**DATA MINING**

**COSC 5371**

**ASSIGNMENT#3**

**DATE:** 04/15/2016

**NAME:** AMAN REDDY GURRAM

**Student ID:** 6001093799

**DETAILS OF THE DATASET: CREDIT APPROVAL**

#Loading the csv file “credit” into data frame named “exp”

>exp<-read.csv("credit.csv",header=TRUE)

#Structure of the data frame

> str(exp)

'data.frame': 690 obs. of 16 variables:

$ A1 : Factor w/ 3 levels "?","a","b": 3 2 2 3 3 3 3 2 3 3 ...

$ A2 : Factor w/ 350 levels "?","13.75","15.17",..: 158 330 91 127 45 170 181 76 312 257 ...

$ A3 : num 0 4.46 0.5 1.54 5.62 ...

$ A4 : Factor w/ 4 levels "?","l","u","y": 3 3 3 3 3 3 3 3 4 4 ...

$ A5 : Factor w/ 4 levels "?","g","gg","p": 2 2 2 2 2 2 2 2 4 4 ...

$ A6 : Factor w/ 15 levels "?","aa","c","cc",..: 14 12 12 14 14 11 13 4 10 14 ...

$ A7 : Factor w/ 10 levels "?","bb","dd",..: 9 5 5 9 9 9 5 9 5 9 ...

$ A8 : num 1.25 3.04 1.5 3.75 1.71 ...

$ A9 : Factor w/ 2 levels "f","t": 2 2 2 2 2 2 2 2 2 2 ...

$ A10: Factor w/ 2 levels "f","t": 2 2 1 2 1 1 1 1 1 1 ...

$ A11: int 1 6 0 5 0 0 0 0 0 0 ...

$ A12: Factor w/ 2 levels "f","t": 1 1 1 2 1 2 2 1 1 2 ...

$ A13: Factor w/ 3 levels "g","p","s": 1 1 1 1 3 1 1 1 1 1 ...

$ A14: Factor w/ 171 levels "?","00000","00017",..: 70 13 98 33 39 117 56 25 64 17 ...

$ A15: int 0 560 824 3 0 0 31285 1349 314 1442 ...

$ A16: Factor w/ 2 levels "-","+": 2 2 2 2 2 2 2 2 2 2 ...

**LOADING AND INSTALLING THE PACKAGES**

> install.packages('e1071', dependencies=TRUE)

> install.packages("caret")

> install.packages("C50")

> require(C50)

**PREDICTION FOR THE 50% DATA**

># 50% of 690 observations : 345

># Shuffling the rows of the data set exp for efficiency

>set.seed(999)

> random<-runif(nrow(exp))

> shuffle<-exp[order(random),]

># Running the C5.0 algorithm for training the first 345 rows of the data set

># Loading all the training data into a vector called trained

>#class attribute number is 16

> trained<-C5.0(shuffle[1:345,-16],shuffle[1:345,16])

>#BELOW IS THE SUMMARY OF WHAT THE MODEL TRAINED HAS LEARNED AND ANALYSIS OF THE PREDICTIONS

> summary(trained)

Call:

C5.0.default(x = shuffle[1:345, -16], y = shuffle[1:345, 16])

C5.0 [Release 2.07 GPL Edition] Fri Apr 15 15:39:01 2016

-------------------------------

Class specified by attribute `outcome'

Read 345 cases (16 attributes) from undefined.data

Decision tree:

A9 = f: - (166/8)

A9 = t:

:...A10 = t: + (113/10)

A10 = f:

:...A15 > 375: + (10)

A15 <= 375:

:...A14 <= 110: + (21/5)

A14 > 110: - (35/11)

Evaluation on training data (345 cases):

Decision Tree

----------------

Size Errors

5 34( 9.9%) <<

(a) (b) <-classified as

---- ----

182 15 (a): class -

19 129 (b): class +

Attribute usage:

100.00% A9

51.88% A10

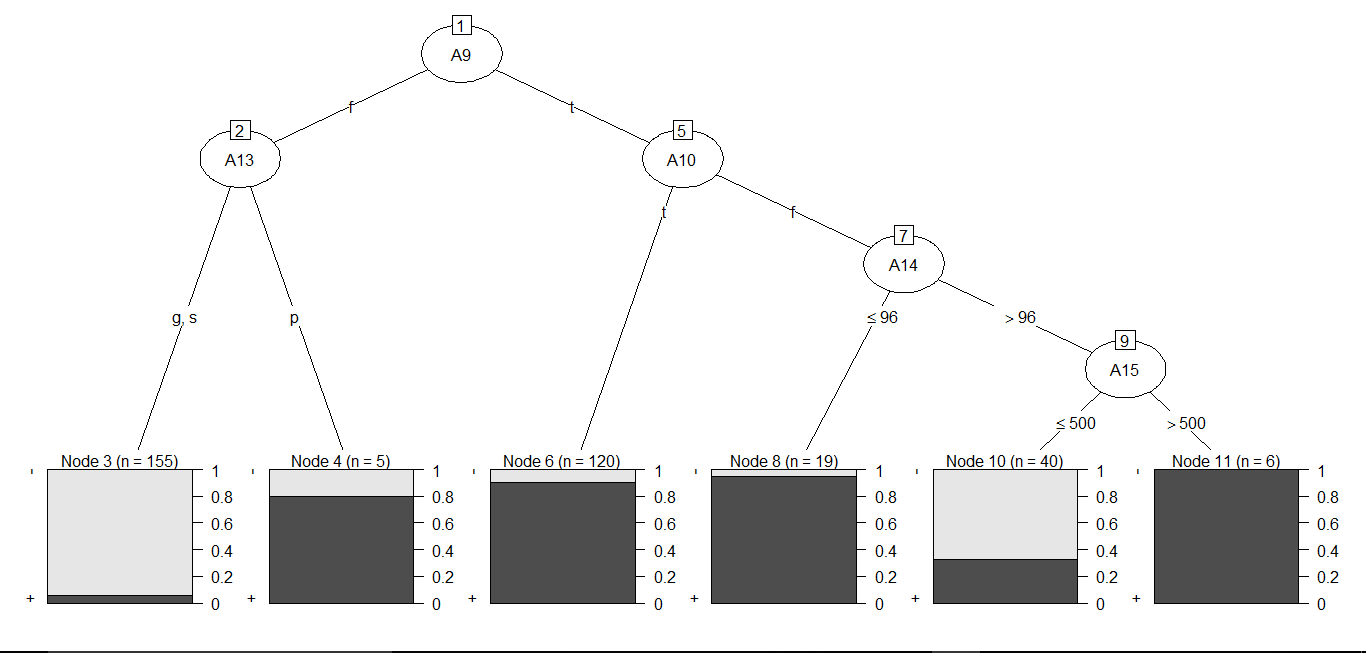
19.13% A15

16.23% A14

Time: 0.0 secs

**>#DECISION TREE**

> plot(trained)



**TESTING THE REMAINING 50% OF THE DATASET**

>#Remaining 50% of 690 observations are 346 to 690

> test<-predict(trained,shuffle[346:690,])

**CONFUSION MATRIX**

> confusionMatrix(shuffle[346:690,16],test)

Confusion Matrix and Statistics

Reference

Prediction - +

- 177 19

+ 32 117

Accuracy : 0.8522

95% CI : (0.8103, 0.8879)

No Information Rate : 0.6058

P-Value [Acc > NIR] : < 2e-16

Kappa : 0.6956

Mcnemar's Test P-Value : 0.09289

Sensitivity : 0.8469

Specificity : 0.8603

Pos Pred Value : 0.9031

Neg Pred Value : 0.7852

Prevalence : 0.6058

Detection Rate : 0.5130

Detection Prevalence : 0.5681

Balanced Accuracy : 0.8536

'Positive' Class : -