FRAMINGHAM HEART STUDY
Three Generations of Dedication
1948
1971
2002
A Project of the National Heart, Lung, and Blood Institute and Boston University

Framingham Heart Study

12/6/21

Kyle Sanborn, Nick Mitchell, Lucas Obrycki, Parth Pusegaonkar
PROFESSOR PACHAMANOVA | QTM2000

Executive Summary

Our team selected a dataset of health characteristics for nearly 4000 patients from the Framingham Heart Study in 2019. Characteristics included: sex, age, current smoker (yes/no), cigarettes smoked per day, blood pressure medication (yes/no), prior history of stroke (yes/no), prevalent to hypertension (yes/no), diabetes (yes/no), total cholesterol, systolic and diastolic blood pressure, body mass index, heart rate, glucose levels, and our target variable, risk of heart disease (yes/no). Our research is critical to help doctors predict if a particular patient is at risk of heart disease due to certain common characteristics within health records.

Our main findings based on the models are that our dataset is incredibly valuable for predicting if a patient is not at risk for heart disease, whereas it is incredibly weak at predicting a patient that is at risk. Within our supervised learning methods, we discovered that the biggest contributing factors to the risk of heart disease are record of prior stroke, cigarettes smoked per day, age and systolic blood pressure. Through unsupervised learning we discovered and reaffirmed that diabetes, prior stroke, and cigarettes smoked per day are the highest contributing factors to the risk of heart disease. Based on these findings, we first recommend that medical professionals take precaution when trying to predict the risk of heart disease, as the models we fit to the data vary in prediction accuracy. As humans get older, the intensity of care from medical professionals increases, so we have no recommendations to combat age as a significant predictor. We recommend that health professionals issue more blood pressure medication, as high systolic blood pressure is a significant contributor to risk of heart disease and there were many observations in the data with systolic blood pressure over 130 (Stage 1 Hypertension) that do not receive medication. Diabetes and prior stroke are uncontrollable, our only recommendations for health professionals is to focus more on other contributing factors for patients with these characteristics, as they are at risk for heart disease by default. Lastly, cigarettes per day is a big contributing factor, but this study was completed in 2019 and the younger generations are now turning towards vaping. Because of this we recommend that health professionals roll out widespread research on vaping as opposed to smoking, as we do not know the implications it could have on risk for heart disease, as well as many other health issues.

Analysis:

Supervised Learning

In preparation of the data for the various models we transformed male, currentSmoker, BPMeds, prevalentStroke, prevalentHyp, diabetes, and TenYearCHD to factors, however for kNN we only converted TenYearCHD to a factor and kept the predictors as numerical, and for clustering we only converted TenYearCHD to a number as a character. We decided to take out the Education variable completely because our focus is on health related metrics to predict the risk of 10 year heart disease. Total cholesterol was also taken out; the information gained from the variable could typically be inferred through the diabetes variable, but also because there are many factors impacting cholesterol that are not in the dataset, so we did not have a reasonable way of replacing the missing data with accuracy. The NAs in CigsPerDay were replaced with 20, the average smoker in our dataset smoked ~18 cigarettes, but we decided to round to 20 given that a typical smoker in the dataset smoked in denominations of packs, or 20s. BPMeds describes whether or not the patient takes blood pressure meds. Because we had the systolic and diastolic blood pressure, we were able to replace missing values based on the rule that if systolic blood pressure is above 120 and diastolic is above 80, the patient is on blood pressure medication. There were also various missing values in the BMI, heartRate, and glucose categories, for which we decided to drop similar to cholesterol, because we had no way of meaningfully predicting these variables based on the other information about the observation. Also, it is important to note that we felt comfortable dropping observations because our dataset was large, containing 3825 observations still after our processing. After we had prepped the data, we dove into the analysis portion of the project.

Logistic Regression is a useful test to run when considering a situation with binary outcomes. In order to run the logistic regression, the appropriate categorical variables had to be converted into factors, and the data had to be randomly split into a training and a test set. While the overall accuracy of 85.29% may suggest the model has strong predicting capabilities, diving into the specificity and sensitivity uncovers a more nuanced conclusion. The sensitivity is 9.62%, meaning 9.62% of at-risk patients were correctly predicted to be at risk for heart disease, whereas the specificity of 99.3% means 99.3% of patients not at risk were predicted correctly by the model. These two metrics show a severe imbalance in strength of the model, it is extremely strong at predicting a patient not at risk for heart disease and extremely weak at predicting those who are at risk. A lift chart for the model shows a fairly flat curve that is close in proximity to the 45º line, as well as an AUC of .736. This shows that: while the model is a stronger predictor than random selection, it is not significantly stronger. This was to be expected however based on the difference between the sensitivity and specificity of the model. Taking a look at the coefficients of the model was interesting, as the coefficient for prior stroke had a coefficient of 3.3016320807, suggesting that is the most influential predictor variable. Another important coefficient was the coefficient of 1.0215558560 for cigarettes smoked per day, meaning for every one cigarette increase per day, the odds of a patient being at risk for heart disease increase by 2.16%, holding all else constant. This is a very significant number as we mentioned smokers typically smoke in denominations of packs, so an increase of one cigarette per day is a realistic possibility for many smokers. We then ran a best subset exhaustive search, which concluded that the best model for logistic regression would exclude variables diabetes, cholesterol, heart rate, and BMI. The first three variables chosen by the exhaustive search were age, cigarettes per day, and systolic blood pressure, implying that they are heavily influential on risk of heart disease, and reaffirming the significance of cigarettes per day. See **Appendix Section E.**

