Untitled

Preet Dabhi

3/17/2022

```
#####################################################
#  Company    : Stevens
#  Project    : HW_05
#  Purpose    : classification and regression tree
#  First Name : Preet
#  Last Name  : Dabhi
#  Id               : 10459151
#  Date       : 03/17/2022

## Delete all the objects from your R- environment.
#####################################################

library(class)
library(rpart)

rm(list=ls())

dataFrame <- read.csv("D:/SEM 3/CS 513/HW_02/breast-cancer-wisconsin.csv",header=TRUE, sep=",")
head(dataFrame, n=5)
```

```
##    Sample F1 F2 F3 F4 F5 F6 F7 F8 F9 Class
## 1 1000025  5  1  1  1  2  1  3  1  1     2
## 2 1002945  5  4  4  5  7 10  3  2  1     2
## 3 1015425  3  1  1  1  2  2  3  1  1     2
## 4 1016277  6  8  8  1  3  4  3  7  1     2
## 5 1017023  4  1  1  3  2  1  3  1  1     2
```

```
#Summarizing each column (e.g. min, max, mean )
summary(dataFrame)
```

```
##      Sample               F1              F2              F3
##  Min.   :   61634   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000
##  1st Qu.:  870688   1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 1.000
##  Median : 1171710   Median : 4.000   Median : 1.000   Median : 1.000
##  Mean   : 1071704   Mean   : 4.418   Mean   : 3.134   Mean   : 3.207
##  3rd Qu.: 1238298   3rd Qu.: 6.000   3rd Qu.: 5.000   3rd Qu.: 5.000
##  Max.   :13454352   Max.   :10.000   Max.   :10.000   Max.   :10.000
##        F4               F5              F6                F7
##  Min.   : 1.000   Min.   : 1.000   Length:699         Min.   : 1.000
##  1st Qu.: 1.000   1st Qu.: 2.000   Class :character   1st Qu.: 2.000
##  Median : 1.000   Median : 2.000   Mode  :character   Median : 3.000
```

```
##  Mean   : 2.807   Mean   : 3.216                    Mean   : 3.438
##  3rd Qu.: 4.000   3rd Qu.: 4.000                    3rd Qu.: 5.000
##  Max.   :10.000   Max.   :10.000                    Max.   :10.000
##        F8               F9             Class
##  Min.   : 1.000   Min.   : 1.000   Min.   :2.00
##  1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.:2.00
##  Median : 1.000   Median : 1.000   Median :2.00
##  Mean   : 2.867   Mean   : 1.589   Mean   :2.69
##  3rd Qu.: 4.000   3rd Qu.: 1.000   3rd Qu.:4.00
##  Max.   :10.000   Max.   :10.000   Max.   :4.00
```

```
#Here we can see that by running summary on the dataframe F6 column there are some missing values in it
n <- as.numeric(as.character(dataFrame$F6))
```

```
## Warning: NAs introduced by coercion
```

```
summary(n, na.rm = TRUE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   1.000   1.000   3.545   6.000  10.000      16
```

```
dataFrame$F6 <- n
summary(n, na.rm = TRUE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   1.000   1.000   3.545   6.000  10.000      16
```

```
#Check the number of rows before removing
nrow(dataFrame)
```

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

```
#Remove the rows with missing values
dataFrame <- na.omit(dataFrame)
View(dataFrame)

#Check the number of rows after removing
nrow(dataFrame)
```

```
## [1] 683
```

```
#Replacing class column 2 and 4 with Benign and Malignant
dataFrame$Class<- factor(dataFrame$Class , levels = c("2","4") , labels = c("Benign","Malignant"))
is.factor(dataFrame$Class)
```

```
## [1] TRUE
```

```r
#Generate train and test in the ratio 70% to 30%
dataFrame<- dataFrame[2:11]
size <- floor(0.70 * nrow(dataFrame))

#Set the seed to make your partition reproducible
set.seed(123)
trainData <- sample(seq_len(nrow(dataFrame)), size = size)

#Loading 70% Breast cancer record in training dataset
training <- dataFrame[trainData, ]

#Loading 30% Breast cancer in test dataset
test <- dataFrame[-trainData, ]

#Implementing CART
cart <- rpart(Class ~ ., data = training, method = "class")

#Predicting class for test set
predicted <- predict(cart, test, type = "class")
print(length(predicted))
```

```
## [1] 205
```

```r
print(length(test$Class))
```

```
## [1] 205
```

```r
#Confusion Matrix
conf_matrix <- table(predicted,test$Class)
print(conf_matrix)
```

```
##
## predicted   Benign Malignant
##   Benign       136         9
##   Malignant      3        57
```

```r
#Accuracy
accuracy <- function(x){sum(diag(x)/(sum(rowSums(x)))) * 100}
accuracy(conf_matrix)
```

```
## [1] 94.14634
```