

PROYECTO FINAL - SEGUNDA ENTREGA

JUAN SEBASTIÁN ALVAREZ ERASO  
RAFAEL CAMILO TEJON ROJAS  
OSCAR JAVIER ÁNGEL BALCÁZAR

CIENCIA DE DATOS APLICADA

DEPARTAMENTO DE INGENIERÍA DE SISTEMAS Y COMPUTACIÓN  
FACULTAD DE INGENIERÍA  
UNIVERSIDAD DE LOS ANDES

## TABLA DE CONTENIDO

<b>CONTEXTO</b>	<b>3</b>
Limpieza y tratamiento de datos	3
<b>MODELOS EVALUADOS</b>	<b>4</b>
Modelo 1	4
Preparación de los datos	4
Ventajas	4
Desventajas	4
Resultados obtenidos	4
Modelo 2	4
Preparación de datos	4
Ventajas	5
Desventajas	5
Resultados obtenidos	5
Modelo 3: Seleccionado	5
Preparación de datos	5
Selección del modelo	6
Construcción del modelo	6
Resultados obtenidos	6
Ventajas	6
Desventajas	7
(Extra) Modelo no supervisado usando KMEANS	7
Resultados obtenidos	7
Ventajas	8
Desventajas	8
<b>CONCLUSIONES</b>	<b>8</b>

## CONTEXTO

La organización con la que vamos a trabajar es la Policía. El problema que logramos observar es que existe una percepción de inseguridad general en todas las ciudades. Ya no existe un horario del día o una zona en la ciudad donde las personas se sientan seguras, lo que genera una incertidumbre en la población para sus quehaceres diarios. Puntualmente el problema se ataca con data de la ciudad de Medellín, la cual es la segunda ciudad más poblada del país después de Bogotá, para el 2020 la ciudad contaba con 2.533.424 habitantes. La ciudad se divide en 16 comunas y zonas las cuales agrupan las comunas, como lo muestra la página [gubernamental](#).

Las entidades del gobierno como la Policía tienen como objetivo velar por la seguridad de la comunidad en general. Se busca agregar valor con los servicios ofrecidos, por ejemplo: ¿Dónde y en qué horarios deben solicitar apoyo de más policías ante un potencial evento de hurto? o, ¿Es suficiente con el CAI ubicado en X lugar para dar abasto ante las situaciones que se presentan diariamente?. Con la data histórica de hurtos en Medellín se pretende tener un primer acercamiento para responder estas preguntas. La data histórica fue obtenida de la página web de la [Alcaldía de Medellín](#).

Estos son unos ejemplos de métricas que nos pueden ayudar:

- Porcentaje de cantidad de hurtos por hora del día.
- Zonas donde ocurren la mayoría de hurtos.
- Frecuencia de ocurrencia de hurtos.
- Hurtos denunciados versus hurtos ocurridos.

### **Limpieza y tratamiento de datos**

Como primer paso se realizaron los siguientes ajustes y tratamientos a la data obtenida:

1. Se eliminaron features que no nos agregan valor al modelo y al problema planteado, por ejemplo: el nivel académico de la persona que comete el hurto, o también el color del objeto robado.
2. Se eliminaron features que al hacer un análisis exploratorio encontramos que tenían muchos datos nulos o eran features muy sesgados hacia un solo valor, por ejemplo: “conducta” (Comportamiento de la persona que comete el hurto) o “testigo” (Si hubo o no un testigo del hecho).
3. Decidimos darle un mayor tratamiento a features que vimos nos agregaban valor al problema planteado, por ejemplo: la ubicación del hurto (Comuna y barrio), el medio de transporte donde sucede el hurto (Tipo medio de transporte) y también la modalidad utilizada en el hurto. Para estas 2 últimas creamos una categorización espacial para determinar si el hurto es peligroso o no.
4. La fecha del hecho es un feature muy importante dentro de todos los modelos evaluados por lo que obtuvimos de esta el día de la semana y el mes.
5. Evaluamos también la posibilidad de agregar un feature que nos dijera si el hecho ocurrió en un día festivo, sin embargo vimos que había mucha correlación de estos días festivos con los días lunes.

