590-01 Final Project

Roderick Whang, Roujia Wang

4/6/2021 modified on 04/12/2021

```r
rm(list = ls())
setwd("/Users/rosesharpaywang/Desktop/BME590Spring2021Dunn/FinalProject/590-01-project/Roujia")
library(tidyverse)
library(ggplot2)
library(lubridate)
library(patchwork)
library(gridExtra)
library(psych)
library(corrplot)
library(ggfortify)
library(factoextra)
library(class) #knn
library(gmodels) # CrossTable()
library(caret) # creatFolds()
library(caTools) #sample.split()
library(ROCR) # prediction(), performance()
library(randomForest) # Random Forest
library(caret)
library(e1071) # SVM
set.seed(2021)

df1 <- read.csv("S1.csv")
df2 <- read.csv("S2.csv")
df3 <- read.csv("S3.csv")
df4 <- read.csv("S4.csv")
df5 <- read.csv("S5.csv")
df6 <- read.csv("S6.csv")
df7 <- read.csv("S7.csv")
df8 <- read.csv("S8.csv")
df9 <- read.csv("S9.csv")
df10 <- read.csv("S10.csv")
df11 <- read.csv("S11.csv")
df12 <- read.csv("S12.csv")
df13 <- read.csv("S13.csv")
df14 <- read.csv("S14.csv")
df15 <- read.csv("S15.csv")

df <- rbind(df1, df2, df3, df4, df5, df6, df7, df8, df9, df10, df11, df12, df13, df14, df15)
df$activity <- as.factor(df$activity)
df <- subset(df, select = -c(X) )
df <- subset(df, df$activity == 1 | df$activity == 2 | df$activity == 3 |
               df$activity ==4 | df$activity == 5| df$activity == 6 | df$activity == 7 |
```

```r
               df$activity ==8)

# df$Activity
# low = 1

# medium = 3
df$activity[df$activity== 3 | df$activity== 5 | df$activity == 6 | df$activity == 8] <- 3

# high = 2
df$activity[df$activity==2 | df$activity== 4 | df$activity == 7] <- 2


df$activity <- factor(df$activity)
```

a2 <- which(df$activity == 2)a3 < -which(df$activity ==3)

a2_s <- sample(a2, 6000, replace = FALSE) a3_s <- sample(a3, 26000, replace = FALSE)

