Project STAT 6620

**Chinki**

**California State University, East Bay**

**Forest type mapping Dataset. source: UCI Machine Learning Repository**

**Step-01: Collecting the data**

This data set contains training and testing data from a remote sensing study which mapped different forest types based on their spectral characteristics at visible-to-near infrared wavelengths, using ASTER satellite imagery. The output (forest type map) can be used to identify and/or quantify the ecosystem services (e.g. carbon storage, erosion protection) provided by the forest.

Class: 's' ('Sugi' forest), 'h' ('Hinoki' forest), 'd' ('Mixed deciduous' forest), 'o' ('Other' non-forest land).

b1 - b9: ASTER image bands containing spectral information in the green, red, and near infrared wavelengths for three dates (Sept. 26, 2010; March 19, 2011; May 08, 2011.   
pred\_minus\_obs\_S\_b1 - pred\_minus\_obs\_S\_b9: Predicted spectral values (based on spatial interpolation) minus actual spectral values for the 's' class (b1-b9).   
pred\_minus\_obs\_H\_b1 - pred\_minus\_obs\_H\_b9: Predicted spectral values (based on spatial interpolation) minus actual spectral values for the 'h' class (b1-b9). Our aim is to map forests, so it is classification problem.

**Step-02: Exploring and Preparing the data**

#Reading the data  
Forest\_train=read.csv("C:/Users/chink/Downloads/ForestTypes/testing.csv",na.strings = F)  
Forest\_test=read.csv("C:/Users/chink/Downloads/ForestTypes/training.csv",na.strings = F)

#Examine the structure  
str(Forest\_train)

## 'data.frame': 325 obs. of 28 variables:  
## $ class : Factor w/ 4 levels "d ","h ","o ",..: 1 4 4 1 4 1 2 3 4 1 ...  
## $ b1 : int 67 67 63 63 46 59 83 63 77 57 ...  
## $ b2 : int 51 28 26 42 27 59 28 37 29 44 ...  
## $ b3 : int 68 51 50 63 50 84 54 58 52 65 ...  
## $ b4 : int 115 99 95 97 83 93 117 95 103 107 ...  
## $ b5 : int 69 50 49 66 51 70 51 58 51 59 ...  
## $ b6 : int 111 97 91 108 90 104 96 101 93 104 ...  
## $ b7 : int 136 82 81 111 76 92 105 89 87 98 ...  
## $ b8 : int 31 26 26 28 26 29 27 27 27 26 ...  
## $ b9 : int 67 59 57 59 56 58 64 62 59 59 ...  
## $ pred\_minus\_obs\_H\_b1: num 47.7 47.9 53.1 52.4 68.5 ...  
## $ pred\_minus\_obs\_H\_b2: num -0.27 23.77 25.72 9.76 24.27 ...  
## $ pred\_minus\_obs\_H\_b3: num 29.2 48 48.3 35.7 48.2 ...  
## $ pred\_minus\_obs\_H\_b4: num -16.32 1.76 7.16 4.44 16.37 ...  
## $ pred\_minus\_obs\_H\_b5: num -42.9 -23.9 -22.9 -39.9 -24.9 ...  
## $ pred\_minus\_obs\_H\_b6: num -49 -34.4 -28.4 -45.4 -27.7 ...  
## $ pred\_minus\_obs\_H\_b7: num -58.09 -2.89 -0.69 -31.33 2.19 ...  
## $ pred\_minus\_obs\_H\_b8: num 0.71 4.32 4.16 2.24 4.93 3.15 1.66 3.14 4.8 4.11 ...  
## $ pred\_minus\_obs\_H\_b9: num -9.17 -2.25 -0.44 -2.34 1.25 0.23 -9.18 -5.46 -1.07 -2.38 ...  
## $ pred\_minus\_obs\_S\_b1: num -18.3 -20.1 -17.6 -20.2 -18.6 ...  
## $ pred\_minus\_obs\_S\_b2: num -1.8 -2.11 -1.81 -1.89 -2.17 -1.98 -1.87 -1.74 -2.31 -2.18 ...  
## $ pred\_minus\_obs\_S\_b3: num -6.32 -6.35 -4.7 -5.47 -7.11 -6.48 -5.87 -4.98 -6.72 -6.74 ...  
## $ pred\_minus\_obs\_S\_b4: num -20.9 -21.9 -19.4 -21.6 -21.1 ...  
## $ pred\_minus\_obs\_S\_b5: num -1.63 -1.22 -0.65 -0.99 -1.56 -1.79 -1.83 -0.93 -1.77 -1.21 ...  
## $ pred\_minus\_obs\_S\_b6: num -6.13 -6.13 -5.01 -5.71 -6.35 -6.25 -7.97 -5.59 -6.29 -6.24 ...  
## $ pred\_minus\_obs\_S\_b7: num -22.6 -22.2 -20.9 -22.2 -22.2 ...  
## $ pred\_minus\_obs\_S\_b8: num -5.53 -3.41 -3.96 -3.41 -4.45 -6.5 -2 -3.26 -6.11 -3.06 ...  
## $ pred\_minus\_obs\_S\_b9: num -8.11 -6.57 -6.85 -6.52 -7.32 -8.93 -5.03 -6.37 -8.57 -6.32 ...

str(Forest\_test)

## 'data.frame': 198 obs. of 28 variables:  
## $ class : Factor w/ 4 levels "d ","h ","o ",..: 1 2 4 4 1 2 4 1 4 3 ...  
## $ b1 : int 39 84 53 59 57 85 56 40 53 51 ...  
## $ b2 : int 36 30 25 26 49 28 29 39 27 57 ...  
## $ b3 : int 57 57 49 49 66 56 50 58 49 77 ...  
## $ b4 : int 91 112 99 103 103 120 93 82 95 90 ...  
## $ b5 : int 59 51 51 47 64 52 51 61 49 89 ...  
## $ b6 : int 101 98 93 92 106 98 94 99 92 123 ...  
## $ b7 : int 93 92 84 82 114 101 77 89 63 97 ...  
## $ b8 : int 27 26 26 25 28 27 26 26 25 47 ...  
## $ b9 : int 60 62 58 56 59 65 58 57 54 83 ...  
## $ pred\_minus\_obs\_H\_b1: num 75.7 30.6 63.2 55.5 59.4 ...  
## $ pred\_minus\_obs\_H\_b2: num 14.86 20.42 26.7 24.5 2.62 ...  
## $ pred\_minus\_obs\_H\_b3: num 40.4 39.8 49.3 47.9 32 ...  
## $ pred\_minus\_obs\_H\_b4: num 7.97 -16.74 3.25 -6.2 -1.33 ...  
## $ pred\_minus\_obs\_H\_b5: num -32.9 -24.9 -24.9 -21 -38 ...  
## $ pred\_minus\_obs\_H\_b6: num -38.9 -36.3 -30.4 -30.3 -43.6 ...  
## $ pred\_minus\_obs\_H\_b7: num -14.94 -15.67 -3.6 -5.03 -34.25 ...  
## $ pred\_minus\_obs\_H\_b8: num 4.47 8.16 4.15 7.77 1.83 ...  
## $ pred\_minus\_obs\_H\_b9: num -2.36 -2.26 -1.46 2.68 -2.94 ...  
## $ pred\_minus\_obs\_S\_b1: num -18.4 -16.3 -15.9 -13.8 -21.7 ...  
## $ pred\_minus\_obs\_S\_b2: num -1.88 -1.95 -1.79 -2.53 -1.64 -1.89 -0.55 -2.61 -2.09 -1.76 ...  
## $ pred\_minus\_obs\_S\_b3: num -6.43 -6.25 -4.64 -6.34 -4.62 -5.89 -3.89 -8.38 -5.95 -5.05 ...  
## $ pred\_minus\_obs\_S\_b4: num -21 -18.8 -17.7 -22 -23.7 ...  
## $ pred\_minus\_obs\_S\_b5: num -1.6 -1.99 -0.48 -2.34 -0.85 -1.89 0.02 -1.51 -2.13 -0.93 ...  
## $ pred\_minus\_obs\_S\_b6: num -6.18 -6.18 -4.69 -6.6 -5.5 -8.05 -4.2 -6.68 -8.73 -5.6 ...  
## $ pred\_minus\_obs\_S\_b7: num -22.5 -23.4 -20 -27.1 -22.8 ...  
## $ pred\_minus\_obs\_S\_b8: num -5.2 -8.87 -4.1 -7.99 -2.74 -1.94 -0.22 -3.42 -2.42 -3.28 ...  
## $ pred\_minus\_obs\_S\_b9: num -7.86 -10.83 -7.07 -10.81 -5.84 ...

