SMANDUMU_Assignment_2

smandumu

4/14/2020

```r
library(mlbench)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

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

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

```r
library(caret)
```

```
## Loading required package: lattice

## Loading required package: ggplot2
```

```r
data(BreastCancer)
dim(BreastCancer)
```

```
## [1] 699  11
```

```r
levels(BreastCancer$Class)
```

```
## [1] "benign"    "malignant"
```

```r
str(BreastCancer)
```

```
## 'data.frame':    699 obs. of  11 variables:
##  $ Id             : chr  "1000025" "1002945" "1015425" "1016277" ...
##  $ Cl.thickness   : Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 5 5 3 6 4 8 1 2 2 4 ...
##  $ Cell.size      : Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 1 4 1 8 1 10 1 1 1 2 ...
##  $ Cell.shape     : Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 1 4 1 8 1 10 1 2 1 1 ...
##  $ Marg.adhesion  : Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 1 5 1 1 3 8 1 1 1 1 ...
##  $ Epith.c.size   : Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 2 7 2 3 2 7 2 2 2 2 ...
##  $ Bare.nuclei    : Factor w/ 10 levels "1","2","3","4",..: 1 10 2 4 1 10 10 1 1 1 ...
##  $ Bl.cromatin    : Factor w/ 10 levels "1","2","3","4",..: 3 3 3 3 3 9 3 3 1 2 ...
##  $ Normal.nucleoli: Factor w/ 10 levels "1","2","3","4",..: 1 2 1 7 1 7 1 1 1 1 ...
##  $ Mitoses        : Factor w/ 9 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 5 1 ...
##  $ Class          : Factor w/ 2 levels "benign","malignant": 1 1 1 1 1 2 1 1 1 1 ...
```

```r
length(which(is.na(BreastCancer)))
```

```
## [1] 16
```

```
library(mice)
```

```
##
## Attaching package: 'mice'

## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```
BreastCancer <- na.omit(BreastCancer)
BreastCancer <- select(BreastCancer,-c(1))
set.seed(2020)
library(caTools)#Package has split function which is used to split our dataset into training and test d
split=sample.split(BreastCancer, SplitRatio = 0.7)  # Splitting data into training and test dataset
trg_set=subset(BreastCancer,split==TRUE)  # Training dataset
test_set=subset(BreastCancer,split==FALSE)# Test dataset
# Implementing RandomForest
library(randomForest)
```

```
## randomForest 4.6-14

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

##
## Attaching package: 'randomForest'

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

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

```
model1_rf <- randomForest(Class ~., data = trg_set)
model1_rf
```

```
##
## Call:
##  randomForest(formula = Class ~ ., data = trg_set)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 3.14%
## Confusion matrix:
##           benign malignant class.error
## benign       306        10  0.03164557
## malignant      5       156  0.03105590
```

```
#Sspecifying mtry values as 2,6,8
model2_rf <- randomForest(Class ~., data = trg_set,mtry=c(2,6,8))
model2_rf
```

```
##
## Call:
##  randomForest(formula = Class ~ ., data = trg_set, mtry = c(2,      6, 8))
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
## 
##         OOB estimate of  error rate: 2.73%
## Confusion matrix:
##           benign malignant class.error
## benign       307         9  0.02848101
## malignant      4       157  0.02484472
```

```
probs <-predict(model2_rf,test_set,type="prob")
head(probs)
```

```
##    benign malignant
## 1   1.000     0.000
## 3   1.000     0.000
## 10  1.000     0.000
## 11  1.000     0.000
## 13  0.438     0.562
## 20  1.000     0.000
```

```
pred_class <-predict(model2_rf,test_set)
head(pred_class)
```

```
##         1         3        10        11        13        20
##    benign    benign    benign    benign malignant    benign
## Levels: benign malignant
```

```
(conf_matrix_forest <- table(pred_class,test_set$Class))
```

```
## 
## pred_class  benign malignant
##   benign       126         2
##   malignant      2        76
```

```
confusionMatrix(conf_matrix_forest)
```

```
## Confusion Matrix and Statistics
## 
## 
## pred_class  benign malignant
##   benign       126         2
##   malignant      2        76
## 
##                Accuracy : 0.9806
##                  95% CI : (0.951, 0.9947)
##     No Information Rate : 0.6214
##     P-Value [Acc > NIR] : <2e-16
## 
##                   Kappa : 0.9587
## 
##  Mcnemar's Test P-Value : 1
## 
##             Sensitivity : 0.9844
##             Specificity : 0.9744
##          Pos Pred Value : 0.9844
##          Neg Pred Value : 0.9744
##              Prevalence : 0.6214
##          Detection Rate : 0.6117
##    Detection Prevalence : 0.6214
##       Balanced Accuracy : 0.9794
##
##        'Positive' Class : benign
##
```

```r
library(gmodels)
CrossTable(pred_class,test_set$Class,digits = TRUE)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  206
##
##
##              | test_set$Class
##   pred_class |    benign | malignant | Row Total |
## -------------|-----------|-----------|-----------|
##       benign |       126 |         2 |       128 |
##              |      27.1 |      44.5 |           |
##              |       1.0 |       0.0 |       0.6 |
##              |       1.0 |       0.0 |           |
##              |       0.6 |       0.0 |           |
## -------------|-----------|-----------|-----------|
##    malignant |         2 |        76 |        78 |
##              |      44.5 |      73.1 |           |
##              |       0.0 |       1.0 |       0.4 |
##              |       0.0 |       1.0 |           |
##              |       0.0 |       0.4 |           |
## -------------|-----------|-----------|-----------|
## Column Total |       128 |        78 |       206 |
##              |       0.6 |       0.4 |           |
## -------------|-----------|-----------|-----------|
##
##
```