SEGMENTACIÓN DE CLIENTES

MALL

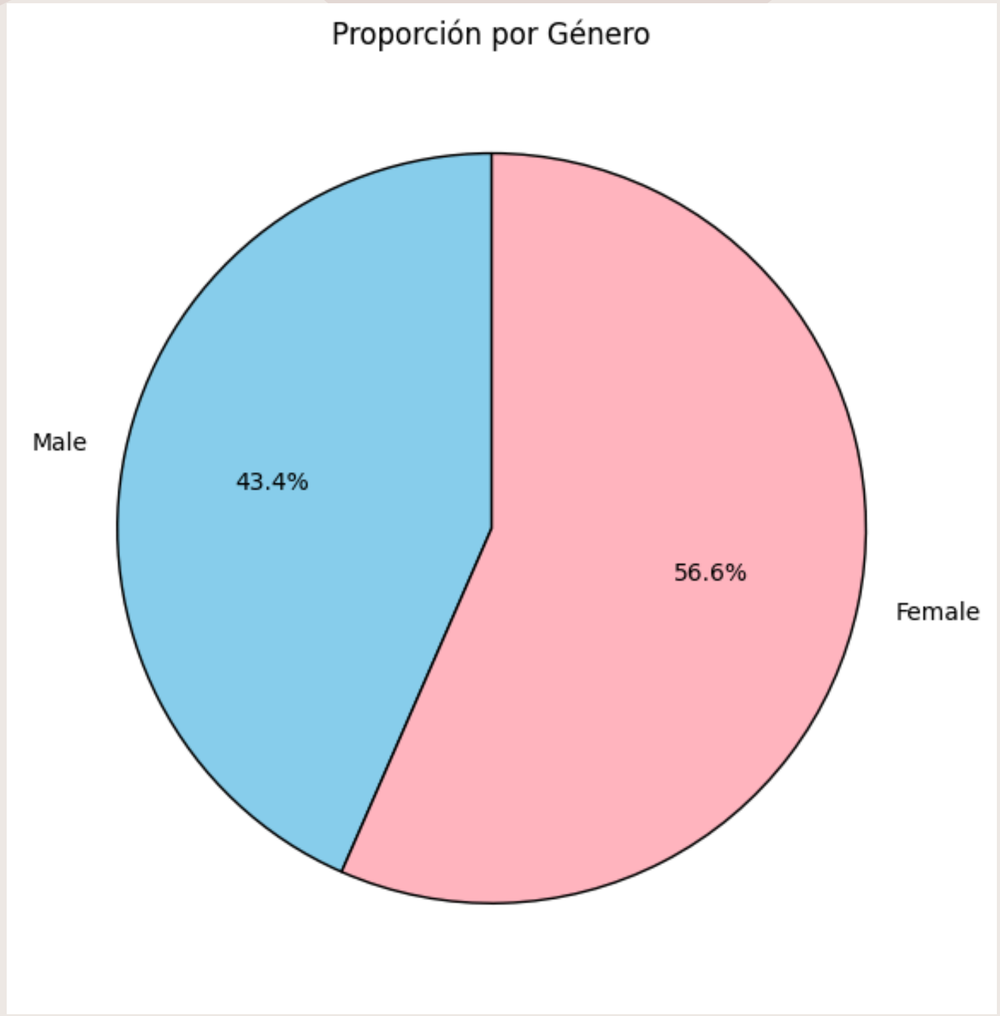
Sebastián Navarro

INTRODUCCIÓN

- Se desea segmentar a los clientes de un centro comercial en grupos basados en sus características demográficas (edad, género) y de gasto (ingreso anual y puntuación de gasto)
- Una estrategia de marketing única no es efectiva para abordar las necesidades específicas de todos los clientes. De esta forma se puede mejorar la experiencia del cliente al ofrecer servicios o promociones específicas para cada grupo
- Se utilizó ténicas de clusterización como k-means, Agglomerative, GMM, DBSCAN, Affinity Propagation
- Para mejorar la clusterización, se utilizó Autoencoder Overcomplete

DATASET

- Filas: 200
- Columnas: 5
  - CustomerID: ID único asignado al cliente.
  - Género: El género del cliente.
  - Edad: La edad del cliente.
  - Ingreso Anual (k$): El ingreso anual del cliente en miles de dólares.
  - Puntuación de Gasto (1-100): La puntuación de gasto del cliente basada en su comportamiento y datos de compra.

