Final Project

S M Mustaquim

2022-12-06

Introduction

In this project, we work on the body signal of smoking data available from Kaggle.

https://www.kaggle.com/datasets/kukuroo3/body-signal-of-smoking?datasetId=2157551

This dataset is a collection of basic health biological signal data. The objective is to determine the presence or absence of smoking through bio-signals.

A brief description of the variables

ID : index

gender

age : 5-years gap

height(cm)

weight(kg)

waist(cm) : Waist circumference length

eyesight(left)

eyesight(right)

hearing(left)

hearing(right)

systolic : Blood pressure

relaxation : Blood pressure

fasting blood sugar

Cholesterol : total

triglyceride

HDL : cholesterol type

LDL : cholesterol type

hemoglobin

Urine protein

serum creatinine

AST : glutamic oxaloacetic transaminase type

ALT : glutamic oxaloacetic transaminase type

Gtp : gamma-GTP

oral : Oral Examination status

dental caries

tartar : tartar status

smoking

On this classification issue, we intend to use some supervised and unsupervised methods.

For the supervised part, we want to apply some predictive modeling on the target variable smoking. First we apply the classification tools we learned in class, namely LDA, logistic regression, random forest, classification trees, and boosting to the training data. Then based on the results of the test data, we draw the ROC curve and calculate the C statistic or C index (area under the ROC curve) to compare the performance of these methods on this dataset.

Among all the classifiers, random forest produced the best AUC value and 'best' logistic regression produced the worst AUC value. The others did a fairly good job in predicting the smoking status.

For the unsupervised portion, we ignore the target variable smoking and use the MDS and t-SNE techniques on the training sample, plotting the results by specifying smokers and non-smokers to see if they can cluster them appropriately.

Both of these methods failed to cluster smokers and nonsmokers.

Data Preparation and Cleaning

First we read the data.

```r
setwd("C:/Users/smustaquim/Documents/Data Mining/Final Project")
data <- read.csv(file="smoking.csv",
                 header = TRUE,  na.string = c("", " ",      "NA", "?"))
dim(data)
```

```
## [1] 55692    27
```

First we check if there are any missing values and impute them accordingly.

```r
colMeans(is.na(data))*100
```

```
##                  ID              gender                 age          height.cm. 
##                   0                   0                   0                   0 
##          weight.kg.           waist.cm.      eyesight.left.     eyesight.right. 
##                   0                   0                   0                   0 
##       hearing.left.      hearing.right.            systolic          relaxation 
##                   0                   0                   0                   0 
## fasting.blood.sugar         Cholesterol        triglyceride                 HDL 
##                   0                   0                   0                   0 
##                 LDL          hemoglobin       Urine.protein    serum.creatinine 
##                   0                   0                   0                   0 
##                 AST                 ALT                 Gtp                oral 
##                   0                   0                   0                   0 
##       dental.caries              tartar             smoking 
##                   0                   0                   0
```

There are no missing values.

Exploratory Data Analysis

We take a look at all the variables in the dataset.

```
str(data)
```

```
## ’data.frame’:    55692 obs. of  27 variables:
##  $ ID                 : int  0 1 2 3 4 5 6 7 9 10 ...
##  $ gender             : chr  "F" "F" "M" "M" ...
##  $ age                : int  40 40 55 40 40 30 40 45 50 45 ...
##  $ height.cm.         : int  155 160 170 165 155 180 160 165 150 175 ...
##  $ weight.kg.         : int  60 60 60 70 60 75 60 90 60 75 ...
##  $ waist.cm.          : num  81.3 81 80 88 86 85 85.5 96 85 89 ...
##  $ eyesight.left.     : num  1.2 0.8 0.8 1.5 1 1.2 1 1.2 0.7 1 ...
##  $ eyesight.right.    : num  1 0.6 0.8 1.5 1 1.2 1 1 0.8 1 ...
##  $ hearing.left.      : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ hearing.right.     : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ systolic           : num  114 119 138 100 120 128 116 153 115 113 ...
##  $ relaxation         : num  73 70 86 60 74 76 82 96 74 64 ...
##  $ fasting.blood.sugar: num  94 130 89 96 80 95 94 158 86 94 ...
##  $ Cholesterol        : num  215 192 242 322 184 217 226 222 210 198 ...
##  $ triglyceride       : num  82 115 182 254 74 199 68 269 66 147 ...
##  $ HDL                : num  73 42 55 45 62 48 55 34 48 43 ...
##  $ LDL                : num  126 127 151 226 107 129 157 134 149 126 ...
##  $ hemoglobin         : num  12.9 12.7 15.8 14.7 12.5 16.2 17 15 13.7 16 ...
##  $ Urine.protein      : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ serum.creatinine   : num  0.7 0.6 1 1 0.6 1.2 0.7 1.3 0.8 0.8 ...
##  $ AST                : num  18 22 21 19 16 18 21 38 31 26 ...
##  $ ALT                : num  19 19 16 26 14 27 27 71 31 24 ...
##  $ Gtp                : num  27 18 22 18 22 33 39 111 14 63 ...
##  $ oral               : chr  "Y" "Y" "Y" "Y" ...
##  $ dental.caries      : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ tartar             : chr  "Y" "Y" "N" "Y" ...
##  $ smoking            : int  0 0 1 0 0 0 1 0 0 0 ...
```

As we can see, gender and tartar are binary categorical variables.

```
unique(data$oral)
```

```
## [1] "Y"
```

Variable oral is unary and variable ID is not predictive. So, we will remove both of them for ensuing analysis.

```
data1 <- data[-c(1,24)]
```

Also, hearing.left. , hearing.right. , and dental.caries are binary variables and every other predictor is either continuous or integer count.

The target variable smoking is binary categorical, with ‘0’ meaning non-smoker and ‘1’ meaning smoker. Let us check the proportion of smokers in the sample.

```
proportion <- sum(data$smoking == 1)/nrow(data)
proportion
```

```
## [1] 0.3672879
```

This seems like the actual prevalence of smokers in the overall population.

Data Partitioning

We partition the data into train and test set with a ratio of 3:1.

```
set.seed(10)

dt = sample(nrow(data1), nrow(data1)*.75)
train = as.data.frame(data1[dt,])
test = as.data.frame(data1[-dt,])
```

We are ready to perform supervised and unsupervised methods to our dataset.

Supervised Learning

We try out the following predictive modeling tools. For each method, use the training set to identify the best model and apply the model to the test set. Then plot the ROC curve and compute the C statistic or C index (area under the ROC curve), all based on the test set performance.

Linear Discriminant Analysis

```
terms <- (names(train))[-(ncol(train))]
formula <- as.formula(paste(c("smoking ~ ", terms), collapse="+"))
formula
```

```
## smoking ~ +gender + age + height.cm. + weight.kg. + waist.cm. +
##     eyesight.left. + eyesight.right. + hearing.left. + hearing.right. +
##     systolic + relaxation + fasting.blood.sugar + Cholesterol +
##     triglyceride + HDL + LDL + hemoglobin + Urine.protein + serum.creatinine +
##     AST + ALT + Gtp + dental.caries + tartar
```

```
library(cvAUC)
library(verification)
library(MASS)
fit.LDA <- lda(formula, data=train)
yhat.LDA <- predict(fit.LDA, newdata = test)$x
```

Now we find the AUC.

```
library(cvAUC)
n <- NROW(test)
yobs <- test$smoking
AUC.LDA <- ci.cvAUC(predictions = yhat.LDA, labels=yobs, confidence=0.95)
AUC.LDA
```

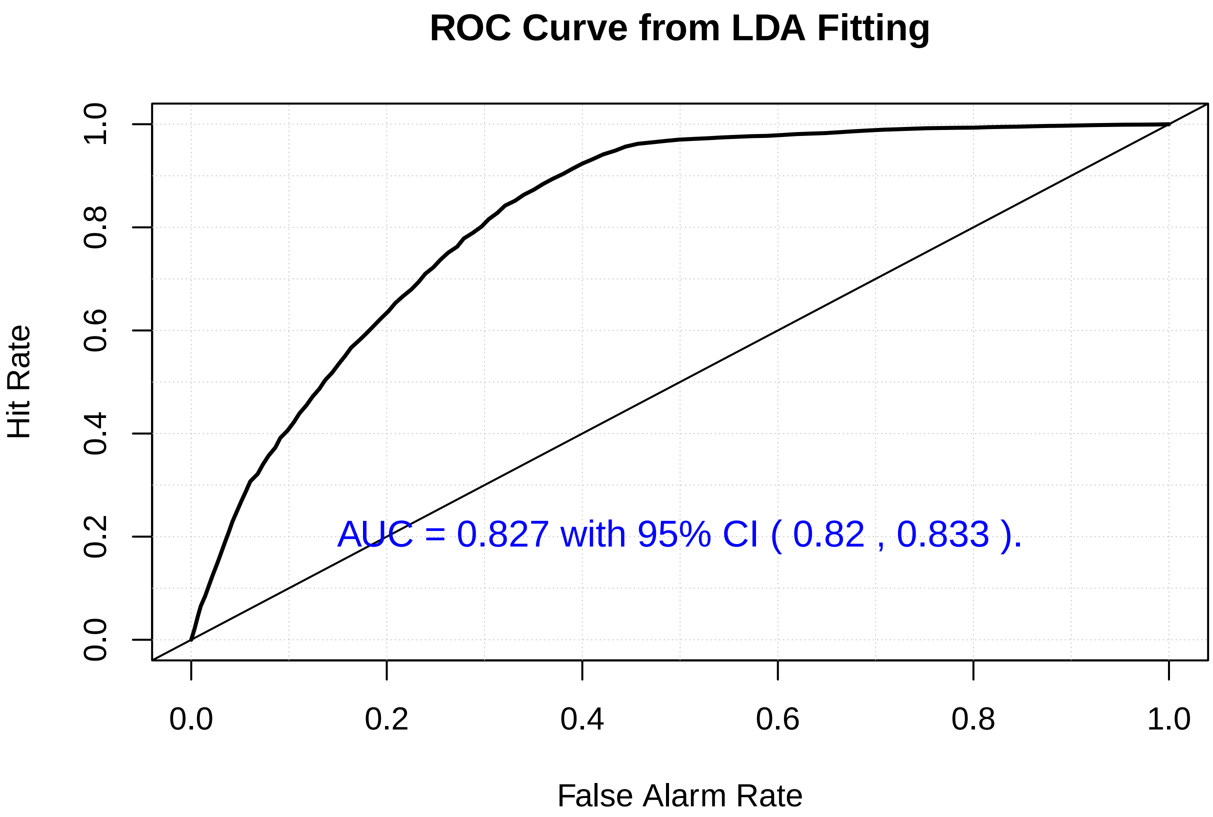
```
## $cvAUC
## [1] 0.8265198
##
## $se
## [1] 0.003416727
##
## $ci
## [1] 0.8198231 0.8332165
##
## $confidence
## [1] 0.95
```