K Nearest Neighbors (kNN) is a numerical data driven approach to classification, which uses the "nearest neighbors" in the data to classify observations by Euclidean distance calculation. Prior to application of the model, the predictor variables all had to be set to numerical values, the target variable had to be set to a factor of two, and the predictor variables had to be normalized. With application of the model, it is evident that a model with 7 nearest neighbors is applied, which generates the highest accuracy. The model has a fairly high accuracy of 84.64%, a very high specificity value of 96.79%, and a fairly high negative predictive value of 86.81%. The model is extremely better at predicting negative values than positive ones, as the sensitivity and positive predictive values are 11.47% and 37.31%, respectively. K Nearest Neighbors suggested the model as beneficial for predicting only whether a patient is not at risk for heart disease. See **Appendix Section A**.

Naive Bayes appears at first to be a good model for predicting whether or not one is at risk for heart disease based on their medical information, but is only beneficial for predicting whether someone is not at risk for heart disease. The overall accuracy of the model is 80.92%, which is much better than random selection, but a look at sensitivity and specificity shows us that the model is only 25.94% accurate when trying to predict whether one is at risk for heart disease, while the model is 91.09% accurate when predicting whether one is not at risk for heart disease. The precision of the model is 35.03%, meaning of the observations the model predicted as at risk for heart disease, 35.03% were actually at risk. The negative predictive value of the model is 86.92%, meaning of the observations the model predicted as not at risk for heart disease, 86.92% were correctly identified. Naive Bayes suggests that the dataset is better used to predict if a patient is not at risk for heart disease. See **Appendix Section B**.

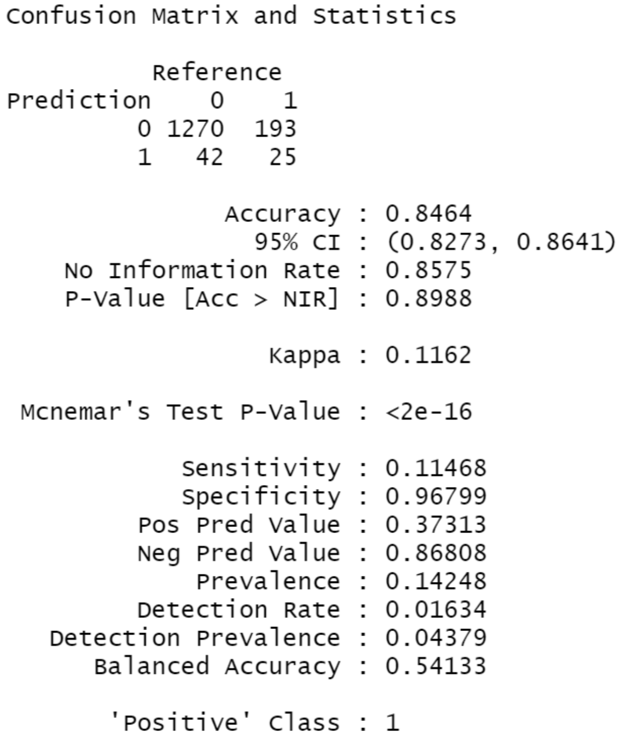
Classification trees are intended to help classify an outcome based on a set of predictors. We intended to classify the variable that contributes the most in causing a ten year risk of heart disease. From our tree we are able to learn that the two leading predictors influencing risk of heart disease, as age is the first classifier in our tree and systolic blood pressure is the second classifier. The tree indicates those most likely to have a ten year risk of heart disease are a person above 47 years old and a person with a systolic blood pressure above 155. We analyzed a confusion matrix to help us better understand our results because our target variable is categorical. From our confusion matrix we were able to learn that 958 of actual at-risk patients were correctly predicted by the algorithm, 145 at-risk patients were incorrectly predicted as not at risk by the algorithm, 25 non risk patients were predicted as at risk by the algorithm, and 20 observations were correctly identified as at risk for heart disease. This model has an accuracy of 85.19% with a sensitivity of 12.12% and a specificity of 97.46%. This sensitivity value tells us that our model is only 12.12% accurate when trying to predict whether one is at risk for heart disease, while the specificity value tells us that the model is 97.46% accurate when predicting whether one is not at risk for heart disease. As with our other supervised learning models, the classification tree is much more accurate in predicting patients not at risk compared to those at risk. **See Appendix Section C**.

