Final Project

Introduction

The goal of this project assignment is to predict the Arthritis in patients having various medical conditions from the dataset which is part of the 2018 BRFSS Survey Data prepared by CDC.

In the course of this assignment, 4 Attribute Selection methods will be used and 6 classifier models will be evaluated for their accuracies for each of the 4 attribute selection methods.

Finally, the based on the highest accuracy determined, best attribute selection method and classification model will be presented with their accuracy.

Outcome of the Experiment

Based on the experiment run over 6 different classification algorithms over 4 different attribute selection methods, we finally arrive at the below accuracies of the models for each attribute selection method.

Steps performed in the experiment

- Pre-requisites
  a. The dataset was split into training and test data sets, keeping the distribution of the class variable same in both training and test data set. Data set was split having -
     Training Data Tuple Count = 7876
     Testing Data Tuple Count = 4057
  b. Attributes of the dataset was examined, and missing values were not altered.
- Over 4 iterations, below actions were performed –
  a. Run an attribute selection method to select 20 attributes out of 108 attributes
  b. Reduce the training and testing data set to data sets with 20 attributes each.
  c. Run 6 classification algorithms to determine the accuracy of them.

Results

Based on the experiment performed, **J48 using One R Classifier attribute selection method** was found to give the best accuracy of 74%. Following is the comparison of the accuracies, obtained from the experiment (outputs of each of the classifiers are presented below) -

Attribute Selection method	**Accuracy (%)**					
	Naïve Bayes	**J48**	**Neural Network**	**One R**	**Random Forest**	**KNN**
Correlation based feature subset selection	72.4	72.91	72.54	73.57	72.03	67.3
One R Classifier Attribute Evaluation	71.97	74.16	70.88	73.57	71.06	68.25
Correlation based on Pearson coefficient	70.84	74.14	72.76	73.57	72.41	70.01
Gain Ratio Attribute Evaluation	70.71	74.16	72.29	73.57	72.04	70.64

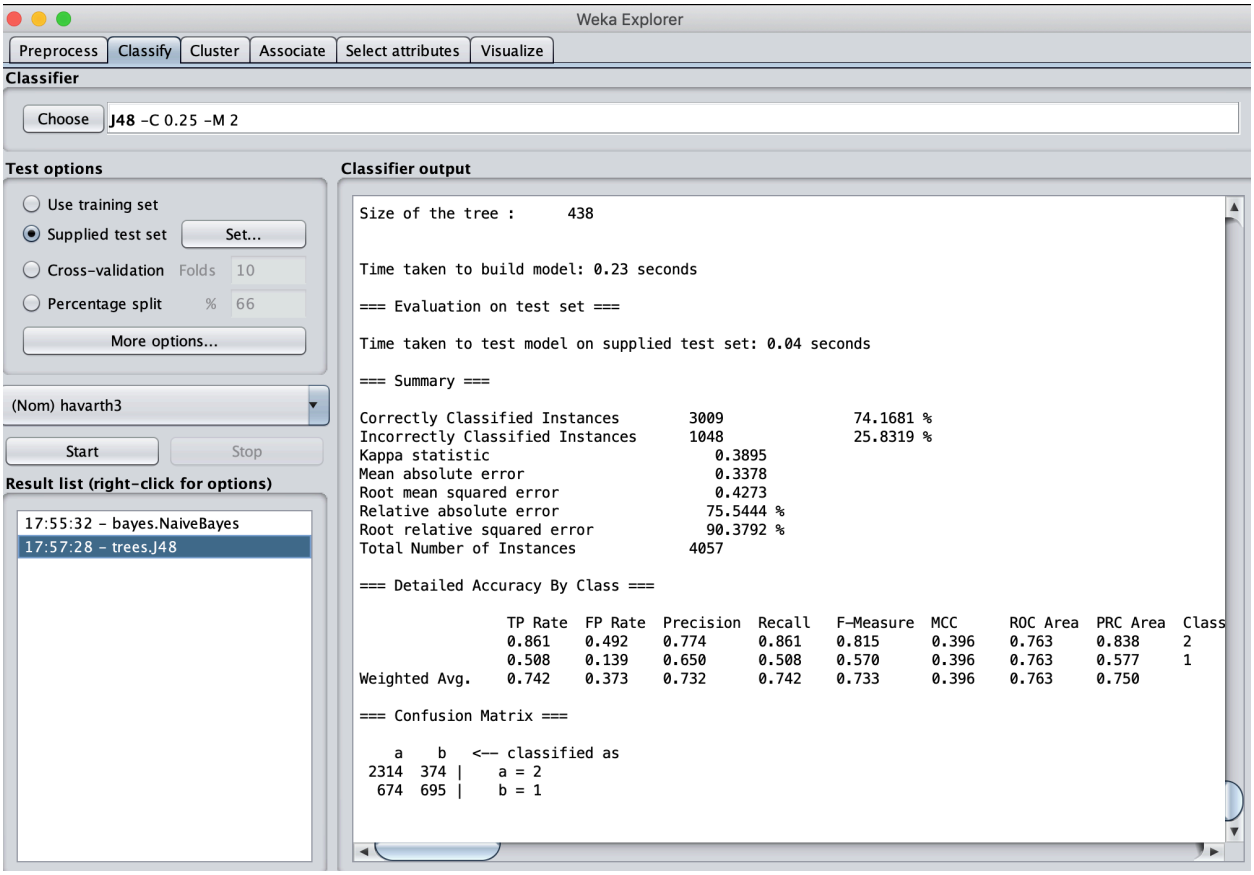
Why J48 with One R Attribute Selection method was chosen?

As we can see the from above grid, we got the maximum accuracy of 74% from this algorithm and attribute selection method combined, as compared to other algorithm and attribute selection methods.

I chose Accuracy as the measure for determining the best algorithm, as the goal for my experiment was to find the classifier which predicts the correct class labels (i.e. patient having arthritis, based on health metrics) for the given data set.

Best training set, test set are available with this submission.

Here are the test results from this classification algorithm –



Confusion Matrix

Actual Labels	Predicted Labels	
	1	2
1 = Diagnosed with arthritis	695 (TP)	674 (FN)
2 = Not diagnosed with arthritis	374(FP)	2314(TN)

Performance measures*
Accuracy = 74.16%
Error rate = 25.83%
Sensitivity = 50.76%
Specificity = 86.08%
F measure = 57.01%

*Note: The Weka output considers not diagnosed as Positive (and vise versa), which returns an inverted confusion matrix.

Five Most relevant attributes

From the 4 attribute selection methods run on the dataset, below attributes were common in all the selection methods –

Attribute Name	Attribute Description
chccopd1	have chronic obstructive pulmonary disease, C.O.P.D., emphysema or chronic bronchitis
diffdres	Do you have difficulty dressing or bathing?
x.age80	Imputed Age value collapsed above 80
physhlth	Number of Days Physical Health Not Good
employ1	Employment Status
diffwalk	Difficulty Walking or Climbing Stairs

Learning from this project

From this experiment, I learnt -

- A complete set of steps involved in data mining.
- Splitting data into training and test data set, maintaining similar class label distribution in both data set
- Became familiar with Attribute selection using Ranking in Weka was also a new learning.
- Comparing the impact of the attribute selection on a predictive model.

Observation from this project

Out of the many observations with the Weka tooling options to run pre-processing, attribute selection, classification, the key observation was the impact of attribute selection based on the ranking. Out of the 4 attribute selection methods used, 3 of them used the Ranking –

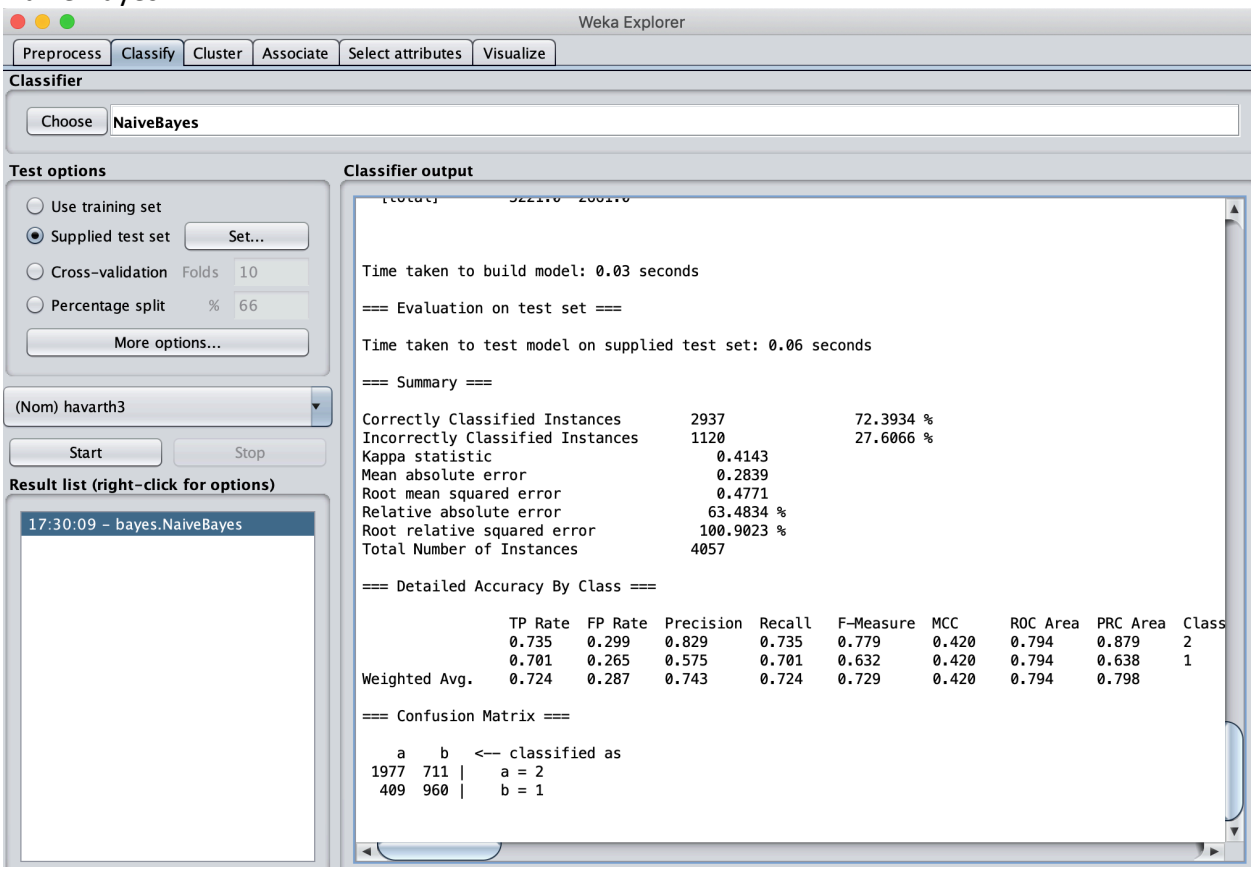
a. One R Classifier Attribute Evaluation
b. Correlation based on Pearson coefficient
c. Gain Ratio Attribute Evaluation

Attribute Selection Method – Correlation based feature subset selection
Attributes selected –

