Clustering de Clientes en un Centro Comercial: Análisis Comparativo de K-means y Clustering Jerárquico

Nombre del Estudiante

Mayo 2025

Abstract

Este informe presenta un análisis de segmentación de clientes utilizando dos algoritmos de clustering: K-means y Clustering Jerárquico, aplicados al conjunto de datos *Mall_Customers.csv*. El objetivo es identificar grupos de clientes basados en su ingreso anual y puntaje de gasto para informar estrategias de marketing. Se describe el procedimiento, incluyendo manejo de datos, preprocesamiento, entrenamiento y evaluación de los modelos. Los resultados muestran cinco segmentos de clientes bien definidos, con métricas de evaluación que indican buena separación y cohesión. Se comparan ambos algoritmos y se proporcionan recomendaciones para su uso en segmentación de clientes.

1 Introducción

El análisis de segmentación de clientes es fundamental en la gestión de centros comerciales, ya que permite identificar grupos homogéneos de clientes para personalizar estrategias de marketing. Este informe utiliza el conjunto de datos *Mall_Customers.csv* para aplicar dos algoritmos de clustering no supervisado: K-means y Clustering Jerárquico. El objetivo es segmentar a los clientes según su ingreso anual y puntaje de gasto, evaluar la calidad de los clusters y comparar el desempeño de ambos algoritmos. Se describen el procedimiento, los resultados, la evaluación y la interpretación, junto con recomendaciones prácticas.

2 Descripción del Conjunto de Datos

El conjunto de datos *Mall_Customers.csv* contiene información de 200 clientes de un centro comercial. Las variables incluidas son:

- **CustomerID**: Identificador único de cada cliente (eliminado para el análisis).
- **Gender**: Género del cliente (Male o Female).
- **Age**: Edad del cliente (en años).
- **Annual Income (k$)**: Ingreso anual del cliente (en miles de dólares).
- **Spending Score (1-100)**: Puntaje de gasto asignado por el centro comercial (1 a 100).

El análisis exploratorio de datos (EDA) reveló que no hay valores faltantes, con distribuciones aproximadamente normales para *Annual Income* y *Spending Score*, y una ligera asimetría en *Age*. La correlación entre las variables numéricas es baja, lo que sugiere que *Annual Income* y *Spending Score* son adecuadas para el clustering.

3 Metodología

El análisis siguió un procedimiento estructurado en siete pasos, utilizando Python con bibliotecas como `pandas`, `scikit-learn`, `seaborn` y `matplotlib`.

3.1 Análisis Exploratorio de Datos (EDA)

Se realizaron visualizaciones para entender las distribuciones y relaciones entre variables:

- Histogramas para *Age*, *Annual Income* y *Spending Score*.
- Gráficos de dispersión para explorar relaciones por género.
- Boxplots para detectar valores atípicos.
- Pair plots y mapas de calor de correlación.
- Violin plots para comparar *Annual Income* y *Spending Score* por género.

No se encontraron valores faltantes, por lo que no fue necesario imputar datos.

3.2 Preprocesamiento de Datos

- Se eliminó la columna *CustomerID*, ya que no aporta al clustering.
- La variable *Gender* se codificó como binaria (Male=0, Female=1).
- Las variables numéricas (*Age*, *Annual Income*, *Spending Score*) se estandarizaron (media=0, desviación estándar=1) usando `StandardScaler`.

3.3 Selección de Características

Basado en el EDA, se seleccionaron *Annual Income (k$)* y *Spending Score (1-100)* como características principales, ya que mostraron patrones claros de segmentación en los gráficos de dispersión.

3.4 Entrenamiento del Modelo

- **K-means**:
  - Se aplicó el método del codo para determinar el número óptimo de clusters ($k = 5$).
  - Se entrenó el modelo con $k = 5$ y semilla aleatoria 42.
- **Clustering Jerárquico**:

- Se generó un dendrograma con el método de enlace *ward* y métrica euclidiana, confirmando 5 clusters.
- Se entrenó el modelo con `AgglomerativeClustering` ($n = 5$, `metric='euclidean'`, `linkage='ward'`).

3.5 Evaluación del Modelo

Se calcularon dos métricas para evaluar la calidad de los clusters:

- **Coeficiente de Silhouette**: Mide la separación y cohesión de los clusters (rango: -1 a 1).
- **Índice de Calinski-Harabasz**: Mide la relación entre la dispersión dentro de los clusters y entre ellos (valores más altos indican mejor clustering).

