Cómo citar este artículo: Maestre-Gongora, G., Acuña-Castellanos, C. A., Londoño-Bedoya, E., & García-García, S. (2023). Análisis de datos sobre los hurtos en la ciudad de Medellín desde un enfoque descriptivo. Revista de Investigación, Desarrollo e Innovación, 13 (1), 173-184
doi: https://doi.org/10.19053/20278306.v13.n1.2023.16059

Análisis de datos sobre los hurtos en la ciudad de Medellín desde un enfoque descriptivo

Data analysis of thefts in the city of Medellin from a descriptive approach

Gina Maestre-Gongora[1]
Camilo Andrés Acuña-Castellanos[2]
Edwar Londoño-Bedoya[3]
Sergio García-García[4]

Recibido: marzo 23 de 2022
Aceptado: octubre 12 de 2022

Resumen

Este artículo tiene por objetivo identificar las tendencias y patrones de hurto en la ciudad de Medellín en el periodo 2014-2020, usando datos abiertos de gobierno. Se utiliza como metodología la inteligencia de negocios para el análisis de datos descriptivo. Se analizan variables como barrios, modalidades, tipo de hurto y se realiza la predicción de la variable modalidad de hurto. Los resultados muestran que históricamente el segundo semestre del año tiene la mayor tendencia de incidencias, donde la mayoría de robos suceden en los lugares públicos con un 60% sin el uso de armas. Se identificó que, debido a la pandemia de COVID, las tendencias históricas presentaron alteraciones notables, pero una vez levantadas las restricciones, estas retomaron las tendencias de alzas en robos en las condiciones de prepandemia. Se concluye que el análisis de datos abiertos brinda información relevante para la toma de decisiones de los ciudadanos.

Palabras clave: datos abiertos, robo, aprendizaje automático, inteligencia de negocios.

Abstract

This article aims to identify trends and patterns of theft in the city of Medellin in the period 2014-2020, using open government data. The methodology used is business intelligence for descriptive data analysis. Variables such as neighborhoods, modalities, type of theft, and the prediction of the theft modality variable are analyzed. The results show that historically the second half of the year has the highest trend of incidences, where most thefts occur in public places 60% without the use of weapons. It is shown that due to the COVID pandemic, historical trends showed significant changes, but once the restrictions were lifted, they resumed the trends of increases in thefts in pre-pandemic conditions. It is concluded that the use of open data analisys gives information to improve the decision-making of the citizens.

Keywords: open data, theft, machine learning, business intelligence.

1 Ingeniera de Sistemas, Doctora en Ingeniería de Sistemas y Computación, Universidad Cooperativa de Colombia, Medellín, Colombia. E-mail: gina.maestre@campusucc.edu.co
Orcid: https://orcid.org/0000-0002-2880-9245

2 Ingeniero de Software, Universidad Cooperativa de Colombia, Medellín, Colombia. E-mail: camilo.acuna@campusucc.edu.co
Orcid: https://orcid.org/0000-0001-5258-2469

3 Ingeniero de Software, Universidad Cooperativa de Colombia, Medellín, Colombia. E-mail: edwar.londono@campusucc.edu.co
Orcid: https://orcid.org/0000-0001-6745-0880

4 Ingeniero de Software, Universidad Cooperativa de Colombia, Medellín, Colombia. E-mail: Sergio.garciag@campusucc.edu.co
Orcid: https://orcid.org/0000-0002-7419-0025

1. Introducción

Movilizarse en una ciudad tiene sus riesgos y la capital antioqueña, Medellín, no es un caso ajeno a esta situación. Uno de estos riesgos son los hurtos que se presentan día a día, es una problemática social con un impacto altamente negativo en la ciudadanía. "El hurto es el delito que más se denuncia en Colombia y el robo a personas es la modalidad más frecuente, con 21 hurtos cada hora, 419 personas afectadas por día y 77.100 denuncias; seguida del robo de celulares, con 15 hurtos cada hora, 272 aparatos robados por día y 49.949 denuncias" (Acuña et al., 2020). Lo anterior hace que los ciudadanos no vayan tranquilos por las calles que transitan, teniendo constantemente la preocupación de que les roben sus pertenencias. Estas cifras son del primer semestre del 2019, mientras que las del 2020 son totalmente distintas al haber estado toda la población en un periodo de cuarentena, disminuyendo excesivamente los hurtos en el país.

Es imposible saber a ciencia cierta el lugar donde ocurrirá un asalto o un robo, pero las técnicas basadas en la tecnología de aprendizaje de máquina brindan un acercamiento a este aspecto. Las predicciones se obtienen a través de un proceso de tratamiento de paquetes con gran cantidad de datos, pero lo más importante es que estos pueden ser públicos y están al alcance de cualquier persona, lo que permite experimentar con ellos y crear soluciones ágiles a problemas cotidianos (Wieczorkowski, 2019). En el caso de Colombia, estos datos son proporcionados por el gobierno nacional, en su plataforma Datos Abiertos.

