UC Berkeley MIDS
W200 Fall 2021 Project 2
December 8, 2021
Submitted By: Jinsoo Chung, Sunitha Haraginadoni, Haibi Lu
Team Github Repo: https://github.com/UC-Berkeley-I-School/Project2_Chung_Haraginadoni_Lu

Project 2 Final Report

What Causes Heart Disease?

Data Source: https://www.kaggle.com/fedesoriano/heart-failure-prediction

Context:

Heart disease was the number one leading cause of death in the United States in 2019 and accounted for over 23% of all deaths [1]. Coming in many forms, such as coronary artery disease, the most common form of heart disease in the US, heart arrhythmias, and heart valve disease, heart disease disrupts the quality of life and creates major health complications in the affected population [2]. In order to effectively prevent and manage the disease, we need to understand common factors that contribute to the disease.

Diagnosis of heart disease comes from multiple angles.Doctors typically use physical exams, electrocardiograms, blood tests, stress tests, cardiac catheterization to understand a patient's condition and diagnose the disease. Our data set included many of the parameters relating to the diagnosis of heart disease. As seen from the data columns, there were 8 specific diagnosis related features in the data set with 2 demographic features. The last column included the diagnosis of heart disease. One of the biggest assumptions in the data set is that except for the demographic features, the other features in the data set are clinically meaningful in relation to heart disease.

In this project, we will be examining multiple variables that may be related to heart disease. We will be examining the correlation between key variables, and try to establish a robust prediction algorithm based on the available sample data set.

Data Variables:

•**Age**: The person's age in years
•**Sex**: The person's sex (1 = male, 0 = female)
•**ChestPainType**:
    0: asymptomatic
    1: atypical angina
    2: non-anginal pain
    3: typical angina
•**RestingBP**: The person's resting blood pressure (mm Hg on admission to the hospital)

•**Cholesterol**: The person's cholesterol measurement in mg/dl
•**FastingBS**: The person's fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)
•**RestingECG**: resting electrocardiographic results
0: showing probable or definite left ventricular hypertrophy by Estes' criteria
1: normal
2: having ST-T wave abnormality
•**MaxHR**: The person's maximum heart rate achieved
•**ExerciseAngina**: Exercise induced angina (1 = yes; 0 = no)
•**Oldpeak**: ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot)
•**ST_Slope**: the slope of the peak exercise ST segment:
0: downsloping; 1: flat; 2: upsloping
• **HeartDisease** (1 = no, 0= yes)

Data Cleaning (Haibi)

1. We first created another version of the dataframe by converting the categorical data types, such as ChestPainType, Exercise Angina, ST_Slope, RestingECG, and sex to nominal values. We also kept the original dataframe, which included the actual name values for categorical variables.

```
RangeIndex: 918 entries, 0 to 917
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Age             918 non-null    int64
 1   Sex             918 non-null    object
 2   ChestPainType   918 non-null    object
 3   RestingBP       918 non-null    int64
 4   Cholesterol     918 non-null    int64
 5   FastingBS       918 non-null    int64
 6   RestingECG      918 non-null    object
 7   MaxHR           918 non-null    int64
 8   ExerciseAngina  918 non-null    object
 9   Oldpeak         918 non-null    float64
 10  ST_Slope        918 non-null    object
 11  HeartDisease    918 non-null    int64
dtypes: float64(1), int64(6), object(5)
```

(Original Dataset)

```
RangeIndex: 918 entries, 0 to 917
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Age             918 non-null    int64
 1   Sex             918 non-null    int64
 2   ChestPainType   918 non-null    int64
 3   RestingBP       918 non-null    int64
 4   Cholesterol     918 non-null    int64
 5   FastingBS       918 non-null    int64
 6   RestingECG      918 non-null    int64
 7   MaxHR           918 non-null    int64
 8   ExerciseAngina  918 non-null    int64
 9   Oldpeak         918 non-null    float64
 10  ST_Slope        918 non-null    int64
 11  HeartDisease    918 non-null    int64
dtypes: float64(1), int64(11)
```

(Cleaned Dataset for Machine Learning)

2. It first appeared that there were no missing values, but as we further looked into the dataset, there were some "0"s in the Cholesterol and RestingBP column.

