A consulting report for prediction of internal thoracic artery and the radial artery based on the risk factors.

Author: Shiming ZHENG
Id: 1149897

Date: 13/11/2022

Contents

Background

Currently, the internal thoracic artery(ITA) and radial artery(RA) are the two coronary artery bypass vessels that can be used to supply blood to the heart during coronary artery bypass surgery(CABG). The potential for pathological changes in ITA and RA is one of the current concerns. There are two types of ITA intimal abnormalities: intimal thickening and atherosclerosis. In particular, as for RA, we only consider whether it has medial calcification or not. For a more detailed background, the client can refer to these two previous studies: [1] and [2].

Problem to be addressed

To investigate the associations and the possibility to make a useful prediction of ITA intimal abnormality and RA medial calcification with the 6 risk factors of the patients undergoing coronary artery bypass graft surgery, including age, gender, presence of diabetes mellitus, history of cigarette smoking, presence of peripheral vascular disease (PVD), and presence of cerebrovascular disease (CVD).

Study Design & Methods

This study was implemented by R version 4.2.1. First, the Chi-square test was selected to determine whether there is a significant difference between the ITA intimal abnormalities and RA medial calcification, after combining the situation of intimal thickening and atherosclerosis of ITA into ITA abnormality. They are significantly different if the P value <0.05.

The correlation between the 5 categorical risk factors, including gender, presence of diabetes mellitus, history of cigarette smoking, presence of PVD, and presence of CVD, of the ITA intimal abnormalities and RA medial calcification, were tested by the Cramer’s V. The closer Cramer's V is to 0, the smaller the association between the two factors, while the closer it is to 1, the greater the association between the two factors.

Two outcomes: ITA intimal abnormalities(yes/no) and RA medial calcification(yes/no) were combined into a 4-label response variable in the statistical modelling analysis. Also, the 6 risk factors (age, gender, diabetes, ever smoked, PVD and CVD) are regarded as the explanatory variables. Multinomial logistic regression and random forest(RF) were selected to explore their associations and the possibility to make a useful prediction of the label of the response variable based on the 6 risk factors. The usefulness of the Multinomial logistic regression model and RF were measured by the accuracy of prediction and the proportion of variation in the prediction explained by the explanatory variables in the fitting model.

Data description

The data used in this report is brought by the client. This dataset contains the morphometric data on the ITA and the RA of 110 patients undergoing coronary artery bypass graft surgery. It has 8 attributes and 110 instances. As shown in Tab.1, the definition of the variables is described.

Variables	Definition
Age	The age in years of the patient
Gender	The sex of the patient
Diabetes	Whether the patient has ever had diabetes
Ever smoked	Whether the patient has ever smoked
PVD	Whether the patient has ever had Peripheral vascular disease (PVD)
CVD	Whether the patient has ever had Cerebrovascular disease (CVD)
RA medial calcification	Whether the RA of the patient has medial calcification
ITA intimal abnormality	Whether the ITA of the patient has intimal thickening, atherosclerosis, or not

Tab.1 The definition of the variables.

Exploratory data analysis

First, for the variable ITA intimal abnormality, we only consider whether it has an exception, however, there are 3 categories in the original data (0 = normal, 1 = intimal thickening, 2 = atherosclerosis). Therefore, we combine the situation of intimal thickening and atherosclerosis into abnormality and use serial number 1 to represent it. However, after checking the dataset, there are no subjects with ITA atherosclerosis, so we don't need to make changes to the dataset. According to the result of the chi-square test, *P*-value > 0.05 indicates that RA medial calcification and ITA intimal abnormality are significantly independent. Also, as Tab.2 shows, the values of Cramer's V indicate that the correlation between the 5 categorical predictors is also relatively low.

	Gender	Diabetes	Ever smoked	PVD	CVD
Gender	1	<0.01	0.24	<0.01	0.08
Diabetes	<0.01	1	0.23	0.07	0.05
Ever smoked	0.24	0.23	1	0.07	0.11
PVD	<0.01	0.07	0.07	1	0.32
CVD	0.08	0.05	0.11	0.32	1

Tab.2 The Cramer's V between the 5 categorical predictors.

In addition, as the Tab.3 shown, the serial numbers 0, 1, 2, and 3 were to represent the four groups of the two outcomes: RA medial calcification(yes/no) and ITA intimal abnormalities(yes/no). Since there are more than 2 levels, we used 0 as the reference level (i.e. the subject who is normal).

	RA medial calcification	ITA intimal abnormalities
0	No	Yes
1	Yes	No
2	No	Yes
3	Yes	Yes

Tab.3 The four groups of the two outcomes: ITA intimal abnormalities and RA medial calcification.

In addition, as the Tab.3 shown, serial numbers 0, 1, 2, and 3 were to represent the four groups of the two outcomes: RA medial calcification(yes/no) and ITA intimal abnormalities(yes/no). Since there are more than 2 levels, we used 0 as the reference level (i.e. the subject who is normal).

Variables	Proportion of each class
Age	range: 42–81; mean±SD: 66±9
Gender	89.1% male, 10.9% female
Diabetes	75.5% no, 24.5% yes
Ever smoked	33.6% never, 66.4% ever
PVD	82.7% no, 17.3% yes
CVD	90.0% no, 10.0% yes
RA + ITA	26.4% normal, 3.6% yes and no, 60.9% no and yes, 9.1% both yes

Tab.4 The descriptive statistics of all the variables

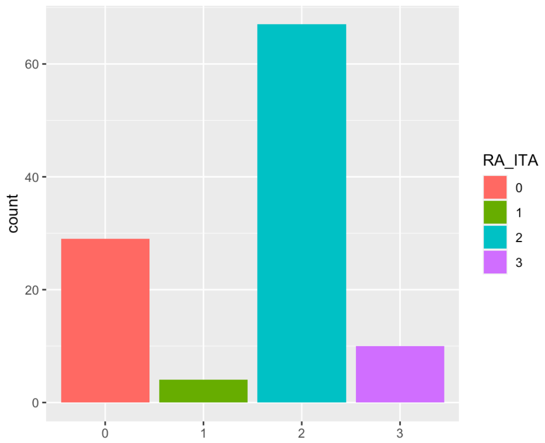


Fig.1 The bar chart of the count of each label in the independent variable.

Prediction of ITA intimal abnormality and RA medial calcification

This prediction is a multi-class classification because there are 4 outcomes (more than 2 categories) of the response variable. Therefore, Multinomial logistic regression, an extension of binary logistic regression, and random forest(RF) were chosen to make the prediction.

MODEL1: MULTINOMIAL LOGISTIC REGRESSION

First, we regarded the variable: age as the continuous variable to fit the stepwise multinomial logistic regression. The results are shown in Tab.5. In this case, age, diabetes, and history of cigarette smoking are the strongest predictors when making the prediction. To determine the proportion of variation in the prediction explained by the explanatory variables in the fitting model, we calculated 3 kinds of Pseudo R-Square: Cox and Snell's R-Square (upper bound<1.0), Nagelkerke's R-Square (modification of Cox and Snell's R-Square; upper bound<1.0), and McFadden's R-Square(upper bound=1.0). According to Cox and Snell's R-Square, the results concluded that there is 22.3% relationship between the predictors and the response variable. Nagelkerke's R-Square indicates that 26.1% of the variation in the predictors is explained by this model. McFadden's R-Square shows that the relationship of 13.1% between the predictors and the response variable. In addition, for this case, the training accuracy is around 68.18% and the testing accuracy is around 40.91%.

In addition, we tried to divide different ages into 4 different age groups (i.e. every 10 years old is a group) as a categorical variable and fit the stepwise multinomial logistic regression again. The results are shown in Tab.6. In this case, only the variable history of cigarette smoking is the strongest predictor when making the prediction. According to Cox and Snell's R-Square, Nagelkerke's R-Square, and McFadden's R-Square, the results concluded that there is 8.6%, 10.0%, and 4.7% relationship between the predictors and the response variable, respectively. For this case, the training accuracy is around 64.77% and the testing accuracy is around 45.45%.

Variable	β Coefficient
RA medial calcification(yes) and ITA intimal abnormalities(no)	
Constant = 0.1672	
Age	1.592
Diabetes[1]	22.262
Ever.smoked[1]	0.175
RA medial calcification(no) and ITA intimal abnormalities(yes)	
Constant = 1.3231	
Age	1.515
Diabetes[1]	2.750
Ever.smoked[1]	2.900
RA medial calcification(yes) and ITA intimal abnormalities(yes)	
Constant = 0.0285	
Age	6.267
Diabetes[1]	9.771
Ever.smoked[2]	7.080
[1] 0 = no, 1 = yes [2] 0 = never, 1 = ever	

Tab.5 Results of stepwise multinomial logistic regression when regarding the variable: Age as the continuous variable.

Variable	*β* Coefficient
RA medial calcification(yes) and ITA intimal abnormalities(no)	
Constant = 0.3	
Ever.smoked[1]	0.333
RA medial calcification(no) and ITA intimal abnormalities(yes)	
Constant = 1.5	
Ever.smoked[1]	2.8
RA medial calcification(yes) and ITA intimal abnormalities(yes)	
Constant = 0.1	
Ever.smoked[2]	6.0
[1] 0 = no, 1 = yes [2] 0 = never, 1 = ever	

Tab.6 Results of Stepwise Multinomial logistic regression when regarding the variable: Age as the categorical variable.

MODEL2: RANDOM FOREST

Besides, a random forest classifier was chosen to make this prediction. The testing accuracy is 77.3% and 54.5% for the RF model(*mtry* = 2, *ntree* = 500) when regarding the variable: age as the continuous variable, and the RF model (*mtry* = 1, *ntree* = 450) when regarding the variable: age as the categorical variable with 4 groups, respectively. As shown in the Fig.2, the plot revealed the variables PVD and history of cigarette smoking are the most important.

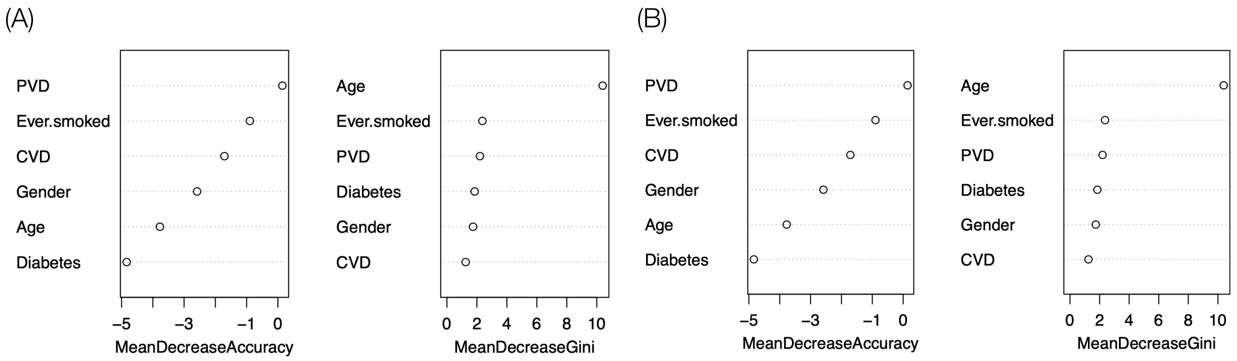


Fig.2 The Importance of variables in the RF model when regarding the variable: age as the continuous variable (A), and the RF mode when regarding the variable: age as the categorical variable (B), respectively.

Model	Testing Accuracy
Multinomial logistic regression model1	40.91%
Multinomial logistic regression model2	45.45%
RF model1	77.3%
RF model2	54.5%

Tab.7 The testing accuracy of all the models.

Conclusions

Our findings investigate that the history of cigarette smoking is the strongest predictor for all the models, indicating that whether the patient has ever smoked is likely to affect the outcome of RA medial calcification and ITA intimal abnormalities. In addition, the variables: age and PVD are significantly correlated with the level of outcome in the multinomial logistic regression model and RF model, respectively, when regarding age as a continuous variable. The predictive model with the highest testing accuracy is the RF model when regarding age as a continuous variable, reaching 77.3%. However, it is also relatively not enough for a predictive model. Therefore, whether it can be effectively predicted needs further discussion.

References

[1] Buxton, B. and Lek, P. (1999) *Study of graft arteries*, *RealStat*. Available at: https://realstat.science.unimelb.edu.au/study-of-graft-arteries/background/ (Accessed: November 11, 2022).

[2] Ruengsakulrach, P., Sinclair, R., Komeda, M., Raman, J., Gordon, I., & Buxton, B. (1999) Comparative histopathology of radial artery versus internal thoracic artery and risk factors for development of intimal hyperplasia and atherosclerosis. *Circulation, 100(19)*, II-139 – II-144.

Appendix

In the next page.