Los datos abiertos son usados como materia prima desde diferentes ámbitos, donde la calidad de los datos implica un conjunto de características, valores y expresiones que se construyen de forma iterativa (Maestre-Gongora et al., 2021). Sin embargo, este conjunto de datos no siempre está organizado y listo para ser usado inmediatamente, lo que requiere un proceso de limpieza de datos, tarea que abarca el 70% del proyecto según algunos autores, resaltando la importancia de tener datos ordenados y de calidad (Pumares-Romero, 2019; Acuña et al., 2020).

Los datos abiertos deben generar capacidades de disponibilidad y uso de los datos públicos por parte de los ciudadanos y las organizaciones, para producir información valiosa y apoyar la toma de decisiones (Maestre-Gongora & Nieto-Bernal, 2019). Los datos abiertos se han usado en análisis del crimen, por ejemplo: en Portland-USA (Nguyen et al., 2017), en la India (Telugu-Maddileti et al., 2020), y en Mississippi-USA (McClendon & Meghanathan, 2015). Treviño y coautores (2020), exponen que el análisis de datos es un puente que conduce hacia la comprensión y la correcta interpretación de las necesidades de información que tienen los consumidores, las cuales se ven reflejadas a través de datos históricos o nuevos, capaces de ser interpretados.

En cuanto al análisis de datos, Pérez-Rave et al. (2019), emplean una metodología dividida en cuatro etapas, donde: preparan, analizan y visualizan datos gubernamentales y abiertos en una web para estudiar casos de accidentalidad vial en Medellín. Pumares-Romero (2019), por su parte, plantea la problemática de la alta accidentalidad en las vías ecuatorianas para monitorear e identificar patrones usando un dataset gubernamental, resaltando que los datos abiertos son un insumo importante para dar soluciones a fenómenos sociales. A su vez, la completitud, trazabilidad y conformidad de los datos de las plataformas colombianas pueden ser testeados como lo muestran López y Mahecha (2017).

La inteligencia de negocios se define como el conjunto de metodologías, aplicaciones, prácticas y capacidades enfocadas a la creación y administración de información que permite tomar mejores decisiones a los usuarios de una organización (Díaz, 2012). La Inteligencia de Negocios o Business Intelligence (BI), es "una herramienta bajo la cual diferentes tipos de organizaciones pueden soportar la toma de decisiones basadas en información precisa y oportuna, garantizando la generación del

conocimiento necesario que permita escoger la alternativa que sea más conveniente para el éxito de la empresa" (Rosado & Rico, 2010, P. 1). Entre las principales ventajas de la inteligencia de negocios, se resaltan: Brinda una perspectiva única, histórica, persistente y de calidad de toda la información; mejora la comprensión y documentación de los sistemas de información en las organizaciones; mejora la competitividad de la organización accediendo más rápido a la información, con mayor agilidad en la toma de las decisiones, monitoreo de los procesos críticos de negocio y las actividades, usando métricas.

Este artículo tiene por objetivo identificar las tendencias y patrones de hurto en la ciudad de Medellín en el periodo 2014-2020, usando datos abiertos de gobierno. Se trata de un análisis descriptivo que utiliza inteligencia de negocios a partir de datos abiertos de gobierno y siguiendo la metodología de análisis de datos, mediante el diseño y construcción de dashboards.

2. Metodología

La metodología está orientada al proceso de análisis y tratamiento de datos abiertos, donde el objetivo es usar la ciencia de datos para estudiar comportamientos y reflejarlos en una visualización que permita inferir sobre una conducta futura. En este caso se empleó la metodología de cuatro fases propuesta por Pérez-Rave et al. (2019), que se resume a continuación:

2.1 Planificación

El caso de estudio se enfoca en la ciudad de Medellín – Colombia, capital del departamento de Antioquia, ubicada en el Valle de Aburrá (Latitud: 6.217 Longitud: -75.567). La ciudad cuenta con una población de 2.533.424 habitantes, dato al año 2020. Se encuentra a una Altitud de 1.495 m.s.n.m y cuenta con una temperatura promedio al año de 22°. La ciudad de Medellín se encuentra zonificada en 16 comunas las cuales fueron tenidas en cuenta para el estudio, junto con los barrios asociados. En la investigación no se incluyeron los corregimientos aledaños, como: Altavista, Santa Elena, San Cristóbal, Palmitas y San Antonio de Prado.

El dataset utilizado se obtuvo de la página web MeData (http://medata.gov.co), perteneciente a la alcaldía de Medellín. El dataset seleccionado fue "hurto a personas" y cuenta con un total de 215.279 registros, desde el año 2003 hasta septiembre del 2020. Inicialmente, se evaluaron cada una de las columnas para determinar cuáles no aportan información al proyecto o no cuentan con suficiente calidad en sus registros, eliminando 12 de las 35 columnas del dataset inicial.

Posteriormente, se renombraron las columnas usando el método 'pandas.DataFrame.rename' para tener mayor facilidad en la gestión de los datos, eliminando: guiones, espacios y palabras innecesarias, entre otras. A manera de ejemplo, se modificó: seguridad.fecha_hecho por fecha; seguridad.latitud por latitud; seguridad.longitud por longitud, para facilitar la gestión de las columnas.

