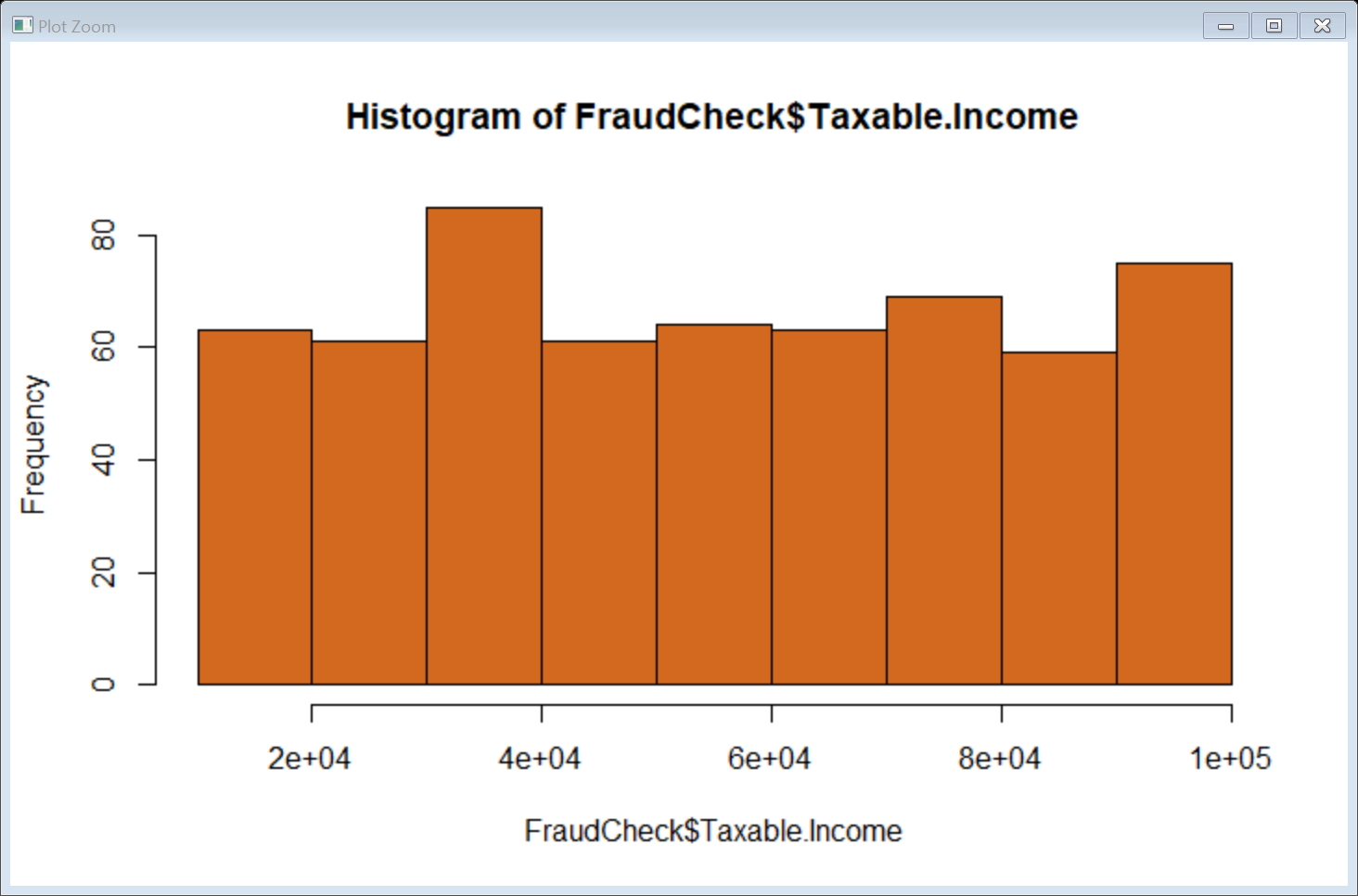
# Random Forest

**Business Objective**: Prepare a model on fraud data treating those who have taxable\_income <= 30000 as "Risky" and others are "Good"

Step 1: Install the package caret, randomforest,mass and read the csv file.

Step 2: According to the given statement if Taxable Income is less than or equal to 30000 then Risky else Good

Step 3: Histogram is created to identify the data distribution.



Step 4: table(FC$Risky\_Good) will show the number of good and risky.

