Introducción

El Breast Cancer Wisconsin Diagnostic Dataset es un conjunto de datos ampliamente utilizado en proyectos de ciencia de datos y aprendizaje automático, especialmente para prácticas de clasificación y agrupamiento. Fue recopilado por el Dr. William H. Wolberg en la Universidad de Wisconsin, y contiene información sobre características de **células tumorales** obtenidas a partir de imágenes digitales de biopsias de tejido mamario.

El objetivo de estos datos es analizar las características físicas de los núcleos celulares para ayudar a distinguir entre tumores malignos y benignos

Objetivo General

Aplicar técnicas de **agrupamiento (clustering)** para explorar y analizar patrones en los datos del cáncer de mama, con el fin de identificar posibles grupos dentro del conjunto de observaciones, basados en características de los tumores.

Metodología

1. Exploración inicial del conjunto de datos: Cargar y visualizar el dataset, Examinar cuántas observaciones y variables hay, Comprobar si existen valores faltantes o inconsistencias**,** Obtener estadísticas básicas para tener una idea general de los datos.

2. Selección de variables: Elegir dos variables numéricas que consideres relevantes para un primer análisis visual, Justificar tu elección en un breve comentario, Realizar una visualización de dispersión (scatter plot) para observar cómo se distribuyen los datos en el espacio de esas dos variables.

3. Aplicación de agrupamiento (Clustering): Aplicar un algoritmo de agrupamiento que permita identificar grupos o segmentos dentro del conjunto de datos. Probar con diferentes cantidades de grupos para explorar cuál podría ser una buena elección. Evaluar visualmente si los grupos tienen sentido con base en su separación o forma.

4. Determinación del número óptimo de grupos: Utilizar un método visual o gráfico que te permita decidir cuántos grupos son más adecuados. Explicar cómo se interpretó ese resultado y qué decisión se tomó

5. Agrupamiento jerárquico: Aplicar un método jerárquico de agrupamiento. Representar visualmente los resultados con un **dendrograma**, que muestre cómo se relacionan las observaciones. Definir una distancia de corte para formar grupos y justificar tu elección. Comparar estos grupos con los obtenidos previamente.

6. Comparación de resultados Representar en una misma figura o en subgráficos los resultados obtenidos con ambos métodos. Comparar visualmente los grupos formados por cada técnica. Reflexionar si ambos métodos produjeron agrupamientos similares o diferentes.

Preguntas reflexivas (a entregar)

1. ¿Qué método te pareció más adecuado para este conjunto de datos? ¿Por qué?
2. ¿Qué criterios utilizaste para elegir el número de grupos?
3. ¿Cómo crees que influye la selección de variables en la formación de los grupos?
4. ¿Qué ventajas observaste en el agrupamiento jerárquico respecto al otro método?
5. ¿Cómo podrían usarse estos agrupamientos en un contexto médico o de investigación?

Entregables

- Entregar un notebook de jupiter (.ipynb) con los requerimientos solicitados
- Entregar un PDF con el análisis solicitado y la respuestas de las preguntas reflexivas

Rúbrica de evaluación

Criterio	Descripción	Puntaje Máximo
1. Exploración inicial de datos	Carga correcta del dataset, revisión de estructura, tipos de datos, valores faltantes, y estadísticas básicas.	**10 puntos**
2. Selección y justificación de variables	Elección de variables adecuada y justificación razonada (no aleatoria). Inclusión de visualización inicial.	**10 puntos**
3. Aplicación del primer método de clustering	Aplicación correcta del método, con resultados visuales claros y análisis mínimo.	**15 puntos**
4. Determinación del número de clusters	Uso de un método gráfico (como el codo) u otro razonable para elegir el número óptimo de clusters.	**10 puntos**
5. Agrupamiento jerárquico y dendrograma	Aplicación correcta del método jerárquico. Dendrograma bien interpretado y corte adecuado.	**15 puntos**
6. Comparación visual y conceptual	Subplots o gráficos comparativos bien presentados. Análisis de similitudes y diferencias entre los métodos.	**15 puntos**
7. Respuestas reflexivas	Respuestas completas, coherentes y con profundidad a las preguntas asignadas. Muestra reflexión personal o grupal.	**15 puntos**
8. Presentación del informe o notebook	Claridad en el formato, uso adecuado de títulos, etiquetas, comentarios y orden lógico del análisis.	**5 puntos**
9. Ortografía, redacción y estilo	Buena redacción, sin errores graves de ortografía o gramática. Lenguaje técnico adecuado.	**5 puntos**