CSE5243
HW 4: Analysis of dataset- Wine_ quality

Overview

This report is concentrated on the classification analysis on wine-red dataset, to test how different classification models preform on wine-red dataset. It covers four parts: data exploration, preprocessing, model development and model evaluation.

1. Preliminary data analysis

1. Summary statistic (For this part, I used SAS)

Tabl1. Summary statistic data for wine dataset

Statistic data for wine

The MEANS Procedure

Variable	N	N Miss	Minimum	Lower Quartile	Upper Quartile	Mean	Median	Maximum	Std Dev	Range
fx_acidity	1599	0	4.6000000	7.1000000	9.2000000	8.3196373	7.9000000	15.9000000	1.7410963	11.3000000
vol_acidity	1599	0	0.1200000	0.3900000	0.6400000	0.5278205	0.5200000	1.5800000	0.1790597	1.4600000
citric_acid	1599	0	0	0.0900000	0.4200000	0.2709756	0.2600000	1.0000000	0.1948011	1.0000000
resid_sugar	1599	0	0.9000000	1.9000000	2.6000000	2.5388055	2.2000000	15.5000000	1.4099281	14.6000000
chlorides	1599	0	0.0120000	0.0700000	0.0900000	0.0874665	0.0790000	0.6110000	0.0470653	0.5990000
free_sulf_d	1599	0	1.0000000	7.0000000	21.0000000	15.8749218	14.0000000	72.0000000	10.4601570	71.0000000
tot_sulf_d	1599	0	6.0000000	22.0000000	62.0000000	46.4677924	38.0000000	289.0000000	32.8953245	283.0000000
density	1599	0	0.9900700	0.9956000	0.9978400	0.9967467	0.9967500	1.0036900	0.0018873	0.0136200
pH	1599	0	2.7400000	3.2100000	3.4000000	3.3111132	3.3100000	4.0100000	0.1543865	1.2700000
sulph	1599	0	0.3300000	0.5500000	0.7300000	0.6581488	0.6200000	2.0000000	0.1695070	1.6700000
alcohol	1599	0	8.4000000	9.5000000	11.1000000	10.4229831	10.2000000	14.9000000	1.0656676	6.5000000
quality	1599	0	3.0000000	5.0000000	6.0000000	5.6360225	6.0000000	8.0000000	0.8075694	5.0000000

From above summary statistic data table, we can know the basic distribution of each attribute (eg. mean and standard deviation). For instance, the total sulfur dioxide attribution spread largely, the range and standard deviation is extremely big comparing to other attributes. Inversely, the standard deviation and range of density is small. That means the density value concentrated on mean value.

