**1: Proposed Methodology**:

The methodology used in this project aims to predict the heart disease in the patients based on their test results by using the different machine learning models.

The complete life cycle of the project is given below

The details of each step are given below

* 1. **: Data Acquisition**

The dataset was sourced from the Kaggle and it was collected in different countries and merged with Stalog heart dataset to make it a single dataset. The final dataset was consisted of 1023 samples with the 12 features. The details of samples taken from different countries are given below

* Cleveland: 303 observations
* Hungarian: 294 observations
* Switzerland: 123 observations
* Long Beach VA: 200 observations
* Stalog (Heart) Data Set: 102 observations

**1.2: Data- Preprocessing**

After obtaining the dataset, it goes through the step of pre-processing. The following operations were performed

* Addressing the missing values
* Looking for duplicated values
* Checking for any outliers
* Normalize the numerical data
* Apply One-Hot encoding technique on Categorical Features
* Understand relationship between variables

**1.3: Exploratory Data Analysis**

In this step, statistical techniques were used to calculate the key information from data. These techniques included mean, variance, different quantiles values etc. Moreover, in this step bar charts, box-plot, line graph and histogram were used to visualize the information from the data which helps to interpret the results.

**1.4: Feature Selection**

The dataset was consisted of 11 features with the 1 target variables. By performing the EDA analysis, the best features were obtained from the dataset for machine learning modelling.

**1.5: Train Test Split**

The dataset was split into train and test set where train set was used for the training of machine learning models and test set was used for the testing purpose. 80 percent of data was used for the training the models and 20 percent was used for the testing purpose.

**1.6: Model Implementation**

Different models of machine learning were used for prediction of heart disease. Each model has its own characteristics that is why different models were tried to check best suitable model. The list is as fellows

* Decision Tree Classifier

It is a tree base model which is used for the classification problem. It selects the best feature as the root node and split the dataset based on that. The criteria to choose the best feature is gini impurity or entropy of the features. The feature which has least entropy is selected for the root and dataset is split into two parts. To conclude, by this process dataset keep splitting until no feature is left. After that testing is performed.

* Random Forest Classifier

It works on the same principle of decision tree but here instead of on one tree multiple trees are used and random samples are used on each tree for training. It is best suited when number of features are large.

* Logistic Regression

It is widely used binary classification model which classify the data into two classes. It used the sigmoid function as the activation function which give the output in range of 0 and 1. By setting the threshold of 0.5 it classified the data into class 0 and class 1.

* Support vector machine also used for binary classification. It used the hyperplane to classify the data into two classes. It used the term of positive gutter and negative gutter which are two closet points of both classes with the hyperplane. Based on calculated score it assigned the class to new datapoint.
* KNN is also used for the classification problem. It works by the measuring the distance of datapoints with their class and datapoints which have least distance with respect to their class, it belongs to that. It is not computational expensive in training.
* Gaussian Navie Bayes is used for classification when dataset have numerical features. It assumed that features follow the normal distribution and they are independent of each other. Therefore, it calculates the prior probabilities of each class and for each datapoint it calculates its probability by using the normal distribution and multiply the all probabilities.

At the end, it assigned the datapoint to that class which has the higher probability.

* Multi-layer perceptron can be used for classification by using single neuron at the output layer with the sigmoid function. Its architect consisted of 11 input neurons with the activation function of relu followed by the hidden layer which consist of 44 neurons with relu activation function and at output layer single neuron is used because it is the binary classification problem. Relu activation function is used because it works well for numerical data and at the output layer, sigmoid function is used which is best suited for binary classification.

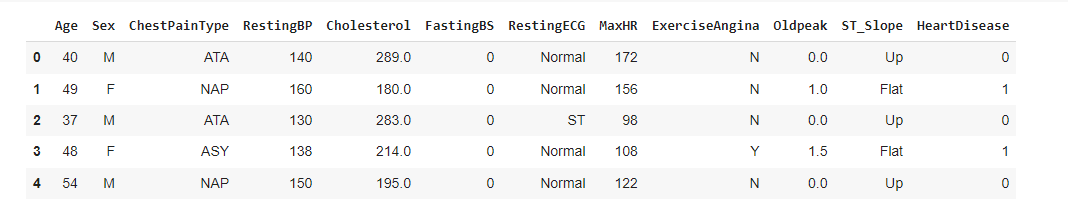
**1.7: Evaluation of model**

The models were evaluated based on their precision, accuracy recall and f score. ROC curve and confusion matrix was also used to visualize the performance of the models

* Precision is correct number of positive predictions with respect to total positive predictions
* Accuracy is total number of correct predictions with respect to all predictions.
* Recall is the total number of correct positive prediction with respect to actual postive values.
* F score is ratio between the precision and recall.
* Confusion matrix helps to visualize the predictions.

**2: Dataset Details:**

The dataset consisted of 1023 samples and it has 11 features with the 1 target variable.

****

The description of each feature is given below

* **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]

TA: Typical angina refers to chest pain or discomfort that has characteristic features typically associated with coronary artery disease.

ATA: Atypical angina refers to chest discomfort that doesn't completely fit the typical pattern of classic angina but is still suggestive of potential heart-related issues.

NAP: Non-anginal pain refers to chest discomfort or pain that is not related to the heart or coronary artery diseases.

ASY: Asymptomatic refers to the absence of any symptoms or noticeable signs.

* **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 wave abnormality

LVH: LVH refers to the thickening of the muscular wall of the heart's left ventricle in response to increased workload or other conditions.

* **MaxHR**: maximum heart rate achieved [Numeric value between 60 and 202]
* **ExerciseAngina**: exercise-induced angina [Y: Yes, N: No]
* **Oldpeak**: Numeric value of ST Slope measured in depression
* **ST\_Slope**:

UP: This is characterized by an upward shift of the ST segment above the baseline. It can indicate conditions such as myocardial infarction (heart attack), pericarditis, or myocarditis.

Down: When the ST segment shifts downward from the baseline, it may suggest myocardial ischemia (inadequate blood supply to the heart)

Flat: Sometimes, the ST segment might appear flat without a distinct elevation or depression. This can be seen in conditions like stable angina, bundle branch blocks, or medications affecting cardiac repolarization

* **HeartDisease**: output class [1: heart disease, 0: Normal]

**3: Results**

**3.1: Target Class**

The dataset consists of 11 feature and 1 target variable. The target variable has the two classes where class 1 belongs to heart patients and class 0 belongs to patients who has no heart disease.

