Crimen y Factores Económicos en Medellín: Un Estudio de Predicción con Machine Learning

Autor

María Camila Ardila Chávarro

Trabajo presentado como requisito para optar por el título de

Magíster en Economía

Tutor

Andrés Felipe García-Suaza

Facultad de Economía

Maestría en Economía

Universidad del Rosario

Bogotá D.C, Colombia

Diciembre 2023

Crimen y Factores Económicos en Medellín: Un Estudio de Predicción con Machine Learning

María Camila Ardila Chávarro

5 de diciembre de 2023

Resumen

La actividad criminal afecta negativamente la calidad de vida y el progreso económico de las personas. Dado el avance en la investigación económica, que aprovecha el aprendizaje automático para detectar patrones y analizar tendencias en campos específicos, estas técnicas se están empleando en diversos contextos, incluyendo la prevención del crimen. El objetivo de este trabajo es estudiar los patrones espaciales de delitos a través de la implementación de técnicas de machine learning, para predecir la probabilidad de ocurrencia de diversos tipos de crímenes a nivel anual con diferencias espaciales en Medellín, Colombia, a partir de datos históricos y sociodemográficos. Para llevar a cabo este objetivo, se utilizaron los modelos de Mínimos Cuadrados Ordinarios, Random Forest y Extreme Gradient Boosting, los cuales obtuvieron niveles aceptables de rendimiento, dada su alta precisión. Un resultado relevante es que las variables socioeconómicas relacionadas con la proporción de hombres, las personas entre 16 y 30 años de edad, proporción de personas desempleadas, personas que pertenecen al SISBEN, proporción de personas con pobreza multidimensional, proporción de personas con déficit cuantitativo de vivienda y que hacen parte del estrato socioeconómico 1 o 2, tanto a nivel de barrio como de grilla tuvieron un alto poder predictivo. Para el propósito de esta investigación, este se empleará en la toma de decisiones y la formulación de políticas públicas destinadas a la reducción de delitos.

Clasificación JEL: C38, C5, H70, K42.
Palabras clave: Machine learning, patrones de delincuencia, modelos de clasificación, predicción del crimen, análisis criminal, políticas públicas, variables sociodemográficas.

*Este artículo se presenta como tesis de Maestría en Economía de la autora. Agradezco a mi asesor Andrés F. García por su orientación y apoyo durante el desarrollo de esta investigación. También agradezco especialmente a mi querida hermana María del mar y Nicolás por su apoyo incondicional y motivación en este proceso, a Dios y a mis padres por concederme este gran sueño.

correo: mariacam.ardila@urosario.edu.co

1. Introducción

La delincuencia es un problema que afecta a todos los países del mundo. Ocupa un lugar destacado en la agenda política de todos los gobiernos debido a que es una preocupación social y tiene un impacto en la economía, la calidad de vida, la seguridad social y en el desarrollo regional (Detotto y Otranto, 2010; Liang et al., 2022; Sun et al., 2023). Afrontarla implica diseñar e implementar estrategias a corto y largo plazo (Bogomolov et al., 2014). Una forma eficiente de reducir la criminalidad es lograr su prevención, lo cual implica prever las acciones criminales antes de su ocurrencia y así generar mecanismos eficientes para las entidades gubernamentales con el fin de asignar recursos pertinentes y rápidos conducentes a la mitigación y reducción de las actividades delictivas.

Los investigadores se esfuerzan por comprender los comportamientos delictivos y sus patrones, pero la gestión de los datos relacionados con el crimen se vuelve compleja debido al rápido crecimiento de la información. La incertidumbre sobre cómo analizar estos datos debido a su falta de consistencia y exhaustividad motivan a los investigadores a indagar y desarrollar métodos más efectivos para abordar este problema persistente (Kshatri y Narain, 2020; Zaidi et al., 2020). La predicción del crimen se ha convertido en una herramienta esencial para la formulación y aplicación de políticas públicas efectivas. Su utilidad radica en la asignación eficiente de recursos, lo que permite dirigir los esfuerzos de prevención y aplicación de la ley hacia áreas y momentos críticos. Además, facilita la prevención temprana al identificar zonas o poblaciones en riesgo antes de que ocurran delitos graves. Este enfoque basado en datos respalda el diseño de políticas más efectivas y la evaluación de estrategias existentes. En última instancia, las políticas públicas fundamentadas en predicciones precisas contribuyen a reducir la delincuencia, mejorar la seguridad ciudadana y el bienestar general, y fomentar la transparencia y la confianza entre los ciudadanos y las autoridades. Estas ventajas hacen de la predicción del crimen una herramienta valiosa en la promoción de comunidades más seguras y prosperas.

De acuerdo con el Índice Global de Crimen Organizado 2023, publicado por Global Initiative, hubo un incremento de la población mundial que vive en países con un alto nivel de criminalidad, pasando del 79 % en 2021 al 83 % en 2023. En Colombia específicamente, la tasa de criminalidad pasó de 7.66 a 7.75, convirtiéndose en el país con mayor índice de criminalidad en América Latina[1]. La alta tasa de delitos y los incidentes violentos de este país tienen una larga historia y a su vez este problema ha sido uno de los principales desafíos de gobernabilidad a nivel nacional y local. Según Mejía et al. (2014) la tasa promedio de homicidios en las principales ciudades del país (Barranquilla,

[1] Global Organized Crime Index 2023, actualizado hasta octubre del 2023, [en línea] Disponible: https://ocindex.net/report/2023/0-3-contents.html

Bogotá, Cali y Medellín) ha aumentado en los últimos cinco años, pasando de ser de 30.6 % en 2008 a 40.1 % en 2011 y 35.3 % en 2013. Sánchez-Torres y Núñez-Méndez (2001) encontraron que la tasa de homicidios en Colombia aumentó considerablemente desde un 16 % entre 1970 y 1974 hasta alcanzar un pico del 89 % en el año de 1991. Datos más recientes analizados por Duarte-Velásquez y Cadavid-Carmona (2020) registraron 584.216 delitos reportados para 2019 en Colombia. Teniendo en cuenta lo anterior, el crimen se ha convertido en una de las principales preocupaciones de los colombianos.

Esta investigación se enfoca en Medellín, una ciudad colombiana que tiene una alta incidencia de crimen. En los años 70 y principios de los 80, fue conocida por ser el epicentro de un poderoso cartel de droga que generó un nivel de violencia sin precedentes (Collazos et al., 2021). Entre tanto, Ramírez (2008) y Martin (2012) en sus investigaciones, señalaron que en los años 1990 a 1993 la tasa de homicidios en la ciudad de Medellín superó los 350 casos por cada 100.000 habitantes[2]. Según Blattman et al. (2020) aproximadamente dos tercios de los barrios de Medellín están bajo el dominio del crimen organizado. Además, Sánchez-Jabba (2013) concluyeron que, a finales del siglo XX, la ciudad experimentó una creciente violencia, exacerbada por la crisis económica de ese período, lo que la sumió en una crisis urbana en múltiples dimensiones políticas, económicas y sociales.

Adicionalmente, Medellín es la segunda ciudad más grande de Colombia con una población de alrededor de 2.5 millones de personas, por lo cual ofrece un entorno propicio para estudiar la difusión espacial del crimen. Durante el período analizado, la ciudad fue una de las más violentas del mundo, lo que permitió observar claramente la disparidad en las tasas de crimen y la segregación de oportunidades económicas, características comunes en grandes centros urbanos (Khanna et al., 2022).

En el contexto previamente expuesto, el objetivo central de esta investigación consiste en estudiar los patrones espaciales de delitos a través de la implementación de las técnicas de Random Forest, Extreme Gradient Boosting y un modelo de Mínimos Cuadrados Ordinarios para predecir la probabilidad de ocurrencia de diversos tipos de delitos, tales como hurto a personas, hurto a residencias, hurto a vehículos, homicidios y extorsiones, haciendo uso de datos reales de eventos delictivos, así como registros históricos y variables sociodemográficas. Estas predicciones se realizan a nivel anual con diferencias espaciales en Medellín, Colombia. Según estudios relevantes, el uso de estos métodos de aprendizaje en conjunto tiene un mejor desempeño que los que no lo son (AL Mansour y Lundy, 2019). Además tanto el XGBoost como el RF no se ven afectados

[2]Registrando una de las tasas de homicidio más significativas a nivel global en ese momento, con un total de 6.809 muertes ese año.

por los problemas de multicolinealidad de las variables, ya que, al ser modelos no paramétricos, no imponen suposiciones acerca de la naturaleza de las relaciones entre las variables. Esta característica les otorga una flexibilidad significativa para detectar patrones complejos (Zhang et al., 2022; Deng et al., 2023). Por otro lado, Maloof (2003) resaltó que RF es eficaz en la predicción de crímenes, especialmente cuando se enfrenta a conjuntos de datos desequilibrados, lo cual es común en la predicción de crimen donde los eventos delictivos son poco frecuentes en comparación con los no delictivos. Es importante destacar que, para los propósitos de este estudio, resulta crucial diferenciar entre los distintos tipos de delitos, ya que todos ellos muestran diversos patrones de comportamiento en relación al tiempo y al espacio, así como una amplia variación en su frecuencia. Esta investigación se destaca por dos contribuciones fundamentales: en primer lugar, se tiene la incorporación de variables adicionales que resultan relevantes para mejorar la precisión de la predicción del delito, tales como: el índice de pobreza multidimensional, la proporción de personas que pertenecen al SISBEN, proporción de personas desempleadas, proporción de personas que no saben leer ni escribir, proporción de personas que hacen parte del estrato socioeconómico 1 o 2, y en segundo lugar, para la predicción se abarcan datos a un nivel de desagregación de barrio y grilla, debido a que esto permitirá identificar patrones y concentraciones de delitos en áreas específicas, lo que resulta fundamental para comprender la dinámica de la criminalidad a nivel local.

Dado el aumento de la delincuencia y la complejidad de los datos criminales, el uso de técnicas de aprendizaje automático se ha convertido en una estrategia destacada para comprender y anticipar patrones delictivos en diversas escalas urbanas (Varian, 2014; Goin et al., 2018; Lima y Delen, 2020). Estas técnicas tienen la capacidad de analizar datos complejos y discernir relaciones no lineales, lo que ha despertado un gran interés en la seguridad ciudadana (Andini et al., 2018; Kounadi et al., 2020). Estas se centran en optimizar la toma de decisiones y asignar recursos eficientemente en la prevención del crimen y en abordar la modelación y anticipación de actividades delictivas en áreas geográficas específicas, lo que contribuye a un enfoque más preciso y estratégico en la protección de entornos urbanos.

La literatura ha identificado diversas variables relevantes para la predicción del crimen. Entre estas se han usado variables socio-económicas como: desigualdad de ingresos, el desempleo total, el porcentaje de mujeres, hombres y jóvenes desempleados, la población total, el número de extranjeros por cada 1.000 habitantes, el tamaño medio de los hogares, la superficie media de edificios habitados (m2), el porcentaje de adultos con estudios secundarios, las características del empleo, los patrones de viaje, el nivel educativo, los indicadores de dificultades financieras, el estado civil y las características de la vivienda (Gerber, 2014; Alves et al., 2018; De Blasio et al., 2022). Adicionalmente se han tenido en cuenta características del entorno, entre ellos: estrato, clima, estaciones

de policía, escuelas, centros comerciales, iglesias, licorerías, bares, analfabetismo, la ubicación, la demografía (Goin et al., 2018; Ingilevich y Ivanov, 2018; Reier-Forradellas et al., 2020; Kajita y Kajita, 2020; Stalidis et al., 2021; Liang et al., 2022).

En este estudio, se incorporaron variables ya utilizadas previamente, como la media de desempleo, estratificación socio-económica de los barrios, la ubicación del delito (latitud y longitud), género y diferentes grupos de edad, las cuales al igual que en el estado del arte, demuestran tener una relación significativa con el crimen en esta investigación. Como contribución adicional, se introdujeron las siguientes variables: percepción del crimen, densidad criminal, proporción de personas analfabetas e indicadores de desigualdad: la pobreza multidimensional, proporción de personas que pertenecen al SISBEN, el déficit cuantitativo y cualitativo de vivienda. La inclusión de estas reviste una importancia significativa para la predicción del delito, ya que estas variables aportan una comprensión más completa de las dinámicas de la delincuencia, teniendo en cuenta que no solo se evalúa la incidencia de crímenes, sino también la vulnerabilidad de las comunidades. Esto tiene un profundo impacto en la toma de decisiones y en la formulación de políticas públicas, ya que permite abordar de manera más efectiva los factores subyacentes que contribuyen a la delincuencia. Así, estas variables no solo son útiles para predecir el crimen, sino que también orientan estrategias específicas de intervención y prevención, promoviendo la seguridad y la calidad de vida en las áreas urbanas. Por ejemplo, la pobreza multidimensional, al identificar carencias en aspectos como educación, vivienda y salud, señala la necesidad de programas de inclusión social que mejoren estas condiciones. El déficit cuantitativo y cualitativo; destacan la importancia de políticas de vivienda y de atención médica que mejoren la calidad de vida y reduzcan la criminalidad. La estratificación socioeconómica permite focalizar recursos y políticas de seguridad en las zonas más afectadas. En conjunto, estas variables orientan acciones concretas que abordan las raíces de la criminalidad y mejoran el bienestar de la población en áreas urbanas.

Adicionalmente, en esta investigación se presentan otras contribuciones entre las cuales se destaca el nivel de desagregación de barrios, reconociendo la necesidad de capturar la variación oculta en áreas geográficas pequeñas, donde la criminalidad puede variar significativamente (Mustard-David, 2010). A diferencia de la mayoría de la literatura nacional sobre predicción del delito, que ha optado por enfoques a nivel municipal o áreas locales, esta investigación se ha centrado en una desagregación detallada. Por ejemplo, investigaciones previas, como las realizadas por Alegría et al. (2020); Mojica-Muñoz (2021); Ferro-Briceño et al. (2021); Bazzi et al. (2022); Rojas-Guerrero et al. (2022), han abordado la predicción del delito a nivel municipal en Colombia, mientras que Ordoñez-Eraso et al. (2020) la implementaron en ciudades colombianas, y Sánchez-Torres y Núñez-Méndez (2001); García et al. (2012); Khanna et al. (2022) se centraron en Medellín, aunque sin

realizar una desagregación a nivel de barrios o comunas. Asimismo, Gelvez-Ferreira et al. (2022) llevaron a cabo su investigación en la ciudad de Bucaramanga a nivel de manzana. Por otro lado, se destaca el uso de grillas como unidad de observación, lo cual representa una decisión desafiante dada la naturaleza cambiante de los delitos (Wang et al., 2020). Este nivel de desagregación ofrece una comprensión más profunda y precisa de la dinámica del crimen en Medellín y tiene implicaciones políticas importantes al mejorar la asignación de recursos, diseñar estrategias de prevención efectivas y promover políticas de seguridad pública más informadas y adaptadas a las necesidades de cada comunidad (Rosser et al., 2017; Lin et al., 2018).

