Chapter 7

Experimental Analysis and Results of classification of unstructured data from online course web pages

This chapter lists the results of the project and the inferences to be made from the testing results. The evaluation metrics have been listed and the results have been accordingly quantified. A real dataset was taken into consideration and the algorithms run on this set. The results are specific to this dataset but give an indication as to the general performance trend of the various classification algorithms. One algorithm each has been chosen from various domains namely decision-tree based, mathematical, probabilistic, lazy learner and ensemble methods.

7.1 Evaluation Metrics

Classifier performance depends greatly on the characteristics of the data to be classified. There is no single classifier that works best on all given problems, a phenomenon that may be explained by the no-free-lunch theorem. Various empirical tests have been performed to compare classifier performance and to find the characteristics of data that determine classifier performance. Determining a suitable classifier for a given problem is however still more an art than a science.

The measures **precision, recall and f-measure** are popular metrics used to evaluate the quality of a classification system. **Receiver Operating Characteristic (ROC)** curves have been used to evaluate the tradeoffs between true- and false-positive rates of classification algorithms.

True positives are the number of items correctly labelled as belonging to the positive class while **false positives** are items incorrectly labelled as belonging to the class. These measures can be expressed as fractions of the entire data set, known as TP rate and FP rate.

Precision (also called positive predictive value) is the fraction of retrieved instances that are relevant, while recall (also known as sensitivity) is the fraction of

relevant instances that are retrieved. Both precision and recall are therefore based on an understanding and measure of relevance. In classification, the precision for a class is the number of true positives (i.e. the number of items correctly labelled as belonging to the positive class) divided by the total number of elements labelled as belonging to the positive class (i.e. the sum of true positives and false positives, which are items incorrectly labelled as belonging to the class) [54].

Recall in this context is defined as the number of true positives divided by the total number of elements that actually belong to the positive class (i.e. the sum of true positives and false negatives, which are items which were not labelled as belonging to the positive class but should have been) [54].

The F-measure is the harmonic mean of precision and recall, giving an overall expression of performance of an algorithm. A Receiver Operating Characteristic (ROC), [55] or simply ROC curve, is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied. It is created by plotting the fraction of true positives out of the total actual positives (TPR = True Positive Rate) vs. the fraction of false positives out of the total actual negatives (FPR = False Positive Rate), at various threshold settings.

7.2 Experimental Dataset

The initial dataset used for bootstrapping consisted of 350 positive and negative samples taken from open course websites, labelled manually and designed to cover all corner case data values that could appear while classifying crawled web pages. This dataset was used to evaluate various classification algorithms and compare their performance with the random forest classifier written for this project. 66% of this data was used as training set while the remaining was used as a test set. The data was randomly segregated into these two sets in order to evaluate the performance of various algorithms.

7.3 Performance Analysis

The data obtained for various algorithms is shown below for the various metrics listed above. The ROC curves and measures for the bootstrap dataset have been

mentioned, as these are accurately labelled data and will provide a clear idea of the performance of these algorithms.

7.3.1 C4.5 Algorithm Evaluation

The C4.5 Algorithm is based on decision tree approach to classification. The algorithm was evaluated on training and test sets as shown below. Table 7.1 and 7.3 contain the detailed accuracy by class for the algorithm on each of these datasets, while table 7.2 and 7.4 describe the confusion matrix for the class labels on the same.

Evaluation on training dataset

Correctly classified instances	161	88.4615%
Incorrectly classified instances	21	11.5385%

Table 7.1: Detailed accuracy by class on training set for C4.5 algorithm

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC area	Class
	0.805	0.57	0.912	0.805	0.855	0.902	no
	0.943	0.195	0.868	0.943	0.904	0.867	yes
Weighted avg.	0.885	0.137	0.887	0.885	0.883	0.882	

Table 7.2: Confusion matrix of training set for C4.5 algorithm

a	b	
62	15	a = no
6	99	b = yes

Re-evaluation on test dataset

Correctly classified instances	100	72.9927%
Incorrectly classified instances	37	27.0073%

Table 7.3: Detailed accuracy by class on test set for C4.5 algorithm

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC area	Class
	0.469	0.41	0.909	0.469	0.619	0.911	no
	0.959	0.531	0.673	0.959	0.791	0.911	yes
Weighted avg.	0.73	0.302	0.783	0.73	0.71	0.911	

Table 7.4: Confusion matrix of test set for C4.5 algorithm

a	b	
30	34	a = no
3	70	b = yes

Evaluation of bootstrap data on training and test data is mentioned above respectively. Accuracy is 72.9927%.

Figure 7.1 depicts the area under the curve for the instances classified as relevant online courses by C4.5 classification algorithm. The area under the curve was observed to be **0.9114.** This is reasonably high but holds scope for improvement.