Now, we plot the ROC Curve.

```
library(verification)
yhat <- scale(yhat.LDA, center=min(yhat.LDA), scale = max(yhat.LDA)-min(yhat.LDA))
mod.glm <- verify(obs=yobs, pred=yhat)
```

```
## If baseline is not included, baseline values  will be calculated from the  sample obs.
```

```
roc.plot(mod.glm, plot.thres = NULL, main="ROC Curve from LDA Fitting")
text(x=0.5, y=0.2, paste("AUC =", round(AUC.LDA$cvAUC, digits=3),
                         "with 95% CI (", round(AUC.LDA$ci, digits=3)[1], ",", round(AUC.LDA$ci, digits=
                         sep=" "), col="blue", cex=1.2)
```



Best Logistic Regression Model via Stepwise Selection

```
fit.full <- glm(formula, family=binomial, data = train)
summary(fit.full)
```

```
##
## Call:
## glm(formula = formula, family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6854  -0.9546  -0.2561   0.9373   4.4451
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -6.6541291  0.4950197 -13.442  < 2e-16 ***
## genderM             2.9321758  0.0591589  49.564  < 2e-16 ***
## age                -0.0006643  0.0012597  -0.527 0.597944
## height.cm.          0.0228437  0.0025606   8.921  < 2e-16 ***
## weight.kg.         -0.0104800  0.0024876  -4.213 2.52e-05 ***
## waist.cm.          -0.0010317  0.0029662  -0.348 0.727980
## eyesight.left.     -0.0239295  0.0254190  -0.941 0.346499
## eyesight.right.    -0.0188501  0.0271306  -0.695 0.487187
## hearing.left.       -0.2602658  0.0949128  -2.742 0.006104 ** 
## hearing.right.      -0.0089539  0.0922978  -0.097 0.922717    
## systolic            -0.0141323  0.0014528  -9.728  < 2e-16 ***
## relaxation           0.0086925  0.0019945   4.358 1.31e-05 ***
## fasting.blood.sugar  0.0039047  0.0006376   6.124 9.14e-10 ***
## Cholesterol         -0.0023798  0.0006242  -3.813 0.000137 ***
## triglyceride         0.0047492  0.0002347  20.237  < 2e-16 ***
## HDL                  0.0021686  0.0011687   1.856 0.063524 .  
## LDL                 -0.0001343  0.0004998  -0.269 0.788117    
## hemoglobin           0.1381994  0.0125660  10.998  < 2e-16 ***
## Urine.protein        0.0266593  0.0314465   0.848 0.396567    
## serum.creatinine    -0.9794953  0.0778032 -12.589  < 2e-16 ***
## AST                 -0.0007512  0.0011386  -0.660 0.509382    
## ALT                 -0.0054841  0.0008957  -6.123 9.19e-10 ***
## Gtp                  0.0072707  0.0003825  19.010  < 2e-16 ***
## dental.caries        0.3206131  0.0297222  10.787  < 2e-16 ***
## tartarY              0.3501379  0.0252312  13.877  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 54993  on 41768  degrees of freedom
## Residual deviance: 39438  on 41744  degrees of freedom
## AIC: 39488
## 
## Number of Fisher Scoring iterations: 5
```

```
library(MASS)
fit <- stepAIC(fit.full, direction = "both", k=log(nrow(train)))
```

```
## Start:  AIC=39703.88
## smoking ~ +gender + age + height.cm. + weight.kg. + waist.cm. + 
##     eyesight.left. + eyesight.right. + hearing.left. + hearing.right. + 
##     systolic + relaxation + fasting.blood.sugar + Cholesterol + 
##     triglyceride + HDL + LDL + hemoglobin + Urine.protein + serum.creatinine + 
##     AST + ALT + Gtp + dental.caries + tartar
## 
##                       Df Deviance   AIC
## - hearing.right.       1    39438 39693
## - LDL                  1    39438 39693
## - waist.cm.            1    39438 39693
## - age                  1    39438 39694
## - AST                  1    39438 39694
## - eyesight.right.      1    39438 39694
## - Urine.protein        1    39439 39694
## - eyesight.left.       1    39439 39694
## - HDL                  1    39441 39697
## - hearing.left.        1    39445 39701
## <none>                      39438 39704
## - Cholesterol          1    39451 39707
## - weight.kg.           1    39456 39711
## - relaxation           1    39457 39712
## - ALT                  1    39475 39730
## - fasting.blood.sugar  1    39477 39732
## - height.cm.           1    39518 39773
## - systolic             1    39533 39788
## - dental.caries        1    39555 39810
## - hemoglobin           1    39561 39816
## - serum.creatinine     1    39605 39860
## - tartar               1    39631 39886
## - triglyceride         1    39749 40005
## - Gtp                  1    39890 40145
## - gender               1    42385 42641
##
## Step:  AIC=39693.25
## smoking ~ gender + age + height.cm. + weight.kg. + waist.cm. +
##     eyesight.left. + eyesight.right. + hearing.left. + systolic +
##     relaxation + fasting.blood.sugar + Cholesterol + triglyceride +
##     HDL + LDL + hemoglobin + Urine.protein + serum.creatinine +
##     AST + ALT + Gtp + dental.caries + tartar
##
##                       Df Deviance   AIC
## - LDL                  1    39438 39683
## - waist.cm.            1    39438 39683
## - age                  1    39438 39683
## - AST                  1    39438 39683
## - eyesight.right.      1    39438 39683
## - Urine.protein        1    39439 39683
## - eyesight.left.       1    39439 39683
## - HDL                  1    39441 39686
## - hearing.left.        1    39448 39693
## <none>                      39438 39693
## - Cholesterol          1    39452 39696
## - weight.kg.           1    39456 39700
## - relaxation           1    39457 39702
## + hearing.right.       1    39438 39704
## - ALT                  1    39475 39720
## - fasting.blood.sugar  1    39477 39721
## - height.cm.           1    39518 39762
## - systolic             1    39533 39778
## - dental.caries        1    39555 39799
## - hemoglobin           1    39561 39806
## - serum.creatinine     1    39605 39849
## - tartar               1    39631 39876
## - triglyceride         1    39749 39994
## - Gtp                  1    39890 40135
## - gender               1    42386 42631
##
## Step:  AIC=39682.68
## smoking ~ gender + age + height.cm. + weight.kg. + waist.cm. +
##     eyesight.left. + eyesight.right. + hearing.left. + systolic +
##     relaxation + fasting.blood.sugar + Cholesterol + triglyceride +
##     HDL + hemoglobin + Urine.protein + serum.creatinine + AST +
##     ALT + Gtp + dental.caries + tartar
##
##                       Df Deviance   AIC
## - waist.cm.            1    39438 39672
## - age                  1    39438 39672
## - AST                  1    39438 39672
## - eyesight.right.      1    39438 39673
## - Urine.protein        1    39439 39673
## - eyesight.left.       1    39439 39673
## - HDL                  1    39443 39677
## - hearing.left.        1    39448 39682
## <none>                      39438 39683
## - weight.kg.           1    39456 39690
## - relaxation           1    39457 39691
## + LDL                  1    39438 39693
## + hearing.right.       1    39438 39693
## - ALT                  1    39475 39709
## - fasting.blood.sugar  1    39477 39711
## - Cholesterol          1    39481 39716
## - height.cm.           1    39518 39752
## - systolic             1    39533 39767
## - dental.caries        1    39555 39789
## - hemoglobin           1    39561 39795
## - serum.creatinine     1    39605 39839
## - tartar               1    39631 39865
## - Gtp                  1    39890 40124
## - triglyceride         1    39954 40188
## - gender               1    42386 42620
##
## Step:  AIC=39672.16
## smoking ~ gender + age + height.cm. + weight.kg. + eyesight.left. +
##     eyesight.right. + hearing.left. + systolic + relaxation +
##     fasting.blood.sugar + Cholesterol + triglyceride + HDL +
##     hemoglobin + Urine.protein + serum.creatinine + AST + ALT +
##     Gtp + dental.caries + tartar
##
##                       Df Deviance   AIC
## - AST                  1    39439 39662
## - age                  1    39439 39662
## - eyesight.right.      1    39439 39662
## - Urine.protein        1    39439 39662
## - eyesight.left.       1    39439 39662
## - HDL                  1    39443 39666
## - hearing.left.        1    39448 39672
## <none>                      39438 39672
## - relaxation           1    39457 39681
## + waist.cm.            1    39438 39683
## + LDL                  1    39438 39683
## + hearing.right.       1    39438 39683
## - ALT                  1    39476 39699
## - fasting.blood.sugar  1    39477 39700
## - Cholesterol          1    39482 39705
## - weight.kg.           1    39496 39719
## - height.cm.           1    39528 39751
## - systolic             1    39534 39757
## - dental.caries        1    39555 39778
## - hemoglobin           1    39561 39784
## - serum.creatinine     1    39605 39828
## - tartar              1    39631 39855
## - Gtp                 1    39891 40115
## - triglyceride        1    39954 40177
## - gender              1    42403 42626
##
## Step:  AIC=39661.94
## smoking ~ gender + age + height.cm. + weight.kg. + eyesight.left. +
##     eyesight.right. + hearing.left. + systolic + relaxation +
##     fasting.blood.sugar + Cholesterol + triglyceride + HDL +
##     hemoglobin + Urine.protein + serum.creatinine + ALT + Gtp +
##     dental.caries + tartar
##
##                       Df Deviance   AIC
## - eyesight.right.      1    39439 39652
## - age                  1    39439 39652
## - Urine.protein        1    39439 39652
## - eyesight.left.       1    39439 39652
## - HDL                  1    39443 39656
## - hearing.left.        1    39449 39661
## <none>                      39439 39662
## - relaxation           1    39458 39670
## + AST                  1    39438 39672
## + waist.cm.            1    39438 39672
## + LDL                  1    39438 39673
## + hearing.right.       1    39438 39673
## - fasting.blood.sugar  1    39477 39690
## - Cholesterol          1    39482 39695
## - weight.kg.           1    39496 39708
## - height.cm.           1    39528 39741
## - ALT                  1    39531 39744
## - systolic             1    39534 39747
## - dental.caries        1    39555 39768
## - hemoglobin           1    39562 39775
## - serum.creatinine     1    39606 39819
## - tartar               1    39632 39845
## - Gtp                  1    39895 40107
## - triglyceride         1    39956 40169
## - gender               1    42403 42616
##
## Step:  AIC=39651.77
## smoking ~ gender + age + height.cm. + weight.kg. + eyesight.left. +
##     hearing.left. + systolic + relaxation + fasting.blood.sugar +
##     Cholesterol + triglyceride + HDL + hemoglobin + Urine.protein +
##     serum.creatinine + ALT + Gtp + dental.caries + tartar
##
##                       Df Deviance   AIC
## - age                  1    39439 39642
## - Urine.protein        1    39440 39642
## - eyesight.left.       1    39440 39643
## - HDL                  1    39444 39646
## - hearing.left.        1    39449 39651
## <none>                      39439 39652
## - relaxation           1    39458 39660
## + eyesight.right.      1    39439 39662
## + AST                  1    39439 39662
## + waist.cm.            1    39439 39662
## + LDL                  1    39439 39662
## + hearing.right.       1    39439 39662
## - fasting.blood.sugar  1    39478 39680
## - Cholesterol          1    39482 39685
## - weight.kg.           1    39496 39698
## - height.cm.           1    39528 39730
## - ALT                  1    39532 39734
## - systolic             1    39535 39737
## - dental.caries        1    39556 39758
## - hemoglobin           1    39562 39764
## - serum.creatinine     1    39606 39808
## - tartar               1    39633 39835
## - Gtp                  1    39895 40097
## - triglyceride         1    39956 40158
## - gender               1    42403 42605
##
## Step:  AIC=39641.61
## smoking ~ gender + height.cm. + weight.kg. + eyesight.left. +
##     hearing.left. + systolic + relaxation + fasting.blood.sugar +
##     Cholesterol + triglyceride + HDL + hemoglobin + Urine.protein +
##     serum.creatinine + ALT + Gtp + dental.caries + tartar
##
##                       Df Deviance   AIC
## - Urine.protein        1    39440 39632
## - eyesight.left.       1    39441 39632
## - HDL                  1    39444 39636
## <none>                      39439 39642
## - hearing.left.        1    39451 39642
## - relaxation           1    39458 39650
## + AST                  1    39439 39652
## + age                  1    39439 39652
## + eyesight.right.      1    39439 39652
## + waist.cm.            1    39439 39652
## + LDL                  1    39439 39652
## + hearing.right.       1    39439 39652
## - fasting.blood.sugar  1    39478 39670
## - Cholesterol          1    39483 39675
## - weight.kg.           1    39496 39688
## - ALT                  1    39532 39723
## - systolic             1    39538 39729
## - height.cm.           1    39540 39732
## - dental.caries        1    39559 39750
## - hemoglobin           1    39569 39761
## - serum.creatinine     1    39609 39800
## - tartar               1    39634 39826
## - Gtp                  1    39896 40088
## - triglyceride         1    39957 40148
## - gender               1    42418 42609
##
## Step:  AIC=39631.65
## smoking ~ gender + height.cm. + weight.kg. + eyesight.left. +
##     hearing.left. + systolic + relaxation + fasting.blood.sugar +
##     Cholesterol + triglyceride + HDL + hemoglobin + serum.creatinine +
##     ALT + Gtp + dental.caries + tartar
##
##                       Df Deviance   AIC
## - eyesight.left.       1    39442 39622
## - HDL                  1    39445 39626
## <none>                      39440 39632
## - hearing.left.        1    39451 39632
## - relaxation           1    39459 39640
## + Urine.protein        1    39439 39642
## + AST                  1    39440 39642
## + age                  1    39440 39642
## + eyesight.right.      1    39440 39642
## + waist.cm.            1    39440 39642
## + LDL                  1    39440 39642
## + hearing.right.       1    39440 39642
## - fasting.blood.sugar  1    39480 39661
## - Cholesterol          1    39484 39665
## - weight.kg.           1    39497 39678
## - ALT                  1    39532 39713
## - systolic             1    39538 39719
## - height.cm.           1    39541 39722
## - dental.caries        1    39560 39740
## - hemoglobin           1    39570 39751
## - serum.creatinine     1    39609 39790
## - tartar               1    39635 39815
## - Gtp                  1    39899 40080
## - triglyceride         1    39957 40138
## - gender               1    42419 42599
##
## Step:  AIC=39622.38
## smoking ~ gender + height.cm. + weight.kg. + hearing.left. +
##     systolic + relaxation + fasting.blood.sugar + Cholesterol +
##     triglyceride + HDL + hemoglobin + serum.creatinine + ALT +
##     Gtp + dental.caries + tartar
##
##                       Df Deviance   AIC
## - HDL                  1    39446 39617
## <none>                      39442 39622
## - hearing.left.        1    39452 39623
## - relaxation           1    39460 39631
## + eyesight.left.       1    39440 39632
## + eyesight.right.      1    39441 39632
## + Urine.protein        1    39441 39632
## + AST                  1    39441 39633
## + age                  1    39441 39633
## + waist.cm.            1    39441 39633
## + LDL                  1    39441 39633
## + hearing.right.       1    39441 39633
## - fasting.blood.sugar  1    39482 39652
## - Cholesterol          1    39486 39656
## - weight.kg.           1    39499 39669
## - ALT                  1    39534 39704
## - systolic             1    39539 39709
## - height.cm.           1    39541 39712
## - dental.caries        1    39561 39731
## - hemoglobin           1    39571 39742
## - serum.creatinine     1    39610 39781
## - tartar               1    39636 39806
## - Gtp                  1    39902 40072
## - triglyceride         1    39958 40128
## - gender               1    42419 42589
##
## Step:  AIC=39616.61
## smoking ~ gender + height.cm. + weight.kg. + hearing.left. +
##     systolic + relaxation + fasting.blood.sugar + Cholesterol +
##     triglyceride + hemoglobin + serum.creatinine + ALT + Gtp +
##     dental.caries + tartar
##
##                       Df Deviance   AIC
## <none>                     39446 39617
## - hearing.left.        1    39458 39617
## + HDL                  1    39442 39622
## - relaxation           1    39465 39625
## + LDL                  1    39445 39626
## + eyesight.left.       1    39445 39626
## + eyesight.right.      1    39445 39626
## + Urine.protein        1    39446 39627
## + age                  1    39446 39627
## + waist.cm.            1    39446 39627
## + AST                  1    39446 39627
## + hearing.right.       1    39446 39627
## - Cholesterol          1    39486 39646
## - fasting.blood.sugar  1    39487 39646
## - weight.kg.           1    39514 39673
## - systolic             1    39542 39702
## - ALT                  1    39543 39702
## - height.cm.           1    39554 39713
## - dental.caries        1    39566 39726
## - hemoglobin           1    39576 39735
## - serum.creatinine     1    39617 39777
## - tartar               1    39639 39799
## - Gtp                  1    39938 40097
## - triglyceride         1    40018 40178
## - gender               1    42434 42594
```

