

Segmentación de clientes para una tienda online

Proyecto de aprendizaje no supervisado: Implementación del algoritmo K-Means

Kevin Iza

1. Introducción

Este informe de business intelligence representa como se diseñó la segmentación de clientes a una tienda online mediante el análisis de su comportamiento histórico. Se utiliza una arquitectura que combina SQL para la gestión de datos y aprendizaje no supervisado para la generación de perfiles estratégicos.

2. Metodología

El proyecto se desarrolló siguiendo un flujo de trabajo de ingeniería de datos y ciencia de datos estructurado en cuatro fases:

1. **Fase 1: Extracción de Datos:** Se extrajo el conjunto de datos crudo desde una plataforma de datos abierta. Los registros incluyen el histórico de transacciones, descripciones de productos y fechas.
2. **Fase 2: Arquitectura SQL:** Se implementó una base de datos SQL local para el almacenamiento y estructuración de la información. Se transformaron los registros transaccionales en una tabla de entidades de cliente basada en el modelo **RFM** (Recency, Frequency, Monetary).
3. **Fase 3: Preprocesamiento de datos:** Para garantizar la convergencia del algoritmo, se aplicó una transformación logarítmica para suavizar el efecto de la asimetría y una normalización (Z-score), asegurando que cada variable contribuya equitativamente a la distancia euclídea.
4. **Fase 4: Validación del Modelo:** Se utilizó el algoritmo **K-Means**. El número de clusters se validó mediante el método del Codo, seleccionando $K = 3$ como el punto óptimo donde la inercia intracluster se estabiliza.

3. Análisis de la distribución de los datos

3.1. Exploración inicial

Antes de proceder con el modelo, se realizó un análisis exploratorio de datos (EDA). Como se observa en la Figura 1, los datos originales presentan una fuerte asimetría positiva. En el contexto comercial, esto indica que la mayoría de los clientes realizan compras pequeñas, mientras que una minoría representa una parte significativa de los ingresos de la tienda.

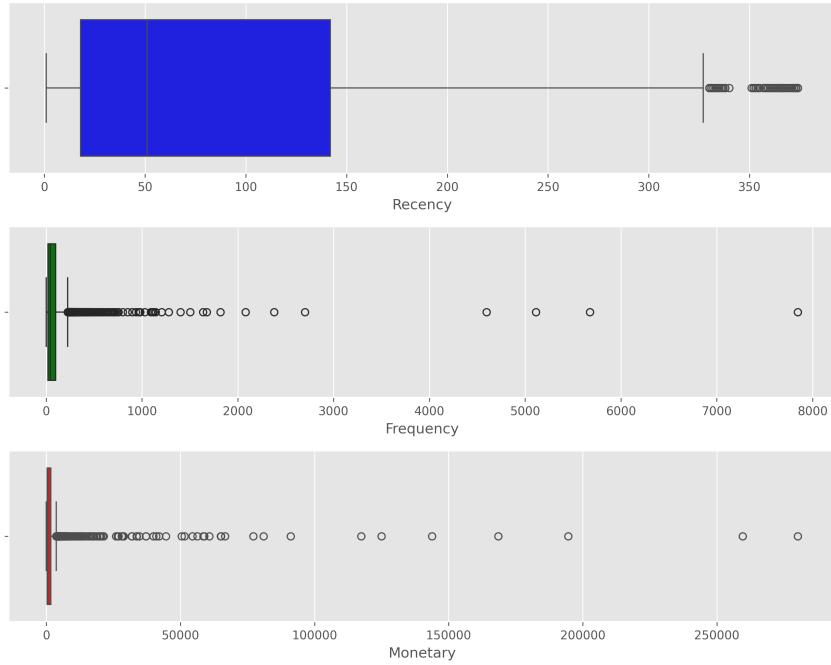


Figura 1: Boxplot de las métricas RFM originales: identificación de valores atípicos.

3.2. Optimización del hiperparámetro K

La selección del número de clusters se basó en la Figura 2. En la gráfica se identificó 3 clusters, suficiente para mantener la interpretabilidad estadística y operativa.