2.2 Preparación de datos

En este paso se eliminaron los registros duplicados usando el método 'pandas.DataFrame.drop_duplicates', obteniendo un nuevo dataset con un total de 194.565 registros. Además, a partir de la columna fecha se generaron otras cinco columnas: día, mes, anio, día, semana y hora, para facilitar el análisis de datos cuantitativo.

Asimismo, se usó la técnica de reducción de dimensionalidad, también llamada agrupación de variables, haciendo uso del método 'pandas.DataFrame.loc'. Esta técnica se aplicó en algunas columnas en donde sus variables comparten la misma naturaleza y solo aumentan el volumen de información innecesaria o repetida. Por ejemplo, las variables Metro y Estación Metro de la columna Lugar fueron agrupadas en una sola variable llamada Metro. Esto debido a dos razones principales: la primera, reduce las dimensiones del dataset al ser dos cantidades

de registros totalmente distintas y mejora el rendimiento al implementar un modelo de aprendizaje automático. La segunda, se debe a que no hay un razonamiento significativo si el hurto ocurrió en la estación o dentro del metro como tal, ya que el hurto ocurrió dentro de las instalaciones de este servicio de transporte público.

Al culminar este proceso, se obtuvo un dataset que cuenta con un total de 194.565 registros, los cuales en una última instancia pasaron por una técnica de homogenización, dándoles un ID único a cada uno de los registros en una nueva columna llamada id, usando para ello el método 'pandas.DataFrame.reindex'. El total de columnas del dataset para este es de 23.

2.3 Análisis descriptivo de datos abiertos de Medellín

Finalizado el proceso de limpieza de datos, se generaron visualizaciones a través del software de Microsoft PowerBI. Se crearon diferentes visualizaciones para evidenciar la información del dataset, usando: gráficos de barras, columnas, barras apiladas, columnas apiladas, gráfico de columnas apiladas y líneas, líneas, gráfico circular, gráfico de dispersión, mapas de calor, outliers y un forecasting. A parte de las gráficas, el software de BI permite conocer la media, la moda, el conteo de cada variable por columna, entre otras. De esta manera, a través de las visualizaciones se puede inferir el comportamiento de los datos fácilmente para que el usuario identifique por cuenta propia comportamientos respecto a los hurtos. Además, el tablero permitió tomar decisiones para el entrenamiento del modelo.

2.4 Modelos de aprendizaje automático

Para el primer acercamiento con los modelos se contó con un dataset más reducido, debido a que se realizó un nuevo proceso de limpieza de datos. El dataset final para el entrenamiento del modelo, después del preprocesamiento de limpieza, cuenta con 178.903 registros y 16 columnas. Por último, seis columnas se convirtieron a tipo numérico usando el método 'pandas.DataFrame.loc' para facilitar el entrenamiento del modelo.

Luego se procedió al entrenamiento del modelo de aprendizaje automático que se realizó con la librería AutoGluon, usando el método 'TabularPredictor', el cual genera un modelo para predicción de los valores de una columna basada en valores de otras columnas. También se empleó la librería Scikit-learn con el método 'train_test_split', en donde el dataset se separó en un 70% para el entrenamiento y en otro conjunto de 30% para prueba: dataset_entrenamiento.csv y dataset_prueba.csv, respectivamente. Luego, autogloun requiere que se le envíe el número de datos para el entrenamiento, donde se utilizó el método 'len()' aplicándolo al dataset de entrenamiento. Después de determinar la cantidad de los datos de entrenamiento y el nombre de la columna a predecir, se hace el llamado al método 'TabularPredictor().fit()' el cual requiere que se le envíe los siguientes atributos; nombre de la columna a predecir, el nombre de la carpeta en donde se guardará el modelo, y en .fit() se le envía el dataset de entrenamiento.

Por último, se realiza la prueba del modelo utilizando el método 'TabularPredictor.load()', el cual recibe como atributo la ruta del modelo. Luego se usa el método 'predictor.evaluate_predictions()', el cual recibe como atributos el conjunto de datos de prueba y el nombre de la columna a predecir. Los resultados obtenidos se evidencian respecto a cada registro y la columna a predecir, que es modalidad.

Las herramientas de procesamiento utilizadas fueron el lenguaje de programación Python 3.8.0 y el software para inteligencia de negocios PowerBI Desktop Version. También se emplearon otras herramientas que permitieron el entrenamiento del modelo y la limpieza de los datos, como: AutoGluon 0.1.0; Pandas 1.2.3;Scikit-lear 0.24.1; XAMPP 3.2.4.

3. Resultados y discusión

Como resultado final del procesamiento en la limpieza de datos, se obtienen las columnas descritas en la Tabla 1. Es de anotar que el dataset cuenta con 23 columnas y 194.565 registros.

Tabla 1. Metadatos Dataset Final.

