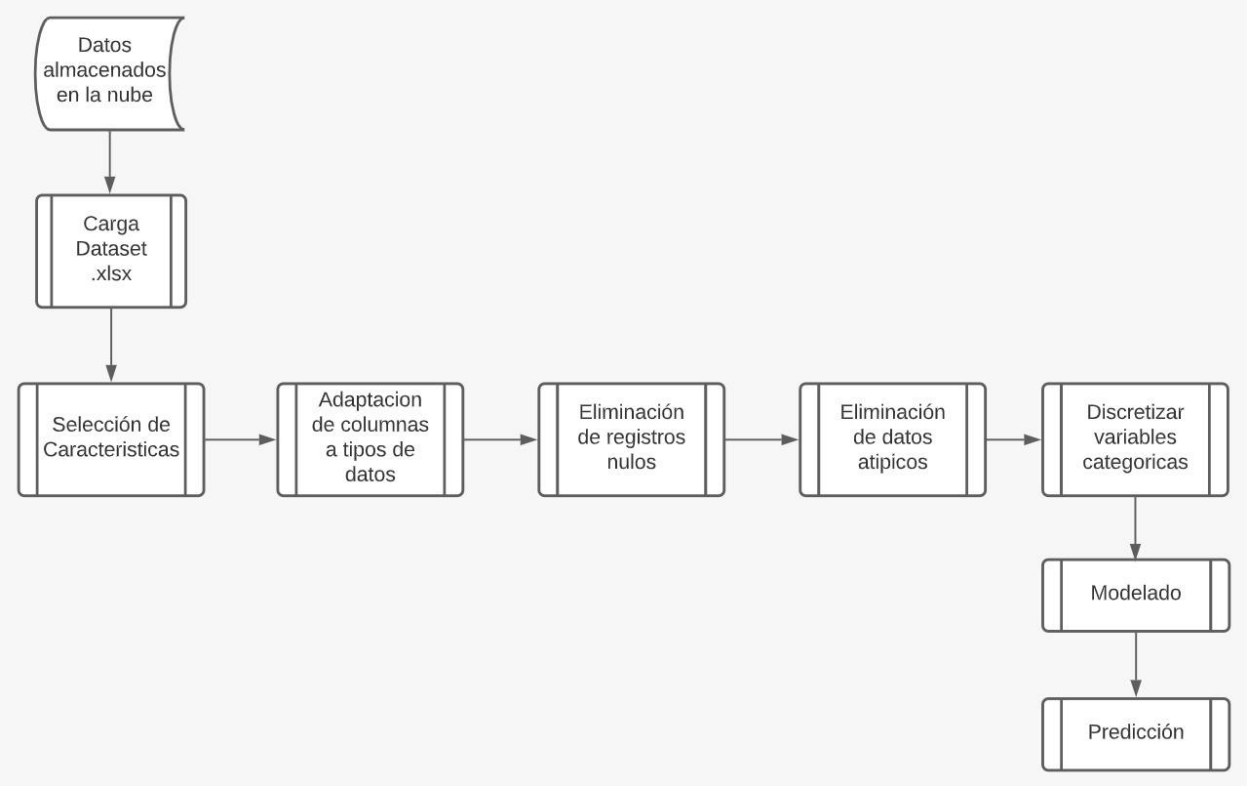


Fuente: Openstreemap

4. PROCESO DE ANALÍTICA

4.1 PIPELINE PRINCIPAL

Imagen 2. Flujo de trabajo



Fuente: elaboración propia

4.2 PREPROCESAMIENTO

El procesamiento de los datos pasó por diferentes etapas que consistieron selección de variables relevantes, transformaciones de datos, eliminación de datos atípicos y nulos. A continuación, se describe brevemente que se hizo en cada una de ellas:

- **Selección de características**: la base de datos contiene algunas variables que no aportan información importante para un posterior análisis, muchas de estas no tienen información. Por lo tanto, deciden eliminarse y no serán tenidas en cuenta en los modelos.

- **Transformación**: la variable "fecha_hecho" que relaciona el día y la hora del hurto está tipo objeto, se realiza la debida transformación a time y a partir de este cambio se crean columnas nuevas (año, mes, día, hora). La mayoría de variables son de tipo categóricas y/o cualitativas. Hay variables con más de 50 etiquetas o niveles, algunas de estas etiquetas pueden estar nombradas como "sin dato" por lo que se decide renombrarla como "otros". Esto aplica para la mayoría de variables de la base de datos.

- **Eliminación de registros nulos:** las variables latitud y longitud no pueden tener datos nulos y estos a su vez no pueden reemplazarse o imputarse dado que son valores previamente definidos en la recolección de los datos. Por lo tanto, se decide eliminar las filas que contienen estos nulos que representan el 5% de la base de datos. Situación similar se realiza con la característica sexo, presenta más de 1000 filas que no tienen asignado ningún sexo por tanto se deben eliminar. Estás muestras representan el 0,66% del conjunto de datos.

- **Eliminación de datos atípicos:** Con el algoritmo LOF cuya metodología para detección de atípicos es muy buena siempre que se use en grandes bases de datos, se encontraron 7.316 muestras con datos atípicos que se procedieron a eliminar de la base de datos.

Después del proceso anterior se discretizan las variables categóricas, esto porque hay muchos niveles en las diferentes características elegidas por lo que se hace necesario reemplazarlas por número enteros que inician desde 1 hasta 100 o 150 dependiendo de que tantos niveles existan en

cada característica. Finalmente, con la base de datos totalmente limpia y transformado se realiza el proceso de escalamiento de las variables para introducirlas en cada modelo. Respecto al conjunto de imágenes no tuvo que hacerse tanta limpieza, dado que en la base de datos los campos de latitud y longitud ya estaban limpios. Solo en el proceso de descarga masivo de las imágenes se hizo un ajuste en la resolución para que estas quedaran de 150 x 150.

4.3 MODELOS

Para el desarrollo de esta investigación se consideraron varios modelos de tipo no supervisado, supervisado y de aprendizaje profundo. Específicamente se utilizó un Mini Batch K-Means mejorado con una red neuronal de auto enconder, el LGBM, la red neuronal convolucional de tipo secuential y/o de arquitectura VGG16.