#Table of class  
table(Forest\_train$class)

##   
## d h o s   
## 105 38 46 136

#Table of class  
table(Forest\_test$class)

##   
## d h o s   
## 54 48 37 59

#Proportion of class variable  
round(prop.table(table(Forest\_train$class))\*100,digits = 1)

##   
## d h o s   
## 32.3 11.7 14.2 41.8

#Proportion of class variable   
round(prop.table(table(Forest\_test$class))\*100,digits = 1)

##   
## d h o s   
## 27.3 24.2 18.7 29.8

#Summary of Forest\_train  
summary(Forest\_train[,c(2:27)])

## b1 b2 b3 b4   
## Min. : 31.00 Min. :23.00 Min. : 47.00 Min. : 69.00   
## 1st Qu.: 50.00 1st Qu.:28.00 1st Qu.: 52.00 1st Qu.: 89.00   
## Median : 57.00 Median :32.00 Median : 55.00 Median : 95.00   
## Mean : 58.02 Mean :38.38 Mean : 61.47 Mean : 96.18   
## 3rd Qu.: 65.00 3rd Qu.:43.00 3rd Qu.: 65.00 3rd Qu.:103.00   
## Max. :107.00 Max. :91.00 Max. :124.00 Max. :141.00   
## b5 b6 b7 b8   
## Min. : 43.0 Min. : 83.0 Min. : 42.00 Min. :19.00   
## 1st Qu.: 51.0 1st Qu.: 93.0 1st Qu.: 73.00 1st Qu.:24.00   
## Median : 54.0 Median : 96.0 Median : 85.00 Median :25.00   
## Mean : 58.1 Mean : 99.2 Mean : 85.86 Mean :27.38   
## 3rd Qu.: 63.0 3rd Qu.:103.0 3rd Qu.: 98.00 3rd Qu.:27.00   
## Max. :100.0 Max. :138.0 Max. :136.00 Max. :84.00   
## b9 pred\_minus\_obs\_H\_b1 pred\_minus\_obs\_H\_b2  
## Min. : 45.00 Min. : 4.95 Min. :-42.83   
## 1st Qu.: 54.00 1st Qu.:48.37 1st Qu.: 8.08   
## Median : 57.00 Median :57.56 Median : 18.87   
## Mean : 58.88 Mean :55.79 Mean : 12.64   
## 3rd Qu.: 60.00 3rd Qu.:64.12 3rd Qu.: 23.02   
## Max. :114.00 Max. :86.08 Max. : 29.90   
## pred\_minus\_obs\_H\_b3 pred\_minus\_obs\_H\_b4 pred\_minus\_obs\_H\_b5  
## Min. :-30.43 Min. :-37.710 Min. :-74.56   
## 1st Qu.: 30.12 1st Qu.: -5.060 1st Qu.:-37.15   
## Median : 40.88 Median : 4.080 Median :-28.90   
## Mean : 34.99 Mean : 1.931 Mean :-32.76   
## 3rd Qu.: 45.40 3rd Qu.: 10.250 3rd Qu.:-25.14   
## Max. : 57.55 Max. : 27.190 Max. :-18.40   
## pred\_minus\_obs\_H\_b6 pred\_minus\_obs\_H\_b7 pred\_minus\_obs\_H\_b8  
## Min. :-77.17 Min. :-58.090 Min. :-54.740   
## 1st Qu.:-42.75 1st Qu.:-20.400 1st Qu.: 2.710   
## Median :-36.33 Median : -8.620 Median : 4.440   
## Mean :-38.92 Mean : -9.218 Mean : 2.353   
## 3rd Qu.:-32.35 3rd Qu.: 2.190 3rd Qu.: 5.760   
## Max. :-23.55 Max. : 34.660 Max. : 10.750   
## pred\_minus\_obs\_H\_b9 pred\_minus\_obs\_S\_b1 pred\_minus\_obs\_S\_b2  
## Min. :-58.280 Min. :-26.79 Min. :-5.510   
## 1st Qu.: -4.660 1st Qu.:-22.25 1st Qu.:-1.750   
## Median : -1.250 Median :-19.95 Median :-1.030   
## Mean : -3.341 Mean :-20.00 Mean :-1.086   
## 3rd Qu.: 1.430 3rd Qu.:-18.25 3rd Qu.:-0.390   
## Max. : 9.580 Max. : -7.76 Max. : 1.780   
## pred\_minus\_obs\_S\_b3 pred\_minus\_obs\_S\_b4 pred\_minus\_obs\_S\_b5  
## Min. :-10.120 Min. :-34.63 Min. :-1.8300   
## 1st Qu.: -5.530 1st Qu.:-24.22 1st Qu.:-1.1900   
## Median : -4.490 Median :-21.04 Median :-0.9900   
## Mean : -4.376 Mean :-21.66 Mean :-0.9798   
## 3rd Qu.: -2.770 3rd Qu.:-19.06 3rd Qu.:-0.7800   
## Max. : 1.040 Max. :-12.07 Max. : 0.2600   
## pred\_minus\_obs\_S\_b6 pred\_minus\_obs\_S\_b7 pred\_minus\_obs\_S\_b8  
## Min. :-7.970 Min. :-29.34 Min. :-6.500   
## 1st Qu.:-5.410 1st Qu.:-21.78 1st Qu.:-2.360   
## Median :-4.670 Median :-18.87 Median :-1.650   
## Mean :-4.633 Mean :-19.00 Mean :-1.702   
## 3rd Qu.:-3.900 3rd Qu.:-16.77 3rd Qu.:-1.030   
## Max. :-0.770 Max. : -8.33 Max. : 2.580

#Summary of test dataset  
summary(Forest\_test[,c(2:27)])

## b1 b2 b3 b4   
## Min. : 34.00 Min. : 25.00 Min. : 47.00 Min. : 54.00   
## 1st Qu.: 54.00 1st Qu.: 28.00 1st Qu.: 52.00 1st Qu.: 92.25   
## Median : 60.00 Median : 31.50 Median : 57.00 Median : 99.50   
## Mean : 62.95 Mean : 41.02 Mean : 63.68 Mean :101.41   
## 3rd Qu.: 70.75 3rd Qu.: 50.75 3rd Qu.: 69.00 3rd Qu.:111.75   
## Max. :105.00 Max. :160.00 Max. :196.00 Max. :172.00   
## b5 b6 b7 b8   
## Min. :44.00 Min. : 84.0 Min. : 54.0 Min. :21.00   
## 1st Qu.:49.00 1st Qu.: 92.0 1st Qu.: 80.0 1st Qu.:24.00   
## Median :55.00 Median : 98.0 Median : 91.0 Median :25.00   
## Mean :58.73 Mean :100.7 Mean : 90.6 Mean :28.69   
## 3rd Qu.:65.00 3rd Qu.:107.0 3rd Qu.:101.0 3rd Qu.:27.00   
## Max. :98.00 Max. :136.0 Max. :139.0 Max. :82.00   
## b9 pred\_minus\_obs\_H\_b1 pred\_minus\_obs\_H\_b2  
## Min. : 50.00 Min. : 7.66 Min. :-112.6000   
## 1st Qu.: 55.00 1st Qu.:40.67 1st Qu.: 0.2725   
## Median : 58.00 Median :53.03 Median : 18.8050   
## Mean : 61.12 Mean :50.82 Mean : 9.8083   
## 3rd Qu.: 63.00 3rd Qu.:59.92 3rd Qu.: 22.2575   
## Max. :109.00 Max. :83.32 Max. : 29.7900   
## pred\_minus\_obs\_H\_b3 pred\_minus\_obs\_H\_b4 pred\_minus\_obs\_H\_b5  
## Min. :-106.12 Min. :-77.010 Min. :-73.29   
## 1st Qu.: 27.20 1st Qu.:-15.922 1st Qu.:-39.77   
## Median : 37.61 Median : -2.180 Median :-29.16   
## Mean : 32.54 Mean : -3.899 Mean :-33.42   
## 3rd Qu.: 43.33 3rd Qu.: 6.657 3rd Qu.:-23.89   
## Max. : 55.97 Max. : 40.820 Max. :-19.49   
## pred\_minus\_obs\_H\_b6 pred\_minus\_obs\_H\_b7 pred\_minus\_obs\_H\_b8  
## Min. :-76.09 Min. :-62.740 Min. :-52.000   
## 1st Qu.:-46.16 1st Qu.:-23.585 1st Qu.: 1.978   
## Median :-37.51 Median :-14.835 Median : 4.140   
## Mean :-40.45 Mean :-13.912 Mean : 1.005   
## 3rd Qu.:-32.94 3rd Qu.: -3.248 3rd Qu.: 5.500   
## Max. :-25.68 Max. : 24.330 Max. : 10.830   
## pred\_minus\_obs\_H\_b9 pred\_minus\_obs\_S\_b1 pred\_minus\_obs\_S\_b2  
## Min. :-53.5300 Min. :-32.95 Min. :-8.8000   
## 1st Qu.: -6.6275 1st Qu.:-23.32 1st Qu.:-1.8600   
## Median : -2.2550 Median :-20.02 Median :-0.9700   
## Mean : -5.5941 Mean :-20.04 Mean :-1.0071   
## 3rd Qu.: 0.2475 3rd Qu.:-17.79 3rd Qu.:-0.0425   
## Max. : 5.7400 Max. : 5.13 Max. :12.4600   
## pred\_minus\_obs\_S\_b3 pred\_minus\_obs\_S\_b4 pred\_minus\_obs\_S\_b5  
## Min. :-11.210 Min. :-40.37 Min. :-3.2700   
## 1st Qu.: -5.790 1st Qu.:-24.09 1st Qu.:-1.2900   
## Median : -4.350 Median :-20.46 Median :-0.9450   
## Mean : -4.356 Mean :-21.00 Mean :-0.9737   
## 3rd Qu.: -2.882 3rd Qu.:-17.95 3rd Qu.:-0.6425   
## Max. : 7.370 Max. : 1.88 Max. : 3.4400   
## pred\_minus\_obs\_S\_b6 pred\_minus\_obs\_S\_b7 pred\_minus\_obs\_S\_b8  
## Min. :-8.730 Min. :-34.14 Min. :-8.870   
## 1st Qu.:-5.747 1st Qu.:-22.24 1st Qu.:-2.370   
## Median :-4.540 Median :-19.20 Median :-1.420   
## Mean :-4.598 Mean :-18.84 Mean :-1.571   
## 3rd Qu.:-3.618 3rd Qu.:-16.23 3rd Qu.:-0.655   
## Max. : 3.940 Max. : 3.67 Max. : 8.840