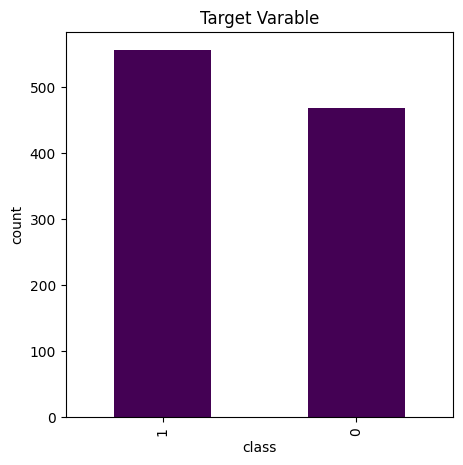
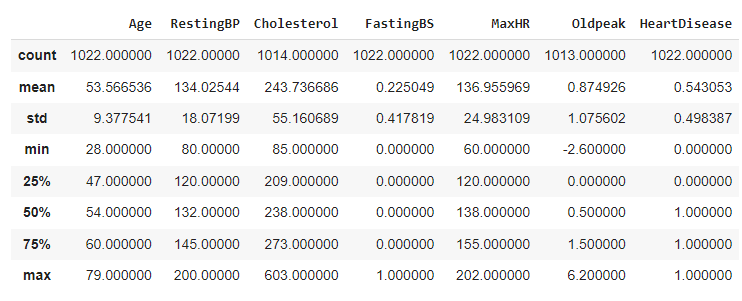


Fig 3.1: Target class

There are 555 patients that are belongs to class 1 and 467 patients that belongs to class 0 as shown in fig 3.1.

**3.2: Exploratory Data Analysis:**

The statistical analysis of the data is given below



The key information that was observed from the EDA are given below

* Age vs Heart Patients

Age factor is closely related with the heart disease. It is noticed from the data that most of heart patients are in age group of 50 to 70 while those patients who has not heart disease are in age group of 40 to 58 as shown in fig 3.4



Fig 3.4: Age vs heart disease

* Resting Blood Pressure vs heart disease

Resting blood pressure is higher in heart disease patients as compared to opposite class.

Most of the heart patients have the resting blood pressure of 110 to 165 and some of them have even higher resting blood pressure which is above 200. In contrast, patients with no heart disease have the normal resting blood pressure with the range of 105 to 145 as shown in fig 3.5

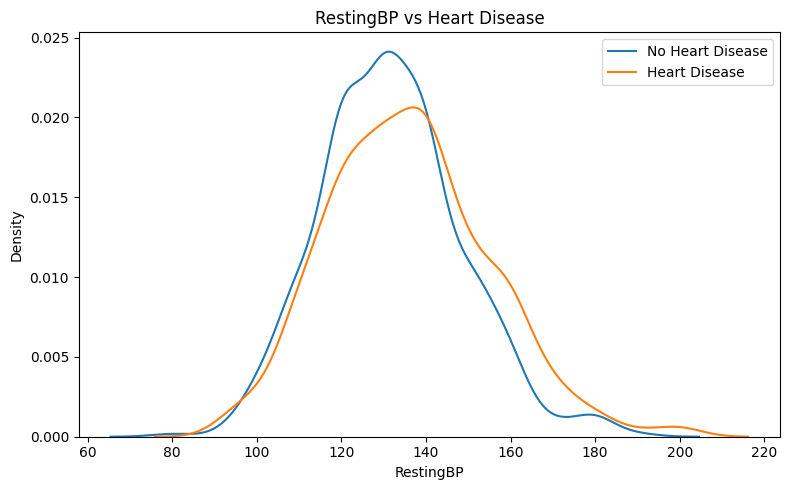


Fig 3.6: RestingBP vs heart disease

* Cholesterol vs Heart Patients

Similarly, heart patients have the higher cholesterol level with the range of 200 to 350 and patients with no heart disease have the cholesterol level in range of 150 to 300 but interesting finding is that most of patients who have not heart disease have the higher cholesterol level which is above 400 and reach at the max value of 600 and same is true for the heart patients as shown in fig 3.7

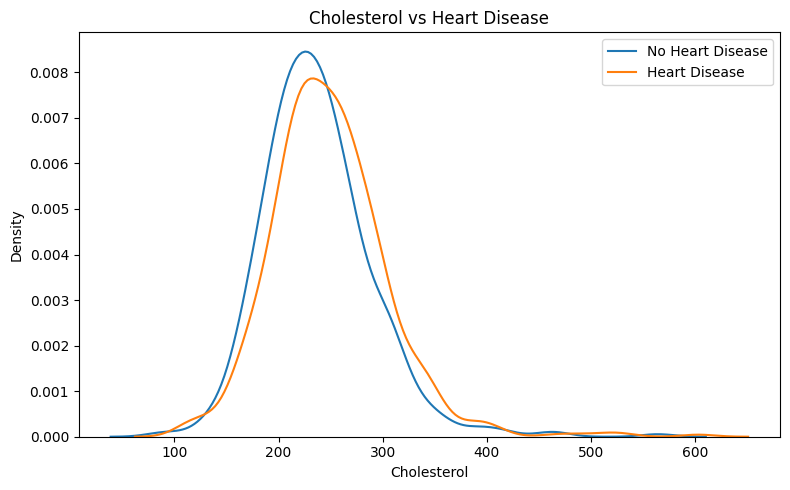


Fig 3.7: Cholesterol vs Heart Patients

* Maximum heart rate vs heart patients

It is noticed that the patients who have no heart disease have the maximum heart rate in range of 125 to 175 as compared to heart patients who have maximum heart rate in range of 100 to 150. It means that healthy patients have the higher maximum heart rate as compare to other as shown in fig 3.8

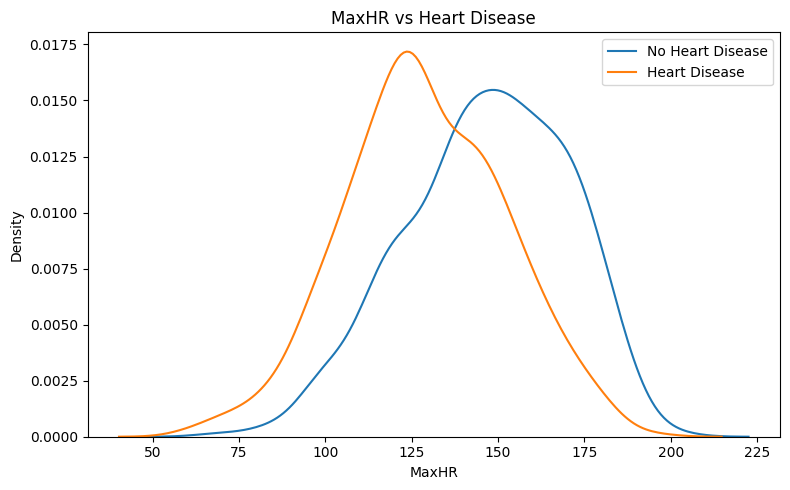


Fig 3.9: Maximum heart rate vs heart disease

* Old peak vs Heart patients

Old peak is numerical value of ST slope when it is downward which indicated that there is not enough blood supply passing through the heart and it is observed that heart patients have the higher old peak value which indicated that there is not enough blood supply towards heart and they have the value in range of -1 to 4. In contrast, healthy patients have the old peak value of -1 to 1 as shown in fig 3.10

