Detección y predicción de patrones de hurto de locales comerciales y viviendas en Cundinamarca mediante modelos de Machine Learning

José Daniel Miranda Díaz
jmirand34441@universidadean.edu.co

Especialización en Machine Learning

Bogotá D.C.

21 de noviembre del 2024

Planteamiento del Problema

En Cundinamarca, el crimen, particularmente el hurto a viviendas y locales comerciales ha emergido como una preocupación significativa, especialmente en las zonas urbanas donde la densidad de población y la actividad económica crean un entorno propicio para actividades delictivas. A medida que las ciudades colombianas continúan expandiéndose y urbanizándose, los retos asociados con la prevención del delito también aumentan. Las autoridades locales y las fuerzas de seguridad enfrentan la creciente dificultad de gestionar recursos limitados de manera eficiente, mientras intentan anticiparse a los crímenes y prevenirlos antes de que ocurran. Tradicionalmente, la estrategia predominante para la prevención del hurto a estos lugares ha involucrado patrullajes policiales regulares y análisis retrospectivos de datos de crímenes. Sin embargo, estas técnicas, aunque útiles, a menudo resultan insuficientes para identificar patrones complejos y predecir de manera precisa dónde y cuándo podrían ocurrir futuros robos. La falta de precisión en estos métodos tradicionales significa que las fuerzas de seguridad pueden no estar posicionadas de manera óptima para responder o prevenir incidentes delictivos, lo que potencialmente deja a áreas vulnerables expuestas a actividades delictivas recurrentes.

Antecedentes del problema

El aumento de robos en las zonas urbanas de Cundinamarca está relacionado con factores socioeconómicos, geográficos, culturales y tecnológicos. En el contexto sectorial, la rápida urbanización no planificada ha generado áreas de alta vulnerabilidad socioeconómica, lo que incrementa la criminalidad. El Banco Mundial (2022) indica que las ciudades en rápido crecimiento carecen de servicios adecuados e infraestructura, contribuyendo al aumento del crimen. Además, las desigualdades económicas entre regiones urbanas y rurales también fomentan la delincuencia. Muggah y García (2019) señalan que las disparidades económicas y sociales promueven tasas más altas de delitos, incluido el robo. El PNUD (2011) también identifica que las ciudades colombianas con mayores niveles de desigualdad tienden a sufrir más delitos violentos.

Culturalmente, el aumento del uso de dispositivos móviles y la economía digital ha hecho que estos sean objetivos comunes para los robos. La UNODC (2020) destaca que el crecimiento en el uso de teléfonos inteligentes ha llevado a un aumento de estos delitos. Según el DANE (2019), la percepción de inseguridad en Colombia ha aumentado debido al robo frecuente de dispositivos móviles en lugares públicos.

Desde una perspectiva tecnológica y de conocimiento, la adopción limitada de métodos avanzados de análisis de datos para la prevención del delito es un factor significativo. Aunque los modelos de *machine learning* han demostrado ser efectivos en predecir actividades delictivas en otras regiones, su implementación en Colombia es incipiente (Johnson & Bogomolov, 2019).

Descripción del problema

Este estudio busca abordar la brecha existente en el uso de tecnologías avanzadas para la prevención del delito en Cundinamarca, con un enfoque particular en la aplicación de modelos de *machine learning* para detectar y predecir patrones de robo en áreas urbanas. La hipótesis central de este trabajo es que mediante la aplicación de técnicas de *machine learning,* es posible no solo mejorar la detección de patrones de hurto a locales comerciales y viviendas , sino también predecir con mayor precisión las áreas más probables de futuros incidentes, lo que permitiría a las fuerzas de seguridad optimizar la asignación de recursos y mejorar sus estrategias de prevención. Para alcanzar este objetivo, se propone un enfoque metodológico que combine datos históricos de incidentes de robo con una variedad de variables contextuales, utilizando modelos de *machine learning* para analizar estos datos y generar predicciones sobre posibles futuros robos. Este enfoque no solo pretende identificar las áreas y los momentos del día más vulnerables a los robos, sino también proporcionar un sistema de alerta temprana que pueda ser utilizado por las autoridades para prevenir proactivamente los incidentes delictivos. Al final, se espera que los resultados de este estudio no solo contribuyan a una mejor comprensión de los patrones de robo en Colombia, sino que también proporcionen una base sólida para el desarrollo de políticas públicas más efectivas y basadas en datos para la prevención del delito.

A partir de los aspectos anteriores se plantea la siguiente pregunta de investigación:

¿De qué manera los modelos de *machine learning* pueden ayudar a identificar y predecir las áreas más propensas a futuros hurtos de locales comerciales y viviendas en Cundinamarca, considerando patrones históricos y datos contextuales?

Objetivos

Objetivo general.

Desarrollar modelos de *machine learning* para analizar datos históricos de robos en Cundinamarca con el fin de identificar patrones de criminalidad y predecir áreas de alto riesgo.

Objetivos específicos.

1. Analizar datos históricos sobre incidentes de robo en Cundinamarca para asegurar la calidad y relevancia de la información utilizada en la investigación.
2. Aplicar técnicas de *machine learning* para identificar patrones y tendencias en los datos históricos de robos, incluyendo la detección de áreas y horarios con mayor incidencia delictiva.
3. Desarrollar un modelo predictivo basado en *machine learning* que pueda prever las áreas y momentos con mayor probabilidad de futuros robos en los municipios de Cundinamarca.
4. Proponer un plan de implementación para el modelo predictivo usando indicadores, metodologías y procesos claves para el mejoramiento continuo del mismo.

Justificación

Este proyecto es clave para enfrentar el problema de los hurtos a locales comerciales y viviendas en Cundinamarca, utilizando tecnologías avanzadas como el *machine learning*. A pesar de que en 2024 se ha registrado una disminución significativa en los hurtos, con un 18% en las viviendas, un 28% en personas y un 25% en vehículos, sigue siendo una preocupación relevante en municipios como Soacha y Chía, donde las autoridades han realizado operativos importantes contra robos en comercios.

El uso de *machine learning* para analizar datos históricos permite predecir incidentes con mayor precisión, lo que facilita una distribución más efectiva de los recursos de seguridad y una mayor presencia policial en las áreas con mayor riesgo. Este enfoque complementa las estrategias actuales, que han logrado resultados positivos, según informa la Gobernación de Cundinamarca en su informe *En el primer trimestre de 2024 disminuyó el hurto en Cundinamarca* presentando una cifra de captura de más de 7,300 delincuentes en lo que va del año, mejorando así la seguridad y calidad de vida de los ciudadanos.

En resumen, este proyecto representa un avance importante hacia una gestión de la seguridad pública más eficiente, que, al integrarse con los esfuerzos en curso, ayudará a reducir aún más la criminalidad en Cundinamarca.

Marco Teórico

En el contexto urbano contemporáneo, la seguridad es una preocupación crucial tanto para las autoridades como para los ciudadanos. Los hurtos en locales comerciales y viviendas son delitos que afectan gravemente la calidad de vida y la economía de las regiones. La implementación de tecnologías avanzadas, como el *machine learning*, ofrece una oportunidad significativa para mejorar las estrategias de prevención y respuesta ante estos delitos.

Estado del Arte

Los modelos predictivos son herramientas valiosas para prever dónde podrían ocurrir delitos, al analizar datos históricos y variables relevantes. Según Santos et al. (2017) y Rodríguez y Pérez (2019), esta capacidad no solo identifica áreas con mayor riesgo de hurto, sino que también proporciona detalles sobre los patrones delictivos, facilitando así una respuesta más precisa por parte de las autoridades.

El análisis espacial complementa esta información al mostrar cómo los delitos están distribuidos y cómo los factores espaciales influyen en su ocurrencia. Gorr y Harries (2004) y Mendoza y López (2020) destacan que esta comprensión es fundamental para identificar zonas con alta incidencia y patrones específicos, lo que resulta crucial para la asignación eficiente de recursos y la implementación de medidas preventivas.

Además, los sistemas de alerta temprana, descritos por Eck et al. (2005) y adaptados por Gutiérrez y Martínez (2021), utilizan datos históricos y socioeconómicos para prever aumentos en la actividad delictiva. Estos sistemas permiten detectar patrones emergentes, lo que capacita a las autoridades para actuar proactivamente antes de que los delitos se materialicen, mejorando la respuesta ante posibles incrementos en la criminalidad.

En conjunto, estos enfoques no solo proporcionan una comprensión más completa de los patrones de hurto, sino que también ayudan a desarrollar estrategias de seguridad más eficaces. La adaptación de estos modelos y sistemas a las condiciones locales de Cundinamarca mejora la precisión de las predicciones, optimiza la gestión de recursos y permite una respuesta más efectiva ante las amenazas emergentes, creando así un marco integral para abordar la seguridad pública.

