Jixing (Jacey) Man CS699 Final Project

Part 1: Pre-Process and File Organization

There are total of 6 classification method I used to complete the project, see below list for the 6 method I used:

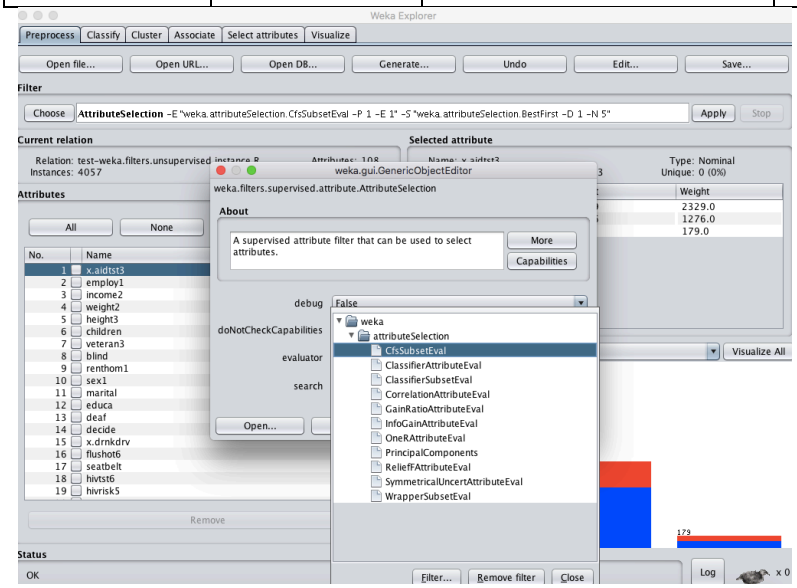
NaiveBayes
Logistic
RandomForest
RandomTree
IBK(KNN)
OneR

There are 4 different attributed selection methods I used, and I have saved train/test files accordingly to each method while run the classification algorithms. I will pact all train/test files and also with the train/test file that is re-named to “best-xxx.arff”.

I actually did not do any pre-processing of the data itself, so there are no extra-step involved here.

See below screenshot of how I selected the attribute selection on weka. I select the filter on the preprocess page, select the method, and hit apply, and do the same for both train and test to reduced-train and reduced-test, and save the files as separate train and test files for each attribute selection method. For example, reduced_test1.arff and reduced_train1.arff would be arff files for CfsSubSetEval.

Train/Test Set	Set 1	Set 2	Set 3	Set 4
Attribute Evaluator	CfsSubSetEval	CorrelationAttributeEval	OneRAttributeEval	ClassifierAttributeEval
Search Method	GreedyStepwise	Ranker	Ranker	Ranker



There are 5 different spreadsheet in the zip file, 4 of which are individual results of each selection and their classification methods along attributed selected.
Example:

The "Man_Jixing_Results_Summary.xlsx" file is a result summary of recall and confutation matrix of all test run.
The reason why I use recall and Corrected instances as the main performance measure matrix for this project is because for havarth3, the data is focused on whether the person was ever told to have some form of arthritis etc, so I believe it is best to focus on Recall, the true positive rate.
Please see screenshot below for the summary and the excel file for detail.

Part 2: Result Analysis Table

I have put the table of the accuracy rate of all test run and also the table of each test run's confusion matrix. All other performance measure and selected attributes from weka's output for the 24 models are at the Part 5, the bottom of this document and is also in the supplement excel files, the excel files are much easier to view, so I suggest to look at the excel file for detailed weka output information instead.
The accuracy and recall table

Overall Accuracy		NaiveBayes	Logistic	RandomForest	RandomTree	IBK	OneR	Average
CfsSubSetEval	Correct Instance	73.9463	74.4392	72.1469	69.2877	68.9426	73.5765	72.0565333
	Incorrect Instance	26.0537	25.5608	27.8531	30.7123	31.0574	26.4235	27.9434667
CorrelationAttributeEval	Correct Instance	70.9884	75.3019	74.1188	64.9988	67.858	73.5765	71.1404
	Incorrect Instance	29.0116	24.6981	25.8812	35.0012	32.142	26.4235	28.8596
OneRAttributeEval	Correct Instance	70.9884	75.3266	74.5625	65.7136	67.858	73.5765	71.3376
	Incorrect Instance	29.0116	24.6734	25.4375	34.2864	32.142	26.4235	28.6624
ClassifierAttributeEval	Correct Instance	70.9884	75.2019	75.0308	65.7875	67.858	73.5765	71.4071833
	Incorrect Instance	29.0116	24.6981	24.9692	34.2125	32.142	26.4235	28.57615

Recall		NaiveBayes	Logistic	RandomForest	RandomTree	IBK	OneR	
CfsSubSetEval	class2	0.788	0.839	0.837	0.817	0.82	0.935	0.83933333
	class1	0.644	0.558	0.494	0.449	0.434	0.344	0.48716667
	Weight	0.739	0.744	0.721	0.693	0.689	0.736	0.72033333
CorrelationAttributeEval	class2	0.703	0.868	0.889	0.751	0.794	0.935	0.82333333
	class1	0.722	0.527	0.451	0.452	0.453	0.344	0.4915
	Weight	0.71	0.753	0.741	0.65	0.697	0.736	0.7145
OneRAttributeEval	class2	0.703	0.869	0.887	0.757	0.794	0.935	0.82416667
	class1	0.722	0.527	0.468	0.461	0.453	0.344	0.49583333
	Weight	0.71	0.753	0.746	0.657	0.679	0.736	0.7135
ClassifierAttributeEval	class2	0.703	0.868	0.897	0.761	0.794	0.935	0.82633333
	class1	0.722	0.527	0.463	0.456	0.453	0.344	0.49416667
	Weight	0.71	0.753	0.75	0.658	0.679	0.736	0.71433333

