Heart Disease prediction

1. Introduction

The dataset contains 133 samples for predicting the occurrence of heart disease. Each sample has 13 features that can be used to predict cardiac failure. The features include demographic details like age and sex and clinical indicators. The cardio vascular disease indicators are: Type of chest pain, Blood pressure, Cholesterol, Fasting blood sugar, Resting electrocardiographic results, Maximum heart rate achieved, Exercise induced angina, Old peak, Slope, Number of major vessels coloured by fluoroscopy and Thalassemia. The target variable is the occurrence or non-occurrence of heart disease. The features are explained in Table 1.

Table 1

Feature	Description	Feature value range
Age	Age in years	29 to 77 years
Sex	Gender	1=male, 0=female
Cp	Chest pain type	0=typical angina, 1=atypical angina 2=non-anginal pain, 3=asymptomatic
Trestbps	Resting blood pressure	94 to 200
Chol	Serum cholesterol	126 to 564
Fbs	Fasting blood sugar > 120mg/dL	1=true, 0=false
Restecg	Resting electrocardiographic results	0=Normal, 1=ST-T wave abnormality 2=Left ventricular hypertrophy
Thalach	Maximum heart rate achieved	71 to 202
Exang	Exercise induced angina	1=Yes, 0=No
Oldpeak	Stress test depression induced by exercise relative to rest	0 to 6.2
Slope	Slope of peak exercise ST segment	0=Upsloping, 1=Flat, 2=Down sloping
Ca	Number of major vessels	0 to 3 coloured by fluoroscopy
Thal	Thallium heart rate	0=Normal, 1=Fixed defect, 2=Reversible defect, 3=Not described
Target	Diagnosis of heart disease	0=No disease, 1=Disease

The objective of data analysis is to predict the prevalence of heart disease from a set of factors represented as features in the data set. The objective also includes an identification of the most important factors for predicting the heart disease.

2. Data preprocessing

First, the data is checked for any missing or null values. The current data does not have any missing values. Next, the categorical variables are identified and converted into datatype 'category'. The categorical data is encoded by using One-Hot encoding. One-Hot encoding is used to encode the categorical variables with more than two categories. For this, the category names are assigned to the categories of each qualitative variable. Next, the implementation of One-Hot encoding transforms each qualitative variable into a set of columns where each column represents the presence or the absence of a class/category of the variable. For instance, '*Chest pain type*' is a categorical variable in the original data with four categories represented as '0', '1', '2' and '3'. The categories in the numerical form are mapped into category names with string data type.

Chest pain type
3
2
1
1
0

Chest pain type
asymptomatic
non-anginal pain
atypical angina
atypical angina
typical angina

	asymptomatic	atypical angina	non-anginal pain	typical angina
0	1.0	0.0	0.0	0.0
1	0.0	0.0	1.0	0.0
2	0.0	1.0	0.0	0.0
3	0.0	1.0	0.0	0.0
4	0.0	0.0	0.0	1.0
...	...	...	...	...
298	0.0	0.0	0.0	1.0
299	1.0	0.0	0.0	0.0
300	0.0	0.0	0.0	1.0

The data is then one hot encoded, converted into a data frame and concatenated with the original data. After transformation, the original column '*Chest pain type*' is dropped. The process of data encoding is repeated for the remaining categorical variables.

After one hot encoding, the quantitative variables are identified and scaled to a fixed range between 0 and 1. *'Minmax scaler'* is used to scale the data and is suitable even when the data distribution is not gaussian. For the given data, variables like age, blood pressure, cholesterol, maximum heart rate achieved, old peak and major vessels coloured by fluoroscopy are scaled using '*Minmax scaler*'. The final dataset in shown in Table 2.

Table 2

	age	sex (male = 1)	Blood pressure	cholestrol	blood sugar (high = 1)	maximum heart rate achieved	exercise induced angina	oldpeak	number of major vessels coloured by fluoroscopy	target	...	ST-T wave abnormality	left ventricular hypertrophy	normal	downsloping	flat
0	0.708333	1	0.481132	0.244292	1	0.603053	0	0.370968	0.0	1	...	0.0	0.0	1.0	0.0	0.0
1	0.166667	1	0.339623	0.283105	0	0.885496	0	0.564516	0.0	1	...	1.0	0.0	0.0	0.0	0.0
2	0.250000	0	0.339623	0.178082	0	0.770992	0	0.225806	0.0	1	...	0.0	0.0	1.0	1.0	0.0
3	0.562500	1	0.245283	0.251142	0	0.816794	0	0.129032	0.0	1	...	1.0	0.0	0.0	1.0	0.0
4	0.583333	0	0.245283	0.520548	0	0.702290	1	0.096774	0.0	1	...	1.0	0.0	0.0	1.0	0.0

3. Exploratory data analysis

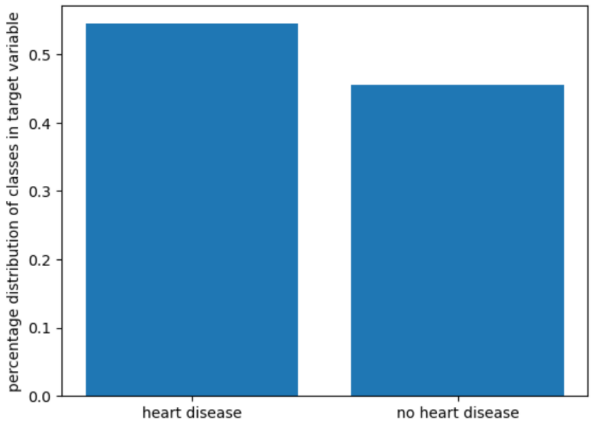


Figure 1

Figure 1 shows that the distribution of target variable is not perfectly balanced. More than 50% of the samples have positive labels. However, the data is not entirely imbalanced as the overall distribution is 55% and 45%.

