Laboratorio (actividad 1): Agrupamiento no supervisado

Objetivo

Aplicar técnicas de *clustering* sobre un conjunto de datos reales para identificar grupos naturales en función de características energéticas de edificios.

Contexto

Trabajaréis como equipos de científicos de datos analizando los datos del rendimiento energético de edificios. El objetivo es segmentar el conjunto en grupos que compartan características similares para proponer mejoras o políticas energéticas adaptadas.

Dataset

Utilizaremos el conjunto de datos *Energy Efficiency Dataset*[1] del UCI *Machine Learning Repository*. Este conjunto de datos contiene variables como:

Variable	Descripción
Relative_Compactness	Grado de compacidad del edificio.
Surface_Area	Área total de la superficie externa (m^2).
Wall_Area	Área total de las paredes exteriores (m^2).
Roof_Area	Área del tejado (m^2).
Overall_Height	Altura total del edificio (m).
Orientation	Orientación cardinal codificada (2-5).
Glazing_Area	Proporción de superficie acristalada.
Glazing_Area_Distribution	Distribución de ventanas por fachadas codificada (0-5).
Heating_Load	Carga térmica de calefacción (kWh/m^2).
Cooling_Load	Carga térmica de refrigeración (kWh/m^2).

Tareas por grupo

1. **Inspección inicial del *dataset*.**
    a. Cargar el *dataset* en un DataFrame de pandas.
    b. Cambiar los nombres de las columnas a nombres legibles.
    c. Verificar los tipos de datos, valores nulos, faltantes y duplicados.
    d. Analizar estadísticas básicas y correlaciones.

[1] https://archive.ics.uci.edu/dataset/242/energy+efficiency

2. **Preprocesamiento e inspección rápida.**
    a. Representar gráficamente las variables principales mediante histogramas, *boxplots*, *pairplots*, etc.
    b. Comprobar si hay *outliers* y decidir cómo tratarlos.
    c. Seleccionar las variables relevantes para el *clustering*.
    d. Estandarizar o normalizar las variables seleccionadas.

3. **Aplicación de algoritmos de *clustering***.
    a. Aplicar al menos dos algoritmos de agrupamiento como podrían ser *k-means* y DBSCAN.
    b. Justificar el número de *clusters* y parámetros usados.
    c. Calcular el *Silhouette score* y el índice de rand ajustado (ARI) para comparar los modelos si coinciden en el número de *clusters*.

4. **Visualización de resultados y conclusiones**.
    a. Mostrar los *clusters* obtenidos en 2D.
    b. Comentar las características comunes en cada grupo (interpretación).
        i. ¿Qué técnica ha generado las agrupaciones de mayor calidad?
        ii. ¿Qué grupos de edificios parecen más eficientes?
        iii. ¿Qué características comparten?