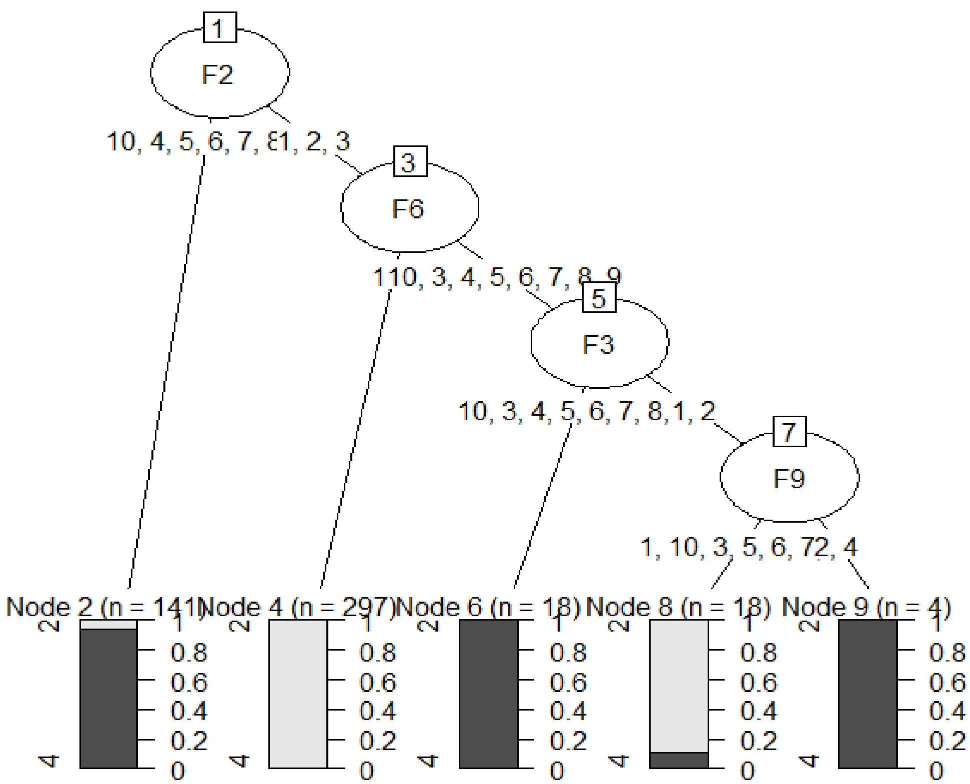
Assignment 6

6.1: Use the C5.0 methodology to develop a classification model for the Diagnosis.

```
#  First Name: Utsav
#  Last Name: Italiya
#  Id : 10475248
rm(list=ls ())
library(C50)

#read data from csv file
df<-read.csv("F:/Sem1/CS513/lecture6/breast-cancer-wisconsin.csv",
             na.strings = c("?") ,
             colClasses=c("Sample"="character",
                    "F1"="factor","F2"="factor",
                    "F3"="factor","F4"="factor",
                    "F5"="factor","F6"="factor",
                    "F7"="factor","F8"="factor",
                    "F9"="factor","Class"="factor"))

df <-na.omit(df)

#70% training and 30% testing data
idx<-sort(sample(nrow(df),as.integer(.70*nrow(df))))
training<-df[idx,]
testing<-df[-idx,]
```

```
#ploting c50 model
c50 <- C5.0(Class~. , training[,-1])
summary(c50)
plot(c50)

#prediction
prediction<-predict(c50,testing[,-1],type="class")
table(testing[,11],prediction)
wrong<-sum(testing[,11]!=prediction)
error_rate<-wrong/length(testing[,11])
```




```
error_rate
```

```
Decision tree:

F2 in {10,4,5,6,7,8,9}: 4 (141/8)
F2 in {1,2,3}:
:...F6 in {1,2}: 2 (297/2)
    F6 in {10,3,4,5,6,7,8,9}:
    :...F3 in {10,3,4,5,6,7,8,9}: 4 (18)
        F3 in {1,2}:
        :...F9 in {1,10,3,5,6,7,8}: 2 (18/2)
            F9 in {2,4}: 4 (4)


Evaluation on training data (478 cases):

            Decision Tree
          ----------------
          Size      Errors

             5   12( 2.5%)   <<


           (a)   (b)    <-classified as
          ----  ----
           311     8    (a): class 2
             4   155    (b): class 4


        Attribute usage:

        100.00% F2
         70.50% F6
          8.37% F3
          4.60% F9


Time: 0.0 secs

> plot(c50)
> #prediction
> prediction<-predict(c50,testing[,-1],type="class")
> table(testing[,11],prediction)
   prediction
      2   4
  2 119   6
  4   1  79
> wrong<-sum(testing[,11]!=prediction)
> error_rate<-wrong/length(testing[,11])
> error_rate
[1] 0.03414634
```