3.6 Visualización e Interpretación

Se generaron gráficos de dispersión de clusters, barras apiladas de distribución por género y boxplots por cluster. Los resultados se interpretaron para caracterizar los segmentos de clientes.

4 Resultados

Ambos modelos identificaron cinco clusters con patrones similares, basados en *Annual Income* y *Spending Score*. A continuación, se describen los resultados clave.

4.1 K-means

- **Método del Codo**: Indicó $k = 5$ como el número óptimo de clusters.
- **Centroides**: Los centroides definieron claramente los segmentos (ver Tabla 1).
- **Gráficos**: Los clusters muestran separación clara en el espacio de *Annual Income* vs. *Spending Score*.

Table 1: Centroides de los Clusters (K-means, estandarizados)

Cluster	Annual Income (k$)	Spending Score (1-100)
0	1.05	-0.44
1	-0.42	-0.42
2	-0.44	1.27
3	1.08	1.27
4	-1.19	-1.18

4.2 Clustering Jerárquico

- **Dendrograma**: Confirmó 5 clusters como una elección razonable.
- **Etiquetas**: Los clusters son consistentes con los de K-means, con pequeñas diferencias en la asignación de puntos.
- **Gráficos**: Similar separación visual a K-means, validada por el dendrograma.

4.3 Distribución por Género

La distribución de género por cluster mostró que no hay una segregación significativa por género, lo que sugiere que los segmentos están más definidos por ingresos y gastos que por género.

5 Evaluación

Las métricas de evaluación para ambos modelos se presentan en la Tabla 2.

Table 2: Métricas de Evaluación de los Modelos

Modelo	Coeficiente de Silhouette	Índice de Calinski-Harabasz
K-means	0.553	248.649
Hierarchical Clustering	0.541	241.357

- **Coeficiente de Silhouette**: Ambos modelos tienen valores cercanos a 0.55, indicando buena separación y cohesión. K-means es ligeramente superior (0.553 vs. 0.541).
- **Índice de Calinski-Harabasz**: K-means también supera ligeramente al modelo jerárquico (248.649 vs. 241.357), sugiriendo clusters más compactos.

6 Interpretación

Los cinco clusters identificados representan segmentos de clientes con características distintas:

1. **Cluster 1 (Premium)**: Alto ingreso, alto gasto. Clientes ideales para productos de lujo.
2. **Cluster 2 (Conservadores)**: Alto ingreso, bajo gasto. Potenciales para promociones que incentiven el gasto.
3. **Cluster 3 (Impulsivos)**: Bajo ingreso, alto gasto. Sensibles a ofertas y descuentos.
4. **Cluster 4 (Económicos)**: Bajo ingreso, bajo gasto. Enfocados en productos de bajo costo.
5. **Cluster 5 (Promedio)**: Ingreso y gasto medios. Clientes versátiles para campañas generales.

La consistencia entre K-means y Clustering Jerárquico valida la robustez de los segmentos. La baja correlación entre variables y la ausencia de segregación por género sugieren que las estrategias de marketing deben centrarse en ingresos y patrones de gasto.

7 Visualización Comparativa

(a) K-means (b) Clustering Jerárquico

Figure 1: Comparación visual de los clusters generados por ambos modelos (imágenes no incluidas en este documento).

Ambos modelos producen clusters visualmente similares, con K-means mostrando bordes más definidos debido a su enfoque basado en centroides, mientras que el Clustering Jerárquico ofrece una perspectiva jerárquica útil para explorar subgrupos.

8 Limitaciones

- El dataset es pequeño (200 clientes), lo que limita la generalización.
- Solo se usaron dos variables (*Annual Income* y *Spending Score*), omitiendo *Age* y *Gender*, que podrían enriquecer los clusters.
- La elección de 5 clusters, aunque justificada, podría ajustarse según objetivos específicos.

9 Conclusiones y Recomendaciones

Ambos algoritmos identificaron cinco segmentos de clientes robustos, con K-means mostrando un rendimiento ligeramente mejor en las métricas de evaluación. Se recomienda:

- Usar K-means para segmentaciones rápidas y efectivas en datasets similares.
- Implementar Clustering Jerárquico cuando se desee explorar relaciones jerárquicas entre clientes.
- Diseñar estrategias de marketing específicas para cada segmento (e.g., promociones para el Cluster 2, productos económicos para el Cluster 4).
- En futuros análisis, incorporar *Age* y más variables para enriquecer los clusters.

10 Referencias

References

[1] Scikit-learn: Machine Learning in Python, https://scikit-learn.org.

[2] Seaborn: Statistical Data Visualization, https://seaborn.pydata.org.