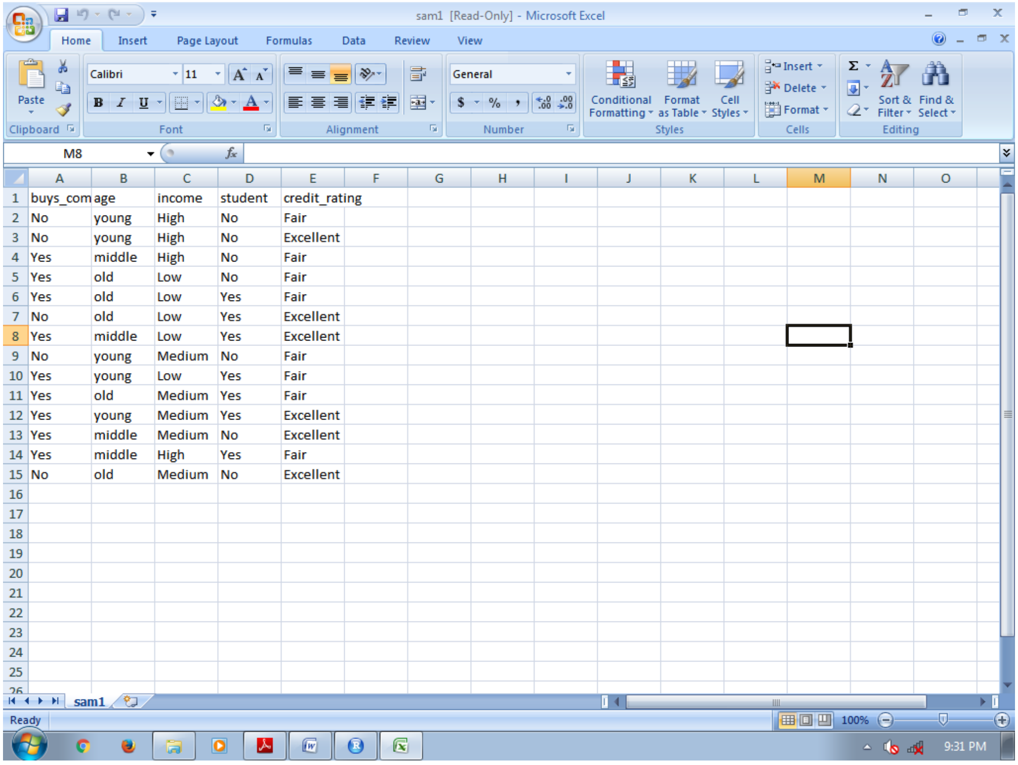
Decision Tree using R

> setwd("C:/Users/home/Desktop")

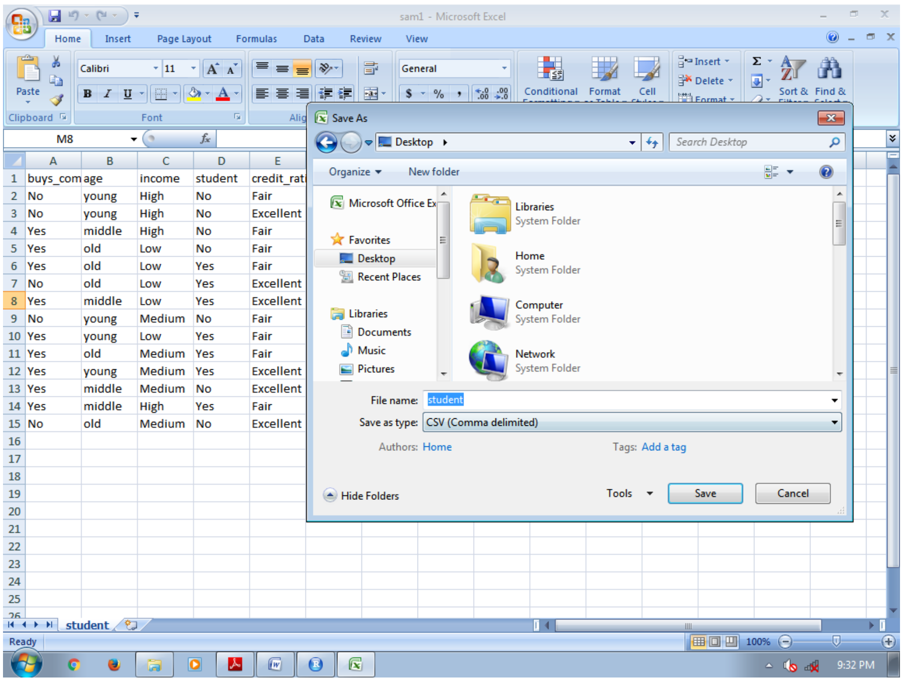
> library("rpart")

