**TAREA 4 Algoritmos de Aprendizaje No Supervisado**

**Israel David Gómez Camacho**

**CÓDIGO: 1127585391**

**202016908 - Análisis de Datos**

**Grupo - 74**

**Docente**

**Jhon Manuel Soto**

**UNIVERSIDAD NACIONAL ABIERTA Y A DISTANCIA**

**ECBTI – Ingeniería de Sistemas**

**09 de Mayo del 2025**

**Introducción**

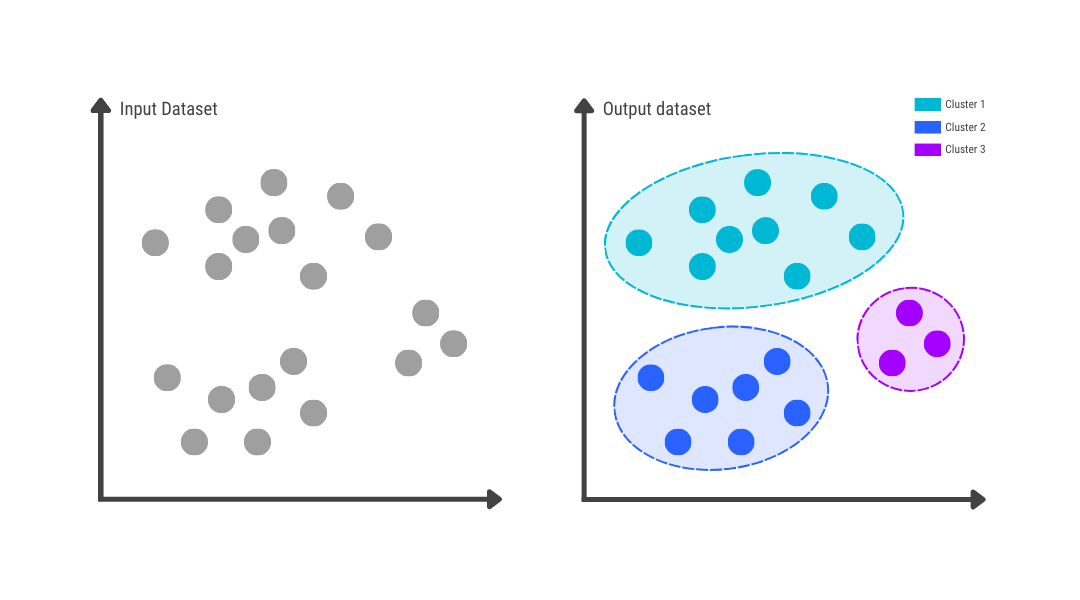
**Objetivos**

**Cuadro sinóptico Aprendizaje no supervisado**

**Listado de definiciones**

A continuación, se explicarán unos conceptos de los términos más importantes en cuanto al aprendizaje no supervisado.

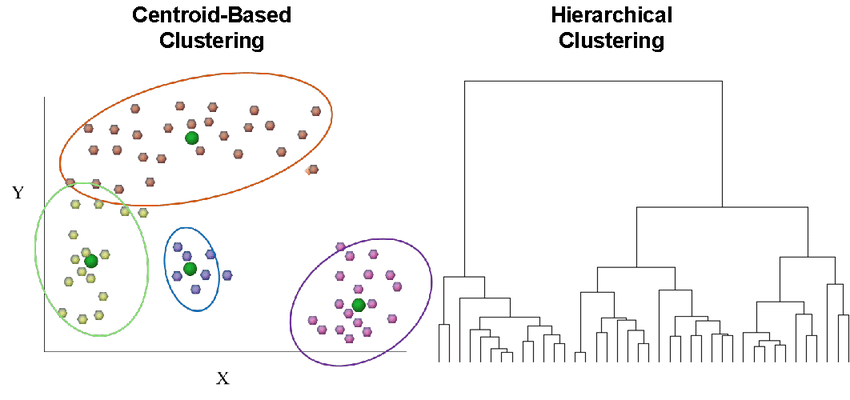
1. **Clustering.**



**Pi.Exchange. (2023, 4 diciembre). Understanding Clustering in Machine Learning: Algorithms and Use Cases. Pi.Exchange.** [**https://www.pi.exchange/blog/clustering-in-machine-learning**](https://www.pi.exchange/blog/clustering-in-machine-learning)

El clustering es la técnica de aprendizaje Arts y aprendizaje no supervisado la cual agrupa los datos que son muy similares en sus clústeres y no es necesario crear las etiquetas. Está es mayormente usada para encontrar estructuras que están ocultas en los datos.

