Heart Disease Prediction Webapp

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS OF

MINI PROJECT

SEMESTER IV

INFORMATION TECHNOLOGY

BY

Frason Francis (201903020)
Dallas Vaz (201903005)
Steffin Sunnymon (201903048)

UNDER THE GUIDANCE OF
Prof. Sulochana Devi & Prof. JAYCHAND UPADHYAY
(Assistant Professor, Department of Information Technology)

INFORMATION TECHNOLOGY DEPARTMENT
XAVIER INSTITUTE OF ENGINEERING
UNIVERSITY OF MUMBAI (2020-2021)

XAVIER INSTITUTE OF ENGINEERING

MAHIM CAUSEWAY, MAHIM, MUMBAI - 400016.

CERTIFICATE

This to certify that

Frason Francis	(201903020)
Dallas Vaz	(201903005)
Steffin Sunnymon	(201903048)

Have satisfactorily carried out the PROJECT work titled “Heart Disease Prediction Webapp” in partial fulfillment of the Mini project of Sem-4 of Information Technology as laid down by the University of Mumbai during the academic year 2020-2021.

Prof. Chhaya Narvekar	**Prof. Jaychand Upadhyay**
Head of Department	**Supervisor/Guide**

DECLARATION

I declare that this written submission represents my ideas in my own words and where other's Ideas or words have been included, I have adequately cited and referenced the original sources.

I also declare that I have adhered to all the principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission.

I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which thus have not been properly cited or from whom proper permission have not been taken when needed.

Project member-1(XIEIT ID) **201903020**

Project member-2(XIEIT ID) **201903005**

Project member-3(XIEIT ID) **201903048**

Date:

Table of Contents

Abstract:

This report represents the mini-project assigned to fourth-semester students for the partial fulfillment of IT 401, Machine Learning, given by the Department of Information Technology and engineering, MU. Cardiovascular diseases are the most common cause of death worldwide over the last few decades in developed and underdeveloped, and developing countries. Early detection of cardiac diseases and continuous supervision of clinicians can reduce the mortality rate. However, it is not possible to accurately monitor patients every day accurately, and consultation of a patient for 24 hours by a doctor is not available since it requires more patience, time, and expertise. In this project, we have developed and researched models for heart disease prediction through the various heart attributes of a patient and detect impending heart disease using Machine learning techniques like Support vector classifier, logistic regression, Random Forest and K-Nearest Neighbour on the dataset available publicly in Kaggle Website, further evaluating optimizing the results using a voting classifier. The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high-risk patients and, in turn, reduce the complications, which can be a significant milestone in the field of medicine.

Introduction:

According to the World Health Organization, every year, 12 million deaths occur worldwide due to Heart Disease. The load of cardiovascular disease has been rapidly increasing all over the world for the past few years. Many kinds of research have been conducted to pinpoint the most influential factors of heart disease and accurately predict the overall risk. Heart Disease is even highlighted as a silent killer, leading to the person's death without apparent symptoms. The early diagnosis of heart disease plays a vital role in making decisions on lifestyle changes in high-risk patients and reducing the complications. This project aims to predict future Heart Disease by analyzing patients' data, which classifies whether they have heart disease or not using machine-learning algorithms.

Problem Definition:

The major challenge in heart disease is its detection. There are instruments available that can predict heart disease, but they are expensive or not efficient to calculate the chance of heart disease in humans. Early detection of cardiac diseases can decrease the mortality rate and overall complications. However, it is impossible to monitor patients every day accurately, and consultation of a patient 24 hours by a doctor is not available since it requires more patience, time, and expertise. Since we have a good amount of data in today's world, we can use various machine learning algorithms to analyze the data for hidden patterns. The hidden patterns can be used for health diagnosis in medicinal data.

Motivation:

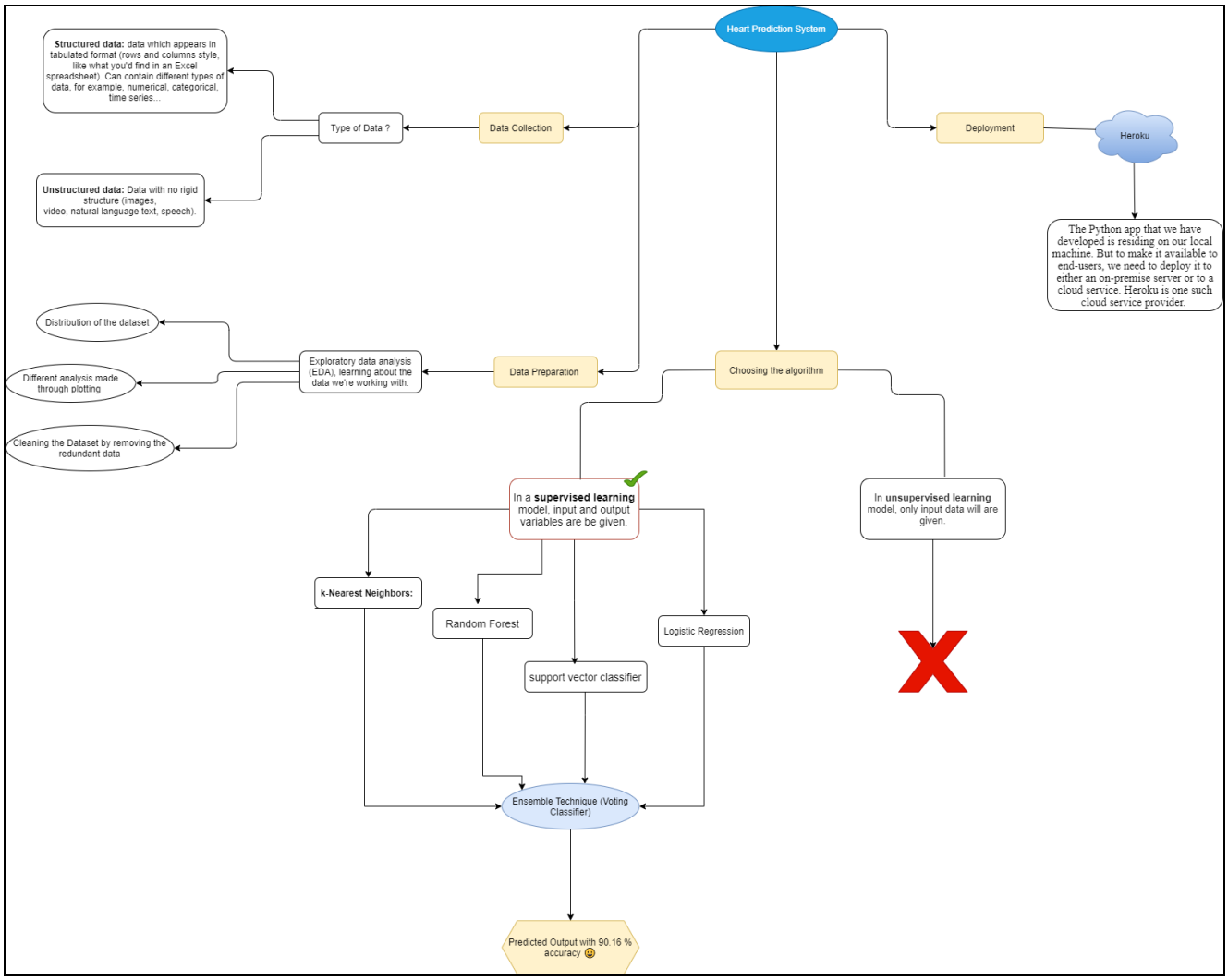
Machine learning techniques have been around us and have been compared and used to analyze many kinds of data science applications. The primary motivation behind this research-based project was to explore the feature selection methods, data preparation, and processing behind the training models in machine learning. With first-hand models and libraries, the challenge we face today is data were beside on their abundance. With our models, the accuracy we see during training, testing, and actual validation has a higher variance. Hence, this project is carried out to explore behind the models and further implement the Logistic Regression model to train the obtained data. Furthermore, as the whole machine learning is motivated to develop an appropriate computer-based system and decision support that can aid to early detection of heart disease, in this project, we have developed a model which classifies if a patient will have heart disease in ten years or not based on various features (i.e., potential risk factors that can cause heart disease) using logistic regression and other algorithms. Hence, the early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high-risk patients and, in turn, reduce the complications, which can be a significant milestone in medicine.

Objectives:

The main objective of developing this project are:

1. To develop a machine learning model to predict future possibility of heart disease by implementing various ML models.
2. To determine significant risk factors based on medical dataset which may lead to heart disease.
3. To analyze selection methods and understand their working principle.
4. To deploy the model on a cloud platform for further use cases.

Heart Prediction Web-App Flowchart:



[Link]

Dataset Information:

The dataset is publicly available on the Kaggle Website from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. It provides patient information, which includes over 300 records and 14 attributes. The attributes include are age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting, sugar blood, resting electrocardiographic results, maximum heart rate, exercise-induced angina, ST depression induced by exercise, the slope of the peak exercise, number of major vessels, and target range is either 0 to1, where 0 is an absence of heart disease. The data set is in CSV (Comma Separated Value) format which is further prepared to data frame as supported by pandas library in python.

```
In [3]: print("Number of rows in data :", data.shape[0])
        print("Number of columns in data :", data.shape[1])

        Number of rows in data : 303
        Number of columns in data : 14

In [4]: data.head()
```

Out[4]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Figure 1: Data Description

METHODS AND ALGORITHMS USED

The primary purpose of designing this system is used to predict the risk of future heart disease. We have used various types as a machine-learning algorithm to train our system and different ensemble techniques to enhance our model to give the most accurate and the best result.

Exploratory Data Analysis

Exploratory data analysis (EDA) is an essential step in any research analysis. The primary aim of the exploratory analysis is to examine the data for distribution, outliers, and anomalies to direct specific testing of your hypothesis. It also provides tools for hypothesis generation by visualizing and understanding the data, usually through graphical representation. At its core, EDA is more of an attitude than it is a step-by-step process. Exploring data with an open mind tends to reveal its underlying nature far more readily than making assumptions about the rules we think (or want) it to adhere to. The dataset provided by kaggle doesn't contain any categorical variable and null values which is a really good thing as we will not have to invest a lot of time on cleaning the data and we can focus more on predictive modelling. Also, the target here is our target vector and all the other are predictors. We have described our dataset explicitly ('age', 'sex', 'chest_pain_type',

'resting_blood_pressure', 'serum_cholesterol', 'fasting_blood_sugar', 'rest_ecg', 'max_heart_rate', 'exercise_angina', 'st_depression', 'st_slope', 'num_major_vessels', 'thalassemia', 'target') to relate to the output variable.