Good Risky

476 124

Step 5: Confusion matrix:

Good Risky class.error

Good 343 1 0.002906977

Risky 1 79 0.012500000

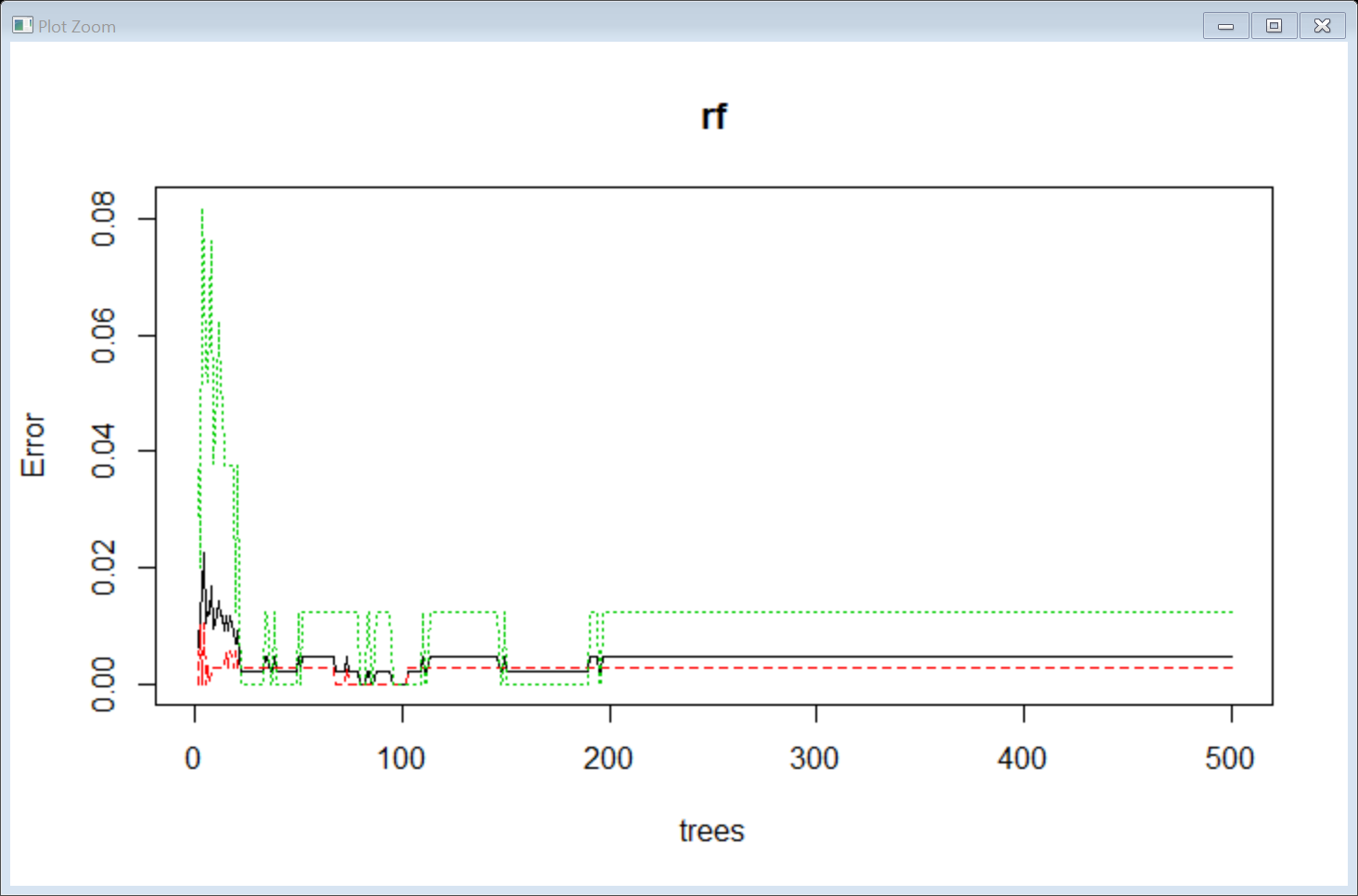
Step 6: head(pred1)

1 3 6 7 9 10

Good Good Good Good Good Good

Levels: Good Risky

Step 7: Plot the Test data of prediction



Step 8: Do the confusion matrix on the train data with proximity = True.