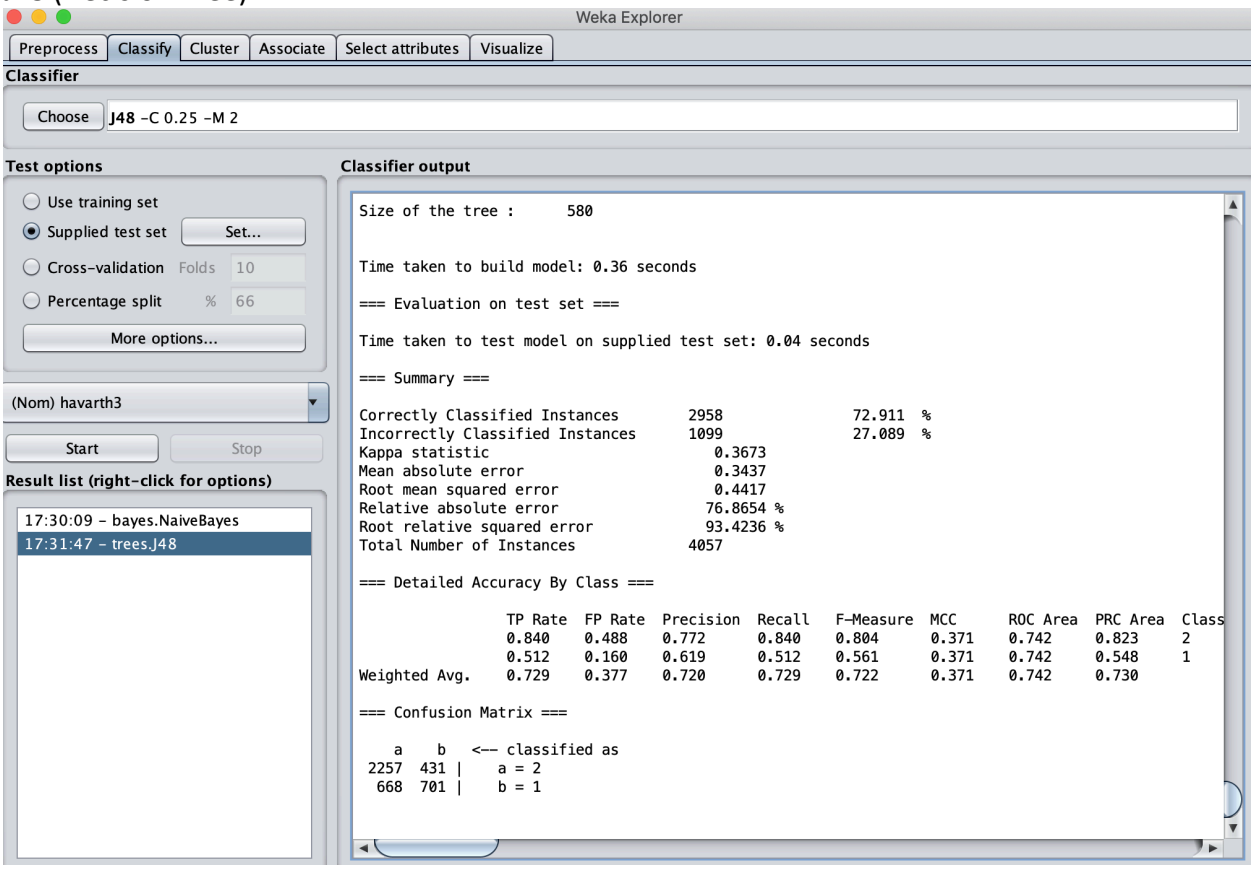
Index	Attribute Name
2	employ1
6	children
13	deaf
20	pneuvac4
22	diffwalk
24	diffdres
29	diabete3
31	physhlth
34	genhlth
41	persdoc2
43	checkup1
45	chcocncr
46	chccopd1
53	cvdcrhd4
62	x.age.g
64	x.age80
67	x.age65yr
87	x.rfhlth
102	x.exteth3

Classification Models –

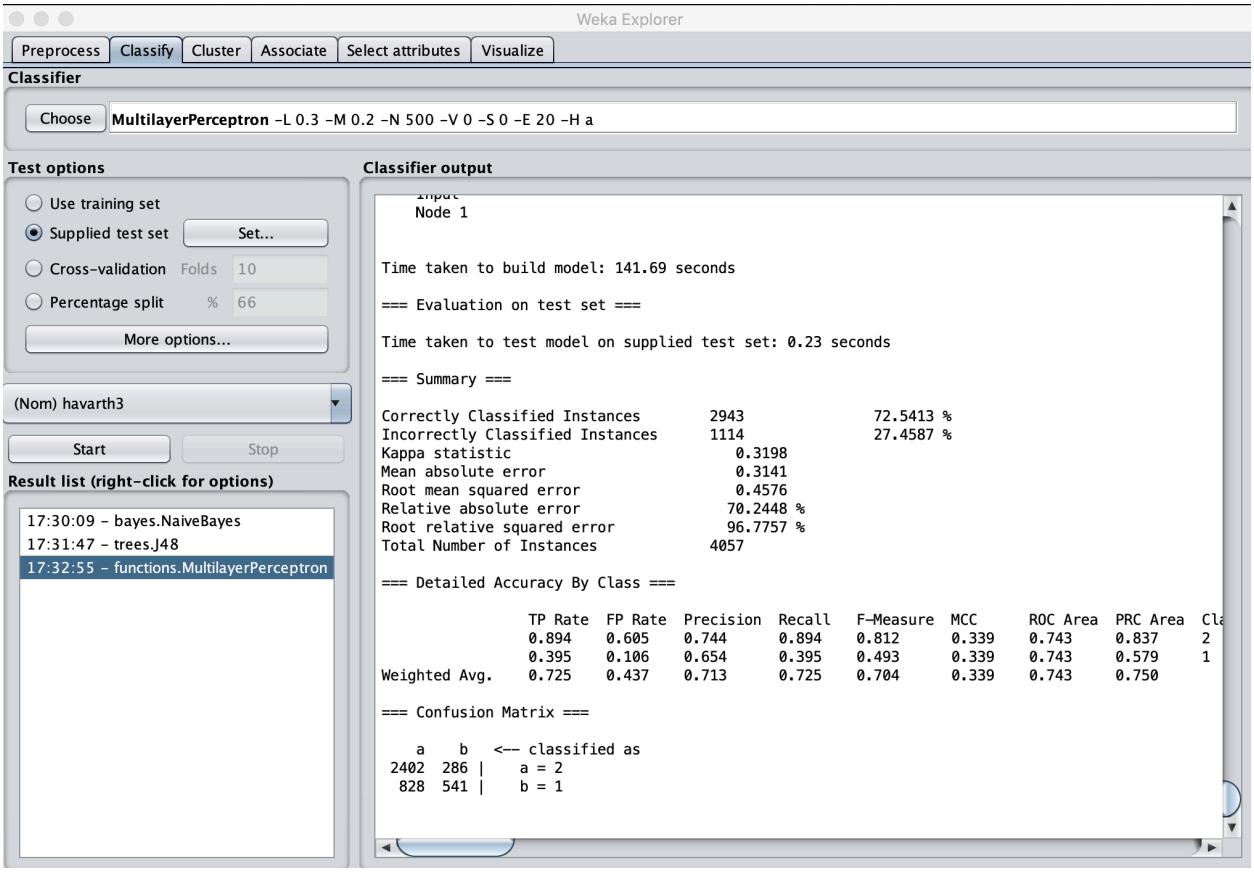
1. Naïve Bayes



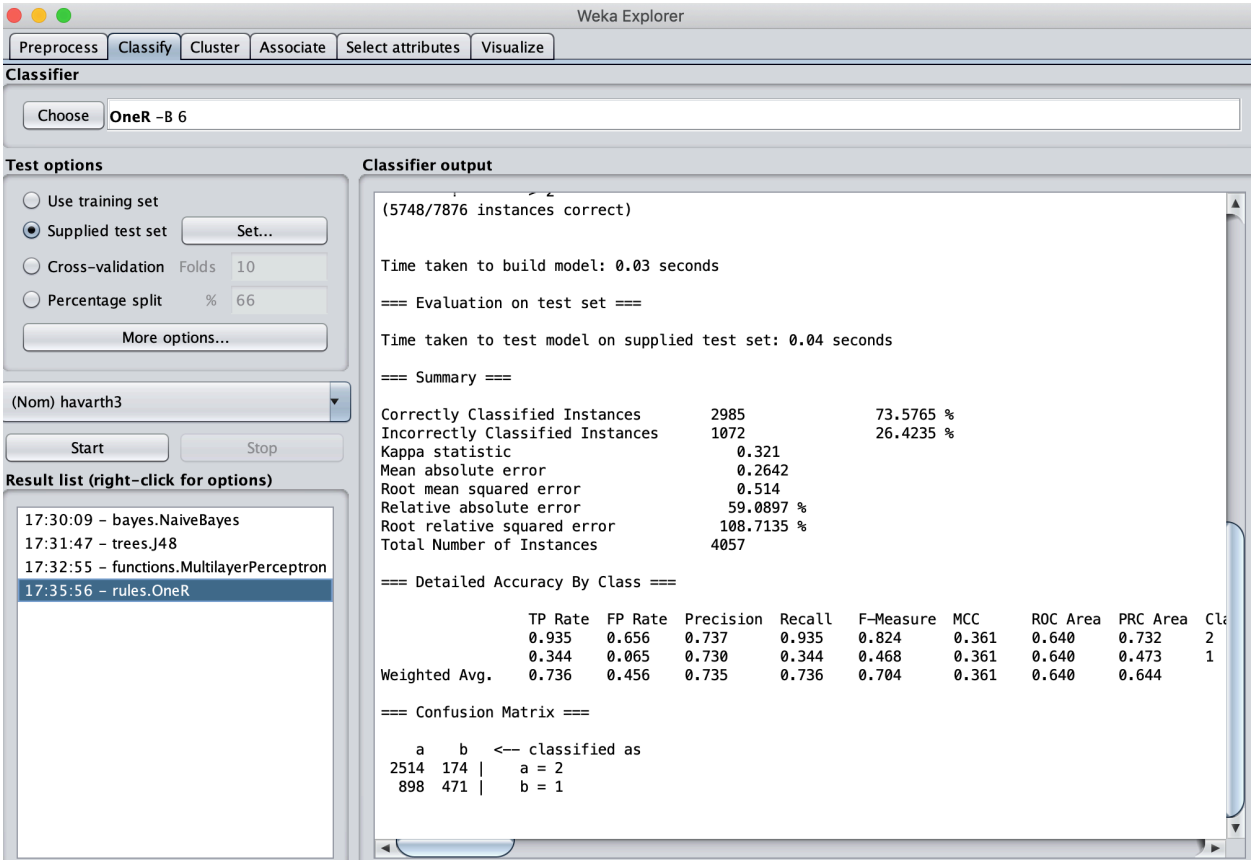
2. J48 (Decision Tree)



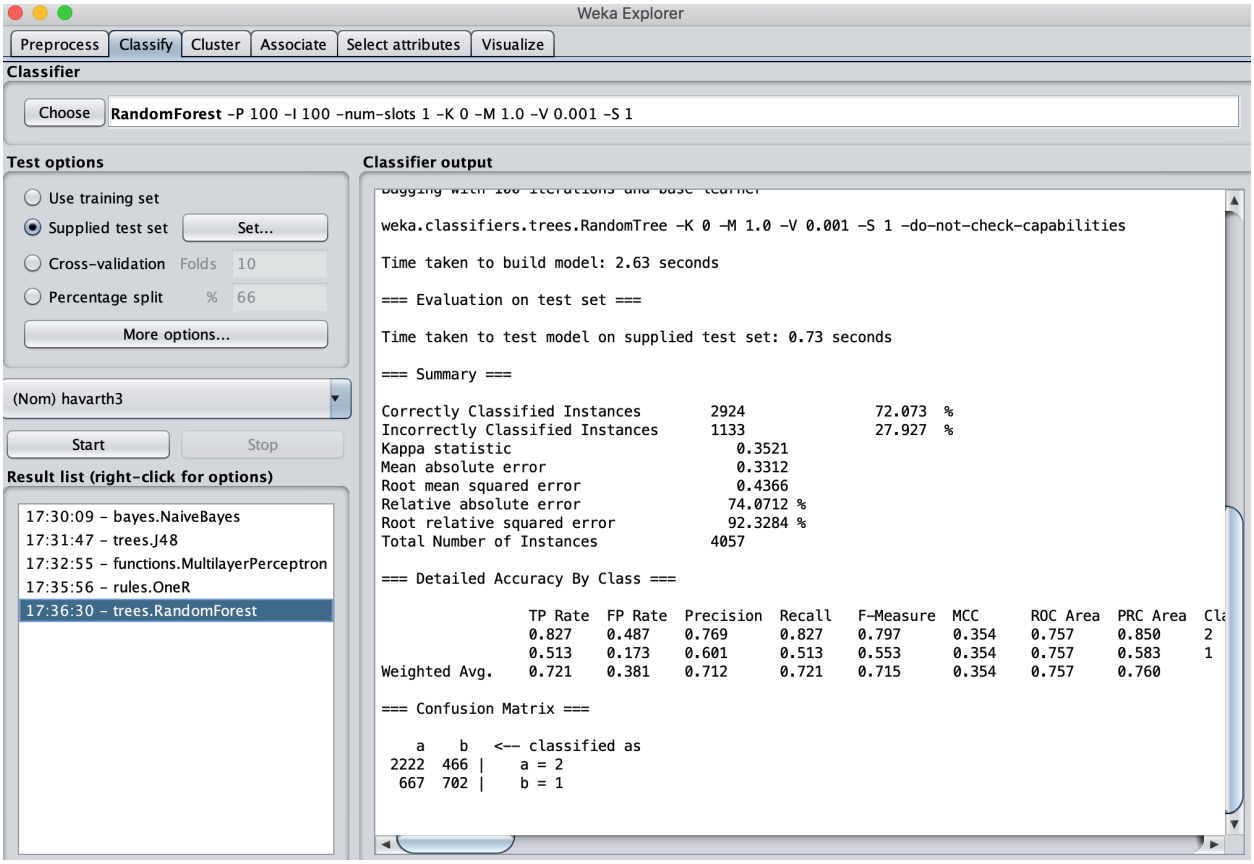
3. Neural Net (Multilayer Perceptron)



