Análisis de agrupamiento de distintos tipos de hurtos en la ciudad de Medellín

Sebastián Franco Franco
Alba Julieth Giraldo Martínez

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

Asesor
Javier Fernando Botia Valderrama, PhD en Ingeniería Electrónica

Universidad de Antioquia
Facultad de Ingeniería
Especialización en Analítica y Ciencia de Datos
Medellín, Antioquia, Colombia
2025

Cita	(Franco Franco & Giraldo Martínez, 2025)
Referencia **Estilo APA 7 (2020)**	Franco Franco, S., & Giraldo Martínez, A. J. (2025). *Análisis de agrupamiento de distintos tipos de hurtos en la ciudad de Medellín* Trabajo de grado especialización]. Universidad de Antioquia, Medellín, Colombia.

Especialización en Analítica y Ciencia de Datos, Cohorte VIII.

Centro de Investigación Ambientales y de Ingeniería (CIA).

Centro de Documentación Ingeniería (CENDOI)

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

Tabla de contenido

Lista de Figuras

Lista de Tablas

Siglas, acrónimos y abreviaturas

APA	American Psychological Association
Esp.	Especialista
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
LGBM	Light Gradient-Boosting Machine
MSc	Magister Scientiae
PhD	Philosophiae Doctor
PostDoc	PostDoctor
RBF	Radial Basis Function
UdeA	Universidad de Antioquia

Análisis de agrupamiento de distintos tipos de hurtos en la ciudad de Medellín

***Resumen*— El presente estudio aborda los datos correspondientes a las denuncias de distintos tipos de hurtos ocurridos en la ciudad de Medellín desde enero de 2003 hasta marzo de 2023. Con el propósito de caracterizar el comportamiento de los hurtos exhibido en los datos, se implementan tres algoritmos de agrupamiento: *K-Means*, *HDBSCAN* y *K-Modes*. Además, se interpretan los rasgos principales de los modelos y de los grupos generados. Se encontró que, desde una perspectiva geográfica, al emplear la longitud y latitud de los puntos referidos, se obtuvieron resultados significativos. De acuerdo con el análisis realizado, el modelo *HDBSCAN* genera los grupos más compactos y separados. Y el modelo que permitió la mejor caracterización de los hurtos fue *K-Modes* que el modelo que permitió la mejor caracterización de los hurtos fue *K-Modes*, probado con distintos conjuntos de variables de los datos. La normalización y la reducción de dimensionalidad en los datos con cada modelo probado tienen efectos disímiles, pero suelen enriquecer las características de los grupos. En lo que respecta a los hurtos, se observó que los casos más denunciados involucran a hombres solteros de mediana edad como víctimas principales. En suma, se observó que la zona del centro urbano registra el mayor número de denuncias, mientras que los corregimientos presentan el menor número de casos.**

Keywords*— *hurtos, aprendizaje no supervisado, análisis de agrupamiento, denuncias.

1. Introducción

Una de las preocupaciones más relevantes tanto a nivel gubernamental como individual es la seguridad, y dentro de ésta, la problemática de los robos. La analítica de datos del crimen puede representar oportunidades relevantes en este caso. Ya que, según Amoako (2021), la identificación de “puntos calientes" y de la distribución del crimen es útil para la selección, implementación y asignación de una respuesta que pueda ser adecuada para la reducción de crímenes, y comenta a su vez, que el estudio de hurtos en zonas urbanas es esencial para que la fuerza pública pueda priorizar acciones que promuevan su prevención.

Durante el periodo comprendido entre 2016 y 2019, han aumentado las denuncias por hurto a personas, comenzando con 13.326 en el 2016, alcanzando un pico de 26.700 en 2019 y reduciéndose considerablemente en el 2020. En cuanto al hurto a establecimientos comerciales, este tuvo un crecimiento significativo entre el 2016 y 2017 con 1.831 y 4.386 respectivamente, y aunque tuvo una reducción relativamente pequeña en el año 2018 y el 2020, aumentó el valor que se declara en las denuncias. Seguidamente, el comportamiento de los robos de motocicletas fue relativamente variado, pero con una tendencia a la baja (Medellín Cómo Vamos, 2021). Adicionalmente, según Medellín Cómo Vamos (2022) en agosto del 2022 se presentó la cifra más alta en los registros de denuncias de hurto a personas, sin embargo, es necesario que se tome en cuenta a su vez el crecimiento de la población: en cuanto a la tasa de hurtos (denuncias por cada 100 mil habitantes) el mayor valor se dio en el 2019 con 665. Además, en el comunicado se presenta la siguiente pregunta: “¿La mayor cantidad de denuncias tiene que ver con el cambio de mecanismos de denuncias o tiene que ver con la victimización, es decir, la percepción de los ciudadanos sobre la inseguridad?”. En donde entra en relevancia que en los últimos años (2019 a 2021) ambos indicadores (índice de victimización y denuncias por hurto) han aumentado.

En un artículo publicado por el periódico El Colombiano (2023), se exponen las declaraciones de Julián Enrique Pinilla, personero de Bogotá y presidente de Procapitales. En estas declaraciones se destaca la relevancia de la problemática de los hurtos en Colombia, señalando que el 80% de los hechos se concentran en las principales ciudades del país. En concordancia con lo anterior, el personero menciona que la forma idónea de contrarrestar la inseguridad debe tomar en cuenta factores como el incremento de la fuerza pública, el uso de tecnologías, la generación de un entorno seguro y una evaluación constante de las políticas públicas dada la naturaleza cambiante del fenómeno.

Desde una perspectiva internacional de la literatura, entre los trabajos que analizan datos de delitos se encuentran estudios como el de Amoako (2021) que investiga ubicaciones de robos en la ciudad de Detroit en un periodo de

cinco años, para identificar puntos “calientes”, “fríos” y patrones espaciales en dos escalas. Hace uso de estadística espacial, incluyendo las siguientes técnicas: *Average Nearest Neighbor Analysis*, *Global Moran’s I*, *Getis-Ord Gi* y *Local Morans’s I*. Se concluye, entre otras cosas, que el agrupamiento de robos en ambas escalas no son resultado de un proceso aleatorio, sino de uno sistemático.

En los trabajos de esta línea se hace hincapié en el entendimiento que puede brindar el estudio de estos patrones y cómo estos se pueden estudiar a partir de distintas técnicas. Un ejemplo de ello es el proyecto de Aprendizaje Automático realizado por Ceballos Sánchez (2023), que analiza distintos tipos de crimen en la ciudad de Nueva York entre 2016 y 2019, esto a partir del uso de modelos de agrupamiento (*K-means* y *K-modes*), con el objetivo de generar conocimiento y estrategias que permitan un uso eficiente de los recursos por parte del estado y las autoridades. Este autor destaca que, al considerar la calidad de los resultados encontrados, la presencia de varias variables categóricas en los datos, como la interpretabilidad de sus centroides, la técnica de *K-modes* es la más adecuada y permite una visión más clara y detallada de los perfiles de criminalidad de cada distrito.

Por otra parte, en cuanto a trabajos realizados con datos de la ciudad de Medellín, entre estos se encuentra el de Arévalo Álvarez & Fernández García (2022) quienes usaron un reporte histórico de hurtos a personas para el periodo 2003-2021, con el objetivo de encontrar las áreas con mayor ocurrencia de estos delitos. Los autores, en un primer momento utilizaron un modelo de tipo *K-means* para encontrar un vector de etiquetas a partir de un proceso de validación interna, el cual sirvió como base para el posterior desarrollo de dos modelos de inteligencia artificial: un modelo supervisado *LGBM*, el cual obtuvo la mayor precisión; y un modelo de red neuronal convolucional secuencial, que no superó el nivel esperado de precisión. Igualmente, también realizaron un análisis tomando como referencia zonas turísticas y de mayor comercio, encontrando sectores cuyo nivel de seguridad tuvo una reducción relativa a lo largo del tiempo. Además, encontraron que la Comuna 10 y la 11 es donde se registran los mayores porcentajes históricos de hurtos a personas.

En el presente trabajo se implementan tres algoritmos diferentes de agrupamiento: *K-Means*, *HDBSCAN* y *K-Modes*, tal que se encuentre entre estos el modelo con mejores resultados, tanto en la definición de los grupos como en la concordancia con la naturaleza de los datos. Es importante notar que, anteriormente ya se han realizado análisis de los hurtos en la ciudad, principalmente de hurtos a personas, sin embargo, en este caso, además de incluirse otras modalidades de robo, se presentan un análisis de grupos a partir de distintos algoritmos y de analizar de forma independiente los grupos que se forman a partir de la longitud y latitud; evolución año a año de los grupos formados por la longitud y latitud con el modelo *K-Modes* y distintos grupos de variables con este último modelo.

Estos modelos de agrupamiento se evalúan a partir de la calidad de los grupos generados, tomando como criterios la concordancia con los datos originales y su nivel de definición y separación. Esto dado que es de gran importancia buscar que los resultados de los modelos sean intuitivos y fáciles de interpretar. Los datos originales provienen del sitio web oficial gubernamental MeData. Este conjunto de datos contiene información sobre denuncias relacionadas a distintos tipos de hurtos ocurridos en la ciudad de Medellín y cuenta con variables claves como lo son: el tipo de hurto, características de la persona hurtada y del hecho, bienes robados, su ubicación, de tiempo, entre otras. Los datos se pueden extraer directamente desde el sitio web mediante su url de descarga.

Este estudio se divide en las siguientes secciones: preparación de los datos, estimación y evaluación de modelos de agrupamiento con: los datos normalizados, no normalizados, reducidos a partir de un método de reducción de dimensionalidad, resultados y discusión y finalmente las conclusiones.

2. Materiales y Métodos

Los datos utilizados provienen del sitio web MEData (2025), que es el repositorio oficial de la Alcaldía de Medellín para la publicación de datos relevantes para la ciudadanía. Esta base de datos contiene información relacionada a las denuncias sobre hurtos desde enero de 2003 hasta marzo de 2023, con una frecuencia de actualización mensual en el origen de datos. El acceso a estos datos es abierto, ya que proviene de fuentes públicas, permitiendo su uso para fines académicos. Estos están publicados en archivos CSV independientes y organizados por modalidad de hurto de manera particionada. La modelación y preparación de los datos se realiza mediante notebooks de Python a través de las siguientes librerías principales: *Scikit-learn, Matplotlib, Seaborn* y *Requests.*

A continuación, en la **Tabla 1** se presenta el listado de archivos disponibles y de las modalidades a los que hacen referencia.

Tabla 1:Listado de archivos disponibles acerca de hurtos en la ciudad de Medellín.

Archivo	Tamaño
hurto_a_entidad_financiera.csv	118 KB
hurto_a_establecimiento_comercial.csv	22 MB
hurto_a_persona.csv	124.5 MB
hurto_a_residencia.csv	17 MB
hurto_de_carro.csv	14.3 MB
hurto_de_moto.csv	32 MB
hurto_de_semoviente.csv	47 KB
hurto_por_pirateria_terrestre.csv	886 KB
hurto_a_persona_transporte_publico.csv	9.4 MB

Fuente: Elaboración Propia a partir de los datos de MEData (2025).

