CARDIOVASCULAR DISEASE ANALYSIS AND PREDICTION

OBJECTIVES AND SIGNIFICANCE

This mini project is a result of insight gained on several subject matter covering a small part of Data Mining. There are majorly two goals we have worked on throughout the span of the project.

i. *To identify rules that can be produced from the variables for cardiovascular disease.*

We as a team had a moto to best predict the effects of several factors on a human heart. In order to bring out meaningful observation from given data one must dig into the patterns and rules followed by the attributes of different records in the dataset.

ii. *To create a model that can predict presence of cardiovascular disease for new cases .*

After fulfilling the first goal, we then kickstarted building the most reliable model for our dataset. Since our data was a mix of both categorical and continuous variables, we had to choose a technique that satisfies incorporation of both.

To better understand the significance of this project, its necessary to understand impact of cardiovascular diseases on our life. The term "cardiovascular disease" is often used interchangeably with the term “heart disease”. According to the World Health Organization Cardiovascular diseases (CVDs) are a group of disorders of the heart and blood vessels and they include - coronary heart disease, cerebrovascular disease, peripheral arterial disease, rheumatic heart disease, congenital heart disease, deep vein thrombosis and pulmonary embolism. These terms may sound futile to a common man who comes with no background in biology, but their impact is life threatening.

CVDs are the number 1 cause of death globally; more people die annually from CVDs than from any other cause. In 2008, over 616,000 people died of heart disease. In 2010, coronary heart disease alone was projected to cost the United States $108.9 billion. More than 159,000 women die each year of congestive heart failure, accounting for 56.3% of all heart failure deaths. These huge figures might give an idea of how important the study of this field is.

Implementing a project that suffice our academic commitment as well as personal benefit to stimulate confidence by carrying out an independent project is the major motivation.

BACKGROUND

There is a lot of research already going on in the field of heart diseases. Only 6 decades ago, we didn't know what caused cardiovascular disease, and many Americans died of heart attacks in their 50s or 60s. By the late 1940s, cardiovascular disease was responsible for half of all U.S. deaths. NIH-funded research, beginning with the Framingham Heart Study in the late 1940s, helped to define the concept of risk factors and changed the course of public health. Although heart disease remains the leading cause of death nationwide, the death rate for heart disease has dropped by more than 60% since 1940.

Now we have new and more effective treatments such as clot-buster drug therapy to open blocked arteries. We also know that heart attack, sudden death, and stroke can often be prevented by quitting smoking, controlling high blood pressure, exercising regularly, and taking therapies like statins, aspirin, and beta-blockers. Increasingly, we can pinpoint those at highest risk for future illness — even before any symptoms appear—and offer them effective prevention strategies. In the course of this project we have tried noticing the effects of several such factors on a human heart.

Many aspiring candidates of health care like us have already build several kernels on different topics of data mining. We have carried out this project differently by not just focusing on part of the subject but taking care of the whole process involved in building a full-fledged model.

METHODS

About the dataset

The dataset was downloaded from the Kaggle repository which is an online community of data scientists and machine learners allowing users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges. The link to our data is here.

Dataset comprised of a total of 70,000 records with 12 attributes and 1 target variable making it 70,000 x 13 dimensional.

There were three types of input features:

Types of Input Features					
Objective		**Subjective**		**Examination**	
Age	in days	**Smoking**	Binary 1 = Yes 0 = No	**Systolic blood pressure**	Integer
Height	in cm	**Alcohol Intake**	Binary 1 = Yes 0 = No	**Diastolic blood pressure**	Integer
Weight	in kgs	**Physical Activity**	Binary 1 = Yes 0 = No	**Cholesterol**	Categorical 1 = Normal 2 = Above Normal 3 = Well Above Normal
Gender	Categorical 1 = Female 2 = Male			**Glucose**	Categorical 1 = Normal 2 = Above Normal 3 = Well Above Normal

Table 1 – shows the different variables and their types

Target Variable was "Cardio" which was Binary type where 1 represented the presence of CVD and 0 represented the absence of CVD.

Methodology

Whole project was carried out on R which is a programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing using RStudio – an open source and enterprise ready professional software for R. Since we were required to carry out ample of visualization, prediction and calculation tasks, we used several libraries such as ggplot2, dplyr, caret, ROCR and many more.

Importing the dataset: The data was in comma separated values (CSV) format. Each cell inside such data file is separated by a special character, which usually is a comma, although other characters can be used as well. We used function read.csv to import our dataset.

Data Preprocessing and Visualization: It is the first step in data analysis and involves summarizing the main characteristics of a data set, including its size, accuracy, initial patterns in the data and other attributes. Using head() and summary() functions () we saw that all our variables had numerical values but given by the description, variables like cholesterol and gluc were categorical, and gender, smoke, alco, active, cardio were binary in nature. Attributes such as alco stands for alcohol consumption, and active stands for physical activity, they were both binary in nature, ap_hi and ap_lo stand for high blood pressure and low blood pressure respectively.

Our dataset had no missing values. We looked for inaccuracies/inconsistency throughout the dataset by creating a duplicate dataset so that we don't alter any client data. With the help of several articles on heart and body standards, we deleted records that were outliers to our dataset and not practically recorded in the world of health care yet such as diastolic blood pressure cannot be higher than systolic pressure, dataset cannot have negative or 0 measurements for blood pressure; 0 means no blood is being pumped, we cannot have systolic pressure more than 300. Also, we converted the age in the duplicated dataset into years, removed any weight lesser than 30 kgs, and a height lesser than 100 cms.

In multivariate analysis, we looked at the correlation matrix(shown in preprocessing results section) and saw that systolic blood pressure and diastolic blood pressure are weakly correlated with our target variable cardio. Cholesterol and age also have some impact on the target variable. Cholesterol and glucose levels as have a moderate correlation. We also created a new variable in our duplicated dataset called Body Mass Index which helped us understand if a person was overweight or underweight.

Visualization results had given us a clearer picture. Female population was dominating male in terms of CVD active. Cases reported had bigger population that has normal cholesterol levels and still bearing CVD. Most of the patients were then indulged in exercising after being detected with

CVDs as we all know that exercising keeps heart healthy. Using box plots, we saw that both male and female population with CVDs tend to have higher BMI which is an indication that weight disproportionate to height can lead to unhealthy functioning of heart.

By splitting our data in 70:30 ratio as training and test respectively we started with different models. We have used two models using logistic regression alone and measured their accuracy, area under the curve to evaluate their performance. Model which gave better results was then applied on the validation data. Variable reduction was done using StepAIC and for better understanding of effects of predictors on the response. After careful analysis of classifiers and removing variables we chose to use second model fit_Test_2 on our Validation dataset. From the summary() of initial model, we clearly saw that gender was not influential in cardiovascular disease. Whereas, glucose, smoking status, alcohol consumption showed a negative relation with respect to people not having cardio diseases. To have more insight on effect of actively exercising on cardio diseases, we opted to have a look at Odds Ratio on classifier 1 and realized that age, height, high BP, low BP, cholesterol, and BMI stood out to be more prominent in deciding if a person had CVD or not.