Confusion matrix:

Good Risky class.error

Good 343 1 0.002906977

Risky 1 79 0.012500000

Step 9: Confusion Matrix and Statistics Train data Accuracy = 100%

Reference

Prediction Good Risky

Good 344 0

Risky 0 80

Step 10: Confusion Matrix and Statistics Test data Accuracy = 100%

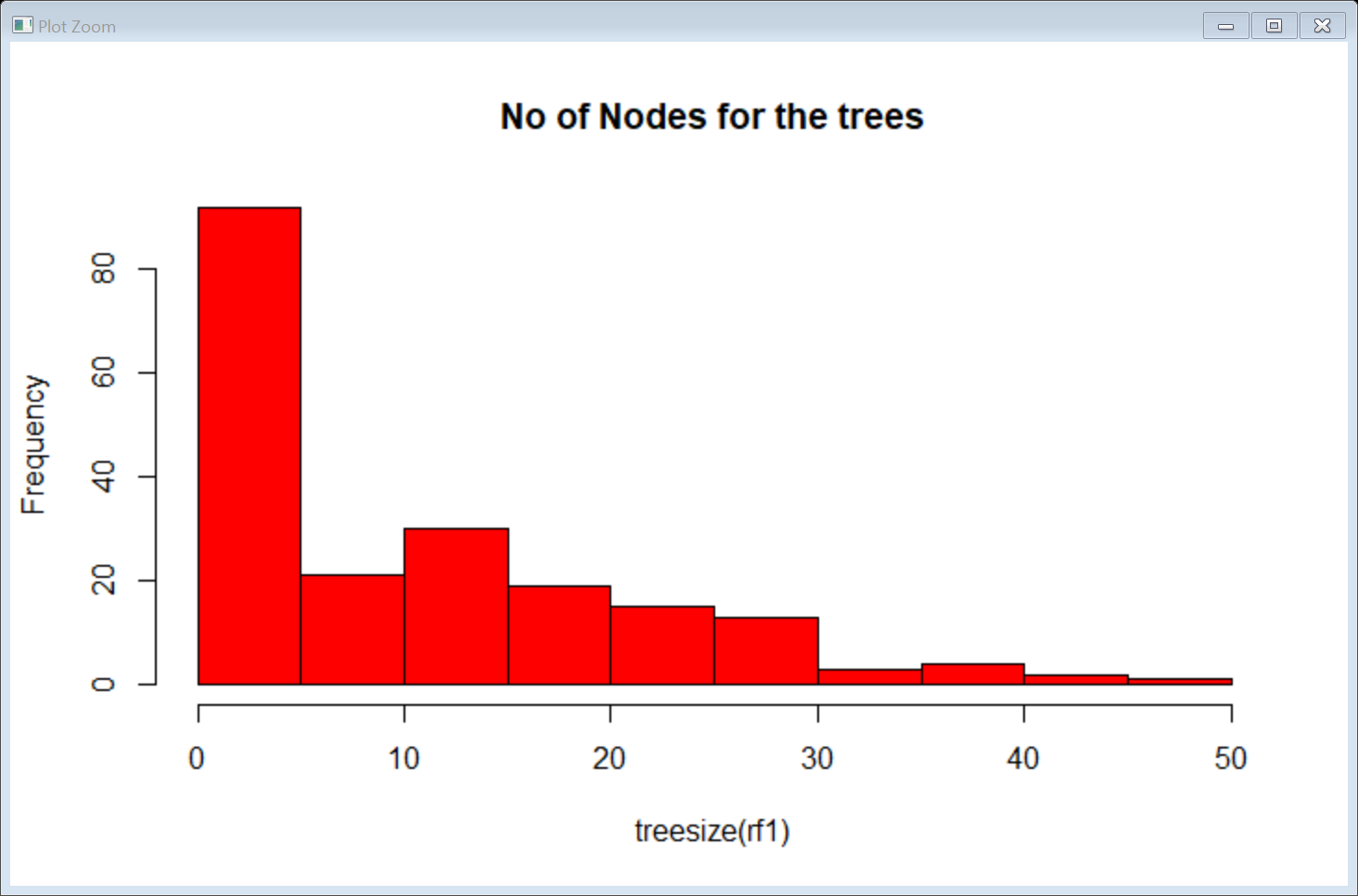
Reference

Prediction Good Risky

Good 132 0

Risky 0 44

Step 11: Histogram for no of nodes for the trees.

