* 1. Which attribute(s) did you pick as the class attribute?  Explain your decision.

ALC is our class attribute, as because from our last assignment, we came to know that it is related to absence, gout, family relation, sex, and other attributes, where in top 5 attributes it become most predominant value to answer the big question, and where it as dependent value , and so DALC, And WALC are the attribute which are main attribute which became more interdependent, so making these two as a single group, we answered the big question, so that it solves the big question.  
  
DALC AND WALC are been grouped, and then made to ALC, to answer the big question.

Alc=(Dalc\*+/Walc\*2)/7

* 1. Which two of these Classification/ Prediction algorithms produced the best results?  Why?

R Code:

Alc=(Dalc\*+/Walc\*2)/7

student<-student\_por1

ind <- sample(2, nrow(student), replace=TRUE, prob=c(0.7, 0.3))

trainDataTree <- student[ind==1,]

testDataTree <- student[ind==2,]

myFormula <- Alc~ school+sex+age+address+famsize+Pstatus+Medu+Fedu+Mjob+Fjob+reason+guardian+traveltime+studytime+failures+schoolsup+famsup+paid+activities+nursery+higher+internet+romantic+famrel+freetime+goout+health+absences

student\_tree <- tree(myFormula, data=trainDataTree)

summary(student\_tree)

print(student\_tree)

plot(student\_tree)

text(student\_tree)

testPred <- predict(student\_tree, newdata = testDataTree)

table(testPred, testDataTree$Alc)

show(testPred)

library(MASS)

studenttree <- tree(myFormula,data=student)

plot(studenttree)

plot(studenttree,col=8)

text(studenttree,digits=2)

summary(studenttree)

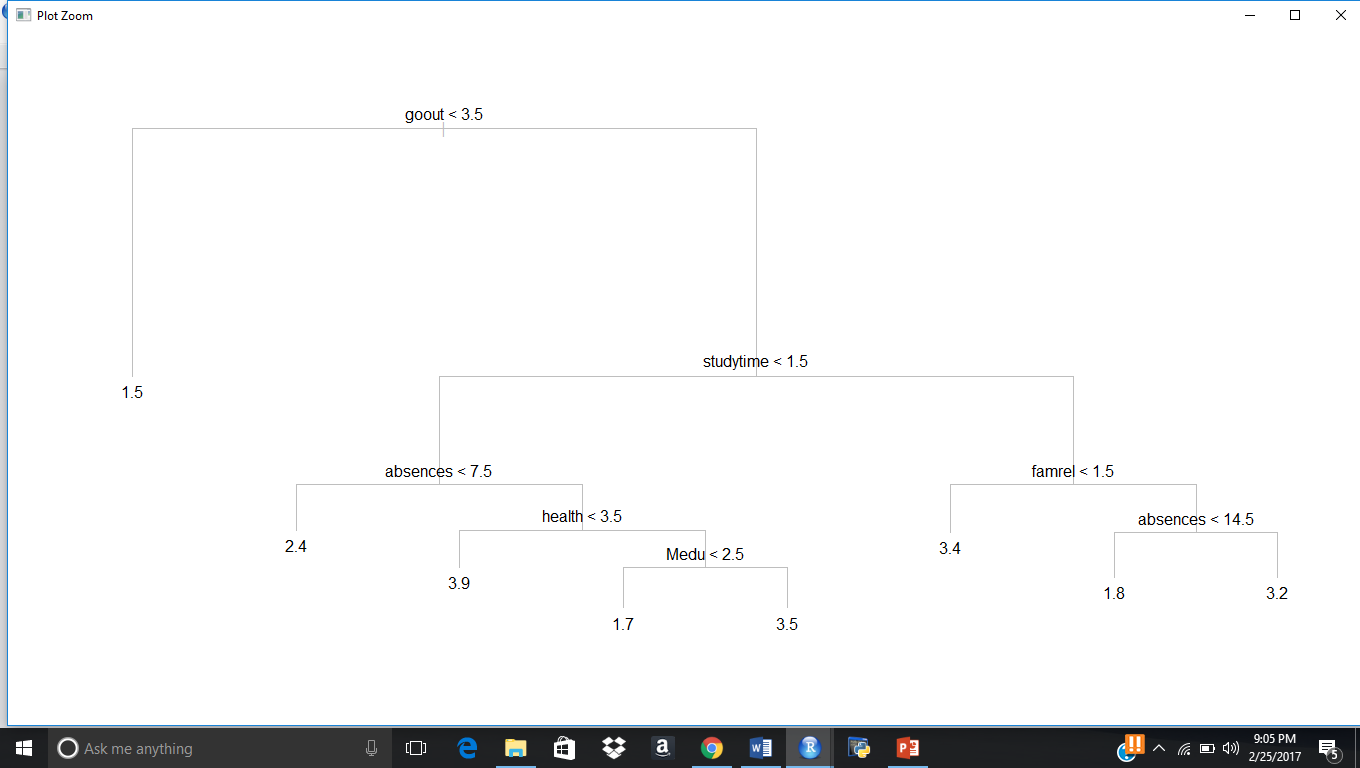
studentsnip=snip.tree(studenttree,nodes=c(7,12))

