Assignment 4: Feature SelectionII

Shaun Pritchard

Florida Atlantic University

CAP 6778

October -27-2021

M. Khoshgoftaar

Assignment 4 Feature SelectionII

This experimentation implements the use of feature selection techniques using NaiveBayes and 5-Nearest Neighbor(KNN) learners. Experimenting with 5, 6, 7, 8, 9, 10, 20, 50, 100, and 200 selected features to discover patterns in terms of FPR, FNR, and AUC. This experiment will use the Information Gain(IG), Chi-Squared(X2) , Gain Ratio (GR), Symmetric Uncertainty (SUA), ReliefF (RF set to False), and ReliefF-W (RFW set to true) feature selection rankers . feature attributes extracted from feature selections will consist of 132 experiment instances being conducted to yield the results.

Note: As a result of the size and amount of data in the appendices, the appendices are included as a separate document.

Part I

This section I will report the patterns, including the optimal number of features in terms of AUC, the evidence that led you to conclude this, and the resulting performance (in terms of FPR, FNR, and AUC) when this number of features is used along with the performance of the classifiers on the full set of attributes for comparison. Also, how these changes are influenced by the choice of classifier and ranker.

Part II

In this section I will 3evlauatre the results of the given experiments and analyze the results in comparison with Assignment ones classifier as follows:

I. I compare the top-performing feature subsets discovered in the previous experiment Part I with each other and with the features selected by the C4.5 classifier in Part 1 of Assignment 1.

II. I evaluate each classifier used in Part 1 of this assignment, and when choosing 6 features with the best feature ranker in terms of AUC.

III. I evaluate which six features were chosen by these best feature rankers in terms of AUC.?

IV. I then compare the six-feature feature subsets chosen by each of these classifier-ranker pairs (e.g., NB-<top ranker with NB> and 5NN-<top ranker with 5NN>) with the features chosen by the decision tree built in Part 1 of Assignment 1.

V. Evaluate the overlap and how many features are in common between the three scenarios.

VI. Compare the use of separate rankers and classifiers (the procedure in this assignment) with the C4.5 decision tree (which has embedded feature selection).

Feature Selection Analysis:
Presented here is the analysis of each learner according to the given feature selection method compared to the number of features selected from 5,6,7,8,9,10,20,50,100, and 200) individually for each learner. The tables display the classification from which the best results for the subset section per learner.

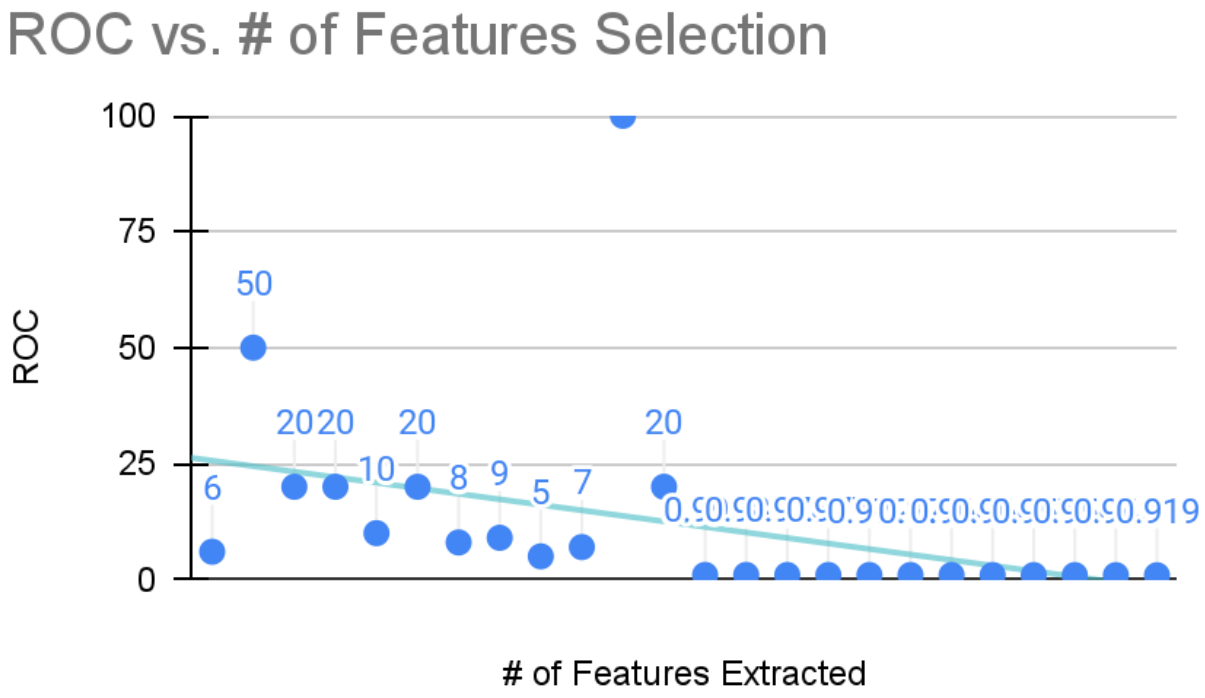
Part I

A summary of the analysis:

This analysis was based on a categorization of key features in the data. Based on the Area Under the ROC curve, Naive Bayes with ReliefF-W set to true performed the best with ROC of 0.992. Among all the experiments, it was the best performer in regards to ROC.

Top AUC - Performance based on Highest Area Under the ROC curve %									
# of Features Extracted	Classifier / Learners	Type I	Type II	Type I Error rate %	Type II Error rate %	Misclassification Rate %	Correctly Classified %	ROC	RMSE
6	NaiveBayes +(RFW-t)	2	3	0.13	0.028	5.20%	94.70%	0.992	0.1944
50	KNN (5) +(RFW-t)	0	8	0.348	0	8.40%	91.50%	0.989	0.2295
20	NaiveBayes +(X2)	8	1	0.043	0.111	9.40%	90.50%	0.978	0.2671
20	NaiveBayes +(SUA)	7	1	0.043	0.097	8.40%	91.50%	0.977	0.2363
10	KNN (5) +(X2)	5	7	0.304	0.069	12.60%	87.30%	0.972	0.2805
20	NaiveBayes +(GR)	4	2	0.087	0.056	6.30%	93.60%	0.97	0.2357
8	NaiveBayes +(IG)	6	1	0.043	0.083	7.30%	92.63%	0.968	0.2652
9	NaiveBayes +(RF-f)	8	2	0.087	0.111	10.50%	89.40%	0.966	0.2975
5	KNN (5) +(SAU)	2	6	0.261	0.028	8.40%	91.50%	0.957	0.258
7	KNN (5)	4	2	0.087	0.056	6.30%	93.60%	0.956	0.2386

	+(IG)								
100	KNN (5) +(GR)	3	4	0.174	0.042	7.3	93%	0.942	0.2506
20	KNN (5) +(RF-f)	0	13	0.565	0	13.60%	86.30%	0.919	0.2956



Comparing the results below in table-1-2 with no feature extraction methods applied, we can see that the performance using feature extraction is about three times better than that compared to just classification learners alone using the full feature set.

Table-1-2: Naive bayes and KNN with no feature selection:

Classifier / Learners	Type I	Type II	Type I Error rate %	Type II Error rate %	Misclassification Rate %	Correctly Classified %	ROC	RMSE
NaiveBayes	11	5	21.70%	15.30%	16.84%	83%	0.844	0.4115
KNN (5)	7	9	39.10%	9.70%	16.8	83%	0.863	0.3573

Following are the overall best performing learners based on feature selection methods with the lowest Type II FPR and FNR. With 20 features selected, the Naive Bayes classifier with Symmetric Uncertainty (SUA) produced the best results. The naive Bayes with Gain Ratio (GR) had the lowest type I and type II errors. With only 7 features selected, KNN had the lowest Type II error rate of 0.00%. As the misclassification rate was 3.10% with an ROC of 0.965, this could point to overfitting.

