MODEL TO PREDICT WINE QUALITY

By Deepak Kumar Singh

INDEX

TOPIC	PAGE NO.
Aim and abstract	02
Understanding of data	02-03
Cleaning of data	03
Visualization	04-07
Machine learning Model	08-09
Conclusion	10

Aim of the project :

The aim of the project is to reduce man power whom wine company hire to taste the quality of wine before launching into the market. They are spending huge amount on hiring healthy volunteers and on their retention policies/strategies.

.

Abstract :

I have developed classification random forest model separately for red and white wine to predict the quality of wine and achieved 94.45 & 93.50 percent accuracy of test data respectively.

Steps involved to make classification model:

1 .understanding of data

2. Cleaning of data

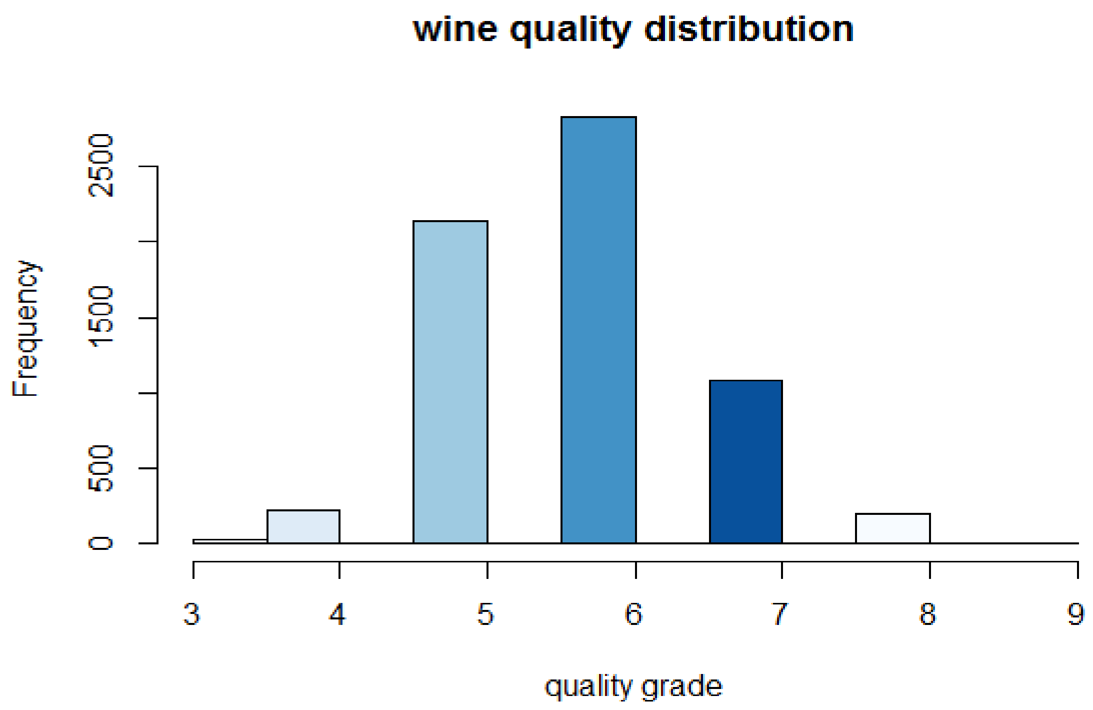
3. visualisation for depth understanding of data

4.Different machine learning approach

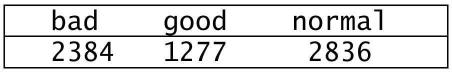
1.UNDERSTANDING OF DATA

I have two sets of data ,one for red wine with dimension (1599,13) and other for white wine with dimension(4898,13). Finally ,I have merged the data for further process.