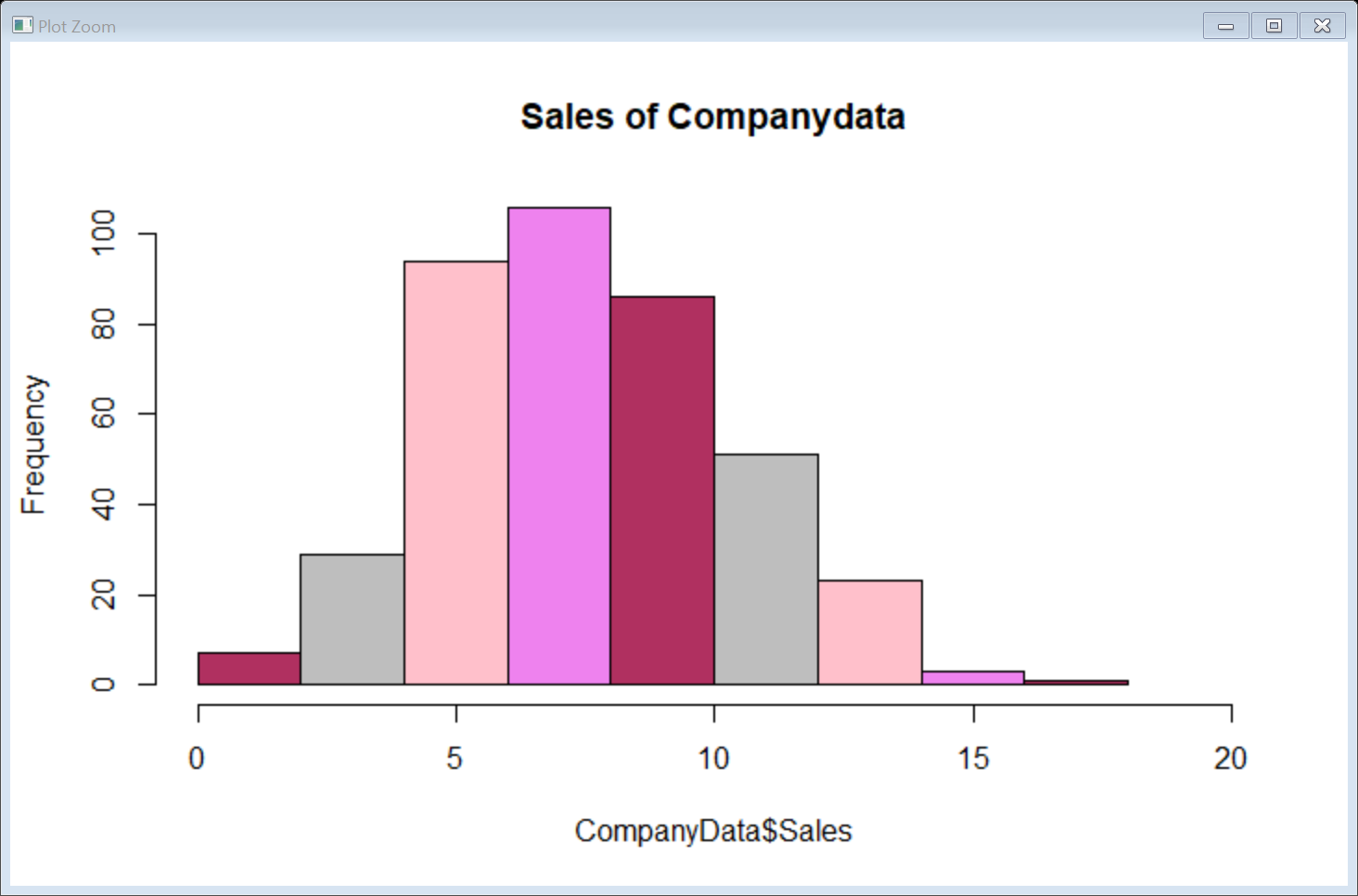


**Conclusion:** Confidence Interval is around 97%

**Business Objective :** A cloth manufacturing company is interested to know about the segment or attributes causes high sale.

Step 1: Install the package caret, randomforest,mass and read the csv file.

Step 2:Histogram represents the sales of companydata.



Step 3: table(CD$highsales)

No Yes

286 114

Step 4: head(pred1)

1 3 6 7 9 10

Yes Yes Yes No No No

Levels: No Yes

Step 5: head(train$highsales)

[1] Yes Yes Yes No No No

Levels: No Yes

Step 6:

Confusion Matrix and Statistics Accuracy= 100% for train data

Reference

Prediction No Yes

No 210 0

Yes 0 75

Step 7:

Confusion Matrix and Statistics Accuracy = 84.35 for test data

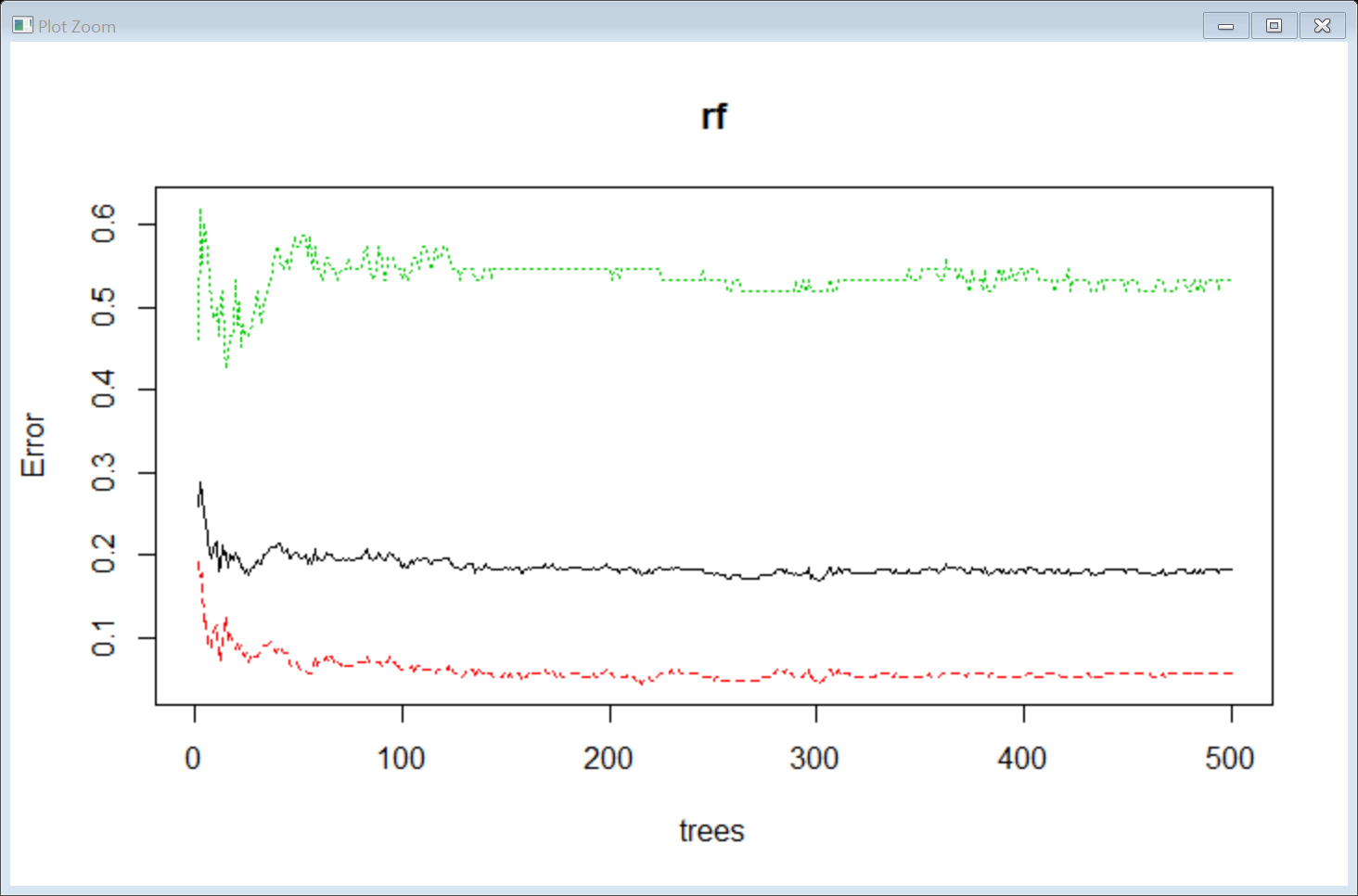
Reference

Prediction No Yes

No 74 16

Yes 2 23

Step 8 :Plot(rf)



Step 9: No. of variables tried at each split: 3

OOB estimate of error rate: 16.49%

Confusion matrix:

No Yes class.error

No 199 11 0.05238095

Yes 36 39 0.48000000

Step 10: Confusion Matrix and Statistics for train data Accuracy = 100%

Reference

Prediction No Yes

No 210 0

Yes 0 75

Step 11:

Confusion Matrix and Statistics for test data Accuracy = 82.61

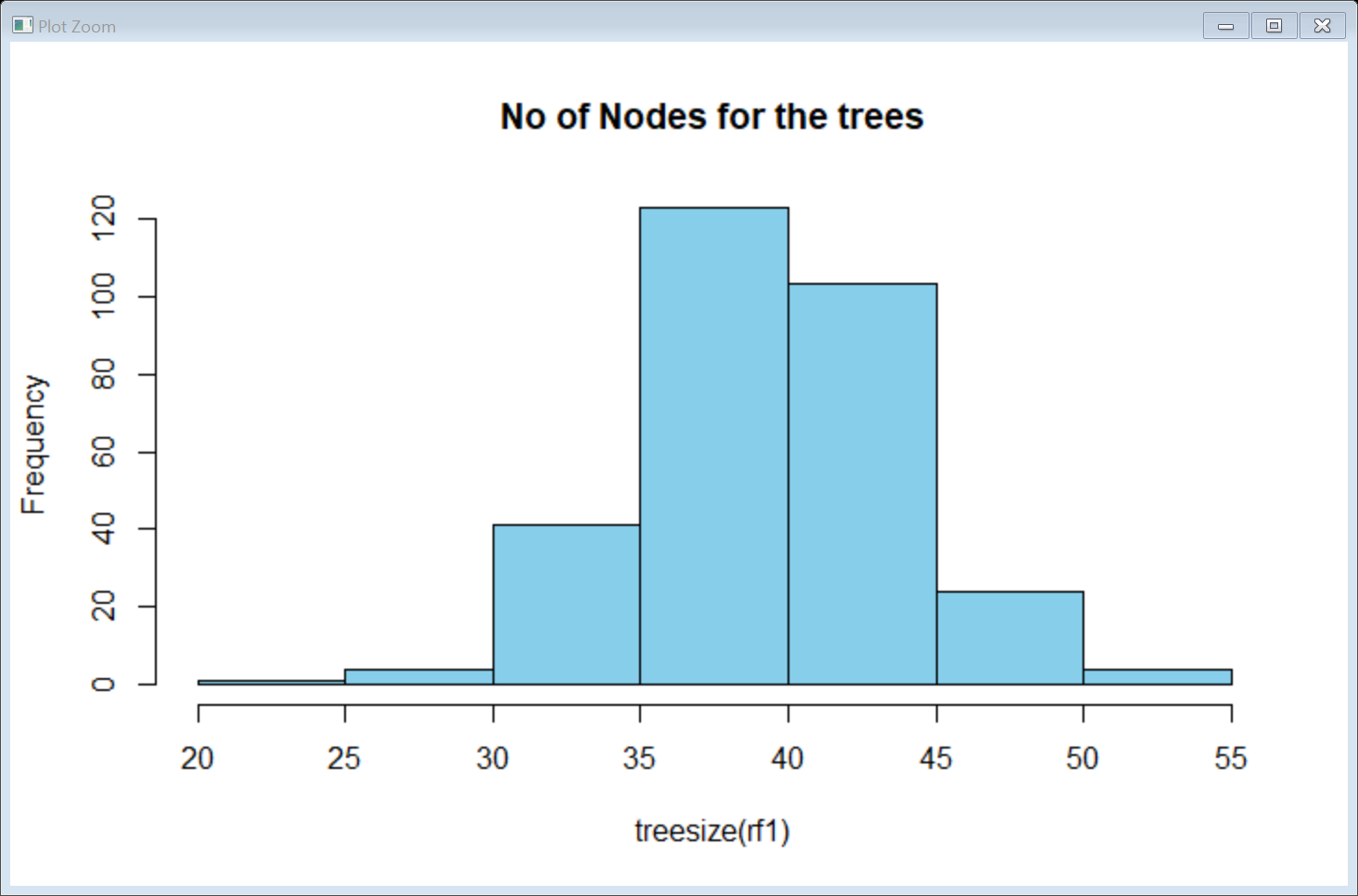
Reference

Prediction No Yes

No 74 18

Yes 2 21

Step 12:Histogram showing no of nodes of tree



Conclusion : Confidence interval is around 90%

**Business Objective:** Build a random forest for the ‘iris’ data

Step 1: Install all the package and use the data iris.

Step 2: Split the data into setosa,versicolor,virginica 50 each.

Step 3:Apply Random forest method on output variable species with all other input variables.