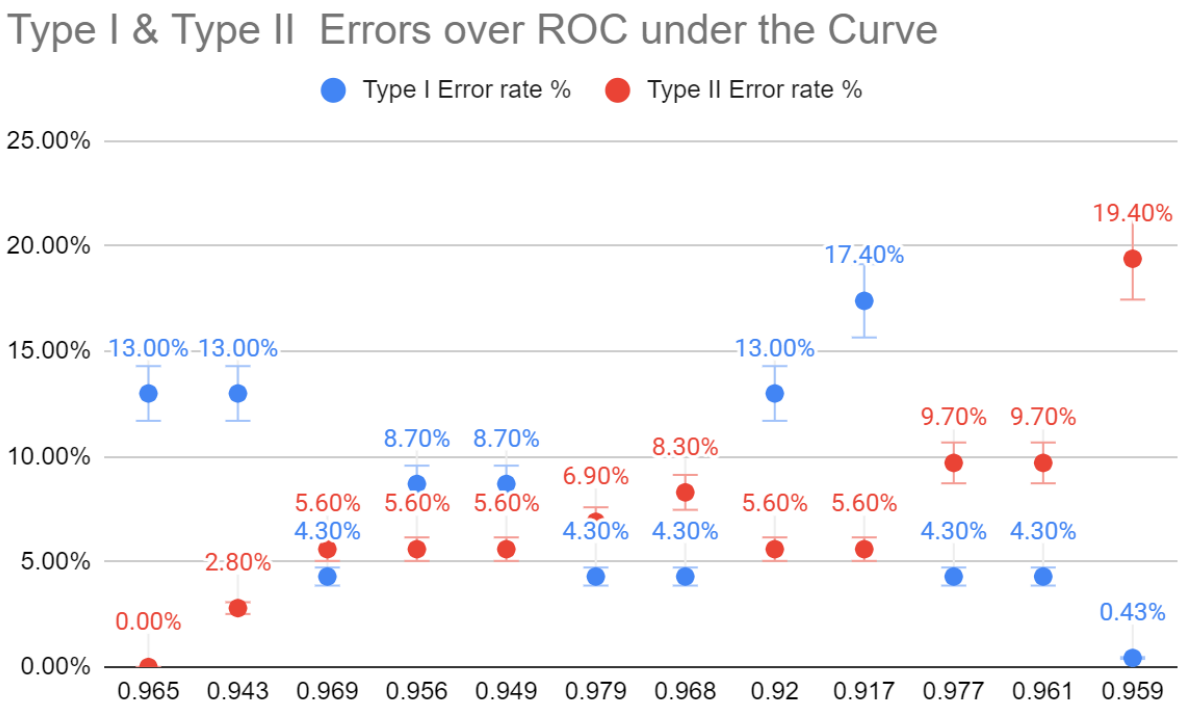
Based on ReliefF (ReliefFAttributeEval with the weightByDistance parameter set to False), naive Bayes yielded a TypeII error percentage of 15.70 with a ROC of 0.959. The KNN with ReliefF-W (RFW set to true) feature selection ranker with only 7 features selected correctly classified the features.

In terms of top performers, KNN with Chi Square at 200 and only 200 features selected seems more adequate and balanced. Due to the amount of data analyzed with this learner and feature selection method, a Type II error percentage of 2.8% with an ROC of 0.943 and only 2 Type I errors would be a more significant identification of overall performance.

Table 1-3, Best performing overall learners based on feature ranking

Best preforming overall learners based on feature ranking									
# of Features Extracted	Classifier / Learners	Type I	Type II	Type I Error rate %	Type II Error rate %	Misclassification Rate %	Correctly Classified %	ROC	RMSE
8	NaiveBayes +(IG)	6	1	4.30%	8.30%	7.30%	92.63%	0.968	0.2652
9	NaiveBayes +(X2)	7	1	4.30%	9.70%	8.40%	91.50%	0.961	0.2886
20	NaiveBayes +(SUA)	7	1	4.30%	9.70%	8.40%	91.50%	0.977	0.2363
20	NaiveBayes +(RFW-t)	5	1	4.30%	6.90%	6.30%	93.60%	0.979	0.2327
50	NaiveBayes +(GR)	4	1	4.30%	5.60%	5.20%	94.70%	0.969	0.2181
100	NaiveBayes +(RF-f)	14	1	0.43%	19.40%	15.70%	84%	0.959	0.3972
7	KNN (5) +(IG)	4	2	8.70%	5.60%	6.30%	93.60%	0.956	0.2386
50	KNN (5) +(SAU)	4	2	8.70%	56.00%	6.30%	93.60%	0.949	0.2367
7	KNN (5) +(RFW-t)	0	3	13.00%	0.00%	3.10%	96.80%	0.965	0.1959

200	KNN (5) +(X2)	2	3	13.00%	2.80%	5.20%	94.70%	0.943	0.2403
200	KNN (5) +(GR)	4	3	13.00%	5.60%	7.30%	92.60%	0.92	0.2645
5	KNN (5) +(RF-f)	4	4	17.40%	5.60%	8.40%	91.50%	0.917	0.2521



In the following **table 1-4**, we present a comparison and contrast of the best performing feature selection techniques compared to the initial analysis, using all available features.

Evaluation for Assignment 4 - Feature selection ALL features								
Classifier / Learners	**Type I**	**Type II**	**Type I Error rate %**	**Type II Error rate %**	**Misclassification Rate %**	**Correctly Classified %**	**ROC**	**RMSE**
NaiveBayes +(IG)	11	5	0.217	0.153	16.80%	83.10%	0.844	0.4115
KNN (5) +(IG)	7	9	0.391	0.097	16.80%	83.10%	0.863	0.3573
NaiveBayes +(X2)	9	1	0.043	0.125	10.50%	89.47%	0.967	0.3228
KNN (5) +(X2)	2	3	0.13	0.028	5.20%	94.70%	0.918	0.2497
NaiveBayes +(GR)	11	5	0.217	0.153	16.80%	83.10%	0.848	0.4115

KNN (5) +(GR)	7	9	0.391	0.097	16.80%	83.10%	0.863	0.3573
NaiveBayes +(SUA)	11	5	0.217	0.153	16.80%	83.10%	0.848	0.4115
KNN (5) +(SUA)	7	9	0.391	0.097	16.80%	83.10%	0.863	0.3573
NaiveBayes +(RF-f)	11	5	0.217	0.153	16.80%	83.10%	0.849	0.4115
KNN (5) +(RF-f)	7	9	0.391	0.097	16.80%	83.10%	0.863	0.3573
NaiveBayes +(RFW-t)	11	5	0.217	0.153	16.80%	83.10%	0.849	0.4115
KNN (5) +(RFW-t)	7	9	0.391	0.097	16.80%	83.10%	0.863	0.3573

Table 1-4

Part II

I. In Part 1 of Assignment 1, I compare the top-performing feature subsets discovered in the previous experiment Part I with each other and with those selected by the C4.5 classifier. According to the results in **table 2-1,** feature selection methods outperform a standard C.45(J48 Weka) tree learner with a cost-sensitive ratio adjusted to altering values. I found that the higher values caused greater type II percentage errors, resulting in a smaller ROC.

Best preforming overall learners compared to Assignment 1 J48 with cost sensitive ratio									
# of FeaturesSelection	**Classifier / Learners**	**Type I**	**Type II**	**Type I Error rate %**	**Type II Error rate %**	**Misclassification Rate %**	**Correctly Classified %**	**ROC**	**RMSE**
7	KNN (5) +(RFW-t)	0	3	13.00%	0.00%	3.10%	96.80%	0.965	0.1959
200	KNN (5) +(X2)	2	3	13.00%	2.80%	5.20%	94.70%	0.943	0.2403
50	NaiveBayes +(GR)	4	1	4.30%	5.60%	5.20%	94.70%	0.969	0.2181
7	KNN (5) +(IG)	4	2	8.70%	5.60%	6.30%	93.60%	0.956	0.2386
50	KNN (5) +(SAU)	4	2	8.70%	5.60%	6.30%	93.60%	0.949	0.2367
200	KNN (5) +(GR)	4	3	13.00%	5.60%	7.30%	92.60%	0.92	0.2645
5	KNN (5) +(RF-f)	4	4	17.40%	5.60%	8.40%	91.50%	0.917	0.2521

20	NaiveBayes +(RFW-t)	5	1	4.30%	6.90%	6.30%	93.60%	0.979	0.2327
	J48 + CS(0.5)-np	6	7	30.40%	8.30%	1360.00%	86.3	0.831	0.3572
8	NaiveBayes +(IG)	6	1	4.30%	8.30%	7.30%	92.63%	0.968	0.2652
20	NaiveBayes +(SUA)	7	1	4.30%	9.70%	8.40%	91.50%	0.977	0.2363
9	NaiveBayes +(X2)	7	1	4.30%	9.70%	8.40%	91.50%	0.961	0.2886
	J48 + CS(2)-np	9	16	69.60%	12.50%	2600.00%	74	0.591	0.4954
	J48 + CS(1)-np	12	9	39.10%	16.70%	2200.00%	78	0.732	0.4496
100	NaiveBayes +(RF-f)	14	1	0.43%	19.40%	15.70%	84%	0.959	0.3972

Table 2-1

II. When choosing 6 features, the best feature ranker in terms of AUC is Naive Bayes with ReliefF-W (ReliefFAttributeEval with the weightByDistance parameter set to True) as displayed below in **Table 2-2**.