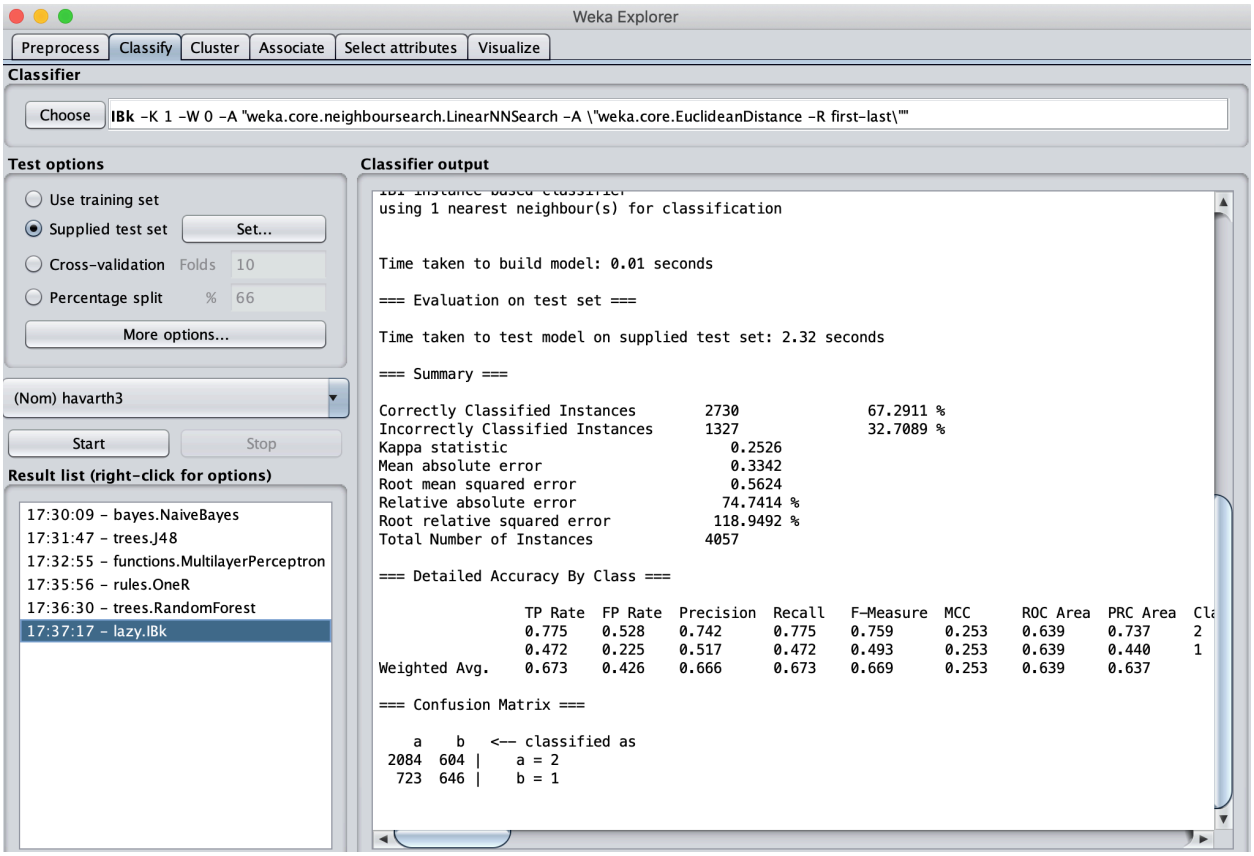
4. One R


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
OneR -B 6
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) havarth3
Start
Stop
Result list (right-click for options)
17:30:09 - bayes.NaiveBayes
17:31:47 - trees.J48
17:32:55 - functions.MultilayerPerceptron
17:35:56 - rules.OneR
Classifier output
(5748/7876 instances correct)
Time taken to build model: 0.03 seconds
=== Evaluation on test set ===
Time taken to test model on supplied test set: 0.04 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
0.935 0.656 0.737 0.935 0.824 0.361 0.640 0.732
0.344 0.065 0.730 0.344 0.468 0.361 0.640 0.473
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

5. Random Forest



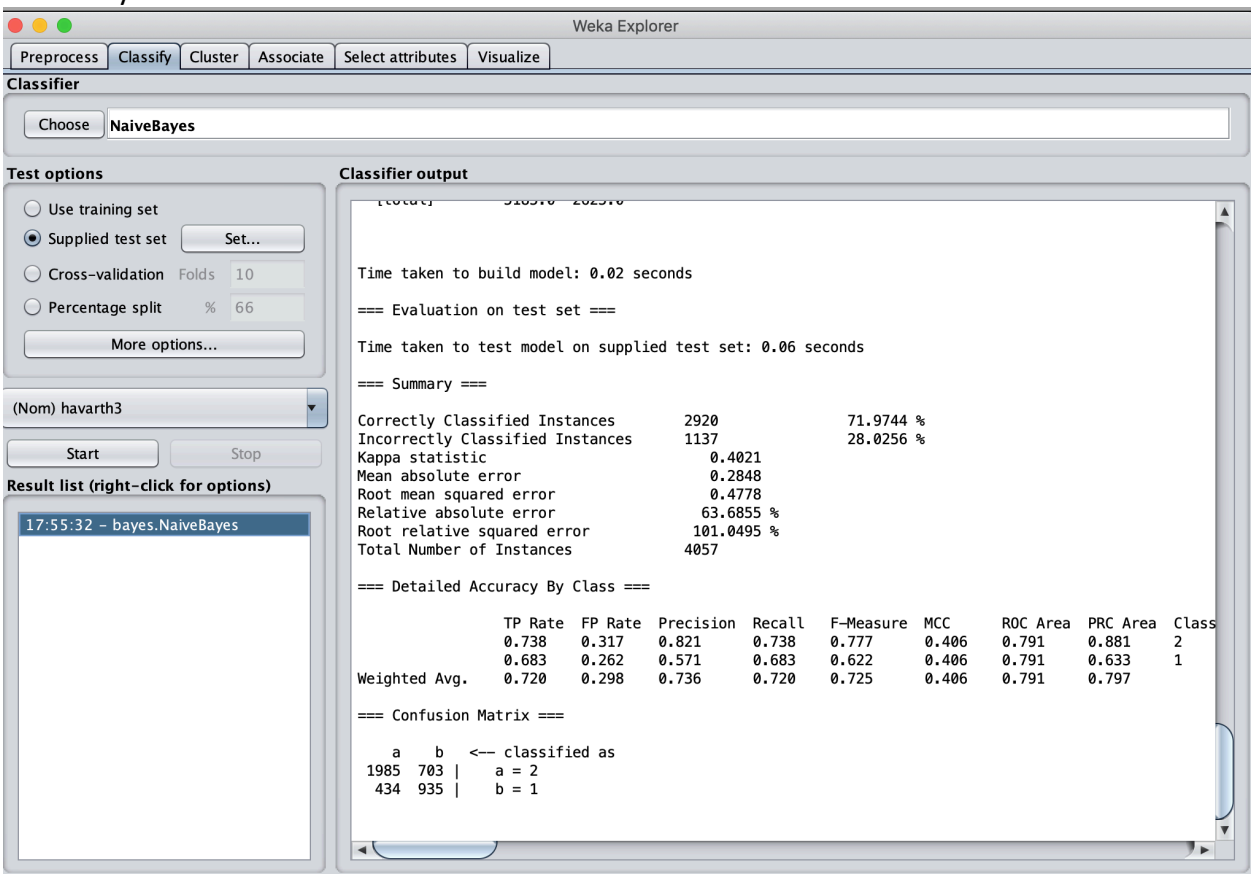
6. KNN


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) havarth3
Start
Stop
Result list (right-click for options)
17:30:09 - bayes.NaiveBayes
17:31:47 - trees.J48
17:32:55 - functions.MultilayerPerceptron
17:35:56 - rules.OneR
17:36:30 - trees.RandomForest
17:37:17 - lazy.IBk
Classifier output
using 1 nearest neighbour(s) for classification
Time taken to build model: 0.01 seconds
=== Evaluation on test set ===
Time taken to test model on supplied test set: 2.32 seconds
=== Summary ===
Correctly Classified Instances 2730 67.2911 %
Incorrectly Classified Instances 1327 32.7089 %
Kappa statistic 0.2526
Mean absolute error 0.3342
Root mean squared error 0.5624
Relative absolute error 74.7414 %
Root relative squared error 118.9492 %
Total Number of Instances 4057
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area
0.775 0.528 0.742 0.775 0.759 0.253 0.639 0.737
0.472 0.225 0.517 0.472 0.493 0.253 0.639 0.440
Weighted Avg. 0.673 0.426 0.666 0.673 0.669 0.253 0.639 0.637
=== Confusion Matrix ===
a b <-- classified as
2084 604 | a = 2
723 646 | b = 1

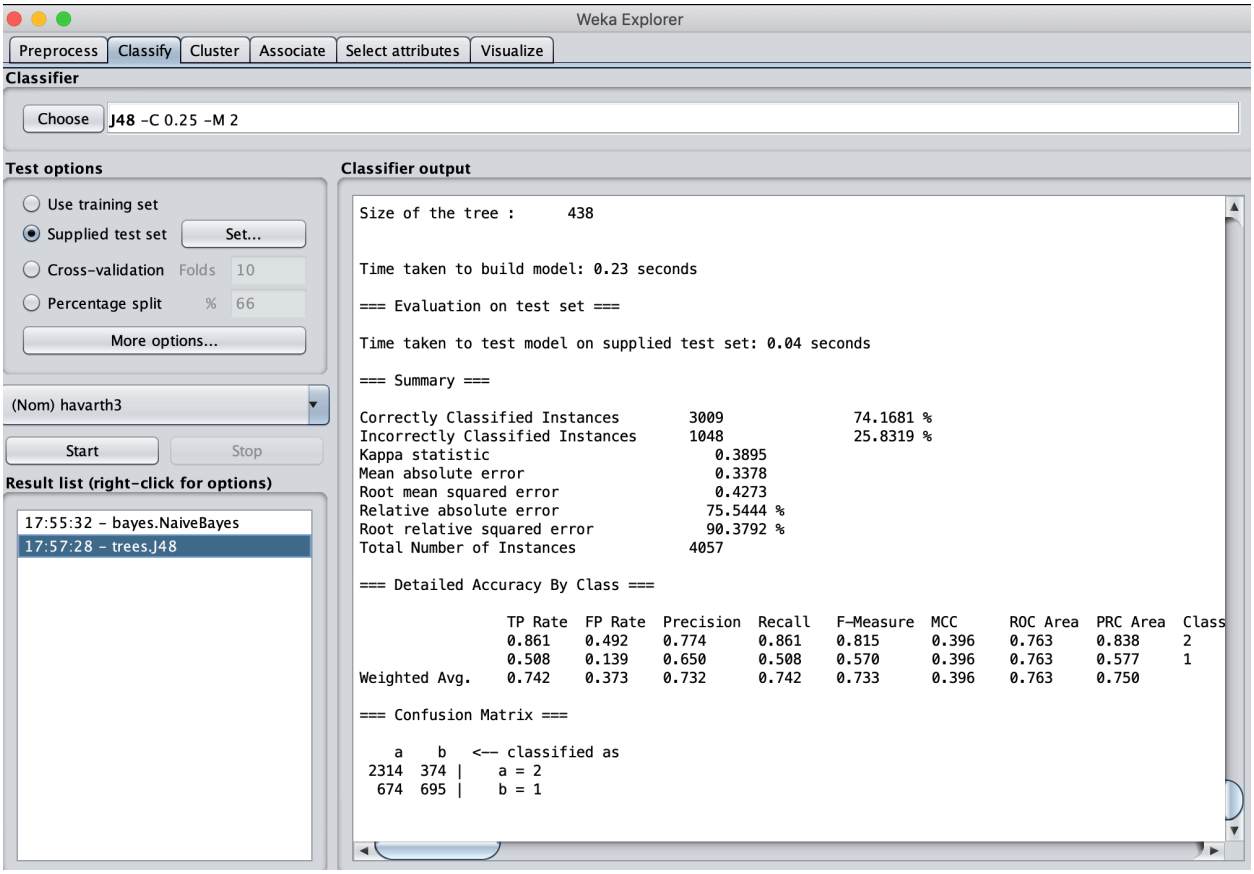
Attribute Selection Method – One R Classifier Attribute Evaluation
Attributes selected –

Index	Attribute Name
22	diffwalk
2	employ1
31	physhlth
95	x.phys14d
87	x.rfhlth
34	genhlth
46	chccopd1
66	x.ageg5yr
25	diffalon
67	x.age65yr
27	rmvteth4
24	diffdres
62	x.age.g
64	x.age80
53	cvdcrhd4
13	deaf
104	x.michd
11	marital
36	chckdny1
52	cvdinfr4