## MODELOS EVALUADOS

### Modelo 1

**Objetivo:** Predecir la variable hurtos de acuerdo a las diferentes features del dataset.

#### Preparación de los datos

Se realizó un primer modelo de regresión lineal en el cual se consideraron las siguientes features tipo\_modalidad, dia\_semana, festivo, tipo\_medio\_transp, tipo\_arma, grupo\_edad y como variable objetivo se estableció la cantidad de hurtos. La anterior variable fue el resultado de realizar un agrupamiento por las anteriores variables y el codigo\_comuna, y codigo\_barrio. A partir de los anteriores features se construyó un modelo de regresión lineal en el cual el objetivo era predecir por comuna y barrio la cantidad de hurtos en un determinado día de la semana. El modelo se entrenó con el 80% de los datos y se evaluó con el porcentaje restante.

#### Ventajas

- Este modelo contempla casi todos los features por medio de los agrupamientos realizados.
- Se categorizaron las variables y el modelo fue base para identificar mejoras que se aplicaron en los siguientes modelos.

#### Desventajas

- No existe suficiente información de hurtos en algunos barrios.
- No es fácil determinar si la categorización es la mejor sin un experto en el área.
- El modelo es muy impreciso ya que no cuenta con los datos suficientes de los diferentes comunas y barrios.

#### Resultados obtenidos

El resultado obtenido del anterior modelo en error de entrenamiento y test fue de: 11.35 y 11.23, concluimos que teníamos un resultado con underfitting y que era muy complejo predecir el número de hurtos en algunos barrios ya que los datos en estos eran pocos por ejemplo algunos barrios para un mes tenían 5 registros con hurtos entre 1 - 5, comparados con otros barrios que alcanzaban a llegar a 38 hurtos.

### Modelo 2

**Objetivo:** Predecir las categorías para cada uno de los barrios, dependiendo del mes y el día de la semana.

#### Preparación de datos

Se realizó un segundo modelo con un Decision Tree Classifier, en el cual se siguieron los siguientes pasos:

- Generación de features: se realizaron conteos de la cantidad de hurtos de la siguiente forma: tipo de *hurtos peligros* y *no peligrosos*, *modalidad de transporte* en la cual se agruparon los diferentes tipos de transporte en categorías como *transporte público*, *particular* y *vía pública*. También se agruparon las *edades* de

forma ordinal en grupos de 15, adicionalmente se categorizaron las *armas en peligrosas y no peligrosas*.

- Agrupamiento: se realizó con el fin de generar los siguientes features *tipo\_mod\_hurtos\_peligrosos*, *tipo\_mod\_hurtos\_no\_peligrosos*, *mod\_transporte\_particular*, *mod\_via\_publica*, *mod\_publico*, *tipo\_arma\_no\_peligrosa*, *tipo\_arma\_peligrosa*. Con los anteriores features se entrenó el modelo en el cual se creó un árbol con una profundidad de 5.

### **Ventajas**

- Este modelo contempla casi todos los features por medio de los agrupamientos realizados.
- Este modelo nos sirvió para generar el modelo 3.

### **Desventajas**

- No existe suficiente información de hurtos en algunos barrios.
- No es fácil determinar si la categorización es la mejor sin un experto en el área.
- Las features del modelo tienen una alta correlación con la variable objetivo, por eso la alta precisión del modelo.
- Por otra parte el modelo también presenta mayor complejidad dada la cantidad de features que se incluyeron en el modelo, para un usuario será más complejo utilizar este modelo.

### **Resultados obtenidos**

Los resultados después de realizar el entrenamiento y evaluar contra el dataset de test fueron los siguientes: precisión: 0.97, recall: 0.97 y F1: 0.97. Aparentemente el modelo tiene un buen comportamiento si vemos los resultados de sin embargo, se da debido a la estrecha relación entre cada una de las features y la variable objetivo, ya que estaban altamente correlacionadas.

