**Análisis Exploratorio de Datos (EDA)**

**Indice:**

1. **Carga y visualización inicial de los datos**
2. **Selección y filtrado de columnas relevantes**
3. **Análisis descriptivo y limpieza de datos**
4. **Visualización de distribuciones y variables categóricas**
5. **Análisis de correlación y visualización de heatmap**
6. **Relación entre diagnóstico y medicación (análisis de contingencia)**
7. **Relación entre variables clínicas y edad**
8. **Análisis de motivo de alta según edad y tratamiento**
9. **Modelado predictivo con Random Forest**
10. **Detección y tratamiento de duplicados y valores nulos**
11. **Identificación y conversión de variables categóricas y numéricas**
12. **Detección de valores atípicos en variables numéricas**
13. **Codificación de variables categóricas**
14. **División de los datos en conjuntos de entrenamiento, validación y test**

Este notebook realiza un análisis exploratorio de datos (EDA) sobre un conjunto de datos de pacientes con enfermedades crónicas, centrándose especialmente en casos de Covid-19. El objetivo es comprender la estructura de los datos, identificar patrones relevantes y preparar el dataset para futuros análisis y modelado predictivo.

**Objetivos**

* Filtrar y analizar datos sensibles relacionados con el Covid-19.
* Explorar la estructura y calidad de los datos sintéticos generados.
* Identificar variables relevantes y relaciones entre ellas (edad, sexo, estancia en UCI, temperatura, saturación de oxígeno, fármacos, diagnósticos y motivo de alta).
* Detectar y tratar valores atípicos, nulos y duplicados.
* Analizar la relación entre diagnósticos, medicación, edad y motivo de alta.
* Preparar los datos para su uso en modelos predictivos, incluyendo codificación de variables categóricas y partición en conjuntos de entrenamiento, validación y test.

**Metodología**

1. **Carga y visualización inicial de los datos**: Se importan los datos reales sintéticos y se realiza una inspección preliminar de las primeras filas.
2. **Identificación de columnas clave**: Se seleccionan y renombran las variables de interés para el análisis clínico.
3. **Análisis descriptivo**: Se exploran las distribuciones de variables numéricas y categóricas, identificando posibles valores anómalos o inconsistentes.
4. **Limpieza de datos**: Se detectan y eliminan filas y columnas con valores nulos, duplicados o constantes. Se rellenan valores faltantes según corresponda.
5. **Visualización**: Se emplean histogramas, diagramas de caja, mapas de calor y gráficos de barras para explorar la distribución y relaciones entre variables.
6. **Análisis de relaciones**: Se estudia la asociación entre diagnósticos y medicación (Cramér's V), así como la relación entre edad, motivo de alta y otros factores clínicos.
7. **Preparación para modelado**: Se codifican variables categóricas, se identifican variables numéricas disfrazadas y se realiza la partición de los datos en conjuntos de entrenamiento, validación y test.
8. **Modelado exploratorio**: Se implementa un modelo de Random Forest para predecir el motivo de alta y se analiza la importancia de las variables.

**Resumen de los análisis realizados**

* Se identificaron y corrigieron valores atípicos y nulos en variables críticas como temperatura y saturación de oxígeno.
* Se analizaron las distribuciones de edad, sexo, estancia en UCI y motivo de alta, observando tendencias clínicas relevantes (por ejemplo, mayor mortalidad en pacientes de mayor edad).
* Se exploró la relación entre diagnósticos y tratamientos farmacológicos, así como la asociación entre motivo de alta y variables clínicas.
* Se preparó el dataset para su uso en modelos predictivos, asegurando la calidad y homogeneidad de las variables.
* Se realizó una partición estratificada de los datos y se evaluó un modelo de clasificación, identificando las variables más influyentes en el motivo de alta.

Este notebook proporciona una base sólida para el análisis clínico y el desarrollo de modelos predictivos en el contexto de enfermedades crónicas y preeclampsia, garantizando la privacidad mediante el uso de datos sintéticos.