Appendix

R code

```
#package installation
#install.packages('dplyr')
#install.packages('jmv')
#install.packages('rcompanion')
#install.packages("DescTools")
```

```
#load and glimpse the dataset
artery_data<-read.table('/Users/jasmyn/Desktop/artery-1.csv',header=T,sep=",")
library(dplyr)
```

```
##
##    'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
glimpse(artery_data)
```

```
## Rows: 110
## Columns: 28
## $ Age                          <int> 74, 64, 44, 74, 68, 66, 48, 71, 72, 65, 7~
## $ Gender                       <int> 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ Diabetes                     <int> 0, 0, 1, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0,~
## $ Ever.smoked                  <int> 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 0, 1, 0, 1,~
## $ PVD                          <int> 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0,~
## $ CVD                          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ Hypercholesterolemia         <int> 1, 1, 1, 0, 1, 0, 1, 0, 0, 0, 1, 0, 1, 0,~
## $ RA.intimal.abnormality       <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ RA.medial.calcification      <int> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,~
## $ ITA.intimal.abnormality      <int> 0, 0, 1, 1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 1,~
## $ RA...luminal.narrowing       <dbl> 22.0825, 6.7462, 14.0704, 18.8055, 19.633~
## $ RA..Intimal.thickness.index  <dbl> 0.20720, 0.08769, 0.21533, 0.30040, 0.164~
## $ RA.Intima.to.media.ratio     <dbl> 0.7950, 0.1872, 0.3626, 0.5359, 0.2959, 0~
## $ ITA...luminal.narrowing      <dbl> 8.2780, 7.4851, 9.5488, 8.8865, 8.9493, 1~
## $ ITA..Intimal.thickness.index <dbl> 0.063398, 0.107537, 0.174526, 0.127243, 0~
## $ ITA.Intima.to.media.ratio    <dbl> 0.11139, 0.03331, 0.29730, 0.40423, 0.409~
## $ RA.DLI                       <dbl> 2.02739, 1.83587, 2.39934, 2.68417, 1.591~
## $ RA.IEL.area                  <dbl> 3.229539, 2.648179, 4.523225, 5.660906, 1~
## $ RA.Intimal.area              <dbl> 0.713163, 0.178651, 0.636436, 1.064560, 0~
## $ RA.Medial.area               <dbl> 3.441862, 2.037250, 2.955581, 3.543850, 2~
## $ RA.Intimal.width             <dbl> 0.31505, 0.05405, 0.12281, 0.28172, 0.113~
## $ RA.Medial.width              <dbl> 0.396277, 0.288752, 0.338693, 0.525706, 0~
## $ ITA.DLI                      <dbl> 1.13633, 1.64756, 1.98807, 2.11451, 1.529~
## $ ITA.IEL.area                 <dbl> 1.014551, 2.132777, 3.105489, 3.513047, 1~
## $ ITA.Intimal.area             <dbl> 0.083985, 0.159641, 0.296538, 0.312185, 0~
## $ ITA.Medial.area              <dbl> 1.324714, 1.484521, 1.699110, 2.453464, 1~
## $ ITA.Intimal.width            <dbl> 0.039363, 0.013333, 0.106796, 0.108547, 0~
## $ ITA.Medial.width             <dbl> 0.353395, 0.400233, 0.359222, 0.268529, 0~
```

Because we only need to predict radial artery (RA) medial calcification and internal thoracic artery (ITA) intimal abnormality bases on the age, gender, diabetes, ever smoked, presence of Peripheral vascular disease (PVD) and presence of Cerebrovascular disease (CVD) of the patients, we extract these variables from the original dataset as a new dataset.

```
#extract the features
artery_data2 <- artery_data[,c(1:6,9,10)]
glimpse(artery_data2)
```

```
## Rows: 110
## Columns: 8
## $ Age                     <int> 74, 64, 44, 74, 68, 66, 48, 71, 72, 65, 73, 72~
## $ Gender                  <int> 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ Diabetes                <int> 0, 0, 1, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1~
## $ Ever.smoked             <int> 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 0, 1, 0, 1, 1, 0~
## $ PVD                     <int> 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0~
## $ CVD                     <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ RA.medial.calcification <int> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0~
## $ ITA.intimal.abnormality <int> 0, 0, 1, 1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 1~
```

In particular, for independent variable ITA intimal abnormality, we only consider whether it has an exception, However, there are 3 categories in the original data (0 = normal, 1 = intimal thickening, 2 = atherosclerosis). Therefore, we combine the situation of intimal thickening and atherosclerosis into abnormality and use serial number 1 to represent.

```
# Transformation:
# ITA intimal abnormality (0 = normal, 1 = intimal thickening, 2 = atherosclerosis)
# --> ITA intimal abnormality (0 = normal, 1 = abnormal)
which(artery_data2$ITA.intimal.abnormality==2)
```

```
## integer(0)
```

However, after checking the dataset, there are no subjects with ITA atherosclerosis, so we don't need to make changes to the dataset.

Then,we use chi-square test to test the whether the two outcomes are independent with each other.

```
#chi-square test in two outcomes
table(artery_data2$RA.medial.calcification, artery_data2$ITA.intimal.abnormality)
```

```
##
##      0  1
##   0 29 67
##   1  4 10
```

```
chisq.test(artery_data2$RA.medial.calcification, artery_data2$ITA.intimal.abnormality, correct=FALSE)
```

```
## Warning in chisq.test(artery_data2$RA.medial.calcification,
## artery_data2$ITA.intimal.abnormality, : Chi-squared
```

```
##
```

```
##  Pearson's Chi-squared test
##
## data:  artery_data2$RA.medial.calcification and artery_data2$ITA.intimal.abnormality
## X-squared = 0.01559, df = 1, p-value = 0.9006
```

Because p-value > 0.05 significance level, indicating that RA.medial.calcification and ITA.intimal.abnormality are significantly independent.

According to the result, Cramer's V between two dependent variables is 0.01, indicating that there is just a negligible correlation between them.

```
#test the correlation btw. independent variables
library(rcompanion)
cramerV(artery_data2$Gender, artery_data2$Diabetes,bias.correct = FALSE)
```

```
## Cramer V
## 0.003696
```

```
cramerV(artery_data2$Gender, artery_data2$Ever.smoked,bias.correct = FALSE)
```

```
## Cramer V
##   0.2446
```

```
cramerV(artery_data2$Gender, artery_data2$PVD,bias.correct = FALSE)
```

```
## Cramer V
##  0.00561
```

```
cramerV(artery_data2$Gender, artery_data2$CVD,bias.correct = FALSE)
```

```
## Cramer V
##  0.07776
```

```
cramerV(artery_data2$Diabetes, artery_data2$Ever.smoked,bias.correct = FALSE)
```

```
## Cramer V
##   0.2272
```

```
cramerV(artery_data2$Diabetes, artery_data2$PVD,bias.correct = FALSE)
```

```
## Cramer V
##  0.07468
```

```
cramerV(artery_data2$Diabetes, artery_data2$CVD,bias.correct = FALSE)
```

```
## Cramer V
##  0.04929
```

```
cramerV(artery_data2$Ever.smoked, artery_data2$PVD,bias.correct = FALSE)
```

```
## Cramer V
##   0.0708
```

```
cramerV(artery_data2$Ever.smoked, artery_data2$CVD,bias.correct = FALSE)
```

```
## Cramer V
##    0.109
```

```
cramerV(artery_data2$PVD, artery_data2$CVD,bias.correct = FALSE)
```

```
## Cramer V
##   0.3287
```

Also, the correlation of the independent variables between each other also relatively low.

Then, we use four groups with the serial number 0, 1, 2, and 3 to represent four groups of the two outcomes (0 = RA medial calcification: no & ITA intimal abnormality: no, 1 = RA medial calcification: yes & ITA intimal abnormality: no, 2 = RA medial calcification: no & ITA intimal abnormality: yes, 3 = RA medial calcification: yes & ITA intimal abnormality: yes)

```
# add a new column to represent the four groups of the two outcomes
artery_data3 <- artery_data2 %>%
  mutate(RA_ITA = ifelse(RA.medial.calcification == 0 &
                           ITA.intimal.abnormality == 0, 0,
                         ifelse(RA.medial.calcification == 1 &
                                  ITA.intimal.abnormality == 0, 1,
                                ifelse(RA.medial.calcification == 0 &
                                         ITA.intimal.abnormality == 1, 2, 3))))

#test multicollinearity (acceptable if vif < 2)
library(car)
```

```
##      carData
```

```
##
##    'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
m <- lm(RA_ITA ~., data = artery_data3)
vif(m)
```

```
## Warning in summary.lm(object, ...): essentially perfect fit: summary may be
## unreliable
```

```
##                     Age                  Gender                Diabetes
##                1.181270                1.124713                1.128153
##             Ever.smoked                     PVD                     CVD
##                1.188842                1.199235                1.191038
## RA.medial.calcification ITA.intimal.abnormality
##                1.080789                1.134882
```

```
artery_data3$RA_ITA <- as.factor(artery_data3$RA_ITA)
```

```
#check missing value
sum(is.na(artery_data3))
```

```
## [1] 0
```

```
artery_data4 <- artery_data3[,c(1:6,9)]
str(artery_data4)
```

```
## 'data.frame':    110 obs. of  7 variables:
##  $ Age        : int  74 64 44 74 68 66 48 71 72 65 ...
##  $ Gender     : int  1 1 1 1 1 0 1 1 1 1 ...
##  $ Diabetes   : int  0 0 1 0 0 1 1 1 0 0 ...
##  $ Ever.smoked: int  1 0 1 0 1 1 1 0 1 1 ...
##  $ PVD        : int  0 0 0 0 1 0 0 1 0 0 ...
##  $ CVD        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ RA_ITA     : Factor w/ 4 levels "0","1","2","3": 1 1 3 3 3 3 1 2 3 1 ...
```

```
# Scale the continous variables
artery_data4$Age <- scale(artery_data$Age)

# Formatted the categorical variables
artery_data4$Gender = factor(artery_data4$Gender)
artery_data4$Diabetes = factor(artery_data4$Diabetes)
artery_data4$Ever.smoked = factor(artery_data4$Ever.smoked)
artery_data4$PVD = factor(artery_data4$PVD)
artery_data4$CVD = factor(artery_data4$CVD)
str(artery_data4)
```

```
## 'data.frame':    110 obs. of  7 variables:
##  $ Age        : num [1:110, 1] 0.887 -0.191 -2.347 0.887 0.24 ...
##   ..- attr(*, "scaled:center")= num 65.8
##   ..- attr(*, "scaled:scale")= num 9.28
##  $ Gender     : Factor w/ 2 levels "0","1": 2 2 2 2 2 1 2 2 2 2 ...
##  $ Diabetes   : Factor w/ 2 levels "0","1": 1 1 2 1 1 2 2 2 1 1 ...
##  $ Ever.smoked: Factor w/ 2 levels "0","1": 2 1 2 1 2 2 2 1 2 2 ...
##  $ PVD        : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 2 1 1 ...
##  $ CVD        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ RA_ITA     : Factor w/ 4 levels "0","1","2","3": 1 1 3 3 3 3 1 2 3 1 ...
```

In addition, we used 0 as the reference level (i.e. the subject who is normal), because there are more than 2 levels.

```
artery_data4$RA_ITA_ref <- relevel(artery_data4$RA_ITA, ref = "0")
# Give the names to each level
levels(artery_data4$RA_ITA_ref) <- c("normal","RAyes_ITAno","RAno_ITAyes", "RAyes_ITAyes")
```

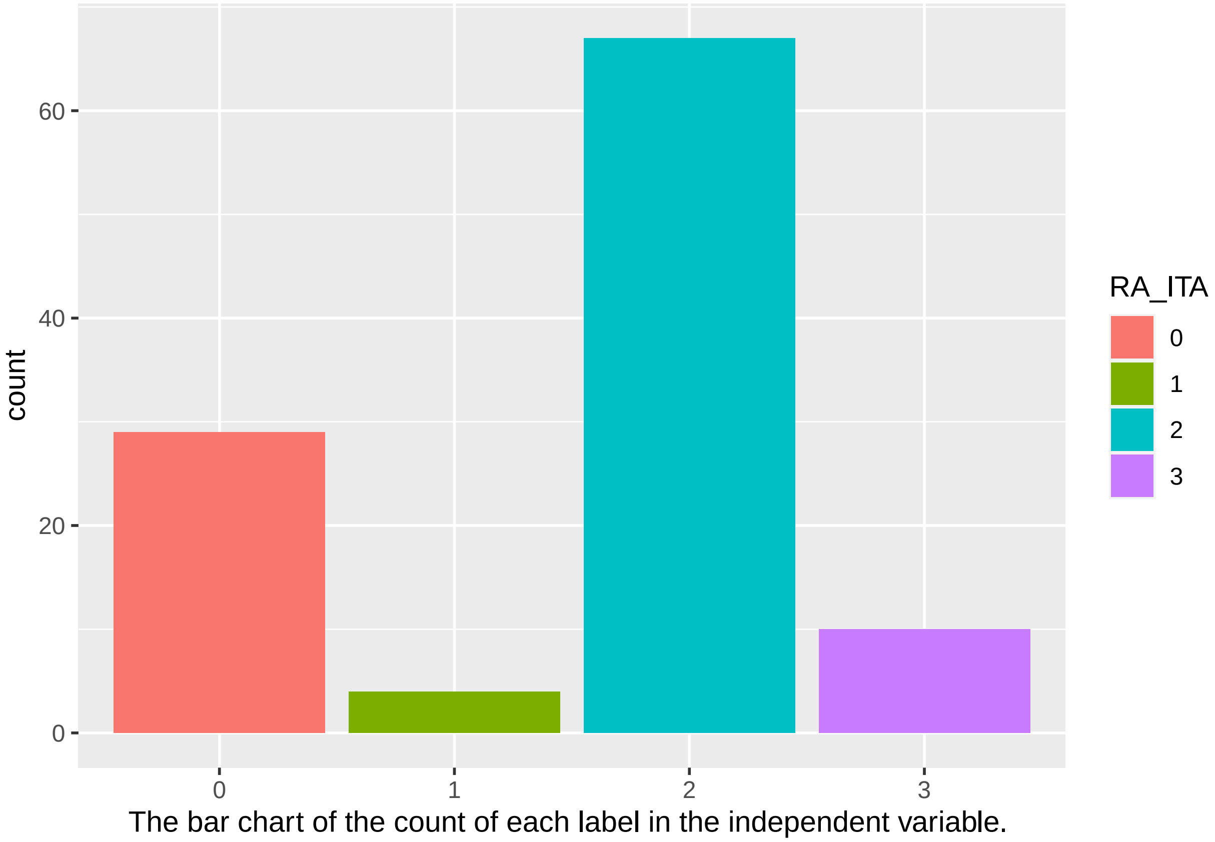
Descriptive statistics of all the variables.

```
#Descriptive statistics
library(jmv)
descriptives(artery_data4, vars = vars(Age, Gender, Diabetes,
                                       Ever.smoked, PVD, CVD, RA_ITA_ref), freq = TRUE)
```

```
##
##  DESCRIPTIVES
##
##  Descriptives
##
##                          Age            Gender    Diabetes    Ever.smoked    PVD    CVD    RA_ITA_r
##
##    N                              110       110         110            110    110    110          11
##    Missing                          0         0           0              0      0      0
##    Mean                  6.938894e-16
##    Median                   0.2939433
##    Standard deviation        1.000000
##    Minimum                  -2.562206
##    Maximum                   1.641184
##
##
##
##  FREQUENCIES
##
##  Frequencies of Gender
```

```
##
##   Levels    Counts    % of Total    Cumulative %
##
##   0             12      10.90909        10.90909
##   1             98      89.09091       100.00000
##
##
##
##  Frequencies of Diabetes
##
##   Levels    Counts    % of Total    Cumulative %
##
##   0             83      75.45455        75.45455
##   1             27      24.54545       100.00000
##
##
##
##  Frequencies of Ever.smoked
##
##   Levels    Counts    % of Total    Cumulative %
##
##   0             37      33.63636        33.63636
##   1             73      66.36364       100.00000
##
##
##
##  Frequencies of PVD
##
##   Levels    Counts    % of Total    Cumulative %
##
##   0             91      82.72727        82.72727
##   1             19      17.27273       100.00000
##
##
##
##  Frequencies of CVD
##
##   Levels    Counts    % of Total    Cumulative %
##
##   0             99      90.00000        90.00000
##   1             11      10.00000       100.00000
##
##
##
##  Frequencies of RA_ITA_ref
##
##   Levels          Counts    % of Total    Cumulative %
##
##   normal              29      26.36364        26.36364
##   RAyes_ITAno          4       3.63636        30.00000
##   RAno_ITAyes         67      60.90909        90.90909
##   RAyes_ITAyes        10       9.09091       100.00000
##
```

```r
# bar chart of dependent variable
library(ggplot2)
p <- ggplot(artery_data4, aes(x = factor(RA_ITA), fill = RA_ITA)) + geom_bar()
p + labs(x = "The bar chart of the count of each label in the independent variable. ")
```