df <- df[-c(a2_s, a3_s),]

```r
head(df)
```

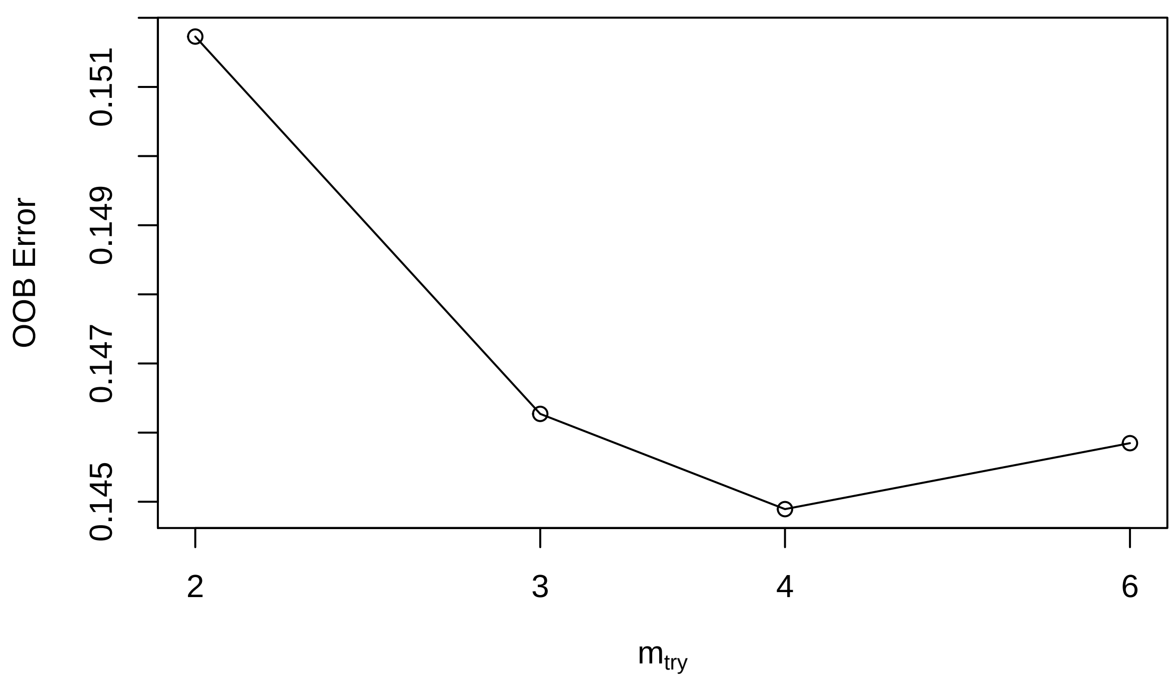
```
##    activity      f1.mean    f2.std    f3.max     f4.min f5.max_position
## 45        1 -0.038429169 0.3478793 0.4465136 -1.0299322     0.070357143
## 46        1 -0.011527297 0.3137633 0.4326159 -0.7738035     0.259821429
## 47        1  0.005274472 0.2966080 0.4326159 -0.7738035     0.009821429
## 48        1 -0.015503497 0.2861047 0.3738340 -0.7559939     0.086071429
## 49        1  0.004943980 0.2922504 0.4119239 -0.8208496     0.980535714
## 50        1 -0.005607502 0.2965154 0.4475430 -0.8208496     0.970714286
##    f6.min_position    f7.hr f8.skewness f9.kurtosis
## 45      0.01375000 46.11024  -0.7152854 -0.51051998
## 46      0.37500000 46.50810  -0.7118847 -0.55704806
## 47      0.12500000 47.03796  -0.8300902 -0.19700263
## 48      0.02946429 46.88382  -0.7594123 -0.36029498
## 49      0.92196429 46.32763  -0.8617181 -0.01984442
## 50      0.67196429 45.61259  -0.8270814 -0.03742761
```

```r
set.seed(2021)

sample <- sample.split(df$activity, SplitRatio = .8) # dataset to split it into 80:20

df_train <- df[sample==TRUE, ]
df_test <- df[sample==FALSE, ]

X_train <- subset(df_train, select = -c(activity) ) # independent variables
y_train <- df_train[,1] # tartget variables

X_test <- subset(df_test, select = -c(activity) ) # independent variables
y_test <- df_test[,1] # tartget variables
```

RF

```
bestmtry <- tuneRF(X_train, y_train, stepFactor=1.5, improve=1e-5, ntree=700)
```

```
## mtry = 3  OOB error = 14.63%
## Searching left ...
## mtry = 2     OOB error = 15.17%
## -0.03731208 1e-05
## Searching right ...
## mtry = 4     OOB error = 14.49%
## 0.009418584 1e-05
## mtry = 6     OOB error = 14.58%
## -0.006582556 1e-05
```



