INTRODUCTION

Heart Disease refers to a condition which affects the normal functioning of the heart. Some of the heart diseases include blood vessels, such as coronary artery disease; rhythmic problems, birth defects related to heart and so on. Heart disease also includes narrowing or blocking of blood vessels leading to heart attack, chest pain or stroke. Some other conditions which involve weakening of heart muscles are also termed as heart disease.

Business Questions to be answered:

- **What are the factors that affect heart disease?**
- **What is the relationship between factors and the possibility of heart disease?**
- **How strong is the relationship between factors and the possibility of heart disease?**

About the dataset:

The dataset is retrieved from the UCI Machine Learning Repository. We have chosen this topic in order to predict whether an individual would face Heart Disease in the future or not. The dataset contains 14 variables namely, age, sex, chest pain, resting blood pressure, fasting blood sugar, cholesterol, resting ECG, maximum heart rate, exercise induces angina, old peak, the slope of peak exercise, number of major vessels, thallium heart scan and result. [1]

The variable definition is as follows:

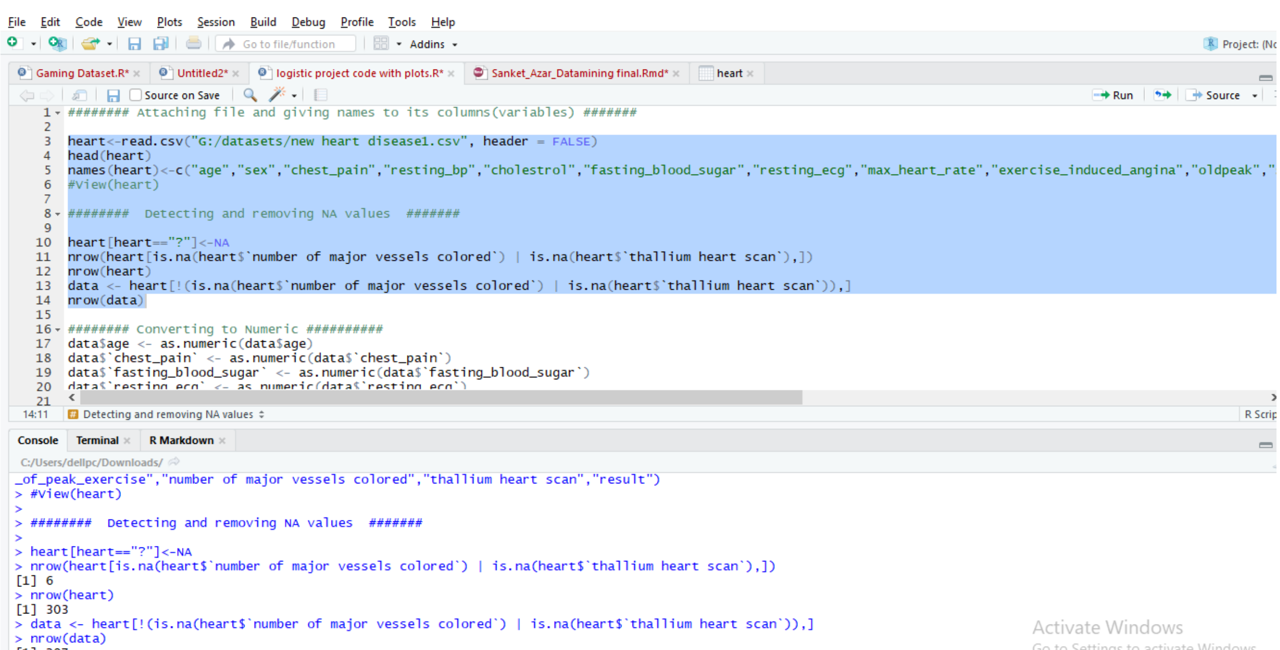
Age	the age group analyzed is 29- 77 with maximum people of 58-59 years old.
Sex	sex=0; individual is female
	sex=1; individual is male
	There are 97 females and 206 males analyzed is this dataset
Chest_pain	It defines the type of chest pain experienced
Resting_bp	Blood pressure in mm Hg on admission to the hospital
Fasting_blood_sugar	fasting blood sugar > 120 mg/dl - 1 = true; 0 = false
Cholesterol	serum cholesterol in mg/dl
Resting_ecg	resting electrocardiographic results
Max_heart_rate	maximum heart rate achieved
Exercise_induced_angina	(1 = yes; 0 = no)

Old peak	ST depression induced by exercise relative to rest
Slope_of_peak_exercise	the slope of the peak exercise ST segment
Number of major vessels colored	number of major vessels (0-3) colored by fluoroscopy
Thallium heart scan	3 = normal; 6 = fixed defect; 7 = reversable defect
Result	The result variable is the prediction which identifies as positive or negative for an individual suffering with any heart disease. If result = 0; no heart disease result = 1; has heart disease

Data Cleaning

It refers to removing the undesired variables and NA values from the dataset. Our dataset had '?' for no value. We converted it to NA And then removed them using is.a() function. We only had 6 NA values which were just 2% of the entire data, so, we removed those 6 rows. Initially, we had 303 observations, after removing those NA values 297 observations are used for analysis.

Also, to facilitate smooth access, the variables were named using names() function.



This is the dataset we have selected.

	age	sex	chest_pain	resting_bp	cholestrol	fasting_blood_sugar	resting_ecg	max_heart_rate	exercise_induced_angina	oldpeak	slope_of_peak_exercise	number of major vessels colored	thallium heart scan	result
1	63	1	1	145	233	1	2	150	0	2.3	3	0.0	6.0	0
2	67	1	4	160	286	0	2	108	1	1.5	2	3.0	3.0	2
3	67	1	4	120	229	0	2	129	1	2.6	2	2.0	7.0	1
4	37	1	3	130	250	0	0	187	0	3.5	3	0.0	3.0	0
5	41	0	2	130	204	0	2	172	0	1.4	1	0.0	3.0	0
6	56	1	2	120	236	0	0	178	0	0.8	1	0.0	3.0	0
7	62	0	4	140	268	0	2	160	0	3.6	3	2.0	3.0	3
8	57	0	4	120	354	0	0	163	1	0.6	1	0.0	3.0	0
9	63	1	4	130	254	0	2	147	0	1.4	2	1.0	7.0	2
10	53	1	4	140	203	1	2	155	1	3.1	3	0.0	7.0	1

Showing 1 to 10 of 303 entries

ANALYSIS

KNN classification

To implement KNN classification, all the desired variables should be converted to numeric form.

```
nrow(data)

######## Converting to numeric form ##########
data$age <- as.numeric(data$age)
data$`chest_pain` <- as.numeric(data$`chest_pain`)
data$`fasting_blood_sugar` <- as.numeric(data$`fasting_blood_sugar`)
data$`resting_ecg` <- as.numeric(data$`resting_ecg`)
data$`exercise_induced_angina` <- as.numeric(data$`exercise_induced_angina`)
data$`slope_of_peak_exercise` <- as.numeric(data$`slope_of_peak_exercise`)
data$`resting_bp` <- as.numeric(data$`resting_bp`)
data$`thallium heart scan` <- as.numeric(data$`thallium heart scan`)
data$cholestrol <- as.numeric(data$cholestrol)
data$sex <- as.numeric(data$sex)
data$`number of major vessels colored` <- as.numeric(data$`number of major vessels colored`)
```

The parameters in the dataset have a different kind of scales, thus we normalized the data.