Unsupervised Learning

Clustering is an exploratory technique used to discover structure within a set of data and to summarize the properties of each cluster. We ran a hierarchical clustering with 3 clusters. Each cluster had its own profile, cluster 1 is the heavy smokers, who do not have diabetes or a previous stroke, cluster 2 is those who have had a previous stroke and do not have diabetes, cluster 3 is those who have never had a stroke before, but do have diabetes. Cluster three represents only patients who have not had a stroke before, but do have diabetes, and has a mean value for risk of heart disease equal to .61, suggesting patients with diabetes are at the highest risk for heart disease. Cluster two has a mean diabetes value of 0.05, meaning the cluster represents mostly patients without diabetes, as well as a prevalentStroke mean value of 1.00, meaning it represents only patients who have previously had a stroke. Cluster two also has the highest average for ten year risk of heart disease with a value of 0.41, suggesting patients without diabetes who have had a stroke are at high risk for heart disease. Cluster one represents patients who have never had a stroke, do not have diabetes, and do not take blood pressure medication. Cluster one has the highest cigs per day average at 9.04 as well as the highest average for current smokers at 0.49. Cluster one has a mean value for risk of heart disease equaling .15. This suggests that smoking is a factor contributing to heart disease, but not the biggest contributor. Based on our model, these are the biggest differentiating factors between each cluster, making diabetes, then previous stroke, then cigarettes per day the three biggest contributing factors to risk of heart disease. See **Appendix Section D**.

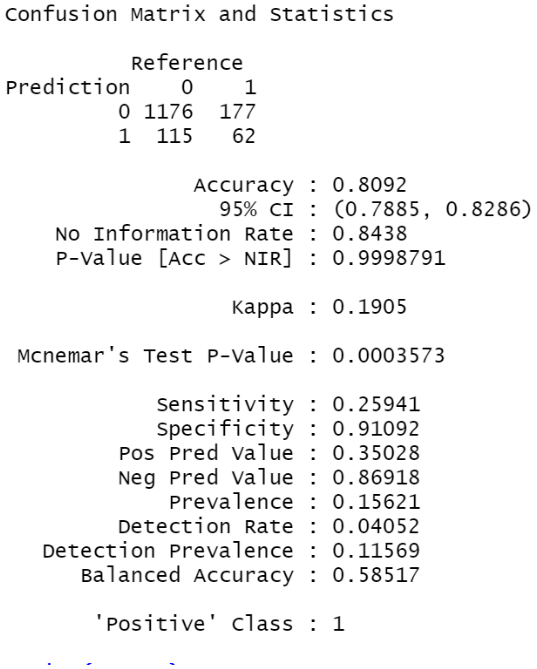
Appendix

- Section A (KNN)



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1270  193
         1   42   25
                                          
               Accuracy : 0.8464          
                 95% CI : (0.8273, 0.8641)
    No Information Rate : 0.8575          
    P-Value [Acc > NIR] : 0.8988          
                                          
                  Kappa : 0.1162          
                                          
 Mcnemar's Test P-Value : <2e-16          
                                          
            Sensitivity : 0.11468         
            Specificity : 0.96799         
         Pos Pred Value : 0.37313         
         Neg Pred Value : 0.86808         
             Prevalence : 0.14248         
         Detection Rate : 0.01634         
   Detection Prevalence : 0.04379         
      Balanced Accuracy : 0.54133         
                                          