In order to find the relationship between the dependent variables and explanatory variables and achieve prediction. According to the Pareto principle, we treated 80% data as the training data, and 20% data as the testing data.

```r
#training / testing dataset
set.seed(123)
s = sort(sample(nrow(artery_data4), nrow(artery_data4)*0.8))
train_data<-artery_data4[s,]
test_data<-artery_data4[-s,]
```

Method 1: Multinomial logistic regression

Also, since there are 4 kinds of classes in the dependent variable (i.e. a multiclass classification), first, a kind of generalization of logistic regression: multinomial logistic regression was selected to do this classification.

```r
library(nnet)
mult_logis_model<-multinom(RA_ITA_ref~Age+Gender+Diabetes+Ever.smoked+PVD+CVD,data = train_data)
```

```
## # weights:  32 (21 variable)
## initial  value 121.993904
## iter  10 value 69.986298
## iter  20 value 68.272077
## iter  30 value 68.211551
## iter  40 value 68.210256
## final  value 68.210252
## converged
```

```
summary(mult_logis_model)
```

```
## Call:
## multinom(formula = RA_ITA_ref ~ Age + Gender + Diabetes + Ever.smoked +
##     PVD + CVD, data = train_data)
##
## Coefficients:
##              (Intercept)        Age    Gender1 Diabetes1 Ever.smoked1
## RAyes_ITAno   -1.1723968 -0.0630500 -1.6778361 2.3575611    -1.219218
## RAno_ITAyes   -0.2351967  0.4636725  0.6666273 0.9110402     1.050832
## RAyes_ITAyes -17.5235137  1.8308741 14.6351213 2.1189413     1.877703
##                      PVD1       CVD1
## RAyes_ITAno   2.673941655  -1.418643
## RAno_ITAyes   0.950851956  -1.340274
## RAyes_ITAyes -0.009632406 -17.132857
##
## Std. Errors:
##              (Intercept)       Age   Gender1 Diabetes1 Ever.smoked1      PVD1
## RAyes_ITAno    1.0925674 0.7915507 1.6765158 1.5313009    1.4514035 1.8924280
## RAno_ITAyes    0.7197148 0.2933625 0.7941392 0.8750735    0.6381356 0.9326195
## RAyes_ITAyes   0.6707565 0.7766314 0.6707566 1.2710267    1.3327522 1.4620293
##                      CVD1
## RAyes_ITAno  2.370741e+00
## RAno_ITAyes  9.342838e-01
## RAyes_ITAyes 2.707484e-07
##
## Residual Deviance: 136.4205
## AIC: 178.4205
```

```
exp(coef(mult_logis_model))
```

```
##               (Intercept)       Age      Gender1 Diabetes1 Ever.smoked1
## RAyes_ITAno  3.096240e-01 0.9388965 1.867777e-01 10.565153    0.2954612
## RAno_ITAyes  7.904154e-01 1.5899023 1.947657e+00  2.486908    2.8600294
## RAyes_ITAyes 2.452645e-08 6.2393379 2.269616e+06  8.322322    6.5384689
##                    PVD1         CVD1
## RAyes_ITAno  14.4969989 2.420424e-01
## RAno_ITAyes   2.5879135 2.617740e-01
## RAyes_ITAyes  0.9904138 3.624890e-08
```

```
# 2-tailed z test
coef<-summary(mult_logis_model)$coefficients
se<-summary(mult_logis_model)$standard.errors
z <- coef/se
(p_value <- (1 - pnorm(abs(z), 0, 1)) * 2)
```

```
##              (Intercept)        Age   Gender1  Diabetes1 Ever.smoked1      PVD1
## RAyes_ITAno    0.2832416 0.93651262 0.3169295 0.12366262    0.4008934 0.1576649
## RAno_ITAyes    0.7438256 0.11398224 0.4012259 0.29782849    0.0996152 0.3079421
## RAyes_ITAyes   0.0000000 0.01840066 0.0000000 0.09549253    0.1588673 0.9947433
##                   CVD1
## RAyes_ITAno  0.5495755
## RAno_ITAyes  0.1514163
## RAyes_ITAyes 0.0000000
```

```
#Test the model fit info.
#The fitting model only with "intercept"
Intercept_model <- multinom(RA_ITA_ref~1,data = train_data)
```

```
## # weights:  8 (3 variable)
## initial  value 121.993904
## final  value 84.470523
## converged
```

```
summary(Intercept_model)
```

```
## Call:
## multinom(formula = RA_ITA_ref ~ 1, data = train_data)
##
## Coefficients:
##              (Intercept)
## RAyes_ITAno    -1.609436
## RAno_ITAyes     1.047317
## RAyes_ITAyes   -1.049830
##
## Std. Errors:
##              (Intercept)
## RAyes_ITAno    0.5477217
## RAno_ITAyes    0.2598918
## RAyes_ITAyes   0.4391559
##
## Residual Deviance: 168.941
## AIC: 174.941
```

```
anova(Intercept_model, mult_logis_model)
```

```
## Likelihood ratio tests of Multinomial Models
##
## Response: RA_ITA_ref
##                                               Model Resid. df Resid. Dev   Test
## 1                                                 1       261   168.9410
## 2 Age + Gender + Diabetes + Ever.smoked + PVD + CVD       243   136.4205 1 vs 2
##      Df LR stat.    Pr(Chi)
## 1
## 2    18 32.52054 0.01906273
```

The p-value < 0.05 indicates that our model fits significantly better than the model without any predictors.

```
#Goodness of fit test
chisq.test(train_data$RA_ITA_ref,predict(mult_logis_model))
```

```
## Warning in chisq.test(train_data$RA_ITA_ref, predict(mult_logis_model)): Chi-
## squared
```

```
##
##  Pearson's Chi-squared test
##
## data:  train_data$RA_ITA_ref and predict(mult_logis_model)
## X-squared = 30.075, df = 9, p-value = 0.000426
```

```r
#Pseudo R-Square of the model
library(DescTools)
```

```
##
##    'DescTools'
```

```
## The following object is masked from 'package:car':
##
##     Recode
```

```r
PseudoR2(mult_logis_model, which = c("CoxSnell","Nagelkerke","McFadden"))
```

```
## Warning in PseudoR2(mult_logis_model, which = c("CoxSnell", "Nagelkerke", :
## Could not find model or data element of multinom object for evaluating PseudoR2
## null model. Will fit null model with new evaluation of 'train_data'. Ensure
## object has not changed since initial call, or try running multinom with 'model =
## TRUE'
```

```
##   CoxSnell Nagelkerke   McFadden
##  0.3089559  0.3620458  0.1924964
```

According to the Cox and Snell's R-Square, the results concluded that there is 31.5% relationship between the predictors and the response variable. Nagelkerke's R-Square indicates that 36.3% of the variation in the predictors is explained by this model. McFadden's R-Square shows that the relationship of 18.8% between the predictors and the response variable.

```r
#significance of predictors by likelihood ratio tests
library(lmtest)
```

```
##      zoo
```

```
##
##    'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
lrtest(mult_logis_model, "Age")
```

```
## # weights:  28 (18 variable)
## initial  value 121.993904
## iter  10 value 74.131549
## iter  20 value 73.345665
## iter  30 value 73.301503
## final  value 73.300894
## converged
```

```
## Likelihood ratio test
##
## Model 1: RA_ITA_ref ~ Age + Gender + Diabetes + Ever.smoked + PVD + CVD
## Model 2: RA_ITA_ref ~ Gender + Diabetes + Ever.smoked + PVD + CVD
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  21 -68.210
## 2  18 -73.301 -3 10.181    0.01709 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lrtest(mult_logis_model, "Gender")
```

```
## # weights:  28 (18 variable)
## initial  value 121.993904
## iter  10 value 70.621830
## iter  20 value 70.020847
## iter  30 value 70.011349
## final  value 70.011277
## converged
```

```
## Likelihood ratio test
##
## Model 1: RA_ITA_ref ~ Age + Gender + Diabetes + Ever.smoked + PVD + CVD
## Model 2: RA_ITA_ref ~ Age + Diabetes + Ever.smoked + PVD + CVD
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  21 -68.210
## 2  18 -70.011 -3 3.6021     0.3078
```

```
lrtest(mult_logis_model, "Diabetes")
```

```
## # weights:  28 (18 variable)
## initial  value 121.993904
## iter  10 value 72.370967
## iter  20 value 70.488979
## iter  30 value 70.443720
## final  value 70.443099
## converged
```

```
## Likelihood ratio test
##
## Model 1: RA_ITA_ref ~ Age + Gender + Diabetes + Ever.smoked + PVD + CVD
## Model 2: RA_ITA_ref ~ Age + Gender + Ever.smoked + PVD + CVD
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  21 -68.210
## 2  18 -70.443 -3 4.4657     0.2154
```

```
lrtest(mult_logis_model, "Ever.smoked")
```

```
## # weights:  28 (18 variable)
## initial  value 121.993904
## iter  10 value 72.443656
## iter  20 value 71.305433
## iter  30 value 71.258140
## final  value 71.257524
## converged
```

```
## Likelihood ratio test
##
## Model 1: RA_ITA_ref ~ Age + Gender + Diabetes + Ever.smoked + PVD + CVD
## Model 2: RA_ITA_ref ~ Age + Gender + Diabetes + PVD + CVD
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  21 -68.210
## 2  18 -71.258 -3 6.0945     0.1071
```

```
lrtest(mult_logis_model, "PVD")
```

```
## # weights:  28 (18 variable)
## initial  value 121.993904
## iter  10 value 71.345436
## iter  20 value 69.822466
## iter  30 value 69.793592
## final  value 69.793426
## converged
```

```
## Likelihood ratio test
##
## Model 1: RA_ITA_ref ~ Age + Gender + Diabetes + Ever.smoked + PVD + CVD
## Model 2: RA_ITA_ref ~ Age + Gender + Diabetes + Ever.smoked + CVD
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  21 -68.210
## 2  18 -69.793 -3 3.1663     0.3667
```

```
lrtest(mult_logis_model, "CVD")
```

```
## # weights:  28 (18 variable)
## initial  value 121.993904
## iter  10 value 70.826946
## iter  20 value 70.052927
## iter  30 value 70.024703
## final  value 70.024570
## converged
```

```
## Likelihood ratio test
##
## Model 1: RA_ITA_ref ~ Age + Gender + Diabetes + Ever.smoked + PVD + CVD
## Model 2: RA_ITA_ref ~ Age + Gender + Diabetes + Ever.smoked + PVD
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  21 -68.210
## 2  18 -70.025 -3 3.6286     0.3045
```

According to the results of the likelihood ratio tests, Age had significant main effects the dependent variable.