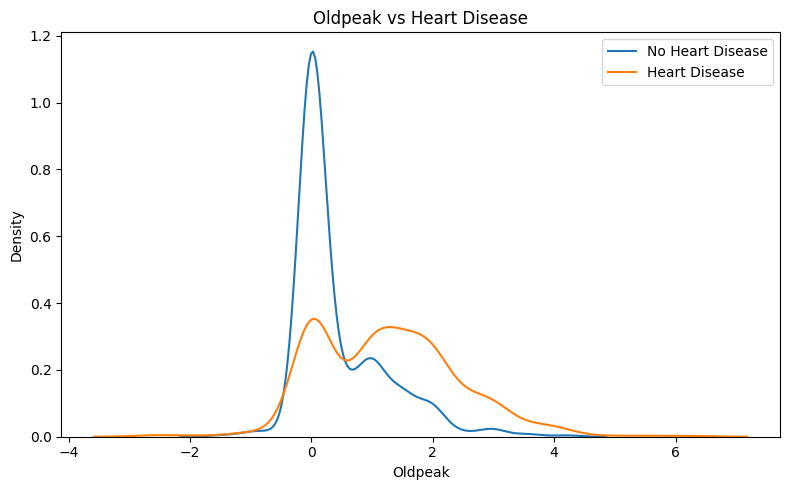


Fig 3.10: Old peak vs Heart Patients

* Fasting Blood Sugar

There are almost 790 patients who have the normal Fasting blood sugar and 220 above patients who have higher fasting blood sugar as shown in fig 3.11

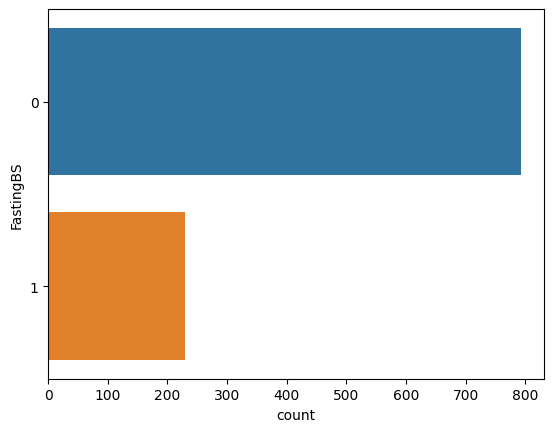


Fig 3.11: Fasting Blood Sugar

* Gender

There are 783 male patients and 239 female patients in the dataset as shown in fig 3.12

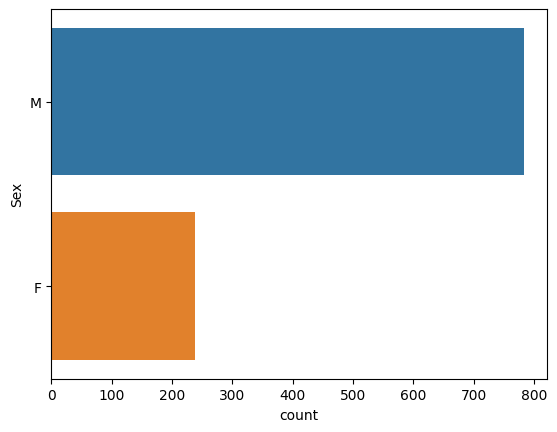


Fig 3.12 Gender

* Exercise Angina

There are almost 600 patients who have no exercise angina as compared to nearly 430 patients who are suffering from exercise angina.

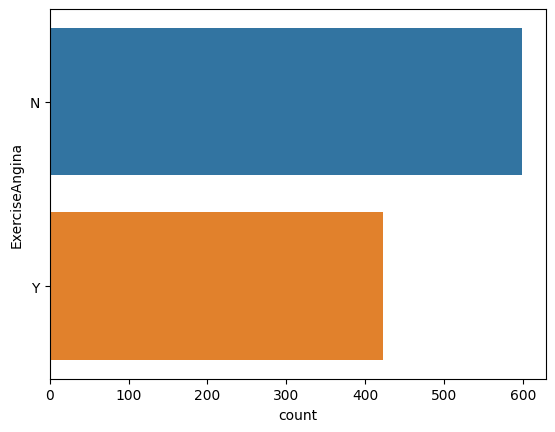


Fig 3.13: Exercise Angina

* ST Slope and Resting ECG

Most of the patients have the flat st slope followed by the up st slope which comes at second and there are few patients who have the down st slope as shown in fig 3.14. Moreover, most of the patients have the normal ECG followed by the ST and LVH indicators which some of the patients have as shown in fig 3.15.

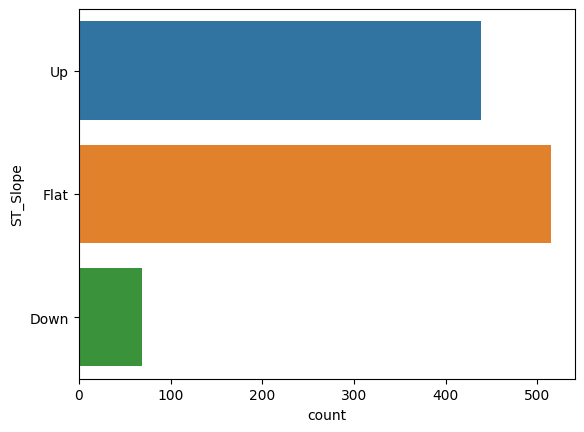


Fig 3.14 : ST Slope

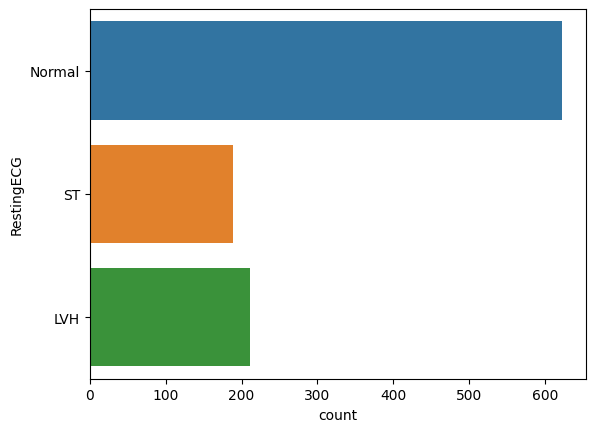


Fig 3.15 Resting ECG

* Gender, Chest pain type, Fasting blood Sugar, Resting ECG, Exercise Angina, ST Slope vs Heart Disease

By plotting the categorical features with the target variable the following graph fig 3.16 is obtain which gives the key information about the heart patients based on their test results.

