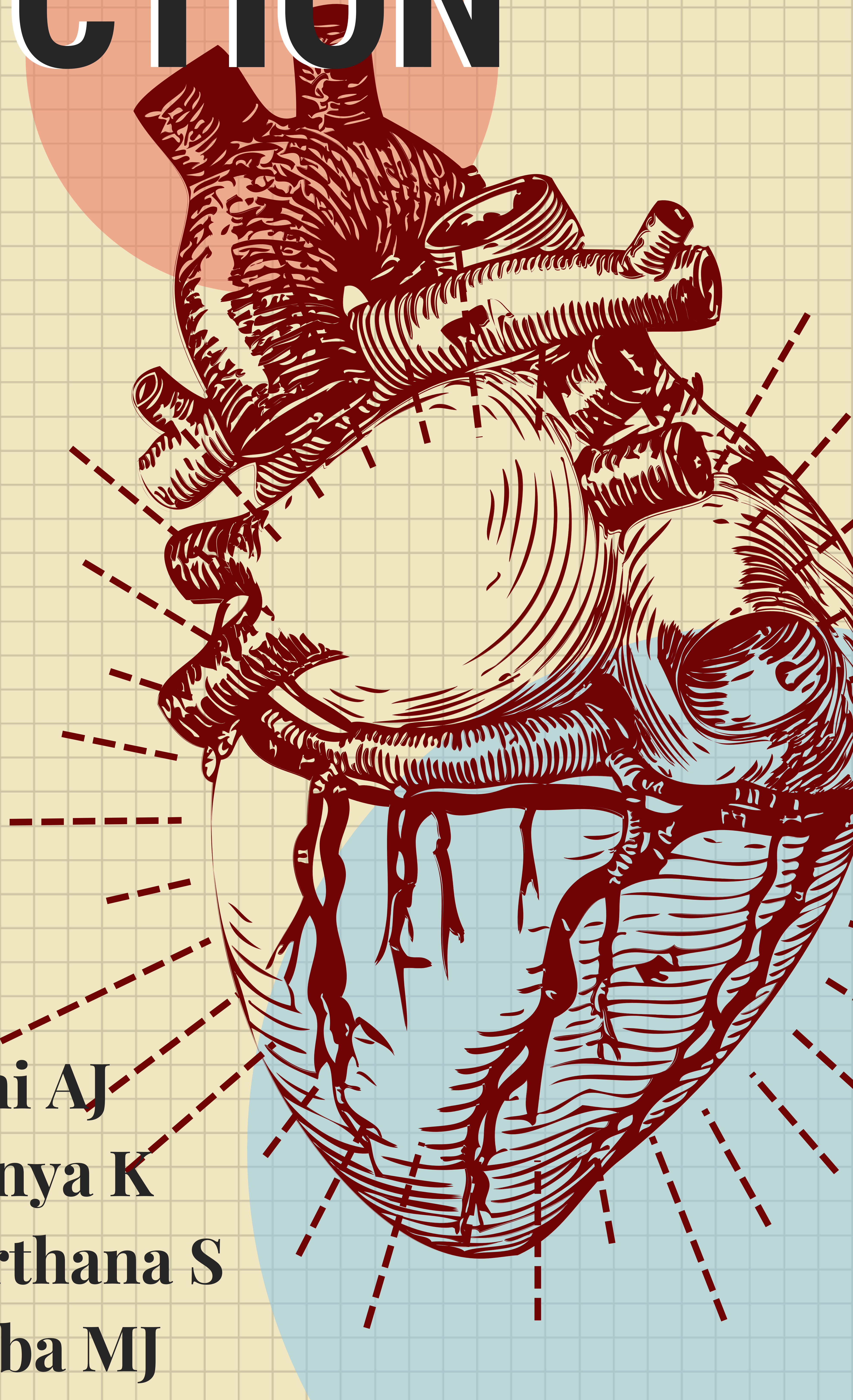
FIDE ET LABORE

ST. JOSEPH'S COLLEGE

HEART DISEASE PREDICTION

SEMESTER- 1

PROJECT REPORT 2020



20BDA27 - Josmi AJ

20BDA68 - Ananya K

20BDA39 - Keerthana S

20BDA13 - Jesalba MJ

Contents

Estimation of Prediction for getting Coronary Heart Disease using Logistic Regression.

INTRODUCTION

The load of cardiovascular diseases is rapidly increasing all over the world. Even if these diseases have been found as the most important source of death, it has been announced as the most manageable and avoidable disease. Mainly, blockage in arteries causes heart stroke. It occurs when heart does not pump the blood around the body efficiently.

Having high blood pressure is also one of the main causes of getting a heart disease. A survey says that, in 2011 to 2014, the commonness of hypertension in the world was about 35%, which is also a cause of heart disease. Similarly, there are many more reasons for getting a heart disease such as obesity, not taking in proper nutrition, increased cholesterol and lack of physical activity. So, prevention is very necessary. For prevention, awareness of heart diseases is important. Around 47% of people die outside the hospital and it shows that they don't act on early warning signs.

Nowadays, lifespan of human beings is reduced because of heart diseases. So, World Health Organization (WHO) developed targets for prevention of non-communicable diseases (NCDs) in 2013, in which, 25% of relative reduction is from cardiovascular diseases and it is being ensured that at least 50% of patients with cardiovascular diseases have access to relevant drugs and medical counselling by 2025.

Around 17.9 million people died due to cardiovascular diseases in 2016, which is 31% of deaths around the world. A major challenge in heart diseases is its detection. It is difficult to predict that a person has a heart disease or not. There are instruments available which can predict heart diseases but they are either expensive or are not efficient to calculate the chance of heart disease in human. A survey of World Health Organization (WHO) says that medical professionals are able to predict just 67% of heart disease, so there is a vast scope of research in this field. In case of India, access to good doctors and hospitals in rural areas is very low. A 2016 WHO report says that, just 58% of the doctors have medical degree in urban areas and 19% in rural areas.

In USA, someone has a heart attack every 40 seconds, that is, more than one person dies in USA due to heart attack. Apart from this, Turkmenistan has the highest rate of deaths till 2012, with 712 deaths per 100,000 people. Kazakhstan has the second highest rate of deaths due to heart diseases. India holds 56th position in this series. Study also shows that, at ages 30-69 years, 1.3 million cardiovascular deaths, 0.9 million (68.4%) were caused by coronary heart disease and 0.4 million (28.0 %) by stroke Heart diseases are a major challenge in medical science, Machine Learning could be a good choice for predicting any heart disease in humans. Heart diseases can be predicted using Neural Network, Decision Tree, KNN, etc. Later in this report, we will see that how Logistic

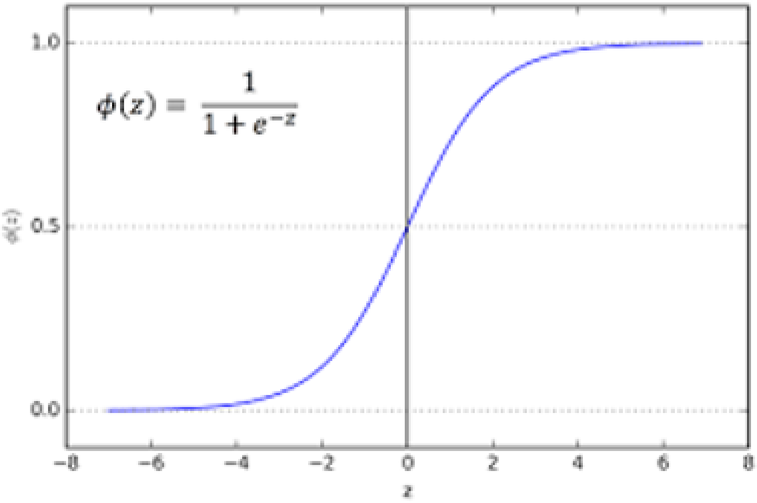
Regression is used to find the accuracy for heart disease. It also shows that how ML will help in our future for heart disease.

Specifically, we focus on the use of Shiny a web-based application framework for R to display our findings.

LOGISTIC REGRESSION

Logistic regression is one of the machine learning classification algorithms for analyzing a dataset in which there are one or more independent variables that determine an outcome. Categorical dependent Linear regression uses output in continuous numeric form whereas logistic regression transforms its output using the logistic sigmoid function to return a probability value which can then be mapped to two or more discrete classes.

Furthermore, logistic regression model uses more complex cost function (known as sigmoid function or logistic function) instead of linear function. Logistic regression limits the cost function between 0 and1.



In the formula, output is between 0 and 1 (probability estimate)

z = input to the function

e = base of natural log

$$\log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \ldots\ldots\ldots\ldots + \beta_n x_n$$

According to the given data set, 1 indicates the high risk of 10-year future coronary heart disease and 0 indicates no heart risks. The independent variables - n in the logistic model are as x1, x2, x3…, xn. Logistic regression achieves this by taking the log odds of the event ln(P/1−P), where, P is the probability of event which is the risk of CHD. Therefore, P always lies between 0 and 1.

DATASET

The dataset which used for the logistic regression analysis is available on the Kaggle website, and it is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal of this study is to predict whether the patient has 10-year risk of future coronary heart diseases. The Framingham dataset consists of 4240 records of patients. 15 independent variables, and predicted value with a total of 645 missing values. ML model is based on identification of dependent variable. It has used binary logistic regression which is one of the classification algorithms used due to target variable being categorical.

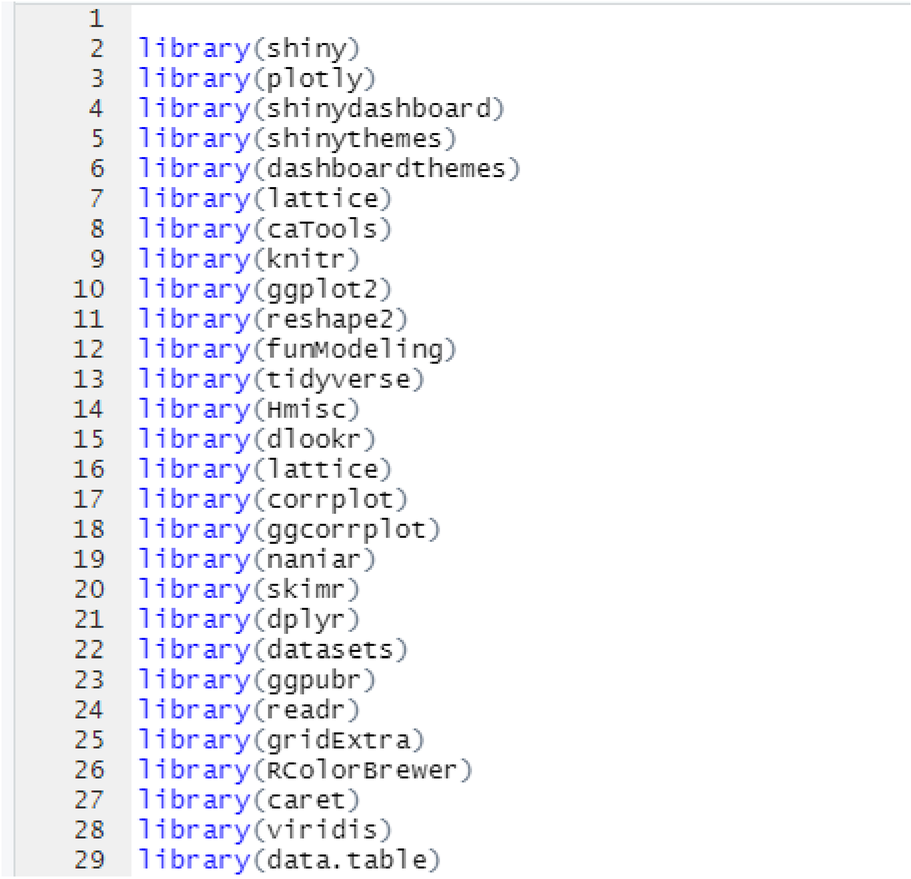
Variables			
Variable Category	Variable Name	Description	Data Type
Demographic	male	Male or female	Nominal
	age	Age of the patient	Continuous
	education	Education level ?	Nominal
Behaviour	currentSmoker	Current smoker or not?	Nominal
	cigsPerDay	Cigarettes per day?	Nominal
Medical History	BPMeds	Blood pressure medication?	Nominal
	prevalentStroke	Whether previously had stroke?	Nominal
	prevalentHyp	Whether was hypertensive?	Nominal
	diabetes	Whether had diabetes?	Nominal
Current Medical Status	totChol	Total Cholesterol Level	Continuous
	sysBP	Systolic Blood Pressure	Continuous
	diaBP	Diastolic Blood Pressure	Continuous
	BMI	Body Mass Index	Continuous
	heartRate	Heart Rate	Continuous
	glucose	Glucose Level	Continuous
Predicted Variable	TenYearCHD	10-year risk of CHD	Binary

DATA ANALYSIS

Data Analysis was carried out in R Studio using R. The following steps were implemented in order to process the logistics regression.

- **Loading Data and Other Required Libraries**

  The heart prediction data was loaded using Framingham CSV file into R studio in Order to build the logistic regression model. In addition to that, required libraries which used as supportive applications were loaded.



```r

library(shiny)
library(plotly)
library(shinydashboard)
library(shinythemes)
library(dashboardthemes)
library(lattice)
library(caTools)
library(knitr)
library(ggplot2)
library(reshape2)
library(funModeling)
library(tidyverse)
library(Hmisc)
library(dlookr)
library(lattice)
library(corrplot)
library(ggcorrplot)
library(naniar)
library(skimr)
library(dplyr)
library(datasets)
library(ggpubr)
library(readr)
library(gridExtra)
library(RColorBrewer)
library(caret)
library(viridis)
library(data.table)
```

PRE PROCESSING

❖ Exploratory Data Analysis

In statistics, exploratory data analysis is an approach to analysing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modelling or hypothesis testing task.

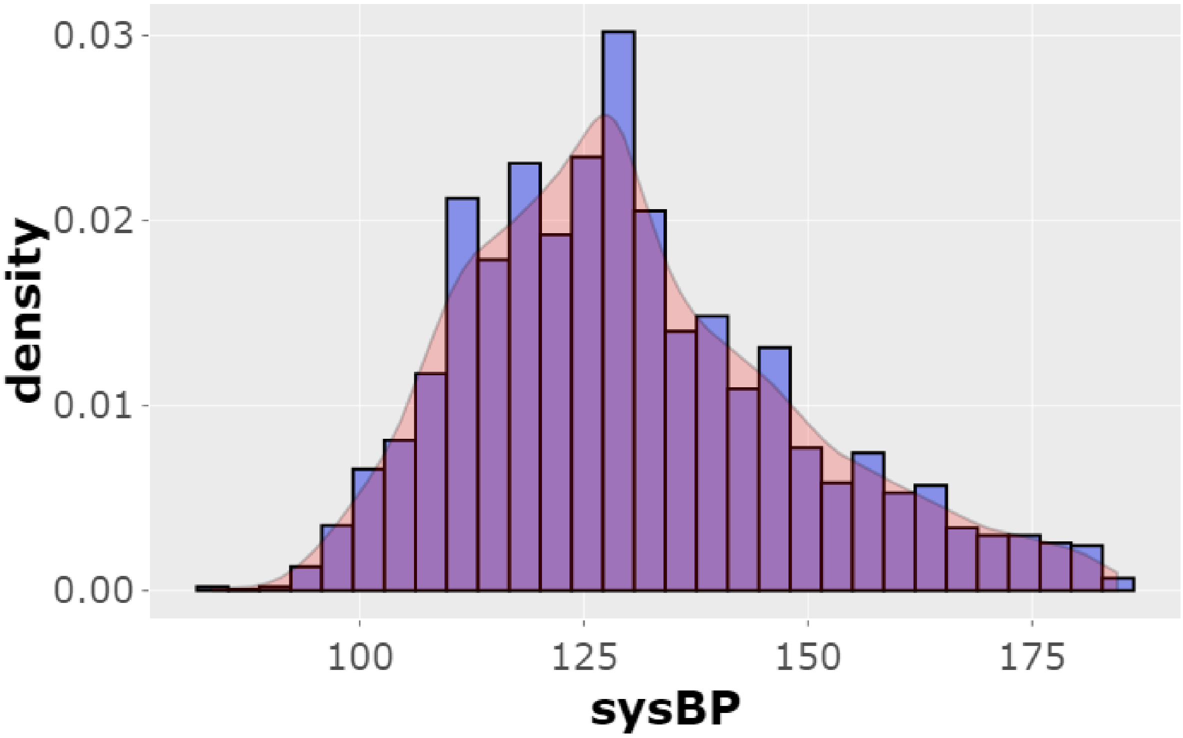
The following visualizations were derived through the RShiny for displaying predicators:

- **Univariate Plots**

A univariate plot shows the data and summarizes its distribution.

- **Histogram**

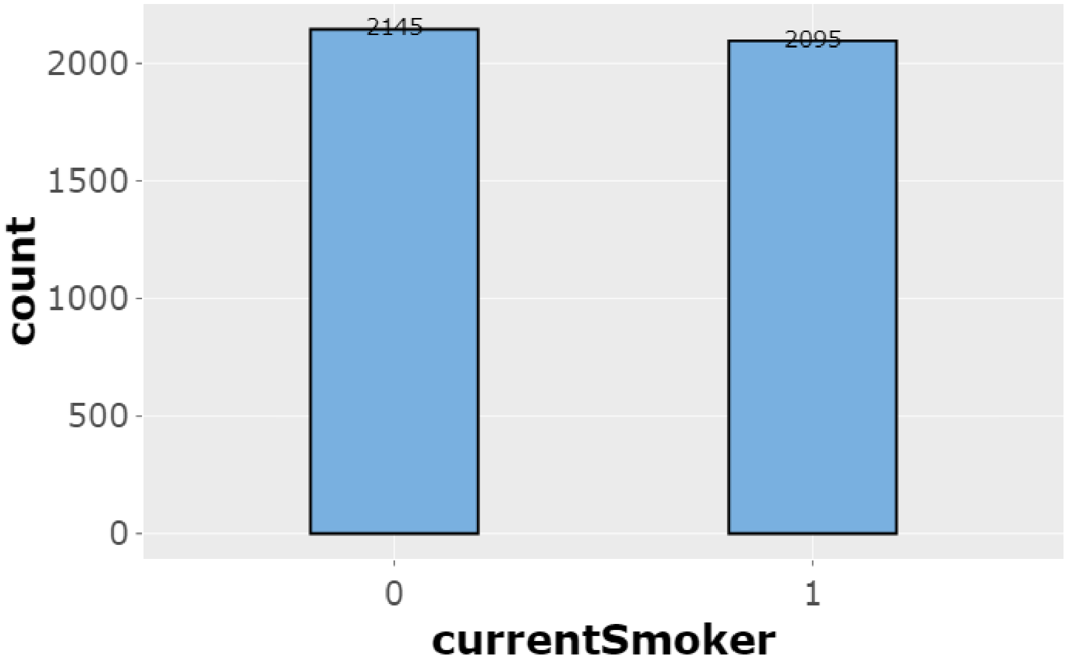
Histograms are a type of bar plot for numeric data that group the data into bins. After you create a Histogram object, you can modify aspects of the histogram by changing its property values.



