Predicción de precios de arriendos de viviendas en la ciudad de Medellín en base a información recolectada a través de Web Scraping

MACHINE LEARNING II

WALTER ARBOLEDA CASTAÑEDA

Planteamiento del problema – Enunciado

Enunciado principal:
Crear un sistema de predicción de precios de arriendos de la ciudad de Medellín en base a información recolectada a través de web scraping

Criterios de aceptación:

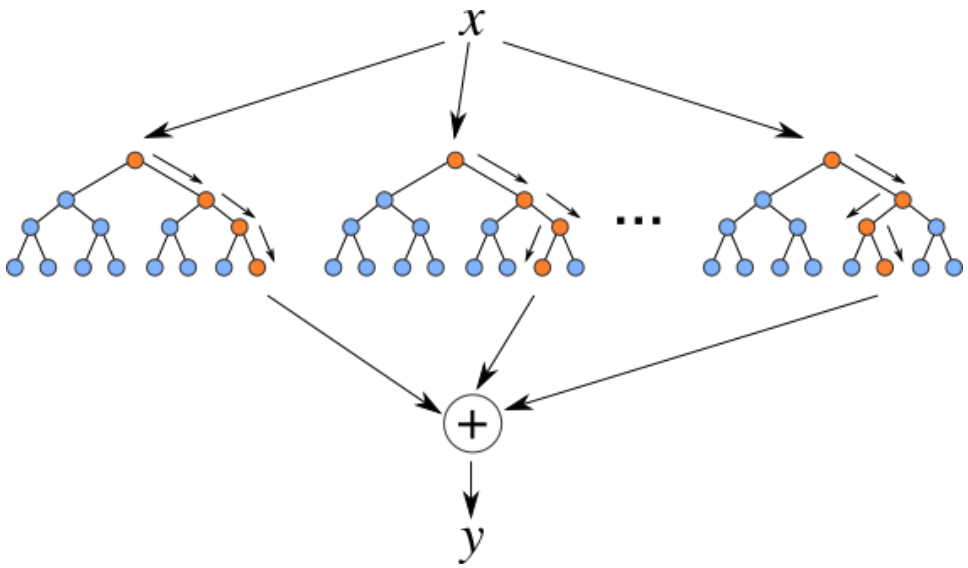
- El proceso de scraping debe implementarse de manera que se garantice una ejecución automática y periódica.
- La información recolectada debe almacenarse en el data lake para mantener un registro histórico
- El modelo debe evaluarse en diferentes zonas de la ciudad dada la siguiente condición *“El modelo es implementable en la zona X si tiene un MAPE menor o igual al 15%”*

Planteamiento del problema – Técnica ML

Técnica ML utilizada:
El sistema de predicción será abordado a través de un Random Forest Regreession dónde se elegirá el mejor modelo resultante de la iteración con la siguiente configuración de parámetros:

- N_estimators: 40, 45 ,50, 60, 70, 100
- Max features: 3,5,7,10
- Max Depth: 3, 5, 7, 10, 15

Se elegirá quien mejor métrica de R2 posea, calculando esta tanto en entrenamiento como en prueba para controlar el overfitting