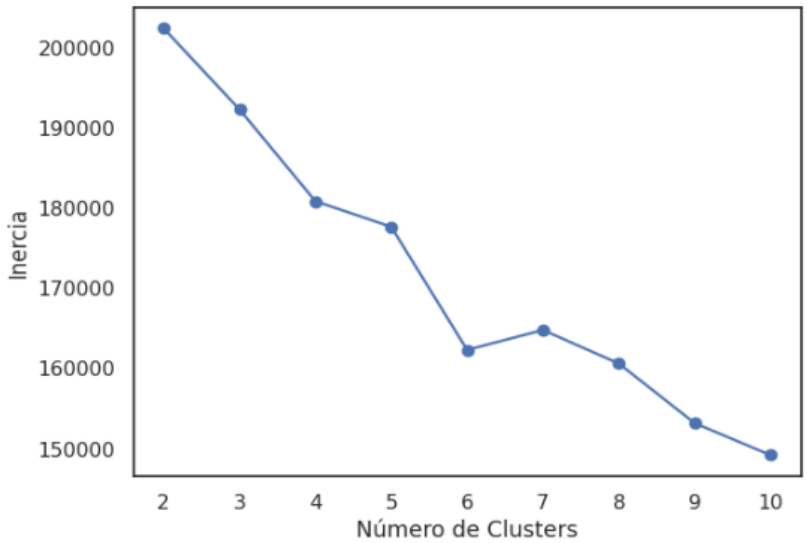
K-Means: la intención de usar un modelo no supervisado tiene como objetivo principal encontrar un vector de clases o una variable de salida que recoge las características generales de las variables, es decir, agrupa en clases las muestras o individuos con comportamientos similares. Al modelo se le pasaron como tal un rango de clusters entre 2 y 10, el algoritmo K-Means generaba un modelo cada vez que se cambiaba el número de clusters. Esta estrategia permite evaluar la calidad del agrupamiento en cada modelo generado por el algoritmo, usando índices de validación de clusters interno. De esta forma, se puede encontrar el número óptimo de clusters para obtener el mejor agrupamiento de datos posible. Adicionalmente, para una mayor eficiencia se aplica una variación del algoritmo K-Means[4] dado el volumen de datos con el que se cuenta.

Los resultados para elegir el número óptimo de clusters se evidencian con el puntaje de la inercia (gráfica 13), el cual sugiere que son seis. Sin embargo, cuatro de ellos eran espurios y al final el agrupamiento quedaba solamente con dos clusters. Se decide entonces escoger 3 clusters dado que el agrupamiento era más estable y no son espurios, entrega una mejor calidad de agrupamiento y finalmente el vector de etiquetas no queda tan desbalanceado. El puntaje de la métrica de Davies Bouldin (gráfico 14) podría sugerir que entre 3 y 4 clusters se encuentra el resultado óptimo.

[4]El MiniBatchKMean es una variante del algoritmo K Means que utiliza mini lotes para reducir el tiempo de cálculo, mientras intenta optimizar la misma función objetivo. Los mini lotes son subconjuntos de los datos de entrada, muestreados aleatoriamente en cada iteración de entrenamiento. Estos mini lotes reducen drásticamente la cantidad de cómputo requerida para converger a una solución local. (SK-learn, 2018)

Finalmente, el puntaje de la silueta[5] es una métrica decisiva para medir la calidad del agrupamiento, este valor fue 0,155 para tres conglomerados, para cinco fue de 0.14

Grafica 13. Selección de clusters métrica inercia



Fuente: elaboración propia

Grafica 14. Selección de clusters métrica Davied Bouldin

Fuente: elaboración propia

5 El gráfico de la silueta se intentó realizar, pero dado el volumen de datos Python no arroja como tal el grafico. Por tanto, se toma solo la función de manera que se pueda obtener lo más importante que es el puntaje (score).

Si bien la calidad del agrupamiento con tres clusters fue la mejor, este resultado se puede superar o mejorar utilizando una red neuronal de Auto Encoder. Para ello fue necesario desplegar una red neuronal de 400 neuronas con 5 capas ocultas. Al final este algoritmo permite realizar una mejora en la calidad del agrupamiento de datos a partir del vector de clases de K-Means y los datos transformados del auto Encoder. Así las cosas, se logró obtener un score del 0.2026 lo cual mejora el resultado del puntaje del Mini Batch K-Means (0.155).

La importancia de estos clústeres radica que son la variable de salida del modelo, en otras palabras, son la clasificación del vector de etiquetas que se usan para interpretar las probabilidades estimada por los modelos de aprendizaje supervisado y aprendizaje profundo, siempre y cuando estos cumplan con el ajuste mínimo de las métricas de desempeño.

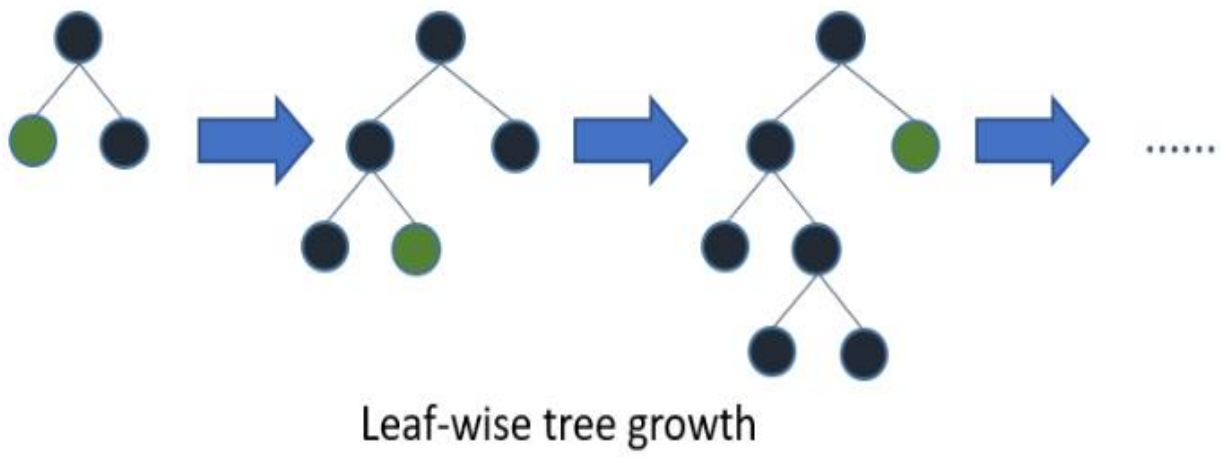
El modelo supervisado multiclase que se eligió fue el LGBM dado que su desempeño en todas de las métricas de desempeño fue superior a 99% con un costo computacional más eficiente respecto al Random Forest y la Regresión Logística, cuyos valores también rodearon el 99%.

El LGBM es un algoritmo de método de ensamble introducido por Microsoft en el año 2017, utiliza una técnica de muestreo basado en el gradiente (GOSS) lo que hace que sea más eficiente. Utiliza además dos técnicas novedosas: el muestreo unilateral basado en el gradiente y la agrupación exclusiva de características (EFB), que suplen las limitaciones del algoritmo basado en el histograma que se utiliza principalmente en todos los marcos GBDT (Gradient Boosting Decision Tree). Las dos técnicas de GOSS y EFB descritas forman las características del algoritmo (LGBM). Estás se combinan para hacer que el modelo funcione de forma eficiente y le proporcionan una ventaja sobre otros marcos GBDT (GeeksforGeeks, 2021)

Así pues, el gradiente representa la pendiente de la tangente de la función de pérdida, por lo que lógicamente si el gradiente de los puntos de datos es grande en algún sentido, estos puntos son importantes para encontrar el punto de división óptimo, ya que tienen un error más alto. (Swalin, 2019).