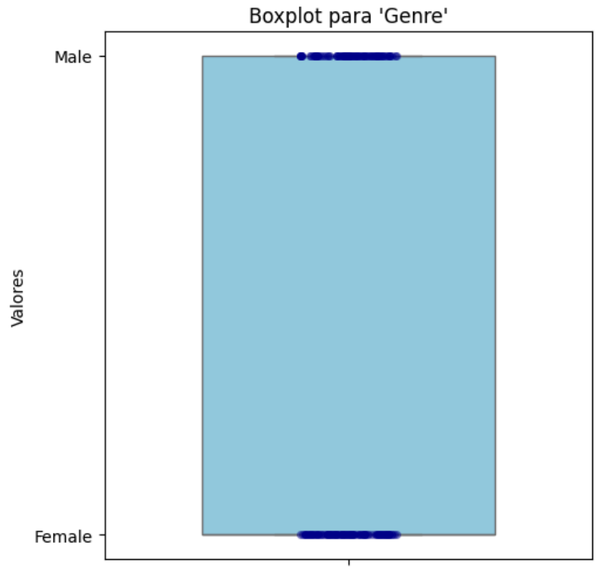


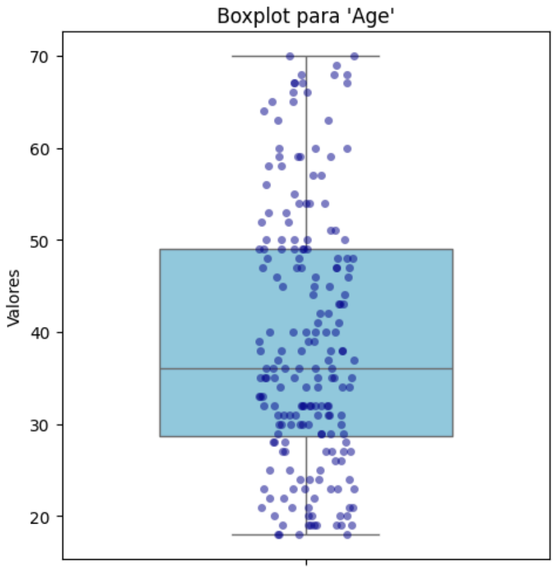
Mall Customers Segmentation

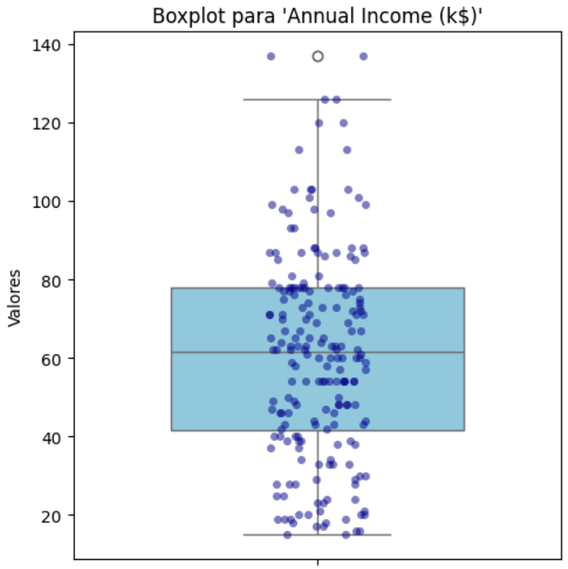
Mall customers data for customer segmentation