To have a second opinion on variable selected and importance we used random forest.
First, we build a default classification tree to see what rules it comes up with.

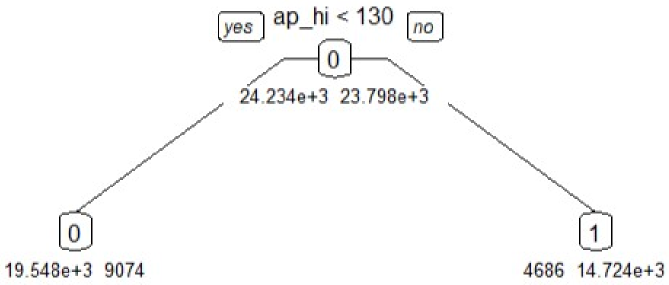
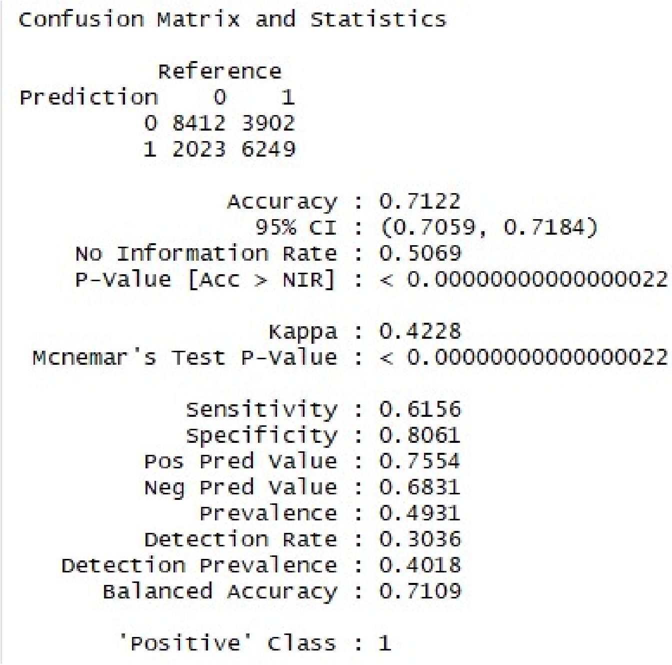


Figure 1 – Rule 1

According to the default classification tree, the rule is IF ap_hi >= 130 THEN class = 1. Which meant if systolic pressure is more than 130 our tree classifies them as having cardiovascular disease.

We checked the accuracy of this model on our validation set.

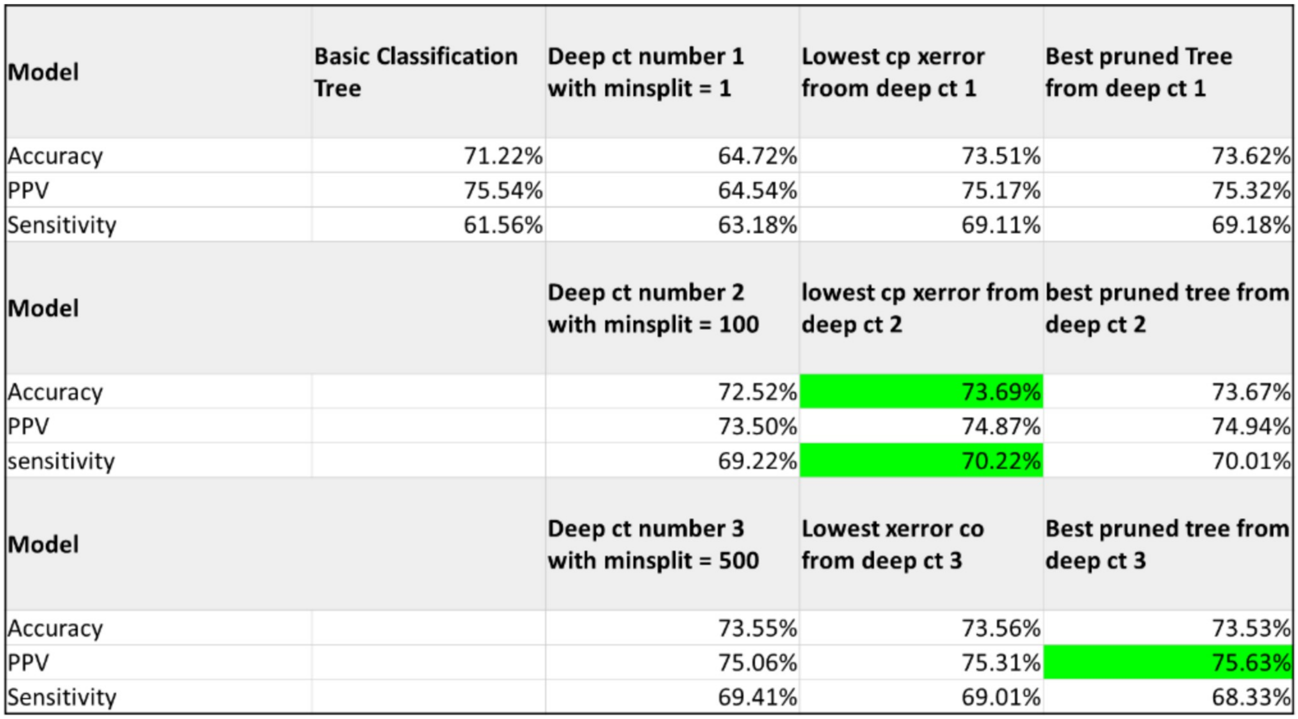


```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 8412 3902
         1 2023 6249

               Accuracy : 0.7122
                 95% CI : (0.7059, 0.7184)
    No Information Rate : 0.5069
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.4228
 Mcnemar's Test P-Value : < 0.00000000000000022

            Sensitivity : 0.6156
            Specificity : 0.8061
         Pos Pred Value : 0.7554
         Neg Pred Value : 0.6831
             Prevalence : 0.4931
         Detection Rate : 0.3036
   Detection Prevalence : 0.4018
      Balanced Accuracy : 0.7109

       'Positive' Class : 1
```

The accuracy was 71.22%, the sensitivity was 61.56% and positive predictive value was 75.54%. The effort was now to grow various deep trees with changing the values of minimum split bucket. The tree rules change for different minimum bucket limits. We wanted a model that is more accurate than our default classification tree and is also deeper so we can get rules based on our predictors. We grew three deep trees with minimum split bucket of 1, 100 and 500. We checked accuracy, sensitivity and positive predictive value for each deep tree.

Then we decided to prune each tree to their lowest x errors to shorten their lengths and avoid overfitting our training data. After pruning, we went up one standard deviation for each tree to account for error.

After running all our various classification models, we decided to check their results on our validation set.