Best preforming overall learners based on feature ranking									
# of FeaturesSelection	**Classifier / Learners**	**Type I**	**Type II**	**Type I Error rate %**	**Type II Error rate %**	**Misclassification Rate %**	**Correctly Classified %**	**ROC**	**RMSE**
6	NaiveBayes +(RFW-t)	2	3	0.13	0.028	5.20%	94.70%	0.992	0.1944
6	NaiveBayes +(SUA)	7	2	0.087	0.097	9.40%	90.50%	0.969	0.2755
6	KNN (5) +(RFW-t)	0	4	0.17%	0.00%	4.20%	95.70%	0.962	0.2073
6	NaiveBayes +(RF-f)	9	3	0.13	0.125	12.60%	87.20%	0.961	0.2994
6	NaiveBayes +(IG)	11	1	0.043	0.153	12.60%	87.36%	0.955	0.3546
6	NaiveBayes +(X2)	10	3	0.13	0.139	13.60%	86.30%	0.953	0.3167
6	KNN (5) +(IG)	7	2	0.087	0.097	9.40%	90.50%	0.938	0.2571
6	KNN (5) +(X2)	3	7	0.304	0.042	10.50%	89.40%	0.918	0.2901
6	KNN (5) +(SAU)	3	6	0.261	0.042	9.40%	90.50%	0.918	0.2856
6	KNN (5) +(RF-f)	1	12	0.522	0.14%	13.60%	86.30%	0.912	0.295
6	KNN (5)	1	16	0.696	0.014	17.80%	82.00%	0.853	0.345

	+(GR)								
6	NaiveBayes +(GR)	12	10	0.435	0.167	23.10%	76.80%	0.817	0.4044

Table 2-2

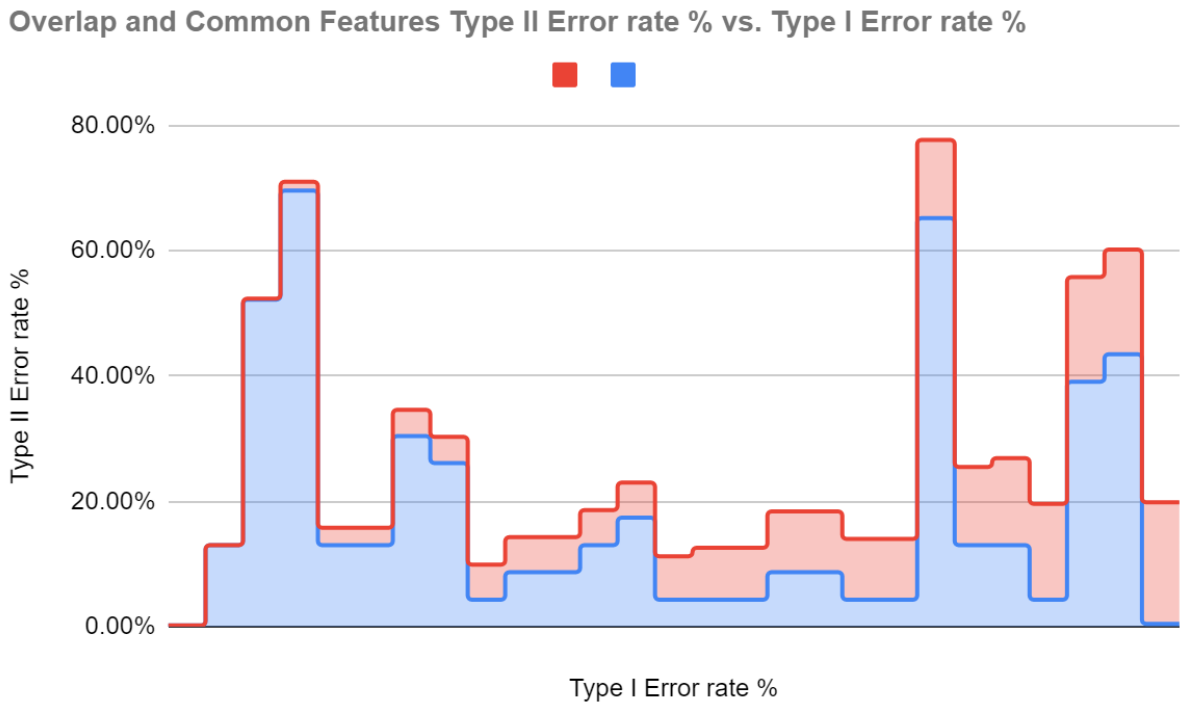
VII. I then compare the six-feature feature subsets chosen by each of these classifier-ranker pairs (e.g., NB-<top ranker with NB> and 5NN-<top ranker with 5NN>) with the features chosen by the decision tree built in Part 1 of Assignment 1. The results show some significance of the C.45(J48 Weka) tree learner compared to Naive Bayes and KNN methods for feature selection. We find that the C.45(J48 Weka) tree learner with cost sensitive ratio of 0.5 out performs Naive Bayes (SAU) KNN(IG). The C.45(J48 Weka) tree learner with cost sensitive ratio of 1 out performs Naive bayes (RF-f), Naive Bayes(X2), and Naive Bayes (IG) C.45(J48 Weka) tree learner with cost sensitive ratio of 0.5 out performs Naive bayes(GR).

Best preforming overall learners based on feature ranking subsets of 6									
# of FeaturesSelection	**Classifier / Learners**	**Type I**	**Type II**	**Type I Error rate %**	**Type II Error rate %**	**Misclassification Rate %**	**Correctly Classified %**	**ROC**	**RMSE**
6	KNN (5) +(RFW-t)	0	4	0.17%	0.00%	4.20%	95.70%	0.962	0.2073
6	KNN (5) +(RF-f)	1	12	0.522	0.14%	13.60%	86.30%	0.912	0.295
6	KNN (5) +(GR)	1	16	0.696	1.40%	17.80%	82.00%	0.853	0.345
6	NaiveBayes +(RFW-t)	2	3	0.13	2.80%	5.20%	94.70%	0.992	0.1944
6	KNN (5) +(X2)	3	7	0.304	4.20%	10.50%	89.40%	0.918	0.2901
6	KNN (5) +(SAU)	3	6	0.261	4.20%	9.40%	90.50%	0.918	0.2856
	J48 + CS(0.5)-np	6	7	30.40%	8.30%	1360.00%	86.3	0.831	0.3572
6	NaiveBayes +(SUA)	7	2	0.087	9.70%	9.40%	90.50%	0.969	0.2755
6	KNN (5) +(IG)	7	2	0.087	9.70%	9.40%	90.50%	0.938	0.2571
	J48 + CS(2)-np	9	16	69.60%	12.50%	2600.00%	74	0.591	0.4954
6	NaiveBayes +(RF-f)	9	3	0.13	12.50%	12.60%	87.20%	0.961	0.2994
6	NaiveBayes +(X2)	10	3	0.13	13.90%	13.60%	86.30%	0.953	0.3167
6	NaiveBayes +(IG)	11	1	0.043	15.30%	12.60%	87.36%	0.955	0.3546
	J48 +	12	9	39.10%	16.70%	2200.00%	78	0.732	0.4496

	CS(1)-np								
6	NaiveBayes +(GR)	12	10	0.435	16.70%	23.10%	76.80%	0.817	0.4044

Table 2-3

VIII. When evaluating the overlap and how many features are in common between the three scenarios the results show 10 learners with different feature selection methods and subsets have common factors with the values for error rates, percentage ROC and RMSE values as follows; NaiveBayes +(RFW-t) with KNN (5) +(X2) with 6 subsets each, KNN (5) +(X2) with KNN (5) +(SAU) with 6 subsets each, KNN (5) +(IG) with KNN (5) +(SAU) 7 and 50 subsets, NaiveBayes +(SUA) with KNN (5) +(IG) with 6 subsets, and NaiveBayes +(SUA) with NaiveBayes +(X2) with 20 and 9 subsets. Alos the results show that NaiveBayes +(RFW-t) has many features overlapping with J48 + CS(0.5)-np as follows in **Table 2-4 and chart 2-1** below**.**



Best preforming overall learners based on feature ranking subsets of 6									
# of FeaturesS	**Classifier / Learners**	**Type I**	**Type II**	**Type I Error rate**	**Type II Error rate**	**Misclassification**	**Correctly Classified**	**ROC**	**RMSE**

election				%	%	Rate %	%		
6	KNN (5) +(RFW-t)	0	4	0.17%	0.00%	4.20%	95.70%	0.962	0.2073
7	KNN (5) +(RFW-t)	0	3	13.00%	0.00%	3.10%	96.80%	0.965	0.1959
6	KNN (5) +(RF-f)	1	12	0.522	0.14%	13.60%	86.30%	0.912	0.295
6	KNN (5) +(GR)	1	16	0.696	1.40%	17.80%	82.00%	0.853	0.345
6	NaiveBayes +(RFW-t)	2	3	0.13	2.80%	5.20%	94.70%	0.992	0.1944
200	KNN (5) +(X2)	2	3	13.00%	2.80%	5.20%	94.70%	0.943	0.2403
6	KNN (5) +(X2)	3	7	0.304	4.20%	10.50%	89.40%	0.918	0.2901
6	KNN (5) +(SAU)	3	6	0.261	4.20%	9.40%	90.50%	0.918	0.2856
50	NaiveBayes +(GR)	4	1	4.30%	5.60%	5.20%	94.70%	0.969	0.2181
7	KNN (5) +(IG)	4	2	8.70%	5.60%	6.30%	93.60%	0.956	0.2386
50	KNN (5) +(SAU)	4	2	8.70%	5.60%	6.30%	93.60%	0.949	0.2367
200	KNN (5) +(GR)	4	3	13.00%	5.60%	7.30%	92.60%	0.92	0.2645
5	KNN (5) +(RF-f)	4	4	17.40%	5.60%	8.40%	91.50%	0.917	0.2521
20	NaiveBayes +(RFW-t)	5	1	4.30%	6.90%	6.30%	93.60%	0.979	0.2327
	J48 + CS(0.5)-np	6	7	4.30%	8.30%	1360.00%	86.30%	0.831	0.3572
8	NaiveBayes +(IG)	6	1	4.30%	8.30%	7.30%	92.63%	0.968	0.2652
6	NaiveBayes +(SUA)	7	2	0.087	9.70%	9.40%	90.50%	0.969	0.2755
6	KNN (5) +(IG)	7	2	0.087	9.70%	9.40%	90.50%	0.938	0.2571
20	NaiveBayes +(SUA)	7	1	4.30%	9.70%	8.40%	91.50%	0.977	0.2363
9	NaiveBayes +(X2)	7	1	4.30%	9.70%	8.40%	91.50%	0.961	0.2886
	J48 + CS(2)-np	9	16	65.20%	12.50%	26.00%	74.00%	0.591	0.4954
6	NaiveBayes +(RF-f)	9	3	0.13	12.50%	12.60%	87.20%	0.961	0.2994
6	NaiveBayes +(X2)	10	3	0.13	13.90%	13.60%	86.30%	0.953	0.3167
6	NaiveBayes	11	1	0.043	15.30%	12.60%	87.36%	0.955	0.3546