Este documento consta de 5 secciones. En la primera, se presenta una introducción al tema junto con su contexto. En la segunda, se tiene la revisión de literatura. La tercera, abarca las fuentes, la descripción detallada de los datos, la muestra utilizada, la descripción de la metodología y las técnicas de machine learning usadas para la predicción de la delincuencia en la ciudad de Medellín. En la cuarta, se incluyen los resultados y finalmente, la última sección contiene las reflexiones y trabajos futuros del estudio.

2. Revisión de literatura

En esta sección se exponen los antecedentes relacionados con el objeto del estudio actual: la predicción de diversos tipos de delitos por barrio y grilla en la ciudad de Medellín, por medio de técnicas de aprendizaje automático. Por tal motivo, se realiza una compilación de investigaciones a nivel internacional y nacional, de tal manera que se pueda construir el conocimiento sobre el asunto, destacando los aportes y vacíos existentes en torno a este.

La conexión entre el estado de la economía y la incidencia del crimen ha sido objeto de estudio en la investigación criminológica durante un largo período. Este interés se remonta a investigadores como Shaw y McKay (1931) y Becker (1968), pionero en la aplicación de modelos económicos de toma de decisiones racionales al estudio del crimen, derivó la función de oferta de delitos. Esta función se basa en relacionar el número de delitos cometidos por una persona con su probabilidad de ser condenada, las consecuencias legales en caso de condena, y otras variables, como sus ingresos legales y ilegales. La propuesta de Merton (1938) complementa este enfoque al señalar que las tasas de criminalidad tienden a ser mayores en sociedades caracterizadas por una mayor desigualdad de oportunidades. Adicionalmente, Cornish y Clarke (2016) mencionan en su investigación la Teoría de la Elección Racional que es similar a la teoría de Becker en algunos aspectos, esta teoría se centra en las decisiones individuales de cometer delitos. Considera que los individuos toman decisiones

racionales basadas en la relación entre los beneficios y los costos de cometer un delito, como la probabilidad de ser capturado y castigado.

Por otro lado, la teoría de la elección racional, desarrollada por Cornish y Clarke (1989), ofrece una perspectiva fundamental sobre cómo las personas evalúan y toman decisiones en relación con el comportamiento delictivo. En el centro de esta teoría está la noción de que los individuos que cometen delitos no son actores irracionales, sino que son tomadores de decisiones racionales que sopesan cuidadosamente sus opciones antes de involucrarse en actividades ilegales. Esta perspectiva psicológica desafía la noción tradicional de que los infractores son inherentemente distintos de las personas que obedecen la ley y sugiere que el comportamiento criminal puede entenderse como parte de un proceso racional de toma de decisiones. Entender cómo los factores económicos influyen en estas decisiones criminales se convierte en un paso clave para anticipar dónde es más probable que ocurra un delito en Medellín. Esta capacidad predictiva podría ser de gran utilidad para los responsables de la formulación de políticas, ya que les proporcionaría información crucial para dirigir estrategias preventivas hacia áreas específicas y situaciones que presenten un mayor riesgo. La identificación proactiva de estas zonas críticas permitiría la implementación de políticas y medidas preventivas más efectivas y dirigidas, trabajando hacia la reducción de la incidencia delictiva en la comunidad.

Por otro lado, en el campo de la criminalística, la minería de datos ha representado un impulso significativo, ya que ha permitido extraer conocimiento, responder preguntas de investigación, desarrollar sistemas de apoyo e identificar características del crimen, por lo que es una herramienta poderosa que permite a las fuerzas del orden descubrir patrones y predecir futuros delitos. Siguiendo esa idea, se encuentran Falade et al. (2019), quienes realizaron una revisión en la que enfatizan la importancia de la predicción del crimen mediante la minería de datos y otras técnicas, concluyeron que es un tema destacado de investigación debido a la influencia del crimen en el desarrollo económico de una nación. De manera similar, diversos investigadores han concluido a partir de sus estudios que la minería de datos es una herramienta en la lucha contra el terrorismo (Thuraisingham, 2004; Okonkwo y Enem, 2011).

En un estudio de Mittal et al. (2019), se utilizaron cuatro algoritmos (Decision Trees, Random Forest, Regresión Lineal y Neural Networks) para predecir la tasa de criminalidad en India. Los resultados revelaron que un alto Índice de Precios al Consumidor (IPC), Producto Interno Bruto (PIB) y el desempleo redujeron la motivación para cometer delitos por necesidad económica. Además, se observó que el modelo de regresión lineal superó a los demás algoritmos, mostrando la mayor precisión en la predicción del crimen. De Blasio et al. (2022) en su investigación a nivel de municipio en Italia, utilizando árboles de clasificación lograron prever con precisión un porcentaje

superior al 70 % de los municipios que experimentarán un aumento en los delitos de corrupción, durante el periodo 2012-2014. Además, concluyeron que características específicas de los mercados laborales y del sector inmobiliario a nivel local, junto con el historial previo de delitos de cuello blanco en la región, desempeñan un papel significativo en la predicción de la corrupción.

Stalidis et al. (2021) y Liang et al. (2022) en sus investigaciones descubrieron que al agregar al modelo variables como latitud, longitud, la categoría delictiva, la semana anterior, la hora del día, año, mes y eventos cercanos tanto en el espacio como en el tiempo obtuvieron una predicción con una precisión más alta al clasificar a nivel de localidad, diversos tipos de crimen como robo, venta de narcóticos, asalto, allanamiento de morada, entre otros; por medio de la implementación de Convolutional Neural Networks (CNN), Decision Trees (C4.5), Naive Bayes, Logit Boost y Random Forest. Concluyendo, que al incorporar tanto la dependencia categórica espacio-temporal como los factores externos en el proceso de predicción son beneficiosos para observar los patrones de crimen. Por otro lado, en el estudio realizado por Goin et al. (2018), emplearon Random Forest y regresión de LASSO para predecir la violencia armada en Estados Unidos. En este caso, identificaron que 18 variables predictoras explicaron el 77.8 % de la variabilidad en las tasas de violencia armada en entornos urbanos. Entre estas variables incluyeron el nivel educativo de los veteranos, factores geográficos, características laborales familiares, variables de estado civil, modos de transporte, ingresos y situación económica, todos con un alto poder predictivo en la incidencia de violencia armada.

Entretanto Ingilevich y Ivanov (2018) y Reier-Forradellas et al. (2020) usaron factores como la cantidad de personas en la ciudad, estaciones de policía, escuelas, centros comerciales, iglesias, licorerías y bares para predecir el crimen, mediante regresión lineal, logística, Gradient Boosting y un modelo Sample, Explore, Modify, Model and Assess (SEMMA). Encontrando que la elección de estos lugares desempeñaron un papel significativo en la predicción del crimen, resaltando su relevancia en el análisis de seguridad y delincuencia. Mientras que en Alves et al. (2018) implementaron un Random Forest Regressor para anticipar la incidencia del delito y medir el impacto de los factores urbanos en los casos de homicidios. Utilizaron como predictores del crimen las variables de desempleo, analfabetismo y la población masculina, encontrando que estos tres están altamente correlacionados con este tipo de delito. Concluyendo que estas variables se relacionan con la dinámica social y la presencia de factores de riesgo asociados con el crimen.

En la investigación de Lin et al. (2018) se destaca su enfoque en el uso de 84 tipos de ubicaciones geográficas para establecer características espaciotemporales en un modelo basado en cuadrículas, a partir del cual emplearon algoritmos de aprendizaje automático (ML) para analizar patrones y predecir el crimen del mes siguiente en cada cuadrícula de la ciudad de Taiwán. Entre los métodos de

ML que usaron, Deep Neural Network (DNN) demostró ser el modelo más efectivo. De igual manera, Wheeler y Steenbeek (2021) usaron cuadrículas de 200 por 200 metros para generar pronósticos de robos a largo plazo por medio de un Random Forest en Dallas, concluyendo que el estudio permite establecer relaciones espacialmente heterogéneas entre los delitos y las variables sociodemográficas, al mismo tiempo detectaron que las variables más relevantes para la predicción del crimen incluyen los lugares como apartamentos, restaurantes y grandes comercios minoristas, junto con la densidad de población.

En adición a la literatura revisada, es esencial enfocarse en investigaciones a nivel nacional. En Colombia, se han realizado diferentes investigaciones y estudios de predicción de actos de corrupción, masacres, accidentes de tránsito, conflictos (Mojica-Muñoz, 2021; Ferro-Briceño et al., 2021; Gallego et al., 2022; Rojas-Guerrero et al., 2022) y víctimas de secuestro (Alegría et al., 2020) mediante la utilización de machine learning. Por ejemplo, en un estudio de Ordoñez-Eraso et al. (2020), pronosticaron las tendencias de homicidios violentos (VH) para los próximos 5 años (2015-2020) utilizando un modelo de Random Forest Regressor. Los resultados del modelo sugieren que el número de homicidios mostrará una tendencia a la baja en 2025, lo que representará una reducción del 60 % en los delitos violentos en Colombia. Mientras que en Mojica-Muñoz (2021) y Gallego et al. (2022) utilizaron algoritmos como XGBoost, Random Forest, regresión lineal, Gradient Boosting Machine, LASSO, Neural Networks y Super Learner, para predecir la corrupción en Colombia. En el primer estudio, se clasificaron en 5 niveles el riesgo de corrupción en la Administración Pública Municipal, encontrando que ciertas variables como procesos penales retrasados, población, establecimientos industriales, índice de conflicto y educación tenían un alto poder predictivo. Estas variables permitieron predecir con un 85 % de precisión el riesgo relativo de corrupción. En el segundo estudio, se identificaron las zonas de mayor riesgo de corrupción, destacando la importancia de las variables relacionadas con el sector financiero y el sector público en la predicción, mientras que las variables vinculadas a conflictos armados, actividades ilícitas y la dependencia de recursos naturales tuvieron un impacto predictivo menor. Concluyeron que la predicción de la corrupción puede ayudar a fortalecer la transparencia y la rendición de cuentas, lo que contribuirá a una gestión pública más eficiente y a la mejora de la confianza de los ciudadanos en las instituciones gubernamentales.

Igualmente, se destaca el estudio de Bazzi et al. (2022), quienes investigaron la predicción de conflictos en Colombia e Indonesia a través LASSO, Random Forest, Adaptive Boosting y Neural Network. Sus hallazgos revelaron que la violencia no mostraba un patrón autorregresivo. En el caso de Colombia, identificaron que el indicador de conflicto se presentaba en un tercio de los municipios cada año, con cinco o más incidentes ocurriendo alrededor del 8 % del tiempo.

En Indonesia, observaron un aumento en la desviación estándar de aproximadamente 4.7 actos de violencia anuales, con variaciones año tras año y entre subdistritos de alrededor del 3.3 %. Del mismo modo, Rojas-Guerrero et al. (2022) emplearon inteligencia artificial para prever los municipios en Colombia que podrían ser susceptibles a masacres entre 2010 y 2020. Descubrieron que el uso de Random Forest logró una precisión media del 76 % al 82 %, superando al modelo binomial que obtuvo una precisión del 27 % al 29 %. También destacaron que el riesgo de masacres aumenta notablemente cuando la tasa de homicidios y los cultivos ilícitos empiezan a crecer.

Otra investigación relacionada es la de Gelvez-Ferreira et al. (2022) en Bucaramanga entre 2016 y 2019, en la cual utilizaron el procesamiento de señales para grafos por semanas, junto con modelos como k-Nearest Neighbors y Support Vector Machine, para predecir delitos. Hallaron que el modelo más efectivo fue KNN. Además, concluyeron que el modelo de predicción de delitos resulta valioso para elaborar estrategias de prevención en grandes ciudades, aunque su utilidad podría ser limitada en ciudades de tamaño mediano con acceso restringido a datos.

Teniendo en cuenta que el presente estudio se realiza para la ciudad de Medellín, a continuación se hace énfasis en los estudios realizados en esta. En Muñoz et al. (2021) analizaron datos de delitos en Medellín, estaciones meteorológicas y tasas de desempleo entre 2015 y 2019 mediante técnicas de machine learning. Descubrieron que eventos como el pago de salarios y festivales tuvieron un impacto significativo en la tasa de delincuencia, y encontraron una correlación del 35 % entre el desempleo y las fluctuaciones en la delincuencia. Contrariamente a algunas teorías, no encontraron una conexión directa entre la temperatura y la incidencia del crimen. Complementando, Khanna et al. (2022) examinaron cómo la participación en actividades delictivas se relacionan con los cambios en los costos de transporte y los incentivos para trabajar en empleos formales en vecindarios. Utilizaron un enfoque de equilibrio general espacial y analizaron datos geocodificados de arrestos en Medellín, junto con registros individuales de empleo y direcciones de encuestas a hogares. Sus resultados mostraron que la reducción de la delincuencia fue más notable en áreas con altas tasas iniciales de delincuencia y falta de oportunidades económicas legítimas. Sin embargo, algunas áreas de baja delincuencia cerca de nuevas estaciones de transporte experimentaron un ligero aumento en la delincuencia.

