# PROJECT REPORT

## Heart Disease Prediction

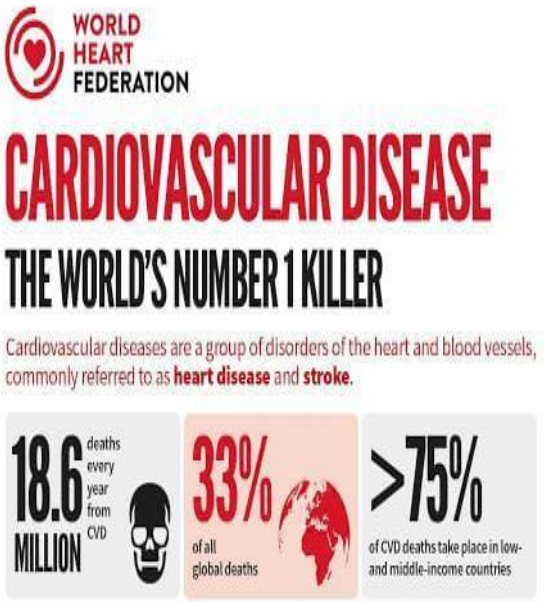
## Abstract :

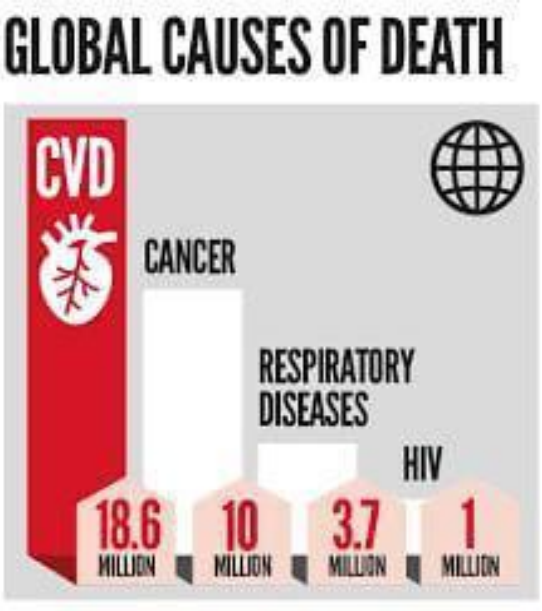
Heart disease is easier to treat when it is detected in the early stages. Machine learning techniques may aid a more efficient analysis in the prediction of the disease. Moreover, this prediction is one of the most central problems in medicine, as it is one of the leading diseases related to an unhealthy lifestyle. So, an early prediction of this disease will be useful for a cure or aversion.

Analyze the heart disease dataset to explore the machine learning algorithms and build a decision tree model to predict the disease. The current study is on Heart disease having medical risk factors such as chest pain experienced by a patient, Blood pressure of the patient, The patient's maximum heart rate and some other risk factors. In this problem, several classification based algorithms from Decision Tree Algorithm are used to build and predict models. The models are evaluated on various performance metrics like confusion matrix, accuracy, sensitivity, f1 score, AUC and ROC curve.


## 1. Introduction :

Analyze heart disease by a dataset which includes the 14 variables which are used to analyze and identify key factors of heart disease. It is also required to build efficient machine learning models which can predict the risk factors of heart disease.




**Source Link : https://world-heart-federation.org/**

The World Heart Federation says that 18.6m People around the world die from Cardiovascular Disease (Heart Disease). Which is 33% of all global deaths. In fact more than 75% of deaths on heart disease take place in Lower-income countries and Middle-income countries. Some Lower-income countries and Middle-income countries are Afghanistan, Somalia, Sudan, Tanzania, Yemen, Bangladesh, India, Pakistan and some more countries. Most deaths based on disease are Heart Disease, about 18.6m globally.


World Heart Federation states that risk factors for heart disease are High BP, High Cholesterol, Overweight, Physical Inactivity, Unhealthy Diet, Diabetes, Kidney disease, using Tobacco and alcohol. The World Heart Federation suggests that people avoid Unhealthy diets, reduce physical inactivity, Control blood pressure and diabetes levels, Monitor cholesterol levels, Avoid Smoking and drinking alcohol, and maintain exercise habits that can reduce the risk of heart disease.

## 2. EDA and Business Implication :

Data Information and Data types

```
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   age                  303 non-null    int64
 1   gender               303 non-null    int64
 2   chest_pain           303 non-null    int64
 3   rest_bps             303 non-null    int64
 4   cholestrol           303 non-null    int64
 5   fasting_blood_sugar  303 non-null    int64
 6   rest_ecg             303 non-null    int64
 7   thalach              303 non-null    int64
 8   exer_angina          303 non-null    int64
 9   old_peak             303 non-null    float64
 10  slope                303 non-null    int64
 11  ca                   303 non-null    int64
 12  thalassemia          303 non-null    int64
 13  target               303 non-null    int64
dtypes: float64(1), int64(13)
```

**Fig : 2.1 Data information and Data Types**

The Variables and their data types are appropriately listed in Fig: 2.1, there are 303 observations and 14 variables including demographic information such as age, gender, chest pain, rest bps, cholesterol, fasting blood sugar, rest ecg, thalach, exer angina, old peak, slope, ca, thalassemia, target. The target variable here is target which indicates whether the patient has heart disease or not.