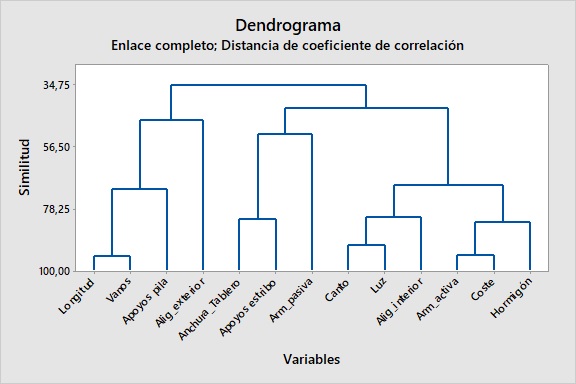
1. **Centroide**



**(S/f). Researchgate.net. Recuperado el 10 de mayo de 2025, de** [**https://www.researchgate.net/figure/An-illustration-comparing-centroid-based-clustering-and-hierarchical-clustering-adapted\_fig2\_328924647**](https://www.researchgate.net/figure/An-illustration-comparing-centroid-based-clustering-and-hierarchical-clustering-adapted_fig2_328924647)

El centroide es el punto central de un clúster en el algoritmo, se puede utilizar como K-means este representa la ubicación promedio en todos los puntos que contiene el clúster.

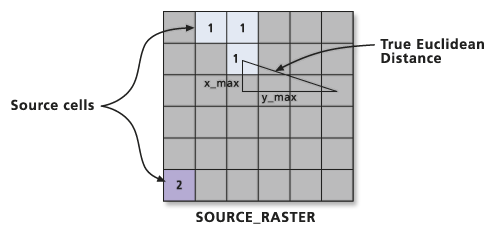
1. **Dendograma**



**dendrograma – El blog de Víctor Yepes. (2020, 21 noviembre).** [**https://victoryepes.blogs.upv.es/tag/dendrograma/**](https://victoryepes.blogs.upv.es/tag/dendrograma/)

El dendograma es un diagrama que es en forma de árbol donde se muestra el tipo de jerarquía de los clústeres que fueron formados durante el clustering jerárquico. Se pueden visualizar los puntos que se agrupan en diferentes niveles de similitud.

