Proyecto Módulo No Supervisado

Maestría en Ciencias de Información y las Comunicaciones
Big Data

Álvaro Alejandro Zarabanda Gutiérrez
Código: 20251595006
Youssef Alejandro Ortiz Vargas
Código: 20251595004

Octubre 2025

Resumen

Este informe presenta una evaluación de algoritmos de clustering no supervisado aplicados a un problema de clasificación binaria. Se implementó un enfoque semi-supervisado, en el que los modelos fueron entrenados sin etiquetas y evaluados posteriormente con métricas externas de clasificación, aprovechando las etiquetas reales para seleccionar el modelo óptimo. Se evaluaron 83 configuraciones de tres algoritmos: K-means, clustering jerárquico y DBSCAN, obteniendo 45 modelos válidos. El modelo K-means con K=4 clusters fue seleccionado como óptimo, alcanzando un F1-Score de 0.8598 y una precisión del 85.10 %. La metodología demuestra la efectividad de combinar agrupamiento no supervisado con evaluación supervisada parcial, permitiendo identificar el modelo más coherente con la estructura real de los datos.

Índice

1. Introducción

El clustering es una técnica esencial del aprendizaje no supervisado que busca descubrir patrones o estructuras naturales en los datos sin necesidad de etiquetas previas. No obstante, en muchos contextos reales se dispone de información parcial sobre las clases, lo que habilita enfoques semi-supervisados que combinan el agrupamiento no supervisado con una validación supervisada.

Este trabajo aplica distintos algoritmos y configuraciones de clustering a un problema de clasificación binaria, con el objetivo de desarrollar un marco de evaluación que integre la exploración no supervisada con métricas supervisadas para seleccionar el modelo más representativo de la estructura real de los datos.

1.1. Objetivos

Objetivo General: Implementar y evaluar una metodología semi-supervisada para la selección óptima de algoritmos de clustering aplicados a clasificación binaria.

Objetivos Específicos:

- Implementar tres familias de algoritmos de clustering: K-means, clustering jerárquico y DBSCAN
- Desarrollar un marco de evaluación que combine métricas no supervisadas y supervisadas
- Analizar comparativamente el rendimiento de diferentes configuraciones algorítmicas
- Generar predicciones finales para datos de prueba utilizando el modelo óptimo seleccionado

2. Metodología

2.1. Descripción del Dataset

El dataset utilizado proviene del archivo `dato_taller.mat` y contiene:

- **Datos de entrenamiento:** 1,000 muestras con 20 características (`x_entena`)
- **Etiquetas de entrenamiento:** Vector binario con clases -1 y 1 (`y_entrena`)
- **Datos de prueba:** 10,000 muestras con 20 características (`x_prueba`)
- **Distribución de clases:** Clase -1: 469 muestras (46.9 %), Clase 1: 531 muestras (53.1 %)