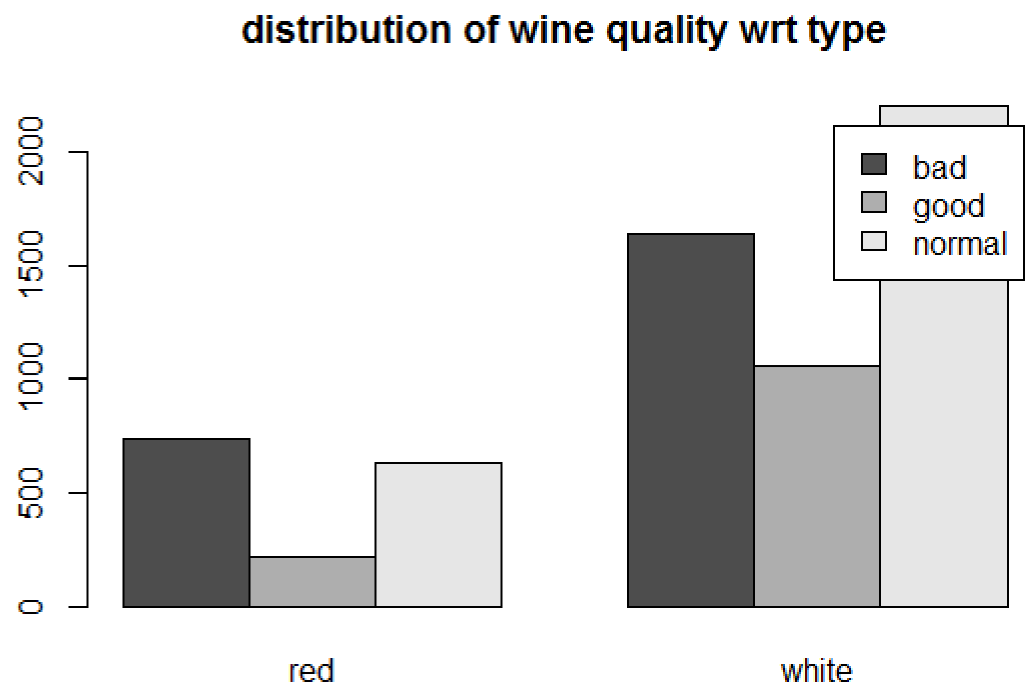
Let's look at the distribution of wine quality



As this graph depicts that quality grade 6 has largest number of wine .there are lot more normal wine than bad and good. So I have classified the wine quality into three categories –bad, good and normal. Total number of wine by categoriewise has been shown in table.



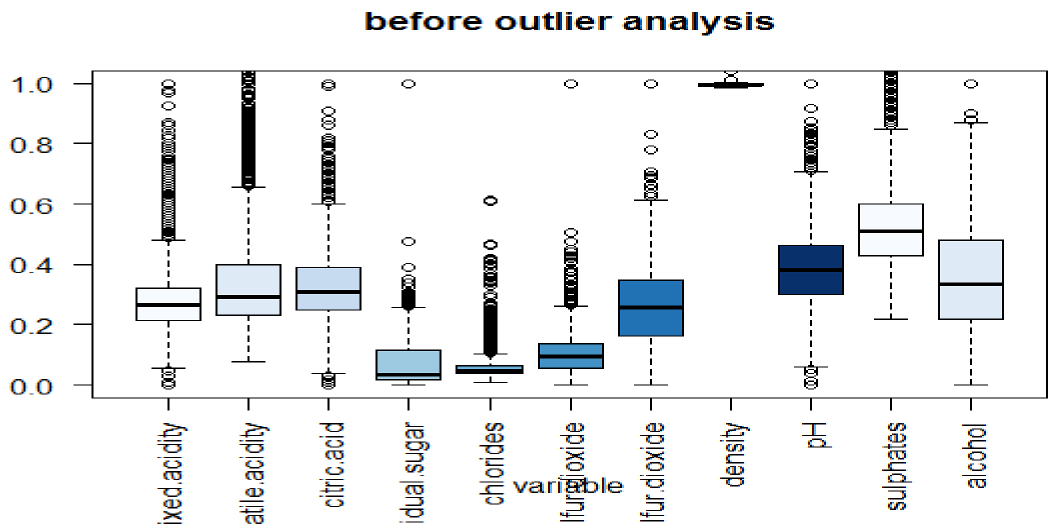
```
   bad    good    normal
  2384    1277     2836
```



This graph depicts the wine quality with wine type. As red wine has the largest number of red wine while white wine has the largest number of normal wine. Good wine holds the least position in both cases.

2. **Cleaning of data**

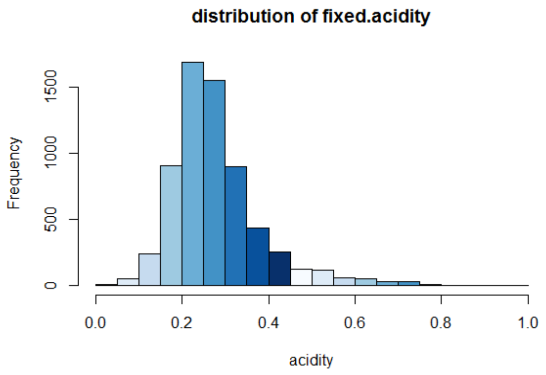
As there is no missing value in data. So I directly shifted to outlier analysis. I have used box plot method to check and remove outliers.