### **Modelo 3: Seleccionado**

**Objetivo:** Predecir las categorías para cada uno de los barrios, dependiendo la fecha.

### **Preparación de datos**

Para este modelo lo primero que se hizo fue agrupar todas las filas de los hurtos por barrio, comuna y fecha. Posteriormente se crearon las categorías que se buscará predecir, estas son:

- 1 si en el barrio en una fecha dada ocurrió un robo como máximo.
- 2 si en el barrio en una fecha dada ocurrieron entre 2 y 10 robos.
- 3 si ocurrieron más de 10 robos.

Además de la variable objetivo se utilizaron los siguientes características:

- *tipo\_mod\_hurtos\_no\_peligrosos*: Se asignó una categoría del 0 al 5, donde 5 indica que ocurrieron más robos no peligrosos en el barrio y 0 que ocurrieron menos.
- *tipo\_mod\_hurtos\_peligrosos*: Se asignó una categoría del 0 al 5, donde 5 indica que ocurrieron más robos peligrosos en el barrio y 0 que ocurrieron menos.

- *dayofweek*: Se realizó un *OneHotEncoding* de el día de la semana en la que se está corriendo el modelo, según la fecha.
- *month*: Se realizó un *OneHotEncoding* del mes en el que se está corriendo el modelo, según la fecha.

### Selección del modelo

Para encontrar el mejor modelo se evaluarán varios modelos de clasificación utilizando una técnica de GridSearch en la cual un modelo y un grupo de hiperparametros con el que quiere probar el modelo, posteriormente para evaluar el modelo, primero se partirá un 90% de los datos para entrenamiento y un 10% para prueba. Después se hará cross validation sobre el dataset de entrenamiento para seleccionar el mejor modelo del GridSearch.

### Construcción del modelo

Para correr el modelo se probaran 2 modelos de clasificación multi categóricos, Decision Trees & Random Forests con los siguiente hiperparametros:

Modelo	Hiper Parámetro	Valores
Decision Trees	max_depth	Enteros del 3 al 15
Decision Trees	criterion	gini o entropy
Random Forest	max_depth	Enteros del 10 al 15
Random Forest	criterion	gini o entropy
Random Forest	n_estimators	Enteros múltiplos de 10 entre 50 y 150

Tabla No. 1

### Resultados obtenidos

Después de correr todos los experimentos utilizando el *GridSearch* se obtuvieron los siguientes [resultados](#) (Dada la cantidad de experimentos se incluye un link a un google sheet). Después de observar los resultados se escogieron los mejores modelos de árboles de decisión y de random forest y se corrió el modelo sobre el dataset de prueba y se obtuvieron los siguientes resultados. (Ver Tabla 2).

Se puede concluir que el mejor modelo es el Random Forest aunque a decir verdad no existe una gran diferencia entre todos los modelos que se probaron. Además, se puede observar en la matriz de confusión que ningún modelo aprende a predecir barrios de categoría 3.

### Ventajas

Este modelo es más simple de utilizar, puesto que solo requiere de un barrio y una fecha y con eso se hace la predicción de la categoría. Por otro lado, es un modelo en el cual es más simple incluir información extra a la que estamos usando actualmente.

## Desventajas

- No existe suficiente información de barrios de categoría 3 para que los modelos aprendan a predecir.
- No utiliza todas las características de los datos.
- No es fácil determinar si la categorización es la mejor sin un experto en el área.

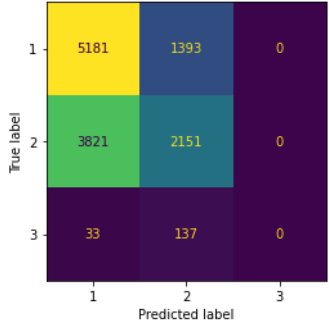
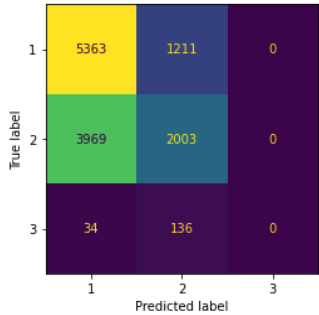
Modelo	Recall	Precisión	F-Score	Matriz de confusión
Árboles de Decisión {'criterion': 'entropy', 'max_depth': 6}	0.5765	0.5765	0.5765	
Random Forest {'criterion': 'entropy', 'max_depth': 10, 'n_estimators': 90}	0.5792	0.5792	0.5792	

