Introduction

**Cardiovascular diseases (CVDs)** are the leading cause of mortality worldwide, killing an estimated 17.9 million people each year, accounting for 31% of all fatalities worldwide. Heart attacks and strokes cause four out of every five CVD fatalities, and one-third of these deaths occur in adults under the age of 70. CVDs are a common cause of heart failure, and this dataset contains 11 variables that can be used to predict heart disease.

People with cardiovascular disease or at high cardiovascular risk (due to the presence of one or more risk factors such as cholesterol,old-age, or pre-existing illness) require early identification and care, which a machine learning model may greatly assist with.



**Attribute Information**

* **Age:** age of the patient [years]
* **Sex:** sex of the patient [M: Male, F: Female]
* **ChestPainType**: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
* **RestingBP:** resting blood pressure [mm Hg]
* **Cholesterol:** serum cholesterol [mm/dl]
* **FastingBS:** fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
* **RestingECG:** resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
* **MaxHR**: maximum heart rate achieved [Numeric value between 60 and 202]
* **ExerciseAngina:** exercise-induced angina [Y: Yes, N: No]
* **Oldpeak: oldpeak** = ST [Numeric value measured in depression]
* **ST\_Slope:** the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
* **HeartDisease**: output class [1: heart disease, 0: Normal]

**Key Finding (After Loading datasets):**

* In the Heart dataset a.k.a **df**, we have 918 rows and 12 coloumns
* Different scales apply to numerical numbers. I must exercise caution while preprocessing since many algorithms call for scaling, and some converge more quickly when we scale all of our numerical variables to the same value. In order for the computer to grasp the catergorical characteristics, we also need to encode them.
* OldPeak's minimum value is -2.60.Resting blood pressure and cholesterol levels are both zero (which appears clinically implausible).
* There are significant class inequalities in various category variables, including 'Sex' and 'ExerciseAngina'. This may result in overfitting, in which our model overperforms on under-represented (unseen) categories. (Sex 'M' has 725 counts and ExerciseAngina: 'N' has '547' counts )

**Key Findings After performing various analysis :**

* Outliers and negative values within particular features ('RestingBP,' 'Cholesterol,' 'Oldpeak') that might contribute noise and distort the model's learning process should be avoided.

**Key Finding (data cleaning):**

* Only one patient has an invalid 'RestingBP'. This record can be deleted from our database. Cholesterol, on the other hand, has 172 incorrect observations.
* It would be preferable to replace the data with mean/median, or even better, regression. However, it would have an effect on the machine learning model. Instead of replacing, we may eliminate 172 clinically impossible data points from the dataset.
* I understand that eliminating this data isn't ideal because we lose a lot of essential information, but we still have a reasonable quantity of data.
* If the 'cholesterol' characteristic is unimportant, we may delete the entire column, which will enhance the model's performance.

***Key Findings in Univariate Analysis:***

* Based on the analysis of the "Gender" categorical variable, it is evident that the dataset has a majority of males (79%).
* The age distribution is relatively symmetrical, with a peak in the middle, suggesting that dataset contains a wide range of ages, but mostly lies in the range between 50 and 60's.
* Asymptomatic (ASY) and non-anginal (NAP) chest pain are the two most typical types. Patients with TA (typical angina) and ATA (atypical angina) are less common.
* When exercising, most people do not develop angina (chest discomfort). Given that there is more information available for individuals without exercise-induced anginal symptoms, the model may perform better in their case.
* Most patients' resting electrocardiographic results are normal. The remaining patients are approximately evenly divided between those who have ST-T wave abnormalities and those who exhibit possible or certain left ventricular hypertrophy.

***Key Findings in Bivariate Analysis:***

* Age & Sex: In the sample, older males are more likely to develop heart disease than older females.
* Heart disease is more likely to affect persons whose resting blood pressure is more than 140.
* Data on cholesterol indicate that heart disease and resting blood pressure have a similar association. It implies that there may be a slender correlation between heart disease and increased cholesterol levels. Several individuals have abnormal blood test results that are above 600 mg/dl, and they are all heart disease patients.
* MaxHR Heart disease is less likely to affect those with greater maximum heart rates.
* Patients who have cardiac disease are more likely to have a greater Oldpeak value (ST depression brought on by activity compared to rest).
* Chest Pain Type: Heart disease is more likely to affect patients with asymptomatic chest pain (ASY) than with non-anginal pain (NAP) or atypical angina (ATA).
* Exercise-Induced Angina: Heart disease is more prevalent in those who have angina while exercising.
* Resting ECG: Heart disease is more likely to affect patients with ST-T wave abnormalities (ST) or probable/definite left ventricular hypertrophy (LVH) than it is to affect individuals with normal ECGs.
* ST Slope: Heart disease is more likely to affect patients with a flat ST slope than those who have an up- or down-sloping ST slope.