2. Boxplot((For this part, I used R)

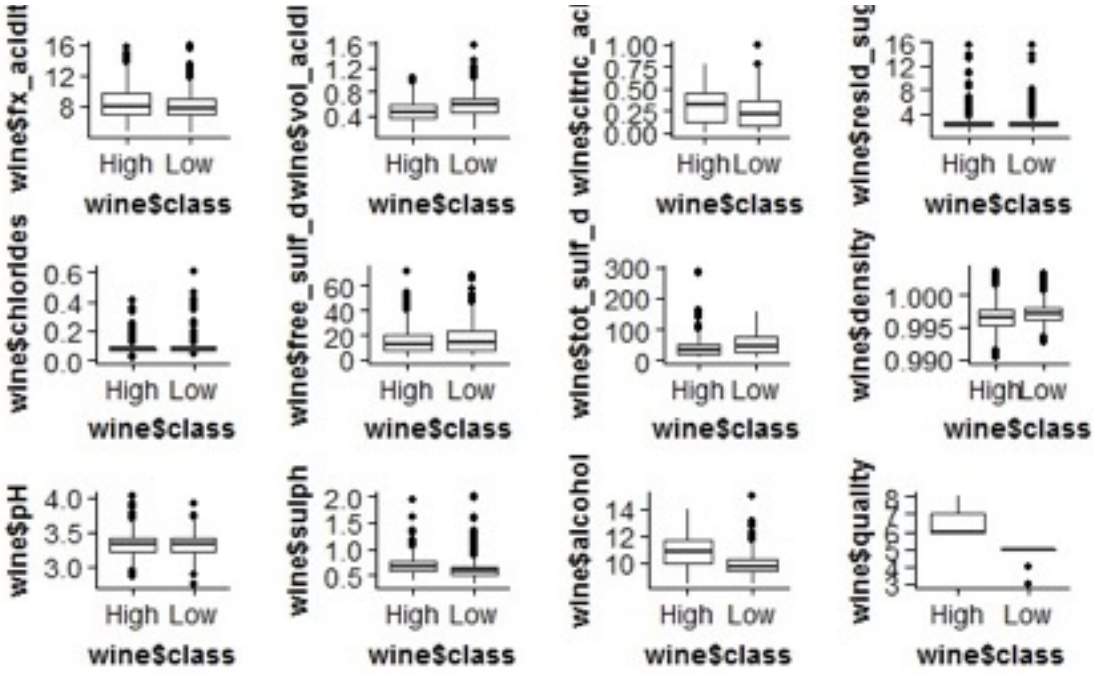


Figure 1 Boxplot Matrix for each attribute.

From the boxplot matrix (Figure1), we can know the distribution of each attributes based on class and outliers.

a. For attribute chlorides, sulph, residual sugar, the spread of the distribution is small. And also, the mean of attribute are similar between two different classes. There are many outliers for attribute chlorides and residual sugar.
b. For attribute PH and free sulfur dioxide, the distribution and mean are almost same between high and low quality. That means, there is no much difference on PH and free sulfur dioxide between high quality wine and low quality wine.
c. For attribute citric acid, the range of distribution is larger comparing to other attributes.
d. For attribute alcohol, the difference of distribution between class high and class low are large. From the boxplot, we can know that higher density of alcohol imply high quality wine.

3. Correlation (For this part, I used R)

Table 2. Correlation matrix of attributes

	fx_acidity	vol_acidity	citric_acid	resid_sugar	chlorides	free_sulf_d	tot_sulf_d	density	pH	sulph	alcohol	quality
fx_acidity	1.000	-0.256	0.672	0.115	0.094	-0.154	-0.113	0.668	-0.683	0.183	-0.062	0.124
vol_acidity	-0.256	1.000	-0.552	0.002	0.061	-0.011	0.076	0.022	0.235	-0.261	-0.202	-0.391
citric_acid	0.672	-0.552	1.000	0.144	0.204	-0.061	0.036	0.365	-0.542	0.313	0.110	0.226
resid_sugar	0.115	0.002	0.144	1.000	0.056	0.187	0.203	0.555	-0.086	0.006	0.042	0.014
chlorides	0.094	0.061	0.204	0.056	1.000	0.006	0.047	0.201	-0.265	0.371	-0.221	-0.129
free_sulf_d	-0.154	-0.011	-0.061	0.187	0.006	1.000	0.668	-0.022	0.070	0.052	-0.069	-0.051
tot_sulf_d	-0.113	0.076	0.036	0.203	0.047	0.668	1.000	0.071	-0.066	0.043	-0.206	-0.185
density	0.668	0.022	0.365	0.555	0.201	-0.022	0.071	1.000	-0.342	0.149	-0.496	-0.175
pH	-0.683	0.235	-0.542	-0.086	-0.265	0.070	-0.066	-0.342	1.000	-0.197	0.206	-0.058
sulph	0.183	-0.261	0.313	0.006	0.371	0.052	0.043	0.149	-0.197	1.000	0.094	0.251
alcohol	-0.062	-0.202	0.110	0.042	-0.221	-0.069	-0.206	-0.496	0.206	0.094	1.000	0.476
quality	0.124	-0.391	0.226	0.014	-0.129	-0.051	-0.185	-0.175	-0.058	0.251	0.476	1.000