Tabla No. 2

## (Extra) Modelo no supervisado usando KMEANS

**Objetivo:** Hacer una clasificación de los barrios de Medellín según la cantidad de hurtos que ocurran.

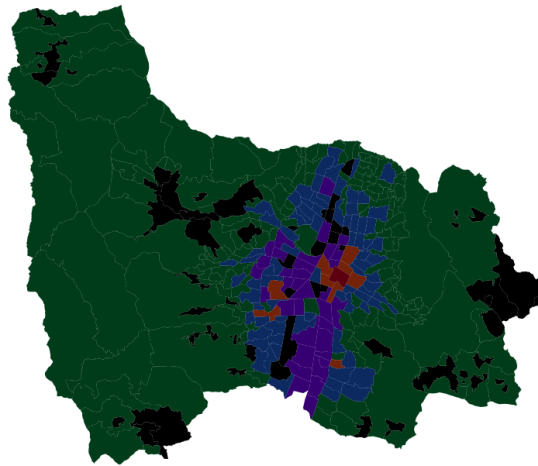
**Preparación de datos:** Para esta sección se agregaron unas nuevas categorías en las que se buscaba calcular la cantidad de hurtos de diferentes clases por barrio, estas son:

- *Hurtos según modalidad:* Cantidad de hurtos de modalidades peligrosas y no peligrosas que ocurrieron en el barrio.
- *Hurtos según género:* Cantidad de hurtos según el género en el barrio.
- *Hurtos según transporte:* Cantidad de hurtos en vía pública, transporte público y transporte particular que ocurrieron en el barrio.
- *Hurtos según edad:* Cantidad de hurtos a diferentes grupos de edad en el barrio.
- *Hurtos:* Cantidad de hurtos en general que ocurrieron en el barrio.

Con las características anteriores se corrió un modelo de KMeans con 5 clusters.

## Resultados obtenidos

Al correr el modelo se obtuvo un mapa en el cual se puede observar la clasificación que hizo el algoritmo de los diferentes barrios y corregimientos de Medellín, según la cantidad de robos se tiene una escala de colores como se puede ver a continuación:



- Negro: No se encontraron datos de estos barrios.



### Ventajas

Este modelo muestra una gran categorización de los barrios que tiene sentido y puede ser utilizado para decidir qué barrios requieren más presencia de la policía.

### Desventajas

Se sale del alcance del curso.

## CONCLUSIONES

1. La cantidad de datos es insuficiente para todos los modelos que se analizaron. Con más datos se pueden obtener resultados más precisos y cumplir mejor el objetivo, por ejemplo:
  - a. Registros de hurtos: Mayor cantidad de hurtos en el dataset.
  - b. Información de las comunas e información de los barrios: El estrato, la cantidad de personas por comuna y barrio para poder categorizarlas, zonas turísticas, categorización por turismo, etc.
  - c. Clima.
  - d. Ubicación del hurto: Aunque ya se tenía el barrio y la comuna se puede contar con información adicional como la dirección del hecho.
  - e. Información urbana: Establecimientos de ocio, bancos etc.
2. Los modelos iniciales que se evaluaron contaban con muchos features que al final aumentaban la complejidad de uso para el usuario final. Esto nos sucedió con el modelo de regresión lineal (Modelo 1) y con los árboles de decisión (Modelo 2).
3. El modelo seleccionado cuenta con los features necesarios que no aumentan la complejidad del mismo y facilitan el uso del API para el usuario final. No se realizaron tantos agrupamientos ni tampoco se utilizó información redundante.
4. El modelo seleccionado es más fácil de evolucionar ya que tiene menos features (Agrupamientos) y esto permite que se puedan agregar más características, por ejemplo: información complementaria de los barrios o comunas.
5. Se aplicaron técnicas de categorización y normalización en los features que facilitaron la construcción de los modelos.