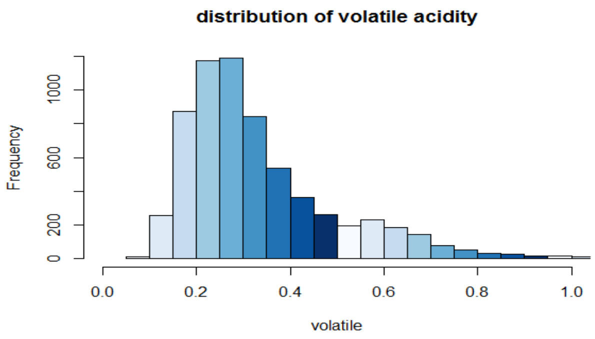


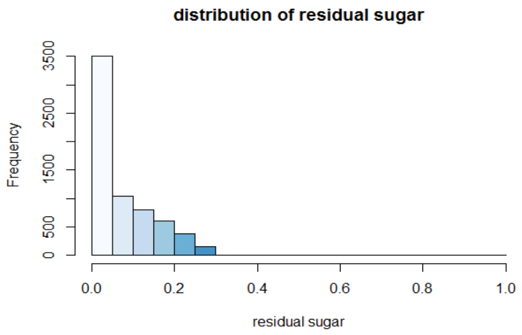
This graph depicts that all variable   have large number of outliers .so I have applied boxplot outlier analysis method to remove outliers.

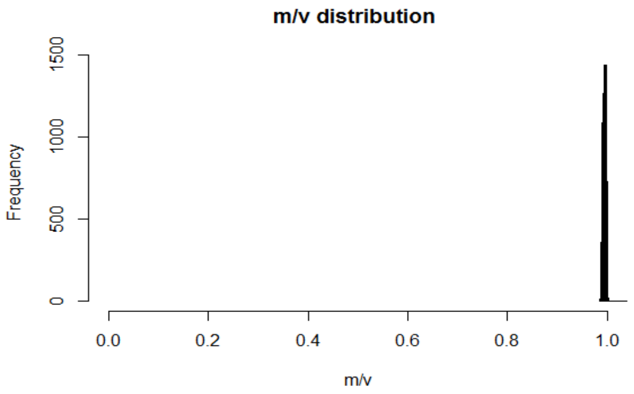
3. visualization

Lets check distribution of some variable using histogram .