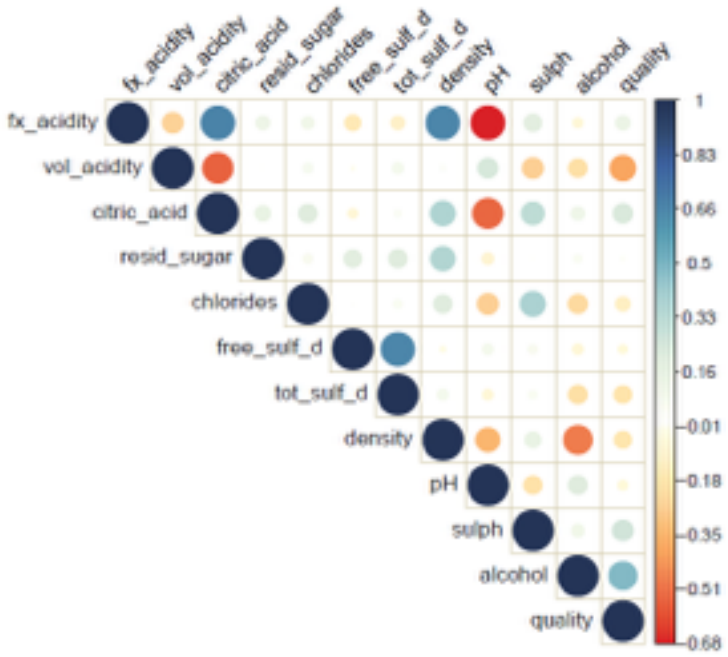


Figure2. Correlation Plot

Figure2, the circle represents the correlation between two different attributes. The size and brightness of circle represents the scale of correlation. The bigger and darker color implies higher correlation value. The red circle means two attributes are negatively related. The blue circle means two attribute are positively related.

From the correlation plot (Figure2) and correlation matrix (table1), we can know that:

a. Attribute citric acid and density has strong positive relation with attribute fixed acidity. PH value is strongly negatively related fixed acidity. The absolute value correlation are larger than 0.6.
b. Other relatively correlated attributes are volatile acidity and citric acid, citric acid and PH value, free sulfur dioxide, total sulfur dioxide.
c. The attributes which affects quality more are alcohol and volatile acidity. Higher alcohol and lower volatile acidity implies high quality wine.

4. Scatterplot

In this part, I select some attributes which have higher correlation based on correlation matrix and correlation plot to analysis the scatterplot individually.