Analysis of associations of different column

Correlation Plot

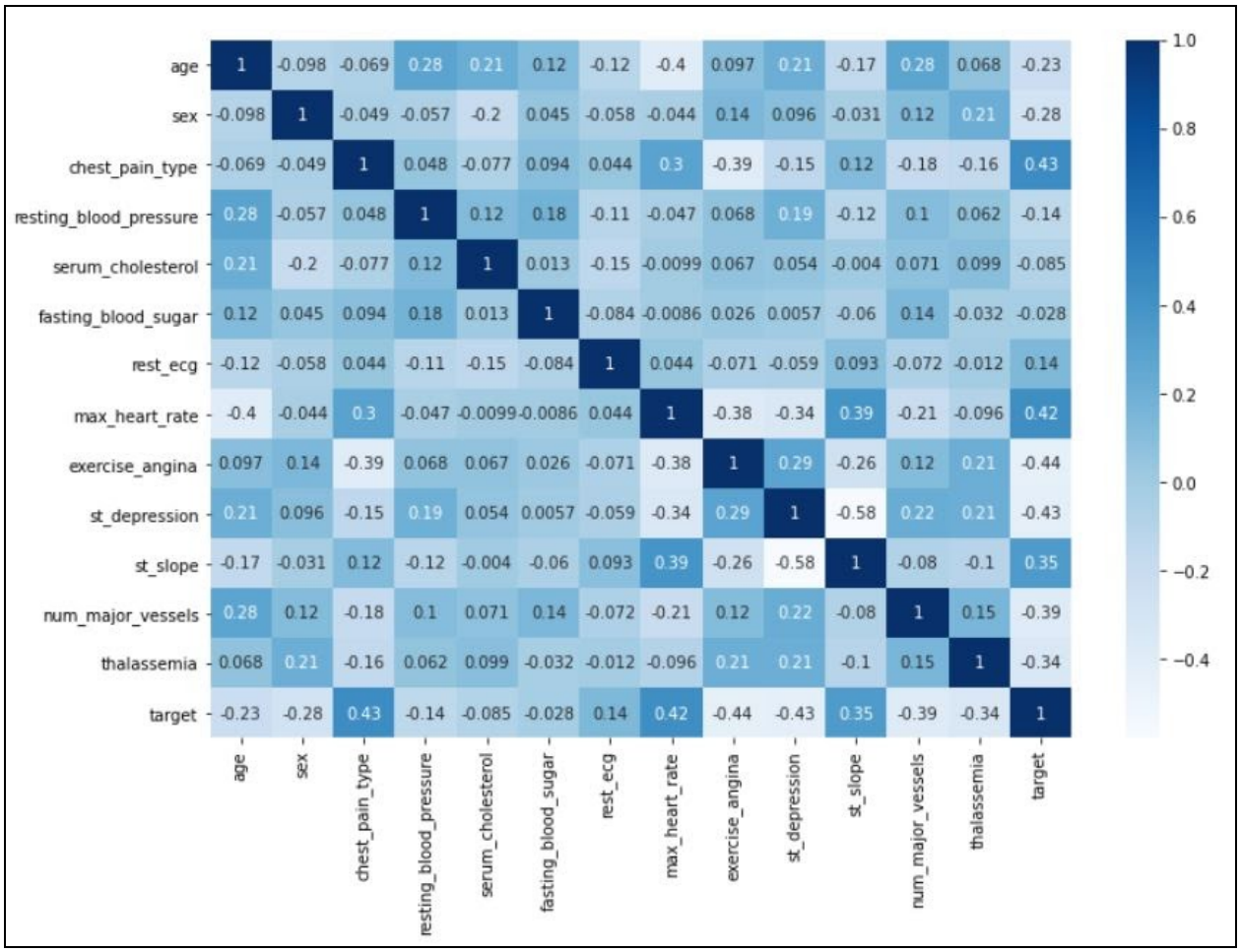


Figure 2: Correlation Plot of Heart prediction Data

The correlation graph shows us that there is a fair correlation between Max_heart_rate, st_slope and chest_pain_type.