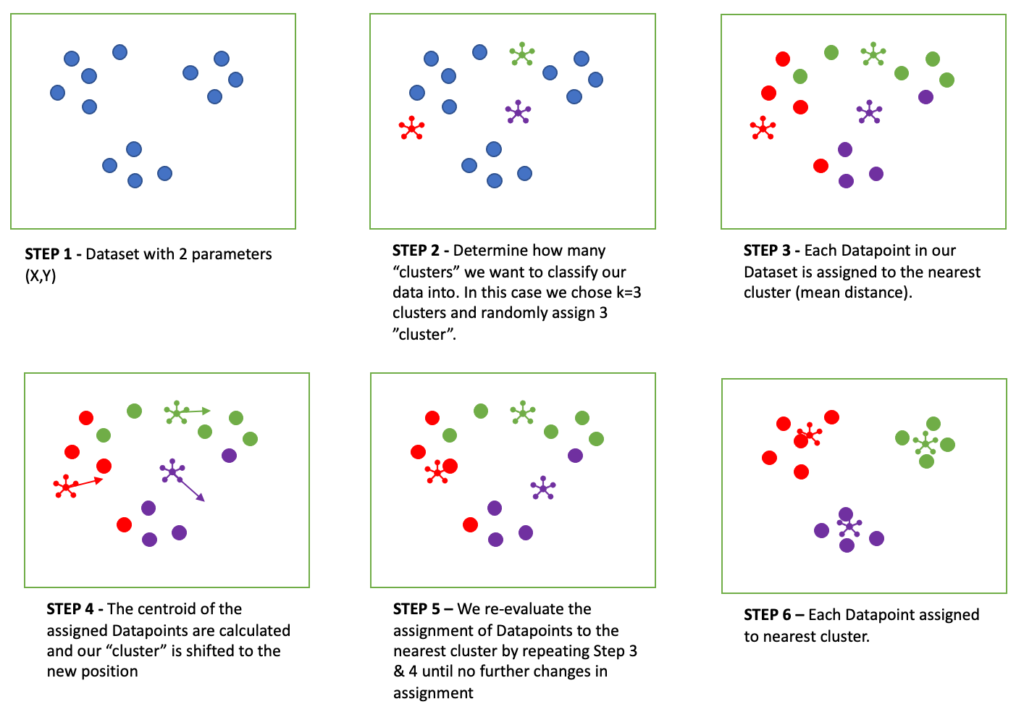
1. **Distancia euclidean**



**Comprender el análisis de distancia euclidiana—ArcMap | Documentación. (s. f.).** [**https://desktop.arcgis.com/es/arcmap/latest/tools/spatial-analyst-toolbox/understanding-euclidean-distance-analysis.htm**](https://desktop.arcgis.com/es/arcmap/latest/tools/spatial-analyst-toolbox/understanding-euclidean-distance-analysis.htm)

Es la medida en línea recta de la distancia entre los 2 puntos que se miden en un espacio multidimensional, la cual se calcula como la raíz cuadrada de la suma de las diferencias al cuadrado de sus coordenadas.

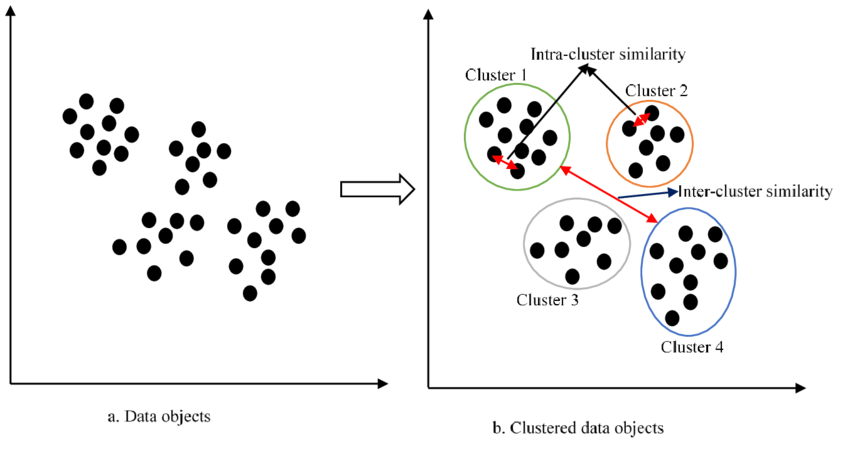
1. **Dispersión intra-cluster**



**Suruchi. (2023, 26 diciembre). Exploring the Art of Data Segmentation with K-Means Clustering. Medium. <https://medium.com/@suruchiarora257/exploring-the-art-of-data-segmentation-with-k-means-clustering-404f34028f5c>**

Es la función que mide cuan compactos son los puntos que se encuentran dentro del clúster, casi siempre se calcula como la suma de cada una de las distancias elevado al cuadrado la cual va desde cada uno de los puntos al centroide del clúster.