As we can see the below summary for ‘data_n’, all the values have been normalized and lie between 0 & 1.

Normalized data variables

```
> normalize <- function(x) {
+         return((x - min(x)) / (max(x) - min(x)))
+ }
>
> data_n <- as.data.frame(lapply(data[1:13], normalize))
> summary(data_n)
      age              sex           chest_pain       resting_bp       cholestrol    
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
 1st Qu.:0.3958   1st Qu.:0.0000   1st Qu.:0.6667   1st Qu.:0.2453   1st Qu.:0.1941  
 Median :0.5625   Median :1.0000   Median :0.6667   Median :0.3396   Median :0.2671  
 Mean   :0.5321   Mean   :0.6768   Mean   :0.7194   Mean   :0.3556   Mean   :0.2771  
 3rd Qu.:0.6667   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.4340   3rd Qu.:0.3425  
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000  
 fasting_blood_sugar  resting_ecg     max_heart_rate   exercise_induced_angina
 Min.   :0.0000      Min.   :0.0000   Min.   :0.0000   Min.   :0.0000         
 1st Qu.:0.0000      1st Qu.:0.0000   1st Qu.:0.4733   1st Qu.:0.0000         
 Median :0.0000      Median :0.5000   Median :0.6260   Median :0.0000         
 Mean   :0.1448      Mean   :0.4983   Mean   :0.6000   Mean   :0.3266         
 3rd Qu.:0.0000      3rd Qu.:1.0000   3rd Qu.:0.7252   3rd Qu.:1.0000         
 Max.   :1.0000      Max.   :1.0000   Max.   :1.0000   Max.   :1.0000         
    oldpeak        slope_of_peak_exercise number.of.major.vessels.colored thallium.heart.scan
 Min.   :0.0000   Min.   :0.0000         Min.   :0.0000                  Min.   :0.0000     
 1st Qu.:0.0000   1st Qu.:0.0000         1st Qu.:0.0000                  1st Qu.:0.0000     
 Median :0.1290   Median :0.5000         Median :0.0000                  Median :0.0000     
 Mean   :0.1703   Mean   :0.3013         Mean   :0.2256                  Mean   :0.4175     
 3rd Qu.:0.2581   3rd Qu.:0.5000         3rd Qu.:0.3333                  3rd Qu.:1.0000     
 Max.   :1.0000   Max.   :1.0000         Max.   :1.0000                  Max.   :1.0000     
> 
```

The dataset is divided into test and train dataset such that 70% of randomly selected data points are in train set and 30% of them in the test set so that the model created with train dataset could be cross-verified with test dataset. We have used the sample() to select the random sample data and the set.seed(1000) so as to fetch the same random sample every time we run the code.

```
set.seed(1000)

# random selection of 70% of data
rand.70 <- sample(1:nrow(data_n),size=nrow(data_n)*0.7,replace = FALSE)

# Training set
train_set <- data_n[rand.70,]    # 70% training data
test_set <- data_n[-rand.70,]    # 30% test data

# Target set
# Creating a data frame for 'defaulter' feature which is our result
train_target <- data[rand.70,14]
test_target <- as.factor(data[-rand.70,14])

```

Implementing KNN- classification

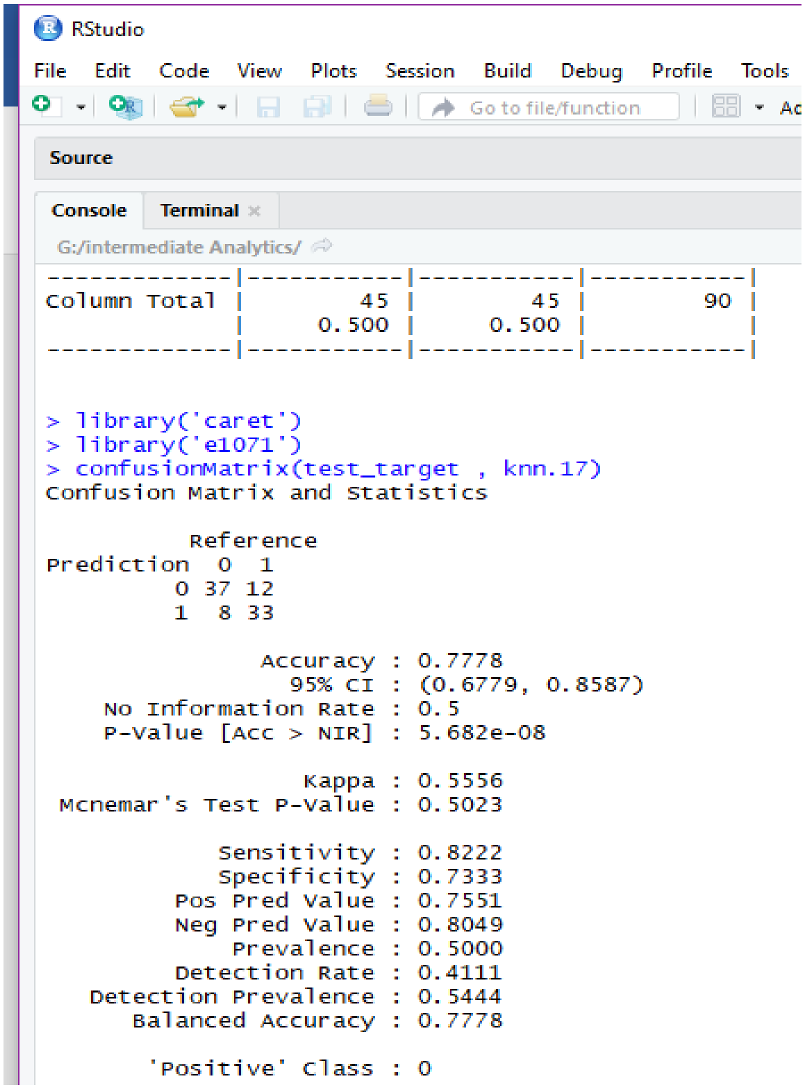
We need to identify the optimum value of k to minimize the error. Generally, we take k as an odd number nearest to the square root of the total number of observations. So, we take k = 17.

```
 Max.   :1.0000   Max.   :1.0000          Max.   :1.0000                   Max.   :1.0000
> library(class)
> sqrt(297) # total observations are 297
[1] 17.23369
> knn.17 <- as.factor(knn(train = train_set, test = test_set, cl = train_target, k = 17))
> table(knn.17, test_target)
      test_target
knn.17  0  1
     0 37  8
     1 12 33
> ACC.173 <- 100 * sum(test_target == knn.17)/NROW(test_target)
> ACC.173
[1] 77.77778
> 
```