irissnip

plot(studentsnip)

text(studentsnip)

summary(studentsnip)



> table(testPred, testDataTree$Alc)

testPred 1 2 3 4 5

1.33333333333333 2 0 0 0 0

1.41463414634146 9 2 3 0 0

1.44827586206897 63 31 4 1 0

1.55 3 1 1 0 0

1.71428571428571 1 0 3 0 0

1.91304347826087 7 4 1 1 0

2.1 2 0 0 0 0

2.28571428571429 1 5 1 0 0

2.31578947368421 3 3 4 1 1

2.66666666666667 1 1 0 1 1

3 4 2 0 2 0

3.33333333333333 4 3 1 1 1

3.42857142857143 1 2 2 0 0

4.16666666666667 1 0 1 0 1

################# C TREE (conditional inference) ###################

set.seed(1234)

ind <- sample(2, nrow(student), replace=TRUE, prob=c(0.7, 0.3))

trainData <- student[ind==1,]

testData <- student[ind==2,]

library(party)

student\_ctree <- ctree(myFormula, data=trainData)

table(predict(iris\_ctree), trainData$Species)

print(student\_ctree)

plot(student\_ctree)

testPred <- predict(student\_ctree, newdata = testData)

table(testPred, testData$Alc)

> table(predict(student\_ctree), trainData$Alc)

1 2 3 4 5

1 227 90 20 6 1

2 10 19 7 1 1

3 14 23 27 11 10

4 0 0 0 0 0

5 0 0 0 0 0

> table(testPred, testData$Alc)

