

PREDICTION OF HEART DISEASE USING MACHINE LEARNING ALGORITHMS

INTERDISCIPLINARY PROJECT (S23AIPROJ)

*Submitted in partial fulfillment of the requirements for the
award of
Bachelor of Technology degree in Biotechnology*

By

DEEKSHA R.V. (39230010)



DEPARTMENT OF BIOTECHNOLOGY

SCHOOL OF BIO AND CHEMICAL ENGINEERING

SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY

JEPPIAAR NAGAR, CHENNAI – 119, INDIA

April 2022



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE

www.sathyabama.ac.in

DEPARTMENT OF BIOTECHNOLOGY

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **DEEKSHA R.V (39230010)** who carried out the project entitled "**PREDICTION OF HEART DISEASE USING MACHINE LEARNING ALGORITHMS**" under my /our supervision from _____ to _____

Internal Guide

Dr. PRAKASH P.

External Guide

Dr. P. AJITHA

Head of the Department

Dr. RAMESH KUMAR V

Submitted for Viva voce Examination held on 29th April

Internal Examiner

Dr. PRAKASH P

External Examiner

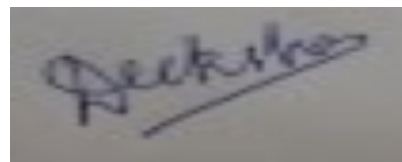
Dr. P. AJITHA

DECLARATION

I, DEEKSHA R.V hereby declare that the Project Report entitled PREDICTION OF HEART DISEASE USING MACHINE LEARNING ALGORITHMS done by me under the guidance of Dr. PRAKASH P (Internal) and Dr. P. AJITHA (External) at SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY, JEPPIAAR NAGAR, CHENNAI-119 is submitted in partial fulfillment of the requirements for the award of Bachelor of Technology degree in Biotechnology.

DATE: 29th April, 2022

PLACE: SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY, CHENNAI

A rectangular box containing a handwritten signature in dark ink. The signature appears to be 'Deeksha' with a horizontal line underneath it.

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management of SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to Dr. -----., **Dean, School of BIO and CHEMICAL ENGINEERING** and **Dr. RAMESH KUMAR V, Head of the Department, Dept. of BIOTECHNOLOGY** for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my **Project Guides Dr. PRAKASH P and Dr. P AJITHA** for their valuable guidance, suggestions and constant encouragement that paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Biotechnology** who were helpful in many ways for the completion of the project.

ABSTRACT:

The cardiovascular, or circulatory, system supplies the body with blood. It consists of the heart, arteries, veins, and capillaries. Any dysfunction of the above could lead to potential cardiovascular diseases or CVD's as they are referred to nowadays. CVD's are now the most common cause of death and generally require quick indicators to predict the onset of a heart-attack or any other related disease. The WHO estimates that by 2030, 23.6 million people will die from CVD conditions annually, mostly due to stroke and heart disease.

According to 2019 worldwide statistics, 18.6 million people died of cardiovascular disease, globally, marking a 17.1% increase in the number of cases over the last decade.

CVD indicators use Machine learning techniques and Deep learning and have come up a big way in terms of providing time-bound indications of defective human hearts around the medical world.

This interdisciplinary project focuses on using big data analysis and machine learning tools to build an efficient prediction algorithm that can predict the occurrence/non-occurrence of heart disease.

Keywords: CVD, heart disease, machine learning, data analytics, big data, deep learning

LIST OF FIGURES AND TABLES:

Figure 4.1.	6
Figure 4.2.	9
Figure 4.3.	10
Figure 4.4.	11
Figure 4.5.	12
Figure 4.6.	13
Figure 4.7.	14
Table 4.1.	15
Table 5.1.	20
Figure 5.1.	21
Figure 5.2.	21
Figure 5.3.	22
Table 5.2.	23
Table 5.3.	23

LIST OF ABBREVIATIONS

CVD Cardiovascular Disease

SVM Support Vector Machine

RF. Random Forest

KNN K-Nearest Neighbors

LDA Linear Discriminant Analysis

QDA Quadratic Discriminant Analysis

TABLE OF CONTENTS

ABSTRACT	iii
LIST OF FIGURES AND TABLES	iv
LIST OF ABBREVIATIONS	v
CHAPTER 1.	1
Introduction	1
CHAPTER 2.	3
Literature Survey	3
CHAPTER 3.	5
3.1 Aim of the Project.	5
3.2 Scope of the Project.	5
CHAPTER 4.	6
4.1 Collection and Importing of Data.	6
4.2 Description of Data.	7
4.3 Pre-Processing and Modification of Data.	9
4.4 Exploratory Data Analysis	10
4.5 Building the Algorithm.	15
4.5.1 Individual Model Testing	16
4.5.2 Ensemble Model Testing	19
CHAPTER 5	20
Results and Discussion	20
CHAPTER 6	24
Conclusions	24
REFERENCES	25

CHAPTER 1

INTRODUCTION

A prediction algorithm is a series of building blocks that are assembled together to make a prediction, or a forecast for an unknown event. Often, prediction algorithms are created to understand how an unknown variable changes with respect to certain known variables, related to the former. This is where the role of Regression steps in. Regression is a powerful statistical tool that helps understand the trends in data and how each variables correlates to the other. This makes predictions easier and more reliable. The most common form of regression used to make prediction algorithms is "Linear Regression". Although it is the easiest tool available, it has its own limitations; linear regression assumes that the dependent and independent variables vary linearly, and it is sensitive to outliers, which is not always a good thing. Hence the algorithm built in this project was created beyond Linear Regression, to get a better accuracy. The following machine-learning models were used for this project:

1. Linear Discriminant Analysis - LDA is typically used when the response variable contains 2 or more classes. It is mainly preferred when the sample size is small, so we don't necessarily have to supply large amounts of data for LDA to make a prediction, making it very handy. Another important point to note is that, LDA assumes that the input data is *normally distributed*. Hence it is always the best idea to normalize/scale our data before using this model.

2. Quadratic Discriminant Analysis - QDA is a variation of LDA, mainly differing such that it separates data in a non-linear manner unlike LDA. The main drawback of this model is that, unlike LDA, it cannot be used as a technique to reduce the dimensions of our data.

3. Logistic Regression - Just like Linear Regression, Logistic regression is a useful statistical tool to predict binary outcomes. It works best when there are no outliers present in the data. Moreover, the output provided by Logistic regression is discrete in nature.

4. Loess - expanded as Locally Weighted Scatterplot Smoothing, the Loess model is yet another tool used in regression analysis for understanding the relationship between variables. Scatterplots are sometimes difficult to interpret due to the presence of some noisy data points and outliers. A loess model helps smooth out data points by fitting a line to the scatterplot and hence shows a clear trend in the data.

5. K-Nearest Neighbors - The KNN model is a distance-based model that computes the distance between the observations of the features, and uses that to relate our dependent and independent variables. The value of "k" selected, controls the accuracy of our estimate; larger values of "k" result in smooth estimates, whereas smaller values of "k" result in flexible estimates. The value of "k" is user-defined.

6. Random Forest - Random Forest model is an improved version of decision trees. The goal of a random forest model is to improve the prediction accuracy by considering the average of multiple decision trees. It has certain parameters which can be tuned by the user, in order to get better accuracy.

7. Support Vector Machines – SVM is the most effective model used for classifying data. It works by bringing the data points into a higher dimensional plane, and then separates the points using a boundary called “hyper plane” that has the maximum distance between the data points of different classes.

CHAPTER 2

LITERATURE SURVEY:

Prediction of heart attack using machine learning [1] Bhardwaj A, Kundra A, Gandhi B, Kumar S, Rehalia A, Gupta M. proposed a Logistic Regression model to predict the early onset of heart disease. They implemented feature engineering, feature transformation and algorithm tuning to improve the accuracy of Logistic Regression. They obtained an accuracy of 87%.

A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms [2] Haq AU, Li JP, Memon MH, Nazir S, Sun R utilized feature selection algorithms such as Relief, mRMR, and LASSO that were used to select important features, and on these selected features, the performance of the classifiers was tested. Machine learning classifiers such as logistic regression, K-NN, ANN, SVM, DT, and NB were used in the system.

An Analytical Model for Prediction of Heart Disease using Machine Learning Classifiers [3] D Roy, M Mahmood, T Joyti Roy used a dataset of 6 features only. They were able to obtain accuracies as high as 96% with random forest and also used new techniques such as ZeroR, Voted Perceptron and KStar, all of which resulted in accuracies greater than 90%.

Prediction of heart disease using machine learning [4] Aditi Gavhane; Gouthami Kokkula; Isha Pandya; Kailas Devadkar proposed to develop an application which can predict the vulnerability of a heart disease given basic symptoms like age, sex, pulse rate etc. The machine learning algorithm neural networks was used in their proposed system due to its high accuracy and reliability.

Heart Disease Prediction Using Machine learning and Data Mining Technique [5]
Jaymin Patel, Prof.TejalUpadhyay, Dr. Samir Patel used J48 algorithm, Logistic Tree and Random Forest algorithms on 12 features to predict heart disease. The J48 algorithm with reduced error pruning resulted in the most acceptable error and hence was concluded as the most reliable algorithm for heart disease prediction.

Prediction of Cardiovascular Disease Using Machine Learning Algorithms [7]
Kumar G Dinesh; K Arumugaraj; Kumar D Santhosh; V Mareeswari compared the accuracies of the individual results of Support Vector Machine, Gradient Boosting, Random forest, Naive Bayes classifier and logistic regression on a dataset taken in a region to present an accurate model of predicting cardiovascular disease.

Improving the Accuracy in Prediction of Heart Disease using Machine Learning Algorithms [8] **B. Keerthi Samhitha; M.R Sarika Priya.; C Sanjana.; Suja Cherukullapurath Mana; Jithina J** proposed a method called “outfit characterization” that combines different classifiers for improving the exactness of frail calculations.

