**Modelo K-Means**

**Enlace GitHub:** <https://github.com/Valenbg17/Mall-Customers_Modelo-K-means>

**Análisis e interpretación de resultados:**

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 distintos perfiles de clientes que permitan diseñar estrategias comerciales más personalizadas.

El dataset contiene información sobre 200 clientes de un mall, con características como Customer ID, Gender, Age, Annual Income y Spending Score. Las variables que se escogieron para analizar con el modelo K-Means fueron Age, Annual Income y Spending Score.

El gráfico de pares demuestra que hay patrones y grupos diferenciales en los datos, especialmente en la relación entre Annual Income y Spending Score. La dispersión de los puntos permite observar que en la variable Age hay una concentración de clientes entre los 25 y 40 años y hay una caída progresiva en clientes mayores. En Annual Income la mayoría de ingresos están entre $40k y $80k, con menos clientes con ingresos más altos o bajos. Finalmente, Spending Score tiende a ser más uniforme, pero tiene una concentración mayor entre 40 y 60, lo que demuestra que hay clientes con un gasto medio.

El análisis de los diagramas de caja de las variables Annual Income y Spending Score revela patrones importantes para la agrupación de clientes. La mayoría de los clientes presentan ingresos anuales entre $40k y $80k, mientras que los niveles de gasto están distribuidos de forma más uniforme a lo largo del rango posible. Sin embargo, se identificó un valor atípico en el ingreso anual, correspondiente a un cliente con ingresos significativamente más altos que el resto. Por otro lado, el Spending Score no presenta outliers y muestra una buena dispersión que refleja la diversidad en los hábitos de consumo dentro de la base de clientes.

La grafica de codo permitió escoger el valor de K, en la gráfica las mejores opciones eran 4 y 5 pero en este caso el valor optimo fue 6, ya que, al momento de analizar el rendimiento del modelo, este número demostró mejores resultados. Este valor es muy importante en el análisis porque permite tomar una decisión fundamentada sobre cuantos grupos o clusters usar al momento de aplicar K-Means, mejorando la precisión.

Al interpretar los gráficos de dispersión, se obtuvo la siguiente información:

Al tener K= 4, se observan cuatro grupos bien definidos con sus respectivos centroides. Aunque los grupos están relativamente separados, hay cierta superposición entre los clientes con ingresos medios y puntuaciones de gasto intermedias. Este número de clusters permite una segmentación general que puede diferenciar entre clientes de alto, medio y bajo gasto e ingreso, pero puede no capturar todos los matices del comportamiento del cliente.

Con K=5, se observa que uno de los clusters grandes del caso anterior se ha subdividido, permitiendo distinguir mejor ciertos patrones, especialmente entre consumidores con ingresos similares, pero niveles de gasto distintos. Esto mejora la precisión en la identificación de perfiles de clientes, como, por ejemplo, quienes ganan mucho pero gastan poco, y viceversa.

Al tener k=6, hay una segmentación aún más detallada, destacando grupos de clientes con patrones de consumo e ingreso similares. Sin embargo, también comienza a haber cierta dispersión de algunos grupos.

Finalmente, al usar métricas para evaluar la calidad del modelo, se obtienen los siguientes resultados:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Número de Clusters** | **Coeficiente Silhouette** | **Índice Calinski-Harabasz** | **Índice Davies-Bouldin** | **Inertia** |
| 4 | 0.41 | 125.1 | 0.87 | 168.2 |
| 5 | 0.41 | 124 | 0.87 | 169.3 |
| 6 | 0.43 | 135.1 | 0.82 | 133.9 |

Entre más cercano sea el coeficiente Silhoutte a 1 es mejor, el índice Calinski Harabasz si es más alto indica una mejor separación, por el contrario, el Índice Davies-Bouldin al igual que Inertia tienen un mayor desempeño cuando sus valores son bajos.  Según las métricas obtenidas, el modelo con 6 clusters demuestra un mejor rendimiento, ya que presenta el mayor coeficiente Silhouette (0.43), el mayor índice Calinski-Harabasz (135.1), y los valores más bajos tanto para el índice Davies-Bouldin (0.82) como para la Inertia (133.9). Esto sugiere que la segmentación con 6 grupos ofrece una estructura más clara, compacta y bien diferenciada, lo cual la convierte en la opción más adecuada para interpretar los patrones de comportamiento de los clientes.