Here, it tells us about the distribution of the variable sysBP. The histogram is normally distributed and right skewed.

- **Bar plot**

A bar graph shows comparisons among discrete categories. One axis of the chart shows the specific categories being compared, and the other axis represents a measured value.



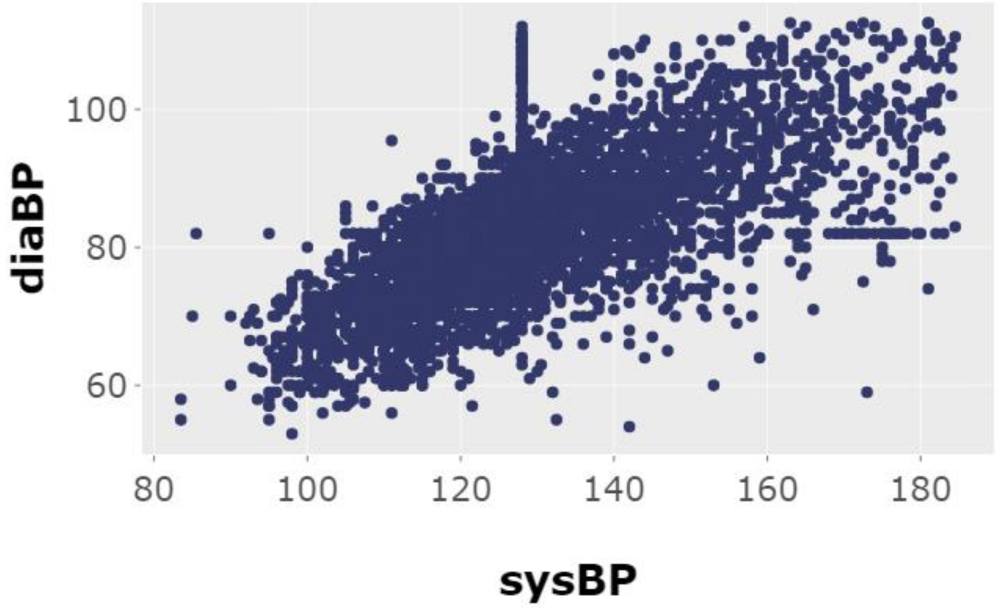
From the bar plot we can say that data is distributed approximately equally for the variable currentSmoker.

- **Bivariate Plots**

A bivariate plot graphs the relationship between two variables that have been measured on a single sample of subjects. Such a plot permits you to see at a glance the degree and pattern of relation between the two variables.

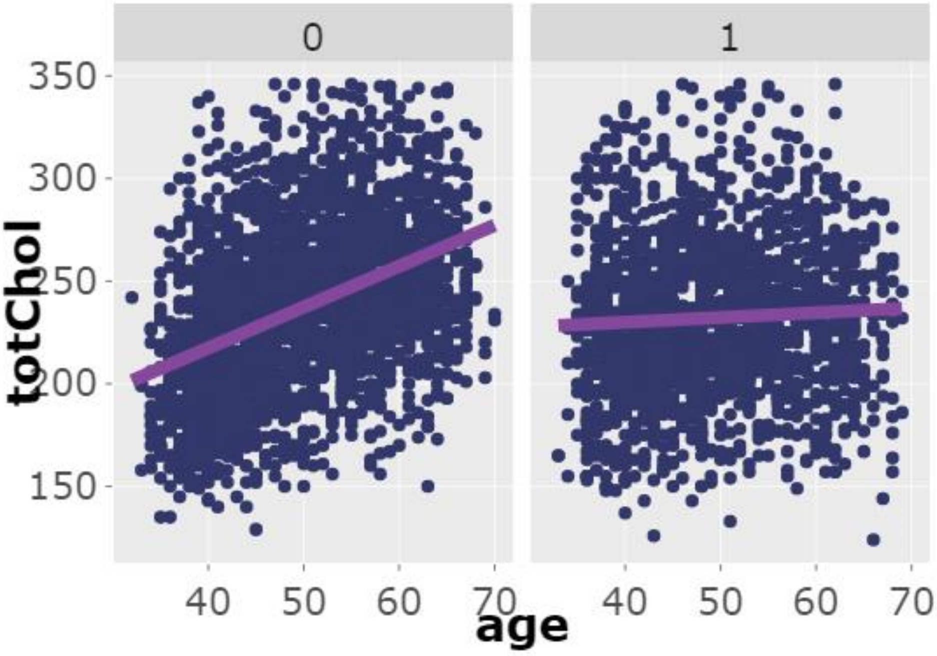
- **Scatter Plots**

A scatter plot uses dots to represent values for two different numeric variables. The position of each dot on the horizontal and vertical axis indicates values for an individual data point. Scatter plots are used to observe relationships between variables.



From the scatter plot we can conclude that there is a correlation between the two variables, sysBP and diaBP.

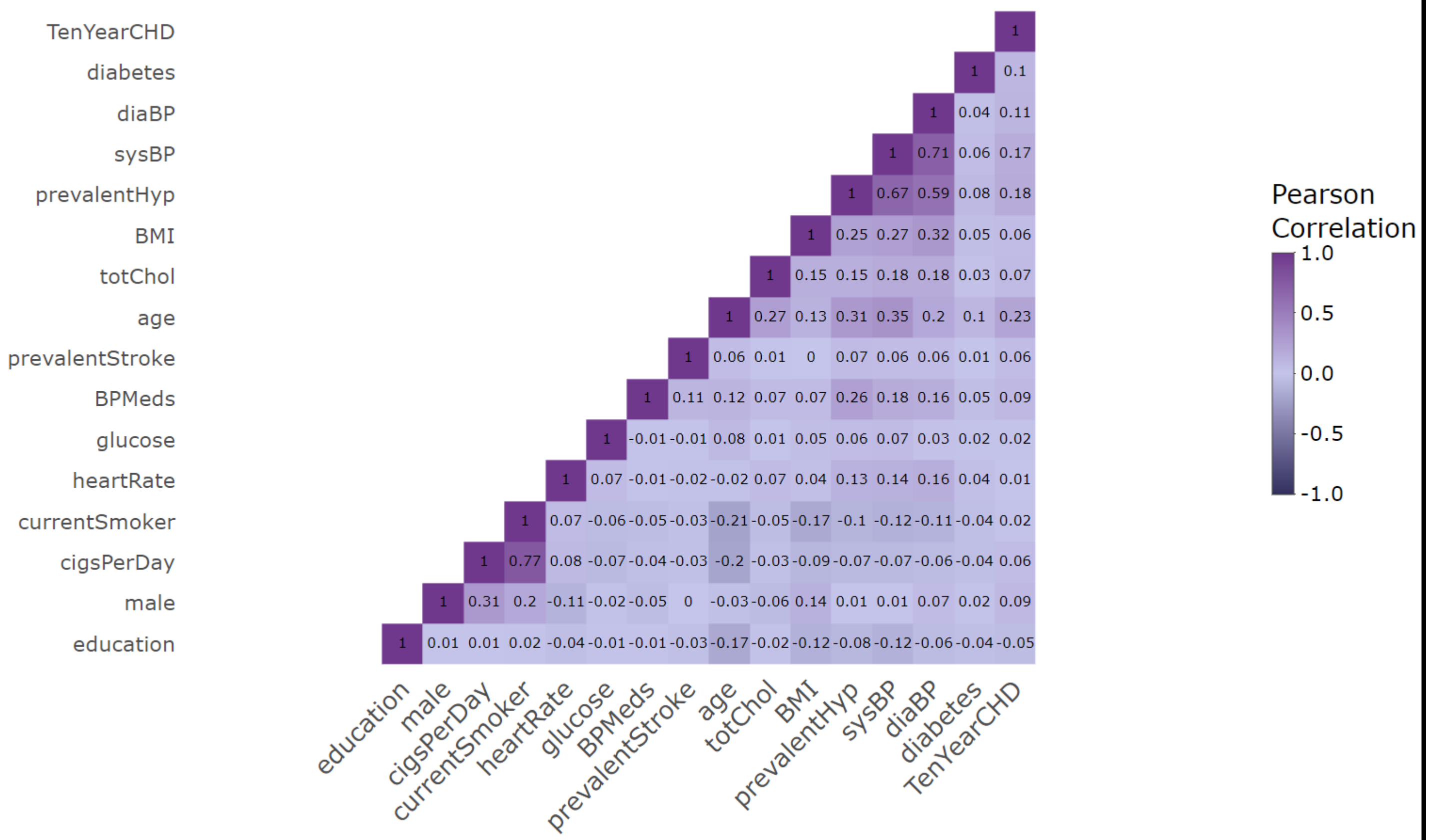
- **Categorical Bivariate Plots**



From the above graph we can infer that for females as age increases totChol also increases but for males it remains constant.

- **Heatmap**

A heat map (or heatmap) is a graphical representation of data where values are depicted by colour. Heat maps make it easy to visualize complex data and understand it at a glance.

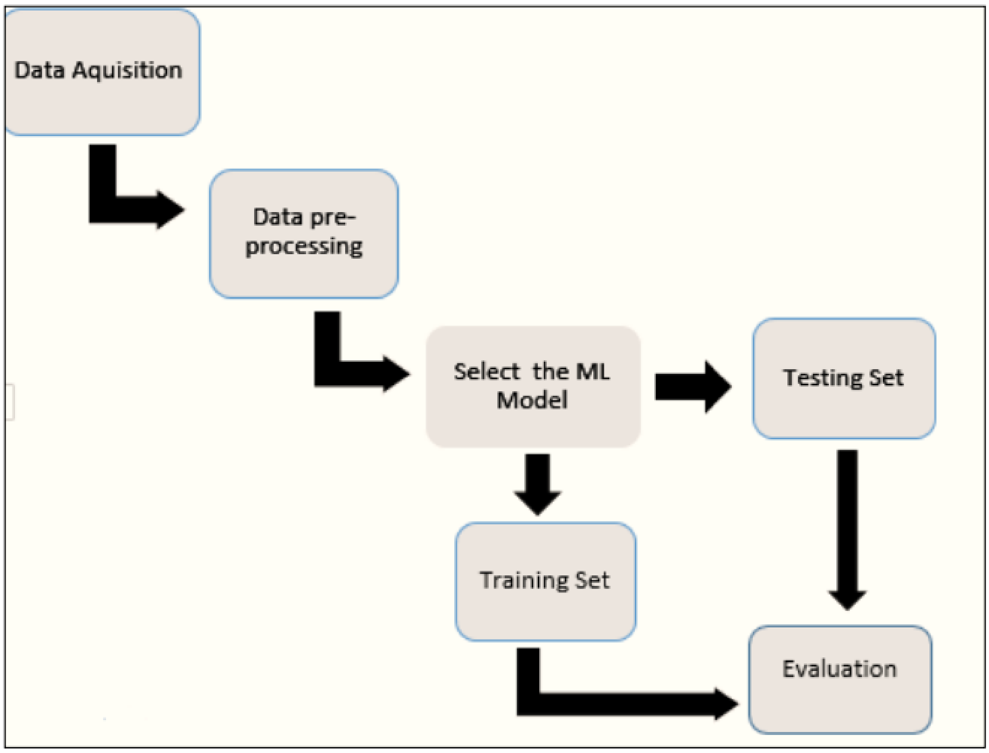


	education	male	cigsPerDay	currentSmoker	heartRate	glucose	BPMeds	prevalentStroke	age	totChol	BMI	prevalentHyp	sysBP	diaBP	diabetes	TenYearCHD
TenYearCHD																1
diabetes															1	0.1
diaBP														1	0.04	0.11
sysBP													1	0.71	0.06	0.17
prevalentHyp												1	0.67	0.59	0.08	0.18
BMI											1	0.25	0.27	0.32	0.05	0.06
totChol										1	0.15	0.15	0.18	0.18	0.03	0.07
age									1	0.27	0.13	0.31	0.35	0.2	0.1	0.23
prevalentStroke								1	0.06	0.01	0	0.07	0.06	0.06	0.01	0.06
BPMeds							1	0.11	0.12	0.07	0.07	0.26	0.18	0.16	0.05	0.09
glucose						1	-0.01	-0.01	0.08	0.01	0.05	0.06	0.07	0.03	0.02	0.02
heartRate					1	0.07	-0.01	-0.02	-0.02	0.07	0.04	0.13	0.14	0.16	0.04	0.01
currentSmoker				1	0.07	-0.06	-0.05	-0.03	-0.21	-0.05	-0.17	-0.1	-0.12	-0.11	-0.04	0.02
cigsPerDay			1	0.77	0.08	-0.07	-0.04	-0.03	-0.2	-0.03	-0.09	-0.07	-0.07	-0.06	-0.04	0.06
male		1	0.31	0.2	-0.11	-0.02	-0.05	0	-0.03	-0.06	0.14	0.01	0.01	0.07	0.02	0.09
education	1	0.01	0.01	0.02	-0.04	-0.01	-0.01	-0.03	-0.17	-0.02	-0.12	-0.08	-0.12	-0.06	-0.04	-0.05

From the heatmap we can say that there is a high correlation among prevalentHyp, sysBP and diaBP. This may lead to multi collinearity which can give a biased model. So we remove some of them.

METHODOLOGY

Workflow of Machine Learning Model Building indicates the steps followed in order to build the logistic regression model in machine learning.



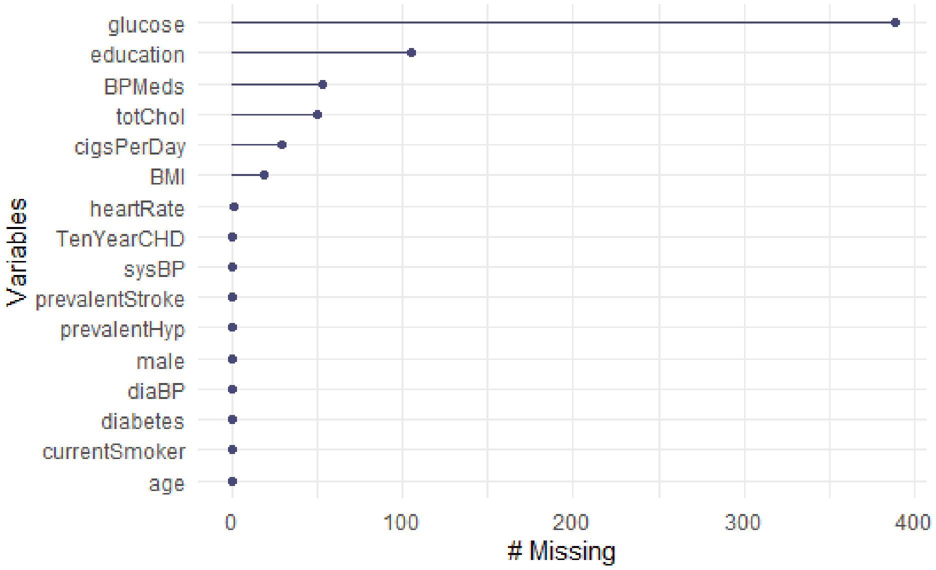
Workflow of Logistic Regression Model

- **DATA PRE-PROCESSING**

In order to build up more accurate ML model, data pre-processing is required. Data pre-process is the process of cleaning the data. This includes identification of missing data, noisy data and inconsistent data.

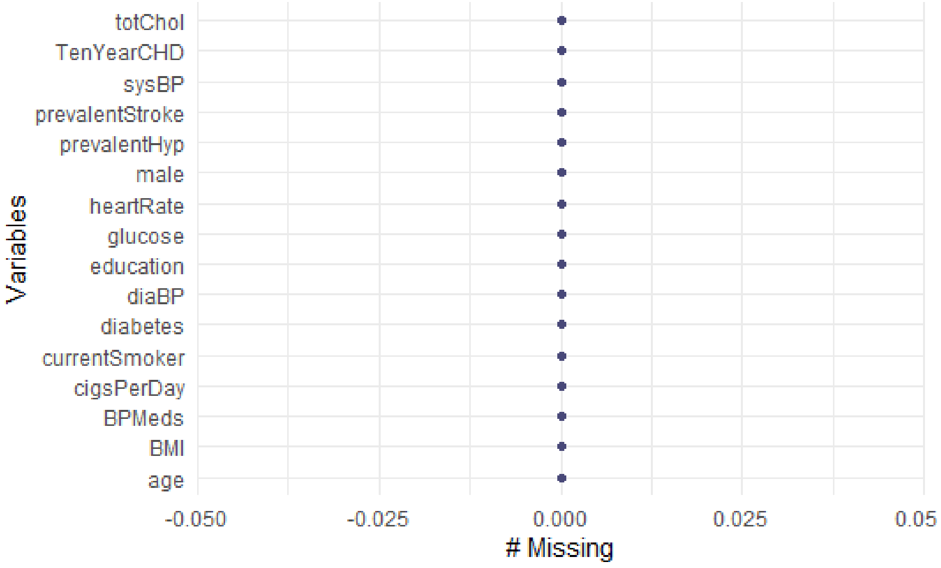
- **Identifying Missing Values**

Further, the number of missing values has identified for cleaning existing dataset. The summarized total number of missing values based on the attributes are given below.



