CS760 Homework1

Data:

To build a classification model based on different models, the dataset is about the information of a certain song, features include genre, composer, librettist, singer, album, language, length, plays, score and release yea, and the label means whether this person like this song or not.

Decision Tree:

Leaves: 13

Size of the tree: 22

Correctly Classified Instances	80	76.1905 %
Incorrectly Classified Instances	25	23.8095 %
Kappa statistic	0.5214	
Mean absolute error	0.2631	
Root mean squared error	0.4692	
Relative absolute error	52.7186 %	
Root relative squared error	93.9248 %	

1-nearest neighbor:

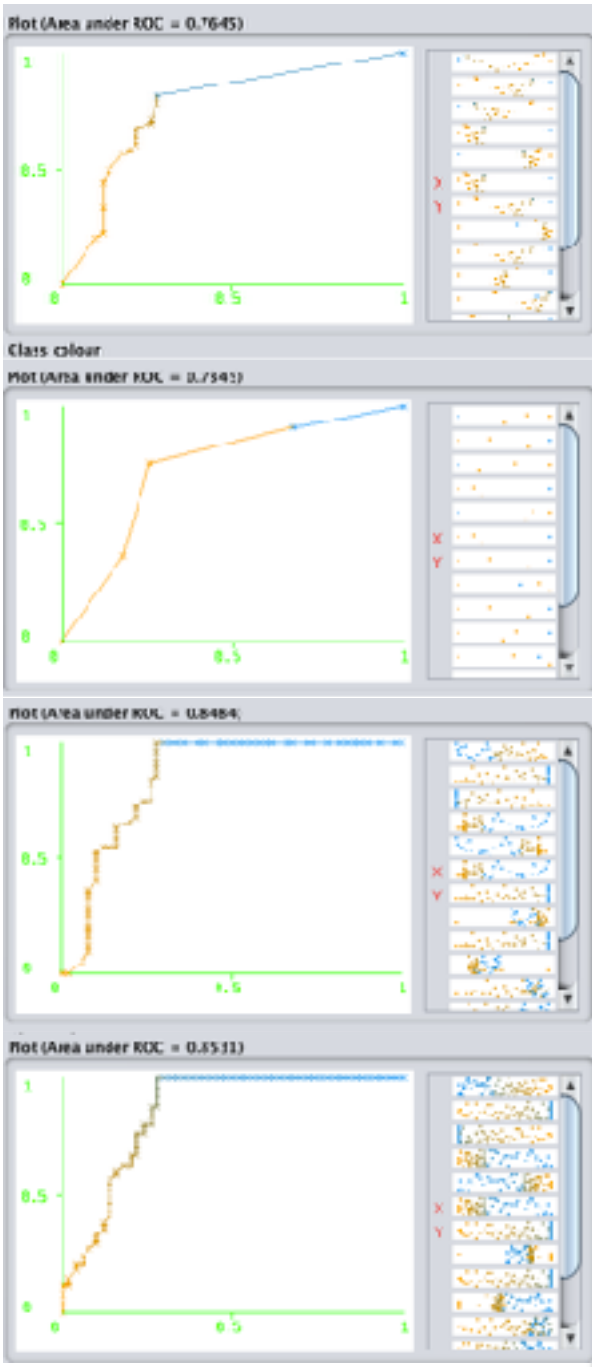
Correctly Classified Instances	79	75.2381 %
Incorrectly Classified Instances	26	24.7619 %
Kappa statistic	0.5036	
Mean absolute error	0.2529	
Root mean squared error	0.4925	
Relative absolute error	50.6714 %	
Root relative squared error	98.5966 %	

Bayes network:

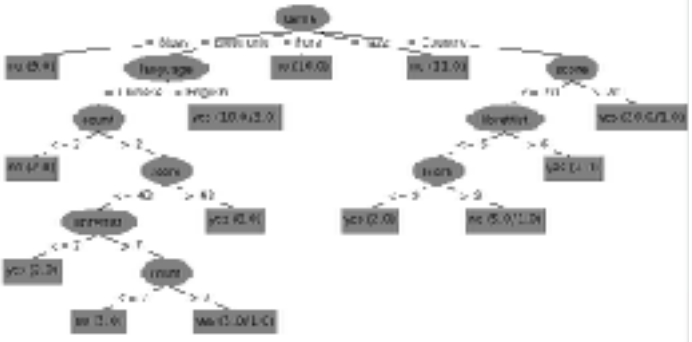
Correctly Classified Instances	91	86.6667 %
Incorrectly Classified Instances	14	13.3333 %
Kappa statistic	0.7293	
Mean absolute error	0.2358	
Root mean squared error	0.3377	
Relative absolute error	47.2541 %	
Root relative squared error	67.6057 %	

Random Forest:

Correctly Classified Instances	86	81.9048 %
Incorrectly Classified Instances	19	18.0952 %
Kappa statistic	0.6343	
Mean absolute error	0.2661	
Root mean squared error	0.3585	
Relative absolute error	53.3233 %	
Root relative squared error	71.7613 %	



Accord to the tree in the right, we can see that the genre is the main factor which influence the preference, and then the average score of the song and recent play frequency will take positive effect, besides the person prefer librettist whose number is larger than 6.



From roc curves we know that the auc corresponding to its classification accuracy. The last two methods give better performance than first two.