Nombre Columna	Descripción
ID	Número de identificador único por registro
dia	Día del robo
mes	Mes del robo
anio	Año del robo
día semana	Día de la semana del robo
hora	Hora exacta del robo
latitud	Localización del lugar del hurto
longitud	Localización del lugar del hurto
sexo	Género de la persona que sufrió el hecho
edad	Edad de la persona
estado civil	Estado civil de la persona
transporte	Forma de transporte en la que se encontraba la persona
modalidad	Tipo de robo
conducta especial	Tipo de robo exacto
arma	Arma que fue usada para el robo
barrio	Barrio donde sucedió el robo
código barrio	Código de barrio único.
código comuna	Código de comuna único.
lugar	Especificación donde sucedió el robo.
fecha	Fecha del hurto en formato (DD-MM-AAAA-HH:MM)
bien	Bien que fue robado.
grupo bien	Agrupación del bien.
categoría	Categorizar el bien.

Con el dataset definitivo se genera un dashboard descriptivo en la plataforma de Power BI, con posibilidad de consultas mediante filtros a variables, como: barrios y comunas con más incidencias, tipo de robo, arma usada, día de la semana y meses de incidencias. Con los datos actualizados se generan las visualizaciones a través de scripts de programación en Python y el software de Microsoft, PowerBI. El informe se segmenta por páginas, con la siguiente información: estadísticas generales, outliers y mapa. De esta manera, a través de las visualizaciones se puede inferir el comportamiento de los datos fácilmente, todo esto a través de técnicas de análisis descriptivo.

Los hurtos en la ciudad de Medellín desde el año 2010 han ido aumentando exponencialmente tal como se ve en la Figura 1, teniendo como punto más alto el año 2019 con un total de 43.411 atracos. Posteriormente y debido a condiciones de la pandemia del COVID 19, los continuos

confinamientos y las restricciones de movilidad, los hurtos disminuyeron en el año 2020 a más de la mitad. Teniendo en cuenta que los datos tienen fecha de corte en el mes de septiembre y que este gráfico puede aumentar a medida que el dataset se actualice, se evidencia una tendencia a la baja.

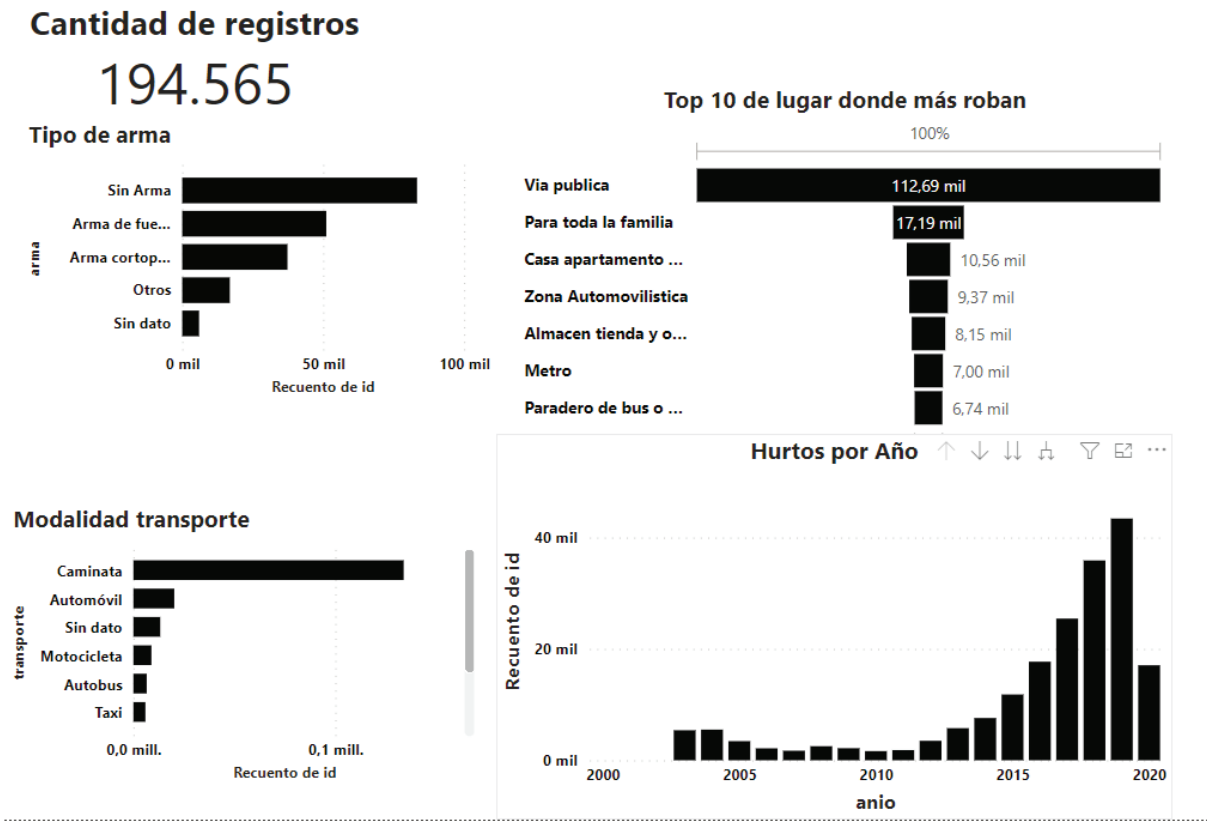

Figura 1. Estadísticas Generales de Hurto en Medellín 2014-2020.