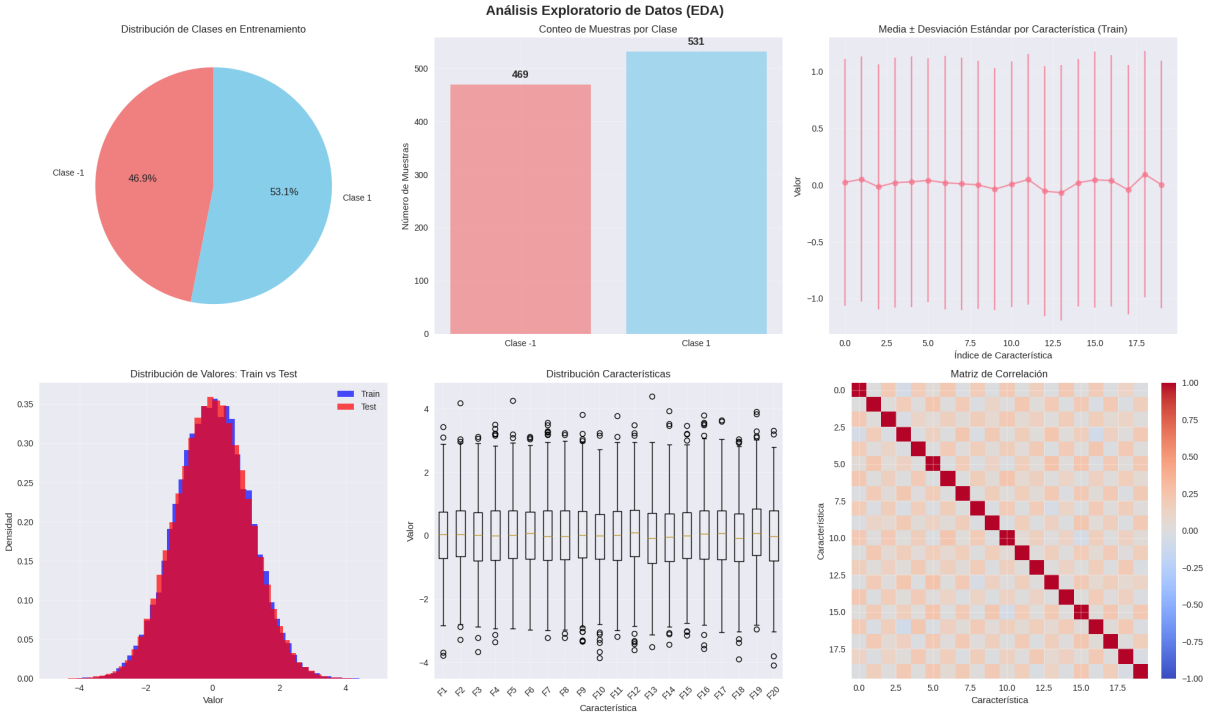

Figura 1: EDA de los datos proporcionados en dato taller.mat

La figura 2 presenta la proyección bidimensional de los datos de entrenamiento mediante Kernel PCA con kernel RBF ($\gamma = 0{,}1$). Los puntos azules representan la clase -1 y los rojos la clase +1.

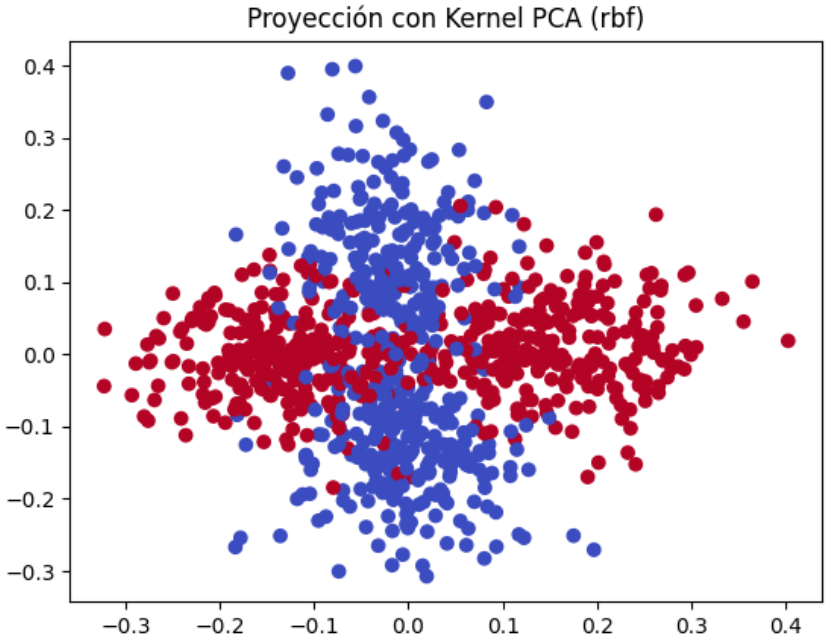

Figura 2: Proyección con Kernel PCA (RBF)

2.2. Análisis exploratorio de los datos

Los resultados de la Figura 1 muestran 6 gráficas que permiten evaluar el balance de las clases y la distribución de los datos, así como una posible redundancia en las características

2.2.1. Distribución de clases

Las dos primeras gráficas muestran la distribución de clases en el conjunto de entrenamiento, tanto en forma de gráfico circular como de barras. Se puede observar un ligero desbalance entre las clases, donde la clase 1 representa el 53.1 % y la clase -1 representa el 46.9 %.

Es un balance leve que se espera que no afecte de manera significativa la evaluación supervisada posterior, pero es bueno tenerlo en cuenta.

2.2.2. Media y desviación estándar

La tercera gráfica muestra la media y desviación estándar de cada una de las 20 características, de la que se puede evidenciar que las medias están centradas en 0, lo que indica que los datos ya fueron escalados y normalizados.