***Key Findings in Multivariate Analysis:***

* Heart Disease Distribution: The combined box plots provide an overview of how various numerical variables and categorical factors relate to heart disease presence or absence. Heart disease cases are generally represented in orange, while non-heart disease cases are in blue.
* Age and Gender: Males are more likely than females to experience the middle 50% of heart disease cases at a younger age, suggesting that male typically begin suffering heart disease at a younger age.
* Exercise and MaxHR: The lower MaxHR readings in this group may be attributable to people with angina who are unable to exercise at their maximum heart rate because of discomfort or pain they suffer.
* Oldpeak and ST\_Slope: They are significant predictors of heart disease. Particularly, patients with a downward ST\_Slope and high Oldpeak (above 2.5) are very likely to have heart disease.
* Maximum Heart Rate and Heart Disease: Patients without heart disease tend to have slightly higher maximum heart rates (MaxHR) than those with heart disease. However, there is overlap between the two groups, and MaxHR alone may not be a definitive predictor of heart disease.

***Key Findings in Correlation Analysis:***

* age & maxhr : Age and maximal heart rate (MaxHR) have a somewhat negative correlation. The maximal heart rate is often a little lower in older persons.
* oldpeak & maxhr : These two feature have a moderate negative correlation suggesting that as "Oldpeak" values slightly increase (indicating more ST depression), "MaxHR" tends to decrease slightly. This indicates that, correlation between them is very weak, indicating that there is no strong linear relationship between these variables.
* heart-disease : Features ( **age , oldpeak, fastinbs, restingbs** ) shows a strong correlation with heart disease.

**Reason to choose logistic reason**

* Efficiency and Simplicity: Logistic regression is computationally effective, enabling quick predictions that are crucial, particularly in the beginning phases.
* Interpretability: In logistic regression, a coefficient is given to each feature. In healthcare contexts where interpretability might affect clinical choices, this quantitative metric offers clear insights into the effect of the attribute.
* Binary Classification: The method is built for binary classification, which perfectly matches our need to categorise categorical information into two different groups (0 & 1).

**ROC Curve**

The genuine positive rate (sensitivity) is shown against the false positive rate (1-specificity) on the Receiver Operating Characteristic (ROC) curve. The top left corner of a ROC curve is where the optimal point should be. The blue dashed line shows that our model outperforms a random prediction by a large margin.

**Project Key Summary**

**Numerical Anlaysis**

* Age: Heart disease is more likely to affect older people. It's interesting to note that men tend to get heart disease earlier than women do.
* MaxHR: Lower risk of heart disease is connected with higher maximum heart rates. Age and occurrences of chest discomfort both cause a drop in MaxHR.
* Oldpeak: Elevated Oldpeak levels suggest an increased risk of heart disease.
* RestingBP & Cholesterol: While typically regarded as risk factors, resting blood pressure and cholesterol had rather minimal associations with heart disease in our sample. Only outliers with abnormally high cholesterol levels are linked to heart disease.

**Categorical Analysis**

* Sex: There is a pronounced gender disparity, with males being not only more numerous but also more likely to suffer from heart disease.
* Asymptomatic chest discomfort is a key signal, and people who experience it are more likely to develop heart disease.
* ExerciseAngina: Angina triggered by exercise is a strong predictor of heart disease.
* Resting ECG: The majority of individuals have normal readings, although cardiac disease is more frequently seen in those with left ventricular hypertrophy or ST-T wave abnormalities.
* ST\_Slope: A flat ST slope is concerning since it suggests an increased risk of heart disease.

**Additional Key Insight**

* Interactions: Age exhibits interactions with a number of characteristics, including MaxHR and RestingBP, offering possible words for modelling interactions.
* Potential Biases: The dataset shows imbalances in the categories for characteristics like Sex and ChestPainType, which may cause biases in model predictions.
* Features such as ( Oldpeak, MaxHR, RestingECG ) were the most important feature for prediction heart disease.