	Age	RestingBP	Cholesterol	FastingBS	MaxHR	Oldpeak	HeartDisease
count	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000
mean	53.510893	132.396514	198.799564	0.233115	136.809368	0.887364	0.553377
std	9.432617	18.514154	109.384145	0.423046	25.460334	1.066570	0.497414
min	28.000000	0.000000	0.000000	0.000000	60.000000	-2.600000	0.000000
25%	47.000000	120.000000	173.250000	0.000000	120.000000	0.000000	0.000000
50%	54.000000	130.000000	223.000000	0.000000	138.000000	0.600000	1.000000
75%	60.000000	140.000000	267.000000	0.000000	156.000000	1.500000	1.000000
max	77.000000	200.000000	603.000000	1.000000	202.000000	6.200000	1.000000

3. There was only one “0” in the Resting BP column, but there were 172 “0”s in the Cholesterol column. To decide what to do with the “0”s , we split the dataset by Cholesterol = 0 and Cholesterol != 0, and compared the mean values of all other numeric columns, especially the HearDisease (0/1) column. The mean value between the two datasets were pretty similar, so we decided to replace all the “0” with the mean value.

   Similarly, since there was only one missing Resting BP value, we also replaced the “0” with the mean value.

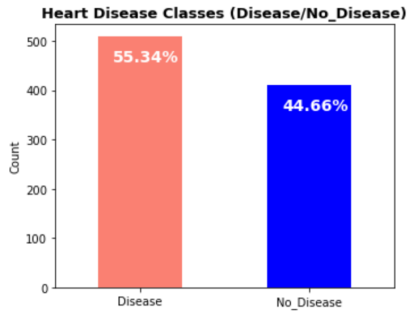
Exploring the Demographic Variables

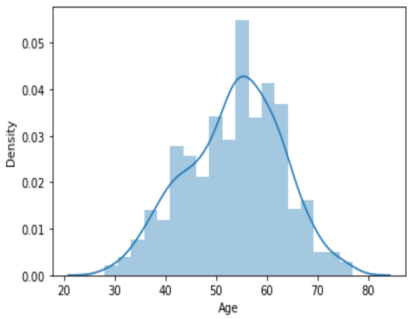
We started exploring some demographic variables in the dataset to understand the data distribution.

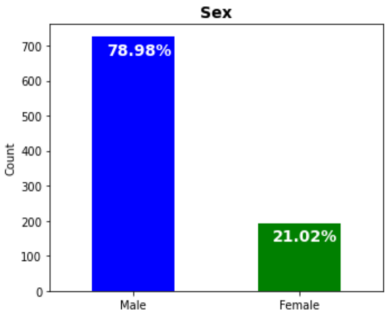
Target variable distribution (Heart Disease): there were more patients with heart disease compared to those without.

Age distribution: Good age representation in dataset with a range of 22 to 77 yrs. Mean age was 54 years.

Sex distribution: Our dataset had more male patient represented than female. Having a disproportionate representation of male and female was a limitation in our prediction model.