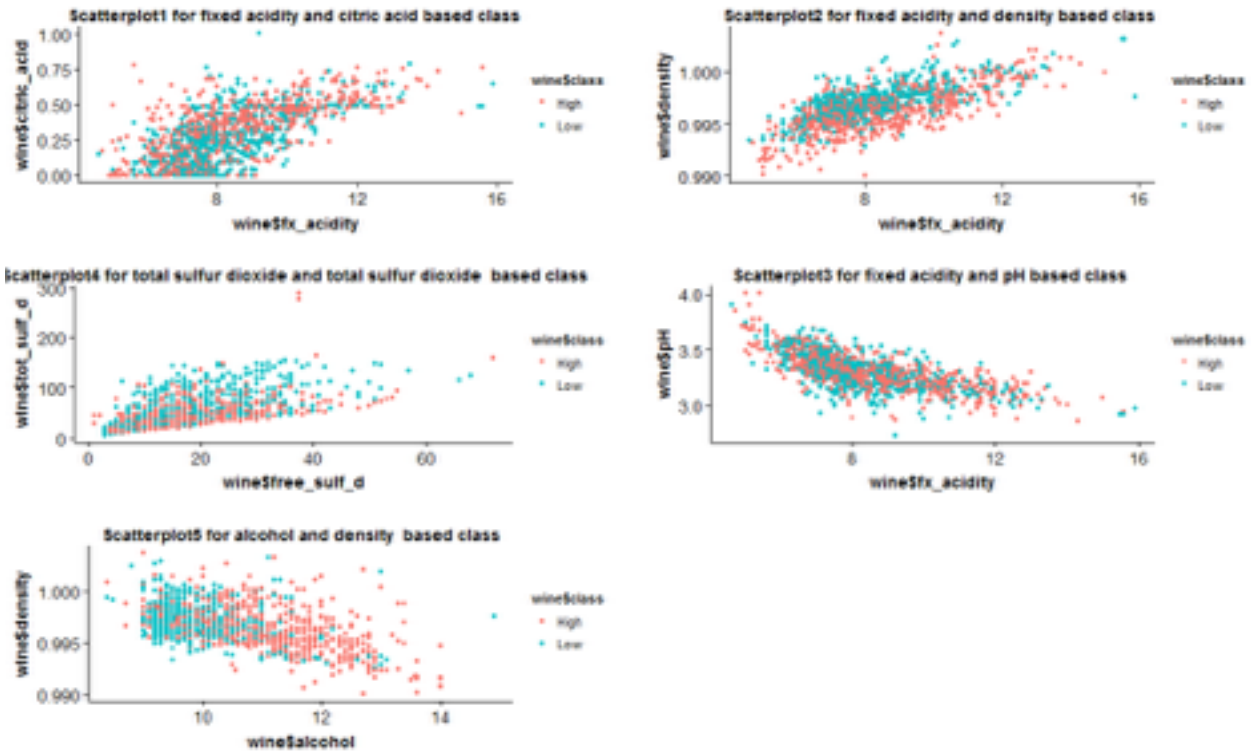


Figure 3 Scatterplots for some attributes

a. From scatterplot1, citric acid is positively related with fixed acidity. Citric acid grows as fixed acidity increase for both high quality wine and low quality wine.
b. From scatterplot2, density is positively related with fixed acidity. Density grows as fixed acidity increase for both high quality wine and low quality wine.
c. From scatterplot3, free sulfur dioxide is positively related with total sulfur dioxide. Free sulfur dioxide grows as total sulfur dioxide increase for both high quality wine and low quality wine.
d. From scatterplot4, fixed acidity is negatively related with PH. Fixed acidity decrease as PH grow for both high quality wine and low quality wine.
e. From scatterplot5, density is negatively related with alcohol. Higher density wine has lower alcohol. And, lower alcohol wine implies higher quality wine.

2. Data transformation

2.1 Missing data

In Table1, from column NMISS, i.e., Number of missing data, we can know that there is no missing data in this dataset.

2.2 Outlier

According to boxplot matrix, there are some outliers in each attributes. I won't decide to discard them. Because they might affect the result of classification and the number of outliers is not small.

2.3 Duplicate data

```
x<-which(duplicated(wine)==TRUE)
length(x)
```

```
## [1] 240
```

There are 240 duplicated data in Wine.csv dataset. I decide not discard these duplicated data, because the percentage of duplicate data is over15%. That's not a small portion and we are not sure whether these duplicated data represent same object or distinct object.

2.4 Feature subset

Remove column ID. This attribute is irrelevant feature. It contains useful information for classification.

Remove column quantity. This attribute display the duplicated information with class attribute.

Redundant feature: Free sulfur dioxide and total sulfur dioxide. Total sulfur dioxide is measured by free sulfur dioxide and bound sulfur dioxide. Thus, they contain same duplicated information. So I decide not to remove feature free sulfur dioxide, because the relation of free sulfur dioxide and total sulfur dioxide with other attribute are not same. For instance, from the correlation plot, we can know that the relation between total sulfur dioxide and alcohol or quality is stronger than the relation between free sulfur dioxide and alcohol or quality

2.5 Data transformation.

For model decision tree, a rules-based classifier, Naïve Bayes, Artificial Neural Network, Ensemble learner, there is no need of normalization.

For Naïve Bayes model, I discretized the continuous data before creating model, which can improve the accuracy of the model performance.