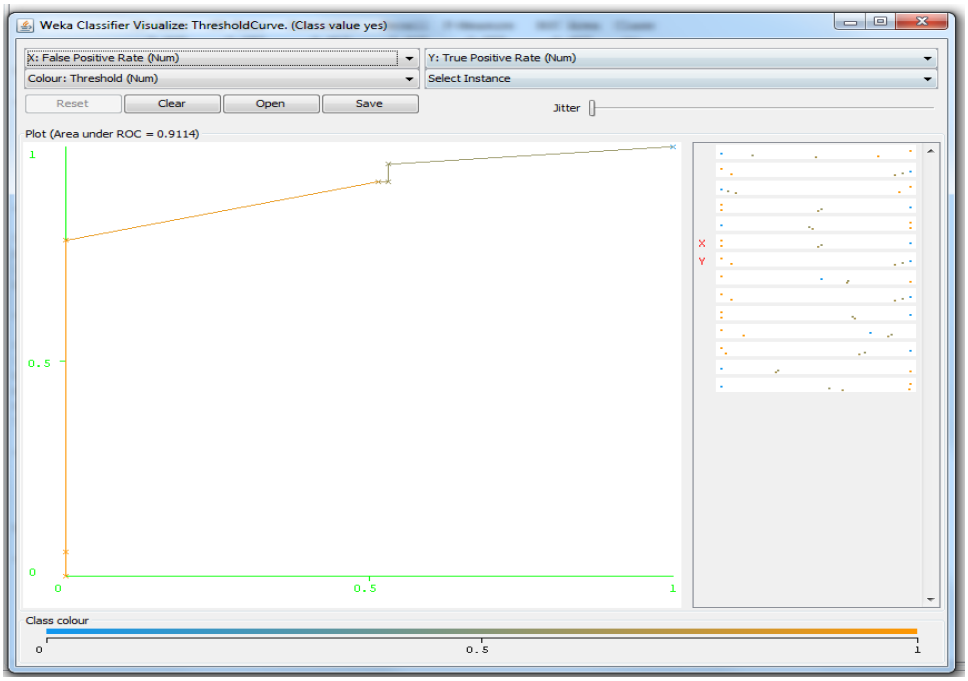

Figure 7.1: ROC curve of C4.5 classification algorithm for the class value 'yes'

Figure 7.2 depicts the area under the curve for the instances classified as non-relevant online courses by C4.5 classification algorithm. The area under the curve was observed to be 0.9114. This is reasonably high but holds scope for improvement.

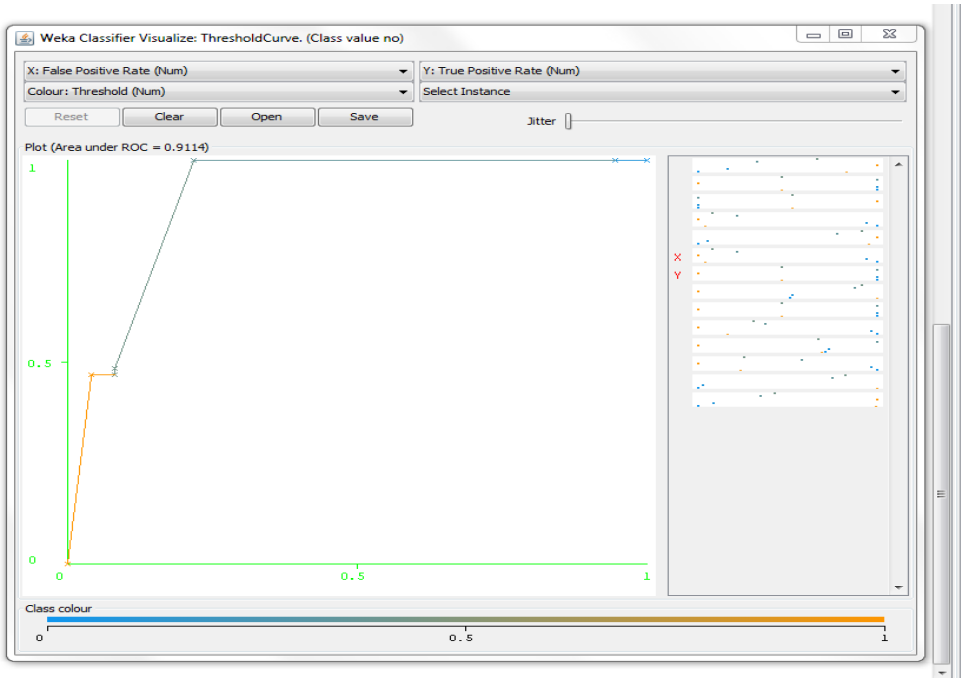

Figure 7.2: ROC curve of C4.5 classification algorithm for the class value 'no'

7.3.2 Logistic Regression Algorithm Evaluation

The Logistic Regression algorithm is based on mathematical approach to classification. The algorithm was evaluated on training and test sets as shown below. Table 7.5 and 7.7 contain the detailed accuracy by class for the algorithm on each of these datasets, while table 7.6 and 7.8 describe the confusion matrix for the class labels on the same.