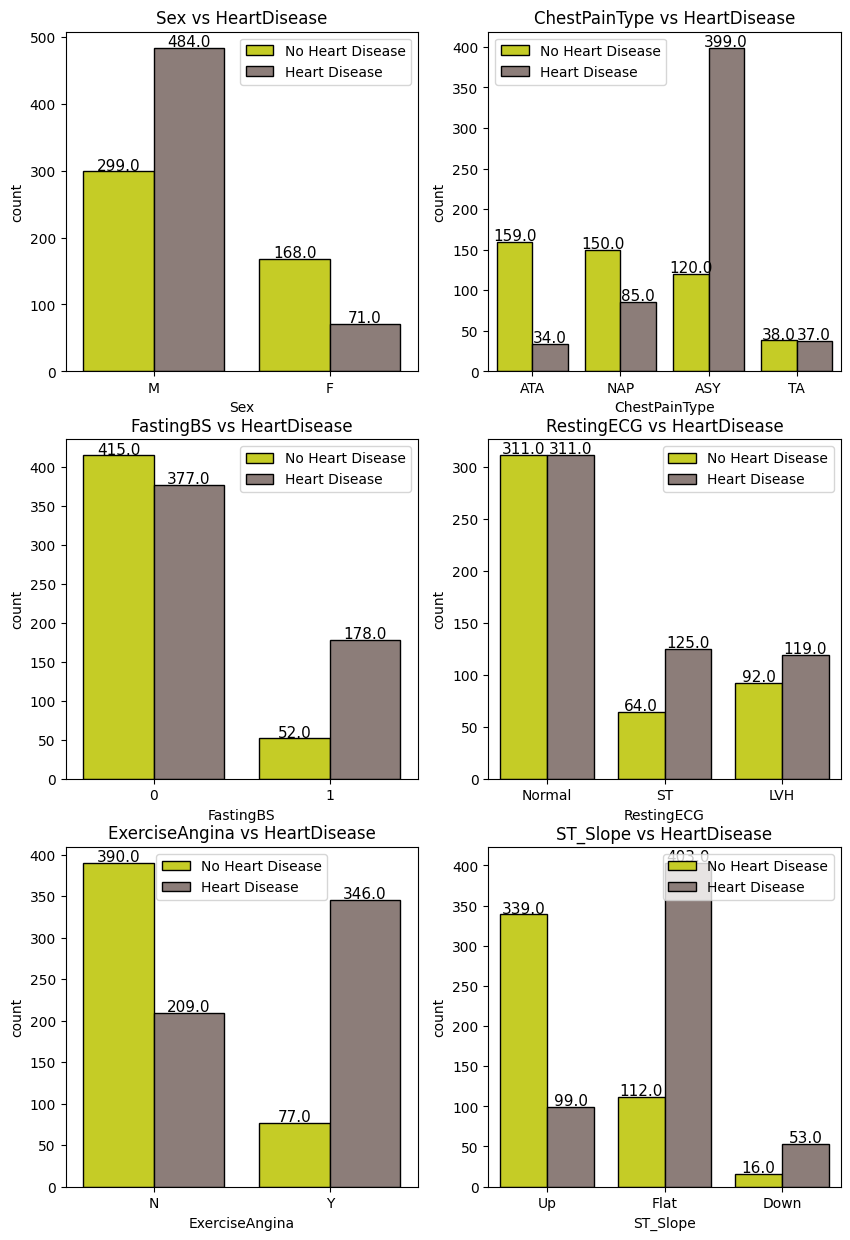


Fig 3.16: Comparison between heart patients vs healthy patients

By observing the graph following information are obtained.

1. From the given data it is observed that male person is more likely to have heart disease in old age.

|  |  |  |
| --- | --- | --- |
|  | MALE | FEMALE |
| Heart disease | 484 | 71 |
| No Heart disease | 299 | 168 |

Tabel 3.1: Gender vs heart disease

1. From the chest pain type graph, it is observed that number of heart patients and healthy patients who suffering from typical angina are same. There are larger number of healthy patients who have ATA and NAP chest pain type. Most of the heart patients are suffering from the ASY chest pain type.

|  |  |  |
| --- | --- | --- |
|  | Heart Disease | No Heart Disease |
| TA | 37 | 38 |
| ASY | 399 | 120 |
| NAP | 85 | 150 |
| ATA | 34 | 159 |

Table 3.2: Chest Pain type vs heart disease

1. Most of the heart disease patients have the normal fasting blood sugar same is true for healthy patients while some heart patients have the higher fasting blood sugar.

|  |  |  |
| --- | --- | --- |
|  | Heart Disease | No Heart Disease |
| Normal | 377 | 415 |
| High | 178 | 52 |

Table 3.3: Fasting Blood Sugar vs heart disease

1. Some of the patients who have the normal resting ECG are heart patients which indicates the underlying heart disease and there are few heart patients that have abnormal ECG as compared to healthy patients. In contrast, some of healthy patients also have abnormal resting ECG but they are not heart patients. Therefore, it is concluded that Resting ECG can be normal for heart patients and healthy patients which indicates that further testing needs to confirm heart disease.

|  |  |  |
| --- | --- | --- |
|  | Heart Disease | No Heart Disease |
| Normal | 311 | 311 |
| ST | 125 | 64 |
| LVH | 119 | 92 |

Tabel 3.4: Resting ECG vs heart disease

1. Most of the heart patients were suffering from the exercise angina while some of the heart patients were not suffering from exercise angina same is true for healthy patients.

|  |  |  |
| --- | --- | --- |
|  | Heart Disease | No Heart Disease |
| Exercise Angina | 346 | 77 |
| No Exercise Angina | 209 | 390 |

Table 3.5: Exercise angina vs heart disease

1. Most of the heart patients have the flat and down slope and few of them have the up slope while mostly healthy patients have the up slope and few of them have the flat and down slope.

|  |  |  |
| --- | --- | --- |
|  | Heart Disease | No Heart Disease |
| Up | 99 | 399 |
| Flat | 403 | 112 |
| Down | 53 | 16 |

Table 3.6: ST Slope vs heart disease.

* **Box Plot**

The box plot of the numerical features is given below

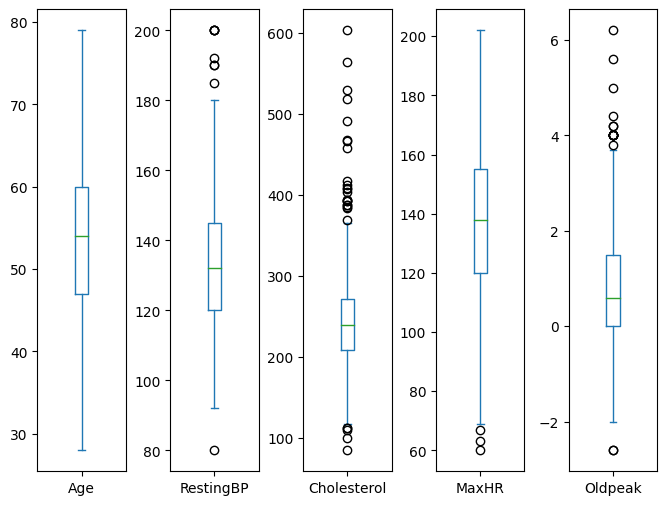


Fig 3.17: Box Plot

Cholesterol feature have the outliers which indicates that some of the patients have too low or too high cholesterol level and same is true for the old peak. Few of the heart patients have the maximum heart rate quite low.

* Correlation matrix

The correlation matrix indicates that some of the features have weak positive relation with each other and most of numerical features have no correlation with each other but from medical point of view all these features are valuable and are correlated with each other.