Obtaining the Cross Table

```
RStudio
File  Edit  Code  View  Plots  Session  Build  Debug  Profile  Tools  Help
Go to file/function    Addins
Source
Console   Terminal ×
G:/intermediate Analytics/
[1] 77.77778
> library(gmodels)
> CrossTable(x = test_target, y = knn.17, prop.chisq = FALSE)


   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  90


             | knn.17 
 test_target |         0 |         1 | Row Total | 
-------------|-----------|-----------|-----------|
           0 |        37 |        12 |        49 | 
             |     0.755 |     0.245 |     0.544 | 
             |     0.822 |     0.267 |           | 
             |     0.411 |     0.133 |           | 
-------------|-----------|-----------|-----------|
           1 |         8 |        33 |        41 | 
             |     0.195 |     0.805 |     0.456 | 
             |     0.178 |     0.733 |           | 
             |     0.089 |     0.367 |           | 
-------------|-----------|-----------|-----------|
Column Total |        45 |        45 |        90 | 
             |     0.500 |     0.500 |           | 
-------------|-----------|-----------|-----------|
```

Obtaining the Confusion matrix




```
-------------|-----------|-----------|-----------|
Column Total |        45 |        45 |        90 |
             |     0.500 |     0.500 |           |
-------------|-----------|-----------|-----------|


> library('caret')
> library('e1071')
> confusionMatrix(test_target , knn.17)
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 37 12
         1  8 33

               Accuracy : 0.7778
                 95% CI : (0.6779, 0.8587)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : 5.682e-08

                  Kappa : 0.5556
 Mcnemar's Test P-Value : 0.5023

            Sensitivity : 0.8222
            Specificity : 0.7333
         Pos Pred Value : 0.7551
         Neg Pred Value : 0.8049
             Prevalence : 0.5000
         Detection Rate : 0.4111
   Detection Prevalence : 0.5444
      Balanced Accuracy : 0.7778

       'Positive' Class : 0
```

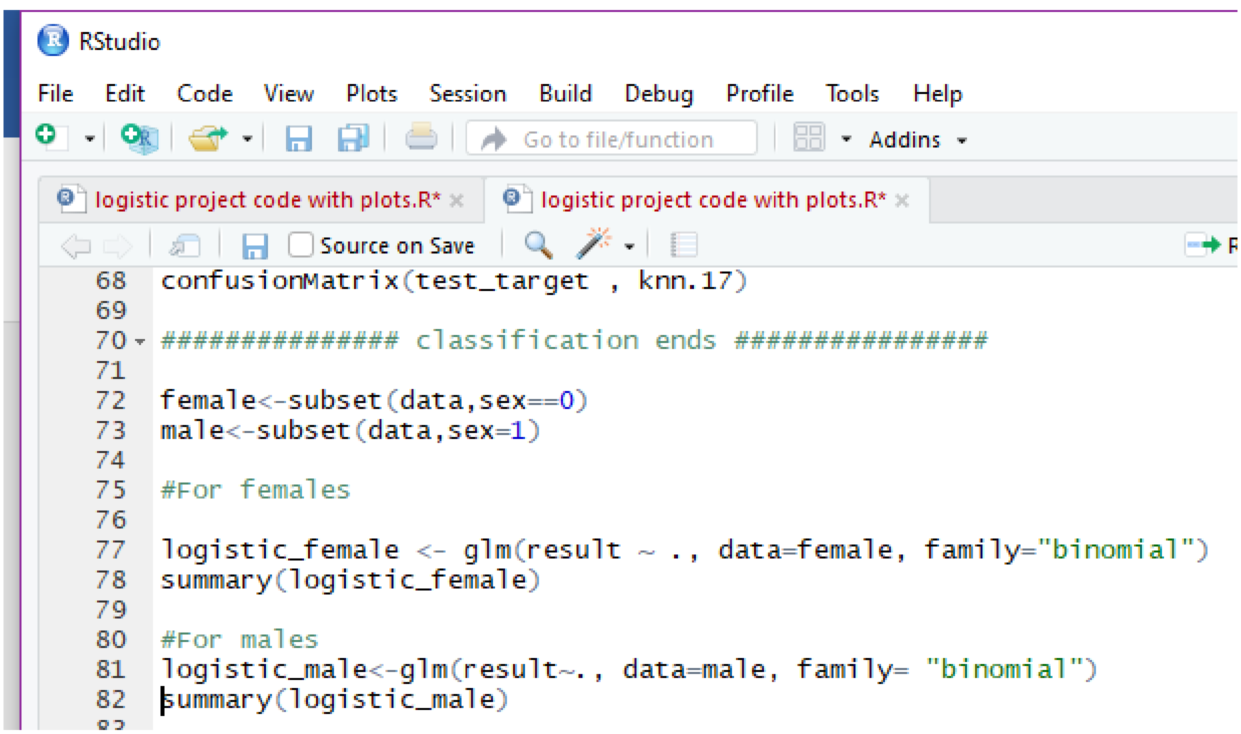
Interpretation

- From the cross table, we can infer that our Test data consisted of 90 observations.
- Out of which37 cases have been accurately predicted (True Negatives) as patients without heart disease. Also, 33 cases out of 90 were accurately predicted (True Positives) as the patients with Heart Disease. While 20 cases were incorrectly predicted, that is, 12 of them were predicted to have heart disease when they did not have and 8 were not predicted of having heart disease while they had the disease.

- Our KNN prediction classification model has an accuracy of 77.78% as shown in the above Confusion Matrix at a confidence level of 95%
- Moreover, sensitivity (proportion of people with the disease and positive result) of the test is 82.2% and the specificity (proportion of people without disease and negative result) of the test is 73.3%.
- Balanced accuracy and actual accuracy are the same indicating that the accuracy cannot be improved than the acquired 78% value.