Respecto al comportamiento de los hurtos y los lugares donde estos ocurren, se realizó una gráfica de embudo que permite ver los 5 lugares donde más delitos de este tipo ocurren. El más alto es la vía pública, con más de 112.000 registros (60%). Se destaca que el metro y los paraderos de buses, a pesar de estar en los últimos lugares, siguen sumando entre los dos más de 13.000 hurtos. Se entiende que la vía pública sea la categoría con más registros y que en la Figura 1, donde se identifica el medio de transporte en el que se encontraba la persona al momento de ocurrir el robo, sea la variable caminata la que lidera esta gráfica. En este sentido, se puede inferir que los peatones que transitan por la vía pública son los más suceptibles a ser despojados de sus pertenencias. Cabe aclarar que la vía pública comprende calles, puentes y callejones, entre otros.

En la Figura 2 se indica el tipo de arma que más se usó en cada barrio para el hurto. En este caso, se observa que el barrio predominante con más datos en todos los tipos de arma es La Candelaria, ubicado en el centro de Medellín. No obstante, es de anotar que predominan los hurtos donde no se evidencia el empleo de ningún tipo de arma, correspondiente al top 3: barrios con más registros.

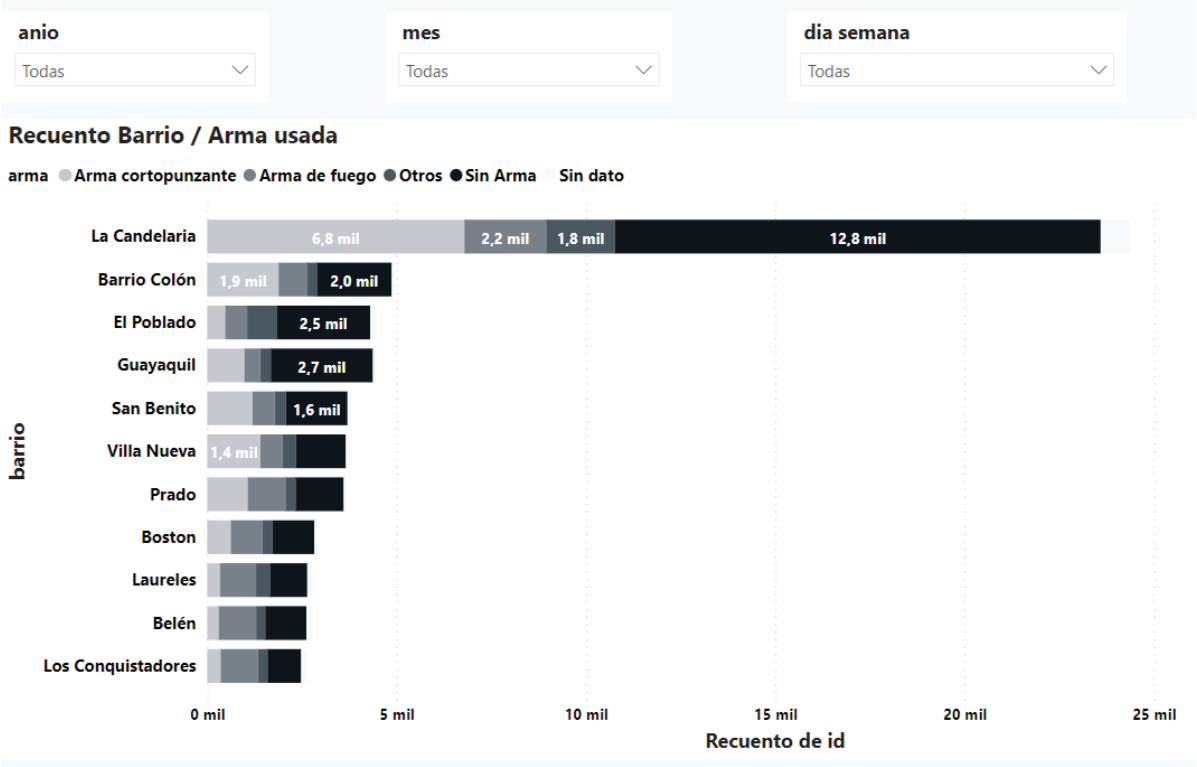

Figura 2. Barrios y tipo de arma empleada en hurtos en Medellín 2014-2020.

Además, se puede observar que el arma cortopunzante ocupa el segundo lugar como arma más usada en los registros. Esto significa que es más probable sufrir un ataque con arma blanca que ser hurtado con un arma de fuego. Para concluir esta visualización, se puede observar que en los cuatro últimos barrios de la figura 2 se igualan los registros de Sin Arma junto con Arma de fuego, destacando que, en barrios como Belén, Laureles y Boston, el arma de fuego es la más usada a la hora de cometer estos delitos.

