Part 1 – Patient Data

Overview:

In this part of the final project, we were given two files which contain data about a large number of patients (approximately 20000). The first file contains some demographic and simple measurements (height, weight, etc). The second contains the results of 5 medical tests. It also has an indicator of whether the patient has a disease or not (0=yes, 1=no). Both files have an id for each patient which uniquely identifies the patient, but the id numbers were not in order and there were a couple of patients missing from each file (but not the same patient). Also, there were some missing values in the first file that need to be estimated. Apart from that, there were some noises in the second file too. After merging both files and cleaning the data, we have created a few models starting with summary statistics and using 10-fold cross validation on this data and tried to find out the best model based on the accuracy of each model. And then, our task was to discover some interesting patterns in the data. To identify the patterns, we have analyzed some of the tools we have learned in this semester (i.e., exploratory analysis, classification, clustering). The R Markdown knit file has been attached as a separate attachment.

Data Pre-Processing with Summary Statistics:

After loading the data, we have seen that both the first file (projData1a) and second file (projData have 19998 instances but 9 attributes for the first one and 6 attributes for the second one. Each of the data file has 1 class (disease or no disease) for all the instances. And while we have merged both of the files based on the "id" number, we have seen that the number of instances has decreased by 2 (19996) and the attributes summed up to 15 with 1 class (disease or no disease) for all the merged instances.

After merging the files, we have counted the missing values and identified the noises from the merged data. We have noticed that the 10th column (heartRate) has 735 missing values where 460 missing values belong to class 0 and 275 missing values belong to class 1. Then missing values were estimated by the mean of the individual class. Please see the summary of the missing values of 10th column (heartRate) below:

Class	Number of missing values	Replaced by the mean of the individual class
0 (with disease)	460	64.0587055606199
1 (without disease)	275	69.9373712901272

For the noises, we have noticed that the column 12 (testB), 13 (testC) and 15 (testE) have some values which are less than 0 or greater than 100. As per the instruction of the project, the mentioned columns should have values from 0 to 100. Hence, the values which are not in this range was detected as noises and removed them from the merged data set. Then we have validated the whole data as per the specific type of data. A summary of the cleaned data is given below:

```
Summary Analysis of the cleaned data
##        id              age        ethnic   income   marital      occGroup
##  Min.   :31938   Min.   :11.00   0:8321   0:1934   0: 5930   6      :2262
##  1st Qu.:36941   1st Qu.:37.00   1:3648   1:4084   1:14030   8      :2257
##  Median :41938   Median :43.00   2:2346   2:7982             3      :2233
##  Mean   :41939   Mean   :42.83   3:2733   3:3982             1      :2231
##  3rd Qu.:46937   3rd Qu.:49.00   4:2182   4:1978             7      :2230
##  Max.   :51937   Max.   :76.00   5: 730                      4      :2196
##                                                             (Other):6551
```

```
##  gender        weight         height         heartRate          testA
##  0:11391   Min.   :107   Min.   :53.00   Min.   :38.00   Min.   :13.00
##  1: 8569   1st Qu.:169   1st Qu.:64.00   1st Qu.:60.00   1st Qu.:44.00
##            Median :180   Median :65.00   Median :66.00   Median :50.00
##            Mean   :175   Mean   :65.04   Mean   :66.57   Mean   :50.13
##            3rd Qu.:187   3rd Qu.:67.00   3rd Qu.:75.00   3rd Qu.:56.00
##            Max.   :220   Max.   :76.00   Max.   :98.00   Max.   :96.00
##      testB           testC           testD           testE        disease
##  Min.   : 3.00   Min.   : 8.00   Min.   :11.00   Min.   : 0.00   0:11430
##  1st Qu.:43.00   1st Qu.:44.00   1st Qu.:44.00   1st Qu.:31.00   1: 8530
##  Median :50.00   Median :51.00   Median :50.00   Median :43.00
##  Mean   :51.39   Mean   :51.83   Mean   :50.12   Mean   :41.51
##  3rd Qu.:59.00   3rd Qu.:59.00   3rd Qu.:56.00   3rd Qu.:52.00
##  Max.   :99.00   Max.   :99.00   Max.   :92.00   Max.   :88.00
```

Best Classifier analysis:

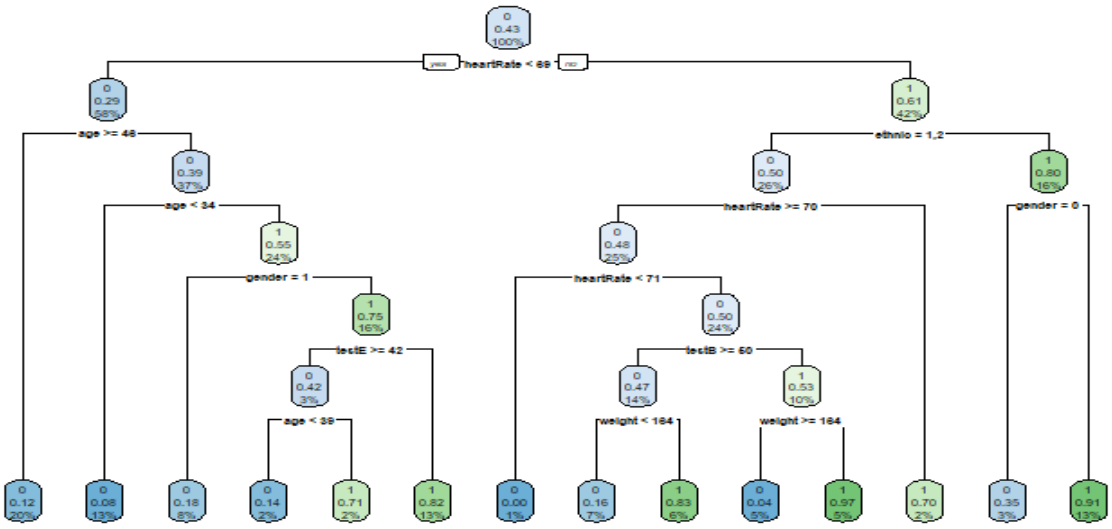
Before fitting the data into different models, we have partitioned our data into 10 different samples and then applied 10-Fold Cross Validation on the data where 1st column (id) was dropped to fit the following models:

- Decision Tree (DT)
- Naïve Bayes (NB)
- ANN
- SVM

Comparison of the models			
Model	Accuracy (%)	Precision (%)	Recall (%)
DT	86.73	81.95	86.32
NB	66.26	60.88	58.92
ANN	57.26	NA	0
SVM	Unable to Determined: Very time consuming		

Decision Tree: After fitting into the decision tree model, we have found that the main determinants for the prediction are heartrate, age and gender and this classifier has the highest accuracy (86.73%) among all the models. The confusion matrix and tree are shown below:

Confusion Matrix	Actual (0)	Actual (1)
Prediction (0)	10322	1540
Prediction (1)	1108	6990