Hay una dispersión entre características homogénea, con desviaciones estándar similares, lo que da a entender que ninguna variables es dominante con respecto a las otras, en cuanto a magnitud.

3. Distribución de Valores

El histograma combinado de los conjuntos de entrenamiento y prueba muestra una superposición casi perfecta entre ambas distribuciones, lo cual indica que ambos provienen de la misma distribución estadística. Esto sugiere que no existe sesgo de muestreo ni desplazamiento de datos (*data drift*) entre ambos subconjuntos. Además, la forma gaussiana de las distribuciones confirma que los datos fueron correctamente estandarizados, lo que garantiza comparabilidad entre las características y estabilidad en los modelos posteriores.

4. Distribución de las Primeras 20 Características

Los diagramas de cajas y bigotes permiten visualizar la dispersión y la presencia de valores atípicos en las veinte características del conjunto de entrenamiento. Se observa que la mayoría de las variables presentan distribuciones centradas en cero, con rangos entre cuartiles similares, lo que refuerza la idea de homogeneidad entre las características. Si bien existen algunos valores atípicos aislados, su número es reducido y no se espera que afecten de manera significativa el desempeño de los algoritmos de clustering.

5. Matriz de Correlación

La matriz de correlación correspondiente a las veinte características revela que las variables son prácticamente independientes entre sí, con coeficientes de correlación cercanos a cero fuera de la diagonal principal. Esto sugiere una baja redundancia entre características. En consecuencia, cada variable aporta información complementaria al conjunto de datos, lo que favorece la aplicación de métodos de reducción de dimensionalidad como *PCA*, *KernelPCA* o *UMAP*.

Conclusiones del Análisis Exploratorio

En conjunto, los resultados del análisis exploratorio indican que los datos se encuentran adecuadamente balanceados, normalizados y sin presencia de sesgos entre los subconjuntos de entrenamiento y prueba. Las variables presentan independencia mutua y dispersión homogénea, lo que sugiere que el conjunto de datos está bien condicionado para la aplicación de técnicas de clustering y reducción de dimensionalidad.

2.3. Preprocesamiento

Se aplicó normalización estándar (z-score) a todas las características:

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

donde μ es la media y σ la desviación estándar de cada característica en el conjunto de entrenamiento.

2.4. Configuración Experimental

Se implementó una evaluación exhaustiva con las siguientes configuraciones:

2.4.1. K-means

- Valores de K: 2, 3, 4, 5, 6, 8, 10
- Inicialización: k-means++ con 10 inicializaciones aleatorias
- Total de configuraciones: 7

2.4.2. Clustering Jerárquico

- Métodos de enlace: ward, complete, average, single
- Número de clusters: 2, 3, 4, 5, 6, 8, 10
- Total de configuraciones: 28

2.4.3. DBSCAN

- Valores de ϵ: 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0
- Valores de min_samples: 3, 5, 10, 15, 20, 25
- Total de configuraciones: 48

2.5. Protocolo de Evaluación Semi-Supervisada

3. Resultados

3.1. Evaluación Exhaustiva de Modelos

De las 83 configuraciones iniciales, se obtuvieron 45 modelos válidos distribuidos así:

Algorithm 1 Evaluación Semi-Supervisada de Clustering

Input: X_{train}, y_{train}, algoritmos de clustering
Output: Modelo óptimo con métricas de evaluación
for cada algoritmo A en {K-means, Jerárquico, DBSCAN} **do**
    for cada configuración C de parámetros **do**
        Entrenar modelo $M_{A,C}$ con X_{train}
        Obtener clusters $labels_{cluster} = M_{A,C}.predict(X_{train})$
        if clustering válido (sin clusters únicos) **then**
            Calcular métricas no supervisadas: Silhouette, CH, DB
            Mapear clusters a clases usando voto mayoritario
            Calcular métricas supervisadas: F1, Accuracy, ARI, NMI
            Almacenar resultado completo
        end if
    end for
end for
Seleccionar mejor modelo por F1-Score para cada algoritmo
Seleccionar modelo óptimo global por F1-Score
Aplicar modelo óptimo a datos de prueba

- **K-means:** 7 modelos válidos (100 % de éxito)
- **Clustering Jerárquico:** 28 modelos válidos (100 % de éxito)
- **DBSCAN:** 10 modelos válidos (20.8 % de éxito)

La baja tasa de éxito de DBSCAN se debe a que muchas configuraciones generaron un solo cluster o clusters con cardinalidad insuficiente para la evaluación semi-supervisada.