The confusion matrix table

CfsSubSetEval			
Confusion M	a	b	
NaiveBayes	2118	570	a=2
	487	882	b=1
Logistic	a	b	
	2256	432	
	605	764	
RandomFore	a	b	
	2251	437	
	693	676	
RandomTree	a	b	
	2197	491	
	755	614	
IBK	a	b	
	2203	485	
	775	594	
OneR	a	b	
	2514	174	
	898	471	

CorrelationAttributeEval			
Confusion M	a	b	
NaiveBayes	1891	797	a=2
	380	989	b=1
Logistic	a	b	
	2334	354	
	648	721	
RandomFore	a	b	
	2390	298	
	752	617	
RandomTree	a	b	
	2018	670	
	750	619	
IBK	a	b	
	2133	555	
	749	620	
OneR	a	b	
	2514	174	
	898	471	

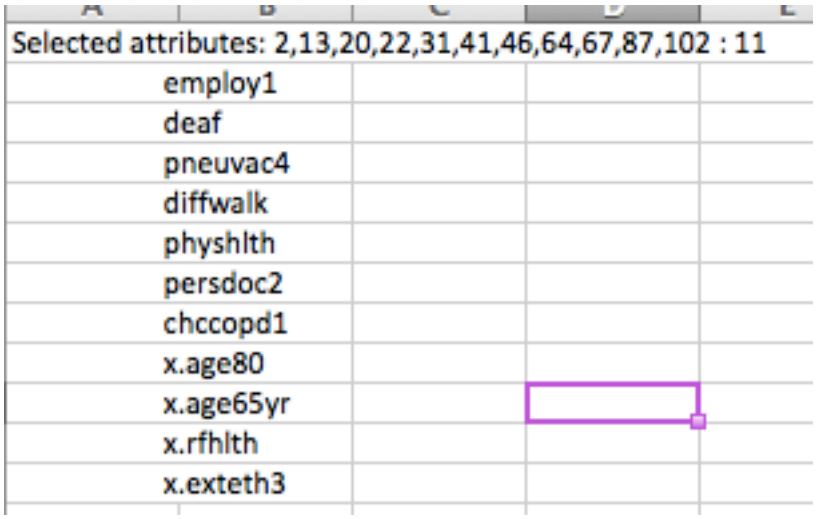
OneRAttributeEval			
Confusion M	a	b	
NaiveBayes	1891	797	a=2
	380	989	b=1
Logistic	a	b	
	2335	353	
	648	721	
RandomFore	a	b	
	2384	304	
	728	641	
RandomTree	a	b	
	2035	653	
	738	631	
IBK	a	b	
	2133	555	
	749	620	
OneR	a	b	
	2514	174	
	898	471	

ClassifierAttributeEval			
Confusion M	a	b	
NaiveBayes	1891	797	a=2
	380	989	b=1
Logistic	a	b	
	2334	354	
	648	721	
RandomFore	a	b	
	2410	278	
	735	634	
RandomTree	a	b	
	2045	643	
	745	624	
IBK	a	b	
	2133	555	
	749	620	
OneR	a	b	
	2514	174	
	898	471	

Part 3: Best Results and Discussion

The best test result I believe I achieved with all the 24 runs are ***with CfsSubSetEval Attribute selection and Logistic classifier.*** The specific detail for this test run is below. The reason why I believe this is the best run, is because out of all the test results, it gives me the accuracy rate of 74.4392 and incorrect instance of 25.5608. The reason why I believe that ***CfsSubSetEval Attribute selection and Logistic classifier*** is the best performance is because even though out of all 24 test, this one actually did not gave me the highest correct instance, I believe OneAttribute-Logistic actually has a higher percentage of correct instance. However, overall, the average correct instance of 72.0565% and average recall 0.83933 for CfSubSetEval is the highest out of all 4 attribute selection, and logistics is the highest out of the 6 classifier used with CfsSubaSetEval. As I mentioned before, I think for this project, if we are trying to test out the true positives for havarth3, then the test run with the highest recall rate should be the best performance test. Therefore I believe, in-combination of the attribute selection and classification methods CfsSubaSetEval –Logistic gives the best results.

```
Time taken to test model on supplied test set: 3.44 seconds

=== Summary ===

Correctly Classified Instances        3020               74.4392 %
Incorrectly Classified Instances      1037               25.5608 %
Kappa statistic                          0.4101
Mean absolute error                      0.3343
Root mean squared error                  0.4111
Relative absolute error                 74.7508 %
Root relative squared error             86.9479 %
Total Number of Instances             4057

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.839    0.442    0.789      0.839    0.813      0.412    0.801     0.881     2
                 0.558    0.161    0.639      0.558    0.596      0.412    0.801     0.654     1
Weighted Avg.    0.744    0.347    0.738      0.744    0.740      0.412    0.801     0.805

=== Confusion Matrix ===

    a    b   <-- classified as
 2256  432 |    a = 2
  605  764 |    b = 1
```



Part 4: Attribute and other Observations Discussions

From the above list of attribute, I believe that employ1, deaf, pneuvac4, diffwalk and physhlth are the 5 most relevant to the class attribute. First of all, when I was doing rank search method with other attribute selections, diffwalk and physhlth do rank fairly high on some of the other attribute selections, such as for **ClassififerAttributeEva**l etc. Also for my best performing model, these are the top 5 attribute.