After fitting the multinomial logistic regression model, prediction and validation based on train data and test data would be implemented in this part.

```
#predict the outcome based on train data
train_predict = predict(mult_logis_model, train_data, "class")
train_class_table = table(train_data$RA_ITA_ref, train_predict)
round((sum(diag(train_class_table))/sum(train_class_table))*100,2)
```

```
## [1] 67.05
```

```
#predict the outcome based on test data
test_predict = predict(mult_logis_model, test_data, "class")
test_class_table = table(test_data$RA_ITA_ref, test_predict)
round((sum(diag(test_class_table))/sum(test_class_table))*100,2)
```

```
## [1] 45.45
```

```
#stepwide model
library(MASS)
```

```
##
##    'MASS'
```

```
## The following object is masked from 'package:dplyr':
```

```
##
##     select
```

```
mult_logis_model2 <- mult_logis_model %>% stepAIC(trace = FALSE)
```

```
## # weights:  28 (18 variable)
## initial  value 121.993904
## iter  10 value 74.131549
## iter  20 value 73.345665
## iter  30 value 73.301503
## final  value 73.300894
## converged
## # weights:  28 (18 variable)
## initial  value 121.993904
## iter  10 value 70.621830
## iter  20 value 70.020847
## iter  30 value 70.011349
## final  value 70.011277
## converged
## # weights:  28 (18 variable)
## initial  value 121.993904
## iter  10 value 72.370967
## iter  20 value 70.488979
## iter  30 value 70.443720
## final  value 70.443099
## converged
## # weights:  28 (18 variable)
## initial  value 121.993904
## iter  10 value 72.443656
## iter  20 value 71.305433
## iter  30 value 71.258140
## final  value 71.257524
## converged
## # weights:  28 (18 variable)
## initial  value 121.993904
## iter  10 value 71.345436
## iter  20 value 69.822466
## iter  30 value 69.793592
## final  value 69.793426
## converged
## # weights:  28 (18 variable)
## initial  value 121.993904
## iter  10 value 70.826946
## iter  20 value 70.052927
## iter  30 value 70.024703
## final  value 70.024570
## converged
## # weights:  28 (18 variable)
## initial  value 121.993904
## iter  10 value 71.345436
## iter  20 value 69.822466
## iter  30 value 69.793592
## final  value 69.793426
## converged
## # weights:  24 (15 variable)
## initial  value 121.993904
## iter  10 value 75.533447
## iter  20 value 74.945110
## iter  30 value 74.931153
## final  value 74.931117
## converged
## # weights:  24 (15 variable)
## initial  value 121.993904
## iter  10 value 71.508229
## iter  20 value 71.247952
## iter  30 value 71.244692
## final  value 71.244676
## converged
## # weights:  24 (15 variable)
## initial  value 121.993904
## iter  10 value 74.089662
## iter  20 value 72.723617
## iter  30 value 72.710076
## final  value 72.710039
## converged
## # weights:  24 (15 variable)
## initial  value 121.993904
## iter  10 value 73.746790
## iter  20 value 73.236135
## iter  30 value 73.229825
## final  value 73.229817
## converged
## # weights:  24 (15 variable)
## initial  value 121.993904
## iter  10 value 72.446426
## iter  20 value 71.609265
## iter  30 value 71.587107
## final  value 71.587040
## converged
## # weights:  24 (15 variable)
## initial  value 121.993904
## iter  10 value 71.508229
## iter  20 value 71.247952
## iter  30 value 71.244692
## final  value 71.244676
## converged
## # weights:  20 (12 variable)
## initial  value 121.993904
## iter  10 value 76.560137
## iter  20 value 76.441217
## final  value 76.440718
## converged
## # weights:  20 (12 variable)
## initial  value 121.993904
## iter  10 value 74.646486
## iter  20 value 74.212961
## final  value 74.211455
## converged
## # weights:  20 (12 variable)
## initial  value 121.993904
## iter  10 value 76.681010
## iter  20 value 76.411210
## final  value 76.409386
## converged
## # weights:  20 (12 variable)
## initial  value 121.993904
## iter  10 value 73.402311
## final  value 73.361302
## converged
## # weights:  20 (12 variable)
## initial  value 121.993904
## iter  10 value 73.402311
## final  value 73.361302
## converged
## # weights:  16 (9 variable)
## initial  value 121.993904
## iter  10 value 77.850092
## final  value 77.846650
## converged
## # weights:  16 (9 variable)
## initial  value 121.993904
## iter  10 value 76.913360
## final  value 76.905717
## converged
## # weights:  16 (9 variable)
## initial  value 121.993904
## iter  10 value 78.113872
## final  value 78.106615
## converged
```

```
summary(mult_logis_model2)
```

```
## Call:
## multinom(formula = RA_ITA_ref ~ Age + Diabetes + Ever.smoked,
##     data = train_data)
##
## Coefficients:
##              (Intercept)       Age Diabetes1 Ever.smoked1
## RAyes_ITAno   -1.7884095 0.4651306  3.102892    -1.741266
## RAno_ITAyes    0.2799777 0.4156539  1.011448     1.064877
## RAyes_ITAyes  -3.5572372 1.8352399  2.279452     1.957228
##
## Std. Errors:
##              (Intercept)       Age Diabetes1 Ever.smoked1
## RAyes_ITAno    0.8168906 0.5896445 1.4338692    1.4126078
## RAno_ITAyes    0.4206333 0.2647853 0.8528984    0.5779113
## RAyes_ITAyes   1.3386071 0.8226671 1.2125647    1.2582730
##
## Residual Deviance: 146.7226
## AIC: 170.7226
```

```
exp(coef(mult_logis_model2))
```

```
##              (Intercept)      Age Diabetes1 Ever.smoked1
## RAyes_ITAno    0.1672259 1.592222 22.262245    0.1752982
## RAno_ITAyes    1.3231003 1.515361  2.749580    2.9004814
## RAyes_ITAyes   0.0285175 6.266638  9.771324    7.0796748
```

```r
# 2-tailed z test
coef<-summary(mult_logis_model2)$coefficients
se<-summary(mult_logis_model2)$standard.errors
z <- coef/se
(p_value <- (1 - pnorm(abs(z), 0, 1)) * 2)
```

```
##              (Intercept)        Age  Diabetes1 Ever.smoked1
## RAyes_ITAno   0.02857586 0.43021004 0.03046439   0.21770228
## RAno_ITAyes   0.50566049 0.11646689 0.23566363   0.06538306
## RAyes_ITAyes  0.00787420 0.02569162 0.06012714   0.11982999
```

```r
#Test the model fit info.
anova(Intercept_model, mult_logis_model2)
```

```
## Likelihood ratio tests of Multinomial Models
##
## Response: RA_ITA_ref
##                            Model Resid. df Resid. Dev   Test    Df LR stat.
## 1                            1       261   168.9410
## 2 Age + Diabetes + Ever.smoked       252   146.7226 1 vs 2     9 22.21844
##       Pr(Chi)
## 1
## 2 0.008211855
```

```r
anova(mult_logis_model, mult_logis_model2)
```

```
## Likelihood ratio tests of Multinomial Models
##
## Response: RA_ITA_ref
##                                                Model Resid. df Resid. Dev   Test
## 1                      Age + Diabetes + Ever.smoked       252   146.7226
## 2 Age + Gender + Diabetes + Ever.smoked + PVD + CVD       243   136.4205 1 vs 2
##      Df LR stat.  Pr(Chi)
## 1
## 2     9  10.3021 0.3265867
```

```r
#Goodness of fit test
chisq.test(train_data$RA_ITA_ref,predict(mult_logis_model2))
```

```
## Warning in chisq.test(train_data$RA_ITA_ref, predict(mult_logis_model2)): Chi-
## squared
```

```
##
##  Pearson's Chi-squared test
##
## data:  train_data$RA_ITA_ref and predict(mult_logis_model2)
## X-squared = 47.202, df = 6, p-value = 1.706e-08
```

```r
#Pseudo R-Square of the model
PseudoR2(mult_logis_model2, which = c("CoxSnell","Nagelkerke","McFadden"))
```

```
## Warning in PseudoR2(mult_logis_model2, which = c("CoxSnell", "Nagelkerke", :
## Could not find model or data element of multinom object for evaluating PseudoR2
## null model. Will fit null model with new evaluation of 'train_data'. Ensure
## object has not changed since initial call, or try running multinom with 'model =
## TRUE'
```

```
##   CoxSnell Nagelkerke   McFadden
##  0.2231300  0.2614719  0.1315160
```

```
#significance of predictors by likelihood ratio tests
lrtest(mult_logis_model2, "Age")
```

```
## # weights:  16 (9 variable)
## initial  value 121.993904
## iter  10 value 77.850092
## final  value 77.846650
## converged
```

```
## Likelihood ratio test
##
## Model 1: RA_ITA_ref ~ Age + Diabetes + Ever.smoked
## Model 2: RA_ITA_ref ~ Diabetes + Ever.smoked
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  12 -73.361
## 2   9 -77.847 -3 8.9707    0.02968 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lrtest(mult_logis_model2, "Diabetes")
```

```
## # weights:  16 (9 variable)
## initial  value 121.993904
## iter  10 value 76.913360
## final  value 76.905717
## converged
```

```
## Likelihood ratio test
##
## Model 1: RA_ITA_ref ~ Age + Diabetes + Ever.smoked
## Model 2: RA_ITA_ref ~ Age + Ever.smoked
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  12 -73.361
## 2   9 -76.906 -3 7.0888    0.06912 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lrtest(mult_logis_model2, "Ever.smoked")
```

```
## # weights:  16 (9 variable)
## initial  value 121.993904
## iter  10 value 78.113872
## final  value 78.106615
## converged
```

```
## Likelihood ratio test
##
## Model 1: RA_ITA_ref ~ Age + Diabetes + Ever.smoked
## Model 2: RA_ITA_ref ~ Age + Diabetes
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  12 -73.361
## 2   9 -78.107 -3 9.4906    0.02343 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#predict the outcome based on train data
train_predict = predict(mult_logis_model2, train_data, "class")
train_class_table = table(train_data$RA_ITA_ref, train_predict)
round((sum(diag(train_class_table))/sum(train_class_table))*100,2)
```

```
## [1] 68.18
```

```
#predict the outcome based on test data
test_predict = predict(mult_logis_model2, test_data, "class")
test_class_table = table(test_data$RA_ITA_ref, test_predict)
round((sum(diag(test_class_table))/sum(test_class_table))*100,2)
```

```
## [1] 40.91
```