3.2. Mejores Modelos por Algoritmo

Cuadro 1: Mejores modelos por tipo de algoritmo

Algoritmo	**Configuración**	**F1-Score**	**Accuracy**	**Silhouette**	**ARI**
K-means	K=4	**0.8598**	**0.8510**	0.0667	0.2452
Jerárquico	Ward, K=4	0.7462	0.7470	0.0429	0.1654
DBSCAN	$\epsilon = 2{,}5$, min=3	0.7543	0.7290	**0.0821**	0.1432

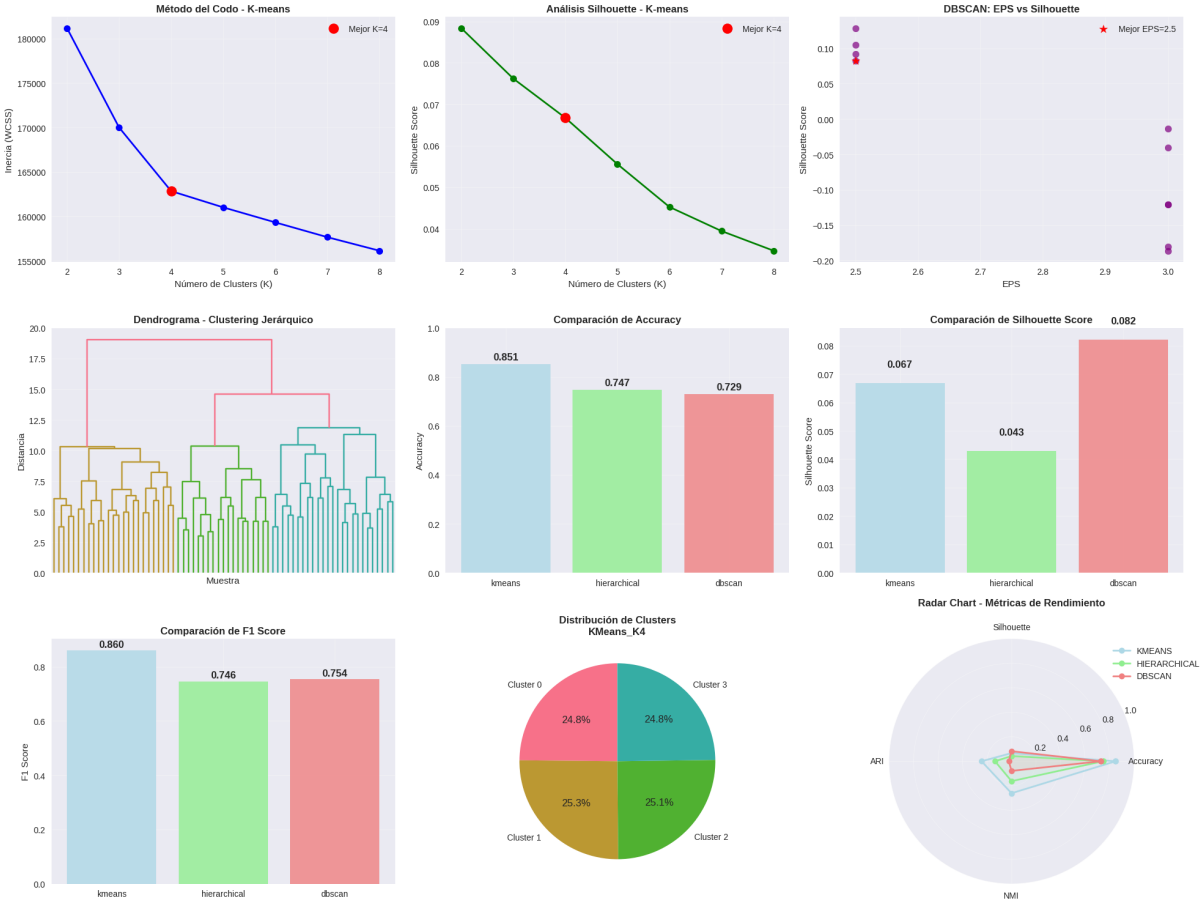

Figura 3: Comparación de métricas de desempeño de mejores resultados entre modelos