What I learned from this project is that, for all the attribute selections I have chosen, ***CfsSubSetEval*** actually has the least attribute, but it actually gaves the best overall performance compared to other methods. It is easy to make assumption that the more attribute you have, or the "more data" you have, you should be able to make better predication, but my results from this project proves this is not true. I believe the reason

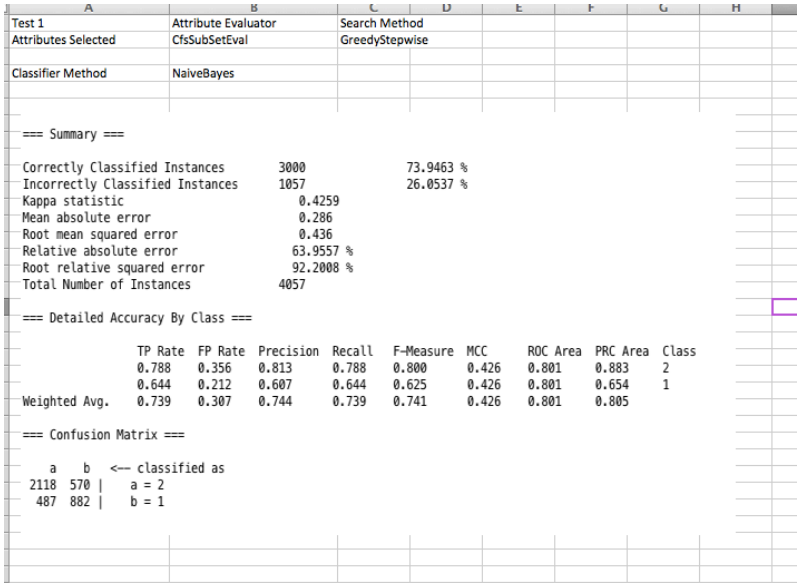
for this, is when you add more attribute or “more data” when you are doing predictive modeling, the large amount of less relevant attribute or data can actually become “noise” that would negatively affect your predication outcome. For better performance, it is actually better just to find out and select the most relevant attribute or data (the relevant data must be correct though, otherwise making prediction with incorrect smaller amount data can cause bigger mistakes) , this way it actually reduced the “noise” in the data or with smaller relevant data, it even reduced the possibility of bad data with error.

There are some other interesting things I noticed in my test results; For OneR classifer, my output results and my confusion matrix are the same across the board. At first I thought this is a mistake in the way I processed the test runs, but no matter how I adjust the methods, I always got the same results for OneR. I think this is due to how OneR is processed, that it will always use only one rule with the smallest total error. Since I am running test on the same BRFSS data over and over again, this is probably why I am getting the same results for all OneR test runs.

I also noticed that RandomTree and IBK produced the worst performance out of the 6 classifier I used. I think this could be there are many attributes (more than 10) when I am processing the test runs. However RandomTree and IBK(KNN) are better to be used to predicate smaller amount of attributes.

Part 5: All results screenshots of all Weka Output Window and selected attributes.

CfsSubSetEval



Test 1	Attribute Evaluator	Search Method							
Attributes Selected	CfsSubSetEval	GreedyStepwise							
Classifier Method	NaiveBayes								

```
=== Summary ===

Correctly Classified Instances        3000               73.9463 %
Incorrectly Classified Instances      1057               26.0537 %
Kappa statistic                          0.4259
Mean absolute error                      0.286
Root mean squared error                  0.436
Relative absolute error                 63.9557 %
Root relative squared error             92.2008 %
Total Number of Instances             4057

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.788    0.356    0.813      0.788    0.800      0.426    0.801     0.883     2
               0.644    0.212    0.607      0.644    0.625      0.426    0.801     0.654     1
Weighted Avg.  0.739    0.307    0.744      0.739    0.741      0.426    0.801     0.805

=== Confusion Matrix ===

    a    b   <-- classified as
 2118  570 |    a = 2
  487  882 |    b = 1
```

Test 1	Attribute Eva	Search Method
Attributes Se	CfsSubSetEva	GreedyStepwise
Classifier Me	Logistic	

```
Time taken to test model on supplied test set: 3.44 seconds

=== Summary ===

Correctly Classified Instances        3020               74.4392 %
Incorrectly Classified Instances      1037               25.5608 %
Kappa statistic                          0.4101
Mean absolute error                      0.3343
Root mean squared error                  0.4111
Relative absolute error                 74.7508 %
Root relative squared error             86.9479 %
Total Number of Instances             4057

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.839    0.442    0.789      0.839    0.813      0.412    0.801     0.881     2
                 0.558    0.161    0.639      0.558    0.596      0.412    0.801     0.654     1
Weighted Avg.    0.744    0.347    0.738      0.744    0.740      0.412    0.801     0.805

=== Confusion Matrix ===

    a    b   <-- classified as
 2256  432 |    a = 2
  605  764 |    b = 1
```