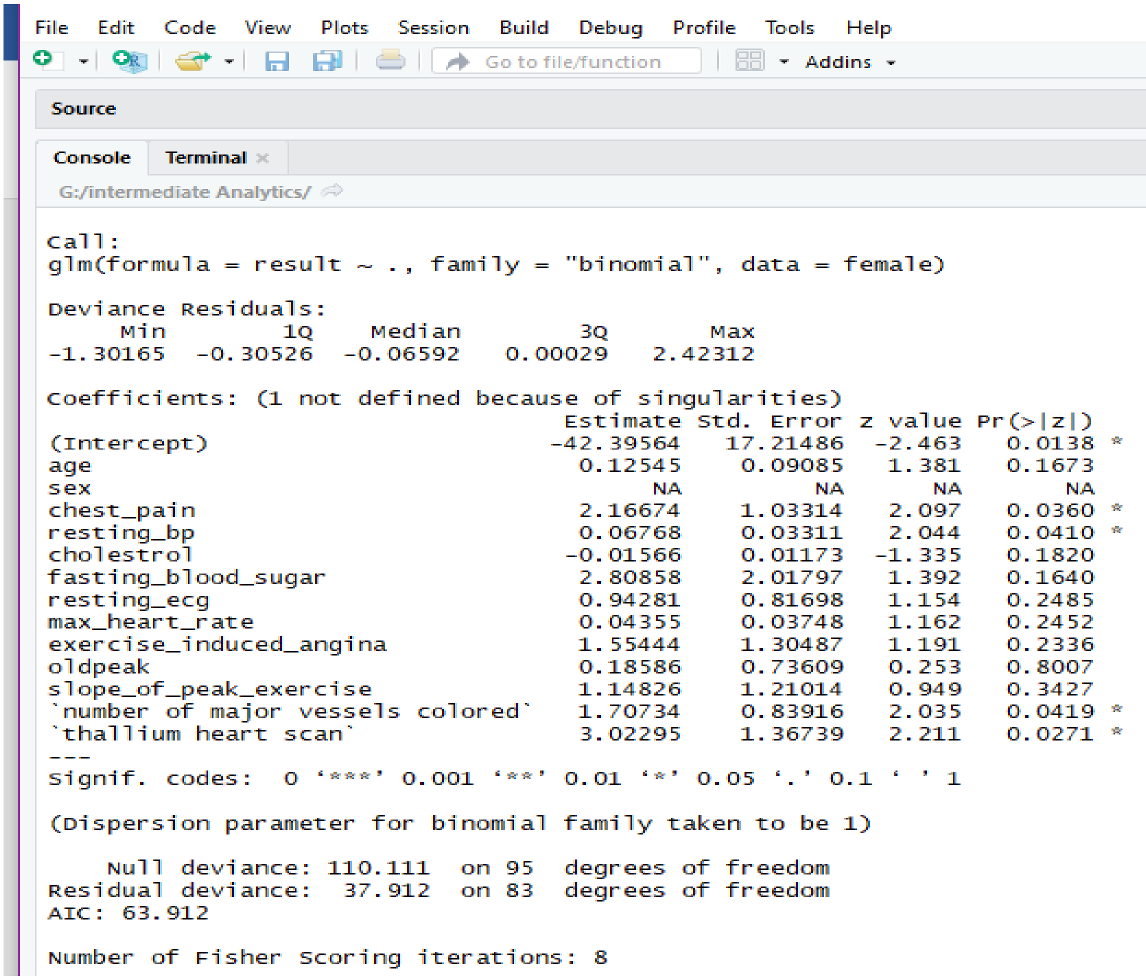
Implementing Logistics Regression

Logistic Regression is considered for males and females separately (considering sex to be a dominant factor)




```r
confusionMatrix(test_target , knn.17)

############### classification ends ################

female<-subset(data,sex==0)
male<-subset(data,sex=1)

#For females

logistic_female <- glm(result ~ ., data=female, family="binomial")
summary(logistic_female)

#For males
logistic_male<-glm(result~., data=male, family= "binomial")
summary(logistic_male)
```

Output for Females dataset




```
Call:
glm(formula = result ~ ., family = "binomial", data = female)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.30165  -0.30526  -0.06592   0.00029   2.42312

Coefficients: (1 not defined because of singularities)
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                      -42.39564   17.21486  -2.463   0.0138 *
age                                0.12545    0.09085   1.381   0.1673
sex                                     NA         NA      NA       NA
chest_pain                         2.16674    1.03314   2.097   0.0360 *
resting_bp                         0.06768    0.03311   2.044   0.0410 *
cholestrol                        -0.01566    0.01173  -1.335   0.1820
fasting_blood_sugar                2.80858    2.01797   1.392   0.1640
resting_ecg                        0.94281    0.81698   1.154   0.2485
max_heart_rate                     0.04355    0.03748   1.162   0.2452
exercise_induced_angina            1.55444    1.30487   1.191   0.2336
oldpeak                            0.18586    0.73609   0.253   0.8007
slope_of_peak_exercise             1.14826    1.21014   0.949   0.3427
`number of major vessels colored`  1.70734    0.83916   2.035   0.0419 *
`thallium heart scan`              3.02295    1.36739   2.211   0.0271 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 110.111  on 95  degrees of freedom
Residual deviance:  37.912  on 83  degrees of freedom
AIC: 63.912

Number of Fisher Scoring iterations: 8
```

Regression Model (Females)