Aplicaciones de Machine Learning (ML) en la seguridad publica

Modelos predictivos

Los modelos predictivos utilizan algoritmos de ML, como la Regresión Logística, Máquinas de Soporte Vectorial (SVM) y Redes Neuronales, para prever la ocurrencia de delitos. Estos modelos se basan en datos históricos y variables relevantes para calcular la probabilidad de futuros incidentes delictivos. Santos et al. (2017) destacan la eficacia de los modelos de ML en la predicción de crímenes mediante técnicas avanzadas de minería de datos. Además, Rodríguez y Pérez (2019) investigan cómo los modelos predictivos de ML pueden ser adaptados al contexto latinoamericano, destacando su capacidad para mejorar la precisión en la predicción de delitos. En América Latina, los desafíos incluyen la variabilidad en la calidad y disponibilidad de datos, así como las diferencias en las características

socioeconómicas y culturales de las regiones. Los autores subrayan la importancia de ajustar los modelos predictivos para tener en cuenta estas diferencias, lo que puede incluir la incorporación de variables específicas regionales y la adaptación de los algoritmos a las particularidades locales.

Por ejemplo, en el contexto colombiano, donde la violencia y el crimen pueden estar influenciados por factores como el conflicto armado o el narcotráfico, es crucial incorporar estos elementos en los modelos predictivos para obtener resultados más precisos y útiles. El uso de técnicas como la regresión logística y las redes neuronales ha mostrado ser eficaz en la predicción de delitos en este entorno, permitiendo a las autoridades enfocar sus esfuerzos en las áreas de mayor riesgo.

Análisis Espacial

El análisis espacial de datos delictivos, que utiliza técnicas como el Análisis de Espacio-Tiempo y los patrones espaciales (Moran's I, K-function), permite identificar áreas con alta incidencia delictiva. Este enfoque ayuda a las autoridades a concentrar sus recursos en las zonas más afectadas. Investigaciones han demostrado que la combinación de modelos espaciales con técnicas de ML puede mejorar significativamente la precisión de las predicciones de delitos (Gorr & Harries, 2004).

Como elementos relacionados a estos, los autores proponen dos como los principales:

Análisis de Espacio-Tiempo

Esta técnica examina cómo los delitos se distribuyen no solo en el espacio, sino también en el tiempo. Permite identificar patrones temporales en la ocurrencia de crímenes y cómo estos patrones se relacionan con factores

espaciales. Por ejemplo, el análisis de espacio-tiempo puede revelar que ciertos delitos son más frecuentes en ciertas épocas del año o en días específicos de la semana (Gorr & Harries, 2004).

Patrones Espaciales

Métodos como el Índice de Moran (Moran's I) y la Función K son utilizados para identificar la aglomeración espacial de crímenes. El Índice de Moran mide la autocorrelación espacial, indicando si las áreas con alta incidencia delictiva están agrupadas o dispersas. La Función K ayuda a entender la densidad de crímenes en diferentes áreas, facilitando la identificación de "*hot spots*" o zonas de alta incidencia (Gorr & Harries, 2004).

Un claro ejemplo de lo anterior es lo que muestran los autores Laghari y Bhayo (2024), los cuales, en su investigación sobre el robo de energía, implementan metodologías de ML para la identificación de patrones;

Los autores mencionados emplean Máquinas de Soporte Vectorial para analizar datos de consumo de electricidad. Utilizan un conjunto de datos que incluye patrones de consumo normales y anómalos para entrenar el modelo de SVM, con el objetivo de clasificar los patrones de consumo y detectar irregularidades que puedan sugerir robos de electricidad. (Abro, S. A., Hua, L. G., Laghari, J. A., Bhayo, M. A., & Memon, A. A, 2024).

Sistema de alerta temprana / Identificación de Patrones

Los sistemas de alerta temprana que combinan datos históricos de crímenes con información socioeconómica y demográfica ayudan a anticipar posibles incrementos en la actividad delictiva. Eck et al. (2005) discuten cómo estos sistemas pueden alertar sobre patrones emergentes y mejorar la capacidad de respuesta. En el contexto latinoamericano, Gutiérrez y

Martínez (2021) exploran cómo adaptar estos sistemas a las características específicas de la región para maximizar su efectividad.

Se encuentran dos herramientas importantes:

Combinación de datos: Eck (2005) define la combinación de datos como: *Los sistemas de alerta temprana utilizan una variedad de datos, incluyendo registros históricos de crímenes, datos socioeconómicos y demográficos, y en algunos casos, datos en tiempo real de sensores y cámaras de seguridad. La integración de estos datos ayuda a identificar patrones y predecir posibles aumentos en la actividad delictiva* (Eck et al., 2005).Tomando como referencia la calidad de la información como lo mencionan los autores Junde Chen, Y.A. Nanehkaran, Weirong Chen, Yajun Liu, Defu Zhang, en el libro *Data-Driven Machine-Learning Theft, 2019 Detection*, la calidad y cantidad de los datos es de vital importancia, dado que sin estos no se tendría pistas ni caminos para el desarrollo de modelos predictivos y sus distintas herramientas como las regresiones, clasificaciones, clustering y redes neuronales, además de demostrar su gran versatilidad en la aplicación en distintos campos, tales como el retail, el fraude o sectores públicos.

En el libro *Theft Detection and Monitoring System Using Machine Learning*, se exploran varias técnicas de *machine learning*, como algoritmos de clasificación y detección de anomalías, para diseñar sistemas que puedan identificar comportamientos sospechosos y patrones de robo. La versatilidad de las herramientas de ML, ayudan a que la variedad de los datos permita percibir de formas variadas de un solo problema, depende de cómo se desarrolle, es la forma en que veremos la conclusión. (J., Bangroo, A., Garg, S., Nalini, N., Nagaraj, H. C., Patnaik, L. M., Hamsavath, P. N., & Shetty, N. R, 2021)

Teorías y modelos

Teoría de la ventana rota

La Teoría de las Ventanas Rotos, desarrollada por Wilson y Kelling (1982), sugiere que el desorden en un área puede fomentar la criminalidad. Esta teoría puede integrarse en modelos de ML que utilicen variables relacionadas con el estado físico de las áreas. Rodríguez y Pérez (2019) aplican esta teoría en el contexto latinoamericano, demostrando cómo la presencia de señales visibles de desorden puede ser un factor relevante en la predicción de delitos.

Teoría del crimen oportunista

La Teoría del Crimen Oportunista, propuesta por Clarke (1992), se basa en la idea de que los delitos ocurren cuando existen oportunidades favorables y falta de vigilancia. Modelos de ML pueden incorporar datos sobre medidas de seguridad y vigilancia para ajustar sus predicciones. Mendoza y López (2020) discuten cómo esta teoría puede ser aplicada en América Latina, considerando las características locales y la infraestructura de seguridad existente.

Marco Legal

Se logró identificar algunas de las normativas que aplicarían dentro del presente trabajo, esto debido a que cualquier desarrollo se debe regir bajo los límites sociales.

Legislación Nacional y Regional

Ley 1266 de 2008: Regula la protección de datos personales en el ámbito financiero, estableciendo principios para la recolección y uso de datos (Ley 1266 de 2008).

Ley 1581 de 2012: Establece disposiciones generales para la protección de datos personales en Colombia, incluyendo la autorización del titular de los datos y los derechos relacionados con el manejo de su información (Ley 1581 de 2012).

Código Penal Colombiano: Incluye disposiciones sobre delitos y medidas de seguridad, regulando la prevención y persecución de crímenes (Código Penal Colombiano).

Consideraciones Éticas

Privacidad: Es esencial proteger la privacidad de los individuos mediante la implementación de medidas de seguridad en el tratamiento de datos personales. (Ethics of Artificial Intelligence and Robotic, 2020)

Sesgo Algorítmico: Los modelos de ML pueden reflejar sesgos presentes en los datos, lo que puede llevar a decisiones injustas. Es importante evaluar y mitigar estos sesgos para garantizar la equidad. (The Ethics of Machine Learning: An Overview, 2018)

Transparencia: Los procesos de toma de decisiones basados en ML deben ser transparentes para asegurar la confianza del público en las herramientas utilizadas. (AI Ethics: A Guide to the Principles and Practices, 2018)

En resumen, usar técnicas avanzadas de *machine learning* para detectar y predecir hurto en viviendas y locales comerciales en Cundinamarca es clave para mejorar la seguridad en la región. Los estudios muestran que identificar y prever patrones delictivos mediante modelos predictivos, análisis espacial y sistemas de alerta temprana es fundamental para hacer más efectivas las estrategias de seguridad.

Metodología

Enfoque

Con el fin de aplicar el modelo de la manera de predicción de patrones en las casas y locales de Cundinamarca, se debe establecer una estructura clara, en primer lugar, tener en

cuenta el objetivo principal; "*Desarrollar modelos de machine learning para analizar datos históricos de robos en Cundinamarca y Bogotá con el fin de identificar patrones de criminalidad y predecir áreas de alto riesgo"* es importante establecer las formas que se usan para predecir estos patrones, luego, se debe recopilar datos históricos sobre robos, además de ser posible buscar variables externas que puedan influir en el modelo.

Posteriormente, se realiza un preprocesamiento de los datos, limpiándolos y transformándolos para facilitar el análisis. A continuación, se lleva a cabo un análisis exploratorio para identificar patrones y tendencias. Una vez identificados, se selecciona y entrena un modelo de ML adecuado, como árboles de decisión o bosques aleatorios, validando su rendimiento.

Finalmente, se implementa el modelo y se monitorea su eficacia, utilizando los resultados para mejorar la asignación de recursos y la planificación de medidas preventivas en la lucha contra el delito.