Para corroborar lo anterior, se hizo la comparación de las comunas de la ciudad de Medellín versus la modalidad de hurto con la que se hurtó a la víctima, siendo Atraco la modalidad que más se presenta en la comuna con más registros (La Candelaria, comuna 10). Si se sumase la cantidad de registros que tiene la variable Descuido o cosquilleo y la variable de Raponazo o forcejeo, se superaría en cantidad de registros a Atraco. Esto reafirma que la categoría Sin Arma es la que más se presenta en barrios céntricos de la ciudad, porque los ladrones usan técnicas que no emplean la intimidación, sino que buscan despojar rápidamente a la víctima de su bien.

Adicionalmente, se diseñó un mapa que permite visualizar los puntos donde han ocurrido los hurtos. Al haber más de 190.000 registros, se diseñó un mapa interactivo que permite el filtro por año, mes, modalidad y comuna y la geolocalización del hurto, como muestra la Figura 3. Si varios hurtos acontecen en el mismo lugar, el punto graficado tiene un radio más grande.

Figura 3. Ejemplo de Consulta de Mapa de Hurtos de Medellín, por año, mes, modalidad y arma.

Continuando el análisis a nivel descriptivo, se buscaron anomalías en los datos como se observa en la Figura 4, registros que por tendencia se espera que sean de una manera, pero resultaron de otra. En este caso se analiza el año 2019 ya que es el que cuenta con el número más alto de hurtos. En la Figura 4 se observan distintos picos que se presentaron a lo largo de ese año, siendo los tres más altos de estos en el mes de agosto, correspondiente a la Feria de las Flores. Esta feria es un evento en el que, por la incidencia del alto número de visitantes y festividades, se convierte en un escenario donde las condiciones de ruido, caos y descuido de los habitantes, facilita la práctica del hurto.

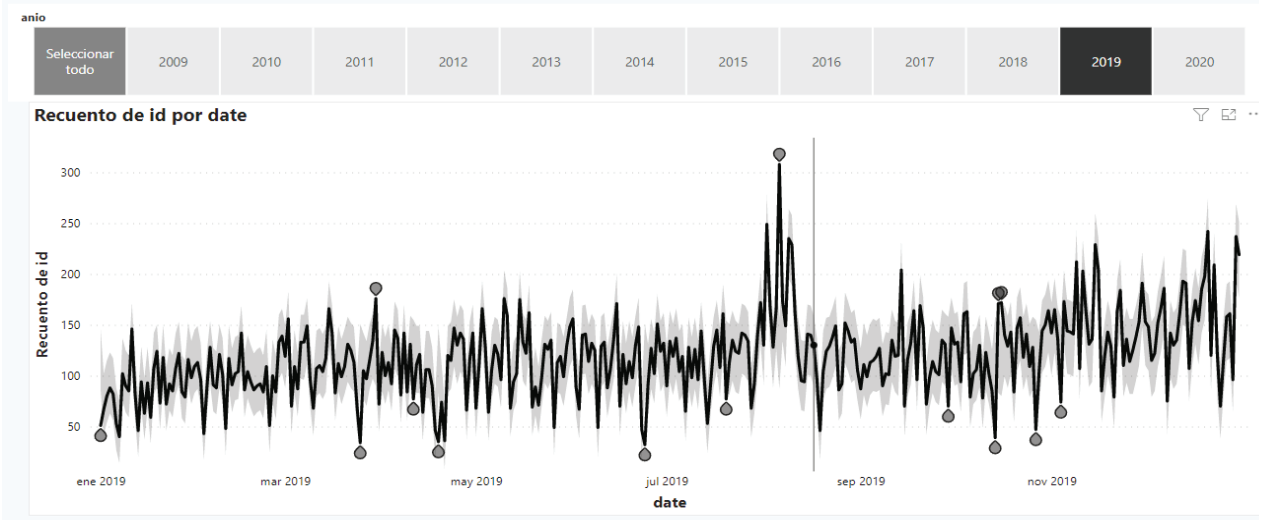

Figura 4. Número de hurtos del año 2019.

Por ejemplo, en 2019 las ferias iniciaron el día 2 de agosto, ese mismo día se presentó la primera alza de los hurtos con una cantidad total de 249. Según la gráfica, una posible explicación para este crecimiento repentino en los registros sería el aumento de hurtos, donde el arma registrada era Otros, es decir, armas poco convencionales pero que por la naturaleza de la festividad que se celebra aumentó, tales como: tóxico o químico, escopolamina, objeto pesado o contundente.

En la tabla 2 se observa una matriz de meses del año frente a días de la semana, donde el tono más oscuro es donde ocurrieron más hurtos y el tono más claro significa menos hurtos. Se observa una tendencia sin filtrar por año pero con todos los meses, evidenciando que cuando ocurren más hurtos son los jueves, viernes y sábado. El domingo en todos los meses tiene un comportamiento constante, infiriendo que ocurren menos atracos por ser el día de descanso para las zonas de trabajo y estudiantil. El comportamiento para los 12 meses indica que los picos altos del año se presentan en el segundo semestre, debido a las épocas definidas como "temporadas altas", las cuales son: mitad de año época de vacaciones de estudiantes, agosto donde se celebra la fiesta local (Feria de las Flores), y por último el mes de diciembre, época de navidad y fin de año. Asimismo, la época del año en la que ocurren menos hurtos corresponde a los meses donde la ciudad se encuentra "vacía": es decir, los periodos de vacaciones tradicionales para los estudiantes y trabajadores, usualmente los primeros meses del año.