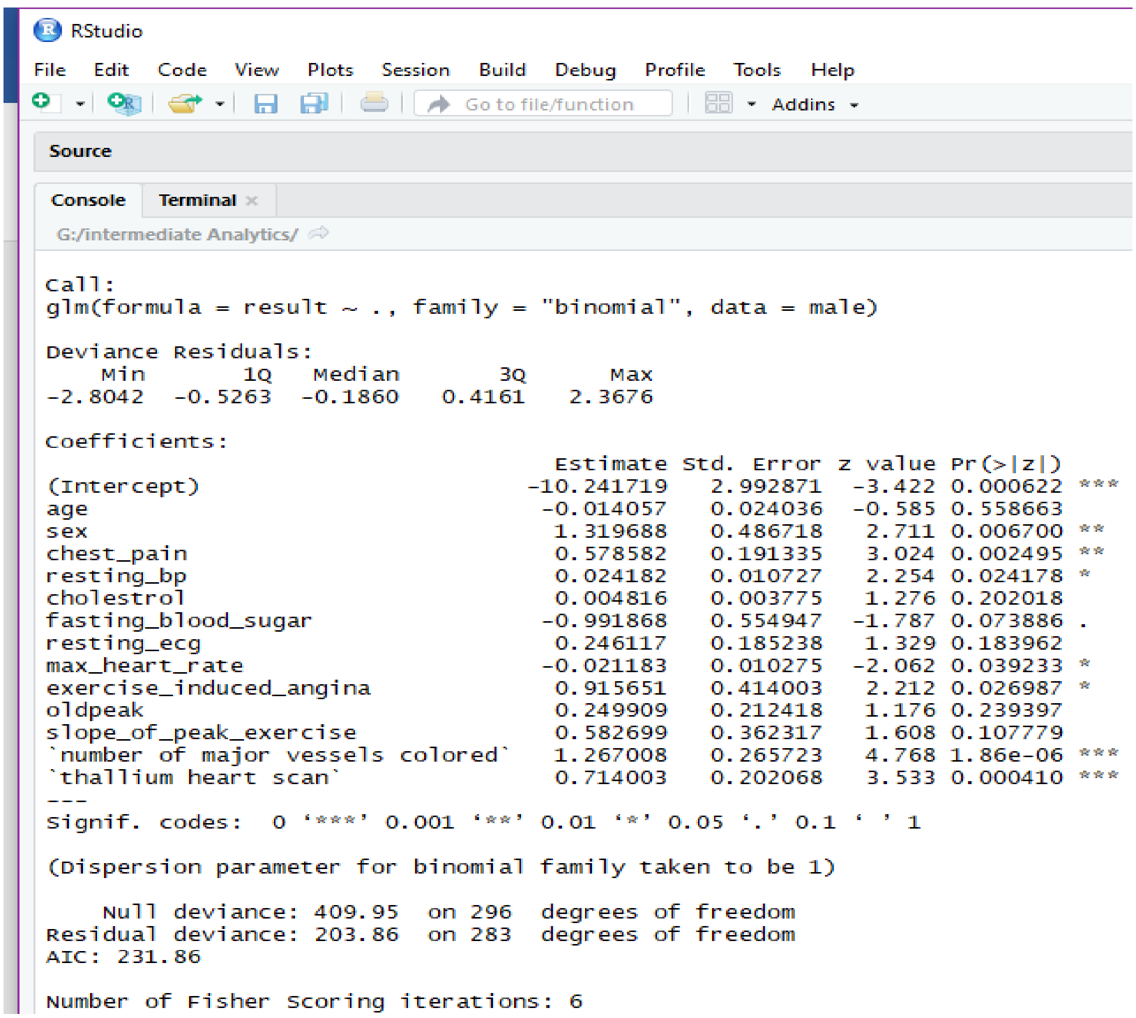
Y=2.166*(chest_pain) +0.067*(resting_bp)+1.71*(number of major vessels colored) + 3.02*(thallium heart scan) – 42.39

P(Y)=1/1+e^(-Y) (model equation for result)

- Logistic regression is done for the female subset to find out the best predictor variables for our given dataset.
- It has been observed that 4 out of 13 variables form the optimum predictor variables for the female subset, namely chest pain, resting blood pressure, number of major vessels colored and thallium heart rate. The regression model has been shown above using the coefficients obtained.

- The variables affecting can be determined from their p-values obtained after performing z-test. For logistic regression, the Null hypothesis is: the result variable is independent upon the variable considered.
  Alternate hypothesis: the result variable is dependent upon the variable considered.
  For the given four variables the p-values are less than 0.05 for the 95% significance level. Therefore, we reject the null hypothesis for all these variables and accept the alternate hypothesis and create a model considering these four variables.
- From the above output we also obtain the min= (-1.30); max= (2.42); median= (-0.0659); quantile1= (-0.305); quantile3= (0.00029) and the degrees of freedom =83 for the residual deviance.

Output for Males dataset



RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins
Source
Console Terminal
G:/intermediate Analytics/

```
Call:
glm(formula = result ~ ., family = "binomial", data = male)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8042  -0.5263  -0.1860   0.4161   2.3676

Coefficients:
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                       -10.241719   2.992871  -3.422 0.000622 ***
age                                -0.014057   0.024036  -0.585 0.558663
sex                                 1.319688   0.486718   2.711 0.006700 **
chest_pain                          0.578582   0.191335   3.024 0.002495 **
resting_bp                          0.024182   0.010727   2.254 0.024178 *
cholestrol                          0.004816   0.003775   1.276 0.202018
fasting_blood_sugar                -0.991868   0.554947  -1.787 0.073886 .
resting_ecg                         0.246117   0.185238   1.329 0.183962
max_heart_rate                     -0.021183   0.010275  -2.062 0.039233 *
exercise_induced_angina             0.915651   0.414003   2.212 0.026987 *
oldpeak                             0.249909   0.212418   1.176 0.239397
slope_of_peak_exercise              0.582699   0.362317   1.608 0.107779
`number of major vessels colored`   1.267008   0.265723   4.768 1.86e-06 ***
`thallium heart scan`               0.714003   0.202068   3.533 0.000410 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 409.95  on 296  degrees of freedom
Residual deviance: 203.86  on 283  degrees of freedom
AIC: 231.86

Number of Fisher Scoring iterations: 6
```

Regression Model (Males)

