Segmentación de clientes para un E-commerce

Implementación de algoritmo K-Means para una E-commerce

Kevin Iza

Proyecto de Aprendizaje no supervisado

8 de enero de 2026

1. Introducción

Este informe detalla la creación de un sistema de inteligencia de negocios diseñado para categorizar a los clientes de un e-commerce mediante el análisis de su comportamiento histórico. Se utiliza una arquitectura robusta que combina SQL para la gestión de datos y aprendizaje no supervisado para la generación de conocimiento.

2. Análisis Exploratorio y Distribución

Antes del modelado, se analizaron las métricas RFM (Recency, Frecuency, Monetary). La Figura 1 muestra la distribución de estas variables. Se identificó un alto sesgo hacia la derecha y la presencia de valores atípicos, lo que justificó la aplicación de transformaciones logarítmicas para normalizar los datos y mejorar la convergencia del algoritmo de clustering.

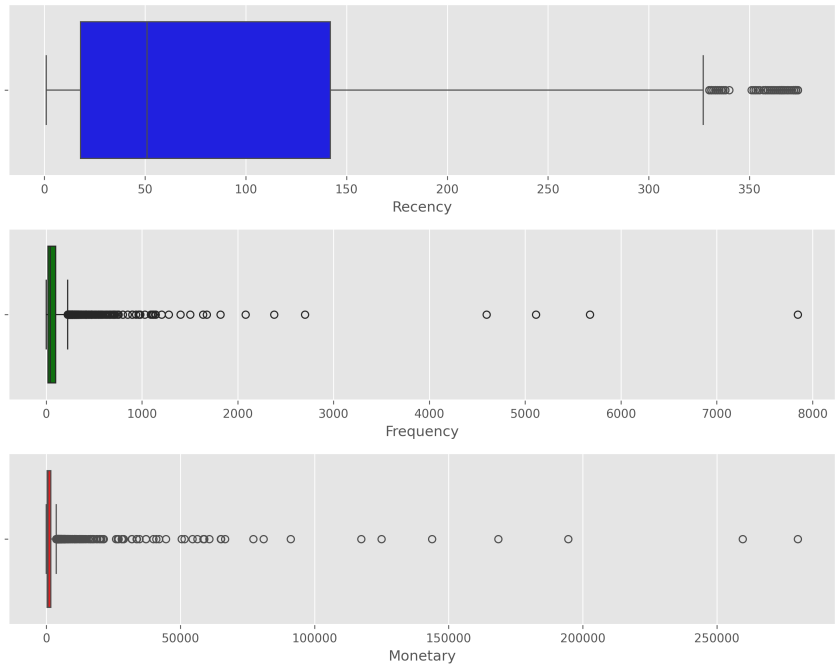

Figura 1: Distribución estadística de las métricas RFM originales.

3. Metodología y Validación

Se utilizó el algoritmo **K-Means**. El número óptimo de grupos se validó mediante el **Método del Codo** (Figura 2), seleccionando $K = 3$ como el punto de inflexión donde la inercia (SSE) comienza a disminuir de forma marginal.

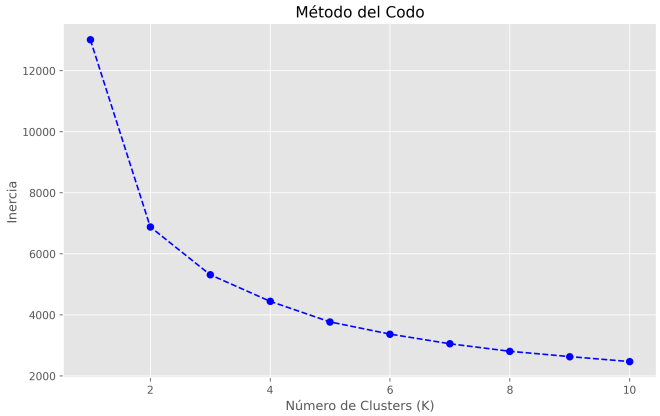

Figura 2: Validación del número de clusters mediante la suma de errores cuadráticos.

4. Interpretación de Segmentos

La segmentación permitió identificar tres perfiles estratégicos claramente diferenciados (Figura 3):

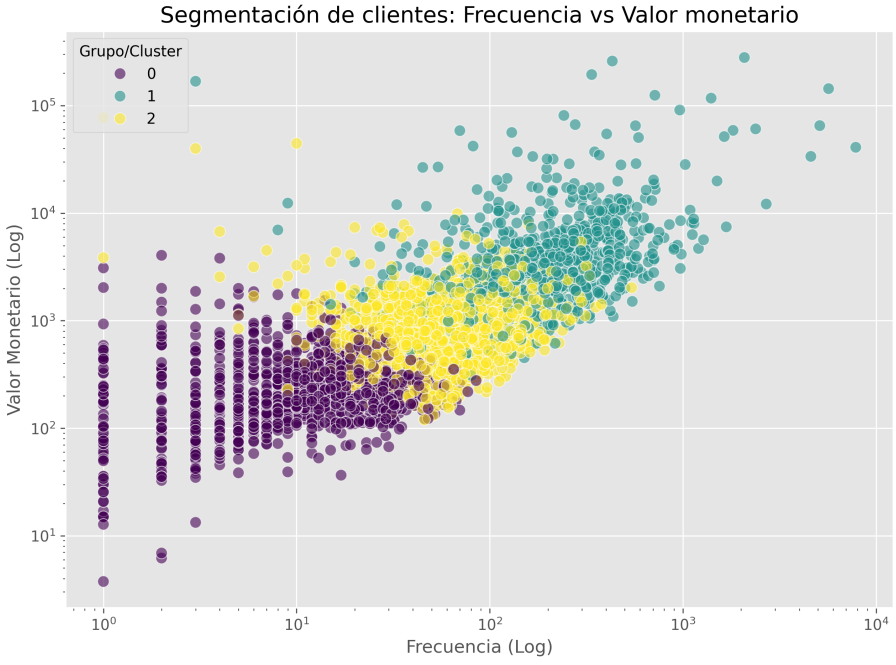

Figura 3: Visualización de Clusters: Frecuencia vs. Valor Monetario (Escala Logarítmica).

- **Cluster 1 (VIP - Verde):** Clientes con alta frecuencia y gasto. Poseen una recencia promedio de solo 13 días, demostrando ser el grupo más leal y rentable.
- **Cluster 2 (Potencial - Amarillo):** Clientes con actividad moderada. Representan la mayor oportunidad de crecimiento mediante estrategias de *upselling*.

- **Cluster 0 (En Riesgo - Morado):** Clientes con baja actividad y una recencia superior a 170 días. Requieren campañas urgentes de reactivación.

Perfil de Cliente	Recencia Prom.	Frecuencia Prom.	Monetario Prom.
VIP	13.1 días	261.8	$6,523.9
Potencial	69.4 días	66.1	$1,169.8
En Riesgo	171.3 días	14.9	$294.2

Cuadro 1: Resumen estadístico de los segmentos identificados.

5. Consideraciones Éticas y Conclusión

El sistema garantiza la **privacidad** mediante la anonimización de datos y asegura la **equidad** al basar sus predicciones únicamente en hechos transaccionales, evitando sesgos demográficos. En conclusión, el modelo proporciona una herramienta objetiva para optimizar la retención de clientes y maximizar el valor de vida del consumidor de forma ética.