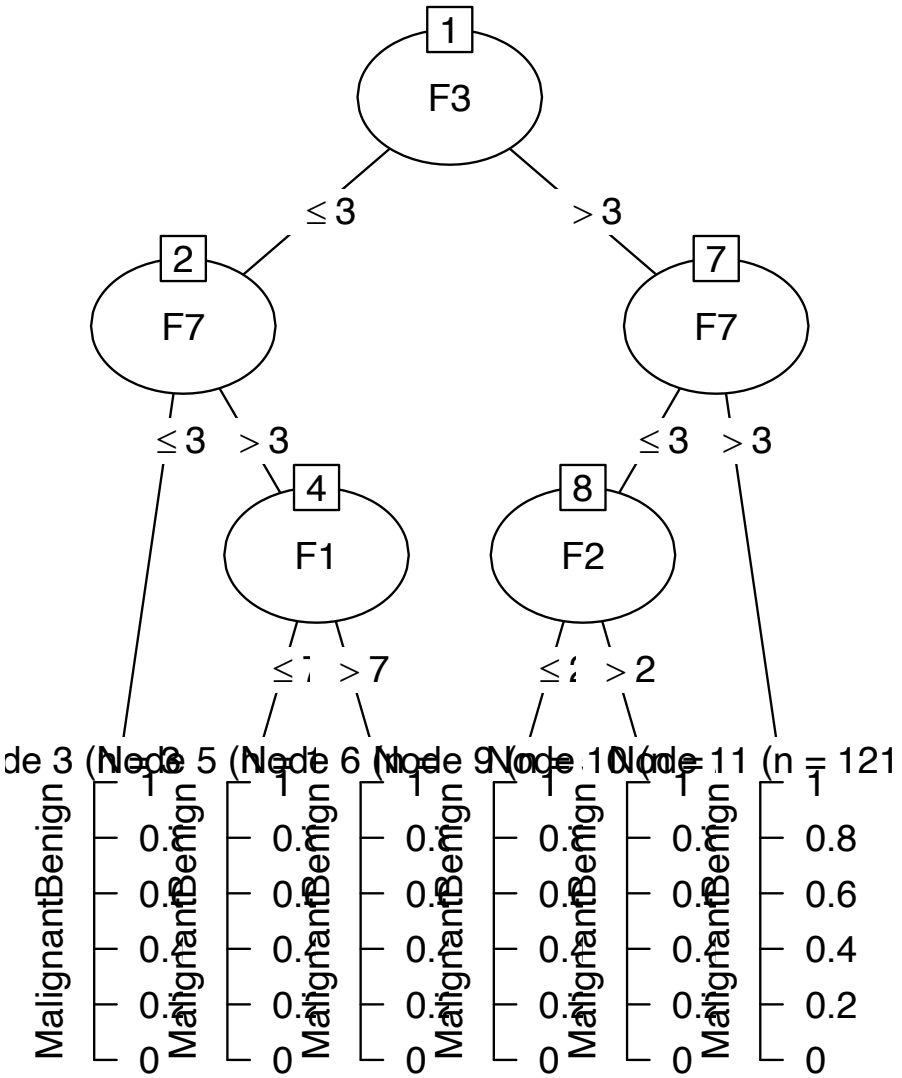
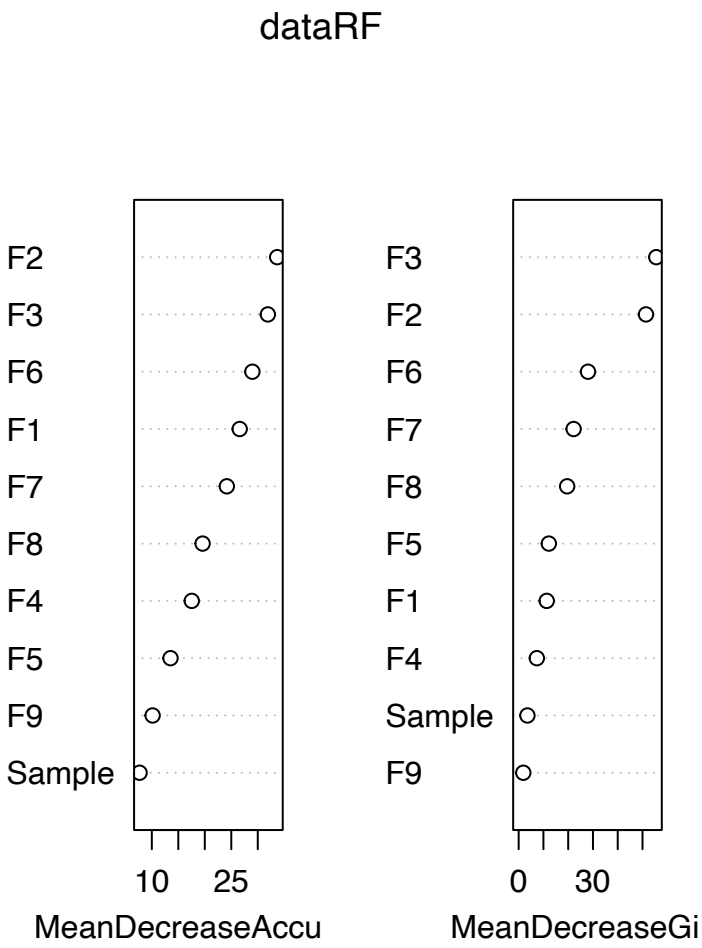
```
######################################################################
#UNIVERSITY : STEVENS INSTITUE OF TECHNOLOGY
#Project : HW 06 C50 and RF Decision Tree
#Purpose : Use the C50 & Decision Tree methodology to develop a
classification models
#First Name : Sarthak
#Last Name : Ahir
#CWID : 10479028
#Date : 11/15/2021
######################################################################
rm(list=ls())


#installing the required libraries
library(C50)

#Load the “breast-cancer-wisconsin.csv” from canvas into R and perform the
analysis

newDataSet = read.csv("~/Downloads/breast-cancer-wisconsin.csv",na.strings
= '?')