Model	Basic Classification Tree	Deep ct number 1 with minsplit = 1	Lowest cp xerror froom deep ct 1	Best pruned Tree from deep ct 1
Accuracy	71.22%	64.72%	73.51%	73.62%
PPV	75.54%	64.54%	75.17%	75.32%
Sensitivity	61.56%	63.18%	69.11%	69.18%
Model		**Deep ct number 2 with minsplit = 100**	**lowest cp xerror from deep ct 2**	**best pruned tree from deep ct 2**
Accuracy		72.52%	73.69%	73.67%
PPV		73.50%	74.87%	74.94%
sensitivity		69.22%	70.22%	70.01%
Model		**Deep ct number 3 with minsplit = 500**	**Lowest xerror co from deep ct 3**	**Best pruned tree from deep ct 3**
Accuracy		73.55%	73.56%	73.53%
PPV		75.06%	75.31%	75.63%
Sensitivity		69.41%	69.01%	68.33%

Table 2 – Comparison of all models obtained from random forest

We can see that model pruned from deep tree 2 gave us the best accuracy and sensitivity and model pruned to best tree from deep tree 3 gave us the best positive predictive value.

Both trees performed better than our default tree.

We ploted the pruned tree from deep tree 2 to check for rules produced.

Classification rules:

IF ap_hi >= 136 THEN class = 1

IF ap_hi >= 130 AND ap_hi < 136 AND cholesterol = 3 THEN class = 1

IF ap_hi >= 130 AND ap_hi < 136 AND cholesterol = 1, 2 AND age >= 59 class = 1

IF ap_hi >= 130 AND ap_hi <136 AND cholesterol = 1, 2 AND age < 59 AND ap_lo > = 90 AND smoke = 0 then class = 1

IF ap_hi >= 130 AND ap_hi <136 AND cholesterol = 1,2 AND age < 59 AND ap_lo > = 90 AND smoke = 1 THEN class = 0

IF ap_hi >= 130 AND ap_hi <136 AND cholesterol = 1,2 AND age < 59 AND ap_lo < 90 AND age > = 41 AND height < 176 AND bmi > = 32 THEN class = 1

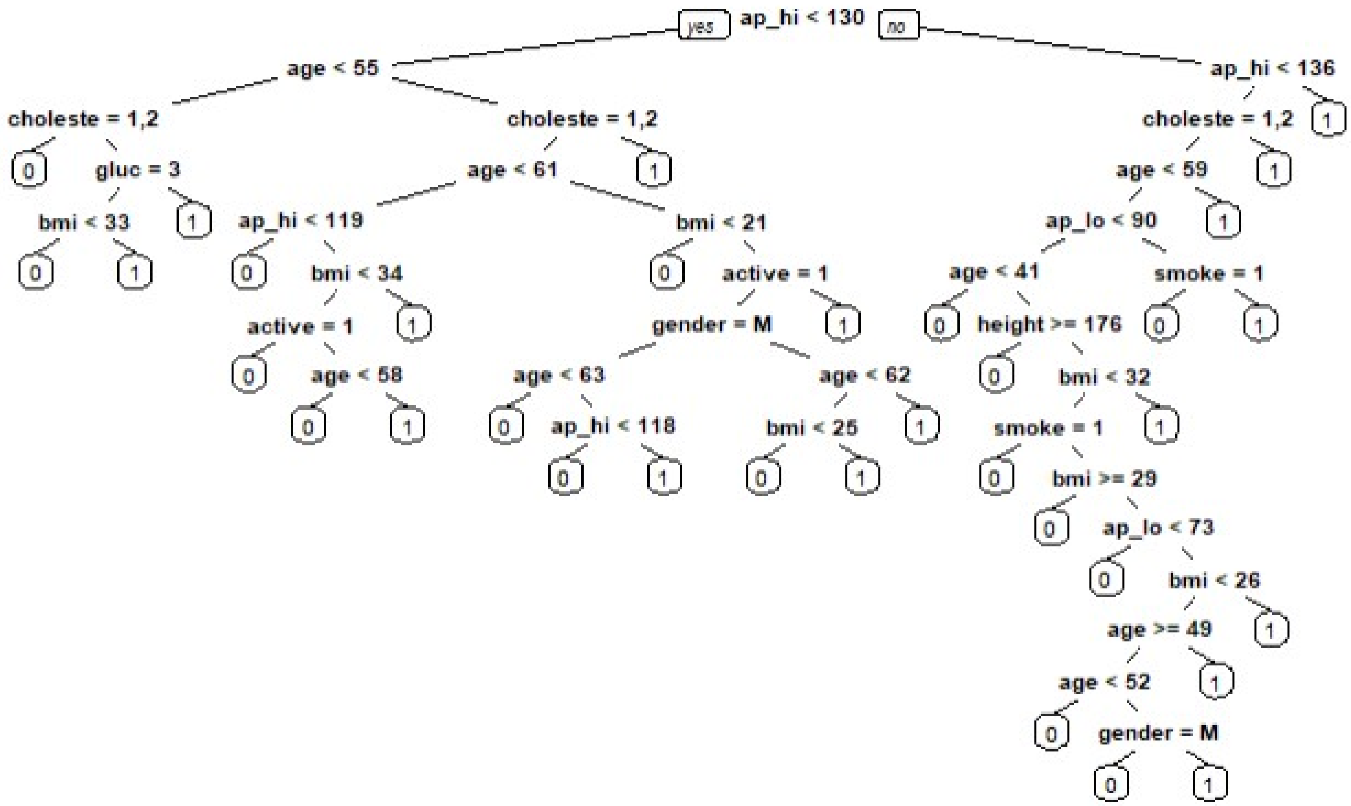


Figure 2 – Classification rules

IF ap_hi >= 130 AND ap_hi <136 AND cholesterol = 1,2 AND age < 59 AND ap_lo 90 AND age > = 41 AND height < 176 AND bmi < 32 AND smoke = 0 AND ap_lo >= 73 AND bmi >= 26 THEN class = 1

IF ap_hi >= 130 AND ap_hi <136 AND cholesterol = 1,2 AND age < 59 AND ap_lo < 90 AND height < 176 AND bmi < 32 AND smoke = 0 AND ap_lo >= 73 AND age >= 52 AND gender = F then class = 1

IF ap_hi >= 130 AND ap_hi <136 AND cholesterol = 1,2 AND age < 59 AND ap_lo < 90 AND height < 176 AND bmi < 32 AND smoke = 0 AND ap_lo >= 73 THEN class = 0

IF ap_hi >= 130 AND ap_hi <136 AND cholesterol = 1,2 AND age < 59 AND ap_lo < 90 AND height < 176 AND bmi < 32 AND smoke = 0 THEN class = 0

IF ap_hi >= 130 AND ap_hi <136 AND cholesterol = 1,2 AND age < 59 AND ap_lo < 90 AND height < 176 AND bmi < 32 AND smoke = 0 AND bmi >= 29 THEN class = 0