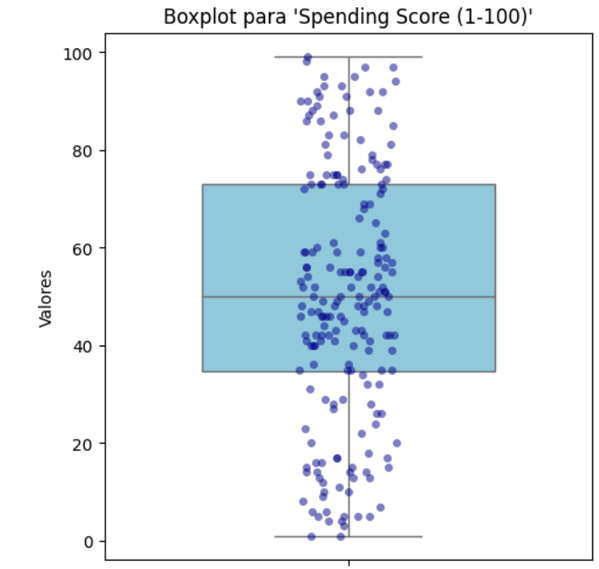
kaggle.com

BOXPLOTS





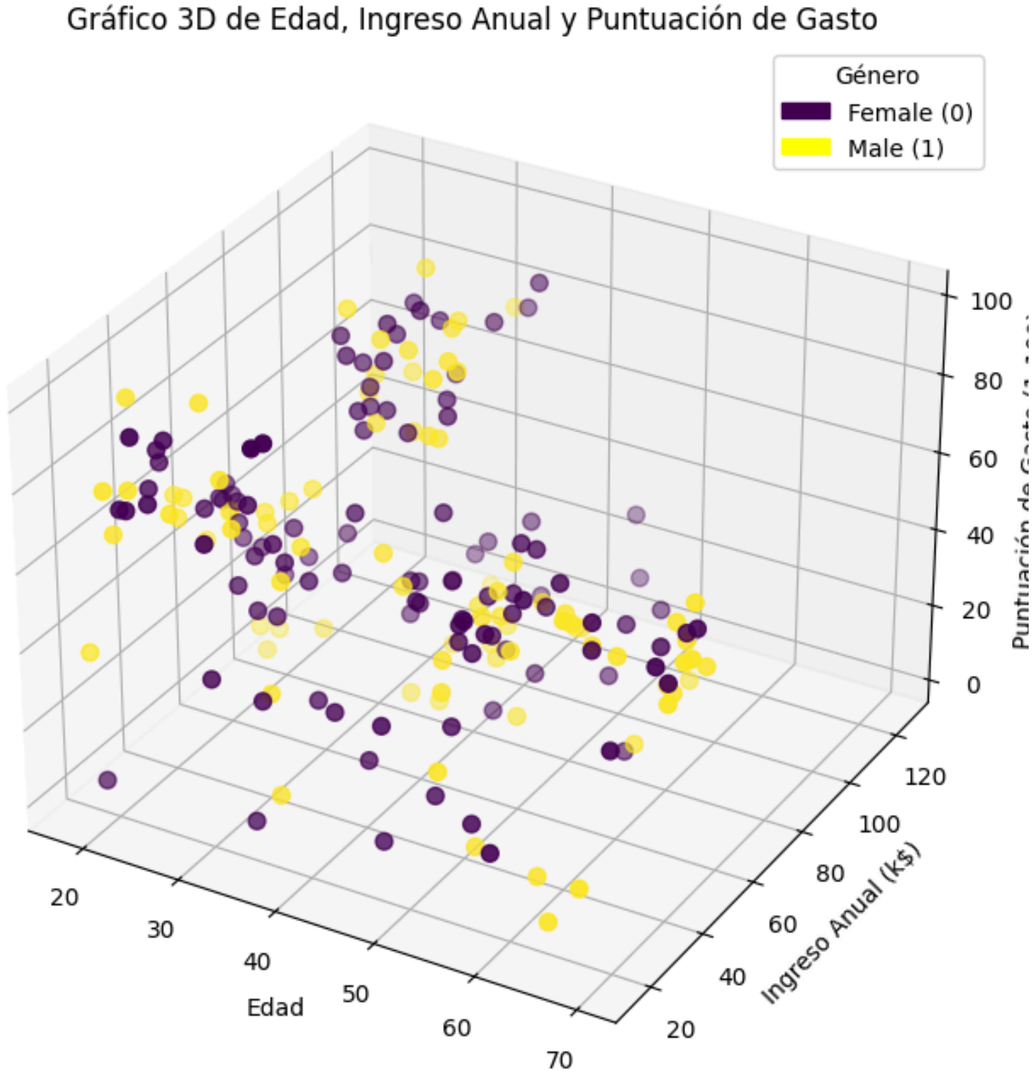


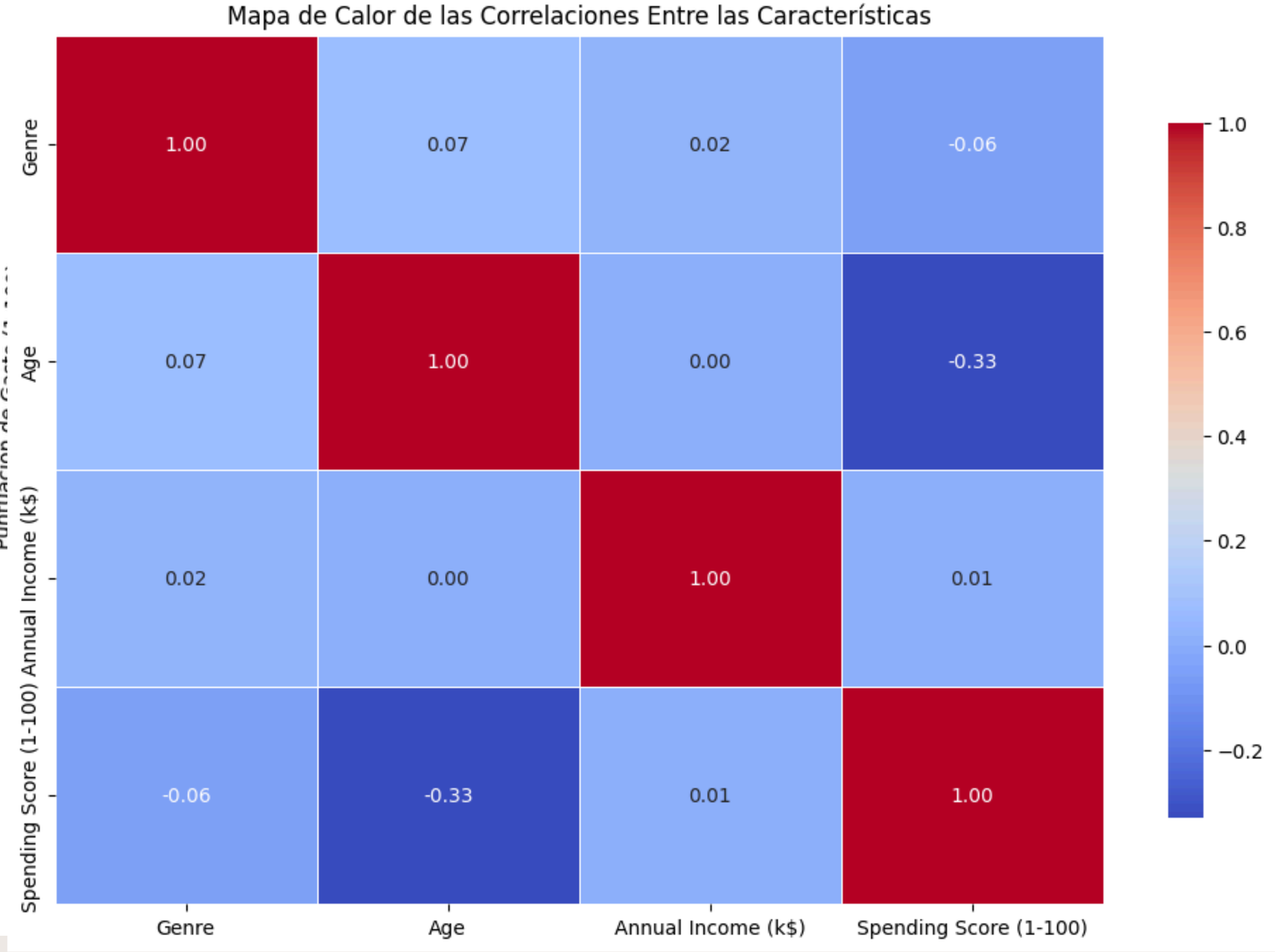


En la variable de ingresos anuales, la mayor parte de los ingresos se encuentran en el rango de 20-80k, con algunos valores atípicos de altos ingresos por encima de los 100k

Genre	Age	Annual Income (k$)	Spending Score (1-100)
Male	32	137	18
Male	30	137	83
Female	45	126	28
Male	32	126	74
Female	47	120	16