#Summarizing each column
summary(newDataSet)
table(newDataSet$Class)

View(newDataSet)


#Convert labels to factor class
newDataSet$Class<- factor(newDataSet$Class , levels = c("2","4") , labels
= c("Benign","Malignant"))

#Omitting the NA values
newDataSet<-na.omit(newDataSet)


# Splitting the newDataSet Data to test and training

Data<-sort(sample(nrow(newDataSet),as.integer(.70*nrow(newDataSet))))

training<-newDataSet[Data,]

test<-newDataSet[-Data,]

dev.off()

# Implementing C - 5.0

C50<-C5.0(Class~.,training[,-1])

summary(C50)

plot(C50)

#predicting using the created test data
```

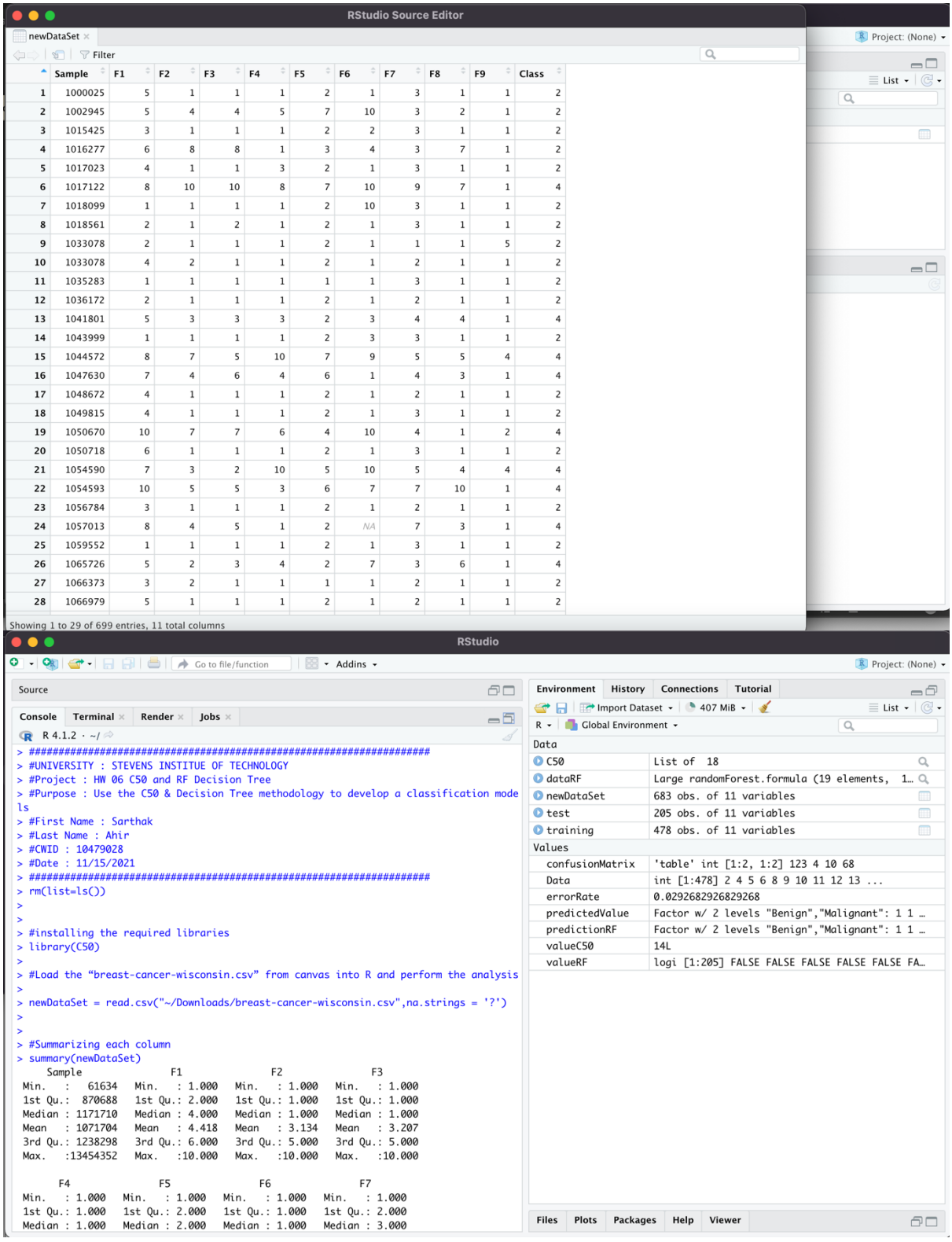
```
predictedValue<-predict(C50,test[,-1],type="class")

#Generating the confusion matrix

confusionMatrix<-table(test[,11],predictedValue)

confusionMatrix
str(predictedValue)

# error rate

valueC50<-sum(test[,11]!=predictedValue)
errorRate<-valueC50/length(test[,11])

errorRate
# error rate in percent
print(paste("the error rate is",errorRate*100))

#######   Random Forest   #######

#installing the required libraries