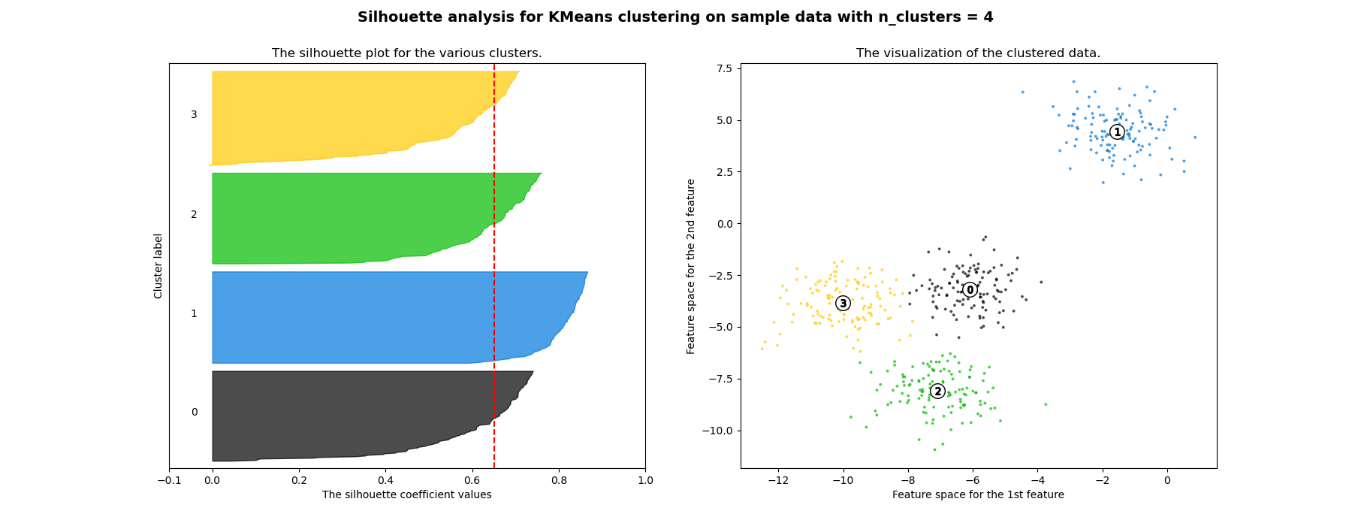
1. **Dispersión inter-cluster**



**(S/f). Researchgate.net. Recuperado el 10 de mayo de 2025, de** [**https://www.researchgate.net/figure/Clustering-example-with-intra-and-inter-clustering-illustrations\_fig1\_344590665**](https://www.researchgate.net/figure/Clustering-example-with-intra-and-inter-clustering-illustrations_fig1_344590665)

La dispersión inter-cluster es la función que mide la distancia o cuan están separados los clústeres entre sí, es decir, la distancia que se encuentra entre cada segmentación, esta se basa en la distancia de los centroides de cada uno de los clústeres.

1. **Coeficiente de Silhouette**

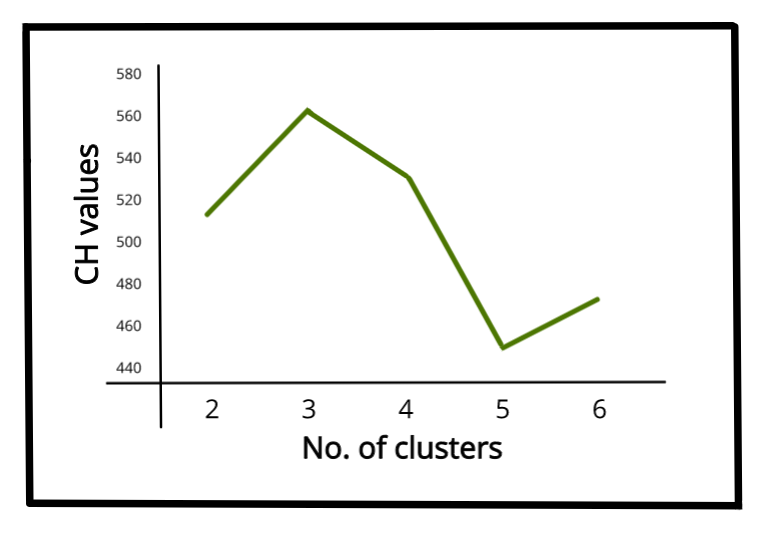


**Selección del número de conglomerados con el análisis de silueta en el agrupamiento KMedias (KMeans) — documentación de scikit-learn - 0.24.1. (s. f.).** [**https://qu4nt.github.io/sklearn-doc-es/auto\_examples/cluster/plot\_kmeans\_silhouette\_analysis.html**](https://qu4nt.github.io/sklearn-doc-es/auto_examples/cluster/plot_kmeans_silhouette_analysis.html)

Es la métrica la cual evalúa la calidad de un clustering la cual mide que tan similar es un punto a su propio clúster que en términos simple es la distancia de a la cohesión frente a clústeres de la separación.