Missing data in the training data set can reduce the power / fit of a model or can lead to a biased model because we have not analysed the behaviour and relationship with other variables correctly. It can lead to wrong prediction or classification.

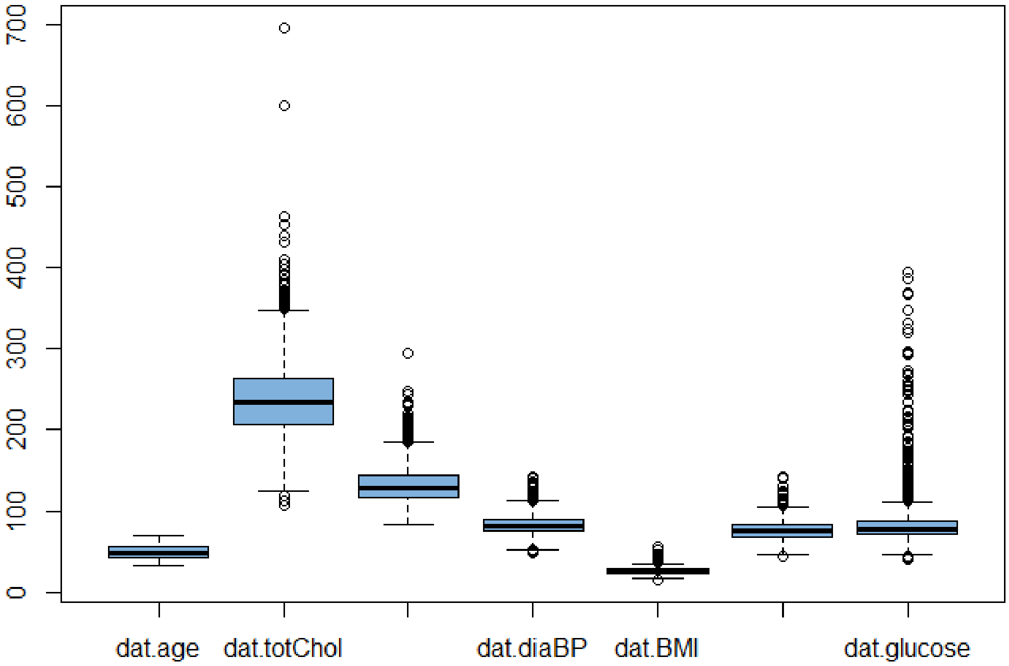
Mean/ Mode/ Median Imputation: Imputation is a method to fill in the missing values with estimated ones. The objective is to employ known relationships that can be identified in the valid values of the data set to assist in estimating the missing values. Mean / Mode / Median imputation is one of the most frequently used methods. It consists of replacing the missing data for a given attribute by the mean or median (quantitative attribute) or mode (qualitative attribute) of all known values of that variable.



- **Outlier Treatment**

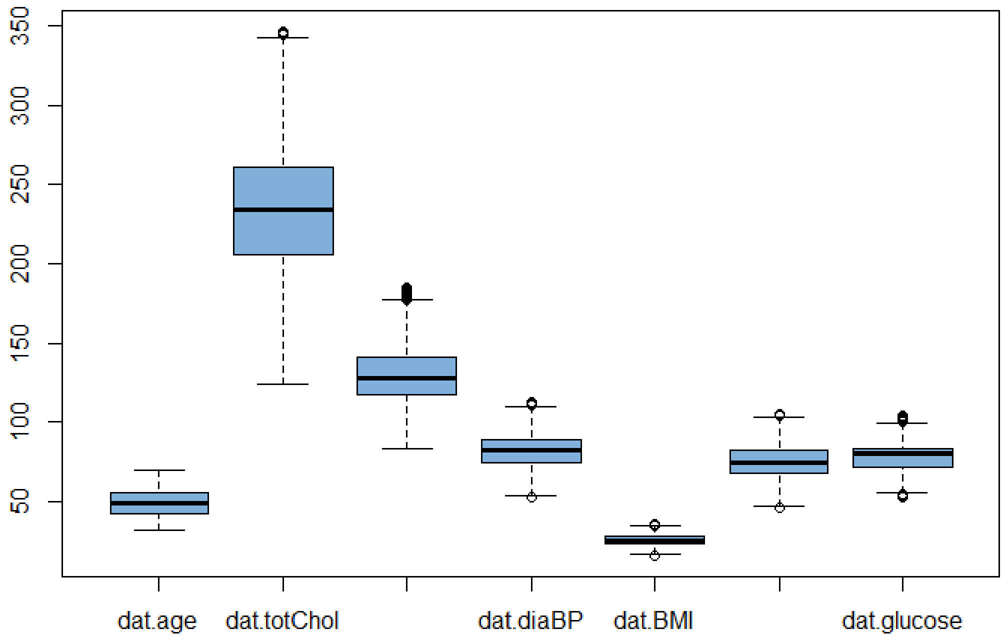
  Outliers can drastically change the results of the data analysis and statistical modelling. There are numerous unfavourable impacts of outliers in the data set:

  - It increases the error variance and reduces the power of statistical tests
  - If the outliers are non-randomly distributed, they can decrease normality
  - They can bias or influence estimates that may be of substantive interest
  - They can also impact the basic assumption of Regression, ANOVA and other statistical model assumptions.



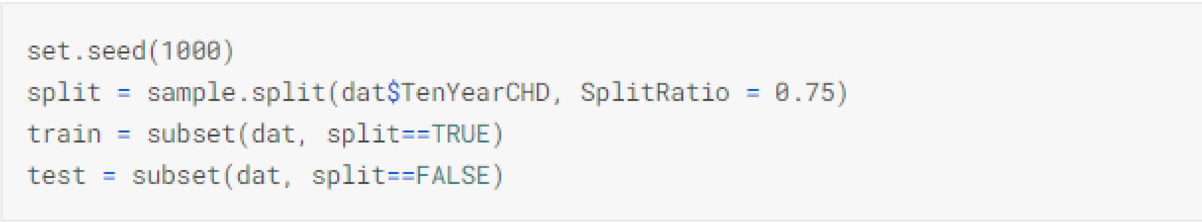
There are different methods used for detecting outliers, and we have made use of box plots. Box plots make use of the median and the lower and upper quartile.

Capping/Flooring Approach A value is identified as outlier if it exceeds the value of the 95th percentile of the variable by some factor, or if it is below the 5th percentile of given values by some factor. The factor is determined after considering the variable distribution and the business case. The outlier is then capped at a certain value above the P95 value or floored at a factor below the P5 value. The factor for capping/flooring is again obtained by studying the distribution of the variable and also accounting for any special business considerations.



- **MODEL**
  - **Splitting The Data**

The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model. It is a fast and easy procedure to perform, the results of which allow you to compare the performance of machine learning algorithms for your predictive modeling problem. Although simple to use and interpret, there are times when the procedure should not be used, such as when you have a small dataset and situations where additional configuration is required, such as when it is used for classification and the dataset is not balanced.



```
set.seed(1000)
split = sample.split(dat$TenYearCHD, SplitRatio = 0.75)
train = subset(dat, split==TRUE)
test = subset(dat, split==FALSE)
```

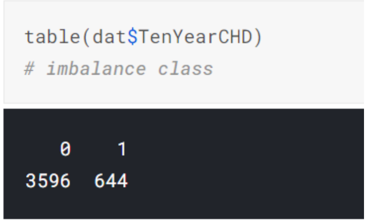
- **Up Sampling**

Imbalanced classes are a common problem in machine learning classification where there is a disproportionate ratio of observations in each class. Class imbalance can be found in many different areas including medical diagnosis, spam filtering, and fraud detection.

Most machine learning algorithms work best when the number of samples in each class are about equal. This is because most algorithms are designed to maximize accuracy and reduce error.

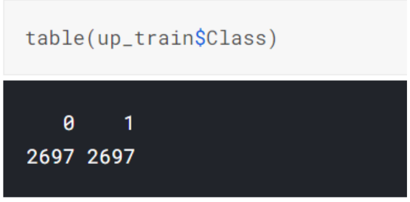
We can get an accuracy score of 90% and without even training a model!

Up-sampling is the process of randomly duplicating observations from the minority class in order to reinforce its signal. There are several heuristics for doing so, but the most common way is to simply resample with replacement.



```
table(dat$TenYearCHD)
# imbalance class
```

```
   0    1
3596  644
```

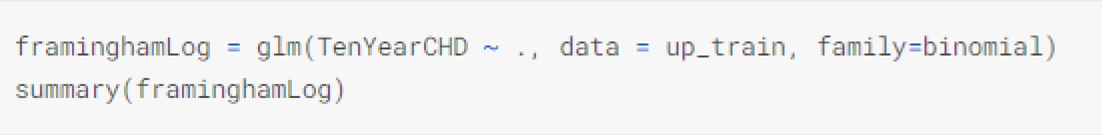


```
table(up_train$Class)
```

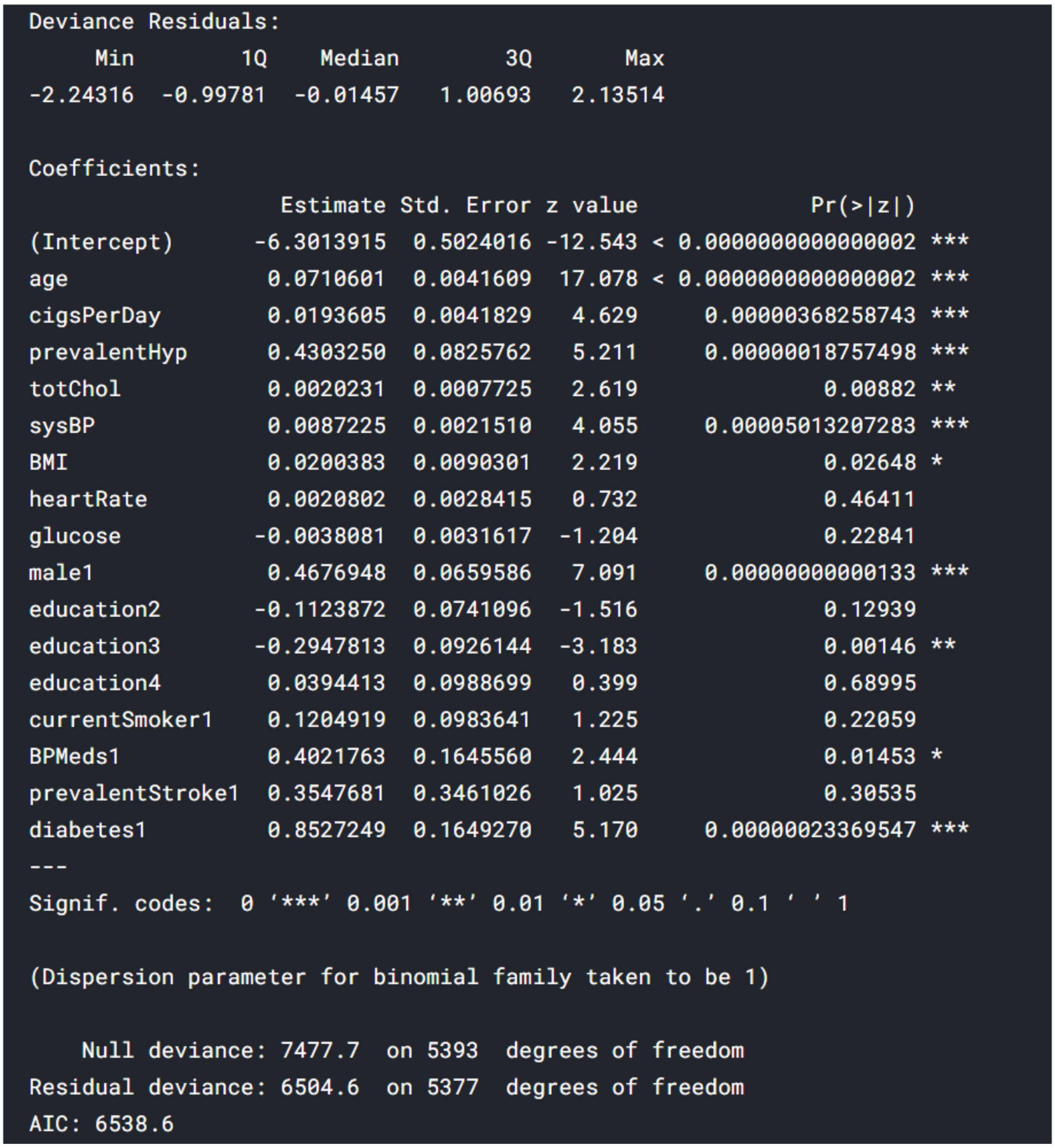
```
   0    1
2697 2697
```

Up sampling can be a good choice when you don’t have a ton of data to work with. So we have chosen up sampling instead of down sampling.

- **Creating the Model**

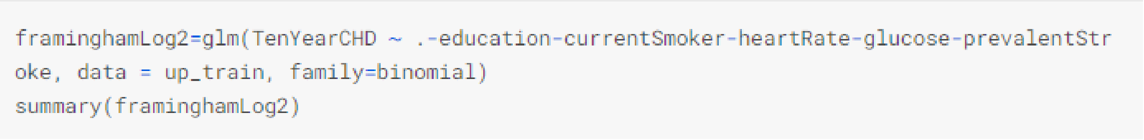


```
framinghamLog = glm(TenYearCHD ~ ., data = up_train, family=binomial)
summary(framinghamLog)
```

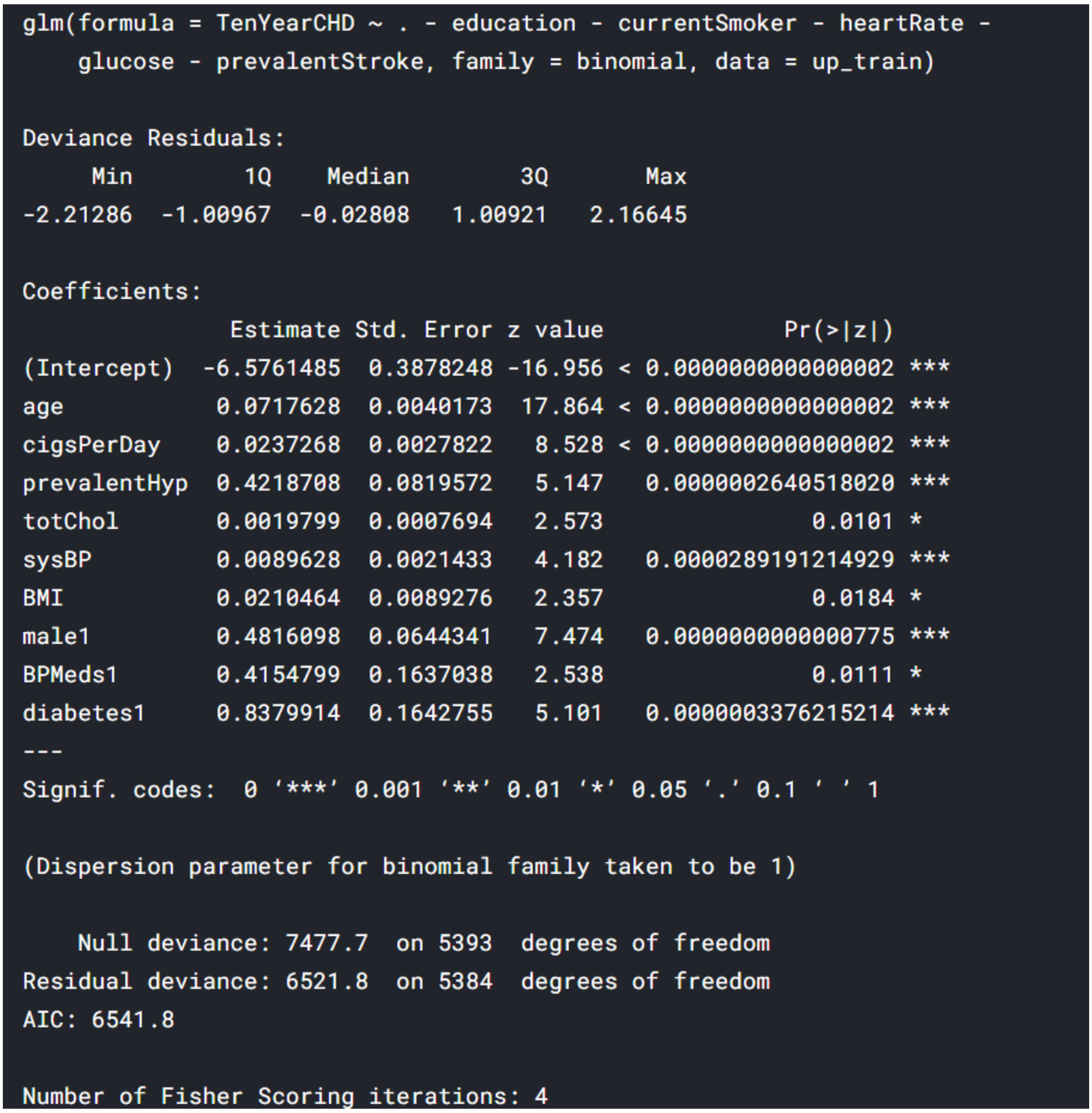


```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.24316  -0.99781  -0.01457   1.00693   2.13514

Coefficients:
                   Estimate Std. Error z value             Pr(>|z|)
(Intercept)      -6.3013915  0.5024016 -12.543 < 0.0000000000000002 ***
age               0.0710601  0.0041609  17.078 < 0.0000000000000002 ***
cigsPerDay        0.0193605  0.0041829   4.629     0.0000036825874 ***
prevalentHyp      0.4303250  0.0825762   5.211     0.0000001875749 ***
totChol           0.0020231  0.0007725   2.619             0.00882 **
sysBP             0.0087225  0.0021510   4.055     0.0000501320728 ***
BMI               0.0200383  0.0090301   2.219             0.02648 *
heartRate         0.0020802  0.0028415   0.732             0.46411
glucose          -0.0038081  0.0031617  -1.204             0.22841
male1             0.4676948  0.0659586   7.091     0.0000000000133 ***
education2       -0.1123872  0.0741096  -1.516             0.12939
education3       -0.2947813  0.0926144  -3.183             0.00146 **
education4        0.0394413  0.0988699   0.399             0.68995
currentSmoker1    0.1204919  0.0983641   1.225             0.22059
BPMeds1           0.4021763  0.1645560   2.444             0.01453 *
prevalentStroke1  0.3547681  0.3461026   1.025             0.30535
diabetes1         0.8527249  0.1649270   5.170     0.0000002336954 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7477.7  on 5393  degrees of freedom
Residual deviance: 6504.6  on 5377  degrees of freedom
AIC: 6538.6
```

According to the above logistic results P >= 0.05 show low statistically significance relationship with probability of heart disease. Therefore, backward elimination approach has been used to remove the attributes with highest P values. We remove variables one by one by comparing the AIC value of each model. If the AIC value is reduced then only we remove the variable.