# create normalization function  
normalize <- function(x) {  
 return ((x - min(x)) / (max(x) - min(x)))  
}

Forest\_train\_n=as.data.frame(lapply(Forest\_train[2:27],normalize))  
Forest\_test\_n=as.data.frame(lapply(Forest\_test[2:27],normalize))

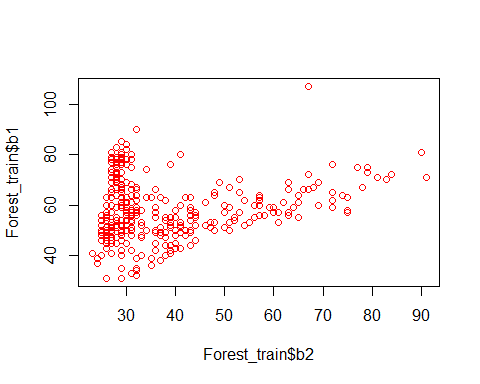
# Summary of test dataset  
summary(Forest\_train\_n$b2)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00000 0.07353 0.13240 0.22620 0.29410 1.00000

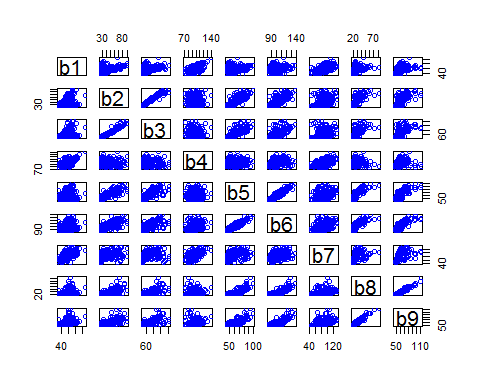
#Creating lables for training and test data  
Forest\_train\_labels=Forest\_train[,1]  
Forest\_test\_labels=Forest\_test[,1]

visualize the data using labels

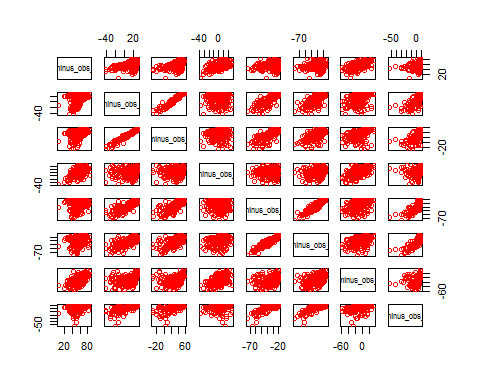
#Sactterplot of bi & b2  
plot(Forest\_train$b1~Forest\_train$b2,col="red")



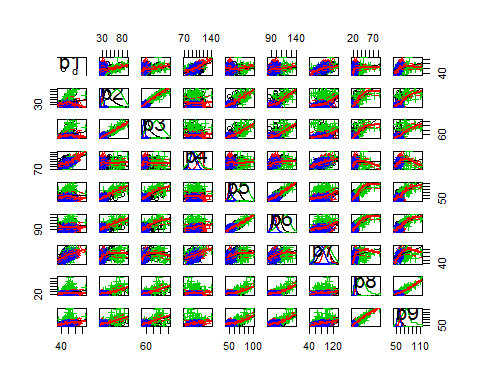
pairs(~b1+b2+b3+b4+b5+b6+b7+b8+b9,data=Forest\_train,col="Blue ")



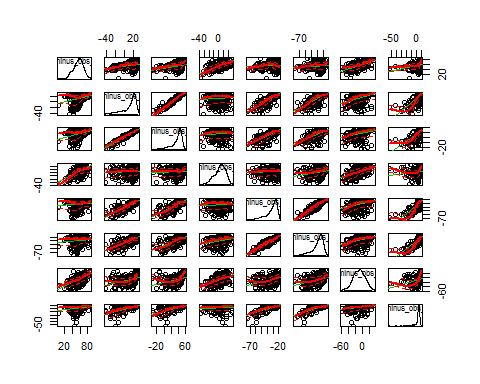
#Sactterplot   
pairs(~pred\_minus\_obs\_H\_b1+pred\_minus\_obs\_H\_b2+pred\_minus\_obs\_H\_b3+pred\_minus\_obs\_H\_b4+pred\_minus\_obs\_H\_b5+pred\_minus\_obs\_H\_b6+pred\_minus\_obs\_H\_b7+pred\_minus\_obs\_H\_b8,data=Forest\_train,col="red")



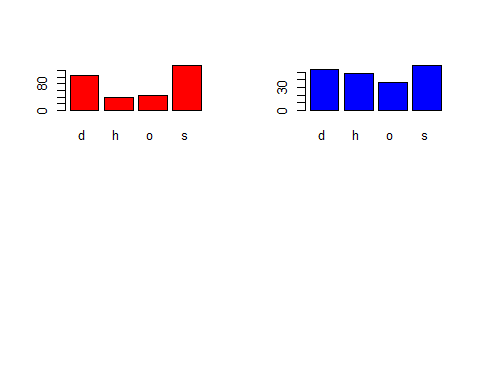
library(car)  
scatterplotMatrix(~b1+b2+b3+b4+b5+b6+b7+b8+b9 | class,data=Forest\_train)



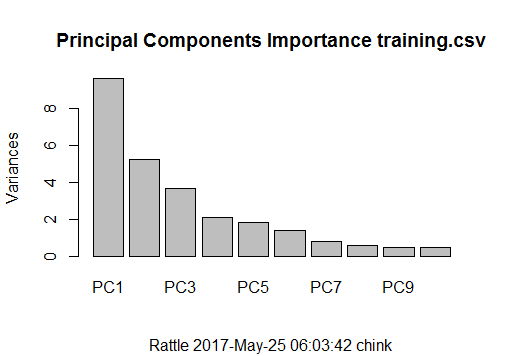
scatterplotMatrix(~pred\_minus\_obs\_H\_b1+pred\_minus\_obs\_H\_b2+pred\_minus\_obs\_H\_b3+pred\_minus\_obs\_H\_b4+pred\_minus\_obs\_H\_b5+pred\_minus\_obs\_H\_b6+pred\_minus\_obs\_H\_b7+pred\_minus\_obs\_H\_b8,data=Forest\_train)

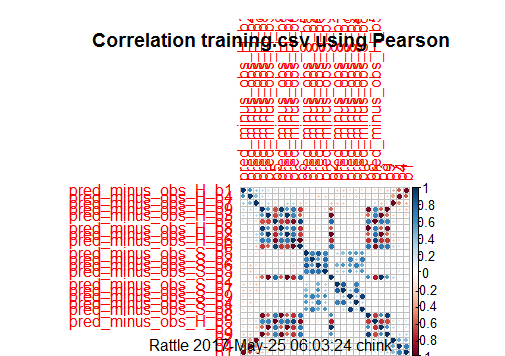


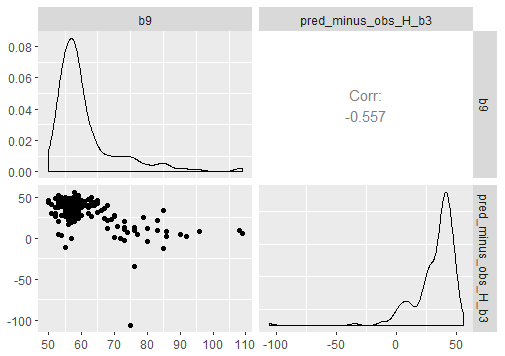
#Boxplot of forest type in training and test data  
par(mfrow=c(2,2))  
plot(Forest\_train$class,col="red")  
plot(Forest\_test$class,col="blue")

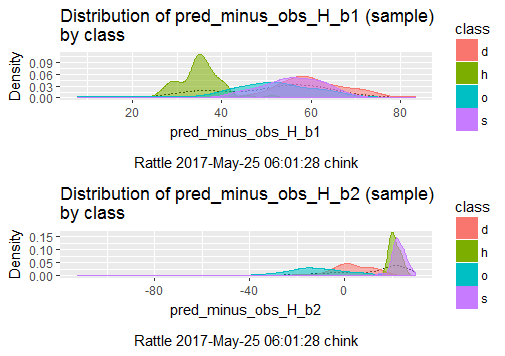


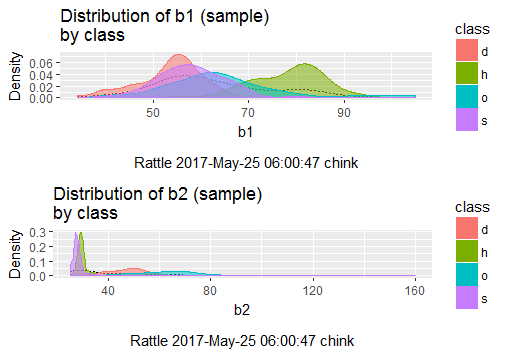
Some graphical representation from the rattle :