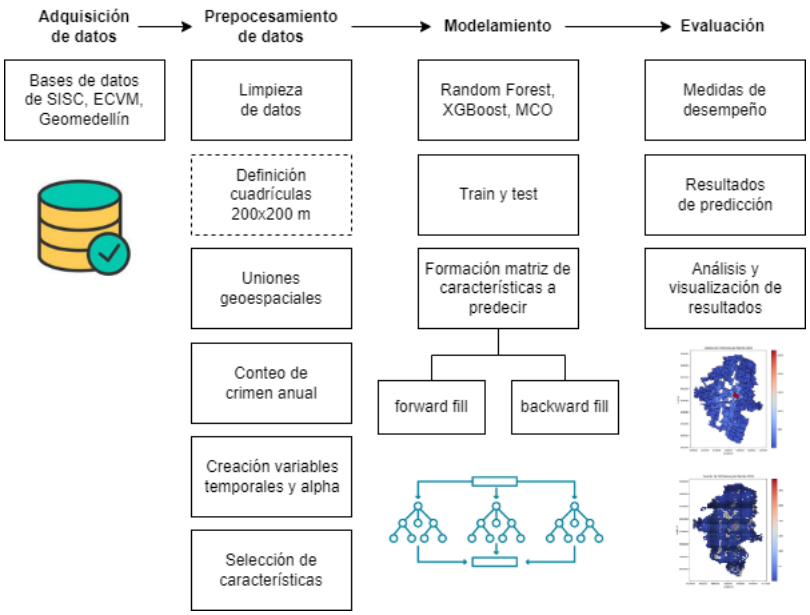
La revisión de literatura presentada resalta el interés frente al tema de la predicción criminal. A su vez, posibilita reunir una serie de información relevante que, además de constituir la sustentación teórica de este estudio, aporta un sistema coordinado y coherente de propuestas en el campo de la criminalística, la minería de datos y el aprendizaje automático, que permiten desarrollar sistemas de apoyo e identificar características del crimen.

3. Datos y Metodología

En la actualidad, el análisis y predicción del crimen ha emergido como un desafío crucial en la búsqueda de estrategias eficaces para garantizar la seguridad y el bienestar de las comunidades urbanas. En este contexto, el presente estudio abarca la totalidad del área urbana de Medellín, Colombia, conocida por su historia de transformación y resiliencia, con el propósito de desarrollar un enfoque innovador para la predicción del crimen utilizando machine learning. El objetivo central de esta investigación es estudiar los patrones espaciales de delitos a través de la implementación de varios modelos de machine learning para predecir la probabilidad de ocurrencia de diversos tipos de delitos en un barrio y grilla en la ciudad de Medellín anualmente, haciendo uso de datos reales de eventos delictivos, así como registros históricos y variables sociodemográficas. De tal manera que se brinde información valiosa a las autoridades encargadas de la seguridad ciudadana.

En resumen, la presente sección delineará de manera detallada los datos y la ruta a seguir para la implementación de los modelos de machine learning que se usarán en el presente trabajo. La metodología contiene una serie de pasos rigurosos y sistemáticos, que incluyen la preparación y transformación de los datos en un conjunto estructurado y apto para la construcción y entrenamiento de los modelos, de tal manera que finalmente se obtenga un análisis predictivo de patrones delictivos en la ciudad de Medellín. La metodología se ilustra en la Figura 1, otorgando una visión clara y detallada del flujo de trabajo desarrollado.

Figura 1: Metodología propuesta



Fuente: Elaboración propia.

3.1. Datos

En esta sección, se presentarán los datos y fuentes empleadas para llevar a cabo la predicción del crimen en la ciudad de Medellín.

3.1.1. Sistema de Información para la Seguridad y la Convivencia de Medellín (SISC)

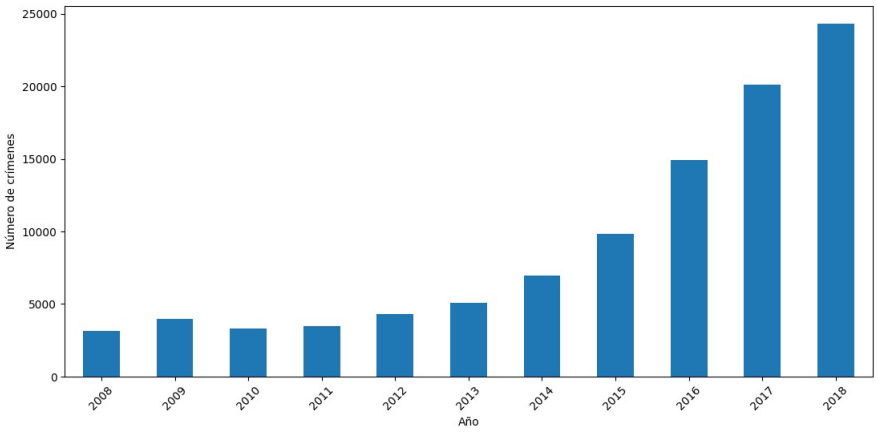
Para esta investigación, se usaron datos proporcionados por el Sistema de Información para la Seguridad y la Convivencia de Medellín (SISC), que tiene como finalidad monitorear y analizar las dinámicas de seguridad y convivencia que se desarrollan en la ciudad de Medellín[3]. Esta base recopila datos georreferenciados de diversos tipos de delitos denunciados por los habitantes de Medellín; los cuales contienen además, sexo, edad, estado civil, modalidad de robo (tipo de arma o medio) y ocupación de las víctimas. La temporalidad va desde el 1 de enero del 2008 hasta el 31 de diciembre 2018, con 99.410 observaciones sobre reportes de crímenes diarios. Para su análisis, las fechas de las denuncias se transformaron en años, y se registró el número de delitos ocurridos, además fueron georreferenciados y analizados mediante métodos de análisis espacial, lo que permitió identificar patrones espaciales y puntos críticos relevantes, asociándolos finalmente con los respectivos barrios para cada año[4]. De esta base de datos se seleccionaron las siguientes variables: latitud, longitud, tipo de crimen (conducta) y código de barrio.

Los datos del 2008 a 2018 muestran un aumento anual en el número de reportes de delitos en la ciudad de Medellín, como se puede observar en la Figura 2, en el año del 2008 se tenían 3.138 reportes de crímenes en la ciudad, llegando a 24.327 denuncias delictivas en el 2018. Este incremento en las denuncias sugiere que la comunidad está cada vez más afectada por la delincuencia y siente la necesidad de reportar los incidentes. Estos datos son una clara señal de que la seguridad pública es una preocupación apremiante en la ciudad y requiere una atención inmediata por parte de las autoridades y los encargados de la formulación de políticas. La magnitud de este aumento refuerza la importancia de investigar y comprender las causas subyacentes de la delincuencia y desarrollar estrategias efectivas de prevención y control para abordar este desafío creciente en Medellín.

[3] Alcaldía de Medellín (s.f.). Sistema de Información para la Seguridad y Convivencia – SISC https://www.medellin.gov.co/Sistema-de-informacion-para-la-Seguridad-y-Convivencia-SISC

[4] Tomado de la página del Geoportal del Departamento Administrativo Nacional de Estadística de Colombia (DANE): https://www.dane.gov.co/files/geoportal-provisional/index.html

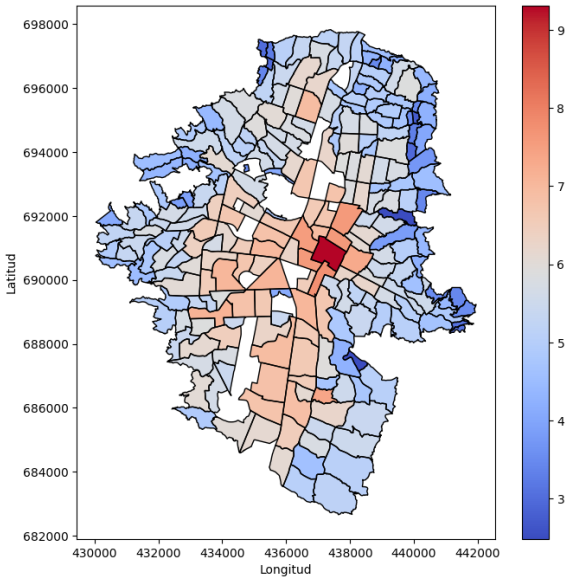
Figura 2: Número de delitos anuales en la ciudad de Medellín, Colombia (2008-2018)



Fuente: Cálculo propio usando la base de datos del Sistema de Información para la Seguridad y la Convivencia de Medellín (SISC).

En la Figura 3 se presenta un mapa de calor en escala logarítmica que ilustra la cantidad de reportes de crímenes hechos por los ciudadanos de Medellín durante el período de estudio. En este mapa, las áreas rojas indican una mayor cantidad de reportes de crímenes en los respectivos barrios. Se observa que existe una concentración significativa de crímenes en el centro de la ciudad, en particular en el barrio La Candelaria, donde se registraron un total de 11.112 reportes de crímenes entre 2008 y 2018. Le siguen los barrios Colón (2.233 reportes), Guayaquil (2.010 reportes) y Villa Nueva (2.003 reportes), mientras que el barrio con la menor cantidad de reportes de crímenes entre los 249 analizados es Los Ángeles, con un total de 101 reportes. Estos hallazgos demuestran que la incidencia delictiva varía notablemente de un barrio a otro en Medellín.

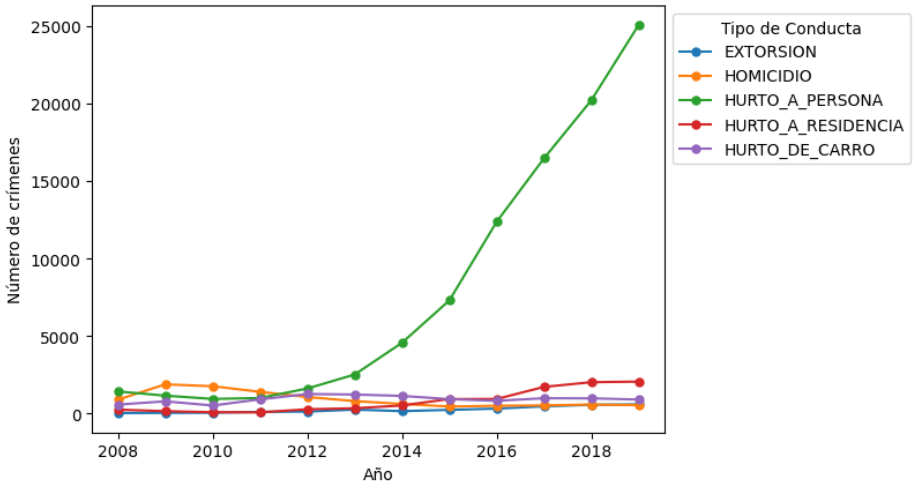
Figura 3: Representación en escala logarítmica del número de reportes de delitos en los barrios de la ciudad de Medellín, Colombia, durante el período comprendido entre 2008 y 2018



Fuente: Cálculo propio usando la base de datos del Sistema de Información para la Seguridad y la Convivencia de Medellín (SISC).

De la Figura 4 se observa el comportamiento de los cinco tipos de crímenes más denunciados en la ciudad de Medellín (hurto a personas, a residencias, a carro, extorsión y homicidios) entre los años del 2008 hasta el 2018; donde la tendencia ha sido creciente para todos estos tipos de delitos. Durante los 11 años analizados los reportes a crímenes de los delitos de hurto a residencias, extorsión y homicidios se mantienen, en cambio en el caso de hurto a personas del año 2011 en adelante, hubo un crecimiento del número de reportes de este crimen en la ciudad de Medellín, volviéndose más común, en comparación a los otros tipos de delitos.

Figura 4: Número de reportes de diferentes tipos de delito en la ciudad de Medellín, Colombia (2008-2018)



Fuente: Elaboración propia usando la base de datos del Sistema de Información para la Seguridad y la Convivencia de Medellín (SISC).

3.1.2. La Encuesta de Calidad de Vida de Medellín (ECVM)

Los datos empleados pertenecen al período desde 2008 hasta 2018 y se basaron en una muestra de hogares representativos de las 16 comunas y 5 corregimientos de Medellín. Esta muestra se obtuvo mediante un proceso de muestreo aleatorio, estratificado y conglomerado de varias etapas. Esta encuesta proporciona datos detallados a nivel de barrio, lo que permitirá un análisis más específico de los hogares. La ECVM es una herramienta diseñada para monitorear y medir la situación socioeconómica de los habitantes de Medellín. De igual manera, es un recurso estadístico que permite a las personas conocer de primera mano indicadores importantes relacionados con temas como población, vivienda, familia, educación, trabajo, salud y seguridad, entre otros[5].

Para los modelos se tuvieron en cuenta las siguientes variables proporcionadas por esta encuesta: edad media en cada uno de los barrios de Medellín, media de edad de personas entre 0 y 15 años, entre 16 a 30 años, mayores de 30 años, proporción de personas que son hombres por barrio,

[5]Alcaldía De Medellín, Banco de documentos, octubre de 2023, [en línea] Disponible: https://www.medellin.gov.co/es/centro-documental/encuesta-calidad-de-vida/

media de personas que presentan estrato 1 y 2, estrato 3 y 4, media de personas por barrio que no tienen ningún nivel de estudios, proporción de personas desempleadas, media de personas que no saben leer y/o escribir, media de personas que tienen índice de pobreza multidimensional, media de personas que tienen déficit cuantitativo de vivienda, déficit cualitativo de vivienda, media de personas que manifestaron que habían sido atracadas en el barrio donde vivían en el último año y media de personas que dijeron que se sentían inseguras en el barrio donde actualmente vivían.

3.1.3. Geomedellín

Adicionalmente, los registros en formato shapefiles (.shp) para cada barrio y comuna de la ciudad se obtuvieron a través de la página web de Geomedellín[6]. A partir de los cuales se usaron variables como: código de barrio y área en metros cuadrados de cada barrio de Medellín, junto a la geometría del límite de la parte urbana de la ciudad.

3.2. Preprocesamiento de los datos

En primer lugar, se lleva a cabo una exhaustiva recopilación y limpieza de datos relacionados con incidentes criminales en Medellín. Después de tener en una sola base de datos todas las variables que se mencionaron en la sección 3.1 (99.410 observaciones comprendidos entre 2008-01-01 al 2018-12-31), se eliminan los datos inconsistentes (NA's), incluidos campos vacíos (missing values), datos ruidosos o incompletos y datos duplicados, teniendo en cuenta que lo anterior es clave para mejorar la calidad de los datos previamente al entrenamiento y prueba del algoritmo.

Seguidamente, se realiza una intersección geoespacial para obtener los registros de crímenes correspondientes solo al casco urbano de Medellín, teniendo en cuenta los límites comunales de este encontrados en el shapefile de Geomedellín descrito en la sección 3.1.3. En este caso, ya se tiene la información a nivel de barrio, por tanto, para el caso de grillas, se definen cuadrículas de 200x200 metros[7], las cuales se generan iterando sobre la extensión del área de estudio y se filtran solo las que pertenecen a la zona urbana de Medellín, para luego realizar una unión espacial para asociar cada registro de crimen con el cuadrado de la grilla en el que se encuentra. Luego, se realiza un conteo de crímenes por barrio y grilla anualmente utilizando como variables: la fecha que hicieron la denuncia, el código de cada barrio, el identificador de grilla, y el tipo de conducta (hurto a persona, hurto a carro, hurto a residencia, extorsión y homicidio).