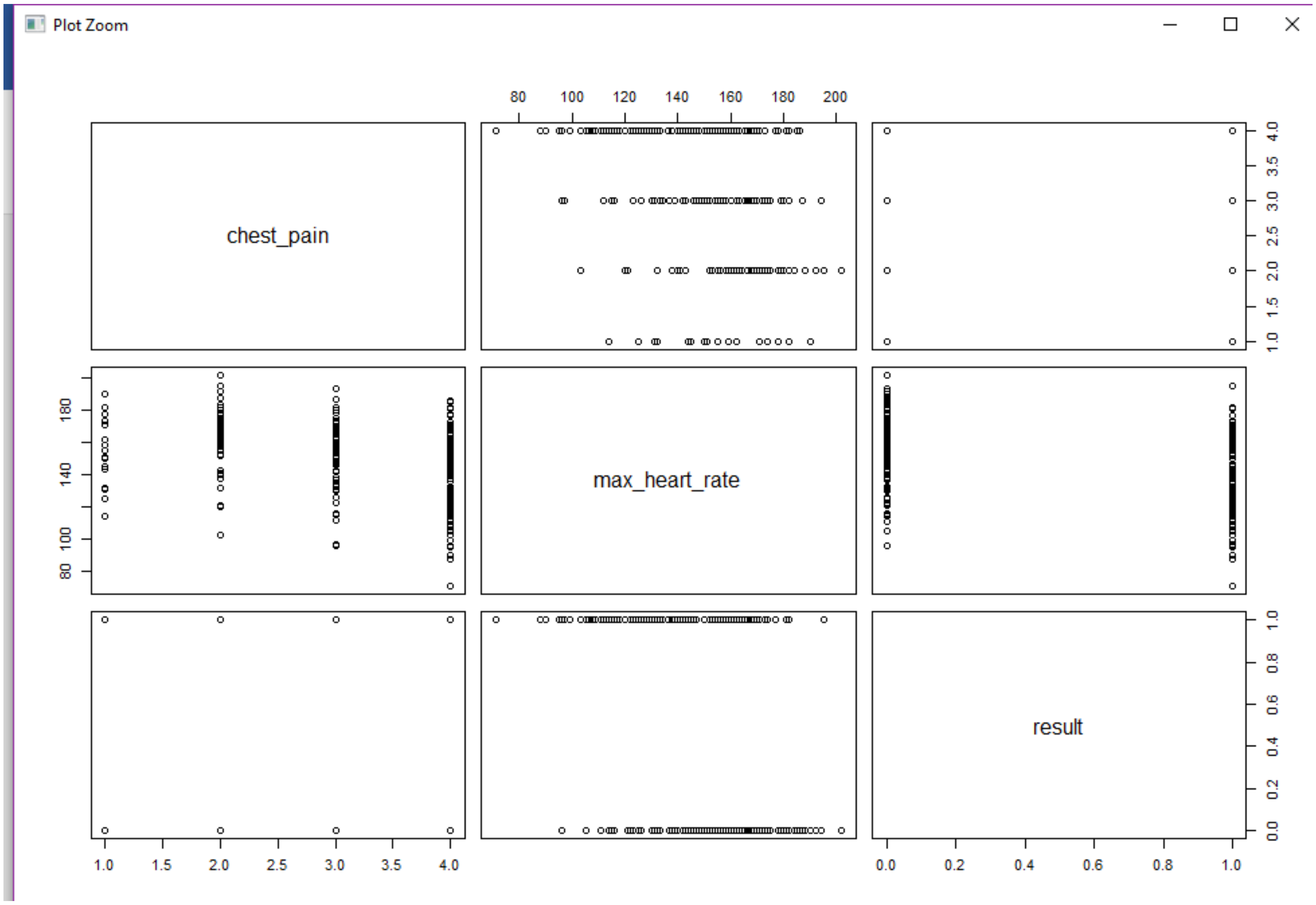
Y=0.578*(chest_pain) + 1.26*(number of major vessels colored) + 0.71*(thallium heart scan) +0.02418*(resting_bp) – 0.02118*(max_heart_rate) + 0.9156*(exercise_induced_angina) – 0.9918*(fasting_blood_sugar) - 10.2

(Y)=1/1+e^(-Y) (model equation)

Logistic regression is done for the male subset to find out the best predictor variables for our given dataset.

- It has been observed that 6 out of 13 variables form the optimum predictor variables for the male subset, namely chest pain, resting blood pressure, a number of major vessels colored and thallium heart rate, maximum heart rate, and angina induced due to heavy exercise. The regression model has been shown above using the coefficients obtained.
- The variables affecting can be determined from their p-values obtained after performing z-test. For logistic regression, the Null hypothesis is: the result variable is independent upon the variable considered.
  Alternate hypothesis: the result variable is dependent upon the variable considered.
  For the given four variables the p-values are less than 0.05 for the 95% significance level. Therefore, we reject the null hypothesis for all these variables and accept the alternate hypothesis and create a model considering these four variables.
- From the above output we also obtain the min= (-2.80); max= (2.36); median= (-0.186); quantile1= (-0.526); quantile3= (0.416) and the degrees of freedom =283 for the residual deviance.
- The effect of fasting_blood_sugar on the result is not very significant. But, in order to obtain an accurate model, all the affecting factors are considered.

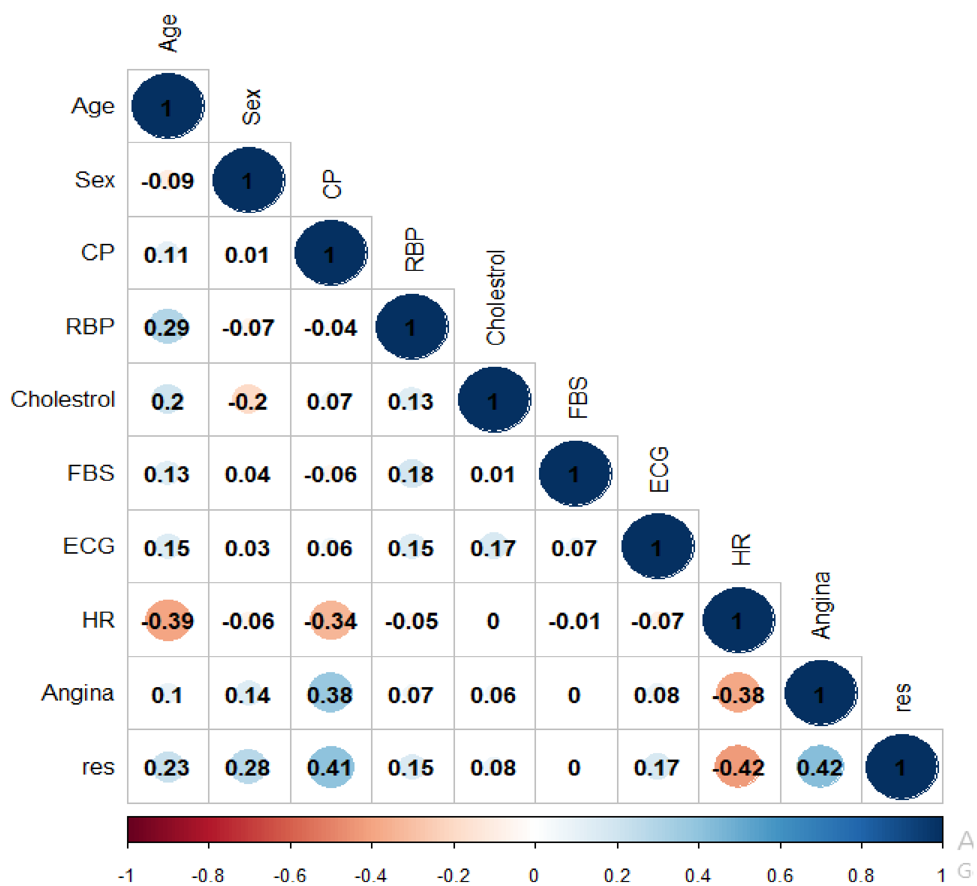
Plotting the scatter plot matrix and correlation matrix



The above is a matrix of scatter plots which shows the relationship between result with chest_pain and max_heart_rate (on separate plots).