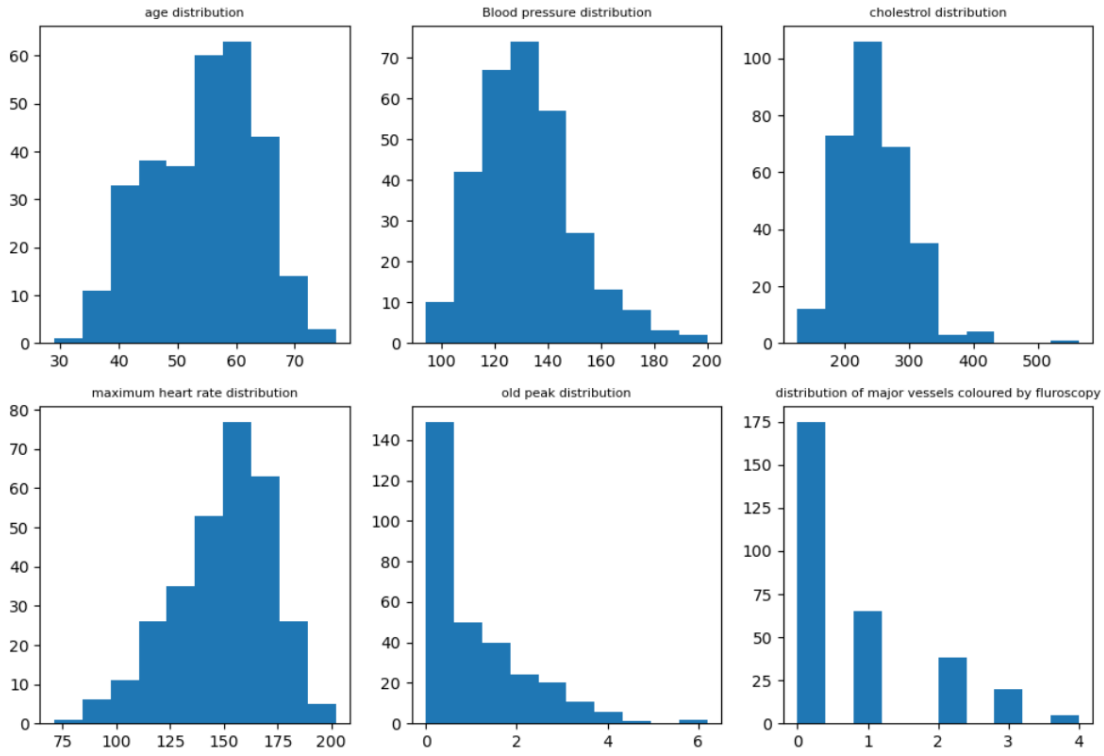


Figure 2

Figure 2 shows the data distribution of quantitative features. None of the quantitative variables has an ideal gaussian distribution. The distribution of '*Age'* is slightly gaussian with mean = 0.53. '*Blood pressure*' and '*Cholesterol'* are slightly skewed towards left with means < 0.5 (0.36 and 0.27 respectively). This shows that 75% of the use cases have blood pressure levels lower than 140 and cholesterol less than 274. The *Maximum heart rate* distribution is skewed towards right with mean rate = 150 and maximum = 202. Most the use cases have old peak values lower than 1.6 and the maximum recorded old peak is 6.2. The *Number of major vessels coloured by fluoroscopy* is fewer than 1 for majority of the data samples. The descriptive statistics of quantitative features is shown in Table 3.

Table 3

	age	Blood pressure	cholestrol	maximum heart rate achieved	oldpeak	number of major vessels coloured by fluoroscopy
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
mean	54.366337	131.623762	246.264026	149.646865	1.039604	0.729373
std	9.082101	17.538143	51.830751	22.905161	1.161075	1.022606
min	29.000000	94.000000	126.000000	71.000000	0.000000	0.000000
25%	47.500000	120.000000	211.000000	133.500000	0.000000	0.000000
50%	55.000000	130.000000	240.000000	153.000000	0.800000	0.000000
75%	61.000000	140.000000	274.500000	166.000000	1.600000	1.000000
max	77.000000	200.000000	564.000000	202.000000	6.200000	4.000000

