Last Name: **Pandey**
First Name: **Aditya**
Student ID: **1001405034**
Course: **Data Mining CSE 5334 Fall 2016**
Topic: **Home Work Assignment 6**

Step 1: Description of the Dataset

Name
Survival of passengers on the Titanic

Description
This data set provides information on the fate of passengers on the fatal maiden voyage of the ocean liner 'Titanic', summarized according to economic status (class), sex, age and survival.

Usage
Titanic

Format
A 4-dimensional array resulting from cross-tabulating 2201 observations on 4 variables. The variables and their levels are as follows:

No	Name	Levels
1	Class	1st, 2nd, 3rd, Crew
2	Sex	Male, Female
3	Age	Child, Adult
4	Survived	No, Yes

Details
The sinking of the Titanic is a famous event, and new books are still being published about it. Many well-known facts—from the proportions of first-class passengers to the 'women and children first' policy, and the fact that that policy was not entirely successful in saving the women and children in the third class—are reflected in the survival rates for various classes of passenger.
These data were originally collected by the British Board of Trade in their investigation of the sinking. Note that there is not complete agreement among primary sources as to the exact numbers on board, rescued, or lost.
Due in particular to the very successful film 'Titanic', the last years saw a rise in public interest in the Titanic. Very detailed data about the passengers is now available on the Internet, at sites such as *Encyclopedia Titanica* (http://www.rmplc.co.uk/eduweb/sites/phind).

Source

Dawson, Robert J. MacG. (1995), The 'Unusual Episode' Data Revisited. *Journal of Statistics Education*, **3**.https://www.amstat.org/publications/jse/v3n3/datasets.dawson.html

The source provides a data set recording class, sex, age, and survival status for each person on board of the Titanic, and is based on data originally collected by the British Board of Trade and reprinted in:

British Board of Trade (1990), *Report on the Loss of the 'Titanic' (S.S.)*. British Board of Trade Inquiry Report (reprint). Gloucester, UK: Allan Sutton Publishing.

Step 2: Initializing the datasets

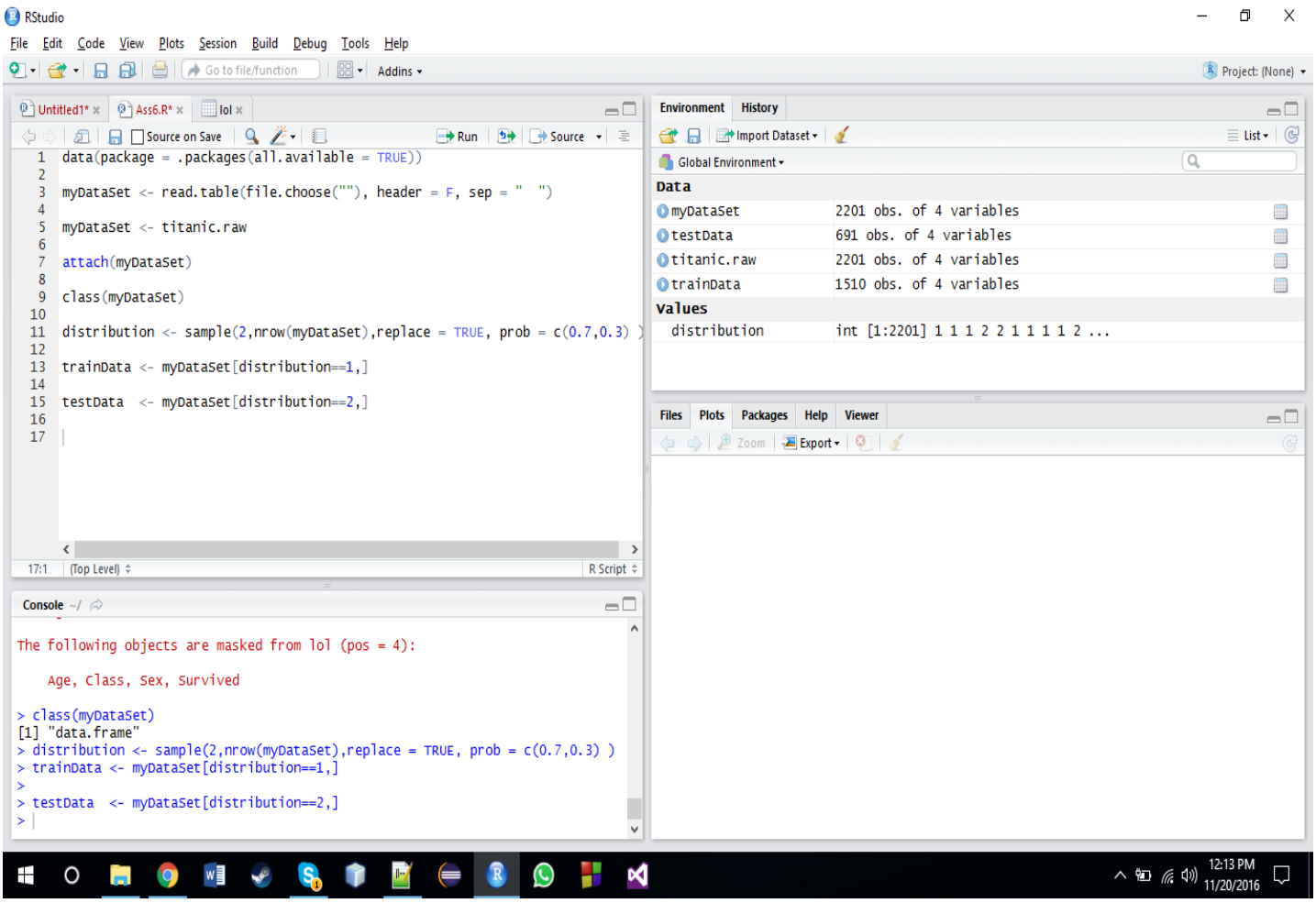

Figure 1: Initializing our datasets

In this we have created, 4 environment variables.

Titanic.raw -> It is the Titanic dataset.

MyDataSet -> It is the copy of titanic dataset which we are going to use in our function.

TrainData -> This is the training dataset which is randomly 70% of the original dataset.

TestData -> This is the testing dataset which is randomly 30% of the original dataset.

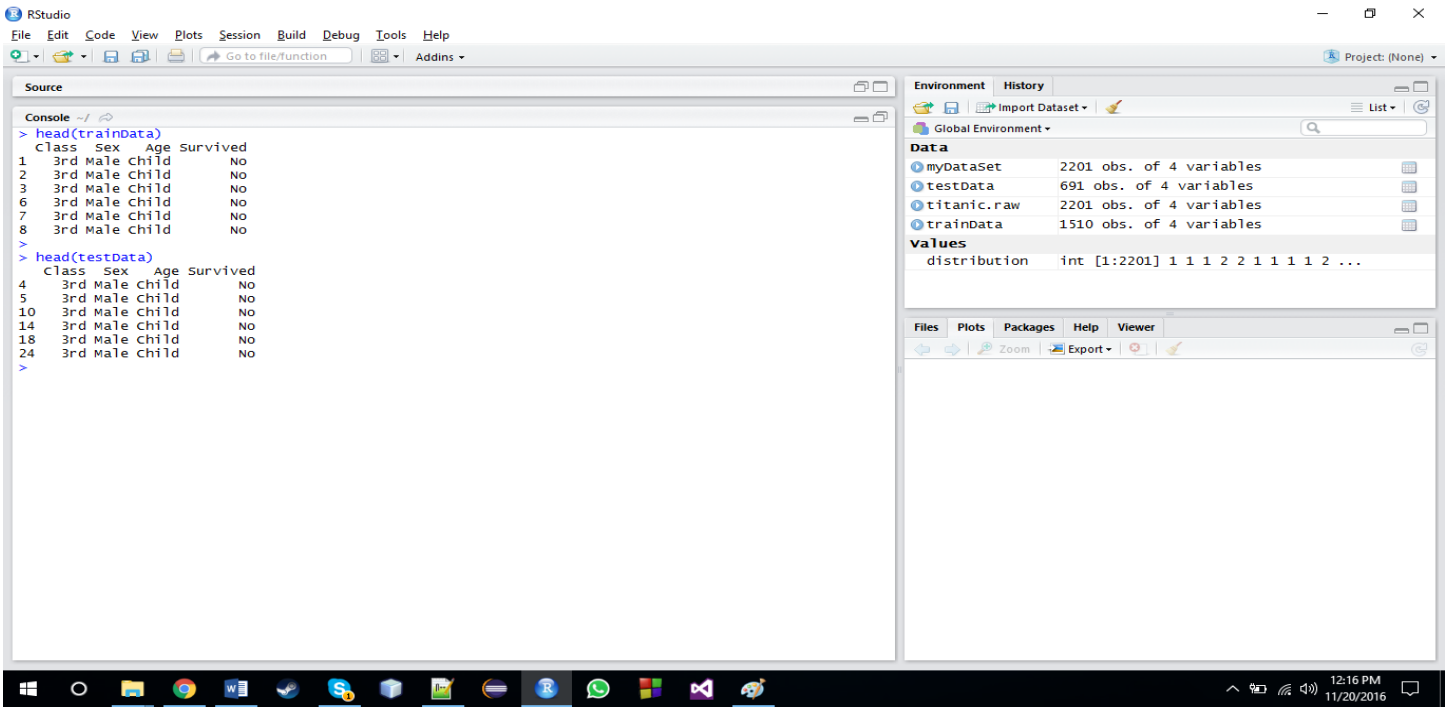

Figure 2: Output of the head command

Using Head command, we display the 1st 5 tuples of both our dataset, training, and testing.