For SVM model, the normalizing method inside WEKA is used. The formula is shown below. This method is more robust to outlier than standard normalization. It convert all attributes in the same scales, avoiding the attribute which has big value dominate in calculation.

```
value = (vals[j] - m_MinArray[j]) / (m_MaxArray[j] - m_MinArray[j])
              * m_Scale + m_Translation;
```

For part 3 and 4, WEKA is used for classification.

3. Model development

In this part, the cross validation is used for define the test data instead of using conventional validation dataset (partitioning the data set into training dataset and test dataset) ,because this method generate independent dataset which can limit overfitting problem and it can provide the average of generalization error which is more correct than the performance measure only derived from training data. In general, if the fold number increase, the performance of model might better. But the running time for building model become longer.

3.1 Decision tree

In this case, I selected J48 decision tree. This algorithm create decision tree based on information gain. Figure 4 showed the part of the decision tree. (The fold number of cross validation is 20)

From figure 5, we can know the size of tree generated through J48 algorithm is 211. It is a big tree. The complexity is high.

```
alcohol <= 10.5
|   tot_sulf_d <= 98
|   |   sulph <= 0.57
|   |   |   alcohol <= 9.7
|   |   |   |   alcohol <= 9
|   |   |   |   |   fx_acidity <= 7.8: Low (4.0)
|   |   |   |   |   fx_acidity > 7.8
|   |   |   |   |   |   citric_acid <= 0.55: High (5.0)
|   |   |   |   |   |   citric_acid > 0.55: Low (2.0)
|   |   |   |   alcohol > 9: Low (195.0/28.0)
|   |   |   alcohol > 9.7
|   |   |   |   vol_acidity <= 0.735
|   |   |   |   |   sulph <= 0.47: Low (17.0/2.0)
|   |   |   |   |   sulph > 0.47
|   |   |   |   |   |   vol_acidity <= 0.575: High (47.0/15.0)
|   |   |   |   |   |   vol_acidity > 0.575
|   |   |   |   |   |   |   fx_acidity <= 7.3: Low (13.0/1.0)
|   |   |   |   |   |   |   fx_acidity > 7.3
|   |   |   |   |   |   |   |   sulph <= 0.56
```

Figure 4 Part of J48 decision tree.

```
Number of Leaves  :     106

Size of the tree :      211
```

Figure 5Size of decision J48 decision tree.

For folds number of cross validation is 10, 20, 30, and 100, the size of the tree are all 211. The complexity of tree is same. The accuracy of decision tree is 73.9212 %, 75.4221 %, 75.6723%, 75.5472 %, respectively. Thus after fold number is bigger than 20, the accuracy of decision trees won't change much. So, for this case, the fold number is set as 20.

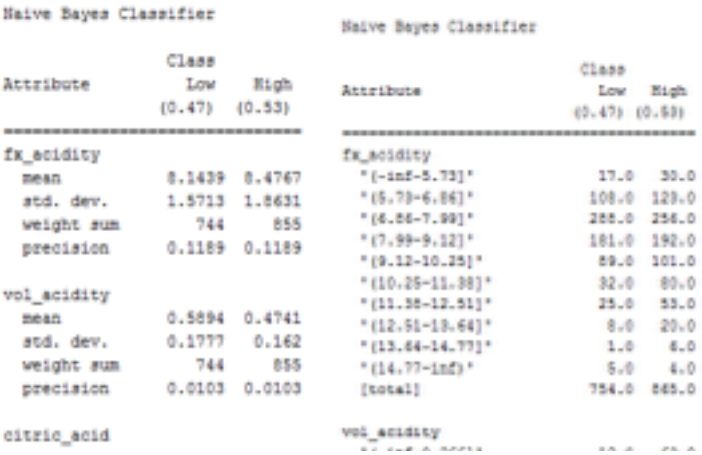
3.2 A rules-based classifier.

In this part, I selected DTNB algorithm, and set folds of cross validation as 20. For decision table algorithm, the accuracy is 72.5453 %, for Naïve Bayes algorithm, classification accuracy is 73.9837%. However, when I use DTNB algorithm which is combination of decision table and Naïve Bayes, the accuracy is improved up to 75.0469 %.