Alcance

Temporalidad: El análisis abarcará datos desde 2010 hasta 2024, permitiendo identificar tendencias a largo plazo y patrones estacionales en la ocurrencia de robos.

Localización: Cundinamarca

Tipología de los delitos: Robos y hurtos a casas y locales comerciales

Tipo de Investigación

Cuantitativa: Se centrará en el análisis de datos numéricos y en la construcción de modelos predictivos a partir de datos históricos.

Etapas de la investigación

Etapa 1: Análisis de datos históricos sobre incidentes de robo en Cundinamarca

En esta etapa se validará los datos claves, las proporciones y las formas en que se visualizan los datos, además de dar una apertura inicial al contexto en el cual se realizara el desarrollo propuesto.

Definición de Variables

El conjunto de datos sobre robos a viviendas y locales comerciales en Cundinamarca, que abarca desde 2014 hasta 2024, es una fuente fundamental para el análisis de la criminalidad en la región. Este *dataset* permite examinar patrones temporales y espaciales, lo que facilita la identificación de puntos críticos donde se concentran los robos. Además, ofrece la posibilidad de correlacionar estos incidentes con otros factores, como el nivel socioeconómico de las zonas afectadas, las políticas de seguridad implementadas y la presencia de infraestructura urbana.

Al estudiar las tendencias de este período de tiempo, se pueden identificar posibles cambios en las modalidades delictivas, como el incremento de robos en áreas urbanas o comerciales en crecimiento. Esto brinda a las autoridades la oportunidad de anticipar situaciones futuras, ajustando sus estrategias de prevención. Asimismo, estos datos permiten el desarrollo de modelos predictivos mediante *machine learning*, lo que ayudaría a optimizar los recursos de seguridad, concentrando esfuerzos en las áreas más vulnerables y de alto riesgo. El análisis exhaustivo de estos datos puede contribuir significativamente a mejorar la seguridad pública en la región.

Variables

La base de datos incluye una variedad de variables clave para el análisis, como:

Departamento

Esta variable indica el departamento donde se registró el delito. Al agrupar la información geográficamente, se puede observar la distribución de la criminalidad en diferentes regiones, para nuestro caso específico Cundinamarca.

Municipio

Especifica el municipio en el que ocurrió el robo. Esta información es esencial para realizar un análisis más detallado a nivel local, permitiendo a las autoridades comprender mejor las dinámicas delictivas en contextos más restringidos y adaptar estrategias de seguridad a las particularidades de cada municipio como lo es Soacha, Facatativá, etc.

Tiene los siguientes valores únicos;

```
Valores únicos en la columna 'municipio':
['BOGOTÁ D.C. (CT)' 'SOACHA' 'FUNZA' 'CHÍA' 'TIBACUY' 'GIRARDOT'
 'LA CALERA' 'MOSQUERA' 'FACHO' 'VILLA DE SAN DIEGO DE UBATE' 'VILLETA'
 'FACATATIVÁ' 'NIMAIMA' 'TOCANCIPÁ' 'GUASCA' 'FUSAGASUGÁ' 'ZIPAQUIRÁ'
 'UBALÁ' 'TABIO' 'AGUA DE DIOS' 'GACHETÁ' 'EL PEÑÓN' 'SUESCA' 'EL COLEGIO'
 'GAMA' 'MADRID' 'SIBATÉ' 'GUAYABETAL' 'NEMOCÓN' 'TENJO' 'CAJICÁ' 'UBAQUE'
 'GUADUAS' 'LA PALMA' 'SILVANIA' 'TENA' 'CHIPAQUE' 'CHOACHÍ' 'LA VEGA'
 'SAN JUAN DE RÍO SECO' 'PUERTO SALGAR' 'CAQUEZA' 'GACHANCIPÁ' 'ALBÁN'
 'FOMEQUE' 'COTA' 'SOPÓ' 'CHAGUANÍ' 'QUETAME' 'CHOCONTÁ'
 'SAN ANTONIO DEL TEQUENDAMA' 'FOSCA' 'BOJACÁ' 'SAN FRANCISCO'
 'VILLAPINZÓN' 'TOCAIMA' 'ANAPOIMA' 'LA MESA' 'SESQUILÉ' 'COGUA' 'SUSA'
 'QUEBRADANEGRA' 'PARATEBUENO' 'ARBELÁEZ' 'MACHETA' 'JERUSALÉN' 'ANOLAIMA'
 'VIOTÁ' 'LENGUAZAQUE' 'YACOPÍ' 'MEDINA' 'EL ROSAL' 'RICAURTE' 'NILO'
 'CAPARRAPÍ' 'SUPATÁ' 'CUCUNUBÁ' 'GUACHETÁ' 'SIMIJACA' 'SUBACHOQUE'
 'ZIPACÓN' 'ÚTICA' 'MANTA' 'CACHIPAY' 'SAN BERNARDO' 'GRANADA' 'PANDI'
 'SASAIMA' 'TAUSA' 'CARMEN DE CARUPA' 'GUAYABAL DE SIQUIMA' 'BITUIMA'
 'VIANÍ' 'SAN CAYETANO' 'APULO' 'PASCA' 'TOPAIPÍ' 'GUTIÉRREZ' 'CABRERA'
 'FÚQUENE' 'SUTATAUSA' 'UNE' 'VILLAGÓMEZ' 'JUNÍN' 'QUIPILE' 'GUATAVITA'
 'LA PEÑA' 'GUATAQUÍ' 'VERGARA' 'NOCAIMA' 'GACHALA' 'NARIÑO' 'BELTRÁN'
 'VENECIA' 'TIBIRITA' 'PULÍ' 'PAIME']
```

Figura 1 – Valores únicos [Municipios]

Los cuales, dentro de las más de 100000 denuncias, la proporción se distribuye en más del 70% en menos de 10 municipios.

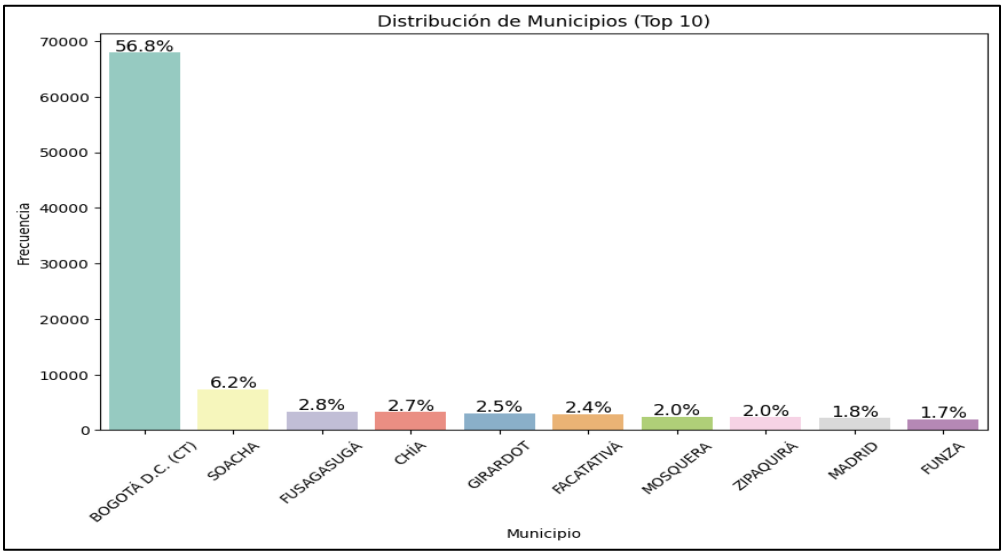

Figura 2 - Distribución de denuncias en el Top 10 de los municipios

Armas Medios

Indica el tipo de armas o métodos utilizados por los delincuentes durante el hurto, como armas de fuego, cuchillos o la ausencia de armas. Esta variable ayuda a evaluar el nivel de violencia asociado con los delitos y puede ser clave para desarrollar medidas de prevención que aborden las tácticas empleadas por los delincuentes.