[6] Alcaldía De Medellín, Límite Catastral de Comunas y Corregimientos , marzo de 2021, [en línea] Disponible: https://www.medellin.gov.co/geomedellin

[7] Se definió este tamaño de celda debido a que si el tamaño de la cuadrícula es demasiado pequeño, la probabilidad de que ocurra un delito en un lugar específico es extremadamente improbable (Lin et al., 2018).

Finalmente, para la selección de variables relevantes se realiza un análisis (presentado en la sección de Resultados) que muestra qué grupos de variables son los más importantes para predecir la ocurrencia de crímenes. En el contexto de los modelos basados en árboles, la importancia de las variables se evalúa mediante la ganancia de información lograda al realizar divisiones basadas en cada variable. Este criterio de importancia se mide en una escala de 0 a 100, donde 100 es el valor que toma la variable que obtiene la mayor importancia hacia la variable predicha, y los demás valores de las variables se toman a partir de la variable que ha obtenido la mayor importancia en el modelo[8].

Además, se calcularon dos variables que representan la densidad de eventos delictivos en días α dentro de un barrio (Ecuación 1). En este caso, se consideraron por la literatura que eran importantes para mejorar la predicción de tipos de crímenes con una granularidad espaciotemporal más fina. Por consiguiente, para cada barrio NH, se realiza un análisis de la tendencia de los eventos delictivos, en el cual se utiliza una ventana móvil con una duración de α días, con valores de $\alpha = 7$ y $\alpha = 30$. Inicialmente, la ventana móvil comienza en el primer día de los datos históricos y, posteriormente, se desplaza día a día hasta llegar al punto en que su inicio es d-α. En cada punto inicial de la ventana móvil, se calcula la densidad de delitos como la proporción entre el número de delitos que ocurrieron dentro de la ventana móvil y la cantidad de días que abarca esta ventana. Además, dado que el área de cada barrio no es uniforme, se normaliza la densidad de eventos delictivos por tamaño de área. Este enfoque permite evaluar y comprender cómo la densidad de delitos varía a lo largo del tiempo en cada barrio NH, proporcionando una visión detallada de las tendencias a largo plazo en la ocurrencia de eventos delictivos en esa área específica (Tomado del artículo de Rumi et al. (2018)).

$$DA(NH, \Delta t) = \frac{\sum_{j=\mathrm{d}-\alpha}^{\mathrm{d}} Cr_j(NH)}{A(NH)} \tag{1}$$

donde $Cr_j(NH)$ es el número de delitos j que ocurren en un barrio NH durante el intervalo de tiempo Δt, en el cual se tiene $\alpha = 7$, $\alpha = 30$ días, y $A(NH)$ es el área del barrio.

Por otro lado, a partir de la variable fecha, se obtienen características como fin de semana, día de la semana y estación del año, para agregar información adicional sobre las fechas en las que ocurrieron estos crímenes, y así permitir análisis más detallados y sofisticados sobre los patrones de criminalidad.

[8]En el caso de Random Forest Classifier la métrica utilizada para medir la importancia de las características es la reducción en la impureza (Gini impurity o entropía) y la reducción en la función de pérdida (logloss para clasificación) para XGBoost.

3.3. Estadísticas descriptivas

En esta sección, se presentarán las estadísticas descriptivas de las variables tanto independientes como dependientes que forman parte de la implementación de los modelos de aprendizaje automático. Adicionalmente, se dará una visión completa de su distribución y características clave en la ciudad de Medellín durante el período 2008-2018.

Se observa que la edad promedio de la población en este lapso se sitúa en torno a los 41 años, y que el 71 % de los residentes se sienten seguros en sus barrios, en comparación con el 22 % que manifiesta sentirse inseguro[9] y el 8 % que se considera muy seguro. La mayoría de la población pertenece al grupo de mayores de 30 años, representando el 62 %, mientras que los hombres conforman el 45 % de la población total (la población mayor de 30 años puede contribuir a la prevención del delito, ya que es menos probable que participe en actividades delictivas. Sin embargo, un mayor porcentaje de hombres en un área podría estar relacionado con tasas de delincuencia más altas, ya que los hombres tienen una participación delictiva más alta). Respecto a las condiciones socioeconómicas, el 82 % enfrenta pobreza multidimensional, en contraste con el 41 % que experimenta déficit cuantitativo y el 7 % con déficit cualitativo en vivienda[10], lo cual presenta una incidencia significativamente menor. En términos de estratificación, la mayoría de los residentes se ubican en el estrato 3 o 4, caracterizados como medio y medio-bajo (64 %)[11]. Además, solo un reducido 12 % se encuentra desempleado, mientras que un considerable 44 % no está actualmente involucrado en actividades de estudio. Estos datos proporcionan una panorámica fundamental de la población de la ciudad de Medellín y sus condiciones durante el período analizado (veáse en el apéndice Tabla A.2).

Adicionalmente, en la Tabla A.3 del apéndice, se muestran las correlaciones que se han identificado entre las variables independientes y la variable dependiente, lo que permitirá determinar la fuerza de una relación lineal o no lineal entre las variables[12]. Se observa en primer lugar, la presencia de una mayor proporción de hombres en un área podría estar relacionada con tasas de criminalidad más altas, y esto ocurre tal vez por la mayor participación de hombres en actividades delictivas actualmente. Además, áreas donde menos residentes se sienten inseguros tienden a tener niveles de criminalidad bajos, lo que refleja un ambiente pacífico y una menor incidencia de delitos. La correlación positiva con la población joven de 16 a 30 años podría implicar que esta franja

[9] La seguridad percibida puede afectar la delincuencia, ya que áreas con altos niveles de inseguridad pueden experimentar mayores tasas de criminalidad.

[10] El déficit cuantitativo y cualitativo de vivienda pueden estar relacionados con la delincuencia, ya que las condiciones precarias de vivienda pueden influir en la calidad de vida y la propensión al crimen.

[11] Los estratos socioeconómicos y la pobreza multidimensional pueden influir en la delincuencia debido a las disparidades socioeconómicas.

[12] Esto arrojará luz sobre la influencia de las variables predictoras en la variable que se busca predecir. El cálculo del valor de la correlación se realiza a través del coeficiente de correlación de Pearson, el cual oscila entre -1 y 1.

de edad tiene un impacto en el aumento de la actividad delictiva, posiblemente relacionado con desafíos socioeconómicos y características propias de la juventud. Además, la correlación con la población que presenta déficit cuantitativo y cualitativo de vivienda podría señalar la influencia de limitaciones educativas y de calidad de vida en la participación en actividades delictivas. Las áreas con una mayor proporción de población en estratos socioeconómicos bajos (1 o 2) podrían experimentar tasas de criminalidad más altas, relacionadas con desafíos económicos y sociales. Además, la falta de educación, el desempleo y la afiliación al SISBEN[13] también podrían estar vinculados a tasas de criminalidad más altas, sugiriendo que limitaciones en oportunidades educativas, laborales, desigualdades económicas y sociales desempeñan un papel importante en la delincuencia.

En segundo lugar, se observa una correlación negativa entre el conteo de crímenes y la proporción de personas que pertenecen a los estratos 3 o 4 (medio-bajo o medio) de la población. Esto podría deberse a que la pertenencia a estos estratos socioeconómicos suele estar asociada con una mayor estabilidad económica y social. Esta mayor estabilidad, a su vez, puede ejercer una influencia indirecta en la reducción de la delincuencia en comparación con áreas donde los niveles socioeconómicos son más bajos.

Por otro lado, un aumento en la pobreza multidimensional (IPM) podría correlacionarse con tasas más altas de criminalidad debido que las áreas con un alto IPM a menudo albergan a comunidades en situación de vulnerabilidad socioeconómica, lo que incluye la falta de acceso a servicios básicos como educación, atención médica, la falta de oportunidades de empleo y desarrollo en estas áreas también puede empujar a individuos, especialmente a los jóvenes, hacia actividades delictivas como medio de subsistencia. Por último, una mayor densidad criminal, particularmente en un período de 7 días, llevaría a un aumento de la criminalidad en un área; debido a que menudo estas áreas atraen a más delincuentes, ya que pueden esperar encontrar más oportunidades para llevar a cabo sus actividades delictivas y pasar desapercibidos en la multitud.

Es fundamental recordar que estas interpretaciones son generales y que la relación real entre estas variables y el conteo de crímenes puede variar según el contexto específico y otros factores influyentes, ya que la correlación no implica causalidad y se necesita un análisis más profundo para comprender completamente las relaciones económicas y sociales en una región determinada. Adicionalmente, los modelos carecen de la capacidad para sugerir reformas o intervenciones específicas que contribuirían al control de la delincuencia.

[13] El SISBEN es el Sistema de Identificación de Potenciales Beneficiarios de Programas Sociales, organización que se encarga de clasificar la población según sus niveles de vida e ingresos, permitiendo así una asignación más precisa de la inversión social, con el objetivo de dirigirla hacia aquellos que tienen mayores necesidades.

3.4. Modelos de machine learning

En esta sección, se exponen los modelos de aprendizaje automático empleados para pronosticar la probabilidad que ocurra cierto tipo de crimen. Se describe además, el proceso de entrenamiento y las medidas utilizadas para evaluar el rendimiento de dichos modelos.

3.4.1. Random Forest (RF)

El Random Forest (RF) es una técnica de aprendizaje automático que presenta diversas ventajas en comparación con los algoritmos de aprendizaje profundo (Lu y Li, 2019). En los estudios más recientes sobre la predicción de puntos críticos de delincuencia, el algoritmo de Random Forest ha demostrado obtener resultados prometedores en términos de predicción (Lim, 2007; Yao et al., 2020). Adicionalmente, en Levine (2008) y Bogomolov et al. (2014), demuestran que el RF al ser comparado con otros métodos de aprendizaje automático (como las Neural Networks), también logra obtener resultados predictivos satisfactorios. Además, es uno de los algoritmos más populares debido a su simplicidad y su capacidad para ser utilizado tanto en tareas de clasificación como en tareas de regresión (Raza y Victor, 2021).

3.4.2. Extreme Gradient Boosting (XGBoost)

El Extreme Gradient Boosting (XGBoost) es una técnica de aprendizaje automático ampliamente elogiada por su versatilidad y alto rendimiento. Su principal característica es su capacidad para mejorar constantemente la precisión de los modelos a través de un proceso de impulso, en el que combina múltiples modelos débiles, generalmente árboles de decisión, convirtiéndolo en un modelo fuerte y preciso. Esto garantiza una mejor capacidad de generalización y previene el sobreajuste a los datos de entrenamiento. Además, XGBoost incorpora técnicas de regularización para evitar el sobreajuste y maneja eficientemente los datos faltantes, lo que facilita su uso en una variedad de aplicaciones. Su velocidad y eficiencia son notables, lo que lo convierte en una elección adecuada para conjuntos de datos extensos y aplicaciones en tiempo real.

3.4.3. Modelo de Mínimos Cuadros Ordinarios (MCO)

La regresión de Mínimos Cuadrados Ordinarios es una técnica estadística fundamental utilizada para modelar y comprender la relación entre una variable dependiente y una o más variables independientes (Gordon, 2010; Alves et al., 2018). Su importancia radica en su capacidad para

ajustar modelos que minimizan la diferencia entre los valores observados y los predichos, lo que facilita la predicción de valores futuros y la inferencia sobre la influencia de las variables predictoras en la variable de respuesta (Zou et al., 2003).

Se utiliza el MCO como punto de referencia para evaluar el rendimiento de otros métodos, como RF y XGBoost, en la predicción del crimen en Medellín. La elección de estos enfoques avanzados en lugar del MCO proporciona ventajas significativas en términos de precisión y capacidad predictiva. Mientras que el MCO se basa en supuestos de linealidad y normalidad, lo que puede limitar su capacidad para capturar relaciones complejas en los datos, tanto RF como XGBoost son métodos de aprendizaje automático que pueden manejar de manera más efectiva patrones no lineales y relaciones no convencionales, de tal manera que se tendrá una mayor precisión en la predicción de eventos delictivos. Esta adopción de enfoques más sofisticados representa una mejora sustancial en la capacidad de modelado y predicción, lo que resulta en una mayor eficacia en la comprensión y prevención de la delincuencia en Medellín.

3.5. Entrenamiento y evaluación

Teniendo la base de datos completa (Tabla A.1), para el modelo se definieron los conjuntos de entrenamiento y de prueba, con una distribución del 80 % y 20 % respectivamente. Es decir, se tiene para el entrenamiento los años del 2008 al 2016, y para prueba los años 2017 y 2018, con el fin de predecir el año 2018 y 2019.

Para que las características estén en un formato específico para un rendimiento óptimo, antes de entrenar el modelo, se escalan o normalizan las características numéricas y se codifican las características categóricas. Finalmente, se importan las librerías necesarias para la implementación del modelo de RF, XGBoost[14] y MCO.

Se crea una matriz de características completa y preparada para usar en un modelo de ML, imputando características basadas en datos históricos, sociodemográficos, como identificadores de grilla y códigos de barrios, características relacionadas con la fecha, densidad criminal y categorías o

[14]Para encontrar la combinación óptima de hiperparámetros del modelo de aprendizaje automático, se empleó "RandomizedSearchCV". Esta herramienta realiza una búsqueda aleatoria de hiperparámetros y evalúa su rendimiento utilizando validación cruzada. En este caso, se aplicó una validación cruzada de 3-folds, dividiendo los datos en subconjuntos de prueba y entrenamiento. Se realizaron 10 iteraciones para identificar la configuración que maximizará el rendimiento según la métrica seleccionada. Para la implementación del modelo de Random Forest classifier (RFC), se seleccionan los siguientes hiperparámetros de entrada, teniendo en cuenta que, durante el entrenamiento, el modelo aprenderá a realizar predicciones basadas en las características de entrada para clasificar las categorías: $radom\_state = 42$; $n\_jobs = 1$; $class\_weight =$ '$balanced$'; $max\_depth = 69$; $n\_estimators = 64$ y por último, se definió $min\_samples\_leaf = 132$. Lo mencionado previamente se llevó a cabo con el objetivo de potenciar la capacidad predictiva del modelo RFC. Para el modelo de XGBoost se tomaron como parámetros: el 0.3 de la tasa de aprendizaje, seguido de $n\_estimators$ con un valor de 100; un $max\_depth = 6$, $min\_child\_weigth = 1$, un gamma con valor de cero; valores de 0 para el *alpha* y 1 para el *lambda* con el fin de controlar la complejidad del modelo y finalmente, se tomó un *objetive* de 'binary:logistic' para clasificación binaria por defecto.

tipos de crímenes[15]. El resultado es una predicción de probabilidad (entre 0 y 1) para las 5 categorías de delitos (hurto a persona, hurto a residencias, extorsiones, hurto a carros y homicidios), para una fecha determinada, dentro de un barrio y grilla específico.