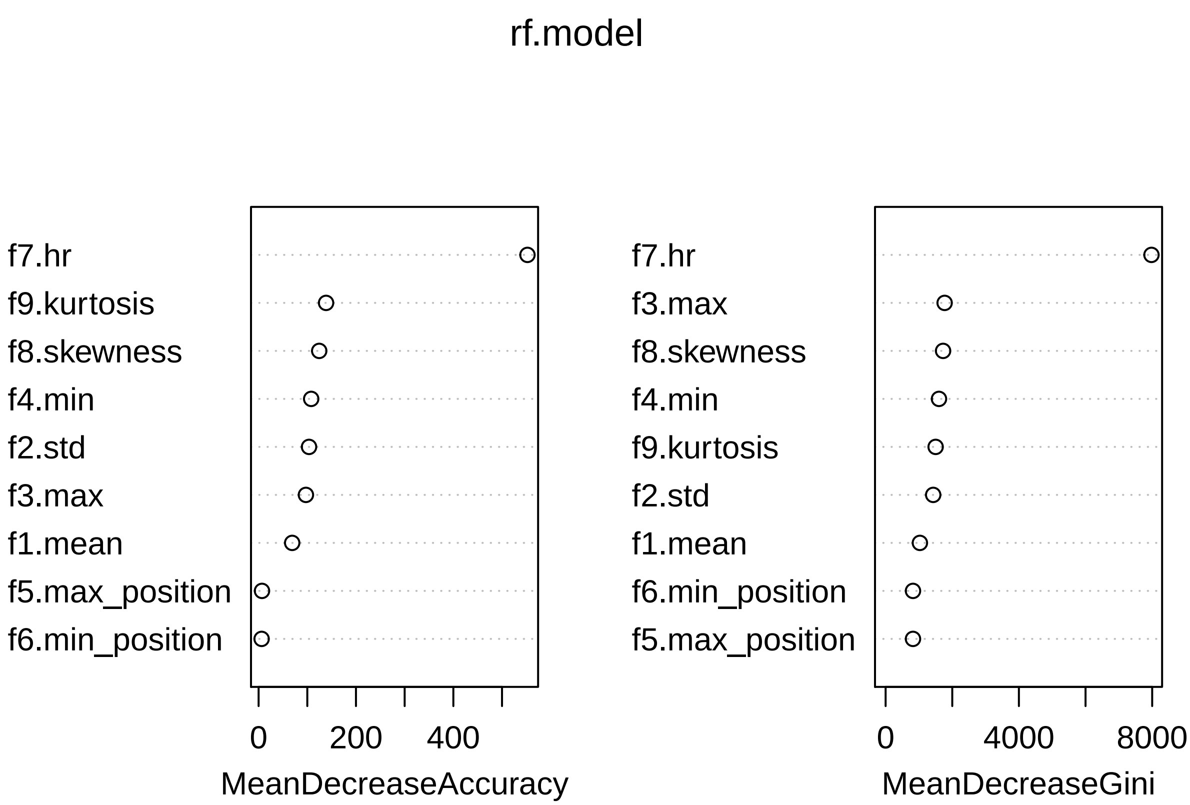

```
print(bestmtry)
```

```
##       mtry   OOBError
## 2.OOB    2 0.1517287
## 3.OOB    3 0.1462710
## 4.OOB    4 0.1448934
## 6.OOB    6 0.1458471
```

```r
rf.model <- randomForest(formula = activity ~ ., data = df_train, ntree=700, mtry=3, importance = TRUE,
rf.model
```

```
##
## Call:
##  randomForest(formula = activity ~ ., data = df_train, ntree = 700,      mtry = 3, importance = TRUE
##                Type of random forest: classification
##                      Number of trees: 700
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 14.65%
## Confusion matrix:
##      1    2     3 class.error
## 1 2887   35   733  0.21012312
## 2    5 6174  2948  0.32354552
## 3  300 1508 23155  0.07242719
```

```r
varImpPlot(rf.model)
```