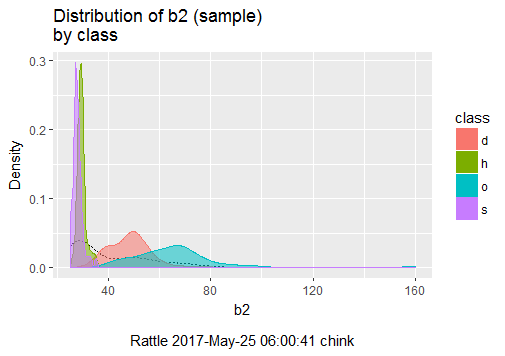


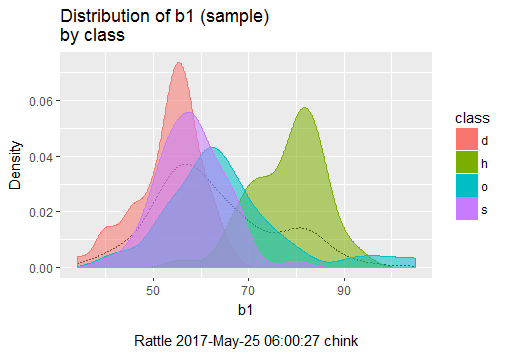












**Step 3: Training a model on the data**

library(class)  
Forest\_test\_pred=knn(train = Forest\_train\_n,test = Forest\_test\_n,cl=Forest\_train\_labels,k=5)  
head(Forest\_test\_n)

## b1 b2 b3 b4 b5 b6  
## 1 0.07042254 0.081481481 0.06711409 0.3135593 0.27777778 0.3269231  
## 2 0.70422535 0.037037037 0.06711409 0.4915254 0.12962963 0.2692308  
## 3 0.26760563 0.000000000 0.01342282 0.3813559 0.12962963 0.1730769  
## 4 0.35211268 0.007407407 0.01342282 0.4152542 0.05555556 0.1538462  
## 5 0.32394366 0.177777778 0.12751678 0.4152542 0.37037037 0.4230769  
## 6 0.71830986 0.022222222 0.06040268 0.5593220 0.14814815 0.2692308  
## b7 b8 b9 pred\_minus\_obs\_H\_b1 pred\_minus\_obs\_H\_b2  
## 1 0.4588235 0.09836066 0.1694915 0.8992863 0.8951471  
## 2 0.4470588 0.08196721 0.2033898 0.3029342 0.9341948  
## 3 0.3529412 0.08196721 0.1355932 0.7340735 0.9782990  
## 4 0.3294118 0.06557377 0.1016949 0.6328311 0.9628485  
## 5 0.7058824 0.11475410 0.1525424 0.6843775 0.8091860  
## 6 0.5529412 0.09836066 0.2542373 0.3632038 0.9553339  
## pred\_minus\_obs\_H\_b3 pred\_minus\_obs\_H\_b4 pred\_minus\_obs\_H\_b5  
## 1 0.9036338 0.7212085 0.7503717  
## 2 0.9004257 0.5114996 0.8990706  
## 3 0.9587266 0.6811508 0.8996283  
## 4 0.9502128 0.6009505 0.9723048  
## 5 0.8522426 0.6422813 0.6561338  
## 6 0.9156024 0.5128575 0.8895911  
## pred\_minus\_obs\_H\_b6 pred\_minus\_obs\_H\_b7 pred\_minus\_obs\_H\_b8  
## 1 0.7373537 0.5489836 0.8987745  
## 2 0.7887324 0.5405995 0.9575044  
## 3 0.9067645 0.6792236 0.8936814  
## 4 0.9087483 0.6628000 0.9512972  
## 5 0.6451101 0.3272080 0.8567563  
## 6 0.8321762 0.5201562 0.8527773  
## pred\_minus\_obs\_H\_b9 pred\_minus\_obs\_S\_b1 pred\_minus\_obs\_S\_b2  
## 1 0.8633373 0.3818277 0.3254939  
## 2 0.8650245 0.4380252 0.3222013  
## 3 0.8785220 0.4472164 0.3297272  
## 4 0.9483719 0.5036765 0.2949200  
## 5 0.8535515 0.2943803 0.3367827  
## 6 0.7297115 0.1777836 0.3250235  
## pred\_minus\_obs\_S\_b3 pred\_minus\_obs\_S\_b4 pred\_minus\_obs\_S\_b5  
## 1 0.2572659 0.4577515 0.2488823  
## 2 0.2669537 0.5107692 0.1907601  
## 3 0.3536060 0.5358580 0.4157973  
## 4 0.2621098 0.4340828 0.1385991  
## 5 0.3546825 0.3936095 0.3606557  
## 6 0.2863294 0.1289941 0.2056632  
## pred\_minus\_obs\_S\_b6 pred\_minus\_obs\_S\_b7 pred\_minus\_obs\_S\_b8  
## 1 0.20126283 0.3078551 0.20722756  
## 2 0.20126283 0.2837874 0.00000000  
## 3 0.31886346 0.3747686 0.26933936  
## 4 0.16811365 0.1861941 0.04968944  
## 5 0.25493291 0.2991272 0.34613213  
## 6 0.05367009 0.1169003 0.39130435

head(Forest\_test\_labels)

## [1] d h s s d h   
## Levels: d h o s

**Step 4: Evaluating model performance**

library(gmodels)  
CrossTable(x=Forest\_test\_labels,y=Forest\_test\_pred,prop.chisq = FALSE)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 198   
##   
##   
## | Forest\_test\_pred   
## Forest\_test\_labels | d | h | o | s | Row Total |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## d | 46 | 0 | 0 | 8 | 54 |   
## | 0.852 | 0.000 | 0.000 | 0.148 | 0.273 |   
## | 0.793 | 0.000 | 0.000 | 0.099 | |   
## | 0.232 | 0.000 | 0.000 | 0.040 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## h | 0 | 34 | 0 | 14 | 48 |   
## | 0.000 | 0.708 | 0.000 | 0.292 | 0.242 |   
## | 0.000 | 0.971 | 0.000 | 0.173 | |   
## | 0.000 | 0.172 | 0.000 | 0.071 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## o | 12 | 0 | 24 | 1 | 37 |   
## | 0.324 | 0.000 | 0.649 | 0.027 | 0.187 |   
## | 0.207 | 0.000 | 1.000 | 0.012 | |   
## | 0.061 | 0.000 | 0.121 | 0.005 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## s | 0 | 1 | 0 | 58 | 59 |   
## | 0.000 | 0.017 | 0.000 | 0.983 | 0.298 |   
## | 0.000 | 0.029 | 0.000 | 0.716 | |   
## | 0.000 | 0.005 | 0.000 | 0.293 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## Column Total | 58 | 35 | 24 | 81 | 198 |   
## | 0.293 | 0.177 | 0.121 | 0.409 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
##   
##

Accuracy=(46+35+25+58)/198\*100=82.83%

**Step 5: Improving model performance**

#Using Z score to improve model  
Forest\_train\_z=as.data.frame(scale(Forest\_train[-1]))  
Forest\_test\_z=as.data.frame(scale(Forest\_test[-1]))

# confirm that the transformation was applied correctly  
summary(Forest\_train\_z$b5)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -1.4170 -0.6662 -0.3847 0.0000 0.4600 3.9330

# re-classify test cases  
Forest\_test\_pred=knn(train = Forest\_train\_z,test=Forest\_test\_z,cl=Forest\_train\_labels,k=5)  
# Create the cross tabulation of predicted vs. actual  
CrossTable(x=Forest\_test\_labels,y=Forest\_test\_pred,prop.chisq = FALSE)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 198   
##   
##   
## | Forest\_test\_pred   
## Forest\_test\_labels | d | h | o | s | Row Total |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## d | 51 | 0 | 1 | 2 | 54 |   
## | 0.944 | 0.000 | 0.019 | 0.037 | 0.273 |   
## | 0.785 | 0.000 | 0.034 | 0.024 | |   
## | 0.258 | 0.000 | 0.005 | 0.010 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## h | 3 | 19 | 0 | 26 | 48 |   
## | 0.062 | 0.396 | 0.000 | 0.542 | 0.242 |   
## | 0.046 | 1.000 | 0.000 | 0.306 | |   
## | 0.015 | 0.096 | 0.000 | 0.131 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## o | 9 | 0 | 28 | 0 | 37 |   
## | 0.243 | 0.000 | 0.757 | 0.000 | 0.187 |   
## | 0.138 | 0.000 | 0.966 | 0.000 | |   
## | 0.045 | 0.000 | 0.141 | 0.000 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## s | 2 | 0 | 0 | 57 | 59 |   
## | 0.034 | 0.000 | 0.000 | 0.966 | 0.298 |   
## | 0.031 | 0.000 | 0.000 | 0.671 | |   
## | 0.010 | 0.000 | 0.000 | 0.288 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## Column Total | 65 | 19 | 29 | 85 | 198 |   
## | 0.328 | 0.096 | 0.146 | 0.429 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
##   
##

Accuracy=(50+19+28+57)/198\*100=77.78% Accuracy is not improving with the use of Z score. I am trying different k values with normalised data.

