

# Heart Disease



Este es un conjunto de datos multivariado, involucra una variedad de variables matemáticas o estadísticas separadas, análisis de datos numéricos multivariados. Está compuesto por 14 atributos que son edad, sexo, tipo de dolor en el pecho, presión arterial en reposo, colesterol sérico, azúcar en la sangre en ayunas, resultados electrocardiográficos en reposo, frecuencia cardíaca máxima alcanzada, angina inducida por ejercicio, depresión del segmento ST inducida por ejercicio en relación al reposo, la pendiente del segmento ST de ejercicio máximo, número de vasos principales y talasemia. Esta base de datos incluye 76 atributos, pero todos los estudios publicados se relacionan con el uso de un subconjunto de 14 de ellos. La base de datos de Cleveland es la única utilizada por los investigadores en aprendizaje automático hasta la fecha. Una de las principales tareas en este conjunto de datos es predecir, basándose en los atributos dados de un paciente, si esa persona en particular tiene enfermedad cardíaca o no, y la otra es la tarea experimental de diagnosticar y descubrir varias ideas de este conjunto de datos que podrían ayudar a comprender mejor el problema.

## Descripciones de Variables:

- Id: Identificación única para cada paciente.
- age: Edad del paciente en años.
- origin: lugar de estudio.
- Sex: Masculino/Femenino.
- cp tipo de dolor en el pecho ([angina típica, angina atípica, no anginal, asintomática])
- trestbps presión arterial en reposo (presión arterial en reposo (en mm Hg al ingresar al hospital))
- chol (colesterol sérico en mg/dl)
- fbs (si el azúcar en la sangre en ayunas es  $> 120$  mg/dl)
- restecg (resultados electrocardiográficos en reposo)
- Values: [normal, anormalidad de stt, hipertrofia ventricular izquierda]
- thalach: frecuencia cardíaca máxima alcanzada
- exang: angina inducida por ejercicio (Verdadero/ Falso)
- oldpeak: depresión del segmento ST inducida por ejercicio en relación al reposo
- slope: la pendiente del segmento ST de ejercicio máximo
- ca: número de vasos principales (0-3) coloreados por fluoroscopia
- thal: [normal; defecto fijo; defecto reversible]

- num: el atributo predicho

A continuación, se presentan algunas fórmulas de cálculos de índices de salud que han servido como referencia en los cálculos de los índices para este conjunto de datos:

## Fórmulas para calcular algunos índices de salud

### 1. Índice de Colesterol (IC)

El Índice de Colesterol se calcula dividiendo el colesterol sérico entre el colesterol HDL:

$$IC = \frac{\text{colesterol\_serico}}{\text{colesterol\_HDL}}$$

### 2. Frecuencia Cardíaca Máxima Estimada (FCM)

La Frecuencia Cardíaca Máxima Estimada se puede calcular utilizando la fórmula de Tanaka, que estima la FCM en función de la edad del paciente:

$$CM = 208 - (0.7 \times edad)$$

### 3. Índice de Presión Arterial (IPA)

El Índice de Presión Arterial se calcula dividiendo la presión arterial sistólica entre la presión arterial diastólica:

$$IPA = \frac{\text{presion\_arterial\_sistolica}}{\text{presion\_arterial\_diastolica}}$$

Acceda a la información de los datos a través del siguiente enlace:

[https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data?select=heart\\_disease\\_uci.csv](https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data?select=heart_disease_uci.csv)

## Resuelva las siguientes preguntas:

1. Calcular la edad promedio de los pacientes en el conjunto de datos y su desviación estándar.
2. ¿Cuál es el porcentaje de pacientes de sexo masculino y femenino en el conjunto de datos?
3. Comparar la presión arterial promedio entre pacientes con y sin enfermedad cardíaca.
4. Calcular el índice de masa corporal (IMC) promedio de los pacientes y compararlo entre aquellos con y sin enfermedad cardíaca.

$$IMC = \frac{\text{peso}}{\text{altura}^2}$$

5. Reportar las estadísticas descriptivas de las variables relevantes del conjunto de datos.
6. ¿Cuál de las variables relacionadas con la salud cardiovascular (por ejemplo, colesterol sérico, presión arterial en reposo, etc.) tiene el mayor rango?
7. ¿Cuántas categorías diferentes de enfermedades cardíacas se observan en hombres y

cuántas en mujeres?

8. Calcular el porcentaje de pacientes con niveles de colesterol sérico por encima del promedio, según el nivel de riesgo de enfermedad cardíaca (por ejemplo, bajo, medio, alto) basado en otras variables como la presión arterial y el nivel de glucosa en sangre.
9. Crear una nueva columna llamada 'average\_risk' que indique si un paciente tiene un riesgo cardiovascular promedio, calculado en base a múltiples variables de riesgo.