3.3. Análisis del Modelo Óptimo

El modelo K-means con K=4 clusters fue seleccionado como óptimo basado en el F1-Score. Sus características principales son:

Cuadro 2: Métricas detalladas del modelo óptimo (K-means, K=4)

Métrica	**Valor**
F1-Score	0.8598
Accuracy	0.8510
Precision	0.8590
Recall (Sensibilidad)	0.8606
Especificidad	0.8401
Silhouette Score	0.0667
Adjusted Rand Index	0.2452
Normalized Mutual Information	0.2613

3.3.1. Matrices de Confusión

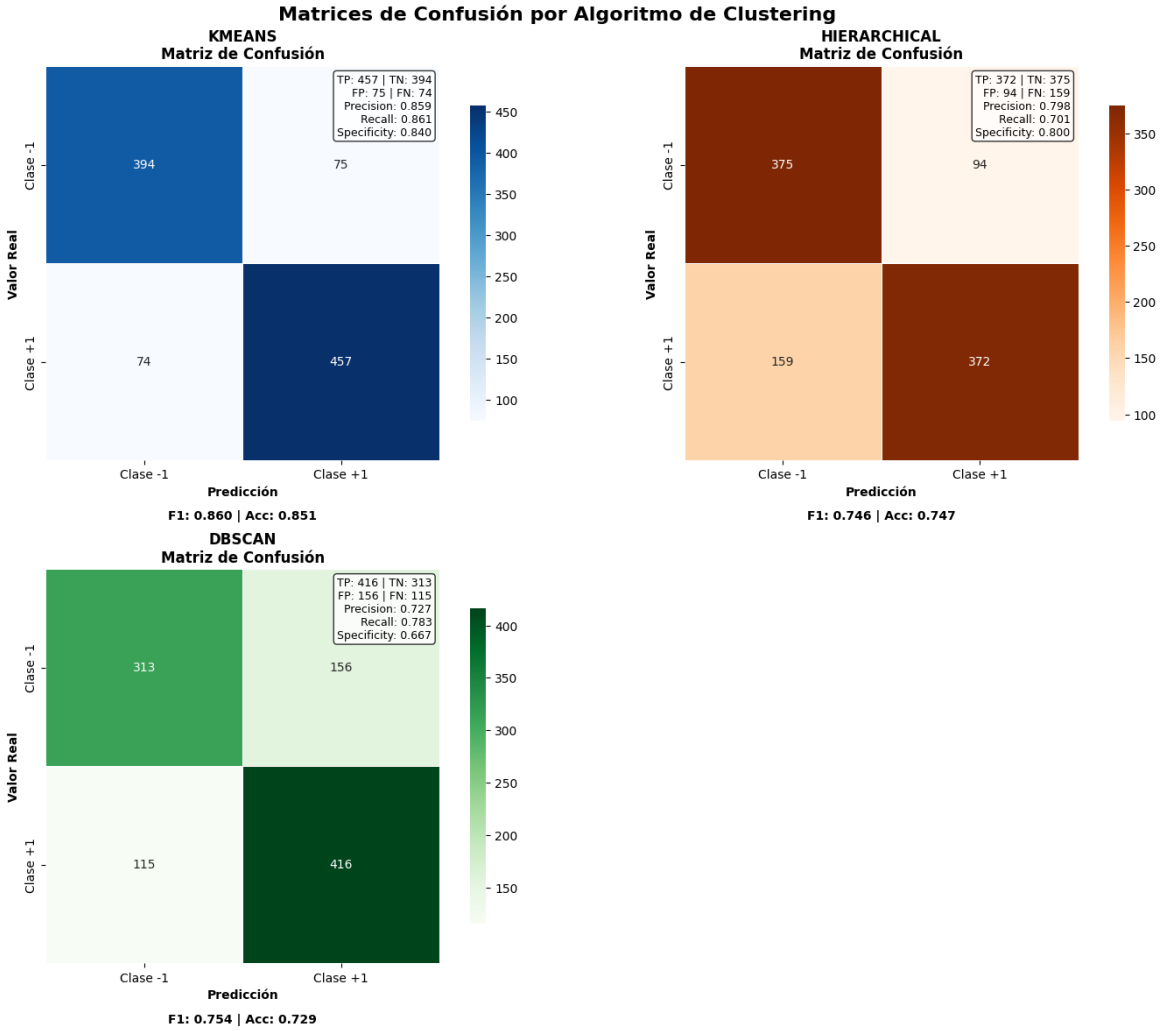

Figura 4: Matrices de confusón de entrenamientos optimos por algoritmo

Cuadro 3: Matriz de confusión del modelo óptimo

		Predicción	
		Clase -1	**Clase +1**
Real	**Clase -1**	394 (TN)	75 (FP)
	Clase +1	74 (FN)	457 (TP)