#KNN model with K=1  
Forest\_test\_pred=knn(train = Forest\_train\_n,test = Forest\_test\_n,cl=Forest\_train\_labels,k=1)  
CrossTable(x=Forest\_test\_labels,y=Forest\_test\_pred,prop.chisq = FALSE)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 198   
##   
##   
## | Forest\_test\_pred   
## Forest\_test\_labels | d | h | o | s | Row Total |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## d | 47 | 0 | 1 | 6 | 54 |   
## | 0.870 | 0.000 | 0.019 | 0.111 | 0.273 |   
## | 0.855 | 0.000 | 0.033 | 0.077 | |   
## | 0.237 | 0.000 | 0.005 | 0.030 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## h | 0 | 33 | 0 | 15 | 48 |   
## | 0.000 | 0.688 | 0.000 | 0.312 | 0.242 |   
## | 0.000 | 0.943 | 0.000 | 0.192 | |   
## | 0.000 | 0.167 | 0.000 | 0.076 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## o | 8 | 0 | 29 | 0 | 37 |   
## | 0.216 | 0.000 | 0.784 | 0.000 | 0.187 |   
## | 0.145 | 0.000 | 0.967 | 0.000 | |   
## | 0.040 | 0.000 | 0.146 | 0.000 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## s | 0 | 2 | 0 | 57 | 59 |   
## | 0.000 | 0.034 | 0.000 | 0.966 | 0.298 |   
## | 0.000 | 0.057 | 0.000 | 0.731 | |   
## | 0.000 | 0.010 | 0.000 | 0.288 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## Column Total | 55 | 35 | 30 | 78 | 198 |   
## | 0.278 | 0.177 | 0.152 | 0.394 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
##   
##

#KNN model with K=3  
Forest\_test\_pred=knn(train = Forest\_train\_n,test = Forest\_test\_n,cl=Forest\_train\_labels,k=3)  
CrossTable(x=Forest\_test\_labels,y=Forest\_test\_pred,prop.chisq = FALSE)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 198   
##   
##   
## | Forest\_test\_pred   
## Forest\_test\_labels | d | h | o | s | Row Total |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## d | 47 | 0 | 1 | 6 | 54 |   
## | 0.870 | 0.000 | 0.019 | 0.111 | 0.273 |   
## | 0.825 | 0.000 | 0.036 | 0.077 | |   
## | 0.237 | 0.000 | 0.005 | 0.030 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## h | 1 | 34 | 0 | 13 | 48 |   
## | 0.021 | 0.708 | 0.000 | 0.271 | 0.242 |   
## | 0.018 | 0.971 | 0.000 | 0.167 | |   
## | 0.005 | 0.172 | 0.000 | 0.066 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## o | 9 | 0 | 27 | 1 | 37 |   
## | 0.243 | 0.000 | 0.730 | 0.027 | 0.187 |   
## | 0.158 | 0.000 | 0.964 | 0.013 | |   
## | 0.045 | 0.000 | 0.136 | 0.005 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## s | 0 | 1 | 0 | 58 | 59 |   
## | 0.000 | 0.017 | 0.000 | 0.983 | 0.298 |   
## | 0.000 | 0.029 | 0.000 | 0.744 | |   
## | 0.000 | 0.005 | 0.000 | 0.293 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## Column Total | 57 | 35 | 28 | 78 | 198 |   
## | 0.288 | 0.177 | 0.141 | 0.394 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
##   
##

#KNN model with K=5  
Forest\_test\_pred=knn(train = Forest\_train\_n,test = Forest\_test\_n,cl=Forest\_train\_labels,k=5)  
CrossTable(x=Forest\_test\_labels,y=Forest\_test\_pred,prop.chisq = FALSE)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 198   
##   
##   
## | Forest\_test\_pred   
## Forest\_test\_labels | d | h | o | s | Row Total |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## d | 46 | 0 | 0 | 8 | 54 |   
## | 0.852 | 0.000 | 0.000 | 0.148 | 0.273 |   
## | 0.807 | 0.000 | 0.000 | 0.099 | |   
## | 0.232 | 0.000 | 0.000 | 0.040 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## h | 0 | 34 | 0 | 14 | 48 |   
## | 0.000 | 0.708 | 0.000 | 0.292 | 0.242 |   
## | 0.000 | 0.971 | 0.000 | 0.173 | |   
## | 0.000 | 0.172 | 0.000 | 0.071 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## o | 11 | 0 | 25 | 1 | 37 |   
## | 0.297 | 0.000 | 0.676 | 0.027 | 0.187 |   
## | 0.193 | 0.000 | 1.000 | 0.012 | |   
## | 0.056 | 0.000 | 0.126 | 0.005 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## s | 0 | 1 | 0 | 58 | 59 |   
## | 0.000 | 0.017 | 0.000 | 0.983 | 0.298 |   
## | 0.000 | 0.029 | 0.000 | 0.716 | |   
## | 0.000 | 0.005 | 0.000 | 0.293 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## Column Total | 57 | 35 | 25 | 81 | 198 |   
## | 0.288 | 0.177 | 0.126 | 0.409 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
##   
##

#KNN model with K=7  
Forest\_test\_pred=knn(train = Forest\_train\_n,test = Forest\_test\_n,cl=Forest\_train\_labels,k=7)  
CrossTable(x=Forest\_test\_labels,y=Forest\_test\_pred,prop.chisq = FALSE)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 198   
##   
##   
## | Forest\_test\_pred   
## Forest\_test\_labels | d | h | o | s | Row Total |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## d | 48 | 0 | 1 | 5 | 54 |   
## | 0.889 | 0.000 | 0.019 | 0.093 | 0.273 |   
## | 0.842 | 0.000 | 0.036 | 0.060 | |   
## | 0.242 | 0.000 | 0.005 | 0.025 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## h | 0 | 28 | 0 | 20 | 48 |   
## | 0.000 | 0.583 | 0.000 | 0.417 | 0.242 |   
## | 0.000 | 0.966 | 0.000 | 0.238 | |   
## | 0.000 | 0.141 | 0.000 | 0.101 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## o | 9 | 0 | 27 | 1 | 37 |   
## | 0.243 | 0.000 | 0.730 | 0.027 | 0.187 |   
## | 0.158 | 0.000 | 0.964 | 0.012 | |   
## | 0.045 | 0.000 | 0.136 | 0.005 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## s | 0 | 1 | 0 | 58 | 59 |   
## | 0.000 | 0.017 | 0.000 | 0.983 | 0.298 |   
## | 0.000 | 0.034 | 0.000 | 0.690 | |   
## | 0.000 | 0.005 | 0.000 | 0.293 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## Column Total | 57 | 29 | 28 | 84 | 198 |   
## | 0.288 | 0.146 | 0.141 | 0.424 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
##   
##

#KNN model with K=9  
Forest\_test\_pred=knn(train = Forest\_train\_n,test = Forest\_test\_n,cl=Forest\_train\_labels,k=9)  
CrossTable(x=Forest\_test\_labels,y=Forest\_test\_pred,prop.chisq = FALSE)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 198   
##   
##   
## | Forest\_test\_pred   
## Forest\_test\_labels | d | h | o | s | Row Total |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## d | 50 | 0 | 0 | 4 | 54 |   
## | 0.926 | 0.000 | 0.000 | 0.074 | 0.273 |   
## | 0.806 | 0.000 | 0.000 | 0.049 | |   
## | 0.253 | 0.000 | 0.000 | 0.020 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## h | 0 | 29 | 0 | 19 | 48 |   
## | 0.000 | 0.604 | 0.000 | 0.396 | 0.242 |   
## | 0.000 | 0.967 | 0.000 | 0.232 | |   
## | 0.000 | 0.146 | 0.000 | 0.096 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## o | 12 | 0 | 24 | 1 | 37 |   
## | 0.324 | 0.000 | 0.649 | 0.027 | 0.187 |   
## | 0.194 | 0.000 | 1.000 | 0.012 | |   
## | 0.061 | 0.000 | 0.121 | 0.005 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## s | 0 | 1 | 0 | 58 | 59 |   
## | 0.000 | 0.017 | 0.000 | 0.983 | 0.298 |   
## | 0.000 | 0.033 | 0.000 | 0.707 | |   
## | 0.000 | 0.005 | 0.000 | 0.293 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## Column Total | 62 | 30 | 24 | 82 | 198 |   
## | 0.313 | 0.152 | 0.121 | 0.414 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
##   
##