> library("rpart.plot")

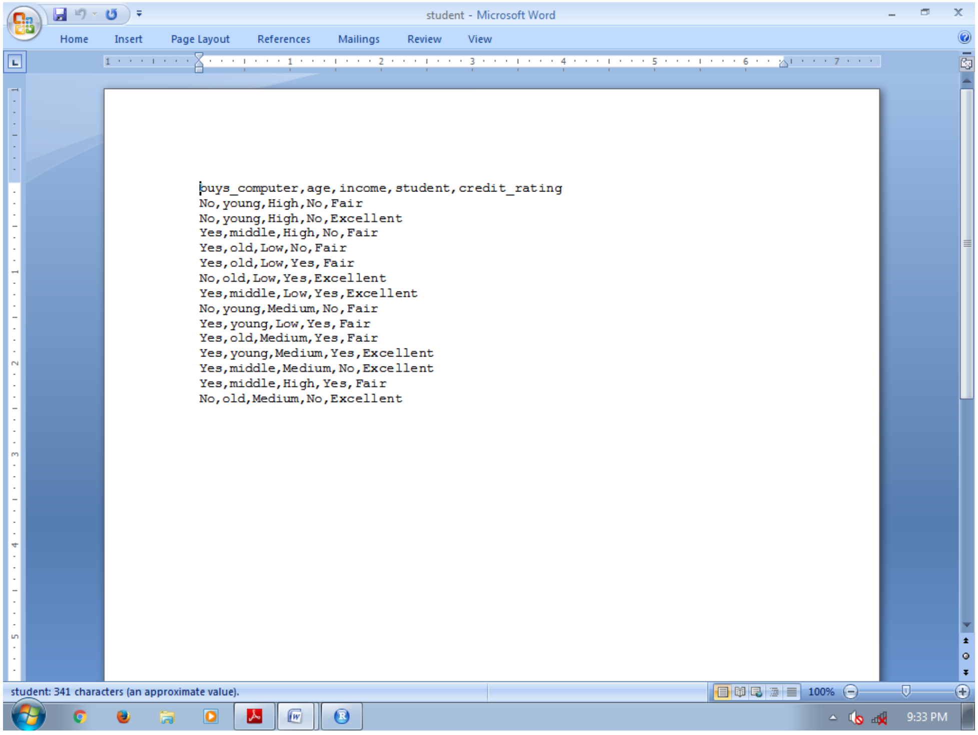
Step 1: **Making as an Excel Format**



Step 2: **Saving the File as CSV** - File->save as ->student.csv(comma delimited) as shown below.