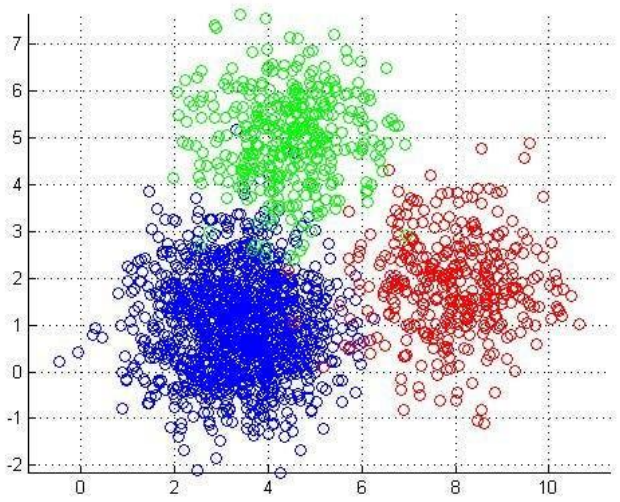


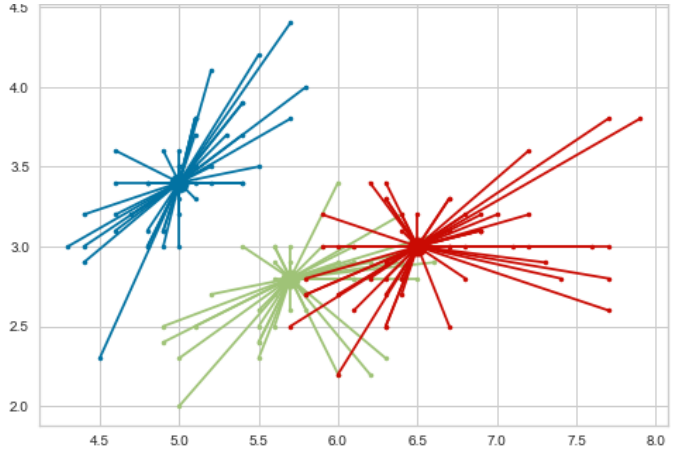
Técnicas del curso KMEANS – KMODES

Evaluación del modelo:
Dado que el criterio de aceptación del negocio es brindar garantía sobre la cobertura del modelo en la ciudad de Medellín, se usará la estrategia de clustering para calcular diferentes zonas en las cuales es optimo dividir la ciudad y se evaluará el modelo en cada una de estas.

Estrategias para clustering

- Aplicación de kmeans sobre variables numéricas
- Aplicación de kmodes sobre variables categóricas





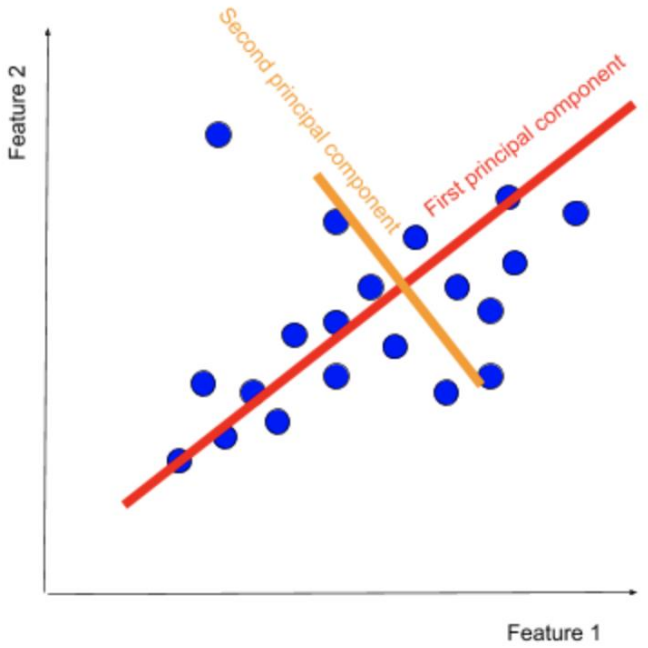
Técnicas del curso PCA – FAMD

Coponentes principales:
Dado el objetivo de lograr un MAPE general de no mayor a 15%, se experimentará usando reducción de dimensionalidad en busca de obtener resultados que acerquen el modelo a la meta deseada

Estrategia de reducción de dimensionalidad:

- PCA sobre la base haciendo encoding sobre los datos categóricos para obtener todas las variables numéricas
- FAMD sobre la base mixta de datos numéricos y categóricos

Con ambos métodos se hará un entrenamiento del modelo y evaluar si hay mejora en las métricas



Resultados Random Forest

Hiperparámetros:

Parámetros que dieron el mejor estimador

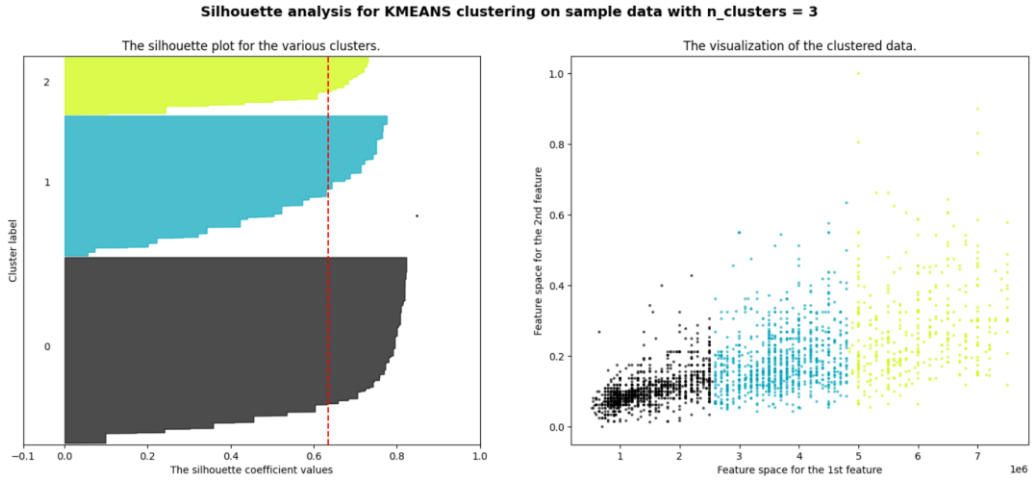
- Max Depth: 7
- Max Features: 5
- N Estimators: 50

Resultados:

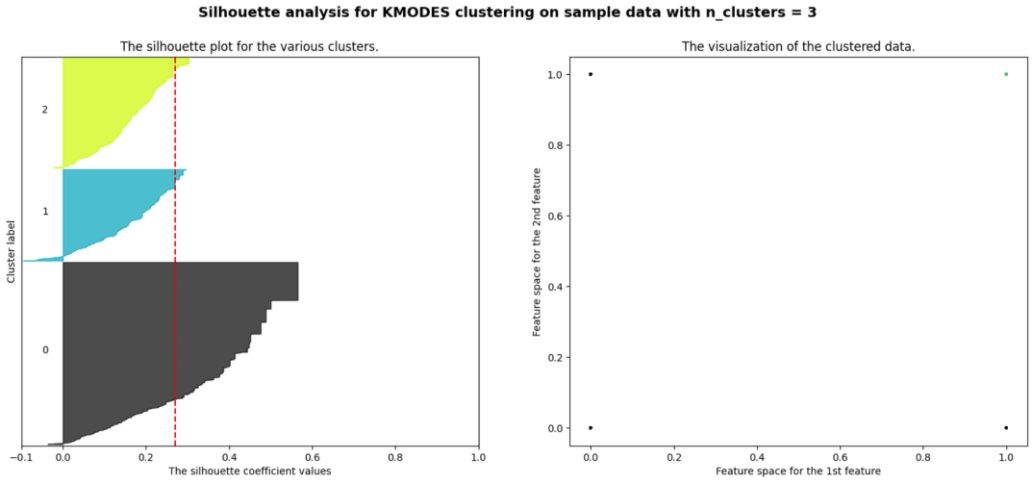
- R2 Train: 0.84
- R2 Tes: 0.78

Métrica	Valor Test
MSE	672129997994,2396
RMSE	819835,348
MAE	577805,586
R2	0,780
MAPE	0,23