Step 3: Plotting the Training Data

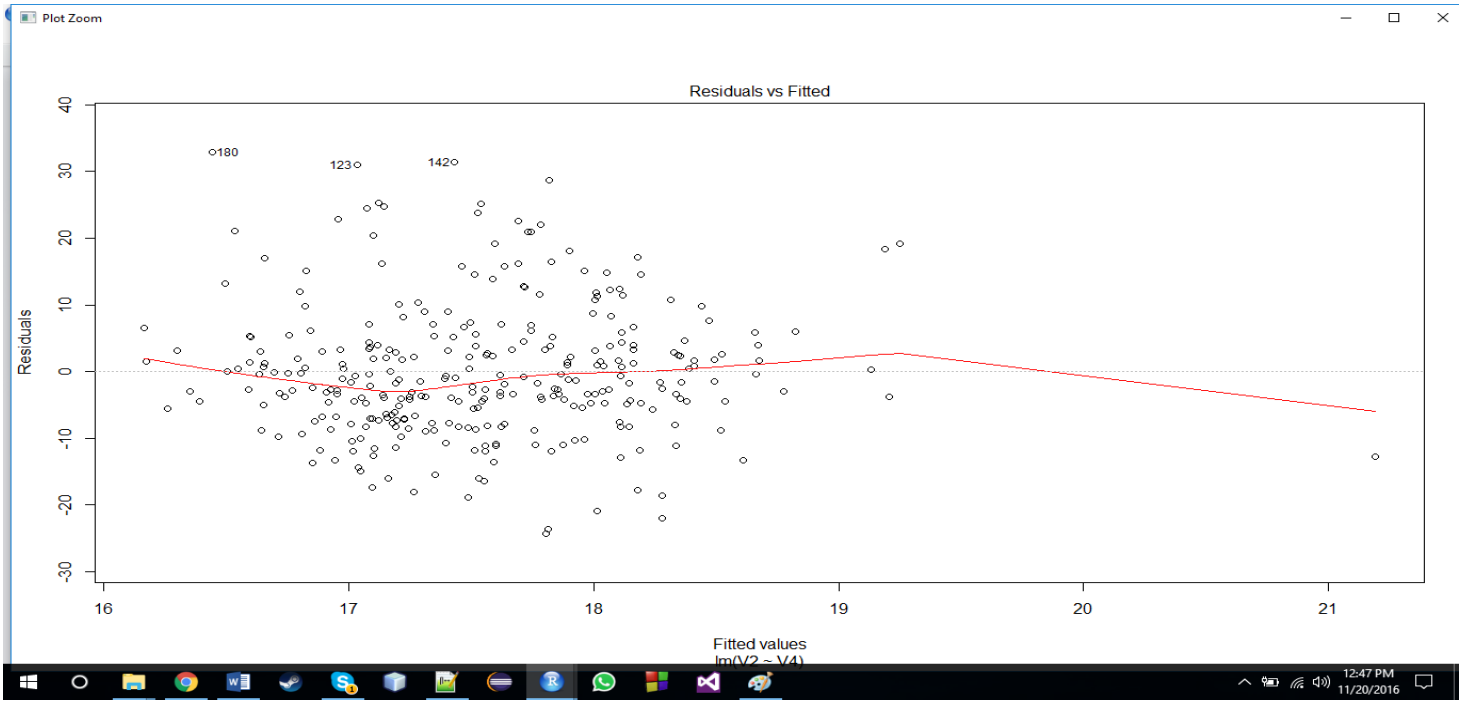

Figure 3: This is the residual v/s fitted values curve

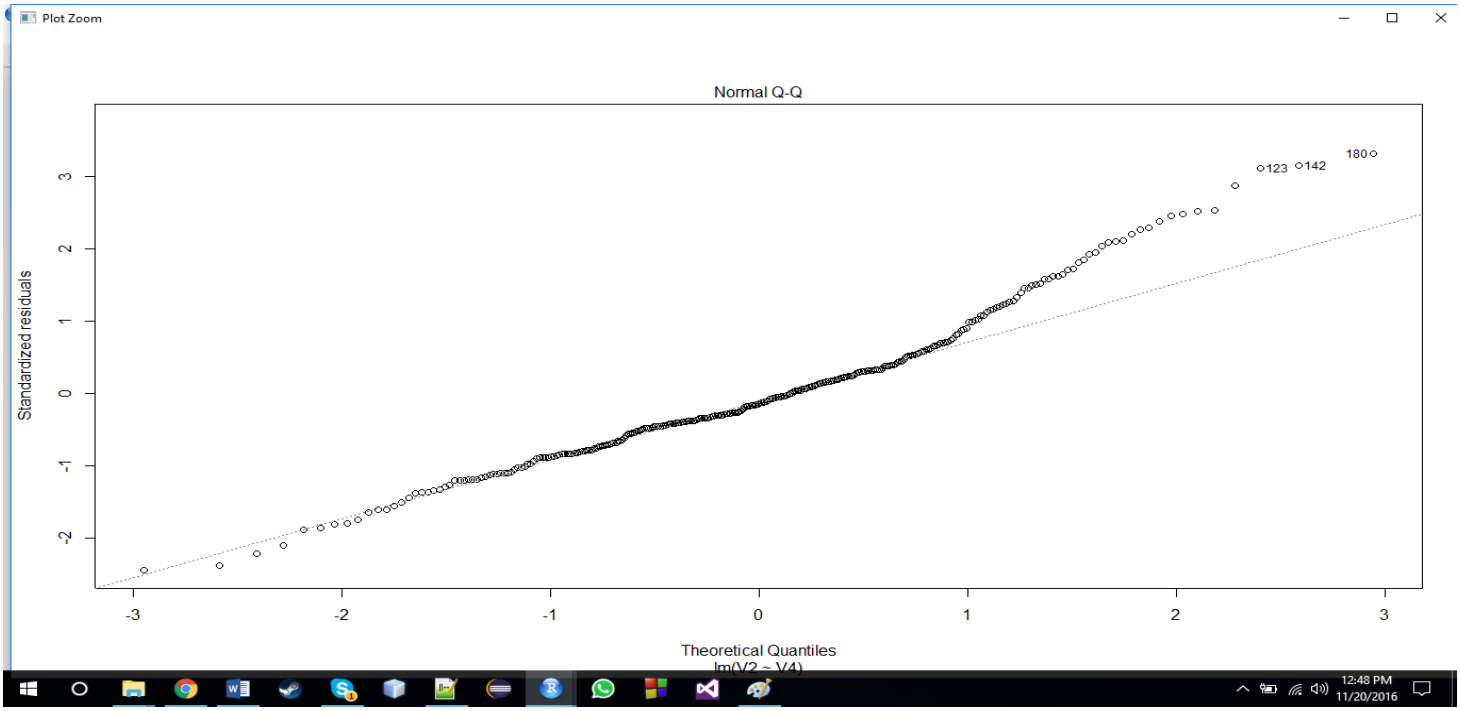

Figure 4: Normal Q-Q graph

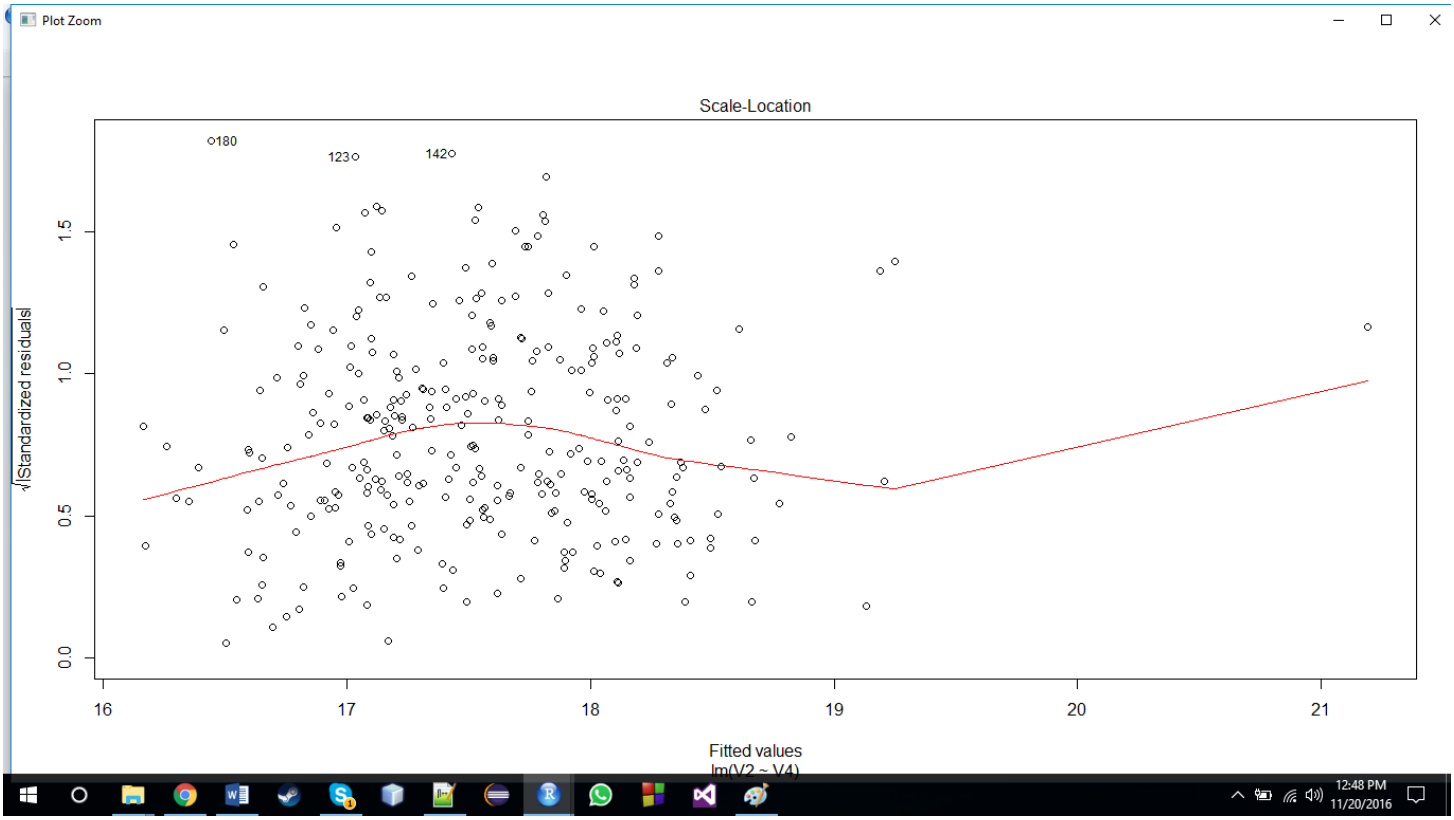