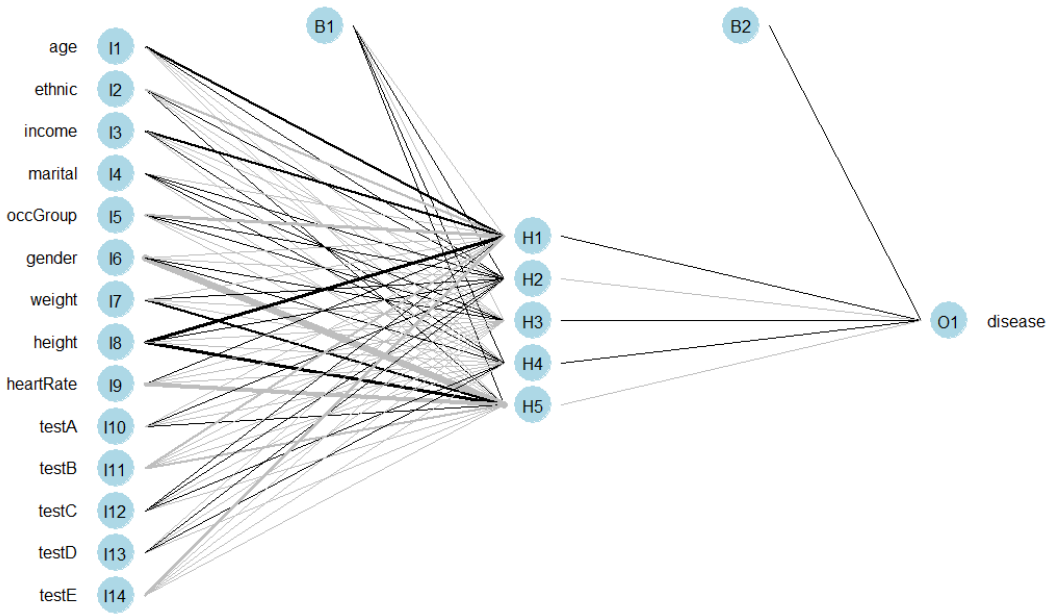


Naïve Bayes: While we fitted our data into the naïve bayes model, we observed that the accuracy, precision and recall of this model are 66.26%, 60.88% and 58.92% respectively which indicates that still the decision tree is better classifier than naïve bayes for our dataset. The confusion matrix of this model is shown below:

Confusion Matrix	Actual (0)	Actual (1)
Prediction (0)	8200	3504
Prediction (1)	3230	5026

ANN: For ANN, we have used 5 hidden nodes and set the initial weights as 0.1 in the model. After fitting the dataset into this, we have noticed that this model is not predicting anything for class value 1. Though, we started with 3 hidden nodes initially but when the 2x2 matrix was not coming, we changed the value 3 to 5 and later we increased the number of hidden nodes till 30 with a step size of 5 to see whether it gives a square matrix or not. But as it was not changing, we have set the value of hidden nodes as 5. The confusion matrix and the neural network plot are shown below:

Confusion Matrix	Actual (0)	Actual (1)
Prediction (0)	11430	8530
Prediction (1)	0	0

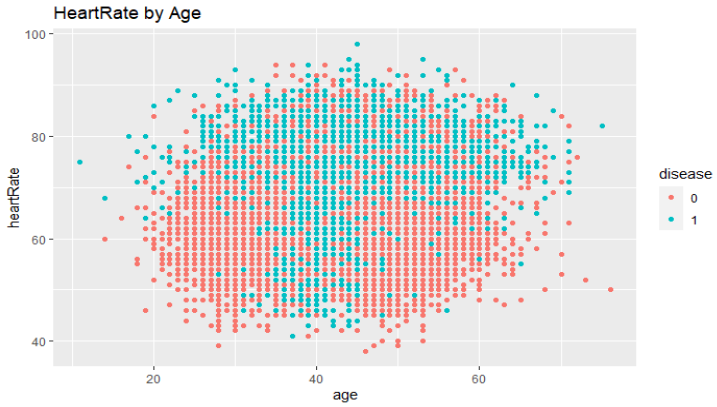


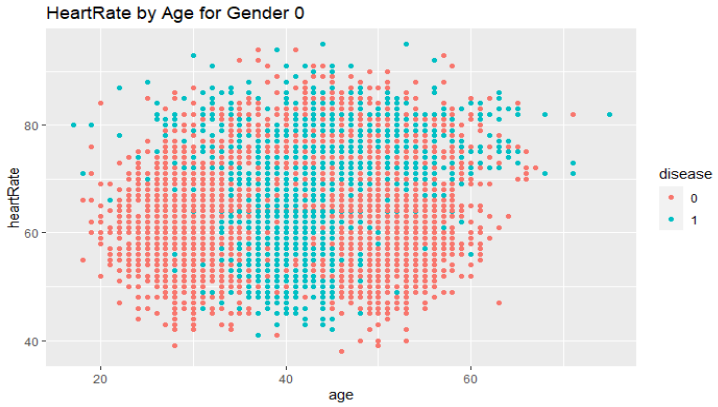
SVM: We have tried to fit our dataset into the SVM model, but it is so time consuming that we didn't get any result even after a long period of time of running the code.

So, from all the above models, the decision tree model indicates that it gives the highest accuracy along with the precision and recall. So, we can use this model as the best classifier for our dataset.

