



Proyecto Final

Integrantes:

- Oscar Javier Ángel Balcázar
- Rafael Camilo Tejon Rojas
- Juan Sebastian Alvarez Eraso

Contexto



Basados en información histórica de hurtos en la ciudad de Medellín planteamos un modelo de machine learning para clasificar la peligrosidad de un barrio/comuna. Lo anterior con el fin de brindar una herramienta a la policía para facilitar la toma de decisiones.

Se plantearon y evaluaron varios modelos supervisados y no supervisados que se describen a continuación.

Fase No 1



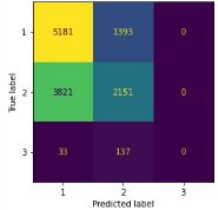
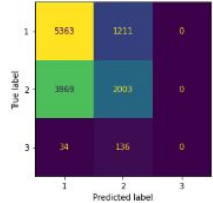
Opción	Modelo	Objetivo	Conclusiones
1	Regresión lineal	Predecir la cantidad de hurtos	<ul style="list-style-type: none">- Es muy complejo utilizar los datos que tenemos- Datos insuficientes.
2	Árboles de decisión	Predecir categorías de barrios como peligrosos o no peligrosos	<ul style="list-style-type: none">- Es muy complejo.- Tiene muchos features.
3	Random forest + Árboles de decisión	Predecir categorías de barrios como peligrosos o no peligrosos	<ul style="list-style-type: none">- Se utilizaron menos features.- Se evaluaron muchas más combinaciones de hiperparámetros.- Simple de evolucionar.
4	KMEANS	Encontrar información en los datos, categorizar los barrios	<ul style="list-style-type: none">- Se sale del conocimiento del curso.

Fase No 2

GridSearch + Árboles de decisión + Random forest

Features: *tipo_mod_hurtos_no_peligrosos*, *tipo_mod_hurtos_peligrosos*, *OneHotEncoding del mes y del día de la semana*.

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

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	

Fase No 3



Modelo seleccionado - KMEANS

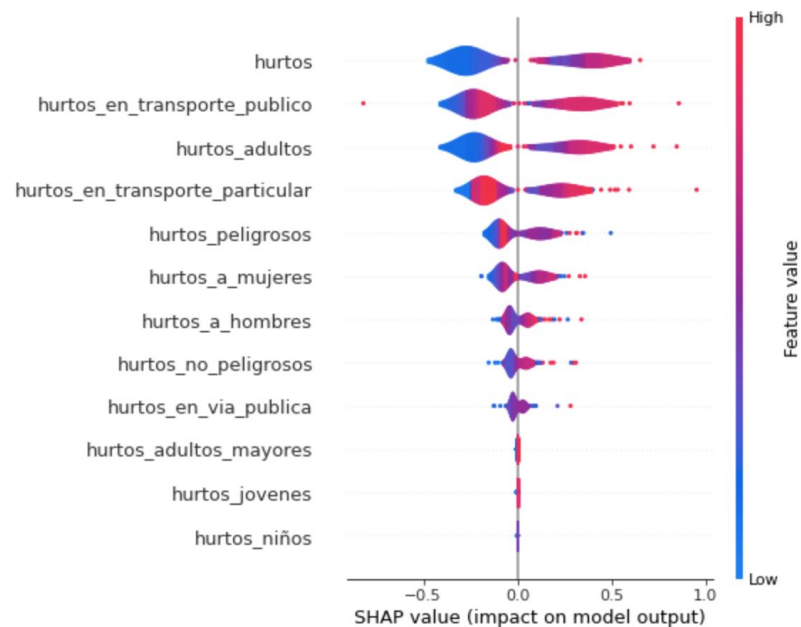
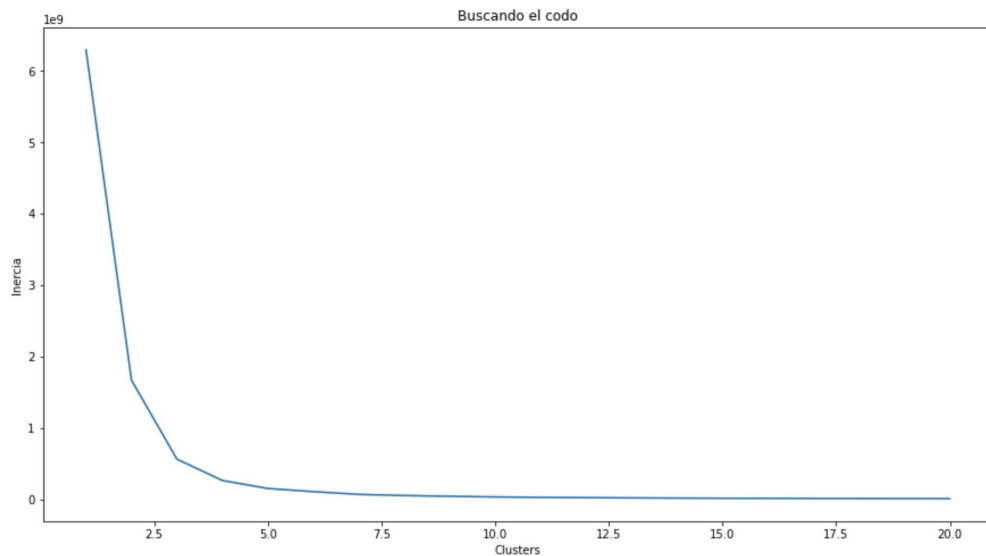
Objetivo: Hacer una clasificación de los barrios de Medellín según la cantidad de hurtos que ocurran.