## Checking For Class Imbalance :

Class imbalance Class imbalance is a situation that occurs when the distribution of classes in the training data used to build a machine learning model is not equal. This can result in the model being biased towards the majority class and performing poorly on the minority

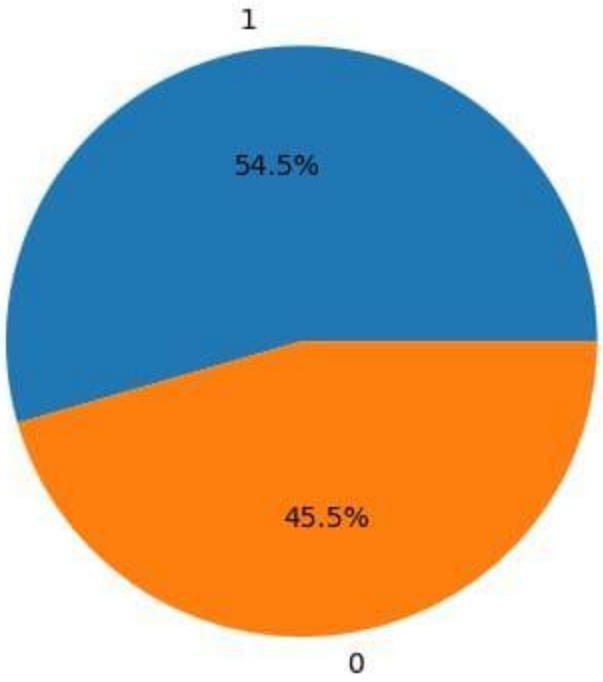



**Fig : 2.2 Checking Class Imbalance**

In this class variable where 0 is not having heart disease and 1 is having heart disease. For the model building our model will perform good even if it is imbalanced because 54.5% of data belongs to persons having heart disease. Model will predict a person having heart disease without any error.

## 3.Data Analysis using Plots :

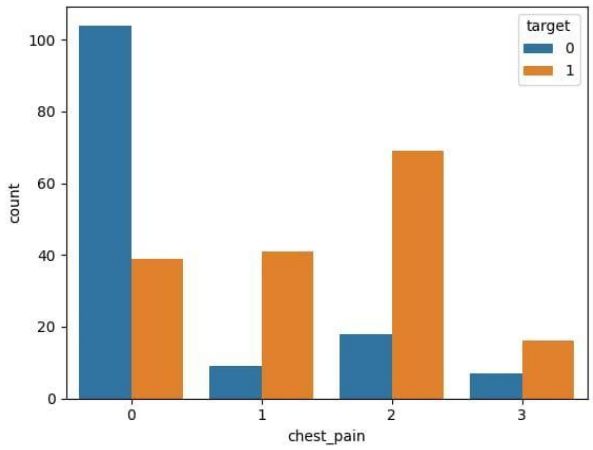



**Fig : 3.1 Chest Pain**

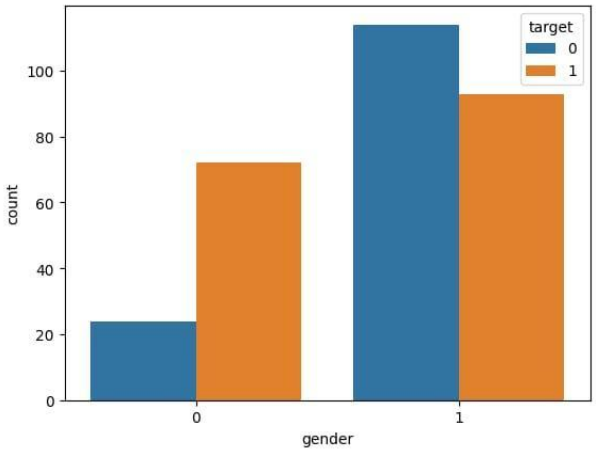



**Fig : 3.2 Gender**

Experiencing chest pain can be a concerning symptom associated with heart disease from Fig **3.1** various levels of the chest are ranked as 0,1,2,3 based on the level of chest pain from the initial stage to extreme stage risk factor is getting gradually increased.

The risk of heart disease can vary depending on gender, Historically, male have been found to have a higher risk of heart disease compared to female from the **Fig 3.2** we can also observe that male is at high risk of having heart disease.

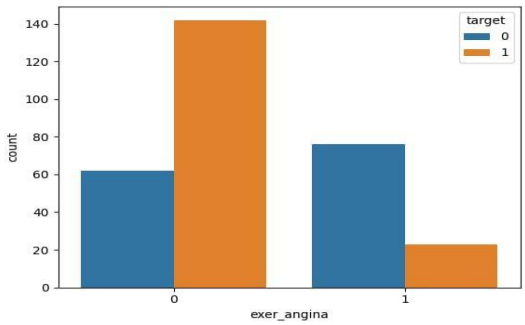



**Fig : 3.3 EXER ANGINA**

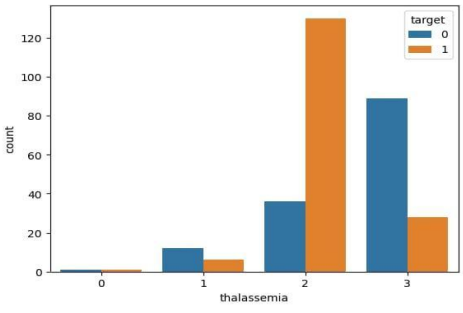



**Fig : 3.4 THALASSEMIA**

Exertional angina is a type of chest pain that occurs during physical activity or exercise and is typically associated with coronary artery disease (CAD). CAD is a type of heart disease that involves the narrowing or blockage of coronary arteries, which supply blood and oxygen to the heart muscle**.** The presence of exertional angina can indicate an increased risk of heart disease from Fig **3.3** we can also observe that a person having exertional angina has a higher risk of heart disease**.**