Estos serían los valores únicos que se presentan en la columna;

```
Valores únicos en la columna 'armas_medios':
['CONTUNDENTES' 'SIN EMPLEO DE ARMAS' 'ARMA BLANCA / CORTOPUNZANTE'
 'NO REPORTADO' 'ARMA DE FUEGO' 'PUNZANTES' 'CORTANTES' 'ESCOPOLAMINA'
 'OTROS' 'LLAMADA TELEFONICA']
```

Figura 3 – Valores Únicos [Armas_medios]

Estos valores se distribuyen de la siguiente manera;

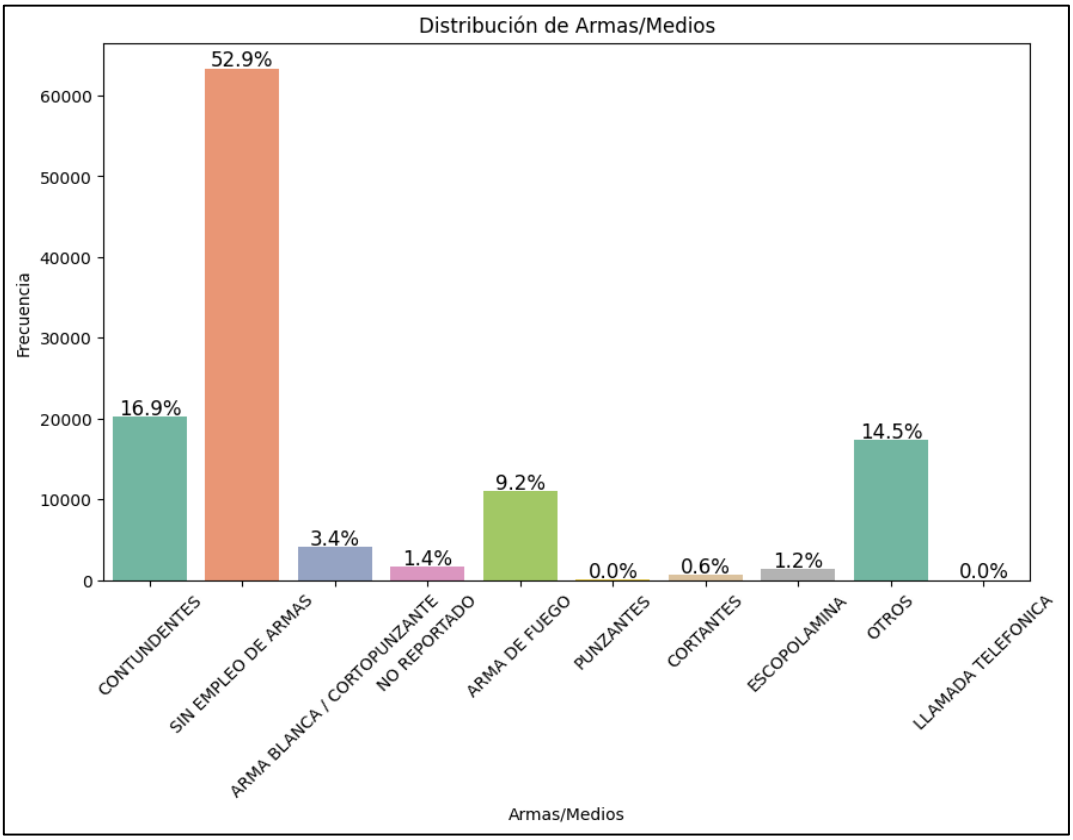

Figura 4 – Distribución de denuncias por tipo de arma usada.

Fecha Hecho

Registra el día en que se produjo el robo, lo que permite realizar análisis temporales para identificar tendencias, como picos de actividad delictiva en ciertas épocas del año, días de la semana o incluso horas del día. Esta información puede ser fundamental para optimizar la asignación de recursos policiales y prevenir futuros delitos.

Para este apartado, se toma en cuenta los meses y años que se evidencia en el *dataset* por lo tanto los valores únicos para esta variable los meses desde enero representado por el numero 1 hasta diciembre representando por el numero 12, en cuanto la distribución de denuncias se obtuvo el siguiente gráfico.

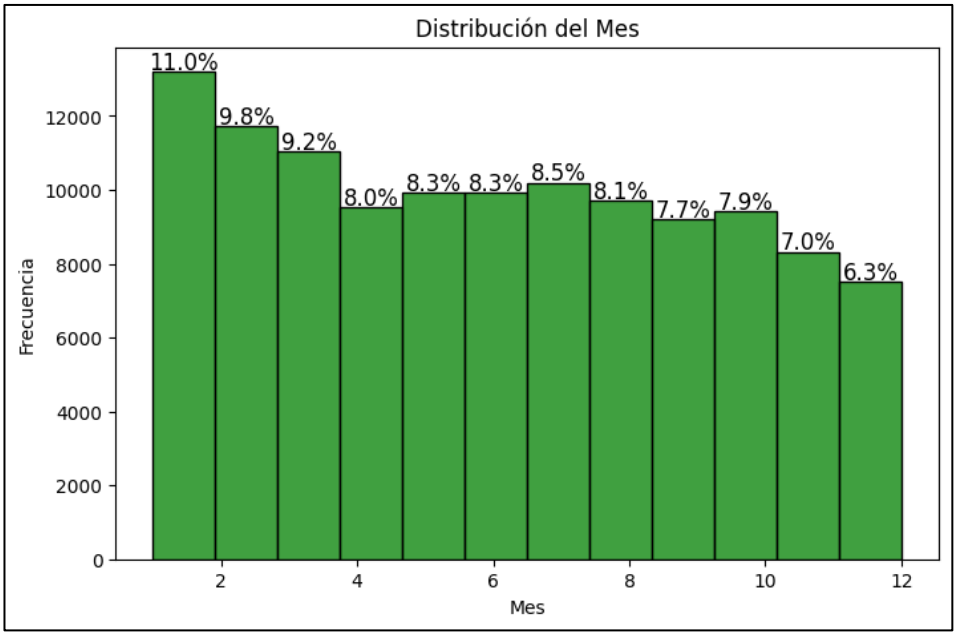

Figura 5 – Distribución de denuncias por mes

Género

Esta variable determina el género de las personas implicadas en el delito, ya sean delincuentes o víctimas. Al analizar cómo se distribuyen los géneros en los casos de robo, se pueden identificar patrones que reflejan las dinámicas de género en la criminalidad, lo que puede ayudar a diseñar campañas de prevención más inclusivas y efectivas.

Esta variable se compone de los siguientes valores únicos;

```
Valores únicos en la columna 'genero':
['FEMENINO' 'MASCULINO' 'NO REPORTADO' 'NO APLICA']
```

Figura 6 – Valores únicos [genero]

En cuanto su distribución se encuentra de la siguiente forma:

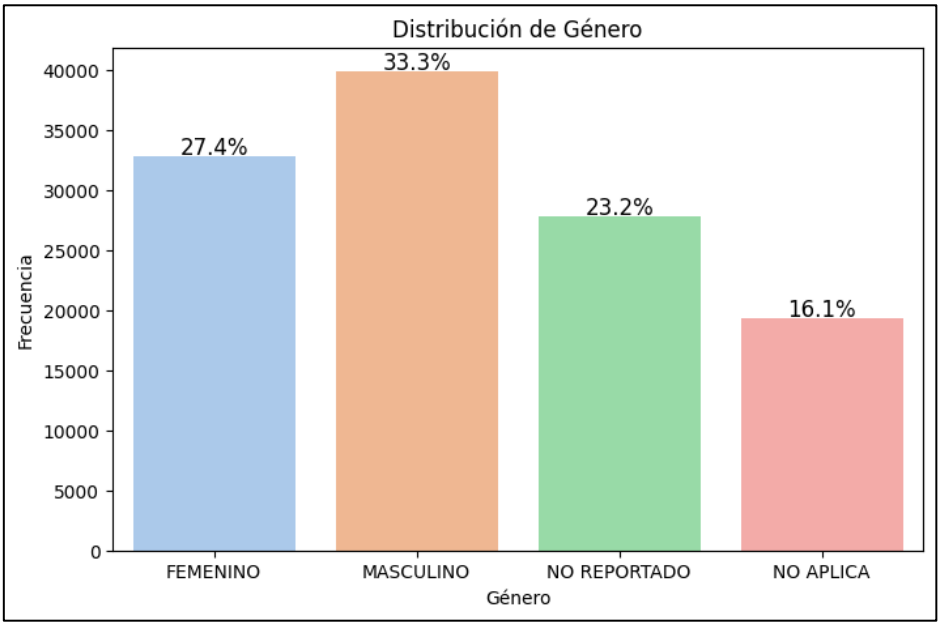

Figura 7 – Distribución de denuncias por Genero

Grupo Etario

Clasifica a los involucrados en diferentes rangos de edad, como menores de 18 años, jóvenes adultos los cuales incluyen a las personas mayores de 18 años hasta los 26 y adultos mayores que se entiende como el rango de edad desde los 27 hasta los 60. Comprender la distribución etaria de los delincuentes y las víctimas es importante para detectar tendencias específicas en la criminalidad y

para desarrollar programas de prevención dirigidos a grupos de edad particulares.

