Práctica 2
CLUSTERING (AGRUPACIÓN) DE SEMILLAS

2,5 puntos

INTRODUCCIÓN

En esta segunda práctica aplicaremos técnicas de agrupamiento/clustering sobre un conjunto de datos con atributos sobre semillas, para comprobar si aparecen agrupaciones significativas y qué técnicas de clustering funcionan mejor sobre estos datos.

CONSIDERACIONES GENERALES

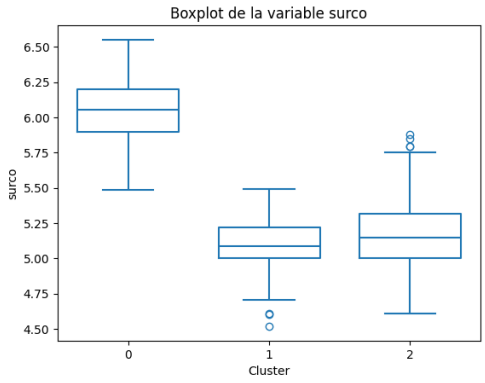
1. Para realizar la práctica, los estudiantes emplearán un repositorio de código en GitHub. Para ello, cada grupo debe crear un repositorio de código privado y agregar como «colaborador» al profesor de prácticas (que indicará a los estudiantes su nombre de usuario en GitHub). Se espera que cada grupo haga un commit semanal del código de la práctica. Esta parte de la práctica se valorará con **0.5 puntos**. Además, también habrá que entregar el cuaderno (notebook) final a través de Aula Global.
2. Los resultados deben ser reproducibles. Por lo tanto, hay que fijar la semilla de números aleatorios en los lugares adecuados. Se usará como semilla el NIA de uno de los miembros del grupo o bien el número del grupo de prácticas. Si hay que utilizar más semillas se usan números consecutivos al NIA de base.

QUÉ HACER

1. (**0.25 puntos**) Comprobar visualmente cuál de los 3 scalers es más apropiado para el problema (MinMaxScaler, RobustScaler, StandardScaler). Para poder visualizar los datos, se aplicará Principal Component Analysis (PCA), extrayendo dos componentes. Se puede usar una pipeline que incluya el scaler y después PCA(n_components=2), la cual hace una transformación no supervisada a 2 dimensiones, de tal manera que los datos puedan ser visualizados en 2D.
2. (**1.25 puntos**) Usando los datos transformados a 2D, aplicar las tres técnicas de clustering explicadas en clase (K-Means, Hierarchical Clustering/Dendrogramas, DBSCAN), comparando y discutiendo los resultados que se obtienen de ellos y decidiendo sus hiper-parámetros más importantes (k-means: determinar número de clusters con codo y/o silueta, aglomerativo/jerárquico: probar varias funciones de linkage y determinar

número de clusters adecuado, DBSCAN: ajustar minpts y eps). Recordar que al ser un problema no supervisado, no se puede utilizar la variable de respuesta “clase” en este apartado.

3. (**0.5 puntos**) Análisis:
    - Usando una comprobación visual, ¿cuál de los métodos de clustering captura mejor la estructura de clusters de este problema?
    - ¿Hay relación entre los clusters obtenidos y las clases de semillas originales? (variable de respuesta “clase”)
    - Interpretar los clusters obtenidos usando por ejemplo boxplots como los de la figura:



4. (**0.5 puntos**) Recordar que es necesario hacer un commit semanal en github para obtener los 0.5 puntos.

QUÉ ENTREGAR

Entregar un único notebook con el código, las descripciones y justificaciones apropiadas, y la comparación y discusión de diferentes algoritmos de clustering, y de los resultados obtenidos en los distintos apartados de la práctica.