testPred 1 2 3 4 5

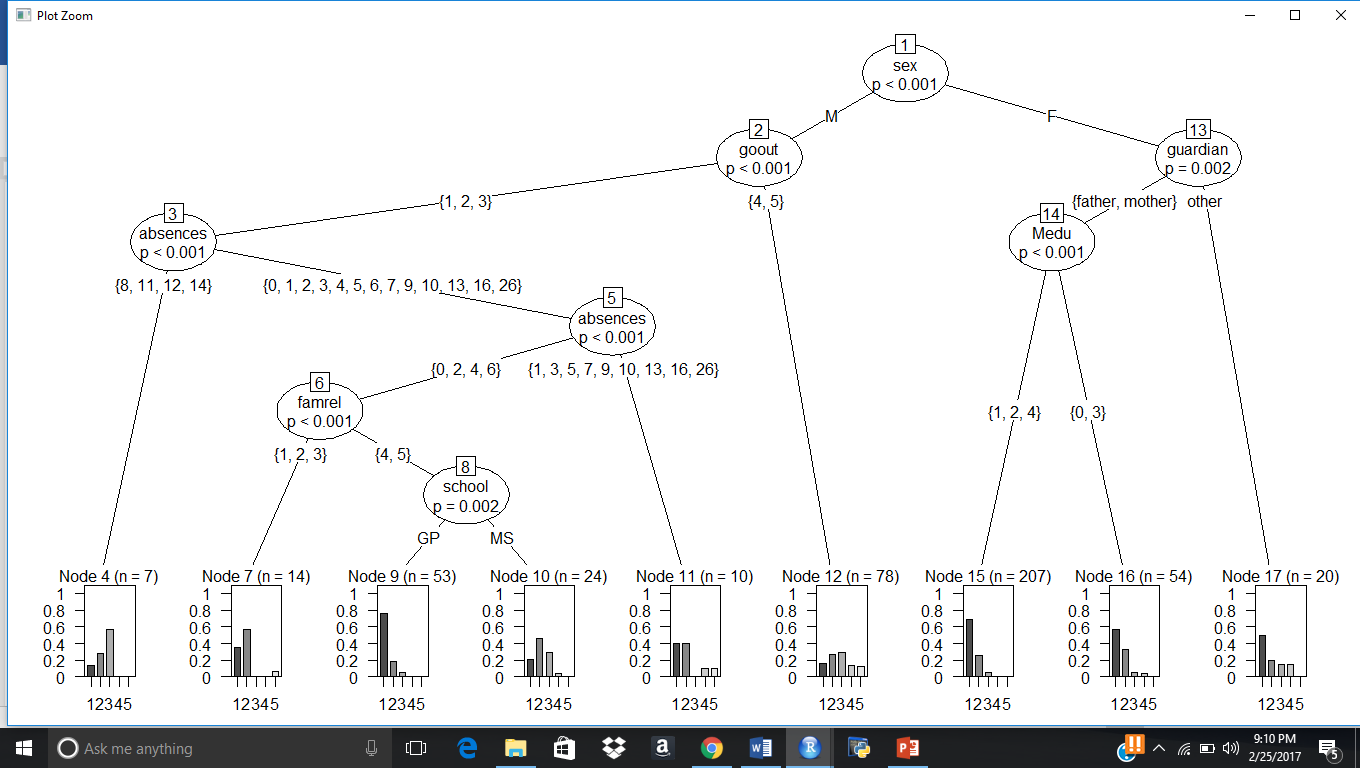
1 85 31 5 5 2

2 10 8 2 0 0

3 8 7 12 6 1

4 0 0 0 0 0

5 0 0 0 0 0



#################### R Part Package ###########################

library(rpart.plot)

attributes(student)

set.seed(1234)

ind <- sample(2, nrow(student), replace=TRUE, prob=c(0.7, 0.3))

student.train <- student[ind==1,]

student.test <- student[ind==2,]

library(rpart)

student\_rpart <- rpart(myFormula, data = student.train, control = rpart.control(minsplit = 10), method="anova")

attributes(student\_rpart)

print(student\_rpart$cptable)

print(student\_rpart)

rpart.plot(student\_rpart,type = 1,digits = 2)

plot(student\_rpart)

text(student\_rpart, use.n=T)

opt <- which.min(student\_rpart$Alc[,"xerror"])

cp <- student\_rpart$Alc[opt, "CP"]

student\_prune <- prune(student\_rpart, cp = cp, method="anova")

print(student\_prune)