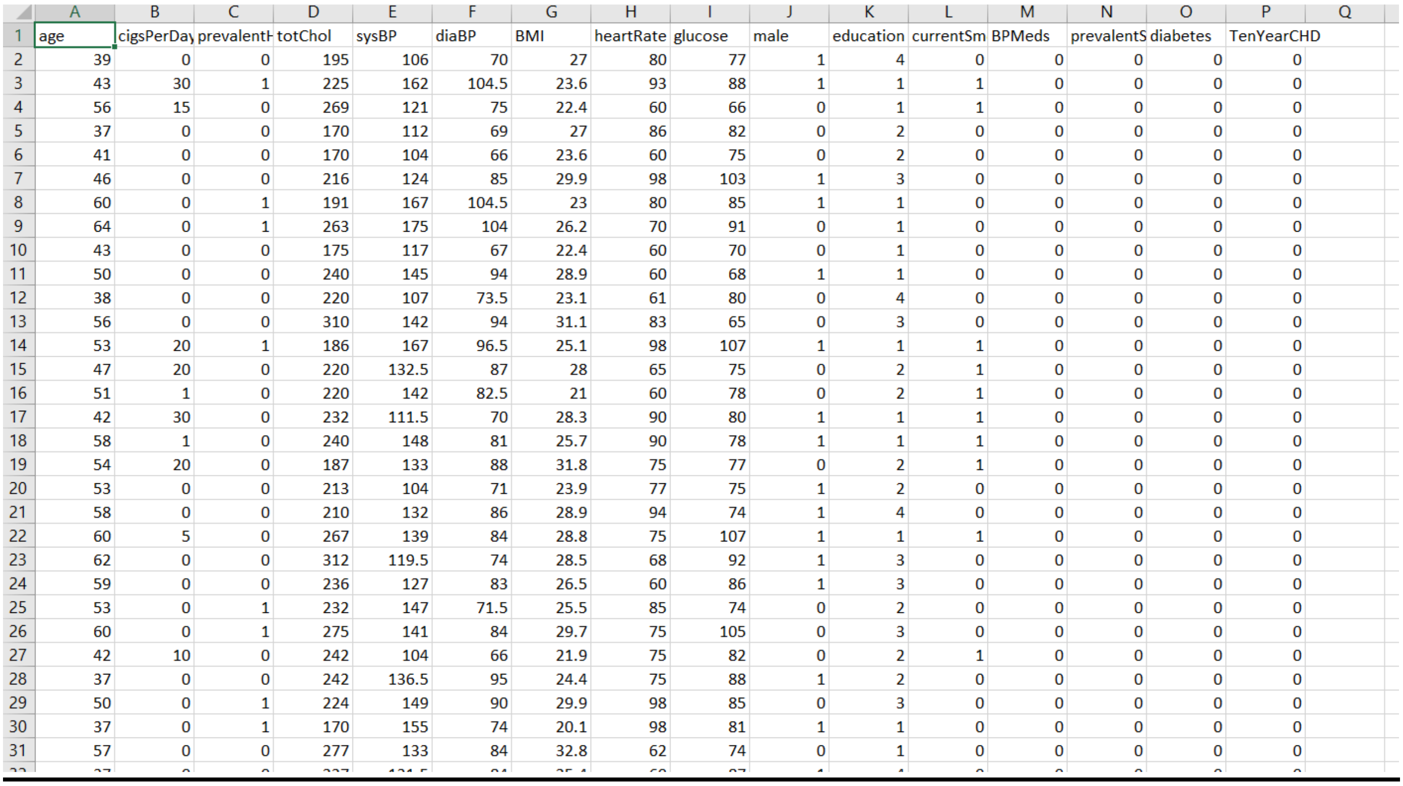
```
framinghamLog2=glm(TenYearCHD ~ .-education-currentSmoker-heartRate-glucose-prevalentStr
oke, data = up_train, family=binomial)
summary(framinghamLog2)
```



```
glm(formula = TenYearCHD ~ . - education - currentSmoker - heartRate -
    glucose - prevalentStroke, family = binomial, data = up_train)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.21286  -1.00967  -0.02808   1.00921   2.16645

Coefficients:
               Estimate Std. Error z value             Pr(>|z|)
(Intercept)  -6.5761485  0.3878248 -16.956 < 0.0000000000000002 ***
age           0.0717628  0.0040173  17.864 < 0.0000000000000002 ***
cigsPerDay    0.0237268  0.0027822   8.528 < 0.0000000000000002 ***
prevalentHyp  0.4218708  0.0819572   5.147   0.0000002640518020 ***
totChol       0.0019799  0.0007694   2.573               0.0101 *
sysBP         0.0089628  0.0021433   4.182   0.0000289191214929 ***
BMI           0.0210464  0.0089276   2.357               0.0184 *
male1         0.4816098  0.0644341   7.474   0.0000000000000775 ***
BPMeds1       0.4154799  0.1637038   2.538               0.0111 *
diabetes1     0.8379914  0.1642755   5.101   0.0000003376215214 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7477.7  on 5393  degrees of freedom
Residual deviance: 6521.8  on 5384  degrees of freedom
AIC: 6541.8

Number of Fisher Scoring iterations: 4
```

The above output indicates the result after using backward elimination.

DATASET



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	age	cigsPerDay	prevalentH	totChol	sysBP	diaBP	BMI	heartRate	glucose	male	education	currentSm	BPMeds	prevalentS	diabetes	TenYearCHD	
2	39	0	0	195	106	70	27	80	77	1	4	0	0	0	0	0	
3	43	30	1	225	162	104.5	23.6	93	88	1	1	1	0	0	0	0	
4	56	15	0	269	121	75	22.4	60	66	0	1	1	0	0	0	0	
5	37	0	0	170	112	69	27	86	82	0	2	0	0	0	0	0	
6	41	0	0	170	104	66	23.6	60	75	0	2	0	0	0	0	0	
7	46	0	0	216	124	85	29.9	98	103	1	3	0	0	0	0	0	
8	60	0	1	191	167	104.5	23	80	85	1	1	0	0	0	0	0	
9	64	0	1	263	175	104	26.2	70	91	0	1	0	0	0	0	0	
10	43	0	0	175	117	67	22.4	60	70	0	1	0	0	0	0	0	
11	50	0	0	240	145	94	28.9	60	68	1	1	0	0	0	0	0	
12	38	0	0	220	107	73.5	23.1	61	80	0	4	0	0	0	0	0	
13	56	0	0	310	142	94	31.1	83	65	0	3	0	0	0	0	0	
14	53	20	1	186	167	96.5	25.1	98	107	1	1	1	0	0	0	0	
15	47	20	0	220	132.5	87	28	65	75	0	2	1	0	0	0	0	
16	51	1	0	220	142	82.5	21	60	78	0	2	1	0	0	0	0	
17	42	30	0	232	111.5	70	28.3	90	80	1	1	1	0	0	0	0	
18	58	1	0	240	148	81	25.7	90	78	1	1	1	0	0	0	0	
19	54	20	0	187	133	88	31.8	75	77	0	2	1	0	0	0	0	
20	53	0	0	213	104	71	23.9	77	75	1	2	0	0	0	0	0	
21	58	0	0	210	132	86	28.9	94	74	1	4	0	0	0	0	0	
22	60	5	0	267	139	84	28.8	75	107	1	1	1	0	0	0	0	
23	62	0	0	312	119.5	74	28.5	68	92	1	3	0	0	0	0	0	
24	59	0	0	236	127	83	26.5	60	86	1	3	0	0	0	0	0	
25	53	0	1	232	147	71.5	25.5	85	74	0	2	0	0	0	0	0	
26	60	0	1	275	141	84	29.7	75	105	0	3	0	0	0	0	0	
27	42	10	0	242	104	66	21.9	75	82	0	2	1	0	0	0	0	
28	37	0	0	242	136.5	95	24.4	75	88	1	2	0	0	0	0	0	
29	50	0	1	224	149	90	29.9	98	85	0	3	0	0	0	0	0	
30	37	0	1	170	155	74	20.1	98	81	1	1	0	0	0	0	0	
31	57	0	0	277	133	84	32.8	62	74	0	1	0	0	0	0	0	

RESULT

The logistic regression equation for the heart prediction data is as follows.
logit(p)=log(p/(1−p))=β0 + β1∗male + β2∗age + β3∗cigsPerDay + β4∗totChol + β5∗sysBP + β6∗BMI + β7*BPMeds + β8*diabetes

Accuracy is not the best metric to use when evaluating imbalanced datasets as it can be very misleading. Metrics that can provide better insight include:

- **Confusion Matrix:** a table showing correct predictions and types of incorrect predictions. The segments of the confusion matrix indicate the following parameters.

  True Positives (TP): cases which are predicted yes (they have the disease), and they do have the disease.

  True Negatives (TN): cases which are predicted no, and they do not have the disease.

  False Positives (FP): cases which are predicted yes, but they do not actually have the disease (Type I error).

  False Negatives (FN): cases which are predicted no, but they actually do have the disease (Type II error).

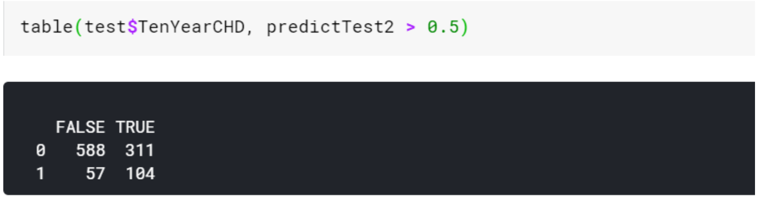
- **Precision:** the number of true positives divided by all positive predictions. Precision is also called Positive Predictive Value. It is a measure of a classifier's exactness. Low precision indicates a high number of false positives.

- **Recall:** the number of true positives divided by the number of positive values in the test data. Recall is also called Sensitivity or the True Positive Rate. It is a measure of a classifier's completeness. Low recall indicates a high number of false negatives.

- **F1: Score:** the weighted average of precision and recall.

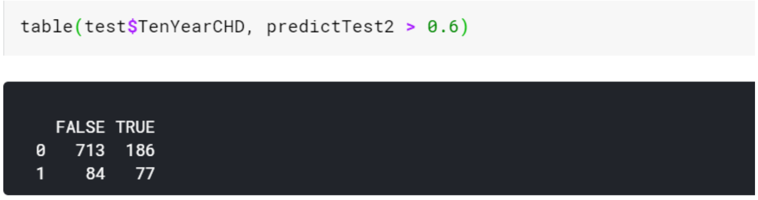
Since the model is predicting Heart disease too many type II errors is not advisable. A False Negative (ignoring the probability of disease when there actualy is one) is more dangerous than a

False Positive in this case. The model is more specific than sensitive. Hence inorder to increase the sensitivity(recall value), threshold can bealtered.



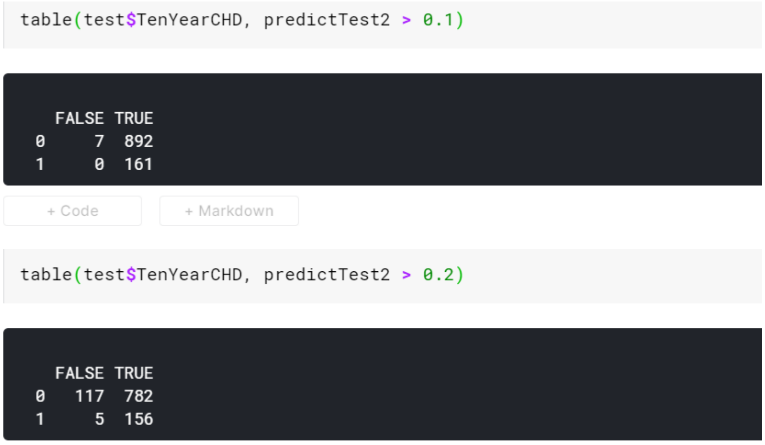
```
table(test$TenYearCHD, predictTest2 > 0.5)
```

```
  FALSE TRUE
0   588  311
1    57  104
```



```
table(test$TenYearCHD, predictTest2 > 0.6)
```

```
  FALSE TRUE
0   713  186
1    84   77
```

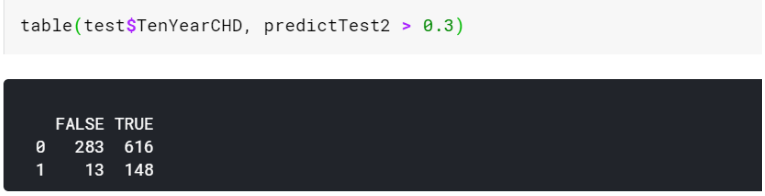


```
table(test$TenYearCHD, predictTest2 > 0.1)
```

```
  FALSE TRUE
0     7  892
1     0  161
```

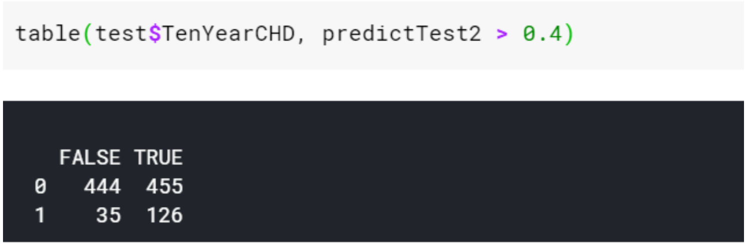
```
table(test$TenYearCHD, predictTest2 > 0.2)
```

```
  FALSE TRUE
0   117  782
1     5  156
```



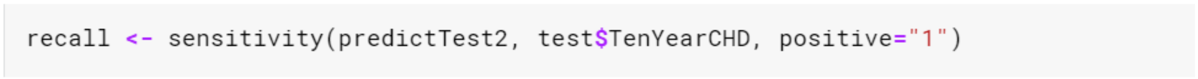
```
table(test$TenYearCHD, predictTest2 > 0.3)
```

```
  FALSE TRUE
0   283  616
1    13  148
```

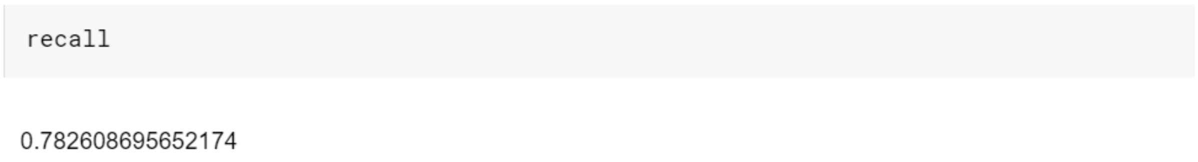


```
table(test$TenYearCHD, predictTest2 > 0.4)
```

```
  FALSE TRUE
0   444  455
1    35  126
```

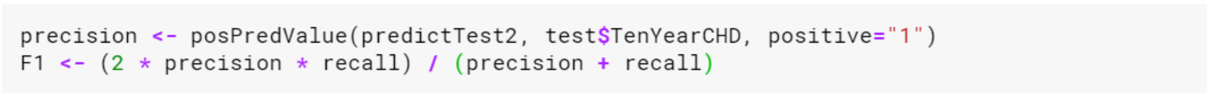


```
recall <- sensitivity(predictTest2, test$TenYearCHD, positive="1")
```

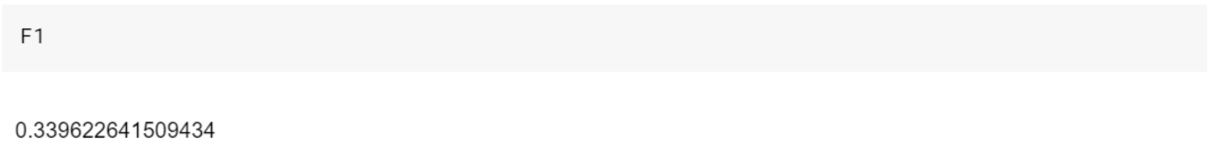


```
recall
```

```
0.782608695652174
```

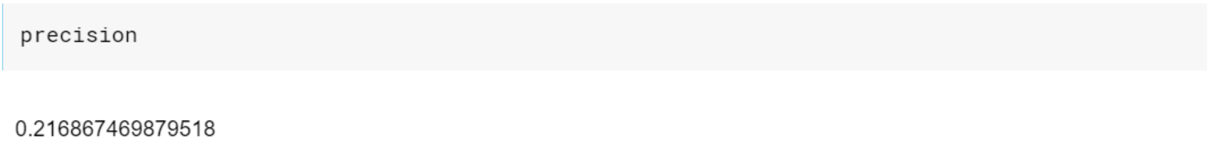


```
precision <- posPredValue(predictTest2, test$TenYearCHD, positive="1")
F1 <- (2 * precision * recall) / (precision + recall)
```



```
F1
```

```
0.339622641509434
```



```
precision
```

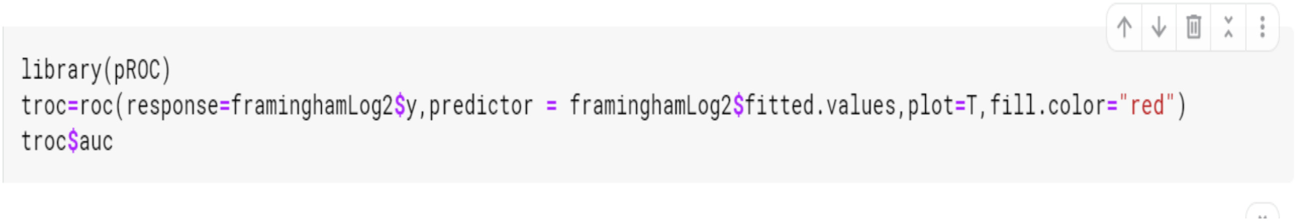
```
0.216867469879518
```

The above confusion matrix has a high recall value hence we can solve our problem statement in better way.