Para el desarrollo de las actividades, se obtuvieron los datos directamente de MeData (2025) mediante la librería *Requests* de Python y se consolidaron para obtener un solo conjunto de datos que contiene 579.182 registros y 36 columnas. Cada registro representa un incidente de hurto en la ciudad para las diferentes modalidades, detallando características del suceso y del entorno. Los datos originales recibieron las siguientes transformaciones para ser utilizados en la estimación de los modelos de agrupamiento:

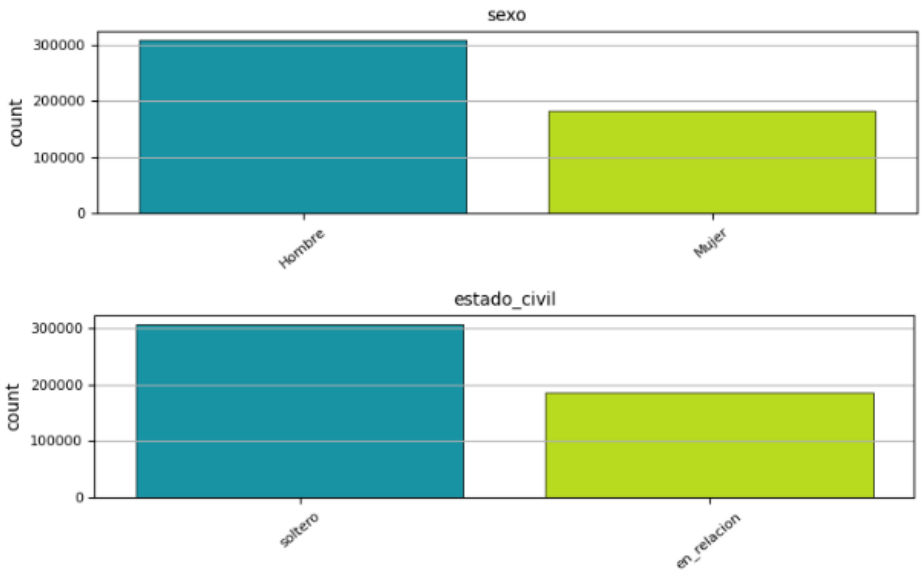
- *Limpieza de datos*: Donde en primer lugar se visualizaron los valores únicos de las columnas para identificar inconsistencias (por ejemplo, valores como “Sin dato” y “SIN DATO”, que fueron reemplazados por valores nulos) o valores faltantes, para posteriormente unificar los datos que sean necesarios y eliminar las inconsistencias; se formatea también la columna de ‘*fecha_hecho’* y se eliminan columnas redundantes o que no fueron consideradas útiles para el estudio.
- *Transformaciones*: Al notar que ciertas columnas tienen una gran cantidad de categorías, se estandarizan y se agrupan estos datos para obtener un grupo de categorías principales para cada columna según sea el caso. Lo cual también fue útil para aumentar el balance en ciertas columnas.
- *Creación de nuevas columnas*: Se crearon variables derivadas de ‘*fecha_hecho*’ que aportan contexto adicional a los datos (año, mes y rango hora) y una que hace referencia a la diferentes zonas de la ciudad: zona.
- *Manejo de datos nulos*: En las variables de latitud y longitud se decidió eliminar estos valores dada la impropiedad de la imputación. Para la columna de edad, que es la única numérica, se realizó una imputación de valores utilizando la mediana. Para el resto de las columnas, las cuales son categóricas, los valores se

imputaron con base en una selección aleatoria basada en la distribución existente de las categorías en cada una de estas, con el objetivo de que este proceso no afecte el balance existente de las categorías.

- *Manejo de datos duplicados*: Tomando en cuenta que los datos dan cuenta de denuncias de robos, cada registro es relevante, dado que, si ocurren robos en un mismo lugar o se denuncian en una misma ubicación y con las mismas características, estos datos pueden ser útiles para la caracterización de los grupos y zonas.
- *Manejo de valores atípicos*: Este manejo se dio en la columna de edad y se utilizó tanto el rango intercuartil como una winsorización al 5% superior.
- *Codificación de las variables categóricas*: En este caso se usó *LabelEncoder* de la librería *Scikit-learn* para convertir los datos de las variables categóricas en etiquetas numéricas.
- *Normalización*: Usando *StandardScaler* se normaliza una copia de los datos limpios, con el objetivo de comparar el efecto de la normalización en el comportamiento de los modelos.

A continuación, una breve descripción del comportamiento de los datos luego de las transformaciones anteriores:

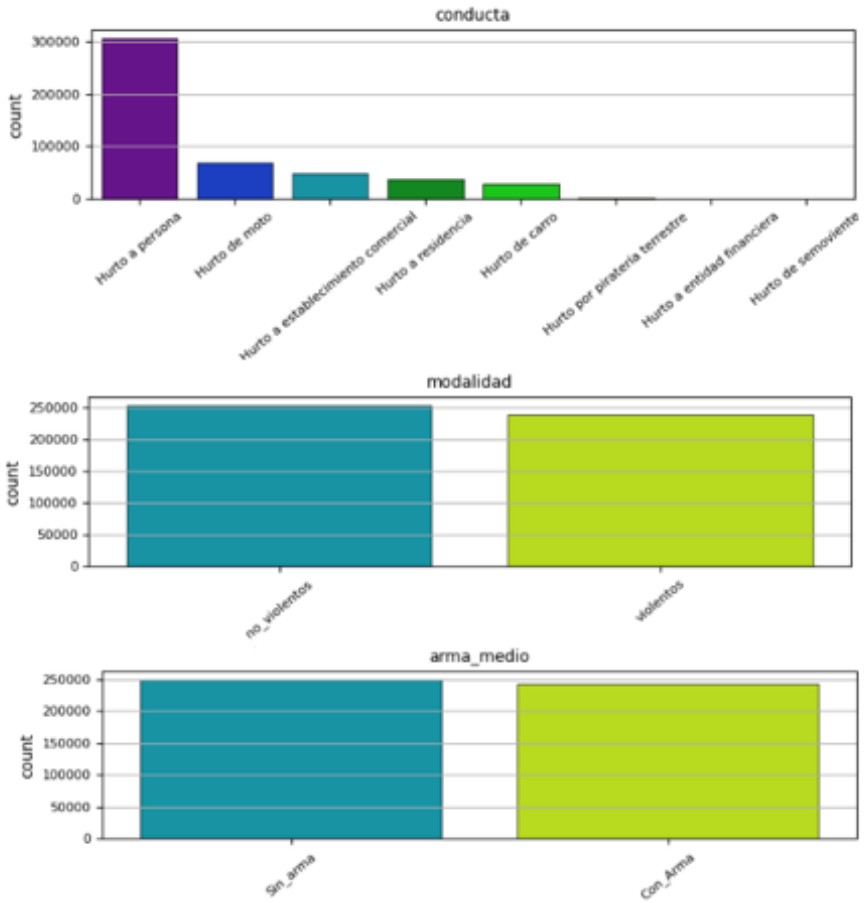
Figura 1:Características de las víctimas/denunciantes.



Fuente: Elaboración propia.

Teniendo en cuenta que el conjunto de datos contiene registros de las denuncias de distintas modalidades de robos en la ciudad de Medellín, la **Figura 1** muestra que las denuncias realizadas en la ciudad son principalmente realizadas por hombres y por personas solteras.

Figura 2:Características de los hurtos.

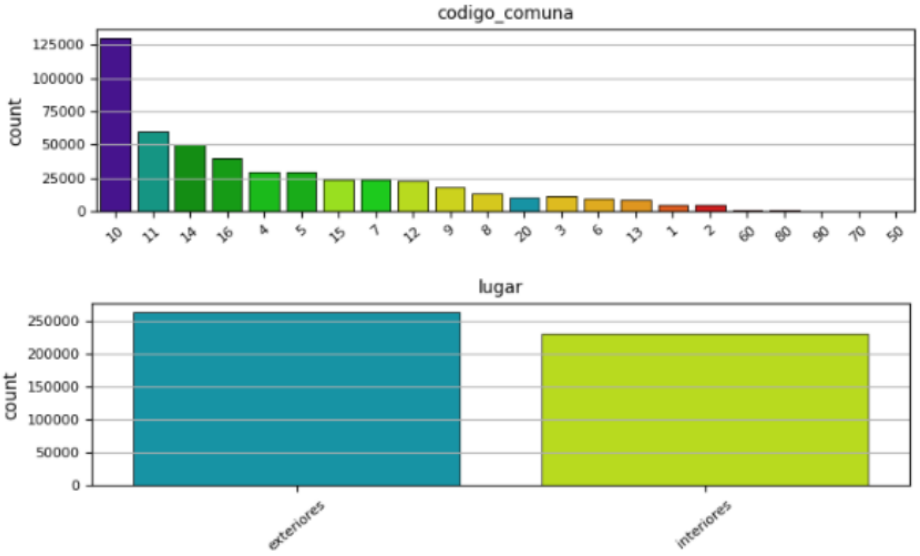


Fuente: Elaboración propia.

Se evidencia que la gran mayoría de los datos están conformados por denuncias de hurtos a personas (ver **Figura 2**). Teniendo en cuenta que se busca que los grupos formados por los modelos reflejen la realidad de la ciudad, no se realizarán cambios para balancear la variable de conducta. Por otra parte, la variable de modalidad está balanceada, pero con un número ligeramente mayor de robos categorizados como no violentos. Lo cual es un comportamiento muy similar al de la variable que indica si hubo un arma como medio del robo.

En cuanto a los lugares donde ocurren los robos, la siguiente gráfica muestra el comportamiento de las denuncias por comuna y si estos ocurrieron en interiores o exteriores:

Figura 3:Comportamiento por comuna y lugar de los hurtos.

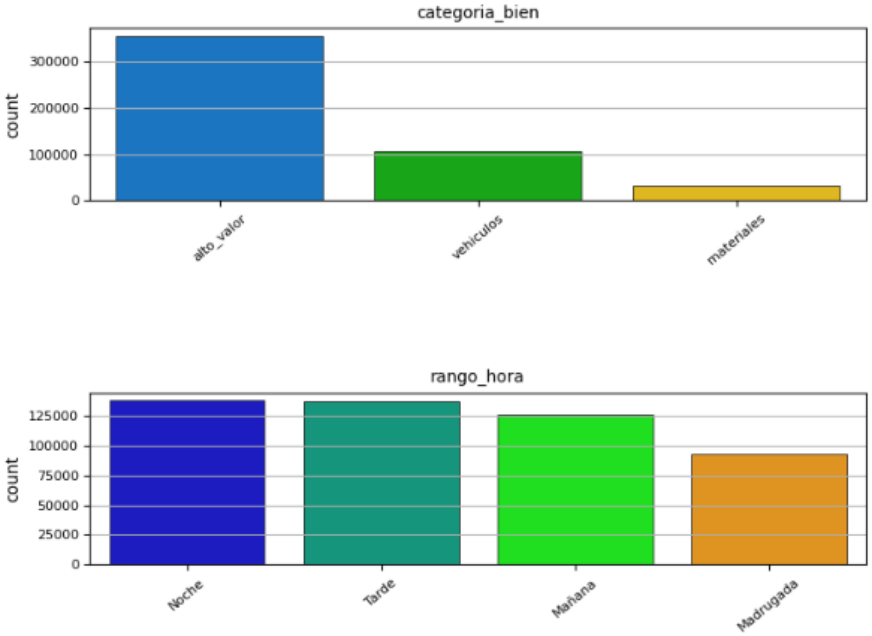


Fuente: Elaboración propia.

La comuna 10 de Medellín es donde más se registraron denunciadas, superando significativamente a las demás comunas. Por el contrario, se presentan menos denuncias es en los corregimientos. Respecto al lugar, los hurtos ocurren principalmente en exteriores, pero la diferencia entre ambas categorías es relativamente pequeña (ver **Figura 3**).

Luego, la siguiente gráfica muestra el comportamiento de otras variables del conjunto de datos que hacen referencia al rango de hora en el que ocurrió el hecho delictivo y a la categoría del bien que fue denunciado como robado:

Figura 4:Categoría del bien hurtado y rango de hora del hecho.



Fuente: Elaboración propia.

Los bienes que son denunciados como robados son principalmente de alto valor. Frente al rango de hora, estos hechos ocurren principalmente en las horas de noche y la tarde, con menor ocurrencia en las madrugadas (ver **Figura 4**).

Posteriormente a estos pasos, se procede a exportar las copias del conjunto de datos final a archivos csv (datos normalizados y no normalizados).

2.1. Estimación de modelos de agrupamiento

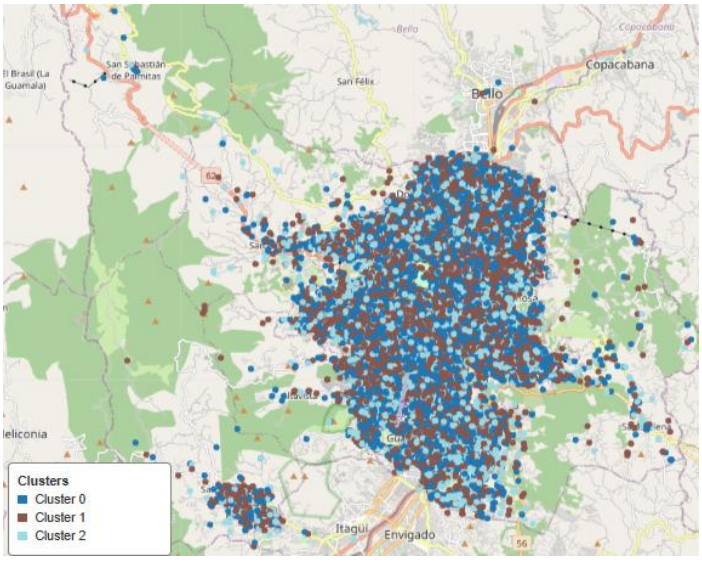
Para la estimación de los modelos de agrupamiento se prueban los modelos tanto con los datos normalizados como no normalizados, para comparar cómo cambia la distribución y compactación de los grupos. En el ámbito de investigación científica, se ha observado la necesidad de desarrollar modelos con datos con reducción de dimensionalidad. Fenómeno que se produce en casos en los que se cuenta con un número elevado de variables. En este estudio, se ha optado por implementar el *Kernel PCA*, específicamente con el kernel "*rbf*", ya que este permite capturar relaciones más complejas que el *PCA* tradicional y tiene en cuenta la distancia entre los puntos. Esta consideración es relevante para el conjunto de datos, ya que las denuncias de hurtos están posicionadas geográficamente. Este proceso resultó en cinco componentes principales, los cuales capturan la mayor cantidad de variabilidad de la información. Por otra parte, tomando en cuenta que los datos contienen distintos tipos de variables (asociadas a la ubicación, al tiempo, características del hecho delictivo, características de las víctimas, entre otras) se realizaron ensayos en la estimación de los modelos tal que, se evidencien los efectos de algunas de estas y así aislar los distintos patrones que se le asocian, principalmente para el caso de las variables de longitud y latitud.