Confusion matrix:

setosa versicolor virginica class.error

setosa 25 0 0 0.00

versicolor 0 24 1 0.04

virginica 0 2 23 0.08

Step 4:

Prediction and Confusion Matrix - Training data

head(pred1)

1 2 3 4 5 6

setosa setosa setosa setosa setosa setosa

Levels: setosa versicolor virginica

Step 5: Traing data - head(iris\_train$Species)

[1] setosa setosa setosa setosa setosa setosa

Levels: setosa versicolor virginica

Step 6: Confusion matrix for train data Accuracy = 100%

Confusion Matrix and Statistics

Reference

Prediction setosa versicolor virginica

setosa 25 0 0

versicolor 0 25 0

virginica 0 0 25

Step 7: Prediction with test data - Test Data Accuracy = 94.67%

Confusion Matrix and Statistics

Reference

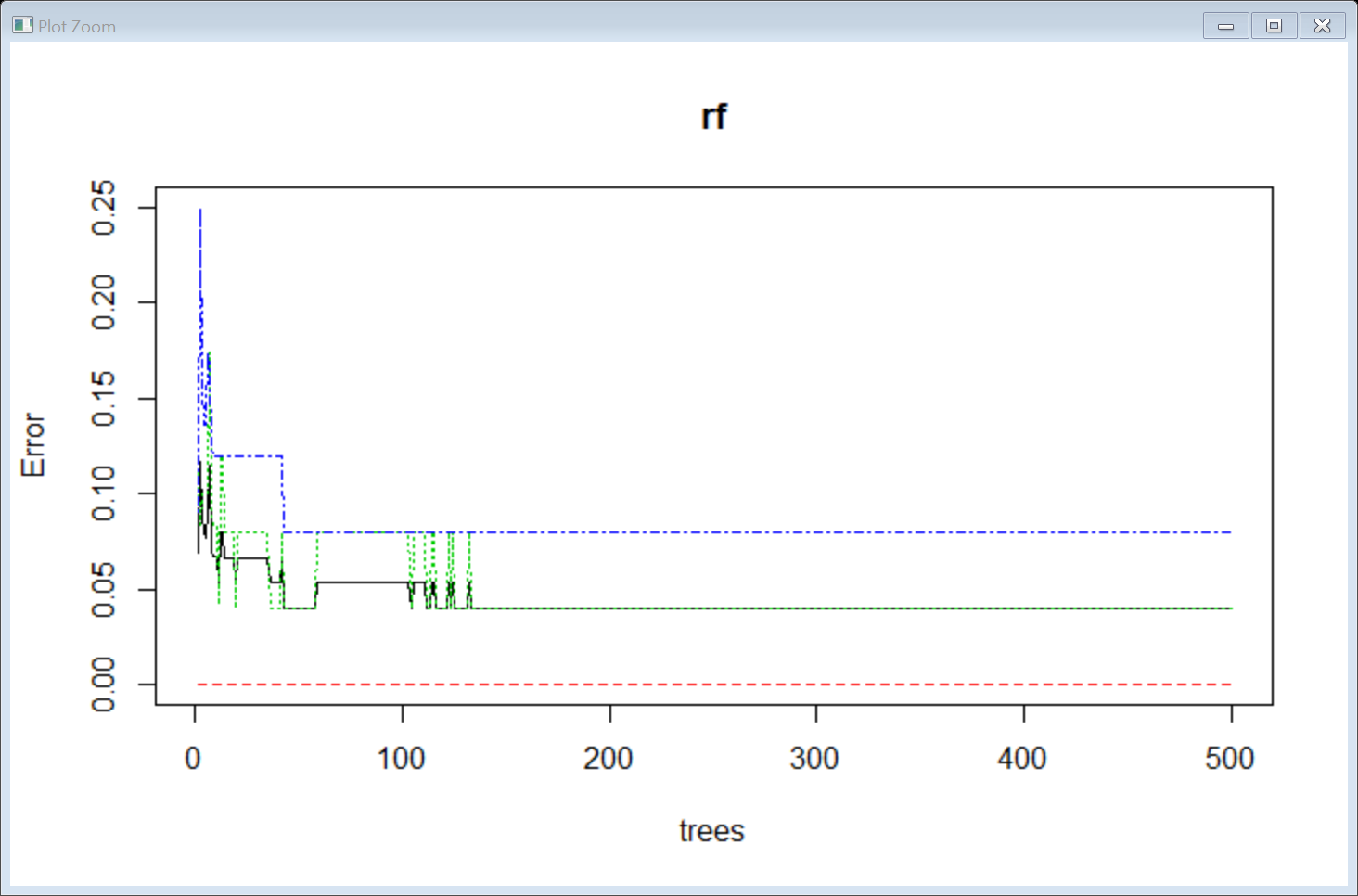
Prediction setosa versicolor virginica

setosa 25 0 0

versicolor 0 23 2

virginica 0 2 23

Step 8: Error Rate in Random Forest Model



Step 9: OOB estimate of error rate: 4%

Confusion matrix:

setosa versicolor virginica class.error

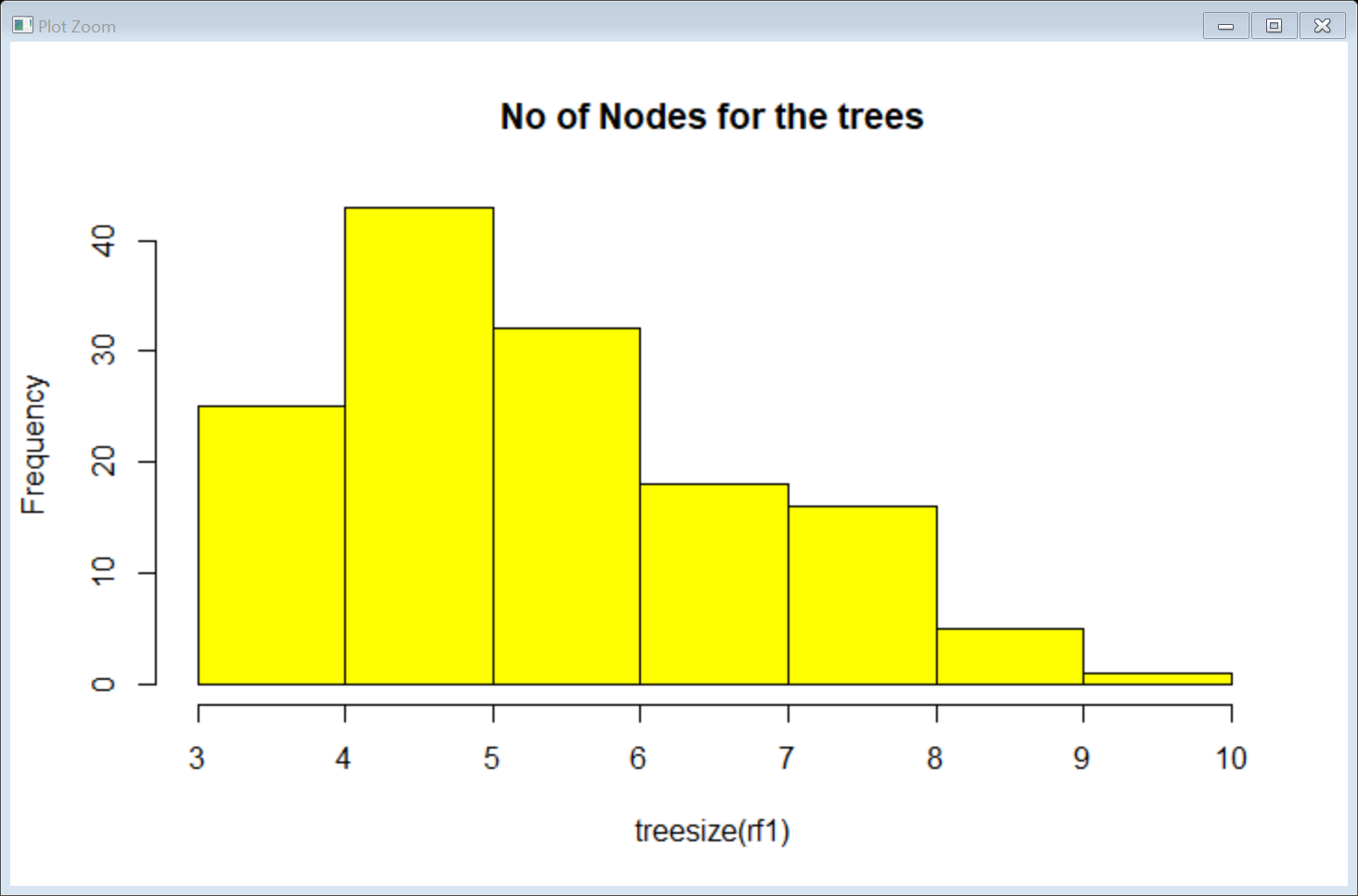
setosa 25 0 0 0.00

versicolor 0 24 1 0.04

virginica 0 2 23 0.08

Step 10: confusion matrix for rf1 model train and test data remains same as rf.

Step 10: Histogram for the no of nodes for the trees.



**Conclusion** : Species setosa = 100% accurate, Versicolor = 92% accurate, Verginica = 96% accurate.

Confidence Interval is between 88% to 99%