CHAPTER 3

3.1 Aim of the Project:

The objective of this project is to build an ensemble model of different machine learning algorithms, including Logistic Regression, Loess model, Random Forest, KNN and Support Vector Machines, compare their performances, and use the best model to predict the onset of cardiovascular disease from a patient's data.

3.2 Scope of the Project:

The prediction model with highest accuracy can be incorporated into a mobile application or a website, to help patients know beforehand if they are likely to be diagnosed with a cardiovascular disease. The model can also be further enhanced to use the patient's current data to predict how many years it could take for them to be diagnosed with a CVD.

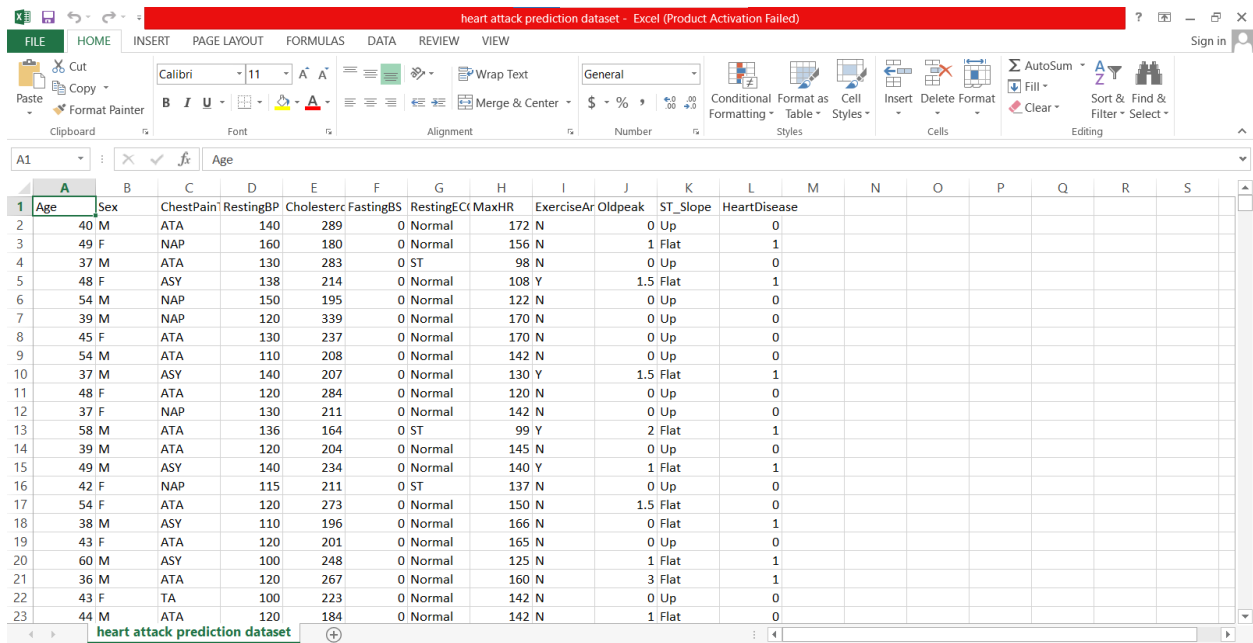
CHAPTER 4

METHODOLOGY AND ALGORITHMS USED

4.1 COLLECTION AND IMPORTING OF DATA

The dataset used for this project was directly obtained from Kaggle, an open source repository. The URL is given below:

<https://www.kaggle.com/fedesoriano/heart-failure-prediction?select=heart.csv>



Age	Sex	ChestPain	RestingBP	Cholesterc	FastingBS	RestingECG	MaxHR	ExerciseAr	Oldpeak	ST_Slope	HeartDisease
40	M	ATA	140	289	0	Normal	172	N	0	Up	0
49	F	NAP	160	180	0	Normal	156	N	1	Flat	1
37	M	ATA	130	283	0	ST	98	N	0	Up	0
48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
54	M	NAP	150	195	0	Normal	122	N	0	Up	0
39	M	NAP	120	339	0	Normal	170	N	0	Up	0
45	F	ATA	130	237	0	Normal	170	N	0	Up	0
54	M	ATA	110	208	0	Normal	142	N	0	Up	0
37	M	ASY	140	207	0	Normal	130	Y	1.5	Flat	1
48	F	ATA	120	284	0	Normal	120	N	0	Up	0
37	F	NAP	130	211	0	Normal	142	N	0	Up	0
58	M	ATA	136	164	0	ST	99	Y	2	Flat	1
39	M	ATA	120	204	0	Normal	145	N	0	Up	0
49	M	ASY	140	234	0	Normal	140	Y	1	Flat	1
42	F	NAP	115	211	0	ST	137	N	0	Up	0
54	F	ATA	120	273	0	Normal	150	N	1.5	Flat	0
38	M	ASY	110	196	0	Normal	166	N	0	Flat	1
43	F	ATA	120	201	0	Normal	165	N	0	Up	0
60	M	ASY	100	248	0	Normal	125	N	1	Flat	1
36	M	ATA	120	267	0	Normal	160	N	3	Flat	1
43	F	TA	100	223	0	Normal	142	N	0	Up	0
44	M	ATA	120	184	0	Normal	142	N	1	Flat	0

Fig 4.1: snapshot of the dataset in excel file

The dataset was imported into RStudio (4.0.0) as a csv file. The following libraries were downloaded for the project:

- tidyverse – for cleaning and tidying data
- ggplot2 – for data visualization and exploratory data analysis
- readxl – to read excel files into Rstudio
- knitr – to create quick tables

- caret – to split data and create machine learning models
- gam
- MASS
- RColorBrewer – to choose a palette of colors for visualization plots
- kableExtra – to generate tables in R
- e1071 – to utilize the support vector machine algorithm
- GGally – to create correlation plots for multivariate analysis

4.2 DESCRIPTION OF DATA

The dataset consists of 918 rows and 12 columns. Out of 12 columns, 11 predictors namely Age, Sex, Chest Pain Type, Resting BP, Cholesterol, Fasting Blood Sugar, Resting ECG, Maximum Heart Rate, Exercise Angina, Old Peak and ST Slope, have been used to predict the response variable, HeartDisease. The HeartDisease variable has a binary outcome; 1 indicating the presence of heart disease and 0 indicating its absence. A detailed description of each predictor is given below:

1. Age - Age of a person in years

2. Sex - M for male, F for female

3. Chest Pain Type -

ASY = Asymptomatic

ATA = Atypical Angina

NAP = Non Anginal Pain

TA = Typical Angina

4. Resting BP - Systolic, in mm Hg, at time of admission in hospital

5. Cholesterol - Serum cholesterol in mg/dl

6. Fasting BS - Fasting Blood Sugar > 120 mg/dl, (1 = true; 0 = false)

7. Resting ECG - Resting Electro-Cardio graphic results

Normal = normal

ST = having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)

LVH = showing probable or definite left ventricular hypertrophy by Estes' criteria

8. Max HR - Maximum Heart Rate achieved

9. Exercise Angina - Exercise induced Angina

Y = yes

N = no

10. Old peak- ST depression induced by exercise relative to rest

11. ST Slope - the slope of the peak exercise ST segment

Up = up sloping

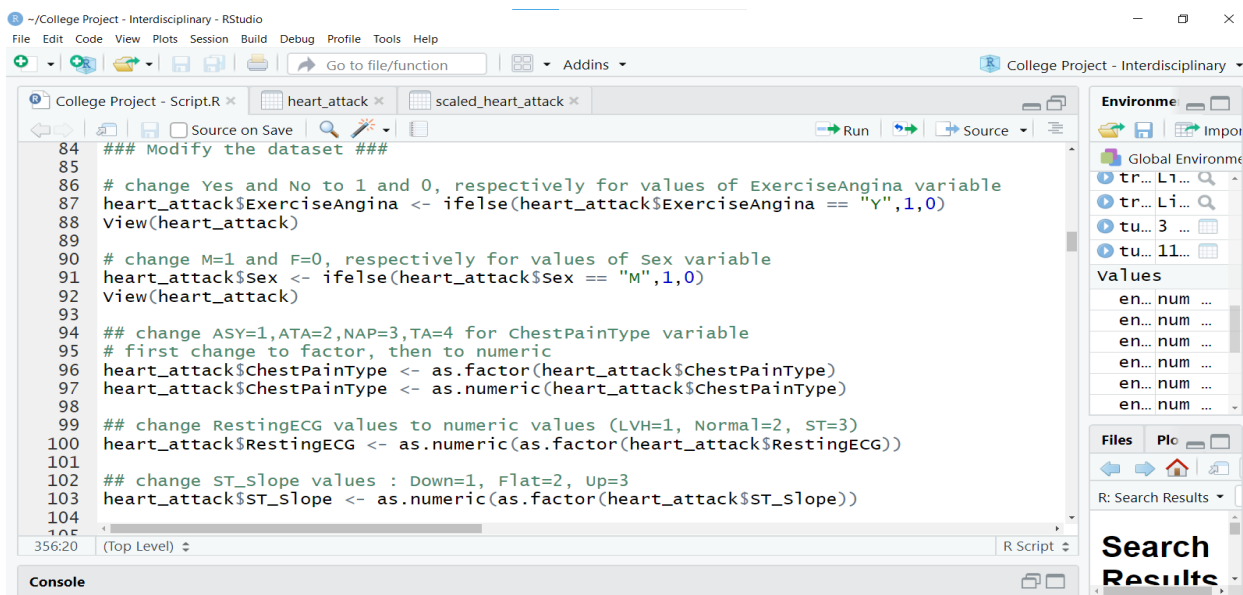
Flat = flat

Down = down sloping

There are no missing values in this dataset.