So, the variables have been selected.

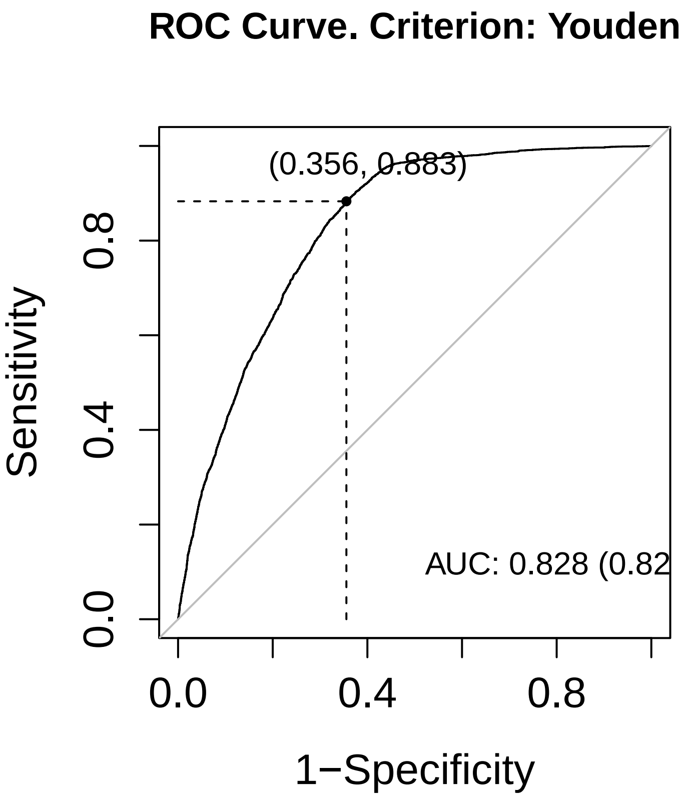
```
fit.step <- glm(smoking ~ gender + height.cm. + weight.kg. + hearing.left. +
                  systolic + relaxation + fasting.blood.sugar + Cholesterol +
                  triglyceride + hemoglobin + serum.creatinine + ALT + Gtp +
                  dental.caries + tartar, family = binomial, data = train)

summary(fit.step)
```

```
##
## Call:
```

