

PROYECTO FINAL - TERCERA 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

CONTEXTO	3
Limpieza y tratamiento de datos	3
MODELOS EVALUADOS	4
Modelo 1	4
Construcción del modelo	4
Modelo 2	4
Construcción del modelo	4
Modelo 3	4
Construcción del modelo	4
Modelo no supervisado usando KMEANS	6
Resultados obtenidos	6
Ventajas	7
Desventajas	7
Evaluación del modelo	7
DESPLIEGUE DEL MODELO	9
CONCLUSIONES	11

CONTEXTO

La organización con la que vamos a trabajar es la Policía de Medellín. 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: ¿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 esta pregunta. La data histórica fue obtenida de la página web de la [Alcaldía de Medellín](#).

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.
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.
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. 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.

Construcción del modelo

Se realizó un primer modelo de regresión lineal y tenía como variable objetivo la cantidad de hurtos. El objetivo era predecir por comuna y barrio la cantidad de hurtos en un determinado día de la semana. El resultado obtenido 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.

Construcción del modelo

Se realizó un segundo modelo con un Decision Tree Classifier, en el cual se generaron features: se realizaron conteos de la cantidad de hurtos y se realizaron agrupamientos por diferentes features, adicionalmente se categorizaron las *armas en peligrosas y no peligrosas*. 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

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

Construcción del modelo

Para este modelo lo primero que se hizo fue agrupar todas las filas. 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. Para encontrar el mejor modelo se utilizó 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 hizo cross validation sobre el dataset de entrenamiento para seleccionar el mejor modelo del GridSearch como se describe en la siguiente tabla:

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

Se obtuvieron los siguientes [resultados](#), se escogieron los mejores modelos de árboles de decisión y de random forest y se corrió el modelo sobre el dataset de prueba, el mejor modelo es el Random Forest aunque a decir verdad no existe una gran diferencia entre todos los modelos que se probaron. Los resultados del mejor modelo se observan en la siguiente tabla:

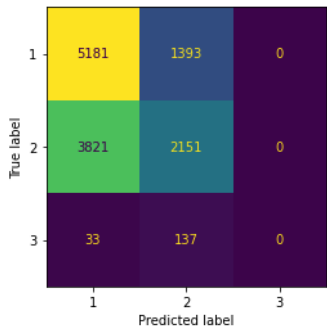
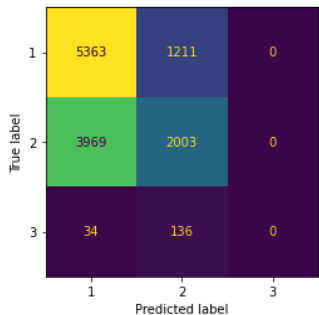
Modelo	Recall	Precisión	F-Score	Matriz de confusión
Árboles de Decisión {'criterion': 'entropy', 'max_depth': 6}	0.5765	0.5765	0.5765	
Random Forest {'criterion': 'entropy', 'max_depth': 10, 'n_estimators': 90}	0.5792	0.5792	0.5792	