From the above scatter matrix we also infer that there is a negative correlation between the result and maximum heart rate and a positive correlation between result and chest pain. This indicates that when chest pain increases the chances of the person having heart disease also increase.

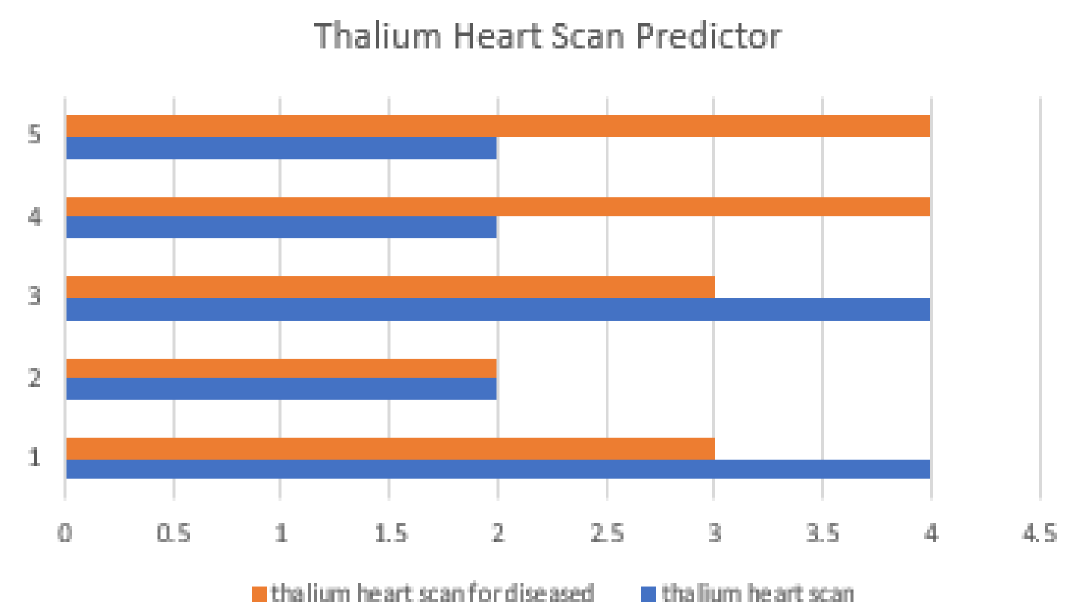
```r
#######Correlation matrix ############

#install.packages('corrplot')
#install.packages('sqldf')
library('corrplot')
library('sqldf')
#names(data)
#str(data)
data1<-sqldf("SELECT age as Age, sex as Sex, chest_pain as CP, resting_bp as RBP,
             cholestrol as Cholestrol, fasting_blood_sugar as FBS, resting_ecg as ECG,
             max_heart_rate as HR, exercise_induced_angina as Angina, result as res FROM data")

corMatrix <- cor(data1)

########Correlation matrix##########

corrplot(corMatrix)
par(mfrow=c(1,1))
corrplot(corMatrix, method="circle", type="lower", addCoef.col = "black", # Add coefficient of correlation
        tl.col="black", tl.srt=90, #Text label color and rotation
        diag=TRUE, sig.level = 0.05, insig = "blank")
```

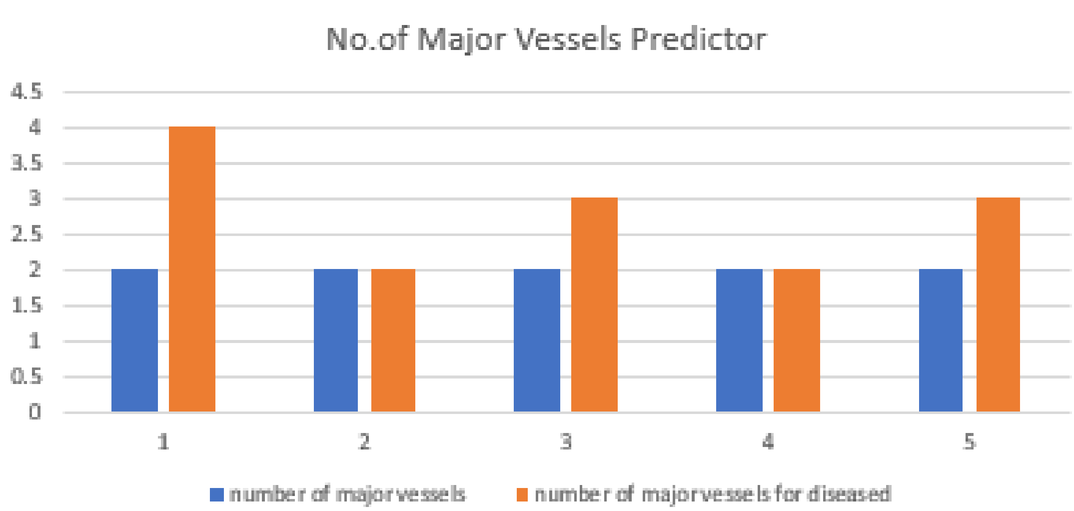


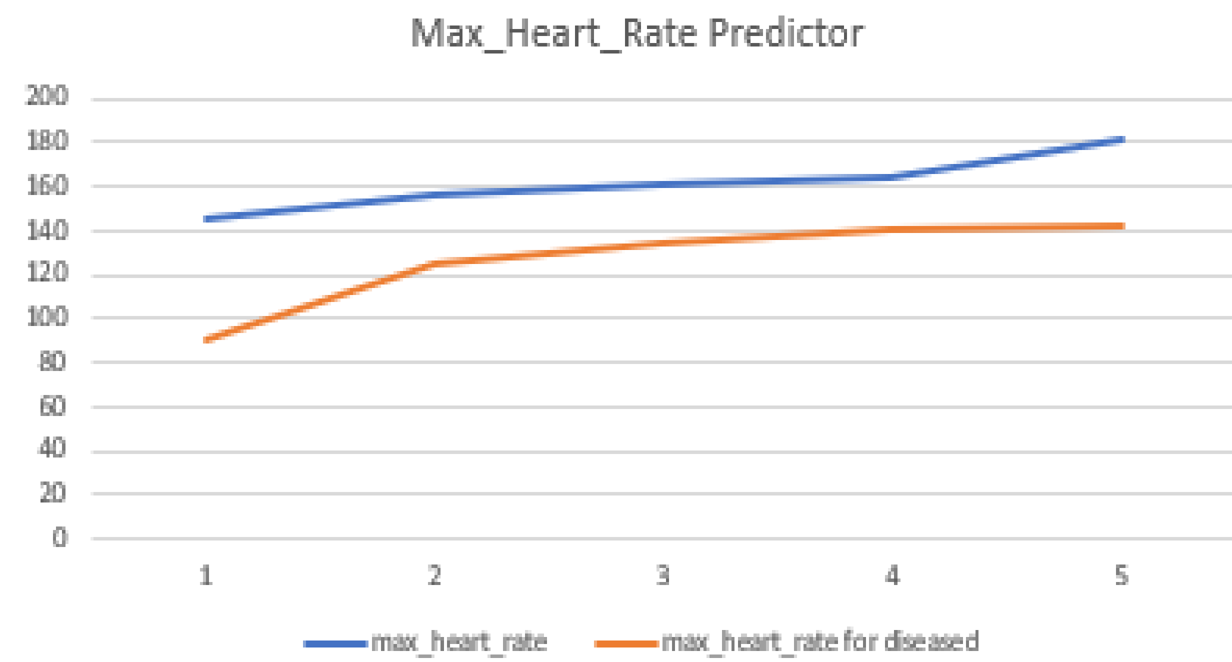
- From the above 2 plots, we find out the correlation between our predicted variables and our prediction result, which is whether an incoming patient has heart disease or not.