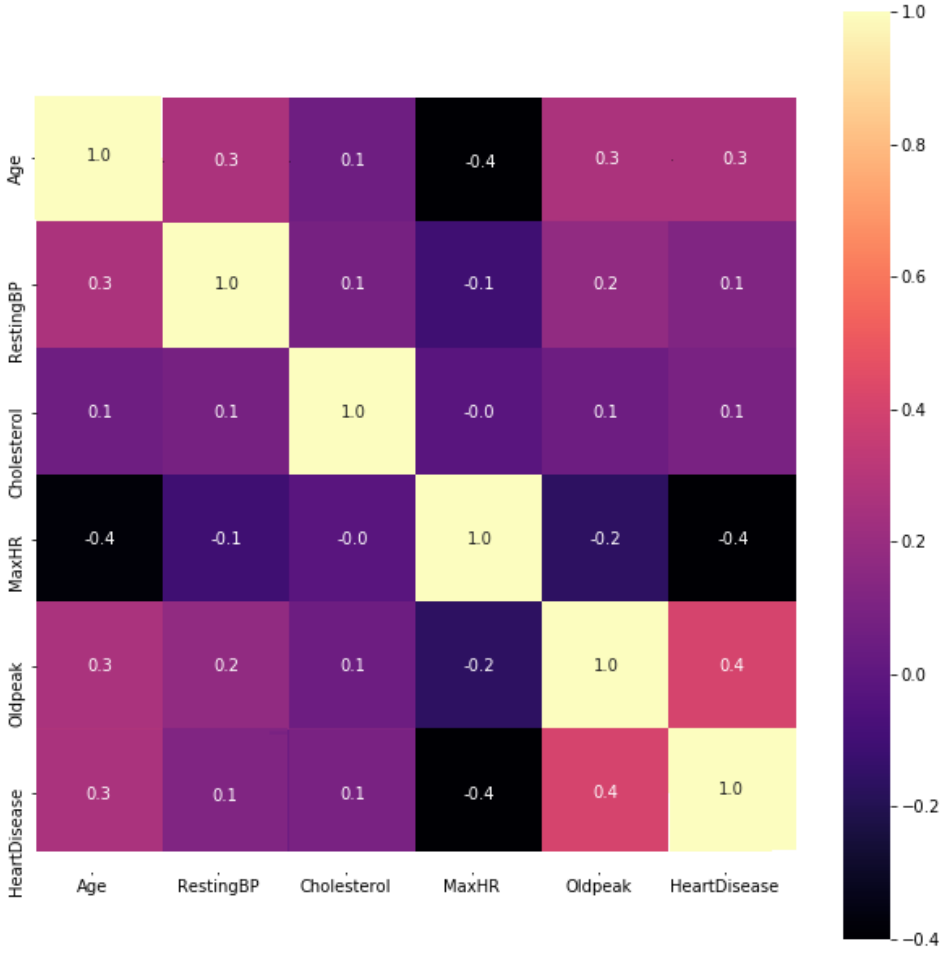


Understanding the Relationship between Variables

Correlation Heatmap for Numeric Variables

As we completed the initial data exploration, we used the seaborn package to explore the correlation between the different variables.

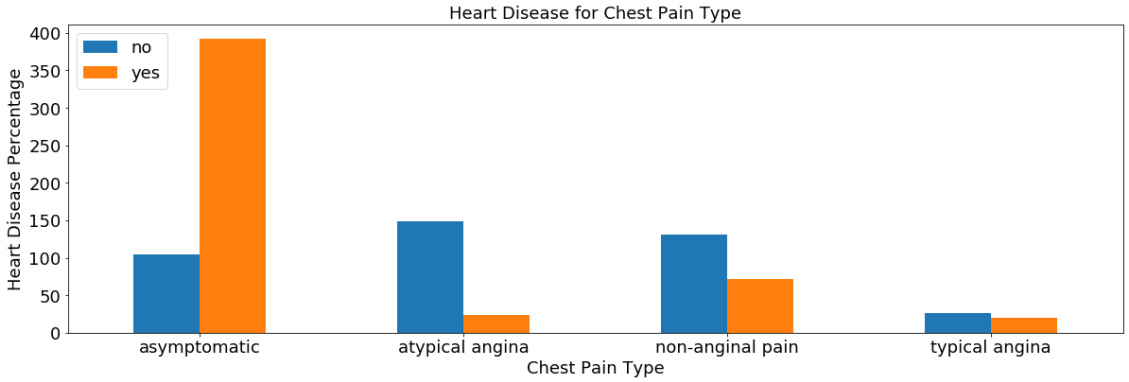
We started with the heatmap function from the seaborn package to take a look at the overall correlation. The initial observations from this chart suggested that oldpeak had the highest positive correlation with HeartDisease. MaxHR had the highest negative correlation.



Based on this observation, we further explored the correlation between these individual variables and HeartDisease. Then we plotted these variables to explore the relationship between each variable type and heart disease.

