# Introduction:

The CDC reports that heart disease is a prominent cause of death in the United States across various racial groups, including African Americans, American Indians and Alaska Natives, and white individuals. Nearly half of all Americans, accounting for 47%, possess at least one of three primary risk factors for heart disease: elevated blood pressure, high cholesterol levels, and tobacco usage. Additional significant indicators encompass diabetes status, obesity indicated by a high body mass index (BMI), insufficient physical activity, and excessive alcohol consumption. Identifying and preventing these factors, which exert a substantial influence on heart disease, is of paramount importance in the realm of healthcare.

In this context, the advancement of computational techniques has opened the door to the utilization of machine learning methodologies for discerning patterns within data. These patterns can be harnessed to make predictions regarding a patient's health condition, thereby aiding in the early detection and prevention of heart disease.

# Objective of this Project:

In this project we have tried to assess the likelihood of having heart disease by an individual based on 13 parameters which are:

1. Age: Numerical data
2. Sex: Categorical Variable (Male or Female)
3. Chest Pain Type: Categorical data (Typical angina, Atypical angina, Non-anginal Pain, Asymptomatic)
4. Blood pressure: Numerical data
5. Cholesterol: Numerical data
6. Fasting Blood Sugar: Categorical data (Lower than 120 mg/l or higher than that)
7. ECG result: Categorical data (ST-T wave abnormality, Normal, Left ventricular hypertrophy
8. Max Heart Rate: Numerical data
9. Exercised Induced Angina: Categorical data (Yes or No)
10. Old peak: ST depression induced by exercise relative to rest
11. Slope: The slope of peak exercise ST segment
12. Vessel colored by Fluroscopy: Categorical data (No of major vessels colored by Fluroscopy 0-3)
13. Thalassemia: Categorical data (Reversable Defect, Fixed Defect, Normal, No)

Depending upon the data collected on these parameters we have tried a form a model which can predict whether that individual is likely to have heart disease or not.

We have used python (version 3.10.10) and have tried to predict using 4 machine learning techniques which are Logistic Regression, K-nearest neighbor classifier, Random Forest classifier, and Support Vector Machine Classifier.

Exploratory data analysis has been done prior to prediction I order to understand the relevance of several parameters and their effect on the target parameter. During prediction hyperparametric tuning has also been done for several models in order to increase the accuracy of prediction.

# Motivation for the project:

The core motivation behind this project is to combat the pervasive issue of heart disease, a leading cause of death in the U.S., impacting various demographic groups. The project's primary objectives are:

1. **Early Detection:** Identifying heart disease risk early to enable timely intervention and prevention.
2. **Personalized Care**: Tailoring healthcare recommendations based on an individual's unique risk profile for more effective interventions.
3. **Resource Efficiency:** Optimizing healthcare resource allocation to ensure those at highest risk receive necessary care.
4. **Public Health Impact**: Reducing the overall societal and economic burden of heart disease by enhancing healthcare outcomes and quality of life.

This project aims to improve public health, save lives, and alleviate the strain on healthcare systems by utilizing machine learning to predict and prevent heart disease.

# Exploratory Data Analysis:

* **Finding Missing values for each parameter:**

No missing values have been found for any parameters

* **Finding of categorical variable and their cardinality:**

Following results have been obtained-

* sex has 2 nos of distinct values
* chest\_pain\_type has 4 nos of distinct values
* fasting\_blood\_sugar has 2 nos of distinct values
* rest\_ecg has 3 nos of distinct values
* exercise\_induced\_angina has 2 nos of distinct values
* slope has 3 nos of distinct values
* vessels\_colored\_by\_flourosopy has 5 nos of distinct values
* thalassemia has 4 nos of distinct values

From this we can conclude that As many categorical variables are present so we have to one-hot encode those variables. We have done that subsequently. Here we might note that while doing one-hot encoding we have dropped one of the categories in order to avoid dummy trap, also we have not label encode the data as we did not want to give higher weightage to a particular category.

* **Outlier Identification for Numerical Variables:**

As we know outlier presents within the dataset can give erroneous results so we have eliminated those prior to the analysis. We have used both visual and numerical technique in order to identify the outliers. Numerically, if the value greater than **Q3 + 1.5\*IQR** or lesser than **Q1 – 1.5\* IQR** (where Q3, Q1, IQR represents 3rd quartile, 1st quartile and inter-quartile range respectively) and for visual representation we have used Box-Whisker plot. Both results show that for the parameters resting blood pressure and cholesterol we have got outliers. But also as outlying values for both the parameters do not occur simultaneously we have not omitted any values here. Results have been shown below:

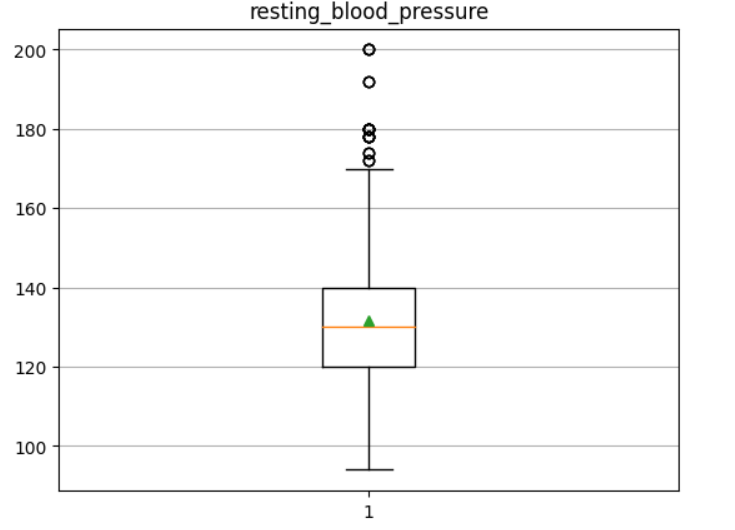
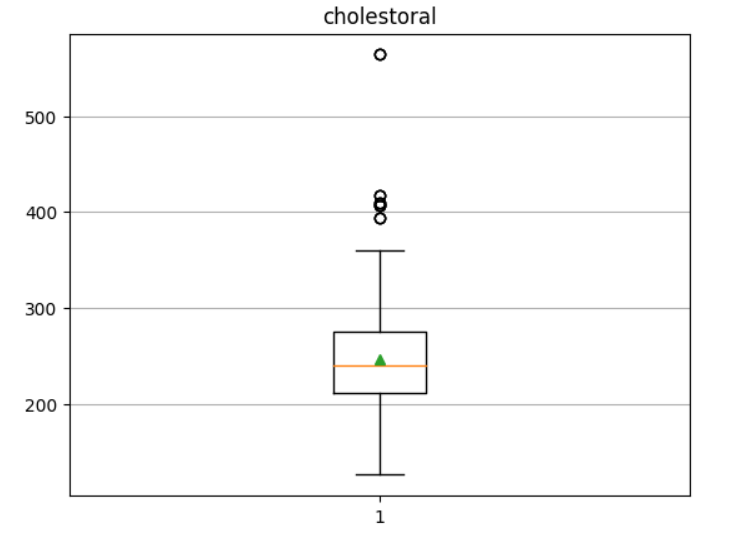
 

Fig 1: Box-Whisker plot showing the distribution of values for resting blood pressure and cholesterol

Numerically the data obtained for these two parameters are the upper limit of resting\_blood\_pressure is 170.0 mmHg and lower limit of resting\_blood\_pressure is 90.0 mmHg the upper limit of cholestoral is 371.0 and lower limit of cholestoral is 115.0

* **Distribution Of Patients Having Heart Disease And Not Having Heart Disease:**

The following graph shows the distribution:

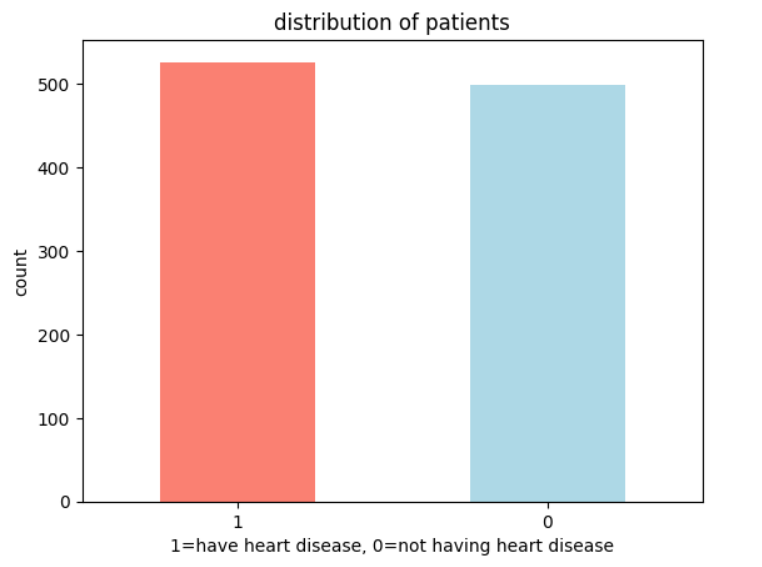


Fig2: Distribution of patients

Here we can see the no of patients having heart disease and not having heart disease are almost same. Hence, we can conclude that our data is balanced hence no preprocessing technique like up sampling, down sampling or threshold manipulation is not strictly required.