Test 1	Attribute Eva	Search Method
Attributes Se	CfsSubSetEva	GreedyStepwise
Classifier Me	RandomForest	

```
=== Evaluation on test set ===

Time taken to test model on supplied test set: 6.41 seconds

=== Summary ===

Correctly Classified Instances        2927               72.1469 %
Incorrectly Classified Instances      1130               27.8531 %
Kappa statistic                          0.3471
Mean absolute error                      0.3423
Root mean squared error                  0.4265
Relative absolute error                 76.5485 %
Root relative squared error             90.2028 %
Total Number of Instances             4057

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.837    0.506    0.765      0.837    0.799      0.351    0.768     0.861     2
                 0.494    0.163    0.607      0.494    0.545      0.351    0.768     0.609     1
Weighted Avg.    0.721    0.390    0.712      0.721    0.713      0.351    0.768     0.776

=== Confusion Matrix ===

    a    b   <-- classified as
 2251  437 |    a = 2
  693  676 |    b = 1
```

Test 1	Attribute Eva	Search Method
Attributes Se	CfsSubSetEva	GreedyStepwise
Classifier Me	RandomTree	

```
=== Evaluation on test set ===

Time taken to test model on supplied test set: 3.66 seconds

=== Summary ===

Correctly Classified Instances        2811               69.2877 %
Incorrectly Classified Instances      1246               30.7123 %
Kappa statistic                          0.279
Mean absolute error                      0.3442
Root mean squared error                  0.4867
Relative absolute error                 76.9741 %
Root relative squared error            102.9319 %
Total Number of Instances             4057

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.817    0.551    0.744      0.817    0.779      0.282    0.687     0.772     2
                 0.449    0.183    0.556      0.449    0.496      0.282    0.687     0.522     1
Weighted Avg.    0.693    0.427    0.681      0.693    0.684      0.282    0.687     0.688

=== Confusion Matrix ===

    a    b   <-- classified as
 2197  491 |    a = 2
  755  614 |    b = 1
```

Test 1	Attribute Eva	Search Method
Attributes Se	CfsSubSetEva	GreedyStepwise
Classifier Me	IBK	

```
Time taken to test model on supplied test set: 7.41 seconds

=== Summary ===

Correctly Classified Instances        2797               68.9426 %
Incorrectly Classified Instances      1260               31.0574 %
Kappa statistic                          0.2674
Mean absolute error                      0.3339
Root mean squared error                  0.5241
Relative absolute error                 74.6691 %
Root relative squared error            110.8434 %
Total Number of Instances             4057

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.820    0.566    0.740      0.820    0.778      0.271    0.682     0.786     2
                 0.434    0.180    0.551      0.434    0.485      0.271    0.682     0.496     1
Weighted Avg.    0.689    0.436    0.676      0.689    0.679      0.271    0.682     0.688

=== Confusion Matrix ===

    a    b   <-- classified as
 2203  485 |    a = 2
  775  594 |    b = 1
```

Test 1	Attribute Eva	Search Method
Attributes Se	CfsSubSetEva	GreedyStepwise
Classifier Me	OneR	

```
Time taken to test model on supplied test set: 3.88 seconds

=== Summary ===

Correctly Classified Instances        2985               73.5765 %
Incorrectly Classified Instances      1072               26.4235 %
Kappa statistic                          0.321
Mean absolute error                      0.2642
Root mean squared error                  0.514
Relative absolute error                 59.0897 %
Root relative squared error            108.7135 %
Total Number of Instances             4057

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.935    0.656    0.737      0.935    0.824      0.361    0.640     0.732     2
                 0.344    0.065    0.730      0.344    0.468      0.361    0.640     0.473     1
Weighted Avg.    0.736    0.456    0.735      0.736    0.704      0.361    0.640     0.644

=== Confusion Matrix ===

    a    b   <-- classified as
 2514  174 |    a = 2
  898  471 |    b = 1
```

Selected attributes: 2,13,20,22,31,41,46,64,67,87,102 : 11
employ1
deaf
pneuvac4
diffwalk
physhlth
persdoc2
chccopd1
x.age80
x.age65yr
x.rfhlth
x.exteth3

CorrelationAttributeEval

A	B	C	D	E	F	G	H	I
Test 1	Attribute Evaluator	Search Method						
Attributes Selected	CorrelationAttributeEval	Ranker						
Classifier Method	NaiveBayes							

```
=== Summary ===

Correctly Classified Instances        2880               70.9884 %
Incorrectly Classified Instances      1177               29.0116 %
Kappa statistic                          0.3963
Mean absolute error                      0.2914
Root mean squared error                  0.5131
Relative absolute error                 65.1718 %
Root relative squared error            108.512  %
Total Number of Instances             4057

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.703    0.278    0.833      0.703    0.763      0.406    0.776     0.859     2
                 0.722    0.297    0.554      0.722    0.627      0.406    0.776     0.591     1
Weighted Avg.    0.710    0.284    0.739      0.710    0.717      0.406    0.776     0.769

=== Confusion Matrix ===

    a    b   <-- classified as
 1891  797 |    a = 2
  380  989 |    b = 1
```