In addition, we tried to divide different age into 4 different age groups as a categorical variable.

```
artery_data4$Age <- artery_data3$Age
summary(artery_data4$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   42.00   60.00   68.50   65.77   72.75   81.00 
```

```
artery_data5 <- mutate(artery_data4,
                     Age_group = ifelse(Age <= 50 ,'0',
                                        ifelse(Age > 50 & Age <=60 ,'1',
                                                    ifelse(Age > 60 & Age <=70, '2','3'))))
```

```
set.seed(123)
s = sort(sample(nrow(artery_data5), nrow(artery_data5)*0.8))
train_data<-artery_data5[s,]
test_data<-artery_data5[-s,]

#model3
mult_logis_model3<-multinom(RA_ITA_ref~Age_group+Gender+Diabetes+Ever.smoked+PVD+CVD,
                            data = train_data)
```

```
## # weights:  40 (27 variable)
## initial  value 121.993904 
## iter  10 value 68.705439
## iter  20 value 66.624480
## iter  30 value 66.527948
## iter  40 value 66.522905
## final  value 66.522885 
## converged
```

```
summary(mult_logis_model3)
```

```
## Call:
## multinom(formula = RA_ITA_ref ~ Age_group + Gender + Diabetes + 
##     Ever.smoked + PVD + CVD, data = train_data)
## 
## Coefficients:
##              (Intercept)  Age_group1 Age_group2 Age_group3    Gender1 Diabetes1
## RAyes_ITAno   -16.041390 15.16399908  14.952031 14.3956142 -1.1344855  3.193248
## RAno_ITAyes    -1.235343  0.06491754   1.403178  0.9889106  0.9540978  0.830664
## RAyes_ITAyes  -38.529993 -4.97595008  19.474478 19.8999715 17.5040485  1.851117
##              Ever.smoked1       PVD1       CVD1
## RAyes_ITAno    -1.5763450  2.019179  -1.163310
## RAno_ITAyes     0.9640962  1.002400  -1.327261
## RAyes_ITAyes    1.6845784 -0.155266 -17.867467
##
## Std. Errors:
##              (Intercept)   Age_group1 Age_group2 Age_group3   Gender1 Diabetes1
## RAyes_ITAno    0.8218037 1.064676e+00  1.0183107  1.0259682 1.6320071  1.787245
## RAno_ITAyes    1.1789508 9.577295e-01  1.0476168  0.9686602 0.8649831  0.919818
## RAyes_ITAyes   0.4712693 1.382583e-10  0.7031914  0.5765850 0.4712693  1.293769
##              Ever.smoked1      PVD1         CVD1
## RAyes_ITAno     1.7846857 1.8114979 2.405262e+00
## RAno_ITAyes     0.6395444 0.9212534 9.410983e-01
## RAyes_ITAyes    1.3062755 1.4766502 1.141924e-07
##
## Residual Deviance: 133.0458
## AIC: 187.0458
```

```
exp(coef(mult_logis_model3))
```

```
##                (Intercept)   Age_group1   Age_group2   Age_group3      Gender1
## RAyes_ITAno  1.079724e-07 3.851600e+06 3.115906e+06 1.786224e+06 3.215875e-01
## RAno_ITAyes  2.907351e-01 1.067071e+00 4.068110e+00 2.688304e+00 2.596327e+00
## RAyes_ITAyes 1.847723e-17 6.901958e-03 2.868522e+08 4.389831e+08 3.998634e+07
##              Diabetes1 Ever.smoked1      PVD1         CVD1
## RAyes_ITAno  24.367454    0.2067293 7.5321361 3.124502e-01
## RAno_ITAyes   2.294842    2.6224165 2.7248125 2.652025e-01
## RAyes_ITAyes  6.366927    5.3901777 0.8561874 1.738833e-08
```

```
# 2-tailed z test
coef<-summary(mult_logis_model3)$coefficients
se<-summary(mult_logis_model3)$standard.errors
z <- coef/se
(p_value <- (1 - pnorm(abs(z), 0, 1)) * 2)
```

```
##              (Intercept) Age_group1 Age_group2 Age_group3   Gender1  Diabetes1
## RAyes_ITAno    0.0000000  0.0000000  0.0000000  0.0000000 0.4869629 0.07398806
## RAno_ITAyes    0.2947159  0.9459586  0.1804403  0.3072992 0.2700164 0.36648644
## RAyes_ITAyes   0.0000000  0.0000000  0.0000000  0.0000000 0.0000000 0.15248927
##              Ever.smoked1      PVD1      CVD1
## RAyes_ITAno     0.3770948 0.2650022 0.6286328
## RAno_ITAyes     0.1316893 0.2765587 0.1584415
## RAyes_ITAyes    0.1971881 0.9162588 0.0000000
```

```
#Test the model fit info.
anova(Intercept_model, mult_logis_model3)
```

```
## Likelihood ratio tests of Multinomial Models
##
## Response: RA_ITA_ref
##                                                     Model Resid. df Resid. Dev
## 1                                                       1       261   168.9410
## 2 Age_group + Gender + Diabetes + Ever.smoked + PVD + CVD       237   133.0458
##     Test    Df LR stat.    Pr(Chi)
## 1
## 2 1 vs 2    24 35.89528 0.05618454
```

```
anova(mult_logis_model, mult_logis_model3)
```

```
## Likelihood ratio tests of Multinomial Models
##
## Response: RA_ITA_ref
##                                                     Model Resid. df Resid. Dev
## 1       Age + Gender + Diabetes + Ever.smoked + PVD + CVD       243   136.4205
## 2 Age_group + Gender + Diabetes + Ever.smoked + PVD + CVD       237   133.0458
##     Test    Df LR stat.   Pr(Chi)
## 1
## 2 1 vs 2     6 3.374733 0.7605544
```

```
anova(mult_logis_model2, mult_logis_model3)
```

```
## Likelihood ratio tests of Multinomial Models
##
## Response: RA_ITA_ref
##                                                     Model Resid. df Resid. Dev
## 1                            Age + Diabetes + Ever.smoked       252   146.7226
## 2 Age_group + Gender + Diabetes + Ever.smoked + PVD + CVD       237   133.0458
##     Test    Df LR stat.   Pr(Chi)
## 1
## 2 1 vs 2    15 13.67683 0.5501665
```

```
#Goodness of fit test
chisq.test(train_data$RA_ITA_ref,predict(mult_logis_model3))
```

```
## Warning in chisq.test(train_data$RA_ITA_ref, predict(mult_logis_model3)): Chi-
## squared
```

```
##
##  Pearson's Chi-squared test
##
## data:  train_data$RA_ITA_ref and predict(mult_logis_model3)
## X-squared = 35.644, df = 6, p-value = 3.232e-06
```

```
#Pseudo R-Square of the model
PseudoR2(mult_logis_model3, which = c("CoxSnell","Nagelkerke","McFadden"))
```

```
## Warning in PseudoR2(mult_logis_model3, which = c("CoxSnell", "Nagelkerke", :
## Could not find model or data element of multinom object for evaluating PseudoR2
## null model. Will fit null model with new evaluation of 'train_data'. Ensure
## object has not changed since initial call, or try running multinom with 'model =
## TRUE'
```

```
## Warning: Using formula(x) is deprecated when x is a character vector of length > 1.
##   Consider formula(paste(x, collapse = " ")) instead.
```

```
##   CoxSnell Nagelkerke   McFadden
##  0.3349552  0.3925127  0.2124722
```

```
#significance of predictors by likelihood ratio tests
lrtest(mult_logis_model3, "Age_group")
```

```
## # weights:  28 (18 variable)
## initial  value 121.993904
## iter  10 value 74.131549
## iter  20 value 73.345665
## iter  30 value 73.301503
## final  value 73.300894
## converged
```

```
## Likelihood ratio test
##
## Model 1: RA_ITA_ref ~ Age_group + Gender + Diabetes + Ever.smoked + PVD +
##     CVD
## Model 2: RA_ITA_ref ~ Gender + Diabetes + Ever.smoked + PVD + CVD
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  27 -66.523
## 2  18 -73.301 -9 13.556      0.139
```

```
lrtest(mult_logis_model3, "Gender")
```

```
## # weights:  36 (24 variable)
## initial  value 121.993904
## iter  10 value 70.454075
## iter  20 value 68.567023
## iter  30 value 68.529947
## iter  40 value 68.529557
## final  value 68.529555
## converged
```

```
## Likelihood ratio test
##
## Model 1: RA_ITA_ref ~ Age_group + Gender + Diabetes + Ever.smoked + PVD +
##     CVD
## Model 2: RA_ITA_ref ~ Age_group + Diabetes + Ever.smoked + PVD + CVD
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  27 -66.523
## 2  24 -68.530 -3 4.0133       0.26
```

```
lrtest(mult_logis_model3, "Diabetes")
```

```
## # weights:  36 (24 variable)
## initial  value 121.993904
## iter  10 value 70.787515
## iter  20 value 69.023937
## iter  30 value 68.971179
## iter  40 value 68.968044
## final  value 68.968031
## converged
```

```
## Likelihood ratio test
##
## Model 1: RA_ITA_ref ~ Age_group + Gender + Diabetes + Ever.smoked + PVD +
##     CVD
## Model 2: RA_ITA_ref ~ Age_group + Gender + Ever.smoked + PVD + CVD
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  27 -66.523
## 2  24 -68.968 -3 4.8903       0.18
```

```
lrtest(mult_logis_model3, "Ever.smoked")
```

```
## # weights:  36 (24 variable)
## initial  value 121.993904
## iter  10 value 71.780767
## iter  20 value 69.304498
## iter  30 value 69.146318
## iter  40 value 69.144255
## final  value 69.144247
## converged
```

```
## Likelihood ratio test
##
## Model 1: RA_ITA_ref ~ Age_group + Gender + Diabetes + Ever.smoked + PVD +
##     CVD
## Model 2: RA_ITA_ref ~ Age_group + Gender + Diabetes + PVD + CVD
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  27 -66.523
## 2  24 -69.144 -3 5.2427     0.1549
```

```
lrtest(mult_logis_model3, "PVD")
```

```
## # weights:  36 (24 variable)
## initial  value 121.993904
## iter  10 value 70.519283
## iter  20 value 68.008454
## iter  30 value 67.977782
## iter  40 value 67.976999
## final  value 67.976994
## converged
```

```
## Likelihood ratio test
##
## Model 1: RA_ITA_ref ~ Age_group + Gender + Diabetes + Ever.smoked + PVD +
##     CVD
## Model 2: RA_ITA_ref ~ Age_group + Gender + Diabetes + Ever.smoked + CVD
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  27 -66.523
## 2  24 -67.977 -3 2.9082      0.406
```

```
lrtest(mult_logis_model3, "CVD")
```

```
## # weights:  36 (24 variable)
## initial  value 121.993904
## iter  10 value 70.015440
## iter  20 value 68.270621
## iter  30 value 68.220327
## iter  40 value 68.219459
## final  value 68.219457
## converged
```

```
## Likelihood ratio test
##
## Model 1: RA_ITA_ref ~ Age_group + Gender + Diabetes + Ever.smoked + PVD +
##     CVD
## Model 2: RA_ITA_ref ~ Age_group + Gender + Diabetes + Ever.smoked + PVD
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  27 -66.523
## 2  24 -68.219 -3 3.3931     0.3349
```

```r
#predict the outcome based on train data
train_predict = predict(mult_logis_model3, train_data, "class")
train_class_table = table(train_data$RA_ITA_ref, train_predict)
round((sum(diag(train_class_table))/sum(train_class_table))*100,2)
```

```
## [1] 70.45
```

```r
#predict the outcome based on test data
test_predict = predict(mult_logis_model3, test_data, "class")
test_class_table = table(test_data$RA_ITA_ref, test_predict)
round((sum(diag(test_class_table))/sum(test_class_table))*100,2)
```

```
## [1] 40.91
```

```r
#stepwide model: model4
mult_logis_model4 <- mult_logis_model3 %>% stepAIC(trace = FALSE)
```

```
## # weights:  28 (18 variable)
## initial  value 121.993904
## iter  10 value 74.131549
## iter  20 value 73.345665
## iter  30 value 73.301503
## final  value 73.300894
## converged
## # weights:  36 (24 variable)
## initial  value 121.993904
## iter  10 value 70.454075
## iter  20 value 68.567023
## iter  30 value 68.529947
## iter  40 value 68.529557
## final  value 68.529555
## converged
## # weights:  36 (24 variable)
## initial  value 121.993904
## iter  10 value 70.787515
## iter  20 value 69.023937
## iter  30 value 68.971179
## iter  40 value 68.968044
## final  value 68.968031
## converged
## # weights:  36 (24 variable)
## initial  value 121.993904
## iter  10 value 71.780767
## iter  20 value 69.304498
## iter  30 value 69.146318
## iter  40 value 69.144255
## final  value 69.144247
## converged
## # weights:  36 (24 variable)
## initial  value 121.993904
## iter  10 value 70.519283
## iter  20 value 68.008454
## iter  30 value 67.977782
## iter  40 value 67.976999
## final  value 67.976994
## converged
## # weights:  36 (24 variable)
## initial  value 121.993904
## iter  10 value 70.015440
## iter  20 value 68.270621
## iter  30 value 68.220327
## iter  40 value 68.219459
## final  value 68.219457
## converged
## # weights:  28 (18 variable)
## initial  value 121.993904
## iter  10 value 74.131549
## iter  20 value 73.345665
## iter  30 value 73.301503
## final  value 73.300894
## converged
## # weights:  24 (15 variable)
## initial  value 121.993904
## iter  10 value 75.228056
## iter  20 value 74.984073
## iter  30 value 74.981707
## final  value 74.981704
## converged
## # weights:  24 (15 variable)
## initial  value 121.993904
## iter  10 value 75.757521
## iter  20 value 74.931644
## iter  30 value 74.900449
## final  value 74.900338
## converged
## # weights:  24 (15 variable)
## initial  value 121.993904
## iter  10 value 76.151000
## iter  20 value 75.657173
## iter  30 value 75.637127
## final  value 75.637056
## converged
## # weights:  24 (15 variable)
## initial  value 121.993904
## iter  10 value 75.533447
## iter  20 value 74.945110
## iter  30 value 74.931153
## final  value 74.931117
## converged
## # weights:  24 (15 variable)
## initial  value 121.993904
## iter  10 value 75.057882
## iter  20 value 74.657466
## iter  30 value 74.644913
## final  value 74.644890
## converged
## # weights:  24 (15 variable)
## initial  value 121.993904
## iter  10 value 75.057882
## iter  20 value 74.657466
## iter  30 value 74.644913
## final  value 74.644890
## converged
## # weights:  20 (12 variable)
## initial  value 121.993904
## iter  10 value 76.554567
## final  value 76.508427
## converged
## # weights:  20 (12 variable)
## initial  value 121.993904
## iter  10 value 76.910235
## iter  20 value 76.484046
## final  value 76.478882
## converged
## # weights:  20 (12 variable)
## initial  value 121.993904
## iter  10 value 77.170378
## iter  20 value 76.803243
## final  value 76.799554
## converged
## # weights:  20 (12 variable)
## initial  value 121.993904
## iter  10 value 76.602981
## iter  20 value 76.054792
## iter  30 value 76.046028
## iter  30 value 76.046027
## iter  30 value 76.046027
## final  value 76.046027
## converged
## # weights:  20 (12 variable)
## initial  value 121.993904
## iter  10 value 76.602981
## iter  20 value 76.054792
## iter  30 value 76.046028
## iter  30 value 76.046027
## iter  30 value 76.046027
## final  value 76.046027
## converged
## # weights:  16 (9 variable)
## initial  value 121.993904
## iter  10 value 77.850092
## final  value 77.846650
## converged
## # weights:  16 (9 variable)
## initial  value 121.993904
## iter  10 value 79.040442
## iter  20 value 78.712495
## iter  30 value 78.712155
## iter  40 value 78.710248
## final  value 78.710241
## converged
## # weights:  16 (9 variable)
## initial  value 121.993904
## iter  10 value 78.748844
## iter  20 value 78.478585
## iter  30 value 78.478394
## final  value 78.477320
## converged
## # weights:  16 (9 variable)
## initial  value 121.993904
## iter  10 value 77.850092
## final  value 77.846650
## converged
## # weights:  12 (6 variable)
## initial  value 121.993904
## iter  10 value 80.525802
## final  value 80.525316
## converged
## # weights:  12 (6 variable)
## initial  value 121.993904
## iter  10 value 81.944429
## final  value 81.944380
## converged
## # weights:  12 (6 variable)
## initial  value 121.993904
## iter  10 value 80.525802
## final  value 80.525316
## converged
## # weights:  8 (3 variable)
## initial  value 121.993904
## final  value 84.470523
## converged
```

```
summary(mult_logis_model4)
```

```
## Call:
## multinom(formula = RA_ITA_ref ~ Ever.smoked, data = train_data)
##
## Coefficients:
##              (Intercept) Ever.smoked1
## RAyes_ITAno   -1.2039706    -1.098603
## RAno_ITAyes    0.4054649     1.029620
## RAyes_ITAyes  -2.3025740     1.791750
##
## Std. Errors:
##              (Intercept) Ever.smoked1
## RAyes_ITAno    0.6582802    1.2382738
## RAno_ITAyes    0.4082484    0.5389586
## RAyes_ITAyes   1.0488038    1.1690407
##
## Residual Deviance: 161.0506
## AIC: 173.0506
```

```
exp(coef(mult_logis_model4))
```

```
##              (Intercept) Ever.smoked1
## RAyes_ITAno    0.3000007    0.3333365
## RAno_ITAyes    1.4999998    2.8000026
## RAyes_ITAyes   0.1000011    5.9999428
```

```
# 2-tailed z test
coef<-summary(mult_logis_model4)$coefficients
se<-summary(mult_logis_model4)$standard.errors
z <- coef/se
(p_value <- (1 - pnorm(abs(z), 0, 1)) * 2)
```

```
##              (Intercept) Ever.smoked1
## RAyes_ITAno   0.06740504   0.37496856
## RAno_ITAyes   0.32062130   0.05608321
## RAyes_ITAyes  0.02813286   0.12535795
```

```
#Test the model fit info.
anova(Intercept_model, mult_logis_model4)
```

```
## Likelihood ratio tests of Multinomial Models
##
## Response: RA_ITA_ref
##         Model Resid. df Resid. Dev   Test    Df LR stat.    Pr(Chi)
## 1           1       261   168.9410
## 2 Ever.smoked       258   161.0506 1 vs 2     3 7.890415 0.04833162
```

```
anova(mult_logis_model, mult_logis_model4)
```

```
## Likelihood ratio tests of Multinomial Models
##
## Response: RA_ITA_ref
##                                               Model Resid. df Resid. Dev   Test
## 1                                       Ever.smoked       258   161.0506
## 2 Age + Gender + Diabetes + Ever.smoked + PVD + CVD       243   136.4205 1 vs 2
##      Df LR stat.    Pr(Chi)
## 1
## 2    15 24.63013 0.05513663
```

```
anova(mult_logis_model2, mult_logis_model4)
```

```
## Likelihood ratio tests of Multinomial Models
##
## Response: RA_ITA_ref
##                          Model Resid. df Resid. Dev   Test    Df LR stat.
## 1                  Ever.smoked       258   161.0506
## 2 Age + Diabetes + Ever.smoked       252   146.7226 1 vs 2     6 14.32803
##      Pr(Chi)
## 1
## 2 0.02617901
```

```
anova(mult_logis_model3, mult_logis_model4)
```

```
## Likelihood ratio tests of Multinomial Models
##
## Response: RA_ITA_ref
##                                                     Model Resid. df Resid. Dev
## 1                                             Ever.smoked       258   161.0506
## 2 Age_group + Gender + Diabetes + Ever.smoked + PVD + CVD       237   133.0458
##     Test    Df LR stat.   Pr(Chi)
## 1
## 2 1 vs 2    21 28.00486 0.1400135
```

```
#Goodness of fit test
fisher.test(train_data$RA_ITA_ref,predict(mult_logis_model4))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  train_data$RA_ITA_ref and predict(mult_logis_model4)
## p-value = 1
## alternative hypothesis: two.sided
```

```
#Pseudo R-Square of the model
PseudoR2(mult_logis_model4, which = c("CoxSnell","Nagelkerke","McFadden"))
```

```
## Warning in PseudoR2(mult_logis_model4, which = c("CoxSnell", "Nagelkerke", :
## Could not find model or data element of multinom object for evaluating PseudoR2
## null model. Will fit null model with new evaluation of 'train_data'. Ensure
## object has not changed since initial call, or try running multinom with 'model =
## TRUE'
```

```
##   CoxSnell Nagelkerke   McFadden
## 0.08576151 0.10049846 0.04670514
```

```
#significance of predictors by likelihood ratio tests
lrtest(mult_logis_model4, "Ever.smoked")
```

```
## # weights:  8 (3 variable)
## initial  value 121.993904
## final  value 84.470523
## converged
```

```
## Likelihood ratio test
##
## Model 1: RA_ITA_ref ~ Ever.smoked
## Model 2: RA_ITA_ref ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   6 -80.525
## 2   3 -84.471 -3 7.8904    0.04833 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#predict the outcome based on train data
train_predict = predict(mult_logis_model4, train_data, "class")
train_class_table = table(train_data$RA_ITA_ref, train_predict)
round((sum(diag(train_class_table))/sum(train_class_table))*100,2)
```

```
## [1] 64.77
```

```
#predict the outcome based on test data
test_predict = predict(mult_logis_model4, test_data, "class")
test_class_table = table(test_data$RA_ITA_ref, test_predict)
round((sum(diag(test_class_table))/sum(test_class_table))*100,2)
```

```
## [1] 45.45
```