* **Frequency Of Heart Disease As Per Sex:**

Following graph shows the distribution of heart disease as per sex:

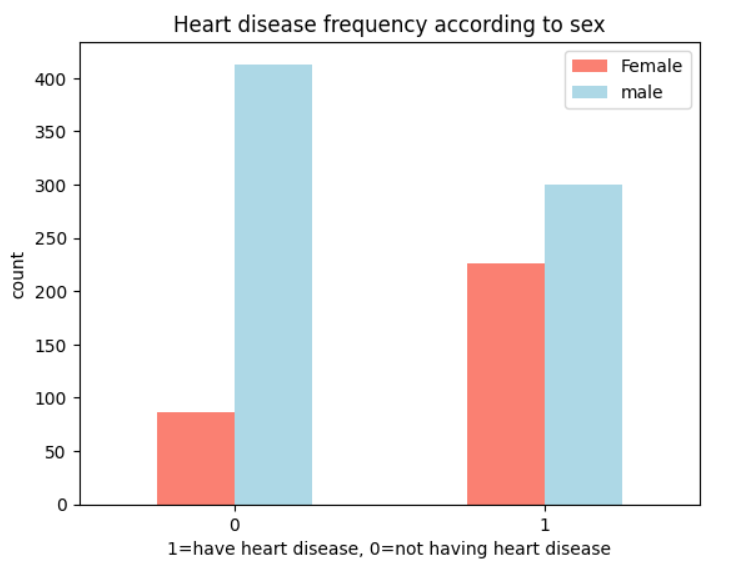
****

Fig3: Distribution Of Heart Disease As Per Sex

Here from the data we have found that about 72% of the female under our study have heart disease and 42% of male have heart disease. This data shows us a considerable higher risk of heart disease in female patients.

* **Max Heart Rate With Respect To Age:**

Following graph shows the distribution of maximum heart rate with respect to age and sex:

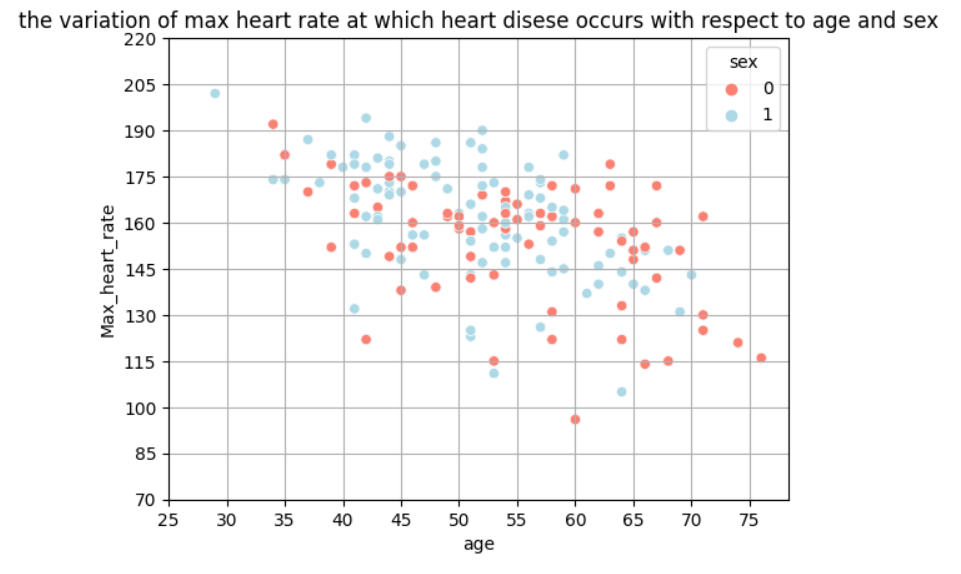


Fig4: Distribution Of Maximum Heart Rate With Respect To Age And Sex

Here we can see that there is a downward trend of max heart rate with respect to age which is quite expected.

* **Comparison Between Male And Female Based On Blood Sugar Vs Cholesterol:**

Following graph shows the comparison between male and female based on blood sugar vs cholesterol:

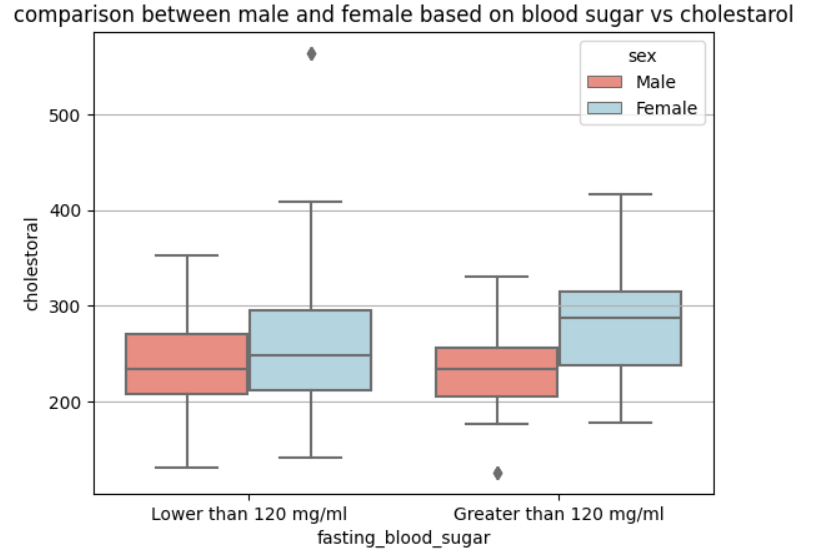
****

Fig5: Comparison Between Male And Female Based On Blood Sugar Vs Cholesterol

Here we can notice a significant variation in blood cholesterol level in female based on their fasting blood sugar which is not very significant for male.

* **The Variation Of Cholesterol At Which Heart Disease Occurs With Respect To Age And Sex:**

Following graph shows the variation of cholesterol at which heart disease occurs with respect to age and sex:

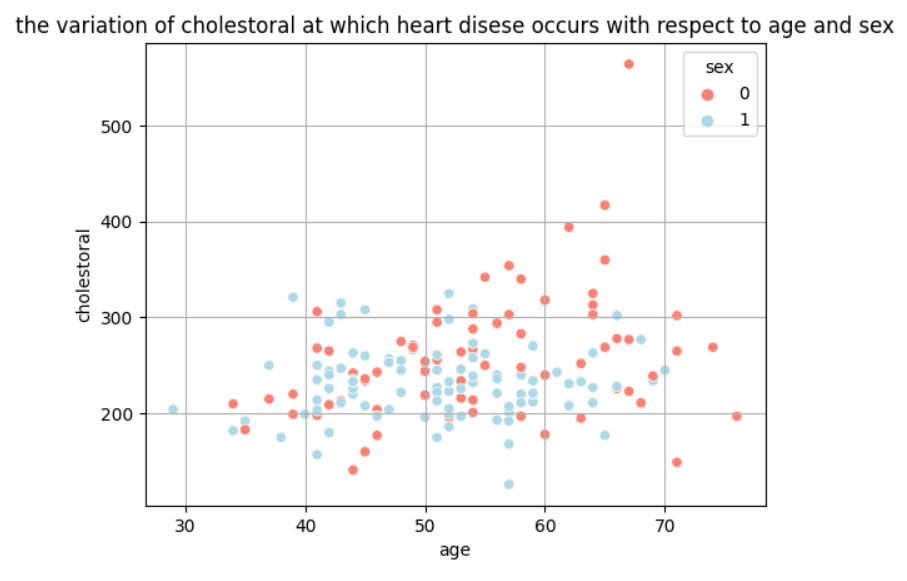
****

Fig6: Variation Of Cholesterol At Which Heart Disease Occurs With Respect To Age And Sex

Here we can see there is not any specific correlation between age and cholesterol level causing heart disease. Cholesterol is a prime factor for causing heart disease but its independent of age. This observation is quite significant as it indicates impact of juvenile cholesterol disease in US.

* **Distribution Of Age of patients Having Heart Diseases:**

Following graph shows distribution of age of patients having heart diseases:

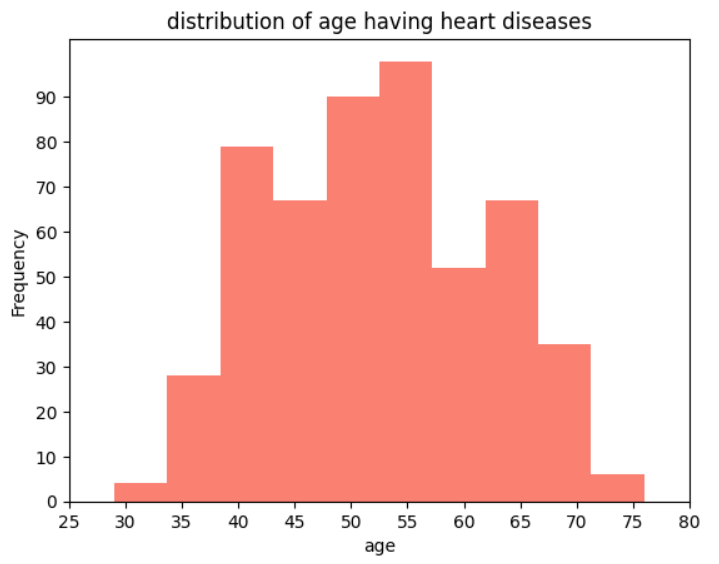
****

Fig7: Distribution Of Age Of Patients Having Heart Diseases

We can see within the age range of 45 to 60 the heart disease frequency is most.