LGBM divide el árbol por hojas, a diferencia de otros algoritmos de refuerzo que hacen crecer el árbol por niveles. Elige la hoja con la máxima pérdida delta para crecer. Dado que la hoja es fija, el algoritmo por hojas tiene menos pérdidas que el algoritmo por niveles. El crecimiento del árbol a nivel de hoja puede aumentar la complejidad del modelo y puede llevar a un sobreajuste en conjuntos de datos pequeños. (GeeksforGeeks, 2021)

Gráfica 15. Diagrama de árbol - LGBM



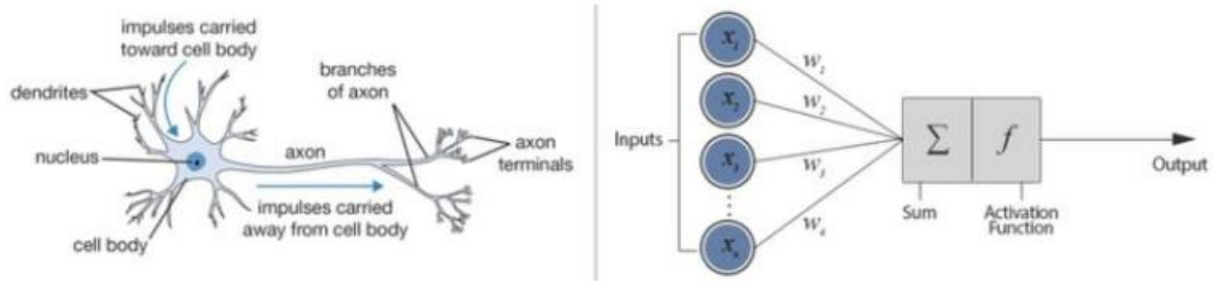
Fuente: Towars Data Science

En el modelo final multiclase LGBM (one vs Rest Classifier) se desarrolló bajo los siguientes parámetros: se usaron 100 árboles con profundidad de -1, el tipo de Boosting es el traditional Gradient Boosting Decision con un máximo de 31 hojas en cada árbol, el número de muestras para la construcción de los contenedores fue 200 mil.

La importancia del modelo supervisado en esta investigación reside en la estimación de las probabilidades, estás se asignan a cada registro de la base de datos por lo que es posible clasificar cada una de las muestras en el vector de etiquetas generadas por el agrupamiento. Es decir, se pueden obtener las probabilidades de que se presente un robo dado determinadas características (sexo, edad, estado civil, medio transporte, modalidad, arma medio, nombre barrio, código comuna, lugar, bien, categoría bien, año, mes, día, hora) de los individuos con buena precisión.

Redes Neuronales: las redes neuronales artificiales se basan en la estructura de una red neuronal biológica. Estas utilizan una serie de unidades de procesamiento interconectadas, llamadas neuronas artificiales, para calcular las entradas recibidas y producir los resultados de una suma ponderada de entradas. El resultado luego es alimentado a una función no lineal, llamada función de activación, para generar el resultado final.

Grafica 16. Modelo genérico de una neurona artificial de un perceptrón



Fuente: https://www.datacamp.com/community/tutorials/deep-learning-python

Como se puede observar en la gráfica 16 las entradas o inputs en la neurona artificial son las variables que se encuentren en el conjunto de datos. Estas representan a las dendritas. Luego, se tienen distintos pesos que multiplican a cada variable, para ser posteriormente sumados. Esto representa al núcleo y cuerpo de la neurona biológica. Finalmente, existe una función de activación que de sobrepasar un cierto umbral da como resultado una cierta salida, lo cual emula el funcionamiento de la neurona cerebral que recibe impulsos eléctricos. Si estos son lo suficientemente potentes y sobrepasan un cierto umbral de excitación, se transmitirá un nuevo impulso a través del axón, lo cual llegará a otras neuronas a través de sus terminales. (Roell, 2017)

Lo anterior se conoce como red neuronal artificial de una sola capa o perceptrón, y se refiere al caso en el cual se tiene una sola neurona. Para problemas más complejos, se utilizan múltiples neuronas y una o más capas.

Hay varios tipos de funciones de activación y su utilización depende del objetivo de la red y de la capa en la misma. Relu (Rectified Linear Unit) es la función de activación más común en capas

intermedias. Relu es una función que convierte en cero los valores negativos, definida por la ecuación 2 se publicó por primera vez Hahnloser et. al, 2000 (Digital selection and analog amplification co-exist in an electronic circuit inspired by neocortex)

(2)

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

Para los casos donde la red es utilizada para clasificar, la capa final utiliza una función sigmoide, de modo que el resultado de la red pueda ser interpretado como una probabilidad. Una función sigmoide convierte todos los valores de entrada al intervalo 0-1 y se define por medio de la ecuación 3. Comúnmente ente las neuronas se organizan en capas, y a su vez, una red neuronal puede estar compuesta por múltiples capas.

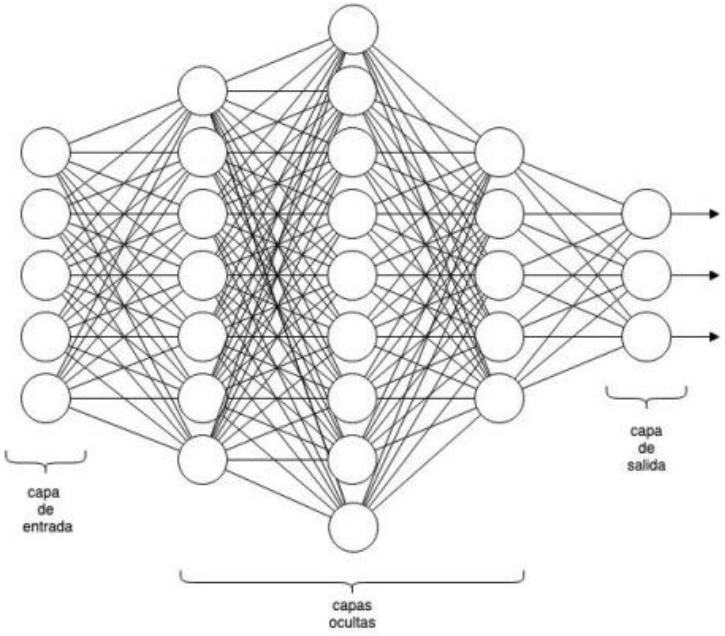
(3)

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

La capa de entrada simplemente pasa los valores de entrada a la próxima capa. Las capas siguientes a la capa de entrada son llamadas capas ocultas por no estar directamente expuestas a los valores de entrada. A la última capa se le llama 'capa de salida' y es la responsable de devolver el, o los, resultados calculados por la red. A las redes organizadas de esta forma, con una capa de entrada, una o más capas ocultas y una de salida se las conoce también como perceptrón multicapa (MLP, por su sigla en inglés). Cuando todas las neuronas de una capa están conectadas a cada neurona de la capa anterior se le llama capa densamente conectada. (Kock A. B. & Teräsvirta T, 2016)

Hasta el momento se ha realizado una breve descripción de las redes neuronales más simples, sin embargo, para el presente trabajo hemos utilizado redes que van más allá de las características básicas de una red neuronal. Surge entonces, la necesidad de trabajar con las redes neuronales convolucionales –CNN.