Tabla No. 2

Modelo no supervisado usando KMEANS

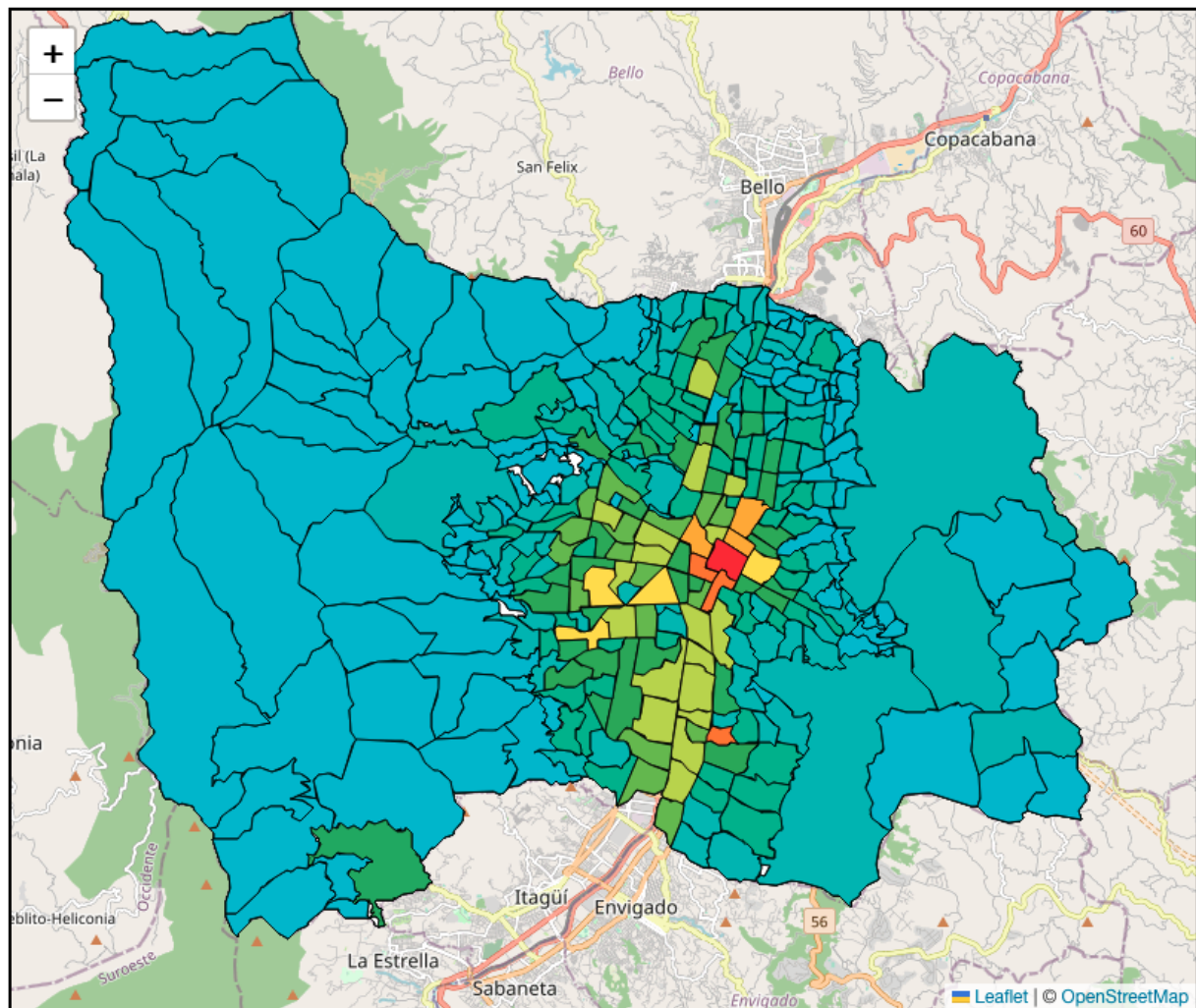
Objetivo: Hacer una clasificación de los barrios de Medellín según la cantidad de hurtos que ocurran.

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.
 - **Modalidades Peligrosas:** atraco, rompimiento de ventana, escopolamina, fleteo, tóxico o agente químico, miedo o terror, vandalismo, forcejeo, halado y rompimiento cerradura.
 - **Modalidades No Peligrosas:** descuido, raponazo,descuido, cosquilleo, clonación de tarjeta, engaño, retención de dinero, informático, suplantación, llamada millonaria, paquete chileno, abuso de confianza, halado, simulando necesidad,comisión de delito, llave maestra
- *Hurtos según género:* Cantidad de hurtos según el género en el barrio.
 - **Hombres**
 - **Mujeres**
- *Hurtos según transporte:* Cantidad de hurtos en vía pública, transporte público y transporte particular que ocurrieron en el barrio.
 - **Transporte Público:** taxi, autobús y metro.
 - **Transporte Particular:** automóvil, motocicleta y bicicleta.
 - **Vía Pública:** motociclista con parrillero, caminata.
- *Hurtos según edad:* Cantidad de hurtos a diferentes grupos de edad en el barrio.
 - **Niños:** Menores de 12 años.
 - **Jóvenes:** Mayores de 12 años y menores de 18 años.
 - **Adultos:** Mayores de 18 años y menores de 61 años.
 - **Adultos Mayores:** Mayores de 61 años.
- *Hurtos:* Cantidad de hurtos en general que ocurrieron en el barrio.