* **Anginal Pain Vs Heart Disease:**

It’s a common misconception that anginal pain is related to heart disease. Following graph shows the distribution of anginal pain vs heart disease:

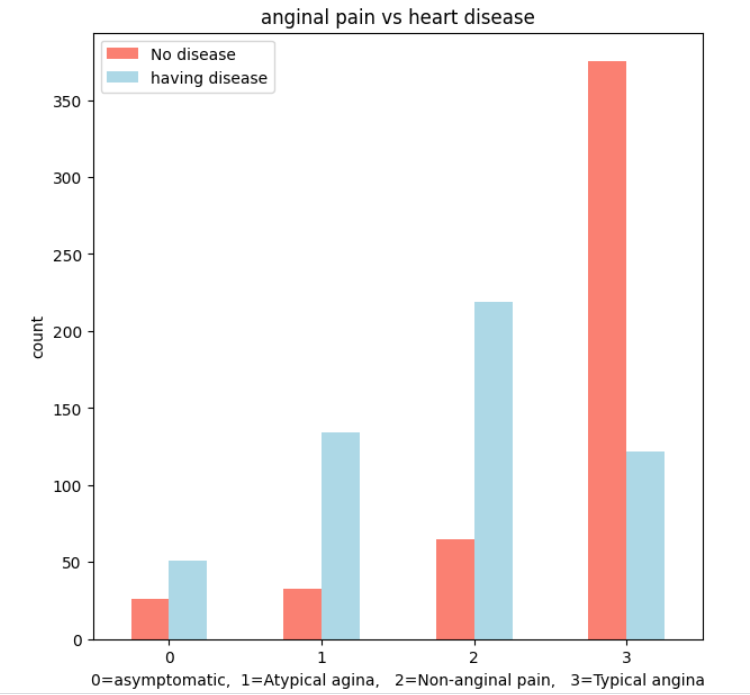


Fig8: Distribution Of Anginal Pain Vs Heart Disease

Here we can see that primary cause of heart disease is not anginal pain, even the patient which are not having any anginal pain are more frequent to have heart disease. This observation hence is quite significant.

* **Distribution Of Blood Sugar Level Between Male And Female As Per Age For Patients Having Heart Disease:**

Following graph shows the variation of blood sugar level between male and female as per age for patients having heart disease:

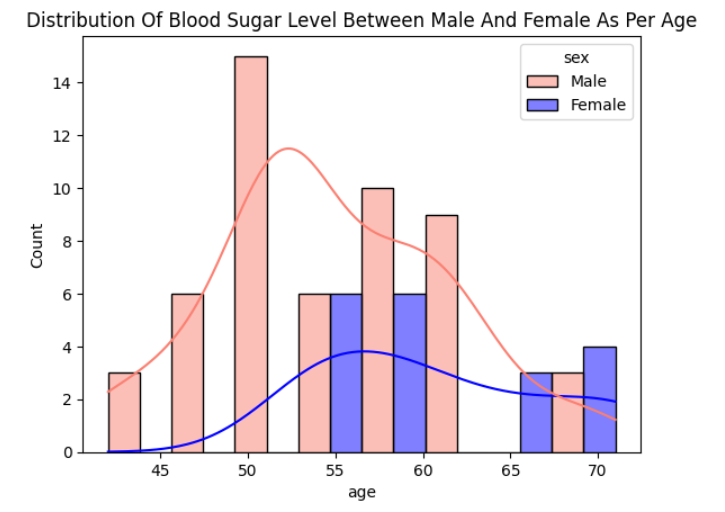


Fig9: Variation Of Blood Sugar Level Between Male And Female As Per Age

Here we can see that between age group of 40 to 50, blood sugar in male is significantly higher, after that it becomes gradually equal. Also, juvenile diabetes is rare which is quite understandable.

* **Relation between blood sugar and heart disease with respect to sex:**

Following graph shows the variation of relation between blood sugar and heart disease with respect to sex:

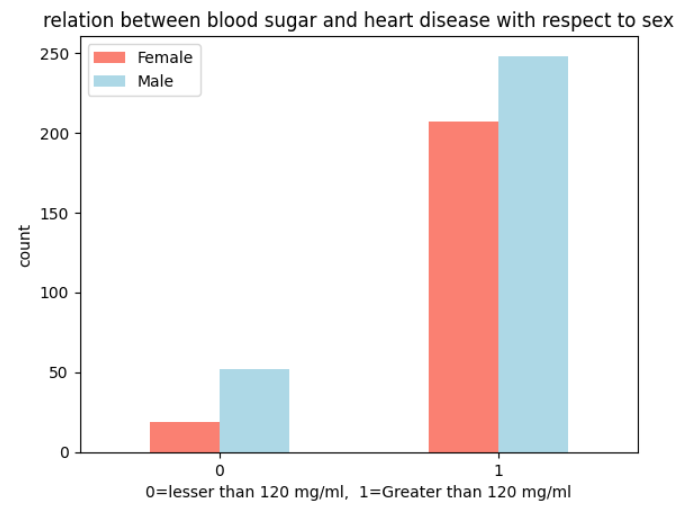


Fig10: The Variation Of Relation Between Blood Sugar And Heart Disease With Respect To Sex

