CODE USED:

```
############ CODE FOR DECISION TREE USING TREE PACKAGE ######################

install.packages("tree")
library(tree)
Bank = read.csv2("bank-full.csv")
temp = Bank
head(Bank)
setSize <- floor(0.67 * nrow(Bank))
set.seed(123) #set a seed for being able to replicate
rowIndices <- sample(seq_len(nrow(Bank)), size = setSize)
trainBank <- Bank[rowIndices, ]
testBank <- Bank[-rowIndices, ]
dtreeModel = tree(y ~., data = trainBank, split = c("gini"))
summary(dtreeModel)
names(dtreeModel)
dtreeModel$y
plot(dtreeModel)
text(dtreeModel, pos=3, cex=0.7, col = 'blue')

############## CODE FOR DECISION TREE USING RPART PACKAGE ###################

library(rpart)
install.packages("rpart.plot")
library(rpart.plot)
Bank = read.csv2("bank-full.csv")
head(Bank)
setSize <- floor(0.67 * nrow(Bank))
set.seed(123) #set a seed for being able to replicate
rowIndices <- sample(seq_len(nrow(Bank)), size = setSize)
trainBank <- Bank[rowIndices, ]
testBank <- Bank[-rowIndices, ]
dtreeModel2 = rpart(y ~., data = trainBank, method = 'class', parms = list(split="gini"))
summary(dtreeModel2)
names(dtreeModel2)
dtreeModel2$variable.importance
rpart.plot(dtreeModel2,extra=1, varlen=0)
```

```
###################### CODE FOR RANDOM FOREST ############################


install.packages("randomForest")
library(randomForest)
Bank = read.csv2("bank-full.csv")
rfModel = randomForest(formula = y~., data = Bank, ntree = 250, importance = TRUE,
replace=TRUE)
summary(rfModel)
names(rfModel)
rfModel$confusion


#########################################################################
setSize <- floor(0.67 * nrow(Bank))
set.seed(123) #set a seed for being able to replicate
rowIndices <- sample(seq_len(nrow(Bank)), size = setSize)
trainBank <- Bank[rowIndices, ]
testBank <- Bank[-rowIndices, ]

rfModel2 = randomForest(formula = y~., data = trainBank, ntree = 500, mtry = 2, importance
= TRUE, replace=TRUE, proximity=TRUE, sampsize=c(500,400))
rfModel2pred <- predict(object = rfModel2, newdata = testBank[,-4])
table(observed = testBank$y, predicted = rfModel2pred)
rfModel2$confusion
rfModel2
par(mfrow=c(1,2))
varImpPlot(rfModel2,main='Variable Importance Plot: Final Model',pch=16,col='blue')

#########################################################################
```

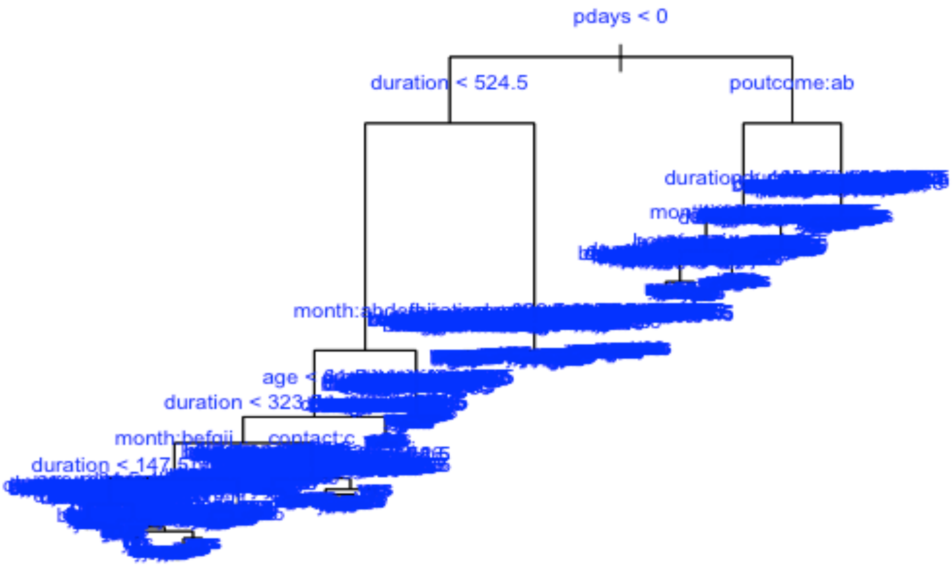
RESULTS:

1) **SUMMARY OF THE DECISION TREE MODEL FOR VARIABLE "Y" USING TREE PACKAGE**

```
Classification tree:
tree(formula = y ~ ., data = trainBank, split = c("gini"))
Variables actually used in tree construction:
 [1] "pdays"     "duration"  "month"     "age"
 [5] "education" "balance"   "housing"   "job"
 [9] "day"       "contact"   "campaign"  "marital"
[13] "loan"      "poutcome"  "previous"
Number of terminal nodes:  1354
Residual mean deviance:  0.2284 = 6610 / 28940
Misclassification error rate: 0.05946 = 1801 / 30291
```