	+(IG)								
	J48 + CS(1)-np	12	9	39.10%	16.70%	22.00%	78.00%	0.732	0.4496
6	NaiveBayes +(GR)	12	10	0.435	16.70%	23.10%	76.80%	0.817	0.4044
100	NaiveBayes +(RF-f)	14	1	0.43%	19.40%	15.70%	84%	0.959	0.3972

IX. Below is a comparison of the separate rankers and classifiers with the C4.5 decision tree (which has embedded feature selection). The J48 + CS(0.5)-np, J48 + CS(0.5)-np,J48 + CS(0.5)-npr represent the C.45 learner and the Knn(5) and Naive Bayes represent the models ran without feature selection below.

For the subsets of 5 selected features compared with the C.45 the data shows KNN (5) +(RF-f) has the best Area under the ROC curve and KNN (5) +(RF-f) has the least Type II errors.

Evaluation for Assignment 4 - 5 Feature subsets								
Classifier / Learners	**Type I**	**Type II**	**Type I Error rate %**	**Type II Error rate %**	**Misclassification Rate %**	**Correctly Classified %**	**ROC**	**RMSE**
KNN (5) +(RF-f)	1	8	0.348	0.014	9.40%	90.50%	0.908	0.2879
KNN (5) +(X2)	2	3	0.13	0.028	5.20%	94.70%	0.943	0.2403
KNN (5) +(GR)	4	3	0.13	0.056	7.30%	92.60%	0.92	0.2645
KNN (5) +(RFW-t)	4	6	0.261	0.056	10.50%	89.40%	0.927	0.276
KNN (5) +(IG)	5	4	0.174	0.069	9.40%	90.50%	0.927	0.2902
J48 + CS(0.5)-np	6	7	30.40%	8.30%	14.00%	86.3	0.831	0.3572
KNN (5)	7	9	39.10%	9.70%	16.8	83%	0.863	0.3573
J48 + CS(2)-np	9	16	69.60%	12.50%	26.00%	74.00%	0.591	0.4954
NaiveBayes +(RFW-t)	9	3	0.13	0.125	12.60%	.87.3	0.94	0.3519
NaiveBayes +(IG)	9	2	0.087	0.125	11.50%	88.40%	0.944	0.296
NaiveBayes +(X2)	9	1	0.043	0.125	10.50%	89.40%	0.959	0.3244
NaiveBayes +(GR)	9	1	0.043	0.125	10.50%	89.40%	0.967	0.3271
NaiveBayes +(SUA)	10	1	0.043	0.139	11.50%	88.40%	0.964	0.3395
NaiveBayes	11	5	21.70%	15.30%	16.84%	83%	0.844	0.4115
J48 + CS(1)-np	12	9	39.10%	16.70%	22.00%	78.00%	0.732	0.4496
NaiveBayes +(RF-f)	16	2	0.087	0.222	18.90%	81.00%	0.931	0.4272

KNN (5) +(SUA)	3	3	0.13	42	6.30%	93.60%	0.944	0.2447

For the subsets of 6 selected features compared with the C.45 the data shows NaiveBayes +(SUA) has the best Area under the ROC curve value of 97% and KNN (5) +(X2) has the least Type II errors of 0.028.

Evaluation for Assignment 4 - Feature selection of 6 features								
Classifier / Learners	**Type I**	**Type II**	**Type I Error rate %**	**Type II Error rate %**	**Misclassification Rate %**	**Correctly Classified %**	**ROC**	**RMSE**
KNN (5) +(X2)	2	3	0.13	0.028	5.2	95%	0.918	0.2497
KNN (5) +(GR)	3	4	0.174	0.042	7.3	93%	0.942	0.2506
KNN (5) +(RF-f)	3	11	0.478	0.042	14.7	85%	0.931	0.2886
KNN (5) +(IG)	4	5	0.217	0.056	9.4	91%	0.96	0.2669
KNN (5) +(RFW-t)	4	5	0.217	0.056	9.4	91%	0.94	0.2538
KNN (5) +(SUA)	5	2	0.087	0.069	7.3	93%	0.954	0.234
J48 + CS(0.5)-np	6	7	30.40%	8.30%	1360.00%	86.3	0.831	0.3572
KNN (5)	7	9	39.10%	9.70%	16.8	83%	0.863	0.3573
NaiveBayes +(GR)	8	1	0.043	0.111	9.4	91%	0.97	0.2989
NaiveBayes +(SUA)	8	1	0.043	0.111	9.40%	90.50%	0.972	0.2946
NaiveBayes +(RFW-t)	8	1	0.043	0.111	9.4	91%	0.968	0.2852
NaiveBayes +(X2)	9	1	0.043	0.125	10.5	89%	0.968	0.3228
J48 + CS(2)-np	9	16	69.60%	12.50%	2600.00%	74.00%	0.591	0.4954
NaiveBayes +(IG)	10	4	0.174	0.139	14.7	85%	0.947	0.3172
NaiveBayes	11	5	21.70%	15.30%	16.84%	83%	0.844	0.4115
J48 + CS(1)-np	12	9	39.10%	16.70%	2200.00%	78.00%	0.732	0.4496
NaiveBayes +(RF-f)	14	1	0.43%	0.194	15.7	84%	0.959	0.3972

For the subsets of 7 selected features compared with the C.45 the data shows KNN (5) +(RFW-t) has the best Area under the ROC curve of 98% andKNN (5) +(RFW-t) has the least Type II errors.

Evaluation for Assignment 4 - Feature selection of 7 features								
Classifier / Learners	**Type I**	**Type II**	**Type I Error rate %**	**Type II Error rate %**	**Misclassification Rate %**	**Correctly Classified %**	**ROC**	**RMSE**

KNN (5) +(RFW-t)	0	8	0.348	0	8.40%	91.50%	0.989	0.2295
NaiveBayes +(SUA)	8	1	0.043	0.111	9.40%	90.50%	0.971	0.2978
NaiveBayes +(X2)	6	1	0.043	0.083	7.30%	92.60%	0.97	0.2604
NaiveBayes +(GR)	4	1	0.043	0.056	5.20%	94.70%	0.969	0.2181
NaiveBayes +(RFW-t)	0	3	0.13	0.00%	3.10%	96.80%	0.965	0.1959
KNN (5) +(X2)	2	5	0.217	0.028	7.30%	92.63%	0.952	0.2303
NaiveBayes +(IG)	9	2	0.087	0.125	11.50%	88.40%	0.949	0.3116
KNN (5) +(SUA)	4	2	0.087	0.56	6.30%	93.60%	0.949	0.2367
KNN (5) +(RF-f)	1	10	0.435	0.014	11.50%	88.40%	0.943	0.2752
NaiveBayes +(RF-f)	14	1	0.43	0.194	15.70%	84.20%	0.938	0.3893
KNN (5) +(IG)	2	8	0.348	0.028	10.50%	89.40%	0.917	0.2916
KNN (5) +(GR)	1	5	0.217	0.014	6.30%	93.60%	0.912	0.266
KNN (5)	7	9	39.10%	9.70%	16.8	83%	0.863	0.3573
NaiveBayes	11	5	21.70%	15.30%	16.84%	83%	0.844	0.4115
J48 + CS(0.5)-np	6	7	30.40%	8.30%	1360.00%	86.3	0.831	0.3572
J48 + CS(1)-np	12	9	39.10%	16.70%	2200.00%	78.00%	0.732	0.4496
J48 + CS(2)-np	9	16	69.60%	12.50%	2600.00%	74.00%	0.591	0.4954

For the subsets of 8 selected features compared with the C.45 the data shows NaiveBayes +(RFW-t) has the best Area under the ROC curve of 98% and KNN (5) +(RF-f) has the least Type II errors.

