Examen Parcial 1 - Estadística Aplicada III

Diana Isabel Muñoz Castillo

October 4, 2024

Pregunta 2 - Diabetes Data

1. Proceso de Clustering

El clustering es un método **no supervisado**, por lo que la columna 'class' fue eliminada para evitar sesgo en los resultados. Se aplicaron tres algoritmos: **K-Means**, **Clustering Jerárquico** y **DBSCAN**, todos con el objetivo de encontrar la estructura interna de los datos sin utilizar las clases reales.

K-Means

Utilicé el **método del codo** para determinar el número óptimo de clusters, el cual fue 3 (Figura 1). El **Silhouette Score** obtenido fue de 0.37 (Figura 2), indicando una separación moderada entre los clusters.

Clustering Jerárquico

El **dendrograma** generado también sugiere que la mejor partición es con 3 clusters (Figura 3). El Silhouette Score fue de 0.36, muy similar al de K-Means.

DBSCAN

Aunque este algoritmo no requiere predefinir el número de clusters y puede identificar ruido, su desempeño en este conjunto fue limitado, con un Silhouette Score de 0.30. El algoritmo clasificó una parte de los datos como ruido, lo que afectó su precisión.

2. Comparación con Clases Reales

Aunque las clases verdaderas no se usaron en el análisis, se evaluó el desempeño real comparando los clusters con las etiquetas originales. La **precisión** se calculó usando una matriz de confusión:

- **K-Means**: Precisión de 75.17%, con algunas confusiones entre los clusters, especialmente en la clase 3.
- **Clustering Jerárquico**: Mejor precisión con 77.93%, aunque también hubo mal agrupamiento en algunas instancias.
- **DBSCAN**: Menor precisión, 61.70%, debido a la identificación de ruido y menor coherencia en los clusters.

3. Conclusión

El **Clustering Jerárquico** fue el mejor algoritmo en este caso, logrando la mayor precisión al compararlo con las clases reales. Sin embargo, es importante recordar que el objetivo del clustering es descubrir patrones internos, sin depender de etiquetas. Ningún algoritmo logró una separación perfecta, lo que sugiere que las clases no son fácilmente separables con las características actuales de los datos.

Anexos

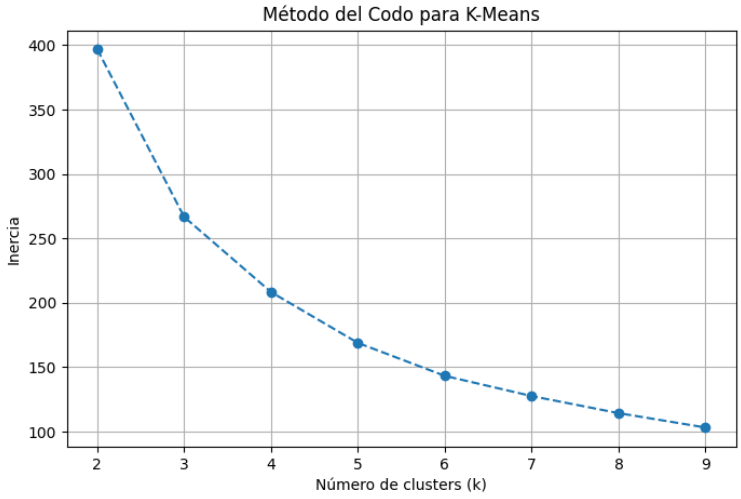

Figure 1: Método del Codo para K-Means

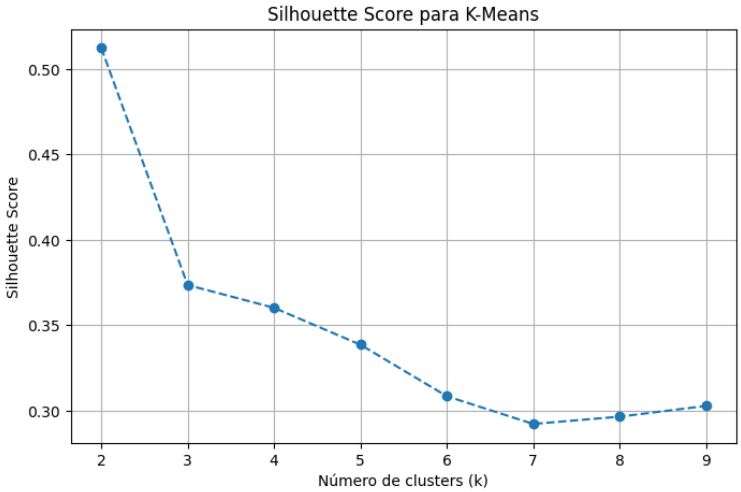

Figure 2: Silhouette Score para K-Means

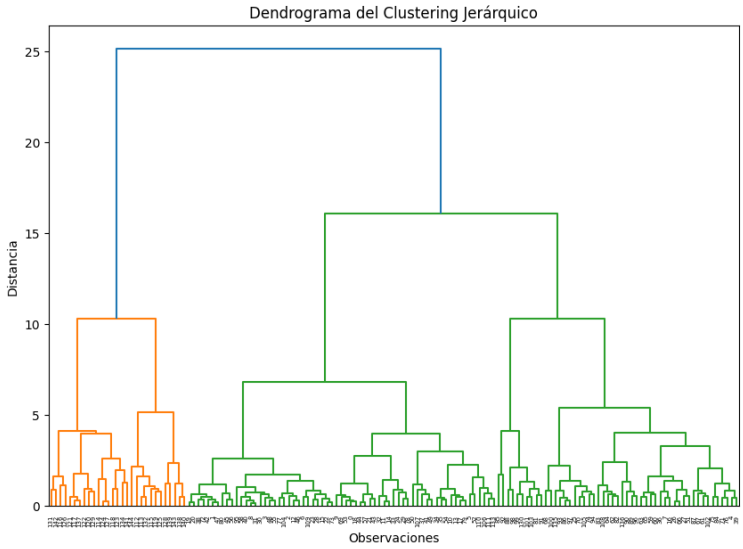

Figure 3: Dendrograma del Clustering Jerárquico

Referencias

- GitHub CRAN MMST - Diabetes Dataset. Recuperado de: `https://github.com/cran/MMST/blob/master/data/diabetes.rda`.