AGE DISTRIBUTION:

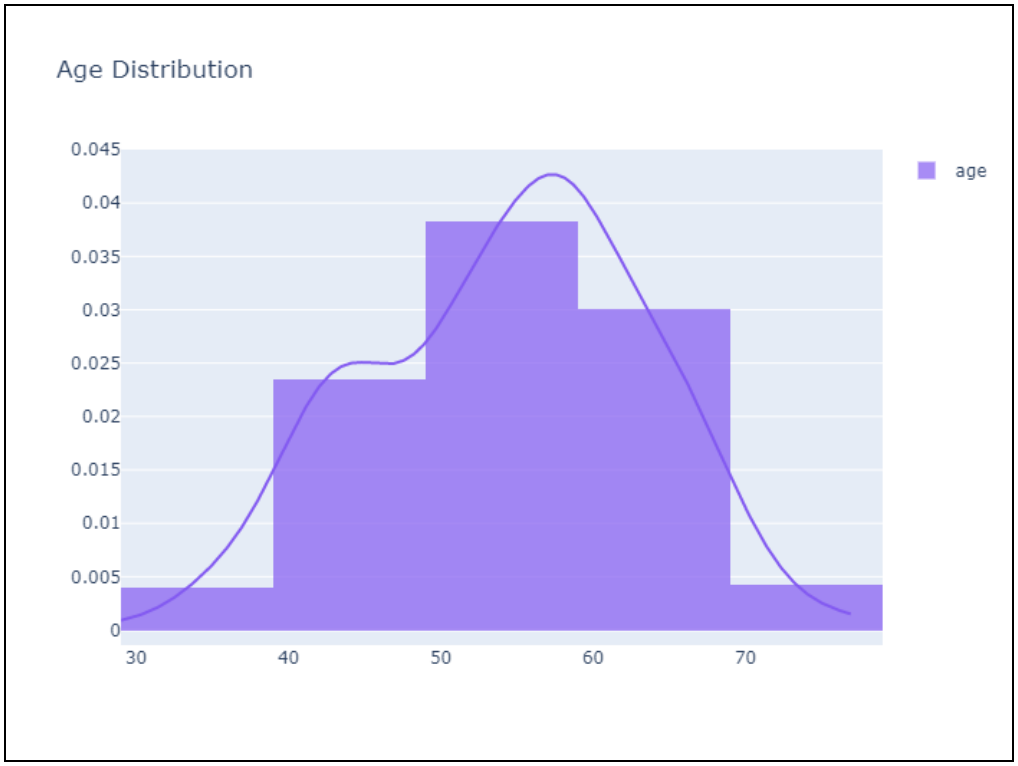


Figure 3: Age Distribution Curve

Observation:

Distribution of age looks close to a Normal Distribution. This data has the age of patients ranging from 29-77 which is good as the data is not biased towards certain kinds of patients. There are 16 young patients, 125 middle aged patients and 151 old aged patients. 16 young patients is quite obvious because heart disease is not very common in the younger population but we have almost a large number of patients in middle age and Old age which is also obvious. (see figure 4 for more detail)

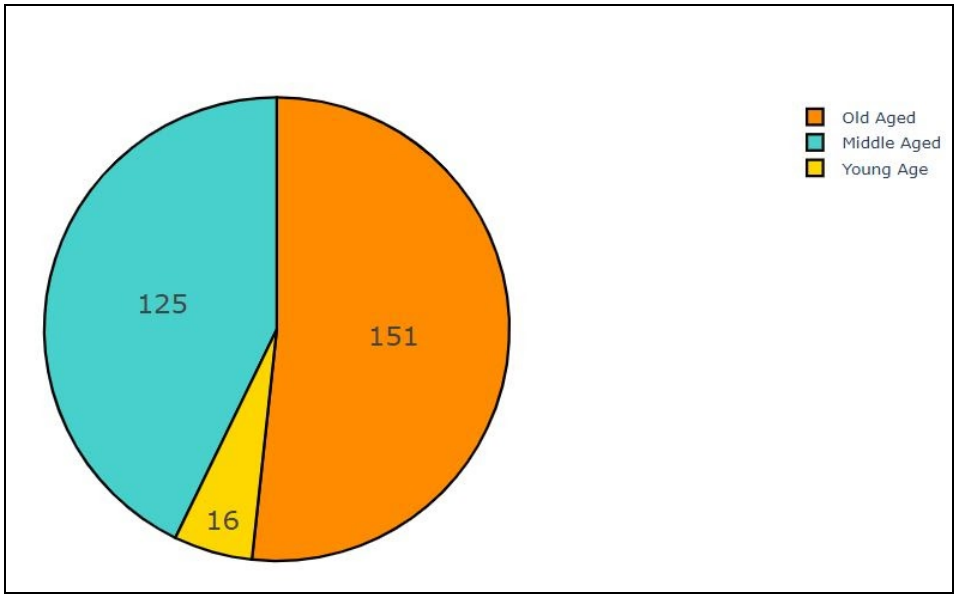


Figure 4: Age Distribution Pie-chart

Resting Blood Pressure:

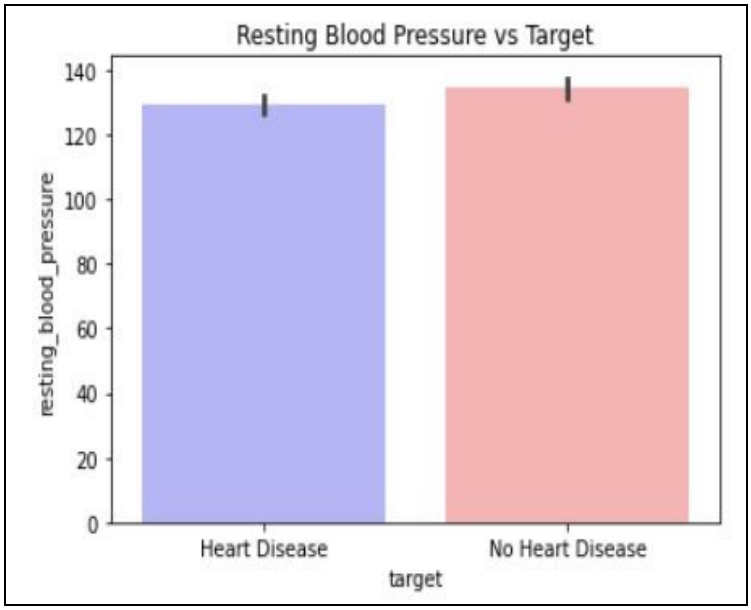


Figure 5: Resting Blood Pressure vs Target

Observation:

- The distribution of resting blood pressure is close to normal distribution.
- The resting blood pressure of both Heart patients and Non-heart patients is almost same.
- Median Resting Blood Pressure with Heart Disease - Male (130) and Female (130)
- Median Resting Blood Pressure without Heart Disease - Male (130) and Female (140)

Thalassemia Analysis

Thalassemia (thal-uh-SEE-me-uh) is an inherited blood disorder that causes your body to have less hemoglobin than normal. Hemoglobin enables red blood cells to carry oxygen. Thalassemia can cause anemia, leaving you fatigued.

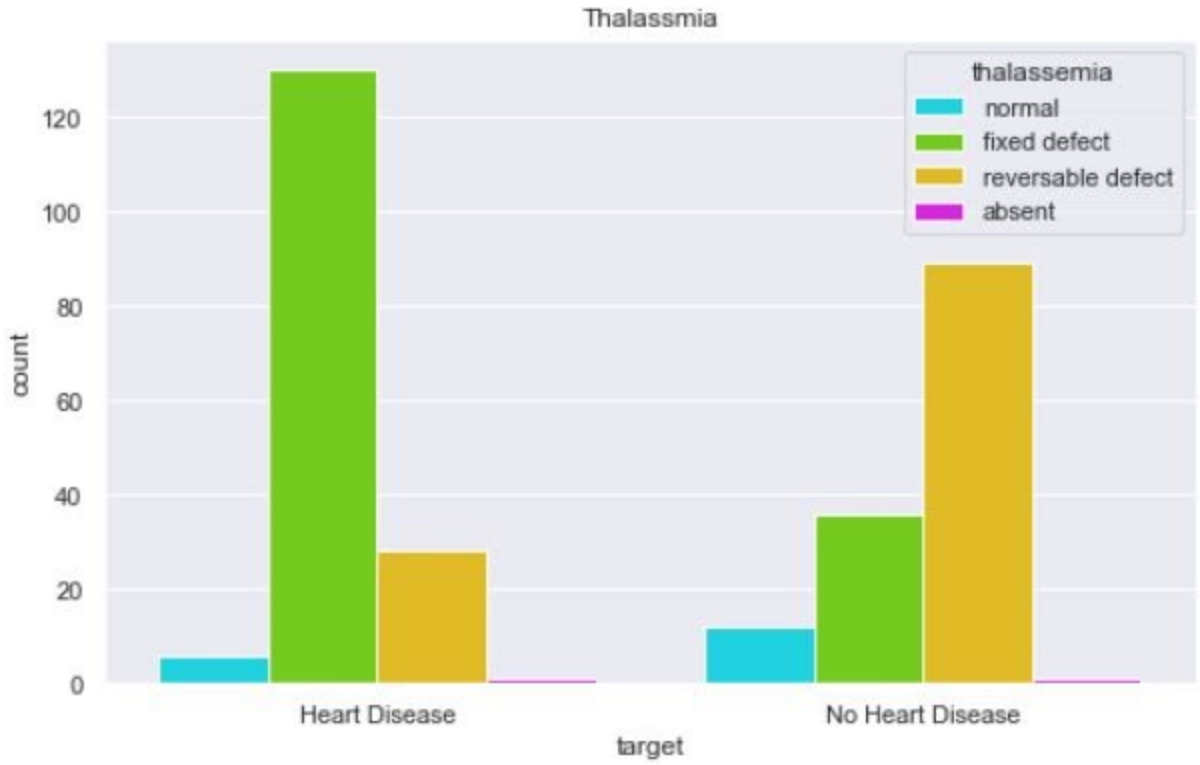
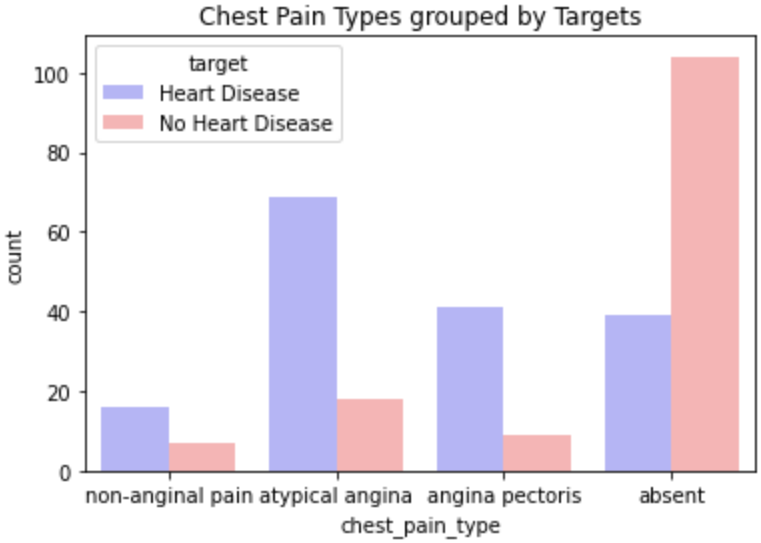