Evaluation for Assignment 4 - Feature selection of 8 features								
Classifier / Learners	**Type I**	**Type II**	**Type I Error rate %**	**Type II Error rate %**	**Misclassification Rate %**	**Correctly Classified %**	**ROC**	**RMSE**
NaiveBayes +(RFW-t)	5	1	0.043	0.069	6.30%	93.60%	0.979	0.2327
NaiveBayes +(X2)	8	1	0.043	0.111	9.40%	90.50%	0.978	0.2671
NaiveBayes +(SUA)	7	1	0.043	0.097	8.40%	91.50%	0.977	0.2363
NaiveBayes +(GR)	4	2	0.087	0.056	6.30%	93.60%	0.97	0.2357
NaiveBayes +(IG)	8	2	0.087	0.111	10.50%	89.40%	0.96	0.3105
KNN (5) +(X2)	1	7	0.304	0.014	8.40%	91.50%	0.95	0.2479

NaiveBayes +(RF-f)	12	2	0.087	0.167	14.70%	85.20%	0.948	0.3594
KNN (5) +(RFW-t)	1	7	0.304	0.014	8.40%	91.50%	0.944	0.2635
KNN (5) +(SUA)	4	5	0.217	0.056	9.4	90.50%	0.941	0.253
KNN (5) +(IG)	2	7	0.304	0.028	9.40%	90.52%	0.929	0.2745
KNN (5) +(RF-f)	0	13	0.565	0	13.60%	86.30%	0.919	0.2956
KNN (5) +(GR)	1	8	0.348	0.014	9.40%	90.50%	0.906	0.2752
KNN (5)	7	9	39.10%	9.70%	16.8	83%	0.863	0.3573
NaiveBayes	11	5	21.70%	15.30%	16.84%	83%	0.844	0.4115
J48 + CS(0.5)-np	6	7	30.40%	8.30%	1360.00%	86.3	0.831	0.3572
J48 + CS(1)-np	12	9	39.10%	16.70%	2200.00%	78.00%	0.732	0.4496
J48 + CS(2)-np	9	16	69.60%	12.50%	2600.00%	74.00%	0.591	0.4954

For the subsets of 9 selected features compared with the C.45 the data shows NaiveBayes +(RFW-t) has the best Area under the ROC curve and KNN (5) +(RF-f) has the least Type II errors.

Evaluation for Assignment 4 - Feature selection of 9 features								
Classifier / Learners	**Type I**	**Type II**	**Type I Error rate %**	**Type II Error rate %**	**Misclassification Rate %**	**Correctly Classified %**	**ROC**	**RMSE**
KNN (5) +(RFW-t)	1	4	0.174	0.014	5.20%	94.70%	0.959	0.2132
KNN (5) +(RF-f)	1	10	0.435	0.014	11.50%	88.40%	0.91	0.2971
KNN (5) +(GR)	2	9	0.391	0.027	11.50%	88.24%	0.921	0.2929
NaiveBayes +(RFW-t)	2	4	0.174	0.028	6.30%	93.60%	0.984	0.199
KNN (5) +(SUA)	2	7	0.304	0.028	9.4	90.50%	0.944	0.2546
KNN (5) +(IG)	3	6	0.261	0.042	9.40%	90.52%	0.924	0.2843
KNN (5) +(X2)	5	7	0.304	0.069	12.60%	87.30%	0.972	0.2805
J48 + CS(0.5)-np	6	7	30.40%	8.30%	1360.00%	86.3	0.831	0.3572
KNN (5)	7	9	39.10%	9.70%	16.8	83%	0.863	0.3573
NaiveBayes +(SUA)	8	1	0.043	0.111	9.40%	90.50%	0.969	0.2895
NaiveBayes +(X2)	9	1	0.043	0.125	10.50%	89.47%	0.965	0.3016
NaiveBayes +(RF-f)	9	2	0.087	0.125	11.50%	88.40%	0.957	0.3142
NaiveBayes +(IG)	9	2	0.087	0.125	11.50%	88.40%	0.944	0.3119
NaiveBayes	9	4	0.174	0.125	13.60%	86.30%	0.923	0.3212

+(GR)								
J48 + CS(2)-np	9	16	69.60%	12.50%	2600.00%	74.00%	0.591	0.4954
NaiveBayes	11	5	21.70%	15.30%	16.84%	83%	0.844	0.4115
J48 + CS(1)-np	12	9	39.10%	16.70%	2200.00%	78.00%	0.732	0.4496

For the subsets of 10 selected features compared with the C.45 the data shows NaiveBayes +(SUA) has the best Area under the ROC curve and NaiveBayes +(RFW-t) has the least Type II errors.

Evaluation for Assignment 4 - Feature selection of 10 features								
Classifier / Learners	**Type I**	**Type II**	**Type I Error rate %**	**Type II Error rate %**	**Misclassification Rate %**	**Correctly Classified %**	**ROC**	**RMSE**
NaiveBayes +(RFW-t)	1	4	0.174	0.014	5.20%	94.70%	0.959	0.2132
KNN (5) +(RFW-t)	1	3	0.13	0.014	4.20%	95.70%	0.958	0.2113
KNN (5) +(GR)	2	10	0.435	0.028	12.60%	87.30%	0.924	0.2922
KNN (5) +(RF-f)	2	9	0.391	0.028	11.50%	88.40%	0.912	0.2878
KNN (5) +(IG)	3	7	0.304	0.042	10.50%	89.40%	0.939	0.1427
KNN (5) +(SUA)	3	8	0.348	0.042	11.50%	88.42%	0.934	0.2722
KNN (5) +(X2)	5	6	0.261	0.069	11.50%	88.40%	0.938	0.2628
J48 + CS(0.5)-np	6	7	30.40%	8.30%	1360.00%	86.3	0.831	0.3572
NaiveBayes +(X2)	7	1	0.043	0.097	8.40%	91.50%	0.961	0.2886
KNN (5)	7	9	39.10%	9.70%	16.8	83%	0.863	0.3573
NaiveBayes +(RF-f)	8	2	0.087	0.111	10.50%	89.40%	0.966	0.2975
NaiveBayes +(IG)	8	2	0.087	0.111	10.50%	89.40%	0.944	0.3094
NaiveBayes +(GR)	8	4	0.174	0.111	12.60%	87.30%	0.926	0.307
NaiveBayes +(SUA)	9	1	0.043	0.125	10.50%	89.40%	0.97	0.2937
J48 + CS(2)-np	9	16	69.60%	12.50%	2600.00%	74.00%	0.591	0.4954
NaiveBayes	11	5	21.70%	15.30%	16.84%	83%	0.844	0.4115
J48 + CS(1)-np	12	9	39.10%	16.70%	2200.00%	78.00%	0.732	0.4496

For the subsets of 20 selected features compared with the C.45 the data shows NaiveBayes +(RFW-t) has the best Area under the ROC curve of 99% and KNN (5) +(GR) has the least Type II errors.

Evaluation for Assignment 4 - Feature selection of 20 features								
Classifier / Learners	**Type I**	**Type II**	**Type I Error rate %**	**Type II Error rate %**	**Misclassification Rate %**	**Correctly Classified %**	**ROC**	**RMSE**
KNN (5) +(GR)	0	16	0.696	0	16.80%	83.10%	0.812	0.3587
NaiveBayes +(RFW-t)	1	2	0.087	0.014	3.10%	96.80%	0.986	0.1637
KNN (5) +(RFW-t)	1	4	0.174	0.014	5.20%	94.70%	0.96	0.2042
KNN (5) +(RF-f)	2	10	0.435	0.028	12..6%	87.30%	0.887	0.3062
KNN (5) +(X2)	5	5	0.217	0.069	10.50%	89.47%	0.934	0.266
KNN (5) +(SUA)	5	5	0.217	0.069	10.50%	89.00%	0.92	0.2835
NaiveBayes +(IG)	6	1	0.043	0.083	7.30%	92.63%	0.968	0.2652
KNN (5) +(IG)	6	3	0.13	0.083	9.40%	90.50%	0.93	0.1427
J48 + CS(0.5)-np	6	7	30.40%	8.30%	1360.00%	86.3	0.831	0.3572
NaiveBayes +(X2)	7	2	0.087	0.097	9.40%	90.50%	0.958	0.2882
KNN (5)	7	9	39.10%	9.70%	16.8	83%	0.863	0.3573
NaiveBayes +(SUA)	9	1	0.043	0.125	10.50%	89.40%	0.966	0.3047
NaiveBayes +(RF-f)	9	2	0.087	0.125	11.50%	88.40%	0.96	0.3021
J48 + CS(2)-np	9	16	69.60%	12.50%	2600.00%	74.00%	0.591	0.4954
NaiveBayes +(GR)	10	10	0.435	0.139	21.00%	79.00%	0.83	0.3929
NaiveBayes	11	5	21.70%	15.30%	16.84%	83%	0.844	0.4115
J48 + CS(1)-np	12	9	39.10%	16.70%	2200.00%	78.00%	0.732	0.4496

For the subsets of 50 selected features compared with the C.45 the data shows NaiveBayes +(RFW-t) has the best Area under the ROC curve of 99% and KNN (5) +(RFW-t) has the least Type II errors.