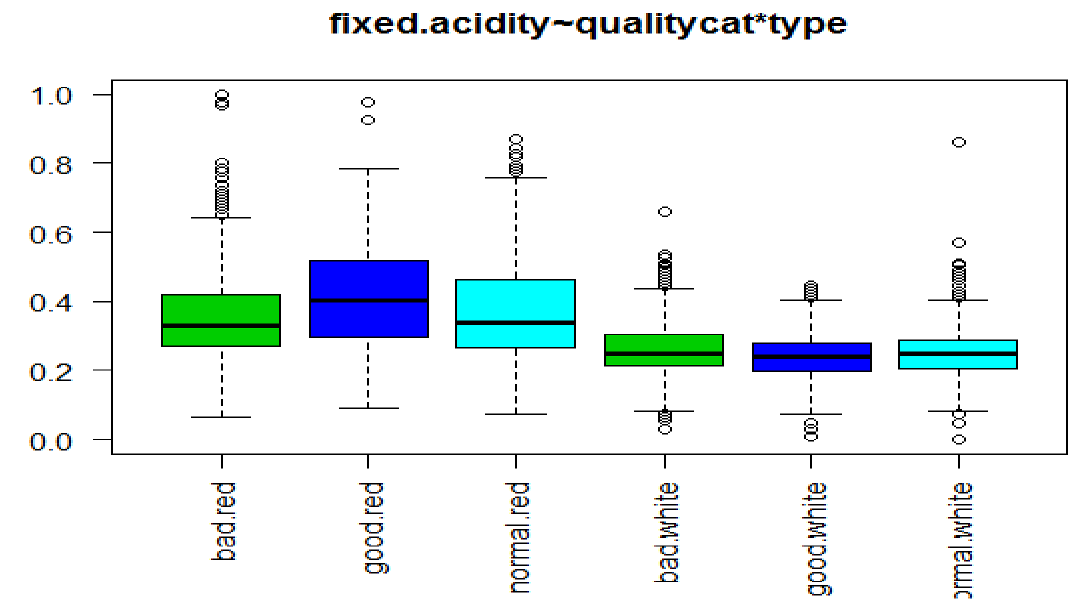




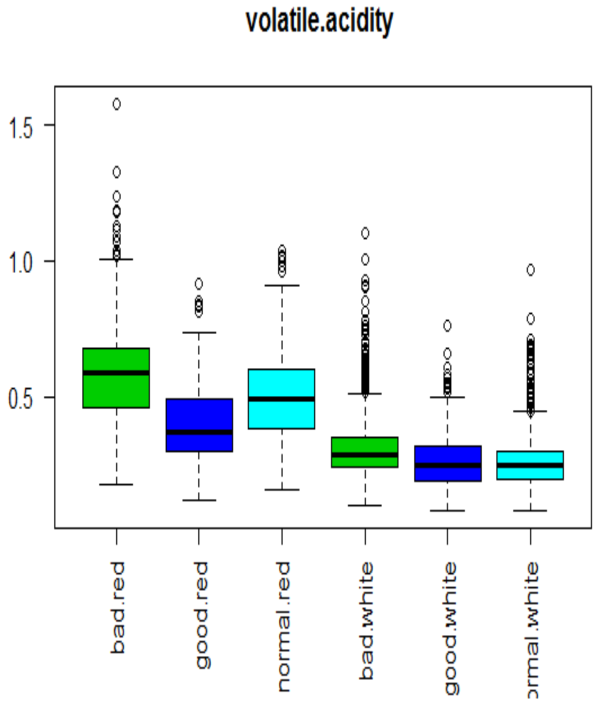


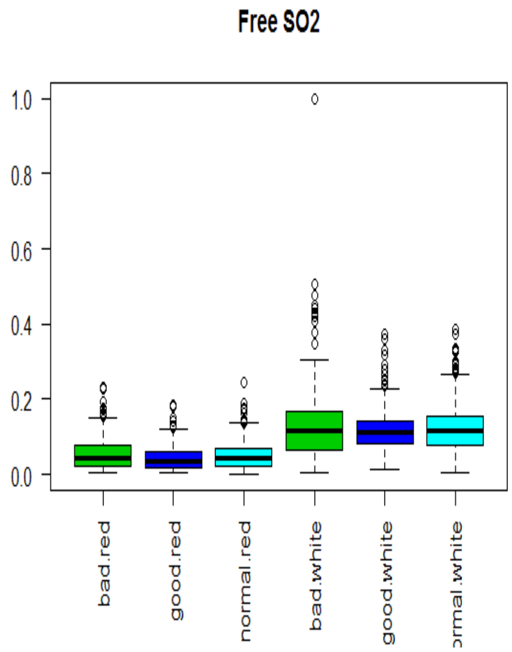
As this graph clearly depicts data are skewed so I will consider median approach method to whole process like distribution of data into train and test. Also I am interested to find relation of target variable with independent variable using median approach by box plot method.

Now let’s use boxplot method to find some   relation .



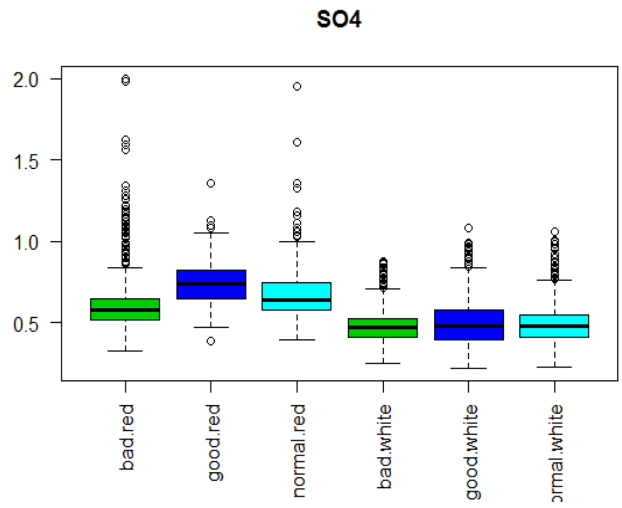
This graph clearly depicts that red wine quality changes proportionally with fixed acidity. But there is no importance of fixed.acidity variable on white wine quality.

