# Implementing Random Forest

**Filename:** RF.R

## Functions used:

* computeRss –
  + computes RSS of two given samples
  + input -> sample dataset A, sample dataset B, column name
  + output -> sum of residual squares
  + formula -> sum ( ( column value – average column value ) ^ 2 )
* optimalSplit –
  + splits given dataset on given attributes and computes the optimum based on least value of RSS
  + input -> attribute list, sample dataset
  + output -> optimum attribute, split and min RSS
* generate –
  + generates binary tree based on dataset, indices passed and depth
* traverse –
  + passes test/ train sample through given tree and calculates prediction based on the same
  + input -> data row, binary tree
  + output -> predicted value

## Algorithm:

Split dataset into train & test sets

Create bootstrapped replicas of training set

Generate binary trees for each BTS (bootstrapped training set) For each train sample:

Create predicted labels by traversing tree Add to matrix

For each test sample:

Create predicted labels by traversing tree Add to matrix

For each train/ test sample:

Create average by calculating mean of all predictions from all trees Calculate Train/ Test MSE using the formula sum((y – y’)^2)/no of rows

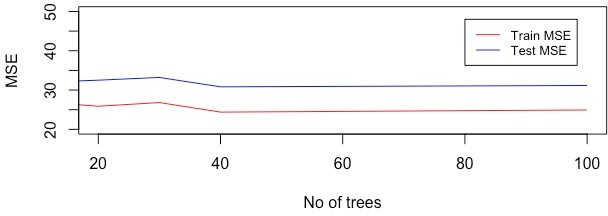
## Observations:

1. Graphs:

a. MSE vs B (10, 20, 30, 40, 100)

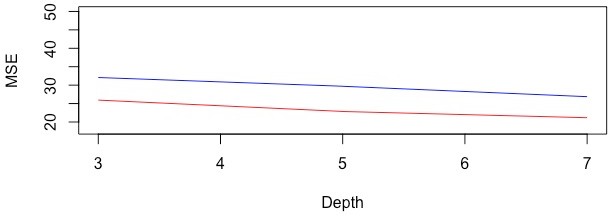
Upon observing the values of train and test MSE against number of decision trees we see that while the Train MSE ranges between 20 and 25, the Test MSE ranges between 30

and 35. The trend indicates a dip in error rates after a B value of 30. For binary trees over 40 the error rates tend to flatten out. This shows us that the optimal value for calculating the MSE is B = 50. The value of MSE decreases gradually showing us that upon increasing the number of bootstrapped samples and decision trees the predictions of the labels tend to be more accurate.



b. MSE vs h (3,5,7)

Upon observing the values of Train and Test MSE for h in [3, 5, 7] we see that upon increasing the depth of our decision trees we create more accurate predictions as the number of filters on our data increases along with the number of predictions. The MSE for the test set ranges from 25 to 35 whereas the MSE for the train set ranges from 15 to 25



1. Train & Test MSE for B = 100, h = 3
   1. Train MSE: 25.32
   2. Test MSE: 31.60