(Here, the decision tree contains 1354 nodes created by means of the Gini index. The Residual Mean Deviance shows how well the response is predicted by the model and the Misclassification error rate seems to be low at 0.059% and only 1801 entries are misclassified out of a total of 30291)

2) **PLOT OF DECISION TREE MODEL USING TREE PACKAGE**



3) SUMMARY OF DECISION TREE MODEL USING RPART PACKAGE

```
Call:
rpart(formula = y ~ ., data = trainBank, method = "class", parms = list(split =
"gini"))
  n= 30291

          CP nsplit rel error    xerror       xstd
1 0.04249930      0 1.0000000 1.0000000 0.01576195
2 0.02420490      3 0.8725021 0.8750352 0.01486617
3 0.01547988      4 0.8482972 0.8528005 0.01469739
4 0.01000000      5 0.8328173 0.8404165 0.01460205

Variable importance
duration poutcome
      60       40

Node number 1: 30291 observations,    complexity param=0.0424993
  predicted class=no   expected loss=0.1172956  P(node) =1
    class counts: 26738  3553
   probabilities: 0.883 0.117
  left son=2 (27019 obs) right son=3 (3272 obs)
  Primary splits:
      duration < 524.5   to the left,   improve=784.8708, (0 missing)
      poutcome splits as  LLRL,         improve=622.8070, (0 missing)
      month    splits as  LLRLLLLRLLRR, improve=365.8911, (0 missing)
      pdays    < 8.5     to the left,   improve=189.2370, (0 missing)
      previous < 0.5     to the left,   improve=186.2596, (0 missing)

Node number 2: 27019 observations,    complexity param=0.0424993
  predicted class=no   expected loss=0.07768607  P(node) =0.8919811
    class counts: 24920  2099
   probabilities: 0.922 0.078
  left son=4 (26142 obs) right son=5 (877 obs)
  Primary splits:
      poutcome splits as  LLRL,         improve=572.5645, (0 missing)
      month    splits as  LLRLLLLRLLRR, improve=361.0196, (0 missing)
      pdays    < 8.5     to the left,   improve=182.7977, (0 missing)
      previous < 0.5     to the left,   improve=180.3616, (0 missing)
      duration < 205.5   to the left,   improve=149.0770, (0 missing)
  Surrogate splits:
      age < 91      to the left,  agree=0.968, adj=0.001, (0 split)

Node number 3: 3272 observations,    complexity param=0.0424993
  predicted class=no   expected loss=0.4443765  P(node) =0.1080189
    class counts:  1818  1454
   probabilities: 0.556 0.444 
  left son=6 (2020 obs) right son=7 (1252 obs)
  Primary splits:
      duration < 807.5   to the left,   improve=78.01715, (0 missing)
      contact  splits as  RRL,          improve=43.62075, (0 missing)
      poutcome splits as  LLRL,         improve=37.69423, (0 missing)
      marital  splits as  RLR,          improve=20.69620, (0 missing)
      month    splits as  LRRLLLLRLLRR, improve=19.20534, (0 missing)
  Surrogate splits:
      balance  < -1170.5 to the right, agree=0.618, adj=0.001, (0 split)
      campaign < 23.5    to the left,  agree=0.618, adj=0.001, (0 split)
      previous < 17.5    to the left,  agree=0.618, adj=0.001, (0 split)

Node number 4: 26142 observations
  predicted class=no   expected loss=0.05883253  P(node) =0.8630286
    class counts: 24604  1538
   probabilities: 0.941 0.059 

Node number 5: 877 observations,    complexity param=0.0242049
  predicted class=yes  expected loss=0.3603193  P(node) =0.02895249
    class counts:   316   561
   probabilities: 0.360 0.640 
  left son=10 (168 obs) right son=11 (709 obs)
  Primary splits:
      duration < 132.5   to the left,   improve=65.054580, (0 missing)
      housing  splits as  RL,           improve=14.114960, (0 missing)
      month    splits as  LRRRLRRRLLRR, improve=12.006420, (0 missing)
      job      splits as  LLLRRRRRRLRR, improve= 8.712319, (0 missing)
      pdays    < 85.5    to the left,   improve= 6.154954, (0 missing)
  Surrogate splits:
      contact  splits as  RRL,         agree=0.814, adj=0.030, (0 split)
      pdays    < 606     to the right, agree=0.811, adj=0.012, (0 split)
      default  splits as  RL,          agree=0.810, adj=0.006, (0 split)
      campaign < 6.5     to the right, agree=0.810, adj=0.006, (0 split)

Node number 6: 2020 observations,    complexity param=0.01547988
  predicted class=no   expected loss=0.3584158  P(node) =0.06668647
    class counts:  1296   724
   probabilities: 0.642 0.358 
  left son=12 (1931 obs) right son=13 (89 obs)
  Primary splits:
      poutcome splits as  LLRL,         improve=37.80239, (0 missing)
      contact  splits as  RRL,          improve=36.53672, (0 missing)
      pdays    < 0       to the left,   improve=21.10700, (0 missing)
      previous < 0.5     to the left,   improve=21.10700, (0 missing)
      job      splits as  RLRLRRRLRRRL, improve=20.54590, (0 missing)
```