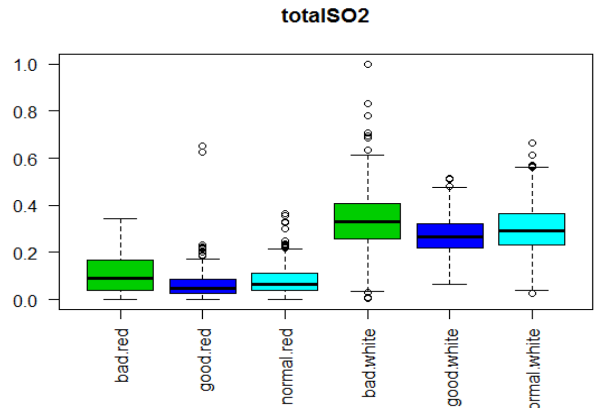


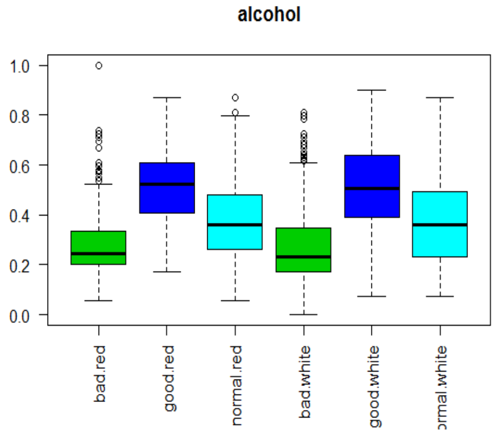


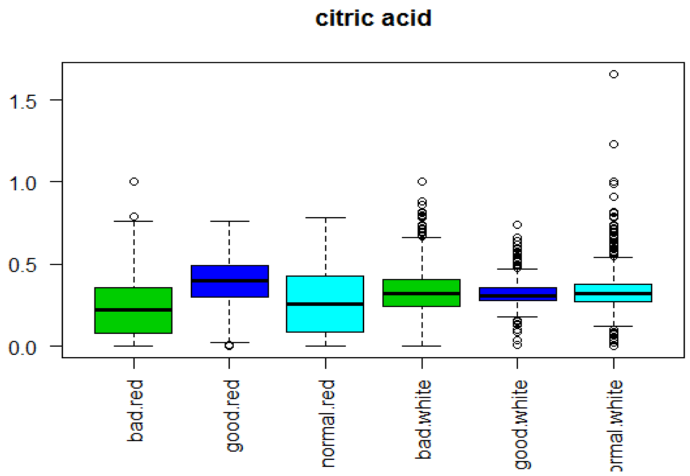
red wine quality changes inversely with volatile.acidity but there is no importance of volatile. acidity on white

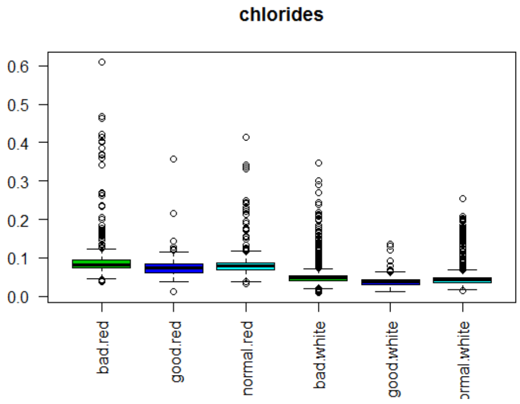
As there is no importance of free So2 on wine quality. But white wine has higher volume than red .

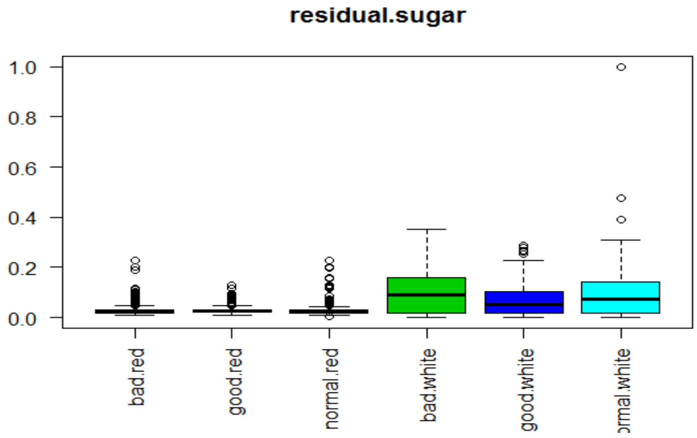

SO4
2.0
1.5
1.0
0.5
bad.red
good.red
normal.red
bad.white
good.white
rmal.white

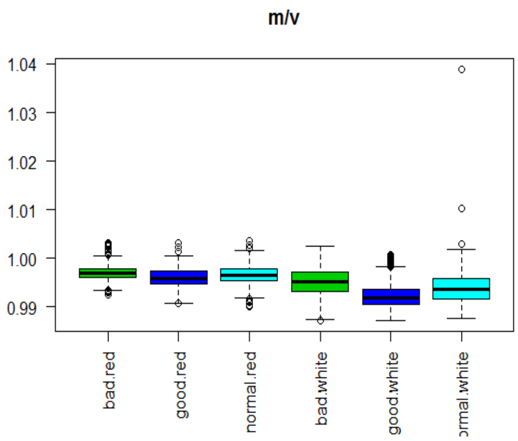

totalSO2
1.0
0.8
0.6
0.4
0.2
0.0
bad.red
good.red
normal.red
bad.white
good.white
rmal.white