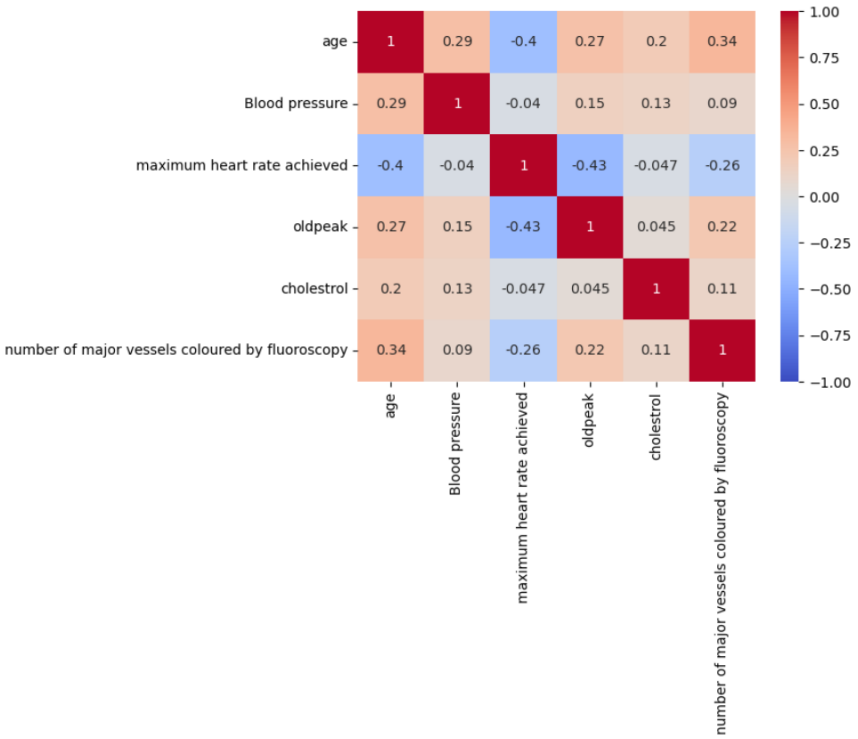


Figure 3

Apart from the data distribution of each quantitative variables, the correlation matrix in Figure 3 shows that nearly all the quantitative features are weakly correlated. The highest correlation is -0.43 between old peak and maximum heart rate achieved. The correlation between maximum heart rate and age is -0.43. Hence, the association of each feature with the target variable can be analysed independent of other features.

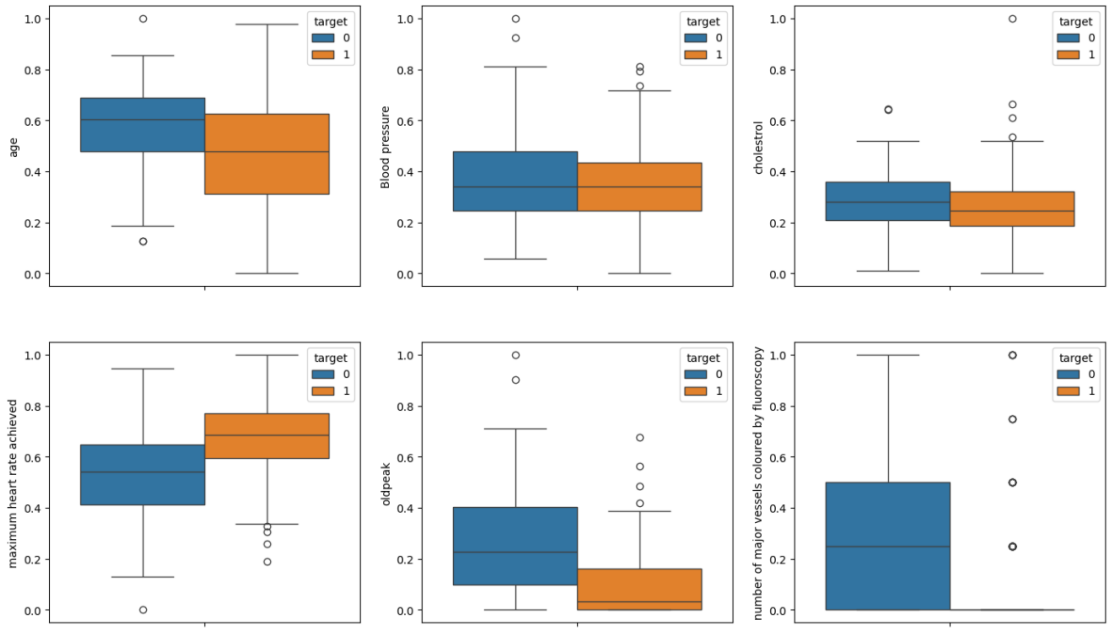


Figure 4

A distribution of the quantitative features for each class of target variables is shown in Figure 4. It is observed than blood pressure and cholesterol levels are not suitable indicators of heart disease, given a large overlap between positive and negative labels of the target variable. A distribution of '*Maximum heart rate achieved*' shows that the maximum heart rate, for nearly 75% of the cases without heart disease, was less than 155. On the other hand, the maximum heart rate was beyond 150 for 75% of the samples with the heart disease. Similarly, there is a small overlap in '*old peak*' values with most of the old peak values less than 1 for the cases having heart disease and more than 1 for cases with no heart disease. The distribution of '*Major vessels coloured by fluoroscopy*' proves that the major vessels of the patients of any cardiovascular disease are blocked and cannot be coloured by fluoroscopy. However, there are few outliers where the major vessels are coloured even in heart patients. It is not difficult to ascertain if age contributes to heart diseases, given small size of 133 samples and an overlap between occurrence and non-occurrence of heart disease.