4.3 PRE-PROCESSING AND MODIFICATION OF DATA

The data variables were modified to make statistical computations on them, easier and more reliable, as follows:



```
## Modify the dataset ##
# change Yes and No to 1 and 0, respectively for values of ExerciseAngina variable
heart_attack$ExerciseAngina <- ifelse(heart_attack$ExerciseAngina == "Y",1,0)
View(heart_attack)

# change M=1 and F=0, respectively for values of Sex variable
heart_attack$Sex <- ifelse(heart_attack$Sex == "M",1,0)
View(heart_attack)

## change ASY=1,ATA=2,NAP=3,TA=4 for ChestPainType variable
# first change to factor, then to numeric
heart_attack$ChestPainType <- as.factor(heart_attack$ChestPainType)
heart_attack$ChestPainType <- as.numeric(heart_attack$ChestPainType)

## change RestingECG values to numeric values (LVH=1, Normal=2, ST=3)
heart_attack$RestingECG <- as.numeric(as.factor(heart_attack$RestingECG))

## change ST_slope values : Down=1, Flat=2, Up=3
heart_attack$ST_slope <- as.numeric(as.factor(heart_attack$ST_slope))
```

Fig 4.2: Modifying the Dataset

First, the binary values “Yes” and “No” of the Exercise Angina variable were converted to 1 and 0 respectively. Similarly, the Sex variable which initially held the values of “M” and “F” corresponding to Male and Female sexes, were changed to 1 and 0 respectively.

The Chest Pain Type variable contains 4 classes namely ASY (Asymptomatic), ATA (Atypical Angina), NAP (Non Anginal Pain) and TA (Typical Angina). These abbreviations were changed to 1, 2, 3 and 4 respectively, as shown above in the figure. The values of Exercise Angina and Resting ECG variables were converted to numerical class. Finally, the values of ST Slope, which were initially Down, Flat and Up, were correspondingly changed to numeric values of 1, 2 and 3. The dataset after modification contained only integer and numerical values.

4.4 EXPLORATORY DATA ANALYSIS

Going further into the data, it was observed that the population of patients consists of 725 men and 193 women. Also, 55.3% of the population in the dataset have been diagnosed with a cardiovascular disease. To explore the data visually, the first step performed was to visualize the distribution of ages in the data's population:

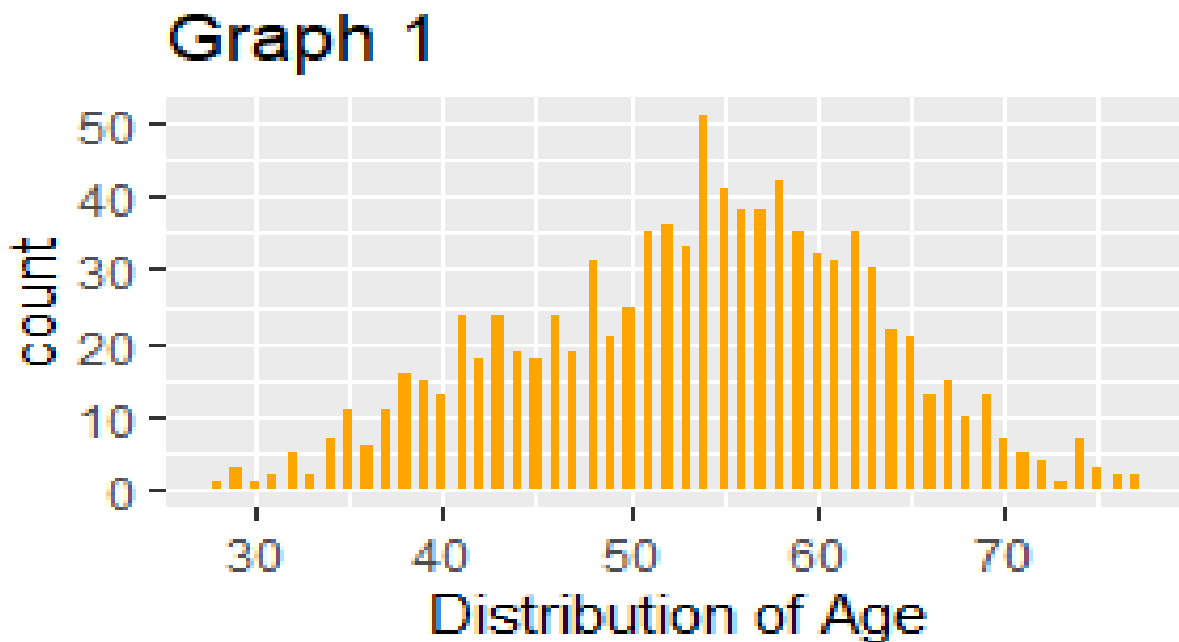


Fig 4.3: Distribution of ages of population across the dataset

The histogram is roughly bell-shaped, implying that the data follows an approximately normal distribution. The majority of the population is aged between 50-60 years. The least age of a person is 28 years, and the oldest person is 77 years old. The mean age of a person from this dataset is 53.5 years, and the median age is 54 years.

The distribution of types of chest pain recorded in patients was visualized with a simple bar graph:

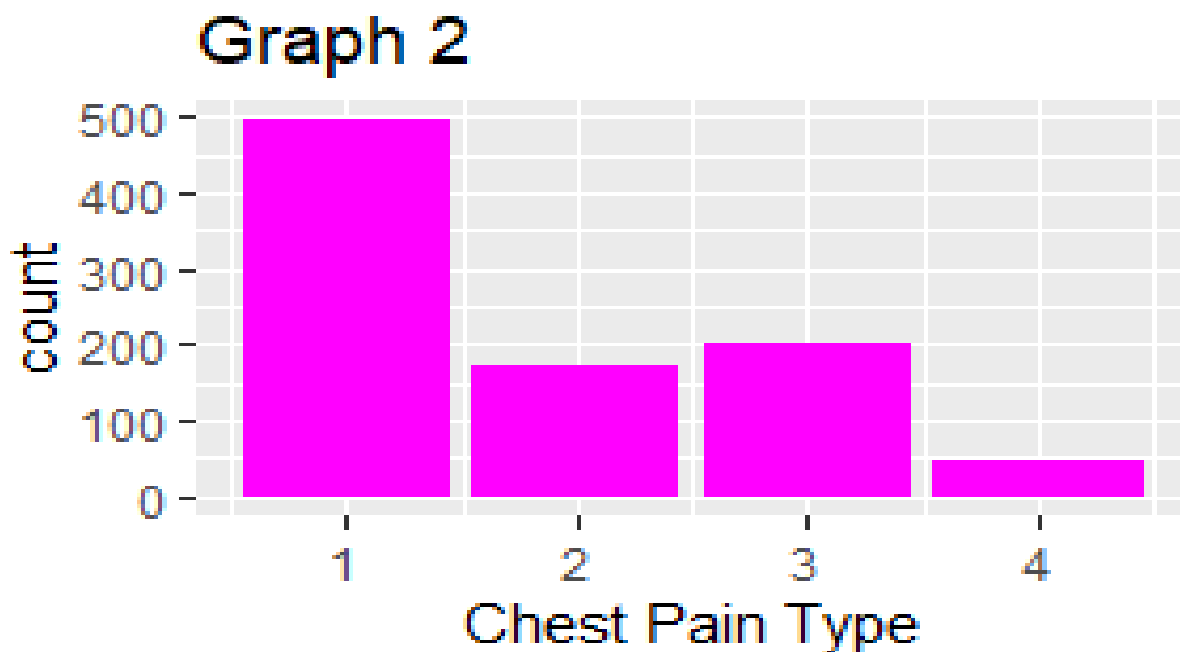


Fig 4.4: Distribution of chest pain type for all patients

From the graph, the majority proportion of the population, 496 people, has *Asymptomatic Chest Pain* (corresponding to a value of *1*), followed by 203 people who have *Non-Anginal pain* (=3*). The least proportion of population, containing 46 patients, has *Typical Anginal Pain* (=4*). This fact that the least number of patients had typical anginal pain was surprising, because the majority of patients were diagnosed with CVD, whereas very few had typical angina pain, which is considered a very common symptom of CVD or heart attack. Anginal pain is a chest pain that occurs due to restricted flow of blood to the heart.

The relation between a patient's age and their fasting blood sugar was analyzed:

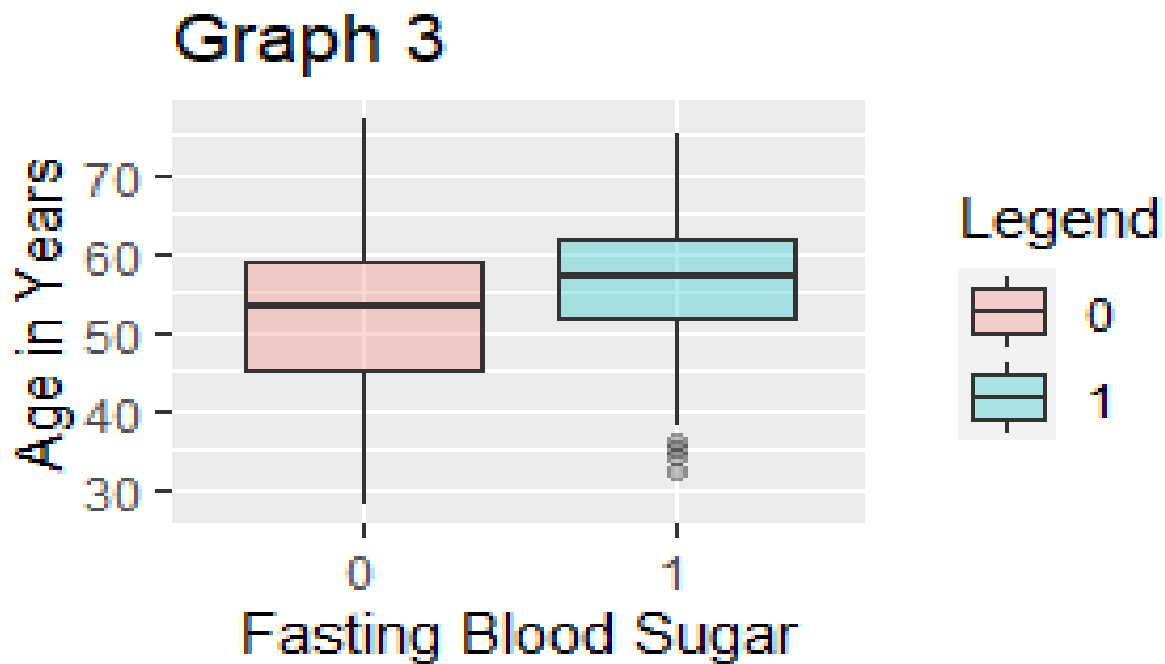


Fig 4.5: Age vs Fasting Blood Sugar of a patient

Fasting blood sugar refers to a test to determine how much glucose (sugar) is in a blood sample after an overnight fast. Generally, a good fasting blood sugar level is 70 – 100 mg/dl. Anything above this value could lead to a potential risk of diabetes, which could further cause damage to blood vessels and nerves that control the heart. Hence this test is an indirect method of checking for the risk of CVD. From the boxplot, a value of 1 for Fasting Blood Sugar indicates a blood sugar level greater than 120 mg/dl. People aged between 52-60 years, seem to have a higher level of Fasting Blood Sugar, the average age being 56 years.

