Análisis de diferentes tipos de hurtos en la ciudad de Medellín mediante modelos de Machine Learning

Seminario de Analítica y Ciencia de Datos

# Resumen Descriptivo del Proyecto

El proyecto consiste en la realización de un análisis descriptivo y predictivo de distintos tipos de hurtos en la ciudad de Medellín, esto por medio de técnicas de *Machine Learning* y de visualización de datos, con el objetivo de presentar de forma clara y accesible los patrones que se presentan desde enfoque individual, colectivo, espacial y en el tiempo.

Marco Teórico:

## Revisión de Literatura.

Desde una perspectiva internacional, entre los trabajos que analizan datos de delitos en distintos espacios geográficos, se encuentran estudios como el de Amoako (2021) que investiga ubicaciones de robos en la ciudad de Detroit en un periodo de cinco años, para identificar puntos “calientes”, “fríos” y patrones espaciales en dos escalas. Su hipótesis nula es que no hay una diferencia estadísticamente significativa entre la tasa de robos y un patrón aleatorio. Hace uso de estadística espacial, incluyendo las siguientes técnicas: *Average Nearest Neighbor Analysis*, *Global Moran’s I*, *Getis-Ord Gi* y *Local Morans’s I*. Se concluye, entre otras cosas, que la clusterización de robos en ambas escalas no son resultado de un proceso aleatorio, sino de uno sistemático. La autora explica, además, que la identificación de “puntos calientes” de crimen es útil para focalizar, implementar y asignar respuestas adecuadas por parte de las autoridades para la reducción del crimen.

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

Paralelamente, también hay trabajos que se interesan por las características más relevantes de los hechos, como en la generación de modelos que brindan perspectivas útiles para autoridades y ciudadanos: Campo Agredo (2017) por su parte, en un principio buscó construir modelos a partir de *algoritmos de clasificación*, *predicción* y *asociación*, para obtener información de hurtos a personas en la ciudad de Bogotá, sin embargo, en el desarrollo de su trabajo estableció una *función de densidad de probabilidad Gaussiana multivariable* para diferentes modelos dependiendo de las variables a tomar en cuenta en la solución del problema; a la vez que por medio de códigos de Matlab, buscó encontrar la ruta más corta entre dos puntos en un mapa, con lo cual, a la par del modelo de probabilidad Gaussiana, presenta una forma de hallar la ruta “más segura” entre dos puntos de la ciudad.

Ahora bien, tomando en cuenta la perspectiva y uso de análisis predictivo y descriptivo en los estudios relevantes, más específicamente para el caso de la ciudad de Medellín, se destacan los siguientes trabajos, los cuales son los más relevantes al proyecto:

* Arévalo Álvarez & Fernández García (2022) realizaron un análisis basado en un reporte histórico de hurtos a personas obtenido de fuentes gubernamentales para el periodo 2003-2021, con el objetivo de encontrar las áreas con mayor ocurrencia. Los autores, en un primer momento utilizaron un modelo de tipo *K-means* para encontrar un vector de etiquetas a partir de un proceso de validación interna, el cual sirvió como base para el posterior desarrollo de dos modelos analíticos de inteligencia artificial: un *modelo supervisado LGBM*, que estima la probabilidad de ser hurtado basándose en un *vector de etiquetas* y en determinadas características, el cual obtuvo la mayor precisión; y un modelo de *red neuronal convolucional secuencial*, que no superó el nivel esperado de precisión. Igualmente, también realizaron un análisis tomando como referencia zonas turísticas y de mayor comercio, encontrando sectores cuyo nivel de seguridad tuvo una reducción relativa. Además, encontraron que la Comuna 10, seguida de la Comuna 11, es donde se registran los mayores porcentajes históricos de hurtos a personas.
* Arboleda Colorado (2023) por su parte, hace uso de modelos de regresión de *Machine Learning* para predecir casos de hurto a personas en diferentes zonas de la ciudad entre 2018 y 2022, esto a partir de información obtenida de entes gubernamentales y complementada con datos de espacios de relevancia obtenidos de la Api de Google Maps (por ejemplo, bancos, tiendas, estaciones de policía, entre otros) y con datos obtenidos de sitios web. Este trabajo tiene como objetivo estudiar el fenómeno desde la relación de los hechos delictivos con las características urbanas que lo rodean, lo que permite la identificación de características que los propician y definir estrategias de prevención e intervención. El autor hace uso del marco de ciclo de vida de un proceso de analítica y para el despliegue de modelos se basó en la metodología de *Machine Learning Operations* (MLOps). Encontró que al analizar el conteo de casos por semana y barrio por medio de *redes neuronales* y su implementación en *Grafos*, la capacidad predictiva del modelo no fue muy alta, por lo que se sugiere que en futuros trabajos se incluyan variables estáticas asociadas a los barrios. Luego, el autor realiza un análisis para determinar el mejor enfoque de predicción de casos a partir de características espaciales, evaluando inicialmente distintas regresiones usando el algoritmo *XGBoost*, sin embargo, estos modelos no se ajustaron adecuadamente a los datos. Por lo tanto, optó por hacer uso de un algoritmo de *Random Forest*, el cual mostró un mejor ajuste y permitió capturar la complejidad e interacciones no lineales presentes en los datos. Paralelamente, construyó un tablero analítico de visualización a través de la herramienta Tableau.
* Maestre-Gongora et al. (2023) en su trabajo buscan identificar tendencias y patrones de hurtos a personas en Medellín en el periodo 2014-2020, usando datos abiertos del gobierno. Estos hacen uso de la metodología de inteligencia de negocios para el análisis descriptivo y realizan una predicción de la variable de modalidad de robo a partir de *modelos de Aprendizaje Automático*, utilizando las siguientes librerías de Python: *AutoGluon* y *Scikit-learn*. Los autores evidenciaron que el mejor modelo para su proyecto fue el de *ExtraTreesGini* (*Árboles de decisión*), el cual obtuvo una precisión mayor a los demás y un tiempo de entrenamiento relativamente menor. Encontraron, además, que el segundo semestre del año tiene la mayor incidencia de hurtos, donde la mayoría de los hechos ocurren en sitios públicos con un 60% y sin el uso de armas. Adicionalmente, hacen uso de PowerBI Desktop para la visualización de los datos.
* Puentes Bolivar (2023) en su investigación busca presentar los resultados preliminares de un modelo predictivo de hurtos a personas en la ciudad de Medellín, el cual hace uso de *Machine Learning*: marco de referencia *MLForecast,* para realizar pronósticos de series de tiempo, validando distintos modelos. El autor usó los criterios del *Error absoluto medio* (MAE por sus siglas en inglés) y *Raíz del error cuadrado medio* (RMSE por sus siglas en inglés) para evaluar los modelos. Se identificó que el modelo con los mejores resultados de predicción fue el de *Random Forest Regressor*, a partir del cual, según el autor, se logró una clasificación con alta precisión de la cantidad de hurtos que se pueden presentar en futuras fechas. En esta investigación se usaron datos abiertos del gobierno sobre hurtos a personas por modalidades en un periodo de 19 años.
* Muñoz Jaramillo (2021) en su tesis evalúa tres modelos de *Machine Learning* para la predicción de crímenes y “zonas calientes” en la ciudad de Medellín: un *Clasificador de Bosques Aleatorios*, una *Regresión Logística* y una *Máquina de Soporte Vectorial*. Su metodología integra el resultado de estudios previos con el proceso estándar de la industria para la minería de datos (*CRISP-DM* por sus siglas en inglés). Más específicamente, el autor formula modelos para la predicción de “zonas calientes” relacionadas a hurtos a personas en distintas modalidades, a partir de información histórica, a la par de datos de la tasa de desempleo de la ciudad. En este caso, se evalúa el desempeño de estos modelos al compararlos con un *modelo basado en reglas*, el cual se realiza con el fin de determinar si una división en el espacio-tiempo es una “zona caliente”. El autor concluye que la elección del mejor modelo se debe realizar basándose en las necesidades y recursos disponibles de las autoridades. Y evidencia que el modelo con mejores características para la predicción de estas zonas es la *Máquina de Soporte Vectorial*, sin embargo, explica que las estrategias que se deriven de sus resultados pueden implicar una alta utilización de recursos.

## Modelos y Métodos ya existentes.

Para analizar los modelos y metodologías utilizados en estudios previos alrededor de la temática, se tomaron como base los trabajos expuestos en la revisión de literatura, específicamente los que usan datos de la ciudad de Medellín. A continuación, se presentan los análisis:

* **Analítica de datos para hurtos a personas en la ciudad de Medellín a través de modelos de Machine Learning y Deep Learning.** (Arévalo Álvarez & Fernández García, 2022).