Grafica 17. Red Neuronal



Fuente: Towards Data Science

La Red Reuronal Convolucional es un tipo de Red Neuronal Artificial que procesa sus capas imitando al cortex visual del cerebro humano para identificar distintas características en las entradas. Para ello, la CNN contiene varias capas ocultas especializadas y con una jerarquía: esto significa que las primeras capas detectan propiedades o formas básicas y se van especializando hasta llegar a capas más profundas capaces de reconocer formas complejas como un rostro o una silueta (PyTorch – Cleverpy, 2018).

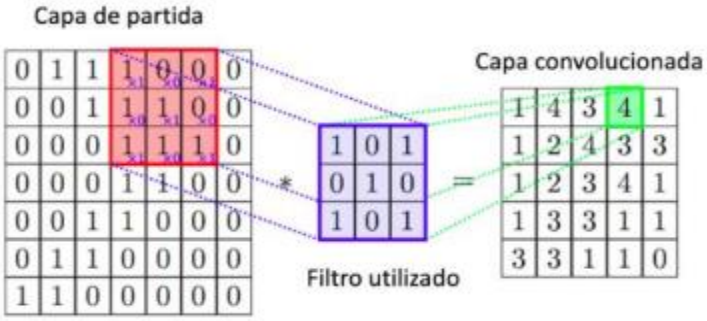
La red neuronal por sí misma debe reconocer una gran cantidad de imágenes para poder captar las características únicas de cada objeto y a su vez poder generalizarlo. Cada imagen se trata de una matriz de pixeles cuyo valor va de 0 a 255 pero se normaliza para la red neuronal de 0 a 1. Como punto de partida la red toma como entrada los píxeles de una imagen. En el caso del presente trabajo, las imágenes serán a color y las entradas serán de 150x150x3 píxeles de alto y ancho, por lo que habrá en total 67500 neuronas.

Luego de estos primeros pasos inicia el procesado distintivo de las CNN, lo que se conoce como las convoluciones. Una convolución consiste en tomar grupos de píxeles cercanos de la imagen de

entrada e ir operando matemáticamente contra una pequeña matriz a la que se denomina kernel conocida como filtro. Ese kernel recorre todas las neuronas de entrada – de izquierda a derecha y de arriba hacia abajo – y genera una nueva matriz de salida, que será la nueva capa de neuronas ocultas, y que también se conoce como la matriz de activación. La convolución será tal si y sólo si el kernel es real y simétrico (Calvo, 2018)

A medida que el kernel se va trasladando obtenemos una nueva imagen filtrada por este. Una vez la imagen realiza una convolución con un kernel, aplica la función de activación. Esta función se encarga de devolver una salida a partir de un valor de entrada, normalmente en un rango determinado como (0,1) o (-1,1) (Calvo, 2018). En general, se buscan funciones cuyas derivadas sean simples para minimizar así la carga computacional.

Grafica 18. – Uso del filtro en la convolución



Fuente: Diego Calvo – Red Convulucional. https://www.diegocalvo.es/red-neuronal-convolucional/

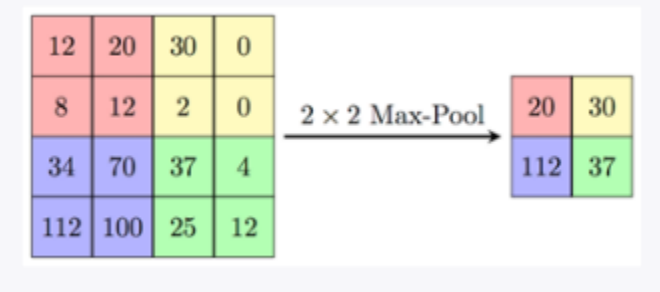
La función de activación llamada Softmax, (función exponencial normalizada) es una generalización de la función sigmoidea y popularizada por las redes neuronales convolucionales. Se utiliza como función de activación de salida para la clasificación multiclase porque escala las entradas precedentes de un rango entre 0 y 1 y normaliza la capa de salida, de modo que la suma de todas las neuronas de salida sea igual a la unidad (FA Softmax – Numerentur.org, 2019).

Una vez culminado el proceso anterior, se reduce la cantidad de neuronas antes de realizar una nueva convolución. Como se comentó anteriormente, partiendo de una imagen de 150x150x3 tenemos una primera capa de entrada de 67500 neuronas y luego de la primera convolución

obtenemos una capa oculta de más de 100.000. Si se llevara a cabo una nueva convolución a partir de esta capa, el número de neuronas de la próxima capa crecería exponencialmente implicando un mayor procesamiento. Para disminuir el tamaño de la próxima capa se realiza el proceso de subsampling en el que se reduce el tamaño de las imágenes filtradas, pero donde prevalecen las características más importantes que detectó cada filtro (Max-Pooling / Pooling - Computer Science Wiki, 2018).

Hay diversos tipos de subsampling, pero el más utilizado es el Max-Pooling. Si suponemos que se realiza una Max-pooling de tamaño 2x2 se recorrerán cada una de las 32 imágenes de características obtenidas anteriormente de 50x50 pixeles de izquierda a derecha, arriba-abajo, pero en lugar de a un solo pixel se toman de 2x2 – 2 de alto por 2 de ancho – y se preserva el valor más elevado de esos 4 pixeles, de ahí el termino Max. En este caso, usando 2x2 la imagen resultante se reduce a la mitad y quedará una de 25x25 pixeles. Después de este proceso de subsampling quedan 32 imágenes de 25x25, pasando de haber tenido 80.000 neuronas a 20.000. El descenso es considerable y teóricamente almacenan la información más importante para detectar las características deseadas (Bagnato, 2020)

Grafica 19. – Maxpooling



Fuente: Aprende Machine Learning https://www.aprendemachinelearning.com/como-funcionan-las-convolutional-neural-networks-vision-por-ordenador/

Como punto de finalización se toma la última capa oculta a la que se le realizó el proceso de subsampling, que se dice que es tridimensional y se procesa para que deje de serlo pasando a ser una capa de neuronas tradicionales. Entonces a esta última se le aplica la función de activación Softmax que conecta contra la capa de salida final que tendrá la cantidad de neuronas