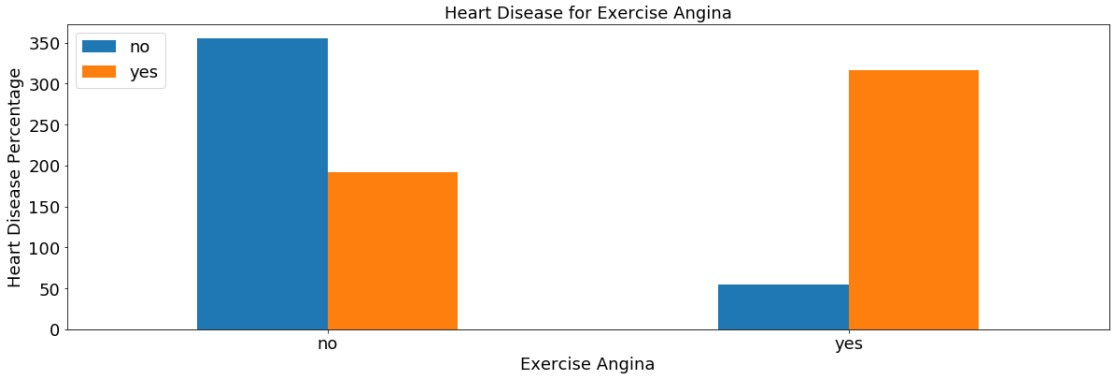
Chest Pain Types vs. Heart Disease

As seen by the plot below, asymptomatic individuals were associated with the highest prevalence of heart disease. This suggested that heart disease can manifest within us without any symptoms, making the condition more dangerous.



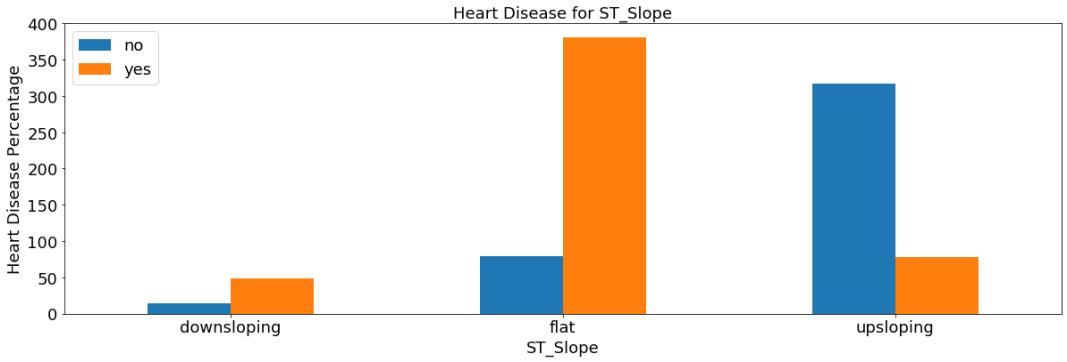
Exercise Angina Vs. Heart Disease

Those who experienced exercise angina showed a higher prevalence of heart disease.



ST Slope Vs. Heart Disease

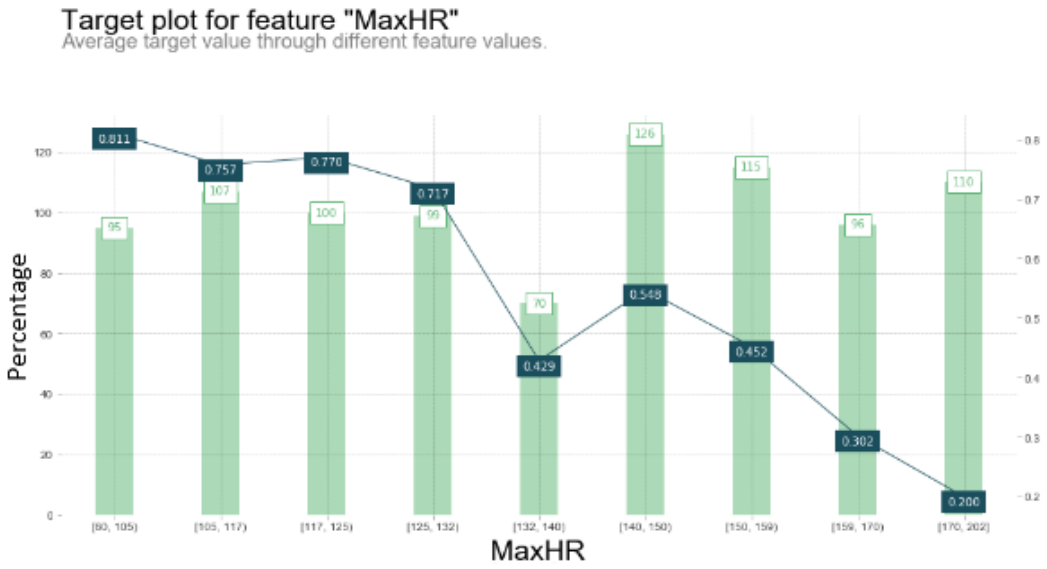
Flat ST_Slope was related to a higher prevalence of heart disease, whereas upsloping ST_Slope seemed to be an indicator of no heart disease.