Las unidades de medición varían de -1 a 1, cuando los valores son cercanos a 1 indican que el clustering está bien desarrollado.

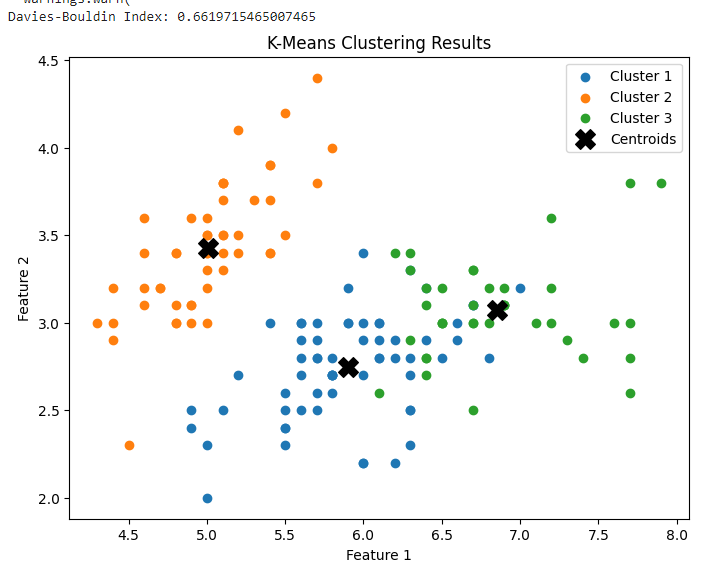
1. **Índice de CalinskiHarabasz**



**GeeksforGeeks. (2022, 25 abril). CalinskiHarabasz Index – Cluster Validity indices | Set 3. GeeksforGeeks.** [**https://www.geeksforgeeks.org/calinski-harabasz-index-cluster-validity-indices-set-3/**](https://www.geeksforgeeks.org/calinski-harabasz-index-cluster-validity-indices-set-3/)

Este índice es una métrica que evalúa la calidad que contiene el clustering la cual se compara con la dispersión entre los clústeres con la dispersión que se encuentra dentro de los clústeres, mientras los valores sean más altos el modelo de clustering estará mejor desarrollado.