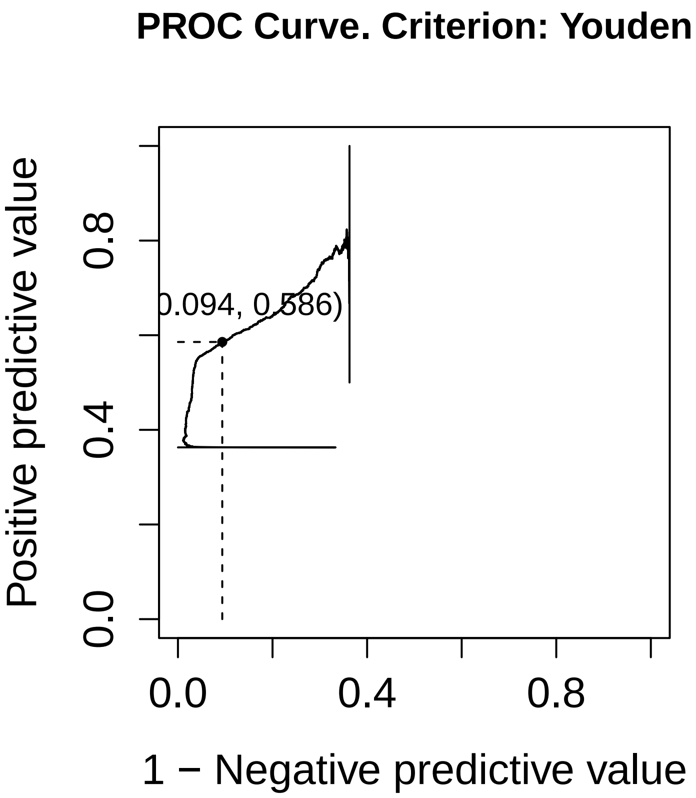
```
## glm(formula = smoking ~ gender + height.cm. + weight.kg. + hearing.left. +
##     systolic + relaxation + fasting.blood.sugar + Cholesterol +
##     triglyceride + hemoglobin + serum.creatinine + ALT + Gtp +
##     dental.caries + tartar, family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.7387  -0.9535  -0.2557   0.9373   4.4276
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -6.8011090  0.4149192 -16.391  < 2e-16 ***
## genderM              2.9099143  0.0583579  49.863  < 2e-16 ***
## height.cm.           0.0240674  0.0023262  10.346  < 2e-16 ***
## weight.kg.          -0.0117118  0.0014268  -8.208 2.24e-16 ***
## hearing.left.       -0.2747379  0.0824394  -3.333  0.00086 ***
## systolic            -0.0140511  0.0014407  -9.753  < 2e-16 ***
## relaxation           0.0086650  0.0019927   4.348 1.37e-05 ***
## fasting.blood.sugar  0.0038941  0.0006215   6.266 3.71e-10 ***
## Cholesterol         -0.0022941  0.0003651  -6.283 3.32e-10 ***
## triglyceride         0.0045811  0.0001952  23.467  < 2e-16 ***
## hemoglobin           0.1395035  0.0123624  11.285  < 2e-16 ***
## serum.creatinine    -0.9848568  0.0773295 -12.736  < 2e-16 ***
## ALT                 -0.0059902  0.0006444  -9.296  < 2e-16 ***
## Gtp                  0.0073715  0.0003745  19.685  < 2e-16 ***
## dental.caries        0.3228461  0.0295834  10.913  < 2e-16 ***
## tartarY              0.3494777  0.0251864  13.876  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 54993  on 41768  degrees of freedom
## Residual deviance: 39446  on 41753  degrees of freedom
## AIC: 39478
##
## Number of Fisher Scoring iterations: 5
```

```r
pred.step <- predict(fit.step, newdata=test)
```

```r
library(OptimalCutpoints)
dat.tmp <- data.frame(pred=pred.step, y = yobs)
result <- optimal.cutpoints(pred~y, data=dat.tmp, tag.healthy=0,
                            methods = "Youden", control=control.cutpoints())
plot(result)
```



```
## Press return for next page....
```



```
bcut <- result$Youden$Global$optimal.cutoff$cutoff; bcut
```

```
## [1] -0.3742594
```

```
yhat.step <- sign(pred.step >= bcut)
```

Now, we plot the ROC curve.

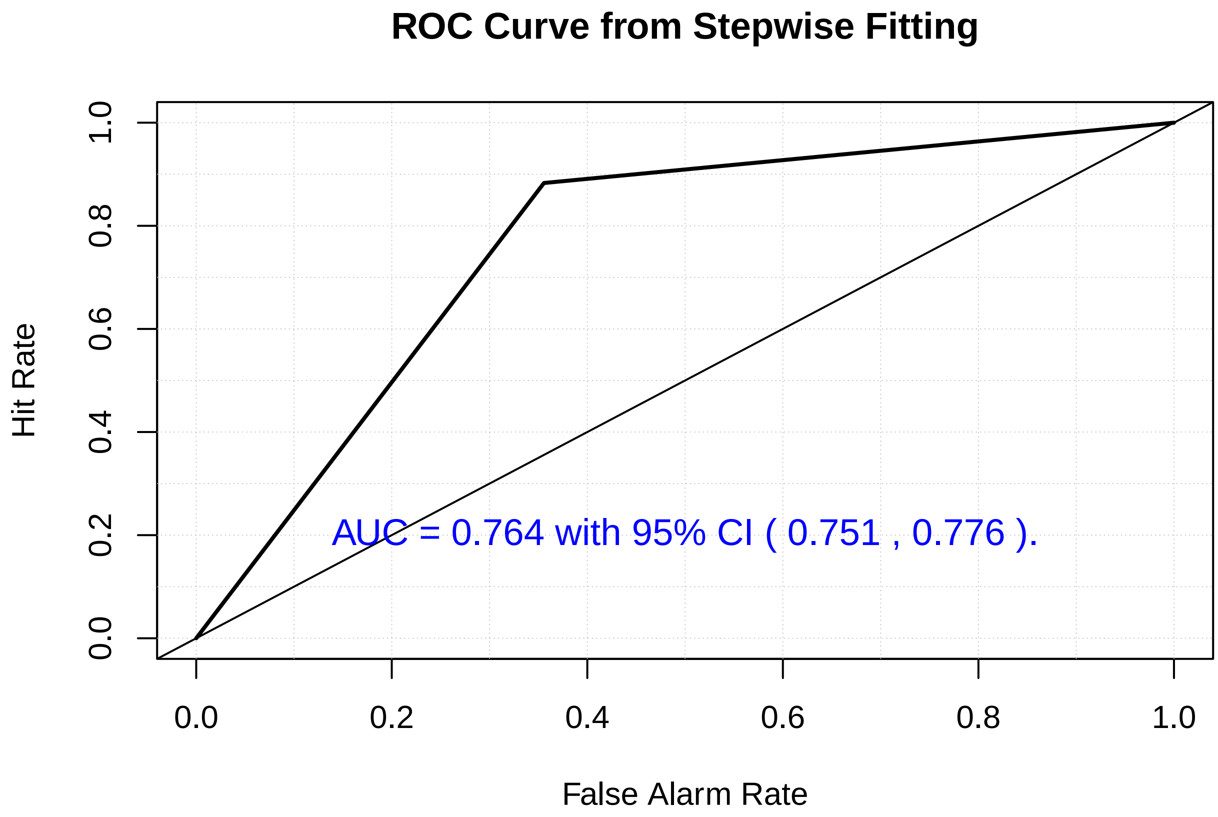
```
AUC.step <- ci.cvAUC(predictions=yhat.step, labels=yobs, confidence=0.95)
AUC.step
```

```
## $cvAUC
## [1] 0.7636817
##
## $se
## [1] 0.00635689
##
## $ci
## [1] 0.7512224 0.7761410
##
## $confidence
## [1] 0.95
```

```
yhat <- scale(yhat.step, center=min(yhat.step), scale = max(yhat.step)-min(yhat.step))
mod.glm <- verify(obs=yobs, pred=yhat)
```

```
## If baseline is not included, baseline values  will be calculated from the  sample obs.
```

```
roc.plot(mod.glm, plot.thres = NULL, main="ROC Curve from Stepwise Fitting")
text(x=0.5, y=0.2, paste("AUC =", round(AUC.step$cvAUC, digits=3),
                         "with 95% CI (", round(AUC.step$ci, digits=3)[1], ",", round(AUC.step$ci, digi
                         sep=" "), col="blue", cex=1.2)
```