3.3 Naïve Bayes

In this part, fold number is also set as 20. Because this no much difference on accuracy between fold number =20 and fold number>20.

In this part I test the Naïve Bayes model based on discretized attribute and continuous attribute. The part of each model is shown as below. The accuracy is 73.9837 % and 72.733%, respectively. So I decide use discretize attribute for this model.



```
Naive Bayes Classifier

                        Class
Attribute                 Low    High
                        (0.47)  (0.53)
================================
fx_acidity
  mean                  8.1439  8.4767
  std. dev.             1.5713  1.8631
  weight sum               744     855
  precision             0.1189  0.1189

vol_acidity
  mean                  0.5894  0.4741
  std. dev.             0.1777   0.162
  weight sum               744     855
  precision             0.0103  0.0103

citric_acid
```

```
Naive Bayes Classifier

                                Class
Attribute                         Low    High
                                (0.47)  (0.53)
========================================
fx_acidity
  "(-inf-5.73]"                   17.0    30.0
  "(5.73-6.86]"                  108.0   129.0
  "(6.86-7.99]"                  288.0   256.0
  "(7.99-9.12]"                  181.0   192.0
  "(9.12-10.25]"                  89.0   101.0
  "(10.25-11.38]"                 32.0    80.0
  "(11.38-12.51]"                 25.0    55.0
  "(12.51-13.64]"                  8.0    20.0
  "(13.64-14.77]"                  1.0     6.0
  "(14.77-inf)"                    5.0     4.0
  [total]                        754.0   865.0

vol_acidity
```

Continous attribute Discretized attribute

3.4 Artificial Neural Network

In this part, folds number of cross validation is set as 20.which is consistent with other model. This algorithm assigns weight to each attribute and adjust the weight to make the output ANN is consistent with the class label of training data. In this case, the activation function of each iteration is nonlinear function- sigmoid. From the output we can know that there are total 7 iterations

```
Sigmoid Node 7
    Inputs    Weights
    Threshold    -1.5339689188643963
    Attrib fx_acidity    5.265701708027778
    Attrib vol_acidity    -3.6135520344571823
    Attrib citric_acid    0.6628342474617536
    Attrib resid_sugar    5.399829569587707
    Attrib chlorides    -5.107083621662072
    Attrib free_sulf_d    7.102768112041244
    Attrib tot_sulf_d    -1.4687138435895448
    Attrib density    -4.767512682078463
    Attrib pH    5.239431143204673
    Attrib sulph    2.62983682851334
    Attrib alcohol    2.4818798004196605
Class Low
    Input
    Node 0
Class High
    Input
    Node 1
```

Figure 5 Part of J48 decision tree.

3.5Support vector machine

In this part, folds number of cross validation is set as 20, which is consistent with other models. For this model, normalization need be executed before classification, because the margin is distance. If we use raw data, the scale of each attribute is not same .The attribute which has large value might dominate the calculation. So, we need normalize them in same scale. According to table 2, we can know that the performance with normalization is better. The accuracy, precision, recall and F-measure are enhanced.

Table2 Comparison between normalization and non-normalization

	Accuracy	Percision	recall	F-measure
without Normalization	71.92%	0.699	0.696	0.698
with normalizaion	73.23%	0.688	0.777	0.73

3.6 Ensemble learner

In this part, all folds of cross validation are also set as 20.which is consistent with other model. Ado boost, Bagging, Random Forest are tested, respectively.

For bagging, there are 10 iterations. In each step, it generates a REP tree. It can improve the performance by reduce the variance of base classifier, and work well than just one decision tree.

Random forest of 100 trees, each constructed while considering 4 random features. Random forest constructs multiple trees and combines the trees, which can reduce the error. There is 1599 instance, and 1290 of them are correctly classified.

4. Model evaluation

These all models are based on 20 folds cross validation, which separates the original data into 20 partitions. 19 partitions are used as training data to construct the classification model and the left one partition is used as testing data to examine the model. The performance of each model is shown in Table3.