Tabla 2. Matriz de correlación entre meses y días

	Lunes	Martes	Miércoles	Jueves	Viernes	Sábado	Domingo
Enero	2008	2372	2652	2438	2755	2414	1599
Febrero	2182	2121	2314	2103	2285	2390	1397
Marzo	1882	2036	2274	2190	2471	2277	1393
Abril	2308	2056	2069	1785	1951	2046	1252
Mayo	1898	2105	2458	2403	2622	2434	1344
Junio	1649	2136	2126	2291	2586	2644	1473
Julio	2270	2436	2799	2461	2580	2239	1327
Agosto	2318	2718	2846	2935	3205	2875	1945
Septiembre	2176	2282	2329	2343	2730	2686	1378
Octubre	2181	2626	2827	2568	2524	2526	1444
Noviembre	1927	2447	2518	2581	2749	2687	1542
Diciembre	2597	2278	2087	2295	2337	2578	1870
			0-1500	1500-2500	2500-3000	Mayor a 3000	

Finalmente, se realizó un ejercicio de predicción mediante la herramienta de Autogloun – TabularPredictor, para análisis predictivo. Como se evidencia en la Tabla 3, Autogloun somete el conjunto de datos a diferentes modelos, entregando como resultado la precisión, el tiempo de entrenamiento y el tiempo de validación de cada modelo. Entre los modelos utilizados por Autogloun, están: Redes neuronales, Bosques Aleatorios, Arboles de decisión y Gradientes.

Tabla 3. Resultados AutoML- AutoGluon.

Modelo	Tipo	Precisión	Tiempo de entrenamiento (s)	Tiempo de validación (s)
NeuralNetFastAI	Redes Neuronales	0.7736	490.29	0.38
KNeighborsUnif	K Vecinos más cercanos	0.5064	0.9	0.1
KNeighborsDist		0.5404	0.92	0.02
RandomForestGini	Bosques Aleatorios	0.7932	11.38	0.07
RandomForestEntr		0.7928	13.54	0.06
ExtraTreesEntr	Arboles de decisión	0.7824	11.78	0.08
ExtraTreesGini		0.7933	3.33	0.04
LightGBM	Gradientes	0.7732	32.63	0.13
LightGBMXT		0.7716	20.32	0.2
CatBoost		0.7556	255.06	0.15
XGBoost		0.7716	125.07	0.78
LightGBMLarge		0.7828	18.41	0.06
Tiempo total (s)			1218.86	

En la Tabla 3 los diversos modelos muestran en la mayoría de los casos un buen comportamiento, ya que por ejemplo los modelos de tipo Redes neuronales, Bosques aleatorios, Arboles de decisión y Gradientes tienen una precisión de más del 70%. En contraste, los modelos de K Vecinos más cercanos muestran un bajo rendimiento, ya que estos tienen una precisión del 50%. En cuanto al tiempo de entrenamiento, los modelos de tipo K Vecinos más cercanos son los de menor tiempo de entrenamiento, lo cual se traduce en un menor costo computacional, mientras que los modelos con más costo computacional son los de Redes neuronales y los de tipo Gradiente.

El uso de AutoGloun evidencia que el mejor modelo para el tipo de proyecto planteado es ExtraTreesGini, ya que su precisión respecto a los demás es mayor, con un valor de 79,33% y tiene un tiempo de entrenamiento mucho menor a los demás, con 3,33 s. Este modelo es mucho mejor que los otros aunque en su naturaleza se trata de un modelo de tipo árboles de decisión, lo que indica que se puede trabajar con cualquier otro modelo del mismo tipo, razón por la que estos modelos obtuvieron una precisión más alta que los demás.

Después de realizar el entrenamiento, se llevó a cabo la prueba del modelo con el dataset_ prueba.csv, como lo muestran las Tablas 4 y 5. En la tabla 4, la columna ID_registro hace referencia al registro de un ciudadano, que cuenta con la siguiente información: día, mes, año, día semana, hora, minutos, sexo, edad, transporte y código comuna. La columna predicción corresponde a la modalidad en que puede ser hurtado.

Tabla 4. Variable Predicción modalidad de hurto

ID_registro	Predicción
0	A Raponazo o Forcejeo
1	Descuido o Cosquilleo
3	Descuido o Cosquilleo
6	Descuido o Cosquilleo
9	Atraco
.....	
53902	Descuido o Cosquilleo
53903	Atraco
53904	Descuido o Cosquilleo
53905	Atraco
53906	Atraco

Tabla 5. Precisión de las variables

Variable	Precisión
Atraco	0.96
Descuido o Cosquilleo	0.64
Raponazo o Forcejeo	0.50
Rompimiento de propiedad	0.53
Otros	0.43

La Tabla 5 hace referencia al porcentaje de precisión de la predicción en cada una de las variables, evidenciando la certeza de la predicción. En el caso de un Atraco, se tiene un 96% de probabilidad de que esta sea la modalidad del hurto de la que se puede ser víctima.