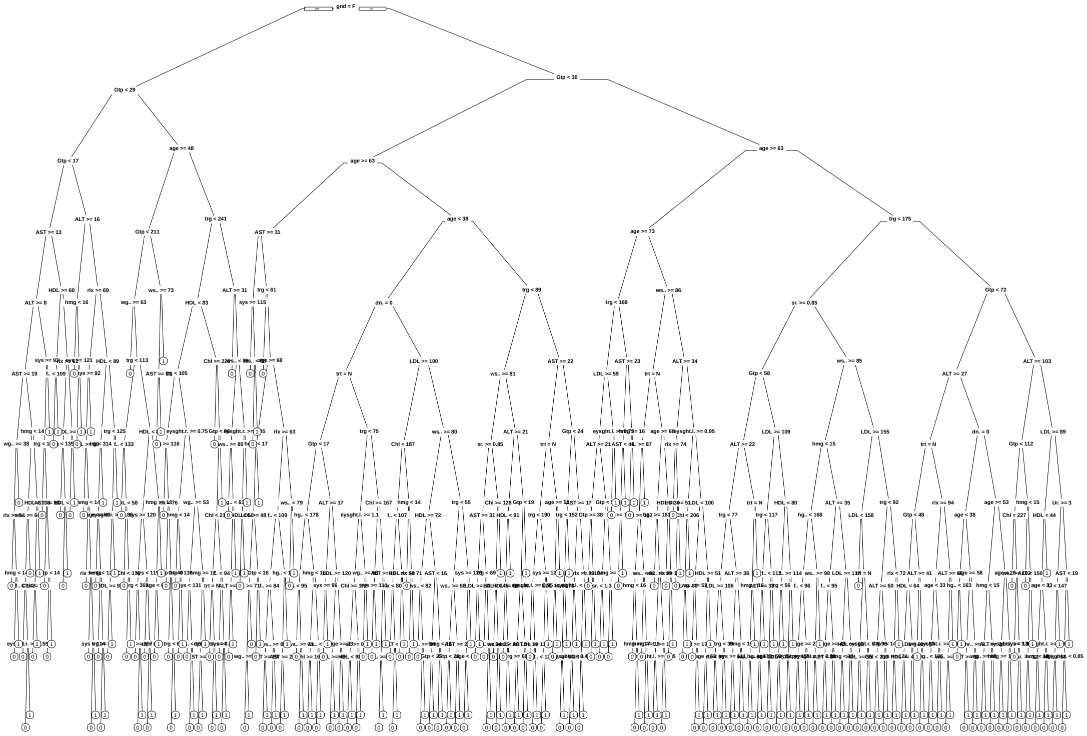


Classification Tree

```
library(rpart)

control0 <- rpart.control(minsplit=10, minbucket=3, maxdepth=10,cp=0, maxcompete=4,maxsurrogate=5, usesu
```

```
tre0 <- rpart(formula, data = train,  method="class", control=control0,
              parms=list(split="gini"))
```

```
library(RColorBrewer)
library(party)
library(rJava)
library(partykit)
library(rpart.plot)
prp(tre0,varlen=3)
```

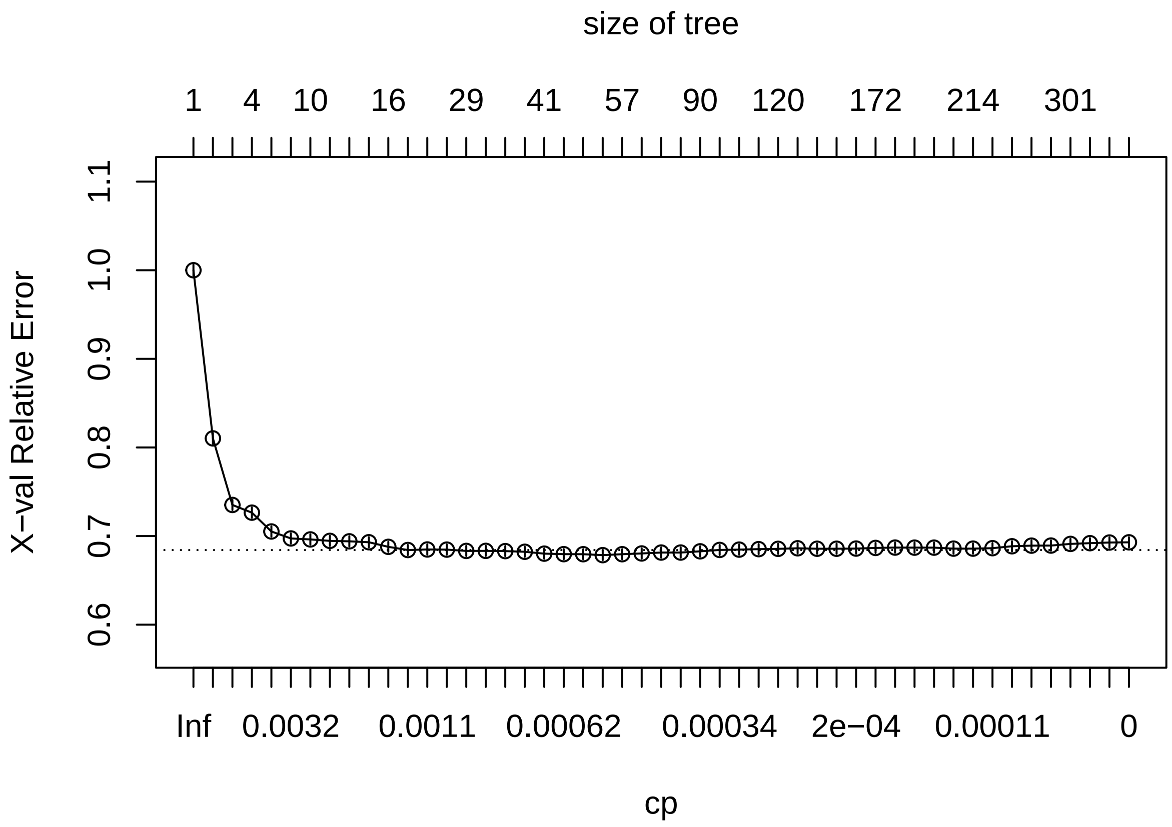


```
printcp(tre0)
```

```
##
## Classification tree:
## rpart(formula = formula, data = train, method = "class", parms = list(split = "gini"),
##     control = control0)
##
## Variables actually used in tree construction:
##  [1] age                 ALT                 AST
##  [4] Cholesterol         dental.caries       eyesight.left.
##  [7] eyesight.right.     fasting.blood.sugar gender
## [10] Gtp                 HDL                 hearing.left.
## [13] height.cm.          hemoglobin          LDL
## [16] relaxation          serum.creatinine    systolic
## [19] tartar              triglyceride        Urine.protein
## [22] waist.cm.           weight.kg.
##
## Root node error: 15403/41769 = 0.36877
```

```
##
## n= 41769
##
##             CP nsplit rel error  xerror      xstd
## 1  1.8977e-01      0   1.00000 1.00000 0.0064017
## 2  7.5115e-02      1   0.81023 0.81023 0.0060733
## 3  1.0193e-02      2   0.73512 0.73512 0.0058981
## 4  9.3921e-03      3   0.72492 0.72642 0.0058760
## 5  3.5383e-03      6   0.69675 0.70519 0.0058204
## 6  2.9215e-03      8   0.68967 0.69733 0.0057992
## 7  2.8566e-03      9   0.68675 0.69623 0.0057962
## 8  2.5644e-03     11   0.68104 0.69460 0.0057918
## 9  1.8828e-03     13   0.67591 0.69409 0.0057903
## 10 1.4283e-03     14   0.67402 0.69305 0.0057875
## 11 1.1816e-03     15   0.67260 0.68785 0.0057732
## 12 1.1037e-03     22   0.66234 0.68428 0.0057632
## 13 1.0388e-03     26   0.65753 0.68487 0.0057649
## 14 9.7384e-04     27   0.65650 0.68474 0.0057645
## 15 8.4399e-04     28   0.65552 0.68324 0.0057603
## 16 7.7907e-04     31   0.65299 0.68331 0.0057605
## 17 7.1415e-04     32   0.65221 0.68305 0.0057598
## 18 6.9251e-04     37   0.64864 0.68227 0.0057576
## 19 6.3299e-04     40   0.64656 0.68019 0.0057518
## 20 6.1676e-04     44   0.64403 0.67961 0.0057501
## 21 5.8430e-04     50   0.64033 0.67954 0.0057500
## 22 5.1938e-04     54   0.63799 0.67850 0.0057470
## 23 4.9774e-04     56   0.63695 0.67954 0.0057500
## 24 4.7610e-04     66   0.63118 0.68039 0.0057523
## 25 4.7069e-04     69   0.62975 0.68143 0.0057553
## 26 3.8953e-04     73   0.62786 0.68143 0.0057553
## 27 3.4625e-04     89   0.62163 0.68279 0.0057591
## 28 3.3904e-04     95   0.61955 0.68435 0.0057634
## 29 3.2461e-04    106   0.61579 0.68474 0.0057645
## 30 2.9215e-04    109   0.61482 0.68526 0.0057660
## 31 2.5969e-04    119   0.61157 0.68558 0.0057669
## 32 2.4346e-04    136   0.60696 0.68623 0.0057687
## 33 2.2723e-04    141   0.60573 0.68571 0.0057672
## 34 2.1100e-04    143   0.60527 0.68571 0.0057672
## 35 1.9477e-04    147   0.60443 0.68591 0.0057678
## 36 1.6231e-04    171   0.59975 0.68662 0.0057697
## 37 1.5149e-04    179   0.59833 0.68701 0.0057708
## 38 1.4608e-04    182   0.59787 0.68694 0.0057707
## 39 1.4283e-04    197   0.59462 0.68694 0.0057707
## 40 1.4067e-04    202   0.59391 0.68584 0.0057676
## 41 1.2984e-04    213   0.59222 0.68584 0.0057676
## 42 9.7384e-05    242   0.58833 0.68636 0.0057690
## 43 9.2746e-05    252   0.58735 0.68850 0.0057750
## 44 8.1153e-05    270   0.58547 0.68922 0.0057769
## 45 6.4922e-05    279   0.58469 0.68928 0.0057771
## 46 5.1938e-05    300   0.58333 0.69116 0.0057823
## 47 3.2461e-05    305   0.58307 0.69201 0.0057846
## 48 2.1641e-05    319   0.58261 0.69272 0.0057866
## 49 0.0000e+00    328   0.58242 0.69298 0.0057873
```

```
plotcp(tre0)
```



