Introduction

Cardiovascular diseases have been the most common cause of death worldwide over the last few decades in developed as well as underdeveloped and developing countries. Early detection of cardiac diseases and continuous supervision of clinicians can reduce the mortality rate. However, it is not possible to monitor patients every day in all cases accurately and consultation of a patient for 24 hours by a doctor is not available since it requires more sapience, time, and expertise.

Every day, the average human heart beats around 100,000 times, pumping gallons of blood through the much less noticeable than the signs of a man. In women, heart attacks may feel uncomfortable squeezing, pressure, fullness, or pain in the center of the chest. It may also cause pain in one or both arms, the back, neck, jaw, stomach, shortness of breath, nausea, and other symptoms.

Men experience typical symptoms of heart attack, such as chest pain, discomfort, and stress. They may also experience pain in other areas, such as arms, neck, back, and jaw, and shortness of breath, sweating, and discomfort that mimics heartburn. It's a lot of work for an organ which is just like a large fist and weighs between 8 and 12 ounces.

Objective of Data

The objective of the UCI Heart Disease dataset is to facilitate research and analysis aimed at developing

Predictive models for the detection and assessment of heart disease. Specifically, the dataset aims to:

1. Enable Prediction: Provide a diverse set of medical attributes and corresponding diagnoses to enable the development of machine learning models capable of predicting the likelihood of heart disease in patients.
2. Support Research: Serve as a valuable resource for researchers and data scientists interested in studying the factors associated with heart disease and exploring novel approaches to its diagnosis and treatment.
3. Promote Healthcare Innovation: Foster innovation in healthcare by empowering healthcare providers, businesses, and policymakers with data-driven insights into heart disease risk assessment and management.
4. Improve Patient Outcomes: Ultimately, the primary objective of the dataset is to contribute to the improvement of patient outcomes by facilitating early detection, intervention, and personalized treatment of heart disease.

How data can help businesses

1) Healthcare Providers: Hospitals and clinics can use these models to assess the risk of heart disease in patients during routine check-ups. This can lead to early detection and intervention, ultimately improving patient outcomes and reducing healthcare costs.

2) Insurance Companies: Insurance companies can utilize these models to assess the risk of heart disease in their policyholders. By identifying high-risk individuals, they can offer targeted interventions or wellness programs to mitigate the risk and reduce claims.

3) Pharmaceutical Companies: Pharmaceutical companies can use predictive models to identify potential candidates for clinical trials of new drugs aimed at preventing or treating heart disease. This can streamline the drug development process and bring new treatments to market more efficiently.

4) Health tech Startups: Startups focused on digital health and wellness can develop applications or wearable devices that utilize heart disease prediction models to provide personalized health recommendations to users. This can empower individuals to take proactive steps toward preventing heart disease.

Real-life Applications

1) Clinical Decision Support: Healthcare professionals can use these models as decision-support tools during patient consultations. By inputting patient data into the model, clinicians can obtain risk score and recommendations for further evaluation or treatment.

2) Public Health Initiatives: Public health authorities can utilize predictive models to identify populations at high risk of heart disease and implement targeted prevention strategies, such as educational campaigns, screening programs, or policy interventions.

3) Remote Monitoring: Remote monitoring devices equipped with heart disease prediction algorithms can continuously monitor individuals at risk and alert them or their caregivers of any significant changes or warning signs, enabling timely medical intervention.

4) Personalized Medicine: Predictive models can facilitate the shift towards personalized medicine by enabling healthcare providers to tailor treatment plans based on an individual's risk profile and genetic predisposition to heart disease.

About Dataset

* This is a multivariate type of dataset which means providing or involving various mathematical or statistical variables, and multivariate numerical data analysis.
* It is composed of 14 attributes which are age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, Old peak-ST depression induced by exercise relative to rest, the slope Of the peak exercise ST segment, number of major vessels and Thalassemia.
* This database includes 76 attributes, but all published studies relate to using a subset of 14 of them One of the major tasks of this dataset is to predict based on the given attributes of a patient whether that particular person has heart disease or not. The other is the experimental task to diagnose and find out various insights from this dataset which could help in understanding the problem more.

Column Descriptions

1) Id: Unique identifier for each patient

2) Age: Age of the patient in years

3) Origin: Place of study

4) Sex: Gender of the patient

5) Chest Pain: Chest pain type (e.g., typical angina, atypical angina, non-anginal, asymptomatic)

6) Trestbps: Resting blood pressure (mm Hg on admission)

7) Chol: Serum cholesterol level (mg/dl)

8) Fbs: Fasting blood sugar (>120 mg/dl)

9) Cestecg: Resting electrocardiographic results

10) Values: normal, ST-T abnormality, left ventricular hypertrophy

11) Thalach: Maximum heart rate achieved

12) Exang: Exercise-induced angina (True/False)

13) Oldpeak: ST depression induced by exercise relative to rest

14) Slope: Slope of the peak exercise ST segment

15) CQ: Number of major vessels colored by fluoroscopy (0-3)

16) Thal: Thalassemia diagnosis (normal, fixed defect, reversible defect)

17) Num: Predicted attribute indicating presence of heart disease

Challenges

1) Ensuring the accuracy and reliability of the medical data is crucial for building effective prediction Incomplete or inaccurate data can lead to biased or unreliable predictions.

2) Feature Selection: Identifying the most relevant features or attributes from the dataset that contribute to the prediction of heart is essential. This requires domain knowledge and careful analysis of the data.

3) Imbalanced Data: Imbalance in the distribution of classes (i.e., presence or absence of heart disease) can affect the performance of machine learning algorithms. Techniques such as oversampling, under- sampling, or using algorithms that handle imbalanced data well are necessary to address this issue.

4) Building models that not only provide accurate predictions but also Offer insights into the factors contributing to the prediction is important for gaining trust from healthcare professionals and patients.

Data Understanding

The size of the dataset is 920 rows and 16 attributes in which num is the dependent variable for which we have to make the prediction.

Dataset Overview

Based on the summary above, it appears that the data consists of a total Of 920 observations. However, many features in this dataset have missing values, including trestbps, Chol, fbs, restecg, thalch, exang, oldpeak, slope, ca, and thal. In addition, the dataset contains both numeric and categorical variables.

Pattern of Missingness

* Based on the heatmap above, missing values appear intensively starting from the 300th row.
* The top three variables with the highest number of observations with missing values are slope, ca, and thal.
* So far, it does not look like the missing values are distributed randomly.

Relationship between Cholesterol Levels and Heart Disease

* The box plot illustrates cholesterol levels across five heart disease categories, showing median values, range variability, and outliers.
* Categories 1 to 4 have similar medians, but the spread and outliers differ, with category 0 showing the most variability

Maximum Heart Rate and Heart Disease

* The plot shows a negative correlation where the maximum heart rate tends to decrease as age increases

The Impact of Exercise-Induced Angina on Heart Disease

* Most cases in category 0 do not report angina, while categories 1 through 4 show a more varied distribution, with both angina and non-angina cases present.
* The data suggests that exercise- induced angina is more commonly reported in individuals with heart disease categories 1 to 4 compared to category 0.