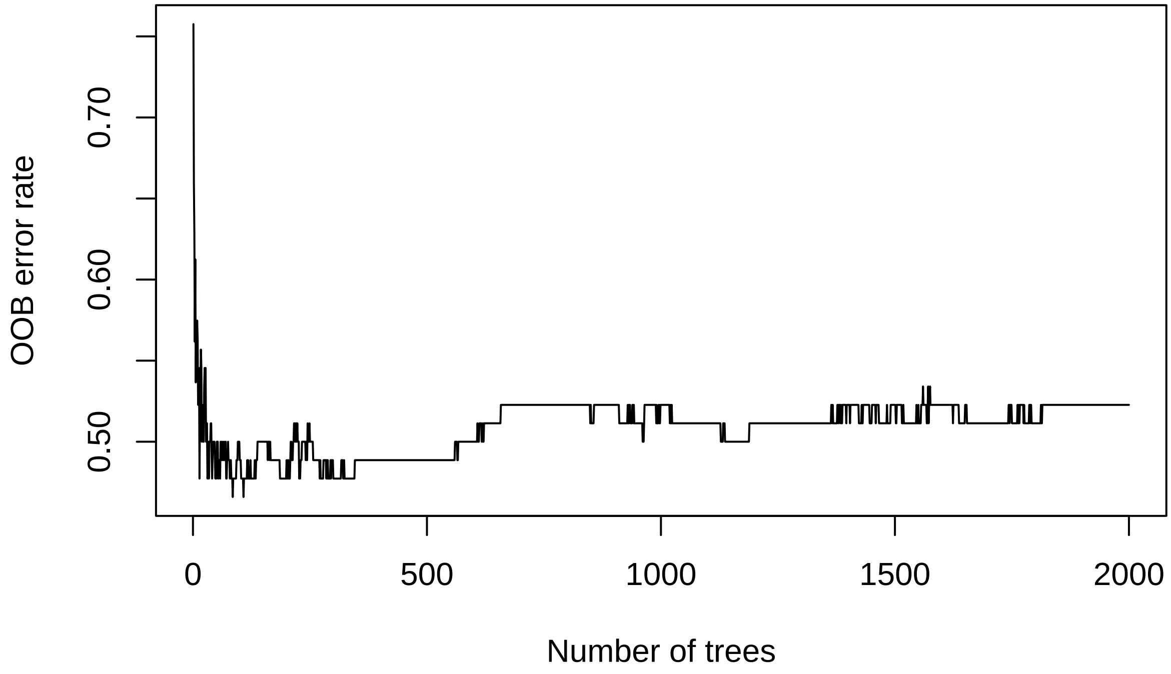
Method 2: Random Forest

Moreover, another model we choose is random forest.

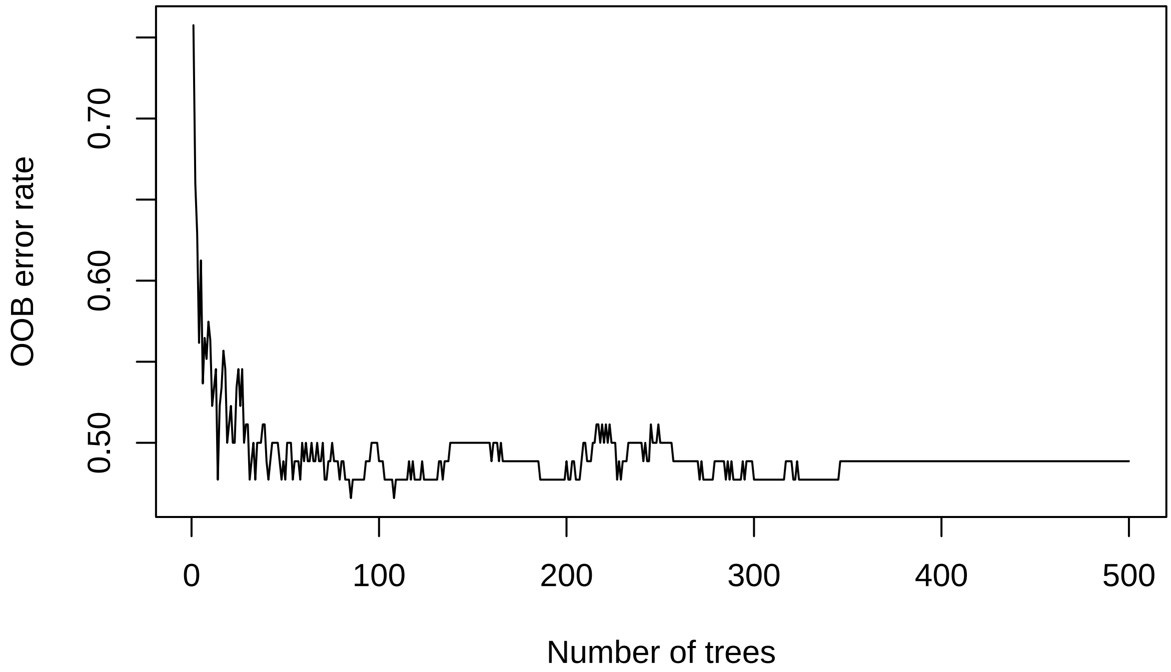
```r
library(randomForest)
```

```
## randomForest 4.7-1.1

## Type rfNews() to see new features/changes/bug fixes.

##
##    'randomForest'

## The following object is masked from 'package:ggplot2':
##
##     margin

## The following object is masked from 'package:dplyr':
##
##     combine
```

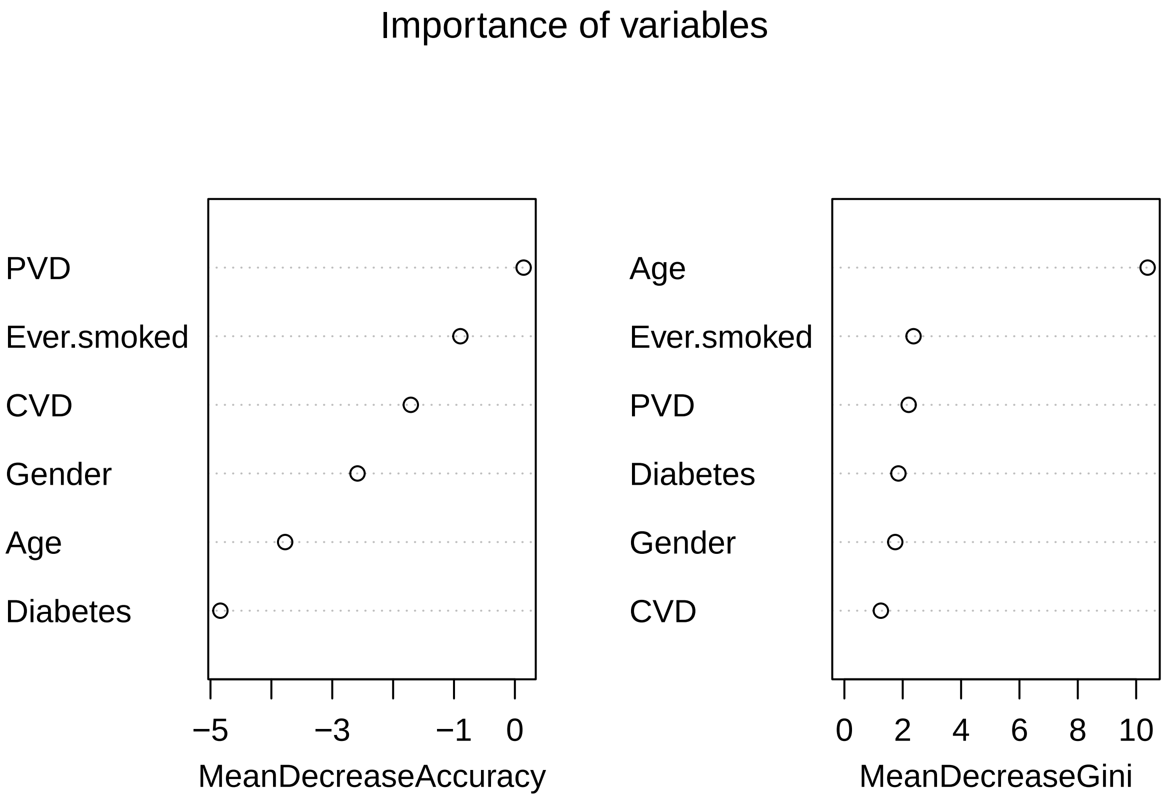
```r
# when the age is continuous
s = sort(sample(nrow(artery_data4), nrow(artery_data4)*0.8))
train_data<-artery_data4[s,]
test_data<-artery_data4[-s,]

# model1
set.seed(12345)
rf_model <- randomForest(RA_ITA_ref~Age+Gender+Diabetes+Ever.smoked+PVD+CVD,
                         data = train_data, ntree = 2000, importance = TRUE)
plot(rf_model$err.rate[,1], xlab = "Number of trees", ylab = "OOB error rate", type = "l")
```



```r
# model2 with 500 trees
set.seed(12345)
rf_model2 <- randomForest(RA_ITA_ref~Age+Gender+Diabetes+Ever.smoked+PVD+CVD,
                          data = train_data, ntree = 500, importance = TRUE)
plot(rf_model2$err.rate[,1], xlab = "Number of trees", ylab = "OOB error rate", type = "l")
```



```
# Importance of variables
varImpPlot(rf_model2, main = "Importance of variables")
```