Thalassemia is a genetic blood disorder that affects the production of hemoglobin, the protein in red blood cells that carries oxygen. The risk of getting heart disease in individuals with thalassemia can vary based on genetic blood disorder. From Fig **3.4** we can also observe that a person having thalassemia will have a chance of getting heart disease.

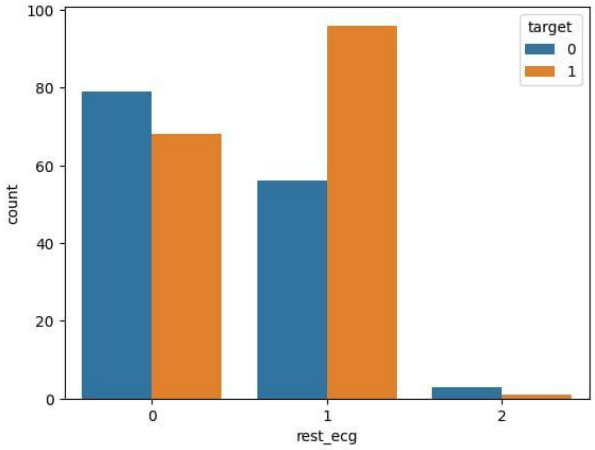



**Fig : 3.5 REST ECG**

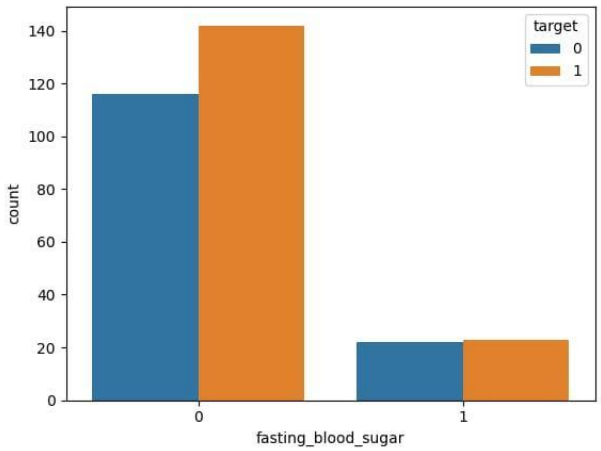



**Fig : 3.6 FASTING BLOOD SUGAR**

The level of potassium in the blood, also known as serum potassium level, is an important electrolyte that plays a critical role in the proper functioning of the heart and other muscles in the body. Abnormal potassium levels, either too high (hyperkalemia) or too low (hypokalemia), can potentially impact heart health and increase the risk of heart disease. From Fig **3.7** we can observe that potassium level in the blood impacts on getting heart disease but we are not able to predict the chances of getting heart disease based on potassium level.

The relationship between blood sugar levels, specifically fasting blood sugar levels, and the risk of heart disease is well-established. High blood sugar levels, also known as hyperglycemia, can contribute to the development of various risk factors for heart disease, which may increase the overall risk of cardiovascular complications. Even in individuals without diabetes, elevated fasting blood sugar levels, even within the normal range, can still be associated with an increased risk of heart disease. From Fig **3.8** we can also observe that fasting blood sugar levels have more impact on getting heart disease.

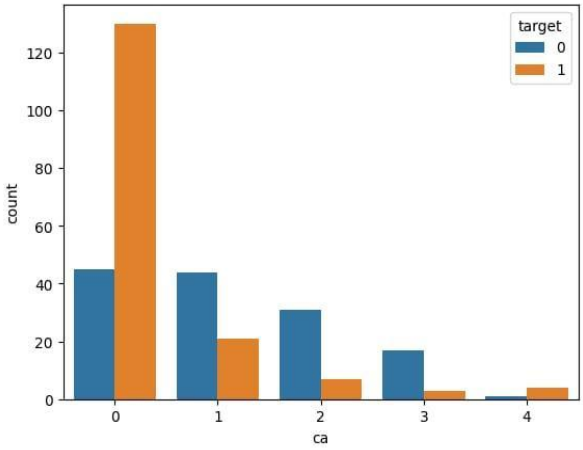



**Fig : 3.7 CA**

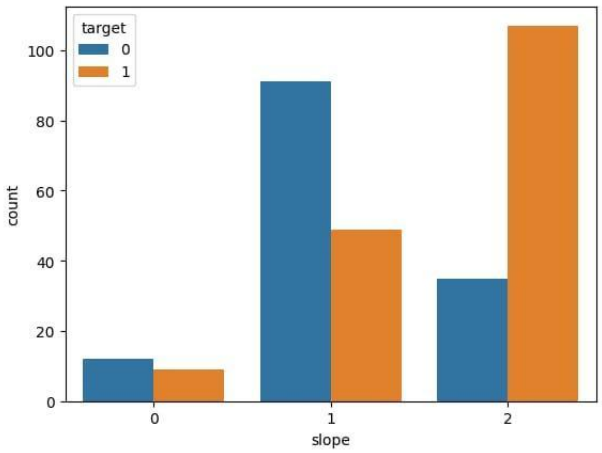



**Fig : 3.8 SLOPE**

The major vessels in the context of heart disease typically refer to the coronary arteries are the blood vessels that supply oxygen and nutrients to the heart muscle. Heart disease, specifically coronary artery disease (CAD), is characterized by the narrowing or blockage of these coronary arteries due to the buildup of fatty deposits called plaques. From Fig **3.7** we can observe that Coronary arteries will have a major impact on getting heart disease.