Esta compuesta por los siguientes valores únicos;

```
Valores únicos en la columna 'grupo_etario':
['ADULTOS' 'NO REPORTADO' 'ADOLESCENTES' 'MENORES' 'NO APLICA' 'OTROS']
```

Figura 8 – Valor únicos [grupo_etario]

Cuya distribución se hace de la siguiente forma:

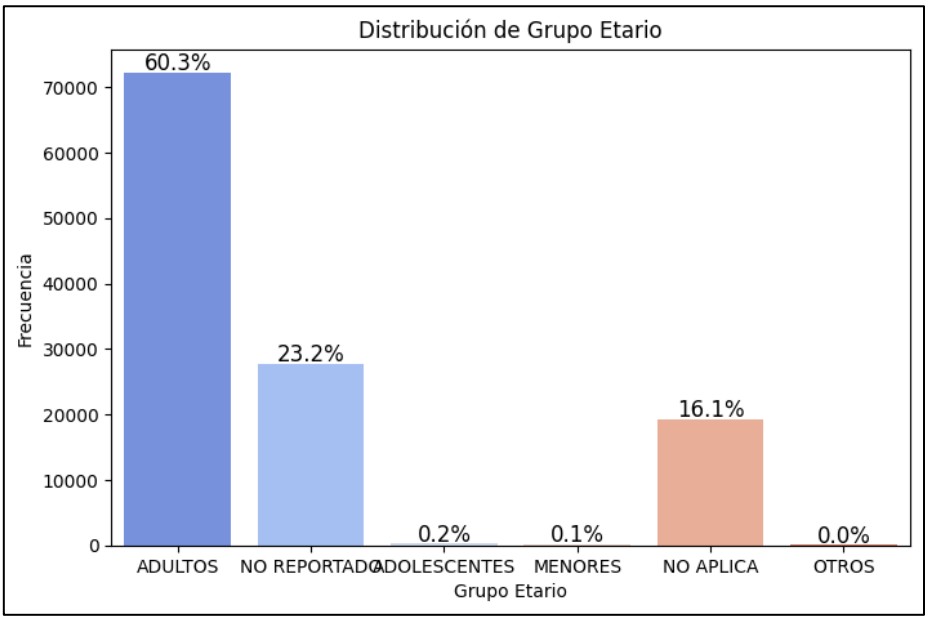

Figura 9 – Distribución de denuncias por Grupo Etario

Tipo de Hurto

Especifica la categoría del robo, como el robo a viviendas, a locales comerciales. Esta variable facilita la identificación de patrones delictivos específicos y permite a las autoridades y a los responsables de la seguridad pública enfocar sus esfuerzos en las áreas más problemáticas.

Se compone de los siguientes valores únicos;

```
Valores únicos en la columna 'tipo_de_hurto':
['HURTO RESIDENCIAS' 'HURTO ENTIDADES COMERCIALES']
```

Figura 10 – Valores únicos [tipo_de_hurto]

Cuya conformación se presenta de la siguiente forma;

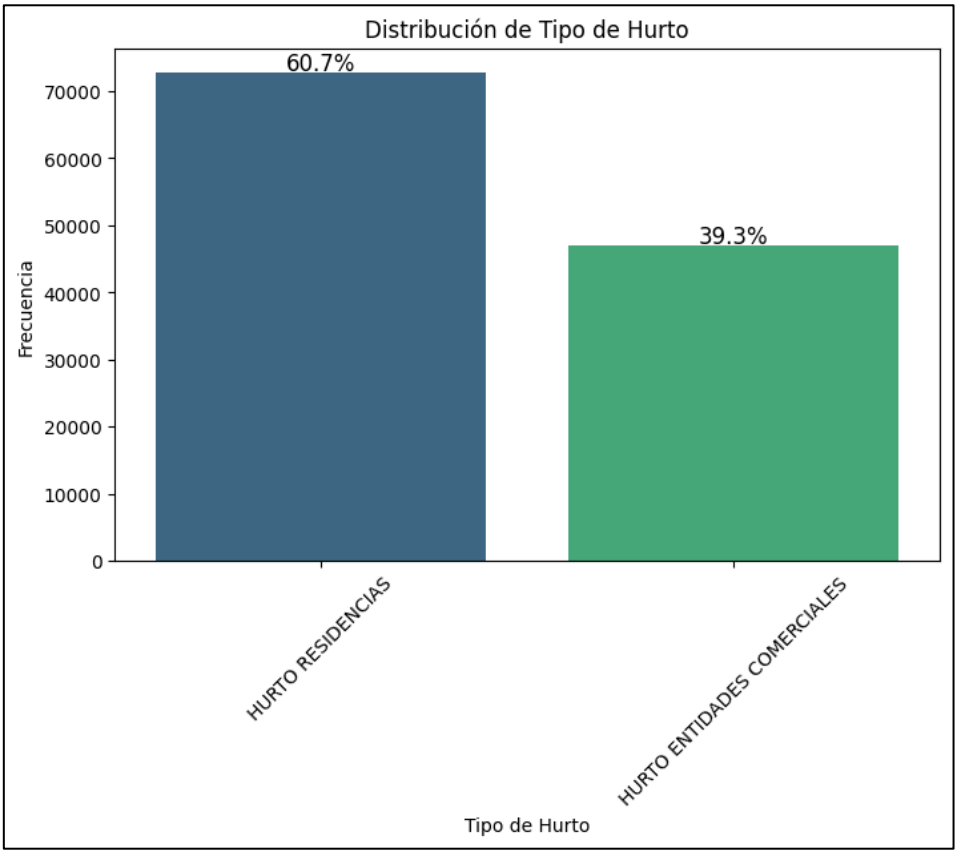

Figura 11 – Distribución de denuncias por Tipo de hurto

Cantidad

Indica la cantidad o el valor economicoestimado de los bienes robados. Esta variable proporciona una medida cuantitativa del impacto económico de los delitos, lo que puede influir en la priorización de recursos para la prevención y la respuesta a la criminalidad.

Métodos para la recolección de información

El conjunto de datos que reúne denuncias de delitos en Cundinamarca, abarcando el periodo de 2014 a 2024 constituye el *dataset* que incluye 119,675 registros, cada uno correspondiente a un caso de personas que han presentado denuncias ante las autoridades. Además, se han contabilizado 291,569 denuncias, lo que sugiere que algunas personas han reportado múltiples incidentes a lo largo del tiempo.

Al seleccionar una muestra de 119,675 registros, se busca garantizar una representación adecuada de las diversas situaciones y contextos en los que ocurren los delitos.

Este conjunto de datos se origina desde las bases de datos público del gobierno de Colombia, por lo tanto la obtención de los datos específicos o particulares se escapa del desarrollo de este documento, por lo tanto dado al contexto se establece que la denuncia es la medida más acertada en la cual se obtiene la información que alimenta la fuente elegida, esta medida puede ser recibida de manera digital o física, ambas contienen la misma cantidad de detalles y son abiertas para las mismas poblaciones.

En cuanto la recolección de información esperada para la ejecución de los modelos, esta se hace mediante uso de líneas de código específicas y la conexión al interfaz de programación de aplicaciones (API´s) publicas ofrecidas por el Gobierno Colombiano.

Se obtendrán los datos entre 2010 y 2024 y de esta manera tener un modelamiento continuo.

En la figura 1 se muestra un ejemplo de lo mencionado anteriormente;

Figura 1. Flujo tradicional de Data

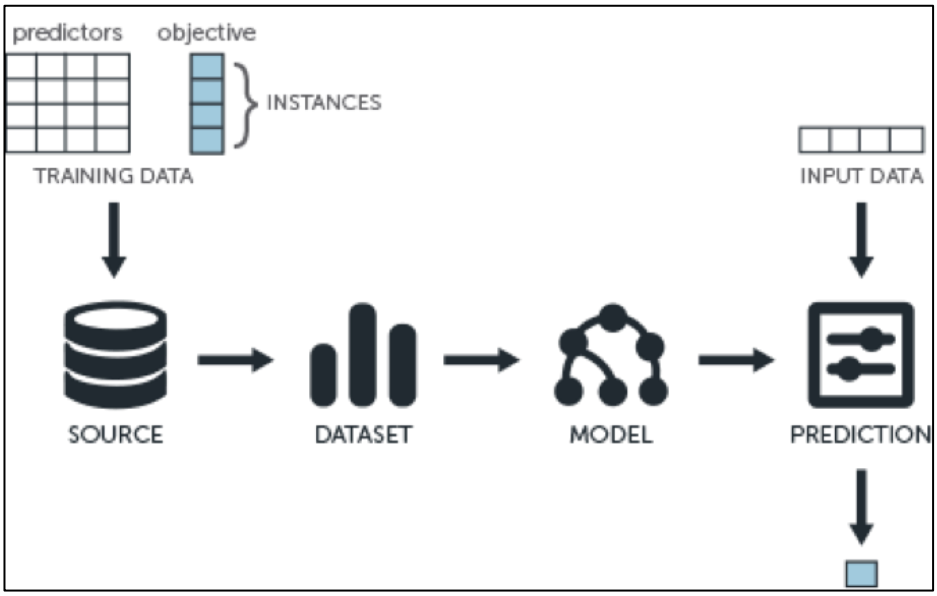

Figura rescatada de: https://www.bbvaapimarket.com/es/mundo-api/apis-y-machine-learning-asi-se-predice-el-exito-de-una-empresa/

Etapa 2: Aplicación de técnicas de *machine learning* para identificar patrones y tendencias en los datos históricos de robos

Teniendo en cuenta que la mayoría de las variables son categóricas, la forma en que se realizaran procesos con estas será en valores de textos nominales, es decir, valores

que sirvan para la distribución de los valores que se logre encontrar, por ejemplo, para las variables de Departamento y Municipio ya que estas son geográficas, ayudaran a distribuir mediante mapas de calor, los sitios con mayor cantidad de robos, en cuanto la unidad de medida, dado a que esta información es consumida desde un banco de datos libres del Gobierno Colombiano, lo más acertado seria establecer las denuncias como la medida predeterminada. En cuanto la cantidad, se tendría un aproximado de 291.569 denuncias, entre las cuales figuran todos los municipios reportados, esto desde el enero del 2010 hasta el mes de septiembre del 2024, actualmente este sigue alimentándose con las denuncias recibidas mediante la conexión a la interfaz de programación de aplicaciones (API) pública.