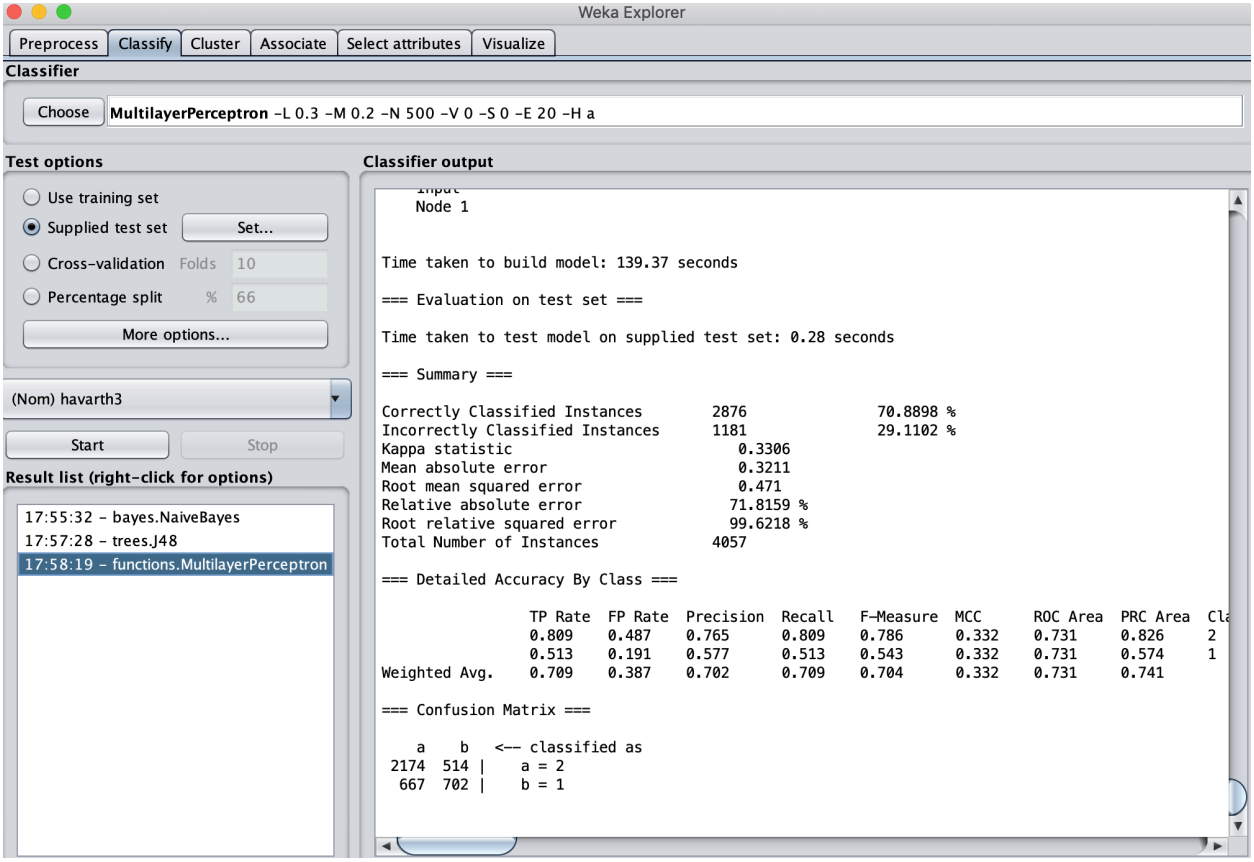
1. Naïve Bayes



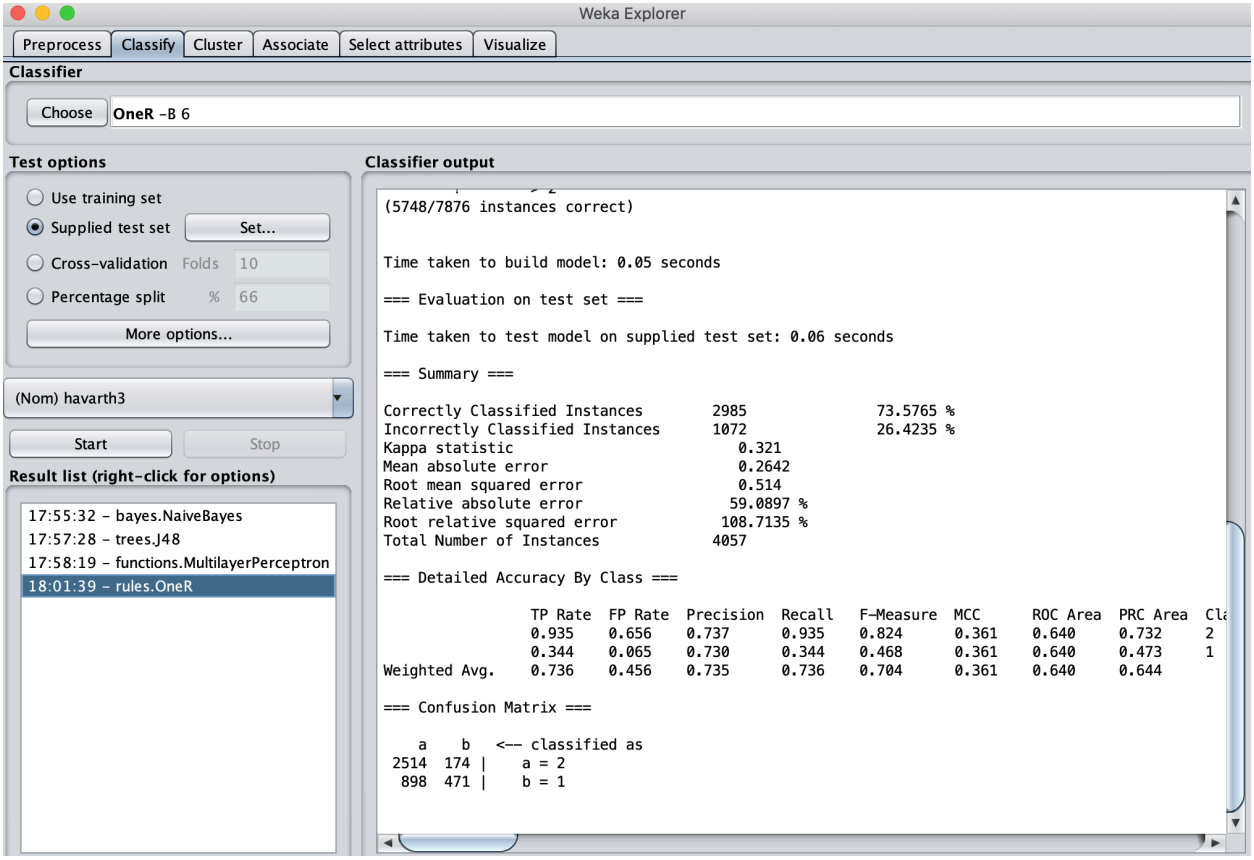
2. J48


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) havarth3
Start
Stop
Result list (right-click for options)
17:55:32 - bayes.NaiveBayes
17:57:28 - trees.J48
Classifier output
Size of the tree : 438
Time taken to build model: 0.23 seconds
=== Evaluation on test set ===
Time taken to test model on supplied test set: 0.04 seconds
=== Summary ===
Correctly Classified Instances 3009 74.1681 %
Incorrectly Classified Instances 1048 25.8319 %
Kappa statistic 0.3895
Mean absolute error 0.3378
Root mean squared error 0.4273
Relative absolute error 75.5444 %
Root relative squared error 90.3792 %
Total Number of Instances 4057
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.861 0.492 0.774 0.861 0.815 0.396 0.763 0.838 2
0.508 0.139 0.650 0.508 0.570 0.396 0.763 0.577 1
Weighted Avg. 0.742 0.373 0.732 0.742 0.733 0.396 0.763 0.750
=== Confusion Matrix ===
a b <-- classified as
2314 374 | a = 2
674 695 | b = 1