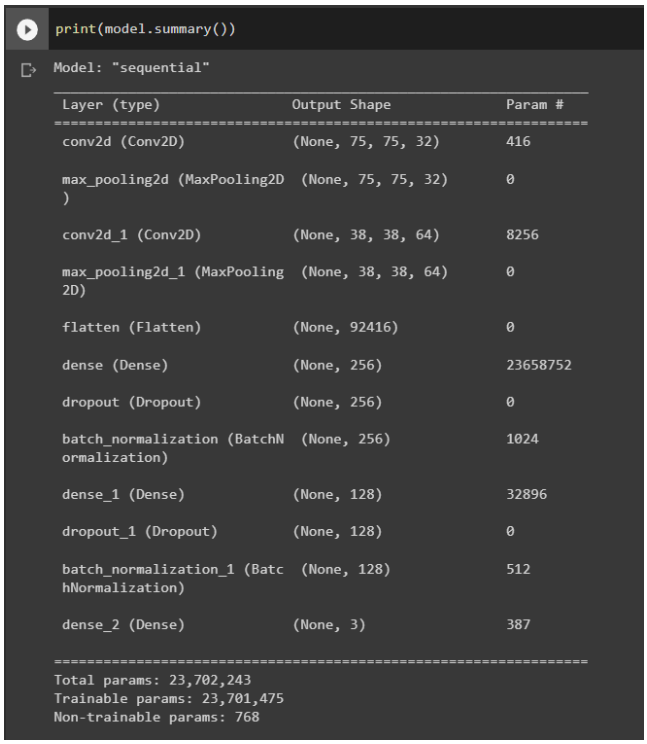
correspondientes con las clases que estamos clasificando. Las salidas en el momento del entrenamiento tendrán el formato conocido como one-hot-encoding que resulta un grupo de bits cuyas combinaciones de valores válidas son solo aquellas en las que, sí aparece un bit a uno, el resto ha de estar a cero.

También se utilizó el método llamado **Dropout,** es un método que desactiva un número determinado de neuronas en una red neuronal de forma aleatoria Las neuronas desactivadas no se tienen en cuenta para la propagación hacia delante ni para atrás, lo que obliga a las neuronas cercanas a no depender tanto de las neuronas desactivadas.

Así las cosas, se construyó una red convolucional de arquitectura sequential, se utilizó la librería Keras de Tensorflow la función keras.models.Sequential que permite crear un modelo de este tipo, al cual se le puede configurar los parámetros necesarios para una red neuronal profunda de clasificación. Como tal la arquitectura del modelo sequential quedo de la siguiente manera: 3 capas convolucionales, una capa Flatten y 3 capas densas, para estas capas densas se utilizó la función de activación llamada “Relu”, adicionalmente se le agregaron 2 capas de MaxPooling de (2,2) y 2 capas Dropout de (0.2), al final el modelo nos arrojó la cantidad de 23.702.243 parámetros para entrenar.

Para compilar el modelo utilizamos el optimizador llamado “Adam” porque es el optimizador que mejor se comporta con este tipo de modelos, para la función de perdida utilizamos la “categorical_crossentropy” porque es la más adecuada para el número de clases que estamos manejando, por último, en el parámetro metrics utilizamos el accuracy (precisión). Para la función de ajuste o cantidad de iteraciones se usaron 50 épocas.

Imagen 2. Arquitectura red convolucional sequential



```
print(model.summary())

Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d (Conv2D)             (None, 75, 75, 32)        416

 max_pooling2d (MaxPooling2D  (None, 75, 75, 32)       0
 )

 conv2d_1 (Conv2D)           (None, 38, 38, 64)        8256

 max_pooling2d_1 (MaxPooling  (None, 38, 38, 64)       0
 2D)

 flatten (Flatten)           (None, 92416)             0

 dense (Dense)               (None, 256)               23658752

 dropout (Dropout)           (None, 256)               0

 batch_normalization (BatchN  (None, 256)              1024
 ormalization)

 dense_1 (Dense)             (None, 128)               32896

 dropout_1 (Dropout)         (None, 128)               0

 batch_normalization_1 (Batc  (None, 128)              512
 hNormalization)

 dense_2 (Dense)             (None, 3)                 387

=================================================================
Total params: 23,702,243
Trainable params: 23,701,475
Non-trainable params: 768
```

Fuente: elaboración propia

4.4 MÉTRICAS

Para el cálculo de las métricas de desempeño se recurrió principalmente a scikit-learn ya que en esta librería se encuentran la mayoría de métricas que se usan en aprendizaje de máquinas. En efecto se usaron: 1) sklearn.metrics para obtener la precisión de la clasificación, la precisión, Recall, F1 el AUC. Estas medidas se obtuvieron a partir del resultado de la predicción hecha en el conjunto de datos entrenamiento que posteriormente se compara con el conjunto de datos de prueba. 2) Model selection que se usó principalmente para calcular la validación cruzada y la búsqueda de hiper parámetros. 3) De nuevo con sklearn.metrics se importa classification_report con lo cual se obtiene un reporte resumido de la validación del modelo. 4) para la red neuronal se utiliza la precisión como métrica de desempeño principal del modelo.

5. METODOLOGÍA

5.1 BASELINE

Se hicieron varios laboratorios o iteraciones para encontrar el mejor resultado en los diferentes modelos desarrollados. Para el agrupamiento se probaron varias metodologías, por ejemplo, se probó el modelo no supervisado DB SCAN el cual no fue eficiente por la cantidad datos debido a la alta complejidad computacional del algoritmo y la alta dimensionalidad de los datos que se trabajaron, el mismo caso paso con el clúster aglomerativo (jerárquico). Así pues, las metodologías que mejor se ajustaron y se llevaron a cabo fueron los modelos no supervisados de Fuzzy C Means y Mini Batch K-Means. Entre estos dos modelos supervisados se optó por escoger el de K-Means dado que el puntaje de la silueta era más alto que el obtenido con el Fuzzy C Means.

Ahora bien, para el modelo supervisado multiclase se probaron 3 diferentes modelos. En la primera iteración se probó un Random Forest que tuvo un excelente desempeño y la regresión logística la que también tuvo un exactitud y precisión bastante bueno, no obstante, el costo computacional era muy alto y se demoraba mucho la convergencia del algoritmo. Finalmente, el LGBM fue muy eficiente respecto a la convergencia y a los resultados obtenidos por lo tanto se escogió este modelo de ensamble.

Se probaron dos arquitecturas diferentes en el modelo de aprendizaje profundo. La red neuronal convolucional de arquitectura VGG16 que en la primera iteración arrojó un resultado muy bajo, ya que el resultado de la precisión no superó el 20% a pesar que probaron varios parámetros. A raíz de esto, se optó por emplear una red convolucional de tipo secuencial la cual arrojos mejores resultados, sin embargo, no alcanzó la precisión mínima propuesta. En cuanto a problemas técnicos, debido a la cantidad de imágenes a procesar, tuvimos inconvenientes con el tiempo de ejecución del modelo ya que la potencia de hardware que teníamos era muy limitada.

5.2 VALIDACIÓN

Para el modelo supervisado se destinó un conjunto de entrenamiento que contiene el 70% de las muestras de las bases de datos, el 30% restante para el conjunto de prueba. Cabe mencionar que la partición se realizó utilizando un muestreo estratificado de los datos utilizando siempre la misma semilla para que los resultados de dicha partición no varíen.