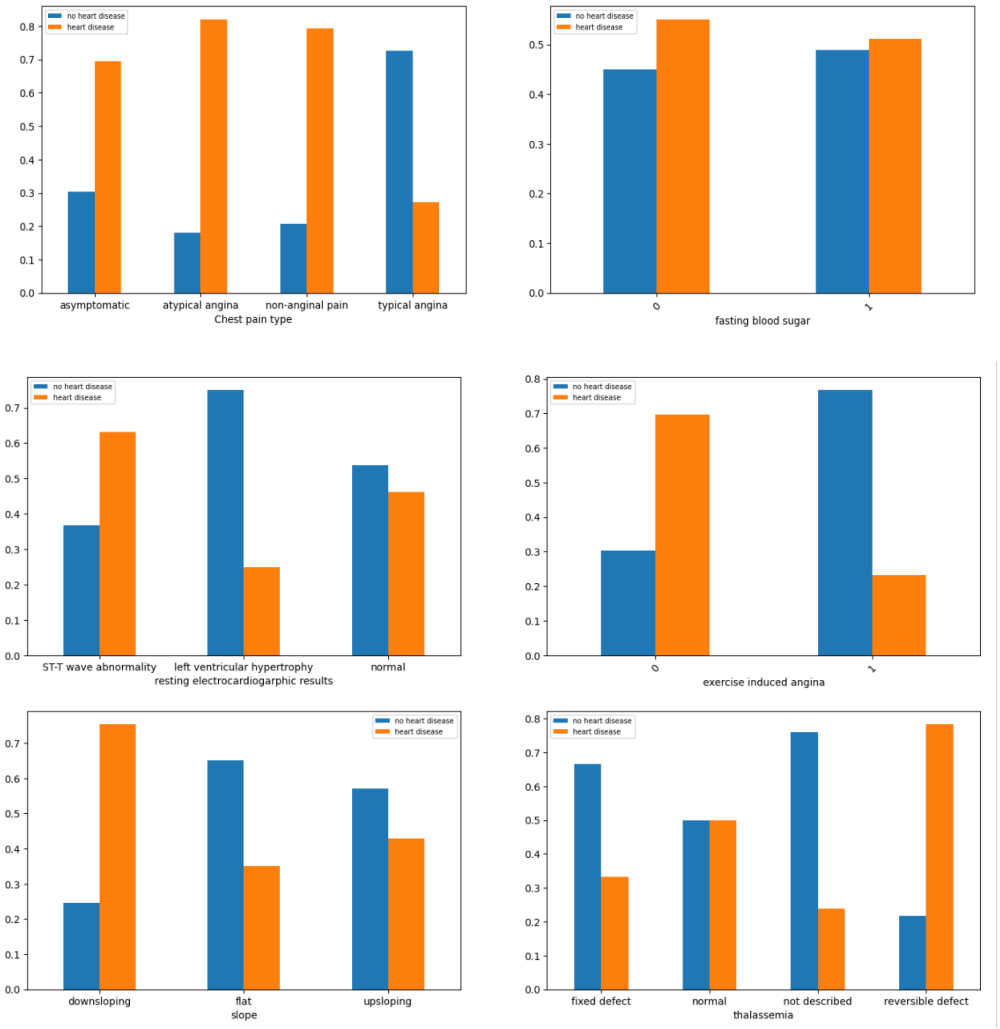


Figure 5

The relationship between categorical data and the target variable was gauged by observing the distribution of different categories of each categorical feature against the target, as shown in Figure 5. The chest pain distribution shows that 60% to 80% each of asymptomatic, atypical angina and non-anginal pain samples had heart disease. Similarly, most of the cases with symptoms of down sloping, reversible defect in thalassemia and the absence of exercise induced angina had no heart diseases.

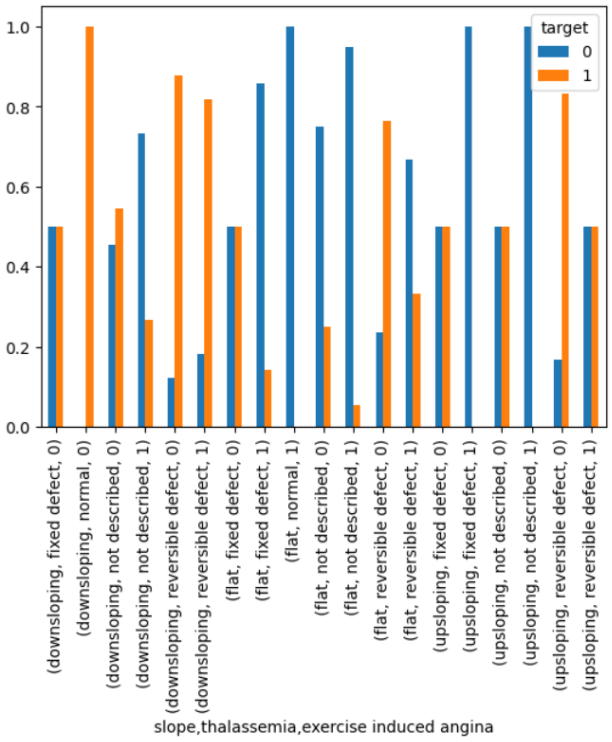


Figure 6

The combined impact of slope, thalassemia and exercise induced angina in Figure 6 shows that the reversible defect in thalassemia is one of the dominant factors associated with occurrence of heart disease.