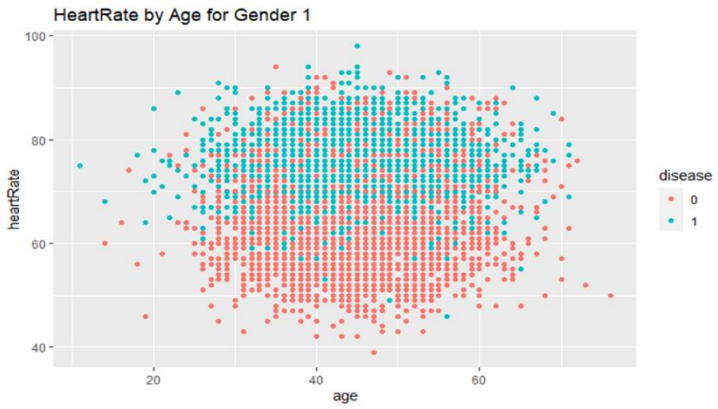
Patterns:

By using the techniques of exploratory analysis, classification and clustering, we have found a few interesting patterns from our dataset. We tried to analyze the relationship among the key determinants which play a significant role to predict the class.

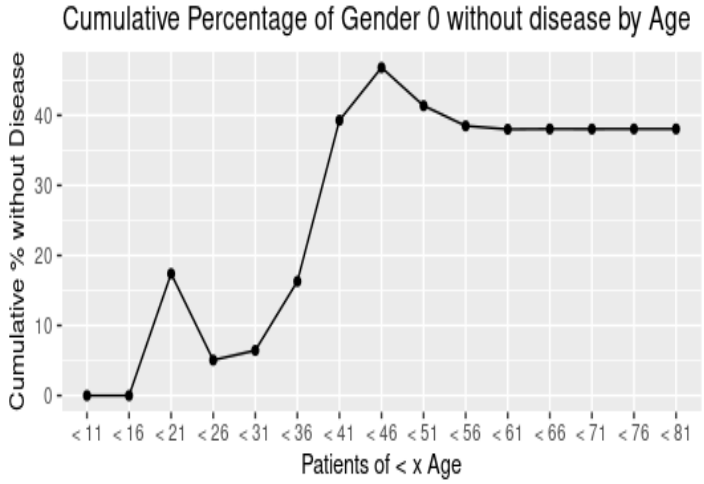


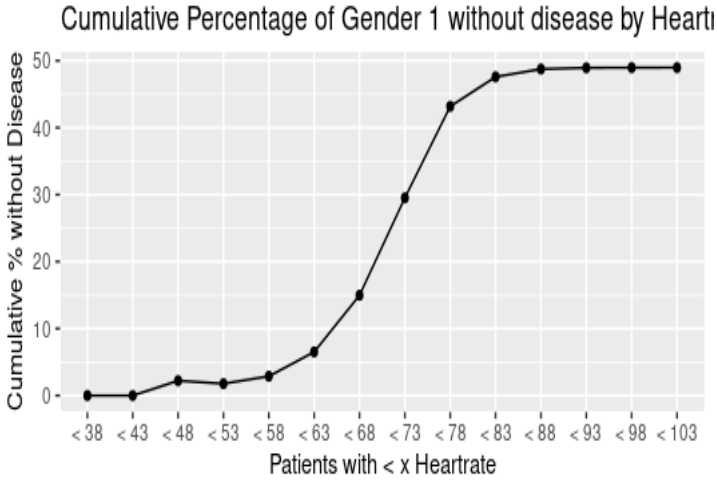


From the first scatter plot, we can say that, for the people with heart rate between 40 and 80 have high chance of getting disease in any age except the age around 40. We can see almost the same scenario for gender 0 but, a different scenario for gender 1. The people of gender 1 with heart rate above 70 have low chance of getting disease in any age. The below line graphs tell us the cumulative percentage of gender 0 and gender 1 without disease based on their age and heart rate respectively.