Here we can see that irrespective of gender high blood sugar can cause higher rate of heart disease. This indicates proper awareness should be taken to control diabetes.

* **Correlation Plot:**

Following figure shows the correlation between different input parameters and target:

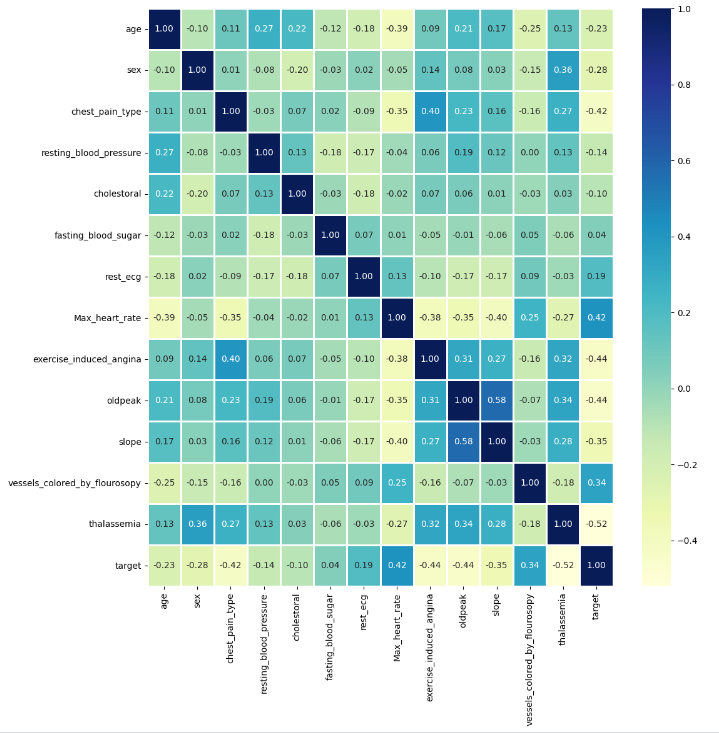


Fig11: correlation between different input parameters and target

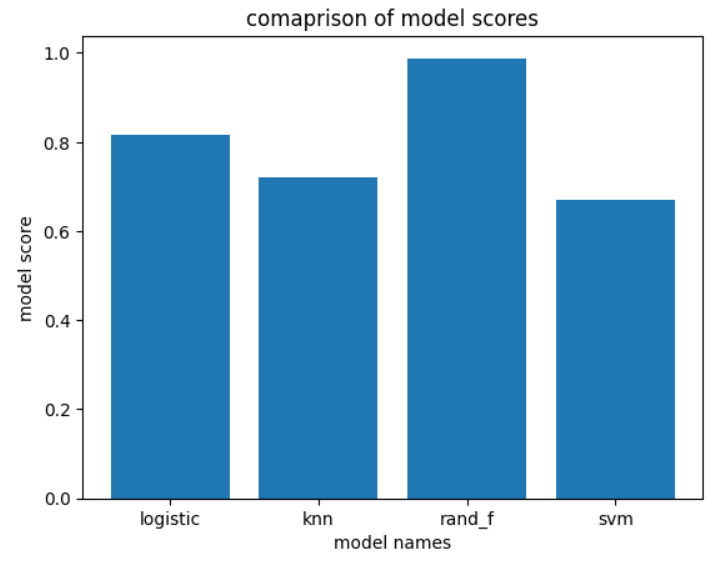
This plot shows that the impact of individual parameters on target is not quite significant but as a whole they can create some impact which has been assessed in the prediction and modelling part of this project. Also here we can see multicollinearity is not that significant.

# Prediction and Modelling:

As discussed earlier we have used 4 machine learning techniques which are Logistic Regression, K-nearest neighbor classifier, Random Forest classifier, and Support Vector Machine Classifier.

* **Initial Modelling:**

At first we have tried to apply these models with default parameters comes with ScikitLearn Module. The dataset has been divided into two part, 75% data has been used for training and 25% data has been used for testing. From this initially accuracy scores have been obtained for each of the models which has been shown in the following graph. Exact scores has also been shown below.