La evaluación de cada modelo se realiza en términos de precisión, exactitud, sensibilidad y puntuación F1-Score.

4. Resultados

En esta sección, se exponen los principales resultados de este estudio. Inicialmente, se tiene el rendimiento general de los modelos predictivos. Después se identifican los mejores predictores y su vínculo con la literatura. Finalmente, se muestra una representación gráfica de los puntos críticos de actividad delictiva en una área determinada.

4.1. Rendimiento de los modelos

En las Tablas 1 y 2 se presentan cuatro métricas: exactitud, sensibilidad, precisión y puntuación F1-Score que permiten evaluar el rendimiento de los modelos de clasificación y la regresión de mínimos cuadrados ordinarios.

Tabla 1: Rendimiento de los modelos a nivel de barrios

Modelos	Precisión	Sensibilidad	F1	Exactitud
Random Forest	0.80	0.76	0.70	0.86
XGBoost	0.79	0.77	0.71	0.83
MCO	0.65	0.57	0.54	0.68

Fuente: Elaboración propia.

Tabla 2: Rendimiento de los modelos a nivel de grillas

Modelos	Precisión	Sensibilidad	F1	Exactitud
Random Forest	0.75	0.78	0.68	0.83
XGBoost	0.74	0.76	0.67	0.81
MCO	0.61	0.56	0.52	0.65

Fuente: Elaboración propia.

A nivel de barrios y grillas, a pesar de que se obtuvieron valores muy similares, se puede observar que a nivel de barrios estos valores fueron más altos. El Random Forest demostró un rendimiento destacado en términos de exactitud, alcanzando un valor de 86 %, lo que significa que es altamente

[15] La imputación consiste en estimar o calcular valores para características faltantes o no disponibles en el conjunto de datos actual utilizando información histórica, para esto se utilizan las técnicas de relleno hacia adelante (forward fill) y el relleno hacia atrás (backward fill), los cuales toman los valores disponibles en las filas cercanas en el mismo conjunto de datos en función del tiempo.

preciso en la identificación de casos positivos. El modelo tiene una precisión alta, lo que sugiere que el modelo es efectivo al prever con exactitud el resultado esperado. En términos prácticos, esto significa que aproximadamente el 80 % de las veces, las predicciones del modelo son acertadas. Por otro lado, el XGBoost también tuvo un buen rendimiento con una exactitud de 83 %. Sin embargo, dado que se obtuvo una precisión más alta con Random Forest a nivel de barrios, este fue el modelo que se aplicó para obtener los mapas de calor que se presentan en la sección 4.3. Al analizar el rendimiento del desempeño del modelo de MCO tanto a nivel de barrios como a nivel de grillas, se observa que alcanzó una precisión significativamente inferior. Esta baja precisión se traduce en un rendimiento global menor en comparación con los modelos de ML implementados[16]. Este análisis de los modelos destaca las diferencias en la eficacia predictiva, proporcionando una visión más completa de su rendimiento en diferentes niveles de granularidad espacial.

4.2. Mejores predictores de incidencia de crímen

Hay factores tanto sociales como económicos que contribuyen a que los delincuentes cometan delitos o que las víctimas sean el blanco de los delitos; estas características son útiles cuando se trata de predecir la ocurrencia futura de delitos, lo que puede ayudar a reducir las tasas de criminalidad en muchas comunidades, entre estas características se incluyen la edad, el género, la ubicación, el número de delincuentes, los ingresos, el arma utilizada, el nivel educativo, factores políticos y económicos, cultura, empleo, conceptos legales, la hora, la fecha, el día de la semana y el mes son algunos factores a considerar.

La Figura A.1 del apéndice, representa las variables individuales que poseen la mayor capacidad predictiva en cada uno de los modelos de aprendizaje automático a nivel de barrio. Es interesante resaltar que en los dos modelos de RF y XGBoost, la proporción de los hombres tuvo gran poder predictivo, al igual que la proporción de desempleados[17]. En el modelo de RF la proporción de personas que tienen pobreza multidimensional (IPM) tuvo el mayor poder predictivo[18] junto con las personas que no saben leer y escribir. En comparación al XGBoost, en el cual, la proporción de personas que hacen parte del SISBEN, y las que pertenecen a los estratos socioeconómicos 1 o 2, tienen un alto poder predictivo[19]. Además, García et al. (2012) observaron en su estudio que la

[16] El hecho de que Random Forest y el XGBoost, aplicado a nivel de barrios, haya superado en precisión al modelo de MCO destaca la capacidad de los modelos de machine learning para capturar patrones complejos y no lineales en los datos, especialmente en entornos más detallados.

[17] Esto debido posiblemente a que las personas que enfrentan dificultades económicas debido al desempleo a menudo experimentan estrés financiero, angustia y desesperación. En algunos casos, estas tensiones pueden llevar a comportamientos delictivos como el robo o el fraude en un intento de aliviar su situación económica.

[18] Goin et al. (2018) en su estudio, afirmaron que la pobreza puede llevar a actividades delictivas como fuente de ingresos y es relevante en la predicción del crimen, concorde a lo que se encontró sobre la pobreza multidimensional.

[19] Una hipótesis sobre este comportamiento sería que por la falta de oportunidades, acceso limitado a educación de calidad, empleos estables y servicios de salud, entre otros, pueden aumentar las posibilidades de que las personas busquen alternativas para sobrevivir o mejorar su calidad de vida, incluyendo actividades delictivas.

mayoría de las víctimas de homicidios en Medellín provenían de entornos socioeconómicos bajos, lo cual justificaría por qué las variables del promedio de personas que hacen parte del estrato bajo aumentan la delincuencia. Explorar en detalle las razones por las cuales éstas variables se convierten en indicadores significativos de la delincuencia queda fuera del alcance de esta investigación, principalmente debido a que el propósito de este análisis es puramente predictivo y no busca establecer relaciones de causalidad.

Adicionalmente, la Figura A.2 del apéndice, ilustra las características que exhiben una alta capacidad predictiva de los modelos implementados a nivel de grilla. Se observaron resultados parecidos, donde las variables de la proporción de hombres, proporción de personas desempleadas, proporción de analfabetas y personas que no estudian tienen un alto nivel predictivo. Un aspecto relevante fue que la proporción de personas que pertenecen al estrato económico medio-bajo (3) o medio (4) y proporción de personas que tienen déficit cuantitativo obtuvieron un poder de predicción alto en el modelo de XGBoost.

En resumen, estos resultados son acordes a lo mencionado por He et al. (2017) quienes en su estudio demuestran que ciertos factores socioeconómicos pueden explicar la presencia de delitos violentos en áreas de alta criminalidad.

4.3. Mapas de calor

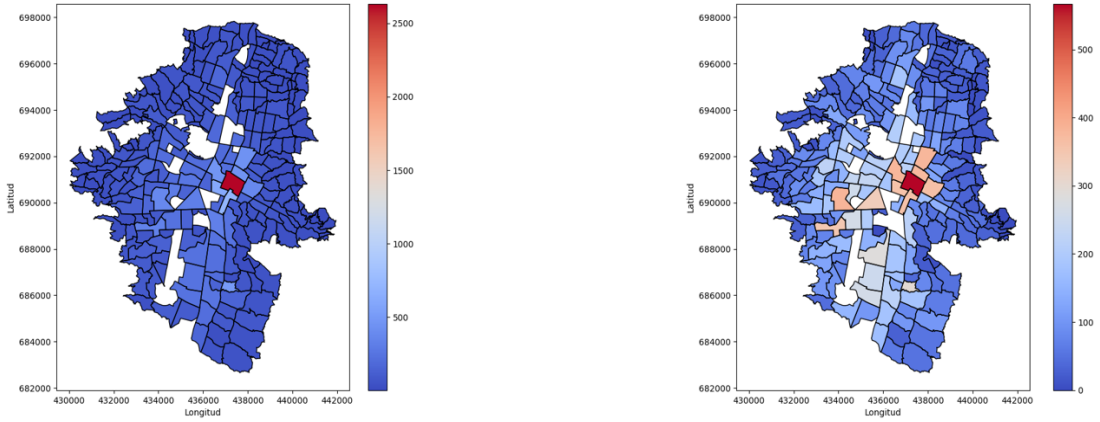
Los resultados de las predicciones de los modelos se materializan como un mapa de calor que muestra la distribución de los delitos según los datos disponibles, en los cuales el color rojo representa los lugares que concentran un alto nivel de delincuencia. Los expertos en análisis delictivo pueden aprovechar esta herramienta para tomar decisiones fundamentadas y elaborar estrategias efectivas.

En las Figuras 5 y 6, se presenta una comparación entre los conteos de crímenes totales en los diversos barrios de Medellín y las predicciones del modelo correspondientes a los años 2018 y 2019, respectivamente. Resulta evidente que tanto en la realidad como en las estimaciones del modelo, el centro de la ciudad se destaca como la zona con la mayor concentración de eventos delictivos (barrio la Candelaria). Las áreas cercanas al centro también registran niveles significativos de criminalidad a lo largo de todo el período de estudio, mientras que las periferias de la ciudad presentan muy poca actividad criminal.

En el caso de la predicción, se puede ver que se presentan eventos delictivos en los barrios cercanos al centro, lo cual se explica por diversas teorías como la dada por Hino y Amemiya (2019)

quienes mencionaron que de acuerdo con la teoría de la victimización repetida, existe una mayor probabilidad de que la conducta delictiva ocurra en el mismo lugar donde ocurrió en el pasado, y esto debido a que se ve una mayor concentración en el centro de Medellín y en sus barrios cercanos, teniendo en cuenta que estos también se ven influenciados por la criminalidad que ocurra ahí.

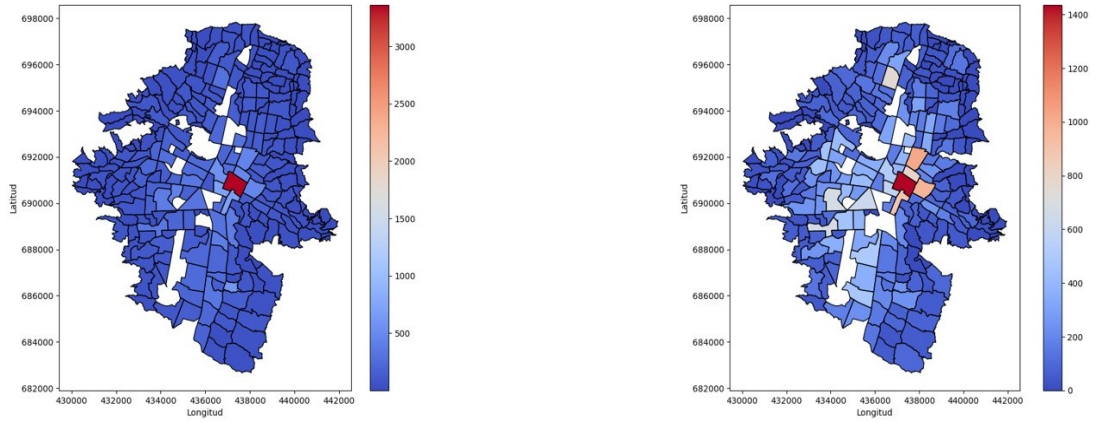
Figura 5: Mapas de calor de los reportes de crímenes del periodo 2018



(a) Real (b) Predicho

Fuente: Cálculo propio usando la base de datos del Sistema de Información para la Seguridad y la Convivencia de Medellín (SISC) y la Encuesta de Calidad de Vida (ECVM).

Figura 6: Mapas de calor de los reportes de crímenes del periodo 2019



(a) Real (b) Predicho

Fuente: Cálculo propio usando la base de datos del Sistema de Información para la Seguridad y la Convivencia de Medellín (SISC) y la Encuesta de Calidad de Vida (ECVM).

De igual manera, Tobler (1979) formuló la primera ley de la geografía, que estableció que todo se relaciona con todo, pero lo que está cerca está más relacionado que lo que está lejos. Esta ley proporciona una base sólida para comprender por qué las áreas cercanas a zonas con alta incidencia de crímenes se ven más afectadas que las áreas más alejadas. Adicionalmente, Zhao y Tang (2017) y Yi et al. (2018) mencionaron que características urbanas similares o proximidad geográfica entre regiones pueden dar lugar a patrones de delincuencia semejantes.

La "Teoría de la Concentración Espacial de la Delincuencia" dada en el investigación de Weis-

burd y Green (1995) y Weisburd et al. (2023) sugieren que los crímenes tienden a agruparse en puntos calientes, donde factores como la oportunidad delictiva y la disponibilidad de objetivos son más favorables. A medida que nos alejamos de estos puntos calientes, es probable que la tasa de criminalidad disminuya. Adicionalmente, la "Teoría de la Difusión Espacial" apunta a que una vez que ocurre un crimen en un área, es más probable que ocurran crímenes similares en áreas cercanas debido a la imitación de comportamientos delictivos y la movilidad de los delincuentes (Braga et al., 2014). Estas teorías son fundamentales para entender cómo los patrones de criminalidad se distribuyen en el espacio y cómo las estrategias de prevención pueden ser efectivas en áreas afectadas.

A pesar de que según los resultados del modelo de predicción, se mantiene la tendencia y ubicación de los crímenes en el centro de la ciudad, el conteo es subestimado dado principalmente por el subregistro de las víctimas[20], lo cual hace que aumente el error al realizar una predicción.