Evaluation on training dataset

Correctly classified instances	181	99.4505%
Incorrectly classified instances	1	0.5495%

Table 7.5: Detailed accuracy by class on training set for Logistic Regression algorithm

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC area	Class
	0.987	0.0	1	0.987	0.855	0.902	no
	1	0.013	0.991	1	0.904	0.867	yes
Weighted avg.	0.995	0.007	0.995	0.995	0.883	0.882	

Table 7.6: Confusion matrix of training set for Logistic Regression algorithm

a	B	
76	1	a = no
0	105	b = yes

Re-evaluation on test dataset

Correctly classified instances	104	75.9124%
Incorrectly classified instances	33	24.0876%

Table 7.7: Detailed accuracy by class on test set for Logistic Regression algorithm:

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC area	Class
	0.484	0.0	1	0.484	0.653	0.963	no
	0.1	0.516	0.689	1	0.816	0.963	yes
Weighted avg.	0.759	0.275	0.834	0.759	0.739	0.963	

Table 7.8: Confusion matrix of test set for Logistic Regression algorithm

a	B	
31	33	a = no
0	73	b = yes

Evaluation of bootstrap data on training and test data is mentioned above respectively. Accuracy is 75.9124%.

Figure 7.3 depicts the area under the curve for the instances classified as non-relevant online courses by Logistic Regression algorithm. The area under the curve was observed to be 0.9631. It is slightly higher than the area given by C4.5 algorithm i.e 0.9114 thereby performing slightly better.

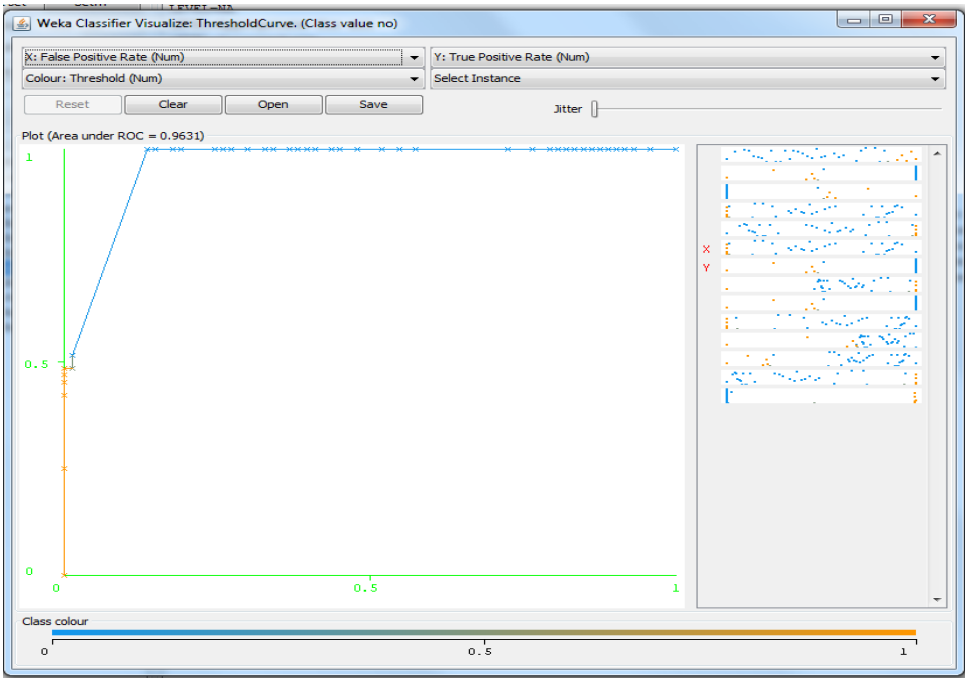

Figure 7.3: ROC curve of Logistic Regression classification algorithm for the class value 'no'

Figure 7.4 depicts the area under the curve for the instances classified as relevant online courses by Logistic Regression algorithm. The area under the curve was observed to be 0.9631. It is slightly higher than the area given by C4.5 algorithm i.e 0.9114 therefore Logistic Regression is better than C4.5.