Para realizar el análisis de los datos en el dataset de denuncias de delitos en Cundinamarca, se pueden aplicar varias técnicas; según se presenta en la tabla 1

Tabla 1 - *Técnicas de análisis de datos*

Técnica	Objetivo	Método
Análisis Descriptivo	Resumir las características básicas de los datos.	Estadísticas descriptivas distribución de frecuencias y visualizaciones como histogramas y gráficos de barras.
Análisis de Tendencias Temporales	Examinar cómo las denuncias varían a lo largo del tiempo.	Series temporales para identificar patrones estacionales y tendencias, usando gráficos de líneas para visualizar la evolución de las denuncias
Modelos Predictivos	Predecir la ocurrencia de delitos futuros.	Regresión Logística
Análisis de Clasificación	Clasificar los delitos en diferentes categorías.	Utilización de técnicas como K-vecinos más cercanos (KNN)

Recuperado de https://openwebinars.net

Etapa 3: Desarrollo de un modelo predictivo basado en *machine learning* que pueda prever las áreas y momentos con mayor probabilidad de futuros robos en los municipios de Cundinamarca.

Dentro del desarrollo de los modelos es de vital importancia entender la consistencia de los datos, por ejemplo para la columna Municipios, dado a que esto se alimenta de manera manual con las denuncias realizadas en la policía, se encontraron varios errores o variantes de un mismo municipios, por ejemplo “Bogota DC” o “Bogotá D.C.” o por ejemplo en la columna armas_medio, se encontraron variantes tales como “NO REPORTA” o “NO REPORTADO” para arreglar esto fue necesario estandarizar estos textos planos, llevando todos los posibles resultados a valores estables.

Como referencia de lo antes mencionado se usaron cadenas de códigos tales como la siguiente;

```
# Validando Columna grupo_etario
df['grupo_etario'] = df['grupo_etario'].replace({'NO REPORTA': 'NO REPORTADO','Otro': 'OTROS'})
✓ 0.1s

valores_unicos_3 = df['grupo_etario'].unique()
print("Valores únicos en la columna 'grupo_etario':")
print(valores_unicos_3)
✓ 0.0s
Valores únicos en la columna 'grupo_etario':
['ADULTOS' 'NO REPORTADO' 'ADOLESCENTES' 'MENORES' 'NO APLICA' 'OTROS']
```

Figura 12 – Referencia de código para estandarizar textos planos.

Con los datos normalizados, hacemos un alistamiento para la ejecución de los códigos de los correspondientes a los modelos, tales alistamientos son la separación de los conjuntos de prueba que para este caso especifico se esta tomando 20% de los datos como conjunto de entrenamiento.

Adicional a esto se crea dos columnas adicionales en el dataset, las cuales corresponden a la separación de los dias, meses y año de la columna fecha_hecho, esto con el fin de facilitar los procesamientos.

Luego de esto, se empieza con el ensamble de los bloques de código para cada uno de los modelos, para este caso en específico se divide en dos segmentos, uno para pronósticos y otro para clasificación.

Tomando referencia de la tabla 1, los modelos a usar son los siguientes;

Pronósticos

Para el correcto procesamiento se debe usar autoencoder dado a que algunas columnas son categóricas, es decir tienen texto plano en sus valores y para la ejecución del código, se necesita que la totalidad del *dataset* sea numérico, por lo tanto, se usa la siguiente cadena de código;

```
# Crear el preprocesador
preprocessor = ColumnTransformer(
    transformers=[
        ('cat', OneHotEncoder(), categorical_cols),  # Categóricas se transforman con OneHotEncoder
        ('num', 'passthrough', numerical_cols)      # Las numéricas no se transforman
    ])
```

Figura 13 – Muestra del autoencoder

Con esto en claro, se realiza el entrenamiento del modelo.

Para la solución de la primera pregunta planteada ¿Cuantos robos tendremos en el siguiente mes en el departamento de Cundinamarca? Se usa regresión lineal, además de realizar las predicciones correspondientes también se calculó el error cuadrático medio (MSE).

Esto mismo se repite en varios procesos variando entre la selección de columnas para el cálculo especifico de cada pregunta, por ejemplo, para la pregunta ¿Que arma se

usará en el próximo robo?, se usa la variante de regresión logística para establecer entre categorías cual sería la más probable que suceda.

Clasificación

Para este caso específico se usa un 30% de los datos como conjunto de prueba, además de usar un encoder distinto para la ejecución como se muestra en la Figura 14

```
le = LabelEncoder()

# Aplicamos el encoding a las columnas categóricas
df['departamento'] = le.fit_transform(df['departamento'])
df['municipio'] = le.fit_transform(df['municipio'])
df['armas_medios'] = le.fit_transform(df['armas_medios'])
df['genero'] = le.fit_transform(df['genero'])
df['grupo_etario'] = le.fit_transform(df['grupo_etario'])
df['tipo_de_hurto'] = le.fit_transform(df['tipo_de_hurto'])
```

Figura 14 – Encoder para la clasificación

Usando el modelo de Random Forest, se realiza el entrenamiento del modelo y se realiza la construcción de los clasificadores.

Etapa 4: Plan de implementación para el modelo predictivo usando indicadores, metodologías y procesos claves para el mejoramiento continuo del mismo.

Este se implementó en cuatro pasos;

a. Conexión e integración de Datos

   Indicadores:

   - Frecuencia de actualización de datos
   - Calidad y completitud de los datos obtenidos
   - Tiempo de respuesta de la API

Metodología:

- Establecer una conexión segura con la API utilizando protocolos como HTTPS y tokens de autenticación.
- Diseñar un proceso ETL (Extract, Transform, Load) automatizado para procesar los datos recibidos.
- Implementar validaciones iniciales para verificar coherencia y precisión de los datos.

b. Desarrollo y validación del modelo

Indicadores:

- Precisión del modelo
- Tasa de error
- Estabilidad en tiempo real

Metodología

- Dividir los datos en conjuntos de entrenamiento, validación y prueba.
- Utilizar algoritmos como regresión logística, árboles de decisión o redes neuronales, según las características del problema.
- Realizar cross-validation para evitar sobreajuste.

c. Despliegue y consumo del modelo

Indicadores:

- Tiempo de respuesta del modelo al procesamiento de una nueva entrada
- Uso de memorias computacionales

Metodología

-

- Implementar el modelo en un servidor escalable o una plataforma en la nube (AWS, Azure, etc.).
- Crear endpoints RESTful para permitir que el modelo sea consumido por otros sistemas o aplicaciones.

d. Mejoramiento continuo

Indicadores:

- Porcentaje de mejora en la precisión del modelo tras cada iteración.
- Velocidad en la detección de datos anómalos o atípicos.

Metodología

- Recopilar retroalimentación constante de los usuarios y clientes del modelo.
- Aplicar análisis de desempeño en períodos regulares.
- Realizar pruebas A/B para evaluar nuevas versiones del modelo.

Análisis y discusión de los resultados

Con el fin de dar contexto y dar solución basada en los objetivos planteados al inicio del documento, se realiza un análisis en tres segmentos;

Resultados del análisis exploratorio;

Dentro de los resultados del análisis exploratorio se encontraron valores interesantes al validar las proporciones, como por ejemplo que 60.3% de las personas que denunciaron se encuentran en un rango de edad desde los 26 hasta los 80 (Véase en la figura 9), lo que poniéndolo en perspectiva es bastante obvia ya que en este rango de edad se encuentran los dueños de locales comerciales y casas.

Otro resultado interesante es que el 52.9% de los casos reportan que el robo fue sin empleo de armas, es decir que se puede especular que la mayoría de estos casos fueron sin violencia y que el modus operandi es el robo cuando los habitantes están dormidos o ausentes, para más información véase la Figura 4.

Entre las denuncias resalta que la mayoría se refieren a los hurtos a residencias que, complementando el anterior hallazgo, se debe presuntamente a banda de "Apartamenteros" o ladrones que aprovechan que las viviendas están vacías o que los habitantes se encuentran dormidos para cometer hurtos sin el uso de la violencia.