Evaluation for Assignment 4 - Feature selection of 50 features								
Classifier / Learners	**Type I**	**Type II**	**Type I Error rate %**	**Type II Error rate %**	**Misclassification Rate %**	**Correctly Classified %**	**ROC**	**RMSE**
KNN (5) +(RFW-t)	0	3	0.13	0.00%	3.10%	96.80%	0.965	0.1959
NaiveBayes +(RFW-t)	2	2	0.087	0.03%	4.20%	95.70%	0.988	0.1947
KNN (5) +(GR)	1	16	0.696	0.014	17.80%	82.10%	0.811	0.365
KNN (5) +(RF-f)	2	10	0.435	0.028	12.60%	87.30%	0.895	0.2964

KNN (5) +(SUA)	3	5	0.217	0.042	8.4	91.50%	0.938	0.2659
KNN (5) +(IG)	4	2	0.087	0.056	6.30%	93.60%	0.956	0.2386
KNN (5) +(X2)	5	6	0.261	0.069	11.50%	88.40%	0.927	0.2745
J48 + CS(0.5)-np	6	7	30.40%	8.30%	1360.00%	86.3	0.831	0.3572
NaiveBayes +(RF-f)	7	2	0.087	0.097	9.40%	90.50%	0.961	0.2887
KNN (5)	7	9	39.10%	9.70%	16.8	83%	0.863	0.3573
NaiveBayes +(X2)	8	3	0.13	0.111	11.50%	88.40%	0.958	0.3072
NaiveBayes +(SUA)	9	2	0.087	0.125	11.50%	88.40%	0.963	0.2888
J48 + CS(2)-np	9	16	69.60%	12.50%	2600.00%	74.00%	0.591	0.4954
NaiveBayes +(IG)	10	1	0.043	0.139	11.50%	88.42%	0.968	0.3298
NaiveBayes +(GR)	10	8	0.348	0.139	18.90%	81.00%	0.827	0.3828
NaiveBayes	11	5	21.70%	15.30%	16.84%	83%	0.844	0.4115
J48 + CS(1)-np	12	9	39.10%	16.70%	2200.00%	78.00%	0.732	0.4496

For the subsets of 100 selected features compared with the C.45 the data shows NaiveBayes +(RFW-t) has the best Area under the ROC curve of 99% and KNN (5) +(RFW-t) has the least Type II errors.

Evaluation for Assignment 4 - Feature selection of 100 features								
Classifier / Learners	**Type I**	**Type II**	**Type I Error rate %**	**Type II Error rate %**	**Misclassification Rate %**	**Correctly Classified %**	**ROC**	**RMSE**
NaiveBayes +(RFW-t)	2	3	0.13	0.028	5.20%	94.70%	0.992	0.1944
NaiveBayes +(SUA)	7	2	0.087	0.097	9.40%	90.50%	0.969	0.2755
KNN (5) +(RFW-t)	0	4	0.17%	0.00%	4.20%	95.70%	0.962	0.2073
NaiveBayes +(RF-f)	9	3	0.13	0.125	12.60%	87.20%	0.961	0.2994
NaiveBayes +(IG)	11	1	0.043	0.153	12.60%	87.36%	0.955	0.3546
NaiveBayes +(X2)	10	3	0.13	0.139	13.60%	86.30%	0.953	0.3167
KNN (5) +(IG)	7	2	0.087	0.097	9.40%	90.50%	0.938	0.2571
KNN (5) +(X2)	3	7	0.304	0.042	10.50%	89.40%	0.918	0.2901
KNN (5) +(SUA)	3	6	0.261	0.042	9.40%	90.50%	0.918	0.2856
KNN (5) +(RF-f)	1	12	0.522	0.14%	13.60%	86.30%	0.912	0.295

KNN (5)	7	9	39.10%	9.70%	16.8	83%	0.863	0.3573
KNN (5) +(GR)	1	16	0.696	0.014	17.80%	82.00%	0.853	0.345
NaiveBayes	11	5	21.70%	15.30%	16.84%	83%	0.844	0.4115
J48 + CS(0.5)-np	6	7	30.40%	8.30%	1360.00%	86.3	0.831	0.3572
NaiveBayes +(GR)	12	10	0.435	0.167	23.10%	76.80%	0.817	0.4044
J48 + CS(1)-np	12	9	39.10%	16.70%	2200.00%	78.00%	0.732	0.4496
J48 + CS(2)-np	9	16	69.60%	12.50%	2600.00%	74.00%	0.591	0.4954

For the subsets of 200 selected features compared with the C.45 the data shows NaiveBayes +(RFW-t) has the best Area under the ROC curve of 98% and KNN (5) +(RFW-t) has the least Type II errors.

Evaluation for Assignment 4 - Feature selection of 200 features								
Classifier / Learners	**Type I**	**Type II**	**Type I Error rate %**	**Type II Error rate %**	**Misclassification Rate %**	**Correctly Classified %**	**ROC**	**RMSE**
KNN (5) +(RFW-t)	0	5	0.217	0.00%	5.20%	94.70%	0.961	0.2083
KNN (5) +(GR)	1	15	0.652	0.014	16.80%	83.10%	0.825	0.3528
NaiveBayes +(RFW-t)	2	7	0.304	0.028	9.40%	90.50%	0.978	0.2564
KNN (5) +(SUA)	2	6	0.261	0.028	8.40%	91.50%	0.957	0.258
KNN (5) +(IG)	4	3	0.13	0.056	7.30%	92.60%	0.944	0.2489
KNN (5) +(RF-f)	4	4	0.174	0.056	8.40%	91.50%	0.917	0.2521
NaiveBayes +(SUA)	5	3	0.13	0.069	8.40%	91.50%	0.961	0.2747
KNN (5) +(X2)	5	6	0.261	0.069	11.50%	88.40%	0.887	0.3184
J48 + CS(0.5)-np	6	7	30.40%	8.30%	1360.00%	86.3	0.831	0.3572
KNN (5)	7	9	39.10%	9.70%	16.8	83%	0.863	0.3573
NaiveBayes +(RF-f)	9	3	0.13	0.125	12.6	87.30%	0.958	0.3199
NaiveBayes +(X2)	9	3	0.13	0.125	12.60%	87.36%	0.951	0.3128
NaiveBayes +(GR)	9	11	0.478	0.125	21.00%	79.00%	0.83	0.3869
J48 + CS(2)-np	9	16	69.60%	12.50%	2600.00%	74.00%	0.591	0.4954
NaiveBayes +(IG)	11	1	0.043	0.153	12.60%	87.30%	0.955	0.3546
NaiveBayes	11	5	21.70%	15.30%	16.84%	83%	0.844	0.4115
J48 + CS(1)-np	12	9	39.10%	16.70%	2200.00%	78.00%	0.732	0.4496

According to the following results, naive bayes and Knn perform better when using feature selection methods compared to the C.45(J48 Weka) tree learner. Based on the given techniques, this NaiveBayes +(RFW-t) learner achieved the highest AUC of 98% - 99%, while the KNN (5) +(RFW-t) learner had the lowest number of Type II errors.

Feature selection analysis

The following are tables that show the data comparing individual feature selection methods with Naive Bayes and Knn for the subsets selected 5,6,7,8,9,10,20,50,100,200. Table 3-1 below compares Naive Bayes learner with Information Gain (IG) learner based on 5,6,7,8,9,10,20,50,100,200 selected features and found that 8 selected features performed the best.

Table 3-1

NaiveBayes +(IG) - Feature selection (5,6,7,8,9,10,20,50,100, and 200)									
FS	**Classifier / Learners**	**Type I**	**Type II**	**Type I Error rate %**	**Type II Error rate %**	**Misclassification Rate %**	**Correctly Classified %**	**ROC**	**RMSE**
200	NaiveBayes +(IG)	9	2	0.087	0.125	11.50%	88.40%	0.944	0.296
100	NaiveBayes +(IG)	10	4	0.174	0.139	14.7	85%	0.947	0.3172
50	NaiveBayes +(IG)	9	2	0.087	0.125	11.50%	88.40%	0.949	0.3116
20	NaiveBayes +(IG)	8	2	0.087	0.111	10.50%	89.40%	0.96	0.3105
10	NaiveBayes +(IG)	9	2	0.087	0.125	11.50%	88.40%	0.944	0.3119
9	NaiveBayes +(IG)	8	2	0.087	0.111	10.50%	89.40%	0.944	0.3094
8	NaiveBayes +(IG)	6	1	0.043	0.083	7.30%	92.63%	0.968	0.2652
7	NaiveBayes +(IG)	10	1	0.043	0.139	11.50%	88.42%	0.968	0.3298
6	NaiveBayes +(IG)	11	1	0.043	0.153	12.60%	87.36%	0.955	0.3546
5	NaiveBayes	11	1	0.043	0.153	12.60%	87.30%	0.955	0.3546

	+(IG)								

Table 3-2 below compares KNN learner with Information Gain (IG) learner based on 5,6,7,8,9,10,20,50,100,200 selected features and found that 7 selected features performed the best.