The overall trends in the dataset was observed from the following heatmap:

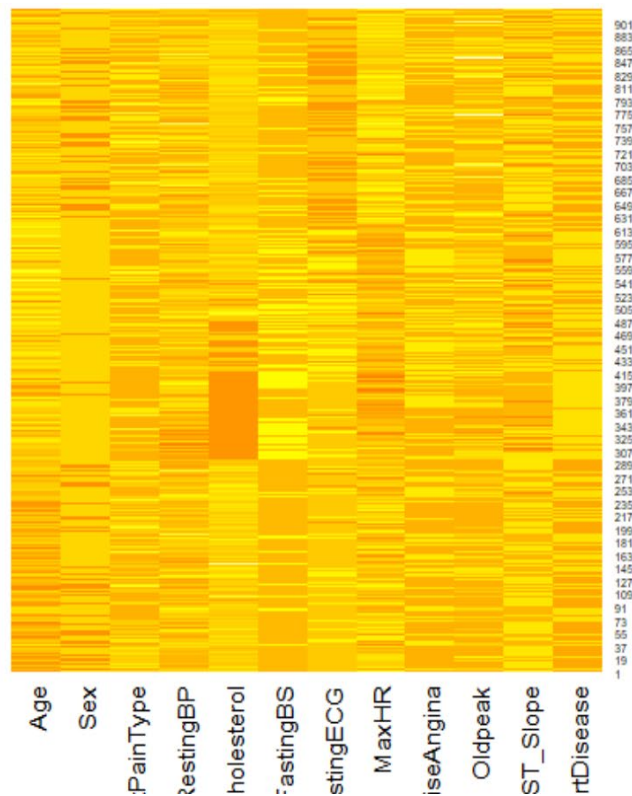


Fig 4.6: heatmap showing trends in the dataset for all variables

The most obvious changes in trends can be observed from the Cholesterol variable's column. Patient samples 300 – 400 have a hike in cholesterol levels compared to the rest of the population. Similarly, majority of patient samples between 600th and 900th observations have been diagnosed with CVD.

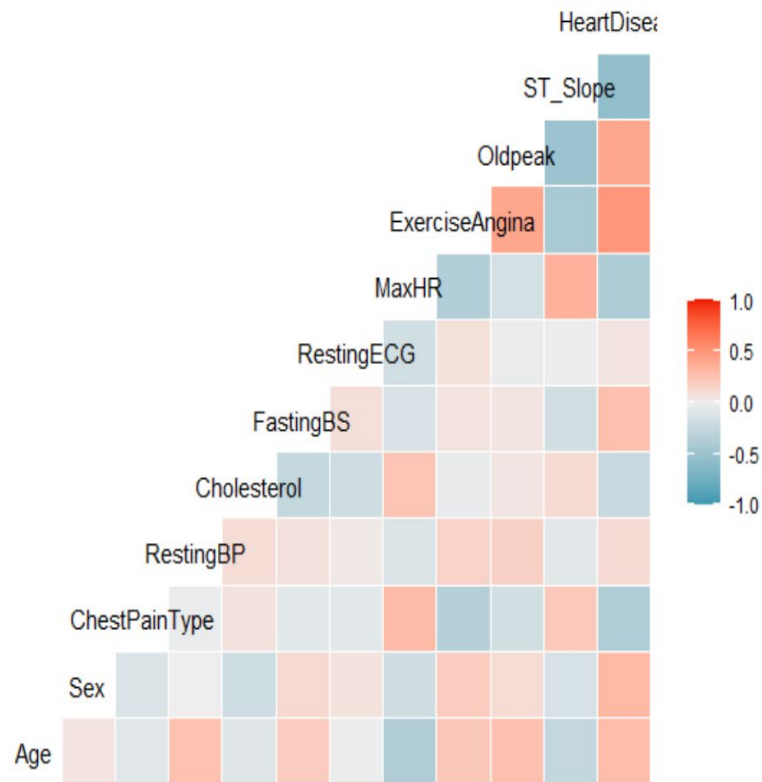


Fig 4.7: heatmap showing correlations between all variables

The correlation heatmap drawn above was perhaps the most important graph drawn for this project. From this, it was observed that the variables Age, Sex, Fasting Blood Sugar, Exercise Angina and Old peak vary linearly with heart disease, showing a correlation greater than 0. Conversely, ST Slope, Chest Pain Type and Maximum Heart Rate show an inverse relationship with Heart Disease. With increasing age, the chances of Exercise angina which is chest pain developed during exercise, also increase. The highest negative correlation with Heart Disease was seen with ST Slope.

The correlations of all variables with Heart Disease was numerically computed using Spearman's rank correlation method, which is robust to outliers and hence a reliable statistical tool.

	Parameter	correlation with heart disease
1	Age	0.28957565
2	Sex	0.30544492
3	ChestPainType	-0.44251603
4	Resting BP	0.11386630
5	Cholesterol	-0.13987309
6	Fasting BS	0.26729119
7	Resting ECG	0.05689404
8	Max HR	-0.40482679
9	Exercise Angina	0.49428199
10	Old peak	0.41904610
11	ST Slope	-0.59191284

Table 4.1: Spearman's rank correlation of all variables with Heart Disease variable

4.5 BUILDING THE ALGORITHM

One of the preliminary steps of model fitting is to first split the data into test and train sets. A vector “y”, holding the binary outcomes under the Heart Disease variable was created. The target variable Heart Disease was removed from the dataset, so it finally contained only the predictor variables. The dataset was then scaled, in order to maintain uniform variance across all predictor variables. The scaled dataset was split into test and train sets. The “y” vector holding the Heart Disease variable's values was also split into test and train sets and converted into factors for easy classification. The splitting percentage used was 75:25.

4.5.1 INDIVIDUAL MODEL TESTING

A number of individual models were fitted onto the data using trial and error method, as follows:

1. LDA model

```
train_lda <- train(train_HA, train_y, method = "lda", family = "binomial")  
lda_preds <- predict(train_lda, test_HA)  
mean(lda_preds == test_y)
```

2. QDA model

```
train_qda <- train(train_HA, train_y, method = "qda", family = "binomial")  
qda_preds <- predict(train_qda, test_HA)  
mean(qda_preds == test_y)
```

3. Logistic regression

```
train_glm <- train(train_HA, train_y, method = "glm", family = "binomial")  
glm_preds <- predict(train_glm, test_HA)  
mean(glm_preds == test_y)
```

4. Loess model

```
train_loess <- train(train_HA, train_y, method = "gamLoess")  
loess_preds <- predict(train_loess, test_HA)  
mean(loess_preds == test_y)
```

5. KNN model

```
tuning <- data.frame(k = seq(21,41,2))  
train_knn <- train(train_HA, train_y,  
                  method = "knn",  
                  tuneGrid = tuning)
```

The best k value used for splitting the data was given by:

```
train_knn$bestTune  
# 29
```

```
knn_preds <- predict(train_knn, test_HA)  
mean(knn_preds == test_y)
```

6. Random Forest model

```
tune_rf <- data.frame(mtry=c(0.1,0.3,0.5))  
train_rf <- train(train_HA, train_y,  
                 method = "rf",  
                 tuneGrid = tune_rf,  
                 importance = TRUE)
```

```
# get best mtry value  
train_rf$bestTune  
# 0.3
```

```
rf_preds <- predict(train_rf, test_HA)  
mean(rf_preds == test_y)
```

7. Support Vector Machine model

```
train_svm <- svm(formula = train_y ~ .,  
                 data = train_HA,  
                 type = 'C-classification',  
                 kernel = 'linear')
```

```
svm_preds <- predict(train_svm, test_HA)
```

```
mean(svm_preds == test_y)
```

8. A modified SVM model was created that was tuned using the tuneGrid parameter:

```
grid <- expand.grid(C = c(0,0.01,0.05,0.1,0.15,0.25,1.5,1.75,3.5))  
new_train_svm <- train(train_HA, train_y,  
                      method = "svmLinear",  
                      preProcess = c("center","scale"),  
                      tuneGrid = grid)  
new_svm_preds <- predict(new_train_svm, test_HA)  
mean(new_svm_preds == test_y)
```

4.5.2 ENSEMBLE MODEL TESTING

A number of ensemble models comprising of various combinations of the individual models were created and fitted onto the dataset. Ensemble models provide an edge over individual models in terms of giving better accuracy as they utilize the methods of more than one algorithm while performing the required task. The best ensemble model can only be found by trial and error method, which was performed as shown below:

1. An ensemble of all models (LDA, QDA, Logistic Regression, Loess, KNN, RF, SVM and Modified SVM)
2. RF and KNN
3. RF, KNN and SVM
4. RF and SVM
5. Loess and SVM
6. Loess, SVM and RF
7. Loess, SVM, RF and KNN
8. Modified SVM and RF

CHAPTER 5

RESULTS AND DISCUSSION

The accuracies of the individual models were tabulated as follows:

MODEL	ACCURACY (%)
Modified SVM	88.70
SVM	88.30
RANDOM FOREST	87.40
LOESS	86.90
LOGISTIC REGRESSION	86.50
LDA	85.60
KNN	84.70
QDA	83.04

Table 5.1: accuracies of individual models tested

The modified SVM gave the highest accuracy, followed by the SVM model and then Random Forest. An important conclusion derived from this result is that tuning always helps in increasing the accuracy and performance of an algorithm. Using a grid of cost values for SVM helped to improve its performance. A method of “feature selection” was also carried out using the modified SVM model to find how well an individual variable could predict Heart Disease. Interestingly, ST Slope alone, could predict Heart Attack with an accuracy of 81%, while Exercise Angina and Chest Pain Type could individually predict Heart Disease with an accuracy of 75.6% and 74% respectively.

