Segmentación de Clientes para Prevención de Lavado de Activos

Solución Analítica E2E para Bancolombia

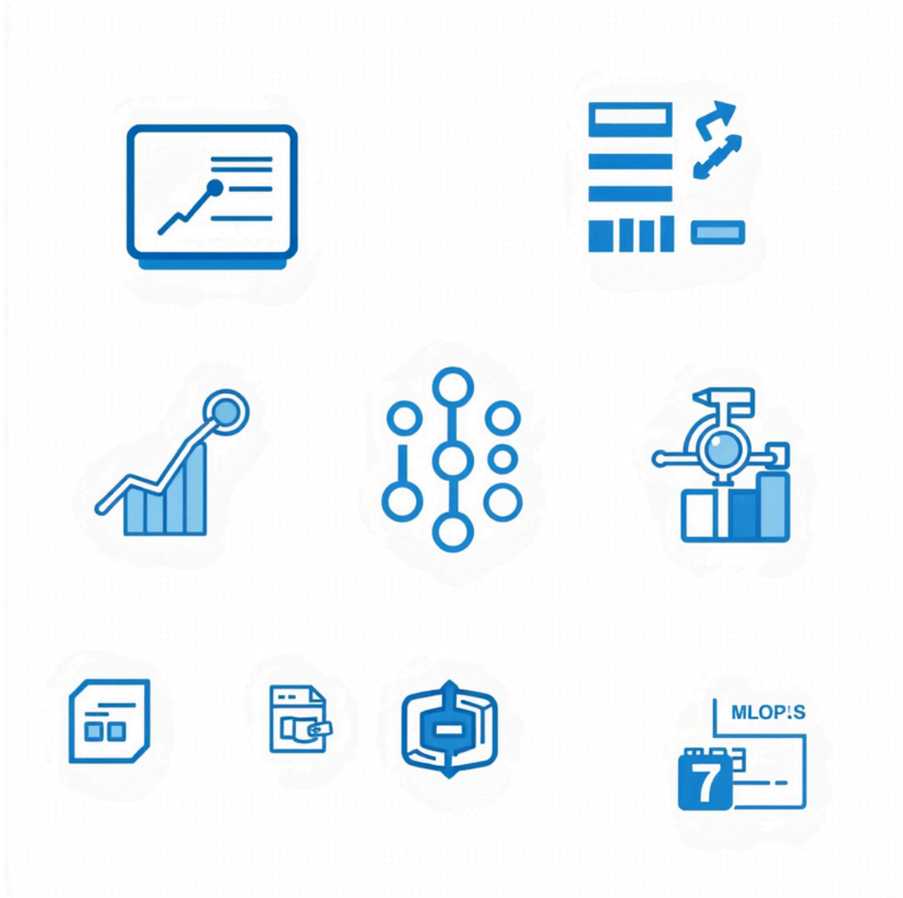
Autor : Cristian Camilo Hidalgo García

03 de marzo de 2025

Contexto y Problema

- Bancolombia como entidad regulada necesita prevenir el lavado de activos y el financiamiento del terrorismo.
- El regulador exige un proceso de segmentación periódico para simplificar la complejidad de tratar con numerosos clientes.
- Actualmente, se mantienen 30 modelos con ejecuciones semestrales, requiriendo alta operatividad.