The slope of the peak of the exercise ST-segment, as observed on an electrocardiogram (ECG) during exercise stress testing, can provide valuable information about the likelihood of heart disease. The ST-segment is a portion of the ECG that reflects the electrical activity of the heart during a specific phase of the cardiac cycle. Changes in the ST-segment can indicate insufficient blood flow to the heart muscle, which may be indicative of coronary artery disease (CAD).From **Fig 3.8** We can also observe that the level of Slope of ST increases the chance of getting heart disease also increases. When compared to level 1 of ST level 3 of ST having more risk of heart disease.

## Data Analysis using Multivariate Analysis :

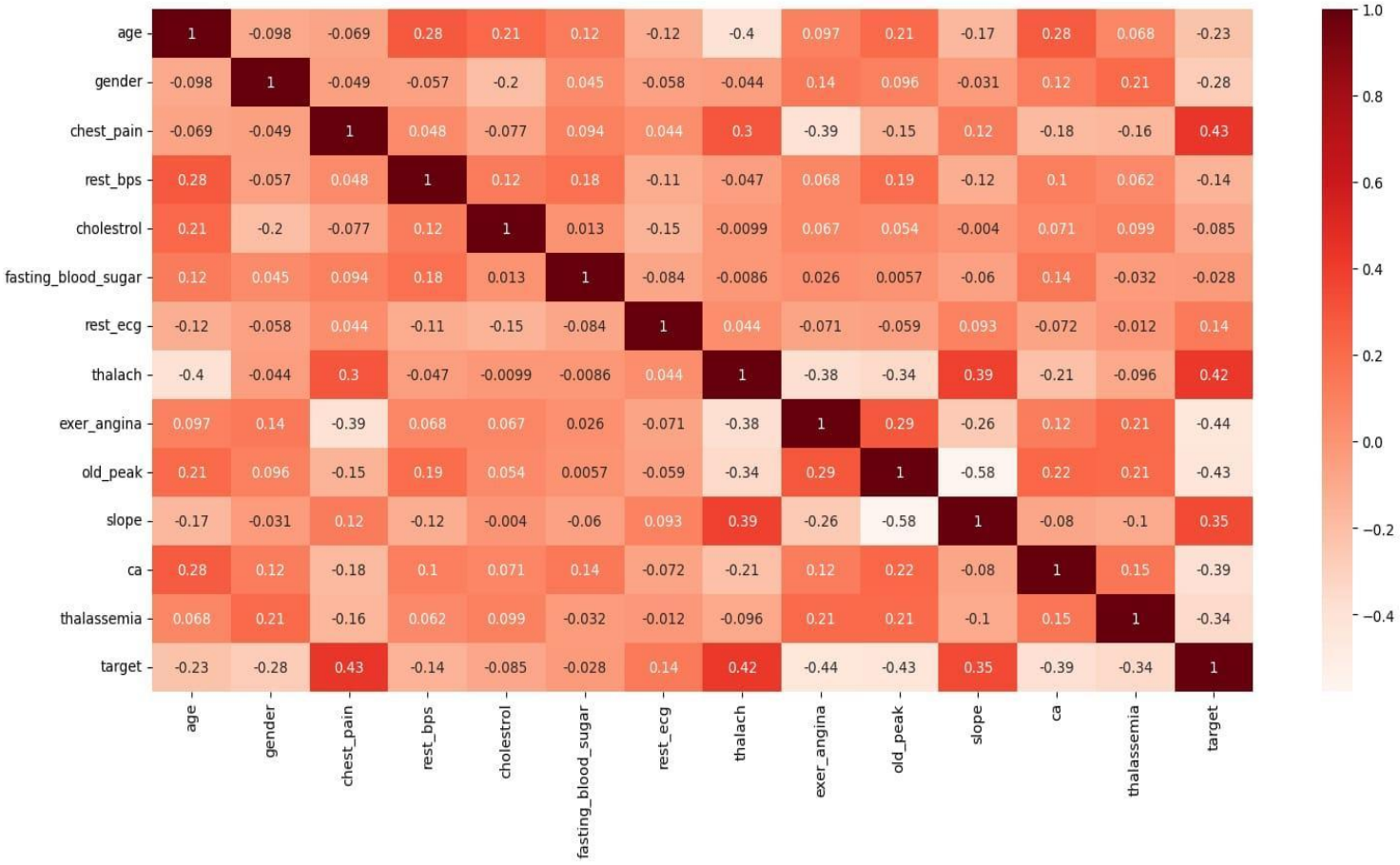



**Fig : 3.9 Correlation Between Variables**

## 4. Data Cleaning and Preprocessing :

Data preprocessing is a crucial step in the data analysis pipeline that involves transforming raw data into a clean and usable format suitable for analysis by machine learning algorithms This step involves removing irrelevant or duplicate data, dealing with missing values, and correcting any inconsistencies in the data.Data transformation step involves transforming the data into a format that is suitable for analysis by machine learning algorithms. This might include scaling or normalizing the data, converting categorical variables to numerical format, or reducing the dimensionality of the data using techniques such as Principal Component Analysis (PCA)