Test 1	Attribute Eva	Search Method
Attributes Se	CorrelationA	Ranker
Classifier Me	Logistic	

```
Time taken to test model on supplied test set: 0.13 seconds

=== Summary ===

Correctly Classified Instances        3055               75.3019 %
Incorrectly Classified Instances      1002               24.6981 %
Kappa statistic                          0.4169
Mean absolute error                      0.3203
Root mean squared error                  0.4088
Relative absolute error                 71.6344 %
Root relative squared error             86.4607 %
Total Number of Instances             4057

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.868    0.473    0.783      0.868    0.823      0.423    0.804     0.880     2
                 0.527    0.132    0.671      0.527    0.590      0.423    0.804     0.655     1
Weighted Avg.    0.753    0.358    0.745      0.753    0.745      0.423    0.804     0.804

=== Confusion Matrix ===

    a    b   <-- classified as
 2334  354 |    a = 2
  648  721 |    b = 1
```

Test 1	Attribute Eva	Search Method
Attributes Se	CorrelationA	Ranker
Classifier Me	RandomForest	

```
=== Summary ===

Correctly Classified Instances        3007               74.1188 %
Incorrectly Classified Instances      1050               25.8812 %
Kappa statistic                          0.3699
Mean absolute error                      0.353
Root mean squared error                  0.4109
Relative absolute error                 78.947  %
Root relative squared error             86.8934 %
Total Number of Instances             4057

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.889    0.549    0.761      0.889    0.820      0.384    0.804     0.886     2
                 0.451    0.111    0.674      0.451    0.540      0.384    0.804     0.661     1
Weighted Avg.    0.741    0.401    0.732      0.741    0.726      0.384    0.804     0.810

=== Confusion Matrix ===

    a    b   <-- classified as
 2390  298 |    a = 2
  752  617 |    b = 1
```

Attribute Eva	Search Method	
CorrelationA	Ranker	
RandomTree		

```
ctly Classified Instances        2637               64.9988 %
rectly Classified Instances      1420               35.0012 %
 statistic                          0.2059
absolute error                      0.3575
mean squared error                  0.5762
ive absolute error                 79.9468 %
relative squared error            121.8664 %
 Number of Instances              4057

etailed Accuracy By Class ===

             TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
             0.751    0.548    0.729      0.751    0.740      0.206    0.610     0.722     2
             0.452    0.249    0.480      0.452    0.466      0.206    0.608     0.414     1
ted Avg.     0.650    0.447    0.645      0.650    0.647      0.206    0.610     0.618

onfusion Matrix ===

    b   <-- classified as
  670 |    a = 2
  619 |    b = 1
```

Test 1	Attribute Eva	Search Method
Attributes Se	CorrelationA	Ranker
Classifier Me	IBK	

```
=== Summary ===

Correctly Classified Instances        2753               67.858  %
Incorrectly Classified Instances      1304               32.142  %
Kappa statistic                          0.2553
Mean absolute error                      0.3215
Root mean squared error                  0.5669
Relative absolute error                 71.8879 %
Root relative squared error            119.8864 %
Total Number of Instances             4057

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.794    0.547    0.740      0.794    0.766      0.257    0.623     0.724     2
                 0.453    0.206    0.528      0.453    0.487      0.257    0.623     0.424     1
Weighted Avg.    0.679    0.432    0.668      0.679    0.672      0.257    0.623     0.623

=== Confusion Matrix ===

   a    b   <-- classified as
2133  555 |    a = 2
 749  620 |    b = 1
```

Test 1	Attribute Eva	Search Method
Attributes Se	CorrelationA	Ranker
Classifier Me	OneR	

```
=== Summary ===

Correctly Classified Instances        2985               73.5765 %
Incorrectly Classified Instances      1072               26.4235 %
Kappa statistic                          0.321
Mean absolute error                      0.2642
Root mean squared error                  0.514
Relative absolute error                 59.0897 %
Root relative squared error            108.7135 %
Total Number of Instances             4057

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.935    0.656    0.737      0.935    0.824      0.361    0.640     0.732     2
                 0.344    0.065    0.730      0.344    0.468      0.361    0.640     0.473     1
Weighted Avg.    0.736    0.456    0.735      0.736    0.704      0.361    0.640     0.644

=== Confusion Matrix ===

   a    b   <-- classified as
2514  174 |    a = 2
 898  471 |    b = 1
```

Attribute Evaluator (supervised, Class (nominal): 108 havarth3):		
Correlation Ranking Filter		
		Selected attributes:
		64,66,22,2,67,87,97,102,20,46,
		95,31,6,25,62,13,29,104,24,69,
		27,53,59,43,26,45,98,44,47,14,
		41,16,36,48,52,105,49,34,106,
		50,78,103,76,84,10,57,61,68,8
		8,8,86,21,33,80,93,9,90,60,58,
		30,71,63,15,54,11,91,81,65,82,
		89,85,56,35,19,55,7,70,12,96,3
		2,42,77,51,100,101,94,28,5,99,
		72,75,18,1,83,74,73,79,4,3,107
Ranked attributes:		,39,40,92,38,17,23,37 : 107