Se realizó además la validación cruzada con 5 pliegues. Esto se hace dividiendo el conjunto de datos de manera aleatoria en dos subconjuntos: entrenamiento y prueba, comúnmente 80% - 20% respectivamente. El conjunto de entrenamiento a su vez se divide nuevamente de manera aleatoria en 5 subconjuntos disyuntos, se usan 5-1 subconjuntos para entrenar y el conjunto restante para validar; dicho proceso se repite 5 veces. El proceso de entrenamiento y validación se utiliza como parte de los hiper parámetros del modelo.

Para el modelo de red neuronal convolucional se definió para el proceso de entrenamiento un total de 15.000 imágenes y para el proceso de validación 5.000 imágenes, es importante hacer esta separación para obtener de la forma más fiable posible la real predictibilidad de los modelos. En esta ocasión el conjunto de validación se utiliza para la evaluación final del modelo ya que los datos pertenecientes a este conjunto no se utilizaron para entrenar la red.

Además, se realizaron otras evaluaciones relacionadas con la relevancia de las variables (entropías de Shannon) y matriz de correlación de variables. Con lo primero, se confirmó que las variables seleccionadas en los modelos aportaban todas información. En segundo lugar, se consiguió revisar que las variables no presentaran una relación fuerte entre sí, de tal manera que no se identificó presencia de colinealidad.

5.3 ITERACIONES y EVOLUCIÓN

Se realiza un conjunto de iteraciones para definir las opciones de preprocesamiento (detección de datos atípicos), análisis estadísticos (mapas de calor, histogramas). Luego se hizo unas iteraciones probando distintos modelos, específicamente se utilizó un Mini Batch K-Means mejorado con una red neuronal de auto enconder, el LGBM y al final ejecutamos una red neuronal convolucional de tipo secuencial.

5.4 HERRAMIENTAS

Google Colab: Colaboratory, o "Colab" para abreviar, es un producto de Google Research. Permite a cualquier usuario escribir y ejecutar código arbitrario de Python en el navegador. Es especialmente adecuado para tareas de aprendizaje automático, análisis de datos y educación.

Python: Python es un lenguaje de programación creado en 1991 por Guido Van Rossum en el Centro para las Matemáticas y la Informática (CWI, Centrum Wiskunde & Informatica) de los Países Bajos. Python ofrece muchos beneficios, lo que significa que poco a poco se está convirtiendo en el lenguaje más utilizado para el aprendizaje profundo.

TensorFlow: es un software de computación numérica creado por Google, orientado a problemas de Deep Learning. Posee interesantes integraciones con otras bibliotecas del ecosistema como Keras, de la que se hablará más adelante. En el proyecto se explotará la utilidadque presenta para construir y entrenar redes neuronales.

Keras: es una API de redes neuronales a alto nivel, escrita en Python y que puede emplearse haciendo uso de TensorFlow, CNTK o Theano. Fue desarrollada con el enfoque de permitir una experimentación rápida, es decir, pasar de la idea al resultado con el menor retraso posible.

Power BI: es un sofware de análisis y visualización de datos creado por Microsoft en el año 2016. Con este programa se realizaron algunas gráficas y algunos análisis del valor esperado.

6. RESULTADOS

6.1 MÉTRICAS

Los resultados del modelo multiclase de aprendizaje profundo en el conjunto de prueba se describe a continuación:

Se logró obtener un puntaje de precisión en la clasificación 0,9973 una precisión balanceada de la clasificación de 0,9958, esto es las etiquetas que logro clasificar correctamente. El resultado de la precisión fue 0,9961, o sea que la precisión es la capacidad que tuvo el clasificador de no etiquetar como positiva una muestra que es negativa. La Memorización que es el Recall, recuerda los verdaderos positivos de los falsos negativos fue de 0,9958, el F1 0.9960 y el área bajo la curva (AUC) es de 0,999. Los anteriores resultados señalan que el modelo tuvo un excelente ajuste y desempeño. En la tabla 2 se resumen los resultados de la validación y en la tabla 3 observa la matriz de confusión.

Tabla 2. Resumen validación del modelo supervisado LGBM

	precision	recall	f1-score	support
0	1.00	1.00	1.00	29124
1	0.99	0.99	0.99	10248
2	1.00	1.00	1.00	28533
accuracy			1.00	67905
macro avg	1.00	1.00	1.00	67905
weighted avg	1.00	1.00	1.00	67905

Fuente: elaboración propia

Tabla 3. Matriz de confusión

```
[[[38684    97]
  [   83 29041]]

 [[57574    83]
  [   97 10151]]

 [[39372     0]
  [    0 28533]]]
```

Fuente: elaboración propia

Ahora bien, con los resultados obtenidos de las probabilidades del modelo multiclase se pudo clasificar a los individuos con una excelente precisión (tal como se muestra en la matriz de confusión), en cada clúster que es el vector de etiquetas del modelo. Lo anterior se podría interpretar como las probabilidades de que sucedan hurtos a personas dado determinadas características teniendo en cuenta las respectivas descripciones de cada clúster.

Así las cosas, los 3 clusters hallados tienen las siguientes características.

Clúster 0: cantidad mayoritaria de hombres que comúnmente les hurtan celulares y dinero con armas de fuego en las comunas 10 y 11 principalmente en los barrios la Candelaria y Villanueva entre las 6pm y 8 pm en vía pública especialmente cuando van caminando.

Clúster 1: cantidad menor de hombres que les roban celulares y dinero sin uso de algún arma en las comunas 10 y 14 particularmente en los barrios La Candelaria y el Poblado especialmente en la mañana tipo 10 am y en la tarde tipo 6 pm especialmente cuando están en alguna estación del metro, almacén o tienda y se desplazan caminando.

Clúster 2: son mujeres que normalmente les hurtan celulares y dinero sin uso de armas en las comunas 10 y 11 frecuentemente en los barrios la Candelaria y barrio Colon entre las 5 pm y 7 pm especialmente cuando se encuentran en vía pública y se desplazan caminando.

Por ejemplo, en el clúster 2 el modelo pudo clasificar correctamente todo, o sea que pudo estimar que las probabilidades de hurtar a una mujer con tales características es en promedio un 95%., este comportamiento es muy similar en los otros clusters.

En las siguientes tablas se observa como fue el desempeño de la clasificación de algunas de las muestras de la base de datos. La columna cluster es el ya mencionado vector de etiquetas y al lado se ve la probabilidad de pertenecer a esa clase.