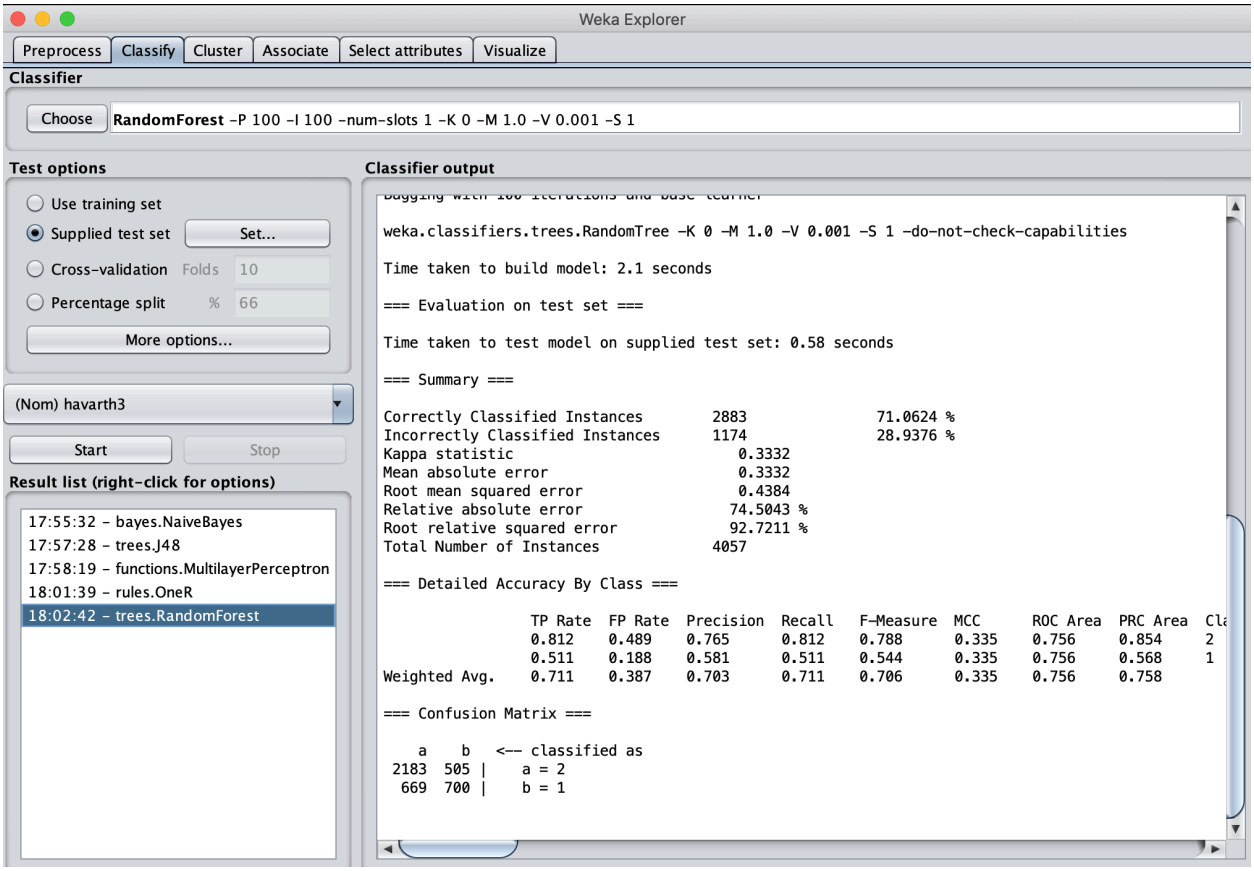
3. Neural Network



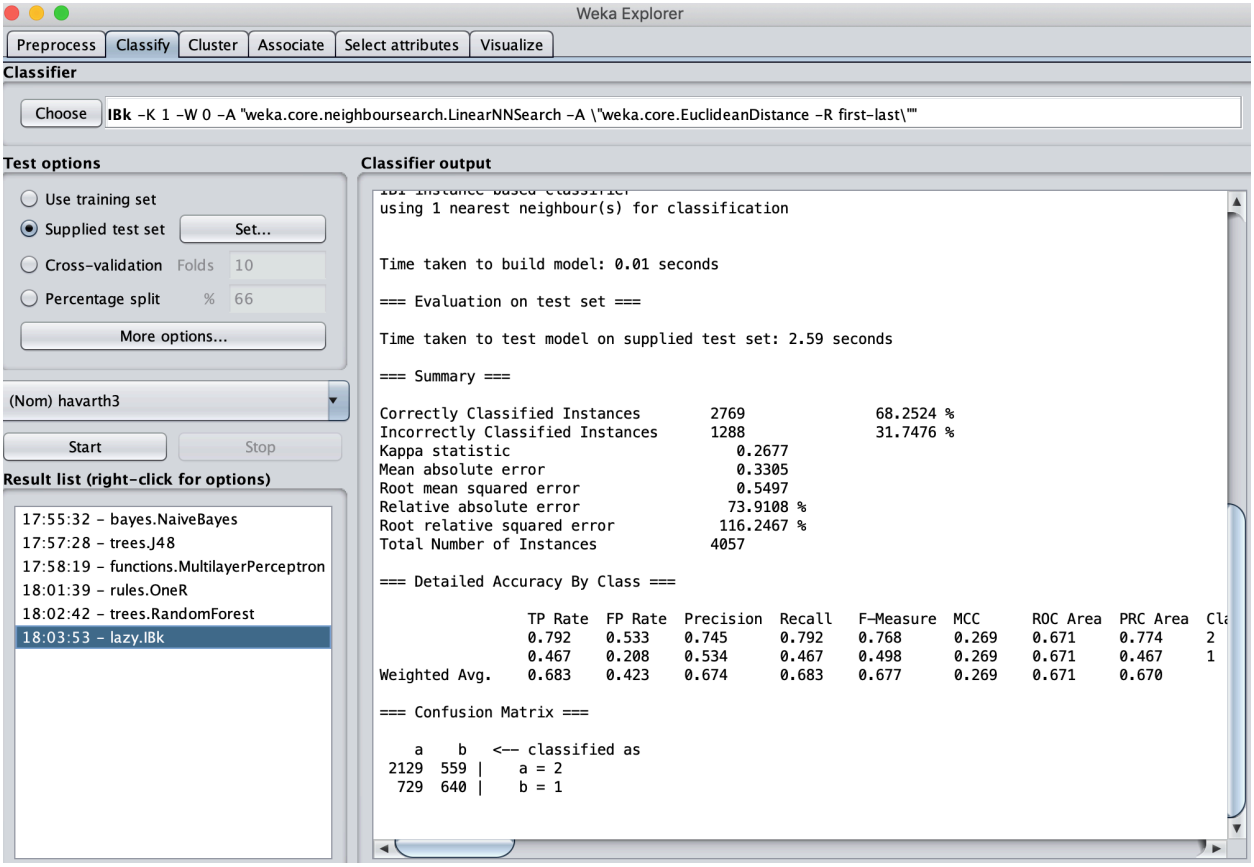
4. One R


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
OneR -B 6
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) havarth3
Start
Stop
Result list (right-click for options)
17:55:32 - bayes.NaiveBayes
17:57:28 - trees.J48
17:58:19 - functions.MultilayerPerceptron
18:01:39 - rules.OneR
Classifier output
(5748/7876 instances correct)
Time taken to build model: 0.05 seconds
=== Evaluation on test set ===
Time taken to test model on supplied test set: 0.06 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
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

5. Random Forest



6. KNN

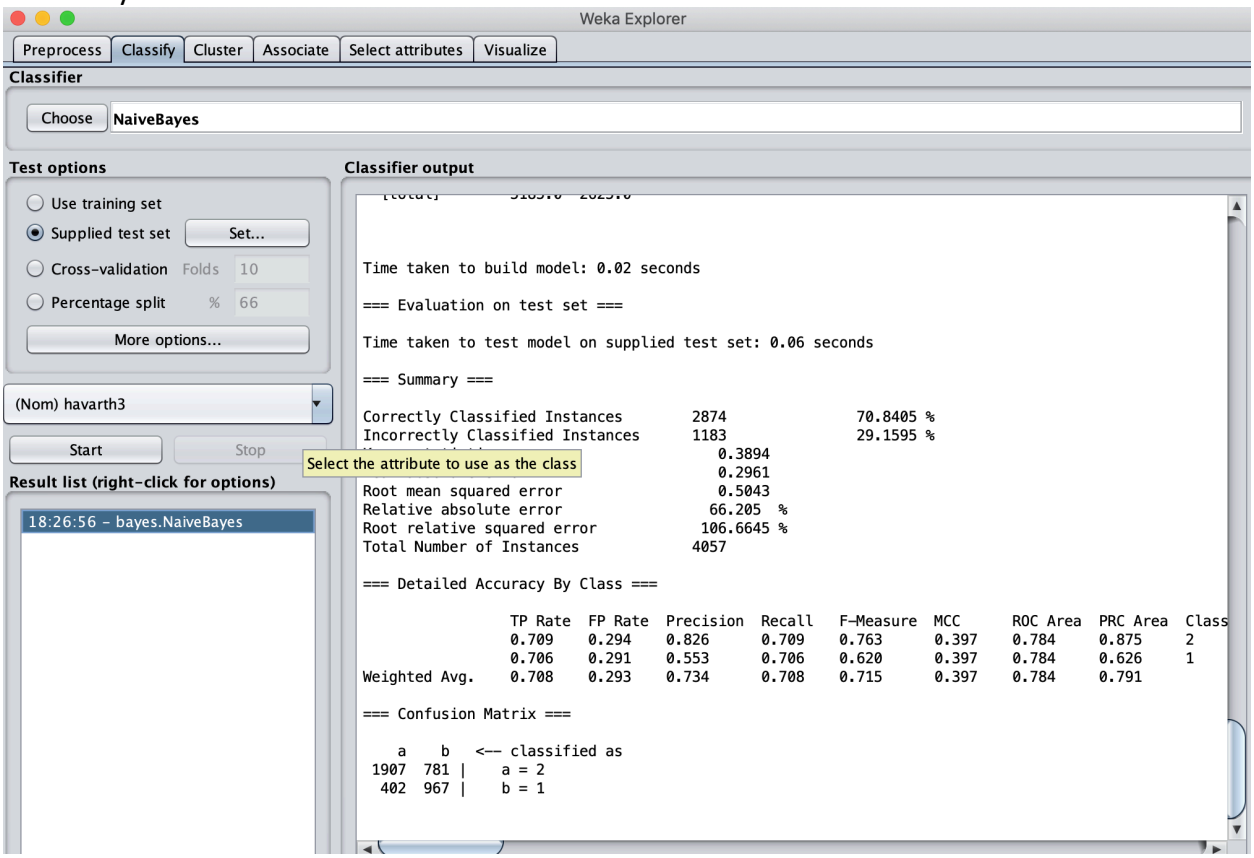

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) havarth3
Start
Stop
Result list (right-click for options)
17:55:32 - bayes.NaiveBayes
17:57:28 - trees.J48
17:58:19 - functions.MultilayerPerceptron
18:01:39 - rules.OneR
18:02:42 - trees.RandomForest
18:03:53 - lazy.IBk
Classifier output
using 1 nearest neighbour(s) for classification
Time taken to build model: 0.01 seconds
=== Evaluation on test set ===
Time taken to test model on supplied test set: 2.59 seconds
=== Summary ===
Correctly Classified Instances 2769 68.2524 %
Incorrectly Classified Instances 1288 31.7476 %
Kappa statistic 0.2677
Mean absolute error 0.3305
Root mean squared error 0.5497
Relative absolute error 73.9108 %
Root relative squared error 116.2467 %
Total Number of Instances 4057
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area
0.792 0.533 0.745 0.792 0.768 0.269 0.671 0.774
0.467 0.208 0.534 0.467 0.498 0.269 0.671 0.467
Weighted Avg. 0.683 0.423 0.674 0.683 0.677 0.269 0.671 0.670
=== Confusion Matrix ===
a b <-- classified as
2129 559 | a = 2
729 640 | b = 1

Attribute Selection Method – Correlation based on Pearson coefficient
Attributes selected –

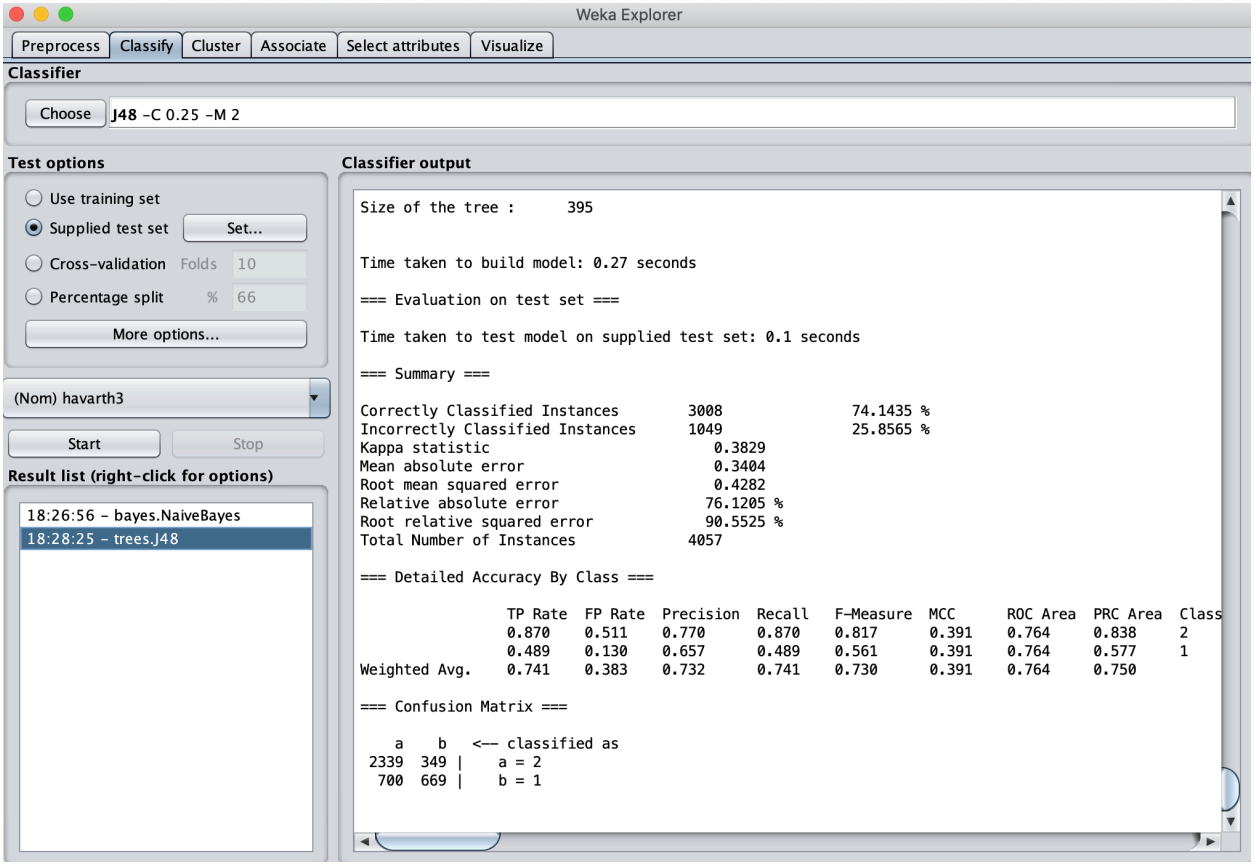
Index	Attribute Name
64	x.age80
66	x.ageg5yr
22	diffwalk
2	employ1
67	x.age65yr
87	x.rfhlth
97	x.hcvu651
102	x.exteth3
20	pneuvac4
46	chccopd1
95	x.phys14d
31	physhlth
6	children
25	diffalon
62	x.age.g
13	deaf
29	diabete3
104	x.michd
24	diffdres
69	x.chldcnt