3.1 Feature selection

Given a weak correlation among quantitative variable, the distribution of each quantitative variable against target variable was analysed independently. It is observed from Figure 4 shows that '*Blood pressure*' and '*Cholesterol*' are not suitable indictors, given a large overlap between positive and negative labels of target variable. The '*Age*' distribution shows 75% of cases of heart disease with age less than 65 and 75% of cases without heart disease falling the age group 20 to 70. Age, alone, may not be a reliable indicator.

Figure 5 shows that nearly 55% of people, with blood sugar levels below 120mg/dL, have heart diseases. Out of those with higher blood sugar levels, nearly 48% do not have heart disease. It is, therefore, difficult to infer any strong association between blood sugar and prevalence of heart disease. Similarly, there are 50% of cases of heart diseases under normal conditions in thalassemia, making it a weak indicator of any heart disease. There is also a little difference between the percentage of cases with and without heart disease under upsloping or normal conditions in resting electrocardiographic results. Figure 6 also shows that the reversible defect in thalassemia and the down sloping condition are dominant factors associated with the prevalence of heart disease.

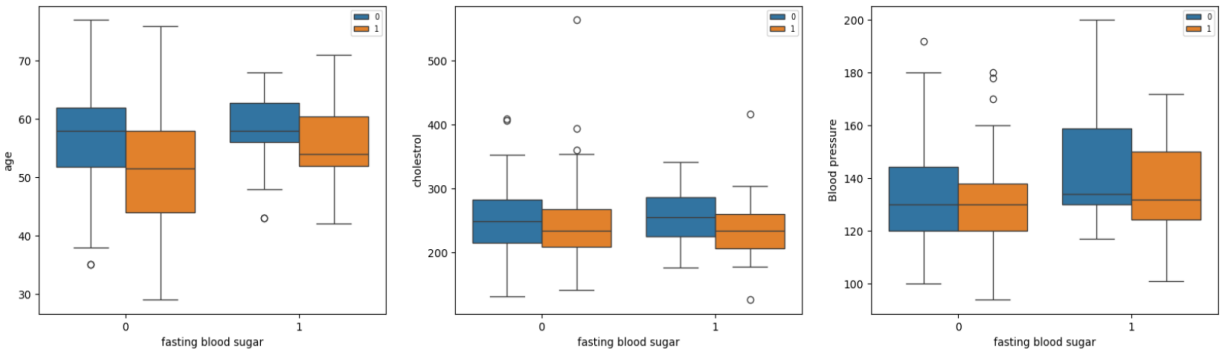


Figure 7

Figure 7 shows distribution of some quantitative variables against '*Blood sugar*' for positive and negative cases of heart disease. It is observed that there are overlaps between '*Age'* and '*Blood sugar'*, '*Age'* and '*Cholesterol* and '*Age'* and '*Blood pressure*' notwithstanding the prevalence of heart disease.

After analysing the relationships among the features and between features and target variable, it was inferred that '*Age'*, '*Cholesterol'*, '*Sex'*, '*Blood sugar*', may have weak association with the target variable. Hence, the features selected for model building are:

'*Typical angina', 'Reversible defect', 'Number of major vessels coloured by fluoroscopy', 'Old peak',' Maximum heart rate achieved', 'Non-anginal pain', 'Atypical angina', 'Asymptomatic', 'Left ventricular hypertrophy', 'Exercise induced angina', 'Down sloping', 'Not described', 'Flat', 'Upsloping', 'Fixed defect', 'ST-T wave abnormality*'.

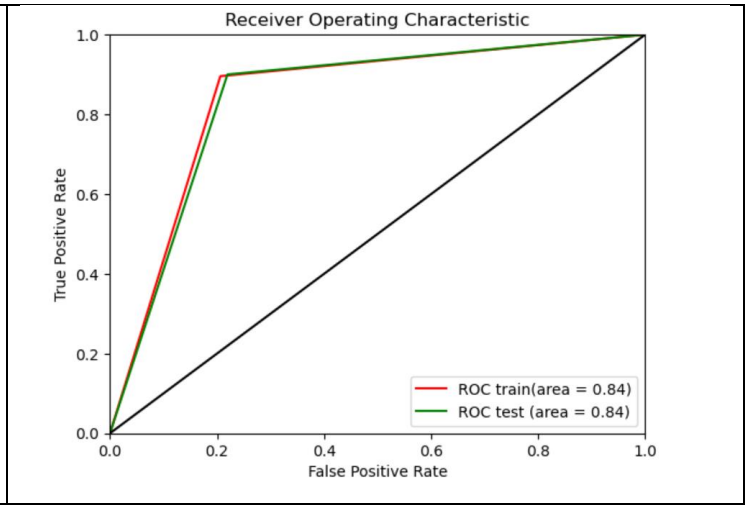
4. Model development

The dataset is split into train and test data. 70% of the data is train. The train data is further subjected to cross validation to select the model with most optimal parameter values. The balance is used to test the model performance.