Por otro lado, en las Figuras A.3, A.4, A.5 y A.6, se presentan los conteos de crímenes que se tienen por cada una de las 5 categorías o tipos de conducta que se tienen en la base de datos de crimen de Medellín. Se presentan tanto los conteos reales como los predichos por el modelo para el año 2018 y 2019, respectivamente. De acuerdo a los mapas, se analiza que dependiendo de las características localizadas de las diferentes modalidades delictivas y los factores circundantes, los modelos de predicción del delito generan una variedad de resultados; donde los delitos de extorsión, homicidio y hurto a persona en el 2018, se concentran en el barrio La Candelaria (que es el centro de la ciudad) y en barrios circundantes, a comparación de hurto a residencia y hurto a carro que predominan al sur y oeste de la ciudad (barrios de Santa Fe, Campo Amor, San Bernardo, La Gloria, entre otros). Adicionalmente, se puede ver que el hurto a personas es el que más va a tener reportes de crímenes en toda la ciudad porque es el crimen que más tiene denuncias. En el período predicho del 2019, se puede observar que tuvo un comportamiento similar al real, donde el delito de extorsión y homicidio se presentan en el centro de la ciudad, en cambio el hurto a persona se puede ver que en el 2019 va a crecer tanto en el centro de la ciudad como en zonas aledañas a este, del hurto a residencia a 2019 puede determinarse que va a disminuir en ese año, finalmente, el hurto a carro se va a mantener constante; es decir se van a presentar en los barrios de Belen, Los Conquistadores, Laureles y Bolivariana.

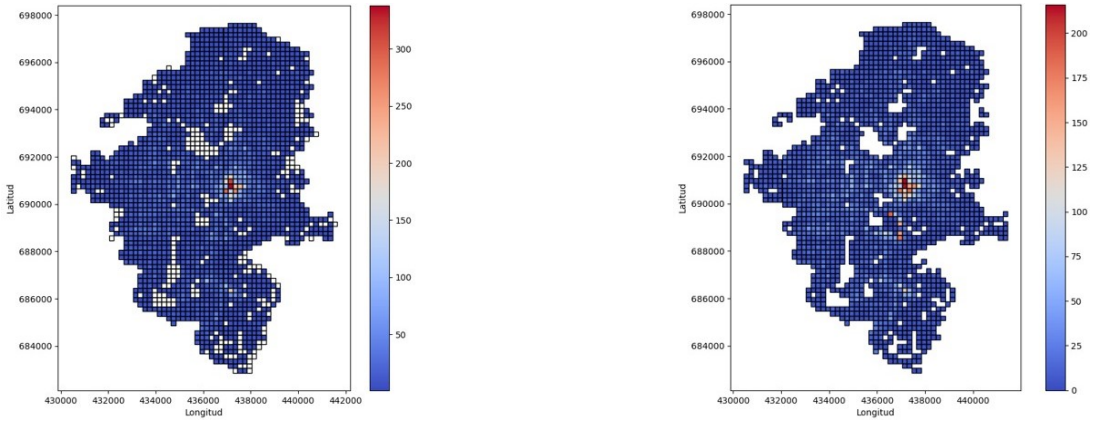
De igual forma, se puede analizar que la delincuencia varía de un barrio a otro, y que además, el algoritmo de RF predijo las categorías de delitos con una alta precisión, lo que permite confiar en el sistema para predecir delitos futuros. Adicionalmente, estos resultados indican que en la zona

[20] Donde las personas que son víctimas de algún delito no denuncian ya sea por limitaciones geográficas o estructurales, miedo a la victimización, desconfianza en la policía, o simplemente no hay incentivos para denunciar.

urbana de Medellín, específicamente en los barrios con baja incidencia delictiva, la predicción del delito puede ser menos precisa debido a la limitada cantidad de eventos registrados. Es importante destacar que los hallazgos específicos para la ciudad de Medellín no pueden extrapolarse a otras ciudades o municipios en el país.

Las grillas permiten dividir un área geográfica en celdas más pequeñas, lo que brinda una mayor granularidad en la identificación de áreas de alto riesgo (Figura 7 y 8). Este enfoque puede ser de utilidad para las autoridades encargadas de hacer cumplir la ley en la optimización de la asignación de sus recursos. Los mapas tienen la capacidad de proporcionar un análisis minucioso de los patrones delictivos, revelando tendencias y conexiones que podrían no ser evidentes a nivel de áreas geográficas más amplias. De manera adicional, esta información puede servir como base para la formulación y aplicación de políticas y estrategias en el ámbito de la seguridad pública.

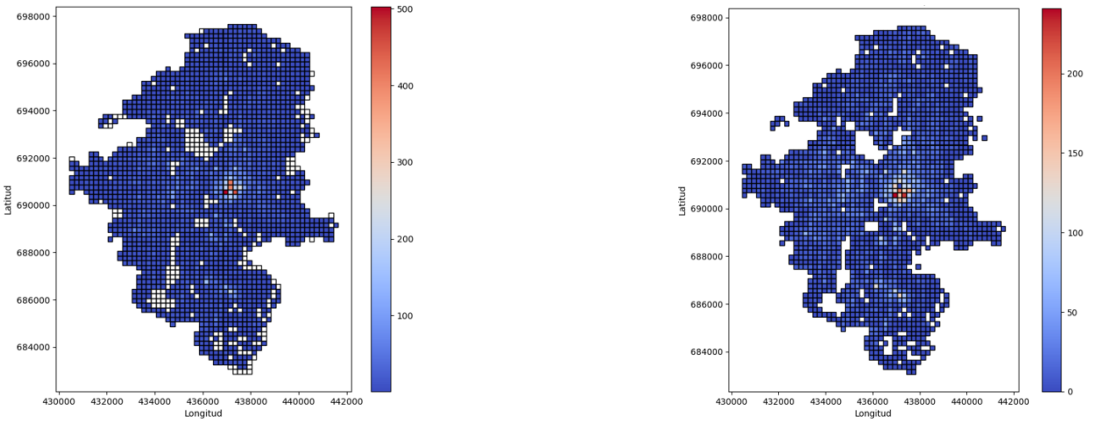
Figura 7: Mapas de calor de los reportes de crímenes del periodo 2018 a nivel de grilla



(a) Real (b) Predicho

Fuente: Cálculo propio usando la base de datos del Sistema de Información para la Seguridad y la Convivencia de Medellín (SISC).

Figura 8: Mapas de calor de los reportes de crímenes del periodo 2019 a nivel de grilla



(a) Real (b) Predicho

Fuente: Cálculo propio usando la base de datos del Sistema de Información para la Seguridad y la Convivencia de Medellín (SISC).

Se percibe que las concentraciones de reportes de crímenes persisten en el centro de la ciudad de Medellín, en consonancia con lo observado a nivel de los barrios. Asimismo, el mapa demuestra que las predicciones correspondientes a los años 2018 y 2019 muestran una notable semejanza con los datos reales.

Los anteriores mapas representan un valioso recurso que puede ser instrumental para la detección de áreas con una mayor vulnerabilidad a la delincuencia o aquellas que han experimentado transformaciones en su perfil delictivo. Utilizando esta información como base, se pueden tomar decisiones estratégicas, como la asignación de recursos adicionales a las comunidades que presentan incrementos significativos en la probabilidad delictiva o la implementación de programas de prevención criminal específicamente diseñados para áreas identificadas.

En resumen, el enfoque presentado en este ejercicio es de naturaleza predictiva en lugar de causal. Esto implica que estas herramientas tienen la capacidad de identificar las ubicaciones con mayor probabilidad de experimentar actos de delincuencia.

5. Conclusiones

Al llegar al cierre de este estudio exhaustivo sobre la predicción del crimen mediante el empleo de técnicas de aprendizaje automático, se abre un panorama esclarecedor en el ámbito de la seguridad ciudadana y la comprensión de los patrones delictivos. A lo largo de esta investigación, se pretendió estudiar los patrones espaciales de delitos a través de la implementación de los métodos de Mínimos Cuadros Ordinarios (MCO), Random Forest (RF) y Extreme Gradient Boosting (XGBoost), al igual que determinar la probabilidad que ocurra cierto tipo de delito en un barrio y grilla en la ciudad de Medellín anualmente. Debido a que se ha destacado la capacidad inherente de los modelos de aprendizaje automático para analizar y procesar conjuntos de datos de delincuencia, permitiendo revelar relaciones no lineales y tendencias sutiles que de otro modo podrían haber pasado desapercibidas.

De acuerdo a los resultados del modelo, se pudo observar que los patrones de delincuencia intraurbana de los diferentes tipos de crímenes exhiben comportamientos distintos relacionados con el tiempo y el espacio. También se evidenció una amplia variabilidad en la frecuencia de los delitos, que se manifiesta de manera específica en los mapas a nivel de barrio y grilla. Otro hallazgo significativo fue la concentración de la mayoría de las denuncias de delitos en el centro de la ciudad de Medellín, particularmente en el barrio La Candelaria, tanto a nivel de desagregación de barrios como en áreas geográficas más específicas (grillas). Esto se explica en parte por la alta afluencia

de turismo, actividad comercial y una densa red vial en esta zona, donde circulan diariamente más de un millón de personas. Esta concentración de actividades ha generado desafíos relacionados con robos, conflictos por el control territorial en plazas de vicio y extorsión económica. Todo lo anterior, apunta a la necesidad de desarrollar políticas públicas y estrategias de intervención específicas para mejorar la seguridad y la calidad de vida en esta área crítica de la ciudad, optimizando al mismo tiempo los recursos tecnológicos y humanos disponibles.

El desempeño de los modelos es bueno, lo que facilita la comprensión de las características de los barrios que tienen una mayor capacidad predictiva para anticipar la ocurrencia de la delincuencia en la ciudad de Medellín. De manera sorprendente, las variables relacionadas con la proporción de hombres, proporción de personas entre 16 y 30 años de edad, proporción de personas desempleadas, proporción de personas que tienen déficit cuantitativo de vivienda, personas que pertenecen al SISBEN y que hacen parte del estrato socioeconómico 1 o 2, aumentan la criminalidad. Mientras que variables como la densidad criminal en las dos ventanas de tiempo en 7 y 30 días, proporción de personas entre 0 y 15 años de edad y proporción de personas que manifestaron sentirse inseguras en el último año en el barrio donde vivían tienen el menor poder predictivo hacia la delincuencia en el período de estudio, en los modelos de Random Forest y XGBoost. Donde se puede afirmar que los datos de las víctimas son fundamentales, ya que resultan necesarios para prever posibles futuras víctimas de delitos (Browning et al., 2010).

Adicionalmente, se observó que los modelos de Random Forest y XGBoost a nivel de barrios tuvieron una precisión similar, con una tasa de acierto del 86 % y del 83 %, respectivamente. Aunque estos niveles no son tan altos, igual pueden considerarse que son valores aceptables, permitiendo que estos modelos puedan ser empleados por las autoridades o policymakers para prevenir la ocurrencia de delitos en Medellín. Por otro lado, en el caso del modelo de Mínimos Cuadrados Ordinarios (MCO), se tuvo una tasa de acierto del 68 % a nivel de barrios. Por tanto, el modelo con el mejor rendimiento fue el de Random Forest.

Asimismo, las políticas públicas pueden beneficiarse al utilizar estos modelos para asignar recursos de manera más eficiente, aumentar la presencia policial en áreas críticas y desarrollar estrategias de prevención del delito focalizadas. Al concentrar esfuerzos en áreas identificadas como vulnerables, Medellín puede lograr una reducción significativa en las tasas de criminalidad, mejorando la seguridad ciudadana y la calidad de vida en las comunidades. Además, estos modelos pueden ser útiles para evaluar el impacto de las intervenciones y ajustar las políticas de seguridad en consecuencia, lo que contribuye a un enfoque más basado en evidencia en la lucha contra la delincuencia en el país. Esta implicación de política pública se centra en la importancia de utilizar modelos predictivos para mejorar la eficacia de las estrategias de seguridad y prevención del delito en Co-

lombia, lo que a su vez puede tener un impacto significativo en la reducción de la criminalidad y el fortalecimiento de la seguridad en las comunidades.

Es esencial resaltar que, si bien los resultados demuestran promesa y validez, persisten desafíos y oportunidades para la mejora continua en este campo. La selección adecuada de atributos, la gestión de datos desequilibrados y la adaptación a contextos cambiantes siguen siendo aspectos fundamentales a considerar en futuras investigaciones. Asimismo, la integración de fuentes de información adicionales y la exploración de enfoques de interpretación de modelos enriquecerán aún más la comprensión de los factores que influyen en la actividad delictiva. Finalmente, para que las predicciones sean más eficaces es necesario que los datos sean actualizados constantemente, incluso se puede trabajar en recopilar y procesar datos en tiempo real, para evitar predicciones sesgadas, teniendo en cuenta que las predicciones se basan en datos históricos.

Referencias

AL Mansour, H. y Lundy, M. (2019). Crime types prediction. In *Advances in Data Science, Cyber Security and IT Applications: First International Conference on Computing, ICC 2019, Riyadh, Saudi Arabia, December 10–12, 2019, Proceedings, Part I 1*, pages 260–274. Springer.

Alegría, S. A. G., Palacios, L. E. O., Guerrero, V. B., y Erazo, H. O. (2020). Modelo de redes neuronales para predecir la tendencia de víctimas de secuestro en Colombia. *Investigación e Innovación en Ingenierías*, 8(3):38–49.

Alves, L. G., Ribeiro, H. V., y Rodrigues, F. A. (2018). Crime prediction through urban metrics and statistical learning. *Physica A: Statistical Mechanics and its Applications*, 505:435–443.

Andini, M., Ciani, E., de Blasio, G., D'Ignazio, A., y Salvestrini, V. (2018). Targeting with machine learning: An application to a tax rebate program in Italy. *Journal of Economic Behavior & Organization*, 156:86–102.

Bazzi, S., Blair, R. A., Blattman, C., Dube, O., Gudgeon, M., y Peck, R. (2022). The promise and pitfalls of conflict prediction: evidence from Colombia and Indonesia. *Review of Economics and Statistics*, 104(4):764–779.

Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of Political Economy*, 76(2):169–217.

Blattman, C., Duncan, G., Lessing, B., Tobón, S., y Mesa-Mejía, J. P. (2020). Gobierno criminal en Medellín: panorama general del fenómeno y evidencia empírica sobre cómo enfrentarlo.

Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., y Pentland, A. (2014). Once upon a crime: towards crime prediction from demographics and mobile data. pages 427–434.

Braga, A. A., Papachristos, A. V., y Hureau, D. M. (2014). The effects of hot spots policing on crime: An updated systematic review and meta-analysis. *Justice Quarterly*, 31(4):633–663.

Browning, C. R., Byron, R. A., Calder, C. A., Krivo, L. J., Kwan, M.-P., Lee, J.-Y., y Peterson, R. D. (2010). Commercial density, residential concentration, and crime: Land use patterns and violence in neighborhood context. *Journal of Research in Crime and Delinquency*, 47(3):329–357.