Tabla 4. Clasificación del modelo con sus respectivas probabilidades (etiqueta 2)

medio_transporte	modalidad	arma_medio	Barrio	codigo_comun	lugar	bien	categoria_bien	Year	Month	Day	hour	Cluster	Probabilidad
Caminata	Descuido	No	San Pablo	1	Hotel, mote	Celular	Tecnología	2015	12	19	12	2	99,9857 %
Caminata	Atraco	Arma cortopunzante	Popular	1	Vía pública	Celular	Tecnología	2017	12	4	21	2	99,9920 %
Caminata	Atraco	Arma cortopunzante	Popular	1	Vía pública	Celular	Tecnología	2017	5	27	3	2	99,9893 %
Caminata	Atraco	Arma cortopunzante	Villa Guadalupe	1	Vía pública	Celular	Tecnología	2018	5	14	6	2	99,9856 %
Caminata	Atraco	Arma cortopunzante	Santo Domingo Savio No.1	1	Vía pública	Celular	Tecnología	2016	4	5	17	2	99,9885 %
Caminata	Atraco	Arma cortopunzante	Moscú No.2	1	Vía pública	Celular	Tecnología	2020	2	12	5	2	99,9856 %
Caminata	Atraco	Arma cortopunzante	Carpinelo	1	Vía pública	Celular	Tecnología	2016	4	4	5	2	99,9862 %
Caminata	Atraco	Arma cortopunzante	Villa Guadalupe	1	Vía pública	Celular	Tecnología	2020	2	22	6	2	99,9893 %
Caminata	Atraco	Arma cortopunzante	Granizal	1	Vía pública	Celular	Tecnología	2017	3	9	11	2	99,9925 %
Caminata	Atraco	Arma cortopunzante	La Avanzada	1	Vía pública	Celular	Tecnología	2015	11	4	6	2	99,9896 %
Caminata	Atraco	Arma cortopunzante	San Pablo	1	Vía pública	Celular	Tecnología	2021	6	6	1	2	99,9906 %
Caminata	Atraco	Arma cortopunzante	Santo Domingo Savio No.1	1	Vía pública	Celular	Tecnología	2020	1	30	15	2	99,9912 %
Caminata	Atraco	Arma cortopunzante	San Pablo	1	Vía pública	Celular	Tecnología	2016	10	30	21	2	99,9897 %
Caminata	Atraco	Arma cortopunzante	Santo Domingo Savio No.1	1	Vía pública	Celular	Tecnología	2016	8	3	15	2	99,9927 %
Caminata	Atraco	Arma cortopunzante	Moscú No.2	1	Vía pública	Celular	Tecnología	2018	5	15	6	2	99,9899 %
Caminata	Atraco	Arma cortopunzante	La Avanzada	1	Vía pública	Celular	Tecnología	2018	11	8	4	2	99,9883 %
Caminata	Atraco	Arma cortopunzante	San Pablo	1	Vía pública	Celular	Tecnología	2018	12	9	0	2	99,9890 %
Caminata	Atraco	Arma cortopunzante	Santo Domingo Savio No.2	1	Vía pública	Celular	Tecnología	2018	11	16	0	2	99,9889 %
Caminata	Atraco	Arma cortopunzante	Santo Domingo Savio No.1	1	Vía pública	Celular	Tecnología	2017	5	9	17	2	99,9925 %
Caminata	Atraco	Arma cortopunzante	Santo Domingo Savio No.2	1	Vía pública	Celular	Tecnología	2017	12	23	0	2	99,9903 %
Caminata	Atraco	Arma cortopunzante	Santo Domingo Savio No.2	1	Vía pública	Celular	Tecnología	2017	4	26	22	2	99,9900 %
Caminata	Atraco	Arma cortopunzante	Santo Domingo Savio No.1	1	Vía pública	Celular	Tecnología	2017	7	31	14	2	99,9915 %
Caminata	Atraco	Arma cortopunzante	Popular	1	Vía pública	Celular	Tecnología	2017	3	15	16	2	99,9922 %
Caminata	Atraco	Arma cortopunzante	Popular	1	Vía pública	Celular	Tecnología	2019	5	2	21	2	99,9916 %
Caminata	Atraco	Arma cortopunzante	Villa Guadalupe	1	Vía pública	Celular	Tecnología	2019	2	26	10	2	99,9914 %

Fuente: elaboración propia

En esta última tabla se ven algunas muestras que arrojan una probabilidad relativamente baja de caer en el clúster 1. De ahí que tal vez en este grupo la clasificación presentó algunos falsos positivos, no obstante, la gran mayoría de las muestras estuvieron bien clasificadas con probabilidades del 99%.

Tabla 5. Clasificación del modelo con sus respectivas probabilidades (etiqueta 1)

medio_transporte	modalidad	arma_medio	Barrio	codigo_comun	lugar	bien	categoria_bien	Year	Month	Day	hour	Cluster	Probabilidad
Caminata	Atraco	Arma cortopunzante	Lorena	11	Edificio	Celular	Tecnología	2019	4	18	20	1	50,0001 %
Sin dato	Otras formas de	Otras armas	Suramericana	11	Parqueader	Radio	Tecnología	2004	10	14	8	1	50,0761 %
Caminata	Atraco	No	Bolivariana	11	Otros lugar	Cédula	Documentos	2009	9	2	16	1	50,1444 %
Caminata	Atraco	No	Bolivariana	11	Otros lugar	Cédula	Documentos	2009	9	2	16	1	50,1444 %
Caminata	Rompimiento ce	Objeto contundente	Barrio Colón	10	Turístico	Bono	Dinero, joyas, piedra	2019	11	6	10	1	50,1514 %
Caminata	Atraco	No	Boston	10	Plaza de me	Sin dato documentos	Documentos	2020	9	1	13	1	50,3475 %
Automóvil	Rompimiento ce	Objeto contundente	La Candelaria	10	Parqueader	Computador	Tecnología	2005	9	1	12	1	50,4071 %
Automóvil	Rompimiento ce	Objeto contundente	La Candelaria	10	Parqueader	Computador	Tecnología	2005	9	1	12	1	50,4071 %
Caminata	Rompimiento ce	No	San Benito	10	Otros lugar	Computador	Tecnología	2021	8	3	11	1	50,5168 %
Caminata	Rompimiento de	Objeto contundente	El Diamante No.2	14	Parque	Sin dato tecnología	Tecnología	2019	7	26	19	1	50,5188 %
Automóvil	Atraco	Arma de fuego	La América	12	Hospital o c	Sin dato tecnología	Tecnología	2020	7	28	12	1	50,5509 %
Automóvil	Rompimiento ce	Objeto contundente	Belén	16	Restaurante	Peso	Dinero, joyas, piedra	2018	9	28	21	1	50,5800 %
Taxi	Atraco	Arma de fuego	La Florida	14	Metro Plus	Sin dato documentos	Documentos	2019	6	15	2	1	50,6458 %
Caminata	Descuido	No	Manila	14	Puesto de t	Autopartes	Autoparte y elemen	2020	11	8	6	1	50,6484 %
Caminata	Atraco	Arma cortopunzante	La Candelaria	10	Iglesia	Celular	Tecnología	2019	11	18	7	1	50,6935 %
Caminata	Descuido	No	El Rincón	16	Parqueader	Otros bienes	Otros elementos	2020	12	22	10	1	50,7640 %
Caminata	Atraco	Arma de fuego	Jardín Botánico	4	Parque	Peso	Dinero, joyas, piedra	2018	12	27	0	1	50,7825 %
Caminata	Rompimiento ce	Objeto contundente	Diego Echavarría	16	Vehículo pa	Peso	Dinero, joyas, piedra	2020	7	1	0	1	50,8537 %
Caminata	Atraco	Arma cortopunzante	La Candelaria	10	Edificio	Celular	Tecnología	2017	4	18	20	1	50,9696 %
Caminata	Rompimiento ce	Objeto contundente	Florencia	5	Vía pública	Tarjeta bancaria	Dinero, joyas, piedra	2019	9	15	4	1	50,9747 %
Caminata	Rompimiento de	Objeto contundente	Santa María de los Ángeles	14	Parqueader	Sin dato tecnología	Tecnología	2019	8	22	20	1	50,9762 %
Caminata	Descuido	No	Jesús Nazareno	10	Institución (	Computador	Tecnología	2018	4	12	11	1	51,0012 %
Caminata	Atraco	Arma cortopunzante	Bomboná No.1	10	Institución (	Celular	Tecnología	2020	10	3	1	1	51,1833 %
Taxi	Atraco	Arma de fuego	Jesús Nazareno	10	Fábrica o en	Celular	Tecnología	2021	1	13	2	1	51,3809 %
Caminata	Atraco	Arma cortopunzante	Gerona	9	Otros lugar	Peso	Dinero, joyas, piedra	2018	12	14	0	1	51,3903 %