6.2: Use the Random Forest methodology to develop a classification model for the Diagnosis and identify important features.

```
######################################################
#random forest
######################################################
rm(list=ls())
library(C50)

#read data from csv file
df<-read.csv("F:/Sem1/CS513/lecture6/breast-cancer-wisconsin.csv",
        na.strings = c("?") ,
        colClasses=c("Sample"="character",
                "F1"="factor","F2"="factor",
                "F3"="factor","F4"="factor",
                "F5"="factor","F6"="factor",
                "F7"="factor","F8"="factor",
                "F9"="factor","Class"="factor"))

df <-na.omit(df)

#70% training and 30% testing data
idx<-sort(sample(nrow(df),as.integer(.70*nrow(df))))
training<-df[idx,]
testing<-df[-idx,]

#ploting random model
library(randomForest)
```

rm <- randomForest(Class~., data=training, importance=TRUE, ntree=1000)

importance(rm)

varImpPlot(rm)

#predictions

prediction<- predict(rm, testing)

table(actual=testing[,11],prediction)

wrong<-sum(testing$Class!=prediction)

error_rate<-wrong/length(testing[,11])

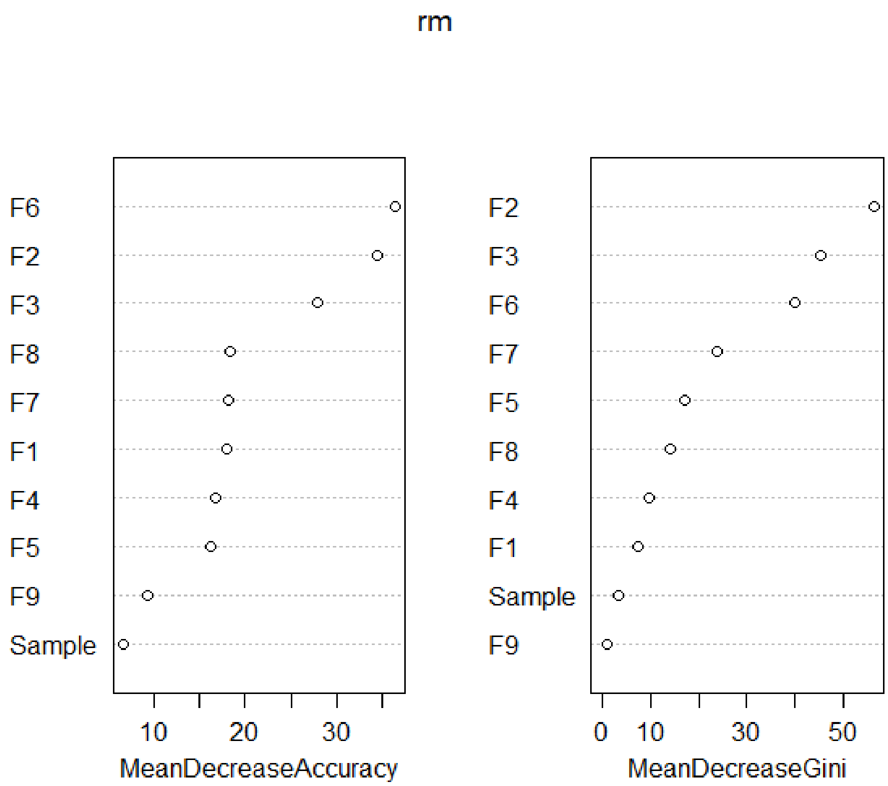
error_rate

#succection rate

successrate <- 1 - error_rate

successrate

```
> importance(rm)
               2         4 MeanDecreaseAccuracy MeanDecreaseGini
Sample  7.631933  3.181955             6.692353        3.1424116
F1     14.408876 15.510997            17.958951        7.4807614
F2     27.310332 22.028779            34.400791       56.3741921
F3     17.163249 22.846787            27.911415       45.2431410
F4     14.509674 10.148260            16.795147        9.5314844
F5     13.891581  9.584173            16.259890       17.1660151
F6     25.456194 29.657522            36.318851       39.8410798
F7     12.499458 14.015139            18.113002       23.7312303
F8     17.022246  9.683699            18.398700       14.1573602
F9      8.830352  1.444527             9.243390        0.9593622
> varImpPlot(rm)
> #predictions
> prediction<- predict(rm, testing)
> table(actual=testing[,11],prediction)
      prediction
actual   2   4
     2 131   4
     4   3  67
> wrong<-sum(testing$Class!=prediction)
> error_rate<-wrong/length(testing[,11])
> error_rate
[1] 0.03414634
> #succection rate
> successrate <- 1 - error_rate
> successrate
[1] 0.9658537
> 
```


rm
F6
F2
F3
F8
F7
F1
F4
F5
F9
Sample
10
20
30
MeanDecreaseAccuracy
F2
F3
F6
F7
F5
F8
F4
F1
Sample
F9
0
10
30
50
MeanDecreaseGini