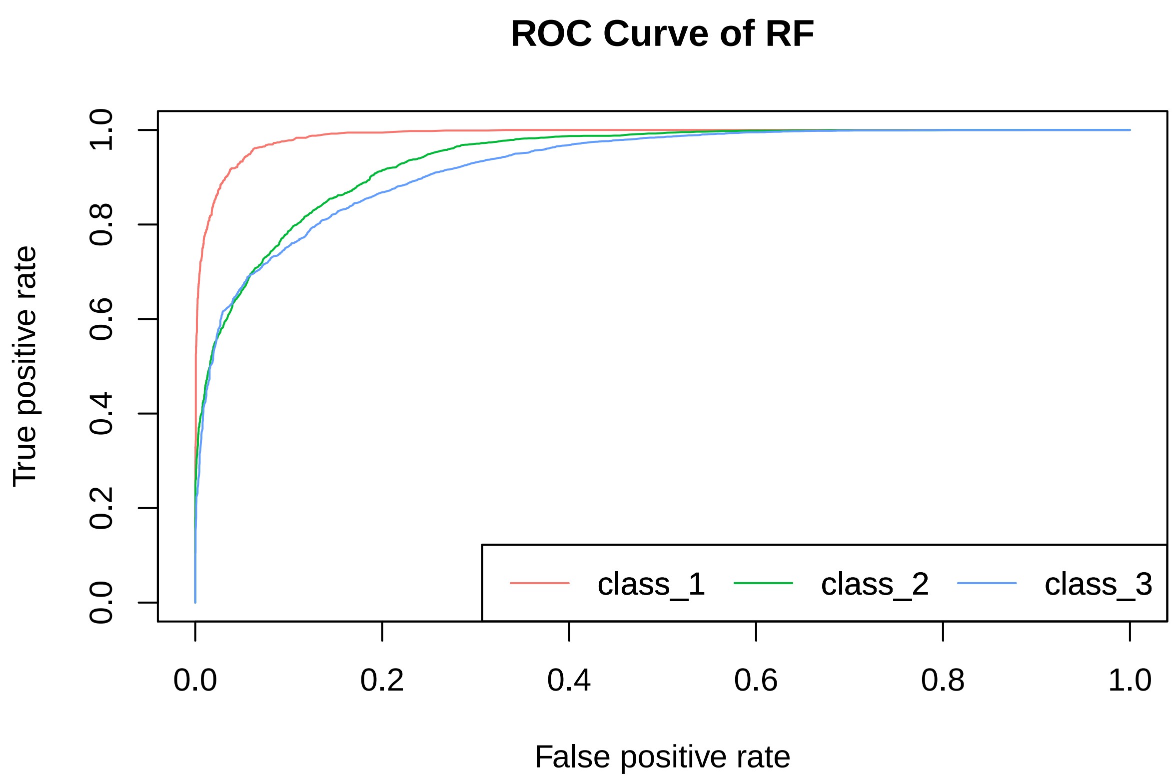
The Mean Decrease Accuracy plot expresses how much accuracy the model losses by excluding each variable. The more the accuracy suffers, the more important the variable is for the successful classification. The variables are presented from descending importance. The mean decrease in Gini coefficient is a measure of how each variable contributes to the homogeneity of the nodes and leaves in the resulting random forest. The higher the value of mean decrease accuracy or mean decrease Gini score, the higher the importance of the variable in the model.

```r
prediction_for_table <- predict(rf.model,X_test)
#table(observed=y_test,predicted=prediction_for_table)
confusionMatrix(prediction_for_table, y_test)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    1    2    3
##          1  713    1   86
##          2   11 1525  367
##          3  190  756 5788
##
## Overall Statistics
##
##                Accuracy : 0.8505
##                  95% CI : (0.8431, 0.8576)
##     No Information Rate : 0.6613
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.6826
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
## Statistics by Class:
##
##                      Class: 1 Class: 2 Class: 3
## Sensitivity           0.78009   0.6683   0.9274
## Specificity           0.98979   0.9472   0.7040
## Pos Pred Value        0.89125   0.8014   0.8595
## Neg Pred Value        0.97673   0.8995   0.8324
## Prevalence            0.09685   0.2418   0.6613
## Detection Rate        0.07555   0.1616   0.6133
## Detection Prevalence  0.08477   0.2017   0.7136
## Balanced Accuracy     0.88494   0.8077   0.8157
```

```r
pred_prob.rf <- predict(rf.model, X_test, decision.values = TRUE, type="prob")
colours <- c("#F8766D","#00BA38","#619CFF")
# Specify the different classes
classes <- levels(df$activity)
# For each class
for (i in 1:3)
{
 # Define which observations belong to class[i]
 true_values <- ifelse(y_test==classes[i],1,0)
 # Assess the performance of classifier for class[i]
 pred <- prediction(pred_prob.rf[,i],true_values)

 perf <- performance(pred, "tpr", "fpr")
 if (i==1)
 {
     plot(perf,main="ROC Curve of RF",col=colours[i])
 }
 else
```

```r
{
    plot(perf,main="ROC Curve of RF",col=colours[i],add=TRUE)
}
legend("bottomright",  c("class_1","class_2","class_3"), col = colours, lty= 1, horiz=TRUE)
# Calculate the AUC and print it to screen
auc.perf <- performance(pred, measure = "auc")
print(paste("AUC of class_",i,":",auc.perf@y.values))
}
```




```
## [1] "AUC of class_ 1 : 0.98880722031337"
## [1] "AUC of class_ 2 : 0.939832315738092"
## [1] "AUC of class_ 3 : 0.925773464226534"
```