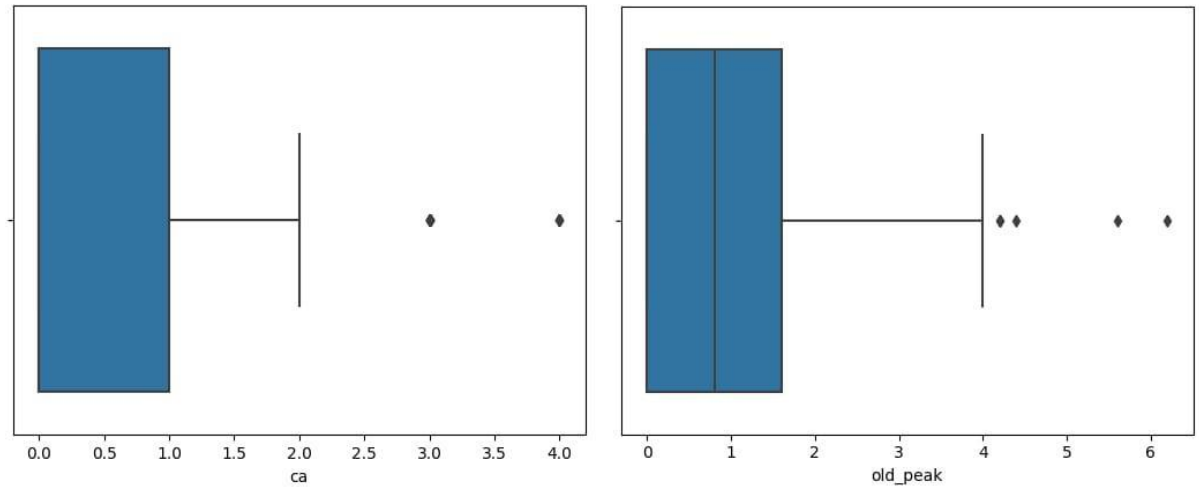
```
age                   0
gender                0
chest_pain            0
rest_bps              0
cholestrol            0
fasting_blood_sugar   0
rest_ecg              0
thalach               0
exer_angina           0
old_peak              0
slope                 0
ca                    0
thalassemia           0
target                0
dtype: int64
```

**Fig : 4.1 Checking For Missing Values**

From Fig **4.1** In this data set there are no null values found. So data cleaning and preprocessing is not needed in this case.

## Checking For Outliers :

In the process of data cleaning and preprocessing we are going to check for Outliers. Outliers are observations that deviate significantly from other observations in a dataset. They can arise due to measurement errors, data entry errors, or other factors, and can have a significant impact on the performance of machine learning models.