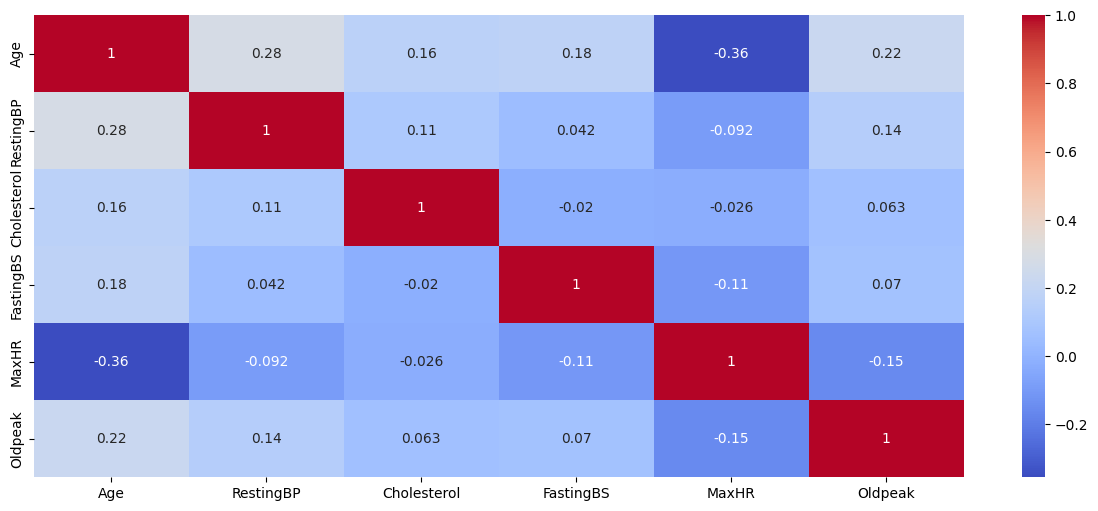


Fig 3.18: Correlation matrix

**3.3: Machine learning models results**

* **Decision Tree**

The classification report of the decision tree is shown in below fig

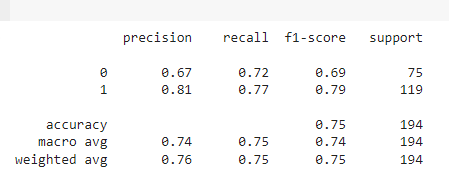


Fig 3.18: decision tree classification report

The overall precision that is obtained from the decision tree model is 0.8142 and area under the curve score is 0.74.

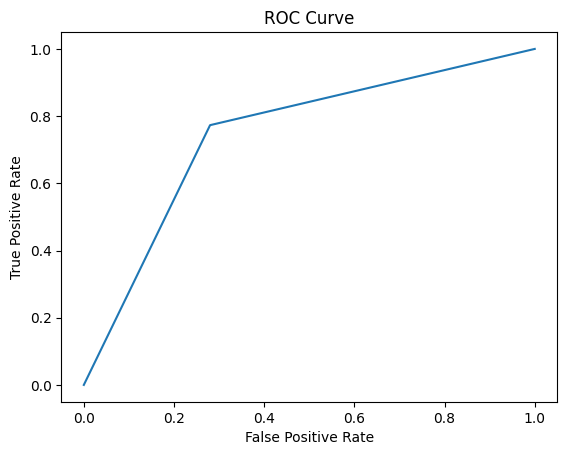


Fig 3.19: ROC Curve Decision Tree

The confusion matrix is given below

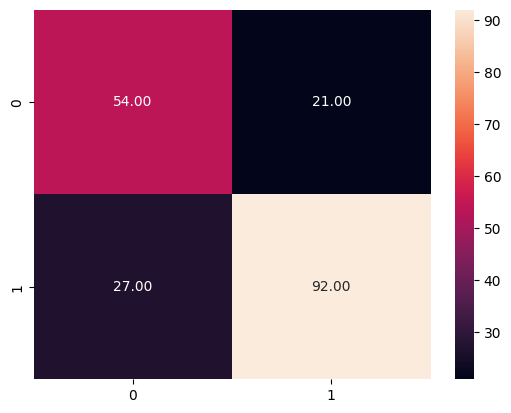


Fig 3.20: Confusion matrix decision tree

* **Random Forest Classifier**

The classification report of the random forest classifier is shown in below fig

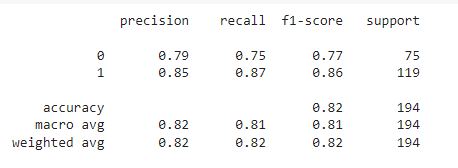


Fig 3.21: Classification report of random forest

The overall precision that is obtained from the decision tree model is 0.8454 and area under the curve score is 0.903.

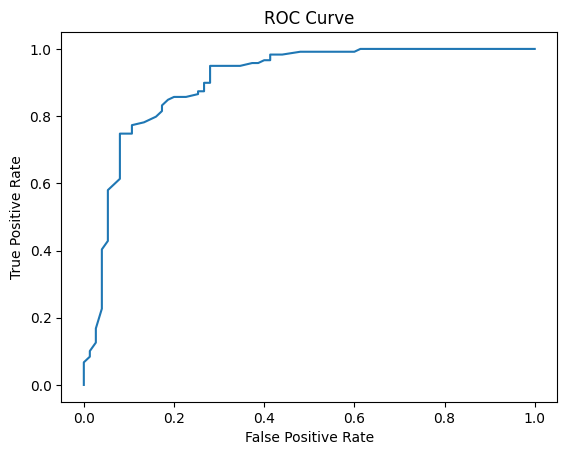


Fig 3.22: ROC curve random forest

The confusion matrix is given below

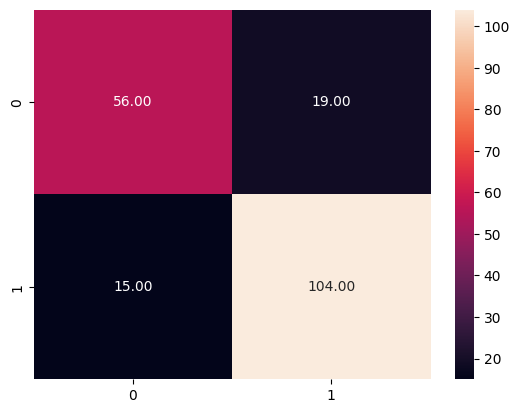


Fig 3.23: Confusion matrix for random forest classifier

* **Logistic Regression**

The classification report of the logistic regression is given below

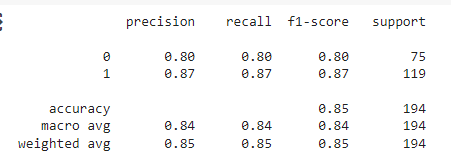


Fig 3.24: classification report for the logistic regression

The overall precision of the logistic regression is 0.8739 and AUC score is 0.9131