Table3 Performance of each model.

	J48 Decision tree	Rule based classifier	Naive bayes	ANN	SVM	Adaboost	Bagging	RandomForest
Accuracy	75.42%	75.05%	73.98%	73.92%	71.92%	72.48%	77.99%	82.36%
Percision	0.731	0.716	0.703	0.713	0.699	0.706	0.759	0.806
recall	0.746	0.767	0.763	0.737	0.696	0.7	0.772	0.819
F-measure	0.739	0.741	0.732	0.724	0.698	0.703	0.765	0.812

According to the table3, we can know Random Forest perform best based on accuracy, precision, recall, and F-measure, following by bagging. SVM perform worst comparing with other model. The value of these four measures is less than others.

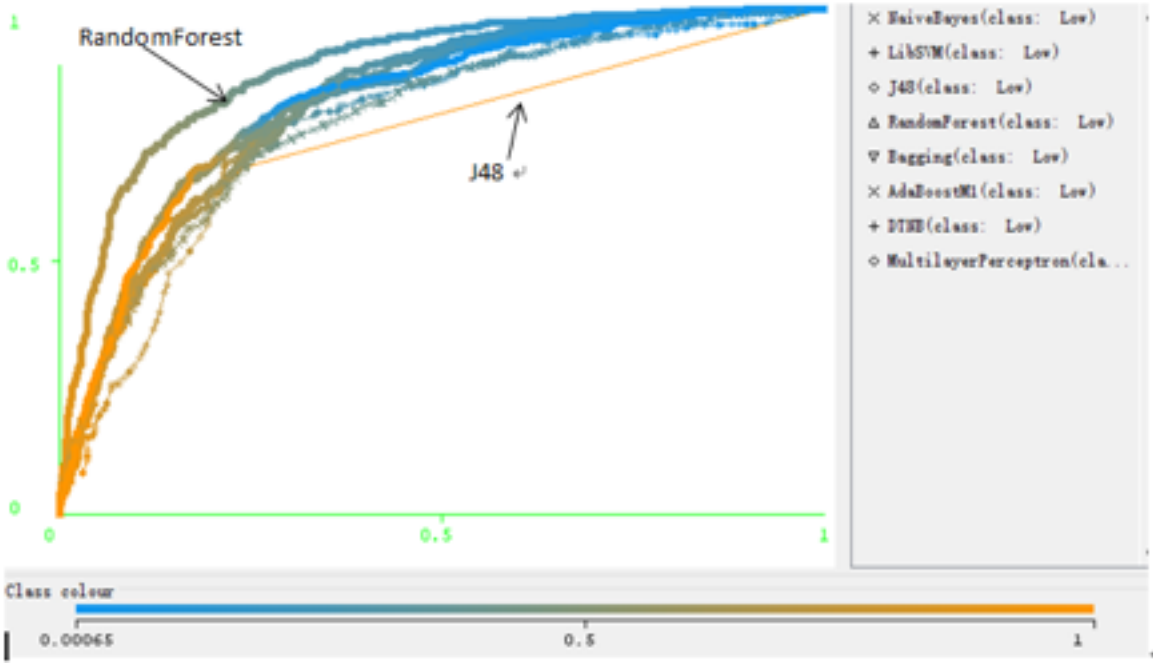


Figure 6 ROC curve for each model performance.

Figure 6 is the ROC curve for each model, which is generated through WEKA. From the figure we can know that all these models are above the diagonal line. They all perform well. By comparing the ROC curve, we can know that Random forest is the best choice, because the area under the curve is bigger than other model. The decision tree created by J48 performs not well as other models. Other models, like Libsvm, ANN, DTNB, Naïve Bayes, Bagging and Adoboost , perform similarly . There is no much difference on accuracy of performance and ROC curve.

5 Conclusions

Based on the resulted displayed by performance table and ROC curve, Random forest works better than other models. But due to the random forest generate large amount trees which might sacrifice the run speed and cost much more time than decision trees mothed.