DISPERSIÓN Y CORRELACIÓN





TÉCNICAS Y METODOLOGÍA

k-means

- Basada en la partición de datos en "k" grupos utilizando distancias euclidianas.
- Métricas: Silhouette, Calinski-Harabasz, Davies-Bouldin, Inercia y método del codo.
- Parámetros: k = 2, 3, 4, 5, 6, 7, 8

Agglomerative

- Método jerárquico que agrupa datos basándose en las similitudes entre puntos.
- Métricas: Silhouette, Calinski-Harabasz, Davies-Bouldin y dendograma.
- Parámetros: k = 2, 3, 4, 5, 6, 7, 8

GMM

- Un enfoque probabilístico que asume que los datos se distribuyen en varias distribuciones gaussianas.
- Métricas: Silhouette, Calinski-Harabasz, Davies-Bouldin.
- Parámetros: k = 2, 3, 4, 5, 6, 7, 8

TÉCNICAS Y METODOLOGÍA

DBSCAN

- Técnica basada en densidades que identifica clusters de cualquier forma y detecta ruido en los datos.
- Métricas: Silhouette, Calinski-Harabasz, Davies-Bouldin y método de k-distancias
- Parámetros:
  - eps = 0.2, 0.5, 0.6, 0.7, 0.8, 1.0
  - min_samples = 3, 5, 10

Affinity Propagation

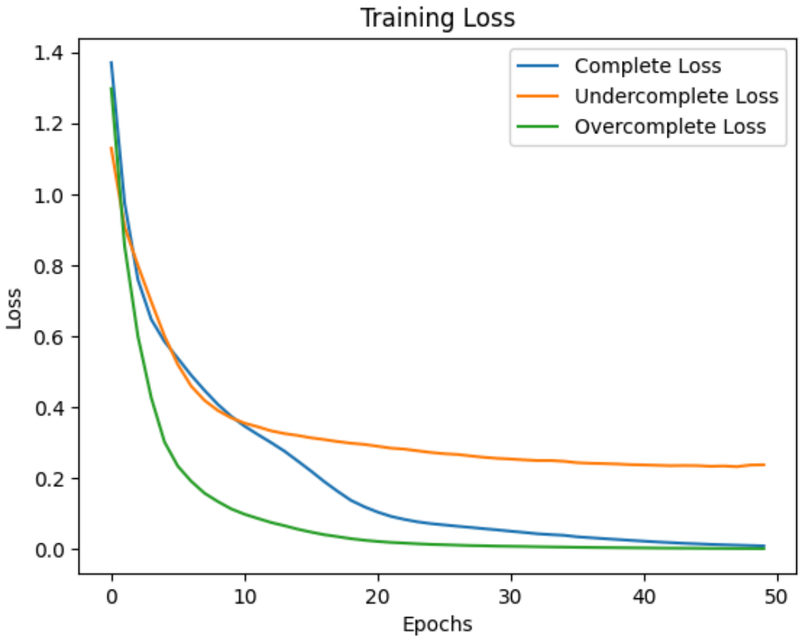
- Un método que determina automáticamente el número de clusters basándose en similitudes entre puntos de datos.
- Métricas: Silhouette, Calinski-Harabasz, Davies-Bouldin y método de k-distancias
- Parámetros:
  - damping = 0.5, 0.7, 0.9
  - preference = -100, -50, None
  - affinity = euclidean

Autoencoder Overcomplete

- Un método que aumenta la dimensionalidad del dataset, logrando una mejor representación de las características subyacentes.
- Métricas: MSE
- Parámetros:
  - Input: input_dim -> Dense(5, ReLU)

AUTOENCODER

Tipo de Autoencoder	Arquitectura de la Capa de Codificación	Arquitectura de la Capa de Decodificación	MSE
Complete Autoencoder	Input: `input_dim` -> Dense(`input_dim` , ReLU)	Dense(`input_dim` , Linear)	MSE: 0.009170
Undercomplete Autoencoder	Input: `input_dim` -> Dense(2, ReLU)	Dense(`input_dim` , Linear)	MSE: 0.235338
Overcomplete Autoencoder	Input: `input_dim` -> Dense(5, ReLU)	Dense(`input_dim` , Linear)	MSE: 0.001973



Se probó reducir la dimensionalidad con 3 Autoencoders:

- **Autoencoder Complete:** Compacta los datos, el número de características de entrada es igual al número de nodos en el espacio latente.
- **Autoencoder Undercomplete:** Reduce la dimensionalidad de manera agresiva, el número de nodos en el espacio latente es menor que el número de características de entrada.
- **Autoencoder Overcomplete:** Aumenta la capacidad del espacio latente, el número de nodos es mayor que el número de características de entrada.