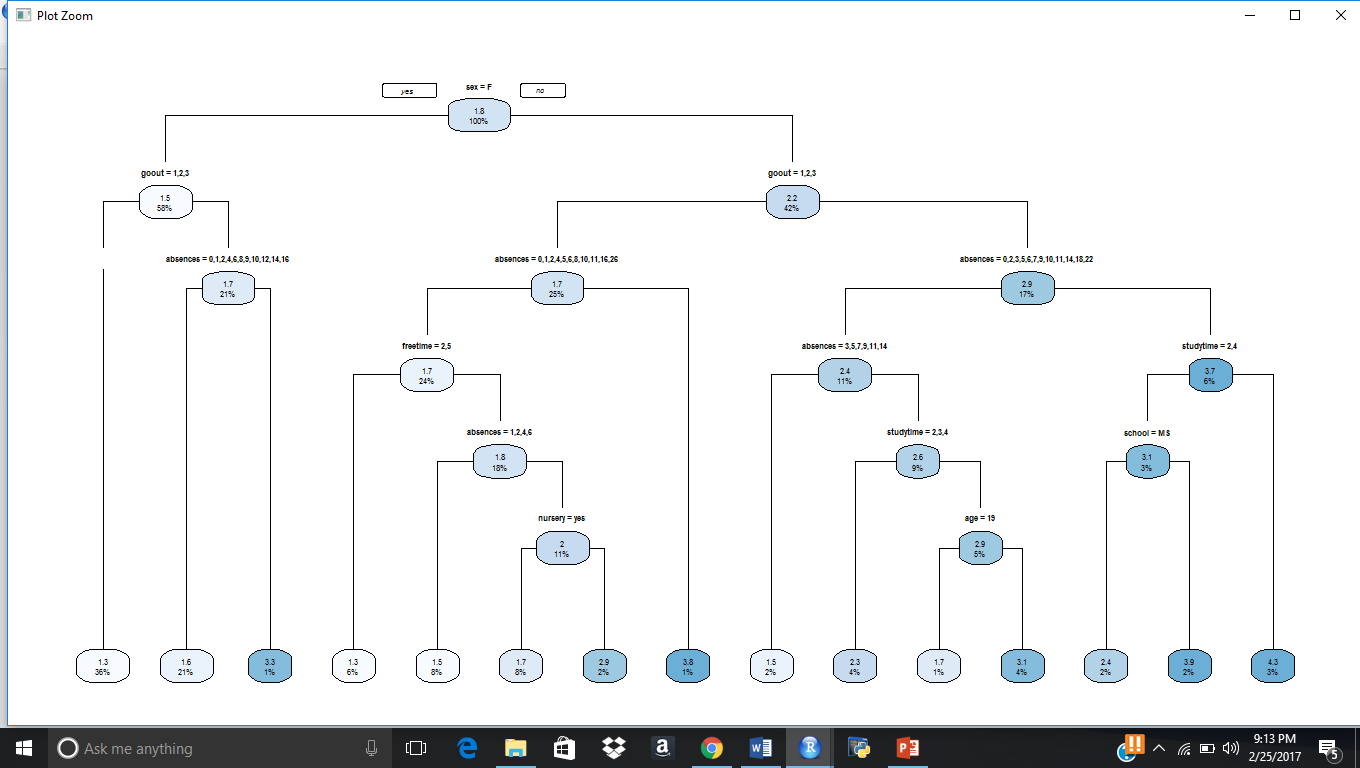
plot(student\_prune)

text(student\_prune, use.n=T)

student\_pred <- predict(student\_prune, newdata=student.test)

rpart.plot(student\_prune)

summary(student\_prune)



> summary(student\_prune)

Call:

rpart(formula = myFormula, data = student.train, method = "anova",

control = rpart.control(minsplit = 10))

n= 451

CP nsplit rel error xerror xstd

1 0.12324065 0 1.0000000 1.0026682 0.08636674

2 0.11871733 1 0.8767593 0.9686184 0.08157060

3 0.05542307 2 0.7580420 0.8468726 0.07535715

4 0.04548228 3 0.7026189 0.8607358 0.07620513

5 0.01891978 4 0.6571367 0.8170927 0.07333165

6 0.01738622 5 0.6382169 0.8692512 0.07898539

7 0.01634287 6 0.6208307 0.8637033 0.07931973

8 0.01579793 8 0.5881449 0.8851357 0.08197791

9 0.01475743 11 0.5407511 0.8904026 0.08209851

10 0.01007596 12 0.5259937 0.9696262 0.08772473

11 0.01000000 14 0.5058418 0.9741168 0.08708042

Variable importance

absences goout sex age freetime studytime Medu

26 17 16 6 6 6 4

nursery school failures health Fjob traveltime Mjob

3 2 2 2 2 2 1

Fedu famrel reason guardian paid

1 1 1 1 1

###################### NAIVE BAYES (conditional probability) ###############################

library(mlbench)

plot(student[,34])

title(main="number of alcohol consumption", xlab="level of consumption", ylab="number of student")

student[,"train"] <- ifelse(runif(nrow(student))<0.80,1,0)

trainColNum <- grep('train', names(student))

trainstudent <- student[student$train==1,-trainColNum]

teststudent <- student[student$train==0,-trainColNum]