Now, we aim to find the optimal subtree.

```
cv.error <- (tre0$cptable)[,4]
a0 <- 1     # IF a0=0, THEN 0SE
SE1 <- min(cv.error) + a0*((tre0$cptable)[,5])[which.min(cv.error)]      # 1SE
position <- min((1:length(cv.error))[cv.error <= SE1])
n.size  <- (tre0$cptable)[,2] + 1  # TREE SIZE IS ONE PLUS NUMBER OF SPLITS.
best.size <- n.size[position]; best.size
```

```
## 15
## 29
```

```
best.cp <-  sqrt(tre0$cptable[position,1] *  tre0$cptable[(position-1),1])
best.cp
```

```
## [1] 0.0009065922
```

```
best.tree <- prune(tre0, cp=best.cp)

# Prediction

pred.tree <- predict(best.tree, newdata = test)
```

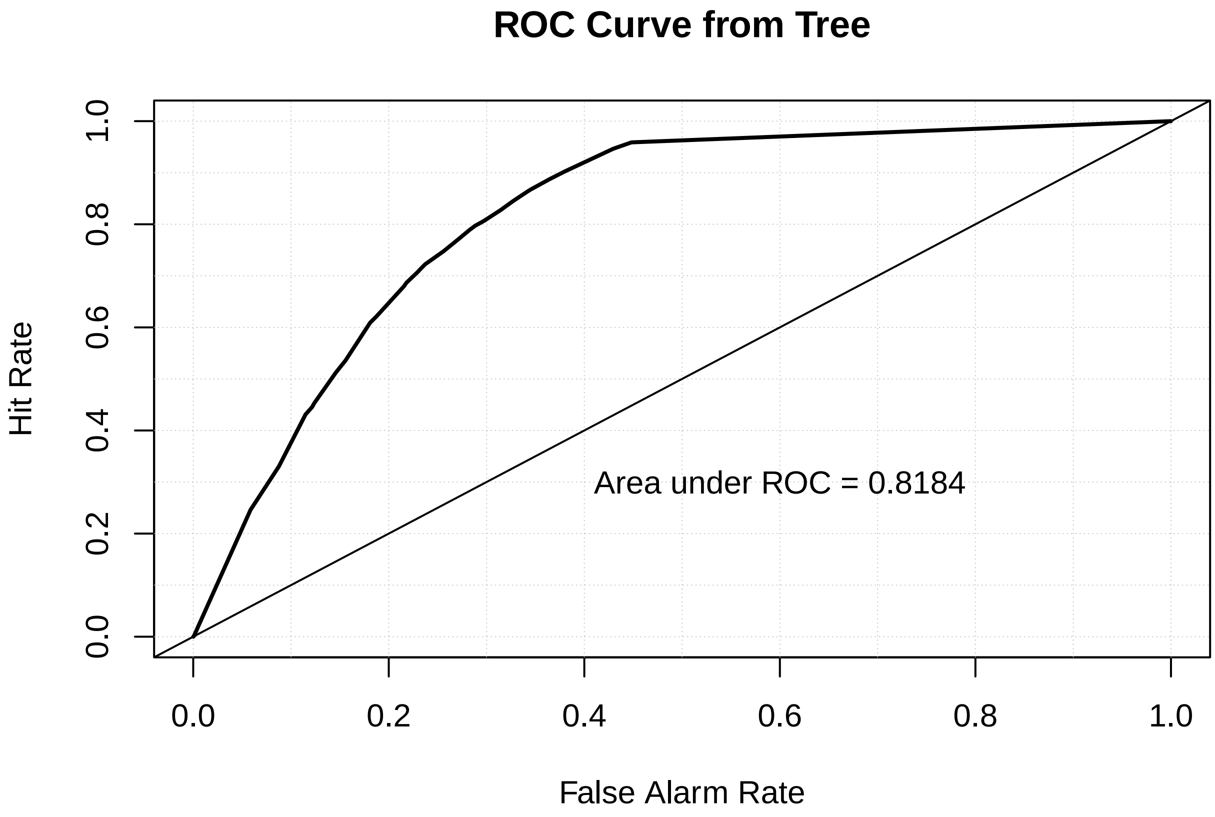
```r
library(verification)
yhat.tree <- pred.tree[,2]
a.ROC.tree <- roc.area(obs=yobs, pred=yhat.tree)$A;
AUC <- round(a.ROC.tree, digits=4); AUC
```

```
## [1] 0.8184
```

```r
par(mfrow=c(1,1), mar=c(4, 4, 4, 4))
mod.glm <- verify(obs=yobs, pred=yhat.tree, bins = FALSE)
```

```
## If baseline is not included, baseline values  will be calculated from the  sample obs.
```

```r
roc.plot(mod.glm, plot.thres = NULL, main ="ROC Curve from Tree")
text(x=0.6, y=0.3, paste("Area under ROC = ", AUC, sep=""))
```



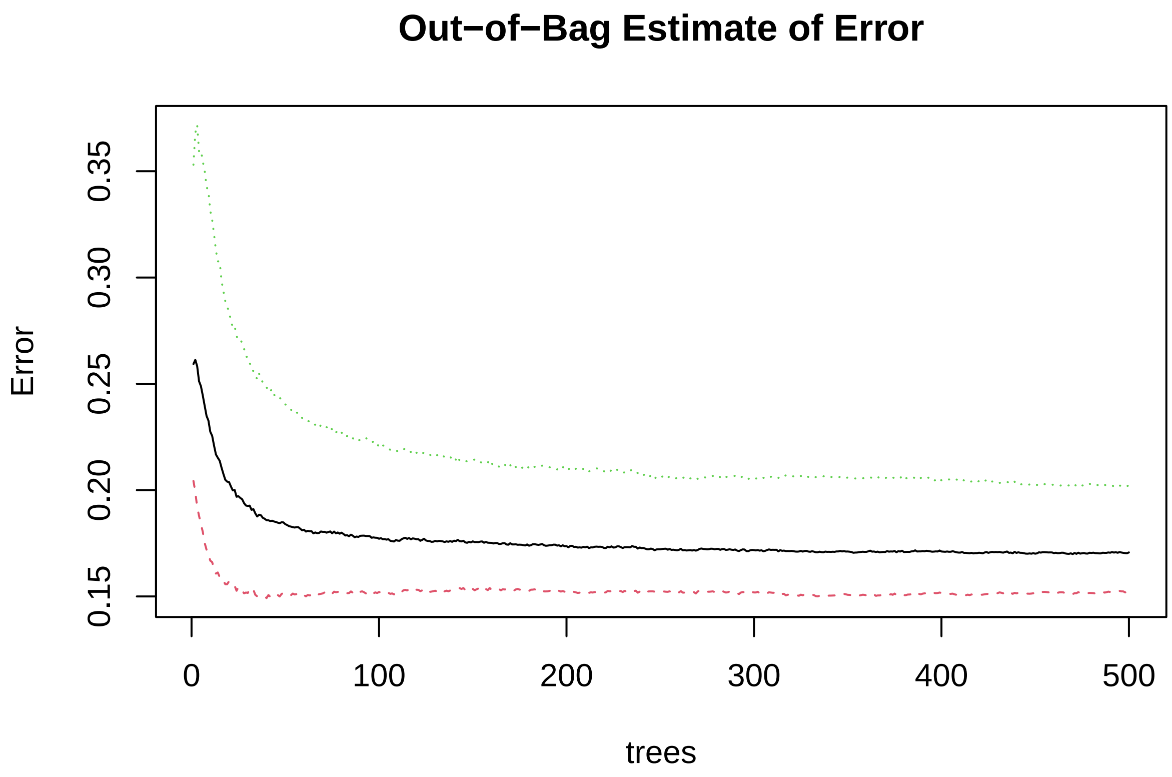
Random Forest

```r
library(randomForest)
fit.rf <- randomForest(factor(smoking) ~ +gender + age + height.cm. + weight.kg. + waist.cm. +
                         eyesight.left. + eyesight.right. + hearing.left. + hearing.right. +
                         systolic + relaxation + fasting.blood.sugar + Cholesterol +
                         triglyceride + HDL + LDL + hemoglobin + Urine.protein + serum.creatinine +
                         AST + ALT + Gtp + dental.caries + tartar, data = train, importance=TRUE,
                       ntree=500)
```

```
fit.rf
```

```
##
## Call:
##  randomForest(formula = factor(smoking) ~ +gender + age + height.cm. +      weight.kg. + waist.cm. +
##                 Type of random forest: classification
##                       Number of trees: 500
## No. of variables tried at each split: 4
##
##         OOB estimate of  error rate: 17.07%
## Confusion matrix:
##       0     1 class.error
## 0 22355  4011   0.1521277
## 1  3119 12284   0.2024930
```

```
plot(fit.rf, main="Out-of-Bag Estimate of Error")
```