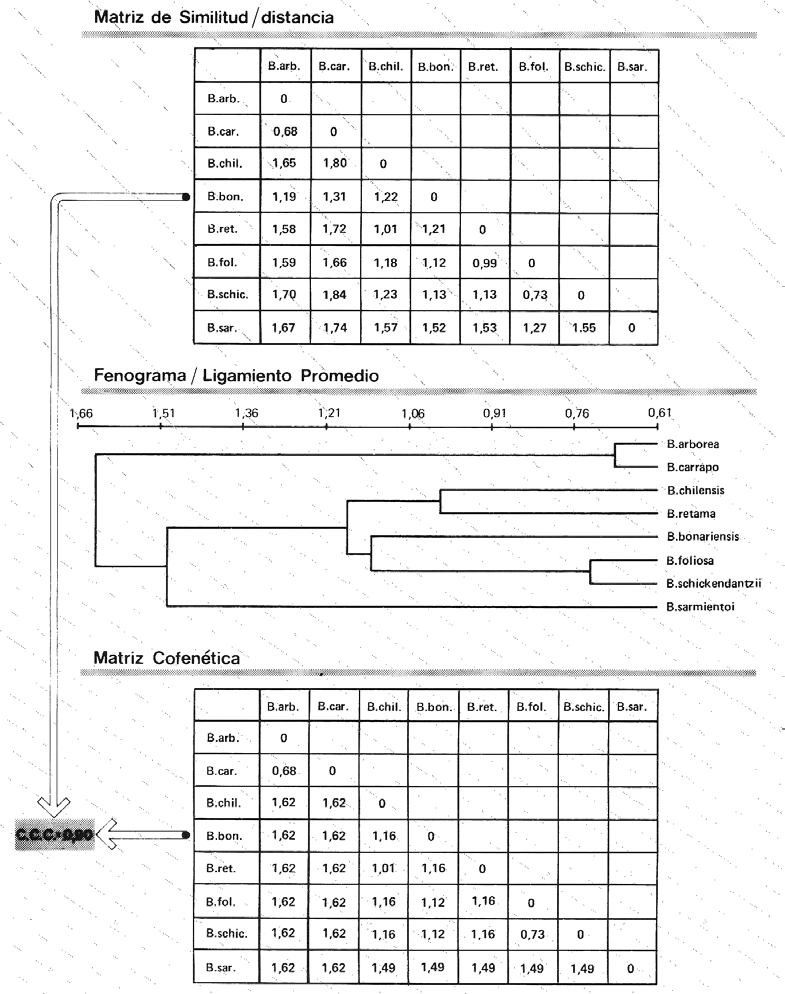
1. **Índice de Davies-Bouldin**



**GeeksforGeeks. (2023, 5 noviembre). DaviesBouldin Index. GeeksforGeeks.** [**https://www.geeksforgeeks.org/davies-bouldin-index/**](https://www.geeksforgeeks.org/davies-bouldin-index/)

Este índice es la métrica que mide la calidad del clustering que va promediando según la relación entre la dispersión intra-cluster y la separación del inter-cluster, este caso es al contrario de CalinskiHarabasz, ya que entre más bajos indican que el clustering está bien desarrollado.

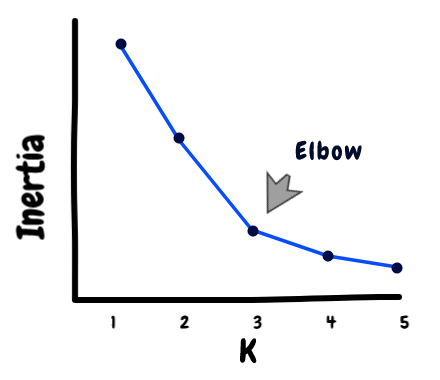
1. **Coeficiente de correlación cofenética**



**(S/f). Researchgate.net. Recuperado el 10 de mayo de 2025, de** [**https://www.researchgate.net/figure/Coeficiente-de-correlacion-cofenetica-matriz-de-similitud-original-distancia-entre\_fig18\_50326234**](https://www.researchgate.net/figure/Coeficiente-de-correlacion-cofenetica-matriz-de-similitud-original-distancia-entre_fig18_50326234)

Este coeficiente es la métrica que mide que tan bien el clustering jerárquico (dendograma), preserva las distancias originales entre cada uno de los puntos, si los valores se acercan a 1 estos informan que hay fidelidad.

1. **Inertia**



**Zhu, A. (2025, 24 enero). 7 Most Asked Questions on K-Means Clustering. Towards Data Science.** [**https://towardsdatascience.com/explain-ml-in-a-simple-way-k-means-clustering-e925d019743b/**](https://towardsdatascience.com/explain-ml-in-a-simple-way-k-means-clustering-e925d019743b/)

Esta es la suma de cada una de las distancias elevadas al cuadrado de cada uno de los puntos hacia el centroide de su clúster en este caso para K-means, para la inertia es recomendable que sea menor en los clústeres ya que son más compactos.

**Link repositorio GitHub códigos de modelo**

**Interpretación de los resultados**

Antes de realizar el análisis e interpretación de los resultados obtenidos por los 2 modelos de clustering, hay que entender que datos tenía el dataset, problemas o situaciones que se podían o se presentaron.