       'Positive' Class : 1               
```

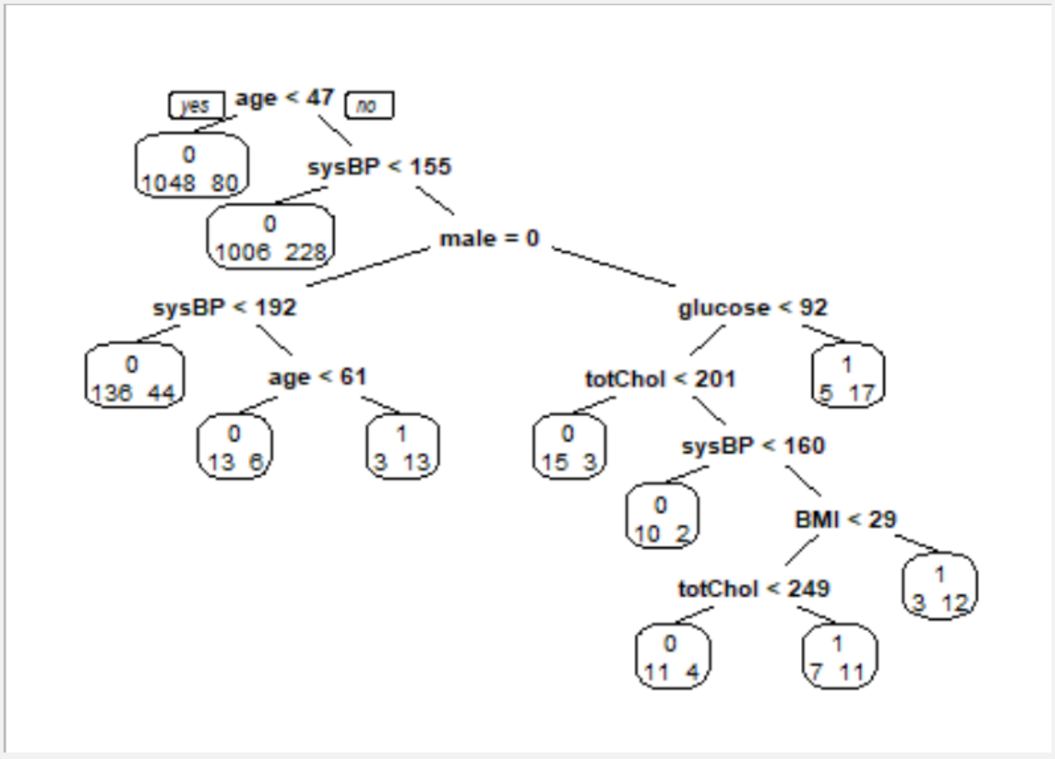
- Section B (Naive Bayes)

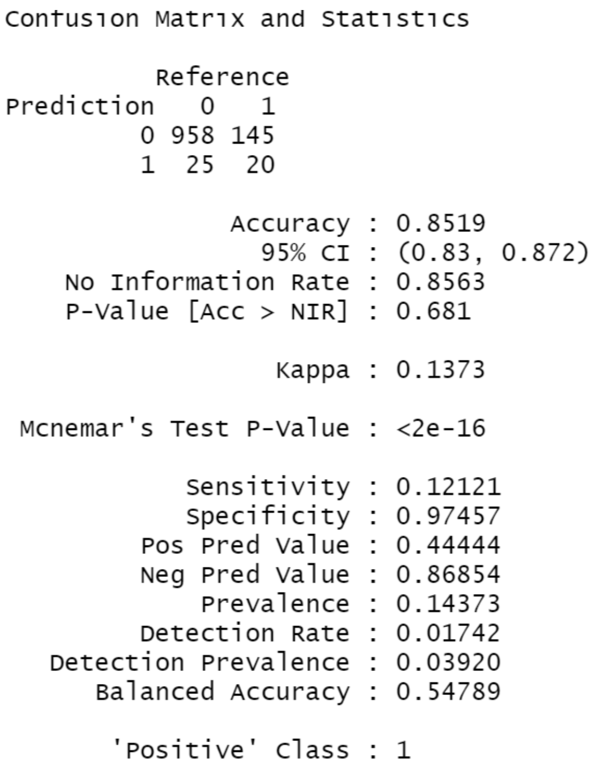


```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1176  177
         1  115   62
                                          
               Accuracy : 0.8092          
                 95% CI : (0.7885, 0.8286)
    No Information Rate : 0.8438          
    P-Value [Acc > NIR] : 0.9998791       
                                          
                  Kappa : 0.1905          
                                          
 Mcnemar's Test P-Value : 0.0003573       
                                          
            Sensitivity : 0.25941         
            Specificity : 0.91092         
         Pos Pred Value : 0.35028         
         Neg Pred Value : 0.86918         
             Prevalence : 0.15621         
         Detection Rate : 0.04052         
   Detection Prevalence : 0.11569         
      Balanced Accuracy : 0.58517         
                                          