The modified SVM model's performance can be visualized here:

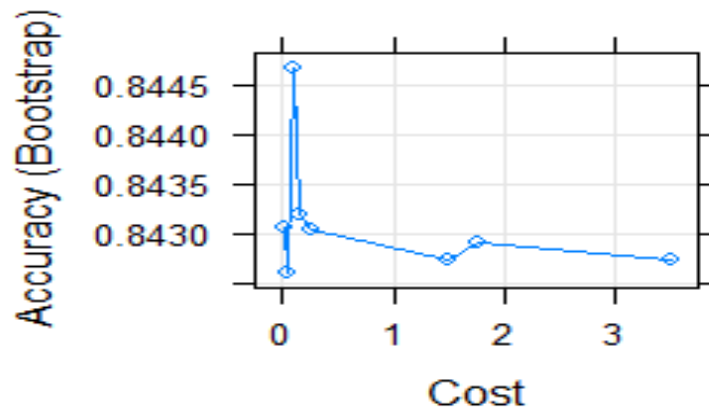


Fig 5.1: graph showing modified SVM's performance for different cost (c) values

However, it is noteworthy to mention that tuning an algorithm results in performing differently each time the model is run; the accuracy changes with every run of the model. The highest obtained accuracies only have been mentioned here. The QDA and LDA models did not fare well because of the large size of the dataset; they work best only when they are supplied with small sizes of data, say, 200 samples. Unfortunately, KNN failed to give a high accuracy; the highest it could provide was 84.7% with a k value of 29, as shown below:

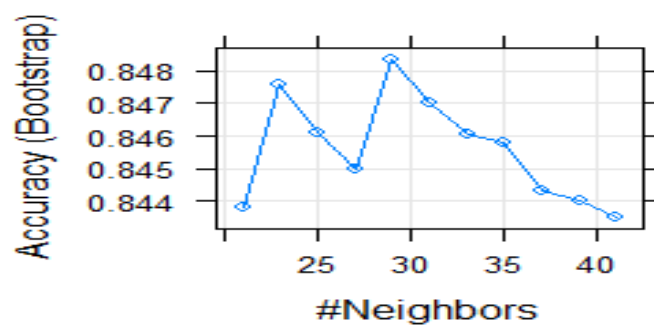


Fig 5.2: graph showing performance of KNN for different k values (neighbors)

In general, the Random Forest model is popularly used for classification problems due to its robust nature of splitting trees. The highest accuracy obtained with the RF model was 87.4%, with a mtry value of 0.3. The mtry value is the number of variables that are used as candidates for each split. It was observed that decreasing the value of mtry drastically, helped improve the model, as shown below:

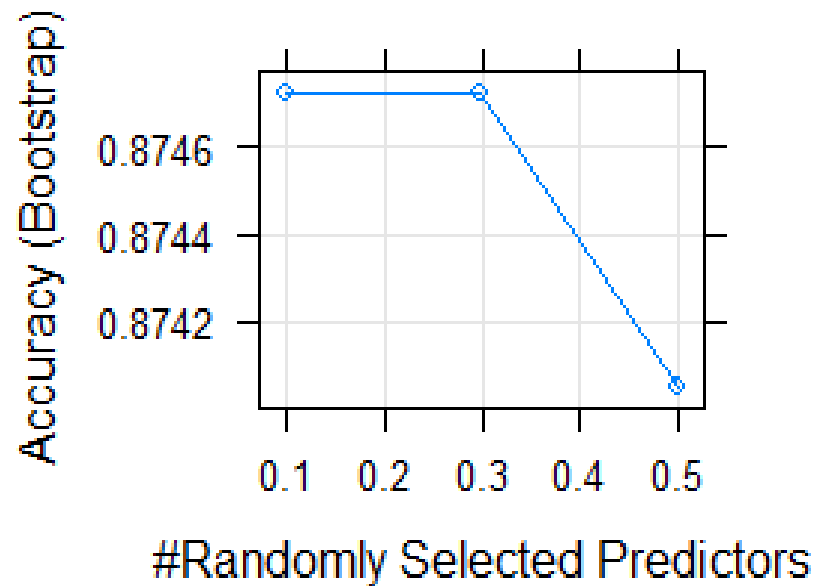


Fig 5.3: performance of Random Forest for different mtry values

The results of the ensemble model are shown below:

MODEL	ACCURACY (%)
Modified SVM and RF	88.70
SVM and RF	88.30
All Models	87.83
SVM, RF and Loess	87.80
SVM, RF and KNN	87.40
SVM, RF, Loess and KNN	87.30
SVM and Loess	86.90
RF and KNN	84.78

Table 5.2: accuracies of ensemble models tests

Overall, it can be observed that ensemble models performed better than individual models. The lowest accuracy of an individual model (QDA) was 83.04%, whereas the lowest accuracy of an ensemble model (RF and KNN) was 84.78%, showing a marginal increase in performance. Most ensemble models resulted in an accuracy of 87%, while the highest accuracy was found with an ensemble of Modified SVM and RF, which is the final model for this project. A confusion matrix of the final ensemble (Modified SVM and RF) was drawn to evaluate its performance in-depth:

	TEST SET	
FINAL MODEL	0	1
0	92	15
1	11	112

Table 5.3: confusion matrix of final ensemble model

CHAPTER 6

CONCLUSIONS

This project concludes with a final ensemble model that utilizes the power of support vector machines and random forest in order to predict the onset of a cardiovascular disease, using the patient's data of their Age, Sex, Chest Pain Type, Resting BP, Cholesterol, Fasting Blood Sugar, Resting ECG, Maximum Heart Rate, Exercise Angina, Old Peak and ST Slope. With regard to the dataset, variables Age, Fasting Blood Sugar levels, Exercise Angina and ST Slope were found most important to affect Heart Disease or CVD. With respect to the patient, it can be concluded that Blood sugar levels for patients over 50 years must be controlled to prevent CVD, and asymptomatic chest pains must be checked for senior citizens at the earliest, for signs of CVD.

REFERENCES

1. Bhardwaj A, Kundra A, Gandhi B, Kumar S, Rehalia A, Gupta M. Prediction of heart attack using machine learning. IITM Journal of Management and IT. 2019;10(1):20-4.
2. Haq AU, Li JP, Memon MH, Nazir S, Sun R. A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. Mobile Information Systems. 2018 Dec 2;2018.
3. Roy D, Mahmood M, Joyti Roy T. An Analytical Model for Prediction of Heart Disease using Machine Learning Classifiers.
4. Gavhane A, Kokkula G, Pandya I, Devadkar K. Prediction of heart disease using machine learning. In 2018 second international conference on electronics, communication and aerospace technology (ICECA) 2018 Mar 29 (pp. 1275-1278). IEEE.
5. Shah D, Patel S, Bharti SK. Heart disease prediction using machine learning techniques. SN Computer Science. 2020 Nov;1(6):1-6.
6. Patel J, Tejal Upadhyay D, Patel S. Heart disease prediction using machine learning and data mining technique. Heart Disease. 2015 Sep;7(1):129-37.
7. Dinesh KG, Arumugaraj K, Santhosh KD, Mareeswari V. Prediction of cardiovascular disease using machine learning algorithms. In 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT) 2018 Mar 1 (pp. 1-7). IEEE.
8. Samhitha BK, Priya MS, Sanjana C, Mana SC, Jose J. Improving the accuracy in prediction of heart disease using machine learning algorithms. In 2020 International Conference on Communication and Signal Processing (ICCSP) 2020 Jul 28 (pp. 1326-1330). IEEE.

**SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY
SCHOOL OF BIO AND CHEMICAL ENGINEERING
DEPARTMENT OF BIOTECHNOLOGY**



**PREDICTION OF HEART DISEASE USING
MACHINE LEARNING ALGORITHMS**

Internal Guide

Dr. P Prakash

Dept of Biotechnology

External Guide

Dr. P Ajitha

Dept of IT

DEEKSHA RV (39230010)

Dept of Biotechnology

PROBLEM STATEMENT

❖ PROBLEM:

CVD's are now the most common cause of death and are in need of quick indicators to predict the onset of a heart-attack or any other related disease.

❖ SOLUTION:

Build an ensemble model of different machine learning algorithms, including Random Forest, Support Vector Machines, Loess, Logistic Regression and KNN, to predict the onset of heart attack from a patient's data.