En cuanto a la variable de tiempo, se encuentra un patrón establecido donde los días donde se comete el mayor número de hurtos son los días entre semana

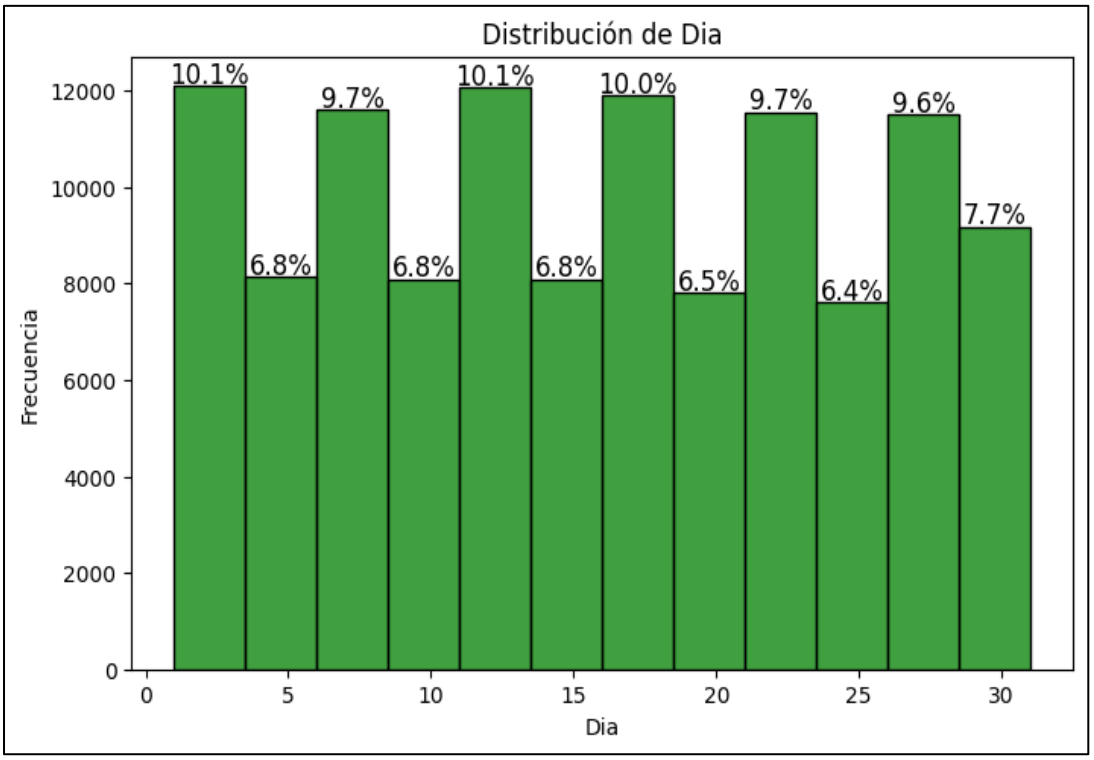

Figura 15 – Distribución de denuncias por día en el mes

Como se ve en la Figura 13, Se encuentran picos entre los días referentes a mitad de semana, por lo tanto, se presume que son los martes, miércoles y jueves los días mas propensos a que suceda un hurto a viviendas o locales comerciales.

En cuanto al mes, se evidencia en la Figura 16 que hay una gran proporción de casos que se comenten iniciando el año, en los meses de enero, febrero y marzo, por lo tanto, en complemento a los hallazgos anteriores, estos meses son de vacaciones, lo que da a lugar a una gran cantidad de viviendas vacías.

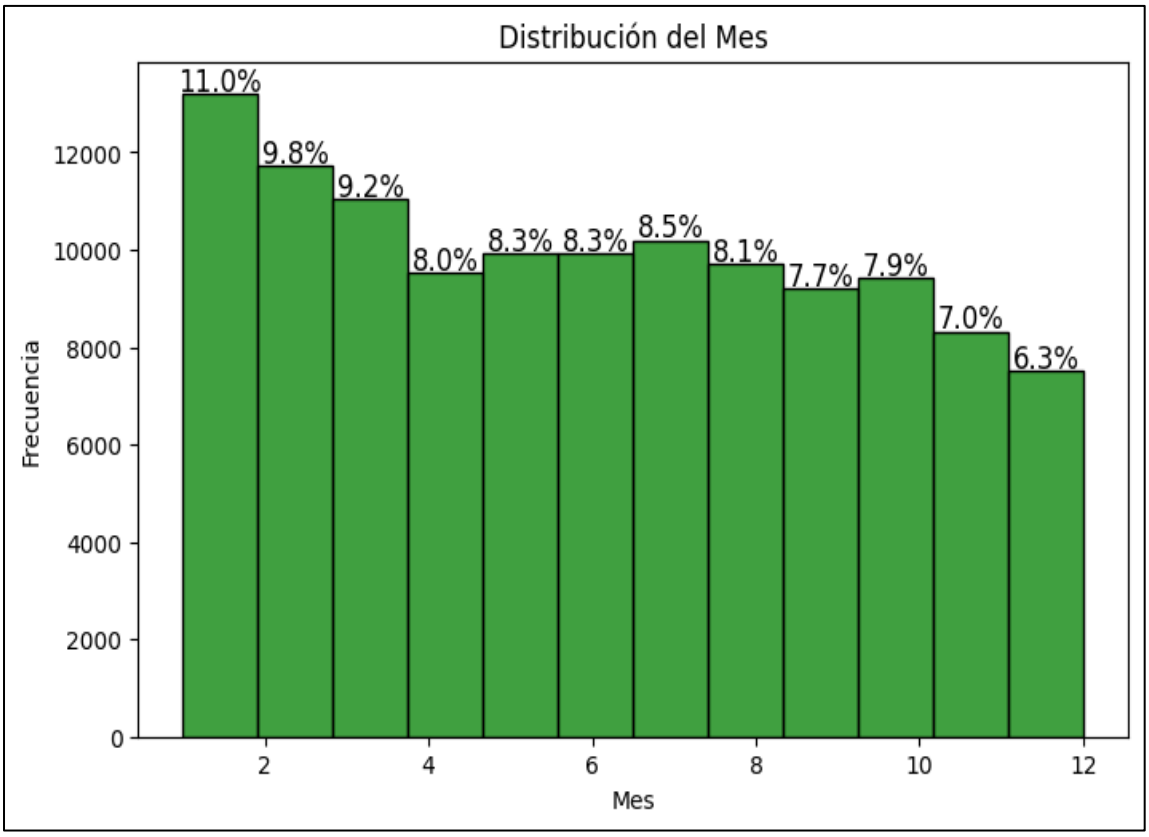

Figura 16 – Distribución de denuncias por mes.

Resultados de modelo de pronóstico (Predicciones)

Se plantearon cinco preguntas con el fin de dar contexto a la solución de los objetivos planteados;

¿Cuántos robos tendremos en el siguiente mes en el departamento de Cundinamarca?

Se plantearon las siguientes condiciones

Departamento: Cundinamarca

Año: 2024

Mes: Octubre

Con estos valores, el pronostico que arroja el modelo construido es;

Error cuadrático medio (MSE): 22.547713441539745

Predicción de robos para el siguiente mes: 1.9155146881256542

Lo que dice que el modelo pronostica aproximadamente 2 robos a locales comerciales o viviendas en el departamento de Cundinamarca en el mes de octubre del 2024, en cuanto al error cuadrado de esta predicción, encontramos 22.4 puntos lo que se refiere a que la predicción se puede desviar 22 puntos del valor real, es decir que la predicción de robos puede ser desde 2 a 24 robos en todo el departamento

.

¿Cuántos robos tendremos en el siguiente día, teniendo en consideración condiciones específicas?

Se plantearon las siguientes condiciones

Departamento: Cundinamarca

Municipio: Bogotá D.C.

Género: Femenino

Grupo Etario: Adulto

Tipo de Hurto: Hurto a Residencias

Dia: 1

Mes: 10

Año: 2024

Con estos valores, el pronóstico que arroja el modelo construido es;

Error cuadrático medio (MSE): 20.22266054164814

Predicción de robos para el siguiente día: 3.4723937649492314

Lo que sugiere que para el 2 de octubre se pronostica 3 robos de viviendas a mujeres adultas en Bogotá, con un error cuadrático medio de 20.2, nos establece un rango entre 3 y 24 robos en las mismas condiciones.

¿Cuántos robos tendremos en los municipios del Departamento de Cundinamarca el próximo mes?

Se plantearon las siguientes condiciones

Departamento: Cundinamarca

Municipio: Bogotá D.C.

Mes: 2

Año: 2024

Con estos valores, el pronóstico que arroja el modelo construido es;

Error cuadrático medio (MSE): 21.32552070414776

Predicción de robos para el siguiente mes: 3.7969686918369234

Lo que sugiere que, para Bogotá, se tendrá un total de 4 casos de hurtos para el mes de octubre, que al tener en cuenta el (MSE) nos establecer que es probable que se presenten entre 4 y 25 casos de hurtos en la ciudad de Bogotá para el mes de octubre.

¿Cuál será el municipio con el siguiente robo?

Se plantearon las siguientes condiciones

Departamento: Cundinamarca

Mes: 2

Año: 2024

Municipio con mayor probabilidad de robos en el siguiente mes:

BOGOTÁ D.C. (CT) - Predicción 3.182877 Casos

El modelo da como resultado a Bogotá como el próximo municipio con mayor probabilidad de que se cometa el siguiente hurto a comercios y hogares.

¿Qué arma se usará en el próximo robo?

Se plantearon las siguientes condiciones

Departamento: Cundinamarca

Mes: 2

Año: 2024

El tipo de arma más probable para el siguiente robo es: SIN EMPLEO DE ARMAS

En base a las observaciones realizadas en el análisis exploratorio, la probabilidad de que el uso de armas sea nulo es más del 50% por lo tanto el resultado que se observa tiene una gran probabilidad de cumplirse.