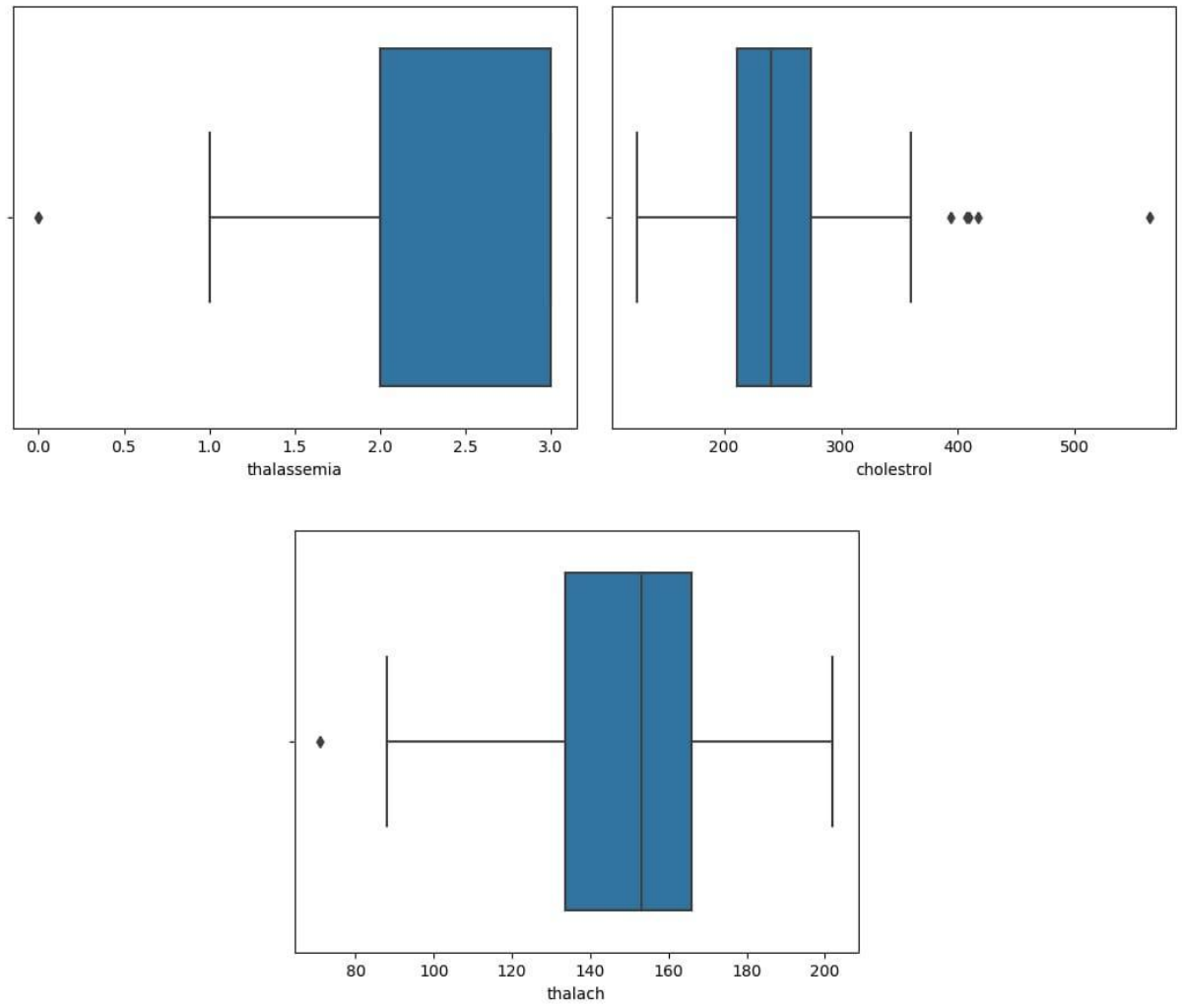



**Fig : 4.2 Checking For Outliers**

From Fig **4.2** we can observe that only the outliers are present in ca, thalassemia, cholesterol, thalach and old peak.

In this case we are going to build a model based on Decision tree Algorithm. Decision trees are more flexible and robust to outliers and noisy data, it can handle outliers and noisy data without affecting the overall performance of the model.

Decision trees are able to handle these types of data because they are based on a hierarchical structure that partitions the feature space into smaller and smaller subsets based on the values of the input features.

## 5.Model Building :

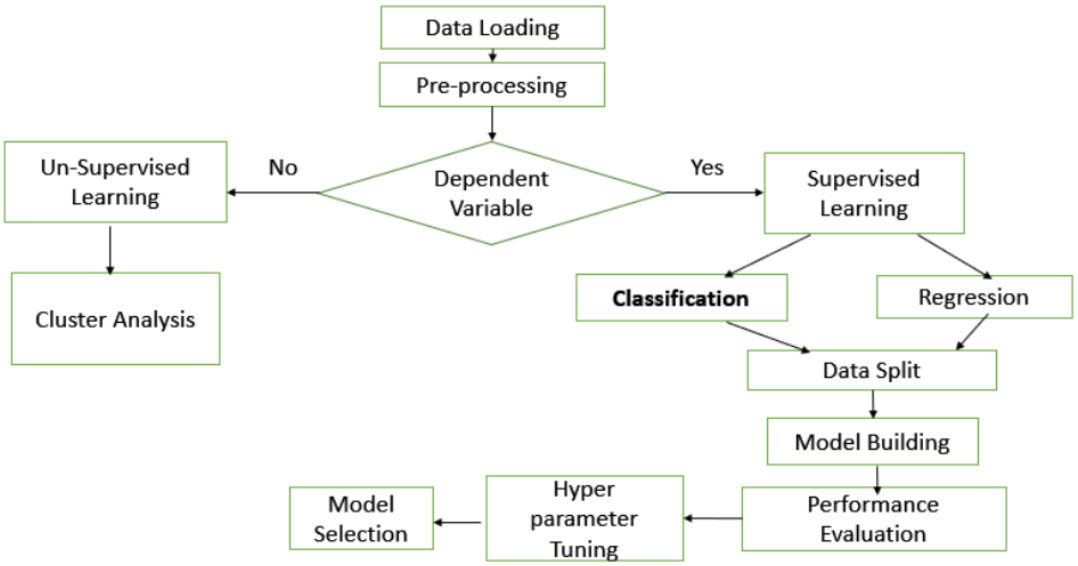



**Fig : 5.1 Process In Model Building**

The process of building a machine learning model involves collecting and cleaning data, exploring and analyzing it to select relevant features, choosing an appropriate algorithm, training the model, evaluating its performance on new data, and deploying it in a production environment.

The objective of the study is to identify the persons getting a chance for heart disease. We have a dependent variable which is the target. Binary class variable where 0 is not having heart disease and 1 is having heart disease. We will use a Supervised learning method to build classification models based on Decision Tree Algorithms. The data is split into trains and tests at a 70:30 ratio. Several classification-based algorithms from the Scikit learn package.