METHOD FOR SOLUTION

1. Analyze data for important features
2. Train individual models for classification
3. Test individual models for classification
4. Train ensemble models with good performance
5. Test ensemble models with good performance
6. Run model with the highest performance

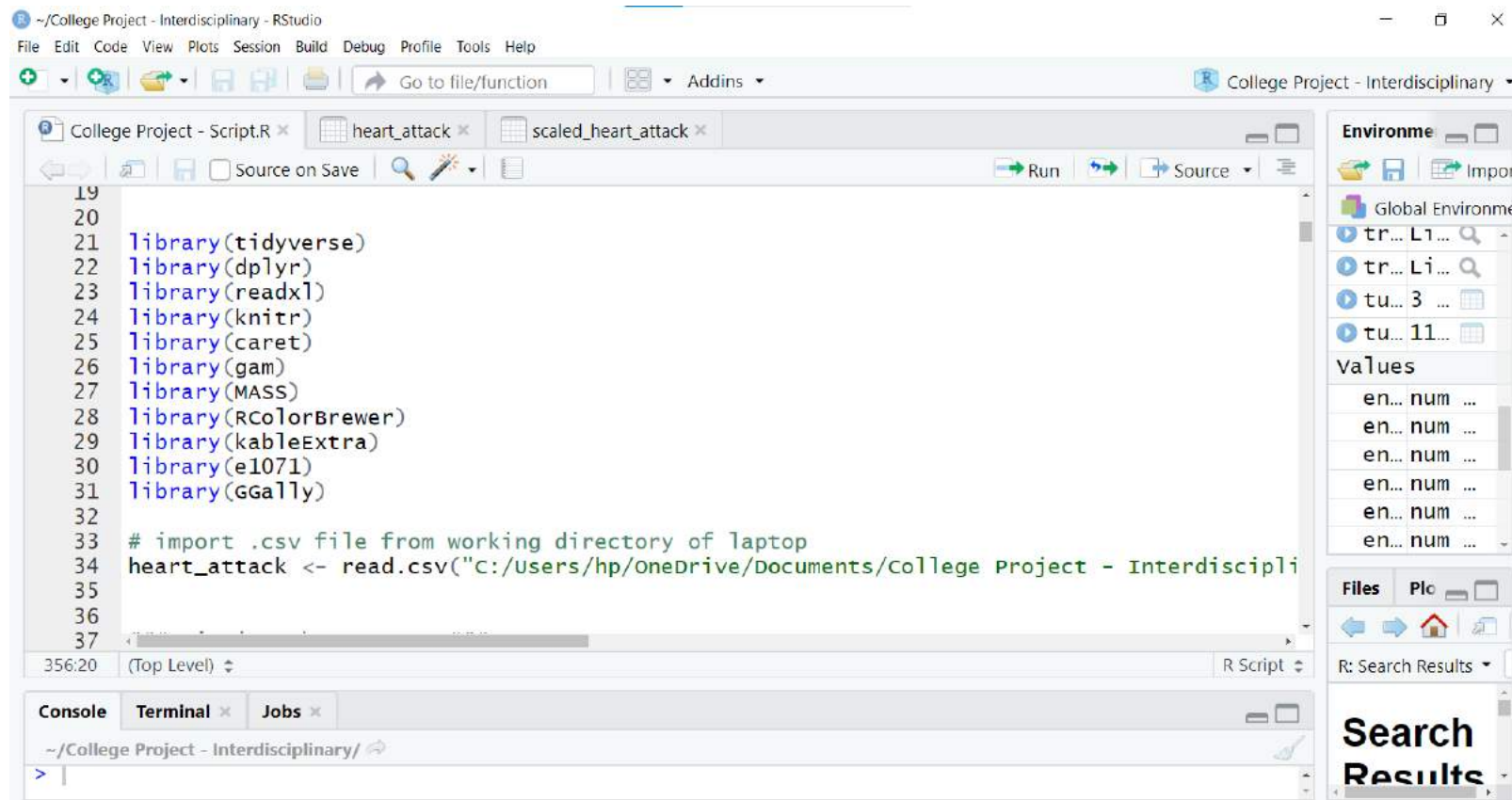
METHODOLOGY

- Data downloading and Importing

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Age	Sex	ChestPain	RestingBP	Cholesterc	FastingBS	RestingECG	MaxHR	ExerciseAr	Oldpeak	ST_Slope	HeartDisease							
2	40	M	ATA	140	289	0	Normal	172	N		0	Up	0						
3	49	F	NAP	160	180	0	Normal	156	N		1	Flat	1						
4	37	M	ATA	130	283	0	ST	98	N		0	Up	0						
5	48	F	ASY	138	214	0	Normal	108	Y		1.5	Flat	1						
6	54	M	NAP	150	195	0	Normal	122	N		0	Up	0						
7	39	M	NAP	120	339	0	Normal	170	N		0	Up	0						
8	45	F	ATA	130	237	0	Normal	170	N		0	Up	0						
9	54	M	ATA	110	208	0	Normal	142	N		0	Up	0						
10	37	M	ASY	140	207	0	Normal	130	Y		1.5	Flat	1						
11	48	F	ATA	120	284	0	Normal	120	N		0	Up	0						
12	37	F	NAP	130	211	0	Normal	142	N		0	Up	0						
13	58	M	ATA	136	164	0	ST	99	Y		2	Flat	1						
14	39	M	ATA	120	204	0	Normal	145	N		0	Up	0						
15	49	M	ASY	140	234	0	Normal	140	Y		1	Flat	1						
16	42	F	NAP	115	211	0	ST	137	N		0	Up	0						
17	54	F	ATA	120	273	0	Normal	150	N		1.5	Flat	0						
18	38	M	ASY	110	196	0	Normal	166	N		0	Flat	1						
19	43	F	ATA	120	201	0	Normal	165	N		0	Up	0						
20	60	M	ASY	100	248	0	Normal	125	N		1	Flat	1						
21	36	M	ATA	120	267	0	Normal	160	N		3	Flat	1						
22	43	F	TA	100	223	0	Normal	142	N		0	Up	0						
23	44	M	ATA	120	184	0	Normal	142	N		1	Flat	0						

METHODOLOGY

2. Data Processing and Modification



The screenshot displays the RStudio environment with the following components:

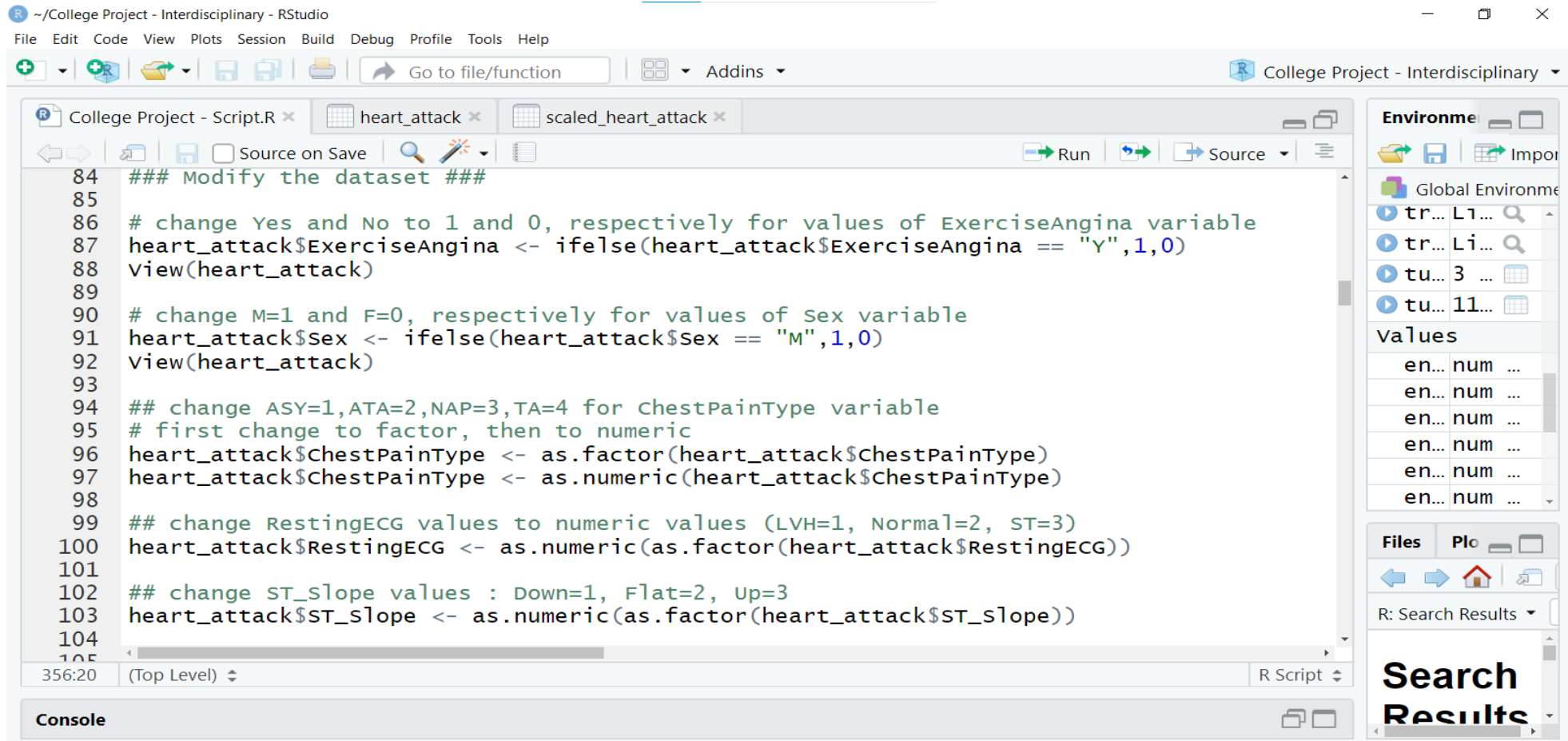
- Menu Bar:** File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Toolbar:** Includes icons for saving, running, and navigating between files.
- Source Editor:** Contains R code for loading libraries and reading a CSV file.

```
19  
20  
21 library(tidyverse)  
22 library(dplyr)  
23 library(readxl)  
24 library(knitr)  
25 library(caret)  
26 library(gam)  
27 library(MASS)  
28 library(RColorBrewer)  
29 library(kableExtra)  
30 library(e1071)  
31 library(GGally)  
32  
33 # import .csv file from working directory of laptop  
34 heart_attack <- read.csv("C:/Users/hp/OneDrive/Documents/College Project - Interdiscipli  
35  
36  
37
```
- Environment Pane:** Shows the Global Environment with variables like 'tr... L1...', 'tr... Li...', 'tu... 3', and 'tu... 11...'. It also displays a table of values with columns 'en...' and 'num...'.

en...	num...
en...	num...
en...	num...
en...	num...
en...	num...
en...	num...
en...	num...
- Console:** Shows the current directory as '~ /College Project - Interdisciplinary/' and a prompt '> |'.
- Search Pane:** Displays 'Search Results'.