Table 3-2

KNN (5) +(IG) - Feature selection (5,6,7,8,9,10,20,50,100, and 200)									
FS	**Classifier / Learners**	**Type I**	**Type II**	**Type I Error rate %**	**Type II Error rate %**	**Misclassification Rate %**	**Correctly Classified %**	**ROC**	**RMSE**
200	KNN (5) +(IG)	5	4	0.174	0.069	9.40%	90.50%	0.927	0.2902
100	KNN (5) +(IG)	4	5	0.217	0.056	9.4	91%	0.96	0.2669
50	KNN (5) +(IG)	2	8	0.348	0.028	10.50%	89.40%	0.917	0.2916
20	KNN (5) +(IG)	2	7	0.304	0.028	9.40%	90.52%	0.929	0.2745
10	KNN (5) +(IG)	3	6	0.261	0.042	9.40%	90.52%	0.924	0.2843
9	KNN (5) +(IG)	3	7	0.304	0.042	10.50%	89.40%	0.939	0.1427
8	KNN (5) +(IG)	6	3	0.13	0.083	9.40%	90.50%	0.93	0.1427
7	KNN (5) +(IG)	4	2	0.087	0.056	6.30%	93.60%	0.956	0.2386
6	KNN (5) +(IG)	7	2	0.087	0.097	9.40%	90.50%	0.938	0.2571
5	KNN (5) +(IG)	4	3	0.13	0.056	7.30%	92.60%	0.944	0.2489

Table 3-3 below compares Naive Bayes learner with Chi Square (X2) learner based on 5,6,7,8,9,10,20,50,100,200 selected features and found that 9 selected features performed the best.

Table 3-3

NaiveBayes +(X2) - Feature selection (5,6,7,8,9,10,20,50,100, and 200)									
FS	**Classifier / Learners**	**Type I**	**Type II**	**Type I Error rate %**	**Type II Error rate %**	**Misclassification Rate %**	**Correctly Classified %**	**ROC**	**RMSE**
200	NaiveBayes	9	1	0.043	0.125	10.50%	89.40%	0.959	0.3244

	+(X2)								
100	NaiveBayes +(X2)	9	1	0.043	0.125	10.5	89%	0.968	0.3228
50	NaiveBayes +(X2)	6	1	0.043	0.083	7.30%	92.60%	0.97	0.2604
20	NaiveBayes +(X2)	8	1	0.043	0.111	9.40%	90.50%	0.978	0.2671
10	NaiveBayes +(X2)	9	1	0.043	0.125	10.50%	89.47%	0.965	0.3016
9	NaiveBayes +(X2)	7	1	0.043	0.097	8.40%	91.50%	0.961	0.2886
8	NaiveBayes +(X2)	7	2	0.087	0.097	9.40%	90.50%	0.958	0.2882
7	NaiveBayes +(X2)	8	3	0.13	0.111	11.50%	88.40%	0.958	0.3072
6	NaiveBayes +(X2)	10	3	0.13	0.139	13.60%	86.30%	0.953	0.3167
5	NaiveBayes +(X2)	9	3	0.13	0.125	12.60%	87.36%	0.951	0.3128

Table 3-4 below compares KNN learner with Chi Square (X2) learner based on

5,6,7,8,9,10,20,50,100,200 selected features and found that 200 selected features performed the best

Table 3-4

KNN (5) +(X2) - Feature selection (5,6,7,8,9,10,20,50,100, and 200)									
FS	**Classifier / Learners**	**Type I**	**Type II**	**Type I Error rate %**	**Type II Error rate %**	**Misclassification Rate %**	**Correctly Classified %**	**ROC**	**RMSE**
200	KNN (5) +(X2)	2	3	0.13	0.028	5.20%	94.70%	0.943	0.2403
100	KNN (5) +(X2)	2	3	0.13	0.028	5.2	95%	0.918	0.2497
50	KNN (5) +(X2)	2	5	0.217	0.028	7.30%	92.63%	0.952	0.2303
20	KNN (5) +(X2)	1	7	0.304	0.014	8.40%	91.50%	0.95	0.2479
10	KNN (5) +(X2)	5	7	0.304	0.069	12.60%	87.30%	0.972	0.2805
9	KNN (5) +(X2)	5	6	0.261	0.069	11.50%	88.40%	0.938	0.2628
8	KNN (5) +(X2)	5	5	0.217	0.069	10.50%	89.47%	0.934	0.266
7	KNN (5) +(X2)	5	6	0.261	0.069	11.50%	88.40%	0.927	0.2745
6	KNN (5)	3	7	0.304	0.042	10.50%	89.40%	0.918	0.2901

	+(X2)								
5	KNN (5) +(X2)	5	6	0.261	0.069	11.50%	88.40%	0.887	0.3184

Table 3-5 below compares Naive Bayes learner with Gain Ratio(GR) learner based on

5,6,7,8,9,10,20,50,100,200 selected features and found that 50 selected features performed the best.

Table 3-5

NaiveBayes +(GR) - Feature selection (5,6,7,8,9,10,20,50,100, and 200)									
FS	**Classifier / Learners**	**Type I**	**Type II**	**Type I Error rate %**	**Type II Error rate %**	**Misclassification Rate %**	**Correctly Classified %**	**ROC**	**RMSE**
200	NaiveBayes +(GR)	9	1	0.043	0.125	10.50%	89.40%	0.967	0.3271
100	NaiveBayes +(GR)	8	1	0.043	0.111	9.4	91%	0.97	0.2989
50	NaiveBayes +(GR)	4	1	0.043	0.056	5.20%	94.70%	0.969	0.2181
20	NaiveBayes +(GR)	4	2	0.087	0.056	6.30%	93.60%	0.97	0.2357
10	NaiveBayes +(GR)	9	4	0.174	0.125	13.60%	86.30%	0.923	0.3212
9	NaiveBayes +(GR)	8	4	0.174	0.111	12.60%	87.30%	0.926	0.307
8	NaiveBayes +(GR)	10	10	0.435	0.139	21.00%	79.00%	0.83	0.3929
7	NaiveBayes +(GR)	10	8	0.348	0.139	18.90%	81.00%	0.827	0.3828
6	NaiveBayes +(GR)	12	10	0.435	0.167	23.10%	76.80%	0.817	0.4044
5	NaiveBayes +(GR)	9	11	0.478	0.125	21.00%	79.00%	0.83	0.3869

Table 3-6 below compares KNN learner with Gain Ratio(GR) learner based on

5,6,7,8,9,10,20,50,100,200 selected features and found that 200 selected features performed the best

Table 3-6

KNN (5) +(GR) - Feature selection (5,6,7,8,9,10,20,50,100, and 200)									
FS	Classifier / Learners	Type I	Type II	Type I Error rate %	Type II Error rate %	Misclassification Rate %	Correctly Classified %	ROC	RMSE
200	KNN (5) +(GR)	4	3	0.13	0.056	7.30%	92.60%	0.92	0.2645
100	KNN (5) +(GR)	3	4	0.174	0.042	7.3	93%	0.942	0.2506
50	KNN (5) +(GR)	1	5	0.217	0.014	6.30%	93.60%	0.912	0.266
20	KNN (5) +(GR)	1	8	0.348	0.014	9.40%	90.50%	0.906	0.2752
10	KNN (5) +(GR)	2	9	0.391	0.027	11.50%	88.24%	0.921	0.2929
9	KNN (5) +(GR)	2	10	0.435	0.028	12.60%	87.30%	0.924	0.2922
8	KNN (5) +(GR)	0	16	0.696	0	16.80%	83.10%	0.812	0.3587
7	KNN (5) +(GR)	1	16	0.696	0.014	17.80%	82.10%	0.811	0.365
6	KNN (5) +(GR)	1	16	0.696	0.014	17.80%	82.00%	0.853	0.345
5	KNN (5) +(GR)	1	15	0.652	0.014	16.80%	83.10%	0.825	0.3528

Table 3-7 below compares Naive Bayes learner with Symmetric Uncertainty (SAU) learner based on 5,6,7,8,9,10,20,50,100,200 selected features and found that 20 selected features performed the best.