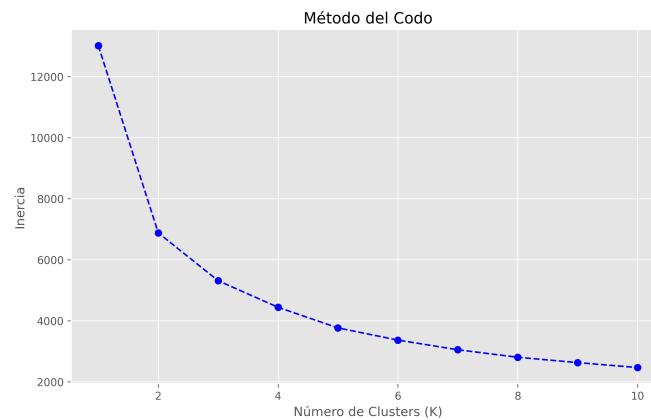


Figura 2: Método del Codo.

4. Interpretación de segmentos

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

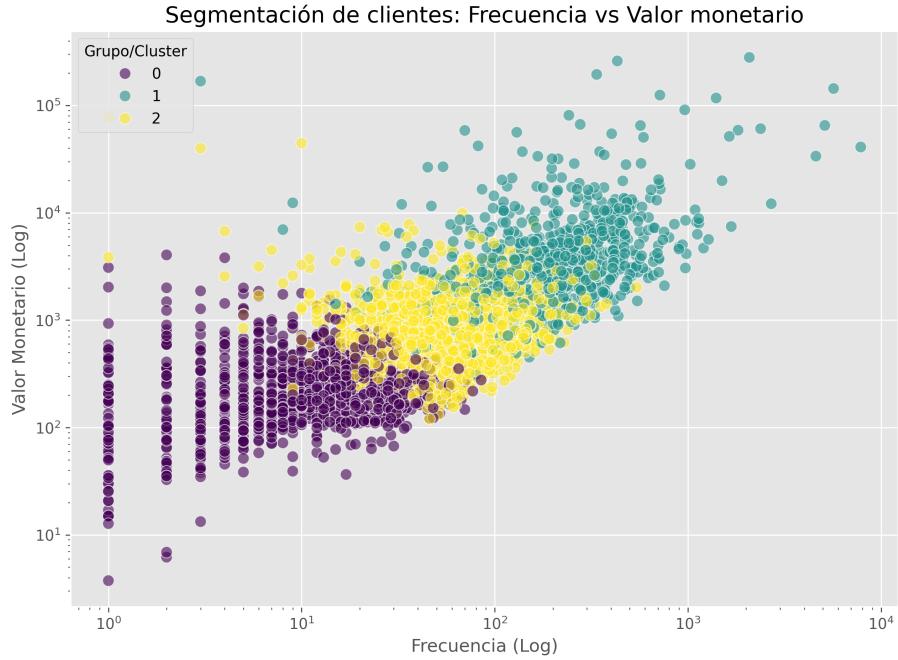


Figura 3: Visualización de clusters (Escala Logarítmica).

- **Cluster 1 (VIP):** Clientes de alto valor con una recencia promedio de 13 días.
- **Cluster 2 (Potencial):** Clientes con actividad moderada.
- **Cluster 0 (Ocasional):** Clientes con baja actividad y alta recencia (> 170 días).

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
Ocasional	171.3 días	14.9	\$294.2

Cuadro 1: Resumen estadístico de los clusters identificados.

5. Aplicaciones estratégicas y recomendaciones

El valor fundamental de este informe radica en la capacidad de convertir los segmentos identificados en acciones comerciales concretas. A continuación, se detallan las estrategias recomendadas:

5.1. Estrategias de marketing y ventas

- **Fidelización VIP (Cluster 1):** Implementar un programa de recompensas exclusivas o acceso anticipado a nuevos productos. El objetivo es mantener su baja recencia y premiar su lealtad para evitar que migren a la competencia.
- **Desarrollo de Clientes Potenciales (Cluster 2):** Utilizar técnicas de Cross-selling y Upselling. Al tener una actividad moderada, estos clientes responden bien a recomendaciones personalizadas basadas en sus compras anteriores.

- **Reactivación de Ocasionales (Cluster 0):** Ejecutar campañas de Win-back mediante cupones de descuento agresivos o encuestas de satisfacción para entender por qué dejaron de comprar.

5.2. Escalabilidad del modelo

Este marco de trabajo no se limita exclusivamente a una tienda online de retail; su arquitectura permite replicarlo en diversos sectores:

- **Sector Bancario:** Segmentación de usuarios según el uso de tarjetas de crédito.
- **SaaS y Suscripciones:** Análisis de usuarios basado en la frecuencia de uso de la plataforma y el nivel del plan contratado.
- **Empresas:** Clasificación de proveedores o clientes corporativos según el volumen de facturación y puntualidad.

6. 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. 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 y basada en evidencia científica.