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:



Menos peligroso  Más peligroso

- Blanco: No se encontraron datos de estos barrios.

Ventajas

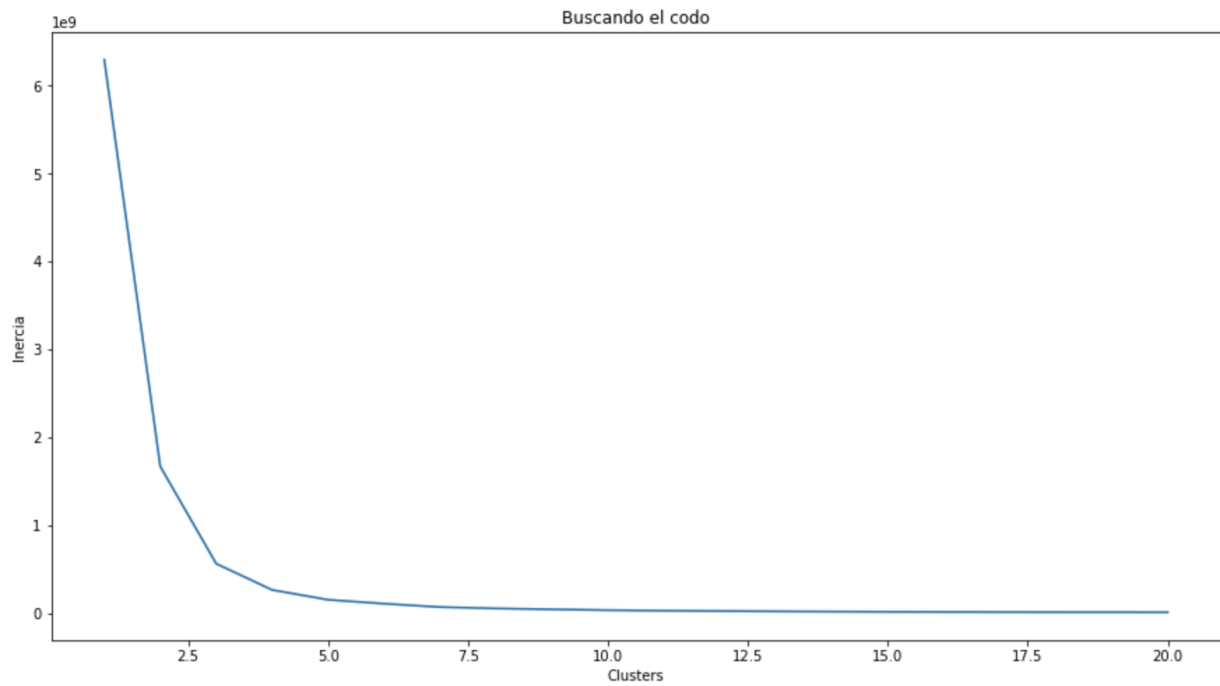
Este modelo muestra una gran categorización de los barrios que tiene sentido y puede ser utilizado para decidir qué barrios requieren más presencia de la policía.

Desventajas

Debido a que no era la temática del curso no pudimos profundizar lo suficiente en esta clase de modelos.