En el caso de los modelos de *K-Means* y *K-Modes*, la evaluación del número óptimo de grupos se lleva a cabo mediante el uso del método del codo y del índice de la silueta. En lo que respecta a *HDBSCAN*, se ha optimizado el valor de "*min_cluster_size*" con el propósito de determinar la cantidad óptima de grupos, además, se decidió que los demás parámetros serán utilizados con sus valores predeterminados. Igualmente, se generaron visualizaciones clave y se tuvieron en cuenta las características de los delitos dentro de cada una de las segmentaciones más relevantes o que presentarán las visualizaciones más separadas y compactas.

2.2. K-Means

Para los datos no normalizados, se obtuvo un número óptimo de tres grupos. La **Figura 5** muestra estos grupos ubicados en un mapa a partir de los datos de latitud y longitud:

Figura 5:Agrupamiento de datos con K-Means con datos no normalizados.



Fuente: Elaboración propia a partir de los datos de MEData (2025).

Para estos datos se obtienen las siguientes características de los grupos (ver **Tabla 2**):

Tabla 2:Características de los grupos con K-Means y datos no normalizados.

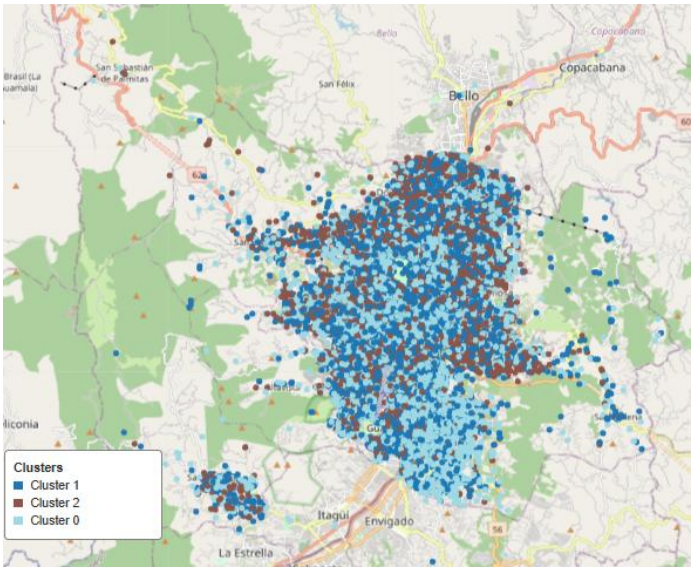
Grupo	Sexo	Edad	Estado civil	Conducta	Modalidad	Arma medio	Código comuna	Lugar	Categoría bien	Día	Mes	Zona	Rango hora	Cantidad
0	hombre	32	soltero	hurto a persona	no violentos	sin arma	10	exteriores	alto valor	1	8	Centro	noche	188103
1	hombre	32	soltero	hurto a persona	no violentos	con arma	10	exteriores	alto valor	28	10	Centro	noche	187827
2	hombre	58	en relación	hurto a persona	no violentos	sin arma	10	exteriores	alto valor	15	11	Centro	tarde	117058

Fuente: Elaboración propia.

Se evidencia que, en promedio todos los grupos exhiben características específicas en cuanto al comportamiento delictivo asociado. Dichos grupos se caracterizan por la participación de actividades delictivas de hurto a personas, que se llevan a cabo principalmente sin el uso de armas y sin violencia. Las víctimas son predominantemente hombres de mediana edad y solteros. Estos eventos se producen predominantemente en los exteriores de la comuna 10 e involucran objetos de alto valor. En este caso, se logran evidenciar patrones de comportamiento en los grupos, sin embargo, se están privilegiando los datos que son mayoritarios, por ejemplo, datos de hurtos a personas y de hombres. En suma, se evidencia una alta similitud entre las rasgos de los distintos grupos.

Luego, para el caso de los datos normalizados, se obtienen tres grupos como el óptimo. La **Figura 6** corresponde a la visualización de las agrupaciones por su localización en los datos normalizados:

Figura 6:Agrupamiento de datos con K-Means con datos normalizados.



Fuente: Elaboración propia a partir de los datos de MEData (2025).

Para ambos casos se obtienen grupos muy entrelazados entre sí, sin algún patrón claro aparente. Sin embargo, con los datos normalizados, se logra una mayor distinción entre estos.

Luego, la **Tabla 3** muestra la siguiente interpretación de los 3 grupos:

Tabla 3:Características de los grupos con K-Means y datos normalizados.

Grupo	Sexo	Edad	Estado civil	Conducta	Modalidad	Arma medio	Código comuna	Lugar	Categoría bien	Día	Mes	Zona	Rango hora	Cantidad

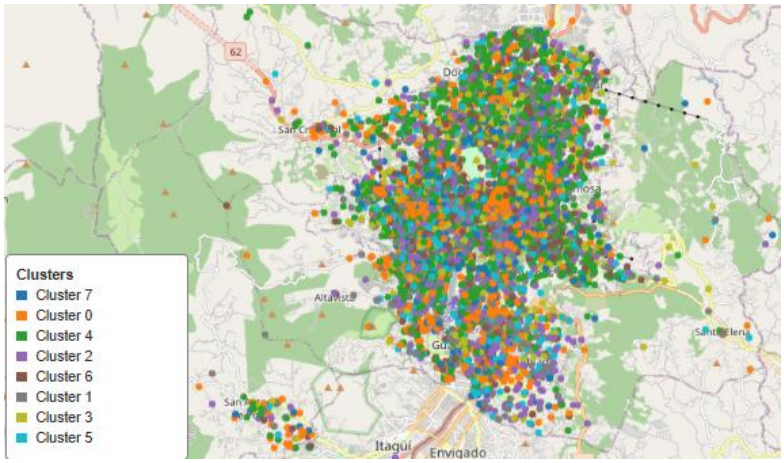
0	hombre	32	soltero	hurto a persona	no violentos	sin arma	10	interiores	alto valor	2	8	Centro	tarde	196208
1	hombre	32	soltero	hurto a persona	violentos	con arma	10	exteriores	alto valor	5	10	Centro	noche	194405
2	hombre	32	soltero	hurto de moto	no violentos	sin arma	10	exteriores	vehículos	19	5	Centro	noche	102375

Fuente: Elaboración propia.

Para este caso, se aborda la asociación de grupos delictivos que principalmente tienen víctimas masculinas, solteras, de 32 años, cuyas actividades delictivas se desarrollan en el centro de Medellín, predominantemente en horario nocturno y vespertino. Dichas actividades se caracterizan por su naturaleza no violenta, aunque en algunos casos se han observado episodios de violencia con armas. Los bienes involucrados suelen ser de alto valor económico, incluyendo vehículos y se componen principalmente de hurtos a personas. No obstante, se ha identificado la aparición de un nuevo grupo de delitos asociados al hurto de motocicletas. En este caso, se evidencia una relativa similitud entre las características de los grupos.

Continuando con la reducción de dimensionalidad, tomando en cuenta la diferencia entre los datos originales y los reconstruidos y el consecuente error obtenido, se opta por tener en cuenta cinco componentes para este modelo. Una vez estimado, se calcula de nuevo K-Means, para el cual en este caso se obtienen ocho agrupaciones. La **Figura 7** muestra la representación de este caso:

Figura 7:Agrupamiento utilizando K-Means con los datos reducidos con Kernel PCA.



Fuente: Elaboración propia a partir de los datos de MEData (2025).

Similarmente a los casos anteriores, los colores que representan los grupos son muy intrincados y no se encuentra un patrón claro o evidente, pero se consiguen un mayor número de grupos respecto a los casos anteriores. A continuación, la **Tabla 4** describe estos grupos y sus características principales:

Tabla 4:Características de los grupos con K-Means y datos normalizados.

Grupo	Sexo	Edad	Estado civil	Conducta	Modalidad	Arma medio	Código comuna	Lugar	Categoría bien	Día	Mes	Zona	Rango hora	Cantidad
0	hombre	32	soltero	hurto a persona	violentos	con arma	10	interiores	alto valor	27	1	Centro	madrugada	1462
1	mujer	32	soltero	hurto a persona	no violentos	sin arma	10	interiores	alto valor	28	8	Centro	mañana	1022
2	hombre	32	en relación	hurto a persona	no violentos	sin arma	10	interiores	alto valor	2	8	Centro	tarde	1436
3	hombre	32	en relación	hurto a persona	violentos	con arma	10	exteriores	alto valor	5	5	Centro	tarde	951

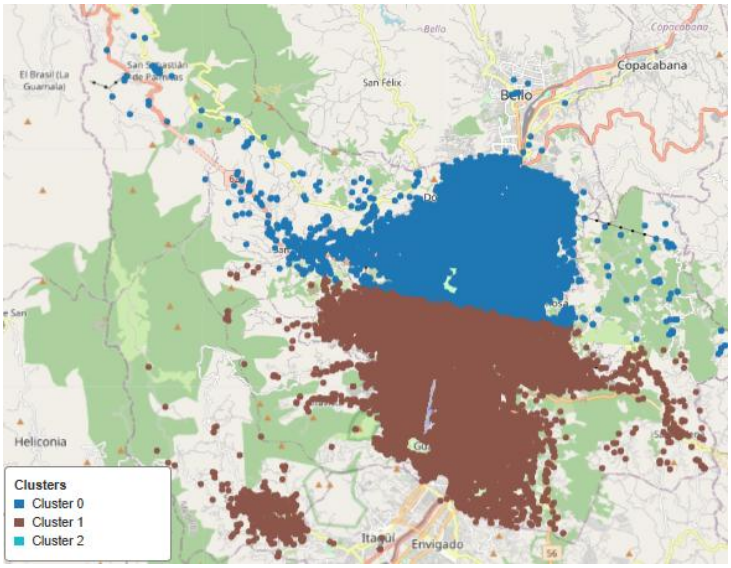
4	hombre	32	soltero	hurto de moto	no violentos	sin arma	10	exteriores	vehículos	18	10	Centro	noche	2139
5	hombre	32	soltero	hurto a persona	no violentos	sin arma	10	interiores	alto valor	13	9	Centro	noche	980
6	hombre	32	soltero	hurto a persona	violentos	con arma	10	exteriores	alto valor	9	8	Centro	noche	1097
7	mujer	32	soltero	hurto a persona	violentos	con arma	10	exteriores	alto valor	17	11	Centro	tarde	913

Fuente: Elaboración propia.

Considerando los ocho grupos obtenidos con el conjunto de datos con reducción de dimensionalidad, se evidencia una asociación predominante de los mismos con la comuna 10, con la zona céntrica de la ciudad y con bienes de alto valor. Además, se ha identificado que la mayoría de los individuos pertenecen al grupo etario alrededor de 32 años. Se ha observado que el 50 % de los individuos son armados, mientras que el otro 50 % no lo es y sólo uno de los grupos presenta datos diferentes a los relacionados con hurtos a personas. En la mayoría de los casos, se observa la presencia de hombres en estado de soltería. Sin embargo, se evidencia la presencia de hombres involucrados en relaciones de pareja, así como datos de mujeres que se clasifican en dos grupos: uno caracterizado por comportamientos violentos durante las horas de la tarde y otro que no presenta tales comportamientos y se manifiesta en la mañana. Además, se ha observado la ausencia de un patrón claro en los días y meses del año, siendo la excepción la aparición del mes de agosto en tres grupos. Así mismo. se evidencia que, para datos reducidos, el número de grupos óptimos y la diversidad de los grupos analizados aumenta para el algoritmo *K-Means*.

En el siguiente paso del proceso, se lleva a cabo un agrupamiento de los datos tomando en consideración únicamente de su ubicación geográfica, junto con los datos que no han sido normalizados (específicamente, los datos relacionados con la latitud y la longitud). Se obtienen tres grupos, que se presentan en la **Figura 8:**

Figura 8:Agrupamiento utilizando K-Means con los datos de latitud y longitud.