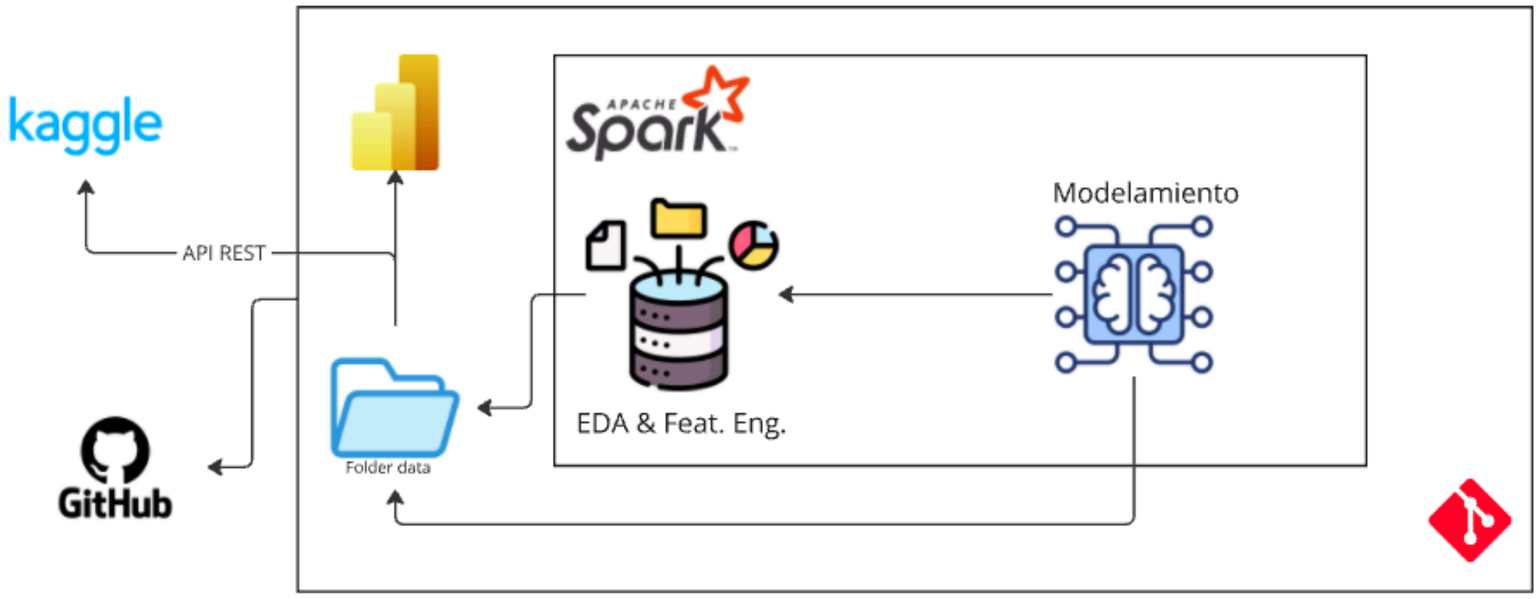
Objetivo de la Prueba Analítica



- Determinar las capacidades analíticas para desarrollar e implementar modelos.
- Diseñar y construir un modelamiento de segmentación a partir de la información dada.
- Proporcionar una solución analítica E2E cumpliendo prácticas de MLOps.

Solución Propuesta

- Preprocesamiento de datos.
- Análisis exploratorio de datos.
- Feature engineering.
- Modelamiento de segmentación.
- Evaluación de clusters.
- Presentación de resultados.
- Blueprint de la solución.