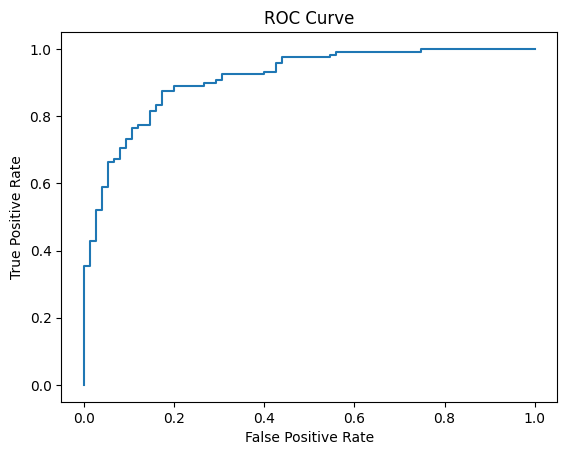


Fig 3.25: ROC curve for the logistic regression

The confusion matrix is given below

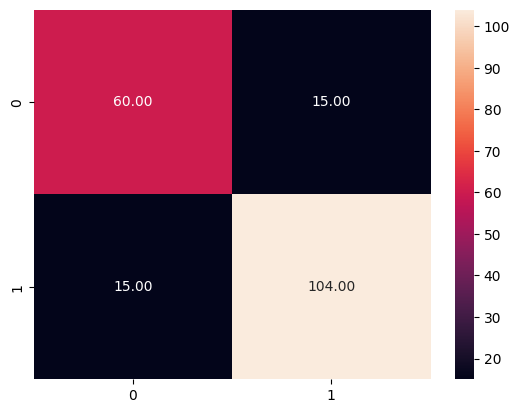


Fig 3.26: Confusion Matrix for logistic regression

* **Support Vector Classifier**

The classification report for the SVC is given below

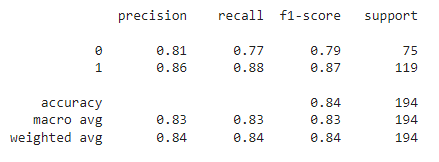


Fig 3.27: Classification report for SVC

The overall precision of SVC is 0.8060 and the AUC score is 0.9143.

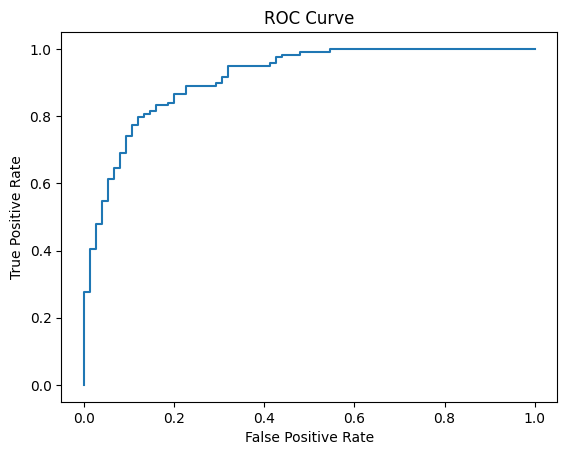


Fig 3.28: ROC Curve for SVC

The Confusion matrix is given below

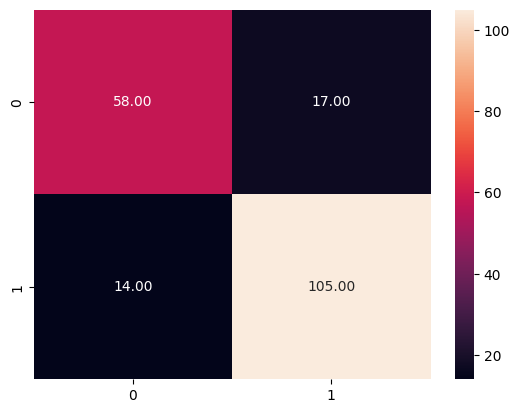


Fig 3.29: Confusion Matrix of SVC

* **KNN**

The classification report of KNN is given below

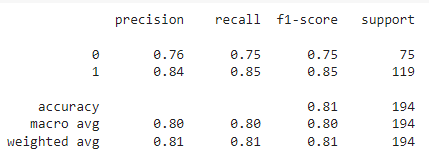


Fig 3.30: Classification report of KNN

The overall precision score of KNN is 0.84166 and the AUC score is 0.8746.

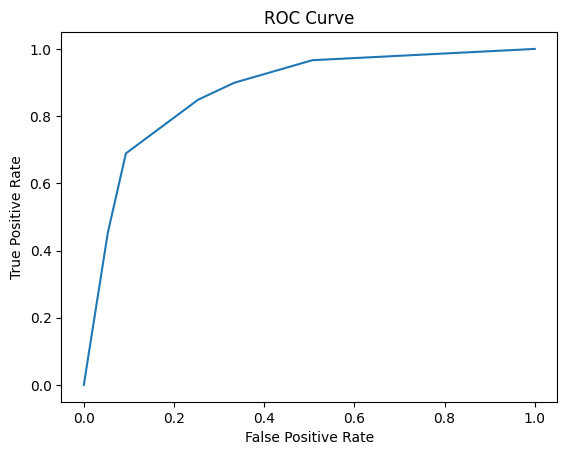


Fig 3.31: ROC Curve of KNN

The confusion matrix is given below

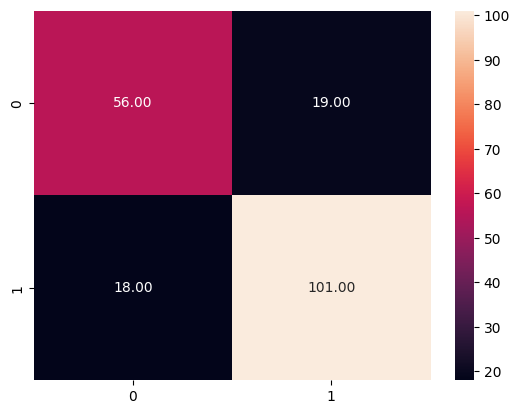


Fig 3.32: Confusion matrix of KNN

* **Gaussian Navie Bayes**

The classification report is given below

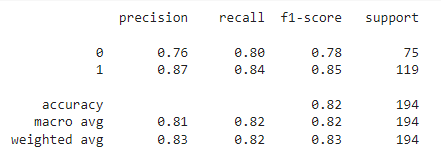


Fig 3.33: Classification report of Gaussian Navie Bayes

The overall precision score is 0.8695 and AUC score is 0.9157.