#KNN model with K=11  
Forest\_test\_pred=knn(train = Forest\_train\_n,test = Forest\_test\_n,cl=Forest\_train\_labels,k=17)  
CrossTable(x=Forest\_test\_labels,y=Forest\_test\_pred,prop.chisq = FALSE)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 198   
##   
##   
## | Forest\_test\_pred   
## Forest\_test\_labels | d | h | o | s | Row Total |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## d | 45 | 0 | 0 | 9 | 54 |   
## | 0.833 | 0.000 | 0.000 | 0.167 | 0.273 |   
## | 0.738 | 0.000 | 0.000 | 0.103 | |   
## | 0.227 | 0.000 | 0.000 | 0.045 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## h | 0 | 29 | 0 | 19 | 48 |   
## | 0.000 | 0.604 | 0.000 | 0.396 | 0.242 |   
## | 0.000 | 0.967 | 0.000 | 0.218 | |   
## | 0.000 | 0.146 | 0.000 | 0.096 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## o | 16 | 0 | 20 | 1 | 37 |   
## | 0.432 | 0.000 | 0.541 | 0.027 | 0.187 |   
## | 0.262 | 0.000 | 1.000 | 0.011 | |   
## | 0.081 | 0.000 | 0.101 | 0.005 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## s | 0 | 1 | 0 | 58 | 59 |   
## | 0.000 | 0.017 | 0.000 | 0.983 | 0.298 |   
## | 0.000 | 0.033 | 0.000 | 0.667 | |   
## | 0.000 | 0.005 | 0.000 | 0.293 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## Column Total | 61 | 30 | 20 | 87 | 198 |   
## | 0.308 | 0.152 | 0.101 | 0.439 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
##   
##

#KNN model with K=35  
Forest\_test\_pred=knn(train = Forest\_train\_n,test = Forest\_test\_n,cl=Forest\_train\_labels,k=35)  
CrossTable(x=Forest\_test\_labels,y=Forest\_test\_pred,prop.chisq = FALSE)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 198   
##   
##   
## | Forest\_test\_pred   
## Forest\_test\_labels | d | h | o | s | Row Total |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## d | 41 | 0 | 0 | 13 | 54 |   
## | 0.759 | 0.000 | 0.000 | 0.241 | 0.273 |   
## | 0.651 | 0.000 | 0.000 | 0.121 | |   
## | 0.207 | 0.000 | 0.000 | 0.066 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## h | 0 | 14 | 0 | 34 | 48 |   
## | 0.000 | 0.292 | 0.000 | 0.708 | 0.242 |   
## | 0.000 | 1.000 | 0.000 | 0.318 | |   
## | 0.000 | 0.071 | 0.000 | 0.172 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## o | 22 | 0 | 14 | 1 | 37 |   
## | 0.595 | 0.000 | 0.378 | 0.027 | 0.187 |   
## | 0.349 | 0.000 | 1.000 | 0.009 | |   
## | 0.111 | 0.000 | 0.071 | 0.005 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## s | 0 | 0 | 0 | 59 | 59 |   
## | 0.000 | 0.000 | 0.000 | 1.000 | 0.298 |   
## | 0.000 | 0.000 | 0.000 | 0.551 | |   
## | 0.000 | 0.000 | 0.000 | 0.298 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
## Column Total | 63 | 14 | 14 | 107 | 198 |   
## | 0.318 | 0.071 | 0.071 | 0.540 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|  
##   
##

Performing knn for different KNN values. It is clear that the k value of 3 and 1 has more accuracy.

Tabular representation:

|  |  |  |  |
| --- | --- | --- | --- |
| **K-Value** | **False Negative** | **False Positive** | **Accuracy** |
| **K=1** | 22 | 10 | 83.84% |
| **K=3** | 20 | 11 | 84.34% |
| **K=5** | 22 | 13 | 82.32% |
| **K=7** | 25 | 10 | 81.82% |
| **K=9** | 23 | 13 | 81.82% |
| **K=17** | 28 | 17 | 76.77% |
| **K=35** | 48 | 23 | 64.14% |

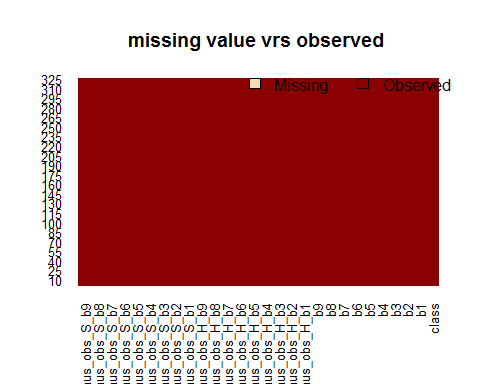
**Now I am going to run for model improvement C5.0 Algorithm.**

#Missing values  
library(Amelia)

## Loading required package: Rcpp

## ##   
## ## Amelia II: Multiple Imputation  
## ## (Version 1.7.4, built: 2015-12-05)  
## ## Copyright (C) 2005-2017 James Honaker, Gary King and Matthew Blackwell  
## ## Refer to http://gking.harvard.edu/amelia/ for more information  
## ##

missmap(Forest\_train, main="missing value vrs observed ")



Data does not have missing values.

# build the simplest decision tree

library(C50)  
Forest\_tree=C5.0(Forest\_train[-1],Forest\_train$class)  
# display simple facts about the tree  
Forest\_tree

##   
## Call:  
## C5.0.default(x = Forest\_train[-1], y = Forest\_train$class)  
##   
## Classification Tree  
## Number of samples: 325   
## Number of predictors: 27   
##   
## Tree size: 18   
##   
## Non-standard options: attempt to group attributes

# display detailed information about the tree  
summary(Forest\_tree)

##   
## Call:  
## C5.0.default(x = Forest\_train[-1], y = Forest\_train$class)  
##   
##   
## C5.0 [Release 2.07 GPL Edition] Thu May 25 05:39:07 2017  
## -------------------------------  
##   
## Class specified by attribute `outcome'  
##   
## Read 325 cases (28 attributes) from undefined.data  
##   
## Decision tree:  
##   
## pred\_minus\_obs\_H\_b9 <= -11.93: o (35/2)  
## pred\_minus\_obs\_H\_b9 > -11.93:  
## :...b2 <= 32:  
## :...b4 > 109: h (29/2)  
## : b4 <= 109:  
## : :...b1 <= 42:  
## : :...b5 <= 51: s (5/1)  
## : : b5 > 51: d (10/1)  
## : b1 > 42:  
## : :...b1 <= 69: s (119/7)  
## : b1 > 69:  
## : :...pred\_minus\_obs\_H\_b3 <= 43.52: h (5)  
## : pred\_minus\_obs\_H\_b3 > 43.52:  
## : :...pred\_minus\_obs\_S\_b4 <= -22.93: h (2)  
## : pred\_minus\_obs\_S\_b4 > -22.93: s (7)  
## b2 > 32:  
## :...pred\_minus\_obs\_H\_b5 > -27.45:  
## :...b9 <= 52: d (2)  
## : b9 > 52: s (7)  
## pred\_minus\_obs\_H\_b5 <= -27.45:  
## :...pred\_minus\_obs\_H\_b2 > 17.71:  
## :...b1 <= 50: d (2)  
## : b1 > 50: s (3)  
## pred\_minus\_obs\_H\_b2 <= 17.71:  
## :...b7 <= 68:  
## :...b2 <= 35: d (2)  
## : b2 > 35: o (4)  
## b7 > 68:  
## :...pred\_minus\_obs\_S\_b8 > -2.61: d (77/4)  
## pred\_minus\_obs\_S\_b8 <= -2.61:  
## :...pred\_minus\_obs\_S\_b3 > -3.62: o (2)  
## pred\_minus\_obs\_S\_b3 <= -3.62:  
## :...b6 <= 111: d (12/1)  
## b6 > 111: o (2)  
##   
##   
## Evaluation on training data (325 cases):  
##   
## Decision Tree   
## ----------------   
## Size Errors   
##   
## 18 18( 5.5%) <<  
##   
##   
## (a) (b) (c) (d) <-classified as  
## ---- ---- ---- ----  
## 99 1 2 3 (a): class d  
## 34 4 (b): class h  
## 4 41 1 (c): class o  
## 2 1 133 (d): class s  
##   
##   
## Attribute usage:  
##   
## 100.00% pred\_minus\_obs\_H\_b9  
## 89.23% b2  
## 54.46% b4  
## 47.08% b1  
## 34.77% pred\_minus\_obs\_H\_b5  
## 32.00% pred\_minus\_obs\_H\_b2  
## 30.46% b7  
## 28.62% pred\_minus\_obs\_S\_b8  
## 4.92% pred\_minus\_obs\_S\_b3  
## 4.62% b5  
## 4.31% b6  
## 4.31% pred\_minus\_obs\_H\_b3  
## 2.77% b9  
## 2.77% pred\_minus\_obs\_S\_b4  
##   
##   
## Time: 0.1 secs

# create a factor vector of predictions on test data  
Forest\_pred\_tree=predict(Forest\_tree,Forest\_test)  
library(gmodels)  
CrossTable(Forest\_test$class,Forest\_pred\_tree,prop.chisq = FALSE, prop.c = FALSE, prop.r = FALSE)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 198   
##   
##   
## | Forest\_pred\_tree   
## Forest\_test$class | d | h | o | s | Row Total |   
## ------------------|-----------|-----------|-----------|-----------|-----------|  
## d | 52 | 0 | 2 | 0 | 54 |   
## | 0.263 | 0.000 | 0.010 | 0.000 | |   
## ------------------|-----------|-----------|-----------|-----------|-----------|  
## h | 0 | 39 | 0 | 9 | 48 |   
## | 0.000 | 0.197 | 0.000 | 0.045 | |   
## ------------------|-----------|-----------|-----------|-----------|-----------|  
## o | 4 | 0 | 33 | 0 | 37 |   
## | 0.020 | 0.000 | 0.167 | 0.000 | |   
## ------------------|-----------|-----------|-----------|-----------|-----------|  
## s | 1 | 2 | 0 | 56 | 59 |   
## | 0.005 | 0.010 | 0.000 | 0.283 | |   
## ------------------|-----------|-----------|-----------|-----------|-----------|  
## Column Total | 57 | 41 | 35 | 65 | 198 |   
## ------------------|-----------|-----------|-----------|-----------|-----------|  
##   
##

Accuracy 90.9%.