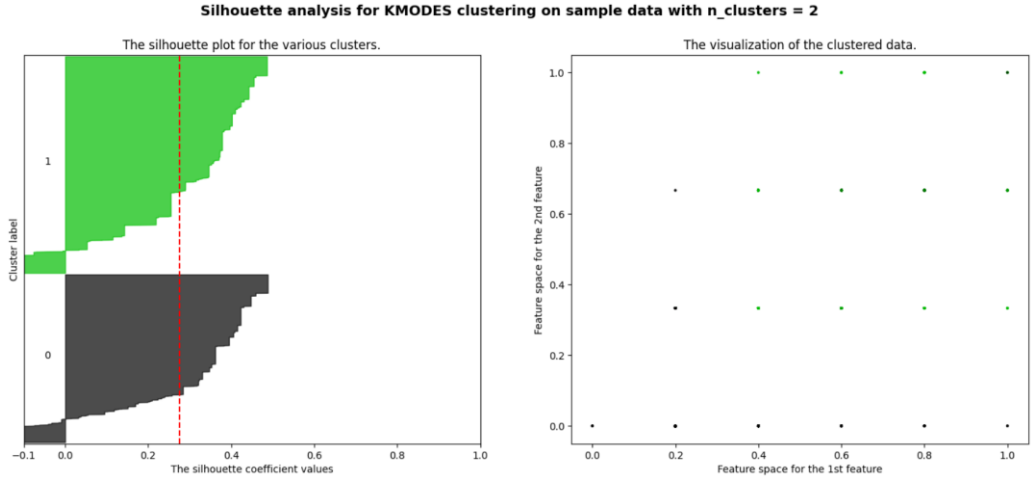
Resultados KMEANS – KMODES



silhouette_score : 0,63



silhouette_score : 0,26

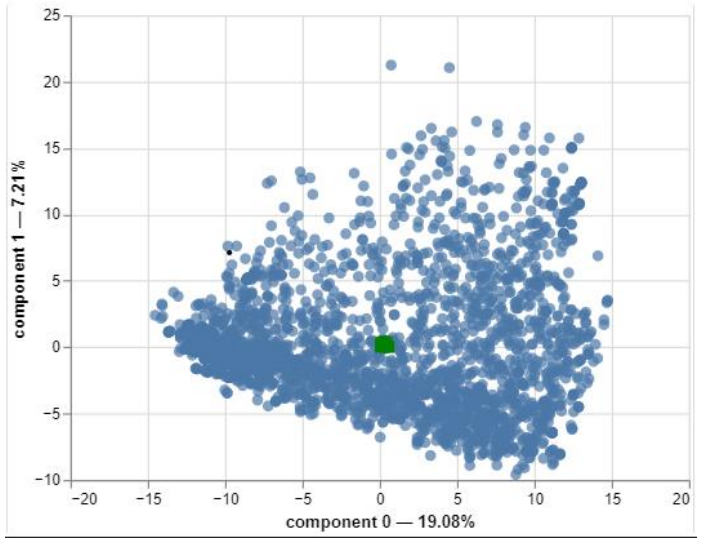


silhouette_score : 0,27

Cluster Precio	Min	Max	MAPE
0	530000.0	2550000.0	0.285
1	2600000.0	4800000.0	0.157
2	4850000.0	7500000.0	0.213

Resultados FAMD

	eigenvalue	% of variance	% of variance (cumulative)
component			
0	62.189	19.08%	19.08%
1	23.518	7.21%	26.29%
2	18.727	5.74%	32.03%
3	14.292	4.38%	36.42%
4	14.151	4.34%	40.76%
5	12.326	3.78%	44.54%
6	11.854	3.64%	48.18%
7	10.621	3.26%	51.43%
8	10.272	3.15%	54.59%
9	10.074	3.09%	57.68%
10	9.802	3.01%	60.68%
11	9.487	2.91%	63.59%
21	5.439	1.67%	86.50%



Hiperparámetros:

Parametros que dieron el mejor estimador

- Max Depth: 7
- Max Features: 3
- N Estimators: 40

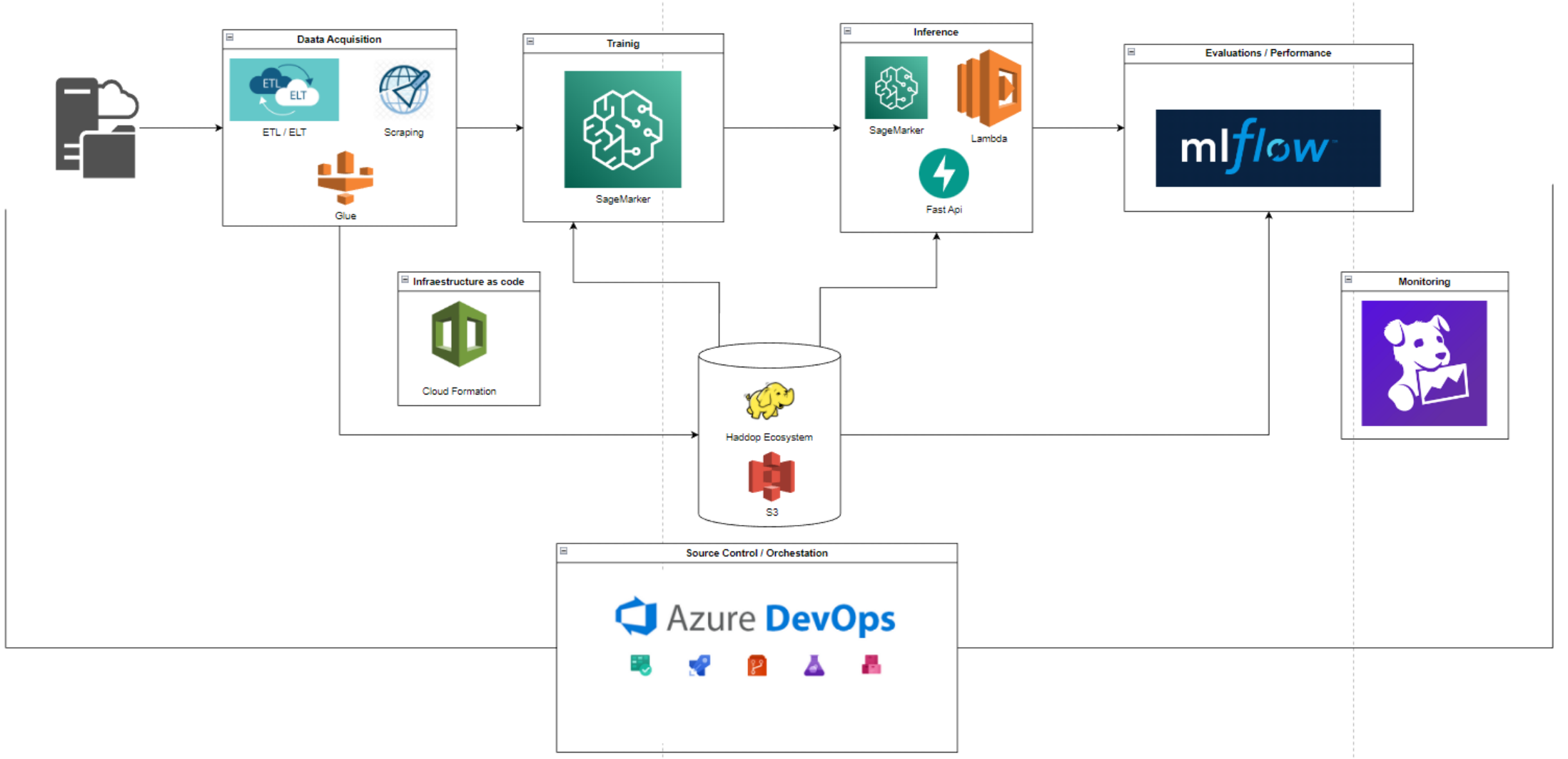
Resultados:

- R2 Train: 0.81
- R2 Tes: 0.73

Resultados modelo con 7 componentes FAMD

Métrica	Valor Test
MSE	824144885367,9
RMSE	907824,259
MAE	654886.65
R2	0,734
MAPE	0,26

Arquitectura ML



Arquitectura Despliegue