IF ap_hi >= 130 AND ap_hi <136 AND cholesterol = 1,2 AND age < 59 AND ap_lo < 90 AND height < 176 AND bmi < 32 AND smoke = 1 THEN class = 0

IF ap_hi >= 130 AND ap_hi <136 AND cholesterol = 1,2 AND age < 59 AND ap_lo < 90 AND height >= 176 THEN class = 0

IF ap_hi >= 130 AND ap_hi <136 AND cholesterol = 1,2 AND age < 59 AND ap_lo < 90 THEN class = 0

IF ap_hi < 130 AND age >= 55 AND cholesterol = 3 THEN class = 0

IF ap_hi < 130 AND age >= 55 AND cholesterol = 3 AND bmi > = 21 AND active = 1 AND gender = F THEN class = 1

IF ap_hi < 130 AND age >= 55 AND cholesterol = 3 AND bmi > = 21 AND active = 1 AND gender = F AND age < 62 AND bmi < 25 then class = 0

IF ap_hi < 130 AND age >= 55 AND cholesterol = 3 AND bmi >= 21 AND active = 1 AND gender = M AND age >= 63 AND ap_hi >= 118 THEN class = 1

IF ap_hi < 130 AND age >= 55 AND cholesterol = 3 AND bmi >= 21 AND active = 1 AND gender = M AND age >= 63 AND ap_hi < 118 THEN class = 0

IF ap_hi < 130 AND age >= 55 AND cholesterol = 3 AND bmi >= 21 AND active = 1 AND gender = M THEN class = 0

IF ap_hi < 130 AND age >= 55 AND cholesterol = 1, 2 AND age < 61 AND ap_hi >= 119 AND bmi >= 34 THEN class = 1

IF ap_hi < 130 AND age >= 55 AND cholesterol = 1, 2 AND age < 61 AND ap_hi >= 119 AND bmi < 34 AND active = 0 THEN class = 1

IF ap_hi < 130 AND age >= 55 AND cholesterol = 1, 2 AND ap_hi >= 119 AND bmi < 34 AND active = 0 AND age < 58 THEN class = 0

IF ap_hi < 130 AND age >= 55 AND cholesterol = 1, 2 AND ap_hi >= 119 AND bmi < 34 AND active = 1 THEN class = 0

IF ap_hi < 130 AND age >= 55 AND cholesterol = 1, 2 AND ap_hi < 119 THEN class = 0

IF ap_hi < 130 AND age < 55 AND cholesterol = 3 AND gluc = 1, 2 THEN class = 1

IF ap_hi < 130 AND age < 55 AND cholesterol = 3 AND gluc = 3 AND bmi >= 33 THEN class = 1

IF ap_hi < 130 AND age < 55 AND cholesterol = 3 AND gluc = 3 AND bmi < 33 THEN class = 0

IF ap_hi < 130 AND age < 55 AND cholesterol = 1,2 THEN class = 0

Some observations we can make from this tree:

i. High systolic pressure will usually lead to CVD.

ii. If cholesterol level is well above normal it can lead to CVD.

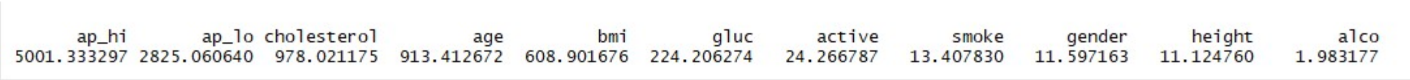
iii. Higher the person's age, more likely they are to develop CVD.

iv. Glucose levels if are well above normal, in combination with cholesterol can lead to CVD.

v. High BMI can lead to CVD.

vi. Not being active in combination with bad cholesterol and age can lead to CVD.

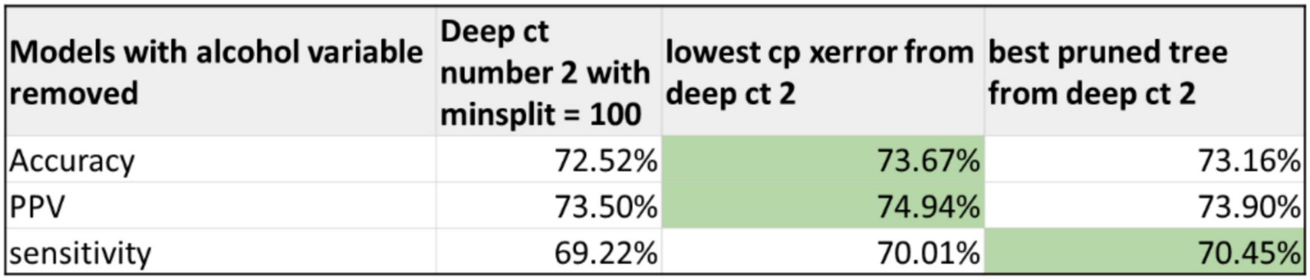
We then checked variable importance from this tree to consider reducing some variables.



```
      ap_hi       ap_lo cholesterol         age         bmi        gluc      active       smoke      gender      height        alco 
5001.333297 2825.060640  978.021175  913.412672  608.901676  224.206274   24.266787   13.407830   11.597163   11.124760    1.983177 
```

With respect to this model, the variable alcohol intake showed the least importance in terms of mean decrease in Gini Index.

We decided to remove this variable and check performance of our classification tree.



Models with alcohol variable removed	Deep ct number 2 with minsplit = 100	lowest cp xerror from deep ct 2	best pruned tree from deep ct 2
Accuracy	72.52%	73.67%	73.16%
PPV	73.50%	74.94%	73.90%
sensitivity	69.22%	70.01%	70.45%

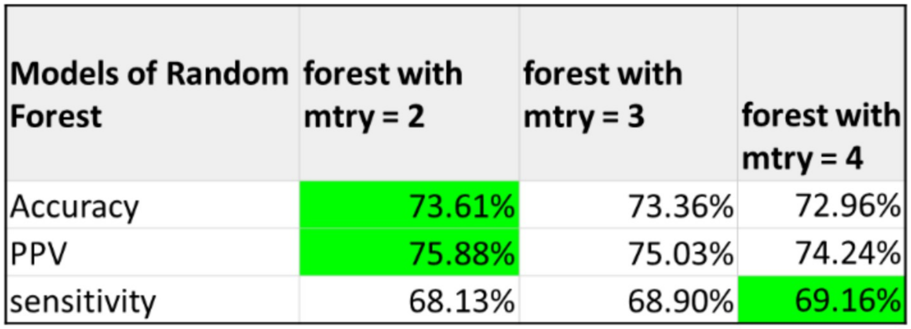
Table 3 – Performance of classification tree

But unfortunately, these results were not better than our previous model hence we decided to not drop the variable alcohol intake.

Random Forest

To accomplish our goal of finding an accurate model we used random forest for modeling. We created models based on the number of splits per decision node (mtry).