OneRAttributeEval

Test 1	Attribute Evaluator	Search Method
Attributes Selected	OneRAttributeEval	Ranker
Classifier Method	NaiveBayes	

```
=== Summary ===

Correctly Classified Instances        2880               70.9884 %
Incorrectly Classified Instances      1177               29.0116 %
Kappa statistic                          0.3963
Mean absolute error                      0.2914
Root mean squared error                  0.5131
Relative absolute error                 65.1718 %
Root relative squared error            108.512  %
Total Number of Instances             4057

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Clas
                 0.703    0.278    0.833      0.703    0.763      0.406    0.776     0.859     2
                 0.722    0.297    0.554      0.722    0.627      0.406    0.776     0.591     1
Weighted Avg.    0.710    0.284    0.739      0.710    0.717      0.406    0.776     0.769

=== Confusion Matrix ===

    a    b   <-- classified as
 1891  797 |    a = 2
  380  989 |    b = 1
```

Test 1	Attribute Eva	Search Method
Attributes Se	OneRAttribu	Ranker
Classifier Me	RandomForest	

```
=== Summary ===

Correctly Classified Instances        3025               74.5625 %
Incorrectly Classified Instances      1032               25.4375 %
Kappa statistic                          0.3843
Mean absolute error                      0.355
Root mean squared error                  0.4119
Relative absolute error                 79.3849 %
Root relative squared error             87.1211 %
Total Number of Instances             4057

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Clas
                 0.887    0.532    0.766      0.887    0.822      0.397    0.802     0.885     2
                 0.468    0.113    0.678      0.468    0.554      0.397    0.802     0.661     1
Weighted Avg.    0.746    0.390    0.736      0.746    0.732      0.397    0.802     0.809

=== Confusion Matrix ===

    a    b   <-- classified as
 2384  304 |    a = 2
  728  641 |    b = 1
```

Test 1	Attribute Eva	Search Method
Attributes Se	OneRAttribu	Ranker
Classifier Me	Logistic	

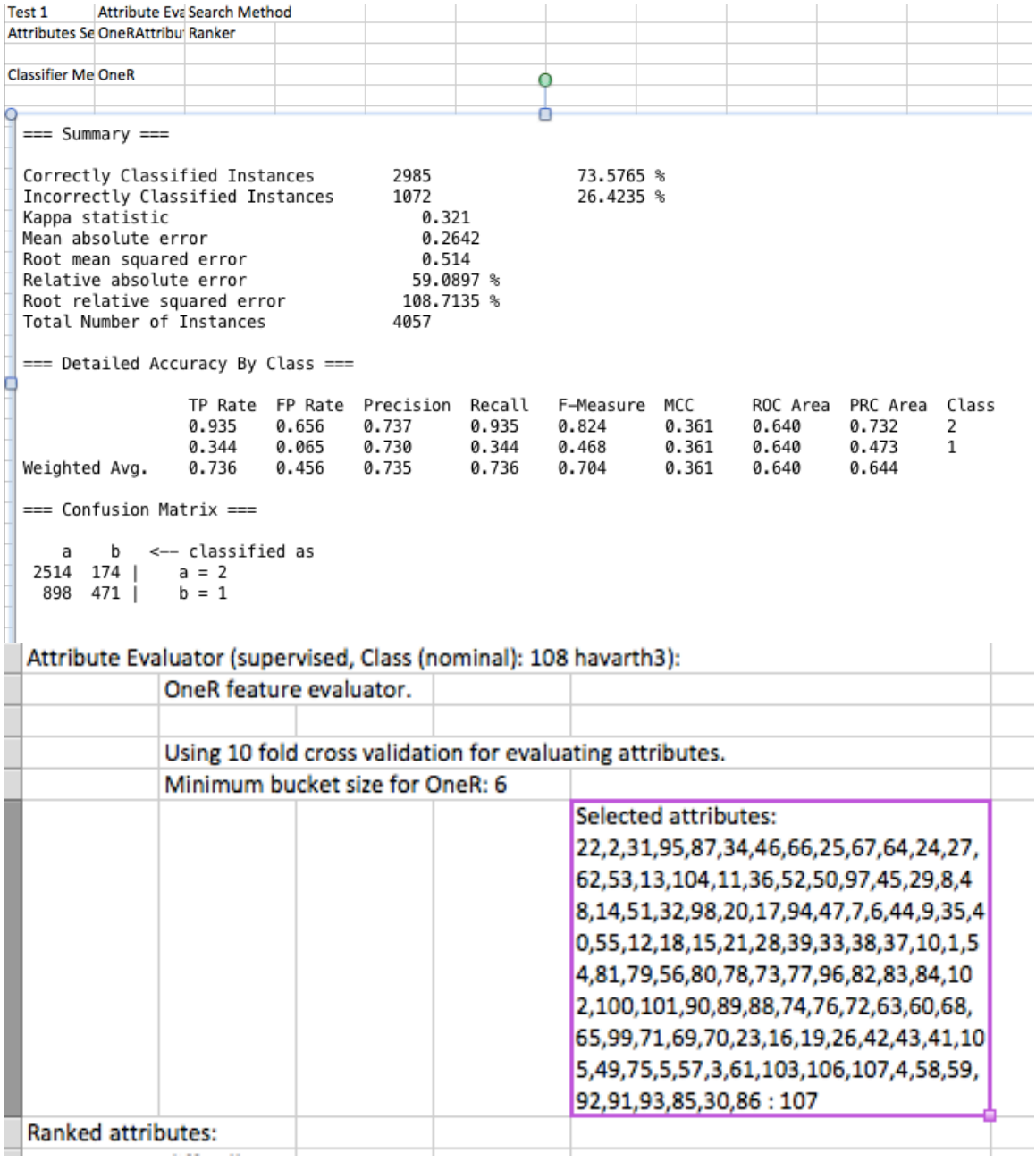
```
=== Summary ===

Correctly Classified Instances        3056               75.3266 %
Incorrectly Classified Instances      1001               24.6734 %
Kappa statistic                          0.4174
Mean absolute error                      0.3203
Root mean squared error                  0.4088
Relative absolute error                 71.6347 %
Root relative squared error             86.4604 %
Total Number of Instances             4057

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Clas
                 0.869    0.473    0.783      0.869    0.823      0.424    0.804     0.880     2
                 0.527    0.131    0.671      0.527    0.590      0.424    0.804     0.655     1
Weighted Avg.    0.753    0.358    0.745      0.753    0.745      0.424    0.804     0.804

=== Confusion Matrix ===

    a    b   <-- classified as
 2335  353 |    a = 2
  648  721 |    b = 1
```

