Análisis Hierical Clustering

Durante el análisis del dataset, el objetivo fue agrupar a los clientes con base en tres características: edad, ingreso anual y puntaje de gasto, con el fin de identificar los distintos perfiles de clientes para permitir diseñar estrategias comerciales más personalizadas.

El dataset contiene información sobre 200 clientes de un centro comercial, con variables como Customer ID, Gender, Age, Annual Income y Spending Score. Para este análisis se descartaron las variables Customer ID y Gender, y se trabajó únicamente con Age, Annual Income y Spending Score, ya que estas variables brindan una base para detectar patrones de comportamiento.

Se procedió a normalizar los datos mediante StandardScaler, para evitar que las diferencias en escalas entre las variables. Luego, se aplicó clustering jerárquico aglomerativo utilizando la distancia euclidiana y el método de enlace ward.

El dendrograma generado permitió visualizar cómo se agrupan los datos y sugirió que un valor óptimo para el número de clusters podría estar entre 4 y 6 grupos, dependiendo del punto de corte. Esta visualización es esencial para interpretar la estructura jerárquica de los datos y tomar decisiones informadas sobre la segmentación.

Posteriormente, se ejecutó el modelo AgglomerativeClustering con 6 clusters, ya que este número ofrecía una segmentación más rica y detallada. Al graficar los resultados, se observó una clara diferenciación entre los grupos, destacando patrones específicos de ingreso y gasto. Por ejemplo:

Un grupo de clientes con alto ingreso y alto gasto, posiblemente los clientes más rentables.

Otro grupo con alto ingreso pero bajo gasto, que representa una oportunidad para fidelización o promoción.

Aunque este tipo de clustering no genera un centroide como K-Means, la separación de los grupos fue clara y coherente con la interpretación visual de los datos.

Evaluación del modelo y justificación del número de clusters

Para validar la calidad del modelo se utilizó el coeficiente de Silhouette, que mide qué tan bien se encuentran separados los grupos. En este caso, el valor fue 0.43, lo que indica una separación aceptable entre clusters (valores mayores a 0.4 suelen ser considerados razonables en problemas reales con datos no perfectamente definidos).

Este resultado, junto con la estructura visual del dendrograma y los patrones claros observados en la dispersión de datos, respalda la decisión de utilizar 6 clusters como la mejor segmentación.

El uso de clustering jerárquico permitió una segmentación efectiva y visualmente interpretativa de los clientes. La elección de 6 grupos no solo se basa en la métrica Silhouette, sino también en el análisis visual del dendrograma y la claridad en los patrones detectados.

El análisis realizado proporciona una base sólida para tomar decisiones comerciales más informadas y segmentadas dentro del mall, maximizando el valor de cada perfil de cliente identificado. ")