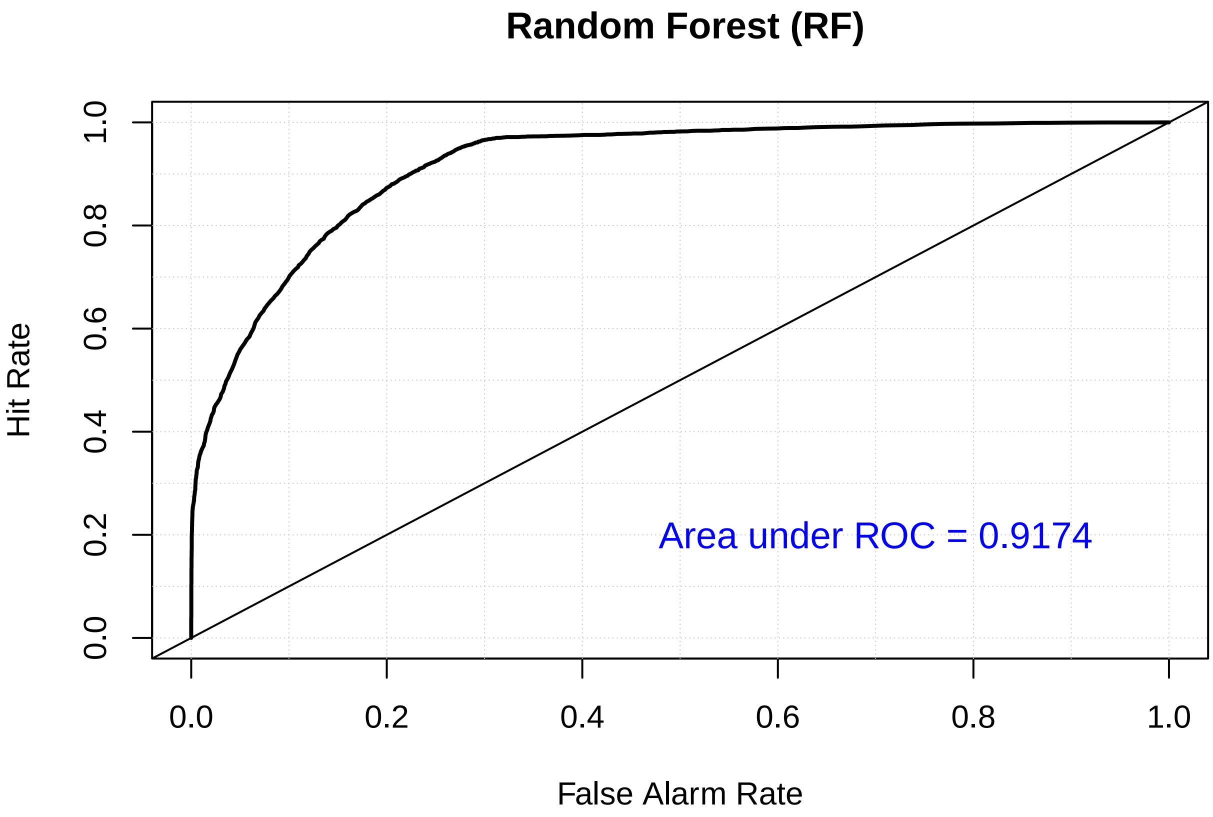
```
yhat.forest <- predict(fit.rf, newdata = test, type="prob")[, 2]

AUC.forest <- roc.area(obs = yobs, pred = yhat.forest)$A; AUC.forest
```

```
## [1] 0.9174073
```

```
modRf.glm <- verify(obs = yobs, pred = yhat.forest)
```

```
## If baseline is not included, baseline values  will be calculated from the  sample obs.
```

```
roc.plot(modRf.glm, plot.thres = NULL, col="red", main=" Random Forest (RF)")
text(x=0.7, y=0.2, paste("Area under ROC =", round(AUC.forest, digits=4),
                         sep=" "), col="blue", cex=1.2)
```



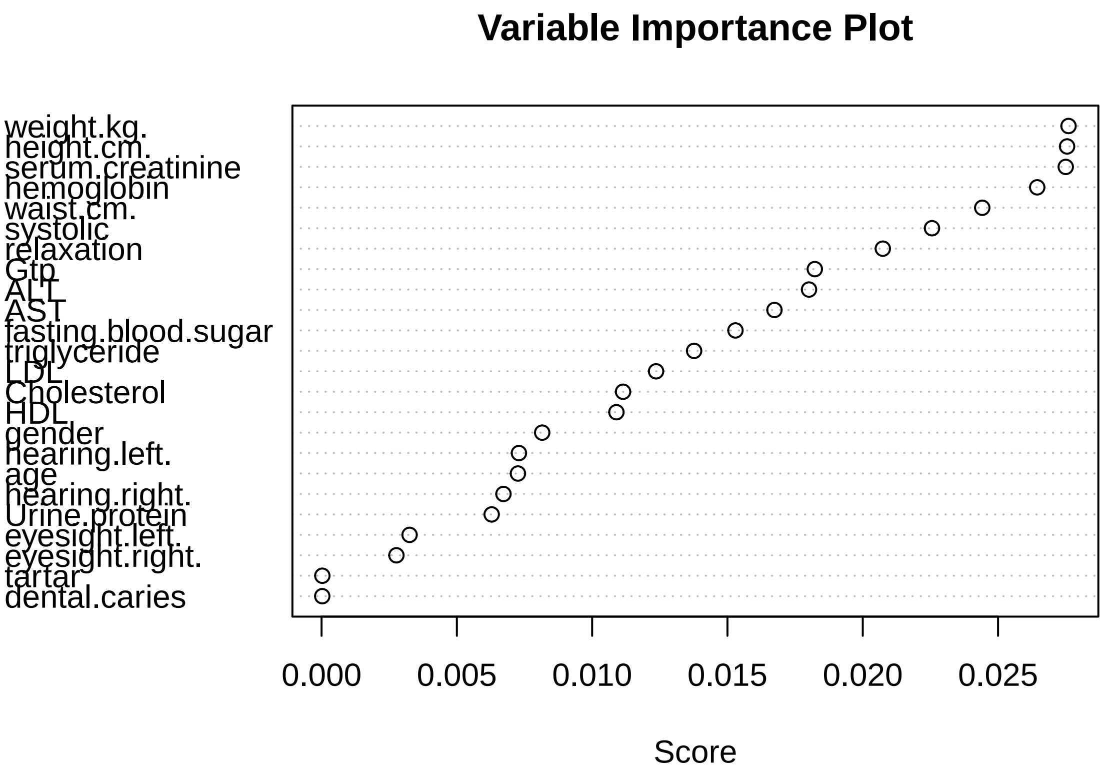
Boosting

```
library("ada")
stump <- rpart.control(cp=-1, maxdepth=1, minsplit=0)
fit.stump <- ada(formula, data=train, iter=500,
                 loss="e", type="discrete",
                 control=stump)

fit.stump
```

```
## Call:
```

```
## ada(formula, data = train, iter = 500, loss = "e", type = "discrete",
##     control = stump)
##
## Loss: exponential Method: discrete   Iteration: 500
##
## Final Confusion Matrix for Data:
##           Final Prediction
## True value     0     1
##          0 20322  6044
##          1  4251 11152
##
## Train Error: 0.246
##
## Out-Of-Bag Error:  0.255  iteration= 496
##
## Additional Estimates of number of iterations:
##
## train.err1 train.kap1
##        499        499
```

```
varplot(fit.stump, plot.it=TRUE,type="scores")
```



```
##          weight.kg.          height.cm.    serum.creatinine          hemoglobin
##        2.760348e-02        2.755085e-02        2.750107e-02        2.644720e-02
##           waist.cm.            systolic          relaxation                 Gtp
##        2.441621e-02        2.255859e-02        2.074120e-02        1.822830e-02
##                 ALT                 AST fasting.blood.sugar        triglyceride
##        1.801405e-02        1.673737e-02        1.529505e-02        1.376768e-02
##                 LDL         Cholesterol                 HDL              gender
##        1.236146e-02        1.114160e-02        1.089600e-02        8.153498e-03
##       hearing.left.                 age      hearing.right.       Urine.protein
##        7.293719e-03        7.256826e-03        6.722353e-03        6.286026e-03
##      eyesight.left.     eyesight.right.              tartar       dental.caries
##        3.253163e-03        2.766246e-03        2.862610e-05        2.599842e-05
```

```
vip <- varplot(fit.stump, plot.it=FALSE, type="scores")
round(vip,4)
```

```
##          weight.kg.          height.cm.    serum.creatinine          hemoglobin
##              0.0276              0.0276              0.0275              0.0264
##           waist.cm.            systolic          relaxation                 Gtp
##              0.0244              0.0226              0.0207              0.0182
##                 ALT                 AST fasting.blood.sugar        triglyceride
##              0.0180              0.0167              0.0153              0.0138
##                 LDL         Cholesterol                 HDL              gender
##              0.0124              0.0111              0.0109              0.0082
##       hearing.left.                 age      hearing.right.       Urine.protein
##              0.0073              0.0073              0.0067              0.0063
##      eyesight.left.     eyesight.right.              tartar       dental.caries
##              0.0033              0.0028              0.0000              0.0000
```

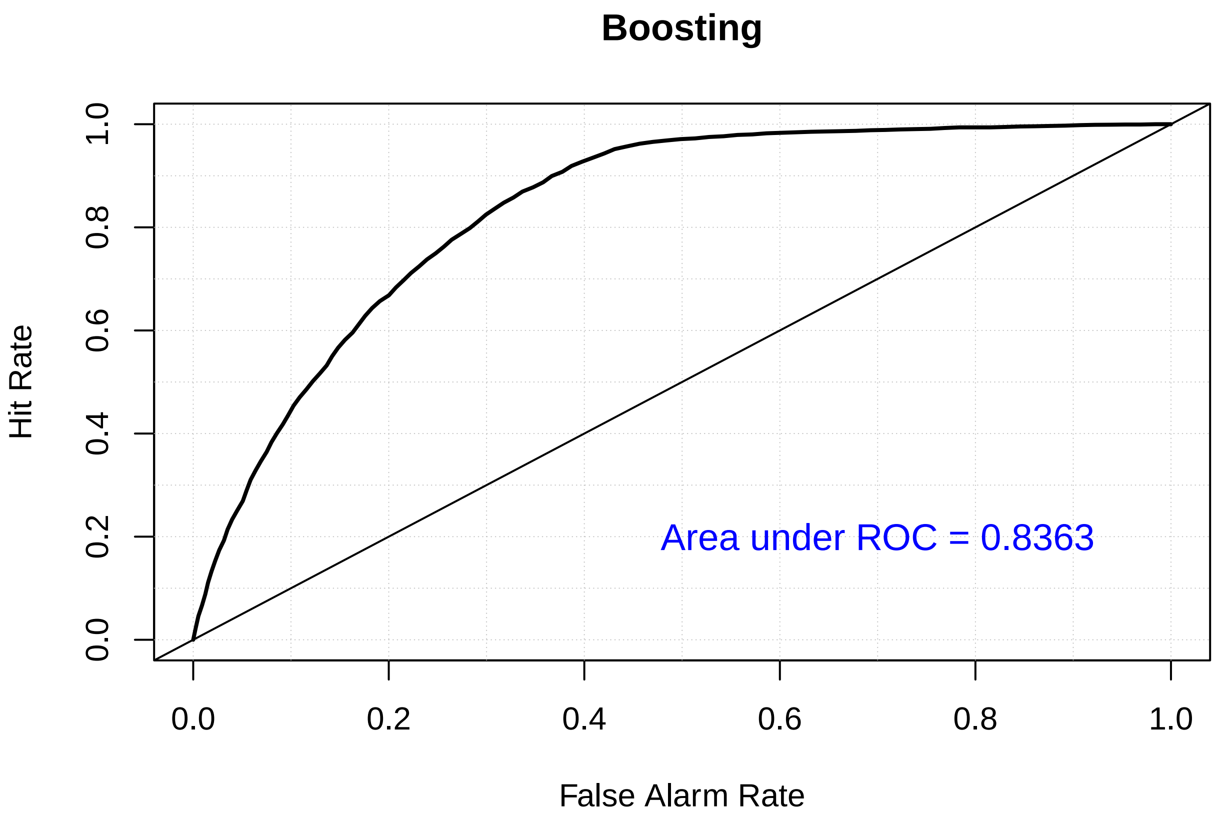
```
yhat.boost <- predict(fit.stump, newdata=test, type="probs")[, 2]

AUC.boost <- roc.area(obs = yobs, pred = yhat.boost)$A; AUC.boost
```

```
## [1] 0.83627
```

```
modBst.glm <- verify(obs = yobs, pred = yhat.boost)
```

```
## If baseline is not included, baseline values  will be calculated from the  sample obs.
```

```
roc.plot(modBst.glm, plot.thres = NULL, col="red", main="Boosting")
text(x=0.7, y=0.2, paste("Area under ROC =", round(AUC.boost, digits=4),
                         sep=" "), col="blue", cex=1.2)
```