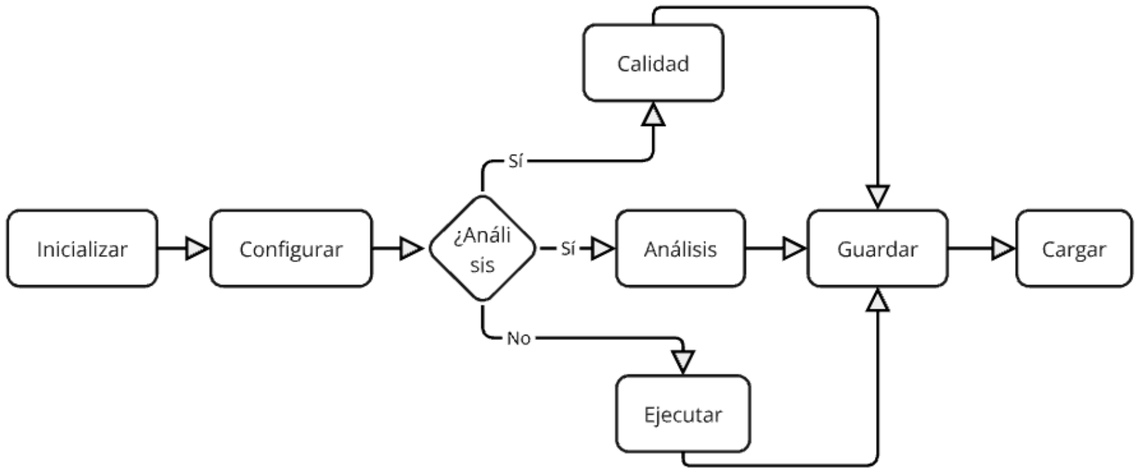


Preprocesamiento de Datos

La limpieza y preparación de datos tuvo las siguientes fases:

- Tratamiento de variables categóricas: homogeneización y conversión a minúsculas.
- Entendimiento de la proporción de los niveles de cada variable.
- Perfil transaccional de los clientes
- Uso de **Spark** para procesamiento debido al volumen de datos.

Análisis Exploratorio de Datos



Comprendiendo los Datos

- Análisis de variables numéricas y categóricas.
- Distribución asimétrica con sesgo a la derecha en datos financieros.
- Conjunto de datos compuesto por características de clientes (1 millón) y transacciones (5 millones).
- Construcción de perfil transaccional de los clientes a través de variables (promedio mes): monto transacciones, número de transacciones, ingresos, egresos, rato ingresos/gastos, etc.

Feature Engineering

Creación de Nuevas Variables

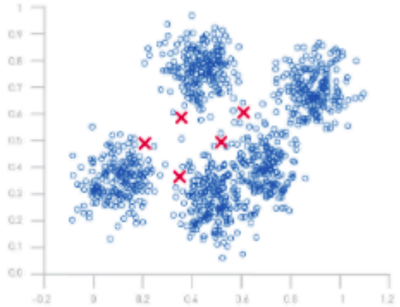
- Dummización de variables categóricas con frecuencia superior al 5%.
- Agrupamiento de categorías para ocupación y profesión.
- Ejemplo de agrupamiento:
  - OCCUPATION_CATEGORIES: Profesionales, Autónomos, No trabajadores, Otros.
  - PROFESSION_CATEGORIES: STEM, Salud, Negocios, Artes.

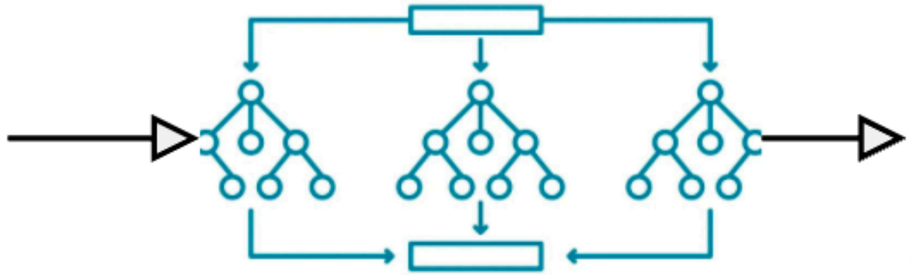
```
OCCUPATION_CATEGORIES = {
"Professionals": ["profesional independiente", "socio o empleado - socio"],
"Self-employed": ["independiente", "comerciante", "rentista de capital", "agricultor", "ganadero"],
"Non-working": ["pensionado", "ama de casa", "estudiante", "desempleado con ingresos", "desempleado sin ingresos"],
"Others": ["None", "otra"]
}
```