       'Positive' Class : 1               
```

- <u>Section C (Classification Tree)</u>

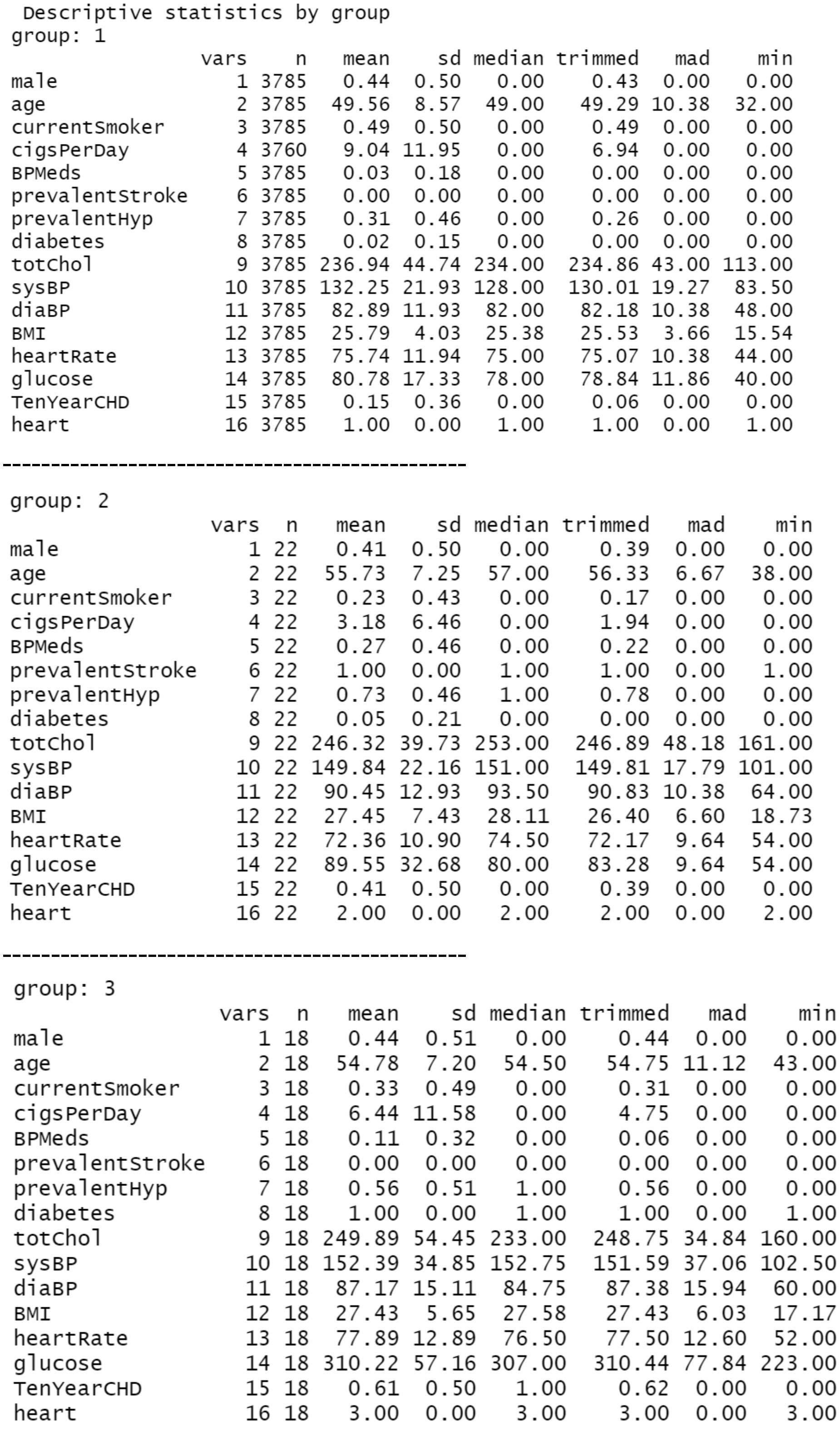
Classification Tree:



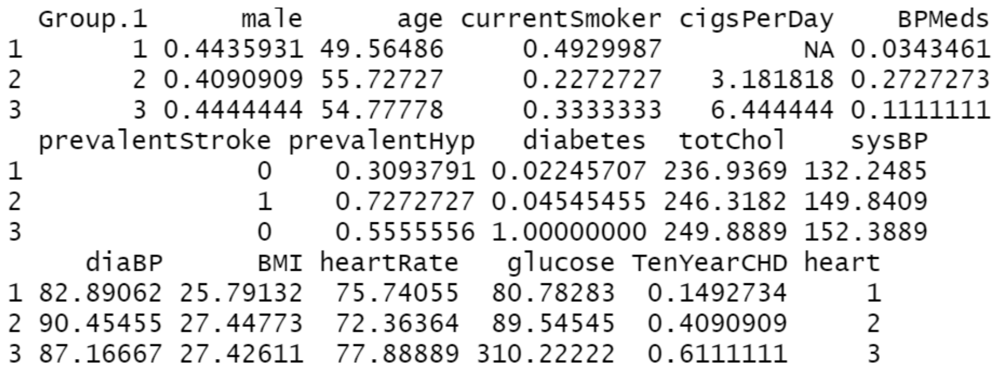


```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 958 145
         1  25  20

               Accuracy : 0.8519
                 95% CI : (0.83, 0.872)
    No Information Rate : 0.8563
    P-Value [Acc > NIR] : 0.681

                  Kappa : 0.1373

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.12121
            Specificity : 0.97457
         Pos Pred Value : 0.44444
         Neg Pred Value : 0.86854
             Prevalence : 0.14373
         Detection Rate : 0.01742
   Detection Prevalence : 0.03920
      Balanced Accuracy : 0.54789