Fuente: elaboración propia

El resultado final de la red neuronal demostró que tal vez por no haber podido emplear un número excesivo de imágenes para entrenar la red a pesar que se aumentaron las épocas los resultados obtenidos no son tan positivos, ya que genera una precisión que solo alcanza hasta el 45% y no supera la precisión mínima del 80%. Es decir, que con el modelo red neuronal convolucional no fue posible clasificar con una buena precisión los lugares donde ocurren hurtos, no fue posible estimar las probabilidades de hurtos en las zonas de mayor y menor frecuencias de robos.

6.2 EVALUACIÓN CUALITATIVA

Para validar si existe sobreajuste o sub ajuste en los resultados del modelo de aprendizaje supervisado se recurre a analizar los resultados obtenidos de la validación cruzada, de la precisión en los conjuntos de entrenamiento y prueba. A partir de lo anterior se obtuvo lo siguiente:

- El puntaje de precisión del modelo en el conjunto entrenamiento fue de 0,9994.
- El puntaje de precisión del modelo en el conjunto entrenamiento fue de 0,9973.
- El puntaje de la validación cruzada en los 5 pliegues fueron los siguientes: 0,99405313 – 0,99515748 – 0,99461341 – 0,99352229 – 0,99195943.

Por lo tanto, se concluye que el modelo no presente ningún tipo de problema.

6.3 CONSIDERACIONES DE PRODUCCIÓN

El modelo de clasificación LGBM tal vez se podría implementar en algún tipo de software que en tiempo real pueda estimar las probabilidades de que las personas puedan ser hurtadas de acuerdo a determinadas características. De hecho, en el mercado ya existen algunos softwares que recurren a modelos de aprendizaje supervisado que se ocupan de recopilar y analizar datos sobre crímenes, el algoritmo intenta predecir dónde y cuándo será más probable que ocurra un cierto tipo de crimen.

7. CONCLUSIONES

En este trabajo se realizó la construcción de un modelo de aprendizaje supervisado con el cual se logró clasificar con una alta precisión las personas hurtadas de acuerdo con determinadas características. También se analizaron los lugares de mayor ocurrencia de hurtos en Medellín tomando como referencias las zonas turísticas y de mayor comercio encontrando que sectores que antes era relativamente seguros actualmente ya no lo son. Los hurtos a personas se han incrementado de una manera acelerada, tanto así que de acuerdo con el indicador calculado durante el 2019 se incrementaron en más de 2900% con respecto al 2011.

Las redes neuronales convolucionales de tipo sequential que permita la predicción del delito de hurto a personas en determinadas zonas de la ciudad de Medellín no obtuvo los resultados esperados. Este modelo fue entrenado con datos extraídos del Sistema de Información para la Seguridad y la Convivencia (SISC), cuyas cifras documentan los reportes de hurto a personas en la ciudad de Medellín desde el año 2003 hasta el año 2020, teniendo en cuenta la ubicación del suceso, objeto robado, localidad o comuna, bien hurtado, tipo de arma, género de la víctima, y fecha, los cuales se encuentran de manera pública en la web http://medata.gov.co/dataset/hurto-persona.

Con respecto al modelo sequential de la red neuronal convolucional podemos concluir que, para el objetivo propuesto en el inicio de este trabajo, este tipo de modelo no es el más idóneo para abordar la problemática, debido a que los resultados obtenidos no son tan positivos, ya que la precisión del modelo no supera el resultado mínimo del 80% que nos habíamos presupuestado, es decir que con estos no fue posible estimar con una buena precisión las probabilidades de hurto en determinadas zonas de la ciudad.

8. REFERENCIAS BIBLIOGRÁFICAS

Aguirre, K. (2017) *Violencia y criminalidad tras la implementación de los acuerdos de paz*. Recuperado de Razón Pública: https://razonpublica.com/violencia-y-criminalidad-tras-la-implementacion-de-los-acuerdos-de-paz/

Bagnato, J (2020, 25 junio). Convolutional Neural Networks: La Teoría explicada en Español. Aprende Machine Learning. https://www.aprendemachinelearning.com/como-funcionan-las-convolutional-neural-networks-vision-por-ordenador/

Calvo, D. (2018, 8 diciembre). *Red Neuronal Convolucional CNN*. Diego Calvo. https://www.diegocalvo.es/red-neuronal-convolucional/#:~:text=Definici%C3%B3n%20de%20Red%20Neuronal%20Convolucional,el%20entrenamiento%20es%20m%C3%A1s%20r%C3%A1pido

Clustering. (2018). Scikit-Learn. https://scikit-learn.org/stable/modules/clustering.html#mini-batch-kmeans

GeeksforGeeks. (2021, 22 diciembre). *LightGBM (Light Gradient Boosting Machine).* https://www.geeksforgeeks.org/lightgbm-light-gradient-boosting-machine/

Gélvez, J. D. (2018). ¿Cuáles determinantes se relacionan con la percepción de inseguridad? Un análisis estadístico y espacial para la ciudad de Bogotá, D.C. *Revista Criminalidad*, 61 (1), 69-84.

FA Softmax – Numerentur.org. (2019). Numerentur. https://numerentur.org/funcion-de-activacion-softmax/

Max-pooling / Pooling - Computer Science Wiki. (2018, 27 febrero). Computer science wiki. https://computersciencewiki.org/index.php/Max-pooling_/_Pooling

Swalin, A. (2019, 12 junio). *CatBoost vs. Light GBM vs. XGBoost - Towards Data Science.* Recuperado de. https://towardsdatascience.com/catboost-vs-light-gbm-vs-xgboost-5f93620723db

Como crear una red convolucional con PyTorch – Cleverpy. (2018, 5 enero). Cleverpy Machine Learning. https://cleverpy.com/2018/01/05/red-convolucional-pytorch/.

9. ANEXOS

https://github.com/mafernga/prediccion_hurto_a_personas