Fuente: Elaboración propia a partir de los datos de MEData (2025).

En este caso, se implementa una segmentación de los grupos en una estructuración de carácter norte-sur. En el análisis realizado, se ha comprobado que únicamente son visibles dos de los tres grupos, lo que sugiere la posibilidad de que dichos grupos se encuentren en una configuración superpuesta. Asimismo, es factible que los resultados de los grupos restantes no puedan ser gráficamente representados con claridad en el mapa.

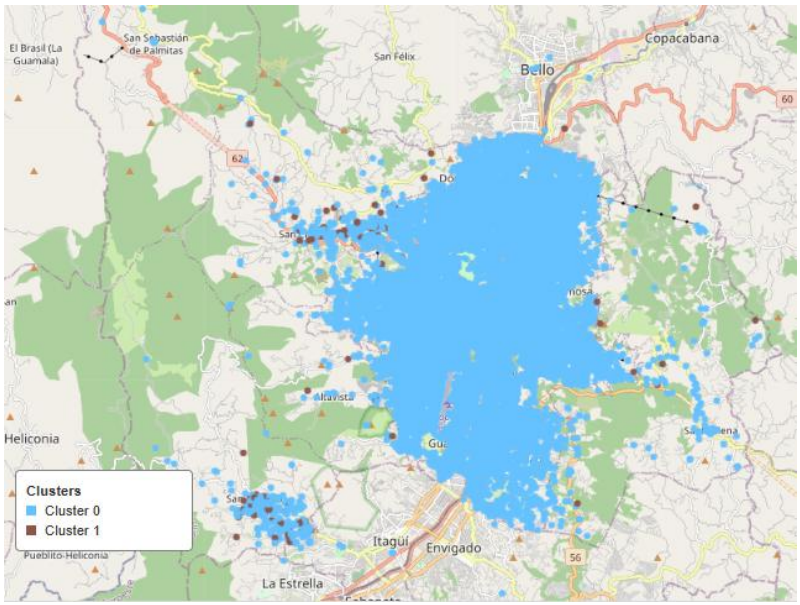
Se destaca que en los dos primeros casos se obtienen datos de tres grupos, los cuales son relativamente más definidos y con características más variadas al normalizar los datos. El tercer caso de agrupamiento muestra una mayor

cantidad de grupos y de variedades, para los cuales especialmente no se evidencian patrones claros, y en cuanto a las características de estos, hay mucha mayor variedad y la representación de datos que anteriormente no aparecían como características generales de un grupo, por ejemplo, la aparición de mujeres y nuevos rangos de horas. El cuarto no brinda conclusiones importantes para este trabajo.

2.3. HDBSCAN

Para el caso de los datos no normalizados, el número óptimo del parámetro *"min_cluster_size"* es 80 y los dos grupos resultantes se ubican en el mapa de Antioquia de la siguiente manera en la **Figura 9**:

Figura 9:Agrupamiento de datos con HDBSCAN con datos no normalizados.



Fuente: Elaboración propia a partir de los datos de MEData (2025).

En este mapa se evidencian dos grupos muy diferentes: uno que ocupa la mayoría del mapa y no tiene un patrón espacial definido y otro que tiene menos puntos, pero se ubican en las afueras de la ciudad y en sus corregimientos, por ejemplo, se evidencian varios puntos en San Antonio de Prado y San Cristóbal. Dado esto, es posible que se estén subdividiendo los hechos delictivos entre ciudad y sus afueras.

La **Tabla 5** muestra las características principales de este caso:

Tabla 5:Características de los grupos con HDBSCAN y datos no normalizados.

Grupo	Sexo	Edad	Estado civil	Conducta	Modalidad	Arma medio	Código comuna	Lugar	Categoría bien	Día	Mes	Zona	Rango hora	Cantidad
-1	hombre	56	soltero	hurto a persona	violentos	con arma	90	exteriores	alto valor	26	5	Centro	tarde	27
0	hombre	32	soltero	hurto a persona	no violentos	sin arma	10	exteriores	alto valor	2	8	Centro	tarde	39845
1	hombre	32	soltero	hurto a persona	no violentos	sin arma	60	exteriores	alto valor	27	12	centro	noche	128

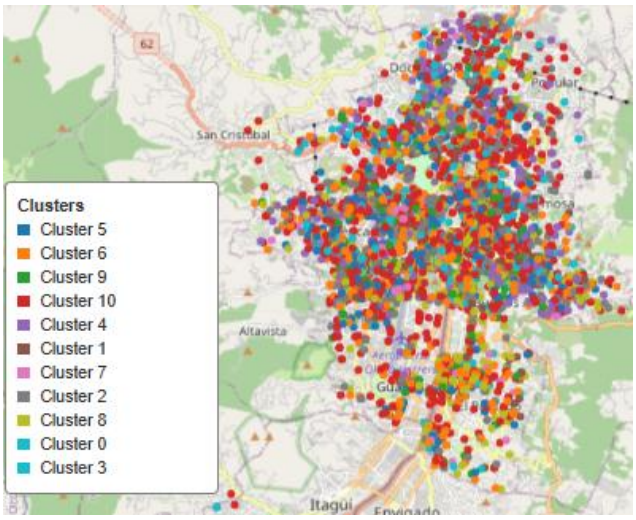
Fuente: Elaboración propia.

Para este caso, el modelo identificó los robos con armas a hombres solteros de la tercera edad en Santa Elena como dato atípico. Los grupos identificados por el modelo son muy similares a diferencia de su ubicación y variables de

tiempo. El grupo más grande es caracterizado por robos en el centro de la ciudad en las horas de la tarde y el otro grupo se caracteriza por hurtos en las noches y principalmente en San Cristóbal.

Para los modelos normalizados, el valor de *"min_cluster_size"* es de 125 y se asocia a once grupos, los cuales tienen el siguiente comportamiento que se muestra en la **Figura 10**:

Figura 10:Agrupamiento de datos con HDBSCAN con datos normalizados.



Fuente: Elaboración propia a partir de los datos de MEData (2025).

Se evidencia una menor cantidad de datos en las afueras de la ciudad y de los corregimientos, hay una cantidad mayor de grupos que en los casos anteriores, sin embargo, no hay un patrón evidente a lo largo de todo el territorio respecto a los grupos.

Las características generales de los once grupos se visualizan en la **Tabla 6:**

Tabla 6:Características de los grupos con HDBSCAN y datos normalizados.

Grupo	Sexo	Edad	Estado civil	Conducta	Modalidad	Arma medio	Código comuna	Lugar	Categoría bien	Día	Mes	Zona	Rango hora	Cantidad
-1	hombre	32	soltero	hurto a persona	no violentos	sin arma	10	exteriores	alto valor	1	11	Centro	tarde	18625
0	hombre	32	soltero	hurto de moto	violentos	con arma	10	exteriores	vehículos	11	5	Centro	noche	323
1	mujer	32	en relación	hurto a persona	no violentos	sin arma	10	interiores	alto valor	9	8	Centro	mañana	213
2	hombre	32	en relación	hurto a persona	violentos	con arma	10	exteriores	alto valor	10	6	Centro	noche	697
3	hombre	32	en relación	hurto a persona	no violentos	sin arma	10	interiores	alto valor	19	8	Centro	mañana	238
4	hombre	25	soltero	hurto de moto	no violentos	sin arma	9	exteriores	vehículos	8	5	Centro	noche	372
5	mujer	32	soltero	hurto a persona	violentos	con arma	10	exteriores	alto valor	15	8	Centro	noche	618
6	Mujer	32	soltero	hurto a persona	no violentos	sin arma	10	interiores	alto valor	19	8	Centro	mañana	982
7	hombre	32	soltero	hurto a persona	no violentos	sin arma	10	exteriores	alto valor	17	9	Centro	noche	246
8	hombre	32	soltero	hurto a persona	no violentos	sin arma	10	interiores	alto valor	11	8	Centro	mañana	746

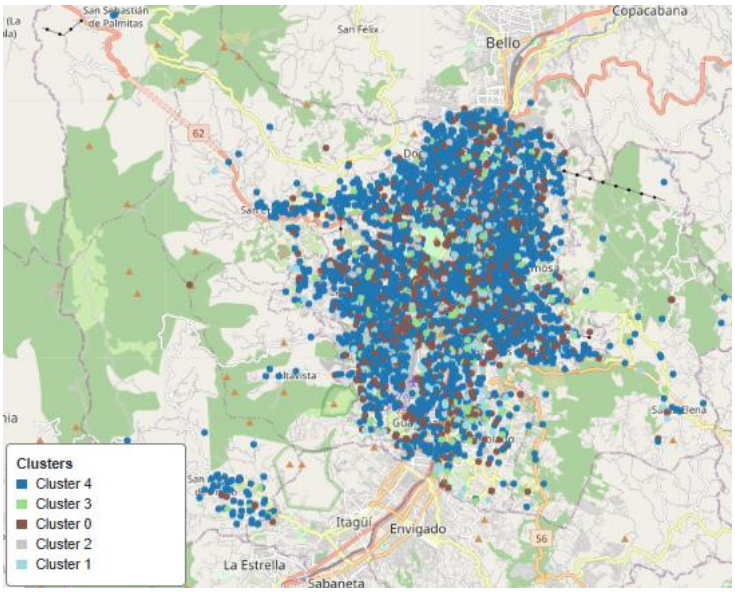
9	hombre	32	soltero	hurto a persona	violentos	con arma	10	interiores	alto valor	9	7	Centro	noche	324
10	hombre	32	soltero	hurto a persona	violentos	con arma	10	exteriores	alto valor	13	9	Centro	noche	1616

Fuente: Elaboración propia.

Se identificó como datos atípicos a los robos cuya diferencia más evidente es que ocurren en las tardes. Los once grupos pertenecen principalmente a hurto a personas, pero también hay dos grupos de hurtos de motocicletas. En su mayoría se asocian a estos datos: hombres, personas solteras, 32 años, exteriores, alto valor, comuna 10, mes de agosto y noches. Las categorías de modalidad y de arma medio están mucho más balanceadas entre sí. Y el grupo más único es el 4, que son hurtos de motos en Buenos Aires a hombres de 25 años. El modelo con datos normalizados permite una mayor cantidad y variedad en los grupos.

Para el caso de reducción de dimensionalidad, se obtiene que el valor del parámetro *"min_cluster_size"* es de 205 y éste se asocia a cinco grupos. Estos tienen la forma que se muestra en la **Figura 11**:

Figura 11:Agrupamiento utilizando HDBSCAN con los datos reducidos con Kernel PCA.



Fuente: Elaboración propia a partir de los datos de MEData (2025).

Se recuperan los datos alrededor de San Antonio de Prado y en las afueras de la ciudad, además hay una distribución relativamente más clara de los grupos, obteniéndose un grupo a lo largo de todo el mapa y otros grupos a lo largo de este. Sin embargo, aún no son del todo evidentes los patrones en los grupos.

Las características generales de los grupos para este caso son:

Tabla 7:Características de los grupos con HDBSCAN y datos reducidos con Kernel PCA.

Grupo	Sexo	Edad	Estado civil	Conducta	Modalidad	Arma medio	Código comuna	Lugar	Categoría bien	Día	Mes	Zona	Rango hora	Cantidad
-1	hombre	32	soltero	hurto a persona	no violentos	sin arma	10	interiores	alto valor	12	8	Centro	tarde	18012
0	hombre	32	soltero	hurto a persona	violentos	con arma	10	exteriores	alto valor	1	3	Centro	noche	1470
1	hombre	32	soltero	hurto a persona	no violentos	sin arma	10	interiores	alto valor	1	12	Centro	mañana	595

2	hombre	32	en relación	hurto a persona	violentos	con arma	10	exteriores	alto valor	3	11	Centro	tarde	218
3	mujer	32	soltero	hurto a persona	no violentos	sin arma	10	interiores	alto valor	3	12	Centro	mañana	439
4	hombre	32	soltero	hurto de moto	no violentos	sin arma	10	exteriores	vehículos	28	1	Centro	noche	4266

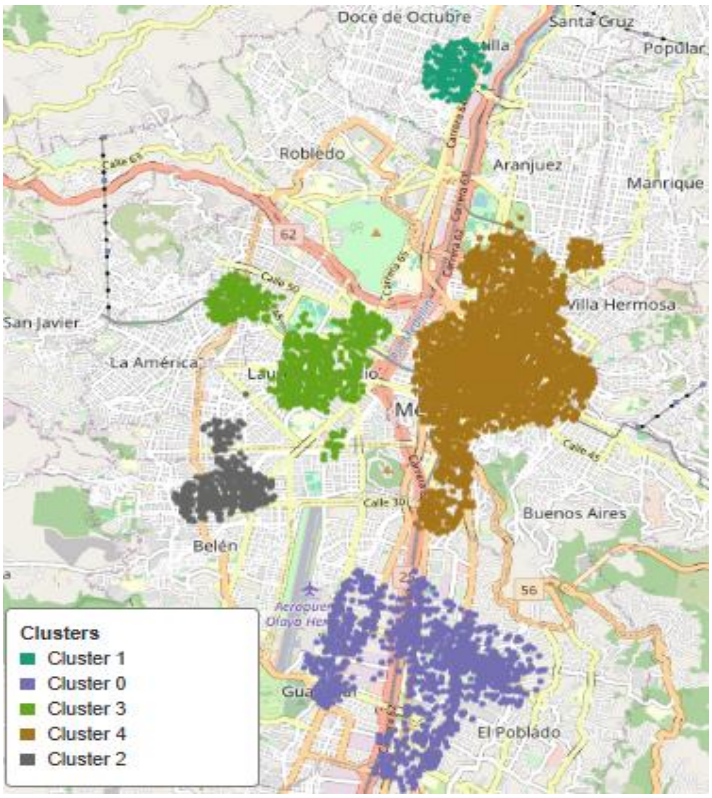
Fuente: Elaboración propia.