       'Positive' Class : 1
```

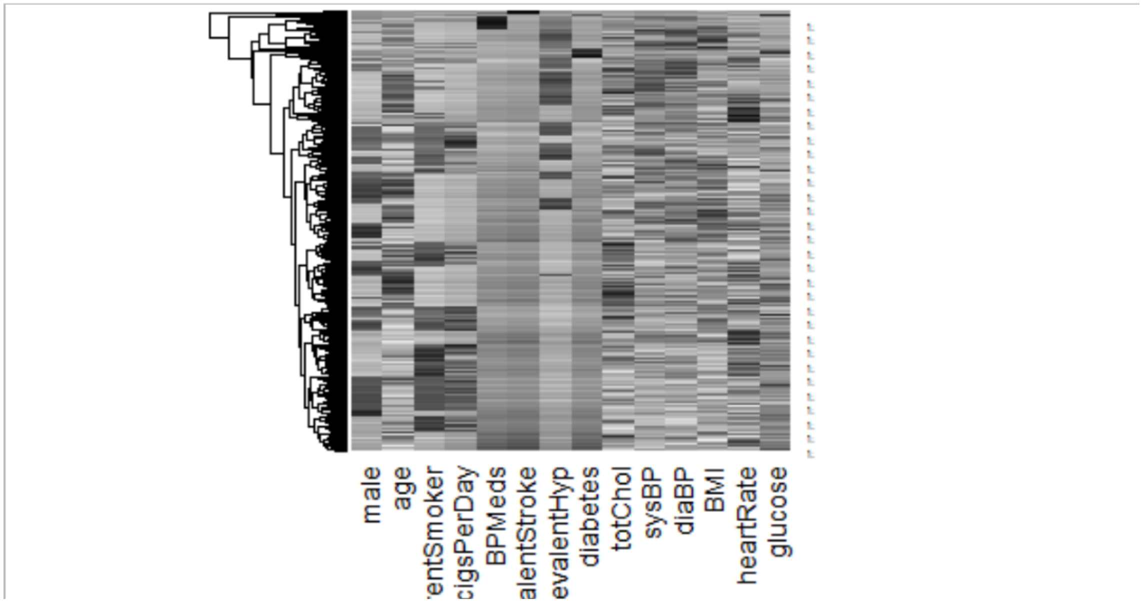
- Section D (Clustering)



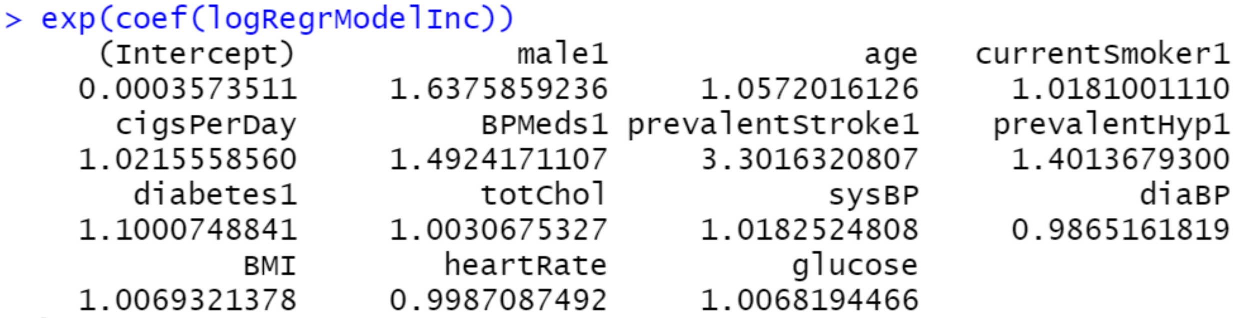
```
 Descriptive statistics by group 
group: 1
                vars    n   mean    sd median trimmed   mad    min
male               1 3785   0.44  0.50   0.00    0.43  0.00   0.00
age                2 3785  49.56  8.57  49.00   49.29 10.38  32.00
currentSmoker      3 3785   0.49  0.50   0.00    0.49  0.00   0.00
cigsPerDay         4 3760   9.04 11.95   0.00    6.94  0.00   0.00
BPMeds             5 3785   0.03  0.18   0.00    0.00  0.00   0.00
prevalentStroke    6 3785   0.00  0.00   0.00    0.00  0.00   0.00
prevalentHyp       7 3785   0.31  0.46   0.00    0.26  0.00   0.00
diabetes           8 3785   0.02  0.15   0.00    0.00  0.00   0.00
totChol            9 3785 236.94 44.74 234.00  234.86 43.00 113.00
sysBP             10 3785 132.25 21.93 128.00  130.01 19.27  83.50
diaBP             11 3785  82.89 11.93  82.00   82.18 10.38  48.00
BMI               12 3785  25.79  4.03  25.38   25.53  3.66  15.54
heartRate         13 3785  75.74 11.94  75.00   75.07 10.38  44.00
glucose           14 3785  80.78 17.33  78.00   78.84 11.86  40.00
TenYearCHD        15 3785   0.15  0.36   0.00    0.06  0.00   0.00
heart             16 3785   1.00  0.00   1.00    1.00  0.00   1.00