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:

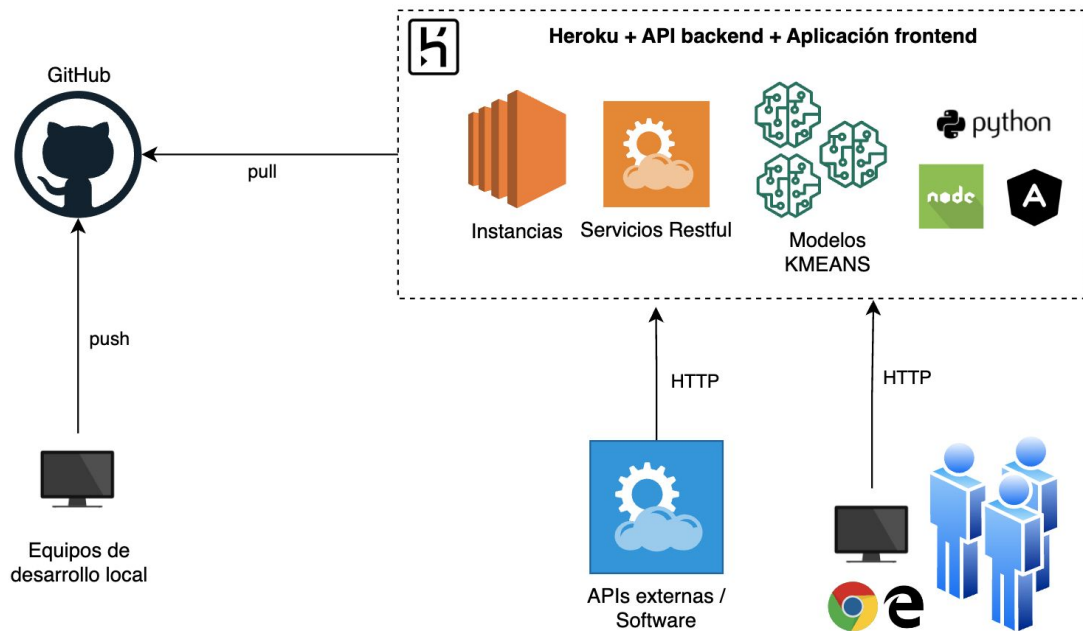
- Hurtos según modalidad
- Hurtos según género
- Hurtos según transporte
- Hurtos según edad
- Hurtos Generales