The model could differentiate between low risk patients and high risk patients pretty well.

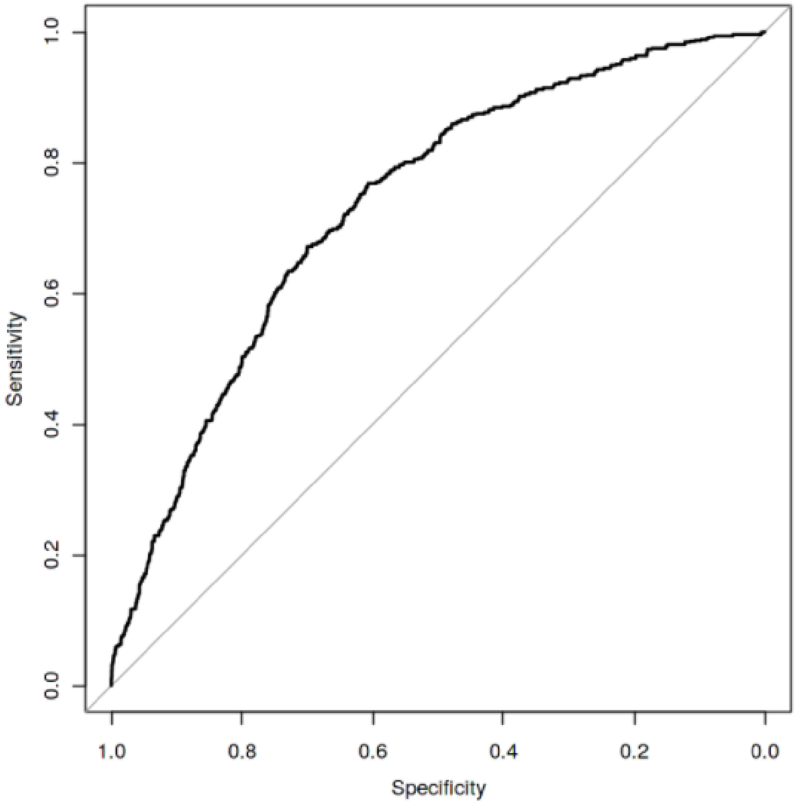
- **AUC-ROC Curve**

  It is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC represents degree or measure of separation. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s. By analogy, higher the AUC, better the model is at distinguishing between patients with disease and no disease.



```
library(pROC)
troc=roc(response=framinghamLog2$y,predictor = framinghamLog2$fitted.values,plot=T,fill.color="red")
troc$auc
```

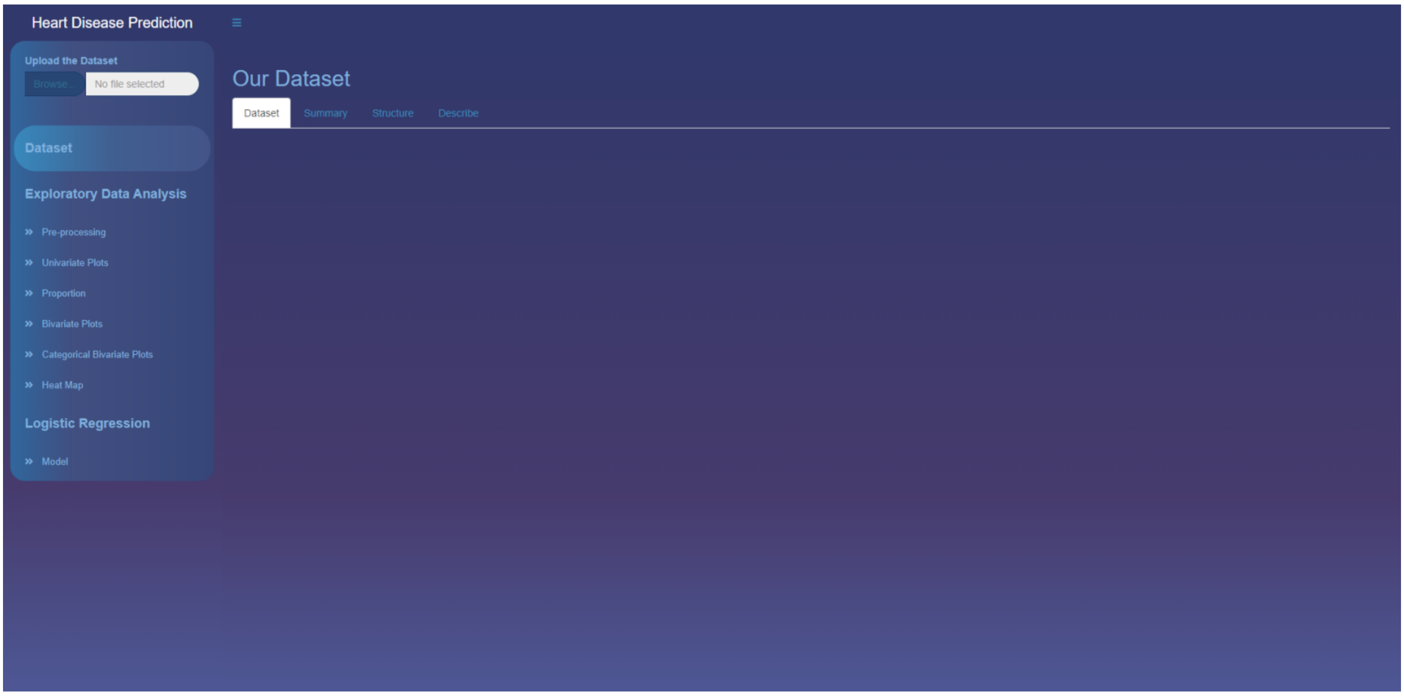
```
0.734939836886011
```



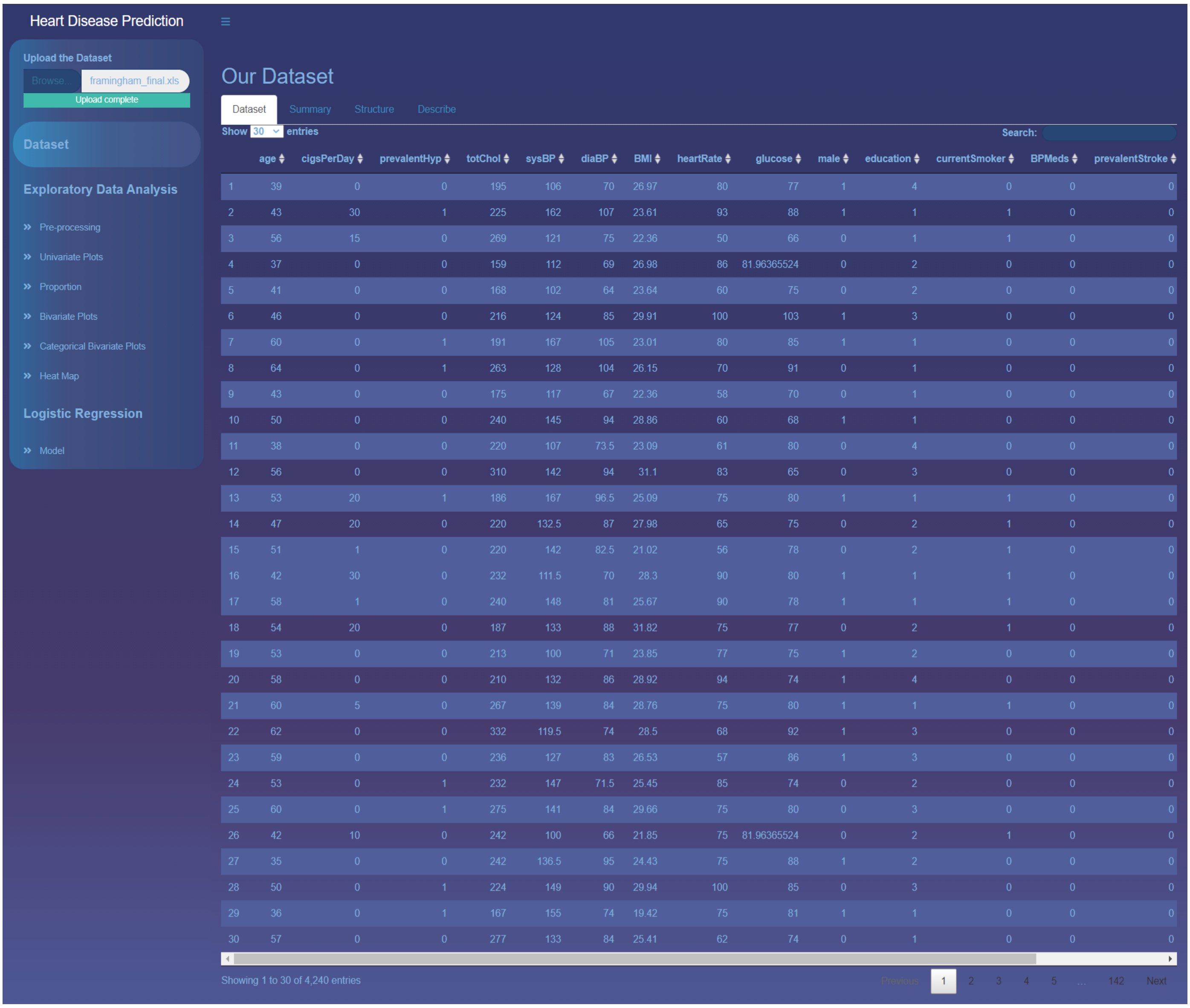
R-SHINY

Shiny is an open source R package that provides an elegant and powerful web framework for building web applications using R. This means we can use all of R's extensive (and extensible) data analysis and visualization features in our app. Essentially, we can take almost any analysis we've done in R, and then make it interactive. We can run our apps locally, within R Studio, make them standalone, either by deploying them to a Shiny server, or to a hosting service, or even including them in a Markdown document. Shiny helps you turn your analyses into interactive web applications without requiring HTML, CSS, or JavaScript knowledge.

Home page



Dataset



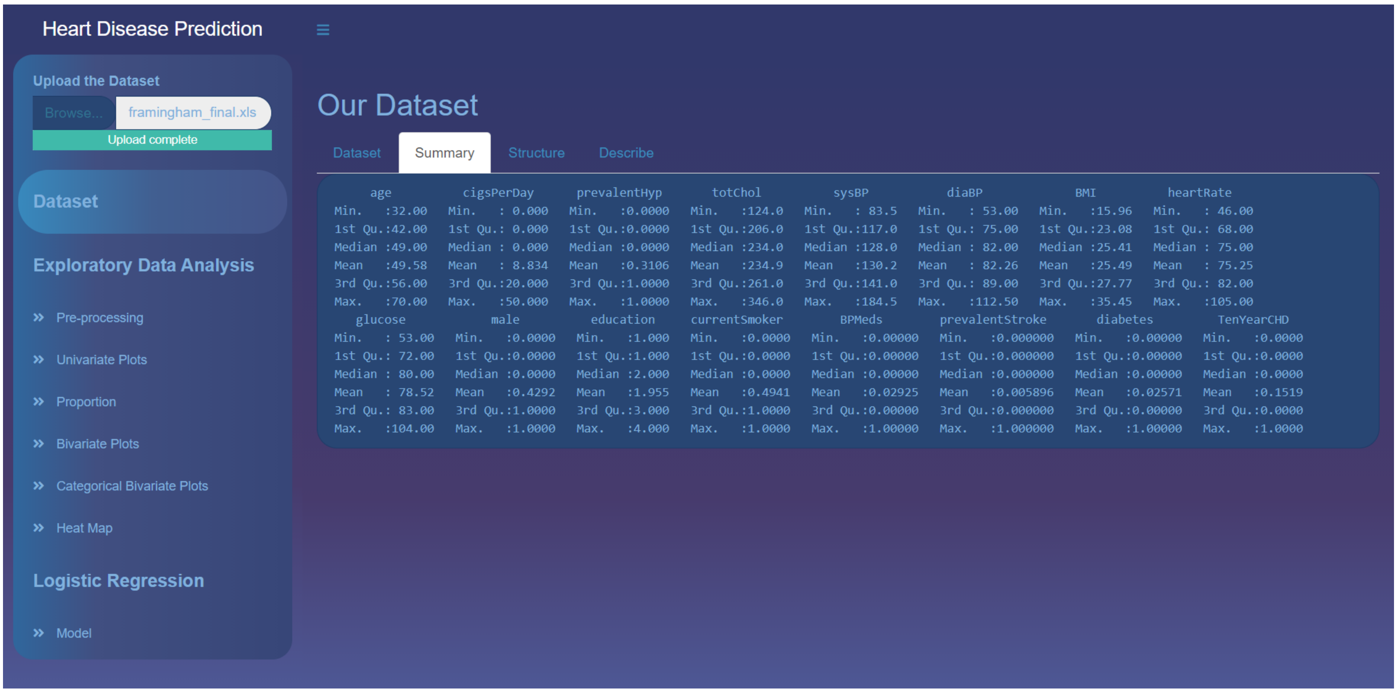
Our Dataset

Show 30 entries

	age	cigsPerDay	prevalentHyp	totChol	sysBP	diaBP	BMI	heartRate	glucose	male	education	currentSmoker	BPMeds	prevalentStroke
1	39	0	0	195	106	70	26.97	80	77	1	4	0	0	0
2	43	30	1	225	162	107	23.61	93	88	1	1	1	0	0
3	56	15	0	269	121	75	22.36	50	66	0	1	1	0	0
4	37	0	0	159	112	69	26.98	86	81.96365524	0	2	0	0	0
5	41	0	0	168	102	64	23.64	60	75	0	2	0	0	0
6	46	0	0	216	124	85	29.91	100	103	1	3	0	0	0
7	60	0	1	191	167	105	23.01	80	85	1	1	0	0	0
8	64	0	1	263	128	104	26.15	70	91	0	1	0	0	0
9	43	0	0	175	117	67	22.36	58	70	0	1	0	0	0
10	50	0	0	240	145	94	28.86	60	68	1	1	0	0	0
11	38	0	0	220	107	73.5	23.09	61	80	0	4	0	0	0
12	56	0	0	310	142	94	31.1	83	65	0	3	0	0	0
13	53	20	1	186	167	96.5	25.09	75	80	1	1	1	0	0
14	47	20	0	220	132.5	87	27.98	65	75	0	2	1	0	0
15	51	1	0	220	142	82.5	21.02	56	78	0	2	1	0	0
16	42	30	0	232	111.5	70	28.3	90	80	1	1	1	0	0
17	58	1	0	240	148	81	25.67	90	78	1	1	1	0	0
18	54	20	0	187	133	88	31.82	75	77	0	2	1	0	0
19	53	0	0	213	100	71	23.85	77	75	1	2	0	0	0
20	58	0	0	210	132	86	28.92	94	74	1	4	0	0	0
21	60	5	0	267	139	84	28.76	75	80	1	1	1	0	0
22	62	0	0	332	119.5	74	28.5	68	92	1	3	0	0	0
23	59	0	0	236	127	83	26.53	57	86	1	3	0	0	0
24	53	0	1	232	147	71.5	25.45	85	74	0	2	0	0	0
25	60	0	1	275	141	84	29.66	75	80	0	3	0	0	0
26	42	10	0	242	100	66	21.85	75	81.96365524	0	2	1	0	0
27	35	0	0	242	136.5	95	24.43	75	88	1	2	0	0	0
28	50	0	1	224	149	90	29.94	100	85	0	3	0	0	0
29	36	0	1	167	155	74	19.42	75	81	1	1	0	0	0
30	57	0	0	277	133	84	25.41	62	74	0	1	0	0	0

