Análisis Exploratorio de Datos en el Sector Biosanitario

Tema: Interpretación de la información contenida en un conjunto de datos.

Objetivo práctico: Interpretar los datos a través de estadísticos básicos y técnicas de visualización.

Descripción: En este ejercicio, se realizará un análisis exploratorio de un conjunto de datos biosanitarios. Deberán calcular estadísticos básicos como la media, mediana y desviación estándar, además de crear visualizaciones como histogramas y diagramas de dispersión para observar la distribución de los datos. Finalmente, aplicarán un algoritmo de *K-means clustering* para identificar posibles grupos o patrones dentro de los datos y graficar los resultados.

Instrucciones:

1. **Obtención del conjunto de datos:** Utiliza el conjunto de datos sintéticos. Este conjunto debe contener características como la edad, el nivel de glucosa en sangre, la presión arterial, entre otras.

2. **Exploración de los datos:**

    - Carga los datos en un entorno de trabajo usando *Pandas*.
    - Inspecciona las primeras filas del conjunto de datos para familiarizarte con su estructura.
3. **Estadísticos básicos:**

    - Calcula la **media**, **mediana** y **desviación estándar** de las principales variables del conjunto de datos, como la edad, el nivel de glucosa y la presión arterial.
    - Comenta sobre la tendencia central y la dispersión de los datos basándote en estos estadísticos.
4. **Visualización de los datos:**

    - Crea un **histograma** para visualizar la distribución de variables como el nivel de glucosa en sangre, la edad y la presión arterial.

Comenta sobre la forma de la distribución (por ejemplo, si es simétrica, sesgada, etc.).
    - Crea un **diagrama de dispersión** para explorar la relación entre dos variables de interés, como la edad y el nivel de glucosa. Observa si parece haber algún patrón o correlación.
5. **Aplicación de K-means clustering:**
    - Normaliza o estandariza las variables si es necesario.
    - Aplica el algoritmo **K-means clustering** para agrupar a los pacientes en clusters. Escoge un valor apropiado para el número de clusters (por ejemplo, 3 o 4 clusters, dependiendo de los datos).
    - Visualiza los resultados de los clusters usando un diagrama de dispersión, donde los puntos de diferentes clusters estén coloreados de manera distinta.
    - Interpreta los resultados de los clusters y explica si hay algún patrón o grupo que se pueda asociar con características específicas de los pacientes.
6. **Conclusiones:**
    - Comenta sobre los hallazgos obtenidos de los estadísticos básicos y las visualizaciones. ¿Cómo afecta la edad y el nivel de glucosa en sangre a la salud de los pacientes?
    - Reflexiona sobre cómo el análisis de clusters puede ayudar a identificar grupos de pacientes con características similares que podrían requerir tratamientos específicos.

Herramientas y bibliotecas recomendadas:

- **Pandas** para la manipulación de datos.
- **Matplotlib** y **Seaborn** para las visualizaciones.
- **Scikit-learn** para la implementación de K-means clustering.