---------------------------------------------

group: 2
                vars  n   mean    sd median trimmed   mad    min
male               1 22   0.41  0.50   0.00    0.39  0.00   0.00
age                2 22  55.73  7.25  57.00   56.33  6.67  38.00
currentSmoker      3 22   0.23  0.43   0.00    0.17  0.00   0.00
cigsPerDay         4 22   3.18  6.46   0.00    1.94  0.00   0.00
BPMeds             5 22   0.27  0.46   0.00    0.22  0.00   0.00
prevalentStroke    6 22   1.00  0.00   1.00    1.00  0.00   1.00
prevalentHyp       7 22   0.73  0.46   1.00    0.78  0.00   0.00
diabetes           8 22   0.05  0.21   0.00    0.00  0.00   0.00
totChol            9 22 246.32 39.73 253.00  246.89 48.18 161.00
sysBP             10 22 149.84 22.16 151.00  149.81 17.79 101.00
diaBP             11 22  90.45 12.93  93.50   90.83 10.38  64.00
BMI               12 22  27.45  7.43  28.11   26.40  6.60  18.73
heartRate         13 22  72.36 10.90  74.50   72.17  9.64  54.00
glucose           14 22  89.55 32.68  80.00   83.28  9.64  54.00
TenYearCHD        15 22   0.41  0.50   0.00    0.39  0.00   0.00
heart             16 22   2.00  0.00   2.00    2.00  0.00   2.00

---------------------------------------------

group: 3
                vars  n   mean    sd median trimmed   mad    min
male               1 18   0.44  0.51   0.00    0.44  0.00   0.00
age                2 18  54.78  7.20  54.50   54.75 11.12  43.00
currentSmoker      3 18   0.33  0.49   0.00    0.31  0.00   0.00
cigsPerDay         4 18   6.44 11.58   0.00    4.75  0.00   0.00
BPMeds             5 18   0.11  0.32   0.00    0.06  0.00   0.00
prevalentStroke    6 18   0.00  0.00   0.00    0.00  0.00   0.00
prevalentHyp       7 18   0.56  0.51   1.00    0.56  0.00   0.00
diabetes           8 18   1.00  0.00   1.00    1.00  0.00   1.00
totChol            9 18 249.89 54.45 233.00  248.75 34.84 160.00
sysBP             10 18 152.39 34.85 152.75  151.59 37.06 102.50
diaBP             11 18  87.17 15.11  84.75   87.38 15.94  60.00
BMI               12 18  27.43  5.65  27.58   27.43  6.03  17.17
heartRate         13 18  77.89 12.89  76.50   77.50 12.60  52.00
glucose           14 18 310.22 57.16 307.00  310.44 77.84 223.00
TenYearCHD        15 18   0.61  0.50   1.00    0.62  0.00   0.00
heart             16 18   3.00  0.00   3.00    3.00  0.00   3.00
```



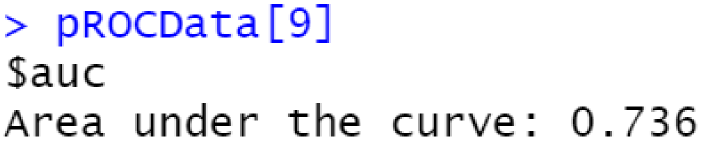
```
  Group.1      male      age currentSmoker cigsPerDay    BPMeds
1       1 0.4435931 49.56486     0.4929987         NA 0.0343461
2       2 0.4090909 55.72727     0.2272727   3.181818 0.2727273
3       3 0.4444444 54.77778     0.3333333   6.444444 0.1111111
  prevalentStroke prevalentHyp   diabetes  totChol    sysBP
1               0    0.3093791 0.02245707 236.9369 132.2485
2               1    0.7272727 0.04545455 246.3182 149.8409
3               0    0.5555556 1.00000000 249.8889 152.3889
     diaBP      BMI heartRate   glucose TenYearCHD heart
1 82.89062 25.79132  75.74055  80.78283  0.1492734     1
2 90.45455 27.44773  72.36364  89.54545  0.4090909     2
3 87.16667 27.42611  77.88889 310.22222  0.6111111     3
```



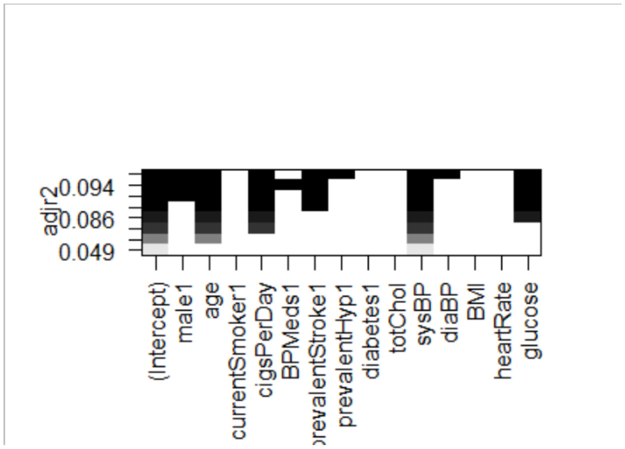
- Appendix Section E:

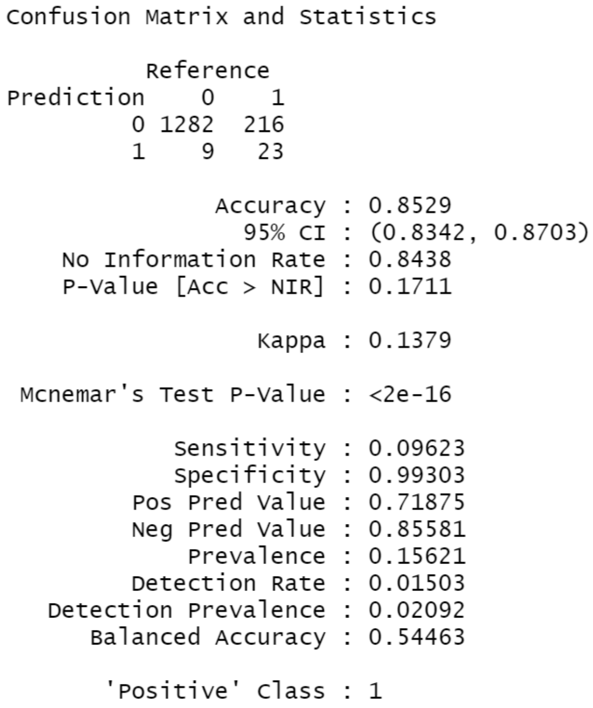