Figure 5: Scale v/s Location graph

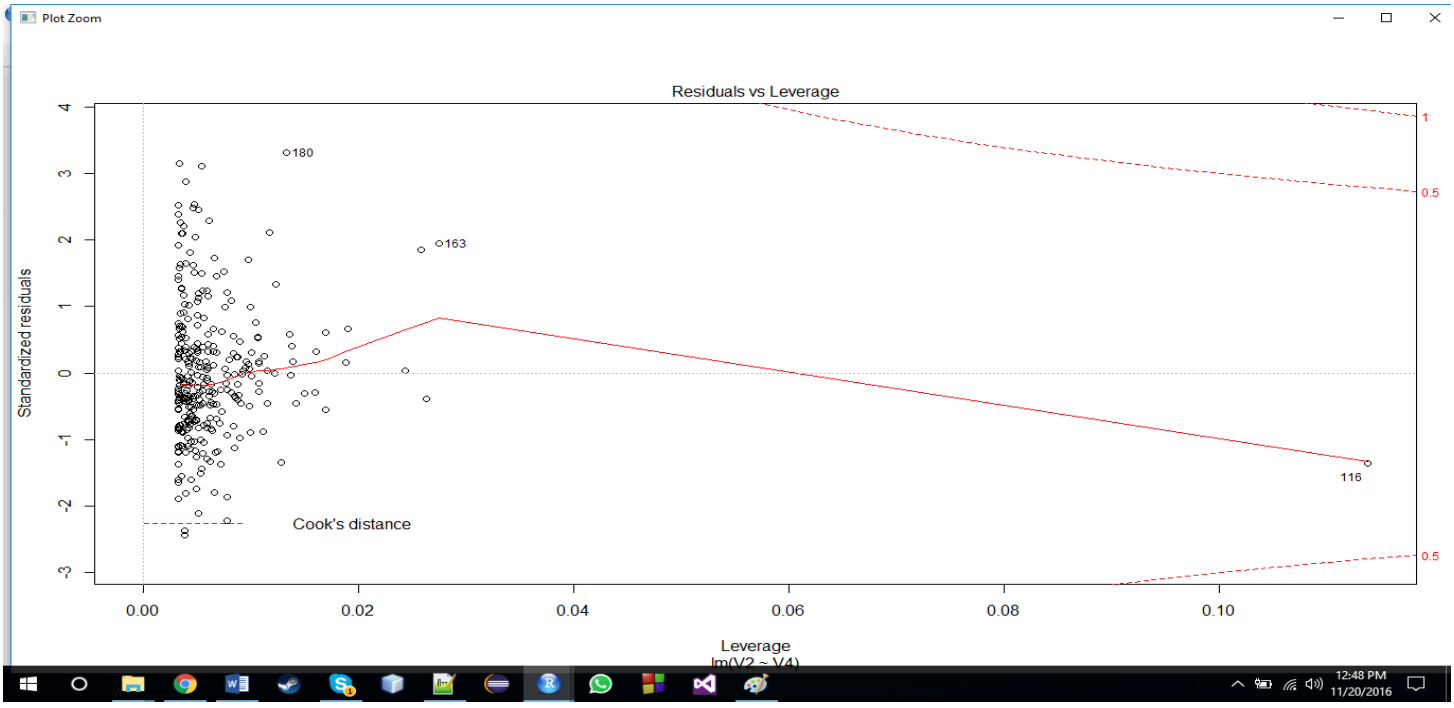

Figure 6: Residual v/s Leverage graph

Step 4: Plotting the Testing Data

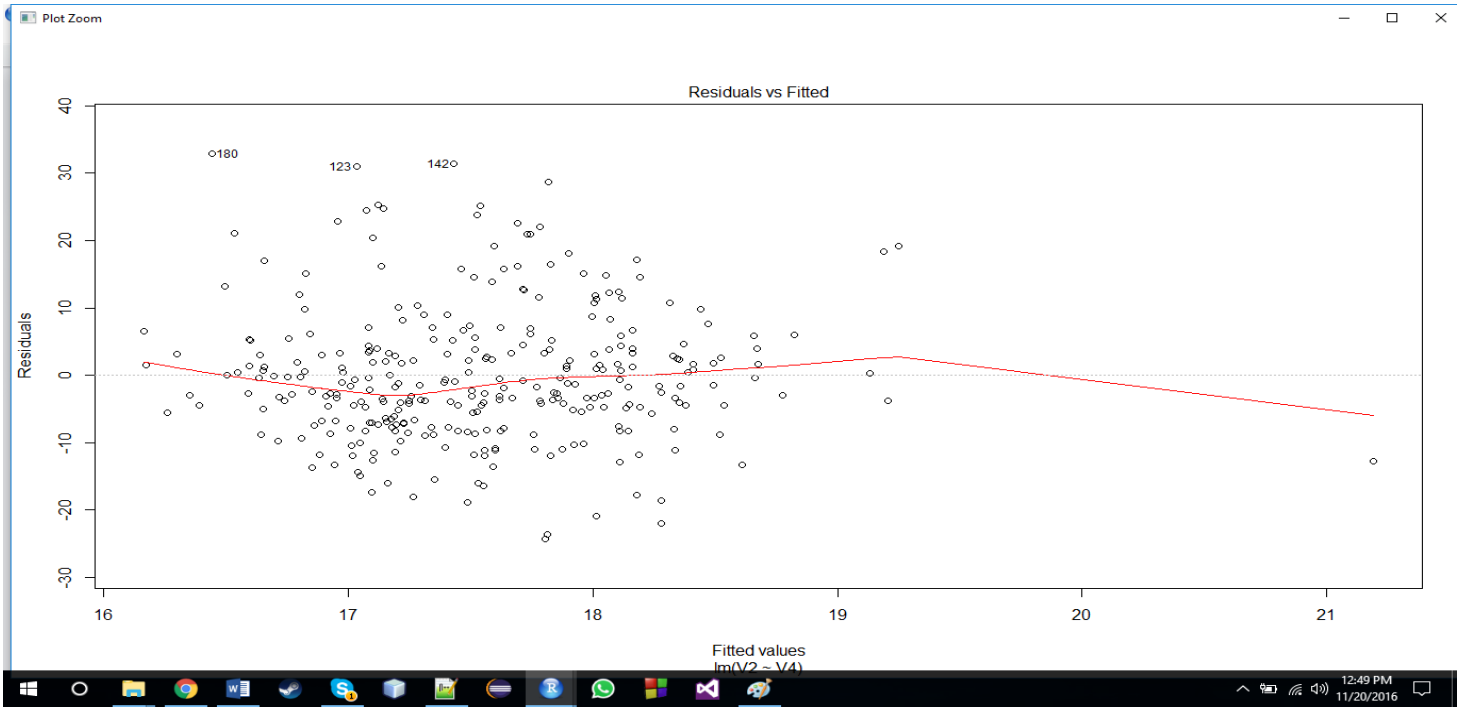

Figure 7: This is the residual v/s fitted values curve

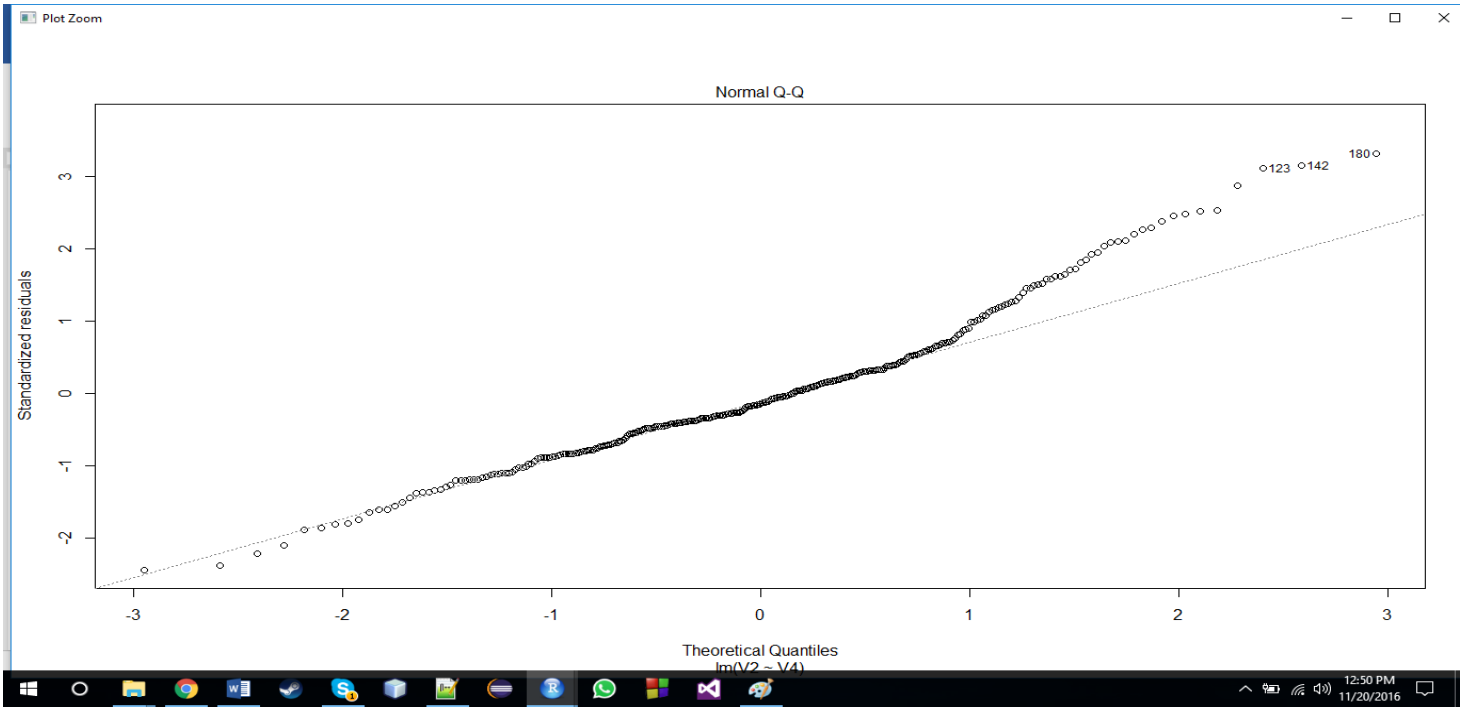