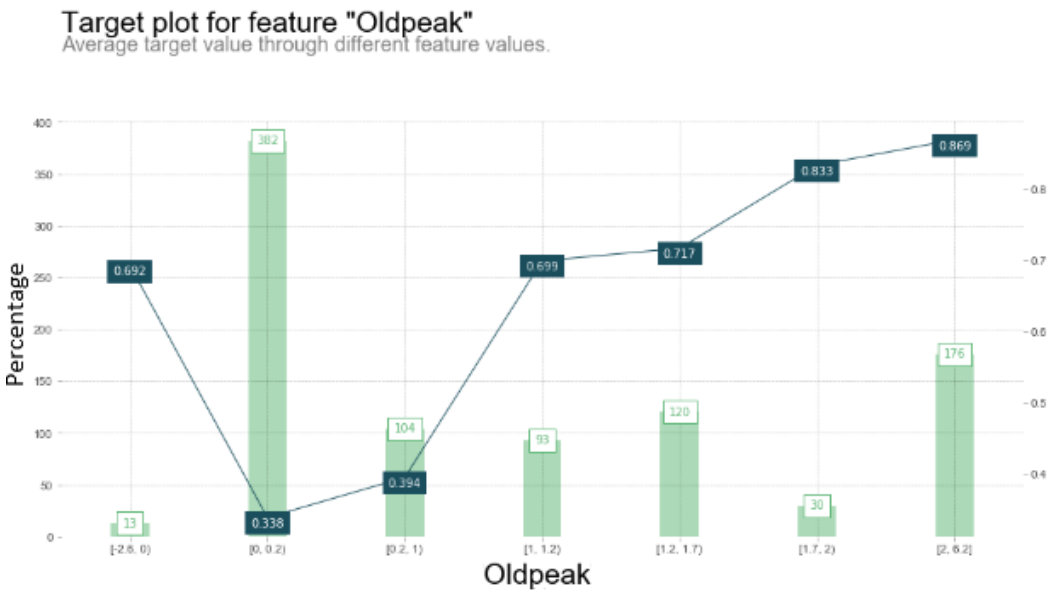
Exploring Striking Continuous Variables (Max HeartRate and OldPeak)

To explore the relationship between MaxHR, Old Peak and Heart Disease, we used the get_dummies function from pandas to deploy the one-hot coding for further data exploration.

First, we took a look at the relationship between Max Heart Rate and Heart Disease. Based on the Graph below, it seemed that the higher MAX Heart Rate, the lower the Heart Disease frequency. This aligned with what the heatmap correlation suggested.



Similarly, we took a look at the oldpeak. The observation from the graph below showed us that higher the oldpeak, higher the frequency of heart disease.



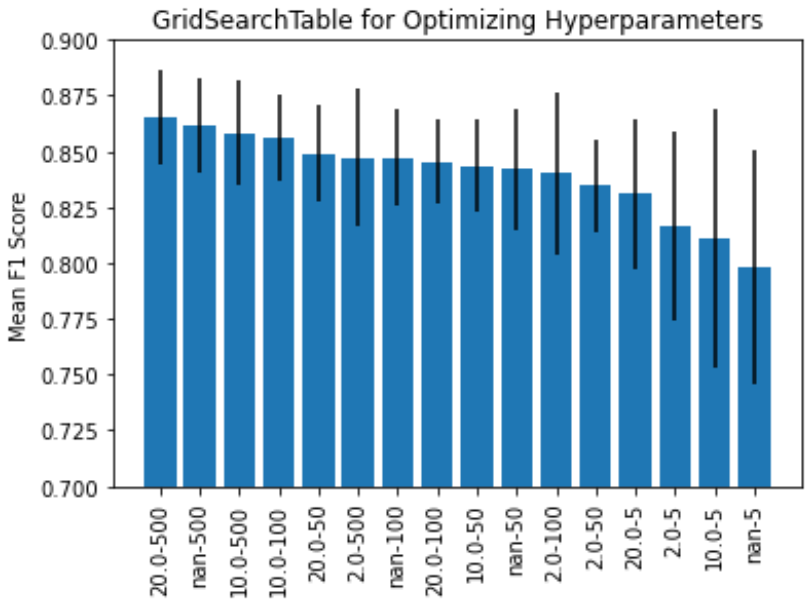
Build Prediction Model for Heart Disease

In order to build a robust prediction model for heart disease, based on the features included in the data set, we preliminarily selected and explored three different Machine Learning models: Logistic Regression, Support Vector Machine, and Random Forest Classifier.