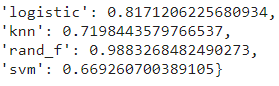




Fig12: Accuracy Scores Obtained From Different Models Using Default Values

from initial investigation we can see the accuracy of the random forest model is highest.

* **Hyperparametric Tuning:**

For the purpose of hyperparametric tuning, GridSearchCV and Randomized SearchCV from ScikitLearn module has been used. It basically takes a range of parameter values for each machine learning model and evaluate scores either for all combination or for a random combination of parameters and gives us the best parameters for maximizing accuracy of prediction. These two techniques have been used for random forest, logistic regression and svm classifier where we have used iterative technique for knn classifier for which we have selected several k values and calculated the accuracy of model. The value of k after which the model accuracy is fairly high can be chosen as the number of neighbours (based on which prediction to be evaluated).

* *Results:*
* Logistic Regression:

C=0.08858667904100823, solver='liblinear' and new accuracy score is 0.82 Which indicates not a significant change in score due to the hyperparametric tuning.

* Random Forest Classifier:

n\_estimators=810, min\_samples\_split=10, min\_samples\_leaf=3 and new accuracy score is 0.93 Which indicates rather to go with default parametrs (Here score has been decreased as RandomizedsearchCV does not look for all probable combination of parameter values)

* SVM classifier:

C= 10000.0,kernel='rbf' and new accuracy score is 0.82 Which indicates a significant change in score due to the hyperparametric tuning.

* KNN classifier:

No of neighbours = 4 and the updated accuracy score is 0.79 Which indicates a significant change in score due to the hyperparametric tuning. (As no of neighbours decreases the accuracy of prediction increases but it leads to overfitting of the model)

* **Prediction Results:**
* *Confusion Matrix:*

Following figure shows the confusion matrix of the results obtained from updated random forest model:

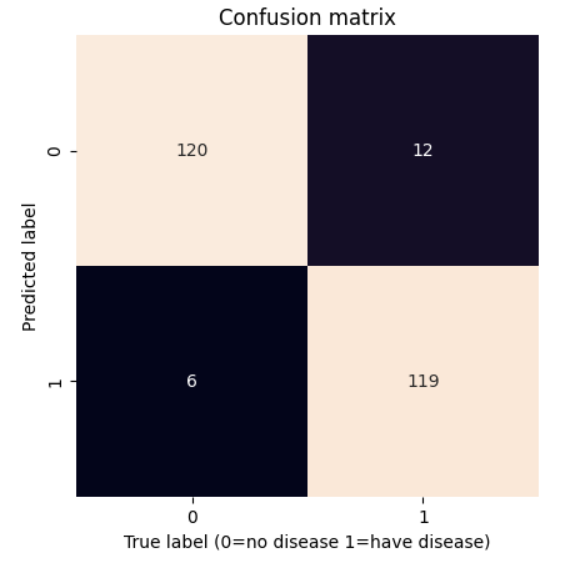


Fig 13: Confusion Matrix

We can see only 18 cases our model fails out of 257 cases, which is quite good !

* *Classification report:*

The classification report for our prediction has been shown below:

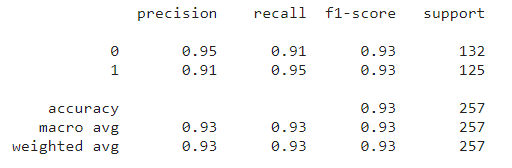


Fig 13: Classification report

Here we can see that both precision and recall for our prediction is quite good and f1 score is also promising. Also it can be noted that these values are for one set of train and test set. Values can be different for different sets. Hence we have shown below the cross validated average score for recall, precession, accuracy and f1 score (cv=5)

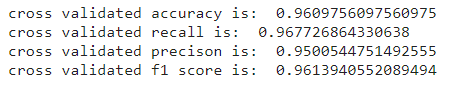


Fig 14: Classification report

* **Feature Engineering:**

We have performed feature engineering for random forest model in order to see whether further medication of the model can be done or not. For this, feature\_importance\_ function has been used which gives the importance of features which basically calculated based on gini impurity value for different parameters which is averaged over several decision trees. Following feagure shows the feature importance values for all parameters:

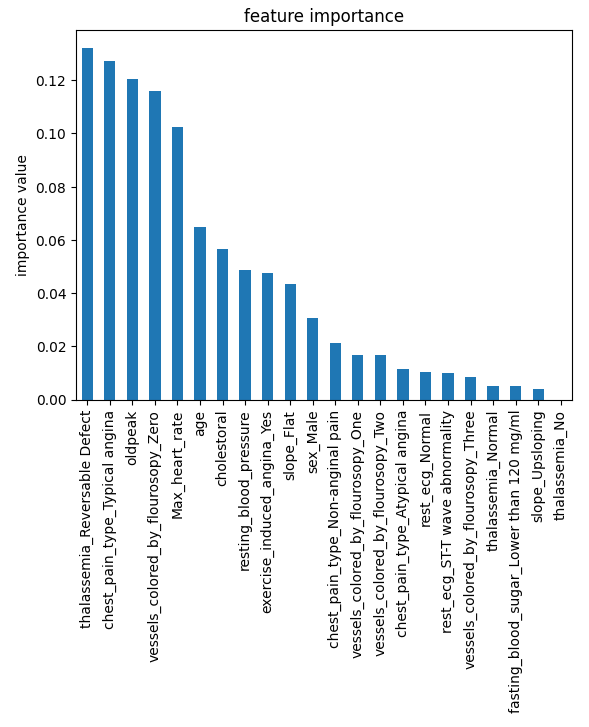


Fig 15: Feature importance values of parameters in sorted manner

So if we take 0.02 as the threshold value for feature importance, then we will end up with 12 parameters (one-hot encoded) shown below:

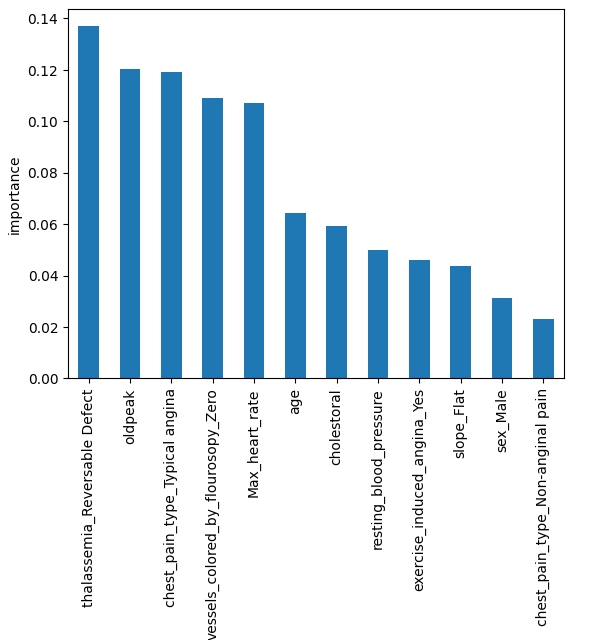


Fig 16: Important features after feature elimination

Based on this updated data we have again run our model which has now given an accuracy of 0.94 which is slightly better than the accuracy obtained initially (0.93). Updated confusion matrix has also been shown below:

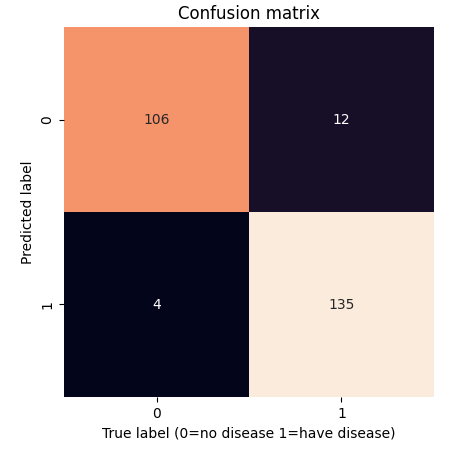


Fig 17: Updated confusion matrix

* **Impact of Feature Scaling:**

Here we have assessed whether scaling the input data can make an impact on the output or not. We have use Standardization as our scaling technique which would make mean of each parameter to be 0 and standard deviation to be 1. And after this the results obtained from each of the model has been become more accurate !

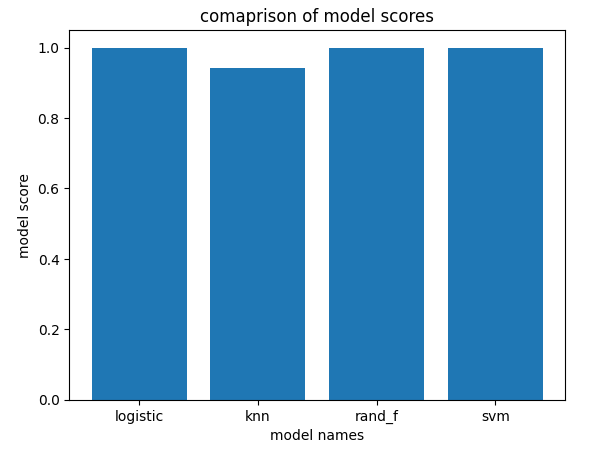


Fig 18: Updated scores after scaling the data

# Conclusion

This project highlights the power of computational techniques in predicting and preventing heart disease. By leveraging machine learning and data analysis, we can improve public health outcomes and save lives. Early detection and personalized care are key in combating heart disease, the leading cause of death in the United States.