Logistic Regression (LR): Logistic Regression is a binary classification technique. The model is trained by using 3-fold cross validation and tuning the hyperparameters with Grid search Cross validation. The parameters used for model tuning are penalty, class weight, maximum number of iterations and inverse of regularisation strength. The metric used to select the best parameters is F1 score. F1 score is the harmonic mean of recall and precision and can handle any imbalance in the target variable. The selected parameters with the highest F1 score were used to fit the final model on the train data. The classification report and ROC-AUC graph are shown in Table 4.

Table 4

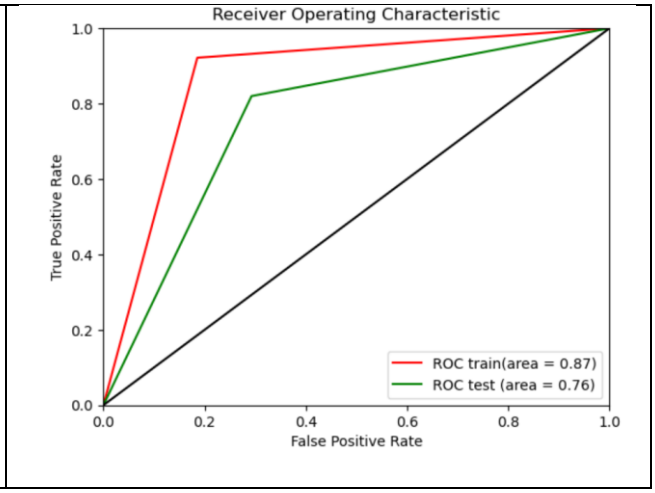
```
classification report of train data is
              precision    recall  f1-score   support

           0       0.87      0.79      0.83        97
           1       0.84      0.90      0.87       115

    accuracy                           0.85       212
   macro avg       0.85      0.84      0.85       212
weighted avg       0.85      0.85      0.85       212

classification report of test data is
              precision    recall  f1-score   support

           0       0.86      0.78      0.82        41
           1       0.83      0.90      0.87        50

    accuracy                           0.85        91
   macro avg       0.85      0.84      0.84        91
weighted avg       0.85      0.85      0.85        91
```



Decision Tree (DT) classifier: Decision tree classifier is suitable for binary classification. The decision tree is used to predict the target value by splitting the data based on one feature at a time. The classifier parameters for hyperparameter tuning are: criterion, maximum depth, minimum number of samples required to split an internal node, maximum number of features to consider for the best split and class weight. Based on F1 score, the best parameters are selected for training the DT model on train data. The classification report and ROC-AUC are shown in Table 5.

Table 5

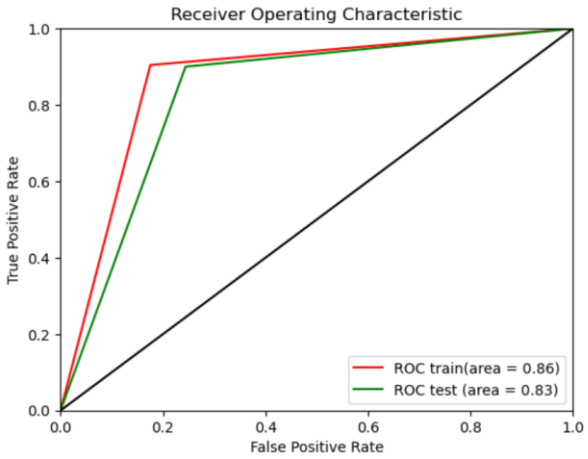
```
classification report of train data is
              precision    recall  f1-score   support

           0       0.90      0.81      0.85        97
           1       0.85      0.92      0.89       115

    accuracy                           0.87       212
   macro avg       0.88      0.87      0.87       212
weighted avg       0.87      0.87      0.87       212

classification report of test data is
              precision    recall  f1-score   support

           0       0.76      0.71      0.73        41
           1       0.77      0.82      0.80        50

    accuracy                           0.77        91
   macro avg       0.77      0.76      0.77        91
weighted avg       0.77      0.77      0.77        91
```



Random Forest (RF) classifier: The Random Forest uses multiple decision trees on various sub-samples of the dataset and uses averaging to predict the target values. The hypermeters for tuning the model and selecting the best Random Forest parameter values are: number of estimators, criterion, maximum depth, minimum number of samples for each internal node, maximum features to consider the best split and class weight. The optimal parameter values are selected with hyperparameter tuning and the final model is trained on the train data (using the most important features). The classification report and ROC-AUC graph are shown in Table 6.