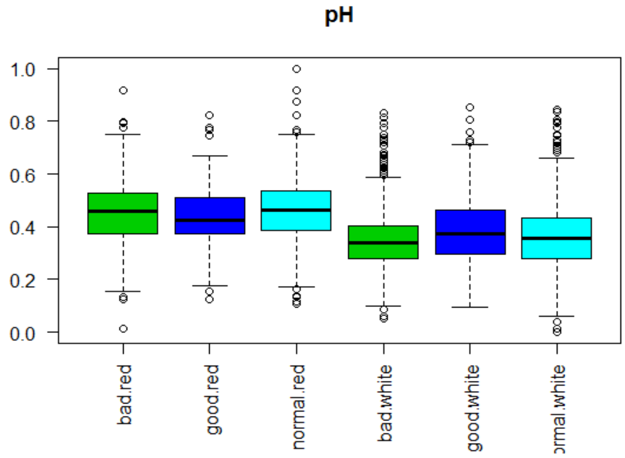

alcohol
1.0
0.8
0.6
0.4
0.2
0.0
bad.red
good.red
normal.red
bad.white
good.white
rmal.white


citric acid
1.5
1.0
0.5
0.0
bad.red
good.red
normal.red
bad.white
good.white
rmal.white


chlorides
0.6
0.5
0.4
0.3
0.2
0.1
0.0
bad.red
good.red
normal.red
bad.white
good.white
rmal.white

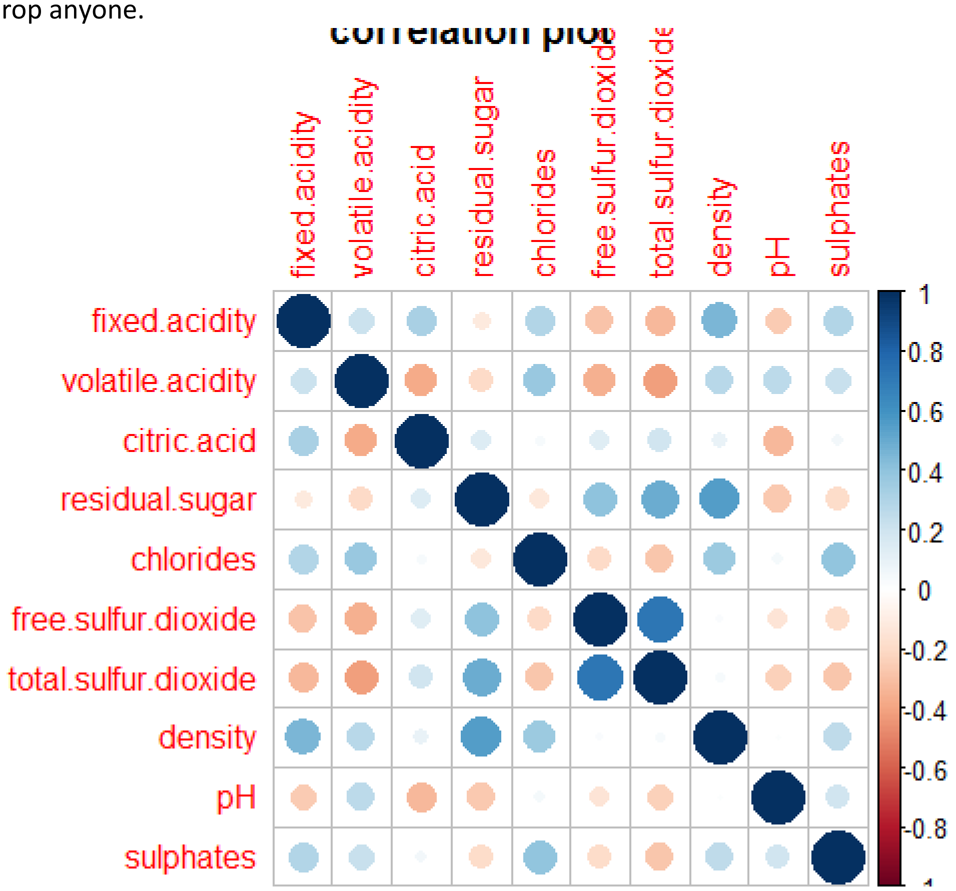

residual.sugar
1.0
0.8
0.6
0.4
0.2
0.0
bad.red
good.red
normal.red
bad.white
good.white
rmal.white





These graph clearly depicts different type of wine has different level of variable importance .As variable alcohol, sulphate , pH ,density and total sulfur dioxide are important for both type of target while fixed,acidity ,volatile.acidity,citric acid are important for only red wine and residual sugar for white wine only.