The data set was first split into three parts, in which 60% of the data was used for training, 20% for validation, and 20% for testing. We used a cross validation scoring method to determine the average fit score of the different models. As seen by the table below, the RandomForestClassifier had the highest mean f1 score, and we thus selected the model to further optimize the hyperparameters.

Model Type	CV1	CV2	CV3	CV4	CV5	Mean
RandomForestClassifier	**0.909**	**0.855**	**0.864**	**0.855**	**0.818**	**0.860**
SupportVectorMachine	**0.745**	**0.700**	**0.682**	**0.700**	**0.609**	**0.687**
LogisticRegression	**0.882**	**0.800**	**0.855**	**0.845**	**0.818**	**0.840**

We iterated over n_estimates, the number of trees, and max_depth, the longest path between root node and leaf node of the trees, with GridSearchCV to tune the hyperparameters on the training data set.



We picked the top three results from the GridSearch to validate our training model. The best result in accuracy, precision, and recall was the hyperparameter with max_depth of 10 and n_estimators of 500.

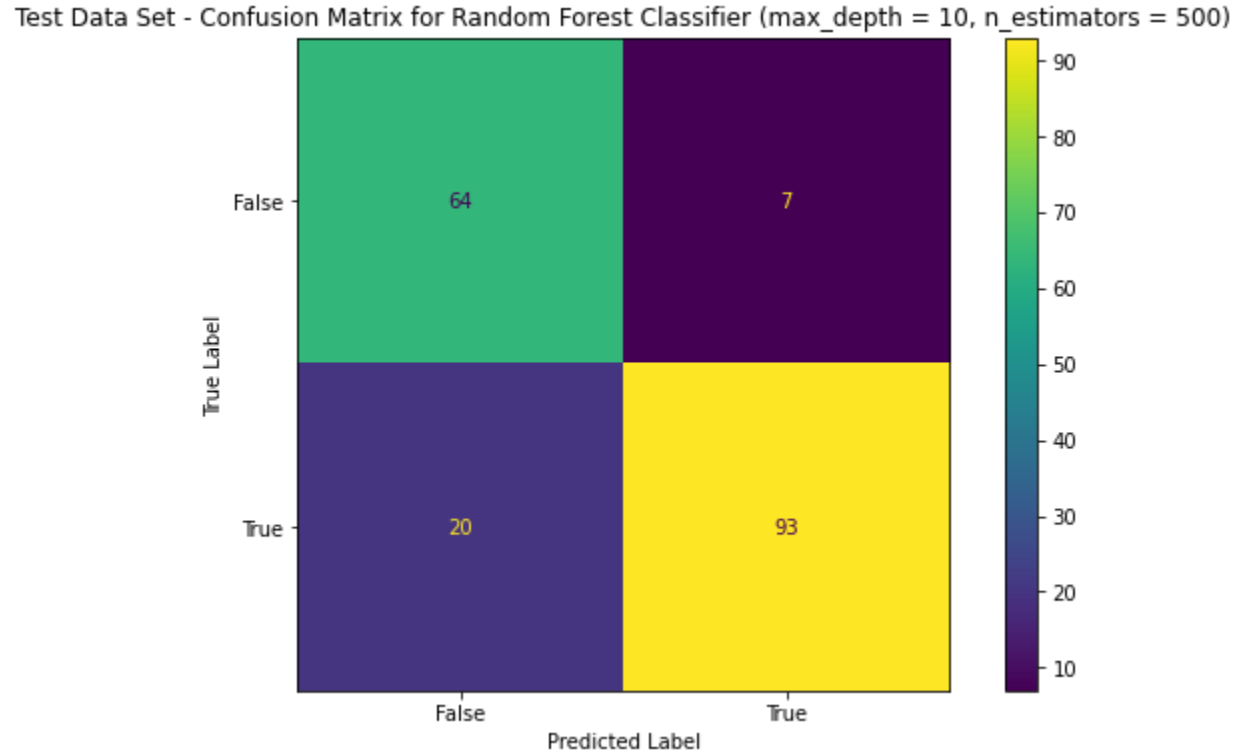
- Accuracy = $\frac{\#\ predicted\ correctly}{total\ \#\ of\ examples}$
- Precision = $\frac{\#\ predicted\ as\ surviving\ that\ actually\ survived}{total\ \#\ predicted\ to\ survive}$
- Recall = $\frac{\#\ predicted\ as\ surviving\ that\ actually\ survived}{total\ \#\ that\ actually\ survived}$