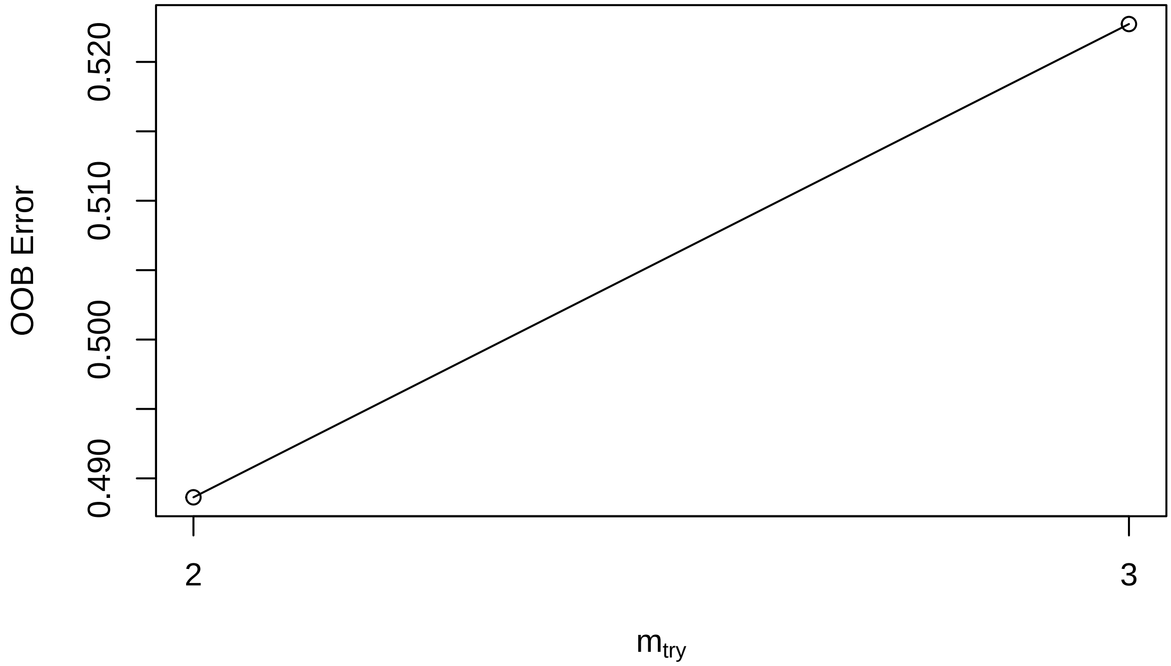


```
# prediction
rf_model_pred<- predict(rf_model2, newdata = test_data[,1:6])
1-(missclass_rate <- sum(rf_model_pred != test_data$RA_ITA_ref)/nrow(test_data))
```

```
## [1] 0.7727273
```

```
# Tune
set.seed(123)
opt_mtry <- tuneRF(train_data[,1:6], train_data[,8], stepFactor=1.5, improve=1e-5, ntree=500)
```

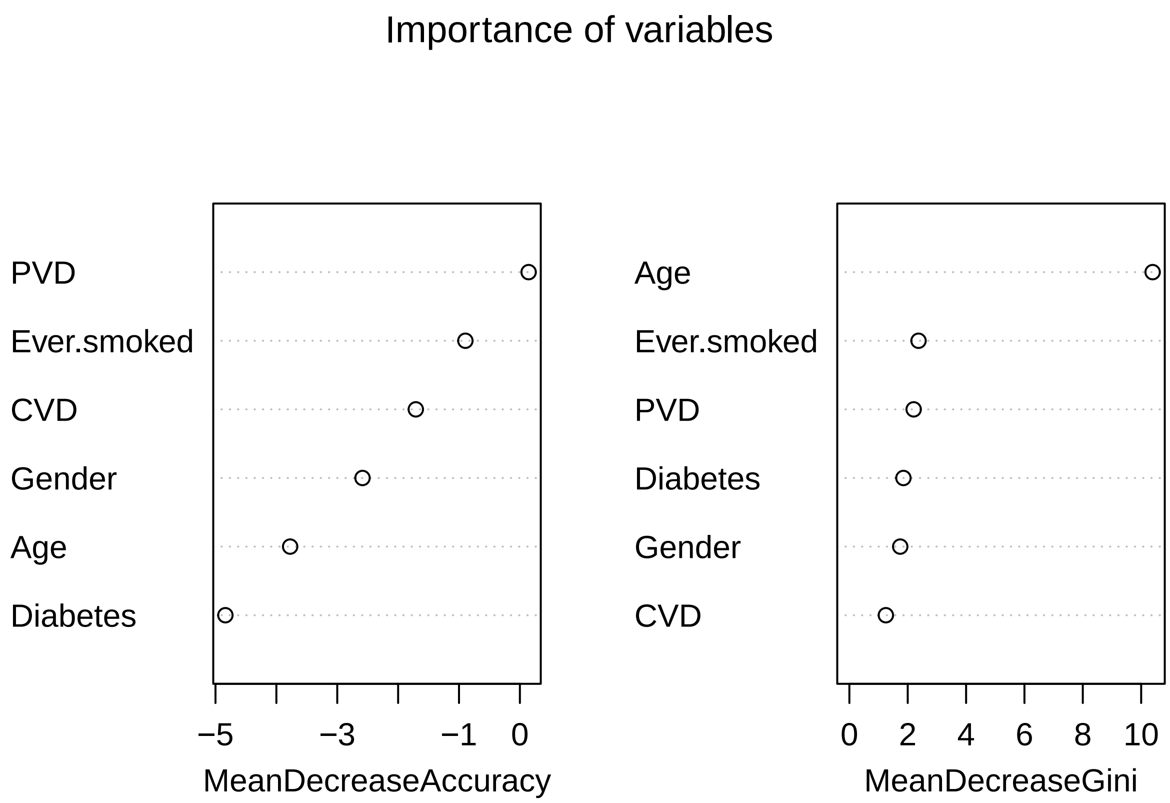
```
## mtry = 2  OOB error = 48.86%
## Searching left ...
## Searching right ...
## mtry = 3     OOB error = 52.27%
## -0.06976744 1e-05
```



```
print(opt_mtry)
```

```
##       mtry  OOBError
## 2.OOB    2 0.4886364
## 3.OOB    3 0.5227273
```

```
# model3 with setting the mtry
set.seed(12345)
rf_model3 <- randomForest(RA_ITA_ref~Age+Gender+Diabetes+Ever.smoked+PVD+CVD,
                          data = train_data, ntree = 500, mtry = 2, importance = TRUE)

# Importance of variables
varImpPlot(rf_model3, main = "Importance of variables")
```



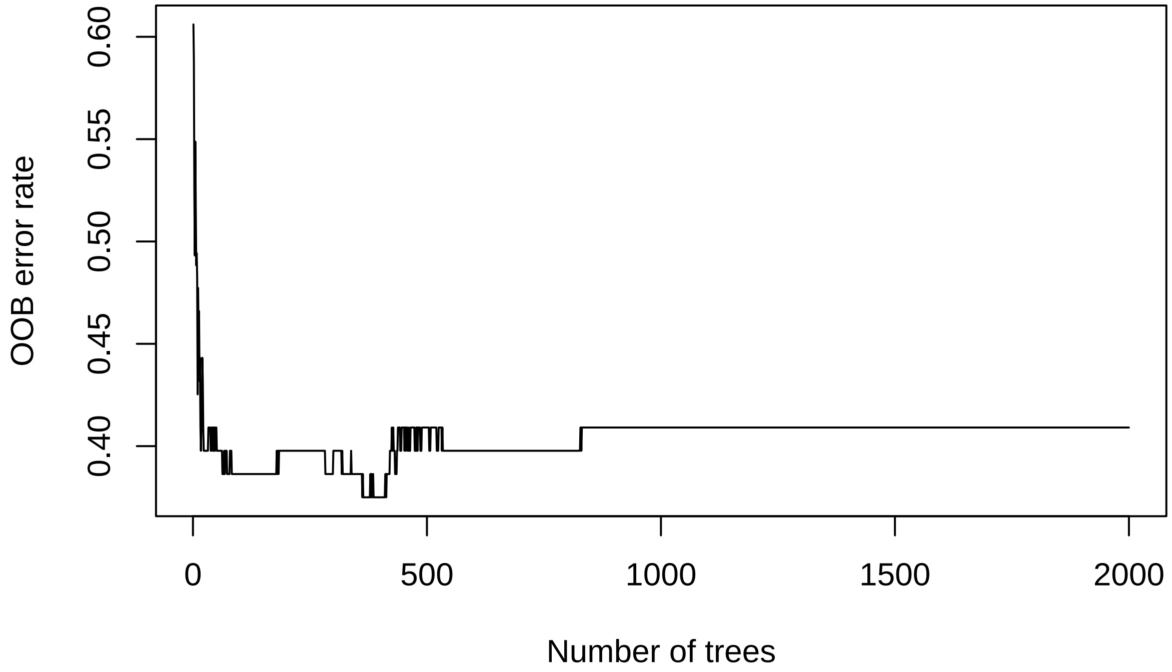
```
# prediction
rf_model_pred<- predict(rf_model3, newdata = test_data[,1:6])
1-(missclass_rate <- sum(rf_model_pred != test_data$RA_ITA_ref)/nrow(test_data))
```

```
## [1] 0.7727273
```

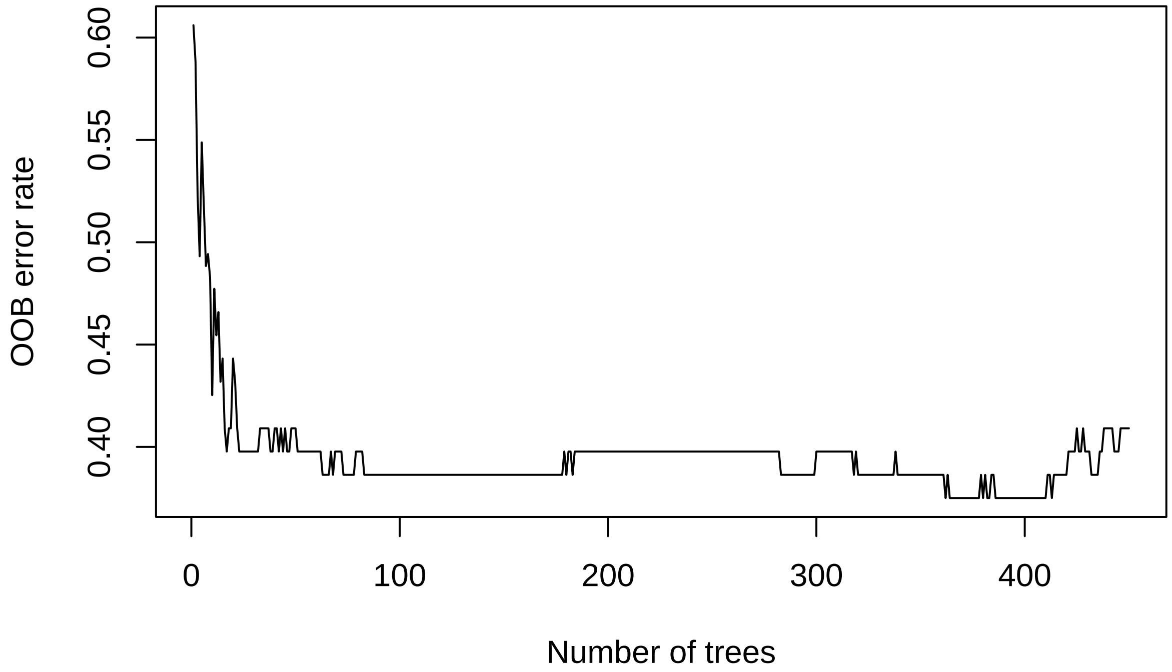
```
# one of the tree
getTree(rf_model3, 1, labelVar=TRUE)
```

```
##    left daughter right daughter   split var split point status   prediction
## 1              2              3         Age        54.0      1         <NA>
## 2              4              5 Ever.smoked         1.0      1         <NA>
## 3              6              7         CVD         1.0      1         <NA>
## 4              8              9         Age        47.5      1         <NA>
## 5              0              0        <NA>         0.0     -1       normal
## 6             10             11 Ever.smoked         1.0      1         <NA>
## 7             12             13 Ever.smoked         1.0      1         <NA>
## 8              0              0        <NA>         0.0     -1       normal
## 9              0              0        <NA>         0.0     -1  RAno_ITAyes
## 10            14             15         PVD         1.0      1         <NA>
## 11            16             17      Gender         1.0      1         <NA>
## 12             0              0        <NA>         0.0     -1  RAyes_ITAno
## 13             0              0        <NA>         0.0     -1  RAno_ITAyes
## 14            18             19      Gender         1.0      1         <NA>
## 15             0              0        <NA>         0.0     -1       normal
## 16             0              0        <NA>         0.0     -1  RAno_ITAyes
## 17            20             21    Diabetes         1.0      1         <NA>
## 18             0              0        <NA>         0.0     -1       normal
## 19            22             23         Age        75.5      1         <NA>
## 20            24             25         PVD         1.0      1         <NA>
## 21            26             27         Age        69.5      1        <NA>
## 22             0              0        <NA>         0.0     -1  RAno_ITAyes
## 23             0              0        <NA>         0.0     -1  RAno_ITAyes
## 24            28             29         Age        73.5      1        <NA>
## 25             0              0        <NA>         0.0     -1  RAno_ITAyes
## 26             0              0        <NA>         0.0     -1  RAno_ITAyes
## 27             0              0        <NA>         0.0     -1  RAno_ITAyes
## 28             0              0        <NA>         0.0     -1 RAyes_ITAyes
## 29            30             31         Age        75.5      1        <NA>
## 30             0              0        <NA>         0.0     -1       normal
## 31             0              0        <NA>         0.0     -1  RAno_ITAyes
```