Table 6

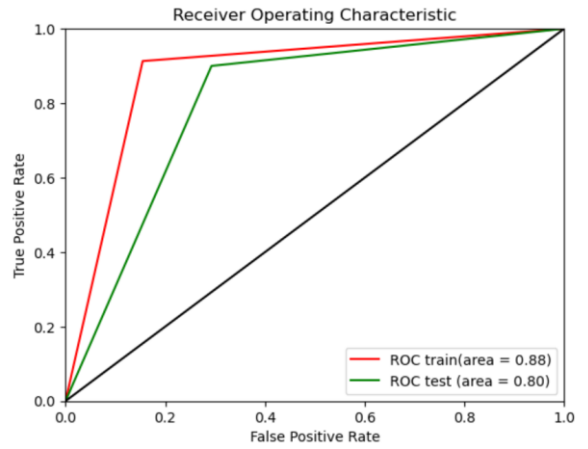
```
classification report of train data is
              precision    recall  f1-score   support

           0       0.88      0.82      0.85        97
           1       0.86      0.90      0.88       115

    accuracy                           0.87       212
   macro avg       0.87      0.86      0.87       212
weighted avg       0.87      0.87      0.87       212

classification report of test data is
              precision    recall  f1-score   support

           0       0.86      0.76      0.81        41
           1       0.82      0.90      0.86        50

    accuracy                           0.84        91
   macro avg       0.84      0.83      0.83        91
weighted avg       0.84      0.84      0.83        91
```



Support Vector Machines: Support Vector Machine (SVM) is classification method and is effective when the relationship between features and target variable is complex and non-linear. Grid Seach CV is used to tune the following hyperparameters: Inverse of regularisation strength, type of kernel, type of kernel coefficient and class weight. The model with the best parameters is fitted on the train data. The classification report and ROC-AUC graph are shown in Table 7.

Table 7

```
classification report of train data is
              precision    recall  f1-score   support

           0       0.89      0.85      0.87        97
           1       0.88      0.91      0.89       115

    accuracy                           0.88       212
   macro avg       0.88      0.88      0.88       212
weighted avg       0.88      0.88      0.88       212

classification report of test data is
              precision    recall  f1-score   support

           0       0.85      0.71      0.77        41
           1       0.79      0.90      0.84        50

    accuracy                           0.81        91
   macro avg       0.82      0.80      0.81        91
weighted avg       0.82      0.81      0.81        91
```



4.1 Model comparison

Table 8 is a comparison of accuracy, recall, precision and F1 score of all the models on test data.

Table 8

Model	Accuracy	Recall	Precision	F1 score
Logistic Regression	0.85	0.90	0.83	0.87
Decision Tree classifier	0.77	0.82	0.77	0.80
Random Forest classifier	0.84	0.90	0.82	0.86
Support Vector Machine	0.81	0.90	0.79	0.84

The Accuracy, Recall, Precision and F1 values of Logistic Regression are the highest. Random Forest performance is better than SVM. The higher Accuracy and F1 values of Logistic Regression make it a suitable candidate for the final model selection. Decision tree is the worst performer, suggesting that DT classifier cannot be selected as the final model.

Figure 8 shows the AUC curve of all the models. The comparison of AUC curves is done separately for train and test data.

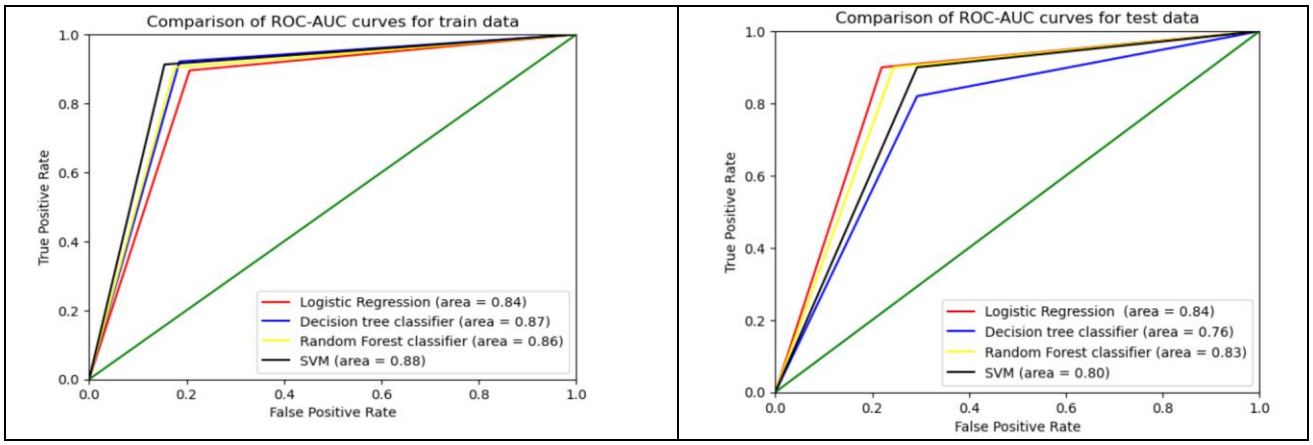


Figure 8

It is observed from Figure 8 that Area Under Curve (AUC) value of SVM is the maximum on train data. However, the performance worsens on test data, indicating the model's ability to pick up train data patterns with higher accuracy. Decision Tree AUC is better than LR and RF classifiers but the performance is the worst on test data among all the models, proving that the tree structure can get altered for any new data. The Decision Tree exhibits the problem of overfitting, as it becomes unstable with slight variations in data. The variations are acute with high dimensional data. The classifier's performance improves for large sample sizes, fewer features and balanced distribution of classes in categorical data. In the given case, the dataset had only 133 samples and large number of features.