RESULTADOS

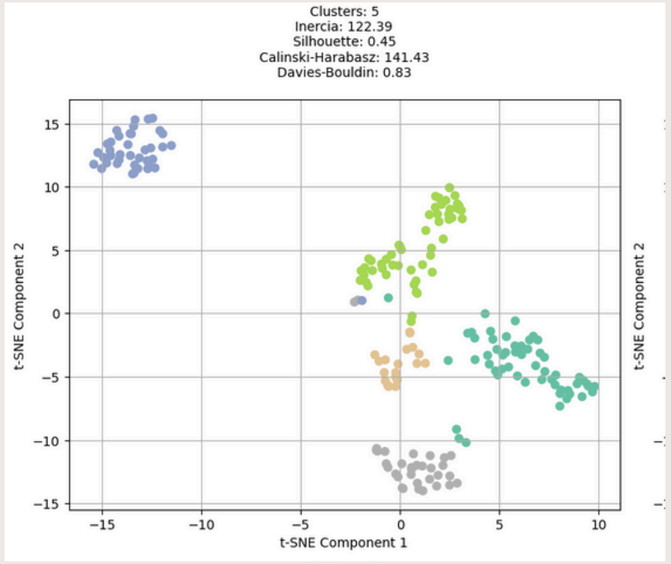
- **K-means** es el mejor método, alcanzando los mejores valores en las tres métricas principales (Silhouette con 0.445 y Calinski-Harabasz con 141.434)

- **Agglomerative Clustering** obtuvo resultados competitivos, especialmente en la métrica Davies-Bouldin, con un valor óptimo de 0.803 en el dataset original.

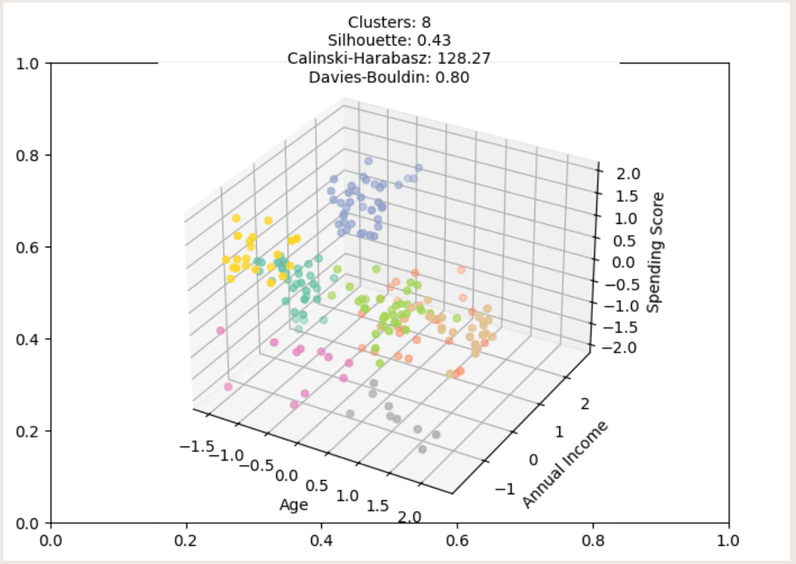
- Los métodos **Gaussian Mixture Model (GMM), DBSCAN y Affinity Propagation** tuvieron un desempeño inferior en comparación, con valores más bajos en las métricas de calidad.
- El dataset aplicado **autoencoder overcomplete** mejoró significativamente los resultados en algunos algoritmos como k-means, DBSCAN, Affinity propagation.