To open the CSV file(student.csv) in Ms-Word



>play_decision <- read.table ("C:/Users/home/Desktop/student.csv", header=TRUE, sep=",")

> play_decision

	buys_computer	**age**	**income**	**student**	**credit_rating**
1	No	young	High	No	Fair
2	No	young	High	No	Excellent
3	Yes	middle	High	No	Fair
4	Yes	old	Low	No	Fair
5	Yes	old	Low	Yes	Fair
6	No	old	Low	Yes	Excellent
7	Yes	middle	Low	Yes	Excellent
8	No	young	Medium	No	Fair
9	Yes	young	Low	Yes	Fair
10	Yes	old	Medium	Yes	Fair
11	Yes	young	Medium	Yes	Excellent
12	Yes	middle	Medium	No	Excellent
13	Yes	middle	High	Yes	Fair
14	No	old	Medium	No	Excellent

> summary(play_decision)

```
 buys_computer     age      income  student   credit_rating
 No :5        middle:4   High  :4   No :7   Excellent:6
 Yes:9        old   :5   Low   :5   Yes:7   Fair     :8
              young :5   Medium:5
```

> fit <- rpart(buys_computer ~ age + income + student + credit_rating,

+ method = "class",

+ data = play_decision,

+ control = rpart.control(minsplit = 1),

+ parms = list(split = 'information'))

> summary(fit)

```
Call:
rpart(formula = buys_computer ~ age + income + student + credit_rating,
    data = play_decision, method = "class", parms = list(split = "information"),
    control = rpart.control(minspli    data = play_decision, method = "class", parms = list(split =
"information"), t = 1))
  n= 14

    CP nsplit rel error xerror      xstd
1 0.30      0       1.0    1.0 0.3585686
2 0.20      2       0.4    1.8 0.3585686
3 0.10      3       0.2    1.0 0.3585686
4 0.01      5       0.0    1.2 0.3703280

Variable importance
          age        income       student credit_rating
           39            32            19            11

Node number 1: 14 observations,    complexity param=0.3
  predicted class=Yes  expected loss=0.3571429  P(node) =1
    class counts:     5     9
   probabilities: 0.357 0.643
  left son=2 (10 obs) right son=3 (4 obs)
  Primary splits:
      age           splits as  RLL, improve=2.1931200, (0 missing)
      student       splits as  LR,  improve=1.4734210, (0 missing)
      credit_rating splits as  LR,  improve=0.4670276, (0 missing)
      income        splits as  LRL, improve=0.4399255, (0 missing)

Node number 2: 10 observations,    complexity param=0.3
  predicted class=No   expected loss=0.5  P(node) =0.7142857
    class counts:     5     5
   probabilities: 0.500 0.500
  left son=4 (5 obs) right son=5 (5 obs)
  Primary splits:
```

```
      student      splits as  LR,  improve=1.9274480, (0 missing)
      income       splits as  LRR, improve=1.6389660, (0 missing)
      credit_rating splits as  LR,  improve=0.8630462, (0 missing)
      age          splits as  -RL, improve=0.2013551, (0 missing)
  Surrogate splits:
      income splits as  LRR, agree=0.7, adj=0.4, (0 split)
      age    splits as  -RL, agree=0.6, adj=0.2, (0 split)

Node number 3: 4 observations
  predicted class=Yes  expected loss=0  P(node) =0.2857143
    class counts:     0     4
   probabilities: 0.000 1.000

Node number 4: 5 observations,    complexity param=0.2
  predicted class=No   expected loss=0.2  P(node) =0.3571429
    class counts:     4     1
   probabilities: 0.800 0.200
  left son=8 (4 obs) right son=9 (1 obs)
  Primary splits:
      income       splits as  LRL, improve=2.5020120, (0 missing)
      age          splits as  -RL, improve=1.1157180, (0 missing)
      credit_rating splits as  LR,  improve=0.5924696, (0 missing)

Node number 5: 5 observations,    complexity param=0.1
  predicted class=Yes  expected loss=0.2  P(node) =0.3571429
    class counts:     1     4
   probabilities: 0.200 0.800
  left son=10 (2 obs) right son=11 (3 obs)
  Primary splits:
      credit_rating splits as  LR,  improve=1.1157180, (0 missing)
      age          splits as  -LR, improve=0.5924696, (0 missing)
      income       splits as  -LR, improve=0.5924696, (0 missing)

Node number 8: 4 observations
  predicted class=No   expected loss=0  P(node) =0.2857143
    class counts:     4     0
   probabilities: 1.000 0.000

Node number 9: 1 observations
  predicted class=Yes  expected loss=0  P(node) =0.07142857
    class counts:     0     1
   probabilities: 0.000 1.000

Node number 10: 2 observations,    complexity param=0.1
  predicted class=No   expected loss=0.5  P(node) =0.1428571
    class counts:     1     1
   probabilities: 0.500 0.500
  left son=20 (1 obs) right son=21 (1 obs)
  Primary splits:
      age    splits as  -LR, improve=1.386294, (0 missing)
      income splits as  -LR, improve=1.386294, (0 missing)

Node number 11: 3 observations
```