Figure 6: Thalassemia types based on target

Observation:

- Almost every patient, having heart disease or not, has Thalassemia.
- Fixed Defect Thalassemia is more common in Patients with Heart Disease whereas Reversible Defect Thalassemia is more common in Patients without Heart Disease.
- There are more patients with Fixed Defect Thalassemia in any type of chest pains whereas patients without chest pain and Reversible Defect Thalassemia are very common.
- Fixed Defect Thalassemia is very common among Male and Female Patients but there are a very large number of Reversible Defect Thalassemia Male Patients as compared to Female Patients.
- Among Horizontal Slope Patients - Fixed Defect Thalassemia is more common as compared to others whereas Upsloping patients have more number of Reversible Defect Thalassemia Patients.

Chest Pain Type Analysis



Observation

- The most common type of chest pain among heart disease patients is "atypical agina" followed by "agina pectoris". Also, we can see around 40 of the patients don't have chest pain but still have heart disease so absence of chest pain does not guarantee that the patient being diagnosed has no Heart Disease.
- There are more male patients with no chest pain as compared to females.

Machine Modelling:

Logistic Regression:

Logistic regression is one such regression algorithm which can be used for performing classification problems. It calculates the probability that a given value belongs to a specific class. If the probability is more than 50%, it assigns the value in that particular class; if the probability is less than 50%, the value is assigned to the other class. Therefore, we can say that logistic regression acts as a binary classifier.

Working of a Logistic Model

For linear regression, the model is defined by:

$$y=\beta 0+\beta 1x \qquad \text{- (i)}$$

and for logistic regression, we calculate probability, i.e. y is the probability of a given variable x belonging to a certain class. Thus, it is obvious that the value of y should lie between 0 and 1.

But, when we use equation (i) to calculate probability, we would get values less than 0 as well as greater than 1. That doesn't make any sense . So, we need to use such an equation which always gives values between 0 and 1, as we desire while calculating the probability.

Advantages of Logistic Regression

- It is very simple and easy to implement.
- The output is more informative than other classification algorithms
- It expresses the relationship between independent and dependent variables
- Very effective with linearly separable data

Disadvantages of Logistic Regression

- Not effective with data which are not linearly separable
- Not as powerful as other classification models
- Multiclass classifications are much easier to do with other algorithms than logistic regression
- It can only predict categorical outcomes

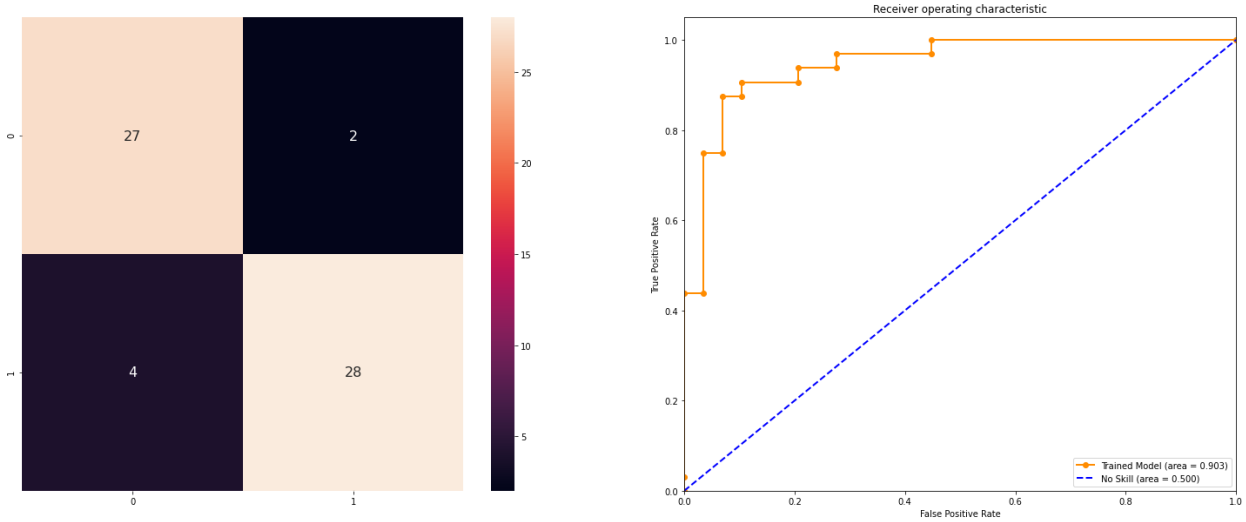


Figure 7: Show's Correlation Matrix and roc_auc_curve

Logistic Regression seems to work well as it is easily able to draw hyperplanes and the reason for that can be that patients with higher ages usually have the problem of high blood pressure thus making it easy to separate it from the low age patients.