Test 1	Attribute Eva	Search Method
Attributes Se	OneRAttribu	Ranker
Classifier Me	RandomTree	

```
=== Summary ===

Correctly Classified Instances        2666               65.7136 %
Incorrectly Classified Instances      1391               34.2864 %
Kappa statistic                          0.2214
Mean absolute error                      0.3505
Root mean squared error                  0.5697
Relative absolute error                 78.3826 %
Root relative squared error            120.4941 %
Total Number of Instances             4057

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.757    0.539    0.734      0.757    0.745      0.222    0.616     0.724     2
                 0.461    0.243    0.491      0.461    0.476      0.222    0.614     0.420     1
Weighted Avg.    0.657    0.439    0.652      0.657    0.654      0.222    0.615     0.622

=== Confusion Matrix ===

    a    b   <-- classified as
 2035  653 |    a = 2
  738  631 |    b = 1
```

Attributes Se	OneRAttribu	Ranker
Classifier Me	IBK	

```
=== Summary ===

Correctly Classified Instances        2753               67.858  %
Incorrectly Classified Instances      1304               32.142  %
Kappa statistic                          0.2553
Mean absolute error                      0.3215
Root mean squared error                  0.5669
Relative absolute error                 71.8879 %
Root relative squared error            119.8864 %
Total Number of Instances             4057

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.794    0.547    0.740      0.794    0.766      0.257    0.623     0.724     2
                 0.453    0.206    0.528      0.453    0.487      0.257    0.623     0.424     1
Weighted Avg.    0.679    0.432    0.668      0.679    0.672      0.257    0.623     0.623

=== Confusion Matrix ===

    a    b   <-- classified as
 2133  555 |    a = 2
  749  620 |    b = 1
```



Test 1	Attribute Eva	Search Method
Attributes Se	OneRAttribu	Ranker
Classifier Me	OneR	

```
=== Summary ===

Correctly Classified Instances        2985               73.5765 %
Incorrectly Classified Instances      1072               26.4235 %
Kappa statistic                          0.321
Mean absolute error                      0.2642
Root mean squared error                  0.514
Relative absolute error                 59.0897 %
Root relative squared error            108.7135 %
Total Number of Instances             4057

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.935    0.656    0.737      0.935    0.824      0.361    0.640     0.732     2
                 0.344    0.065    0.730      0.344    0.468      0.361    0.640     0.473     1
Weighted Avg.    0.736    0.456    0.735      0.736    0.704      0.361    0.640     0.644

=== Confusion Matrix ===

    a    b   <-- classified as
 2514  174 |    a = 2
  898  471 |    b = 1
```

Attribute Evaluator (supervised, Class (nominal): 108 havarth3):

OneR feature evaluator.

Using 10 fold cross validation for evaluating attributes.

Minimum bucket size for OneR: 6

Selected attributes:
22,2,31,95,87,34,46,66,25,67,64,24,27,
62,53,13,104,11,36,52,50,97,45,29,8,4
8,14,51,32,98,20,17,94,47,7,6,44,9,35,4
0,55,12,18,15,21,28,39,33,38,37,10,1,5
4,81,79,56,80,78,73,77,96,82,83,84,10
2,100,101,90,89,88,74,76,72,63,60,68,
65,99,71,69,70,23,16,19,26,42,43,41,10
5,49,75,5,57,3,61,103,106,107,4,58,59,
92,91,93,85,30,86 : 107

Ranked attributes:

ClassifierAttributeEva

Test 1	Attribute Evaluator	Search Method
Attributes Selected	ClassifierAttributeEval	Ranker
Classifier Method	NaiveBayes	

```
=== Summary ===

Correctly Classified Instances        2880               70.9884 %
Incorrectly Classified Instances      1177               29.0116 %
Kappa statistic                          0.3963
Mean absolute error                      0.2914
Root mean squared error                  0.5131
Relative absolute error                 65.1719 %
Root relative squared error            108.512  %
Total Number of Instances             4057

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.703    0.278    0.833      0.703    0.763      0.406    0.776     0.859     2
                 0.722    0.297    0.554      0.722    0.627      0.406    0.776     0.591     1
Weighted Avg.    0.710    0.284    0.739      0.710    0.717      0.406    0.776     0.769

=== Confusion Matrix ===

    a    b   <-- classified as
 1891  797 |    a = 2
  380  989 |    b = 1
```