Then we checked their performance on the validation set.



Models of Random Forest	forest with mtry = 2	forest with mtry = 3	forest with mtry = 4
Accuracy	73.61%	73.36%	72.96%
PPV	75.88%	75.03%	74.24%
sensitivity	68.13%	68.90%	69.16%

Table 4 – Performance of random forest on validation set

We found our model with mtry = 2 to be the best one. We then do variable importance for feature selection from this random forest model.

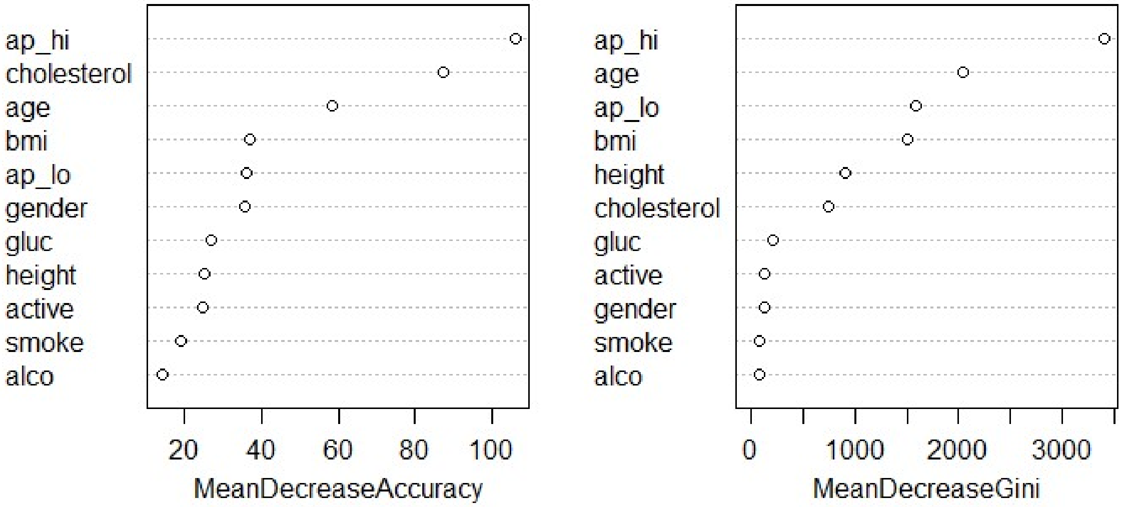
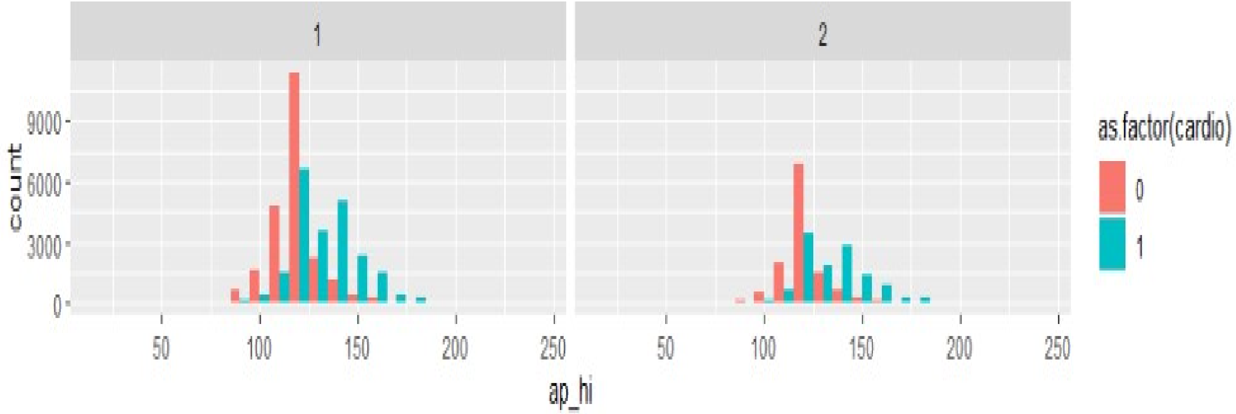


Figure 3 – Variable importance test

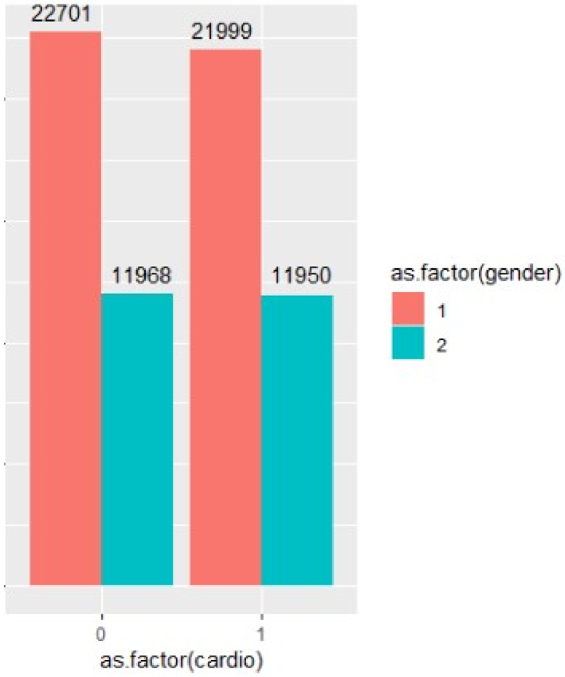
Here as well the variable alcohol had the least importance with respect to mean decrease in Gini Index and Mean decrease in accuracy. We decided to remove this variable and create a random forest model.