En este caso se utilizó el dataset Mall\_Customers.csv, en este dataset se realiza el análisis de la segmentación de clientes, este análisis es muy importante la correcta gestión de los centros comerciales, identificación de grupos y subgrupos para las estrategias de marketing. Para realizar el proceso de identificación y análisis de datos se implementaron 2 modelos de algoritmos no supervisados los cuales son: K-means y hierarchical clustering.

El objetivo principal es la segmentación de los clientes según su ingreso anual y el puntaje en que gasta, evaluación de la calidad de los clusters y la comparación de los 2 algoritmos utilizados.

**Descripción de los conjuntos de datos**

El dataset Mall\_Customers.csv contiene la información de 200 clientes de un centro comercial, se incluyen variables importantes como:

* CustomerID: El cual es el identificador único de cada uno de los clientes
* Age: Edad del cliente en años.
* Gender: Género del cliente si es Male o Female.
* Annual Income (k$): Es el ingreso anual del cliente el valor es en miles de dólares.
* Spending Score (1-100): Es el puntaje de gasto asignado por el centro comercial.

Realizando el análisis exploratorio (EDA) se evidenció que no hay valores faltantes, además, la correlación que existe entre las variables numéricas es baja, lo que sugiere que Annual Income y Spending Score son adecuadas para el uso de clustering.

La metodología utilizada para el procedimiento de la generación de graficas, resultados, entrenamiento del modelo entre otros, se siguió el reglamento estructurado de 7 pasos, además para la generación del código de Python se utilizaron varias bibliotecas como pandas, seaborn, scikit-learn y matplotlib.

1. **EDA.**

Para el análisis exploratorio EDA se realizaron varios gráficos esto con el fin de entender las distribuciones y relaciones que existe entre las variables:

* Se generaron histogramas para: Age, Annual Income y Spending Score.
* Se crearon los gráficos de dispersión y la exploración de las relaciones por género.
* Se generaron boxplots para la detección de valores atípicos que se podían encontrar.
* Se realizan gráficos de pair plots y algunos mapas de calor en cuanto a la correlación.
* Se realizaron varios gráficos de violin plots para comparar las variables Annual Income y Spending Score por su género.

1. **Preprocesamiento de Datos**

Para el procesamiento de los datos se debía realizar un análisis de las variables y como estas podían influir para el algoritmo ya sea positiva o negativamente. Para esto se realizaron varios procedimientos:

* Se eliminó la columna CustomerID, ya que esta no aportaba gran valor al clustering.
* La variable Gender se codificó como binaria (Male=0, Female=1), esto para no tener errores.
* Las variables numéricas como: Age, Annual Income y Spending Score se estandarizaron en categorías: media = 0, desviación estándar = 1 usando la función de Python StandardScaler.

1. **Selección de Características**

Basándonos en el análisis exploratorio (EDA), se seleccionaron las variables Annual Income (k$) y Spending Score (1-100) como las características principales, ya que estas 2 mostraron patrones claros en la segmentación de los gráficos de dispersión. Por tal motivo se utilizaron estas 2 variables.

1. **Entrenamiento del modelo**

Se realizaron 2 entrenamientos para los modelos de los algoritmos, entonces:

* **K-means:** Se aplicó el método del codo con el fin de determinar el número de clusters el cual es Se entrenó el modelo con y aleatorio 42.
* **Hierarchical clustering:** Se creo un dendograma con el método de enlace Ward y la métrica euclidiana, confirmando los 5 clusters. Se entrenaron los modelos con el código de Python: AgglomerativeClustering (n = 5, metric=’euclidean’,linkage=’ward’).

1. **Evaluación del modelo**

Se calcularon dos métricas con el fin de evaluar la calidad de los clusters creados:

* **Coeficiente de Silhouette:** Este mide la separación y la cohesión de los clusters entre los rangos: -1 a 1.
* **Índice de Calinski-Harabasz:** Este mide la relación entre la dispersión dentro de los clusters y entre ellos, mientras los valores sean más altos, esto quiere decir que mejor será clustering.

