**BASIC INFO**

**RECORDS: 303**

**Features:14**

**13 columns are independent variables**

**1 column (target)**

**Feature Breakdown:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **Description** | **Type** | **Impact on Heart Disease** |
| **age** | Age of the patient | Numerical | Older age increases risk |
| **sex** | Gender (0 = Female, 1 = Male) | Categorical | Males have a higher risk |
| **cp** | Chest pain type (0-3) | Categorical | Higher values indicate higher risk |
| **trestbps** | Resting blood pressure (mm Hg) | Numerical | Higher values indicate risk |
| **chol** | Serum cholesterol (mg/dl) | Numerical | High cholesterol increases risk |
| **fbs** | Fasting blood sugar (>120 mg/dl, 1 = True) | Binary | High values indicate risk |
| **restecg** | Resting ECG results (0-2) | Categorical | Abnormal values indicate risk |
| **thalach** | Maximum heart rate achieved | Numerical | Lower values indicate higher risk |
| **exang** | Exercise-induced angina (1 = Yes, 0 = No) | Binary | Positive angina indicates risk |
| **oldpeak** | ST depression induced by exercise | Numerical | Higher values indicate risk |
| **slope** | Slope of peak exercise ST segment | Categorical | Higher values indicate risk |
| **ca** | Number of major vessels (0-3) colored by fluoroscopy | Numerical | Higher values indicate risk |
| **thal** | Thalassemia (0-3) | Categorical | Certain values indicate disease presence |
| **target** | Presence of heart disease (1 = Yes, 0 = No) | Binary | Prediction variable |

Linear features: age, cholesterol, blood pressure

Non linear: chest pain type (cp), ST depression (oldpeak), and thalassemia (thal)

Binary classification dataset: 0 or 1

**Algorithm 1: KNN**

How KNN Works?

1️ Choose a value for "k" (number of nearest neighbors).

2️ Calculate the distance between the new data point and all training points (commonly using Euclidean distance).

3️ Find the k closest neighbors to the new data point.

4️ Majority voting (Classification): The class with the most neighbors is assigned.

Example: If k=5 and 3 neighbors are "Heart Disease" and 2 are "No Disease," it predicts "Heart Disease."

5️ Averaging (Regression): The prediction is the average of k-nearest values.

Disadv:

* slower than random forest
* requires feature scaling

**Algorithm 2: Random forest**

* Handles both cat an numerical data
* Linearity handling

**Algorithm 3: XG boost**

Imagine you are a teacher giving a test to students. After checking their answers, you notice some common mistakes. The next time you teach, you focus more on those **mistakes** so students can improve.

🔹 **XGBoost works similarly!** It learns from mistakes step by step to improve predictions. It does this using **many small decision trees** that fix errors made by previous trees. Each new tree tries to correct the errors made by earlier trees, making the model more accurate.

Datset updated 5 months ago

1. Zip file : plotting of age vs chol
2. Zip file: different algos and feature importance