## Decision Tree :

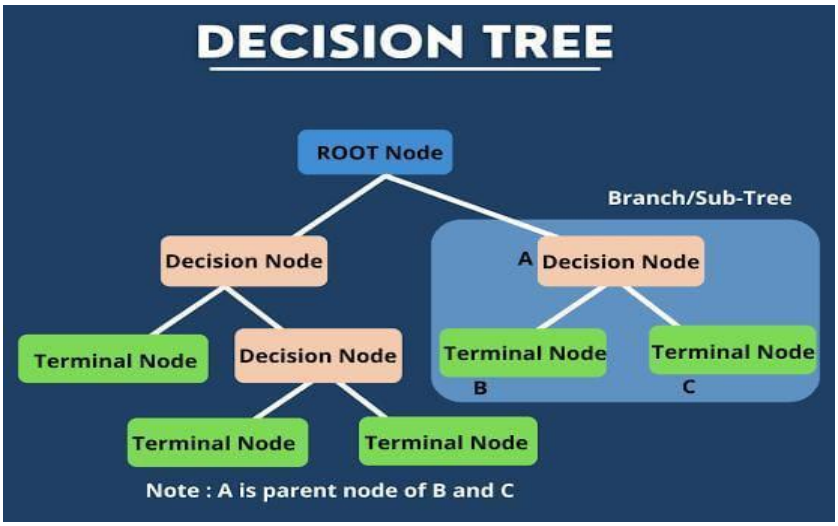



**Fig : 5.2 Decision Tree**

A decision tree model is a supervised machine learning algorithm that is commonly used for classification and regression tasks. The algorithm works by recursively partitioning the data into subsets based on the values of the features, until a stopping criterion is met. At each level of the tree, the algorithm chooses the feature that best separates the data, according to a certain criterion, such as information gain or Gini impurity.

Information gain is a measure of the reduction in entropy (or uncertainty) achieved by partitioning the data based on a particular feature. The entropy of a set of data measures the degree of disorder or unpredictability in the data.

Gini impurity, on the other hand, measures the probability of misclassifying a randomly chosen data point if it is labeled according to the class distribution in the subset.

The resulting tree structure can be visualized as a set of binary splits, where each internal node represents a feature, and each leaf node represents a class label or a continuous value. To make a prediction on a new data point, the algorithm follows the path from the root to a leaf node, based on the values of the features, and returns the label or value associated with that node.

## Model Evaluate :

### AUC And ROC Curve :

AUC (Area Under the Curve) is the area under the ROC curve, which ranges from 0 to 1. A higher AUC indicates better classifier performance.

ROC (Receiver Operating Characteristic) curve is a graphical representation of the performance of a binary classifier system as its discrimination threshold is varied. It is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.

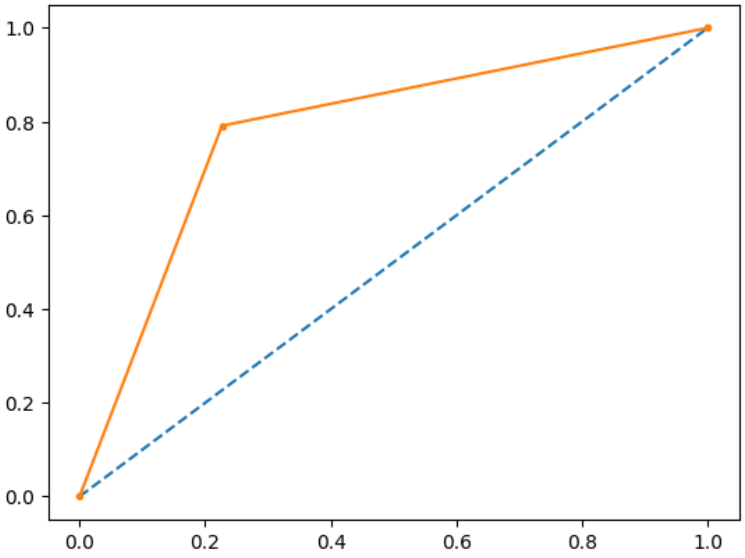



**Fig : 5.3 ROC On Train Data**

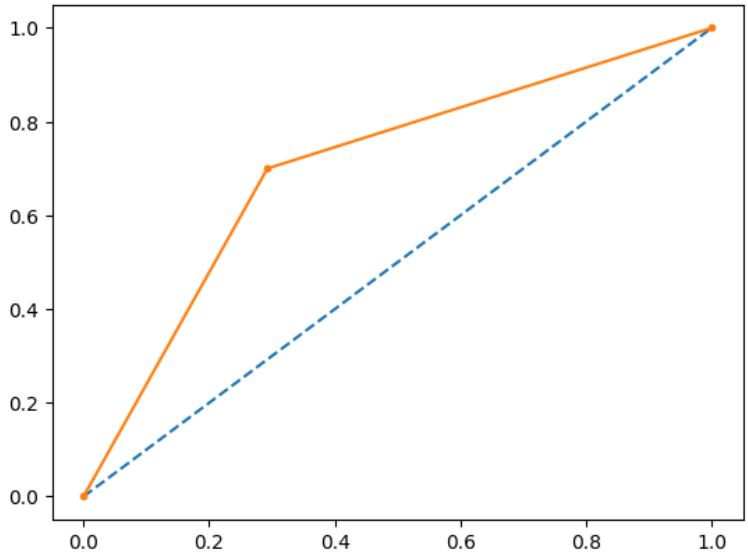



**Fig : 5.4 ROC On Test Data**

**AUC Of Train Data : 0.782**

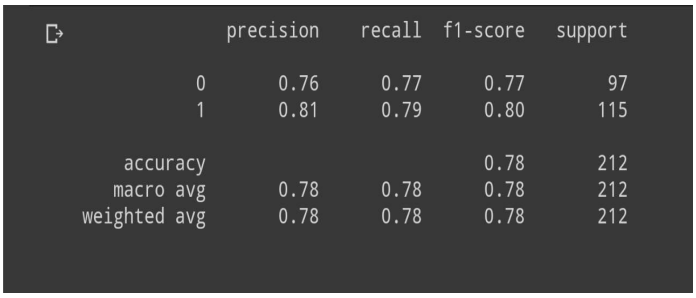
**AUC Of Test Data : 0.704**

The Area under the curve for the test is 70% which means the possibility of distinguishing between class-0 and class-1 is 70%. Receiver operating characteristic curve (ROC) is a graph between True positives and false negatives. From the graph it is visible that the curve is more inclined towards the y axis which are true positives.