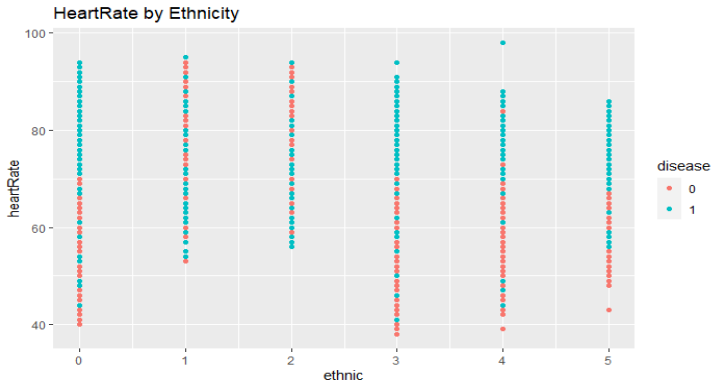


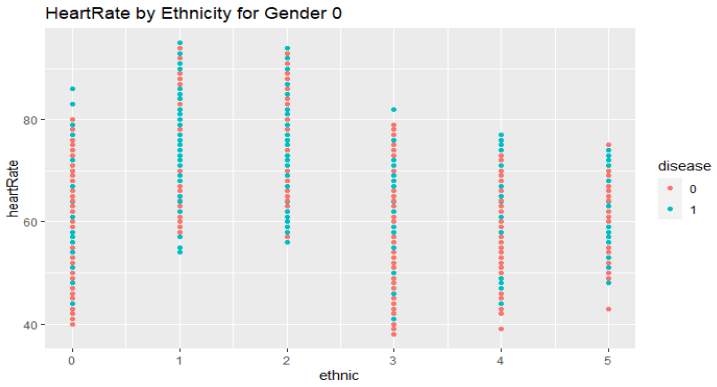
More specifically, if the age goes towards 40 of gender 0, the percentage of getting disease flactuates between 60% and 100%. Similarly, for the gender 0 of age above 50 has also high percentage of getting disease but this time it remains onstant which is arround 62% chance of getting disease. On the otherhand, Gender 1 has an interesting relationship between heart rate and percentage of getting disease. Gender 1 with heart rate below 70 has a high risk of getting disease and the percentage lies between 70% and 100%.

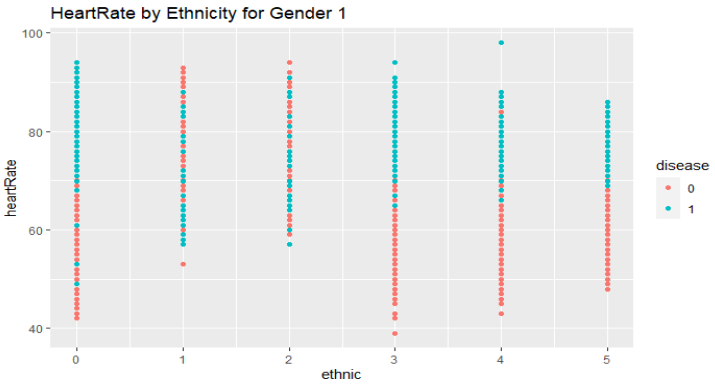


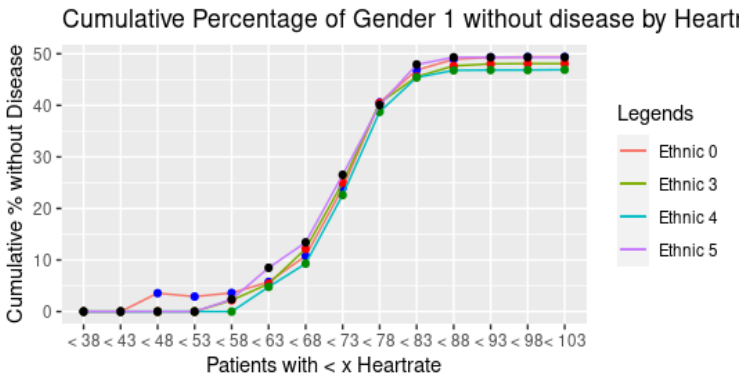


From the below plots, we have found few more interesting patterns among the heart rate, Ethnicity, gender and weight. The people of both gender with the value of Ethnicity [0, 3, 4 & 5] have low chance of getting disease if their heart rates lie above 70. It happen same for gender 1. More specifically, the cumulative percentage of not getting disease for gender 1 is 30% to 50% if their heart rates lie above 70. But, for gender 0, if the value of Ethnicity is 1, theere are high chance of not getting disease if their heart rates lie between 70 and 85.

