Evidencia 2 Módulo Analista de Datos: Documentación.

Grupo Cluster-Fusion - **Repositorio**

Origen de los datos:

El conjunto de datos fue recopilado en el Hospital Zheen, ubicado en Erbil, Irak, durante el período comprendido entre enero de 2019 y mayo de 2019. La recopilación fue parte de un trabajo vinculado a investigaciones médicas, específicamente en el área de enfermedades cardíacas.

Este dataset fue utilizado por la University of Kurdistan Hewler y está publicado en Kaggle bajo el nombre *Heart Attack Dataset*. Está orientado a la clasificación binaria entre pacientes con y sin ataque cardíaco.

Enlace al conjunto de datos en Kaggle: Heart Attack Dataset en Kaggle

El *Dataset* tiene 1351 filas , lo que significa que hay 1351 registros de datos, y tiene 9 columnas.

Descripción de cada columna y tipo de dato.

Variable	Tipo de Variable	Descripción	Rango observado	Rango clínicamente posible
Age	Cuantitativa Discreta (entera)	Edad del paciente en años	14 a 103	0 a 120
Gender	Categórica Dicotómica (binaria)	Género del paciente	1: Mujer, 0: Hombre	Hombre, Mujer
Heart Rate	Cuantitativa Continua (decimal)	Frecuencia cardíaca (latidos por minuto)	20 a 1111 bpm	20 a 250 bpm
Systolic Blood Pressure	Cuantitativa Continua (decimal)	Presión arterial sistólica (mmHg)	42 a 223 mmHg	50 a 300 mmHg
Diastolic Blood Pressure	Cuantitativa Continua (decimal)	Presión arterial diastólica (mmHg)	38 a 154 mmHg	30 a 180 mmHg
Blood Sugar	Cuantitativa Continua (decimal)	Nivel de glucosa en sangre (mg/dL)	35 a 541	30 a 600 mg/dL
CK-MB	Cuantitativa Continua (decimal)	Marcador de daño cardíaco (Creatina quinasa MB)	0.321 a 300	0 a 200 ng/mL
Troponin	Cuantitativa Continua (decimal)	Marcador de daño cardíaco (Troponina)	0.001 a 10.3	0 a 100 ng/mL
Result	Categórica Nominal (texto)	Resultado diagnóstico de ataque cardíaco	negative / positive	positive / negative

¿Qué significa que los datos estén en formato RAW?

En ciencia de datos, RAW se refiere a datos en estado bruto, es decir, no procesados ni modificados desde su fuente original. Esto implica que los datos:

- No han sido limpiados (pueden tener errores, duplicados o valores faltantes).
- No han sido transformados (están tal como fueron medidos o registrados).
- No han sido normalizados o codificados manualmente.
- Generalmente requieren preprocesamiento antes de poder usarse para modelado.

En el caso de nuestro Dataset tomado de Kagle el hecho de que la columna *Gender* esté codificada ya es un indicio de un procesamiento previo. Además, se encuentran valores faltantes. Podemos concluir que está parcialmente procesado.