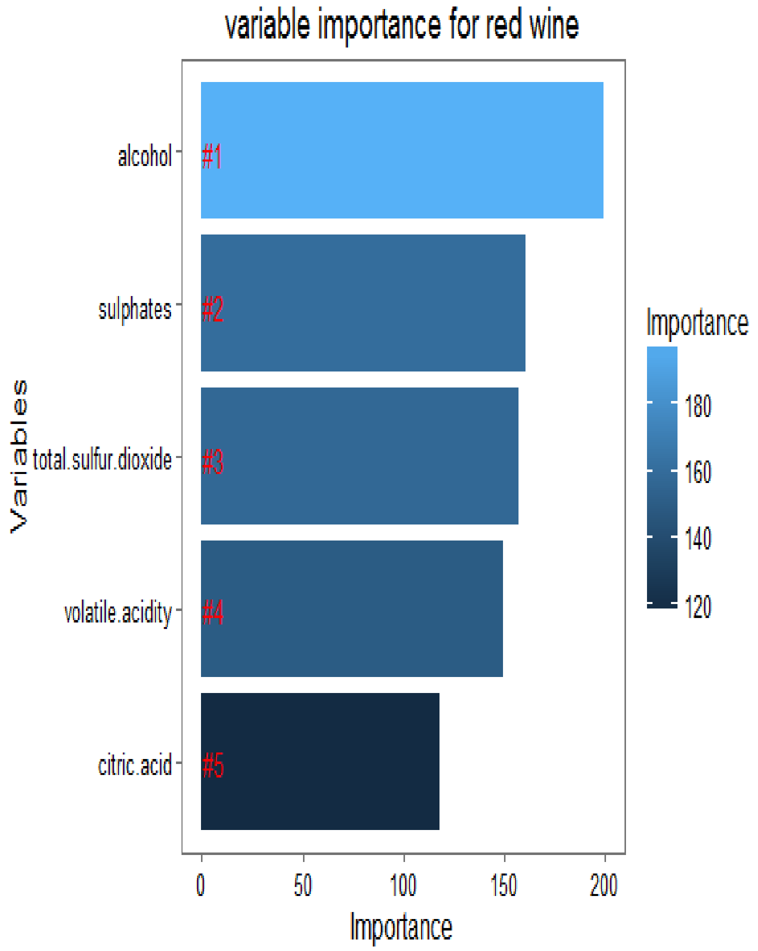
So I have to build the two model for different type of wine . and also none of the variable are highly correlated.so we can't drop anyone.



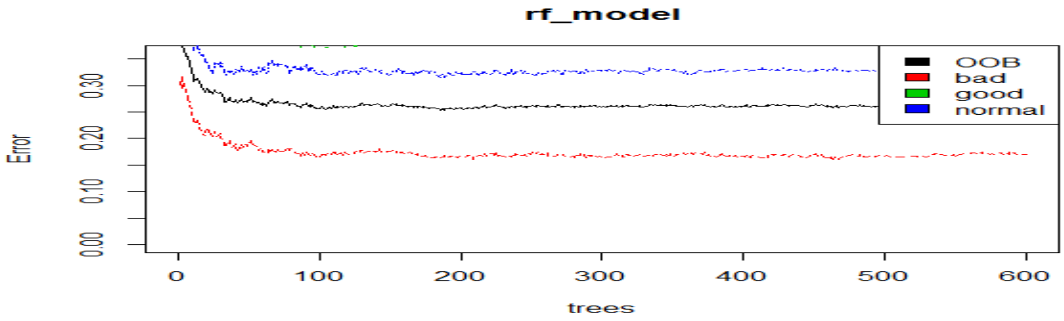
Classification   model

I have used two machine learning method , one is decision tree and other is random forest .random forest build model with high accuracy so I have mentioned here only about random forest method.