Random Forest

Random Forest Classifier is an ensemble algorithm. Ensemble algorithms are those which combine more than one algorithm of the same or different kind for classifying objects. Random forest classifier creates a set of decision trees from randomly selected subsets of the training set. It then aggregates the votes from different decision trees to decide the final class of the test object. Basic parameters to Random Forest Classifier can be the total number of trees to be generated and decision tree related parameters like minimum split, split criteria etc. We use 3 fold cross validation which search across 100 different combinations, and use all available cores to find the best parameter for our model. After getting the best parameter we create an instance of model and fit the parameters for the best accuracy (refer to figure 8)

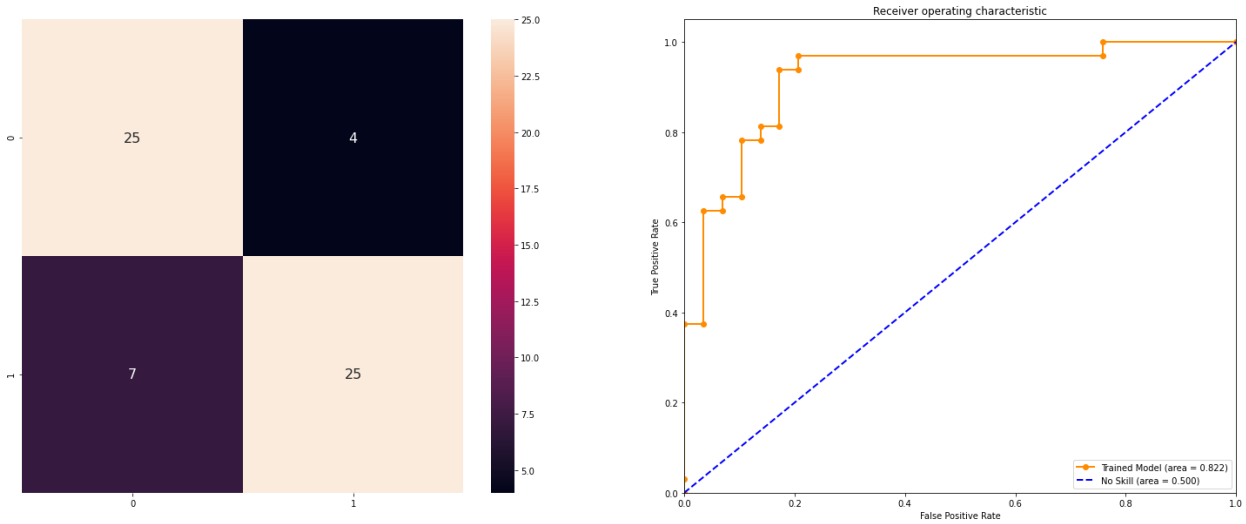
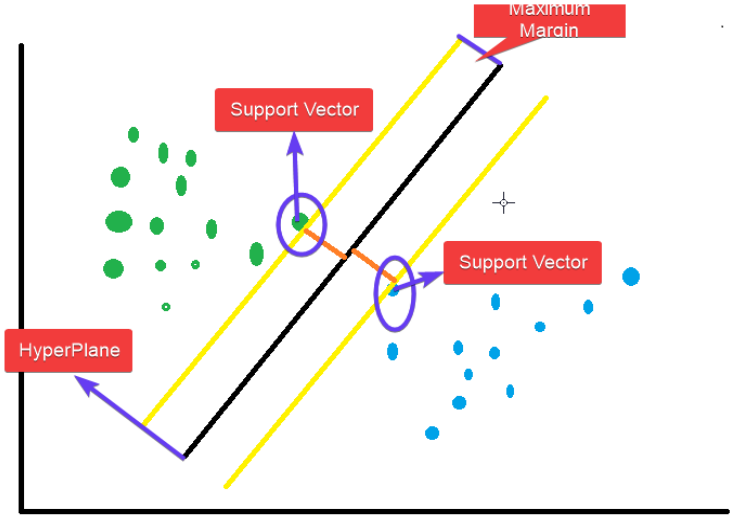


Figure 8: Show's Correlation Matrix and roc_auc_curve

Support Vector Classifier:

Support Vector Machine is a supervised Machine Learning algorithm widely used for solving different machine learning problems. Given a dataset, the algorithm tries to divide the data using hyperplanes and then makes the predictions. SVM is a non-probabilistic linear classifier. While other classifiers, when classifying, predict the probability of a data point to belong to one group or the another, SVM directly says to which group the data point belongs to without using any probability calculation.



The black line in the middle is the optimum classifier. This line is drawn to maximise the distance of the classifier line from the nearest points in the two classes. It is also called a hyperplane in terms of SVM. A Hyperplane is an n-1 dimensional plane which optimally divides the data of n dimensions. Here, as we have only a 2-D data, so the hyperplane can be represented using one dimension only. Hence, the hyperplane is a line here. The two points (highlighted with circles) which are on the yellow lines, are called the support vectors. As it is a 2-D figure, they are points. In a multi-dimensional space, they will be vectors, and hence, the name- support vector machine as the algorithm creates the optimum classification line by maximising its distance from the two support vectors. After applying the model the accuracy we get on the test set is 88.5%.