library(randomForest)

# Applying random forest
dataRF <- randomForest( Class~., data=training, importance=TRUE,
ntree=1000)
importance(dataRF)
varImpPlot(dataRF)

#predicting using the created test data for Random Forest

predictionRF <- predict(dataRF, test)
table(actual=test$Class,predictionRF)

# error rate

valueRF<- (test$Class!=predictionRF )

errorRate<-sum(valueRF)/length(valueRF)
errorRate
# error rate in percent
print(paste("the error rate is",errorRate*100))



##install.packages('XQuartz',repos='http://cran.us.r-project.org')
```

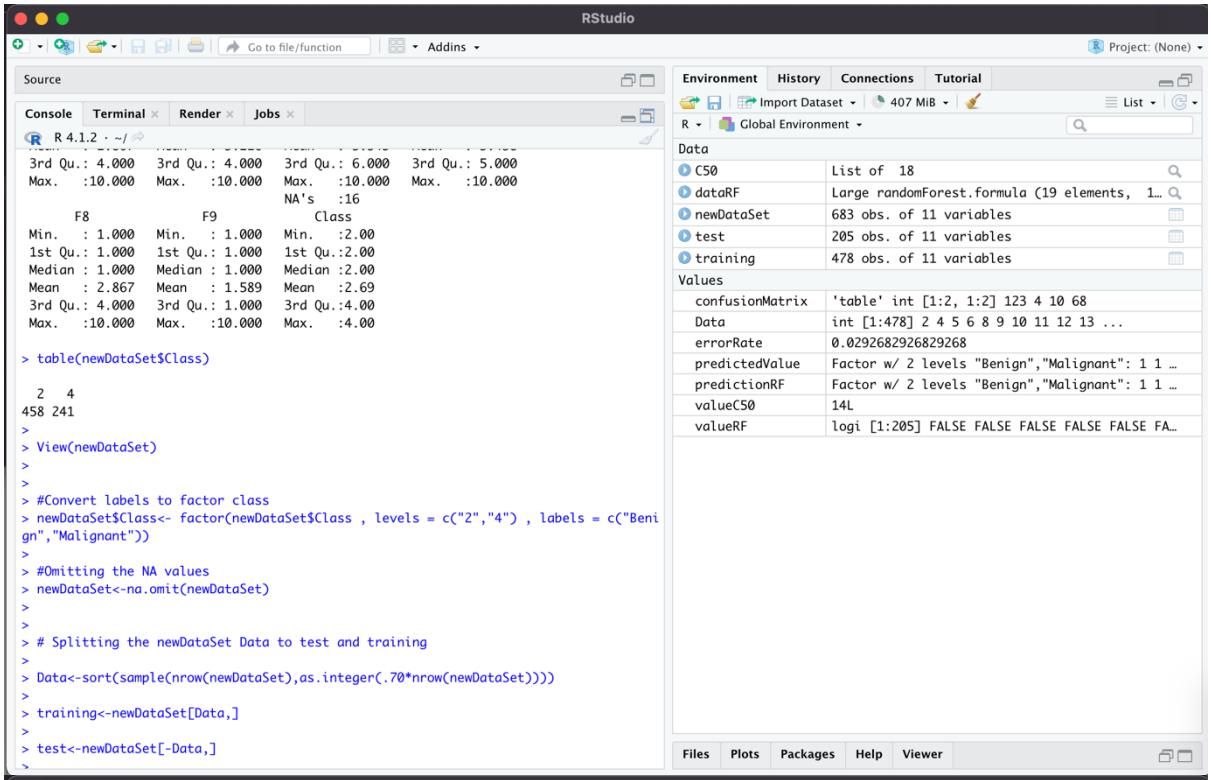
OUTPUT –
C 50

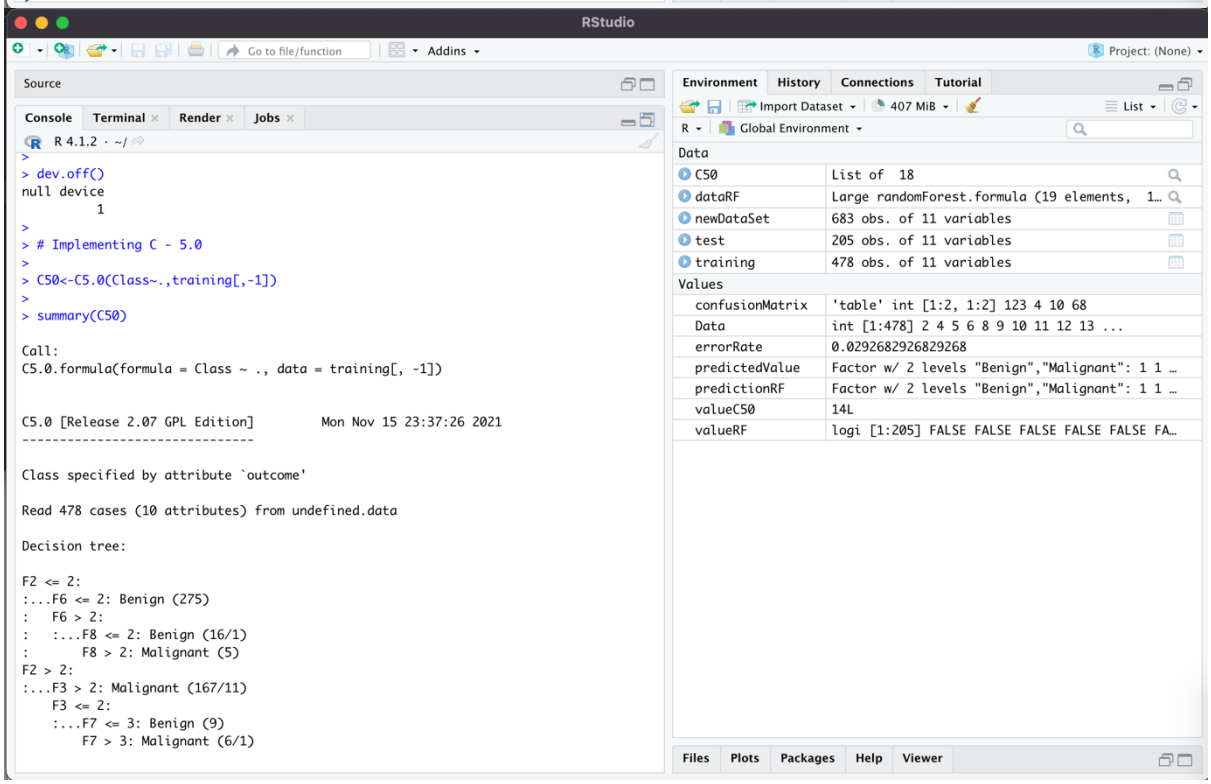


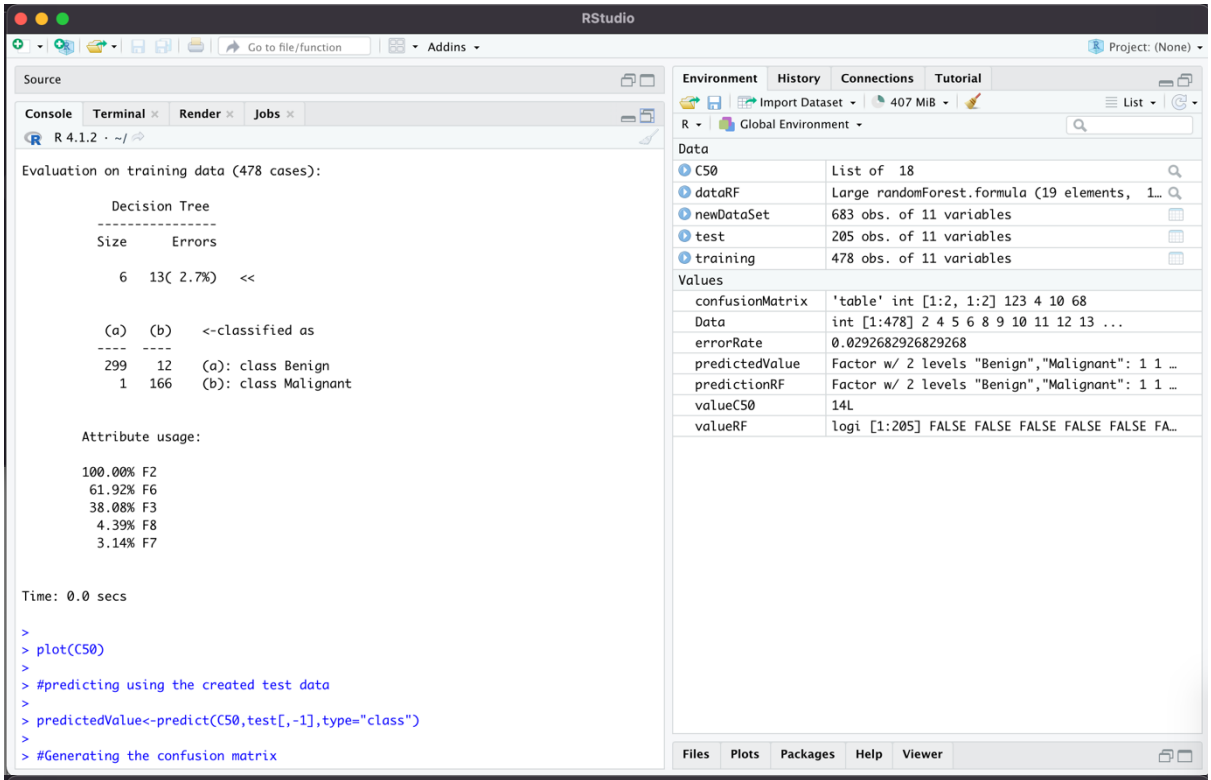
For Random forest Decision tree

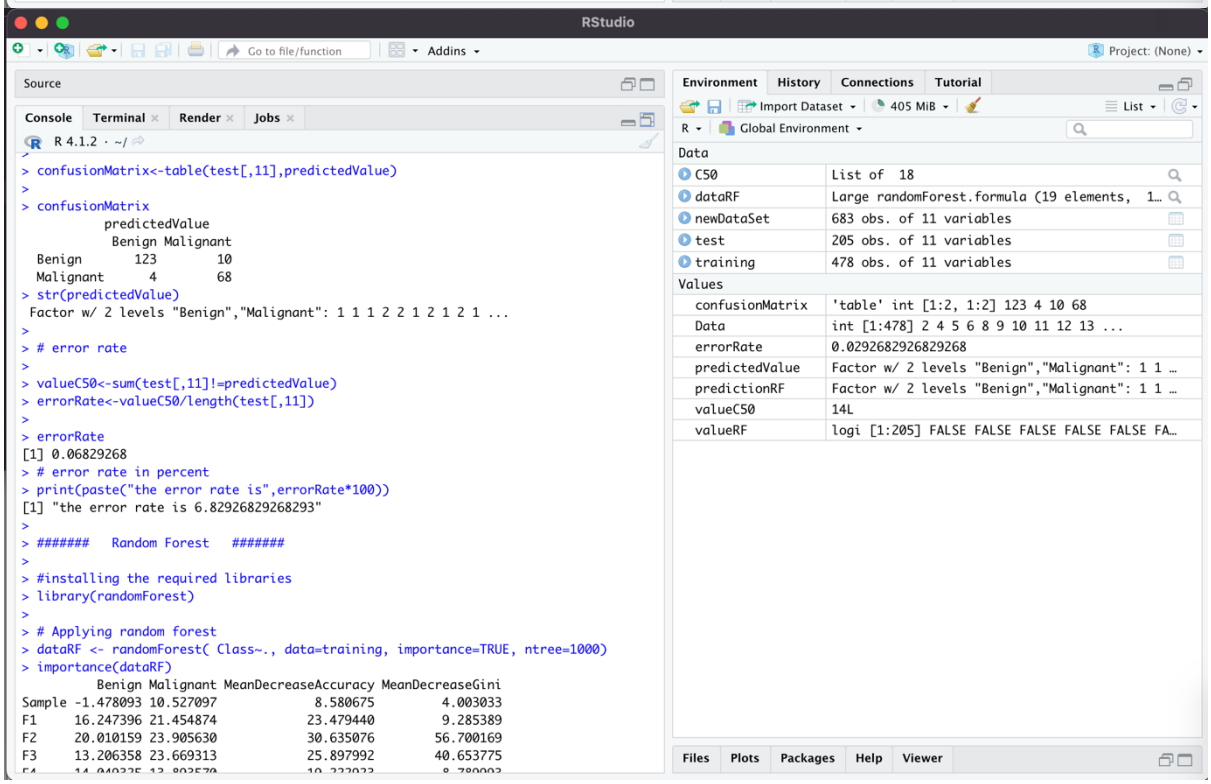

dataRF
F2
F3
F6
F1
F7
F8
F4
F5
F9
Sample
10
25
MeanDecreaseAccu
F3
F2
F6
F7
F8
F5
F1
F4
Sample
F9
0
30
MeanDecreaseGi

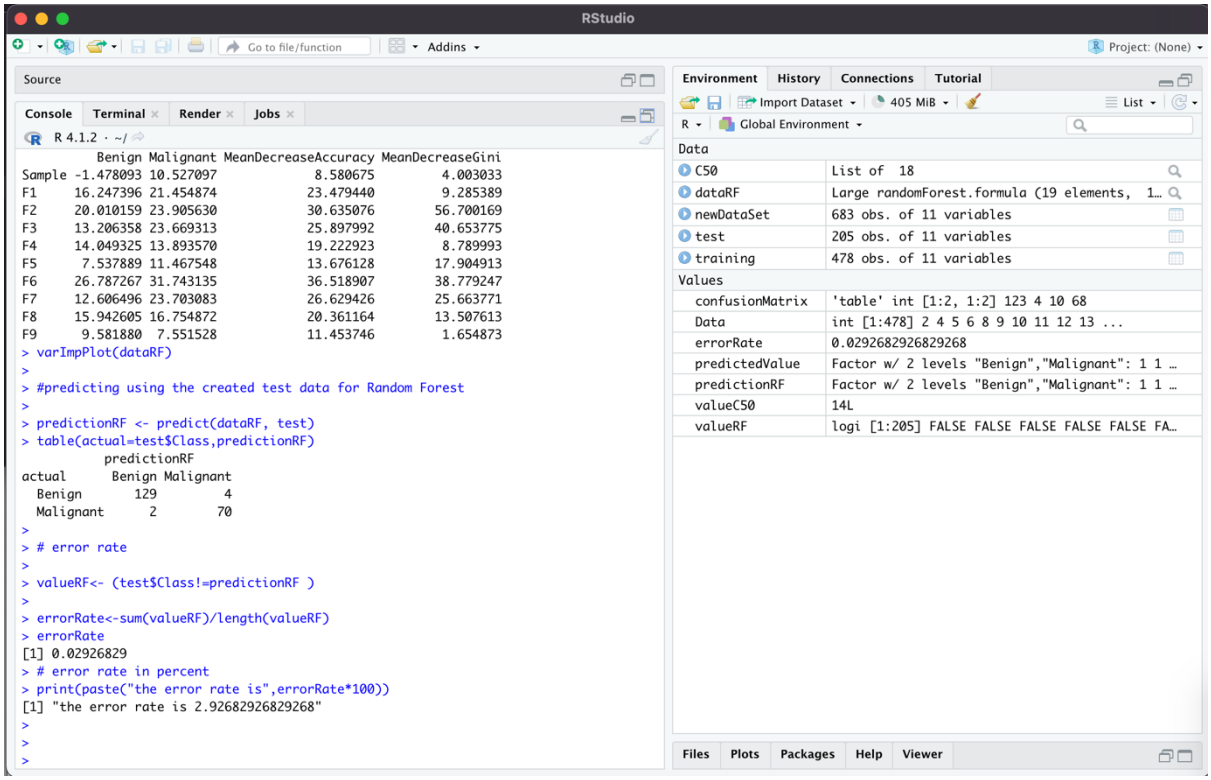

RStudio Source Editor
newDataSet
Filter
Showing 1 to 29 of 699 entries, 11 total columns
RStudio
Project: (None)
Source
Console
Terminal
Render
Jobs
R 4.1.2 · ~/
> ##
> #UNIVERSITY : STEVENS INSTITUE OF TECHNOLOGY
> #Project : HW 06 C50 and RF Decision Tree
> #Purpose : Use the C50 & Decision Tree methodology to develop a classification models
> #First Name : Sarthak
> #Last Name : Ahir
> #CWID : 10479028
> #Date : 11/15/2021
> ##
> rm(list=ls())
> #installing the required libraries
> library(C50)
> #Load the “breast-cancer-wisconsin.csv” from canvas into R and perform the analysis
> newDataSet = read.csv("~/Downloads/breast-cancer-wisconsin.csv",na.strings = '?')