RESULTS

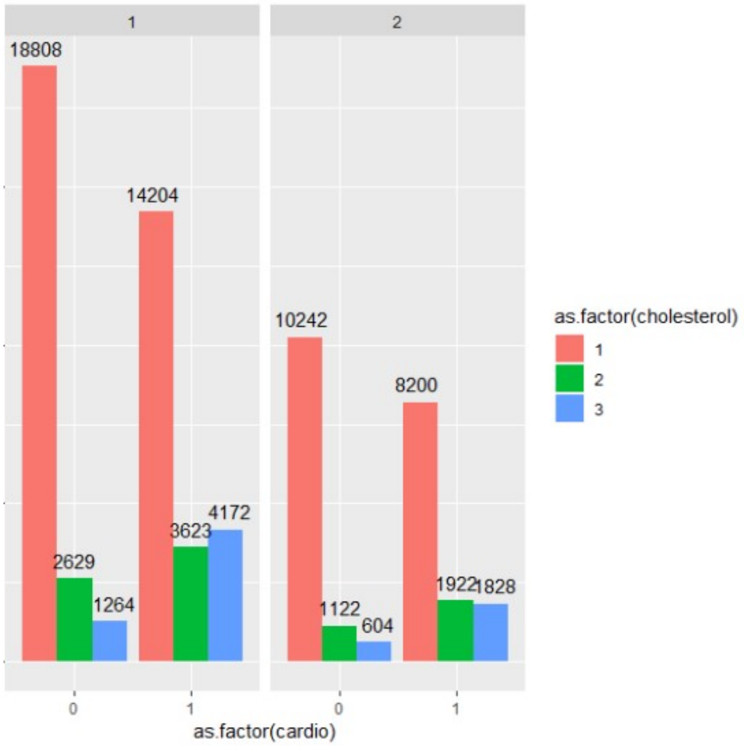
Visualization and preprocessing results



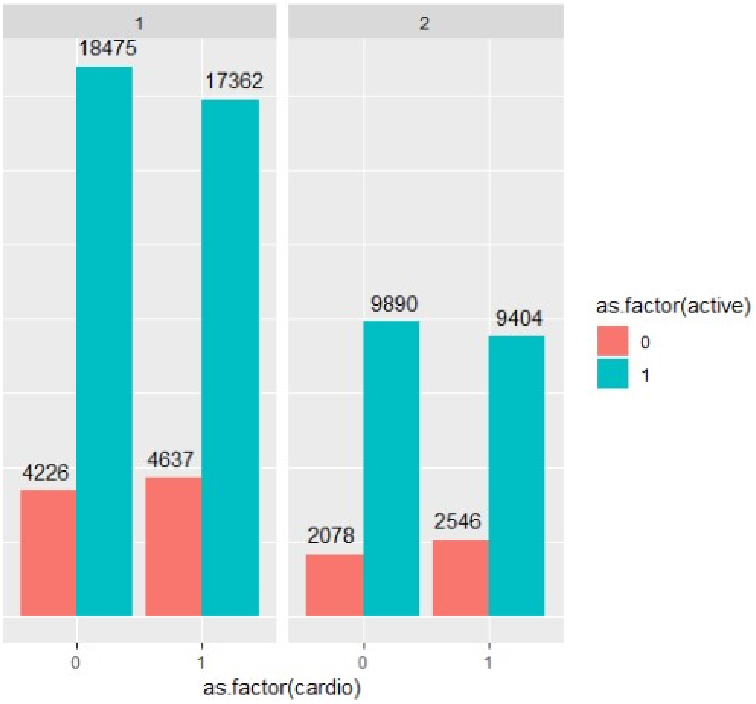
Result 1 – Histogram of high blood pressure against CVD gender wise



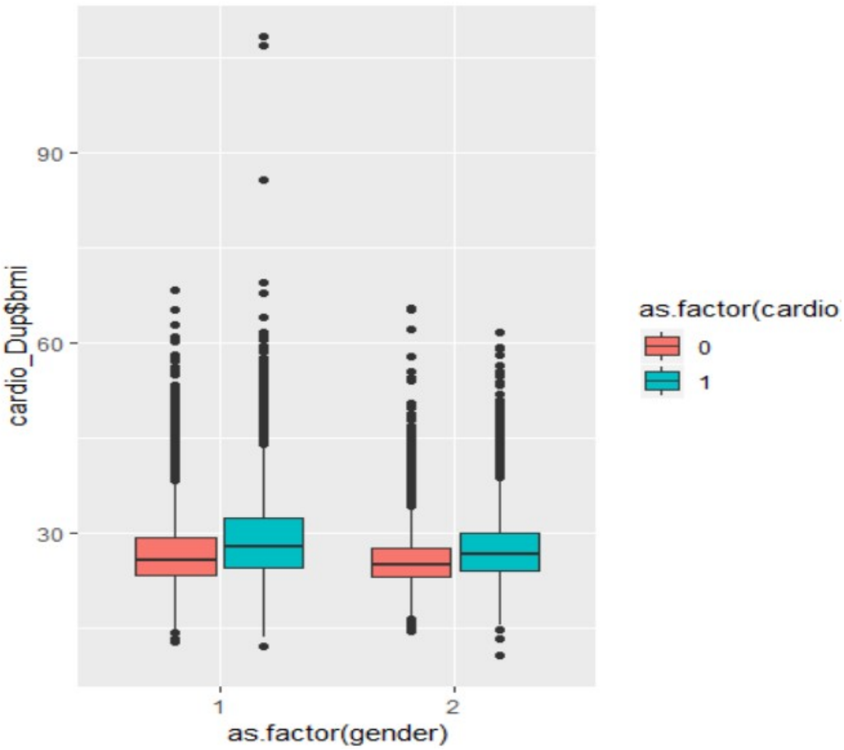
Result 2 – Gender wise frequency of CVD patients



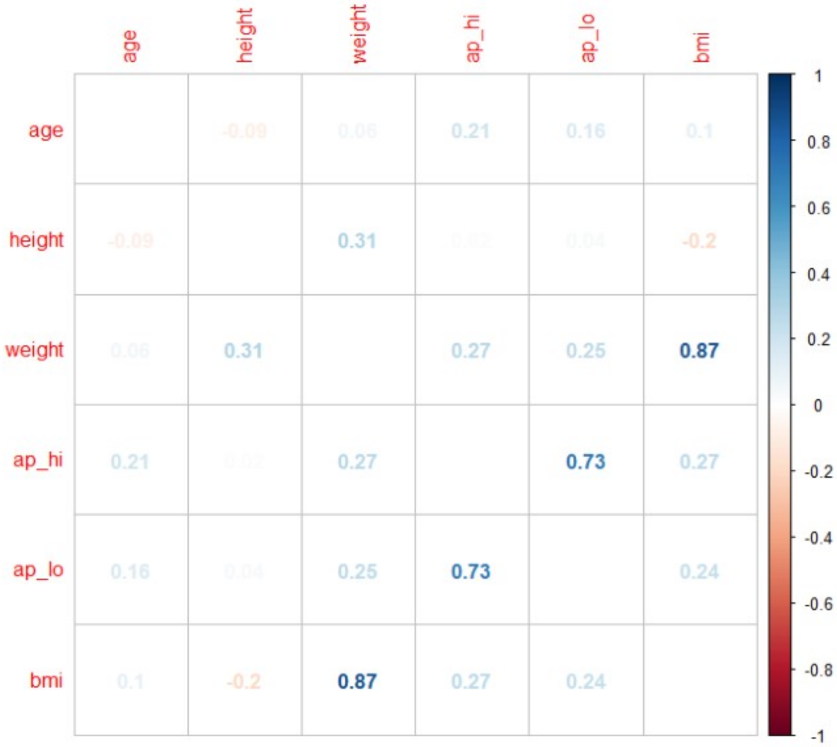
Result 3 – Gender wise frequency of CVD patients against cholesterol



