**Anotaciones sobre EDA realizado basado en dataset: Framinghan\_clean**

- El dataset contiene 4240 filas y 19 columnas.

- Algunas de las columnas incluyen `Unnamed: 0`, `gender`, `age`, `education`, `currentSmoker`, `cigsPerDay`, `BPMeds`, `prevalentStroke`, `prevalentHyp`, `diabetes`, `totChol`, `sysBP`, `diaBP`, `BMI`, `heartRate`, `glucose`, `TenYearCHD`, `age\_groups`, y `heart\_rate\_groups`.

* Exploración General

- El dataset no tiene valores faltantes.

- La columna `BPMeds` contenía valores de texto ("Missing"), que fueron reemplazados por el valor numérico 2.

- La columna `gender` fue convertida a texto con valores "M" y "F".

- La columna `currentSmoker` fue convertida a texto con valores "NO" y "YES".

* Resumen Estadístico

- Algunas estadísticas clave para las variables numéricas:

- Edad (age): Media de 49.58 años, rango de 32 a 70 años.

- Cigarrillos por día (cigsPerDay: Media de 8.94 cigarrillos por día, con muchos valores de 0.

- Presión arterial sistólica (sysBP): Media de 132.35 mmHg.

- Presión arterial diastólica (diaBP): Media de 82.90 mmHg.

- Índice de masa corporal (BMI): Media de 25.80.

- Glucosa (glucose): Media de 81.60 mg/dL.

- Frecuencia cardíaca (heartRate): Media de 75.88 latidos por minuto.

* Análisis de Outliers

- Se observó que algunas variables, como `cigsPerDay`, `BPMeds` y `prevalentStroke`, tienen muchos valores de cero, lo que podría indicar la ausencia de estos factores en una gran parte de la población estudiada.

* Observaciones y Recomendaciones

1. Valores Faltantes:

- No se encontraron valores faltantes, lo cual es positivo y facilita el análisis y el modelado posterior.

2. Distribución de Variables:

- Algunas variables, como `cigsPerDay` y `BPMeds`, tienen muchos valores de cero. Esto podría sugerir que muchos participantes no fuman o no toman medicamentos para la presión arterial.

3. Conversiones de Datos:

- La conversión de `BPMeds`, `gender` y `currentSmoker` a formatos más adecuados (numérico y texto) es un buen paso para asegurar que el análisis y el modelado sean precisos.

4. Análisis Bivariado:

- Se podría realizar un análisis bivariado adicional para explorar cómo las variables independientes (por ejemplo, `age`, `cigsPerDay`, `sysBP`, etc.) están relacionadas con la variable objetivo `TenYearCHD`.

5. Visualizaciones:

- Las visualizaciones adicionales, como gráficos de dispersión, diagramas de caja y violín, pueden proporcionar más información sobre la relación entre las variables y ayudar a identificar patrones y outliers.

* Próximos Pasos (?)

1. Modelado Predictivo:

- Utilizar los datos limpios y preprocesados para construir modelos predictivos, como regresión logística, Random Forest, Gradient Boosting, XGBoost, etc.

- Evaluar el rendimiento de los modelos utilizando métricas como precisión, recall, F1-score y AUC.

2. Validación Cruzada:

- Realizar validación cruzada para asegurar que los modelos no están sobreajustados y que tienen un buen rendimiento en datos no vistos.

3. Feature Engineering:

- Crear nuevas características derivadas de las existentes para mejorar la capacidad predictiva del modelo.