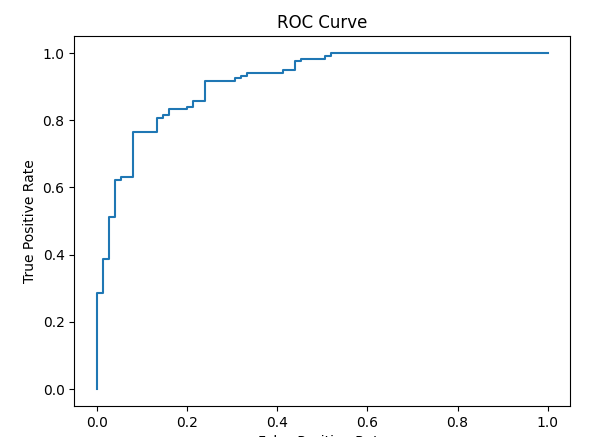


Fig 3.34: ROC curve of Gaussian Navie Bayes

The confusion matrix is given below

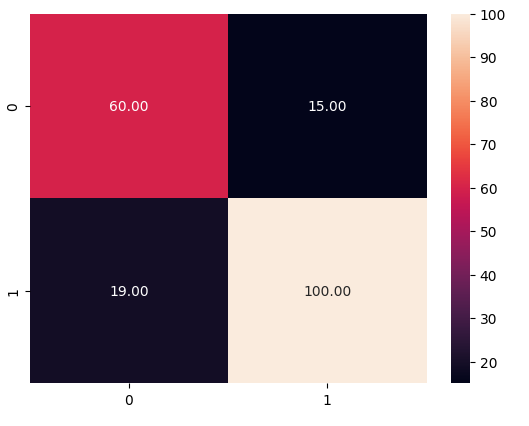


Fig 3.35: Confusion matrix of Gaussian navie bayes

* **Multi-layer perceptron**

The architect of the MLP is given below

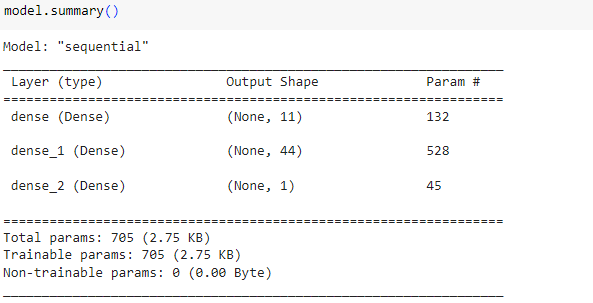


Fig 3.36: Architect of MLP

The precision score of MLP is 0.8429 and the AUC score is 0.80190.

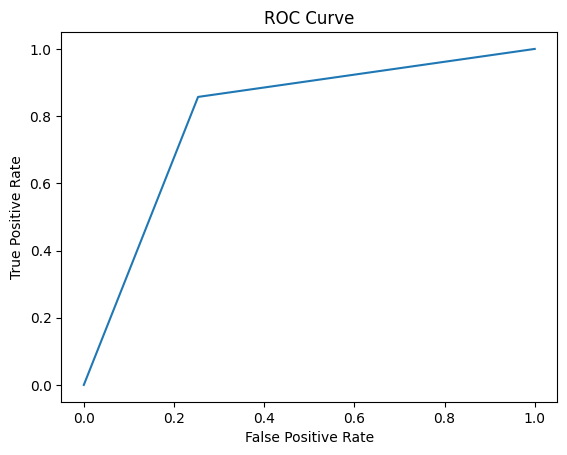


Fig 3.37: ROC Curve of MLP

The confusion matrix is given below

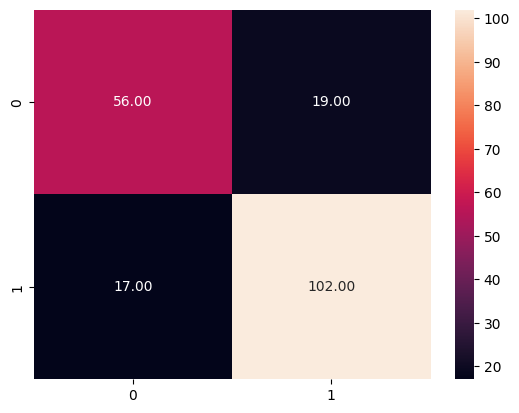


Fig 3.38: Confusion matrix of MLP

**3.4: Comparison of machine learning models**

* The precision scores of all models are given below

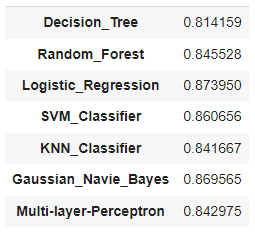


Fig 3.39: Precision of models

precision of logistic regression is higher as compared to all.

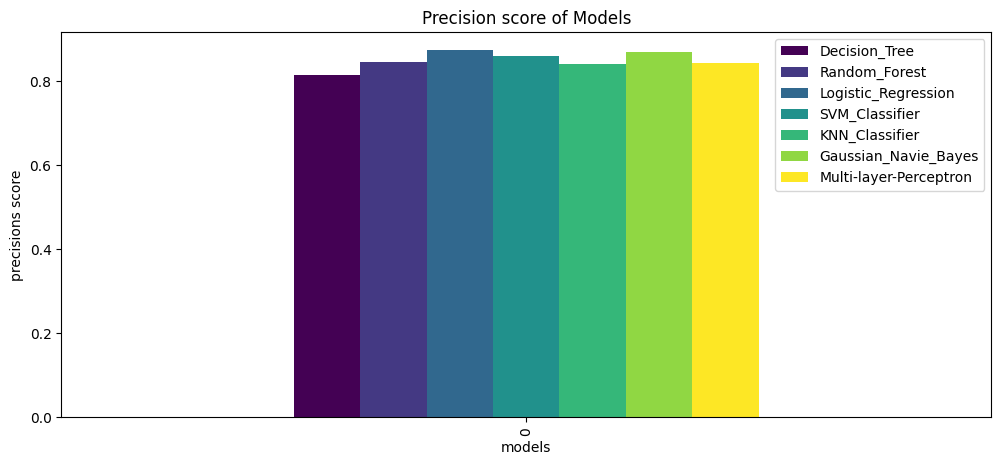


Fig 3.40: chart of precisions

* The accuracy score of all models is given below

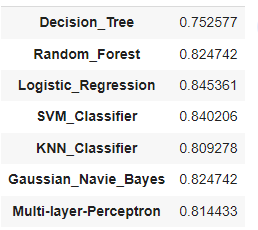


Fig 3.41: accuracy of models

The accuracy of the logistic regression is higher as compared to all.

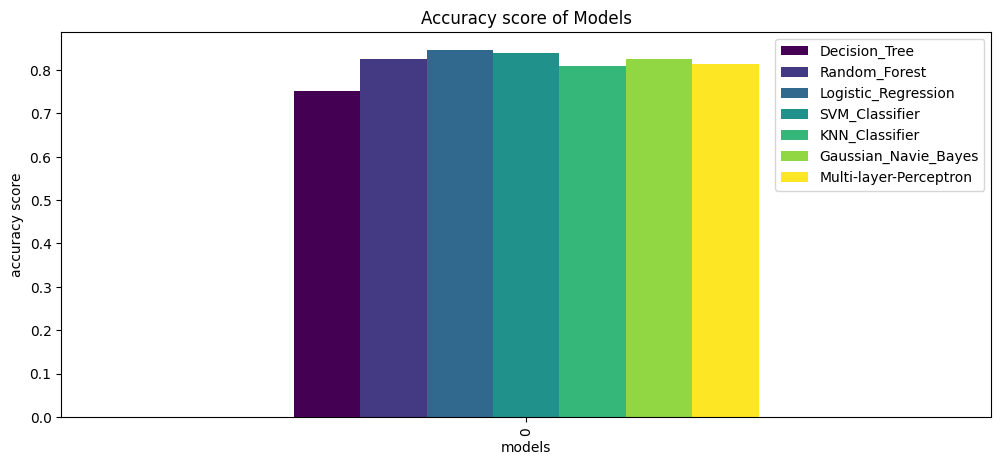


Fig 3.42: accuracy score

* Reall of models are given below

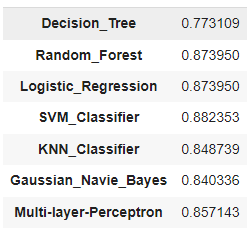


Fig 3.43: Recall score of models

Recall score of SVM Classifier is higher as compared to others.