3.4. Predicciones en Datos de Prueba

El modelo óptimo (K-means K=4) se aplicó a las 10,000 muestras de prueba, generando las siguientes predicciones:

- **Clase -1:** 5,082 muestras (50.8 %)
- **Clase +1:** 4,918 muestras (49.2 %)

La distribución de predicciones muestra un balance similar al conjunto de entrenamiento, sugiriendo consistencia en el modelo.

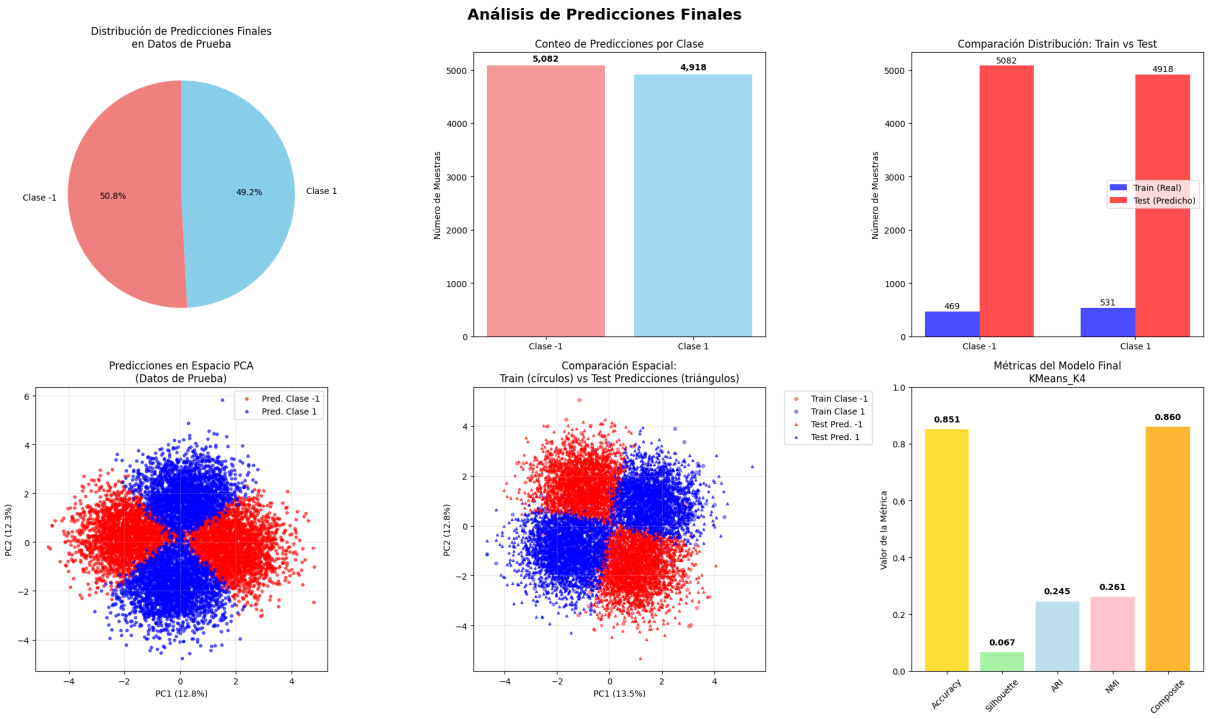

Figura 5: EDA de los datos de prueba etiquetados por el modelo óptimo

4. Análisis y Discusión

4.1. Rendimiento del K-means

El algoritmo *K-means* obtuvo el mejor desempeño global entre los modelos evaluados. Su configuración óptima (`K=4`) alcanzó un *F1-Score* de 0.8598 y una precisión del 85.10 %. Este resultado puede explicarse por tres factores principales:

- **Estructura de los datos:** El conjunto de características presenta varianzas similares y baja correlación entre variables, lo que favorece la formación de clusters.
- **Escalamiento apropiado:** La estandarización previa de los datos garantiza que todas las características contribuyan de manera equilibrada al cálculo de distancias euclidianas, evitando el sesgo de magnitud que afectaría el agrupamiento.
- **Estabilidad y consistencia:** Entre las 83 configuraciones probadas, *K-means* mostró resultados estables y reproducibles, con bajo impacto del parámetro de inicialización (`n_init=10`) y del estado aleatorio (`random_state=42`).