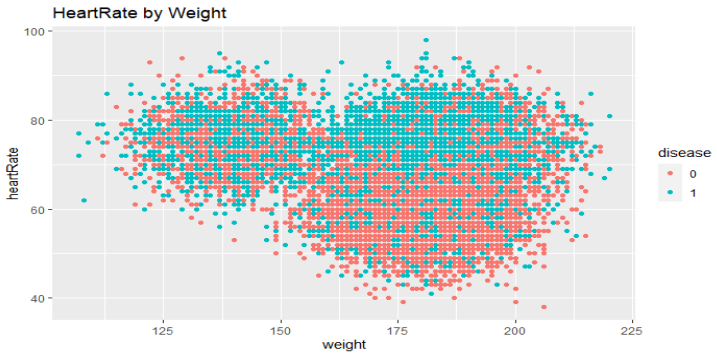


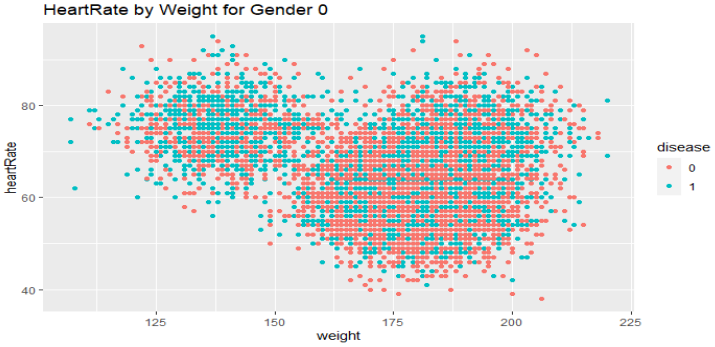


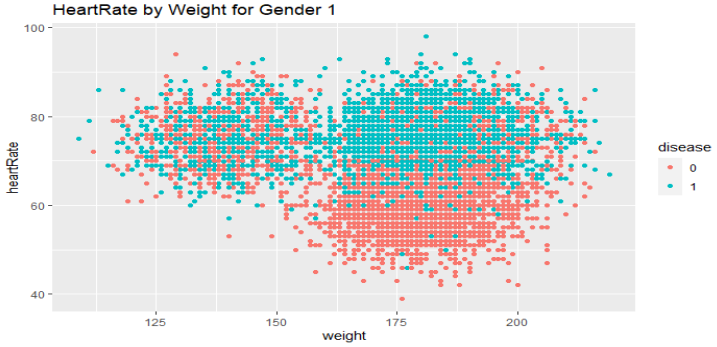




These three scatter plots show the relationship among gender, herat rate and weight. The over weight people of both gender 0 and 1 with heart rate between 40 and 70 have high risk of getting disease and again this relationship is almost the same for gender 1, whereas the low weighted people of gender 0 is comparatively safe if their heart rates lie between 60 and 90.