## Confusion matrix:

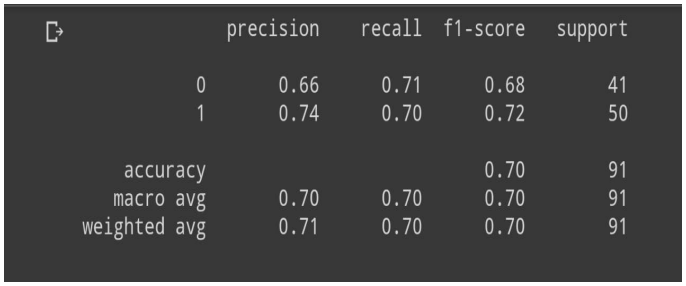
A confusion matrix is a table that is often used to evaluate the performance of a classification model by comparing the actual and predicted values of the target variable. The matrix is constructed by counting the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) produced by the model.

This graph shows the confusion matrix for the training data in the form of a heatmap, depicting the relationship between the actual and predicted training data.



| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.76 | 0.77 | 0.77 | 97 |
| 1 | 0.81 | 0.79 | 0.80 | 115 |
| accuracy | | | 0.78 | 212 |
| macro avg | 0.78 | 0.78 | 0.78 | 212 |
| weighted avg | 0.78 | 0.78 | 0.78 | 212 |

**Fig :5.5 Confusion Matrix of Train Data**



| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.66 | 0.71 | 0.68 | 41 |
| 1 | 0.74 | 0.70 | 0.72 | 50 |
| accuracy | | | 0.70 | 91 |
| macro avg | 0.70 | 0.70 | 0.70 | 91 |
| weighted avg | 0.71 | 0.70 | 0.70 | 91 |

**Fig : 5.6 Confusion Matrix Of Test Data**

The output shows the evaluation metrics for a binary classification model on both the training and testing datasets.

Comparing the metrics between the training and testing datasets, we can see that the performance on the testing dataset is slightly worse than on the training dataset, with a lower f1-score and recall for the positive class. This suggests that the model may be overfitting to the training data.

In conclusion, the model has an overall reasonable performance with a high accuracy, but it struggles to correctly identify instances belonging to the positive class. This may be an issue if the positive class is of particular interest and needs to be identified accurately. It is also important to note that the model may be overfitting to the training data, which may require further investigation and modification of the model.

# Grid Search CV

However, after applying GridSearchCV, the best set of hyper parameters that maximize the model's performance on the validation data are selected. Therefore, the performance of the model with the best hyper parameters may differ from the performance of the model trained on the original data.

In conclusion, while the original training data remains the same, GridSearchCV can identify the best hyperparameters that maximize the model's performance on the validation data, leading to better performance on new data.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.81 | 0.74 | 0.77 | 97 |
| 1 | 0.80 | 0.85 | 0.82 | 115 |
| accuracy | | | 0.80 | 212 |
| macro avg | 0.80 | 0.80 | 0.80 | 212 |
| weighted avg | 0.80 | 0.80 | 0.80 | 212 |

**Fig : 5.7 GridSearchCV On Train Data**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.72 | 0.80 | 0.76 | 41 |
| 1 | 0.82 | 0.74 | 0.78 | 50 |
| accuracy | | | 0.77 | 91 |
| macro avg | 0.77 | 0.77 | 0.77 | 91 |
| weighted avg | 0.77 | 0.77 | 0.77 | 91 |

**Fig : 5.8 GridSearchCV On Test Data**

After GridSearchCV accuracy of the model on test data had increased to 77% and accuracy on train data in 80% . This indicates the model has a better ability to generalize to new data.

Accuracy, AUC, Precision and Recall for test data is almost inline with training data. This proves no overfitting or underfitting has happened, and overall the model is a good model for classification.

## Conclusion :

Based on a dataset with multiple clinical variables, a decision tree model was created to forecast the presence of heart disease. To avoid overfitting, the model was trained on a portion of the data and then verified using a different validation set. The model was then evaluated for generalizability using data from a separate testing group.

The model accuracy on test data increase following regularization, with training accuracy of 80% and testing accuracy of 77%. This suggests that the model can generalize fresh, untested data more effectively.

In general, the decision tree model shows potential in detecting the existence of cardiac disease based on the supplied dataset. Additional features, a larger sample size, and the investigation of other machine learning algorithms can all lead to further advancements.

## General Suggestion To Free From Heart disease :

Eat Healthy Foods

Exercise regularly

Maintain a healthy weight

Don't smoke Don't drink alcohol

Manage stress

Control your blood pressure and cholesterol

Control Blood Sugar Level