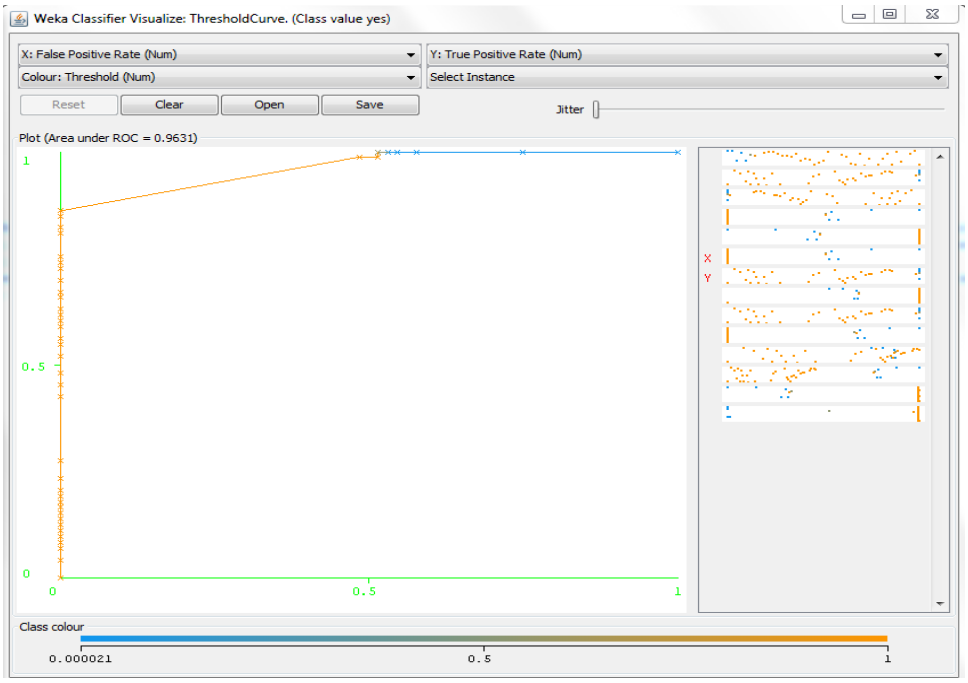

Figure 7.4: ROC curve of Logistic Regression classification algorithm for the class value 'yes'

7.3.3 Naive Bayes Algorithm Evaluation

The Naive Bayes algorithm is based on probabilistic approach to classification. The algorithm was evaluated on training and test sets as shown below. Table 7.9 and 7.11 contain the detailed accuracy by class for the algorithm on each of these datasets, while table 7.10 and 7.12 describe the confusion matrix for the class labels on the same.

Evaluation on training dataset

Correctly classified instances	172	94.5055%
Incorrectly classified instances	10	5.5495%

Table 7.9: Detailed accuracy by class on training set for Naive Bayes algorithm

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC area	Class
	1	0.095	0.855	1	0.939	0.99	no
	0.905	0	1	0.905	0.95	0.983	yes
Weighted avg.	0.945	0.04	0.951	0.945	0.945	0.986	

Table 7.10: Confusion matrix of training set for Naive Bayes algorithm

a	B	
77	0	a = no
10	95	b = yes

Re-evaluation on test dataset

Correctly classified instances	104	88.3212%
Incorrectly classified instances	33	11.6788%

Table 7.11: Detailed accuracy by class on test set for Naive Bayes algorithm

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC area	Class
	1	0.219	0.8	1	0.889	0.957	no
	0.781	0	1	0.781	0.877	0.957	yes
Weighted avg.	0.883	0.102	0.907	0.883	0.883	0.957	

Table 7.12: Confusion matrix of test set for Naive Bayes algorithm

a	B	
64	0	a = no
16	57	b = yes

Evaluation of bootstrap data on training and test data is mentioned above respectively. Accuracy is 88.3212%.

Figure 7.5 depicts the area under the curve for the instances classified as non-relevant online courses by Naive Bayes classification algorithm. The area under the curve was observed to be 0.9571.