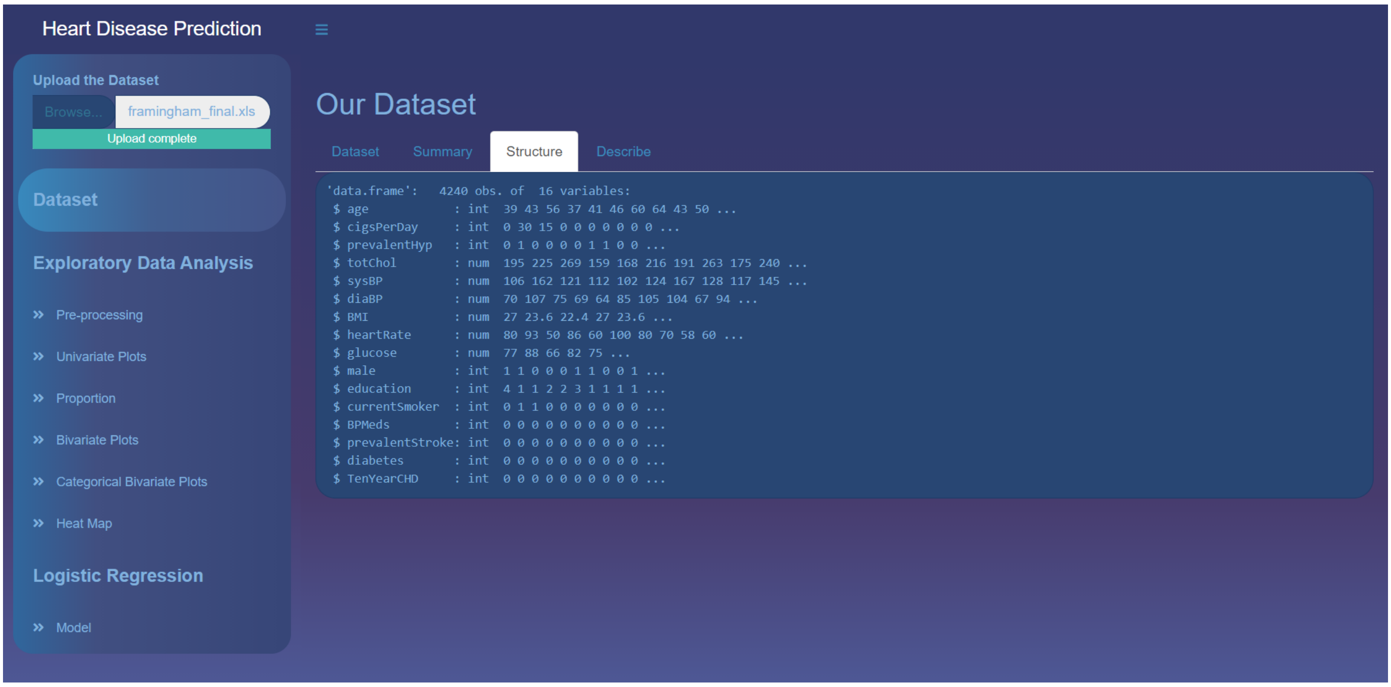
Showing 1 to 30 of 4,240 entries

Summary



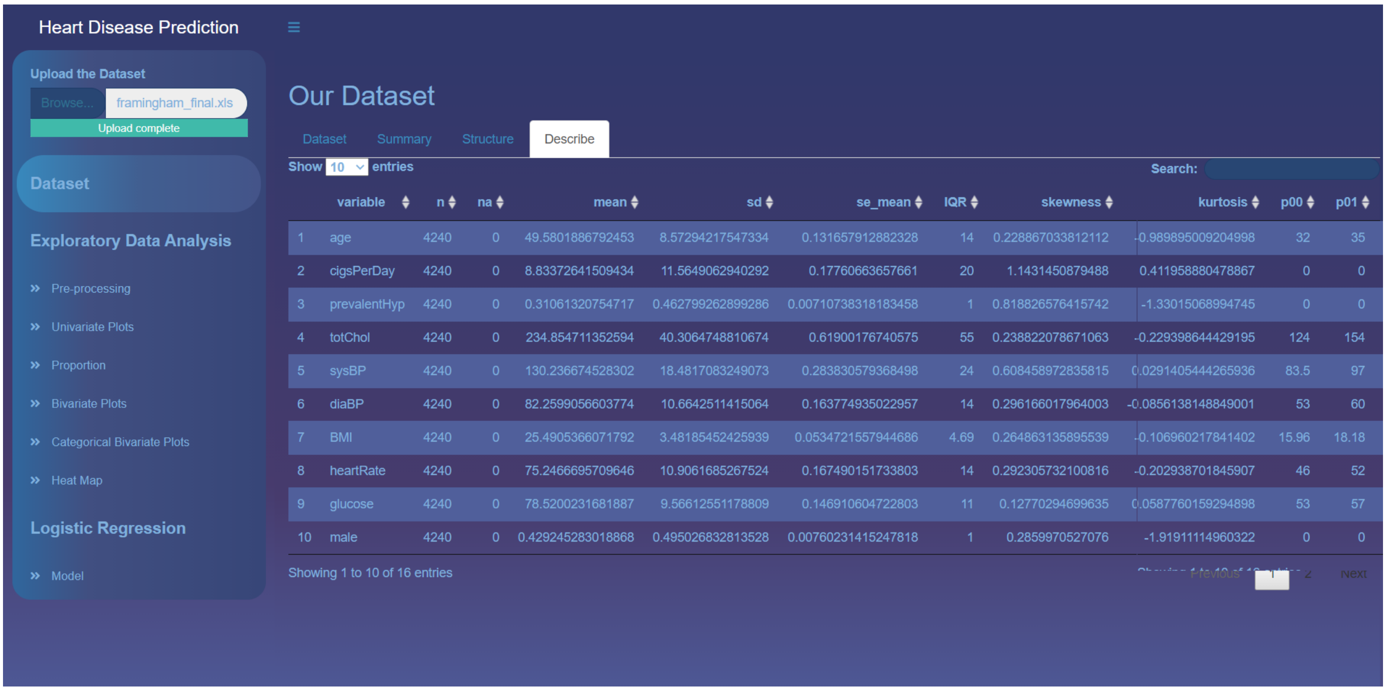
```
      age          cigsPerDay      prevalentHyp       totChol          sysBP           diaBP             BMI          heartRate     
 Min.   :32.00   Min.   : 0.000   Min.   :0.0000   Min.   :124.0   Min.   : 83.5   Min.   : 53.00   Min.   :15.96   Min.   : 46.00  
 1st Qu.:42.00   1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.:206.0   1st Qu.:117.0   1st Qu.: 75.00   1st Qu.:23.08   1st Qu.: 68.00  
 Median :49.00   Median : 0.000   Median :0.0000   Median :234.0   Median :128.0   Median : 82.00   Median :25.41   Median : 75.00  
 Mean   :49.58   Mean   : 8.834   Mean   :0.3106   Mean   :234.9   Mean   :130.2   Mean   : 82.26   Mean   :25.49   Mean   : 75.25  
 3rd Qu.:56.00   3rd Qu.:20.000   3rd Qu.:1.0000   3rd Qu.:261.0   3rd Qu.:141.0   3rd Qu.: 89.00   3rd Qu.:27.77   3rd Qu.: 82.00  
 Max.   :70.00   Max.   :50.000   Max.   :1.0000   Max.   :346.0   Max.   :184.5   Max.   :112.50   Max.   :35.45   Max.   :105.00  
    glucose            male          education     currentSmoker        BPMeds        prevalentStroke        diabetes         TenYearCHD    
 Min.   : 53.00   Min.   :0.0000   Min.   :1.000   Min.   :0.0000   Min.   :0.00000   Min.   :0.000000   Min.   :0.00000   Min.   :0.0000  
 1st Qu.: 72.00   1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.00000   1st Qu.:0.0000  
 Median : 80.00   Median :0.0000   Median :2.000   Median :0.0000   Median :0.00000   Median :0.000000   Median :0.00000   Median :0.0000  
 Mean   : 78.52   Mean   :0.4292   Mean   :1.955   Mean   :0.4941   Mean   :0.02925   Mean   :0.005896   Mean   :0.02571   Mean   :0.1519  
 3rd Qu.: 83.00   3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:0.00000   3rd Qu.:0.0000  
 Max.   :104.00   Max.   :1.0000   Max.   :4.000   Max.   :1.0000   Max.   :1.00000   Max.   :1.000000   Max.   :1.00000   Max.   :1.0000  
```

Structure



```
'data.frame':	4240 obs. of  16 variables:
 $ age            : int  39 43 56 37 41 46 60 64 43 50 ...
 $ cigsPerDay     : int  0 30 15 0 0 0 0 0 0 0 ...
 $ prevalentHyp   : int  0 1 0 0 0 0 1 1 0 0 ...
 $ totChol        : num  195 225 269 159 168 216 191 263 175 240 ...
 $ sysBP          : num  106 162 121 112 102 124 167 128 117 145 ...
 $ diaBP          : num  70 107 75 69 64 85 105 104 67 94 ...
 $ BMI            : num  27 23.6 22.4 27 23.6 ...
 $ heartRate      : num  80 93 50 86 60 100 80 70 58 60 ...
 $ glucose        : num  77 88 66 82 75 ...
 $ male           : int  1 1 0 0 0 1 1 0 0 1 ...
 $ education      : int  4 1 1 2 2 3 1 1 1 1 ...
 $ currentSmoker  : int  0 1 1 0 0 0 0 0 0 0 ...
 $ BPMeds         : int  0 0 0 0 0 0 0 0 0 0 ...
 $ prevalentStroke: int  0 0 0 0 0 0 0 0 0 0 ...
 $ diabetes       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ TenYearCHD     : int  0 0 0 0 0 0 0 0 0 0 ...
```

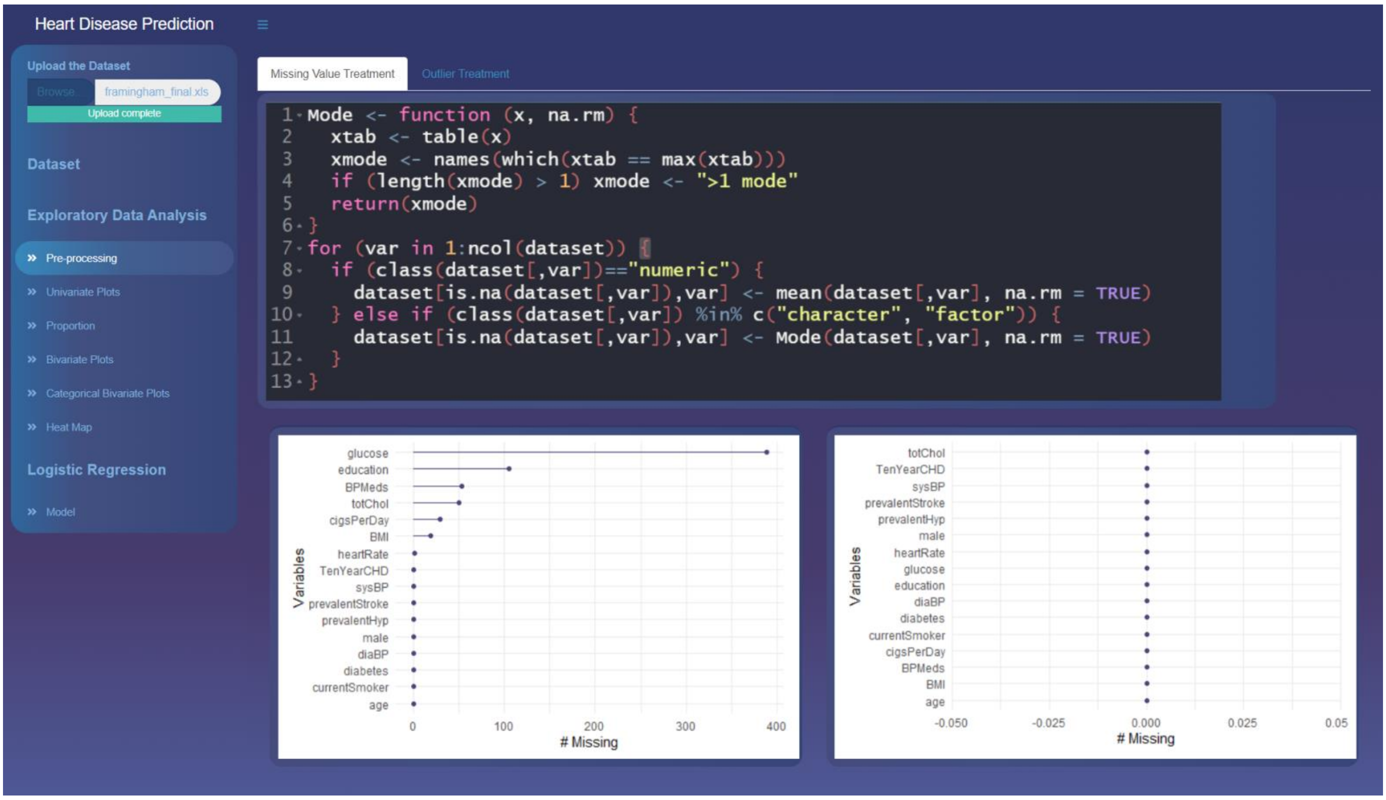
Describe



	variable	n	na	mean	sd	se_mean	IQR	skewness	kurtosis	p00	p01
1	age	4240	0	49.5801886792453	8.57294217547334	0.131657912882328	14	0.228867033812112	-0.989895009204998	32	35
2	cigsPerDay	4240	0	8.83372641509434	11.5649062940292	0.17760663657661	20	1.1431450879488	0.411958880478867	0	0
3	prevalentHyp	4240	0	0.31061320754717	0.462799262899286	0.00710738318183458	1	0.818826576415742	-1.33015068994745	0	0
4	totChol	4240	0	234.854711352594	40.3064748810674	0.61900176740575	55	0.238822078671063	-0.229398644429195	124	154
5	sysBP	4240	0	130.236674528302	18.4817083249073	0.283830579368498	24	0.608458972835815	0.0291405444265936	83.5	97
6	diaBP	4240	0	82.2599056603774	10.6642511415064	0.163774935022957	14	0.296166017964003	-0.0856138148849001	53	60
7	BMI	4240	0	25.4905366071792	3.48185452425939	0.0534721557944686	4.69	0.264863135895539	-0.106960217841402	15.96	18.18
8	heartRate	4240	0	75.2466695709646	10.9061685267524	0.167490151733803	14	0.292305732100816	-0.202938701845907	46	52
9	glucose	4240	0	78.5200231681887	9.56612551178809	0.146910604722803	11	0.12770294699635	0.0587760159294898	53	57
10	male	4240	0	0.429245283018868	0.495026832813528	0.00760231415247818	1	0.2859970527076	-1.91911114960322	0	0