Table 3-7

NaiveBayes +(SUA) - Feature selection (5,6,7,8,9,10,20,50,100, and 200)									
FS	Classifier / Learners	Type I	Type II	Type I Error rate %	Type II Error rate %	Misclassification Rate %	Correctly Classified %	ROC	RMSE
200	NaiveBayes +(SUA)	10	1	0.043	0.139	11.50%	88.40%	0.964	0.3395
100	NaiveBayes +(SUA)	8	1	0.043	0.111	9.40%	90.50%	0.972	0.2946
50	NaiveBayes +(SUA)	8	1	0.043	0.111	9.40%	90.50%	0.971	0.2978
20	NaiveBayes +(SUA)	7	1	0.043	0.097	8.40%	91.50%	0.977	0.2363
10	NaiveBayes +(SUA)	8	1	0.043	0.111	9.40%	90.50%	0.969	0.2895

9	NaiveBayes +(SUA)	9	1	0.043	0.125	10.50%	89.40%	0.97	0.2937
8	NaiveBayes +(SUA)	9	1	0.043	0.125	10.50%	89.40%	0.966	0.3047
7	NaiveBayes +(SUA)	9	2	0.087	0.125	11.50%	88.40%	0.963	0.2888
6	NaiveBayes +(SUA)	7	2	0.087	0.097	9.40%	90.50%	0.969	0.2755
5	NaiveBayes +(SUA)	5	3	0.13	0.069	8.40%	91.50%	0.961	0.2747

Table 3-8 below compares KNN learner with Symmetric Uncertainty (SAU) learner based on 5,6,7,8,9,10,20,50,100,200 selected features and found that 50 selected features performed the best

Table 3-8

KNN (5) +(SAU) - Feature selection (5,6,7,8,9,10,20,50,100, and 200)									
FS	**Classifier / Learners**	**Type I**	**Type II**	**Type I Error rate %**	**Type II Error rate %**	**Misclassification Rate %**	**Correctly Classified %**	**ROC**	**RMSE**
200	KNN (5) +(SAU)	3	3	0.13	42	6.30%	93.60%	0.944	0.2447
100	KNN (5) +(SAU)	5	2	0.087	0.069	7.3	93%	0.954	0.234
50	KNN (5) +(SAU)	4	2	0.087	0.56	6.30%	93.60%	0.949	0.2367
20	KNN (5) +(SAU)	4	5	0.217	0.056	9.4	90.50%	0.941	0.253
10	KNN (5) +(SAU)	2	7	0.304	0.028	9.4	90.50%	0.944	0.2546
9	KNN (5) +(SAU)	3	8	0.348	0.042	11.50%	88.42%	0.934	0.2722
8	KNN (5) +(SAU)	5	5	0.217	0.069	10.50%	89.00%	0.92	0.2835
7	KNN (5) +(SAU)	3	5	0.217	0.042	8.4	91.50%	0.938	0.2659
6	KNN (5) +(SAU)	3	6	0.261	0.042	9.40%	90.50%	0.918	0.2856
5	KNN (5) +(SAU)	2	6	0.261	0.028	8.40%	91.50%	0.957	0.258

Table 3-9 below compares Naive Bayes learner with ReliefF (RF-f) weight by distance parameter set to False learner based on 5,6,7,8,9,10,20,50,100,200 selected features and found that 100 selected features performed the best.

Table 3-9

NaiveBayes +(RF-f) - Feature selection (5,6,7,8,9,10,20,50,100, and 200)									
FS	**Classifier / Learners**	**Type I**	**Type II**	**Type I Error rate %**	**Type II Error rate %**	**Misclassification Rate %**	**Correctly Classified %**	**ROC**	**RMSE**
200	NaiveBayes +(RF-f)	16	2	0.087	0.222	18.90%	81.00%	0.931	0.4272
100	NaiveBayes +(RF-f)	14	1	0.43%	0.194	15.7	84%	0.959	0.3972
50	NaiveBayes +(RF-f)	14	1	0.43	0.194	15.70%	84.20%	0.938	0.3893
20	NaiveBayes +(RF-f)	12	2	0.087	0.167	14.70%	85.20%	0.948	0.3594
10	NaiveBayes +(RF-f)	9	2	0.087	0.125	11.50%	88.40%	0.957	0.3142
9	NaiveBayes +(RF-f)	8	2	0.087	0.111	10.50%	89.40%	0.966	0.2975
8	NaiveBayes +(RF-f)	9	2	0.087	0.125	11.50%	88.40%	0.96	0.3021
7	NaiveBayes +(RF-f)	7	2	0.087	0.097	9.40%	90.50%	0.961	0.2887
6	NaiveBayes +(RF-f)	9	3	0.13	0.125	12.60%	87.20%	0.961	0.2994
5	NaiveBayes +(RF-f)	9	3	0.13	0.125	12.6	87.30%	0.958	0.3199

Table 3-10 below compares KNN learner with ReliefF (RF-f) weight by distance parameter set to False learner based on 5,6,7,8,9,10,20,50,100,200 selected features and found that 5 selected features performed the best

Table 3-10

KNN (5) +(RF-f) - Feature selection (5,6,7,8,9,10,20,50,100, and 200)

FS	Classifier / Learners	Type I	Type II	Type I Error rate %	Type II Error rate %	Misclassification Rate %	Correctly Classified %	ROC	RMSE
200	KNN (5) +(RF-f)	1	8	0.348	0.014	9.40%	90.50%	0.908	0.2879
100	KNN (5) +(RF-f)	3	11	0.478	0.042	14.7	85%	0.931	0.2886
50	KNN (5) +(RF-f)	1	10	0.435	0.014	11.50%	88.40%	0.943	0.2752
20	KNN (5) +(RF-f)	0	13	0.565	0	13.60%	86.30%	0.919	0.2956
10	KNN (5) +(RF-f)	1	10	0.435	0.014	11.50%	88.40%	0.91	0.2971
9	KNN (5) +(RF-f)	2	9	0.391	0.028	11.50%	88.40%	0.912	0.2878
8	KNN (5) +(RF-f)	2	10	0.435	0.028	12..6%	87.30%	0.887	0.3062
7	KNN (5) +(RF-f)	2	10	0.435	0.028	12.60%	87.30%	0.895	0.2964
6	KNN (5) +(RF-f)	1	12	0.522	0.14%	13.60%	86.30%	0.912	0.295
5	KNN (5) +(RF-f)	4	4	0.174	0.056	8.40%	91.50%	0.917	0.2521

Table 3-5 below compares Naive Bayes learner withReliefF-W (RFW-t) weight by distance parameter set to true learner based on 5,6,7,8,9,10,20,50,100,200 selected features and found that 20 selected features performed the best.

Table 3-5

NaiveBayes +(RFW-t) - Feature selection (5,6,7,8,9,10,20,50,100, and 200)									
FS	Classifier / Learners	Type I	Type II	Type I Error rate %	Type II Error rate %	Misclassification Rate %	Correctly Classified %	ROC	RMSE
200	NaiveBayes +(RFW-t)	9	3	0.13	0.125	12.60%	.87.3	0.94	0.3519
100	NaiveBayes +(RFW-t)	8	1	0.043	0.111	9.4	91%	0.968	0.2852
50	NaiveBayes +(RFW-t)	15	5	0.217	0.208	21	79	0.916	0.4483
20	NaiveBayes	5	1	0.043	0.069	6.30%	93.60%	0.979	0.2327

	+(RFW-t)								
10	NaiveBayes +(RFW-t)	2	4	0.174	0.028	6.30%	93.60%	0.984	0.199
9	NaiveBayes +(RFW-t)	1	2	0.087	0.014	3.10%	96.80%	0.986	0.1637
8	NaiveBayes +(RFW-t)	0	2	0.087	0	2.10%	97.80%	0.981	0.1522
7	NaiveBayes +(RFW-t)	2	2	0.087	0.03%	4.20%	95.70%	0.988	0.1947
6	NaiveBayes +(RFW-t)	2	3	0.13	0.028	5.20%	94.70%	0.992	0.1944
5	NaiveBayes +(RFW-t)	2	7	0.304	0.028	9.40%	90.50%	0.978	0.2564

Table 3-6 below compares KNN learner with ReliefF-W (RFW-t) weight by distance parameter set to true learner based on 5,6,7,8,9,10,20,50,100,200 selected features and found that 7 selected features performed the best

Table 3-6

KNN (5) +(RFW-t) - Feature selection (5,6,7,8,9,10,20,50,100, and 200)									
FS	**Classifier / Learners**	**Type I**	**Type II**	**Type I Error rate %**	**Type II Error rate %**	**Misclassification Rate %**	**Correctly Classified %**	**ROC**	**RMSE**
200	KNN (5) +(RFW-t)	4	6	0.261	0.056	10.50%	89.40%	0.927	0.276
100	KNN (5) +(RFW-t)	4	5	0.217	0.056	9.4	91%	0.94	0.2538
50	KNN (5) +(RFW-t)	0	8	0.348	0	8.40%	91.50%	0.989	0.2295
20	KNN (5) +(RFW-t)	1	7	0.304	0.014	8.40%	91.50%	0.944	0.2635
10	KNN (5) +(RFW-t)	1	4	0.174	0.014	5.20%	94.70%	0.959	0.2132
9	KNN (5) +(RFW-t)	1	3	0.13	0.014	4.20%	95.70%	0.958	0.2113
8	KNN (5) +(RFW-t)	1	4	0.174	0.014	5.20%	94.70%	0.96	0.2042
7	KNN (5) +(RFW-t)	0	3	0.13	0.00%	3.10%	96.80%	0.965	0.1959
6	KNN (5) +(RFW-t)	0	4	0.17%	0.00%	4.20%	95.70%	0.962	0.2073
5	KNN (5)	0	5	0.217	0.00%	5.20%	94.70%	0.961	0.2083

	+(RFW-t)								