Los grupos son relativamente más similares entre sí al reducir la dimensionalidad respecto al caso anterior, pero se conserva cierta variedad en las características. Principalmente estos grupos se resumen en hurtos a: personas, hombres, solteros, de 32 años, no violentos, sin arma, bienes de alto valor, en exteriores y en la zona centro de la ciudad (ver **Tabla 7**).

Se procede con el análisis separado de las variables de ubicación en primer lugar y en segundo lugar las variables de tiempo (años). Con el objetivo de encontrar qué tanto inciden en la separación y en las características de los grupos. Este análisis se realiza con los datos no normalizados.

A continuación, se muestra en la **Figura 12** el comportamiento de los grupos a partir de latitud y longitud, en el que el valor de min_cluster_size es de 260 y se asocia a cinco grupos:

Figura 12: Agrupamiento de datos con HDBSCAN y datos geográficos.

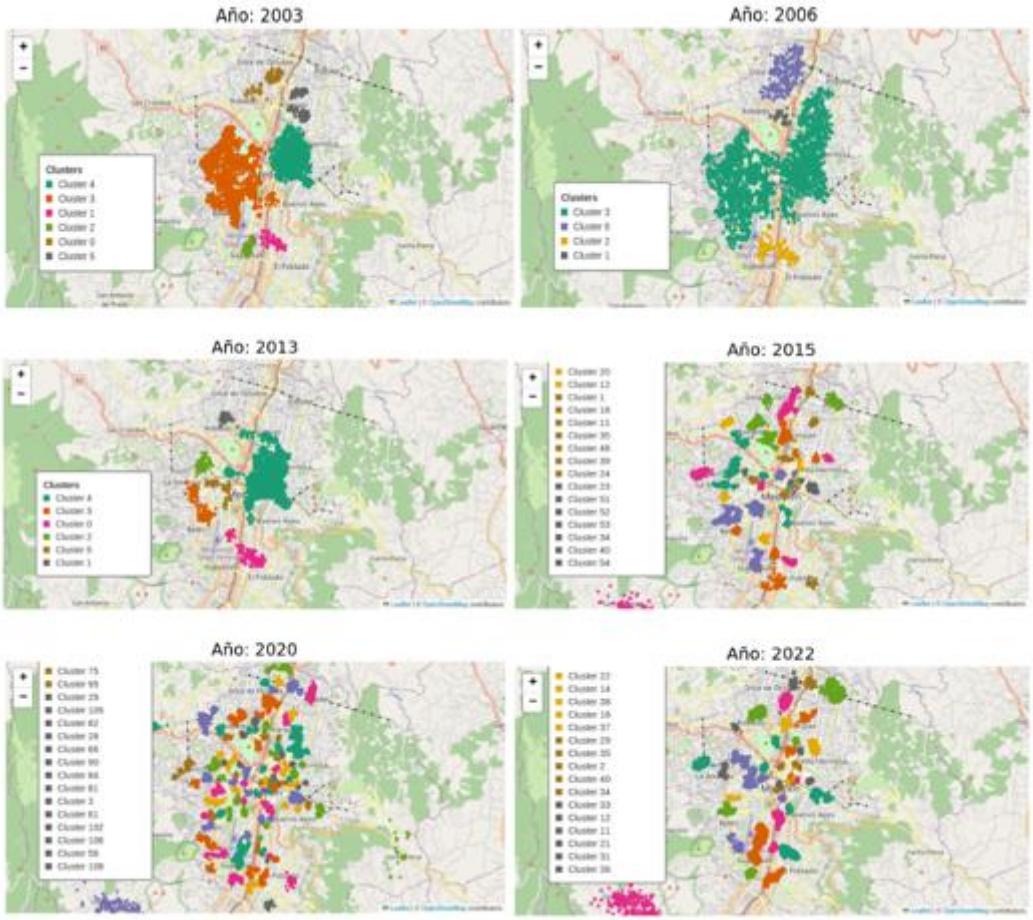


Fuente: Elaboración propia a partir de los datos de MEData (2025).

En este caso, el uso de *HDBSCAN* con los datos de longitud y latitud logran unos grupos más separados y compactos. Donde se destacan las siguientes zonas: la más pequeña en Castilla; la más grande alrededor de las siguientes estaciones del metro de Medellín: Hospital, Universidad, Prado, Parque Berrío, San Antonio, Cisneros, Alpujarra y Exposiciones (centro de la ciudad); otra zona en verde asociada a las Estaciones Estadio, Suramericana y Floresta; la zona en gris oscuro se encuentra cercana al barrio Belén y la zona en morado se encuentra en las zonas aledañas a Poblado, Aguacatala y Guayabal.

Se continua con los grupos asociados a distintos años (2003-2022), donde se obtienen 19 grupos distintos para cada año, cada una con su respectivo *"min_cluster_size"* el cual varía desde 50 hasta 185. A continuación, en la **Figura 13** se muestra en pares algunos de los años más relevantes:

Figura 13:Agrupamiento de datos con HDBSCAN y datos de diferentes años.



Fuente: Elaboración propia a partir de los datos de MEData (2025).

La evolución de los grupos año a año se encuentra en la **Tabla 8**:

Tabla 8:Número de grupos a lo largo de los años.

Año	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
Grupos	6	6	4	4	3	6	5	5	4	7	6	6	55	30	57	56	32	110	47	42

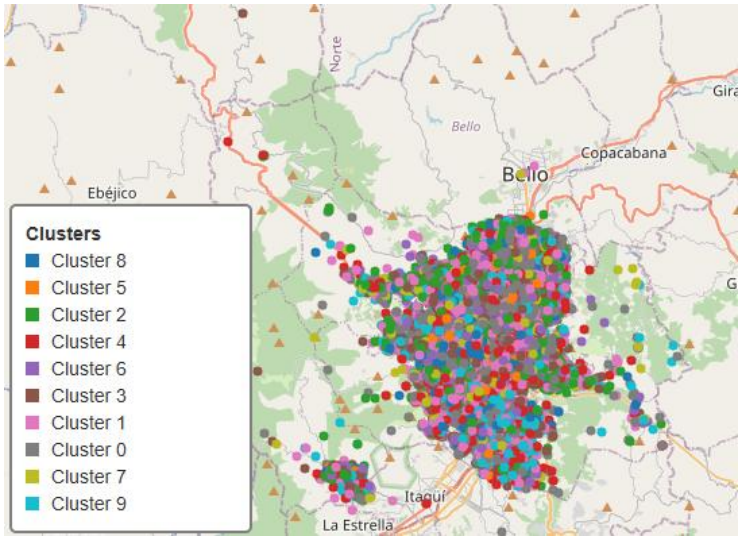
Fuente: Elaboración propia.

A partir de la **Figura 13** y **Tabla 8**, se deduce que hasta 2014 hay un patrón similar en el número de grupos, pero a partir de este, comienzan a aumentar estos números de forma significativa hasta llegar a un pico en el año 2020, pero los números siguen siendo muy altos a comparación de los once primeros años. El caso del año 2020 podría estar relacionado con los efectos de la pandemia, en la que la movilización de las personas era limitada y solía ser en los perímetros de sus lugares de residencia.

2.4. K-Modes

Para este modelo, con el uso de los datos normalizados se encuentra como número óptimo diez grupos, los cuales se ven así en la **Figura 14**:

Figura 14:Agrupamiento de datos con K-Modes con datos reducidos con Kernel PCA.



Fuente: Elaboración propia a partir de los datos de MEData (2025).

Los grupos se encuentran muy entrelazados al usar *K-Modes* con los datos no normalizados. Y a simple vista no se encuentran patrones relevantes en estos.

Luego, las características de los grupos generados en este caso se presentan en la **Tabla 9**:

Tabla 9:Características de los grupos en K-Modes y datos no normalizados.

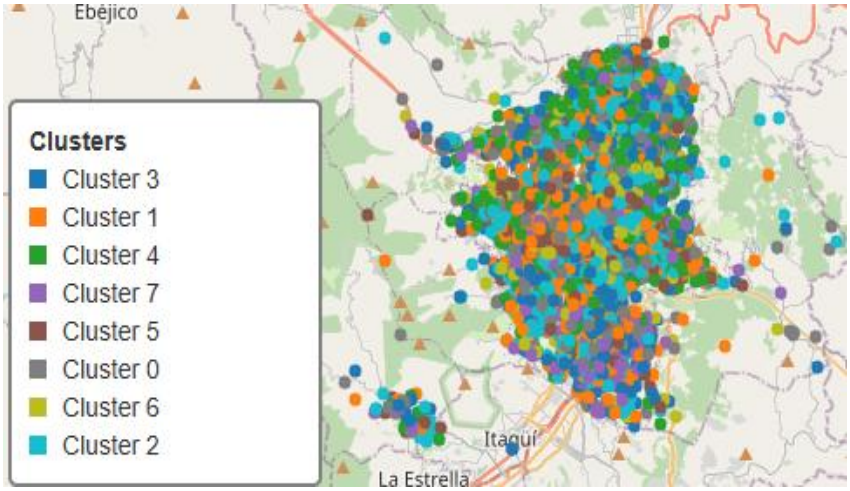
Grupo	Sexo	Edad	Estado civil	Conducta	Modalidad	Arma medio	Código comuna	Lugar	Categoría bien	Día	Mes	Zona	Rango hora	Cantidad
0	hombre	32	en relación	hurto a persona	violentos	con arma	10	exteriores	alto valor	11	7	Centro	tarde	5315
1	hombre	32	soltero	hurto a establecimiento comercial	no violentos	sin arma	10	interiores	alto valor	7	9	Centro	tarde	4895
2	hombre	26	soltero	hurto de moto	no violentos	sin arma	10	exteriores	vehículos	19	3	Centro	noche	4261
3	hombre	29	soltero	hurto a persona	violentos	con arma	10	interiores	alto valor	15	10	Centro	noche	2793
4	mujer	32	en relación	hurto a persona	no violentos	sin arma	10	interiores	alto valor	11	6	Centro	mañana	3524
5	hombre	28	en relación	hurto a persona	violentos	con arma	12	exteriores	alto valor	22	8	Zona Norte	noche	863
6	hombre	19	soltero	hurto a persona	violentos	con arma	10	exteriores	alto valor	3	3	Centro	noche	2686
7	hombre	30	soltero	hurto a persona	violentos	con arma	10	exteriores	alto valor	27	2	Centro	tarde	1759
8	mujer	32	soltero	hurto a persona	violentos	con arma	11	interiores	alto valor	15	3	Centro	noche	1489
9	mujer	58	soltero	hurto a persona	no violentos	sin arma	14	interiores	alto valor	7	4	Centro	mañana	2415

Fuente: Elaboración propia.

Estos grupos son relativamente más variados a comparación de los casos anteriores, pero en su mayoría tienen las siguientes características: Hombres, personas solteras, hurto a personas, violentos, con arma, en la comuna 10 y zona centro y bienes de alto valor. En este caso aparecen nuevas características en los grupos: zona norte, comunas 12 (La América), 11 (Laureles-Estadio) y 14 (Poblado) y Hurtos a establecimientos comerciales.

Usando los datos normalizados se decide por utilizar ocho grupos, los cuales se comportan como lo muestra la **Figura 15**:

Figura 15:Agrupamiento de datos con K-Modes y datos normalizados.