Collazos, D., García, E., Mejía, D., Ortega, D., y Tobón, S. (2021). Hot spots policing in a high-crime environment: An experimental evaluation in Medellín. *Journal of Experimental Criminology*, 17(3):473–506.

Cornish, D. B. y Clarke, R. V. (1989). Crime specialisation, crime displacement and rational choice theory. pages 103–117.

Cornish, D. B. y Clarke, R. V. (2016). The rational choice perspective. pages 48–80.

De Blasio, G., D'Ignazio, A., y Letta, M. (2022). Gotham city. predicting 'corrupted' municipalities with machine learning. *Technological Forecasting and Social Change*, 184:122016.

Deng, Y., He, R., y Liu, Y. (2023). Crime risk prediction incorporating geographical spatiotemporal dependency into machine learning models. *Information Sciences*, 646:119414.

Detotto, C. y Otranto, E. (2010). Does crime affect economic growth? *Kyklos*, 63(3):330–345.

Duarte-Velásquez, Y. A. y Cadavid-Carmona, J. A. (2020). Análisis de umbral: técnica diferencial en la interpretación de los registros de criminalidad en Colombia (2019). *Revista Criminalidad*, 62(2):9–144.

Falade, A., Azeta, A., Oni, A., y Odun-ayo, I. (2019). Systematic literature review of crime prediction and data mining. *Review of Computer Engineering Studies*, 6(3).

Ferro-Briceño, P. V. et al. (2021). Uso de redes neuronales para determinar la influencia del estado del pavimento en siniestros viales de la ciudad de Bogotá.

Gallego, J., Prem, M., y Vargas, J. F. (2022). Predicting politicians' misconduct: Evidence from Colombia. *Data & Policy*, 4:e41.

García, H. I., Giraldo, C. A., López, M. V., Pastor, M. d. P., Cardona, M., Tapias, C. E., Cuartas, D., Gómez, V., y Vera, C. Y. (2012). Treinta años de homicidios en Medellín, Colombia, 1979-2008. *Cadernos de Saude Pública*, 28:1699–1712.

Gelvez-Ferreira, J.-D., Nieto-Rodríguez, M.-P., y Rocha-Ruiz, C.-A. (2022). Prediciendo el crimen en ciudades intermedias: un modelo de "machine learning" en Bucaramanga, Colombia. *URVIO Revista Latinoamericana de Estudios de Seguridad*, (34):82–98.

Gerber, M. S. (2014). Predicting crime using twitter and kernel density estimation. *Decision Support Systems*, 61:115–125.

Goin, D. E., Rudolph, K. E., y Ahern, J. (2018). Predictors of firearm violence in urban communities: a machine-learning approach. *Health & Place*, 51:61–67.

Gordon, M. B. (2010). A random walk in the literature on criminality: A partial and critical view on some statistical analyses and modelling approaches. *European Journal of Applied Mathematics*, 21(4-5):283–306.

He, L., Páez, A., y Liu, D. (2017). Persistence of crime hot spots: an ordered probit analysis. *Geographical analysis*, 49(1):3–22.

Hino, K. y Amemiya, M. (2019). Spatiotemporal analysis of burglary in multifamily housing in Fukuoka City, Japan. *Cities*, 90:15–23.

Ingilevich, V. y Ivanov, S. (2018). Crime rate prediction in the urban environment using social factors. *Procedia Computer Science*, 136:472–478.

Kajita, M. y Kajita, S. (2020). Crime prediction by data-driven green's function method. *International Journal of Forecasting*, 36(2):480–488.

Khanna, G., Medina, C., Nyshadham, A., Ramos, D., Tamayo, J., y Tiew, A. (2022). Spatial mobility, economic opportunity, and crime.

Kounadi, O., Ristea, A., Araujo, A., y Leitner, M. (2020). A systematic review on spatial crime forecasting. *Crime Science*, 9:1–22.

Kshatri, S. S. y Narain, B. (2020). Analytical study of some selected classification algorithms and crime prediction. *International Journal of Engineering and Advanced Technology*, 9(6):241–247.

Levine, N. (2008). The "hottest" part of a hotspot: comments on "the utility of hotspot mapping for predicting spatial patterns of crime". *Security Journal*, 21(4):295–302.

Liang, W., Wang, Y., Tao, H., y Cao, J. (2022). Towards hour-level crime prediction: A neural attentive framework with spatial–temporal-categorical fusion. *Neurocomputing*, 486:286–297.

Lim, N. (2007). Classification by ensembles from random partitions using logistic regression models. *State University of New York at Stony Brook*.

Lima, M. S. M. y Delen, D. (2020). Predicting and explaining corruption across countries: A machine learning approach. *Government Information Quarterly*, 37(1):101407.

Lin, Y.-L., Yen, M.-F., y Yu, L.-C. (2018). Grid-based crime prediction using geographical features. *ISPRS International Journal of Geo-Information*, 7(8):298.

Lu, R. y Li, L. (2019). Crime prediction model based on Random Forest. *Journal of China Interpol Academy*, 5(3):108–112.

Maloof, M. A. (2003). Learning when data sets are imbalanced and when costs are unequal and unknown. 2:2–1.

Martin, G. (2012). Medellín, tragedia y resurrección: mafias, ciudad y Estado, 1975-2013.

Mejía, D., Ortega, D., y Ortiz, K. (2014). Un análisis de la criminalidad urbana en Colombia. *Technical report, CAF - Banco de Desarrollo de America Latina*.

Merton, R. K. (1938). Social structure and anomie. *American sociological review*, 3(5):672–682.

Mittal, M., Goyal, L. M., Sethi, J. K., y Hemanth, D. J. (2019). Monitoring the impact of economic crisis on crime in India using machine learning. *Computational Economics*, 53:1467–1485.

Mojica-Muñoz, K. S. (2021). Inteligencia artificial para detectar corrupción en la administración pública municipal de Colombia (artificial intelligence to detect corruption in Colombia's municipal public administration). *Documentos CEDE*, (31).

Mustard-David, B. (2010). How do labor markets affect crime? new evidence on an old puzzle. *Handbook On The Economics Of Crime. Elsevier*.

Muñoz, V., Vallejo, M., y Aedo, J. E. (2021). Exploratory analysis of crime behavior in the city of Medellín. pages 1–5.

Okonkwo, R. O. y Enem, F. O. (2011). Combating crime and terrorism using data mining techniques.

Ordoñez-Eraso, H.-A., Pardo-Calvache, C.-J., y Cobos-Lozada, C.-A. (2020). Detection of homicide trends in Colombia using machine learning. *Learning*, 29(54):e11740.

Ramírez, J. G. (2008). La conflicto armado urbano y violencia homicida. el caso de medellín. *URVIO: Revista Latinoamericana de Estudios de Seguridad*, (5):99–113.

Raza, D. M. y Victor, D. B. (2021). Data mining and region prediction based on crime using Random Forest. pages 980–987.

Reier-Forradellas, R. F., Náñez Alonso, S. L., Jorge-Vazquez, J., y Rodriguez, M. L. (2020). Applied machine learning in social sciences: Neural Networks and crime prediction. *Social Sciences*, 10(1):4.

Rojas-Guerrero, M., Grautoff Laverde, M., et al. (2022). Predicción de las masacres en Colombia empleando inteligencia artificial.

Rosser, G., Davies, T., Bowers, K. J., Johnson, S. D., y Cheng, T. (2017). Predictive crime mapping: Arbitrary grids or street networks? *Journal of Quantitative Criminology*, 33:569–594.

Rumi, S. K., Deng, K., y Salim, F. D. (2018). Crime event prediction with dynamic features. *EPJ Data Science*, 7(1):43.

Sánchez-Jabba, A. (2013). La reinvención de Medellín. *Lecturas de Economía*, (78):185–227.

Sánchez-Torres, F. J. y Núñez-Méndez, J. A. (2001). Determinantes del crimen violento en un país altamente violento: el caso de Colombia.

Shaw, C. R. y McKay, H. D. (1931). Report on the causes of crime. *Government Printing Office*.

Stalidis, P., Semertzidis, T., y Daras, P. (2021). Examining deep learning architectures for crime classification and prediction. *Forecasting*, 3(4):741–762.

Sun, Y., Chen, T., y Yin, H. (2023). Spatial-temporal meta-path guided explainable crime prediction. *World Wide Web*, pages 1–27.

Thuraisingham, B. (2004). Data mining for counter-terrorism. *Data Mining: Next Generation Challenges and Future Directions*, pages 157–183.

Tobler, W. R. (1979). Cellular geography. *Springer*, pages 379–386.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2):3–28.

Wang, S., Cao, J., y Philip, S. Y. (2020). Deep learning for spatio-temporal data mining: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 34(8):3681–3700.

Weisburd, D. y Green, L. (1995). Policing drug hot spots: The Jersey City drug market analysis experiment. *Justice Quarterly*, 12(4):711–735.

Weisburd, D., Maher, L., Sherman, L., Buerger, M., Cohn, E., y Petrosino, A. (2023). Contrasting crime general and crime specific theory: The case of hot spots of crime. *Routledge*, pages 45–70.

Wheeler, A. P. y Steenbeek, W. (2021). Mapping the risk terrain for crime using machine learning. *Journal of Quantitative Criminology*, 37:445–480.

Yao, S., Wei, M., Yan, L., Wang, C., Dong, X., Liu, F., y Xiong, Y. (2020). Prediction of crime hotspots based on spatial factors of Random forest. pages 811–815.

Yi, F., Yu, Z., Zhuang, F., Zhang, X., y Xiong, H. (2018). An integrated model for crime prediction using temporal and spatial factors. pages 1386–1391.

Zaidi, N. A. S., Mustapha, A., Mostafa, S. A., y Razali, M. N. (2020). A classification approach for crime prediction. pages 68–78.

Zhang, X., Liu, L., Lan, M., Song, G., Xiao, L., y Chen, J. (2022). Interpretable machine learning models for crime prediction. *Computers, Environment and Urban Systems*, 94:101789.

Zhao, X. y Tang, J. (2017). Modeling temporal-spatial correlations for crime prediction. pages 497–506.

Zou, K. H., Tuncali, K., y Silverman, S. G. (2003). Correlation and simple linear regression. *Radiology*, 227(3):617–628.

A. Apéndice

Tabla A.1: Descripción de variables

Variables	Descripción
Fecha	Contiene día, mes y año (2008-2018).
Latitud	Es la latitud del punto georreferenciado (x)
Longitud	Es la longitud del punto georreferenciado (y)
Código de barrio	Contiene el número de identificación único para cada barrio de Medelín.
Área	Contiene el área de cada barrio en metros cuadrados.
Conducta	Hurto a persona, hurto a persona, hurto a residencias, extorsión y homicidios.
Media de edad	Contiene el media de edad de las personas por barrio.
Proporción de Hombres	Media de hombres en un barrio.
Proporción de niños	Media de personas entre 0 a 15 años por barrio.
Proporción de adolescentes	Media de personas entre 16 a 30 años por barrio.
Proporción de adultos	Media de personas con más de 30 años por barrio.
Media de personas que se sienten inseguras	Media de personas que respondieron que se sentían inseguras en el barrio donde viven actualmente.
Media de personas que se sienten seguras	Media de personas que respondieron que se sentían seguras en el barrio donde viven actualmente.
Media de personas que se sienten muy seguras	Media de personas que respondieron que se sentían muy seguras en el barrio donde viven actualmente.
Proporción de desempleados	Media de personas desempleadas.
Proporción de personas analfabetas	Media de personas analfabetas; es decir que no saben leer ni escribir.
Proporción de personas que no estudian	Media de personas que no estudian.
Media de personas que son de estrato 1 o 2	Media de personas que pertenecen al estrato socio-económico 1 o 2 (1 Bajo-bajo y el 2 es Bajo).
Media de personas que son de estrato 3 o 4	Media de personas que pertenecen al estrato socio-económico 3 o 4 (3 - Medio-bajo y 4 - Medio).
Media de personas que son de estrato 5 o 6	Media de personas que pertenecen al estrato socio-económico 5 o 6 (5 - Medio-alto y 6 - Alto).
Media de personas que hacen parte del SISBEN	Media de personas que hacen parte del Sistema de Identificación de Potenciales Beneficiarios de Programas Sociales.
Media de personas que tiene pobreza multidimensional	Media de personas en un barrio que tienen pobreza multidimensional; es decir es una medición de la pobreza que refleja las múltiples carencias que enfrentan las personas pobres al mismo tiempo en áreas como educación, salud, entre otros.
Media de personas que tienen déficit cualitativo	Media de personas en un barrio que tienen déficit cualitativo de vivienda. la insuficiencia de viviendas disponibles para atender las necesidades habitacionales de la población. Esto se refiere a la falta de viviendas; en términos de cantidad, lo que puede llevar a la sobrepoblación o la ocupación de viviendas inadecuadas
Media de personas que tienen déficit cuantitativo	Media de personas en un barrio que tienen déficit cuantitativo de vivienda; se refiere a la calidad de las viviendas y a la falta de servicios básicos, la presencia de condiciones insalubres o inseguras en las viviendas, y la falta de acceso a agua potable, saneamiento y otros aspectos que afectan negativamente la calidad de vida de los residentes en esas viviendas.
Media de densidad criminal (7 días)	Variable de densidad de delincuencia de conteo de 7 días, tomando como referencia la variable "Fecha".
Media de densidad criminal (30 días)	Variable de densidad de delincuencia de conteo de 30 días, tomando como referencia la variable "Fecha".
Día de la semana	Esta variable representa el número del día de la semana, tomando como referencia la variable "Fecha".
Estación	Esta variable es como una aproximación o proxy para la estación del año en la que ocurrió el evento registrado de crimen en la columna "Fecha", tendrá valores enteros del 1 al 4, representando los trimestres del año (por ejemplo, 1 para enero a marzo, 2 para abril a junio, y así sucesivamente).
Fin de semana	Esta variable es una indicación de si el día correspondiente fecha en a la la columna "Fecha" del dataframe es un fin de semana o no.

Fuente: Elaboración propia.