**Step 5: Improving model performance**

## Boosting the accuracy of decision trees  
# boosted decision tree with 10 trials  
Forest\_boost10=C5.0(Forest\_train[-1],Forest\_train$class,trails=10)  
Forest\_boost10

##   
## Call:  
## C5.0.default(x = Forest\_train[-1], y = Forest\_train$class, trails = 10)  
##   
## Classification Tree  
## Number of samples: 325   
## Number of predictors: 27   
##   
## Tree size: 18   
##   
## Non-standard options: attempt to group attributes

#Getting summary of Forest\_boost10  
summary(Forest\_boost10)

##   
## Call:  
## C5.0.default(x = Forest\_train[-1], y = Forest\_train$class, trails = 10)  
##   
##   
## C5.0 [Release 2.07 GPL Edition] Thu May 25 05:39:08 2017  
## -------------------------------  
##   
## Class specified by attribute `outcome'  
##   
## Read 325 cases (28 attributes) from undefined.data  
##   
## Decision tree:  
##   
## pred\_minus\_obs\_H\_b9 <= -11.93: o (35/2)  
## pred\_minus\_obs\_H\_b9 > -11.93:  
## :...b2 <= 32:  
## :...b4 > 109: h (29/2)  
## : b4 <= 109:  
## : :...b1 <= 42:  
## : :...b5 <= 51: s (5/1)  
## : : b5 > 51: d (10/1)  
## : b1 > 42:  
## : :...b1 <= 69: s (119/7)  
## : b1 > 69:  
## : :...pred\_minus\_obs\_H\_b3 <= 43.52: h (5)  
## : pred\_minus\_obs\_H\_b3 > 43.52:  
## : :...pred\_minus\_obs\_S\_b4 <= -22.93: h (2)  
## : pred\_minus\_obs\_S\_b4 > -22.93: s (7)  
## b2 > 32:  
## :...pred\_minus\_obs\_H\_b5 > -27.45:  
## :...b9 <= 52: d (2)  
## : b9 > 52: s (7)  
## pred\_minus\_obs\_H\_b5 <= -27.45:  
## :...pred\_minus\_obs\_H\_b2 > 17.71:  
## :...b1 <= 50: d (2)  
## : b1 > 50: s (3)  
## pred\_minus\_obs\_H\_b2 <= 17.71:  
## :...b7 <= 68:  
## :...b2 <= 35: d (2)  
## : b2 > 35: o (4)  
## b7 > 68:  
## :...pred\_minus\_obs\_S\_b8 > -2.61: d (77/4)  
## pred\_minus\_obs\_S\_b8 <= -2.61:  
## :...pred\_minus\_obs\_S\_b3 > -3.62: o (2)  
## pred\_minus\_obs\_S\_b3 <= -3.62:  
## :...b6 <= 111: d (12/1)  
## b6 > 111: o (2)  
##   
##   
## Evaluation on training data (325 cases):  
##   
## Decision Tree   
## ----------------   
## Size Errors   
##   
## 18 18( 5.5%) <<  
##   
##   
## (a) (b) (c) (d) <-classified as  
## ---- ---- ---- ----  
## 99 1 2 3 (a): class d  
## 34 4 (b): class h  
## 4 41 1 (c): class o  
## 2 1 133 (d): class s  
##   
##   
## Attribute usage:  
##   
## 100.00% pred\_minus\_obs\_H\_b9  
## 89.23% b2  
## 54.46% b4  
## 47.08% b1  
## 34.77% pred\_minus\_obs\_H\_b5  
## 32.00% pred\_minus\_obs\_H\_b2  
## 30.46% b7  
## 28.62% pred\_minus\_obs\_S\_b8  
## 4.92% pred\_minus\_obs\_S\_b3  
## 4.62% b5  
## 4.31% b6  
## 4.31% pred\_minus\_obs\_H\_b3  
## 2.77% b9  
## 2.77% pred\_minus\_obs\_S\_b4  
##   
##   
## Time: 0.0 secs

Forest\_boost10\_pred=predict(Forest\_boost10,Forest\_test)  
CrossTable(Forest\_test$class,Forest\_boost10\_pred,prop.chisq = FALSE, prop.c = FALSE, prop.r = FALSE)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 198   
##   
##   
## | Forest\_boost10\_pred   
## Forest\_test$class | d | h | o | s | Row Total |   
## ------------------|-----------|-----------|-----------|-----------|-----------|  
## d | 52 | 0 | 2 | 0 | 54 |   
## | 0.263 | 0.000 | 0.010 | 0.000 | |   
## ------------------|-----------|-----------|-----------|-----------|-----------|  
## h | 0 | 39 | 0 | 9 | 48 |   
## | 0.000 | 0.197 | 0.000 | 0.045 | |   
## ------------------|-----------|-----------|-----------|-----------|-----------|  
## o | 4 | 0 | 33 | 0 | 37 |   
## | 0.020 | 0.000 | 0.167 | 0.000 | |   
## ------------------|-----------|-----------|-----------|-----------|-----------|  
## s | 1 | 2 | 0 | 56 | 59 |   
## | 0.005 | 0.010 | 0.000 | 0.283 | |   
## ------------------|-----------|-----------|-----------|-----------|-----------|  
## Column Total | 57 | 41 | 35 | 65 | 198 |   
## ------------------|-----------|-----------|-----------|-----------|-----------|  
##   
##

Accuracy after applying boosting is also 90.9%.

## Boosting the accuracy of decision trees  
# boosted decision tree with 10 trials  
Forest\_boost20=C5.0(Forest\_train[-1],Forest\_train$class,trails=20)  
Forest\_boost20

##   
## Call:  
## C5.0.default(x = Forest\_train[-1], y = Forest\_train$class, trails = 20)  
##   
## Classification Tree  
## Number of samples: 325   
## Number of predictors: 27   
##   
## Tree size: 18   
##   
## Non-standard options: attempt to group attributes

summary(Forest\_boost20)

##   
## Call:  
## C5.0.default(x = Forest\_train[-1], y = Forest\_train$class, trails = 20)  
##   
##   
## C5.0 [Release 2.07 GPL Edition] Thu May 25 05:39:08 2017  
## -------------------------------  
##   
## Class specified by attribute `outcome'  
##   
## Read 325 cases (28 attributes) from undefined.data  
##   
## Decision tree:  
##   
## pred\_minus\_obs\_H\_b9 <= -11.93: o (35/2)  
## pred\_minus\_obs\_H\_b9 > -11.93:  
## :...b2 <= 32:  
## :...b4 > 109: h (29/2)  
## : b4 <= 109:  
## : :...b1 <= 42:  
## : :...b5 <= 51: s (5/1)  
## : : b5 > 51: d (10/1)  
## : b1 > 42:  
## : :...b1 <= 69: s (119/7)  
## : b1 > 69:  
## : :...pred\_minus\_obs\_H\_b3 <= 43.52: h (5)  
## : pred\_minus\_obs\_H\_b3 > 43.52:  
## : :...pred\_minus\_obs\_S\_b4 <= -22.93: h (2)  
## : pred\_minus\_obs\_S\_b4 > -22.93: s (7)  
## b2 > 32:  
## :...pred\_minus\_obs\_H\_b5 > -27.45:  
## :...b9 <= 52: d (2)  
## : b9 > 52: s (7)  
## pred\_minus\_obs\_H\_b5 <= -27.45:  
## :...pred\_minus\_obs\_H\_b2 > 17.71:  
## :...b1 <= 50: d (2)  
## : b1 > 50: s (3)  
## pred\_minus\_obs\_H\_b2 <= 17.71:  
## :...b7 <= 68:  
## :...b2 <= 35: d (2)  
## : b2 > 35: o (4)  
## b7 > 68:  
## :...pred\_minus\_obs\_S\_b8 > -2.61: d (77/4)  
## pred\_minus\_obs\_S\_b8 <= -2.61:  
## :...pred\_minus\_obs\_S\_b3 > -3.62: o (2)  
## pred\_minus\_obs\_S\_b3 <= -3.62:  
## :...b6 <= 111: d (12/1)  
## b6 > 111: o (2)  
##   
##   
## Evaluation on training data (325 cases):  
##   
## Decision Tree   
## ----------------   
## Size Errors   
##   
## 18 18( 5.5%) <<  
##   
##   
## (a) (b) (c) (d) <-classified as  
## ---- ---- ---- ----  
## 99 1 2 3 (a): class d  
## 34 4 (b): class h  
## 4 41 1 (c): class o  
## 2 1 133 (d): class s  
##   
##   
## Attribute usage:  
##   
## 100.00% pred\_minus\_obs\_H\_b9  
## 89.23% b2  
## 54.46% b4  
## 47.08% b1  
## 34.77% pred\_minus\_obs\_H\_b5  
## 32.00% pred\_minus\_obs\_H\_b2  
## 30.46% b7  
## 28.62% pred\_minus\_obs\_S\_b8  
## 4.92% pred\_minus\_obs\_S\_b3  
## 4.62% b5  
## 4.31% b6  
## 4.31% pred\_minus\_obs\_H\_b3  
## 2.77% b9  
## 2.77% pred\_minus\_obs\_S\_b4  
##   
##   
## Time: 0.0 secs