Now, we compare the methods we have implemented so far.

```
cbind(AUC.LDA , AUC.step, AUC.forest, AUC, AUC.boost)
```

```
##            AUC.LDA     AUC.step   AUC.forest AUC    AUC.boost
## cvAUC      0.8265198   0.7636817  0.9174073  0.8184 0.83627
## se         0.003416727 0.00635689 0.9174073  0.8184 0.83627
## ci         numeric,2   numeric,2  0.9174073  0.8184 0.83627
## confidence 0.95        0.95       0.9174073  0.8184 0.83627
```

Among all the classifiers, random forest has the best AUC value and 'best' logistic regression has the worst AUC value.

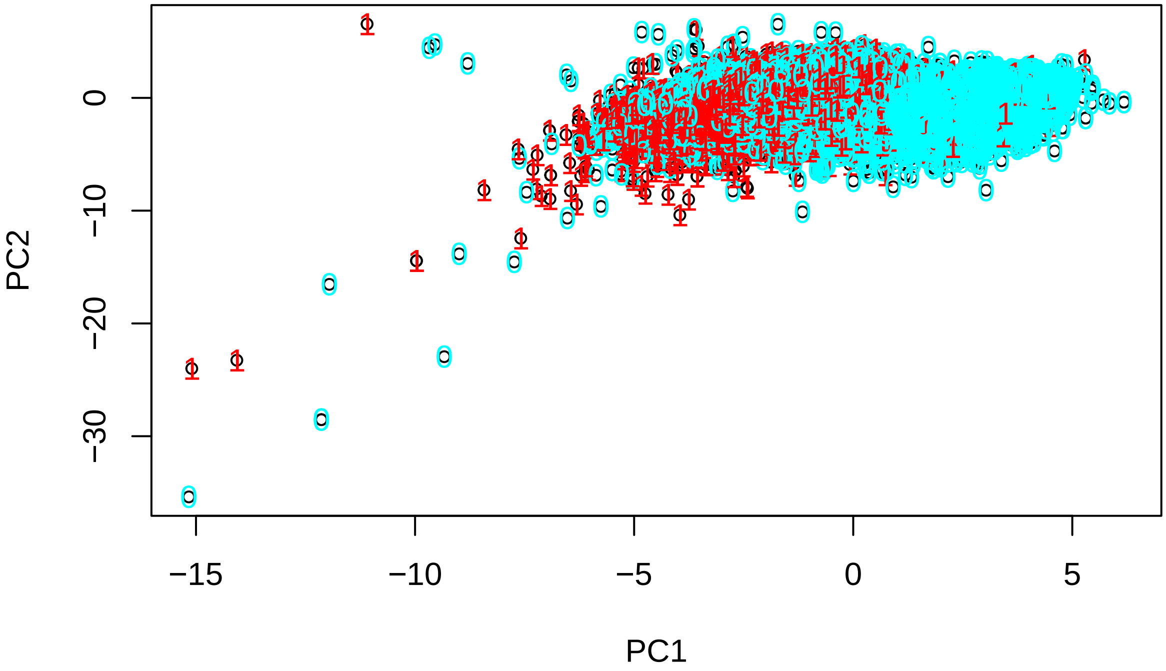
Unsupervised Learning

For the time being, we ignore the response variable smoking and apply multidimensional scaling and t-SNE to the training data.

Classical MDS

```
train1 <- model.matrix(~.-1, data = train[-c(25)])
```

```
# Preparing colors for plotting

label = as.character(train$smoking)
colors = rainbow(2)
names(colors) = unique(label)

result.pca <- prcomp(train1, scale = TRUE)
plot(result.pca$x[,1:2],cex = 0.75)
text(result.pca$x[,1:2], labels = label, col = colors[label])
```

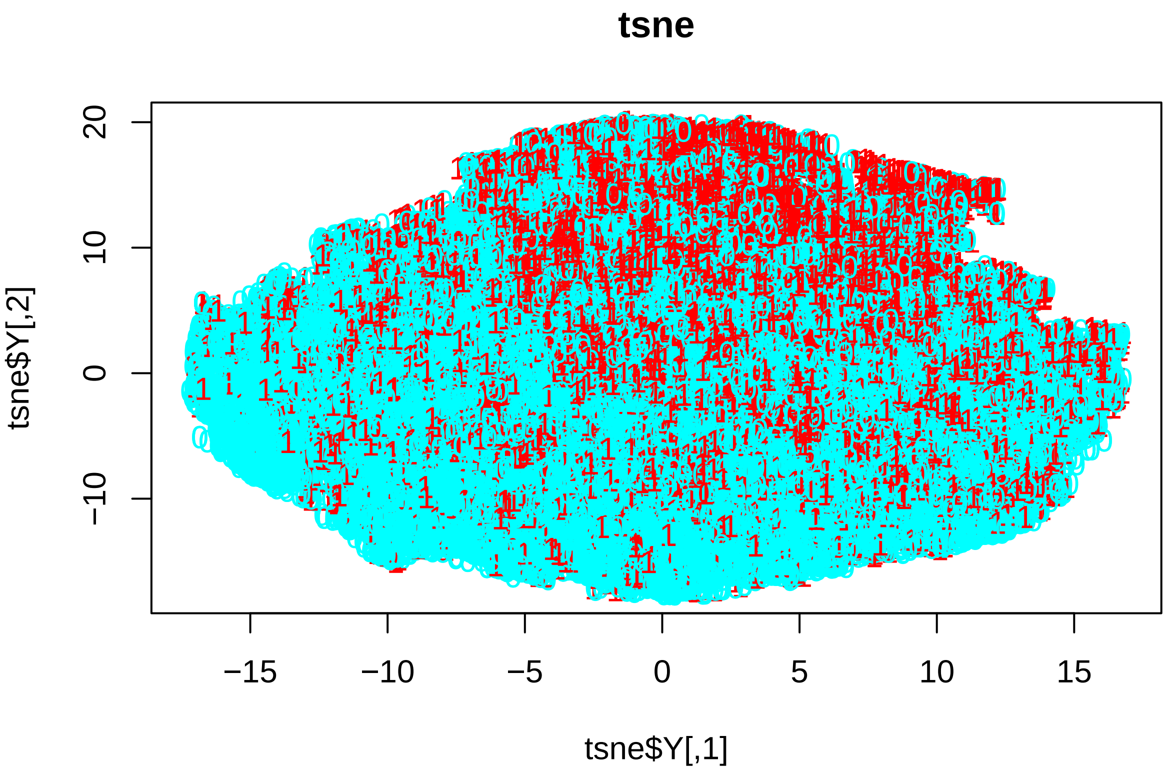


t-SNE

```
library(Rtsne)
tsne <- Rtsne(train1, dims = 2, perplexity=30, verbose=TRUE, max_iter = 500, check_duplicates = FALSE )
```

```
## Performing PCA
## Read the 41769 x 25 data matrix successfully!
## OpenMP is working. 1 threads.
## Using no_dims = 2, perplexity = 30.000000, and theta = 0.500000
## Computing input similarities...
## Building tree...
##  - point 10000 of 41769
##  - point 20000 of 41769
##  - point 30000 of 41769
##  - point 40000 of 41769
## Done in 47.19 seconds (sparsity = 0.003091)!
## Learning embedding...
## Iteration 50: error is 113.980792 (50 iterations in 10.45 seconds)
## Iteration 100: error is 113.980778 (50 iterations in 10.46 seconds)
## Iteration 150: error is 110.296347 (50 iterations in 11.98 seconds)
## Iteration 200: error is 102.413842 (50 iterations in 19.50 seconds)
## Iteration 250: error is 100.643106 (50 iterations in 21.39 seconds)
## Iteration 300: error is 4.628388 (50 iterations in 16.38 seconds)
## Iteration 350: error is 4.303840 (50 iterations in 14.56 seconds)
## Iteration 400: error is 4.113120 (50 iterations in 14.01 seconds)
## Iteration 450: error is 3.972682 (50 iterations in 11.62 seconds)
## Iteration 500: error is 3.859862 (50 iterations in 11.27 seconds)
## Fitting performed in 141.62 seconds.
```

```
plot(tsne$Y, t='n', main="tsne")
text(tsne$Y, labels = label, col = colors[label])
```



Neither MDS nor t-SNE does a good job in clustering smokers and nonsmokers.