Figure 8: Normal Q-Q graph

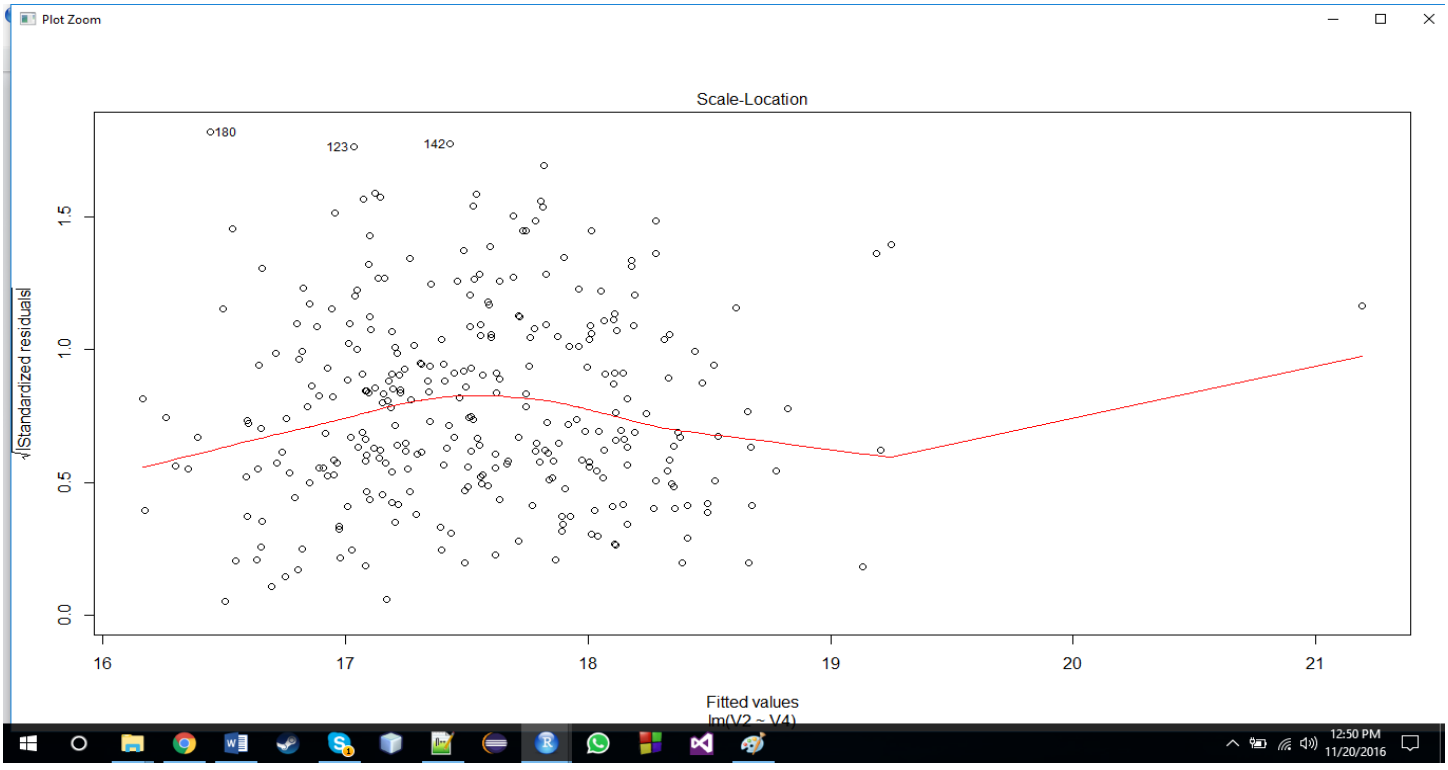

Figure 9: Scale v/s Location graph

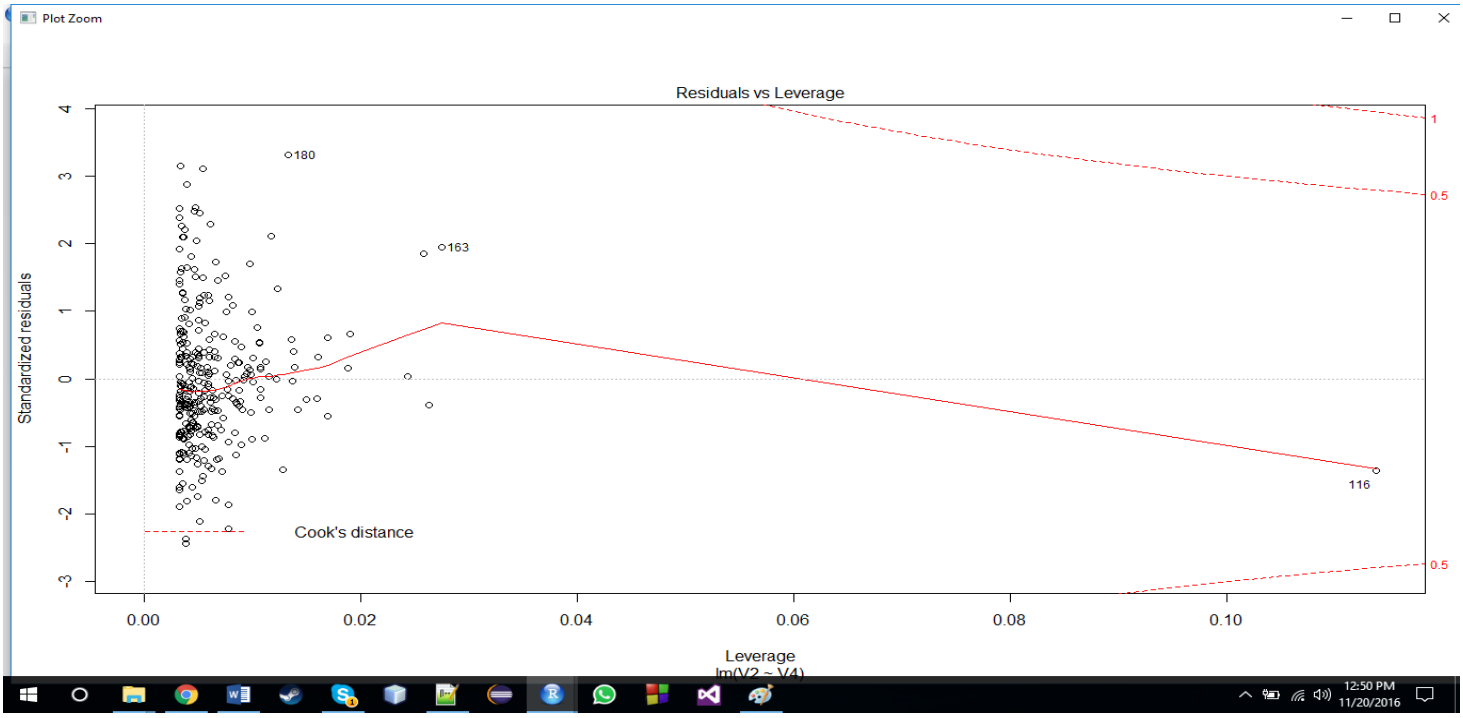

Figure 10: Residual v/s Leverage graph

Step 5: Theoretically comparing Training data and testing data:

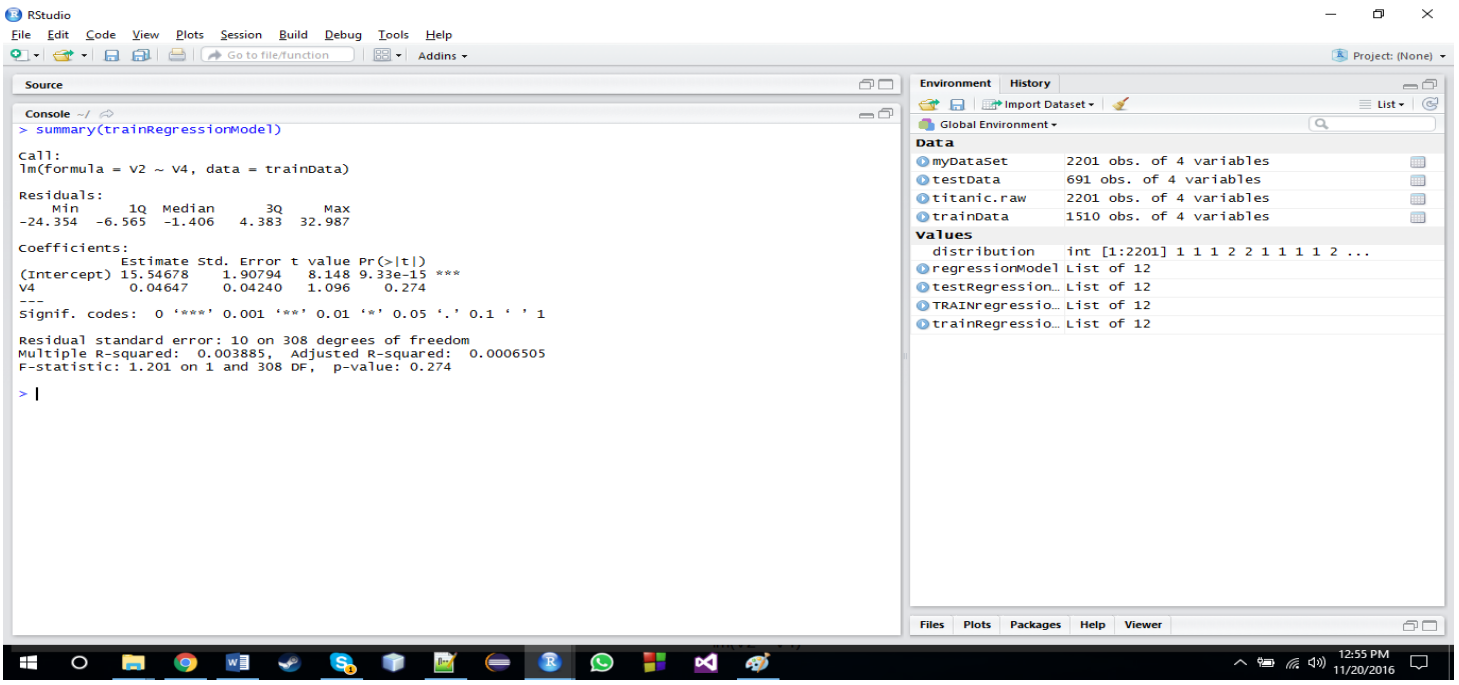

Figure 11: summary of the training Data

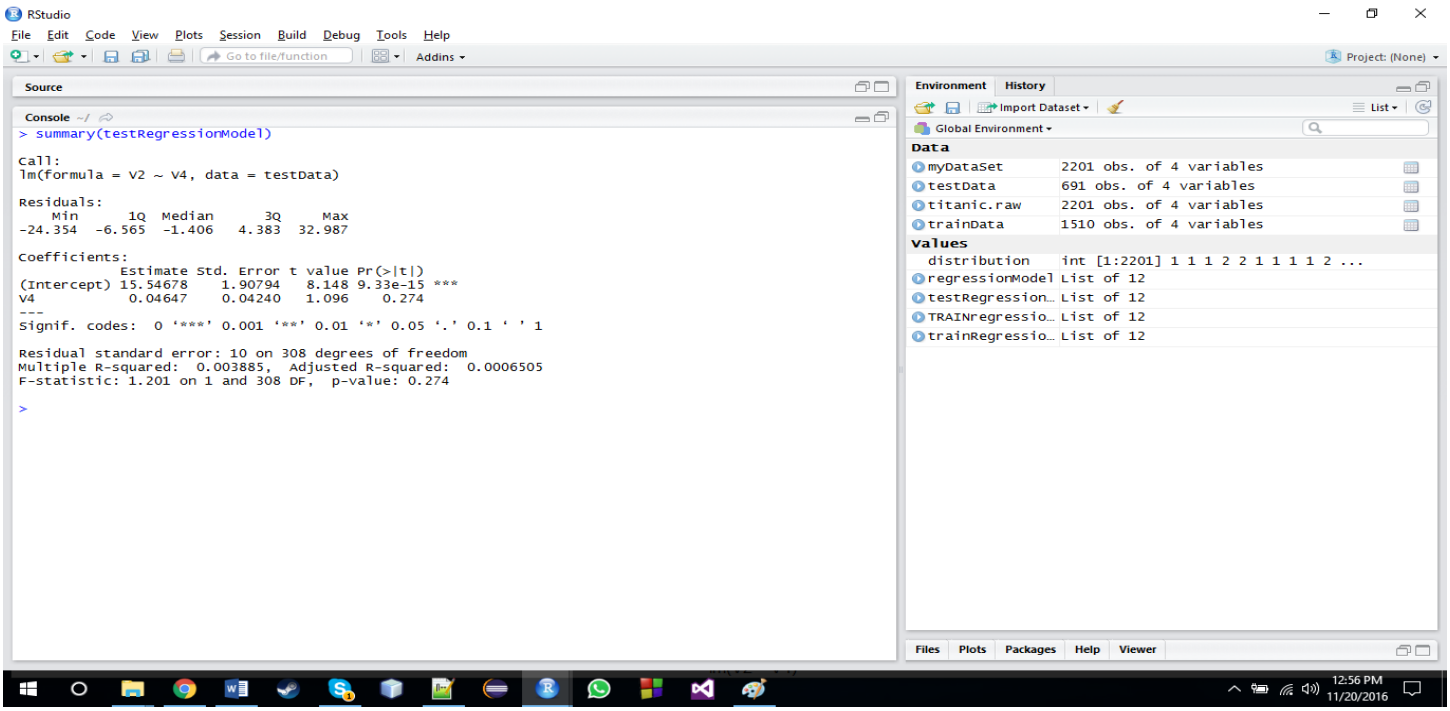

Figure 12: summary of the testing Data

Step 6: Creating decision tree model for the Original data

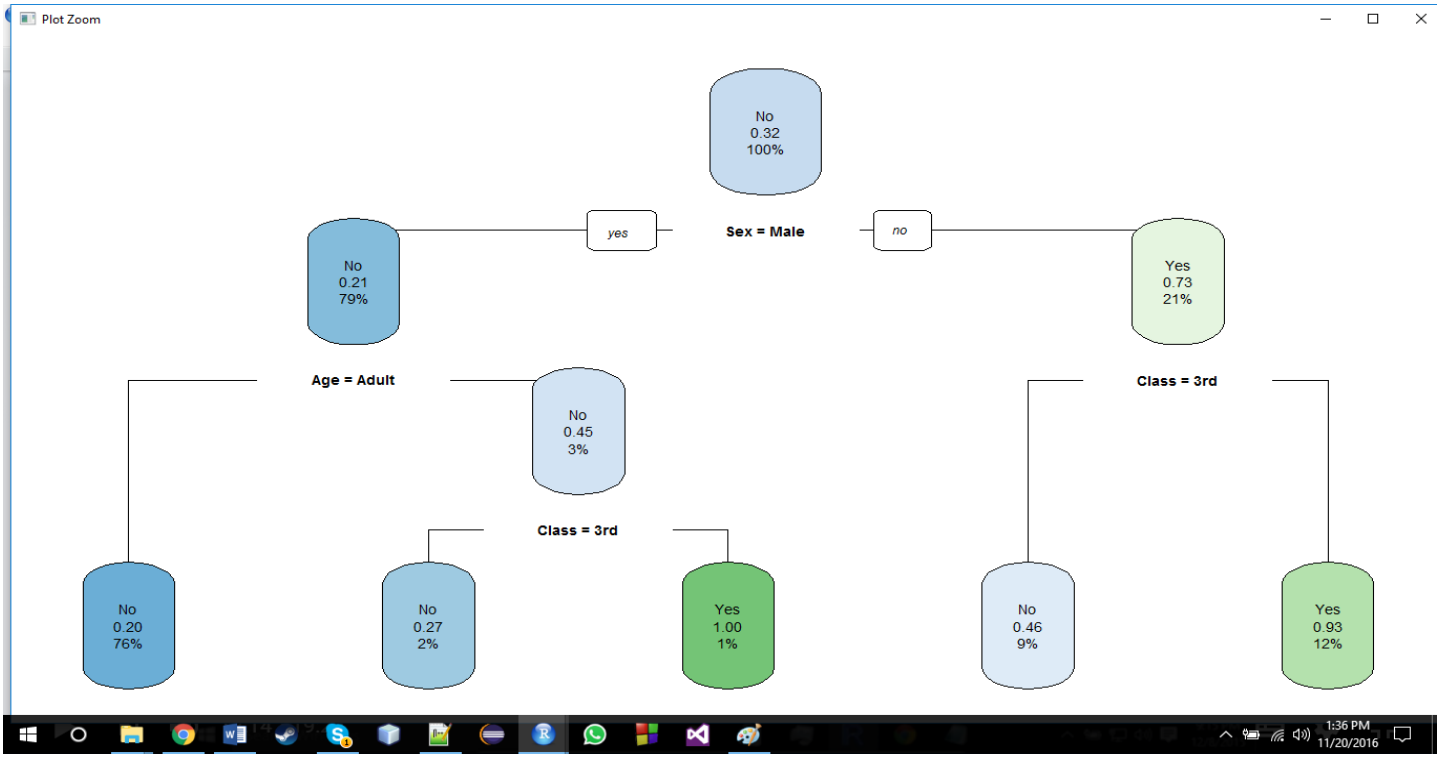

Figure 13: Decision tree classifier on original data

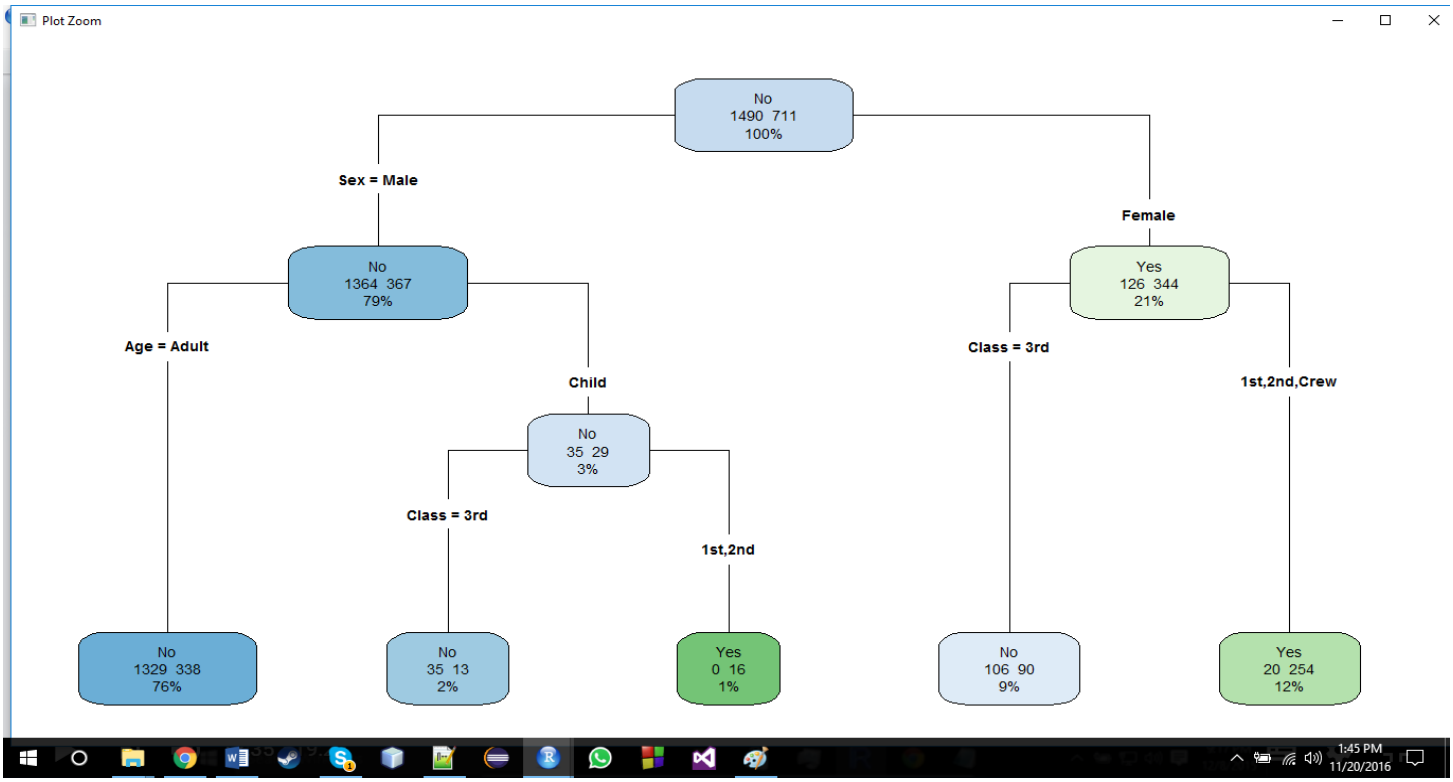

Figure 14: Simplified decision tree classifier on original data

Summary on the decision tree model for the original data

```
summary(decisionTreeModel)
Call:
rpart(formula = Survived ~ ., data = myDataSet, method = "class")
  n= 2201

         CP nsplit rel error    xerror       xstd
1 0.30661041      0 1.0000000 1.0000000 0.03085662
2 0.02250352      1 0.6933896 0.6933896 0.02750982
3 0.01125176      2 0.6708861 0.7018284 0.02762806
4 0.01000000      4 0.6483826 0.6905767 0.02747003

Variable importance
  Sex Class   Age
   73    23     4

Node number 1: 2201 observations,    complexity param=0.3066104
  predicted class=No   expected loss=0.323035  P(node) =1
    class counts:  1490   711
   probabilities: 0.677 0.323
  left son=2 (1731 obs) right son=3 (470 obs)
  Primary splits:
      Sex   splits as  RL,   improve=199.821600, (0 missing)
```

```
     Class splits as  RRLL, improve= 69.684100, (0 missing)
     Age   splits as  LR,   improve=  9.165241, (0 missing)

Node number 2: 1731 observations,    complexity param=0.01125176
  predicted class=No   expected loss=0.2120162  P(node) =0.7864607
    class counts:  1364   367
   probabilities: 0.788 0.212
  left son=4 (1667 obs) right son=5 (64 obs)
  Primary splits:
     Age   splits as  LR,   improve=7.726764, (0 missing)
     Class splits as  RLLL, improve=7.046106, (0 missing)

Node number 3: 470 observations,    complexity param=0.02250352
  predicted class=Yes  expected loss=0.2680851  P(node) =0.2135393
    class counts:   126   344
   probabilities: 0.268 0.732
  left son=6 (196 obs) right son=7 (274 obs)
  Primary splits:
     Class splits as  RRLR, improve=50.015320, (0 missing)
     Age   splits as  RL,   improve= 1.197586, (0 missing)
  Surrogate splits:
     Age splits as  RL, agree=0.619, adj=0.087, (0 split)

Node number 4: 1667 observations
  predicted class=No   expected loss=0.2027594  P(node) =0.757383
    class counts:  1329   338
   probabilities: 0.797 0.203

Node number 5: 64 observations,    complexity param=0.01125176
  predicted class=No   expected loss=0.453125  P(node) =0.02907769
    class counts:    35    29
   probabilities: 0.547 0.453
  left son=10 (48 obs) right son=11 (16 obs)
  Primary splits:
     Class splits as  RRL-, improve=12.76042, (0 missing)

Node number 6: 196 observations
  predicted class=No   expected loss=0.4591837  P(node) =0.08905043
    class counts:   106    90
   probabilities: 0.541 0.459

Node number 7: 274 observations
  predicted class=Yes  expected loss=0.0729927  P(node) =0.1244889
    class counts:    20   254
   probabilities: 0.073 0.927

Node number 10: 48 observations
  predicted class=No   expected loss=0.2708333  P(node) =0.02180827
```

class counts: 35 13
probabilities: 0.729 0.271

Node number 11: 16 observations
predicted class=Yes expected loss=0 P(node) =0.007269423
class counts: 0 16
probabilities: 0.000 1.000