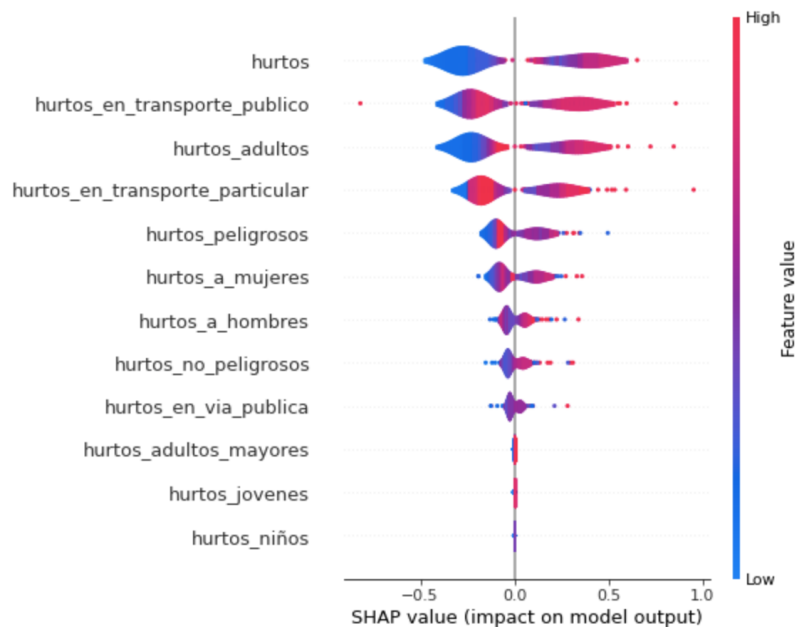
Evaluación del modelo

Utilizamos el método del codo para seleccionar el número de clusters de nuestro modelo. Este consiste en graficar la inercia vs el número de clusters y elegir el valor del cluster que está en el codo de la curva.



Se puede observar que la cantidad de clusters para nuestro modelo está entre 2 y 3.

También evaluamos cuál es el feature más importante de nuestro modelo utilizando shap, nos arrojó que los features más importantes son los hurtos en el transporte público junto con los hurtos a las personas adultas.



Para probar distintos modelos no supervisados se decidió variar la cantidad de clusters, desde 1 hasta 10.

- ¿Qué condiciones considera que deberían tener los datos para obtener mejores resultados? Más datos, mejores características, etc.

En cuanto a los datos se podría mejorar la calidad del lugar donde ocurrieron los hechos con información de latitud y longitud más precisa, también se podría incluir features como los siguientes:

- ❖ Dirección de los hechos, esto con el fin de obtener mayor detalle y quizás poder determinar las calles, carreras, parques o zonas públicas más o menos peligrosas.
 - ❖ Descripción de los hechos.
 - ❖ Información demográfica y de puntos de interés cercanos como bancos, centros comerciales, cajeros, hospitales, parques, centros culturales, estaciones del metro, etc.
- ¿El modelo obtenido es suficiente para soportar la necesidad u oportunidad de negocio identificada?

Es un buen MVP del modelo, ya que se pueden tomar decisiones referente a la distribución de los agentes de policía en la ciudad. Aunque el modelo se podría evolucionar en su nivel de detalle utilizando la información de las calles, relacionando también los puntos de interés general con el fin de obtener una mejor comprensión de las actividades delincuenciales.