> #Summarizing each column
> summary(newDataSet)
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 Min. : 1.000 Min. : 1.000
1st Qu.: 1.000 1st Qu.: 2.000 1st Qu.: 1.000 1st Qu.: 2.000
Median : 1.000 Median : 2.000 Median : 1.000 Median : 3.000
Environment
History
Connections
Tutorial
Import Dataset
407 MiB
List
R
Global Environment
Data
C50 List of 18
dataRF Large randomForest.formula (19 elements, 1…
newDataSet 683 obs. of 11 variables
test 205 obs. of 11 variables
training 478 obs. of 11 variables
Values
confusionMatrix 'table' int [1:2, 1:2] 123 4 10 68
Data int [1:478] 2 4 5 6 8 9 10 11 12 13 ...
errorRate 0.0292682926829268
predictedValue Factor w/ 2 levels "Benign","Malignant": 1 1 …
predictionRF Factor w/ 2 levels "Benign","Malignant": 1 1 …
valueC50 14L
valueRF logi [1:205] FALSE FALSE FALSE FALSE FALSE FA…
Files
Plots
Packages
Help
Viewer


RStudio
Source
Console
Terminal
Render
Jobs
R 4.1.2 · ~/
3rd Qu.: 4.000 3rd Qu.: 4.000 3rd Qu.: 6.000 3rd Qu.: 5.000
Max. :10.000 Max. :10.000 Max. :10.000 Max. :10.000
NA's :16
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
> table(newDataSet$Class)
2 4
458 241
> View(newDataSet)
> #Convert labels to factor class
> newDataSet$Class<- factor(newDataSet$Class , levels = c("2","4") , labels = c("Benign","Malignant"))
> #Omitting the NA values
> newDataSet<-na.omit(newDataSet)
> # Splitting the newDataSet Data to test and training
> Data<-sort(sample(nrow(newDataSet),as.integer(.70*nrow(newDataSet))))
> training<-newDataSet[Data,]
> test<-newDataSet[-Data,]
Environment
History
Connections
Tutorial
Import Dataset
407 MiB
List
R
Global Environment
Data
C50 List of 18
dataRF Large randomForest.formula (19 elements, 1…
newDataSet 683 obs. of 11 variables
test 205 obs. of 11 variables
training 478 obs. of 11 variables
Values
confusionMatrix 'table' int [1:2, 1:2] 123 4 10 68
Data int [1:478] 2 4 5 6 8 9 10 11 12 13 ...
errorRate 0.0292682926829268
predictedValue Factor w/ 2 levels "Benign","Malignant": 1 1 …
predictionRF Factor w/ 2 levels "Benign","Malignant": 1 1 …
valueC50 14L
valueRF logi [1:205] FALSE FALSE FALSE FALSE FALSE FA…
Files
Plots
Packages
Help
Viewer


RStudio
Source
Console
Terminal
Render
Jobs
R 4.1.2 · ~/
> dev.off()
null device
1
> # Implementing C - 5.0
> C50<-C5.0(Class~.,training[,-1])
> summary(C50)
Call:
C5.0.formula(formula = Class ~ ., data = training[, -1])
C5.0 [Release 2.07 GPL Edition] Mon Nov 15 23:37:26 2021
Class specified by attribute `outcome'
Read 478 cases (10 attributes) from undefined.data
Decision tree:
F2 <= 2:
:...F6 <= 2: Benign (275)
: F6 > 2:
: :...F8 <= 2: Benign (16/1)
: F8 > 2: Malignant (5)
F2 > 2:
:...F3 > 2: Malignant (167/11)
F3 <= 2:
:...F7 <= 3: Benign (9)
F7 > 3: Malignant (6/1)
Environment
History
Connections
Tutorial
Import Dataset
407 MiB
List
R
Global Environment
Data
C50 List of 18
dataRF Large randomForest.formula (19 elements, 1…
newDataSet 683 obs. of 11 variables
test 205 obs. of 11 variables
training 478 obs. of 11 variables
Values
confusionMatrix 'table' int [1:2, 1:2] 123 4 10 68
Data int [1:478] 2 4 5 6 8 9 10 11 12 13 ...
errorRate 0.0292682926829268
predictedValue Factor w/ 2 levels "Benign","Malignant": 1 1 …
predictionRF Factor w/ 2 levels "Benign","Malignant": 1 1 …
valueC50 14L
valueRF logi [1:205] FALSE FALSE FALSE FALSE FALSE FA…
Files
Plots
Packages
Help
Viewer


RStudio
Go to file/function
Addins
Project: (None)
Source
Console
Terminal
Render
Jobs
R 4.1.2 · ~/
Evaluation on training data (478 cases):
Decision Tree
Size Errors
6 13(2.7%) <<
(a) (b) <-classified as
299 12 (a): class Benign
1 166 (b): class Malignant
Attribute usage:
100.00% F2
61.92% F6
38.08% F3
4.39% F8
3.14% F7
Time: 0.0 secs
> plot(C50)
> #predicting using the created test data
> predictedValue<-predict(C50,test[,-1],type="class")
> #Generating the confusion matrix
Environment
History
Connections
Tutorial
Import Dataset
407 MiB
List
R
Global Environment
Data
C50 List of 18
dataRF Large randomForest.formula (19 elements, 1…
newDataSet 683 obs. of 11 variables
test 205 obs. of 11 variables
training 478 obs. of 11 variables
Values
confusionMatrix 'table' int [1:2, 1:2] 123 4 10 68
Data int [1:478] 2 4 5 6 8 9 10 11 12 13 ...