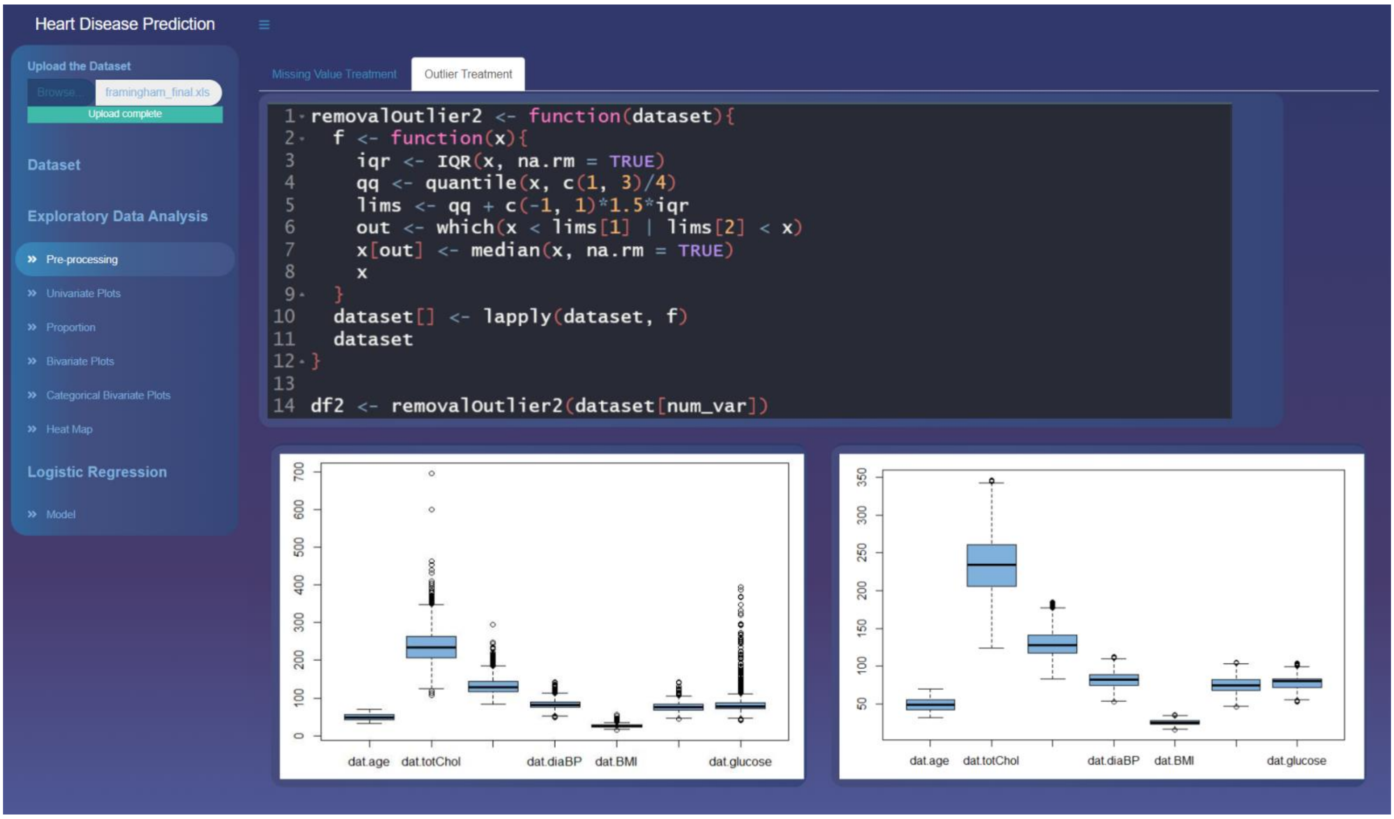
Showing 1 to 10 of 16 entries

Pre-processing – Missing Value Treatment



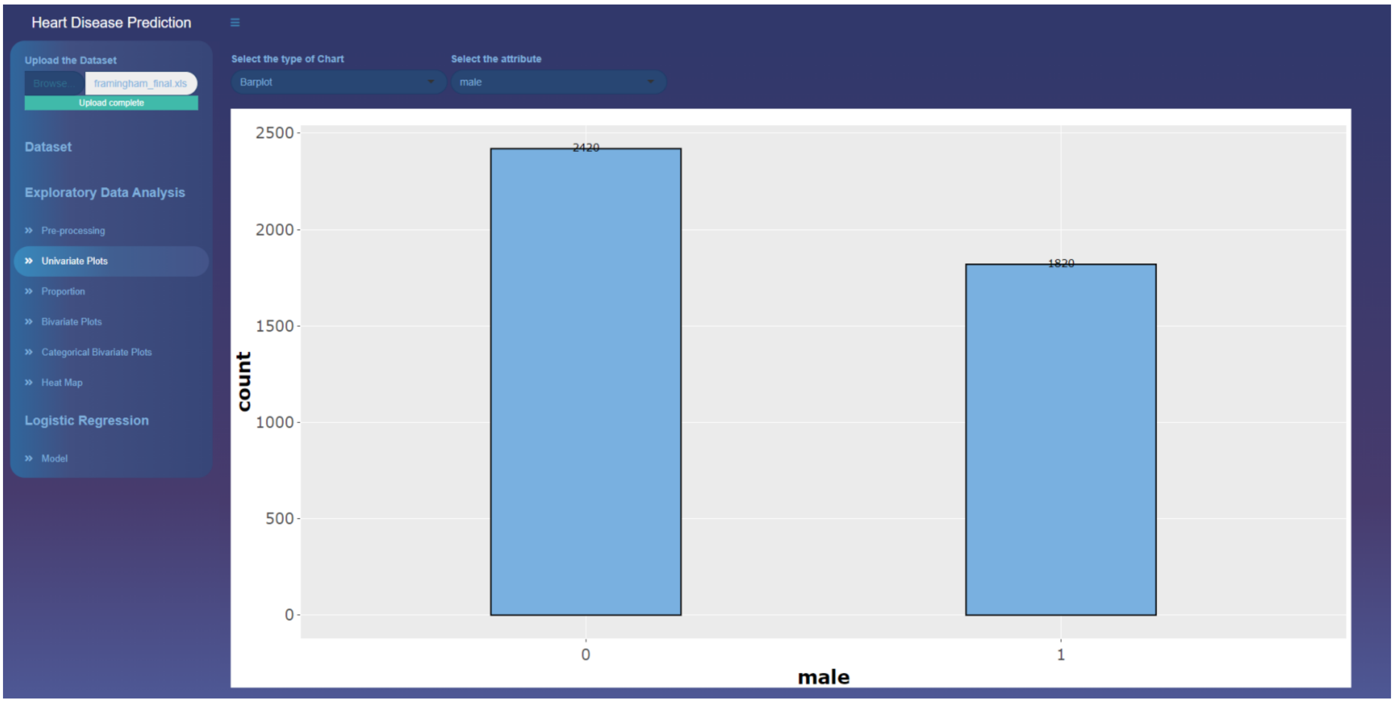
```
Mode <- function (x, na.rm) {
  xtab <- table(x)
  xmode <- names(which(xtab == max(xtab)))
  if (length(xmode) > 1) xmode <- ">1 mode"
  return(xmode)
}
for (var in 1:ncol(dataset)) {
  if (class(dataset[,var])=="numeric") {
    dataset[is.na(dataset[,var]),var] <- mean(dataset[,var], na.rm = TRUE)
  } else if (class(dataset[,var]) %in% c("character", "factor")) {
    dataset[is.na(dataset[,var]),var] <- Mode(dataset[,var], na.rm = TRUE)
  }
}
```

Pre-processing – Outlier Treatment



```
removalOutlier2 <- function(dataset){
  f <- function(x){
    iqr <- IQR(x, na.rm = TRUE)
    qq <- quantile(x, c(1, 3)/4)
    lims <- qq + c(-1, 1)*1.5*iqr
    out <- which(x < lims[1] | lims[2] < x)
    x[out] <- median(x, na.rm = TRUE)
    x
  }
  dataset[] <- lapply(dataset, f)
  dataset
}

df2 <- removalOutlier2(dataset[num_var])
```

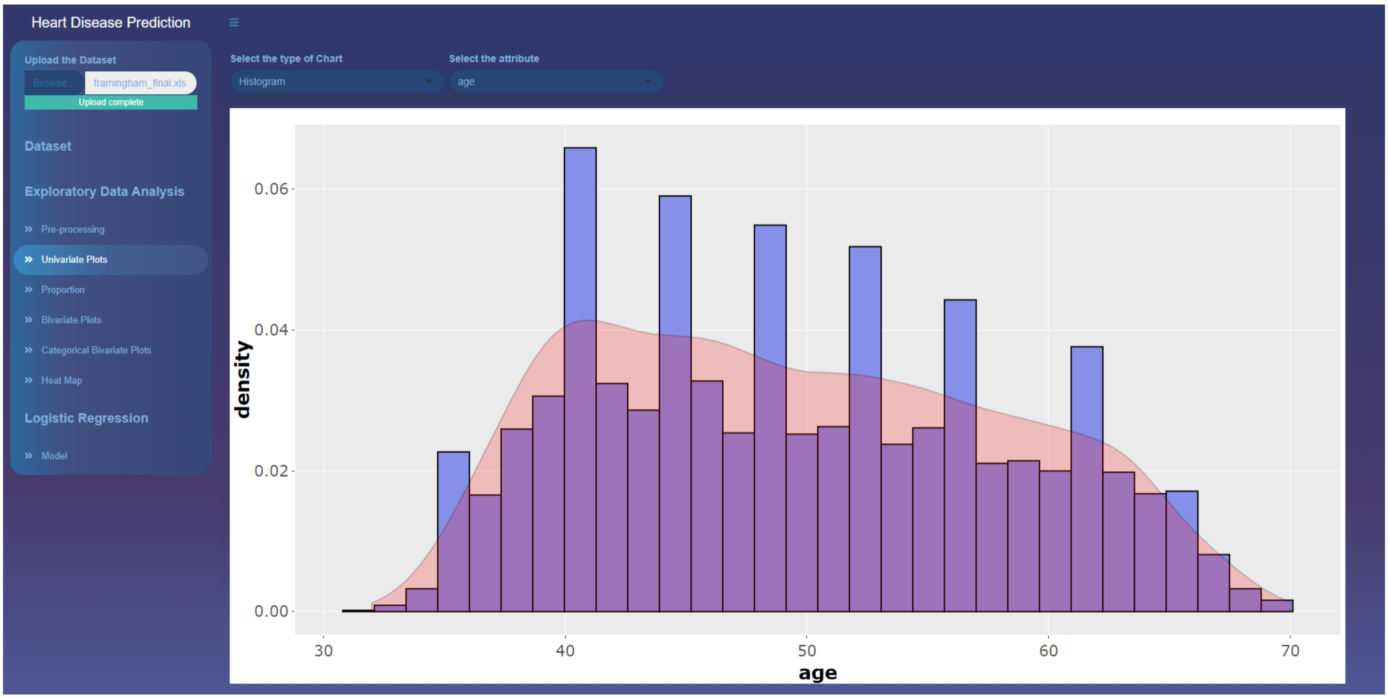
Univariet Plots – Barplot



Select the type of Chart: Barplot

Select the attribute: male

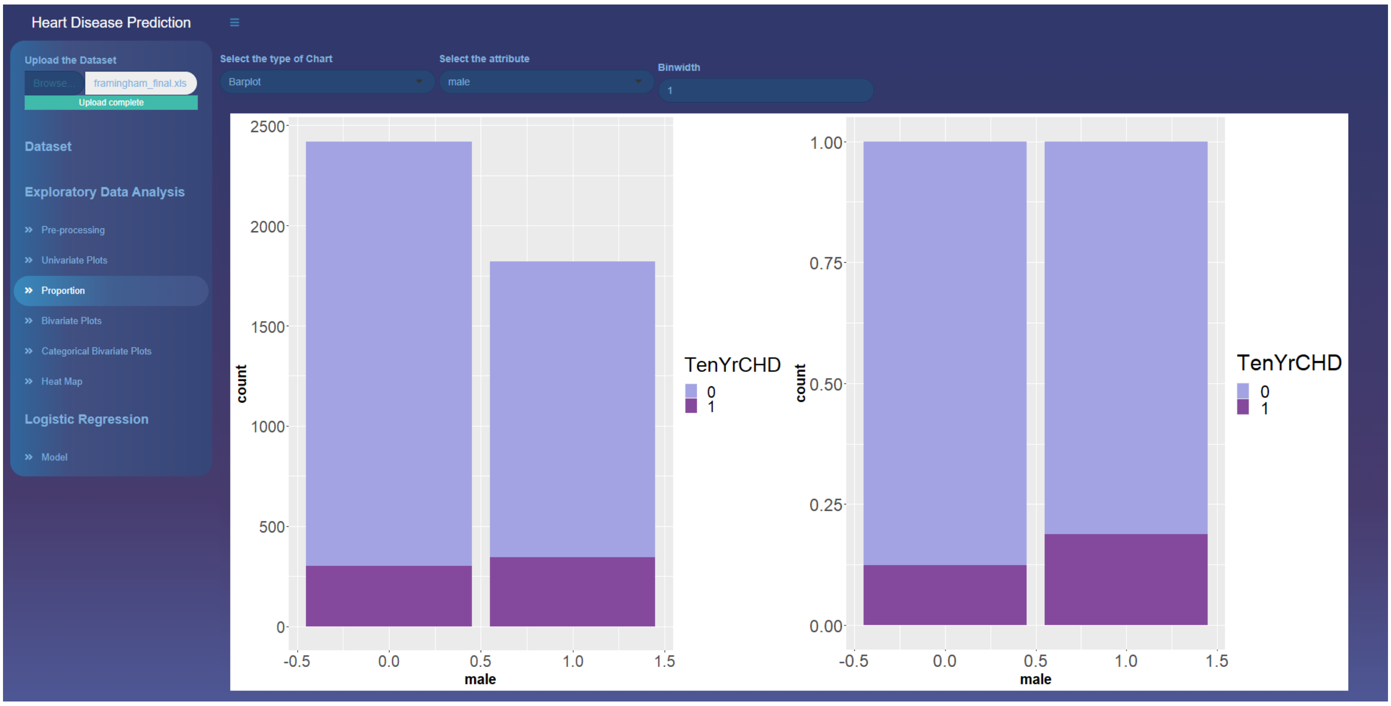
Univariet Plots – Histogram



Select the type of Chart: Histogram

Select the attribute: age

Proportion Plots – Barplot

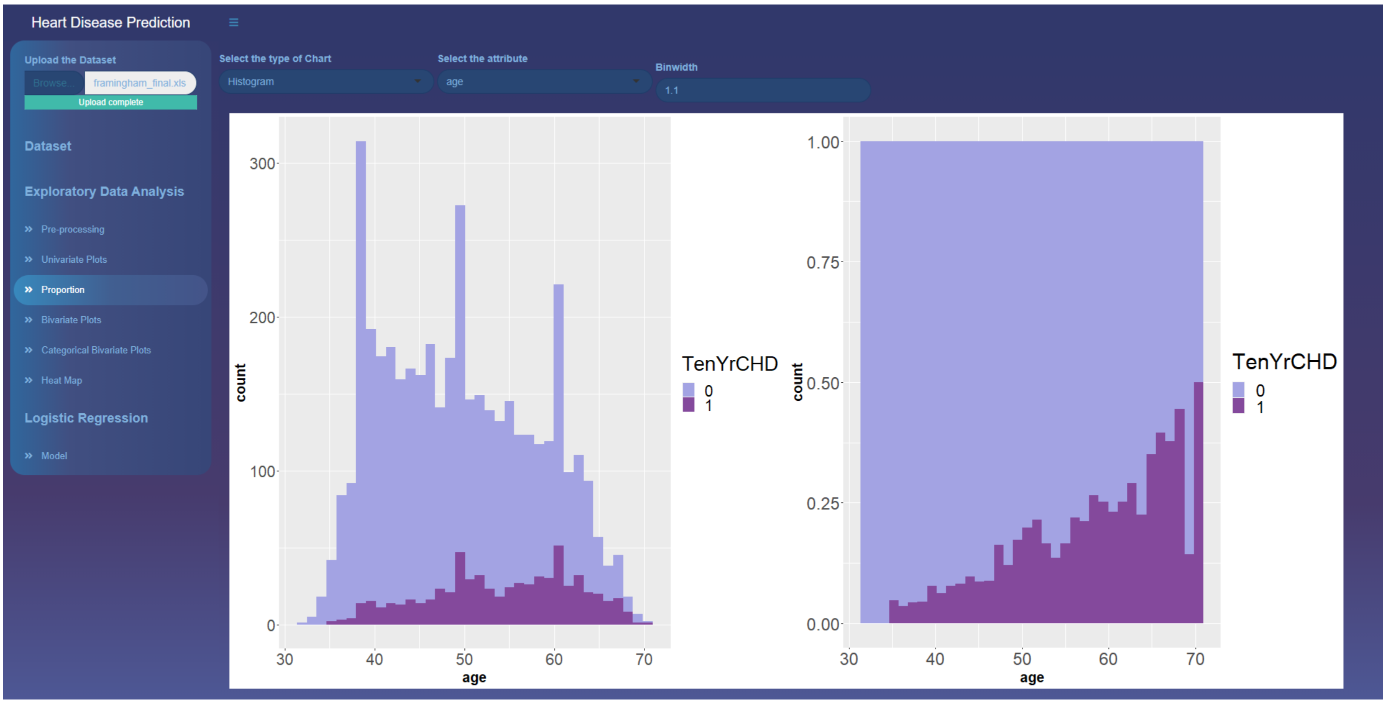


Select the type of Chart: Barplot

Select the attribute: male

Binwidth: 1

Proportion Plots – Histogram



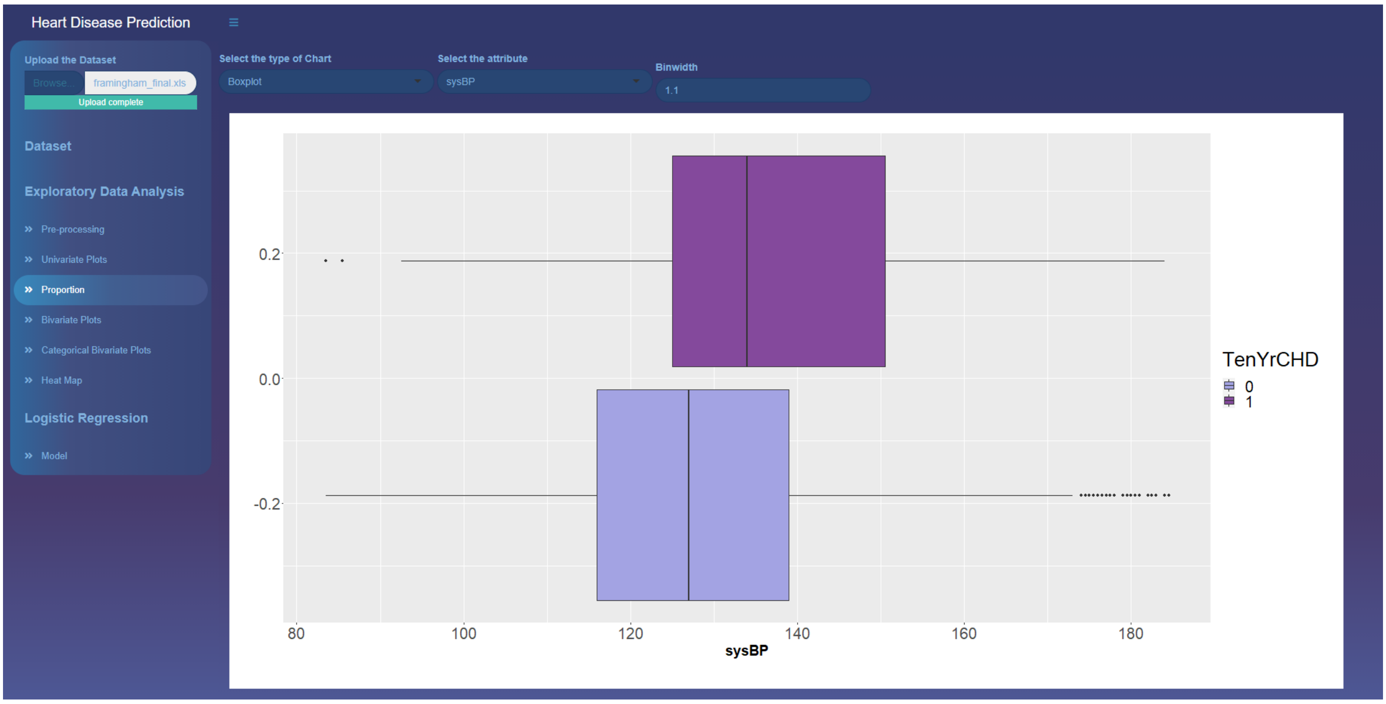
Upload the Dataset
framingham_final.xls
Upload complete

Select the type of Chart
Histogram

Select the attribute
age

Binwidth
1.1

Proportion Plots – Boxplot



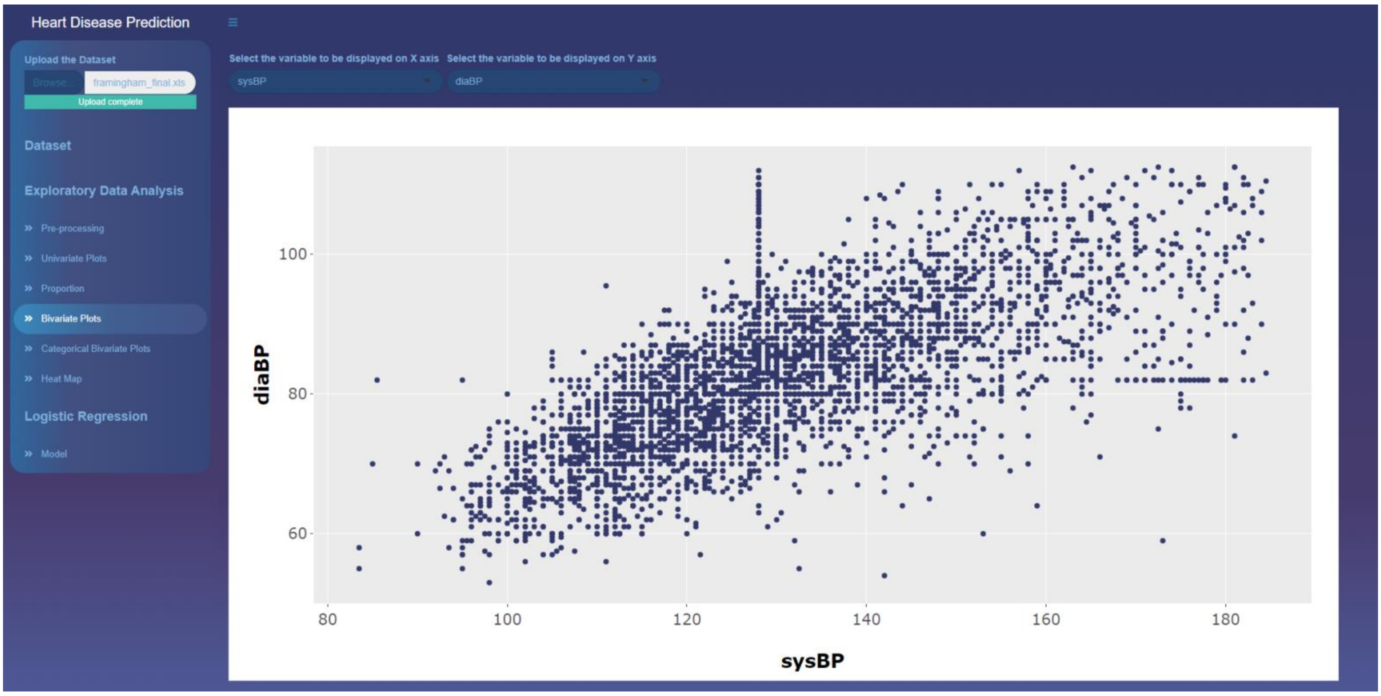
Upload the Dataset
framingham_final.xls
Upload complete