K-Nearest Neighbors

K-nearest neighbors (KNN) is a type of supervised learning algorithm which is used for both regression and classification purposes, but mostly it is used for the later. Given a dataset with different classes, KNN tries to predict the correct class of test data by calculating the distance between the test data and all the training points. It then selects the k points which are closest to the test data. Once the points are selected, the algorithm calculates the probability (in case of classification) of the test point belonging to the classes of the k training points and the class with the highest probability is selected. We got our default value of K as 27, to get the accuracy score of the matrix we apply KFoldCV. This is an important step because if our Hyperplane is giving same accuracy for a wide range of values, it's better to do early stopping and use the lower value which will avoid underfitting or overfitting of the model. (refer to figure 9)

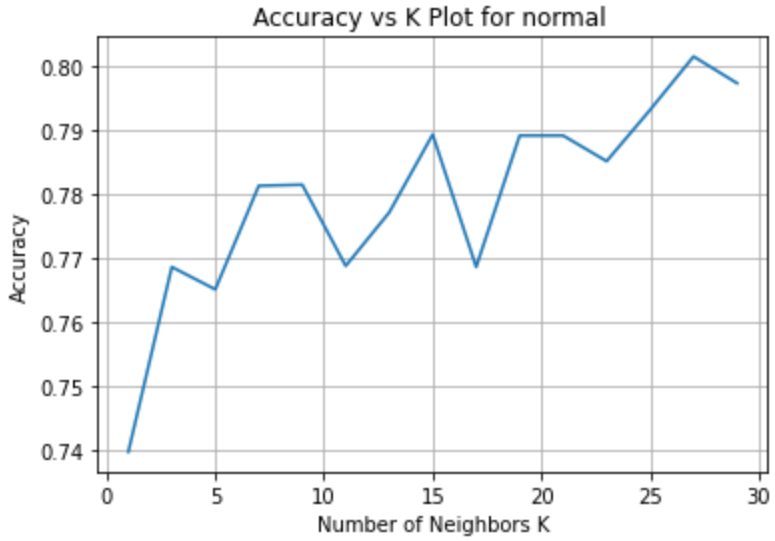


Figure 9: Shows the plot against accuracy and KNN

Ensemble Models - Voting Classifier

This type of ensemble works as an extension of bagging (e.g. RF). The architecture of the voting classifier is based on the 'n' number of ML models, Whose prediction values are in two different ways: hard and soft. In hard, the winning prediction is the one with the most votes. In soft mode, it considers the probability thrown by each ML model; this probability will be weighted and average; consequently, the winning class will be the one with the highest weight and averaged probability. The accuracy on test set using voting classifier is 90.2% (refer figure 10)

```
# Let's look at our accuracy
acc = voting_clf.score(X_test,y_test)


print(f"The accuracy on test set using Logistic Regression is: {np.round(accuracy1, 3)*100.0}%")
print(f"The accuracy on test set using KNN for optimal K = {optimal_k} is {np.round(accuracy2, 3)}%")
print(f"The accuracy on test set using SVC is: {np.round(accuracy3, 3)*100.0}%")
print(f"The accuracy on test set using RandomForest is: {np.round(accuracy4, 3)*100.0}%")
print(f"The accuracy on test set using voting classifier is {np.round(acc, 3)*100}%")
```

```
The accuracy on test set using Logistic Regression is: 90.2%
The accuracy on test set using KNN for optimal K = 27 is 85.246%
The accuracy on test set using SVC is: 88.5%
The accuracy on test set using RandomForest is: 86.9%
The accuracy on test set using voting classifier is 90.2%
```

Figure 10: Compare the accuracy of different models

Training and Testing

Finally, this resulting data split into 80% train and 20% test data, which was further passed to the ensemble technique voting classifier which confirmed that the logistic regression model gave us the highest accuracy to fit, predict and store the model for further prediction.

Result

When performing various methods of feature selection, testing it was found that logistic regression gave us the best result among others. The various methods tried were Random Forest, Decision Tree and KNN. The accuracy that was seen in them ranged around 85.24% to 90.2% being the highest. The accuracy metric for KNN classifier can be seen below.

Precision	0.87
Recall	0.93
f1-score	0.90

Future Work and Conclusion

We have summarized different types of machine learning algorithms for the prediction of heart disease. We elaborated on various machine learning algorithms and worked towards finding the best algorithm by analyzing their features. Every algorithm has given different results in different situations. Further, it is analyzed that only marginal accuracy is achieved for the predictive model of heart disease, and hence more complex models are needed to increase the accuracy of predicting early heart disease. In the future, we will propose a methodology for the early prediction of heart disease with high accuracy and minimum cost and complexity.

Contributions made towards the project

Task \ Members	Frason Francis	Dallas Vaz	Steffin Sunnymon
Data Selection			
Data Cleaning			
EDA			
Building Model			
Result analysis and Accuracy Test			
Web Services			
Documentation			