**Modelos utilizados:**

* ***K-Means:*** Inicialmente para el análisis se utilizó un modelo de agrupación no supervisada con el objetivo de identificar patrones de hurtos. Al realizarlo se encontraron grupos con características similares para diferentes ubicaciones de la ciudad. Los autores utilizaron específicamente una variante del modelo llamada “*Mini Batch K-Means*” para mejorar la eficiencia de los cálculos, ya que se contaba con un gran volumen de datos. Como herramienta adicional para el modelo se utilizó una ***Red Neuronal Auto Encoder*** para realizar una reducción de las variables, haciendo así un agrupamiento de ellas. Esto mejoró significativamente la calidad de los resultados.
* ***LGBM (Light Gradient Boosting Machine)*:** Este modelo supervisado tuvo un desempeño del 99% de precisión en el estudio. Los autores lo eligieron ya que presenta grandes ventajas en cuanto a la eficiencia computacional y consistencia de los resultados. La eficiencia de este modelo viene de las técnicas que utiliza: muestreo basado en gradientes (*GOSS*) y agrupación exclusiva en características (*EFB*).
* ***Convutional Neural Network (CNN)*:** Con este modelo se posibilitó el análisis de imágenes satelitales de áreas con alta incidencia de hurtos. Luego de realizar los diferentes ajustes y evaluaciones, este modelo solo alcanzó un 45% de precisión.

**Métodos rechazados o con bajo desempeño:**

* ***DBSCAN:*** Dado que este modelo de agrupamiento resulta menos eficiente para el análisis y la dimensionalidad de los datos, los autores decidieron descartarlo, optando en su lugar por el modelo *K-Means* debido a sus ventajas.
* ***Random Forest y Regresión Logística:*** Estos dos modelos supervisados lograron buenos resultados, pero su tiempo de procesamiento y costo computacional fue significativamente mayor que el de *LGBM*. Esto llevó a descartarlos como opción principal.
* **Hurto a personas en la ciudad de Medellín: análisis predictivo de la cantidad de casos en diferentes zonas de la ciudad a partir de modelos de Machine Learning implementando técnicas de MLOps.** (Arboleda Colorado, 2023).

**Modelos utilizados:**

***Redes de Convolución Recurrentes:*** Este modelo es la utilización de la combinación de *Redes Neuronales Convolucionales (CNN)* y *Recurrentes (RNN)* para realizar un análisis temporal de los hurtos en barrios de Medellín. Estas suelen ser usadas por su efectividad para modelar la evolución temporal de un fenómeno. El enfoque utilizado está basado en la idea de que los barrios vecinos tienen patrones de comportamiento similares. La relación entre estos se modela mediante una *matriz de adyacencia*.

Sin embargo, para este trabajo los autores encontraron que la capacidad predictiva de estas no fue muy alta, debido a la alta dispersión en las matrices utilizadas y por la falta de variables asociadas a cada espacio territorial.

* ***Random Forest*:** Al igual que en el estudio anterior, el modelo *Random Forest* también fue una opción para realizar la clasificación de los hurtos en los diferentes barrios. Este modelo fue seleccionado dada su capacidad de dar predicciones precisas para el contexto del problema. Y a diferencia del estudio anterior, este modelo sí se seleccionó, ya que se adaptó de manera efectiva a la complejidad de los datos.

**Métodos considerados, pero con desempeño limitado:**

* ***Regresión Lineal:*** Este método se experimentó ya que se contaba con las variables que este necesita, pero se encontró que no ofrecía el mismo nivel de precisión que los dos modelos anteriores.
* ***XGBoots:*** Este modelo fue probado debido a su popularidad en competiciones de *Machine Learning* y por su comportamiento proveniente de los *Árboles de Decisión*, pero su rendimiento fue superado por el modelo de *Random Forest*.
* **Análisis de datos sobre los hurtos en la ciudad de Medellín desde un Enfoque Descriptivo.** (Maestre-Gongora et al., 2023).

**Métodos utilizados:**

* ***AutoML (AutoGluon)*:** Esta librería facilita el uso automático de modelos de predicción seleccionando el mejor en términos de precisión y tiempo de entrenamiento. Entre los resultados obtenidos destacan:
  1. *Redes Neuronales* (*NeuralNetFastAI*) con una precisión del 77.36%.
  2. *Random Forest* con una precisión del 79.32%.
  3. *LightGBM* con una precisión del 77.32%.
  4. *Árboles de Decisión*, es el modelo con mejor rendimiento, con una precisión del 79.33% y un tiempo de entrenamiento de solo 3.33 segundos.

