

Proyecto final - Segunda entrega

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 predecir la peligrosidad de un barrio/comuna en un día de la semana y mes del año. Esta peligrosidad la clasificamos de 1 a 3 (1 = menos peligrosa, 3 = más peligrosa).

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



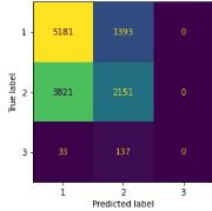
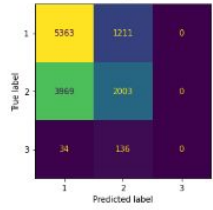
Modelos evaluados

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.

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	



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 (Registros de hurtos, información de las comunas, información de los barrios, clima, ubicación del hurto, información urbana, etc.).
2. Los modelos iniciales que se evaluaron contaban con muchos features que al final aumentaban la complejidad de uso para el usuario final.
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 (plataforma) final.
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.