Result 4 – Gender wise frequency of CVD patients against actively exercising

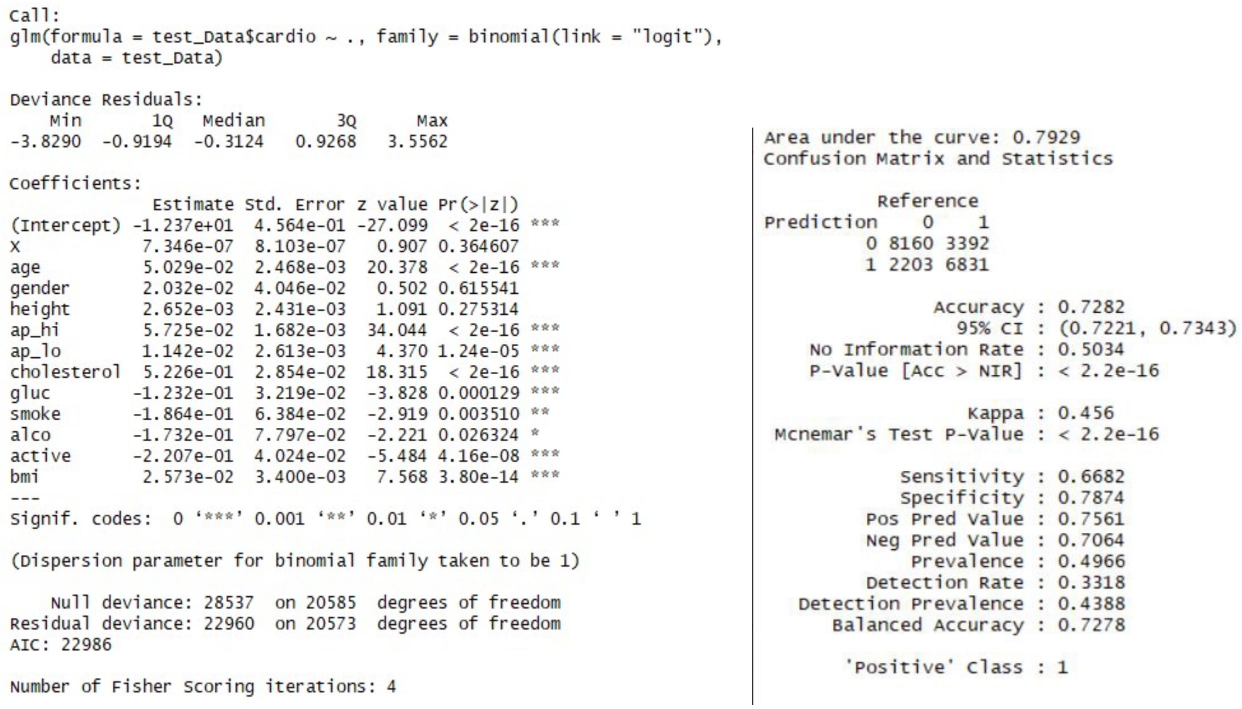


Result 5 – Box plot of BMI against CVD gender wise



Result 6 – Correlation matrix

Logistic Regression results

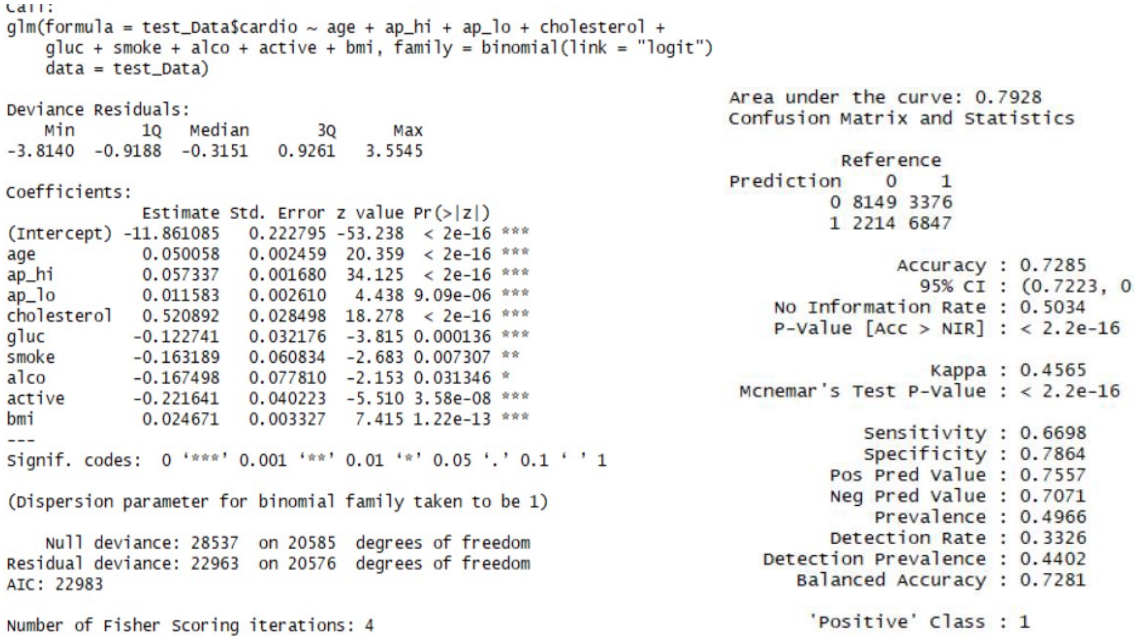


```
Call:
glm(formula = test_Data$cardio ~ ., family = binomial(link = "logit"),
    data = test_Data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.8290  -0.9194  -0.3124   0.9268   3.5562

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.237e+01  4.564e-01 -27.099  < 2e-16 ***
X            7.346e-07  8.103e-07   0.907 0.364607
age          5.029e-02  2.468e-03  20.378  < 2e-16 ***
gender       2.032e-02  4.046e-02   0.502 0.615541
height       2.652e-03  2.431e-03   1.091 0.275314
ap_hi        5.725e-02  1.682e-03  34.044  < 2e-16 ***
ap_lo        1.142e-02  2.613e-03   4.370 1.24e-05 ***
cholesterol  5.226e-01  2.854e-02  18.315  < 2e-16 ***
gluc        -1.232e-01  3.219e-02  -3.828 0.000129 ***
smoke       -1.864e-01  6.384e-02  -2.919 0.003510 **
alco        -1.732e-01  7.797e-02  -2.221 0.026324 *
active      -2.207e-01  4.024e-02  -5.484 4.16e-08 ***
bmi          2.573e-02  3.400e-03   7.568 3.80e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28537  on 20585  degrees of freedom
Residual deviance: 22960  on 20573  degrees of freedom
AIC: 22986

Number of Fisher Scoring iterations: 4
```

```
Area under the curve: 0.7929
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 8160 3392
         1 2203 6831

               Accuracy : 0.7282
                 95% CI : (0.7221, 0.7343)
    No Information Rate : 0.5034
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.456
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.6682
            Specificity : 0.7874
         Pos Pred Value : 0.7561
         Neg Pred Value : 0.7064
             Prevalence : 0.4966
         Detection Rate : 0.3318
   Detection Prevalence : 0.4388
      Balanced Accuracy : 0.7278

       'Positive' Class : 1
```

Result 7 – Model 1 results

From the p-values, gender and height are not at all significant and alcohol consumption is less significant in determining CVDs. This model predicted results with an accuracy of 72.82%. Second model excluded least significant variables and the results are as followed.