De nuevo en este estudio las *Redes Neuronales* y el *Random Forest* aparecen con un buen rendimiento en el análisis. Esto demuestra que son buenas opciones para analizar el comportamiento de los datos de la problemática de hurtos.

**Métodos rechazados o con bajo desempeño**

* ***K-Vecinos más Cercanos (KNN):*** Este modelo en la herramienta utilizada tuvo un tiempo de entrenamiento menor, pero su precisión fue considerablemente baja (alrededor del 50%), por lo que no se consideró útil para este caso.
* ***XGBoost y CatBoost:*** En comparación con los modelos seleccionados como destacados, *XGBoots* y *CatBoots* mostraron un rendimiento intermedio en la precisión, que estuvo en torno al 75%.
* **Modelo predictivo de hurtos para la ciudad de Medellín utilizando Machine Learning.** (Puentes Bolivar, 2023).

**Métodos utilizados:**

* ***MLForecast*:** Fue seleccionado como marco de referencia para la predicción de cantidad de hurtos, ya que: facilita la evaluación de la efectividad de los modelos; cuenta con ocho modelos distintos de *Machine Learning* y permite la estimación simultánea de los mismos. Para el trabajo se probaron los siguientes:
  1. *Regresión lineal.*
  2. *LGBMRegressor.*
  3. *XGBRegressor.*
  4. *RandomForestRegressor.*
  5. *Lasso.*
  6. *Ridge.*

En este caso, para este trabajo el modelo con mejores resultados en las métricas de evaluación (*MAE* y *RMSE*) fue el de *RandomForestRegressor*. En suma, los modelos tipo *Random Forest* son uno de los más utilizados en la revisión de literatura realizada por el autor y suelen tener buenos resultados.

**Métodos rechazados o con bajo desempeño:**

* ***Regresión Lineal***: Para este trabajo, en comparación de los demás modelos, la regresión lineal tuvo el peor desempeño en las métricas de evaluación.
* **Evaluación de modelos de Machine Learning para la predicción de crímenes en la ciudad de Medellín.** (Muñoz Jaramillo, 2021).

**Modelos probados:**

* ***Modelo basado en reglas (modelo base):***Las reglas que lo componen se resumen en que: la zona se clasifica como “caliente” si: el promedio de incidentes en las últimas diez ventanas de tiempo supera el percentil 95; el promedio de incidentes en las últimas cinco ventanas supera el percentil 95; si todas las ventas temporales contienen información de hurtos; en caso contrario, se clasifica como ausencia de “zona caliente”. Este modelo obtuvo valores apropiados en las métricas, por ejemplo, obtuvo la segunda mejor exhaustividad.
* ***Clasificador de Bosques Aleatorios:*** Este modelo obtuvo el segundo mejor valor del F1, tiene el mayor puntaje en términos de exactitud con 76.25% y el segundo mejor desempeño en términos de precisión.
* ***Máquina de Soporte Vectorial:*** Tiene el mejor resultado del valor F1 con 76.06% y el mayor puntaje de exhaustividad con 80.57%.
* ***Regresión Logística:*** Cuenta con la segunda mejor exactitud y el mayor puntaje de precisión con 81.38%.

**Evaluación de los modelos y conclusiones del autor:**

Para los algoritmos de *Aprendizaje Automático* utilizados, se utiliza la técnica de validación: *Nested Cross Validation*, con el objetivo de calibrar los hiperparámetros y validar paralelamente el desempeño de los modelos. A partir de esta validación se evidencia que en los resultados no se presentó un sobreajuste. El autor concluye que no hay un modelo que supere a los demás en todas las métricas, por lo que la elección del mejor modelo debe realizarse con base en las necesidades y recursos disponibles. Por ejemplo, se explica que, si las autoridades buscan abarcar el mayor número de “zonas calientes”, sin importar los recursos utilizados, se recomienda usar el modelo de *Máquina de Soporte Vectorial*, el cual es su vez el modelo más equilibrado, al contar con el mejor valor de F1. Por otra parte, si las autoridades buscan ahorrar recursos, el modelo de *Regresión Logística* es el más apropiado por su precisión, sin embargo, tiene el menor valor en términos de exhaustividad.