Fuente: Elaboración propia a partir de los datos de MEData (2025).

Los grupos están relativamente más definidos, sin embargo, la alta dispersión en los puntos pertenecientes a cada grupo dificulta su análisis. Por lo que se procede a revisar las características de cada uno de estos mostradas en la **Tabla 10**:

Tabla 10:Características de los grupos en K-Modes y datos normalizados.

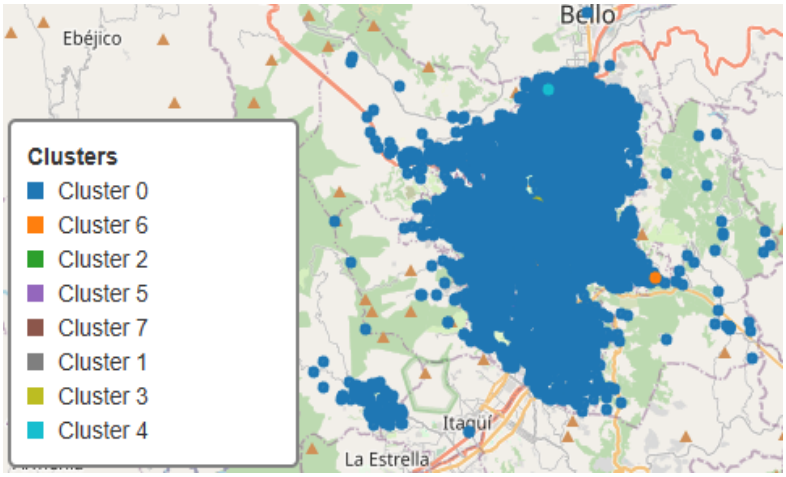
Grupo	Sexo	Edad	Estado civil	Conducta	Modalidad	Arma medio	Código comuna	Lugar	Categoría bien	Día	Mes	Zona	Rango hora	Cantidad
0	hombre	58	soltero	hurto a persona	no violentos	sin arma	10	exteriores	alto valor	16	2	Centro	madrugada	1202
1	hombre	30	soltero	hurto a persona	violentos	con arma	11	interiores	alto valor	29	7	Centro	madrugada	1148
2	hombre	32	soltero	hurto a persona	violentos	con arma	10	exteriores	alto valor	2	10	Centro	tarde	1602
3	mujer	32	soltero	hurto a persona	no violentos	sin arma	14	interiores	alto valor	28	6	Centro	mañana	1255
4	hombre	32	soltero	hurto de moto	no violentos	sin arma	4	exteriores	vehículos	10	11	Centro	noche	1030
5	hombre	22	soltero	hurto a persona	violentos	con arma	10	exteriores	alto valor	12	3	Centro	mañana	671
6	hombre	58	soltero	hurto a persona	no violentos	sin arma	10	interiores	alto valor	18	7	Centro	mañana	581
7	hombre	32	en relación	hurto a establecimiento comercial	no violentos	con arma	14	interiores	alto valor	2	4	Centro	noche	511

Fuente: Elaboración propia.

Los hurtos dentro de estos grupos son principalmente a hombres, personas mayores de 30 años, personas solteras, Hurtos a personas, no violentos, zona centro y bienes de alto valor. Uno de los grupos se ubica principalmente en la comuna 4 (Aranjuez).

Se sigue con la estimación del modelo con los datos reducidos con *Kernel PCA*. Para este caso también se decide utilizar ocho grupos, los cuales se muestran en la **Figura 16**:

Figura 16:Agrupamiento de datos con K-Modes y datos reducidos con Kernel PCA.



Fuente: Elaboración propia a partir de los datos de MEData (2025).

Se destaca que en este caso la reducción de dimensionalidad no tuvo un efecto claro en la separación, ya que sólo se visualizan tres colores (grupos). Aparentemente ayudó a compactar los datos del grupo en azul y a obtener sólo dos puntos para dos grupos visibles a simple vista.

Estos grupos tienen las siguientes características mostradas en la **Tabla 11**:

Tabla 11:Características de los grupos en K-Modes y datos reducidos con Kernel PCA.

Grupo	Sexo	Edad	Estado civil	Conducta	Modalidad	Arma medio	Código comuna	Lugar	Categoría bien	Día	Mes	Zona	Rango hora	Cantidad
0	hombre	32	soltero	hurto a persona	no violentos	sin arma	10	exteriores	alto valor	2	8	Centro	tarde	9991
1	hombre	32	en relación	hurto a establecimiento comercial	no violentos	sin arma	11	interiores	alto valor	23	1	Centro	tarde	1
2	mujer	32	soltero	hurto a establecimiento comercial	violentos	sin arma	6	exteriores	alto valor	21	2	Centro	mañana	1
3	mujer	32	en relación	hurto a establecimiento comercial	no violentos	sin arma	7	exteriores	materiales	2	6	Centro	madrugada	1
4	mujer	43	en relación	hurto a residencia	no violentos	sin arma	6	interiores	alto valor	29	5	Centro	tarde	1
5	mujer	28	soltero	hurto a persona	violentos	con arma	10	exteriores	alto valor	18	9	Centro	noche	1
6	hombre	58	soltero	hurto a residencia	no violentos	sin arma	9	interiores	alto valor	6	9	Centro	noche	3
7	hombre	35	en relación	hurto de carro	no violentos	sin arma	9	exteriores	vehículos	2	7	Centro	mañana	1

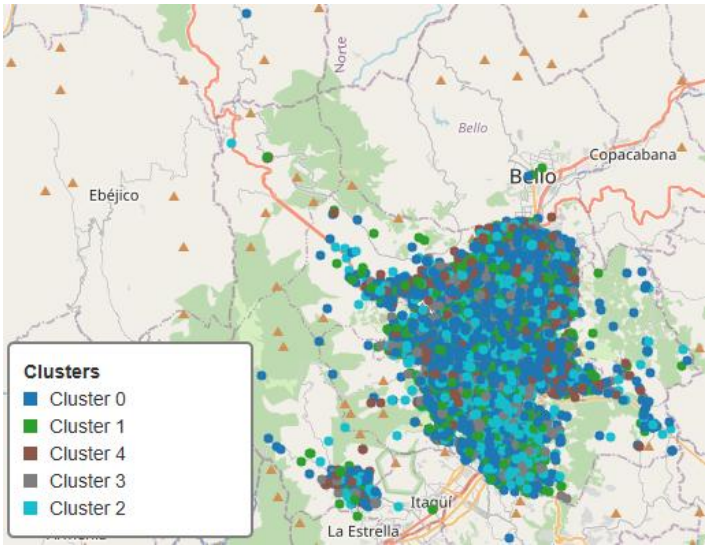
Fuente: Elaboración propia.

Se dan dos grupos relevantes: el grupo más grande, que abarca la gran mayoría de datos con características que son compartidas tanto con los grupos más pequeños de este caso, como en los grupos de otros modelos y casos; y el grupo que contiene 3 observaciones, que se destaca por robos de bienes de alto valor que tienen como víctimas

hombres solteros de la tercera edad en residencias de la comuna 9 (Buenos Aires). A pesar de una disminución en la proporción de representación de cada grupo, se evidencia un incremento en la diversidad de atributos.

Tomando en cuenta lo anterior, se procede a estimar el modelo *K-Modes* con distintos conjuntos de variables y los datos no normalizados. El primer conjunto contiene: conducta, modalidad y arma medio. Para este se toman cinco grupos como óptimo, que se comportan como la muestra la **Figura 17**:

Figura 17:Agrupamiento de datos con K-Modes y el primer conjunto de características.



Fuente: Elaboración propia a partir de los datos de MEData (2025).

Para este caso, no se distinguen patrones relevantes o que puedan ser distinguidos a nivel geográfico. Por lo que se describen las características de estos (ver **Tabla 12**):

Tabla 12:Rasgos de los grupos con K-Modes y en el primer conjunto de características.

Grupo	Conducta	Modalidad	Arma medio	Cantidad
0	hurto a persona	violentos	con arma	13080
1	hurto a persona	violentos	sin arma	2635
2	hurto de moto	no violentos	sin arma	6069
3	hurto a persona	no violentos	sin arma	6845
4	hurto de moto	violentos	con arma	1371

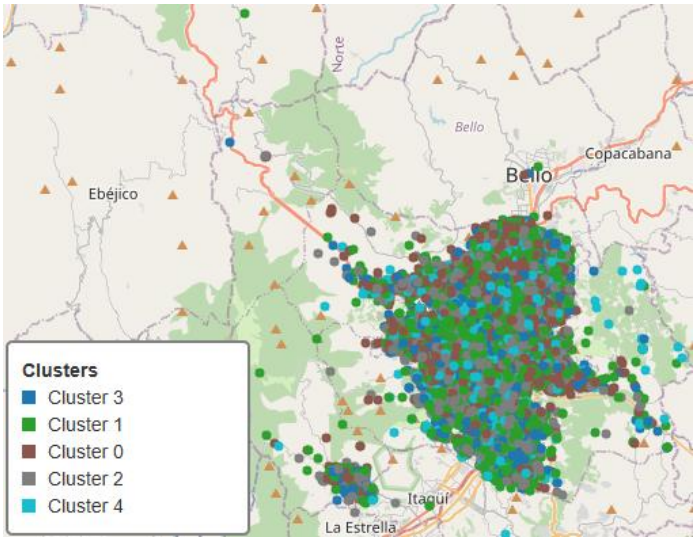
Fuente: Elaboración propia.

En su mayoría, los hurtos asociados son no violentos y sin armas. En este caso este conjunto de características permite una mayor variedad de las características, ya que, aunque algunas de ellas sean compartidas, cada grupo se distingue bien de los demás. Respecto a otras características dentro de estos grupos, se encuentran las siguientes generalidades a lo largo de los datos: En la mayoría de los casos hay más hombres que mujeres o son valores cercanos, más personas solteras, más datos de bienes y de la zona centro y una buena representatividad de datos de los rangos tarde y noche. El grupo 0, la cual tiene la mayoría de los datos en este caso se encuentra conformado principalmente por hombres, personas solteras, ocurren en exteriores, se asocian a la zona centro de la ciudad, los

artículos son de alto valor y ocurren en las tardes y noches. Se tiene a su vez, pocos datos de los corregimientos de Medellín y de bienes materiales. Y en cuanto a los datos de días y meses, estos tienen una gran dispersión a lo largo de los grupos, el único grupo que no cumple completamente estas características es el cuarto que se asocia más a meses cercanos a junio.

El segundo conjunto se conforma por: lugar, categoría bien y rango hora. A su vez, cuenta con cinco grupos distribuidos de esta forma:

Figura 18:Agrupamiento de datos con K-Modes con el segundo conjunto de características.



Fuente: Elaboración propia a partir de los datos de MEData (2025).

El resultado es similar al obtenido con el primer conjunto de características (ver **Figura 18**). Se revisa entonces las características de este último en la **Tabla 13**:

Tabla 13:Rasgos de los grupos en el segundo conjunto de características.

Grupo	Lugar	Categoría bien	Rango hora	Cantidad
0	exteriores	vehículos	noche	6426
1	exteriores	alto valor	noche	9856
2	interiores	alto valor	tarde	6269
3	interiores	alto valor	mañana	3905
4	exteriores	alto valor	tarde	3544

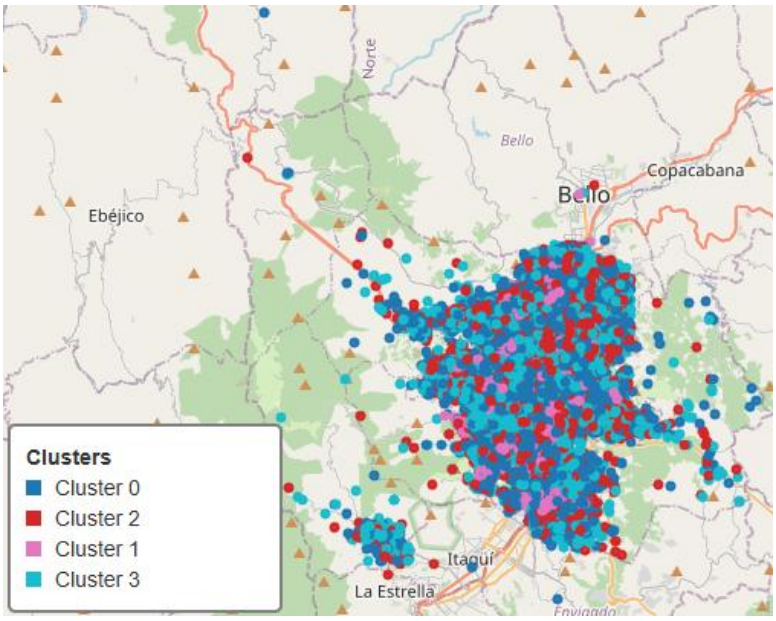
Fuente: Elaboración propia.