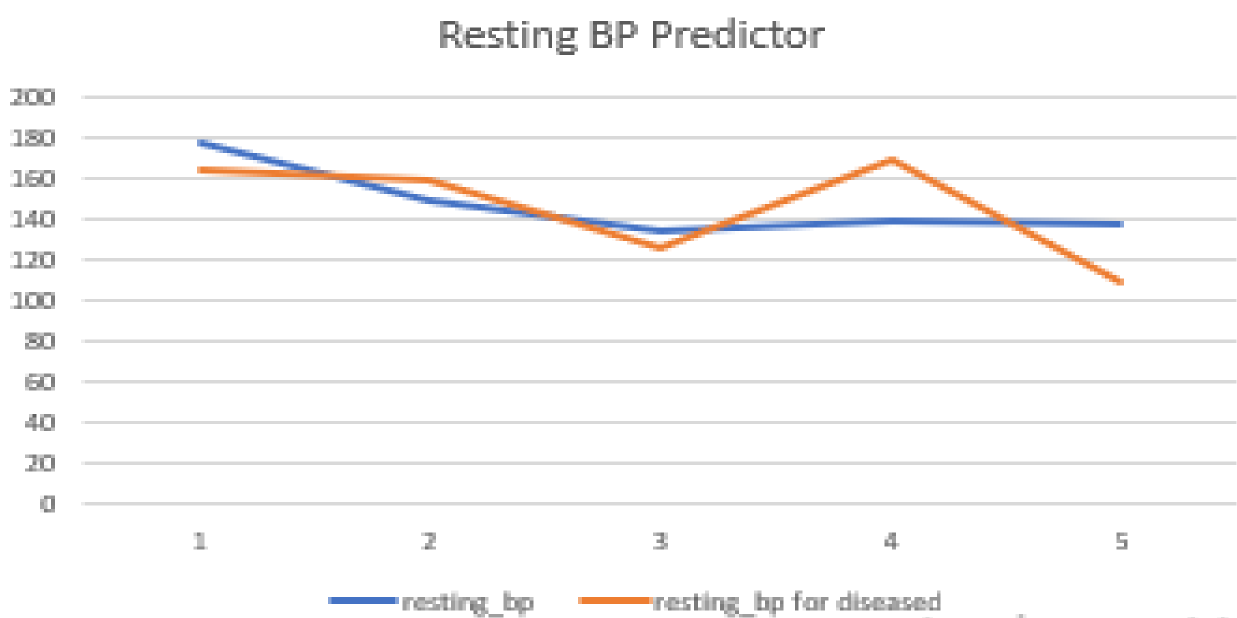
- From the above plots, we infer that there is a huge correlation value for the result with max heart rate, chest pain as declared above using regression methods as well.
- The above two matrices also show that there are some variables which are completely independent of each other, such as HR and cholesterol; FBS and result and angina with FBS.
- Using the above matrix, it can be concluded that the variables are not highly correlated. Thus, making the results of regression acceptable.

Prediction Variables Analysis Plots









Interpretation of the above plots

- To verify our prediction, we took a dataset for a similar age group (59 years) and analyzed the data pattern.

- From the above plots we can see that for max heart rate, people with heart disease have lower max heart rate compared to the people without heart disease proving the negative correlation between result and max heart rate variables.
- For a number of vessels colored we prove a positive correlation as the number is greater for people with heart disease compared to people without.
- One significant observation in the plot is that there are 2 cases where both counts are equal. This is because our prediction model is only 78% accurate which we inferred during the KNN classification.

CONCLUSION

1. Using Logistic Regression, we determine that chest pain, resting blood pressure, a number of major vessels, thallium heart scan are the factors that are significant for prediction of heart disease in females. While for males, the significant factors for prediction of heart disease are chest pain, the number of major vessels, resting blood pressure, maximum heart rate, exercise-induced pain, and thallium heart scan.
2. From the correlation plots we infer that there is a huge correlation value for the result with max heart rate and chest pain which could be verified using logistic regression method as well.
3. For max heart rate, people with heart disease have lower max heart rate compared to the people without heart disease proving the negative correlation between result and max heart rate variables.
4. For a number of vessels colored, we prove a positive correlation exists with the result, as the number is greater for people with heart disease compared to people who do not have heart disease.
5. One significant observation in the plot is that there are 2 cases where both counts are equal. This is because our prediction model is only 78% accurate which we inferred during the KNN classification.
6. Our prediction model states that 4 predictor variables out of the total 13 play a significant role in our prediction model with 78% accuracy.

Therefore, for any new patients if we have their chest pain type, thallium scan output, the number of vessels colored and resting blood pressure values then we could predict whether the person would have heart ailments or not.

REFERENCES

Dataset reference link: [1] http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/heart/heart.dat

1. Mayo Clinic (n.d) ‘Heart Disease’. Retrieved from https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118
2. Idre (n.d) ‘Logit Regression | R Data Analysis Examples’. Retrieved from https://stats.idre.ucla.edu/r/dae/logit-regression/
3. Statistics How to (n.d) ‘Sensitivity vs Specificity vs Predictive Value. Retrieved from https://www.statisticshowto.datasciencecentral.com/sensitivity-vs-specificity-statistics/
4. Data School (March 25, 2014) ‘Simple guides to confusion matrix terminology’ Retrieved from https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/