max_depth	n_estimators	Accuracy	Precision	Recall
10.0	500	0.870	0.920	0.852
NaN	500	0.864	0.919	0.843
20.0	500	0.864	0.919	0.843

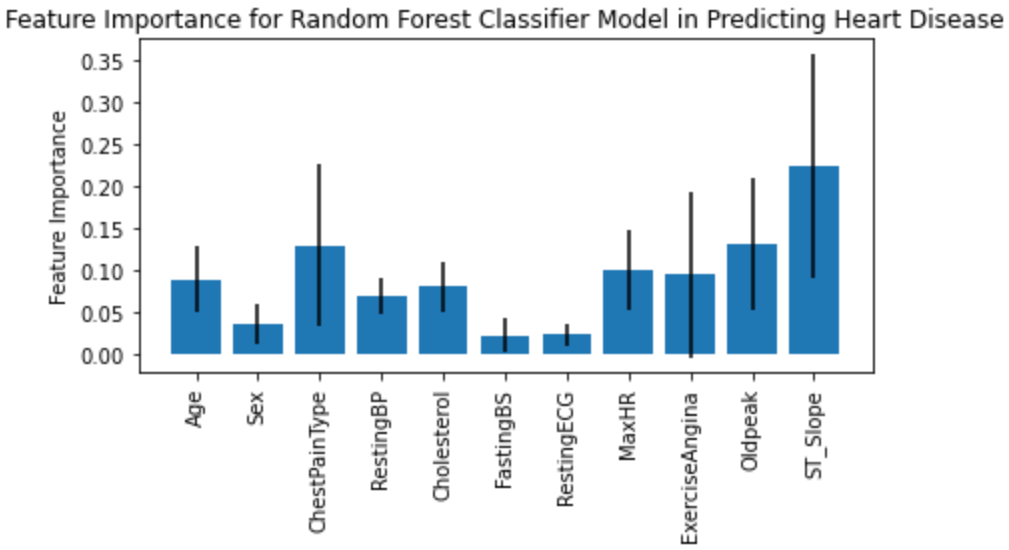
Using the hyperparameters of max_depth of 10 and n_estimators of 500, we evaluated the performance of our model on the test data. The confusion matrix below shows the performance on both the test data set and the total data set provided.

Performance Results on Test Data Set					
max_depth	**n_estimators**	**Accuracy**	**Precision**	**Recall**	**Root Mean Squared Error**
10.0	500	0.853	0.93	0.823	0.383

As seen from the table, the model was accurate 85.3% of the time in predicting heart disease in the test data set. The model was found to have a high positive predictive value (PPV) shown by the high level of precision, which indicates the model's performance in predicting positive values. However, due to the prevalence of false negatives, the model suffered in its recall metrics, suggesting that the model is prone to incorrectly labeling true values as false. This was well visualized using the confusion matrix below, shown by the higher rate of false negatives.



We also evaluated the important features in predicting heart disease. Few outstanding features were ST_slope, ChestPainType, and Oldpeak. The parameters that did not seem to have too much of an effect on the prediction model were Sex, FastingBS, and RestingECG.



Conclusion/Summary

Heart Disease is the major cause of disability and premature death in the world.This damage can be reduced considerably if diagnosed and treated early.Through this project we have predicted top three factors that are related to heart disease : ST_Slope , Old peak and Chest pain type.

We need to further fine-tune our prediction model based on heart disease type and with better datasets. We have considered heart disease as generic, in reality there are multiple types of heart diseases like Blood vessel disease, arrhythmias, congenital heart defects and many more. The prediction model and the contributing factors could differ for a specific heart disease.Our dataset currently has more male patients compared to female patients which affects the prediction model. That's a limitation that will be addressed with a better dataset in future.