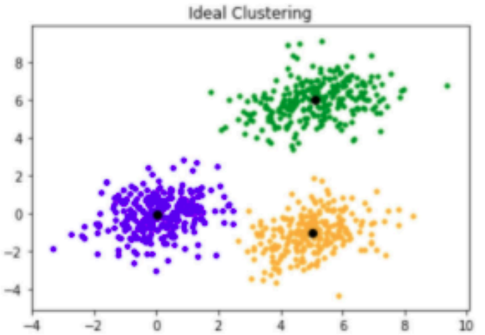
Modelamiento de Segmentación

Técnicas de Clustering

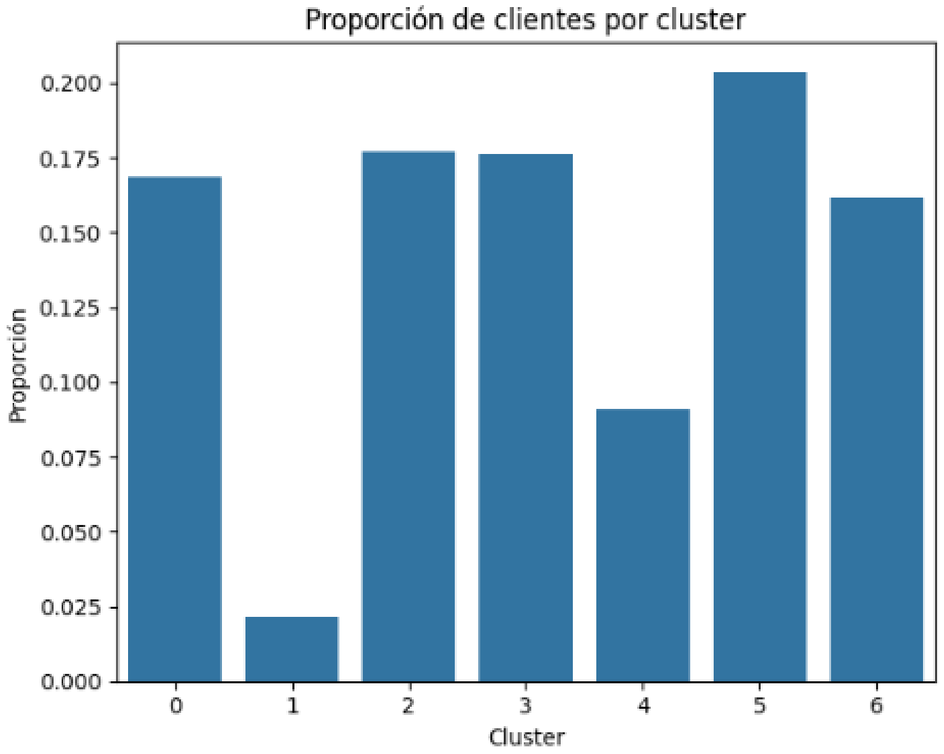
- Uso de KMeans con Spark como modelo base.
- Identificación de variables con mayor variación.
- Aplicación de Random Forest para determinar importancia de variables.
- Reentrenamiento del modelo KMeans con las K variables más importantes.







Evaluación de Clusters



Calidad de los Clusters

- Contenido :
  - Indicador de silueta: **0.5536.**
  - Cantidad de segmentos: **7**
  - Distribución de clusters:
    - Cluster 1: 21,557 clientes (2.16%).
    - Cluster 6: 161,715 clientes (16.17%).
    - Cluster 3: 175,891 clientes (17.59%).
    - Cluster 5: 203,630 clientes (20.36%).
    - Cluster 4: 91,123 clientes (9.11%).
    - Cluster 2: 177,236 clientes (17.73%).
    - Cluster 0: 168,755 clientes (16.88%).

Resultados: Segmentos

cluster	count	proportion
1	21557	0.021559
6	161715	0.161730
3	175891	0.175907
5	203630	0.203649
4	91123	0.091131
2	177236	0.177252
0	168755	0.168771

Perfiles de Segmentos

- Segmento 1: "Profesionales de la Salud de Alto Ingreso".
- Segmento 2: "Ahorradores de Ingreso Medio-Alto".
- Segmento 3: "Trabajadores Tradicionales".
- Segmento 4: "Profesionales Urbanos de Medellín".
- Segmento 5: "Hombres Profesionales".
- Segmento 6: "Trabajadores de Ingreso Básico".
- Segmento 0: "Hombres Trabajadores de Ingreso Medio-Bajo".

¿Preguntas?