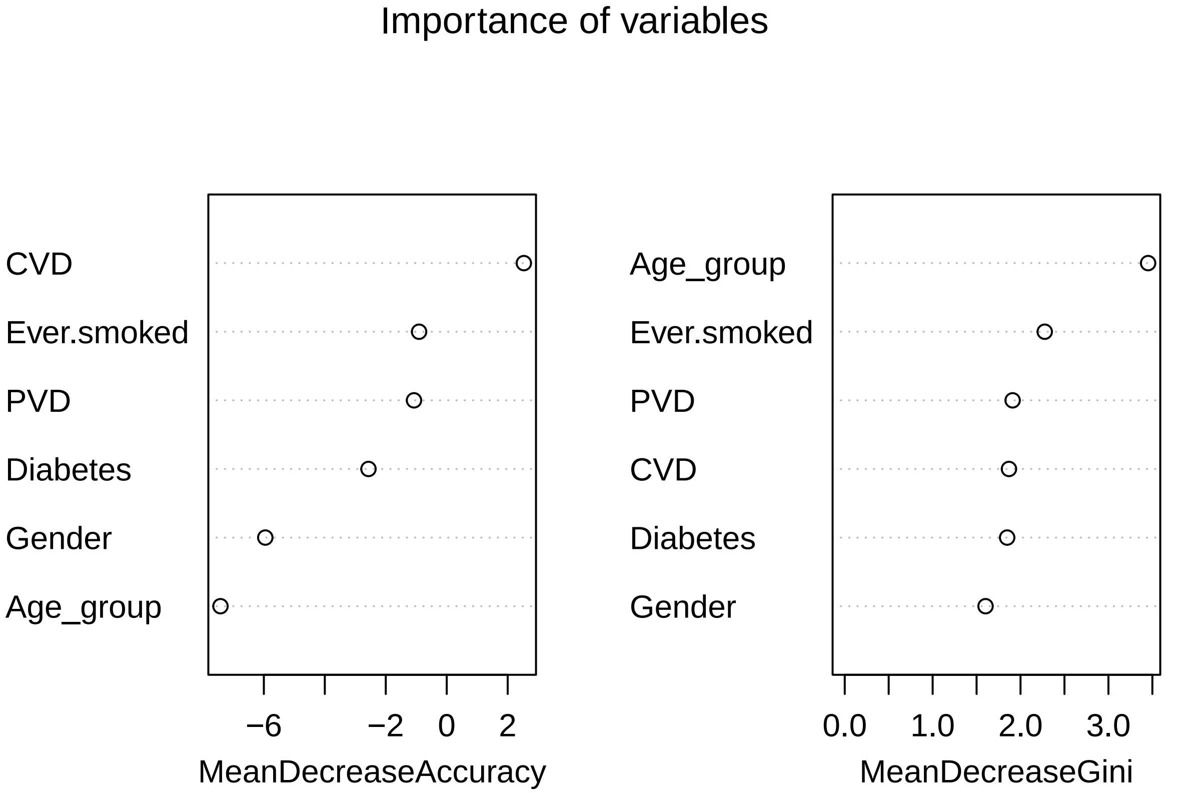
```
# when the age is categorical
s = sort(sample(nrow(artery_data5), nrow(artery_data5)*0.8))
train_data<-artery_data5[s,]
test_data<-artery_data5[-s,]

# model4
set.seed(12345)
rf_model4 <- randomForest(RA_ITA_ref~Age_group+Gender+Diabetes+Ever.smoked+PVD+CVD,
                          data = train_data, ntree = 2000, importance = TRUE)
plot(rf_model4$err.rate[,1], xlab = "Number of trees", ylab = "OOB error rate", type = "l")
```



```
# model5 with 450 trees
set.seed(12345)
rf_model5 <- randomForest(RA_ITA_ref~Age_group+Gender+Diabetes+Ever.smoked+PVD+CVD,
                          data = train_data, ntree = 450, importance = TRUE)
plot(rf_model5$err.rate[,1], xlab = "Number of trees", ylab = "OOB error rate", type = "l")
```



```
# Importance of variables
varImpPlot(rf_model5, main = "Importance of variables")
```



```
# prediction
rf_model_pred<- predict(rf_model5, newdata = test_data[,c(2:6,9)])
1-(missclass_rate <- sum(rf_model_pred != test_data$RA_ITA_ref)/nrow(test_data))
```

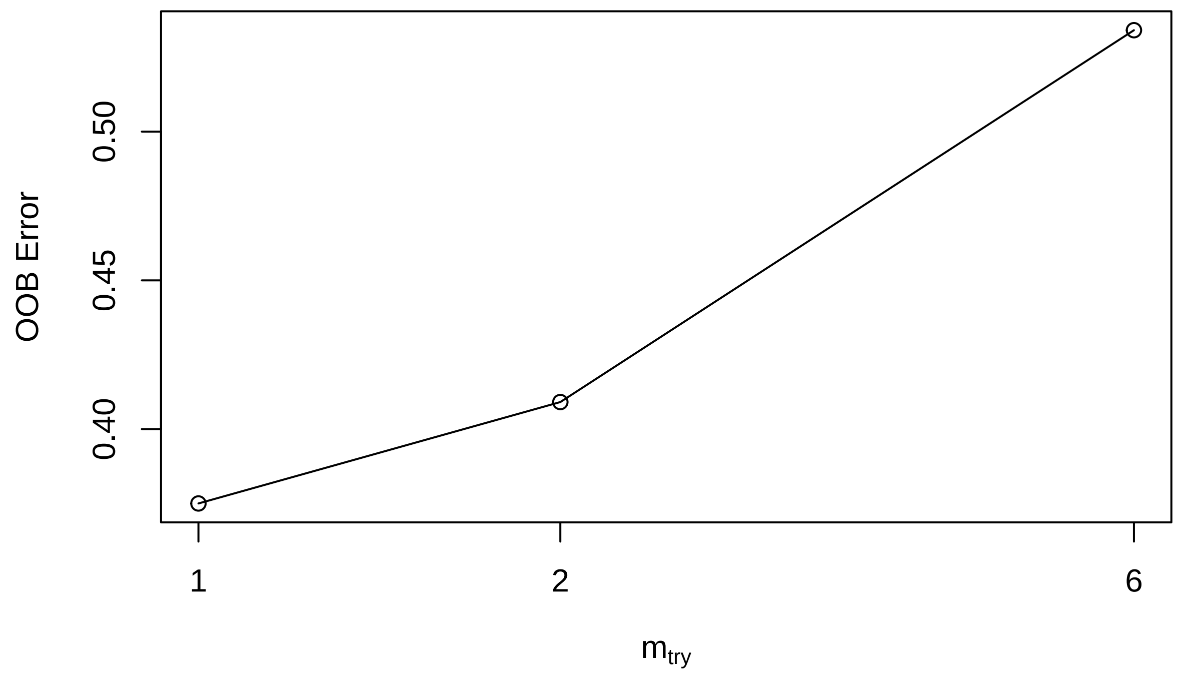
```
## [1] 0.5454545
```

```r
# one of the tree
getTree(rf_model5, 1, labelVar=TRUE)
```

```
##    left daughter right daughter   split var split point status  prediction
## 1              2              3   Age_group         3.5      1        <NA>
## 2              4              5   Age_group         1.5      1        <NA>
## 3              6              7         PVD         1.0      1        <NA>
## 4              8              9 Ever.smoked         1.0      1        <NA>
## 5             10             11      Gender         1.0      1        <NA>
## 6             12             13 Ever.smoked         1.0      1        <NA>
## 7             14             15 Ever.smoked         1.0      1        <NA>
## 8              0              0        <NA>         0.0     -1 RAno_ITAyes
## 9              0              0        <NA>         0.0     -1 RAno_ITAyes
## 10            16             17         CVD         1.0      1        <NA>
## 11            18             19 Ever.smoked         1.0      1        <NA>
## 12             0              0        <NA>         0.0     -1      normal
## 13            20             21    Diabetes         1.0      1        <NA>
## 14            22             23         CVD         1.0      1        <NA>
## 15             0              0        <NA>         0.0     -1 RAno_ITAyes
## 16            24             25 Ever.smoked         1.0      1        <NA>
## 17             0              0        <NA>         0.0     -1      normal
## 18            26             27    Diabetes         1.0      1        <NA>
## 19            28             29   Age_group         2.5      1        <NA>
## 20            30             31         CVD         1.0      1        <NA>
## 21             0              0        <NA>         0.0     -1 RAno_ITAyes
## 22             0              0        <NA>         0.0     -1 RAno_ITAyes
## 23             0              0        <NA>         0.0     -1 RAyes_ITAno
## 24             0              0        <NA>         0.0     -1      normal
## 25             0              0        <NA>         0.0     -1 RAno_ITAyes
## 26            32             33   Age_group         2.5      1        <NA>
## 27             0              0        <NA>         0.0     -1      normal
## 28            34             35         PVD         1.0      1        <NA>
## 29            36             37         CVD         1.0      1        <NA>
## 30             0              0        <NA>         0.0     -1 RAno_ITAyes
## 31             0              0        <NA>         0.0     -1 RAno_ITAyes
## 32             0              0        <NA>         0.0     -1 RAno_ITAyes
## 33             0              0        <NA>         0.0     -1 RAno_ITAyes
## 34             0              0        <NA>         0.0     -1 RAno_ITAyes
## 35             0              0        <NA>         0.0     -1 RAno_ITAyes
## 36             0              0        <NA>         0.0     -1 RAno_ITAyes
## 37             0              0        <NA>         0.0     -1 RAno_ITAyes
```

```r
# Tune
set.seed(123)
opt_mtry <- tuneRF(train_data[,1:6], train_data[,8], stepFactor=3, improve=1e-5, ntree=500)
```

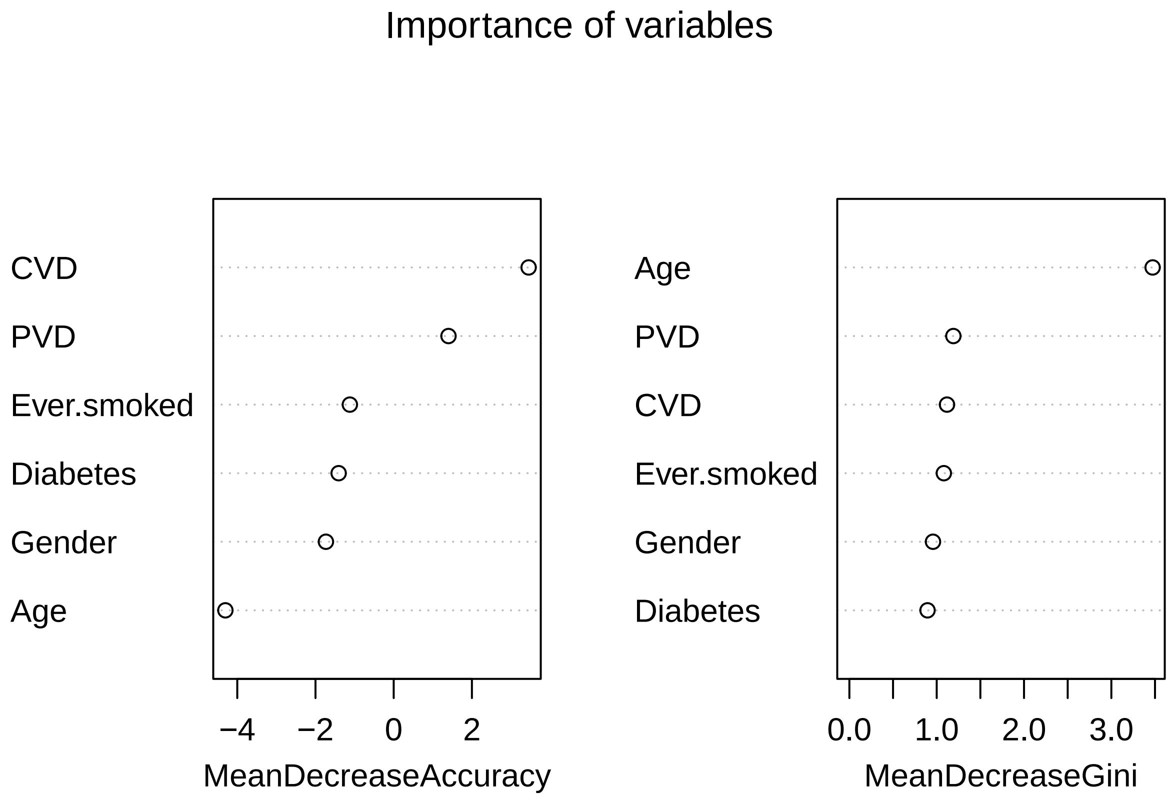
```
## mtry = 2  OOB error = 40.91%
## Searching left ...
## mtry = 1     OOB error = 37.5%
## 0.08333333 1e-05
## Searching right ...
## mtry = 6     OOB error = 53.41%
## -0.4242424 1e-05
```



```
print(opt_mtry)
```

```
##       mtry  OOBError
## 1.OOB    1 0.3750000
## 2.OOB    2 0.4090909
## 6.OOB    6 0.5340909
```

```
# model3 with setting the mtry
set.seed(12345)
rf_model6 <- randomForest(RA_ITA_ref~Age+Gender+Diabetes+Ever.smoked+PVD+CVD,
                          data = train_data, ntree = 500, mtry = 1, importance = TRUE)

# Importance of variables
varImpPlot(rf_model6, main = "Importance of variables")
```



```
# prediction
rf_model_pred<- predict(rf_model6, newdata = test_data[,1:6])
1-(missclass_rate <- sum(rf_model_pred != test_data$RA_ITA_ref)/nrow(test_data))
```

```
## [1] 0.5454545
```