Proyecto final análisis Exploratorio

Leopoldo Muñoz, Valentina Yáñez
Universidad de Talca
Facultad de Ingeniería
leom18@alumnos.utalca.cl, vyanez20@alumnos.utalca.cl

***Abstract*—This study explores the use of exploratory data analysis techniques and unsupervised clustering on a dataset of user sessions collected through Google Analytics. The influence of data points was analyzed using Cook's distance. The quality of the clusters was evaluated using the Hopkins statistic and the silhouette coefficient, applying algorithms such as K-Means, DBSCAN, Agglomerative Clustering, and GMM. The analysis revealed a tendency to form two clusters, enabling the exploration of differentiated strategies to improve conversion rates in e-commerce in the future.**

***Index Terms*—Clustering, PCA, K-Means, E-commerce, Análisis Exploratorio de Datos, Hopkins, DBSCAN**

I. Introducción

El análisis del comportamiento de usuarios en plataformas de e-commerce es un pilar fundamental para la mejora de estrategias comerciales. En este trabajo se analiza un dataset con más de 12.000 sesiones de navegación anónimas, con el objetivo de agruparlas según patrones comunes mediante técnicas de clustering no supervisado.

II. Descripción del Dataset

El dataset fue obtenido del UCI Machine Learning Repository. Contiene 12.330 instancias (una por sesión) y 18 atributos (entre ellos, número de páginas vistas, duración de visitas, tasas de rebote, día de la semana, tipo de visitante, etc.). La variable objetivo Revenue fue eliminada para permitir un análisis no supervisado.

III. Preprocesamiento de Datos

No se encontraron valores faltantes. Las variables categóricas fueron convertidas a tipo `category` y procesadas con one-hot encoding. Las numéricas fueron escaladas. Se utilizó `RobustScaler` para atributos con alta influencia (como `BounceRates` y `SpecialDay`) y `StandardScaler` para el resto.

IV. Tendencia a Agrupamiento

Se aplicó el estadístico de Hopkins, obteniéndose un valor de 0.9785, lo que indica una fuerte tendencia a formar clusters.

V. Reducción de Dimensionalidad

Se utilizó Análisis de Componentes Principales (PCA), conservando que los primeros 5 componentes explican el más del 85% de la varianza acumulada. Estas fueron usadas para visualizar los resultados de clustering.

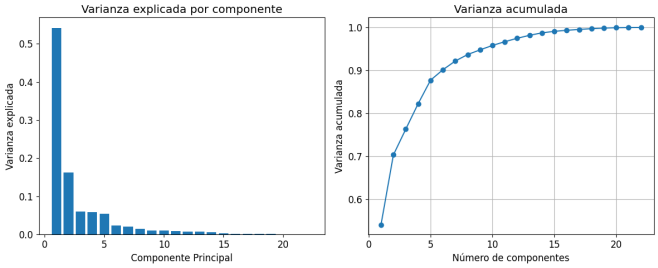

Fig. 1. Gráfico de varianza acumulada y explicada por componente.

VI. Determinación del Número Óptimo de Clusters

El método del codo y el coeficiente de silueta indicaron que $k = 2$ es una opción adecuada para separar los datos en grupos distintos.

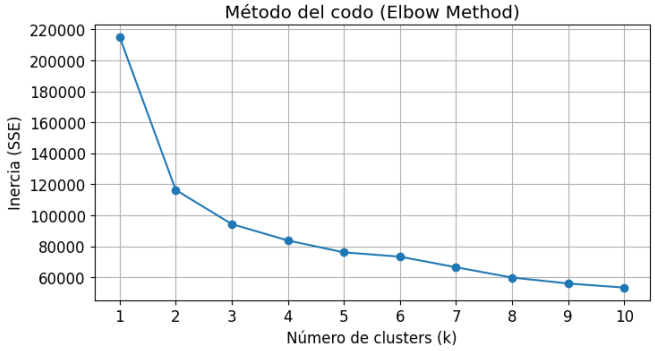

Fig. 2. Método del codo para determinar el número de clusters (k=2).