Técnica	Métrica	Valor	Configuración
K-means	Silhouette	0.445360	Dataset Autoencoder, `k=5` clusters
	Calinski-Harabasz	141.434474	Dataset Autoencoder, `k=5` clusters
	Davies-Bouldin	0.823786	Dataset Original, `k=6` clusters
Agglomerative Clustering	Silhouette	0.428048	Dataset Original, `k=8` clusters
	Calinski-Harabasz	133.092029	Dataset Autoencoder, `k=6` clusters
	Davies-Bouldin	0.803267	Dataset Original, `k=8` clusters
Gaussian Mixture Model	Silhouette	0.405600	Dataset Original, `k=7` clusters
	Calinski-Harabasz	119.797616	Dataset Original, `k=7` clusters
	Davies-Bouldin	0.871108	Dataset Original, `k=7` clusters
DBSCAN	Silhouette	0.289699	Dataset Autoencoder, `eps=0.5`, `min_samples=10`, `clusters=4`
	Calinski-Harabasz	53.798595	Dataset Autoencoder, `eps=0.5`, `min_samples=10`, `clusters=4`
	Davies-Bouldin	1.234629	Dataset Autoencoder, `eps=0.5`, `min_samples=10`, `clusters=4`
Affinity Propagation	Silhouette	0.424915	Dataset Autoencoder, `damping=0.7`, `preference=-50.0`, `clusters=5`
	Calinski-Harabasz	136.562289	Dataset Autoencoder, `damping=0.7`, `preference=-50.0`, `clusters=5`
	Davies-Bouldin	0.863096	Dataset Autoencoder, `damping=0.7`, `preference=-50.0`, `clusters=5`

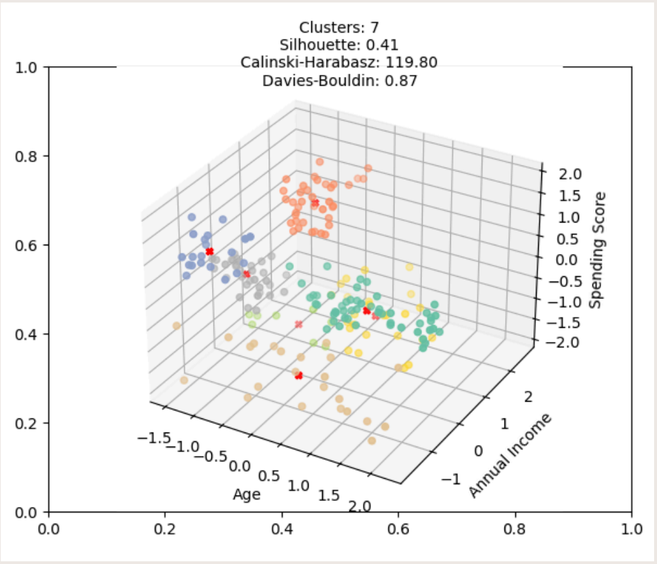
MEJORES RESULTADOS



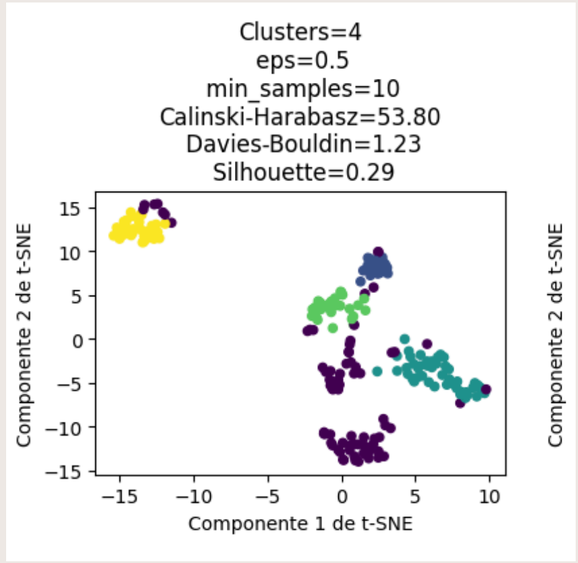
Mejor k-means (AE overomplete)