Estos grupos comparten similitudes, destacando principalmente la característica de poseer bienes de alto valor y que los robos ocurren en exteriores. En lo que respecta a las demás características (las demás variables de los datos), se evidencia una mayor presencia de sujetos masculinos, con edades aproximadas a 30 años, así como un mayor número de individuos en situación de soltería y de hurtos a personas en la zona céntrica de la ciudad. Además, se evidencia una carencia de información respecto a los corregimientos de Medellín. En lo que respecta a los datos concernientes a los días y los meses, se evidencia una amplia dispersión a lo largo de los grupos analizados. Sin

embargo, el único grupo que no se ajusta completamente a estas características es aquel que se asocia predominantemente a meses cercanos a agosto.

El tercer conjunto se conforma por: sexo, estado civil y zona. A su vez, cuenta con cuatro grupos que se ven en la **Figura 19**:

Figura 19:Agrupamiento de datos con K-Modes con el tercer conjunto de características.



Fuente: Elaboración propia a partir de los datos de MEData (2025).

De manera similar a los casos precedentes, los grupos no exhiben una separación evidente. No obstante, en este caso particular, se identifica un número óptimo de 4 grupos. Por consiguiente, se aborda la revisión de las características que definen el caso objeto de análisis, tal y como se muestra en la **Tabla 14**:

Tabla 14:Rasgos de los grupos en el tercer conjunto de características.

Grupo	Sexo	Estado civil	Zona	Cantidad
0	mujer	soltero	Centro	11114
1	hombre	soltero	Zona Norte	3704
2	hombre	en relación	Centro	6877
3	hombre	soltero	Centro	8305

Fuente: Elaboración propia.

En el tercer conjunto de características, se forman grupos que suelen estar principalmente conformados por hombres, personas solteras y se ubican en la zona centro de la ciudad. En este caso, se observa un grupo mayoritario conformado por mujeres como una gran diferencia a los conjuntos anteriores. En cuanto a las demás variables, en general sus valores se encuentran más distribuidos entre las categorías, es decir, hay un menor desbalance, a excepción de, por ejemplo, los bienes y los tipos de hurtos, ya que predominan los de alto valor y los hurtos a personas.

3. Resultados y Discusión

A partir del uso de los algoritmos utilizados, se encontró que, en general, la normalización de los datos genera una mayor diferenciación relativa entre los grupos, además de una mayor variedad en las características subyacentes de estos. Mientras que la reducción de dimensionalidad, aunque aumenta la variedad de las características que describen los grupos, tiene efectos distintos para cada modelo. En suma, el hecho de que el conjunto de datos este conformado principalmente conformado por variables categóricas causó que ciertos modelos tuvieran un desempeño menor. Por ejemplo, el uso de *K-Means* no muestra grupos definidos y compactos a menos de que se usen únicamente variables de ubicación. Sin embargo, para este último caso, el modelo no presentó una segmentación que brindara información relevante sobre las denuncias de hurtos. Por otra parte, este modelo es el que generó el menor número óptimo de grupos.

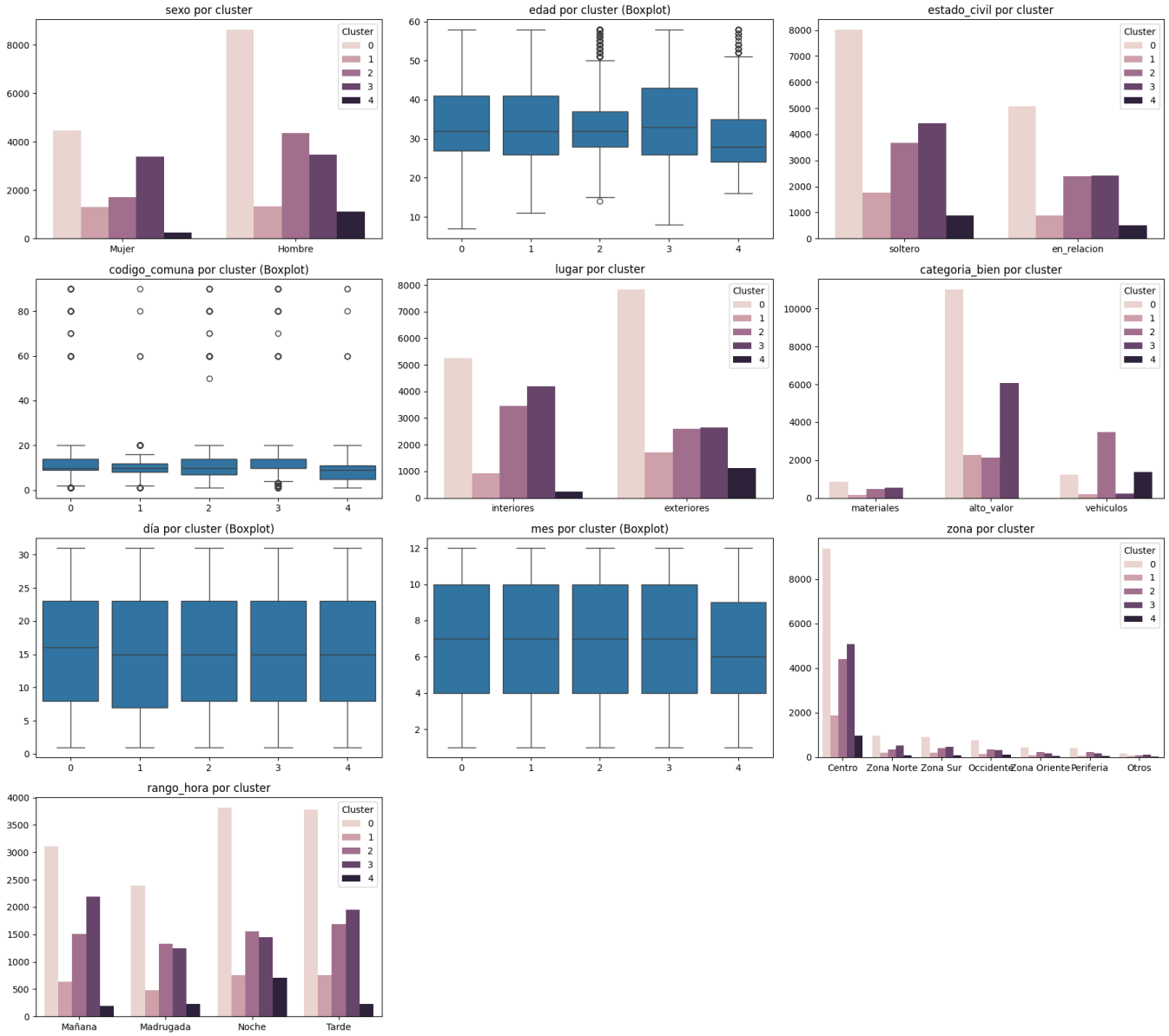
HDBSCAN por su parte, brinda una mejor interpretabilidad, encontrándose una caracterización separada de los robos de la ciudad y de sus corregimientos al usar los datos sin normalizar. Pero los resultados más relevantes de este modelo surgen al analizar el comportamiento de los grupos con las variables de latitud y longitud por si solas y su evolución a través de los años. Se muestra que para el primer caso hay grupos compactos y diferenciables en zonas relevantes de la ciudad. Mientras que, para el segundo caso, a medida que aumenta el tiempo y especialmente de 2015 en adelante, el número óptimo de grupos encontrados aumenta con los años (destacándose el gran aumento en el año 2020). En resumen, este modelo consiguió encontrar anomalías en los grupos y en la caracterización de estos.

Continuando con *K-Modes*, su uso con los datos normalizados y no normalizados no genera grupos separables a nivel geográfico, pero genera una mayor cantidad de grupos respecto a *K-Means*, los cuales tienen una mejor representatividad de otros tipos de hurtos y de otras características que no son observadas en otros modelos. Igualmente, al separar por conjuntos las variables, a pesar de que no se consiguió una separación a nivel geográfico, la caracterización de los grupos brindó los mejores resultados respecto a las variables. Lo cual es relevante dado que es consistente con los resultados de Ceballos Sánchez (2023), quienes mencionan que este algoritmo permite una mejor definición de los perfiles de criminalidad y tiene un buen comportamiento en presencia de una gran cantidad de variables categóricas en el conjunto de datos.

Por lo anterior, se presentan unas gráficas que describen el comportamiento de los grupos generados por los tres conjuntos de variables utilizados en *K-Modes*:

La primera es la **Figura 20** que se asocia al primer conjunto de variables, cuyas características son las siguientes:

Figura 20:Características subyacentes de los grupos con K-MODES (conducta, modalidad y arma_medio).



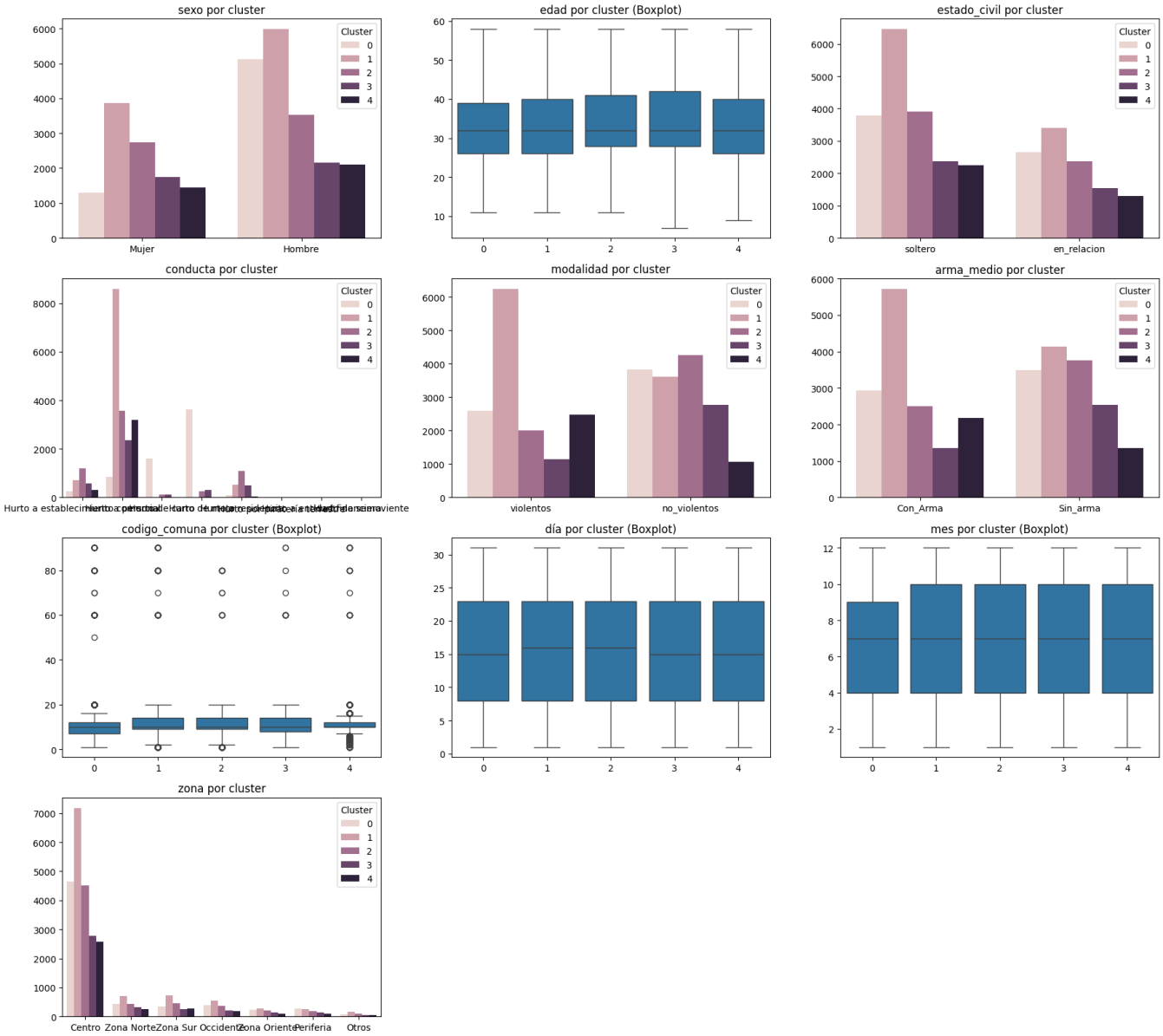
Fuente: Elaboración propia.