1. **Visualización e Interpretación**

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

**Resultados**

Los 2 modelos identificaron cinco clusters con patrones similares los cuales se basaron en Annual Income y Spending Score. Para los 2 algoritmos:

**K-means**

* **Método del Codo:** Se creo el gráfico el cual indicó k = 5 como el número óptimo de los clusters.
* **Centroides:** Los centroides definieron claramente los segmentos.
* **Gráficos:** Los clusters creados muestran la separación clara en el espacio de Annual Income vs. Spending Score.

Texto, Tabla

El contenido generado por IA puede ser incorrecto.

**Hierarchical clustering:**

* **Dendrograma:** Se confirmaron 5 clusters como elección.
* **Etiquetas:** Los clusters creados son consistentes con los de K-means, con pequeñas diferencias en la asignación de puntos.
* **Gráficos:** Es un poco similar a la separación visual de K-means, validada por el dendrograma.

**Distribución del Género**

La distribución de género por cada uno de los clusters, se mostró que no hay segregación significativa o que sea muy exponencial por género, por lo que se sugiere que los segmentos están más definidos por ingresos y gastos que por género.

Evaluación

Las métricas de la evaluación utilizando el Coeficiente de Silhouette y Índice de Calinski-Harabasz se muestran a continuación:

Tabla

El contenido generado por IA puede ser incorrecto.

Para el Coeficiente de Silhouette se puede evidenciar que ambos modelos tienen valores cercanos a lo cual indica una buena separación y cohesión. Aunque se evidencia que K-means es ligeramente mejor que Hierarchical.

Para el Índice de Calinski-Harabasz se evidencia que el modelo de K-means supera a Hierarchical ligeramente lo cual sugiere la utilización de clusters más compactos.

La interpretación de los 5 clusters identificados representan a los clientes en cada uno de sus segmentos los cuales tienen características distintas.

1. **Cluster 1 (Premium):** Este tiene unalto ingreso y un alto gasto. Los clientes son ideales para productos de lujo o muy costosos.
2. **Cluster 2 (Conservadores):** Este tiene un alto ingreso y un bajo gasto. Los clientes potenciales para promociones que incentiven el gasto en los productos.
3. **Cluster 3 (Impulsivos):** Este tiene un bajo ingreso y un alto gasto. Ya que se tienen en cuenta las sensibles a ofertas y descuentos.
4. **Cluster 4 (Económicos):** Este cluster es de bajo ingreso y contiene un bajo gasto. Están enfocados en productos de bajo costo.
5. **Cluster 5 (Promedio):** Tiene un ingreso y sus gastos son medios. Sus clientes son versátiles para campañas generales.

La baja correlación que existe entre las variables y la ausencia de segregación por género sugieren que las estrategias de marketing deben centrarse en los ingresos y patrones de gasto.

Comparativa entre los 2 modelos:

**K-means**

Gráfico, Gráfico de dispersión

El contenido generado por IA puede ser incorrecto.

**Hierarchical clustering**

Gráfico, Gráfico de dispersión

El contenido generado por IA puede ser incorrecto.

Cada uno de los 2 modelos se ven visualmente parecidos con K-means muestra los bordes más definidos debido a su enfoque centrado en centroides, mientras que en el clustering ofrece una perspectiva jerárquica para la exploración de subgrupos.

Aunque el dataset tiene un alto valor y capacidad de análisis, también tiene varias imitaciones como el dataset no están grande lo que imita la generalización, también se utilizaron 2 variables, pero se podían haber utilizado aun más para el enriquecimiento del los gráficos y análisis.

La creación de los 2 modelos de algoritmos identificó con éxito cinco segmentaciones de clientes robustos con el uso de K-means nos muestra un rendimiento un poco mejor en comparación de Hierarchical clustering en las métricas de evaluación, para estos casos se recomienda el suso de K-means para segmentaciones rápidas y efectivas para datasest similares.

Implementaciones de clustering para exploraciones de relaciones jerárquicas entre los clientes, diseños de estrategias de campañas de marketing especificas para cada uno de los segmentos.

**Conclusiones**

**Referencias bibliográficas**

[**https://scikit-learn.org/stable/**](https://scikit-learn.org/stable/)

[**https://seaborn.pydata.org/**](https://seaborn.pydata.org/)