METHODOLOGY

Data Processing and Modification



The screenshot displays the RStudio interface with a script editor containing R code for data processing. The code is as follows:

```
84 ### Modify the dataset ###
85
86 # change Yes and No to 1 and 0, respectively for values of ExerciseAngina variable
87 heart_attack$ExerciseAngina <- ifelse(heart_attack$ExerciseAngina == "Y",1,0)
88 View(heart_attack)
89
90 # change M=1 and F=0, respectively for values of Sex variable
91 heart_attack$Sex <- ifelse(heart_attack$Sex == "M",1,0)
92 View(heart_attack)
93
94 ## change ASY=1,ATA=2,NAP=3,TA=4 for ChestPainType variable
95 # first change to factor, then to numeric
96 heart_attack$ChestPainType <- as.factor(heart_attack$ChestPainType)
97 heart_attack$ChestPainType <- as.numeric(heart_attack$ChestPainType)
98
99 ## change RestingECG values to numeric values (LVH=1, Normal=2, ST=3)
100 heart_attack$RestingECG <- as.numeric(as.factor(heart_attack$RestingECG))
101
102 ## change ST_slope values : Down=1, Flat=2, Up=3
103 heart_attack$ST_slope <- as.numeric(as.factor(heart_attack$ST_slope))
104
105
```

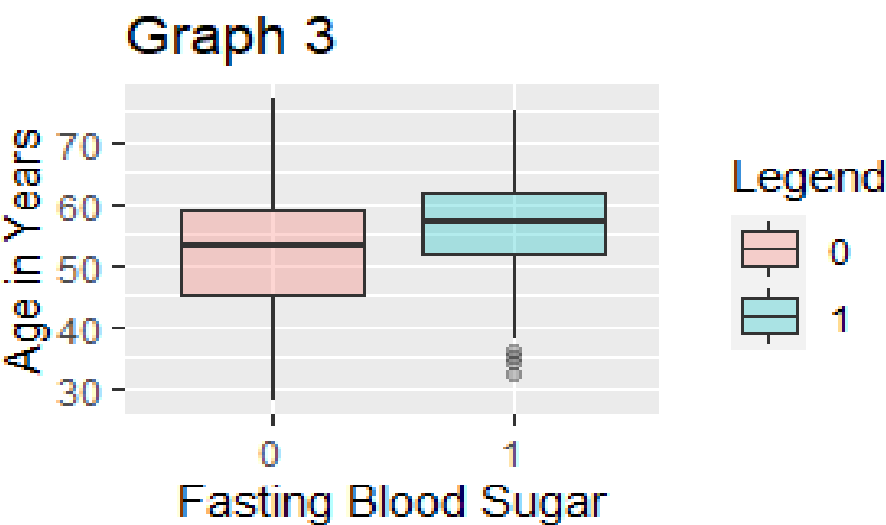
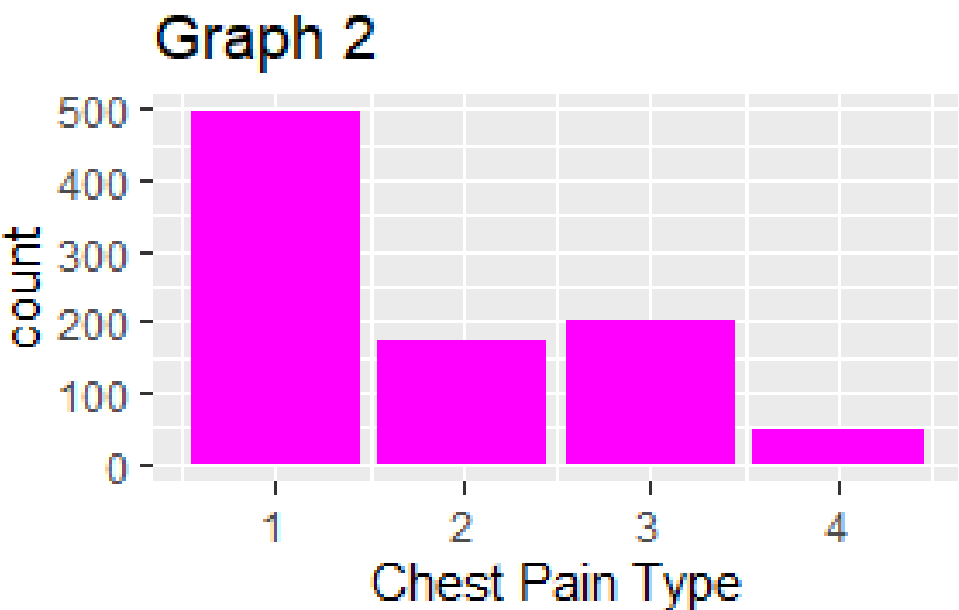
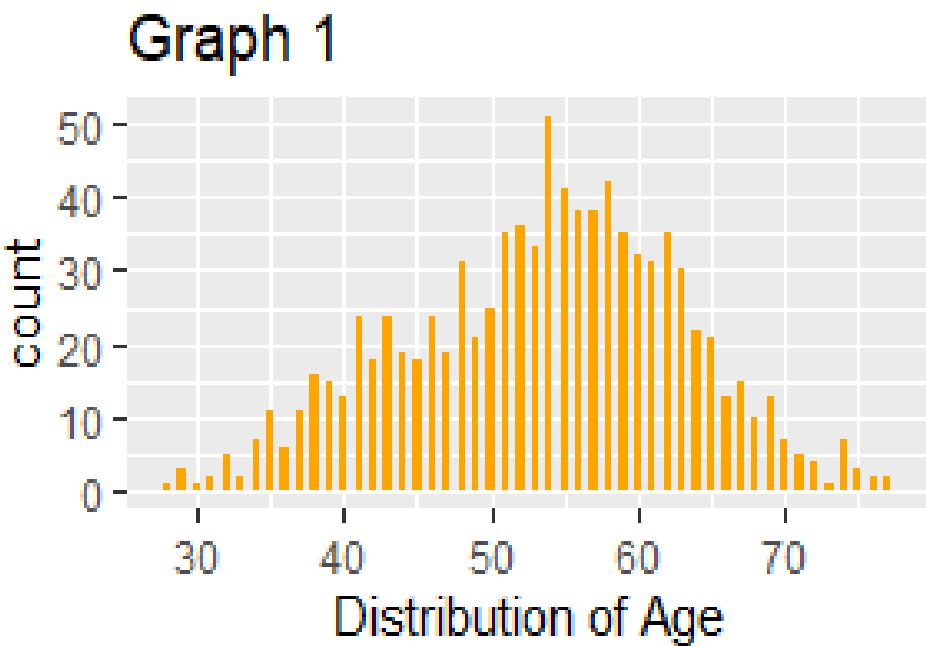
The right sidebar shows the Environment pane with a table of values:

values	
en...	num ...
en...	num ...
en...	num ...
en...	num ...
en...	num ...
en...	num ...

The bottom pane shows the Console with the text "(Top Level)" and "R Script".