Test 1	Attribute Eva	Search Method
Attributes Se	ClassifierAtt	Ranker
Classifier Me	Logistic	

```
=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.16 seconds

=== Summary ===

Correctly Classified Instances        3055               75.3019 %
Incorrectly Classified Instances      1002               24.6981 %
Kappa statistic                          0.4169
Mean absolute error                      0.3203
Root mean squared error                  0.4088
Relative absolute error                 71.6345 %
Root relative squared error             86.4605 %
Total Number of Instances             4057

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.868    0.473    0.783      0.868    0.823      0.423    0.804     0.880     2
                 0.527    0.132    0.671      0.527    0.590      0.423    0.804     0.655     1
Weighted Avg.    0.753    0.358    0.745      0.753    0.745      0.423    0.804     0.804

=== Confusion Matrix ===

    a    b   <-- classified as
 2334  354 |    a = 2
  648  721 |    b = 1
```

Test 1 Attribute Eva Search Method
Attributes Se ClassifierAtt Ranker

Classifier Me RandomForest

```
=== Summary ===

Correctly Classified Instances        3044               75.0308 %
Incorrectly Classified Instances      1013               24.9692 %
Kappa statistic                          0.3918
Mean absolute error                      0.3538
Root mean squared error                  0.4104
Relative absolute error                 79.1157 %
Root relative squared error             86.7983 %
Total Number of Instances             4057

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.897    0.537    0.766      0.897    0.826      0.407    0.806     0.886     2
                 0.463    0.103    0.695      0.463    0.556      0.407    0.806     0.667     1
Weighted Avg.    0.750    0.391    0.742      0.750    0.735      0.407    0.806     0.813

=== Confusion Matrix ===

    a    b   <-- classified as
 2410  278 |    a = 2
  735  634 |    b = 1
```

Test 1 Attribute Eva Search Method
Attributes Se ClassifierAtt Ranker

Classifier Me RandomTree

```
=== Summary ===

Correctly Classified Instances        2669               65.7875 %
Incorrectly Classified Instances      1388               34.2125 %
Kappa statistic                          0.2206
Mean absolute error                      0.3515
Root mean squared error                  0.5631
Relative absolute error                 78.6136 %
Root relative squared error            119.0905 %
Total Number of Instances             4057

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Clas
                 0.761    0.544    0.733      0.761    0.747      0.221    0.619     0.728     2
                 0.456    0.239    0.493      0.456    0.473      0.221    0.619     0.424     1
Weighted Avg.    0.658    0.441    0.652      0.658    0.654      0.221    0.619     0.625

=== Confusion Matrix ===

    a    b   <-- classified as
 2045  643 |    a = 2
  745  624 |    b = 1
```

Test 1	Attribute Eva	Search Method
Attributes Se	ClassifierAtt	Ranker
Classifier Me	IBK	

```
=== Summary ===

Correctly Classified Instances        2753               67.858  %
Incorrectly Classified Instances      1304               32.142  %
Kappa statistic                          0.2553
Mean absolute error                      0.3215
Root mean squared error                  0.5669
Relative absolute error                 71.8879 %
Root relative squared error            119.8864 %
Total Number of Instances             4057

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Cla
                 0.794    0.547    0.740      0.794    0.766      0.257    0.623     0.724     2
                 0.453    0.206    0.528      0.453    0.487      0.257    0.623     0.424     1
Weighted Avg.    0.679    0.432    0.668      0.679    0.672      0.257    0.623     0.623

=== Confusion Matrix ===

    a    b   <-- classified as
 2133  555 |    a = 2
  749  620 |    b = 1
```

Test 1	Attribute Eva	Search Method
Attributes Se	ClassifierAtt	Ranker
Classifier Me	OneR	

```
=== Summary ===

Correctly Classified Instances        2985               73.5765 %
Incorrectly Classified Instances      1072               26.4235 %
Kappa statistic                          0.321
Mean absolute error                      0.2642
Root mean squared error                  0.514
Relative absolute error                 59.0897 %
Root relative squared error            108.7135 %
Total Number of Instances             4057

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.935    0.656    0.737      0.935    0.824      0.361    0.640     0.732     2
                 0.344    0.065    0.730      0.344    0.468      0.361    0.640     0.473     1
Weighted Avg.    0.736    0.456    0.735      0.736    0.704      0.361    0.640     0.644

=== Confusion Matrix ===

    a    b   <-- classified as
 2514  174 |    a = 2
  898  471 |    b = 1
```

Attribute Evaluator (supervised, Class (nominal): 108 havarth3):			
	Classifier feature evaluator		
	Using	Wrapper Subset Evaluator	
	Learning scheme: weka.classifiers.rules.ZeroR		
	Scheme options:		
	Subset evaluation: classification accuracy		
	Number of folds for accuracy estimation: 5		
		Selected attributes: 107,34,36,37,38,35,33,27,32, 29,30,31,39,40,41,42,49,50,5 1,48,47,46,43,44,45,28,26,53, 7,9,10,11,8,6,25,5,2,3,4,12,13 ,14,15,22,23,24,21,20,19,16,1 7,18,52,54,106,88,90,91,92,8 9,87,81,86,83,84,85,93,94,95, 96,103,104,105,102,101,100, 97,98,99,82,80,55,61,63,64,6 5,62,60,79,59,56,57,58,66,67, 68,69,76,77,78,75,74,73,70,7 1,72,1 : 107	
Ranked attributes:			