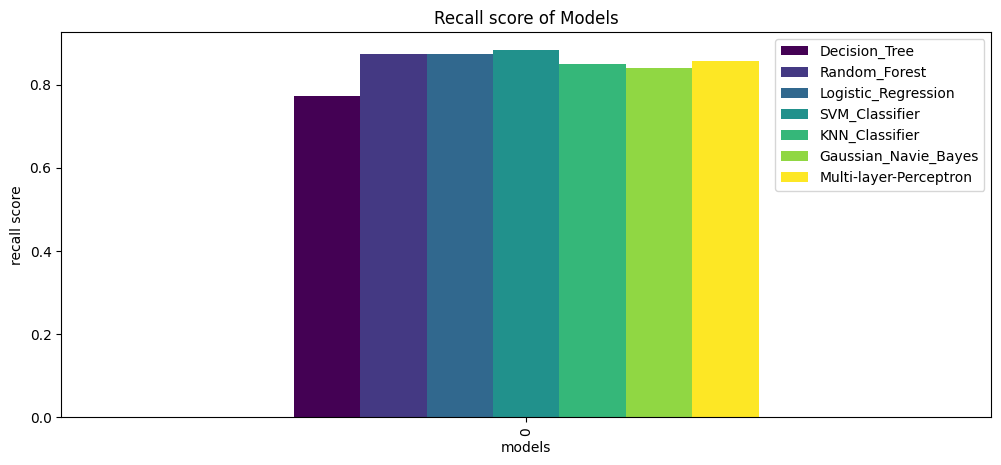


Fig 3.44: Recall scores

* F1 Score of models is given below

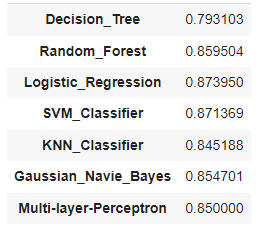


Fig 3.45 F1 Score of models

Logistic Regression have the higher F1 Score.

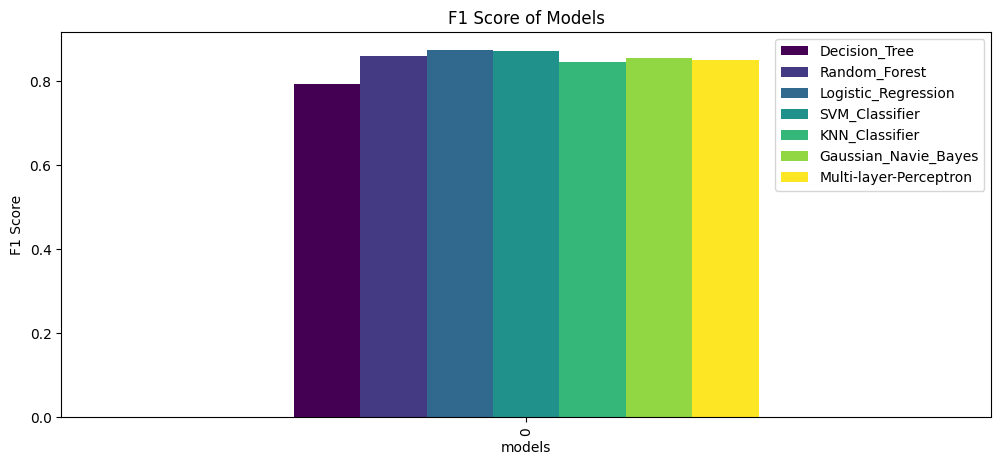


Fig 3.46: F1 Scores

By comparing the all models based on precision, accuracy f1 score and recall it is concluded that logistic regression is best model for prediction of heart disease on given dataset.

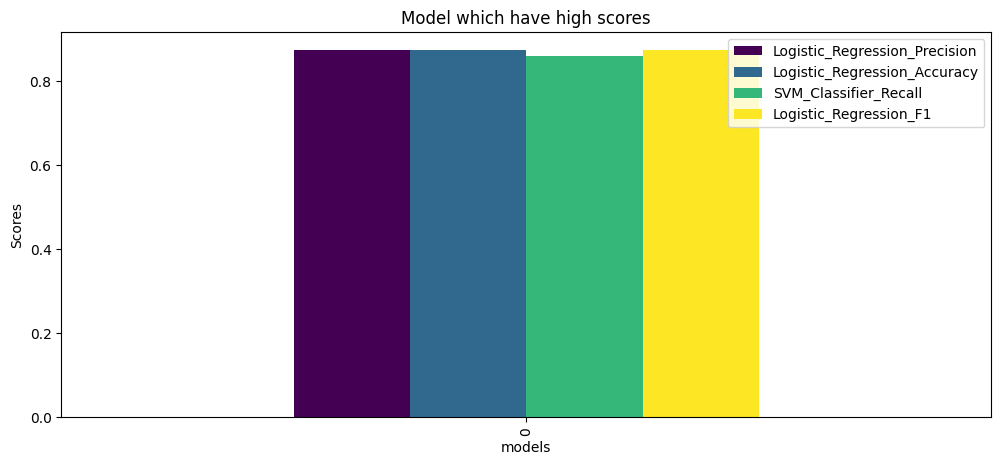


Fig 3.47: Comparing the models

The following comparison is observed between the machine learning models on the given dataset.

* Decision tree performs well on classification problems but it has drawback of overfitting and sensitive to variation in data. Moreover, it is unable to handle complex non-linear relationship between features which might be reason it does not perform well on heart disease dataset.
* Random Forest is suitable for classification when the number of features is large but again it suffers from overfitting problem which might be reason that it did not make correct prediction in this case as expected.
* Logistic regression best suited for binary classification due to its sigmoid function and it is observed that it is best model among all in our case.
* SVM perform well on classification and it inherently provides class predictions without directly giving probability estimates for the target classes which might be reason it did not perform well in this case.
* KNN works well on the numerical data and it assume that datapoint belongs to nearby class which is not always the case and it is sensitive to outliers and irrelevant features. Moreover, larger value of single feature may dominant the distance calculation which may affect the result. Some of our features were having large scale values as compared to other so it might be reason it did not make prediction accurate.
* Gaussian Navie bayes assume that data has the normal distribution and it works well on numerical data when target variable is categorical. It also assumes the linear independence among features and also consider that features follow the normal distribution. In contrast, some of our numerical feature did not follow the normal distribution which is reason it did not perform well.
* Multi-layer perceptron can be used for binary or multi class classification. Selection of number of neurons is crucial in MLP and it requires the large amount of data for training which is the reason it did not perform well.