SVM

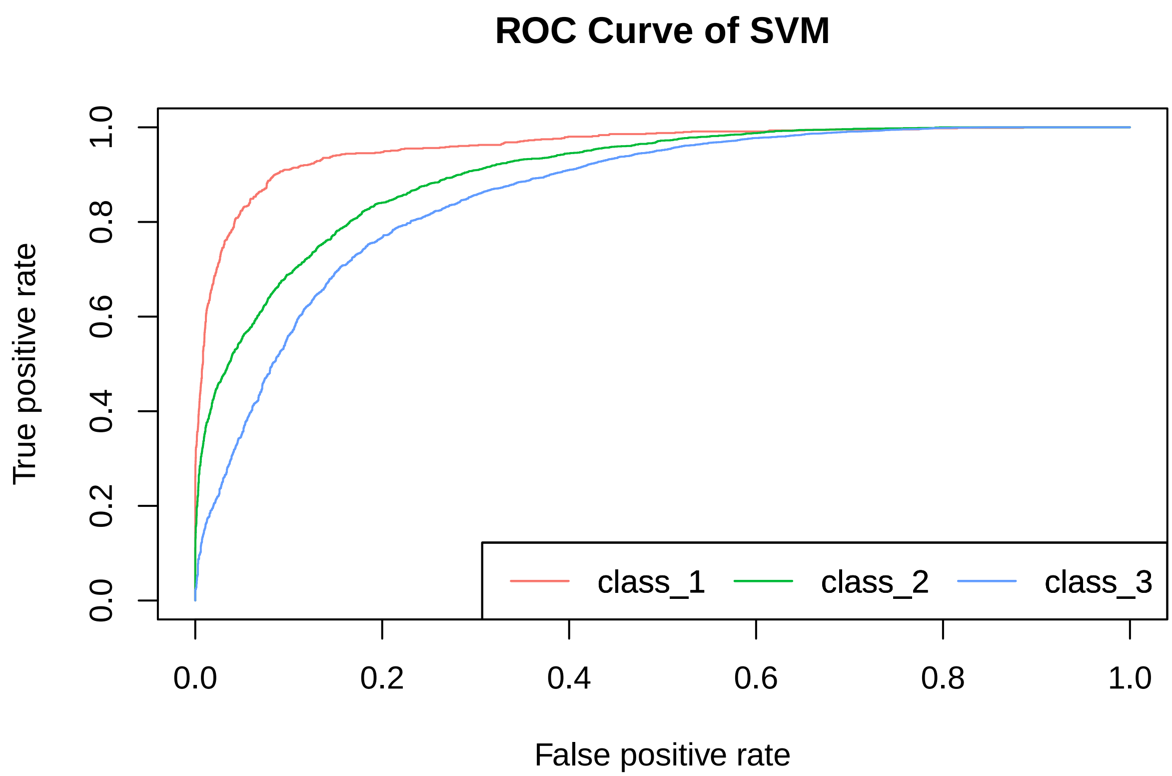
```r
#set.seed(1)
#X <- sample(dim(X_train)[1], 3000, replace=FALSE)
#tune.out <- tune(svm, activity ~., data=df_train[X,],
#                  kernel='radial',
#                  ranges = list(cost=c(0.1,1,10,100,1000),
#                  gamma=c(0.5, 1,2,3,4)))

#summary(tune.out)
```

```r
svm.opt <- svm(activity ~., data=df_train, kernel='radial', type =  'C-classification',
               gamma=0.07, cost=10
               , decision.values=T, probability = TRUE)
pred <- predict(svm.opt, X_test, decision.values = TRUE, probability = TRUE)

confusionMatrix(pred, y_test)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    1    2    3
##          1  566    0  111
##          2    0 1254  353
##          3  348 1028 5777
##
## Overall Statistics
##
##                Accuracy : 0.805
##                  95% CI : (0.7969, 0.813)
##     No Information Rate : 0.6613
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5673
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: 1 Class: 2 Class: 3
## Sensitivity           0.61926   0.5495   0.9257
## Specificity           0.98698   0.9507   0.5695
## Pos Pred Value        0.83604   0.7803   0.8076
## Neg Pred Value        0.96027   0.8687   0.7968
## Prevalence            0.09685   0.2418   0.6613
## Detection Rate        0.05998   0.1329   0.6122
## Detection Prevalence  0.07174   0.1703   0.7580
## Balanced Accuracy     0.80312   0.7501   0.7476
```

```r
pred_prob.svm <- attr(pred, "probabilities")
colours <- c("#F8766D","#00BA38","#619CFF")
# Specify the different classes
classes <- levels(df$activity)
# For each class
for (i in 1:3)
{
 # Define which observations belong to class[i]
 true_values <- ifelse(y_test==classes[i],1,0)
 # Assess the performance of classifier for class[i]
 pred <- prediction(pred_prob.svm[,i],true_values)

 perf <- performance(pred, "tpr", "fpr")
 if (i==1)
 {
```

```r
    plot(perf,main="ROC Curve of SVM",col=colours[i])
 }
 else
 {
    plot(perf,main="ROC Curve of SVM",col=colours[i],add=TRUE)
 }
 legend("bottomright",  c("class_1","class_2","class_3"), col = colours, lty= 1, horiz=TRUE)
 # Calculate the AUC and print it to screen
 auc.perf <- performance(pred, measure = "auc")
 print(paste("AUC of class_",i,":",auc.perf@y.values))
}
```




```
## [1] "AUC of class_ 1 : 0.959684837860562"
## [1] "AUC of class_ 2 : 0.903579681412777"
## [1] "AUC of class_ 3 : 0.860172515756856"
```