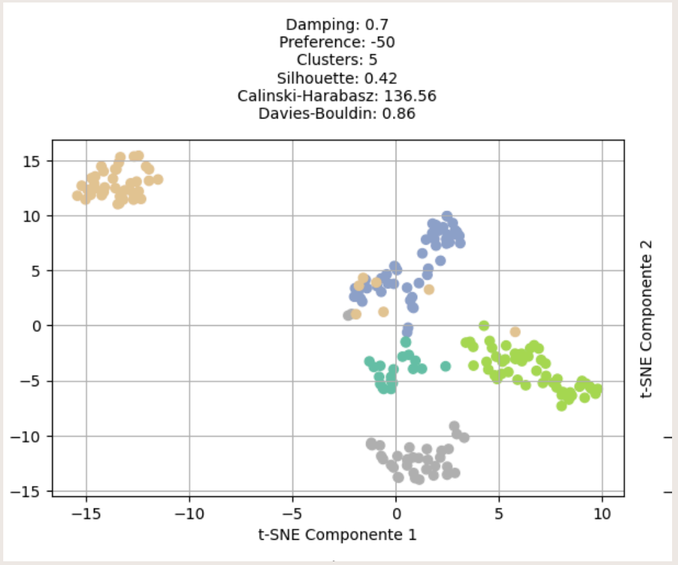
Mejor Agglomerative (DF original)



Mejor GMM (DF original)



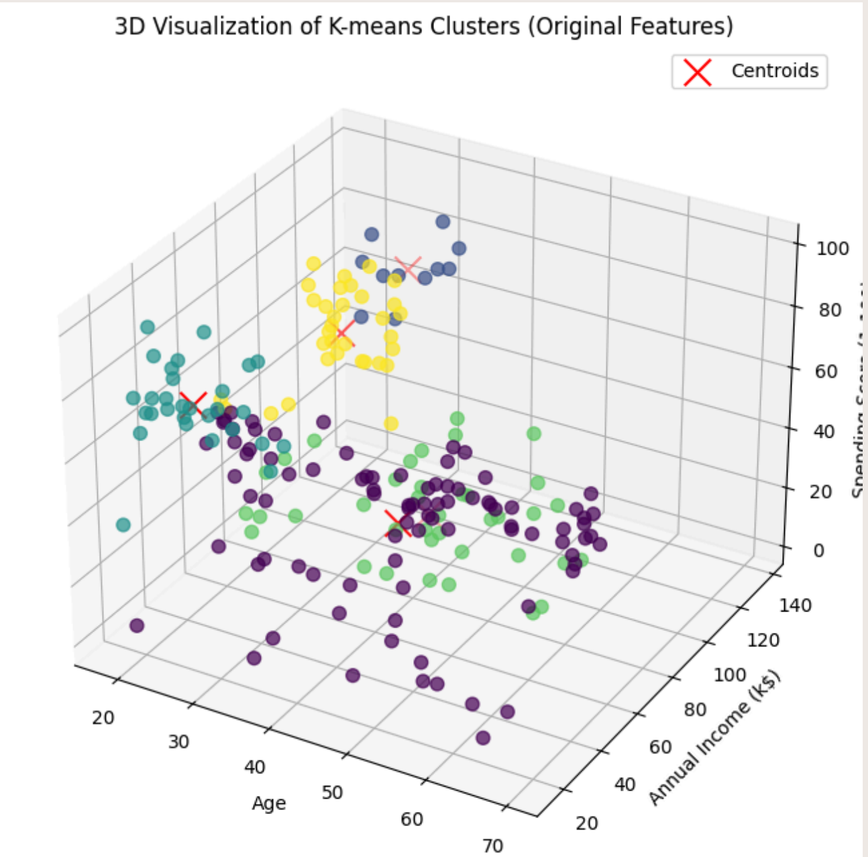
Mejor DBSCAN (AE overcomplete)



Mejor AP (AE overcomplete)

CONCLUSIONES

- **K-means** con 5 clusters es la configuración más robusta y recomendable para este análisis, especialmente con técnicas como Autoencoder Overcomplete



1. **Clúster turquesa:** Jóvenes con bajo ingreso, pero alto gasto.
2. **Clúster amarillo:** Jóvenes con alto ingreso y alto gasto.
3. **Clúster morado:** Adultos mayores con alto ingreso y bajo gasto.
4. **Clúster verde:** Adultos con ingreso medio a alto y bajo gasto.
5. **Clúster azul:** Jóvenes con alto ingreso y alto gasto.