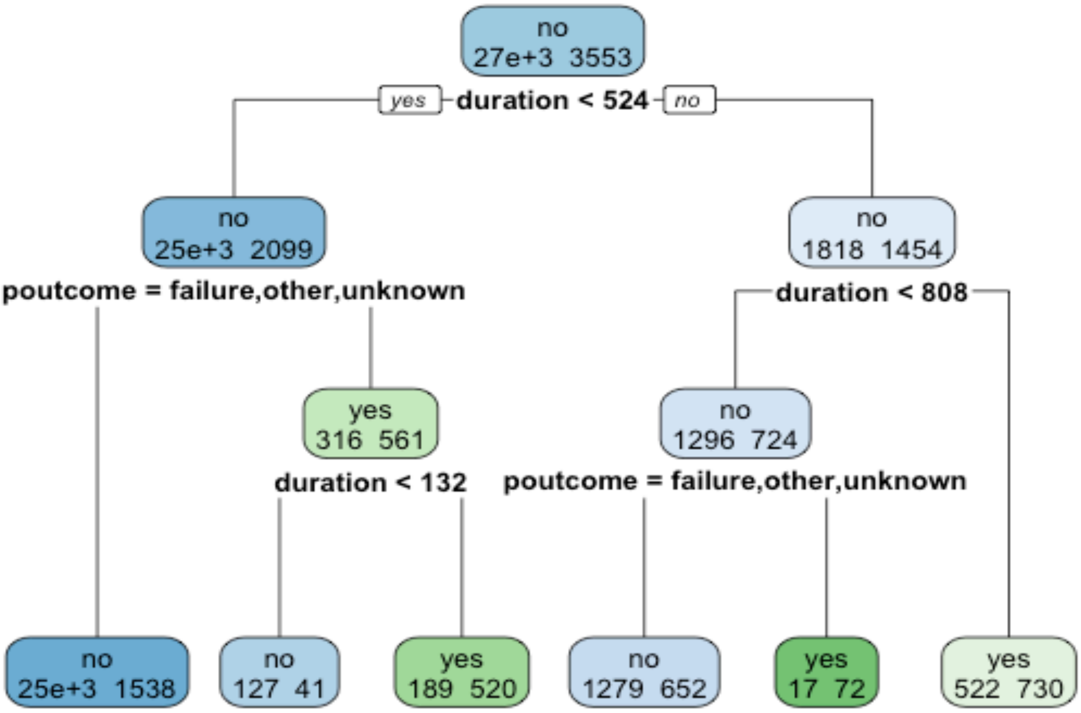
```
Node number 7: 1252 observations
  predicted class=yes  expected loss=0.4169329  P(node) =0.04133241
    class counts:   522   730
   probabilities: 0.417 0.583

Node number 10: 168 observations
  predicted class=no   expected loss=0.2440476  P(node) =0.005546202
    class counts:   127    41
   probabilities: 0.756 0.244

Node number 11: 709 observations
  predicted class=yes  expected loss=0.2665726  P(node) =0.02340629
    class counts:   189   520
   probabilities: 0.267 0.733

Node number 12: 1931 observations
  predicted class=no   expected loss=0.3376489  P(node) =0.06374831
    class counts:  1279   652
   probabilities: 0.662 0.338

Node number 13: 89 observations
  predicted class=yes  expected loss=0.1910112  P(node) =0.002938166
    class counts:    17    72
   probabilities: 0.191 0.809
```

4) THE TABLE DESCRIBING VARIABLE IMPORTANCE

```
duration        poutcome       contact       pdays
927.94251589   610.36689922   1.93614830   0.77445932
        age          campaign        default      balance
  0.65286717     0.44954368     0.38722966   0.06231402
    previous
  0.06231402
```

(We infer that the variable “duration” is the most important, followed by “poutcome” and “contact”)

5) PLOT OF DECISION TREE MODEL USING RPART.PLOT PACKAGE



6) SUMMARY OF RANDOM FOREST MODEL

```
                    Length  Class   Mode
call                   6  - none- call
type                  1   -none-  character
predicted             45211  factor numeric
err.rate              750  -none- numeric
confusion             6  -none- numeric
votes                90422  matrix numeric
oob.times            45211  -none- numeric
classes               2  -none- character
importance            64  -none- numeric
importanceSD          48  -none- numeric
localImportance     0  -none- NULL
proximity              0  -none- NULL
ntree                  1  -none- numeric
mtry                    1  -none- numeric
forest                14  -none- list
y                    45211  factor numeric
test                  0  -none- NULL
inbag                 0  -none- NULL
terms                 3  terms  call
```

7) CONFUSION MATRIX FOR RANDOM FOREST

```
> rfModel$confusion
       no      yes      class.error
no  38417   1505      0.03769851
yes  2666   2623      0.50406504
```