**Justificación de modelos y su utilidad en el análisis de distintos tipos de hurtos en la ciudad de Medellín**

En el desarrollo de un modelo de predicción para la problemática es de vital importancia una elección metodológica que posibilite el manejo de datos complejos y que su precisión y adaptabilidad a los datos sea significativa. Sabiendo esto y teniendo en cuenta los objetivos específicos de la investigación, la evaluación de la pertinencia para la aplicabilidad de los modelos analizados anteriormente se basa en el análisis de los posibles beneficios y resultados que pueden entregar.

Los diferentes enfoques y modelos presentados por los estudios realizados aportan una base sólida para la correcta implementación del proyecto, pero es importante entender que, aunque han sido efectivos en sus contextos, pueden ser optimizados y adaptados a las particularidades de los datos disponibles, ofreciendo una oportunidad de mejora en los resultados.

Uno de los aspectos más valiosos encontrados, es el uso de herramientas automáticas como ***AutoGluon*** y ***AutoML*** para la selección de modelos óptimos. Este proceso es realmente útil para un proyecto como este, que maneja grandes volúmenes de datos históricos. Estas herramientas permiten explorar múltiples modelos de manera eficiente, ayudando a estudiar una mayor cantidad de alternativas metodológicas.

En este sentido, modelos como *XGBoots* complementan esta automatización, ya que son valorados por la capacidad que tienen de aprendizaje a partir de datos de difícil entendimiento y por la eficiencia que presentan para la clasificación y regresión. Luegp, al igual que en los estudios analizados, la técnica de *boosting* que se utiliza, permitirá corregir errores de otros modelos, aumentando la precisión en general. Estos hechos hacen que este modelo sea una opción sólida para el estudio a realizar.

Seguidamente, también se destacaron los modelos asociados a Random Forest, ya que, aunque en uno de los trabajos se haya descartado por su alto costo computacional, se explica que se obtienen buenos resultados a partir de estos. En suma, otros estudios destacan la precisión de estos modelos y su desempeño.

Dada la información de los estudios analizados, las diferentes *Redes Neuronales* pueden ser útiles por su capacidad para modelar interacciones complejas. Por ende, la selección de un tipo de *Red Neuronal* justo con la utilización del modelo *Random Forest* se toma como posibilidad a la hora de la evaluación de modelos en la investigación.

Por otro lado, el modelo de agrupamiento *K-means* puede agregar información al análisis acerca de áreas de segmentación de Medellín según patrones de hurtos, identificando así zonas con mayor impacto de esta problemática. La característica de simplicidad y disidencia lo hacen ideal para un análisis exploratorio previo. Sin embargo, su limitada capacidad para manejar datos de alta dimensionalidad puede causar problemas en etapas posteriores, e incluso, es necesario tomar en cuenta que este modelo de agrupamiento no es adecuado para el caso de bases de datos con una gran cantidad de variables categóricas, por lo que es necesario probar otros modelos, siendo un ejemplo de estos el de *K-modes*.

Finalmente, el resto de los modelos socializados como el de *K-Vecinos* o la *Regresión Lineal* se descartan por su ineficiencia con grandes conjuntos de datos y por su poca capacidad de capturar relaciones no lineales respectivamente. Esto los convierte en modelos inapropiados para el objetivo del estudio ya que los hurtos en la ciudad de Medellín son influenciables por diferentes factores complejos.

## Diferenciación del Proyecto

El proyecto de análisis predictivo de hurtos en la ciudad de Medellín se diferencia, ya que no solo se plantea un enfoque analítico para el entendimiento de un problema relacionado con la seguridad, sino que además integra y amplía las posibilidades de implementación de mejoras en las técnicas de *Machine Learning* y herramientas de visualización de datos, de manera de que sea evidente en el resultado la significancia de la problemática y la interpretación del análisis predictivo.

Como primera línea de innovación se pretende la incorporación de datos de diferentes tipos de hurtos en la ciudad, dado que los estudios previos expuestos para el caso de Medellín solo contemplan los hurtos a personas. Igualmente, se agregarán datos de diversas fuentes, como proyecciones poblacionales, eventos temporales relevantes e indicadores socioeconómicos, que pueden ser apoyo fundamental para el entendimiento de los datos. Esto ampliará el contexto en el cual se analizan los hurtos, proporcionando un enfoque holístico y un valor agregado al estudio de la problemática.