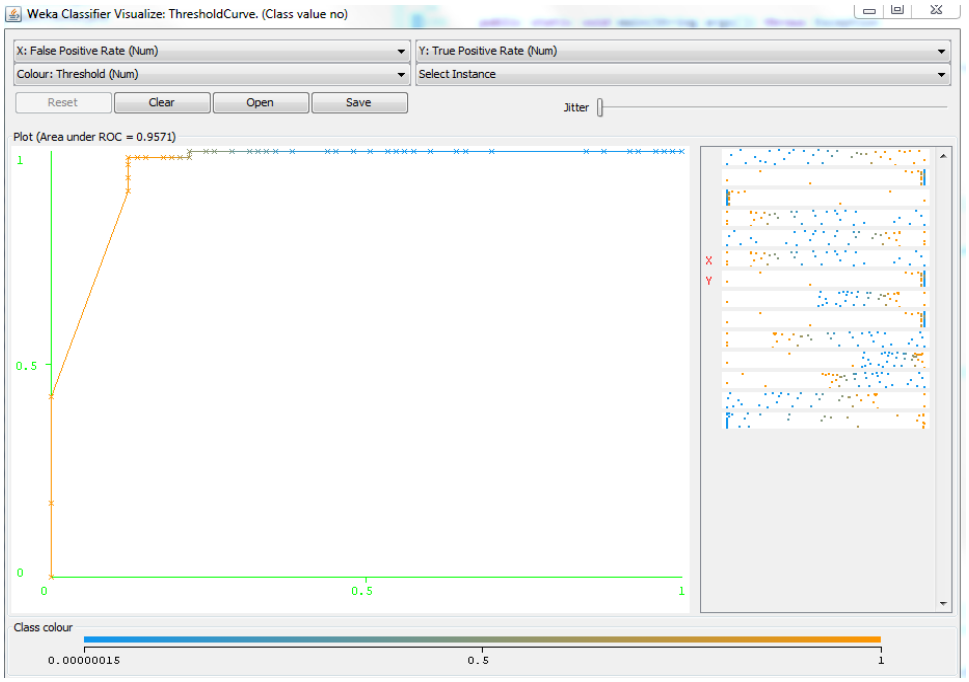

Figure 7.5: ROC curve of Naive Bayes classification algorithm for the class value 'no'

Figure 7.6 depicts the area under the curve for the instances classified as relevant online courses by Naive Bayes classification algorithm. The area under the curve was observed to be 0.9571. It can be observed that the area is slightly higher for Naive Bayes as compared to C4.5.

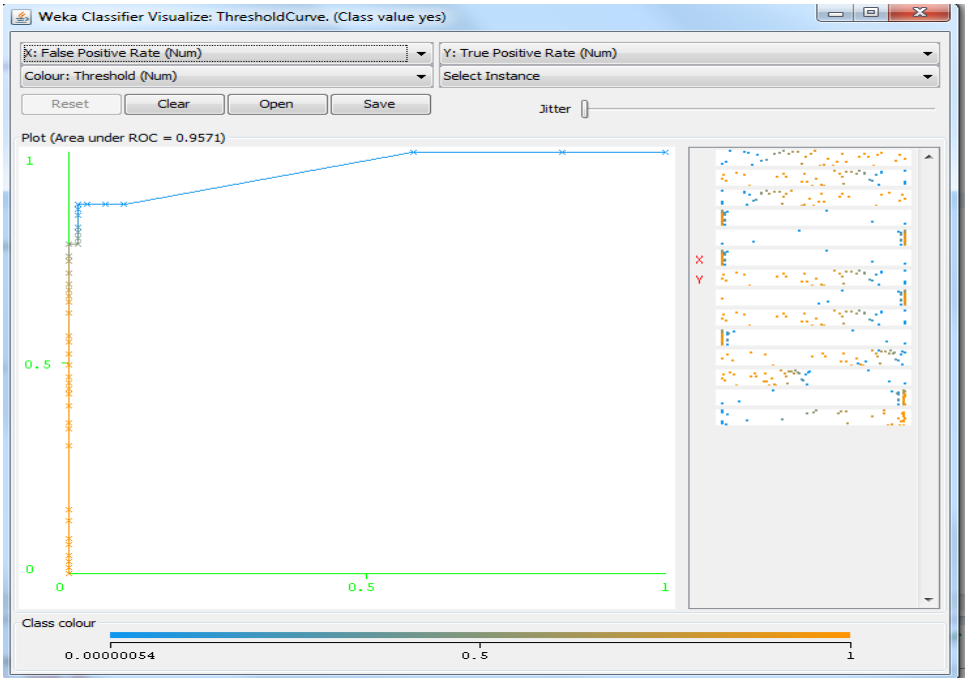

Figure 7.6: ROC curve of Naive Bayes classification algorithm for the class value 'yes'

7.3.4 K Nearest Neighbours Algorithm Evaluation

The K Nearest Neighbours algorithm is based on lazy learning approach to classification. The algorithm was evaluated on training and test sets as shown below. Table 7.13 and 7.15 contain the detailed accuracy by class for the algorithm on each of these datasets, while table 7.14 and 7.16 describe the confusion matrix for the class labels on the same.

Evaluation on training dataset

Correctly classified instances	124	90.5109%
Incorrectly classified instances	13	9.4891%

Table 7.13: Detailed accuracy by class on training set for K Nearest Neighbours algorithm

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC area	Class
	0.984	0.164	0.84	0.984	0.906	0.939	no
	0.836	0.016	0.984	0.836	0.904	0.939	yes
Weighted avg.	0.905	0.085	0.917	0.905	0.905	0.939	

Table 7.14: Confusion matrix of training set for K Nearest Neighbours algorithm

a	B	
63	1	a = no
12	61	b = yes

Re-evaluation on test dataset

Correctly classified instances	121	88.3212%
Incorrectly classified instances	16	11.6788%

Table 7.15: Detailed accuracy by class on test set for K Nearest Neighbours algorithm

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC area	Class
	1	0.219	0.8	1	0.889	0.957	no
	0.781	0	1	0.781	0.877	0.957	yes
Weighted avg.	0.883	0.102	0.907	0.883	0.883	0.957	

Table 7.16: Confusion matrix of test set for K Nearest Neighbours algorithm

a	B	
64	0	a = no
16	57	b = yes

Evaluation of bootstrap data on training and test data is mentioned above respectively. Accuracy is 88.3212%.