Step 7: Creating decision tree model for the Training data

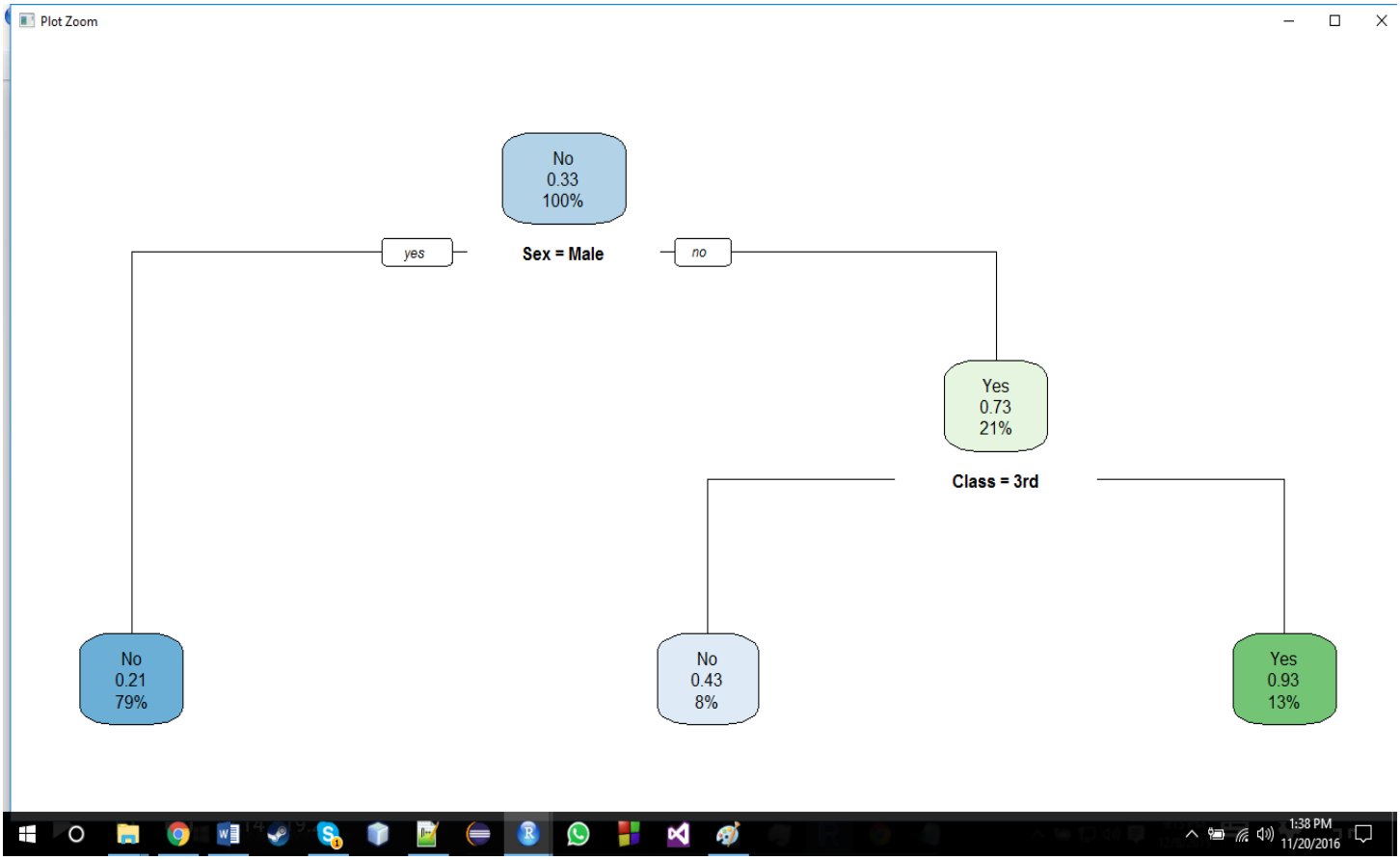

Figure 15: decision tree classifier on Training data

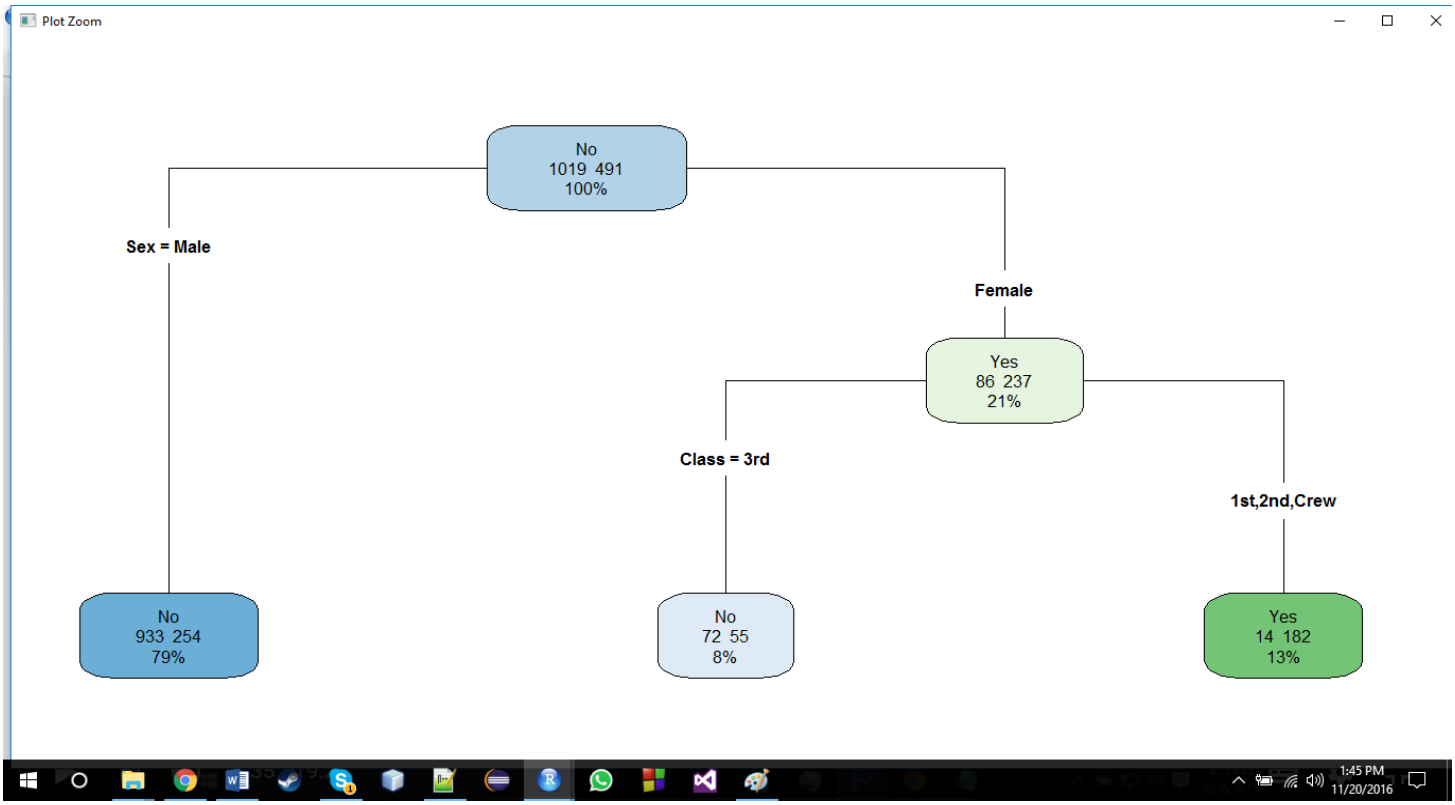

Figure 16: Simplified decision tree classifier on Training data

Summary on the decision tree model for the training data

```
summary(trainingDecisionTreeModel)
Call:
rpart(formula = Survived ~ ., data = trainData, method = "class")
  n= 1510

         CP nsplit rel error    xerror      xstd
1 0.30753564      0 1.0000000 1.0000000 0.03707301
2 0.03462322      1 0.6924644 0.6924644 0.03305692
3 0.01000000      2 0.6578411 0.6578411 0.03245313

Variable importance
  Sex Class   Age
   76    21     3

Node number 1: 1510 observations,    complexity param=0.3075356
  predicted class=No   expected loss=0.3251656  P(node) =1
    class counts:  1019   491
   probabilities: 0.675 0.325
```

left son=2 (1187 obs) right son=3 (323 obs)
Primary splits:
Sex splits as RL, improve=137.187400, (0 missing)
Class splits as RRLL, improve= 54.886010, (0 missing)
Age splits as LR, improve= 3.784003, (0 missing)

Node number 2: 1187 observations
predicted class=No expected loss=0.2139848 P(node) =0.7860927
class counts: 933 254
probabilities: 0.786 0.214

Node number 3: 323 observations, complexity param=0.03462322
predicted class=Yes expected loss=0.2662539 P(node) =0.2139073
class counts: 86 237
probabilities: 0.266 0.734
left son=6 (127 obs) right son=7 (196 obs)
Primary splits:
Class splits as RRLR, improve=37.842130, (0 missing)
Age splits as RL, improve= 2.068422, (0 missing)
Surrogate splits:
Age splits as RL, agree=0.659, adj=0.134, (0 split)

Node number 6: 127 observations
predicted class=No expected loss=0.4330709 P(node) =0.08410596
class counts: 72 55
probabilities: 0.567 0.433

Node number 7: 196 observations
predicted class=Yes expected loss=0.07142857 P(node) =0.1298013
class counts: 14 182
probabilities: 0.071 0.929