library(e1071)

nb\_model <- naiveBayes(Alc~.,data = trainstudent)

nb\_model

summary(nb\_model)

str(nb\_model)

nb\_test\_predict <- predict(nb\_model,teststudent[,-1])

table(pred=nb\_test\_predict,true=teststudent$Alc)

mean(nb\_test\_predict==teststudent$Alc)

nb\_multiple\_runs <- function(train\_fraction,n){

fraction\_correct <- rep(NA,n)

for (i in 1:n){

student[,"train"] <- ifelse(runif(nrow(student))<train\_fraction,1,0)

trainColNum <- grep('train',names(student))

trainstudent <- student[student$train==1,-trainColNum]

teststudent <- student[student$train==0,-trainColNum]

nb\_model <- naiveBayes(Alc~.,data = trainstudent)

nb\_test\_predict <- predict(nb\_model,teststudent[,-1])

fraction\_correct[i] <- mean(nb\_test\_predict==teststudent$Alc)

}

return(fraction\_correct)

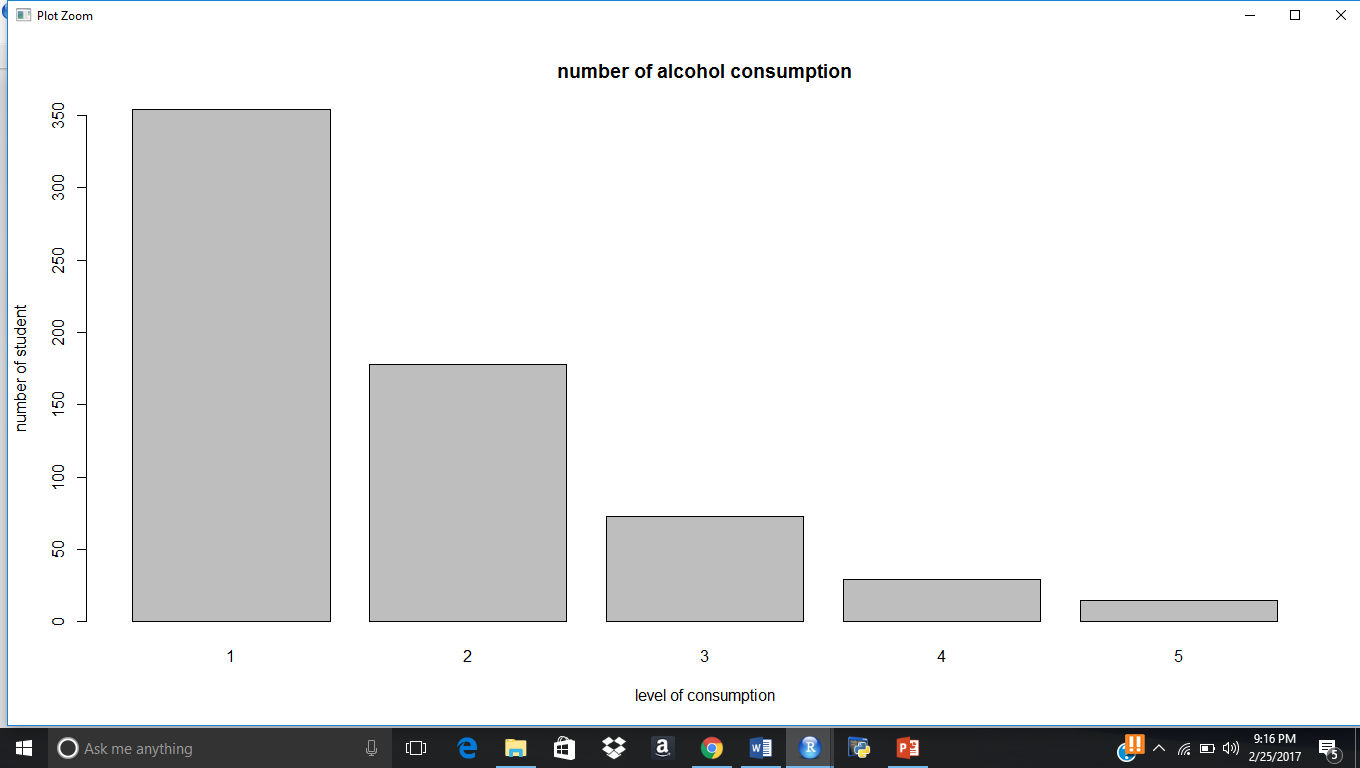
}

fraction\_correct\_predictions <- nb\_multiple\_runs(0.8,20)

fraction\_correct\_predictions

summary(fraction\_correct\_predictions)

sd(fraction\_correct\_predictions)