VII. Métodos de Clustering Aplicados

Se aplicaron los siguientes algoritmos:

- **K-Means:** produjo dos clusters bien separados pero muy polarizados.
- **DBSCAN:** detectó ruido y grupos densos.
- **Agglomerative Clustering:** identificó dos grupos similares a K-Means.
- **Gaussian Mixture Models (GMM):** generó una segmentación probabilística, identificando también patrones en forma de núcleo y siendo mucho más flexible.

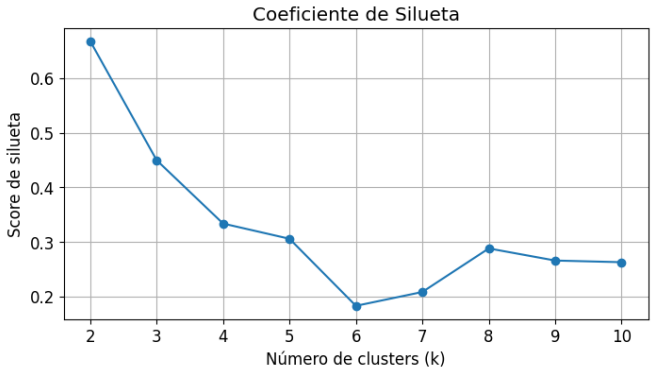

Fig. 3. Método de la silueta para determinar el número óptimo de clusters (k=2).

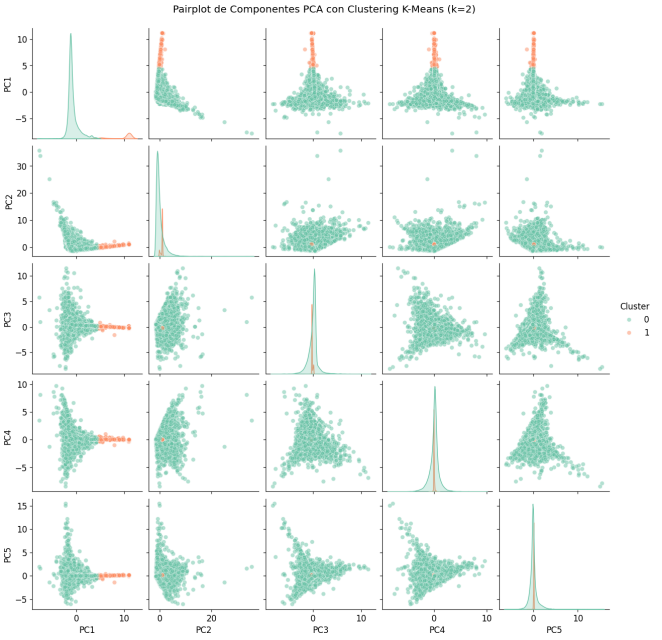

Fig. 4. Resultados de K-Means con k=2.

VIII. Análisis de Resultados

Los clusters encontrados se diferenciaron principalmente en las variables 'BounceRates','ExitRates' y 'SpecialDay'. Por lo que los resultados sugieren que un grupo corresponde a usuarios interesados y el otro a visitantes que abandonan rápidamente.

IX. Conclusiones

El análisis permitió identificar dos perfiles principales de usuario. .

References

[1] Sakar, C. O., et al. "Online Shoppers Purchasing Intention Dataset." UCI Machine Learning Repository, 2018.
[2] Han, J., Kamber, M., & Pei, J. "Data Mining: Concepts and Techniques." Morgan Kaufmann, 2011.