METHODOLOGY

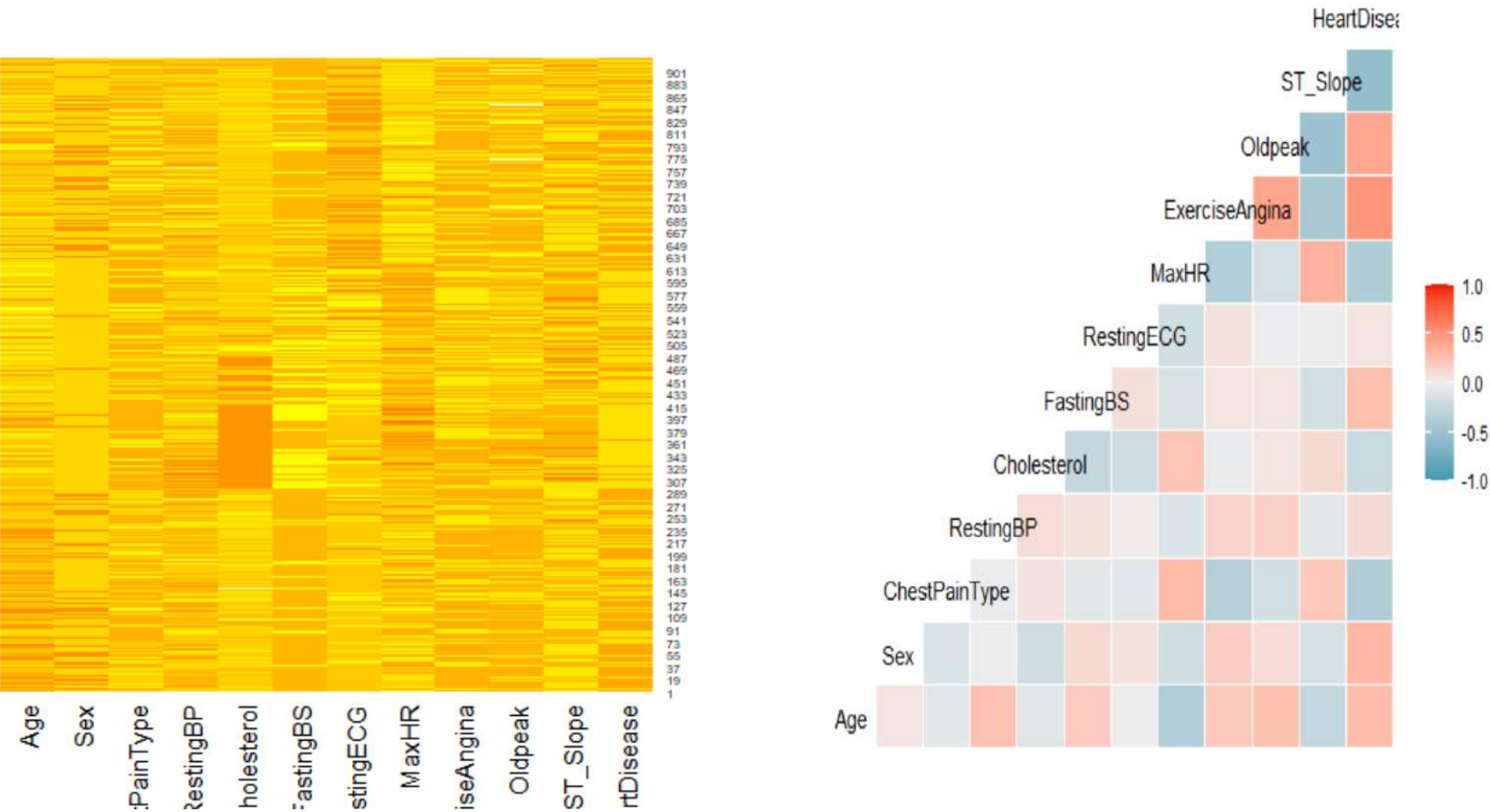
3. Exploratory Data Analysis



METHODOLOGY

3. Exploratory Data Analysis

- Multivariate analysis



	Parameter	correlation with heart disease
1	Age	0.28957565
2	Sex	0.30544492
3	ChestPainType	-0.44251603
4	Resting BP	0.11386630
5	Cholesterol	-0.13987309
6	Fasting BS	0.26729119
7	Resting ECG	0.05689404
8	Max HR	-0.40482679
9	Exercise Angina	0.49428199
10	Old peak	0.41904610
11	ST Slope	-0.59191284

METHODOLOGY

3. Building the Algorithm

➤ Individual models tested

- Linear Discriminant Analysis
- Quadratic Discriminant Analysis
- Logistic Regression
- Loess
- KNN
- Random Forest
- Support Vector Machine

➤ Ensemble models tested

- All models
- RF and KNN
- RF, KNN and SVM
- RF and SVM
- Loess and SVM
- Loess, SVM and RF
- Loess, SVM, KNN and RF
- Modified SVM and RF

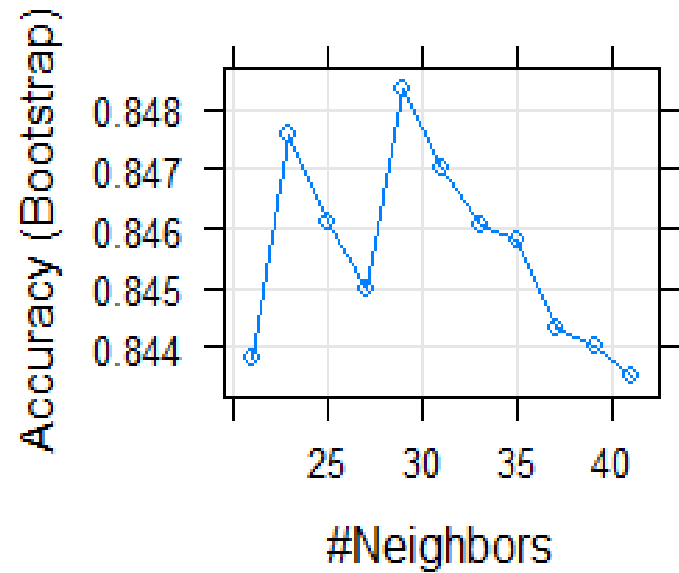
RESULTS

MODEL	ACCURACY (%)
Modified SVM	88.70
SVM	88.30
RANDOM FOREST	87.40
LOESS	86.90
LOGISTIC REGRESSION	86.50
LDA	85.60
KNN	84.70
QDA	83.04

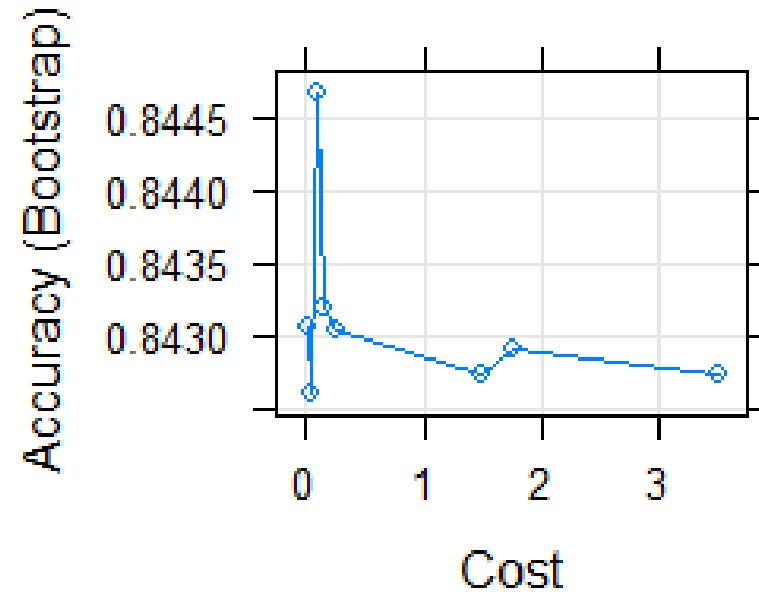
MODEL	ACCURACY (%)
Modified SVM and RF	88.70
SVM and RF	88.30
All Models	87.83
SVM, RF and Loess	87.80
SVM, RF and KNN	87.40
SVM, RF, Loess and KNN	87.30
SVM and Loess	86.90
RF and KNN	84.78

RESULTS

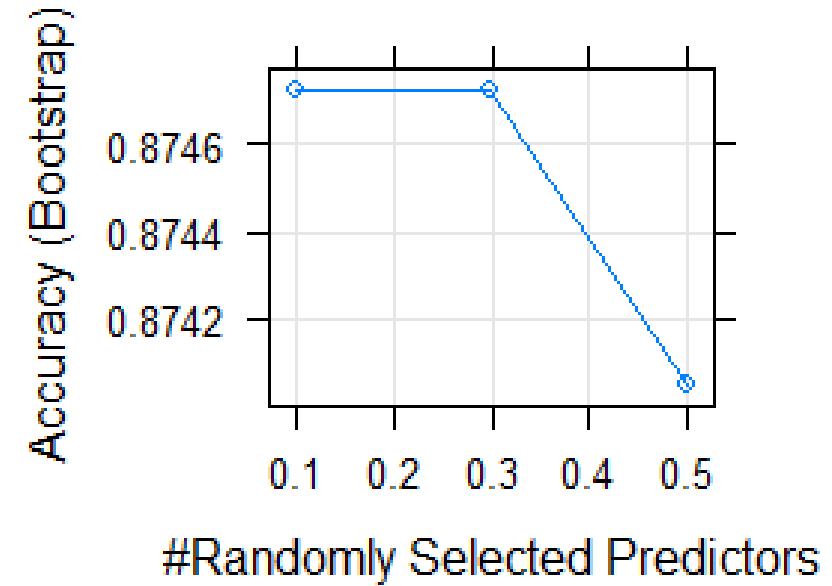
KNN



SVM



RF



RESULTS

Confusion Matrix of Final Model (Modified SVM + RF) vs QDA model

	FINAL MODEL	
TEST SET	0	1
0	92	11
1	15	112

	QDA MODEL	
TEST SET	0	1
0	83	20
1	19	108

~/College Project - Interdisciplinary - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins College Project - Interdisciplinary

College Project - Script.R heart_attack scaled_heart_attack

Source on Save Run Source

```
19
20
21 library(tidyverse)
22 library(dplyr)
23 library(readxl)
24 library(knitr)
25 library(caret)
26 library(gam)
27 library(MASS)
28 library(RColorBrewer)
29 library(kableExtra)
30 library(e1071)
31 library(GGally)
32
33
```

21:1 (Top Level) R Script

Console Terminal Jobs

~/College Project - Interdisciplinary/

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was $c = 0.1$.
> mean(ensemble_8_preds == test_y)
[1] 0.8869565
>

Environment History

Import Dataset

Global Environment

Data

ens...	logi	[1...
ens...	logi	[1...
ens...	logi	[1...
ens...	logi	[1...
ens...	logi	[1...
ens...	logi	[1...
ens...	logi	[1...
ens...	logi	[1...
ens...	logi	[1...
Gra...	List of...	
Gra...	List of...	
Gra...	List of...	
Gra...	List of...	
Gra...	Large g...	

Files Plots Pack

Zoom Exp

34°C 10:10 AM 4/29/2022

CONCLUSIONS

- Age, blood sugar levels, exercise angina and ST Slope were the most important variables that affected the onset of a cardiovascular disease;
- Asymptomatic chest pains must be checked at the earliest, by a medical professional for signs of CVD;
- Blood sugar levels for patients over 50 years must be controlled to prevent CVD;
- The Support Vector Machine algorithm proved to be the most superior model for classification, when tested individually;
- Ensemble models provide better accuracy than individual models, hence the final algorithm combines SVM and RF models;