Forest\_boost20\_pred=predict(Forest\_boost20,Forest\_test)  
CrossTable(Forest\_test$class,Forest\_boost20\_pred,prop.chisq = FALSE, prop.c = FALSE, prop.r = FALSE)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 198   
##   
##   
## | Forest\_boost20\_pred   
## Forest\_test$class | d | h | o | s | Row Total |   
## ------------------|-----------|-----------|-----------|-----------|-----------|  
## d | 52 | 0 | 2 | 0 | 54 |   
## | 0.263 | 0.000 | 0.010 | 0.000 | |   
## ------------------|-----------|-----------|-----------|-----------|-----------|  
## h | 0 | 39 | 0 | 9 | 48 |   
## | 0.000 | 0.197 | 0.000 | 0.045 | |   
## ------------------|-----------|-----------|-----------|-----------|-----------|  
## o | 4 | 0 | 33 | 0 | 37 |   
## | 0.020 | 0.000 | 0.167 | 0.000 | |   
## ------------------|-----------|-----------|-----------|-----------|-----------|  
## s | 1 | 2 | 0 | 56 | 59 |   
## | 0.005 | 0.010 | 0.000 | 0.283 | |   
## ------------------|-----------|-----------|-----------|-----------|-----------|  
## Column Total | 57 | 41 | 35 | 65 | 198 |   
## ------------------|-----------|-----------|-----------|-----------|-----------|  
##   
##

Same level of accuracy in 20 boosting also because dataset is not very big.

**Random Forest.**

library(caret)

## Loading required package: lattice

## Loading required package: ggplot2

library(lattice)  
library(ggplot2)  
library(randomForest)

## randomForest 4.6-12

## Type rfNews() to see new features/changes/bug fixes.

##   
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':  
##   
## margin

#Set.seed for getting the same output  
set.seed(300)  
#Creating random forest  
randomforest\_forest <- randomForest(class ~ ., data = Forest\_train)  
randomforest\_forest

##   
## Call:  
## randomForest(formula = class ~ ., data = Forest\_train)   
## Type of random forest: classification  
## Number of trees: 500  
## No. of variables tried at each split: 5  
##   
## OOB estimate of error rate: 16%  
## Confusion matrix:  
## d h o s class.error  
## d 87 1 3 14 0.17142857  
## h 0 27 0 11 0.28947368  
## o 12 0 33 1 0.28260870  
## s 5 5 0 126 0.07352941

The randomForest() function creates an ensemble of 500 trees & 5 variable at each split.Estimated error rate is 16%

**Evaluating model performance:**

library(caret)  
ctrl <- trainControl(method = "repeatedcv",  
 number = 10, repeats = 10)  
grid\_rf <- expand.grid(.mtry = c(5, 10, 15, 20, 25))  
set.seed(300)  
m\_rf <- train(class ~ ., data = Forest\_train, method = "rf",  
 metric = "Kappa", trControl = ctrl,  
 tuneGrid = grid\_rf)  
m\_rf

## Random Forest   
##   
## 325 samples  
## 27 predictor  
## 4 classes: 'd ', 'h ', 'o ', 's '   
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold, repeated 10 times)   
## Summary of sample sizes: 294, 292, 293, 292, 291, 292, ...   
## Resampling results across tuning parameters:  
## mtry Accuracy Kappa   
## 5 0.8464606 0.7715494  
## 10 0.8424794 0.7659475  
## 15 0.8387643 0.7605602  
## 20 0.8360075 0.7567447  
## 25 0.8348434 0.7550194  
##   
## Kappa was used to select the optimal model using the largest value.  
## The final value used for the model was mtry = 5.

I am using 10 fold cross validation repeated 10 times. Mtry is the tuning parameter in the feature. As in the dataset we have 28 features, I am using 5 tuning parameters.

# auto-tune a boosted C5.0 decision tree  
grid\_c50 <- expand.grid(.model = "tree",  
 .trials = c(10, 20, 30, 40),  
 .winnow = "FALSE")  
set.seed(300)  
library(C50)  
m\_c50 <- train(class ~ ., data = Forest\_train, method = "C5.0",  
 metric = "Kappa", trControl = ctrl,  
 tuneGrid = grid\_c50)

## Loading required package: plyr

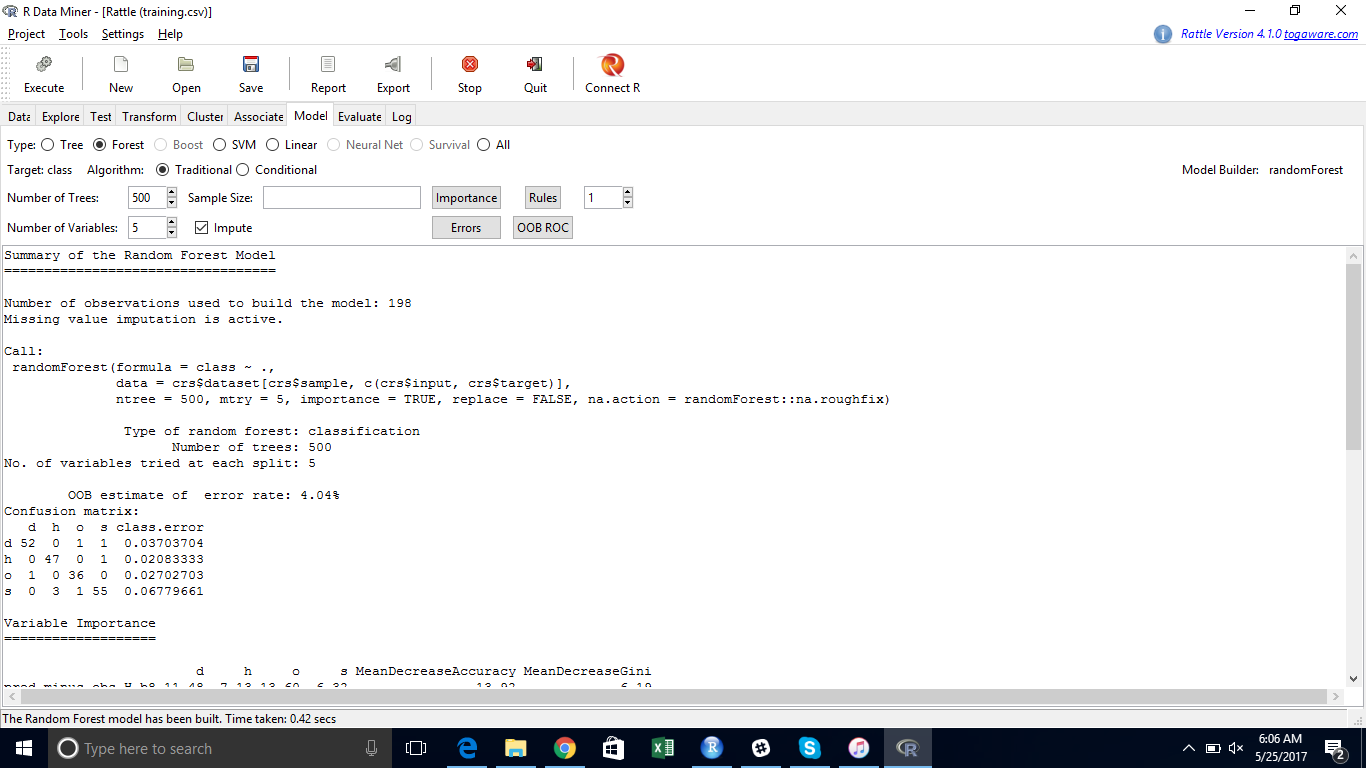
## Warning in Ops.factor(x$winnow): '!' not meaningful for factors

m\_c50

## C5.0   
##   
## 325 samples  
## 27 predictor  
## 4 classes: 'd ', 'h ', 'o ', 's '   
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold, repeated 10 times)   
## Summary of sample sizes: 294, 292, 293, 292, 291, 292, ...   
## Resampling results across tuning parameters:  
##   
## trials Accuracy Kappa   
## 10 0.8450956 0.7706469  
## 20 0.8475281 0.7741926  
## 30 0.8487864 0.7758831  
## 40 0.8475732 0.7740314  
##   
## Tuning parameter 'model' was held constant at a value of tree  
##   
## Tuning parameter 'winnow' was held constant at a value of FALSE  
## Kappa was used to select the optimal model using the largest value.  
## The final values used for the model were trials = 30, model = tree  
## and winnow = FALSE.

The best C5.0 decision tree, which had a kappa of about 0.7758831 for the trail 30 is good among all models in the random forest setup.

I tried to do random forest in the rattle, I used only my training data to perform random forest. I divided training data in 70/30 ration and performed random forest.



**Summary**

The dataset is already divided into 2 training & test data. I initially started from KNN algorithm & tried to improve model accuracy with the help of z-score & from different k values. Then I ran C5.0 classification algorithm and random forest for improving the accuracy of the model. I am getting accuracy 90.9% from C5.0 algorithm for 20 boosting which is good.C5.0 algorithm is working good for this dataset.