errorRate 0.0292682926829268
predictedValue Factor w/ 2 levels "Benign","Malignant": 1 1 …
predictionRF Factor w/ 2 levels "Benign","Malignant": 1 1 …
valueC50 14L
valueRF logi [1:205] FALSE FALSE FALSE FALSE FALSE FA…
Files
Plots
Packages
Help
Viewer


RStudio
Go to file/function
Addins
Project: (None)
Source
Console
Terminal
Render
Jobs
R 4.1.2 · ~/
> confusionMatrix<-table(test[,11],predictedValue)
> confusionMatrix
predictedValue
Benign Malignant
Benign 123 10
Malignant 4 68
> str(predictedValue)
Factor w/ 2 levels "Benign","Malignant": 1 1 1 2 2 1 2 1 2 1 ...
> # error rate
> valueC50<-sum(test[,11]!=predictedValue)
> errorRate<-valueC50/length(test[,11])
> errorRate
[1] 0.06829268
> # error rate in percent
> print(paste("the error rate is",errorRate*100))
[1] "the error rate is 6.82926829268293"
> ####### Random Forest #######
> #installing the required libraries
> library(randomForest)
> # Applying random forest
> dataRF <- randomForest(Class~., data=training, importance=TRUE, ntree=1000)
> importance(dataRF)
Benign Malignant MeanDecreaseAccuracy MeanDecreaseGini
Sample -1.478093 10.527097 8.580675 4.003033
F1 16.247396 21.454874 23.479440 9.285389
F2 20.010159 23.905630 30.635076 56.700169
F3 13.206358 23.669313 25.897992 40.653775
Environment
History
Connections
Tutorial
Import Dataset
405 MiB
List
R
Global Environment
Data
C50 List of 18
dataRF Large randomForest.formula (19 elements, 1…
newDataSet 683 obs. of 11 variables
test 205 obs. of 11 variables
training 478 obs. of 11 variables
Values
confusionMatrix 'table' int [1:2, 1:2] 123 4 10 68
Data int [1:478] 2 4 5 6 8 9 10 11 12 13 ...
errorRate 0.0292682926829268
predictedValue Factor w/ 2 levels "Benign","Malignant": 1 1 …
predictionRF Factor w/ 2 levels "Benign","Malignant": 1 1 …
valueC50 14L
valueRF logi [1:205] FALSE FALSE FALSE FALSE FALSE FA…
Files
Plots
Packages
Help
Viewer


RStudio
Go to file/function
Addins
Project: (None)
Source
Console
Terminal
Render
Jobs
R 4.1.2 · ~/
Benign Malignant MeanDecreaseAccuracy MeanDecreaseGini
Sample -1.478093 10.527097 8.580675 4.003033
F1 16.247396 21.454874 23.479440 9.285389
F2 20.010159 23.905630 30.635076 56.700169
F3 13.206358 23.669313 25.897992 40.653775
F4 14.049325 13.893570 19.222923 8.789993
F5 7.537889 11.467548 13.676128 17.904913
F6 26.787267 31.743135 36.518907 38.779247
F7 12.606496 23.703083 26.629426 25.663771
F8 15.942605 16.754872 20.361164 13.507613
F9 9.581880 7.551528 11.453746 1.654873
> varImpPlot(dataRF)
>
> #predicting using the created test data for Random Forest
>
> predictionRF <- predict(dataRF, test)
> table(actual=test$Class,predictionRF)
predictionRF
actual Benign Malignant
Benign 129 4
Malignant 2 70
>
> # error rate
>
> valueRF<- (test$Class!=predictionRF)
>
> errorRate<-sum(valueRF)/length(valueRF)
> errorRate
[1] 0.02926829
> # error rate in percent
> print(paste("the error rate is",errorRate*100))
[1] "the error rate is 2.92682926829268"
>
>
>
Environment
History
Connections
Tutorial
Import Dataset
405 MiB
List
R
Global Environment
Data
C50 List of 18
dataRF Large randomForest.formula (19 elements, 1…
newDataSet 683 obs. of 11 variables
test 205 obs. of 11 variables
training 478 obs. of 11 variables
Values
confusionMatrix 'table' int [1:2, 1:2] 123 4 10 68
Data int [1:478] 2 4 5 6 8 9 10 11 12 13 ...
errorRate 0.0292682926829268
predictedValue Factor w/ 2 levels "Benign","Malignant": 1 1 …
predictionRF Factor w/ 2 levels "Benign","Malignant": 1 1 …
valueC50 14L
valueRF logi [1:205] FALSE FALSE FALSE FALSE FALSE FA…
Files
Plots
Packages
Help
Viewer