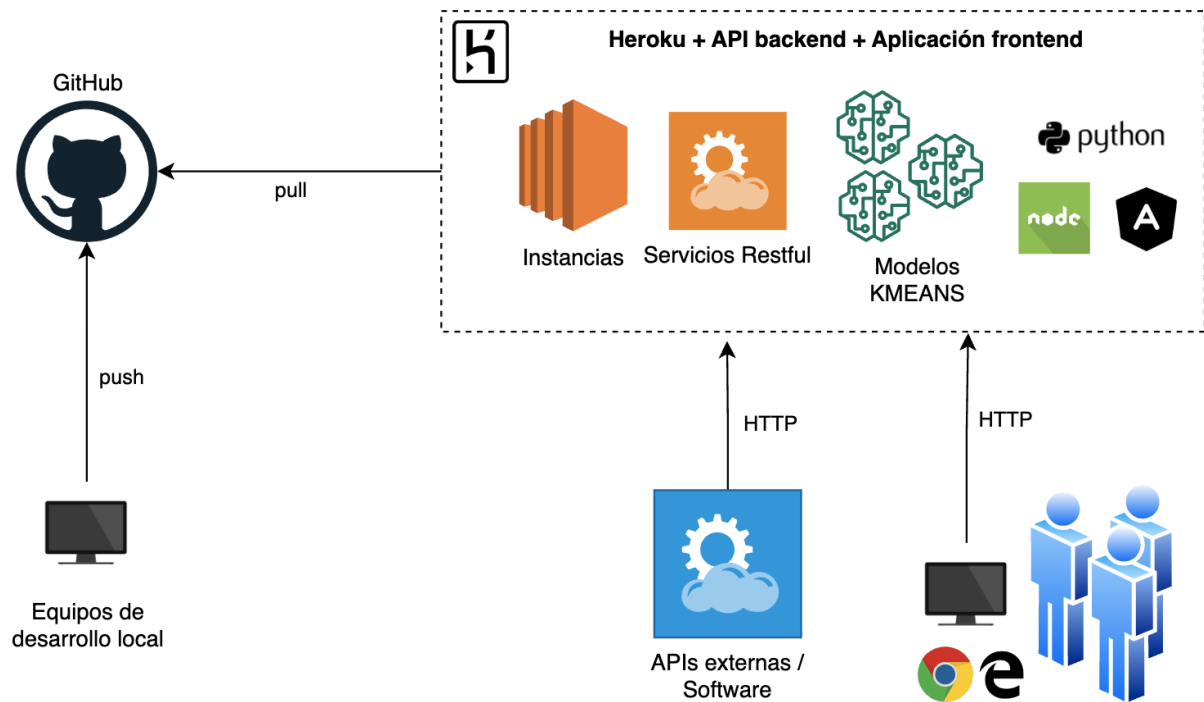
DESPLIEGUE DEL MODELO

Para el despliegue del modelo se construyeron 2 aplicaciones software; una es una API backend construida en Python con Flask y la otra es una aplicación web construida en Angular. Las 2 aplicaciones se encuentran desplegadas en Heroku que permite, dada su facilidad de uso, que con un push de los cambios en la rama “main” se desplieguen automáticamente en producción.

Dentro de la API backend se encuentran los modelos en formato .joblib que son cargados en memoria, no es necesario realizar nuevamente el entrenamiento de estos modelos. La API es capaz de devolver las categorías donde cada modelo ubica a los barrios de acuerdo a su peligrosidad.

La aplicación web renderiza el mapa de la ciudad de Medellín y los barrios con un color de acuerdo a su categoría.

Este es el diagrama de despliegue del proyecto:



Los usuarios finales (personal de la Policía) van a poder acceder al sistema usando un explorador web y la URL de la aplicación frontend.

Repositorios de código:

- [API backend](#).
- [Aplicación frontend](#).

URLs:

- <https://mine-4101-proyecto-final-api.herokuapp.com/ping> → API health status.
- <https://mine-4101-proyecto-final-front.herokuapp.com> → Sitio web.

CONCLUSIONES

1. La evaluación y experimentación con varios modelos de clasificación nos hizo caer en cuenta que estábamos haciendo una clasificación previa de los barrios como peligrosos, no tan peligrosos y menos peligrosos. A su vez esperábamos que el modelo hiciera lo mismo, esto es como si a nuestro modelo lo estuviéramos sesgando a dar una respuesta.
2. Por el feedback recibido decidimos irnos con un modelo no supervisado, que hiciera la clasificación de acuerdo al conjunto de datos pre-procesado.
3. Utilizando el método del codo se encontró que el mejor número de clusters para nuestro modelo estaba entre 2 y 3. Un buen modelo es uno con baja inercia y un bajo número de conglomerados (K). Sin embargo desde la visualización de los datos clasificados en el mapa los modelos con más de 3 cluster también son de gran utilidad porque brindan información detallada de la peligrosidad de los barrios.
4. No se obtuvieron gráficas claras de la separación de clusters en el notebook y esto puede ser debido a que tenemos outlayers en nuestro conjunto de datos.
5. Los features más importantes según la gráfica del shap para nuestro modelo son los hurtos en el transporte público y los hurtos a las personas adultas.
6. Con el modelo de KMEANS se obtuvo una segmentación de los diferentes barrios de la ciudad y por medio de una escala de color se muestran los barrios más peligrosos y menos peligrosos.
7. El modelo de KMEANS puede ser iterado agregando features para obtener otro nivel de detalle y sea de mayor valor para los stakeholders.