Resultados de modelo de clasificación (Patrones)

Teniendo en cuenta las siguientes condiciones, se usaron 5 clusters, los cuales se distribuyeron como se muestra en la siguiente Figura

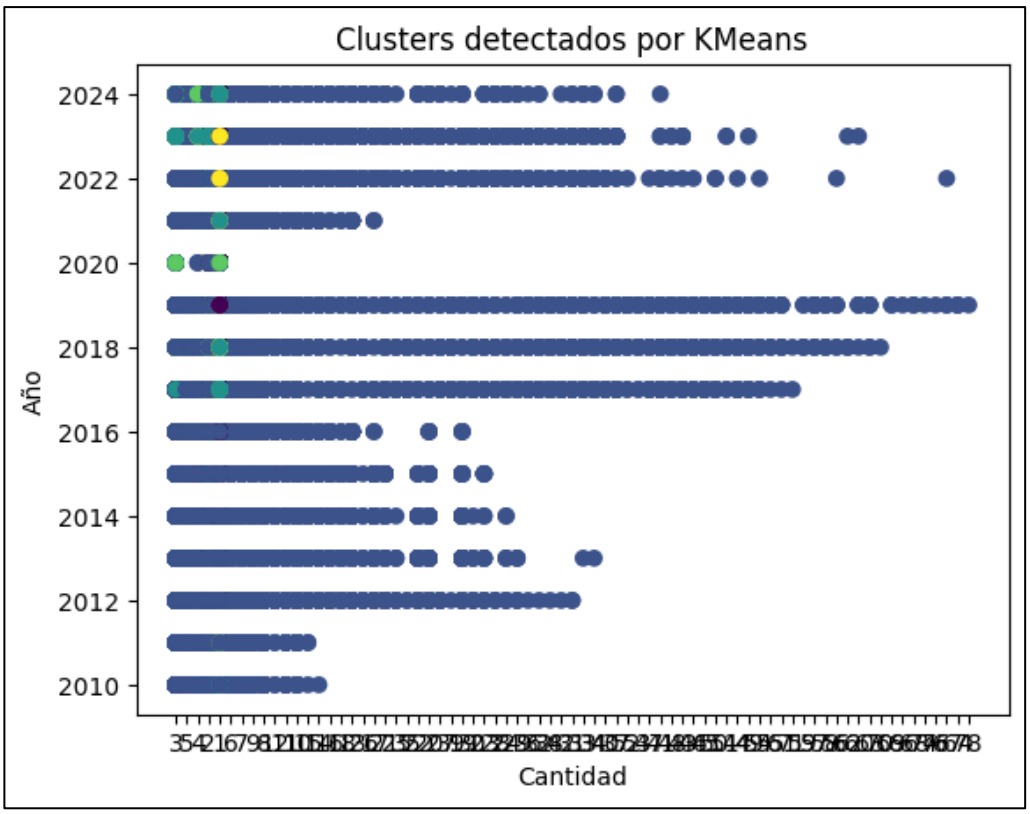

Figura 17 – Clustering denuncias por año

Se observa una mayor densidad de puntos en años como 2016 y 2018, lo que podría reflejar un incremento en la incidencia delictiva durante esos períodos, mientras que, en años recientes, como 2022 y 2024, los datos parecen más dispersos, sugiriendo una mayor variabilidad en los eventos de hurto. Esta dispersión puede ser indicativa de cambios en los patrones delictivos o en la recolección de datos.

Conclusiones

En conclusión, si bien el modelo predictivo desarrollado ofrece una base inicial para estimar eventos como robos en Cundinamarca, las limitaciones evidentes en su precisión reflejadas en el error cuadrático medio (MSE) indican que aún hay margen significativo para mejoras. Estas deficiencias subrayan la importancia de contar con datos de mayor calidad, más completos y actualizados, así como de seleccionar características relevantes que permitan representar mejor las dinámicas del fenómeno estudiado.

El futuro desarrollo del modelo debería centrarse en incorporar técnicas más sofisticadas, como las redes neuronales y métodos de aprendizaje profundo, que permitan identificar patrones complejos dentro de los datos. Asimismo, resulta crucial implementar mecanismos de evaluación y mejora continuos, basados en datos actualizados, para asegurar que las predicciones sean cada vez más precisas y útiles en la toma de decisiones estratégicas.

En este contexto, el trabajo realizado sienta una base importante, pero también resalta la necesidad de seguir avanzando hacia un enfoque más robusto que no solo ofrezca predicciones más precisas, sino que también se convierta en una herramienta efectiva para la planificación estratégica en la seguridad pública y la prevención del delito. Con estas mejoras, el modelo podría convertirse en un recurso clave para la toma de decisiones informadas.

Referencias

Figuras

Bibliografía

1. *En el primer trimestre de 2024 disminuyó el hurto en Cundinamarca*. Gobernación de Cundinamarca. (2024).

El Nogal: Carrera 11 no. 78 - 47 / Centro de contacto: (57-1) 593 6464
Bogotá D.C. - Colombia, Suramérica / www.universidadean.edu.co
UNIVERSIDAD EAN: SNIES 2812, Personería Jurídica Res. n°. 2898 del Minjusticia - 16/05/69

2. Plus Publicación. (2024). *Cifras del Observatorio de Seguridad y Convivencia 2024*.
3. Banco Mundial. (2022). Urbanización y desarrollo económico en América Latina.
4. Departamento Administrativo Nacional de Estadística (DANE). (2019). Informe sobre seguridad y convivencia en Colombia.
5. Machine learning for crime prediction: Principles and challenges. Johnson, S., & Bogomolov, A. (2019).
6. Urban violence and humanitarian action: Engaging the fragile city. Muggah, R., & García, J. (2019).
7. Oficina de las Naciones Unidas contra la Droga y el Delito (UNODC). (2020). Impacto del crimen urbano en América Latina.
8. Programa de las Naciones Unidas para el Desarrollo (PNUD). (2011). Informe Nacional de Desarrollo Humano en Colombia.
9. Crime Analysis with Crime Mapping. Sage Publications. Chainey, S., Tompson, L., & Uhlig, S. (2014).
10. Predictive Policing: Review of the Data Mining and Machine Learning Techniques. *IEEE Transactions on Knowledge and Data Engineering*, 29(4), 1048-1061. Santos, F., et al. (2017).
11. Crime and Spatial Analysis. *Journal of Quantitative Criminology*, 20(3), 223-254. Gorr, W. L., & Harries, K. D. (2004).
12. Social Change and Crime Rate Trends: A Routine Activity Approach. *American Sociological Review*, 44(4), 588-608. Cohen, L. E., & Felson, M. (1979).
13. Ethics of Artificial Intelligence and Robotics, Stanford Encyclopedia of Philosophy (2020)

14. The Ethics of Machine Learning: An Overview" *C. Dastin (2018)*
15. AI Ethics: A Guide to the Principles and Practices *N. Binns (2018))*
16. Eck, J. E., Chainey, S. P., Cameron, J. R., & Wilson, R. E. (2005). Mapping Crime: Understanding Hot Spots. *National Institute of Justice*.
17. Data-Driven Approaches to Crime Prediction and Analysis. *Journal of Crime and Justice*, 43(2), 123-145. García, J., et al. (2020).
18. Broken Windows: The Police and Neighborhood Safety. *The Atlantic Monthly*, 249(3), 29-38. Wilson, J. Q., & Kelling, G. L. (1982).
19. Situational Crime Prevention: Theory and Practice. *Criminal Justice Press*. Clarke, R. V. (1992).
20. Spatial Data Analysis and Crime Prediction. *International Journal of Geographical Information Science*, 22(4), 485-505. Gorr, W. L., et al. (2008).
21. Ethical Implications of AI and Machine Learning in Law Enforcement. *Technology Review*, 121(6), 45-59. Dastin, J. (2018).
22. “Aplicación de Machine Learning en la Predicción de Delitos en Entornos Urbanos”. *Revista Latinoamericana de Ciencias de la Computación*, 15(2), 57-76. Rodríguez, C., & Pérez, M. (2019).
23. “Análisis de Datos Espaciales y Temporal para la Predicción de Crímenes en América Latina”. *Revista de Ciencias Forenses y Seguridad*, 22(3), 89-105. Mendoza, E., & López, A. (2020).

24. “Desafíos y Oportunidades en el Uso de Machine Learning para la Seguridad Pública en Colombia”. *Revista de Seguridad y Justicia*, 30(1), 112-130. Gutiérrez, J., & Martínez, A. (2021).

25. Ley 1266 de 2008. *Por la cual se dictan disposiciones para la protección de datos personales*.
26. Ley 1581 de 2012. *Por la cual se dictan disposiciones generales para la protección de datos personales*.
27. Código Penal Colombiano. *Ley 599 de 2000*.
28. Abro, Safdar Ali ; Hua, Lyu Guang ; Laghari, Javed Ahmed ; Bhayo, Muhammad Akram ; Memon, Abdul Aziz International journal of electrical and computer engineering (Malacca, Malacca), 2024-04, Vol.14 (2), p.1240
29. Data-Driven Machine-Learning Theft Detection (2023) Junde Chen, Y.A. Nanehkaran, Weirong Chen, Yajun Liu, Defu Zhang,
30. Theft Detection and Monitoring System Using Machine Learning. Arora, J., Bangroo, A., Garg, S., Nalini, N., Nagaraj, H. C., Patnaik, L. M., Hamsavath, P. N., & Shetty, N. R. (2021).
31. OpenWebinars. (n.d.). Análisis de datos: Predictivo, descriptivo y prescriptivo. Recuperado de https://openwebinars.net