En este caso, el grupo que más destaca es el 0, el cual se caracteriza por tener hombres de alrededor de 32 años que están solteros. Estos denuncian ser robados en las horas de la tarde o noche. Estos hechos ocurren en exteriores y en la zona centro de la ciudad. En este grupo los hechos delictivos se realizaron con violencia y armas, además, los bienes hurtados son de alto valor. El comportamiento de este grupo y la gran cantidad relativa de registros que lo conforman se ve explicado por el comportamiento promedio de las variables de los datos originales. Similarmente el comportamiento de los datos se ve permeado por la conducta más representativa: hurto a personas.

Otro grupo relevante es el 2, el cual se relaciona a denuncias de hurtos a personas, los cuales no fueron violentos y no hubo armas. A su vez, se identifica por tener hechos delictivos donde se roban más vehículos y ocurren más frecuentemente en interiores. Luego, el grupo cuatro está conformado por denuncias de hurtos de motos, no violentos y sin armas, donde las que más denuncian son las mujeres y les roban principalmente en las noches.

La siguiente, es la **Figura 21** que muestra el comportamiento del segundo conjunto de variables:

Figura 21:Características subyacentes de los grupos con K-MODES (lugar, categoria_bien y rango_hora).

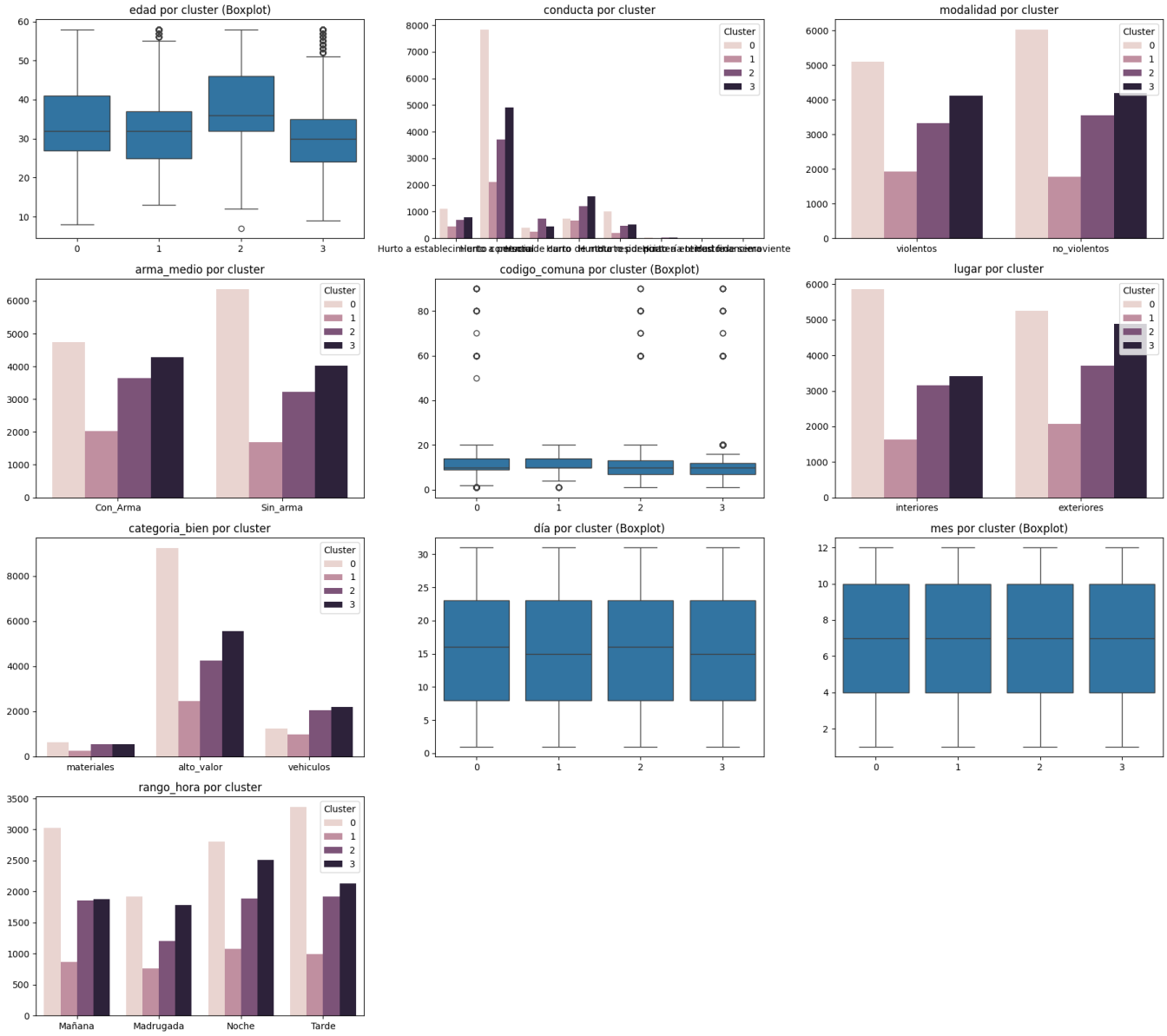


Fuente: Elaboración propia.

Para el segundo conjunto de características, se presentan grupos con características similares a los del conjunto anterior. Lo que indica una consistencia en los resultados. En este caso, se destaca el grupo 0, el cual son denuncias a robos en exteriores en la noche, donde se roban más vehículos. Este conjunto de características permite una mayor variedad de casos distintos a hurto a personas y tiene una menor variabilidad en la variable de mes.

La **Figura 22** muestra el comportamiento de los rasgos de cada grupo obtenido con *K-Modes* y con el tercer conjunto de variables:

Figura 22:Características subyacentes de los grupos con K-MODES (sexo, estado_civil, zona).



Fuente: Elaboración propia.

Para el tercer conjunto de características, hay una mayor variabilidad en todo los grupos en lo que respecta a la edad. En este se destaca el grupo cero, el cual caracteriza a denunciantes mujeres solteras, las cuales son hurtadas en el centro de la ciudad de forma no violenta, sin arma, en interiores, también se relaciona hechos que ocurren principalmente en horas de la tarde y donde los bienes son de alto valor. También es relevante el grupo uno, el cual tiene mayor balance en las categorías respecto a los demás y se ubica en la zona norte.

En cuanto a variables notables, se encuentra que: conducta, modalidad, arma medio, sexo, estado civil, zona, comuna, categoría de bien y rango de hora, tienen una incidencia relativa significativa en el comportamiento de los grupos. Mientras que las variables de día y mes tienen un efecto menor y suelen ser más aleatorias.

Los anteriores resultados muestran tendencias similares a las de los datos limpios y es consistente con los datos originales. Es importante mencionar que, el comportamiento de las agrupaciones obedece a que los datos se encuentran principalmente conformados por denuncias de hurtos a personas, por lo que las peculiaridades de estos hurtos suelen destacarse y en ocasiones permean el comportamiento de todas agrupaciones. Por lo que se recomienda que, en futuros trabajos, se realice un análisis separado de las demás conductas (tipos de hurtos) y zonas

distintas al centro de la ciudad, con el objetivo de evidenciar el efecto del grueso de la información sobre estos. Por otra parte, también se recomienda hacer uso de distintas transformaciones de los datos y de modelos de agrupamiento distintos.

Asimismo, es pertinente señalar que, si bien los registros de denuncias constituyen un indicador relevante de la actividad delictiva, estos no reflejan en su totalidad los robos que tienen lugar en el ámbito urbano. Esto se debe a que, por diversas razones, un porcentaje de robos no es denunciado, como consecuencia del temor de las víctimas. Para elaborar un análisis exhaustivo de los hurtos en el ámbito urbano, resulta imperativo examinar este porcentaje y abordar otras temáticas y datos que puedan ser pertinentes para comprender los hurtos ocurridos en las ciudades.

Todo el procesamiento y modelado realizado en este trabajo se encuentra en el siguiente repositorio GitHub: https://github.com/Sfranco12/cluster_hurtos_medellin (Franco Franco & Giraldo Martínez, 2025)

4. Conclusiones

En lo que respecta a los algoritmos estimados, se ha comprobado que *HDBSCAN* genera los grupos más separados y compactos, resultado que mejora al usar sólo las variables de ubicación. Por su parte, *K-Modes* cuenta con grupos mejor definidos que para el caso de *K-Means*, pero se destaca más en la interpretabilidad de los centroides La normalización de los datos causa efectos diferentes para cada modelo, sin embargo, se evidencia una particularidad general para todos: se incrementa la diversidad de rasgos en los centroides, generando grupos con menor similitud entre sí. En suma, la normalización posibilitó la obtención de mejores resultados en la visualización de los grupos para *K-Means*. En lo que respecta a la reducción de dimensionalidad, esta presenta efectos distintos en función del modelo en cuestión. En consecuencia, la mayoría de los casos analizados presentan diversos números de grupos óptimos, lo que dificulta la caracterización de los hurtos en la ciudad.

En lo que respecta a los datos, considerando la elevada incidencia de hurtos a personas y de las denuncias asociadas al centro urbano, se evidencia la necesidad de optimizar la caracterización de los demás robos y las zonas circundantes. En este sentido, se recomienda la implementación de diversos conjuntos o segmentaciones de los datos, con el propósito de refinar aún más los atributos de los grupos y describir con mayor precisión las denuncias que no se ajustan a estas categorías. Como se desprende del estudio realizado, los grupos con mayor cantidad de datos presentan las siguientes características: sexo masculino, estado civil soltero, ubicación en el centro urbano, hurto a personas y bienes de alto valor. En segundo lugar, se ha de considerar la premisa de que dichos rasgos presenten una alta frecuencia a lo largo de todas las segmentaciones generadas.

Se encontró a su vez, que, respecto a las variables de día y mes del año, no hay un patrón fijo asociado en los resultados de los casos analizados. Pero la variable relacionada a hora muestra que las tardes y las noches tienen una incidencia considerable en los grupos, mientras que las madrugadas tienen menor efecto. Las variables asociadas a zonas muestran que el centro de la ciudad es la zona más relevante del estudio, pero *HDBSCAN* permitió observar otras zonas importantes como: Laureles-Estadio y Poblado-Guayabal. Luego, aunque en varios caso se logra caracterizar los corregimientos de la ciudad, se evidencia que estos tienen una menor cantidad de denuncias.

Finalmente, además de la recomendación de probar distintos subdivisiones de los datos, se recomienda que futuros trabajos tomen en cuenta la incidencia que puede tener la propensión a denunciar o no este tipo de delitos, para tener un panorama más completo de los mismos. A su vez, es importante que se prueben más modelos, técnicas de análisis (por ejemplo, hacer uso de técnicas de estadística espacial usados en trabajos similares) y datos que se relacionen a la naturaleza de los datos y a las dinámicas de la ciudad.

Referencias

Amoako, E. A. (2021). A Spatial Analysis of Robbery Rate in the City of Detroit using Exploratory Data Analysis Approach. *Proceedings of the ICA, 4*, 1–8. https://doi.org/10.5194/ica-proc-4-6-2021

Arévalo Álvarez, J. C., & Fernández García, M. A. (2022). *Analítica de datos 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]. https://hdl.handle.net/10495/29063

Ceballos Sánchez, J. D. (2023). *Clasificación de crímenes por zonas en la ciudad de Nueva York utilizando técnicas de Aprendizaje Automático no Supervisado* [Trabajo de grado especialización, Universidad de Antioquia]. https://hdl.handle.net/10495/35747

El Colombiano. (2023, August 3). *Medellín, segunda capital del país en número de hurtos por día*. https://www.elcolombiano.com/medellin/medellin-es-la-segunda-ciudad-del-pais-en-hurtos-tambien-han-crecido-extorsiones-y-desplazamiento-intraurbano-JH22075804

Franco Franco, S., & Giraldo Martínez, A. (2025). *cluster_hurtos_medellin [Repositorio de código]*. Github. https://github.com/Sfranco12/cluster_hurtos_medellin

MEData. (2025). *Seguridad y Defensa*. https://medata.gov.co/search/?fulltext=hurto&theme=Seguridad%20y%20Defensa

Medellín Cómo Vamos. (2021). *Informe de calidad de vida de Medellín, 2020*. https://www.medellincomovamos.org/system/files/2021-09/docuprivados/Seguridad%20Informe%20de%20Calidad%20de%20Vida%20de%20Medell%C3%ADn%2C%202020.pdf

Medellín Cómo Vamos. (2022, November 25). *Pilas en la calle*. https://www.medellincomovamos.org/hurtos-en-medellin