Figure 7.7 depicts the area under the curve for the instances classified as non-relevant online courses by K Nearest neighbour classification algorithm. The area under the curve was observed to be 0.9389. The area is slightly lower as compared to the area of Naive Bayes.

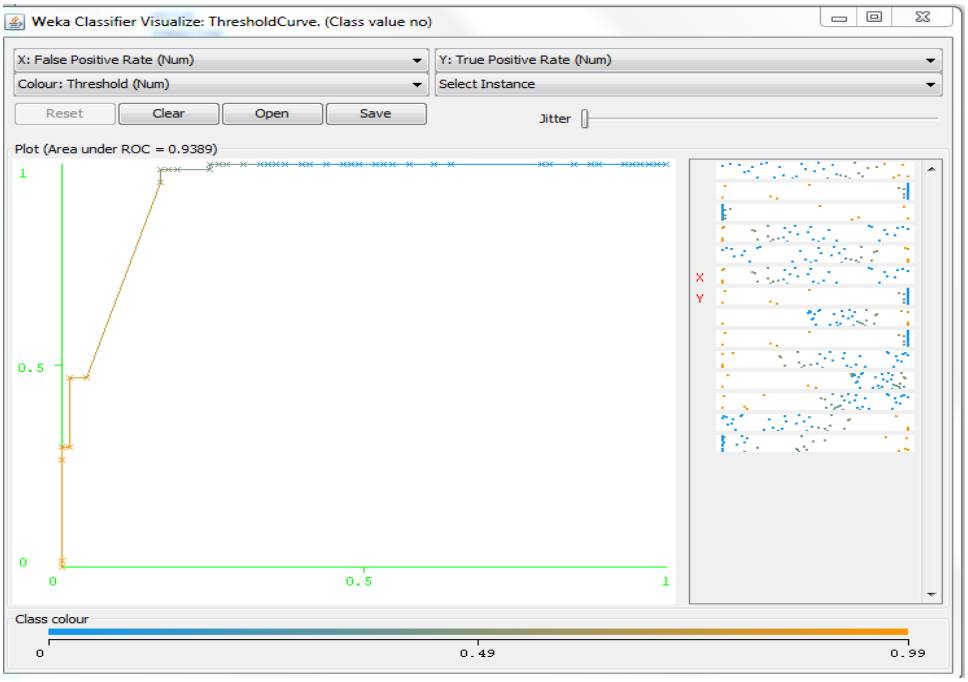

Figure 7.7: ROC curve of K Nearest neighbour classification algorithm for the class value 'no'

Figure 7.8 depicts the area under the curve for the instances classified as relevant online courses by K Nearest neighbour classification algorithm. The area under the curve was observed to be 0.9389. The area is slightly lower as compared to the area of Naive Bayes.

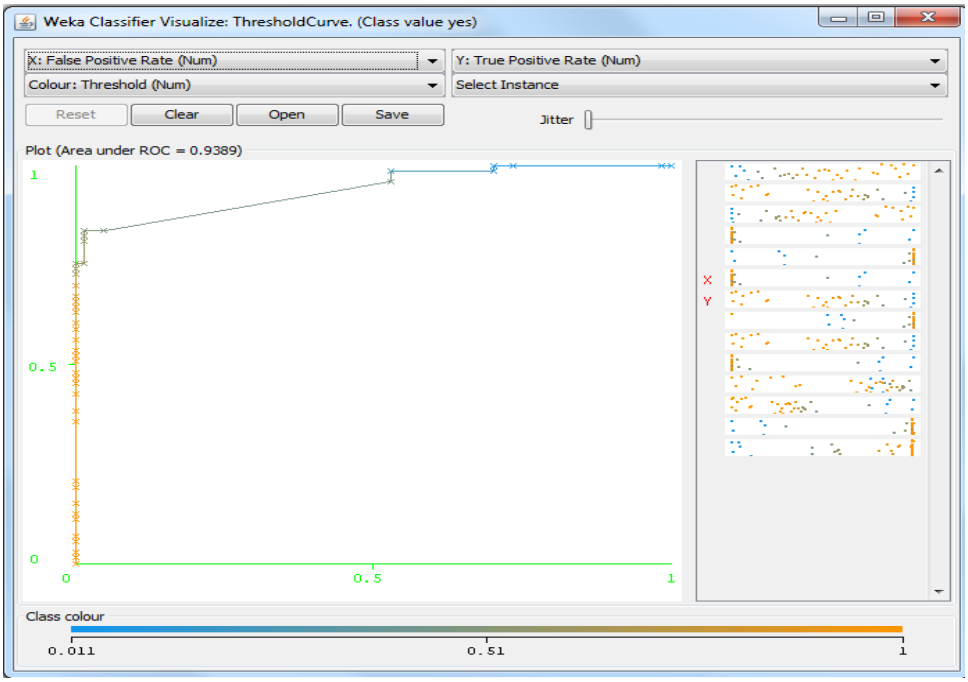

Figure 7.8: ROC curve of K Nearest classification algorithm for the class value 'yes'