Para el proyecto, es importante discriminar los tipos de hurtos y su referenciación geográfica en la ciudad, tal que se entregue una perspectiva particular que pueda no solo predecir volumen de hurtos, sino también evalúe cómo estos patrones evolucionan a lo largo del tiempo y cómo estos se diferencian entre los distintos tipos de robos. En este sentido, se busca entender las diferencias en los resultados de los modelos de predicción de las distintas tipologías de robos, por ejemplo, si estos se deben analizar de forma conjunta, en grupos o discriminada, es decir, ¿Existen tipos de hurtos que tienen patrones lo suficientemente similares como para ser analizados en conjunto en un modelo? En general, se busca entender distintas perspectivas, posibles nuevos aspectos y dinámicas de los hurtos que ocurren en la ciudad.

Seguidamente, la visualización interactiva de los datos es otro aspecto central del proyecto y ésta también añade la capacidad innovadora de no solo incluir predicciones, sino también presentarlas de manera clara y comprensible para que la información pueda ser consumida por un público general e interesado. Esta plataforma permitirá visualizar patrones históricos y futuros de los hurtos de manera visual, georeferenciada y fácil de interpretar. En este aspecto, aprendemos de proyectos como los de Bernal-Urrutia y Campo Agredo, quienes implementaron herramientas de visualización de datos en Bogotá, sin embargo, se busca mejorar sus propuestas y adaptarlas para el caso de Medellín.

## Referencias.

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

Arboleda Colorado, J. S. (2023). *Hurto a personas en la ciudad de Medellín: análisis predictivo de la cantidad de casos en diferentes zonas de la ciudad a partir de modelos de machine learning implementando técnicas de MLOps* [Trabajo de grado maestría, Universidad EAFIT]. http://hdl.handle.net/10784/32909

Arévalo Álvarez, J. C., & Fernández García, M. A. (2022). *Analítica de datos para hurtos a personas en la ciudad de Medellín a través de modelos de Machine Learning y Deep Learning* [Trabajo de grado especialización, Universidad de Antioquia]. https://hdl.handle.net/10495/29063

Bernal-Urrutia, P. A. (2023). *Modelos estocásticos para la predicción y visualización espacial del delito de hurto a personas en la modalidad de atraco en la localidad de Kennedy (Bogotá D.C, Colombia)* [Tesis de grado maestría, Universidad Jorge Tadeo Lozano]. http://hdl.handle.net/20.500.12010/31909

Campo Agredo, C. A. (2017). *Desarrollo de modelos de predicción sobre el hurto a personas en Bogotá a partir del análisis y aplicación de algoritmos en Base de datos* [Tesis de grado, Universidad de los Andes]. http://hdl.handle.net/1992/50785

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

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

Muñoz Jaramillo, V. D. (2021). *Evaluación de modelos de Machine Learning para la predicción de crímenes en la ciudad de Medellín* [Trabajo de grado de maestría, Universidad Nacional de Colombia]. https://repositorio.unal.edu.co/handle/unal/80976

Puentes Bolivar, R. (2023). *Modelo predictivo de hurtos para la ciudad de Medellín utilizando Machine Learning* [Trabajo de grado maestría, Universidad Cooperativa de Colombia]. https://hdl.handle.net/20.500.12494/54260

# Rúbricas

R01: Completitud y consistencia: el estudiante realiza una entrega del momento evaluativo con todos los elementos desarrollados, manteniendo una relación coherente entre los elementos solicitados. El estudiante usa una consistencia argumentativa que incluye el uso de referencias cuando hay mérito de autoría entre otros.

R02: Línea Base: el estudiante es capaz de establecer una clara línea de trabajos previos y teorías basadas alrededor de su proyecto de clase. Esto permite evidenciar que la solución propuesta en el trabajo precio es viable y tiene una base sólida teórica detrás de su posible solución.

R02: Innovación: dentro del ámbito de la analítica y la ciencia de datos el estudiante muestra, basado en la literatura existente, una clara diferenciación de su proyecto frente las soluciones tecnologico-metodológicas encontradas. Se tiene en cuenta la inclusión de la base de datos de CESET para la rúbrica.

R04: Referencias: hay referencias válidas según la descripción del capítulo y están adecuadamente vinculadas al estilo narrativo que permite una clara expresión de los resultados de la investigación de la línea base y el marco tecnológico-metodológico.