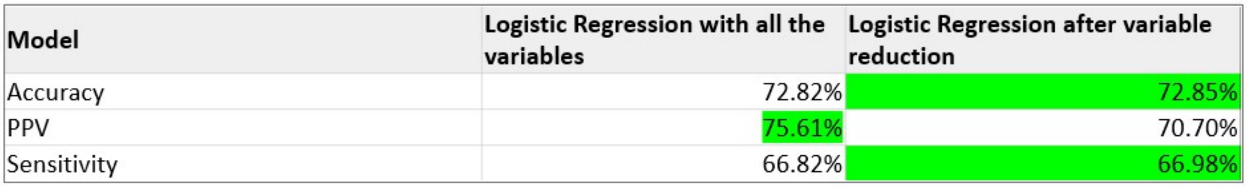


```
Call:
glm(formula = test_Data$cardio ~ age + ap_hi + ap_lo + cholesterol +
    gluc + smoke + alco + active + bmi, family = binomial(link = "logit")
    data = test_Data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.8140  -0.9188  -0.3151   0.9261   3.5545

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -11.861085   0.222795 -53.238  < 2e-16 ***
age           0.050058   0.002459  20.359  < 2e-16 ***
ap_hi         0.057337   0.001680  34.125  < 2e-16 ***
ap_lo         0.011583   0.002610   4.438 9.09e-06 ***
cholesterol   0.520892   0.028498  18.278  < 2e-16 ***
gluc         -0.122741   0.032176  -3.815 0.000136 ***
smoke        -0.163189   0.060834  -2.683 0.007307 **
alco         -0.167498   0.077810  -2.153 0.031346 *
active       -0.221641   0.040223  -5.510 3.58e-08 ***
bmi           0.024671   0.003327   7.415 1.22e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28537  on 20585  degrees of freedom
Residual deviance: 22963  on 20576  degrees of freedom
AIC: 22983

Number of Fisher Scoring iterations: 4
```

```
Area under the curve: 0.7928
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 8149 3376
         1 2214 6847

               Accuracy : 0.7285
                 95% CI : (0.7223, 0.7345)
    No Information Rate : 0.5034
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4565
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.6698
            Specificity : 0.7864
         Pos Pred Value : 0.7557
         Neg Pred Value : 0.7071
             Prevalence : 0.4966
         Detection Rate : 0.3326
   Detection Prevalence : 0.4402
      Balanced Accuracy : 0.7281

       'Positive' Class : 1
```

Result 8 – Model 2 results

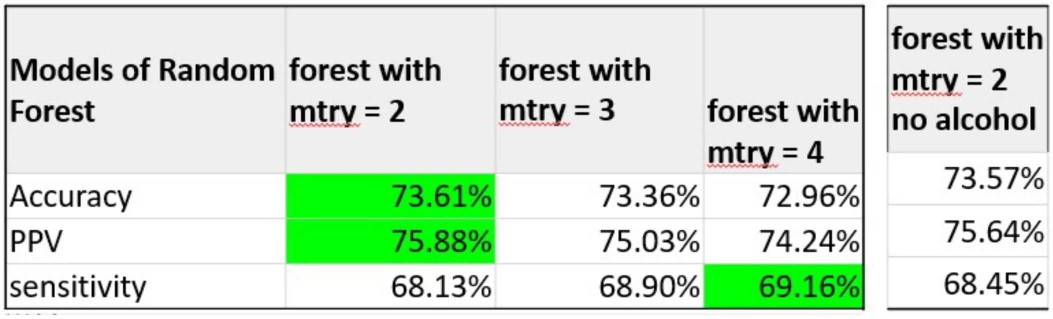
The accuracy did not significantly increased. To deal with this, we ran stepAIC and looked for more important variables but the results of stepAIC suggested that all possible best predictors are already existing in the model and nothing has to be removed. To conclude logistic regression, we managed to get an accuracy of 72.85%.



Model	Logistic Regression with all the variables	Logistic Regression after variable reduction
Accuracy	72.82%	72.85%
PPV	75.61%	70.70%
Sensitivity	66.82%	66.98%

Table 5 – Comparison of both the models resulted from logistic regression

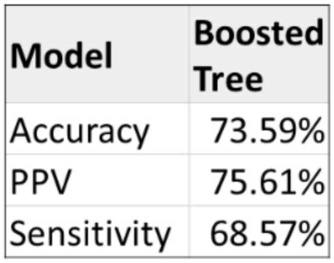
Random Forest results



Models of Random Forest	forest with mtry = 2	forest with mtry = 3	forest with mtry = 4	forest with mtry = 2 no alcohol
Accuracy	73.61%	73.36%	72.96%	73.57%
PPV	75.88%	75.03%	74.24%	75.64%
sensitivity	68.13%	68.90%	69.16%	68.45%

Table 6 – Performance of random forest after variable reduction

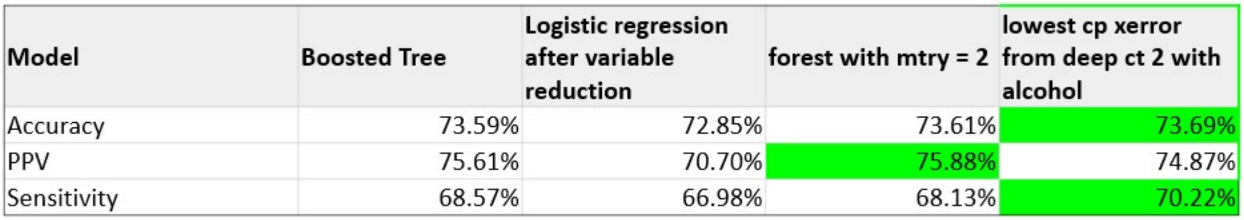
This did not increase performance of our model, hence we decided to keep the variable alcohol for random forest. We used boosted trees to reduce any error from our random forest model and checked for performance on our validation set.



Model	Boosted Tree
Accuracy	73.59%
PPV	75.61%
Sensitivity	68.57%

Table 7 – Boosted trees results

CONCLUSION



Model	Boosted Tree	Logistic regression after variable reduction	forest with mtry = 2	lowest cp xerror from deep ct 2 with alcohol
Accuracy	73.59%	72.85%	73.61%	73.69%
PPV	75.61%	70.70%	75.88%	74.87%
Sensitivity	68.57%	66.98%	68.13%	70.22%

Table 8 – Final result by comparing different models

Removing any variable from classification tree and random forest caused it to reduce in accuracy but removing variable height and gender in logistic regression improved the performance.

We also were able to generate rules from our dataset using the best performing classification tree. The most important variables are high blood pressure, low blood pressure, age, BMI, and cholesterol.