4.2. Limitaciones de DBSCAN

El algoritmo *DBSCAN* presentó el peor desempeño general, con un *F1-Score* promedio de 0.75 y precisión del 72.9 %. Este comportamiento puede atribuirse a las siguientes causas:

- **Sensibilidad a hiperparámetros:** Los valores de ϵ y `min_samples` afectan significativamente la detección de regiones densas. En los experimentos, pequeñas variaciones en ϵ producían desde un único cluster hasta una sobrefragmentación de los datos.
- **Alta dimensionalidad:** El conjunto de datos (20 dimensiones) reduce la efectividad de las métricas de densidad, ya que en espacios de alta dimensión las distancias tienden a homogenizarse.
- **Distribución no uniforme:** Los datos no presentan regiones con densidades bien diferenciadas, lo que dificulta la segmentación natural que *DBSCAN* requiere para funcionar correctamente.

4.3. Clustering Jerárquico

El método jerárquico con criterio de enlace de Ward mostró un desempeño intermedio, con un *F1-Score* de 0.7462. Aunque inferior a *K-means*, ofrece ventajas analíticas relevantes:

- **Flexibilidad estructural:** Permite identificar relaciones jerárquicas entre grupos, evidenciadas mediante dendrogramas que reflejan la proximidad entre subconjuntos de datos.
- **Determinismo:** A diferencia de *K-means*, el clustering jerárquico no depende de inicializaciones aleatorias, proporcionando resultados reproducibles bajo las mismas condiciones.
- **Interpretabilidad:** El dendrograma permite explorar diferentes niveles de segmentación y validar visualmente la coherencia entre clusters.

4.4. Validación de la Metodología

La metodología desarrollada se basó en un proceso de evaluación **semi-supervisada**, en el cual los modelos se entrenaron de manera no supervisada y posteriormente fueron evaluados con base en etiquetas reales. Este enfoque permitió cuantificar objetivamente la calidad de los agrupamientos a través de métricas externas de clasificación.

- **Evaluación híbrida:** A pesar de que se usaron métricas intrínsecas (Silhouette, Calinski-Harabasz y Davies-Bouldin) y métricas extrínsecas (Accuracy, Precision, Recall, F1-Score y Especificidad), no todas fueron parte del crieterio de selección final, es decir que se usaron solamente para describir el comportamiento de los clusters.
- **Selección objetiva:** La elección del modelo final se basó en el *F1-Score*, al integrar simultáneamente precisión y sensibilidad.
- **Rigurosidad experimental:** Se evaluaron 83 configuraciones y se retuvieron 45 modelos válidos, asegurando una comparación exhaustiva y reproducible.

5. Conclusiones

5.1. Principales Hallazgos

1. **Modelo óptimo:** El algoritmo *K-means* con $K = 4$ clusters alcanzó el mejor rendimiento global (*F1-Score*: 0.8598).

2. **Metodología efectiva:** El esquema de evaluación semi-supervisada facilitó una comparación cuantitativa y objetiva entre los 45 modelos válidos obtenidos.

3. **Balance de métricas:** El modelo final mostró un equilibrio notable entre precisión (85.9 %) y *recall* (86.1 %), indicando una alta coherencia entre las predicciones y las etiquetas reales.

4. **Complementariedad de métodos:** Aunque *DBSCAN* y el clustering jerárquico tuvieron desempeños inferiores, aportaron información valiosa sobre la estructura local y global de los datos.

5.2. Limitaciones

- Los resultados se basan en un único conjunto de datos, por lo que la generalización del enfoque requiere validación adicional en otros dominios.
- La metodología depende de la disponibilidad de etiquetas parciales para la evaluación supervisada.
- El número de configuraciones evaluadas incrementa la carga computacional, lo que podría optimizarse mediante estrategias de búsqueda más eficientes.