#FOR RED WINE

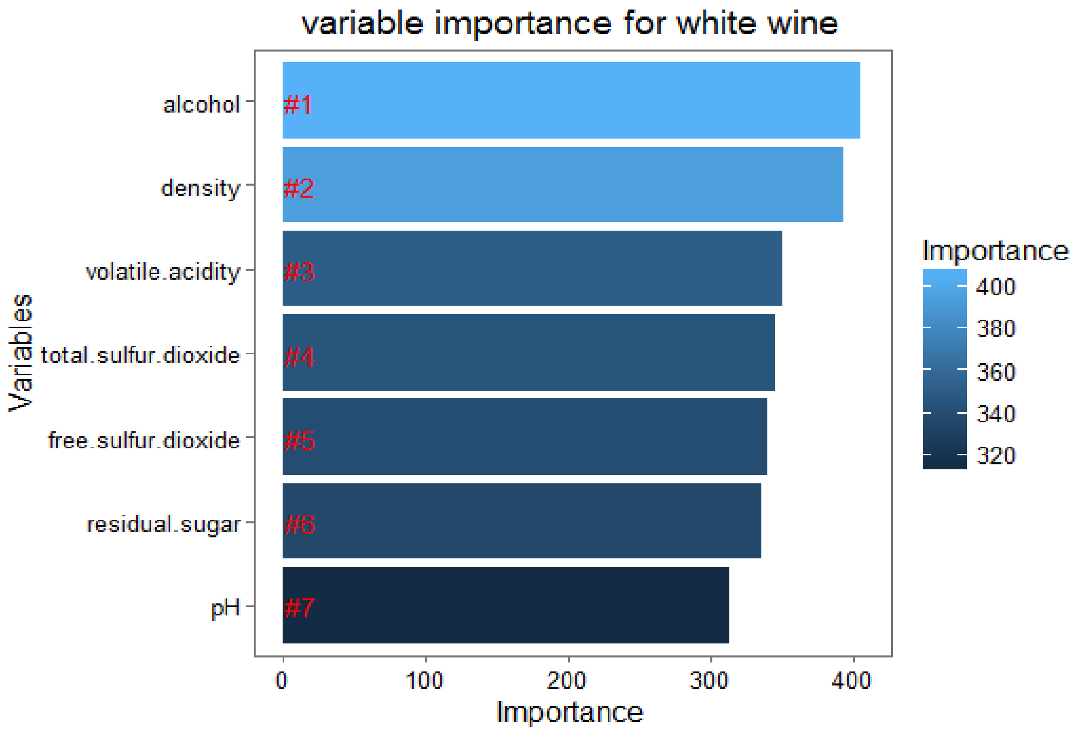


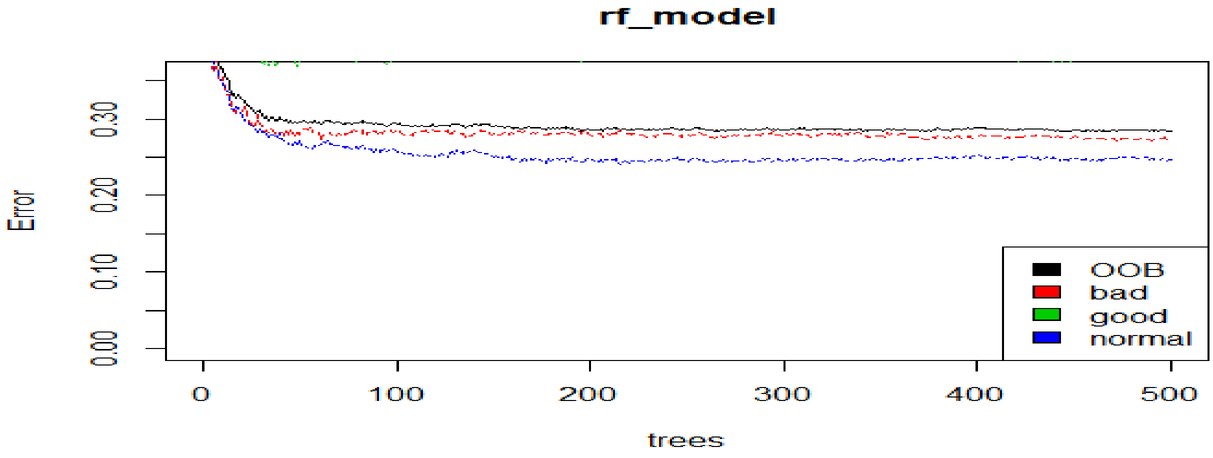
I have used variable alchol,sulphates, total.sulfur.dioxide, volatile.acidity and citric acid importance wise to build the model which predicts the target with 90.97 percent accuracy also from below graph model predicts bad quality wine with high accuracy.



# For white wine

I have used variable alchol, density, ,volatile.acidity , total.sulfur.dioxide ,free.sulfur.dioxide,residual sugar and pH importance wise to build the model which predicts the target with 93.85 percent accuracy also from below graph model predicts normal quality wine with high accuracy.





CONCLUSION :

Variable behaves differently for different type and quality of wine . Random forest model predicts the best result for this case . Model for Red wine predicts the test data with 95.45 % accuracy and the best suited for bad quality wine and Model for white wine predicts the test data with 93.50 % accuracy and the best suited for normal quality Wine .

Variable	Red wine	White wine
1. alcohol	Directly proportional	Directly proportional
2. Sulphates	Directly proportional	No effect
3. Total.sulfur.dioxide	Inversely proportional	Inversely proportional
4. Volatile.acidity	Inversely proportional	Inversely proportional with very little effect
5. Citric acid	Directly Propotional	No effect
6. Residual.sugar	No effect	Good wine has less residual of sugar while bad and normal have the same level.