Classification

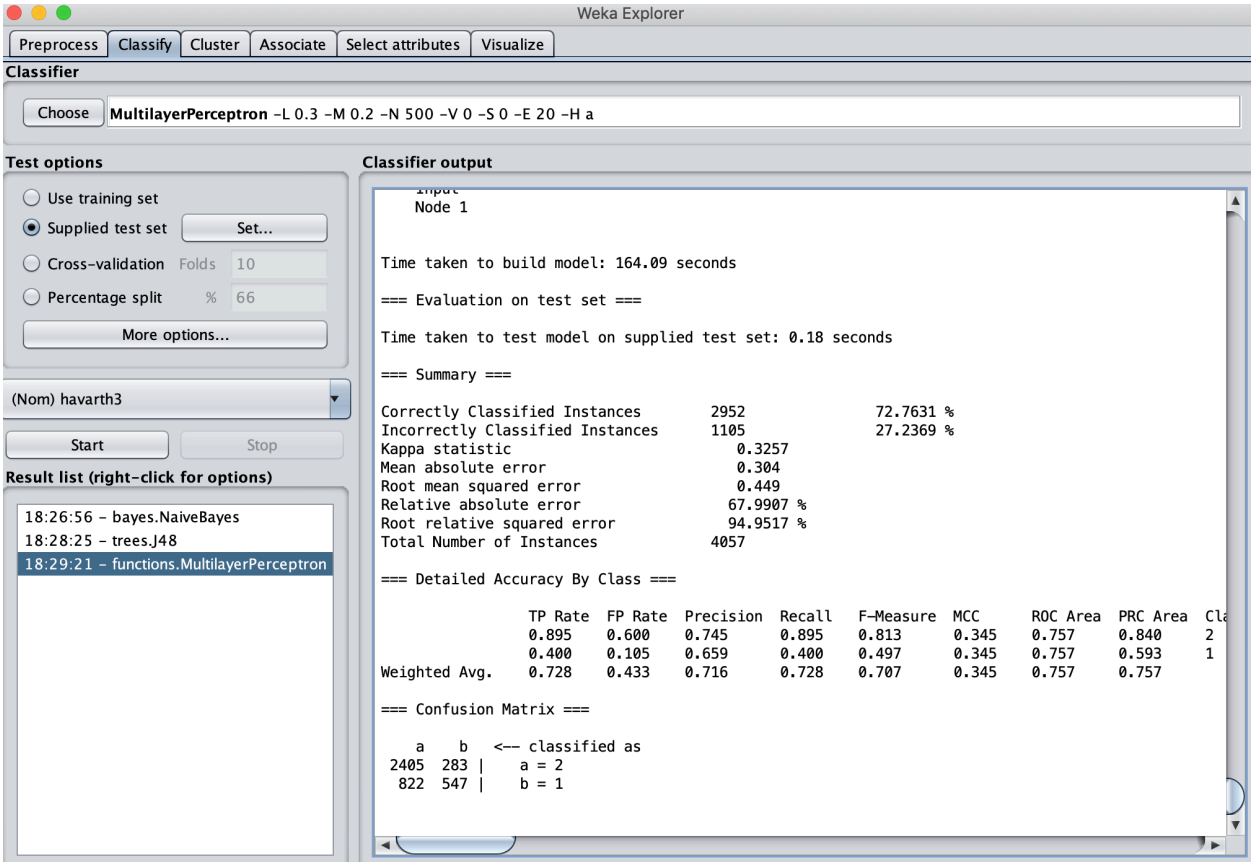
1. Naïve Bayes



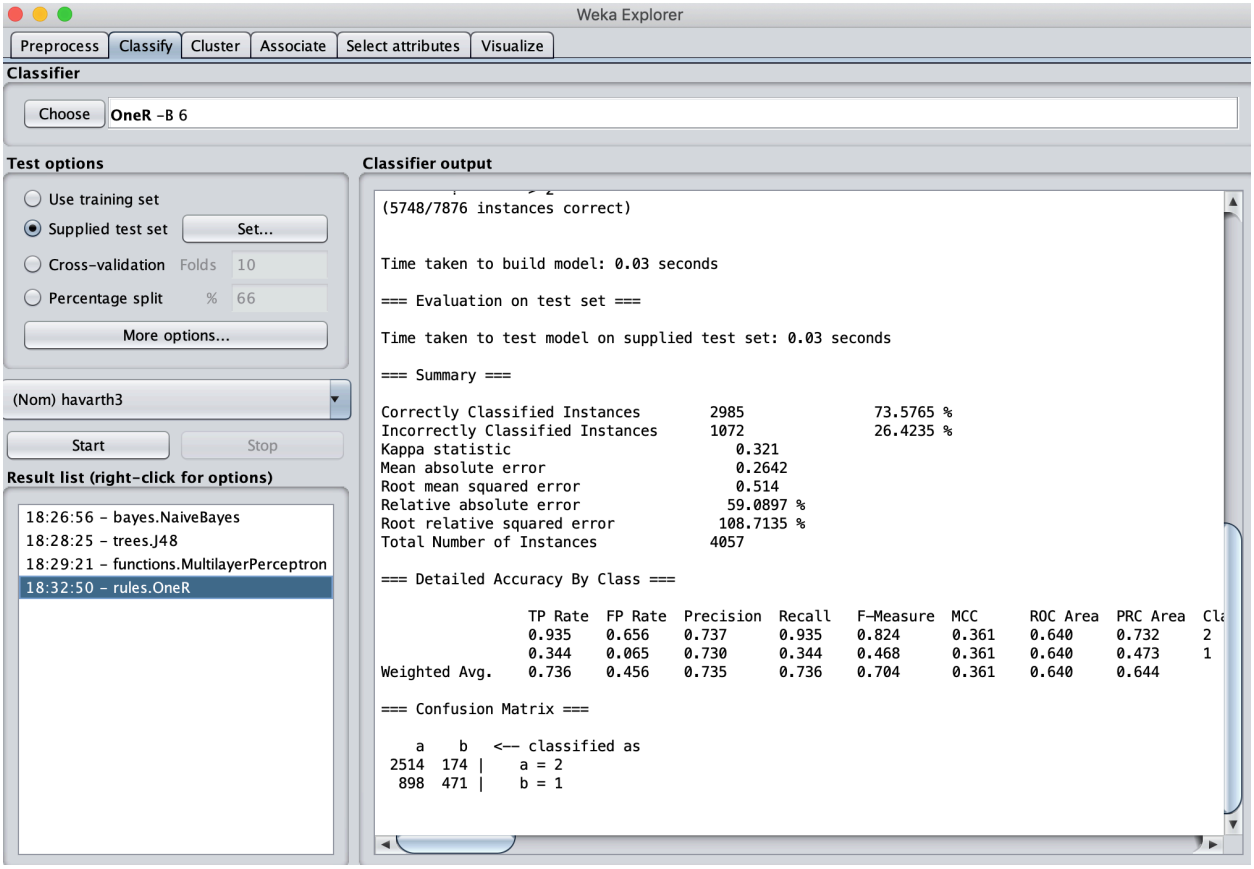
2. J48


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) havarth3
Start
Stop
Result list (right-click for options)
18:26:56 - bayes.NaiveBayes
18:28:25 - trees.J48
Classifier output
Size of the tree : 395
Time taken to build model: 0.27 seconds
=== Evaluation on test set ===
Time taken to test model on supplied test set: 0.1 seconds
=== Summary ===
Correctly Classified Instances 3008 74.1435 %
Incorrectly Classified Instances 1049 25.8565 %
Kappa statistic 0.3829
Mean absolute error 0.3404
Root mean squared error 0.4282
Relative absolute error 76.1205 %
Root relative squared error 90.5525 %
Total Number of Instances 4057
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.870 0.511 0.770 0.870 0.817 0.391 0.764 0.838 2
0.489 0.130 0.657 0.489 0.561 0.391 0.764 0.577 1
Weighted Avg. 0.741 0.383 0.732 0.741 0.730 0.391 0.764 0.750
=== Confusion Matrix ===
a b <-- classified as
2339 349 | a = 2
700 669 | b = 1

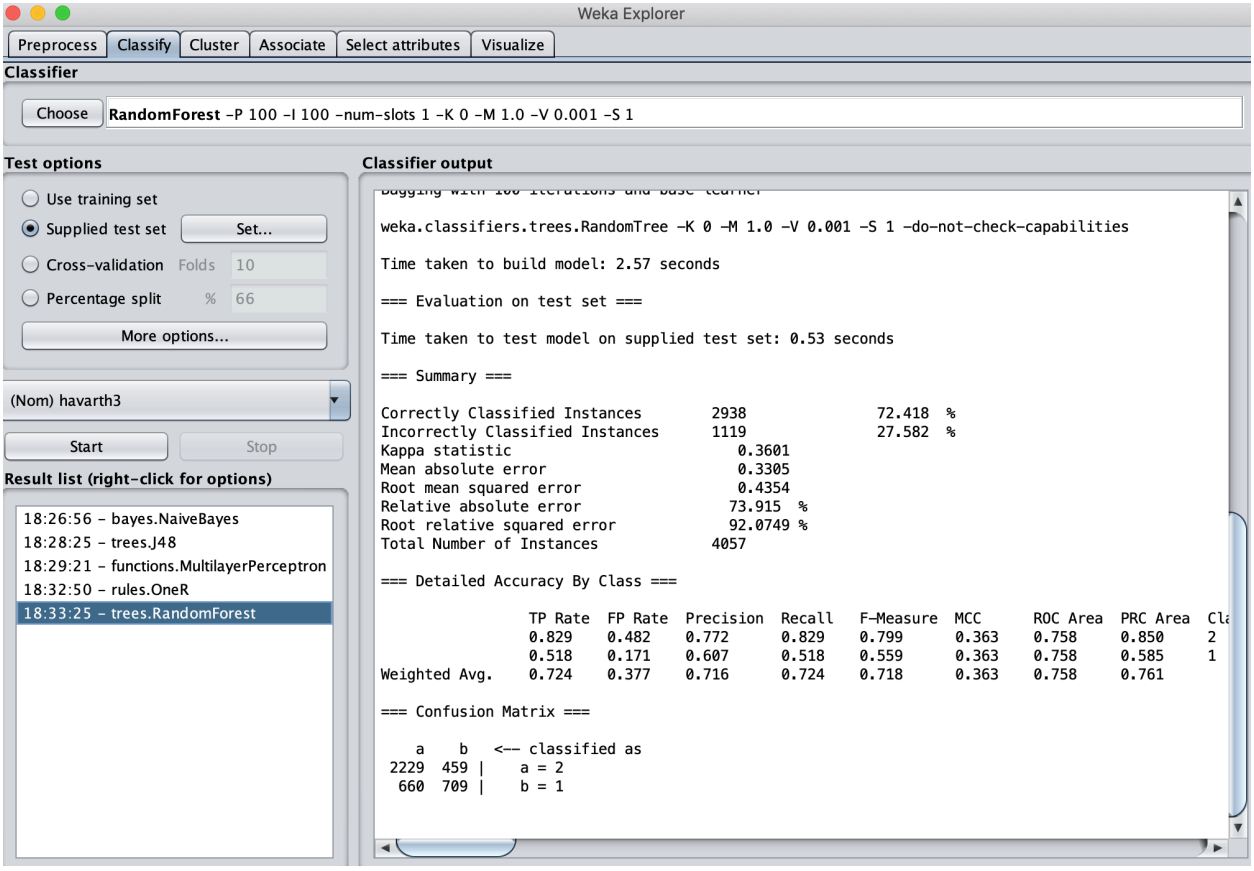
3. Neural Network



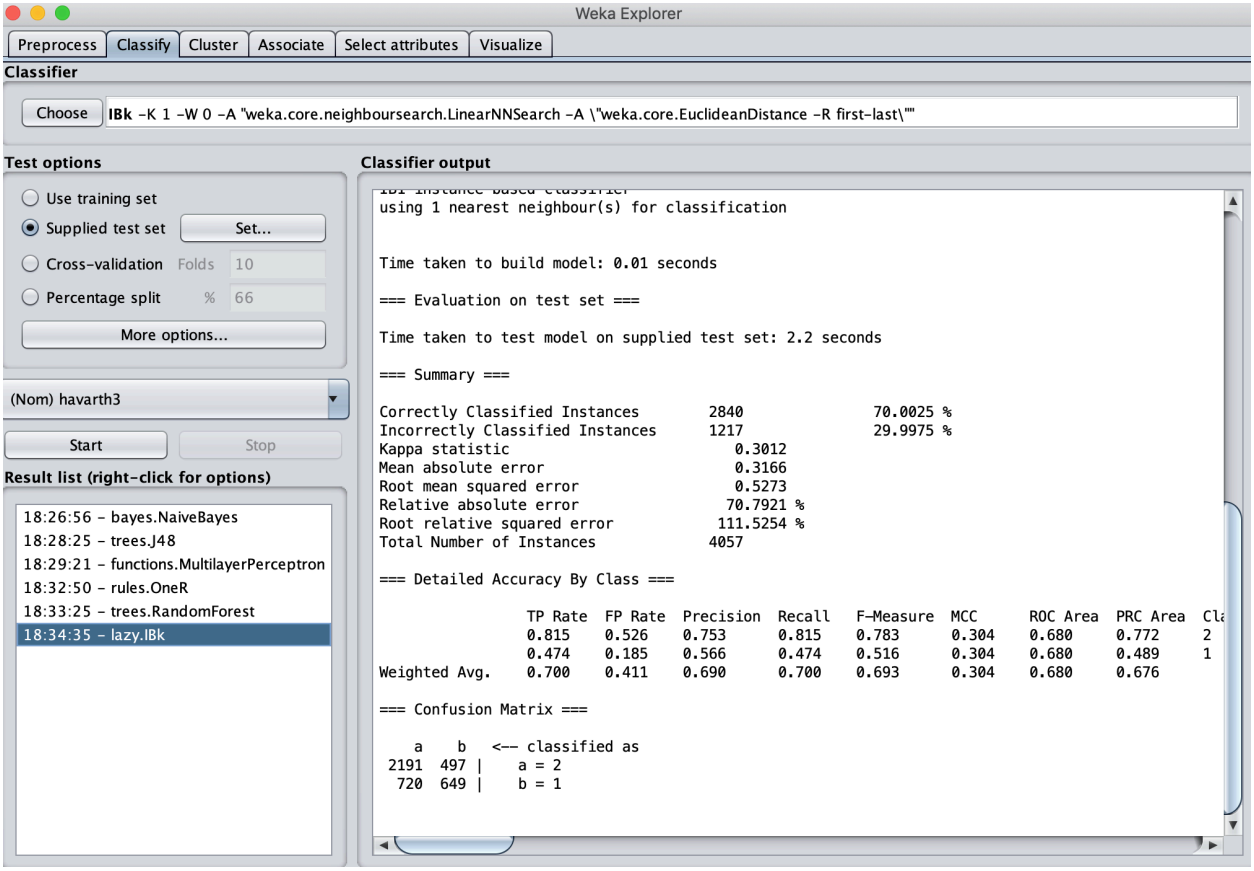
4. One R


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
OneR -B 6
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) havarth3
Start
Stop
Result list (right-click for options)
18:26:56 - bayes.NaiveBayes
18:28:25 - trees.J48
18:29:21 - functions.MultilayerPerceptron
18:32:50 - rules.OneR
Classifier output
(5748/7876 instances correct)
Time taken to build model: 0.03 seconds
=== Evaluation on test set ===
Time taken to test model on supplied test set: 0.03 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
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