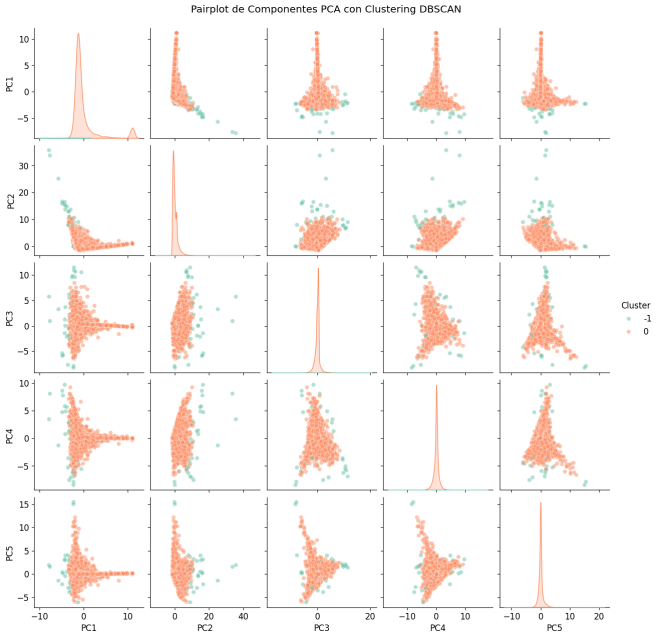

Fig. 5. Resultados de DBSCAN.

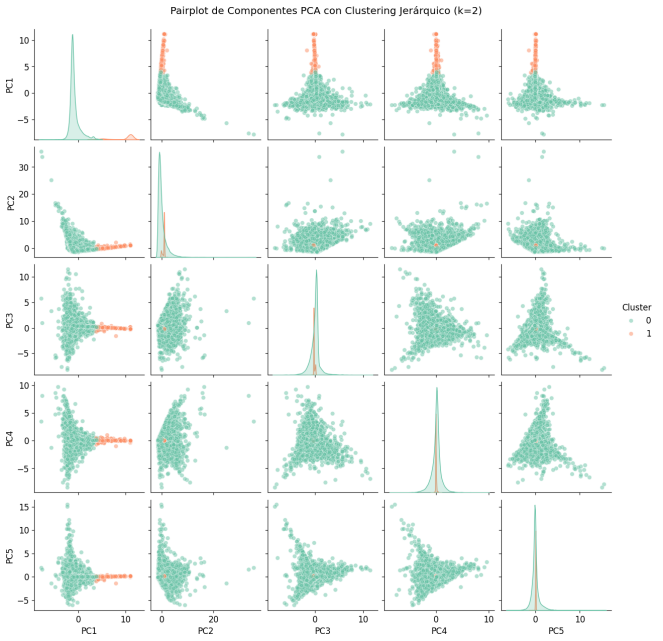

Fig. 6. Resultados de Agglomerative Clustering.

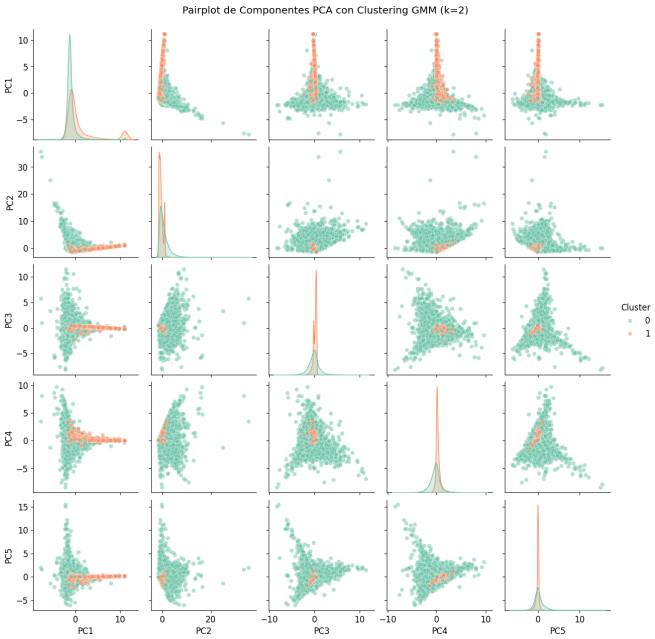

Fig. 7. Resultados de GMM.