The Random Forest performs better than Decision Tree. However, the values of performance metrics like accuracy, F1 score and ROC-AUC are better on train data than on test data. Although RF is robust to overfitting, it can still overfit if there is noise in the train data.

The Logistic Regression AUC values are exactly the same performance for both train and test data. The LR performance in Accuracy, Precision, Recall, F1 and ROC AUC are better than other machine learning algorithms. Logistic Regression is suitable for binary classification and performs better with small sample sizes. Hence, Logistic Regression is selected as the final model.

5. Model deployment

Creating web application: A web application framework named Flask was used for developing an interface that accepts patient input details and generating response for heart disease prediction. First, the trained model is saved as pickle file named '*model.pkl*'. Next, an application file named '*app.py*' is created in python. The file is the core of web application that responds to user requests based on the type of request. The Flask app is initialised and two functions are created. The first function is to display the content of HTML page of web app. The second function 'POST' is used to retrieve information from HTML file, feed the data to the trained model and send the prediction back to the HTML file to display

the result. The third file is HTML file named '*index.html*' stored in a 'Templates' folder. The HTML file consists of HTML code for creating input fields. The value entered in each input field, representing an independent variable, is fed to the model by using the function '*request.form.get*' in the 'app.py' file. The independent variables used for this web app included only the most important features as explained in 'Feature selection' section.

Deployment of web app: Following are the files created for web app deployment:

- model.pkl
- app.py
- Templates/index.html
- Procfile
- Requirements.txt

All the files are uploaded on a Git Hub repository. A new app named '*cvd-prediction*' is created in Heroku (web application hosting platform) and is connected to the Git repository. All the Git files are pushed to Heroku. The web app gets activated and can be accessed at the following link:

https://cvd-prediction-cd2be672a63c.herokuapp.com/

6. Future improvements

It is observed that even the best model cannot predict more than 84% of the labelled responses accurately. The model performance can be improved by increasing the number of data samples. It is unlikely that a new advanced algorithm can improve the performance. The XG Boost classifier did not improve the accuracy of train data beyond 85% and that of test data beyond 80%. Figure 9 shows that AUC of XG Boost classifier is 0.85 for the train data and 0.79 for the test data.

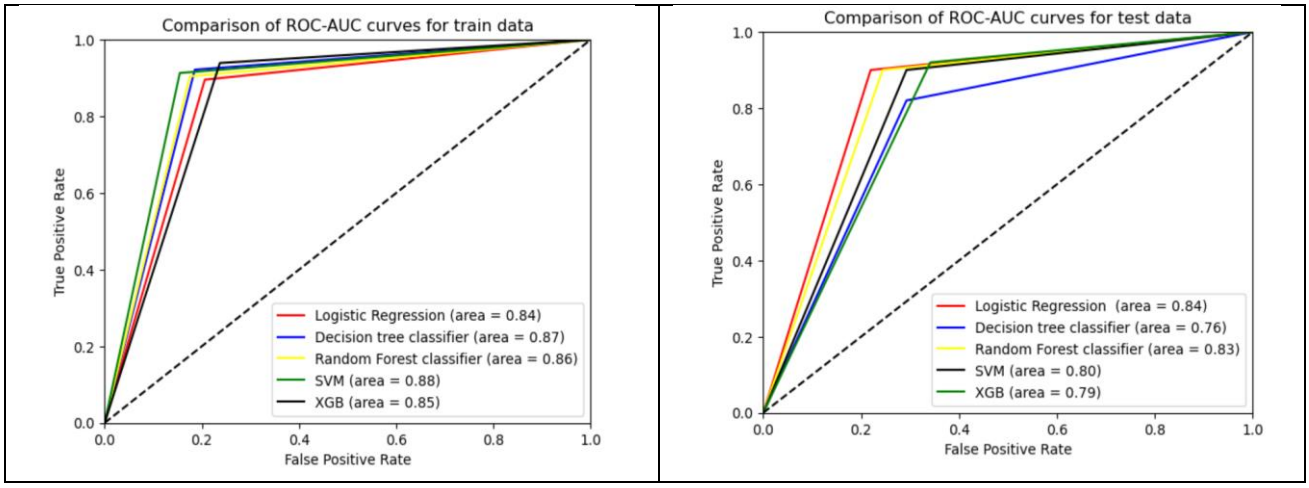


Figure 9

The ambiguity in relationships between the prevalence of heart disease and factors like '*Cholesterol*', '*Blood pressure*', '*Blood sugar*' and '*Age*' can be overcome by applying stratifying sampling. The collection of a large number of samples separately for different age groups and an equal distribution of samples between the two classes of gender can provide a clearer picture of the association of '*Age*' and '*Sex*' with heart disease prediction. A large sample size may also significantly alter the strength of association between '*Blood sugar*', '*Old peak*', '*Exercise induced angina*' and '*Slope*' and the heart disease prediction.