5. Random Forest



6. KNN

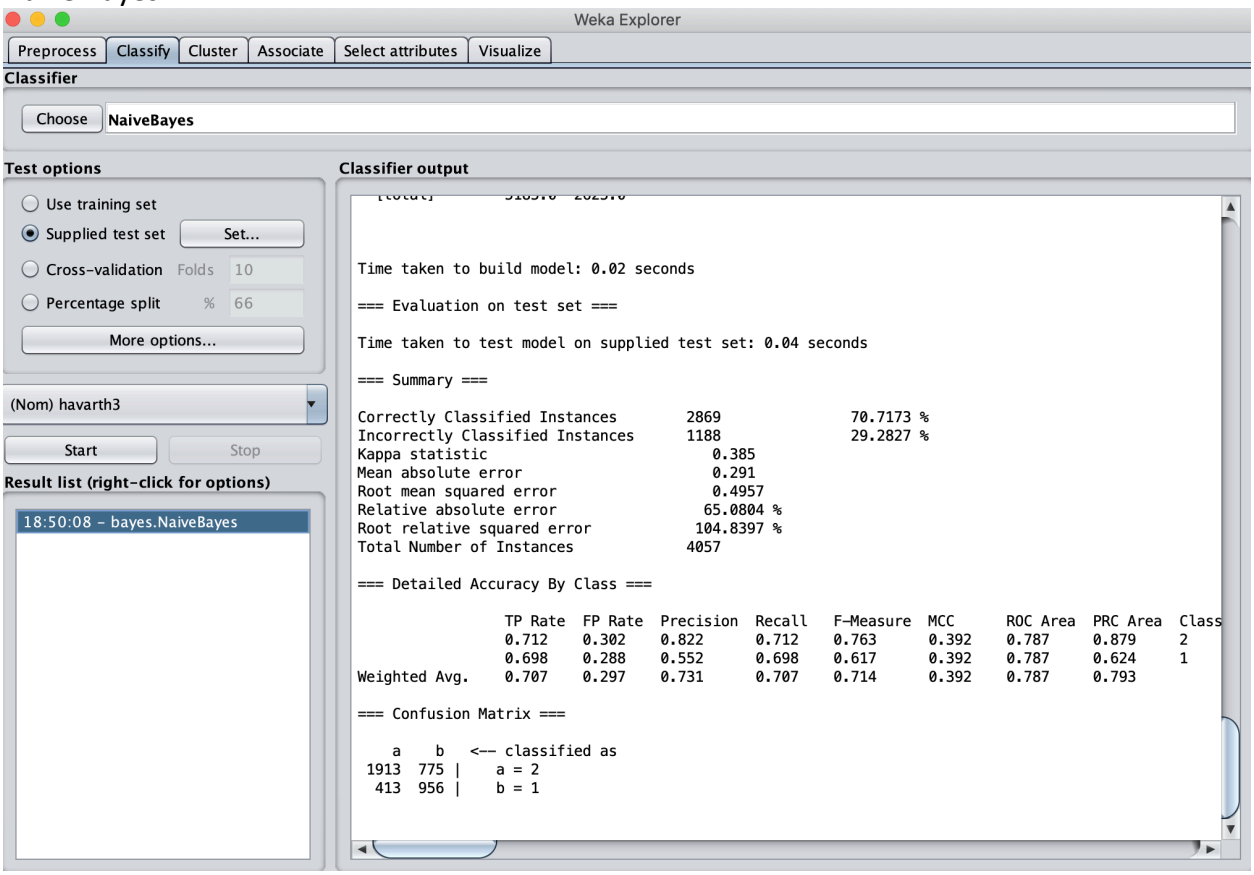

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) havarth3
Start
Stop
Result list (right-click for options)
18:26:56 - bayes.NaiveBayes
18:28:25 - trees.J48
18:29:21 - functions.MultilayerPerceptron
18:32:50 - rules.OneR
18:33:25 - trees.RandomForest
18:34:35 - lazy.IBk
Classifier output
using 1 nearest neighbour(s) for classification
Time taken to build model: 0.01 seconds
=== Evaluation on test set ===
Time taken to test model on supplied test set: 2.2 seconds
=== Summary ===
Correctly Classified Instances 2840 70.0025 %
Incorrectly Classified Instances 1217 29.9975 %
Kappa statistic 0.3012
Mean absolute error 0.3166
Root mean squared error 0.5273
Relative absolute error 70.7921 %
Root relative squared error 111.5254 %
Total Number of Instances 4057
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area
0.815 0.526 0.753 0.815 0.783 0.304 0.680 0.772
0.474 0.185 0.566 0.474 0.516 0.304 0.680 0.489
Weighted Avg. 0.700 0.411 0.690 0.700 0.693 0.304 0.680 0.676
=== Confusion Matrix ===
a b <-- classified as
2191 497 | a = 2
720 649 | b = 1

Attribute Selection Method – Gain Ratio Attribute Evaluation
Attributes selected –

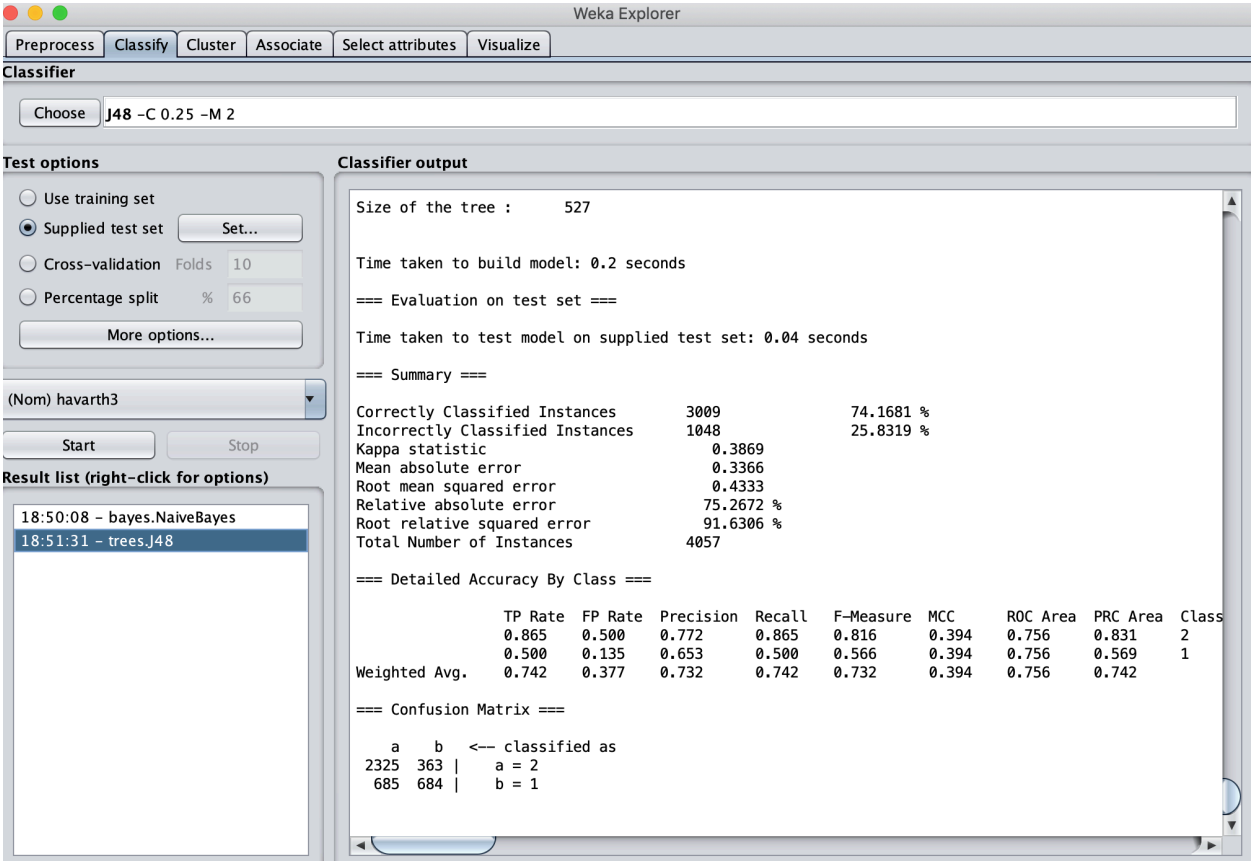
Index	Attribute Name
22	diffwalk
46	chccopd1
24	diffdres
87	x.rfhlth
67	x.age65yr
2	employ1
25	diffalon
62	x.age.g
64	x.age80
97	x.hcvu651
53	cvdcrhd4
36	chckdny1
104	x.michd
66	x.ageg5yr
13	deaf
31	physhlth
95	x.phys14d
20	pneuvac4
102	x.exteth3
50	cvdstrk3