Modelos: Se probaron modelos K MEANS con 1 hasta 10 clusters

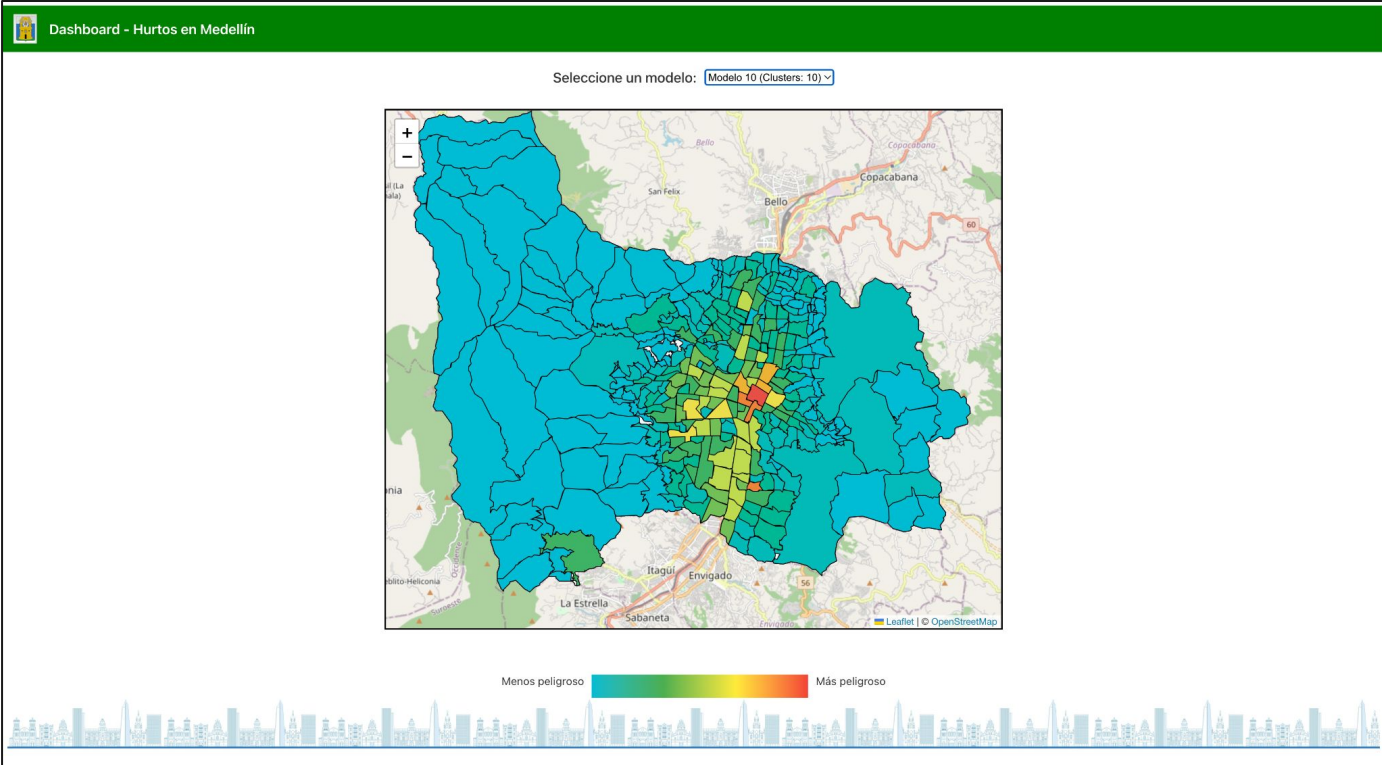
Evaluación del Modelo



Despliegue de la solución



Plataforma



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. 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.
3. Los features más importantes para nuestro modelo son los hurtos en el transporte público y los hurtos a las personas adultas.
4. 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.
5. El modelo de KMEANS puede ser iterado agregando features para obtener otro nivel de detalle y sea de mayor valor para los stakeholders.