Step 8: Creating decision tree model for the testing data

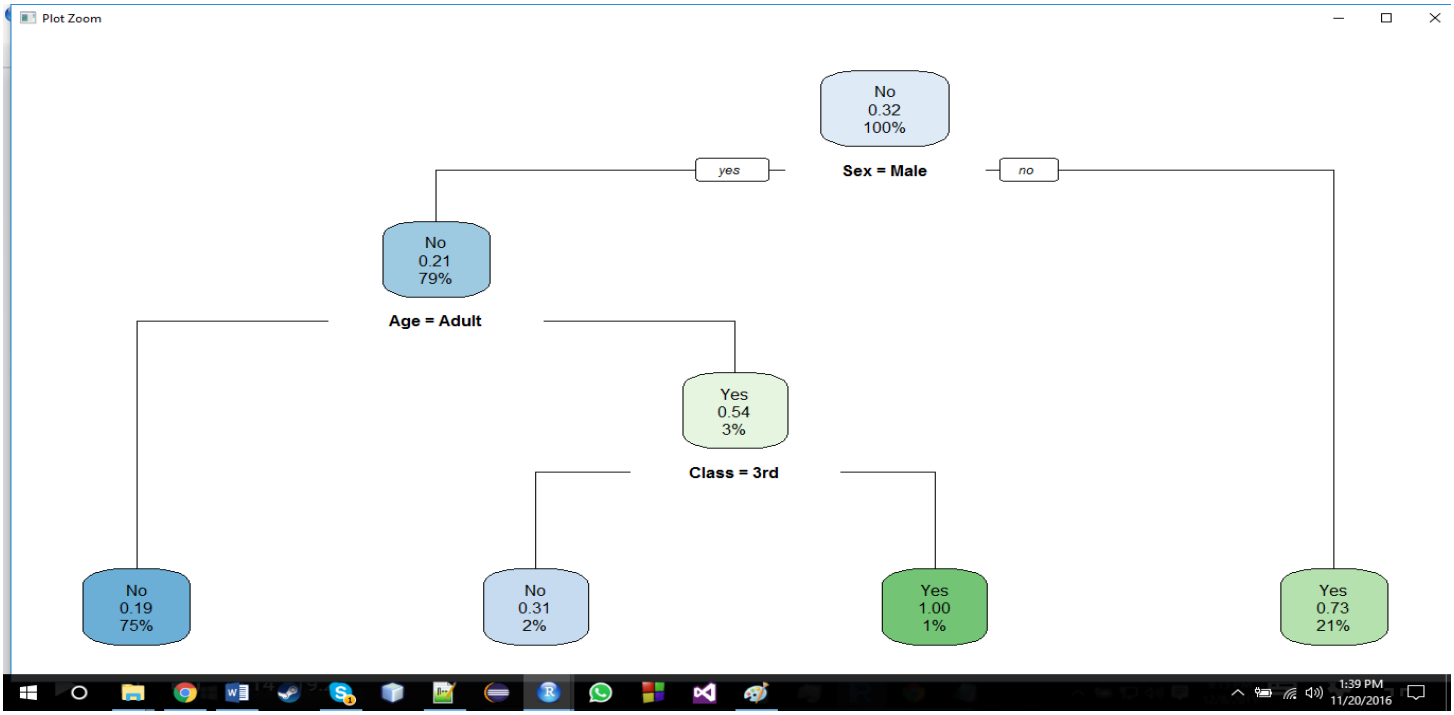

Figure 16: decision tree classifier on testing data

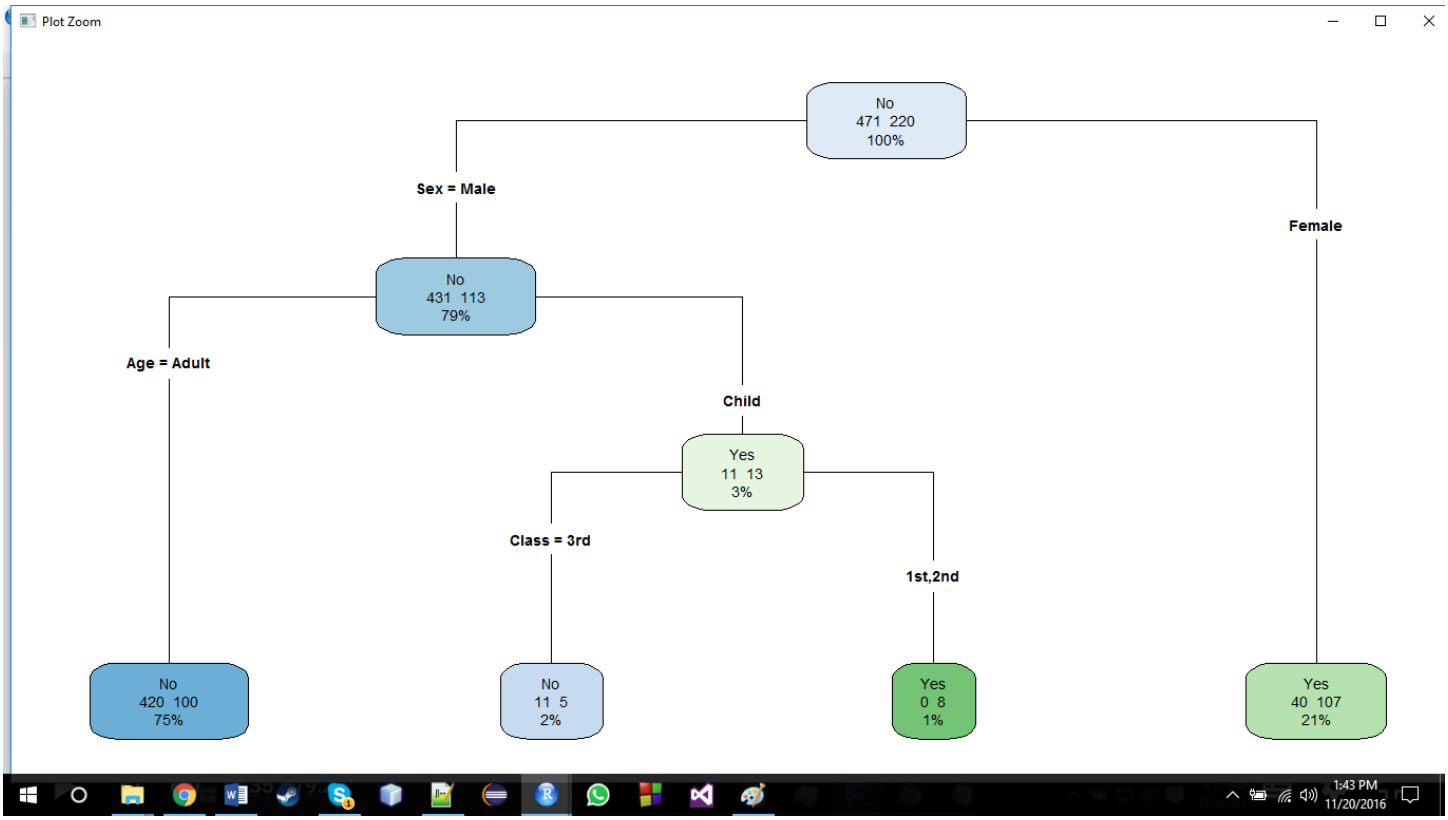

Figure 16: Simplified decision tree classifier on Testing data

Summary on the decision tree model for the testing data

```
summary(testingDecisionTreeModel)
Call:
rpart(formula = Survived ~ ., data = testData, method = "class")
  n= 691

         CP nsplit rel error    xerror       xstd
1 0.30454545      0 1.0000000 1.0000000 0.05566216
2 0.01818182      1 0.6954545 0.6954545 0.04961068
3 0.01000000      3 0.6590909 0.6772727 0.04913954

Variable importance
  Sex   Age Class
   85     8     7

Node number 1: 691 observations,    complexity param=0.3045455
  predicted class=No   expected loss=0.3183792  P(node) =1
    class counts:   471   220
   probabilities: 0.682 0.318
  left son=2 (544 obs) right son=3 (147 obs)
  Primary splits:
      Sex   splits as  RL,   improve=62.626730, (0 missing)
      Class splits as  RLLL, improve=20.180600, (0 missing)
      Age   splits as  LR,   improve= 6.405335, (0 missing)

Node number 2: 544 observations,    complexity param=0.01818182
  predicted class=No   expected loss=0.2077206  P(node) =0.7872648
    class counts:   431   113
   probabilities: 0.792 0.208
  left son=4 (520 obs) right son=5 (24 obs)
  Primary splits:
      Age   splits as  LR,   improve=5.600019, (0 missing)
      Class splits as  RLLL, improve=2.974835, (0 missing)

Node number 3: 147 observations
  predicted class=Yes  expected loss=0.2721088  P(node) =0.2127352
    class counts:    40   107
   probabilities: 0.272 0.728

Node number 4: 520 observations
  predicted class=No   expected loss=0.1923077  P(node) =0.7525326
    class counts:   420   100
```

probabilities: 0.808 0.192

Node number 5: 24 observations, complexity param=0.01818182
predicted class=Yes expected loss=0.4583333 P(node) =0.03473227
class counts: 11 13
probabilities: 0.458 0.542
left son=10 (16 obs) right son=11 (8 obs)
Primary splits:
Class splits as RRL-, improve=5.041667, (0 missing)

Node number 10: 16 observations
predicted class=No expected loss=0.3125 P(node) =0.02315485
class counts: 11 5
probabilities: 0.688 0.312

Node number 11: 8 observations
predicted class=Yes expected loss=0 P(node) =0.01157742
class counts: 0 8
probabilities: 0.000 1.000

REFERENCES:

[1] http://www.rdatamining.com/resources/data

[2] www.quora.com

[3] www.stackoverflow.com

[4] www.youtube.com

[5] www.Wikipedia.com

[6] www.cran.r-project.org