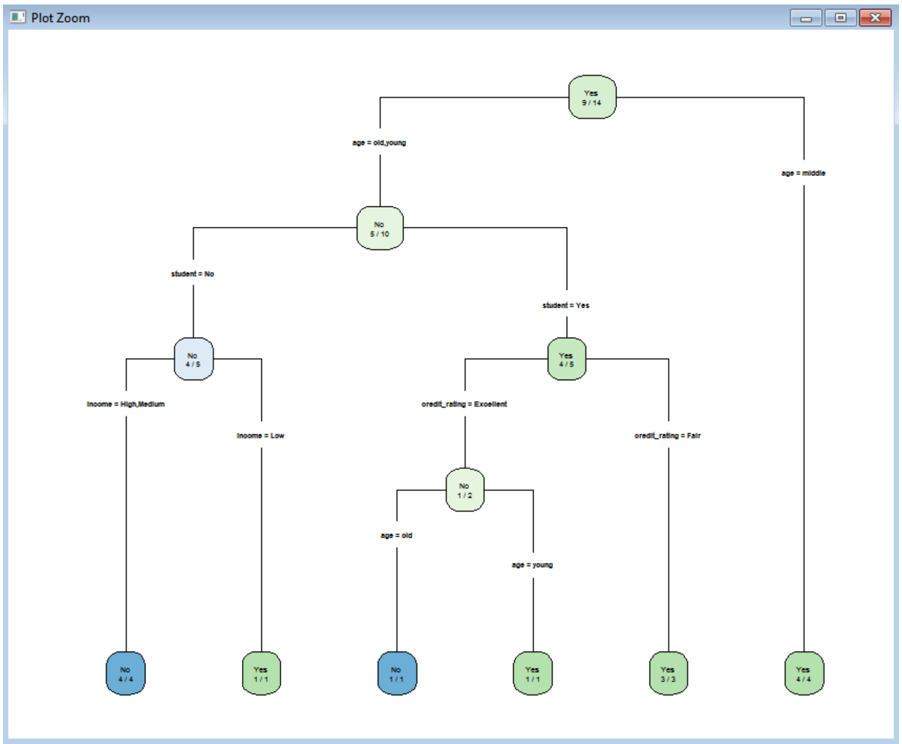
predicted class=Yes expected loss=0 P(node) =0.2142857
  class counts: 0   3
 probabilities: 0.000 1.000

Node number 20: 1 observations
 predicted class=No   expected loss=0 P(node) =0.07142857
  class counts: 1   0
 probabilities: 1.000 0.000

Node number 21: 1 observations
 predicted class=Yes expected loss=0 P(node) =0.07142857
  class counts: 0   1
 probabilities: 0.000 1.000

> rpart.plot(fit,type=4,extra=2,clip.right.labs=FALSE,

+ varlen =0, faclen=0)



<u>Sample data</u>

> newdata <- data.frame(age="middle",income="High",student="No",credit_rating="Fair")
> newdata

	Age	income	student	credit_rating
1	middle	High	No	Fair

> predict (fit,newdata=newdata, type= "prob")

  No Yes
1 0   1

> predict (fit,newdata=newdata, type= "class")

  1
Yes
Levels: No Yes