```
> exp(coef(logRegrModelInc))
   (Intercept)           male1             age   currentSmoker1
  0.0003573511    1.6375859236    1.0572016126     1.0181001110
    cigsPerDay          BPMeds1 prevalentStroke1   prevalentHyp1
  1.0215558560    1.4924171107    3.3016320807     1.4013679300
     diabetes1          totChol            sysBP           diaBP
  1.1000748841    1.0030675327    1.0182524808     0.9865161819
           BMI        heartRate          glucose
  1.0069321378    0.9987087492    1.0068194466
```

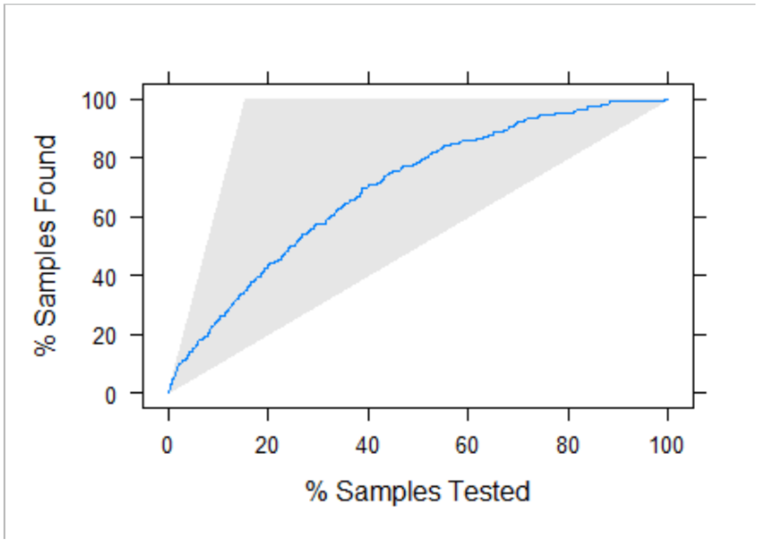


```
> pROCData[9]
$auc
Area under the curve: 0.736
```





```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1282  216
         1    9   23

               Accuracy : 0.8529
                 95% CI : (0.8342, 0.8703)
    No Information Rate : 0.8438
    P-Value [Acc > NIR] : 0.1711

                  Kappa : 0.1379

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.09623
            Specificity : 0.99303
         Pos Pred Value : 0.71875
         Neg Pred Value : 0.85581
             Prevalence : 0.15621
         Detection Rate : 0.01503
   Detection Prevalence : 0.02092
      Balanced Accuracy : 0.54463

       'Positive' Class : 1
```


% Samples Found
% Samples Tested