Select the type of Chart
Boxplot

Select the attribute
sysBP

Binwidth
1.1

Bivariet Plots

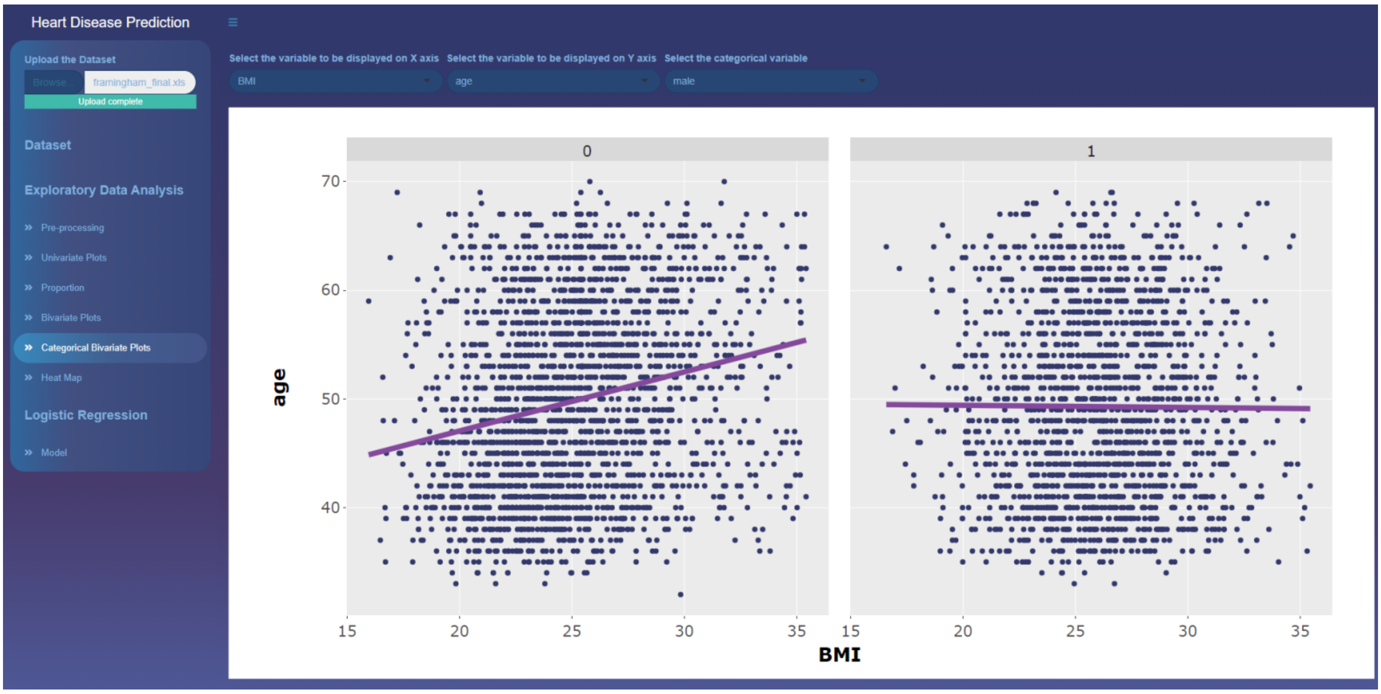


framingham_final.xls

Select the variable to be displayed on X axis: sysBP

Select the variable to be displayed on Y axis: diaBP

Categorical Bivariet Plots



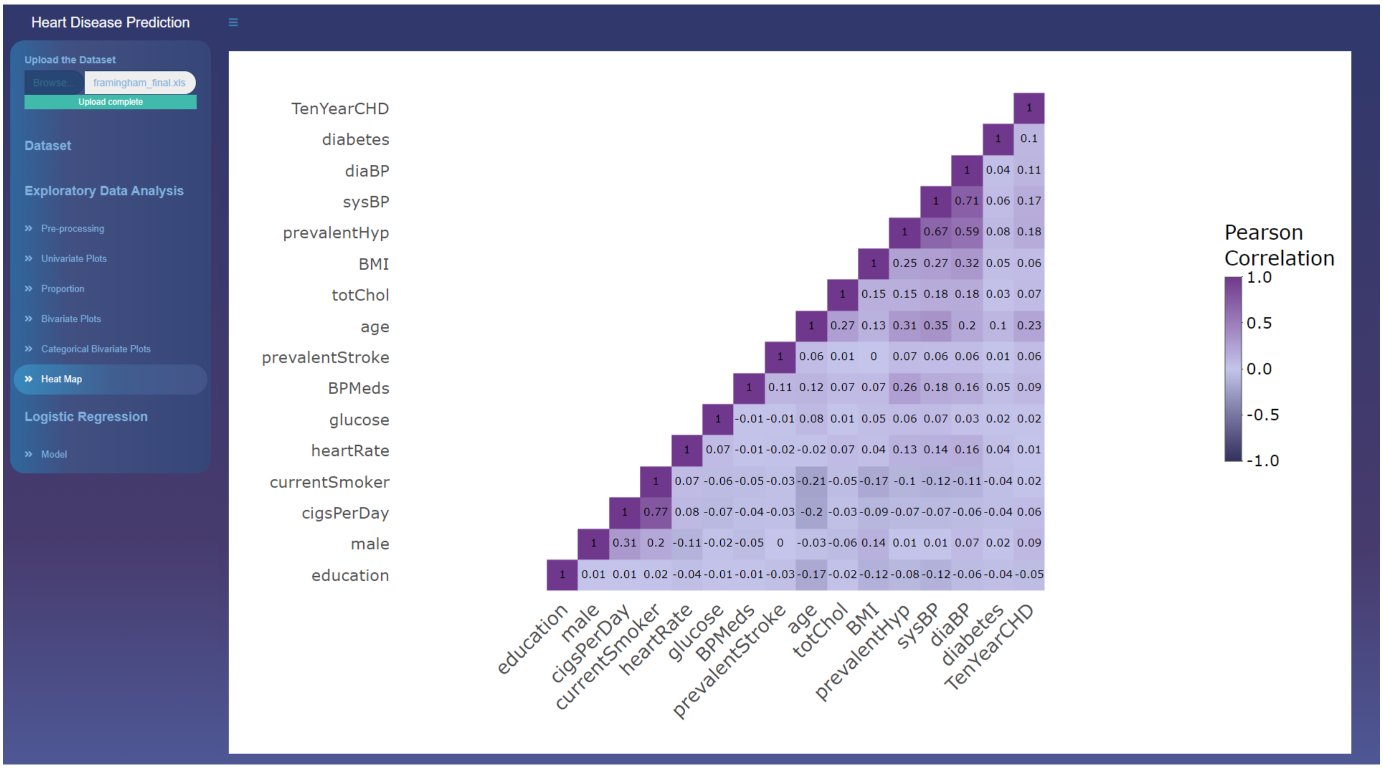
framingham_final.xls

Select the variable to be displayed on X axis: BMI

Select the variable to be displayed on Y axis: age

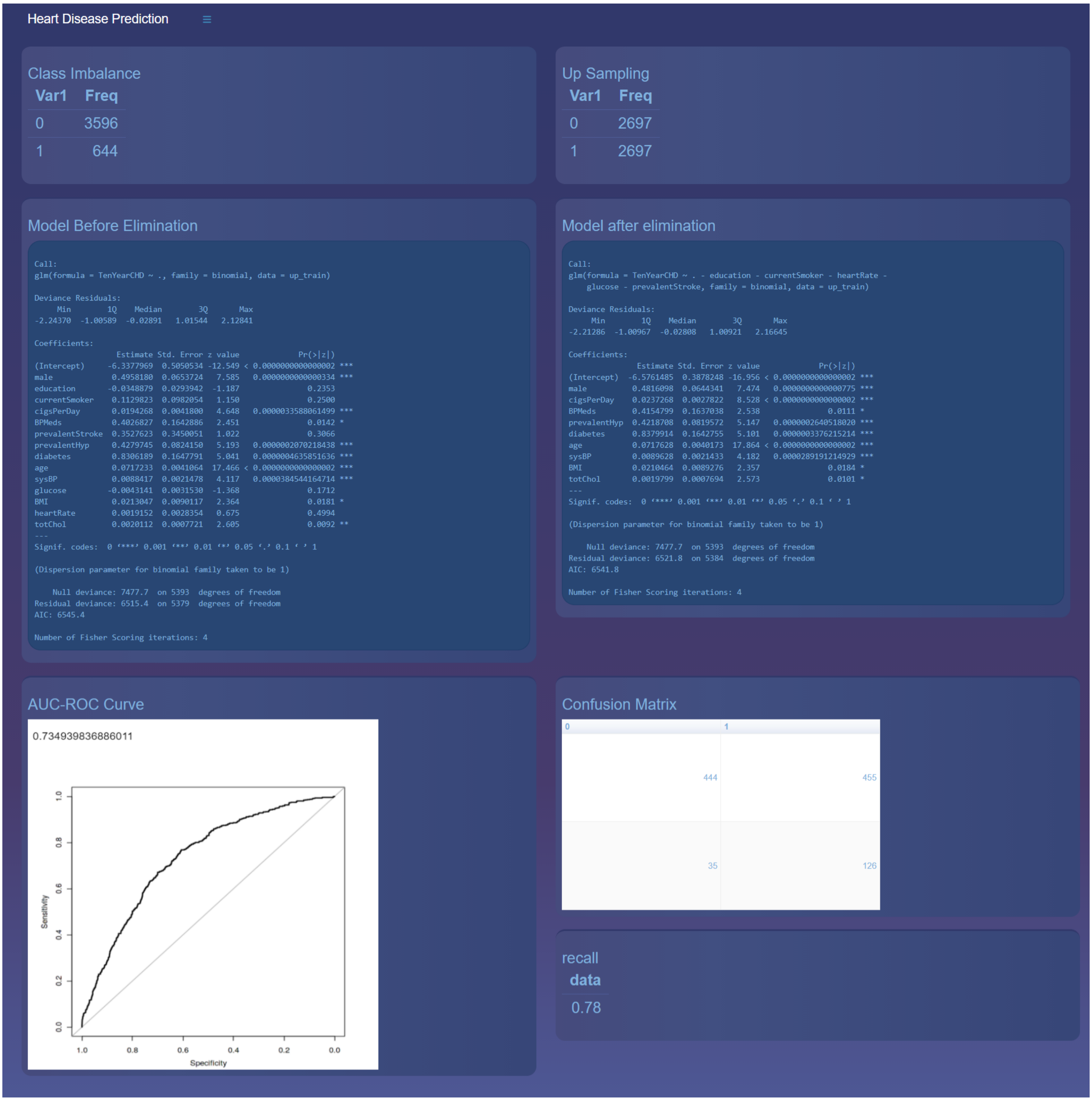
Select the categorical variable: male

Heatmap



	education	male	cigsPerDay	currentSmoker	heartRate	glucose	BPMeds	prevalentStroke	age	totChol	BMI	prevalentHyp	sysBP	diaBP	diabetes	TenYearCHD
TenYearCHD																1
diabetes															1	0.1
diaBP														1	0.04	0.11
sysBP													1	0.71	0.06	0.17
prevalentHyp												1	0.67	0.59	0.08	0.18
BMI											1	0.25	0.27	0.32	0.05	0.06
totChol										1	0.15	0.15	0.18	0.18	0.03	0.07
age									1	0.27	0.13	0.31	0.35	0.2	0.1	0.23
prevalentStroke								1	0.06	0.01	0	0.07	0.06	0.06	0.01	0.06
BPMeds							1	0.11	0.12	0.07	0.07	0.26	0.18	0.16	0.05	0.09
glucose						1	-0.01	-0.01	0.08	0.01	0.05	0.06	0.07	0.03	0.02	0.02
heartRate					1	0.07	-0.01	-0.02	-0.02	0.07	0.04	0.13	0.14	0.16	0.04	0.01
currentSmoker				1	0.07	-0.06	-0.05	-0.03	-0.21	-0.05	-0.17	-0.1	-0.12	-0.11	-0.04	0.02
cigsPerDay			1	0.77	0.08	-0.07	-0.04	-0.03	-0.2	-0.03	-0.09	-0.07	-0.07	-0.06	-0.04	0.06
male		1	0.31	0.2	-0.11	-0.02	-0.05	0	-0.03	-0.06	0.14	0.01	0.01	0.07	0.02	0.09
education	1	0.01	0.01	0.02	-0.04	-0.01	-0.01	-0.03	-0.17	-0.02	-0.12	-0.08	-0.12	-0.06	-0.04	-0.05

Logistic Regression – Model



Heart Disease Prediction

Class Imbalance

Var1	Freq
0	3596
1	644

Up Sampling

Var1	Freq
0	2697
1	2697

Model Before Elimination

```
Call:
glm(formula = TenYearCHD ~ ., family = binomial, data = up_train)

Deviance Residuals: 
     Min        1Q    Median        3Q       Max  
-2.24370  -1.00589  -0.02891   1.01544   2.12841  

Coefficients:
                 Estimate Std. Error z value             Pr(>|z|)    
(Intercept)     -6.3377969  0.5050534 -12.549 < 0.0000000000000002 ***
male             0.4958180  0.0653724   7.585   0.0000000000000334 ***
education       -0.0348879  0.0293942  -1.187               0.2353    
currentSmoker    0.1129823  0.0982054   1.150               0.2500    
cigsPerDay       0.0194268  0.0041800   4.648   0.0000033588061499 ***
BPMeds           0.4026827  0.1642886   2.451               0.0142 *  
prevalentStroke  0.3527623  0.3450051   1.022               0.3066    
prevalentHyp     0.4279745  0.0824150   5.193   0.0000002070218438 ***
diabetes         0.8306189  0.1647791   5.041   0.0000004635851636 ***
age              0.0717233  0.0041064  17.466 < 0.0000000000000002 ***
sysBP            0.0088417  0.0021478   4.117   0.0000384544164714 ***
glucose         -0.0043141  0.0031530  -1.368               0.1712    
BMI              0.0213047  0.0090117   2.364               0.0181 *  
heartRate        0.0019152  0.0028354   0.675               0.4994    
totChol          0.0020112  0.0007721   2.605               0.0092 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7477.7  on 5393  degrees of freedom
Residual deviance: 6515.4  on 5379  degrees of freedom
AIC: 6545.4

Number of Fisher Scoring iterations: 4
```

Model after elimination

```
Call:
glm(formula = TenYearCHD ~ . - education - currentSmoker - heartRate - 
    glucose - prevalentStroke, family = binomial, data = up_train)

Deviance Residuals: 
     Min        1Q    Median        3Q       Max  
-2.21286  -1.00967  -0.02808   1.00921   2.16645  

Coefficients:
               Estimate Std. Error z value             Pr(>|z|)    
(Intercept)  -6.5761485  0.3878248 -16.956 < 0.0000000000000002 ***
male          0.4816098  0.0644341   7.474   0.0000000000000775 ***
cigsPerDay    0.0237268  0.0027822   8.528 < 0.0000000000000002 ***
BPMeds        0.4154799  0.1637038   2.538               0.0111 *  
prevalentHyp  0.4218708  0.0819572   5.147   0.0000002640518020 ***
diabetes      0.8379914  0.1642755   5.101   0.0000003376215214 ***
age           0.0717628  0.0040173  17.864 < 0.0000000000000002 ***
sysBP         0.0089628  0.0021433   4.182   0.0000289191214929 ***
BMI           0.0210464  0.0089276   2.357               0.0184 *  
totChol       0.0019799  0.0007694   2.573               0.0101 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7477.7  on 5393  degrees of freedom
Residual deviance: 6521.8  on 5384  degrees of freedom
AIC: 6541.8

Number of Fisher Scoring iterations: 4
```

AUC-ROC Curve

0.734939836886011

Confusion Matrix

0	1
444	455
35	126

recall

data
0.78

CONCLUSION

The amount of Heart diseases can exceed the current scenario to reach the maximum point. Heart disease are complicated and each and every year lots of people are dying with this disease. It is difficult to manually determine the odds of getting heart disease based on risk factors previously shown. By using this system one of the major drawbacks of this work is that it's main focus is aimed only to the application of classifying techniques and algorithms for heart disease prediction, by studying various data cleaning and mining techniques that prepare and build a dataset appropriate for data mining so that we can use this Machine Learning in that logistic regression algorithms by predicting if patient has heart disease or not. Any non-medical employee can use this software and predict the heart disease and reduce the time complexity of the doctors. It is still an open domain waiting to get implemented in heart disease predication and increase the accuracy. Overall model could be improved with more data.

We could make different prediction model with Neural Network, Decision Tree, KNN, etc.

BIBLIOGRAPHY REFERENCES

- www.google.com
- www.kaggle.com