> summary(nb\_model)

Length Class Mode

apriori 5 table numeric

tables 33 -none- list

levels 5 -none- character

call 4 -none- call

> nb\_test\_predict <- predict(nb\_model,teststudent[,-1])

> table(pred=nb\_test\_predict,true=teststudent$Alc)

true

pred 1 2 3 4 5

1 105 0 0 1 0

2 4 44 5 2 1

3 0 2 20 3 1

4 0 0 0 2 4

5 0 0 0 0 1

> mean(nb\_test\_predict==teststudent$Alc)

[1] 0.8820513

> fraction\_correct\_predictions

[1] 0.9387755 0.8709677 0.8661972 0.8270677 0.9000000 0.8666667 0.8870968 0.9268293

[9] 0.8769231 0.9126984 0.8947368 0.8333333 0.8802817 0.8560000 0.8939394 0.8936170

[17] 0.8661417 0.8769231 0.8768116 0.8906250

> summary(fraction\_correct\_predictions)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.8271 0.8665 0.8786 0.8818 0.8941 0.9388

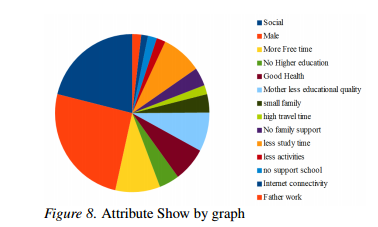
> sd(fraction\_correct\_predictions)

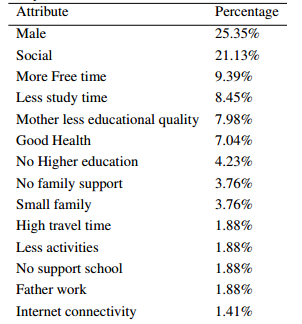
[1] 0.02718371

Bayes algorithm is given best results than the other algorithm which was been used , because of this we are able to identify the percentage of the attribute best inter related, and make it best pre processed forwarded to make a best relation so that it suffice the big question using the class attribute how ever it also give how much trust and relation does ALC is given in compared to other attributes to make it with other attributes, if we can identify the percentage way it gives more information and descriptive format for the client and the owner to make more informative decision to be taken on this.

* 1. What do the results tell you about the data in your dataset?  Are they meaningful and useful (such that a client would pay for them)

Here are some of the inferrences which can be taken from the desired results.   
  
  
Below identified with graph and table, we identify that how best the ttribute are internally related to the attribute, so make it with that we identify that Men are considered to be most drinking people. People who goes out with friends would have more drinking. More free time and less education time is making people to drink more. Students who wont go to university drink more, Family size and mother education is also making student addicted to alcohol. So, by giving these many kind of results, yes, client would definetly would pay us for this results by just pre processing the data, and making a clear cut information to see what best relates, to each other, and what best can be made to the student.





* 1. What business, political, or medical decision could management make on the basis of your results?

So based on this kind of results, yes we can identify some decision, which would make government to take good decision on student as considered as student to be considered as next growth of the country. Some of the decisions said to be as without having time given to student for education, it helps student to drink more, so they can start time in the schooling as study time to make them avigated with some other work for education which make student to decrease intke of alcohol. Family time is something very important, so to build more parks for students so that they can spend more time with family, and events for students so that they can time with family. Health is one factor, to make a regular check up in school to check student health checkup done on monthly basis. Providing more jobs for women would also make student to get inspired from child hood and make them start working.  
  
  
These are some of the important things which we considered as important points which are been found in our results, and also as on we make more reverse interpreted relation to it, it would generate more results for government to take more results.