Tabla A.2: Estadísticas descriptivas

Nombre de las variables	**Obs.**	**Media**	**Desv. Est.**
Área	93,134	422080.09	201,881
Media de edad	93,134	40.86	7.24
Media de hombres	93,134	0.45	0.06
Media de personas que se sienten inseguros en el barrio donde vive	93,134	0.22	0.16
Media de personas que se sienten seguros en el barrio donde vive	93,134	0.71	0.15
Media de personas que se sienten muy seguros en el barrio donde vive	93,134	0.08	0.07
Proporción de personas entre 0 y 15 años de edad	93,134	0.15	0.09
Proporción de personas entre 16 y 30 años de edad	93,134	0.23	0.07
Proporción de personas mayores de 30 años de edad	93,134	0.62	0.13
Media de personas con pobreza multidimensional	93,134	0.81	0.32
Media de personas con déficit cuantitativo de vivienda	93,134	0.41	0.12
Media de personas con déficit cualitativo de vivienda	93,134	0.07	0.43
Media de personas que hacen parte del estrato 1 o 2	93,134	0.18	0.32
Media de personas que hacen parte del estrato 3 o 4	93,134	0.64	0.39
Media de personas que hacen parte del estrato 5 o 6	93,134	0.18	0.33
Proporción de personas que no estudian	93,134	0.44	0.27
Proporción de personas son analfabetas	93,134	0.06	0.18
Proporción de personas desempleadas	93,134	0.12	0.02
Media de personas que pertenecen al SISBEN	93,134	0.05	0.07

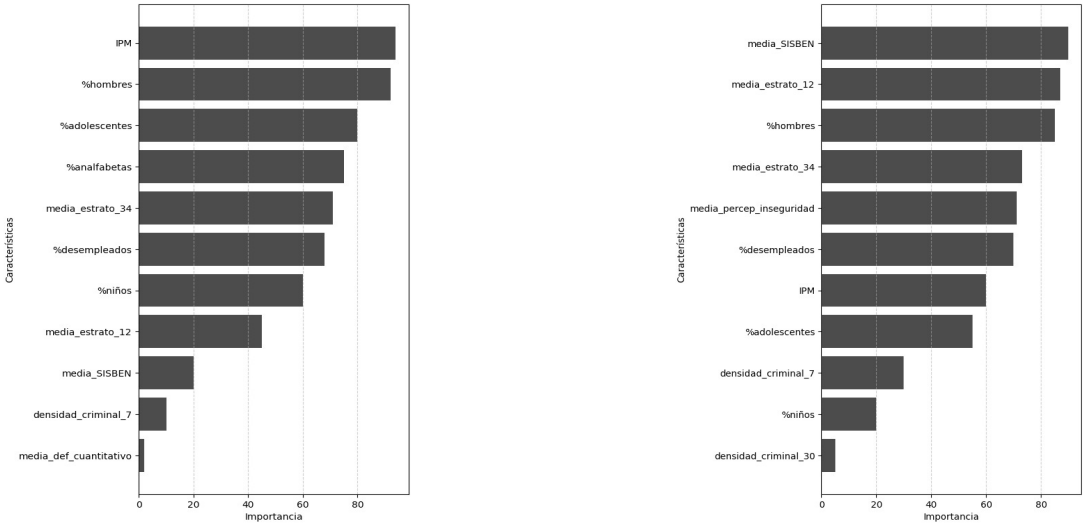
Fuente: Elaboración propia.

Tabla A.3: Correlación entre variables económicas y variable objetivo

Variables económicas	**Cantidad de crímenes**
media_estrato_12	0.22
media_estrato_34	-0.21
%hombres	0.03
%adolescentes	0.19
%desempleados	0.07
%analfabetas	0.11
IPM	0.17
media_def_cuantitativo	0.04
media_def_cualitativo	0.19
media_SISBEN	0.07
media_percep_inseguridad	0.13
densidad_criminal_7	0.12

Fuente: Elaboración propia.

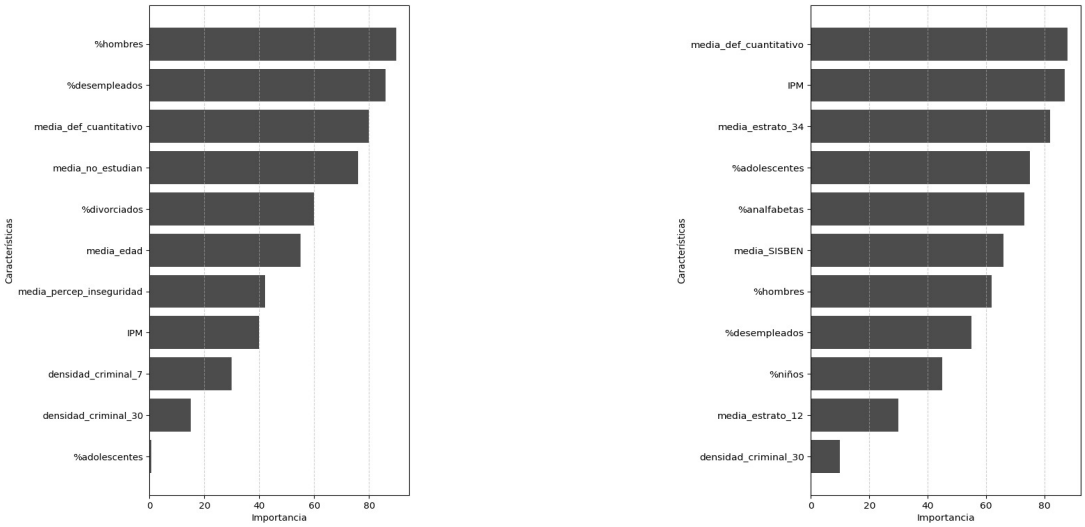
Figura A.1: Importancia de las variables a nivel de barrio



(a) Random Forest (b) XGBoost

Fuente: Cálculo propio.

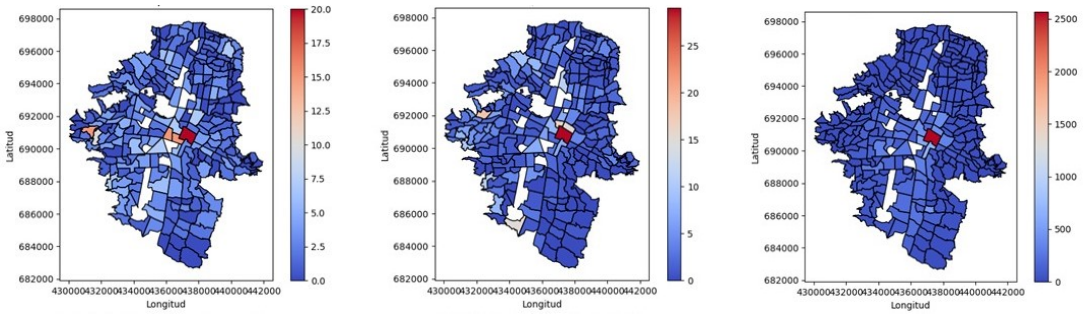
Figura A.2: Importancia de las variables a nivel de grilla



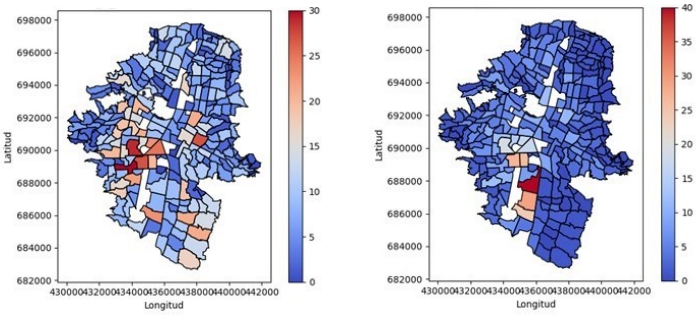
(a) Random Forest (b) XGBoost

Fuente: Cálculo propio.

Figura A.3: Mapa de calor real de los 5 tipos de reportes reales de crímenes del periodo 2018



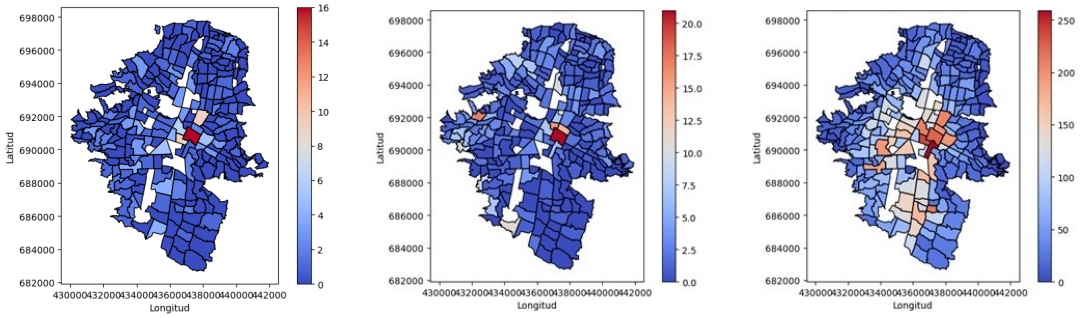
(a) Extorsión (b) Homicidio (c) Hurto a persona



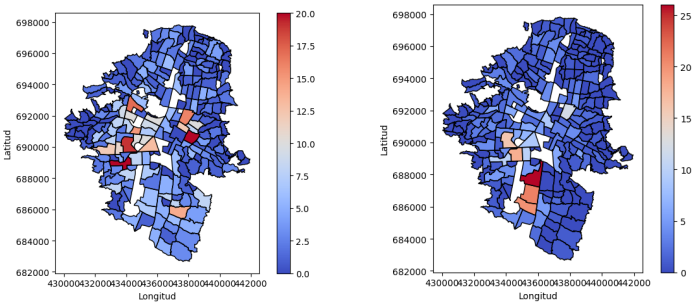
(d) Hurto a residencia (e) Hurto a carro

Fuente: Cálculo propio usando la base de datos del Sistema de Información para la Seguridad y la Convivencia de Medellín (SISC).

Figura A.4: Mapa de calor del modelo predicho de los 5 tipos de reportes de crímenes del periodo 2018



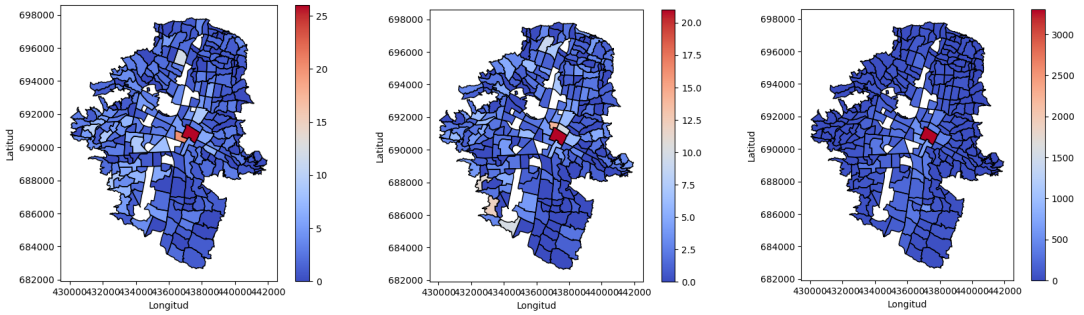
(a) Extorsión (b) Homicidio (c) Hurto a persona



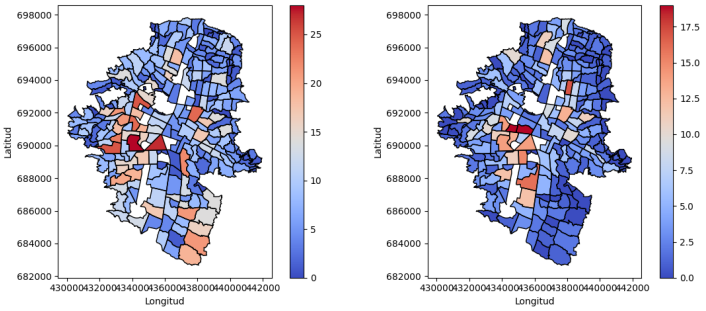
(d) Hurto a residencia (e) Hurto a carro

Fuente: Cálculo propio usando la base de datos del Sistema de Información para la Seguridad y la Convivencia de Medellín (SISC).

Figura A.5: Mapa de calor real de los 5 tipos de reportes reales de crímenes del periodo 2019



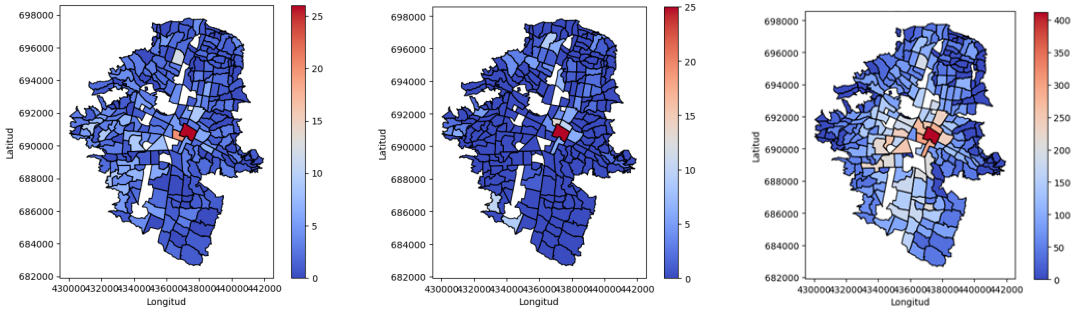
(a) Extorsión (b) Homicidio (c) Hurto a persona



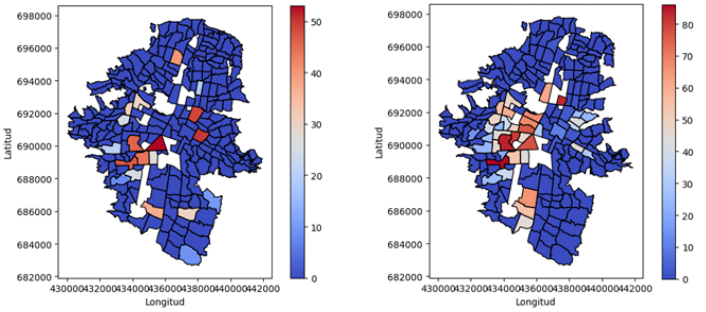
(d) Hurto a residencia (e) Hurto a carro

Fuente: Cálculo propio usando la base de datos del Sistema de Información para la Seguridad y la Convivencia de Medellín (SISC).

Figura A.6: Mapa de calor del modelo predicho de los 5 tipos de reportes de crímenes del periodo 2019



(a) Extorsión (b) Homicidio (c) Hurto a persona



(d) Hurto a residencia (e) Hurto a carro

Fuente: Cálculo propio usando la base de datos del Sistema de Información para la Seguridad y la Convivencia de Medellín (SISC).