4. Conclusiones

Se concluye que los hurtos en la ciudad de Medellín ocurren en su mayoría en el contexto de lugares públicos, orientado a la sustracción de objetos pequeños como tecnología o dinero. En estos casos, si bien puede haber intimidación, en su mayoría no hay uso de armas por lo que el factor de confianza o descuido del afectado podrá incidir significativamente.

La analítica de datos es una técnica que permite identificar el comportamiento de un fenómeno mediante un conjunto de datos, proporcionando información para la toma de decisiones. Aunque estos datos sean publicados por entidades gubernamentales se debe hacer un proceso de limpieza, ajuste y normalización de los mismos según la problemática a ser abordada, con el fin de garantizar resultados confiables y óptimos. Otras líneas de investigación que se pueden explorar a futuro son la comparación del rendimiento de AutoML en diferentes situaciones, así como el estudio de casos relacionados con seguridad, por ejemplo robos residenciales o de automotores, para verificar si su rendimiento es similar en escenarios de robos y seguridad ciudadana.

Finalmente, se busca promover el uso de datos abiertos para la investigación y como trabajo futuro generar modelos predictivos más sofisticados asociados a las problemáticas sociales, económicas y de investigación de los territorios, aprovechando el potencial y la disponibilidad de datos de gobierno. En este sentido, se destaca la importancia de utilizar datos abiertos, ya que posibilitan el desarrollo de aplicaciones que permiten contribuir a generar información, socialización y sensibilización para los ciudadanos en problemas relacionados con las ciudades inteligentes.

Referencias

Acuña, C., Garcia, S., Londoño, E., & Maestre-Gongora, G. (2020). Inteligencia de negocios para analizar hurtos en la ciudad de Medellín: un enfoque desde datos abiertos. *Congreso Internacional de La Sociedad de Doctores e Investigadores de Colombia, 3*.

Díaz, J. (2012). Introducción al business intelligence. Editorial UOC.

López, N., & Mahecha, J. (2017). *Prototipo de software para la evaluación de la calidad de datos abiertos*. Universidad Católica de Colombia.

Maestre-Gongora, G., & Nieto-Bernal, W. (2019). Conceptual model of information technology management for smart cities: Smarticity. *Journal of Global Information Management, 27* (2). https://doi.org/10.4018/JGIM.2019040109

Maestre-Gongora, G., Rangel-Carrillo, A., & Osorio-Sanabria, M. (2021). The value of open data government: a quality assessment approach. *Revista de Investigación, Desarrollo e Innovación, 11* (3), 507–518. https://doi.org/https://doi.org/10.19053/20278306.v11.n3.2021.13348

McClendon, L., & Meghanathan, N. (2015). Using Machine Learning Algorithms to Analyze Crime Data. *Machine Learning and Applications: An International Journal, 2* (1), 1–12. https://doi.org/10.5121/mlaij.2015.2101

Nguyen, T. T., Hatua, A., & Sung, A. H. (2017). Building a Learning Machine Classifier with

Inadequate Data for Crime Prediction. *Journal of Advances in Information Technology,* 141–147. https://doi.org/10.12720/jait.8.2.141-147

Pérez-Rave, J., Correa-Morales, J. C., & González-Echavarría, F. (2019). Metodología para explorar datos abiertos de accidentalidad vial usando Ciencia de Datos: Caso Medellín. *Ingeniare, Revista Chilena de Ingeniería, 27* (3), 495–509. https://doi.org/10.4067/s0718-33052019000300495

Pumares-Romero, A. G. (2019). Minería de datos en el análisis de causas de accidentes de tránsito en el Ecuador. Universidad Israel. http://repositorio.uisrael.edu.ec/handle/47000/2299

Rosado, A. A., & Rico, D. W. (2010). Inteligencia de negocios: Estado del arte. *Scientia et Technica, 1* (44), 321–326. https://doi.org/https://doi.org/10.22517/23447214.1803

Telugu-Maddileti, V., Sai, M., Sai-Sashank, K. V., & Shriphad-Rao, G. (2020). Crime Data Analysis Using Machine Learning Models. *International Journal of Advanced Science and Technology, 29* (9), 3260–3268. http://sersc.org/journals/index.php/IJAST/article/view/15887

Treviño, R., Rivera, F., & Garza, J. (2020). La analítica de datos como ventaja competitiva en las organizaciones. *VinculaTégica, 6* (2), 1063–1074. http://www.web.facpya.uanl.mx/vinculategica/Vinculategica6_2/5_Treviño_Rivera_Garza.pdf

Wieczorkowski, J. (2019). Open data as a source of product and organizational innovations. *Proceedings of the European Conference on Innovation and Entrepreneurship, ECIE,* 2, 1119–1128. https://doi.org/10.34190/ECIE.19.190