Classification

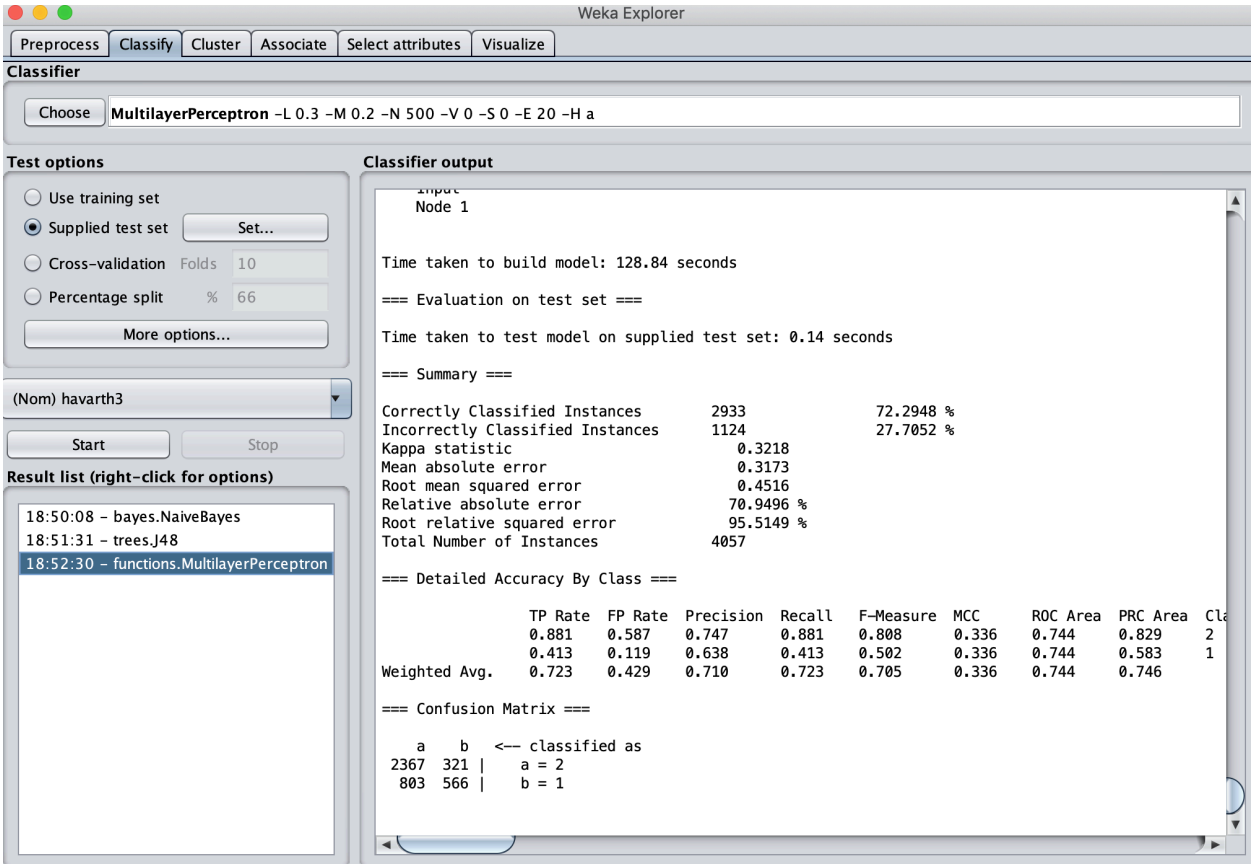
1. Naïve Bayes



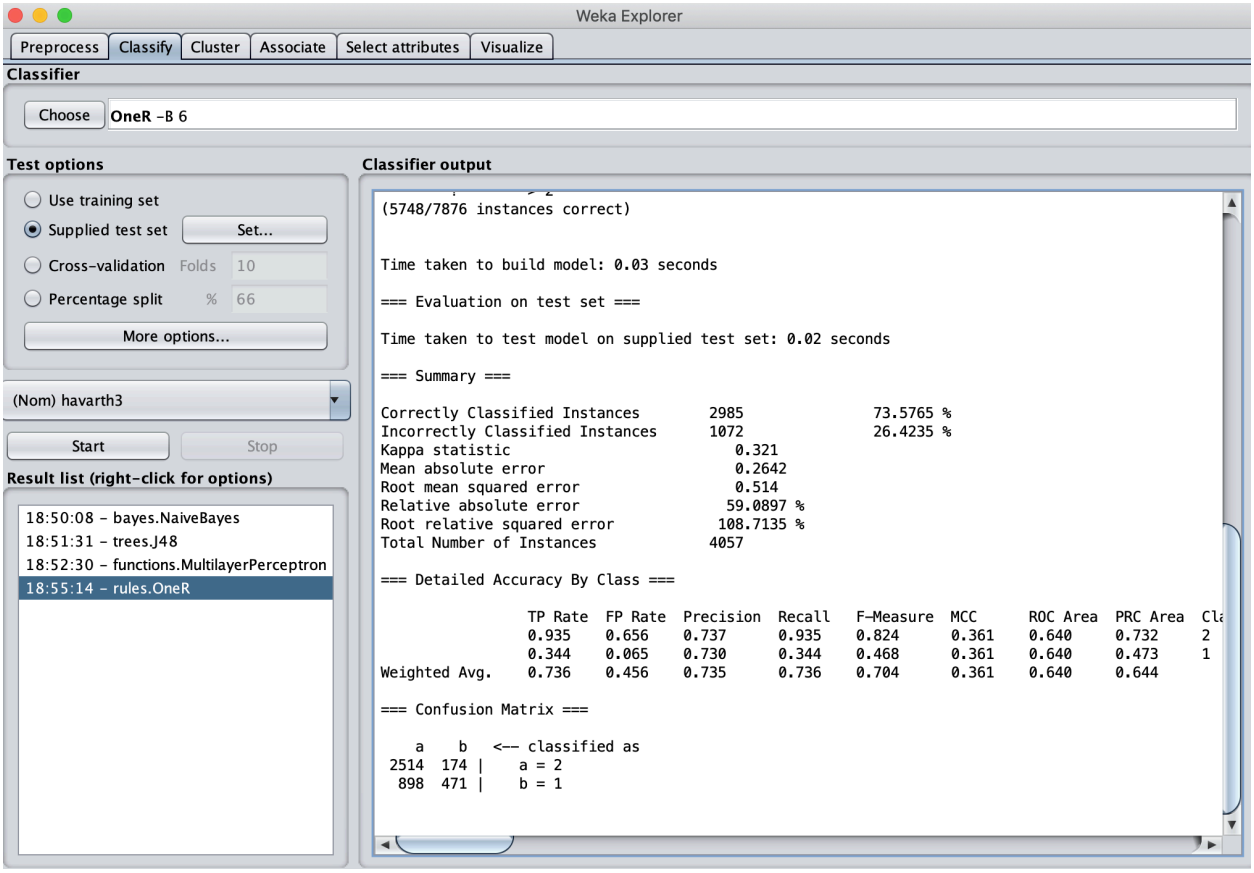
2. J48


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) havarth3
Start
Stop
Result list (right-click for options)
18:50:08 - bayes.NaiveBayes
18:51:31 - trees.J48
Classifier output
Size of the tree : 527
Time taken to build model: 0.2 seconds
=== Evaluation on test set ===
Time taken to test model on supplied test set: 0.04 seconds
=== Summary ===
Correctly Classified Instances 3009 74.1681 %
Incorrectly Classified Instances 1048 25.8319 %
Kappa statistic 0.3869
Mean absolute error 0.3366
Root mean squared error 0.4333
Relative absolute error 75.2672 %
Root relative squared error 91.6306 %
Total Number of Instances 4057
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.865 0.500 0.772 0.865 0.816 0.394 0.756 0.831 2
0.500 0.135 0.653 0.500 0.566 0.394 0.756 0.569 1
Weighted Avg. 0.742 0.377 0.732 0.742 0.732 0.394 0.756 0.742
=== Confusion Matrix ===
a b <-- classified as
2325 363 | a = 2
685 684 | b = 1

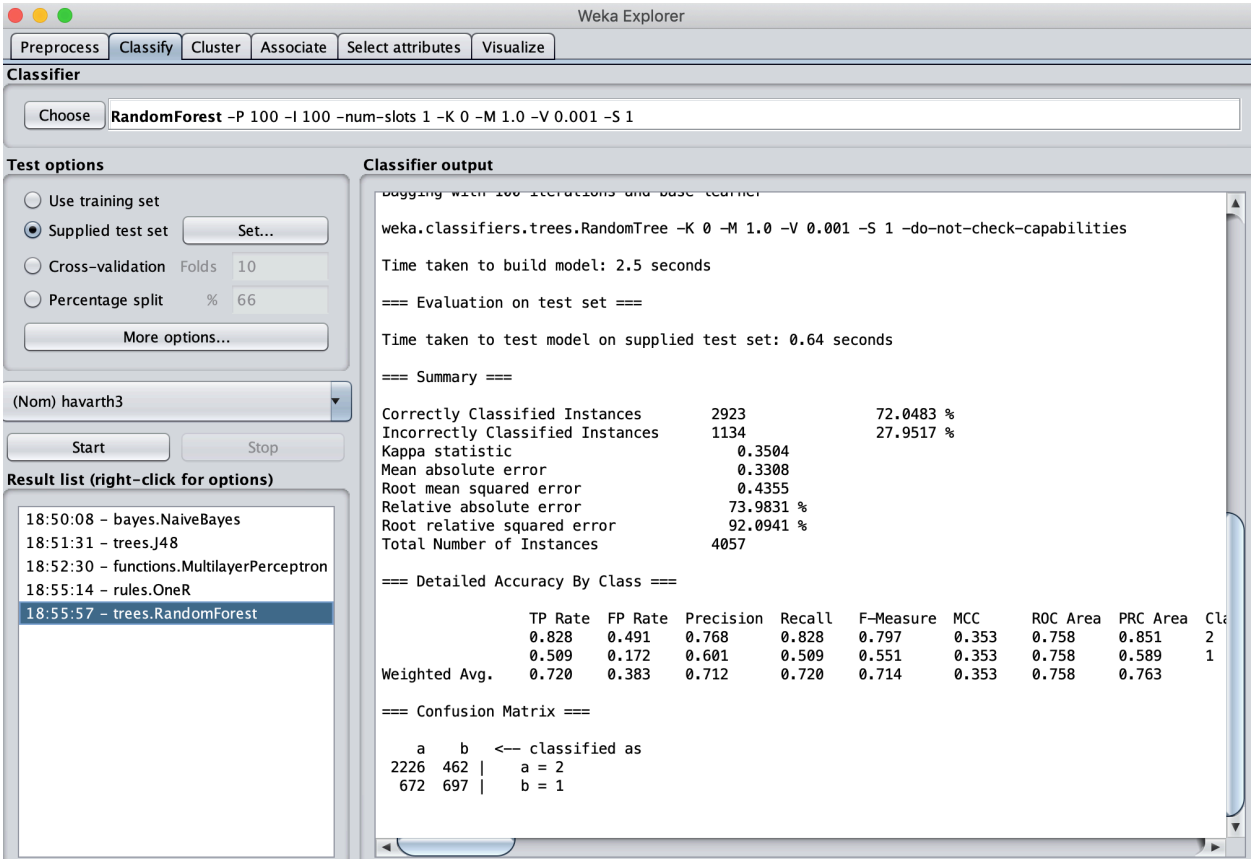
3. Neural Network



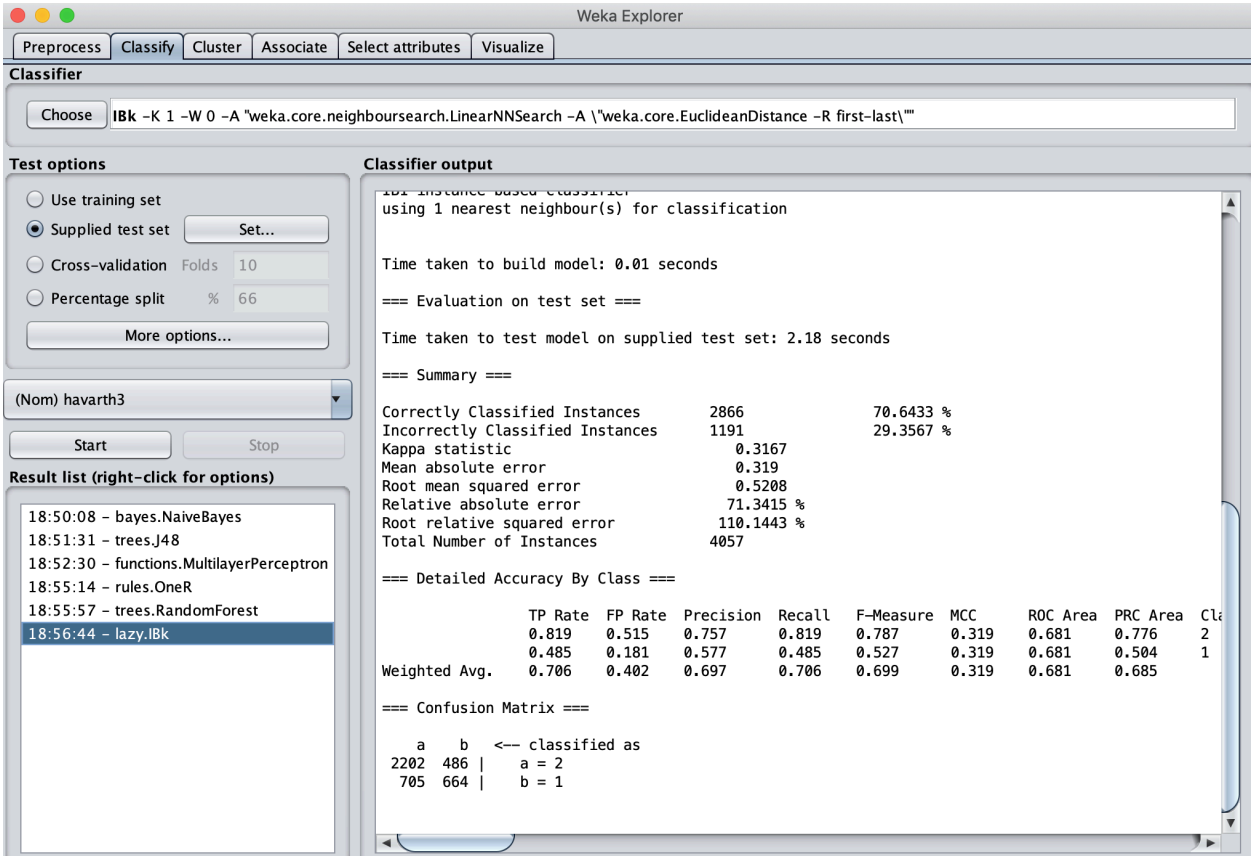
4. One R


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
OneR -B 6
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) havarth3
Start
Stop
Result list (right-click for options)
18:50:08 - bayes.NaiveBayes
18:51:31 - trees.J48
18:52:30 - functions.MultilayerPerceptron
18:55:14 - rules.OneR
Classifier output
(5748/7876 instances correct)
Time taken to build model: 0.03 seconds
=== Evaluation on test set ===
Time taken to test model on supplied test set: 0.02 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
0.935 0.656 0.737 0.935 0.824 0.361 0.640 0.732
0.344 0.065 0.730 0.344 0.468 0.361 0.640 0.473
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

5. Random Forest



6. KNN


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) havarth3
Start
Stop
Result list (right-click for options)
18:50:08 - bayes.NaiveBayes
18:51:31 - trees.J48
18:52:30 - functions.MultilayerPerceptron
18:55:14 - rules.OneR
18:55:57 - trees.RandomForest
18:56:44 - lazy.IBk
Classifier output
using 1 nearest neighbour(s) for classification
Time taken to build model: 0.01 seconds
=== Evaluation on test set ===
Time taken to test model on supplied test set: 2.18 seconds
=== Summary ===
Correctly Classified Instances 2866 70.6433 %
Incorrectly Classified Instances 1191 29.3567 %
Kappa statistic 0.3167
Mean absolute error 0.319
Root mean squared error 0.5208
Relative absolute error 71.3415 %
Root relative squared error 110.1443 %
Total Number of Instances 4057
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area
0.819 0.515 0.757 0.819 0.787 0.319 0.681 0.776
0.485 0.181 0.577 0.485 0.527 0.319 0.681 0.504
Weighted Avg. 0.706 0.402 0.697 0.706 0.699 0.319 0.681 0.685
=== Confusion Matrix ===
a b <-- classified as
2202 486 | a = 2
705 664 | b = 1