Práctica Introductoria al Aprendizaje No Supervisado

Objetivos de la hoja

- Aprender a preparar datos para ML no supervisado.
- Reducir la dimensionalidad para visualizar y entender los datos.
- Aplicar y evaluar técnicas de agrupamiento (*clustering*).
- Elegir el número óptimo de *clusters*.
- Interpretar y visualizar resultados.

Ejercicio 1. Cargando y explorando los datos

a) Descarga el dataset *Iris* usando `sklearn.datasets.load_iris()` o desde este enlace de UCI.

b) Carga el dataset en un DataFrame de pandas.

c) Realiza una **exploración básica**:

- ¿Cuántas filas y columnas hay?
- ¿Qué significa cada columna?
- ¿Hay valores nulos?

Ejercicio 2. Preprocesamiento y estandarización

a) ¿Por qué es importante escalar o estandarizar los datos antes de aplicar *clustering*?

b) Utiliza `StandardScaler` para estandarizar los datos numéricos.

c) Vuelve a mostrar un resumen estadístico y compara con el original.

Ejercicio 3. Reducción de dimensionalidad (PCA)

a) Explica brevemente qué es y para qué sirve **PCA** (en tus propias palabras).

b) Aplica PCA para reducir los datos a **2 dimensiones**.

c) Haz un scatter plot (gráfico de dispersión) con los dos componentes principales.

d) Colorea los puntos según la "especie" real del iris (si la tienes).

Ejercicio 4. Clustering con K-Means

a) Explica qué es el clustering y en qué consiste el algoritmo **K-Means**.

b) Aplica **K-Means** con `n_clusters=3` al dataset **ya estandarizado**.

c) Añade una columna al DataFrame con la etiqueta de cluster asignada a cada muestra.

d) Visualiza los resultados en el plano de los dos primeros componentes principales, coloreando por **cluster** en vez de por especie.

Ejercicio 5. ¿Cómo elegir el número de clusters?

a) Explica brevemente los métodos del "codo" (elbow method) y la "silhouette".

b) Usando el dataset escalado, calcula la inercia (SSE) para valores de K de 1 a 10 y haz la gráfica del codo.

c) Calcula el silhouette score para valores de K entre 2 y 6.

¿Cuál te parece el valor más adecuado de K? Justifica tu elección.

Ejercicio 6. Interpretando resultados

a) Compara los clusters encontrados con las especies reales (puedes usar una tabla de contingencia con pandas).

¿Coinciden perfectamente? ¿Por qué crees que ocurre eso?

b) ¿Qué información útil se puede obtener de un análisis no supervisado como este?

Ejercicio 7. Detección de valores atípicos (bonus)

a) Aplica el algoritmo **Isolation Forest** para buscar valores atípicos en el dataset de iris.

b) Visualiza los datos y resalta los posibles outliers.

¿Hay alguno? ¿Tiene sentido desde el punto de vista biológico?

Recomendaciones

- Puedes hacer todos los ejercicios en un **notebook de Jupyter**.
- Utiliza comentarios en el código para explicar cada paso.
- Usa gráficos en cada ejercicio de visualización.