7.3.5 Random Forest Algorithm Evaluation

The Random Forest algorithm is based on ensemble approach to classification. The algorithm was evaluated on training and test sets as shown below. Table 7.17 and 7.19 contain the detailed accuracy by class for the algorithm on each of these datasets, while table 7.18 and 7.20 describe the confusion matrix for the class labels on the same.

Evaluation on training dataset

Correctly classified instances	182	100%
Incorrectly classified instances	0	0%

Table 7.17: Detailed accuracy by class on training set for Random Forest algorithm

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC area	Class
	1	0	1	1	1	1	no
	1	0	1	1	1	1	yes
Weighted avg.	1	0	1	1	1	1	

Table 7.18: Confusion matrix of training set for Random Forest algorithm

a	B	
77	0	a = no
0	105	b = yes

Re-evaluation on test dataset

Correctly classified instances	127	92.7007%
Incorrectly classified instances	10	7.2993%

Table 7.19: Detailed accuracy by class on test set for Random Forest algorithm:

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC area	Class
	0.984	0.123	0.875	1	0.984	0.926	no
	0.877	0.016	0.985	0.781	0.877	0.928	yes
Weighted avg.	0.927	0.066	0.933	0.883	0.927	0.927	

Table 7.20: Confusion matrix of test set for Random Forest algorithm

a	B	
63	1	a = no
9	64	b = yes

Evaluation of bootstrap data on training and test data respectively. Accuracy is 92.7007%.

Figure 7.9 depicts the area under the curve for the instances classified as non-relevant online courses by Random Forest classification algorithm. The area under the curve was observed to be 0.9605. This value is recorded as the highest value as compared to the area value observed in other algorithms.

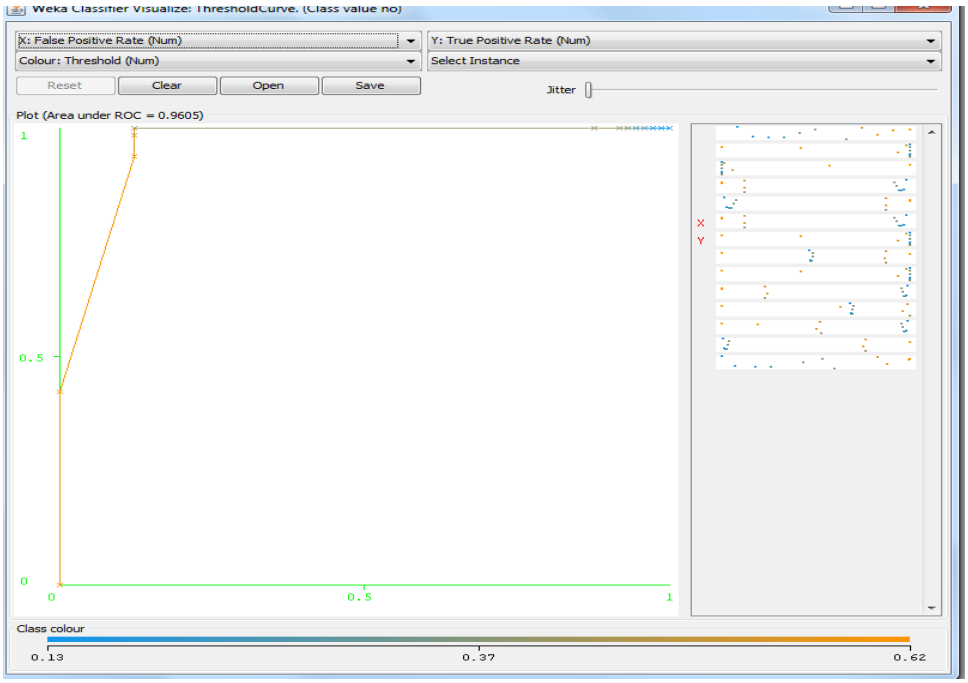

Figure 7.9: ROC curve of Random Forest classification algorithm for the class value 'no'

Figure 7.10 depicts the area under the curve for the instances classified as non-relevant online courses by Random Forest classification algorithm. The area under the curve was observed to be 0.9549. This value is recorded as the highest value as compared to the area value observed in other algorithms.