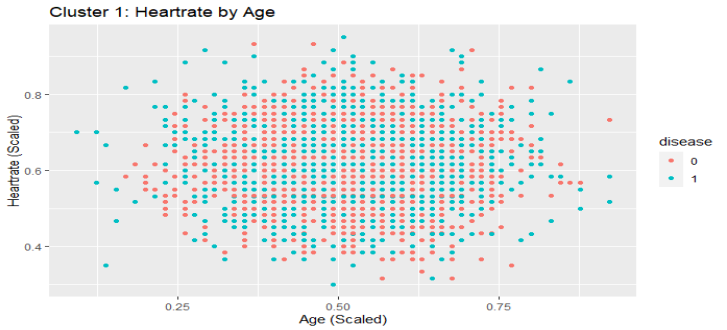


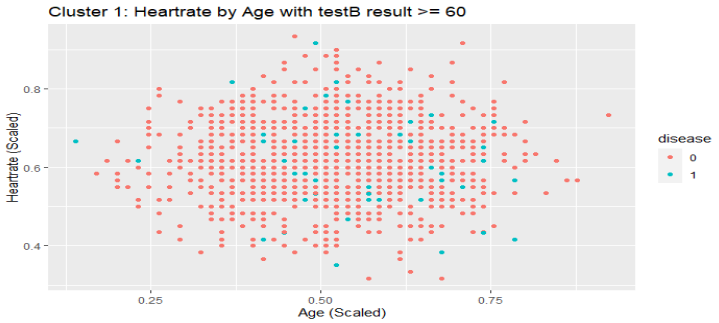


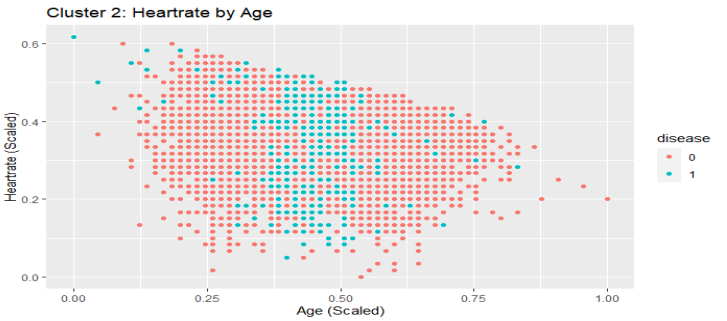


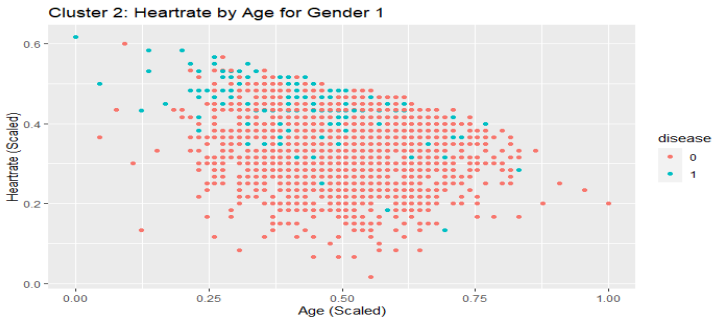
To perform the clustering analysis, we had to normalize the data which we have used for finding the interesting patterns. We have digged the below patterns for the cluster size 3. Using all the clusters, we have analyzed the relationship between heart rate and age. Later on for cluster 1 and 3, we have tried to find out a relationship of heart rate by age with testB result whereas we have found a relationship of heart rate by age with gender 1 for the cluster 2.

In cluster 1, for the relationship between heart rate and age, we have seen that the ratio of getting disease equally distributed but while we have compared this with the testB result, we have found an interesting pattern. More specifically, we have noticed that if the testB result greater than or equal to 60, almost all the patients have chance of getting disease.



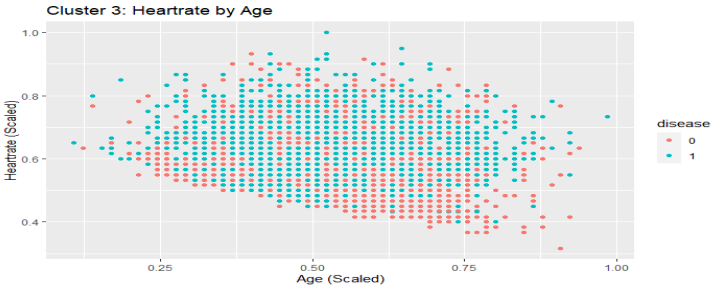


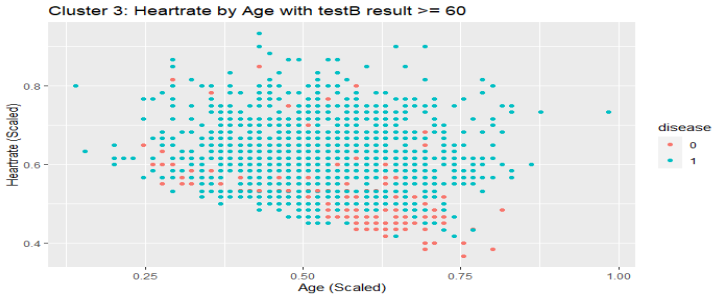




For cluster 2, we have analyzed the relationship between heart rate and age first and observed that maximum case of this cluster have chance of getting disease excpet the middle aged people. And when we have compared the relationship with the gender, we have found much more interesting pattern for gender 1. In other words, almost all the case have chance of getting disease in this cluster for gender 1.

On the otherhand, for cluster 3, we have noticed that maximum cases have chance of not getting disease for the relationship between heart rate and age and the chance increases if this relationship is compared with testB result.





Conclusion: Finally, we can come to a conclusion from the above analysis is that, the decision tree is the best model for the given dataset. And we have seen from the explanatory analysis, classification and clustering that the heart rate, age, gender, Ethnicity and the testB result were the key determinants to predict the class.