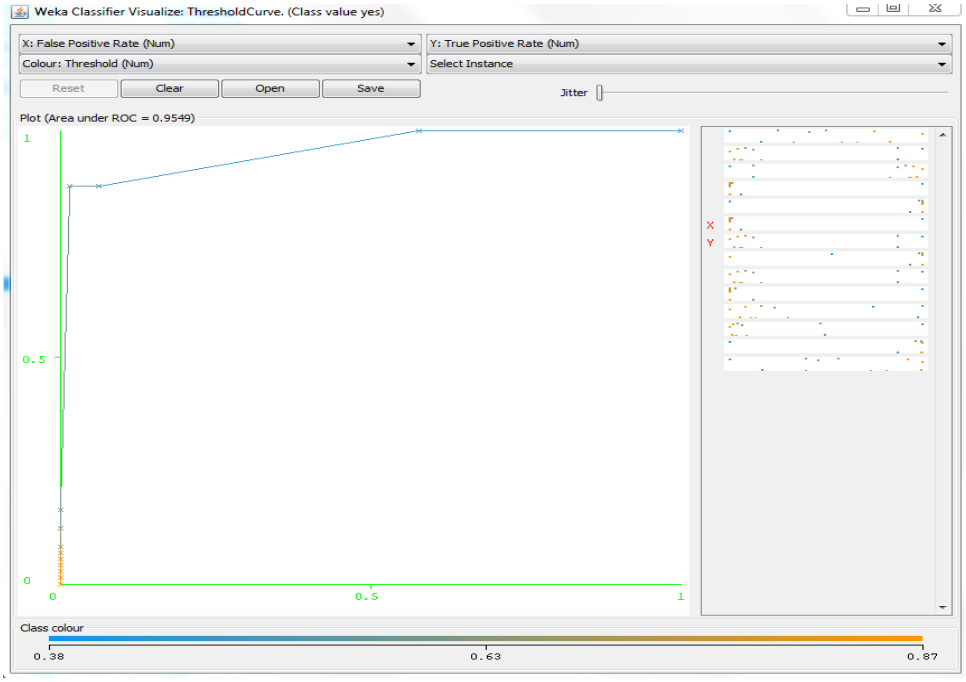

Figure 7.10: ROC curve of Random Forest classification algorithm for the class value 'yes'

7.4 Inference from the Result

The experimental analysis presented above has been classified under various heads such as precision, recall, F Measure and others as defined earlier. Under each of these parameters, it can be seen that the decision-tree based C 4.5 algorithm is the weakest of all. The probabilistic Naive Bayes and the lazy learner K Nearest Neighbour have identical results for the dataset used and the mathematical Logistic Regression is better than both of them. However, the ensemble based Random Forest algorithm has outperformed all the other algorithms.

It is also noticed that for all the algorithms mentioned above, the training set precision is sufficiently high but the test set precision is lesser, with the gap narrowing in respective order between training and test-set precision.

In terms of time and computational resources, all the algorithms had nearly identical performance and since the focus of this research is on the precision of the algorithms, this factor is ignored during the differentiation and comparison of the various algorithms.

Various methods such as random splitting of train and test data, cross-validation and others were considered however they did not substantially order the relative performance of the algorithms or the type of predictor used. This shows us that the

advantage that any one algorithm has over another is not merely for a particular type of test-set, but across all forms of testing and hence it would be safe to conclude a relative order of superiority as proposed. The standard split of 66% training data and 33% test data was adopted for presentation of results.

It is also seen that for the ROC curves, the area under the curve is sufficiently high for almost all algorithms for higher threshold values whereas for lower threshold values, there is a marked difference. This is a clear indication of the behaviour of the algorithm as the dataset size increases, because there will be a natural increase in the number of true positives and corresponding false positives. These curves therefore help us to predict the scalability of the algorithms and the resulting compromise on precision.

It is clear from the experimental analysis that the Random Forest algorithm outperforms all others not only in training set evaluation metrics, but more importantly in test set evaluation metrics. It can be inferred that this algorithms is more capable of dealing with hitherto unseen data type samples and the areas under ROC curves show that for Random Forest algorithm, varying of threshold intensity does not lower the performance significantly as is done in other algorithms and hence this algorithm works better with large datasets as well. Table 7.1 gives the comparison of all the five algorithms.

Table 7.21: Accuracies of the algorithms

Classification Algorithm	**Accuracy (%)**
C4.5	72.9
Logistic Regression	75.9
Naive Bayes	88.3
K Nearest Neighbour	88.3
Random Forest	92.7

Hence Random Forest Classification algorithm has been used in all further stages of development for prediction of class labels for a large data corpus.

7.5 Summary

The seventh chapter describes in detail the experimental results and its analysis of all the five classification algorithms – C4.5, Naïve Byes, K Nearest Neighbor, Logistic Regression and Random Forest. The evaluation metrics like confusion matrix, ROC curves and accuracies of the respective algorithms have been shown. The best of the algorithm has been chosen for the developmental purposes.