

LIS590DT Assignment #5

Jialu Wang

1. Circle

Random Forest

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. The classifier chosen is 'LogitBoost -P 100 -L -1.7976931348623157E308 -H 1.0 -Z 3.0 -O 1 -E 1 -S 1 -I 10 -W weka.classifiers.trees.DecisionStump'. The 'Test options' section shows 'Cross-validation' with 'Folds' set to 10. The 'Classifier output' pane displays the following results:

```
weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities
Time taken to build model: 0.44 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      1693           84.65 %
Incorrectly Classified Instances    307           15.35 %
Kappa statistic                    0.693
Mean absolute error                 0.1883
Root mean squared error             0.3349
Relative absolute error             37.667 %
Root relative squared error         66.9844 %
Total Number of Instances          2000

=== Detailed Accuracy By Class ===

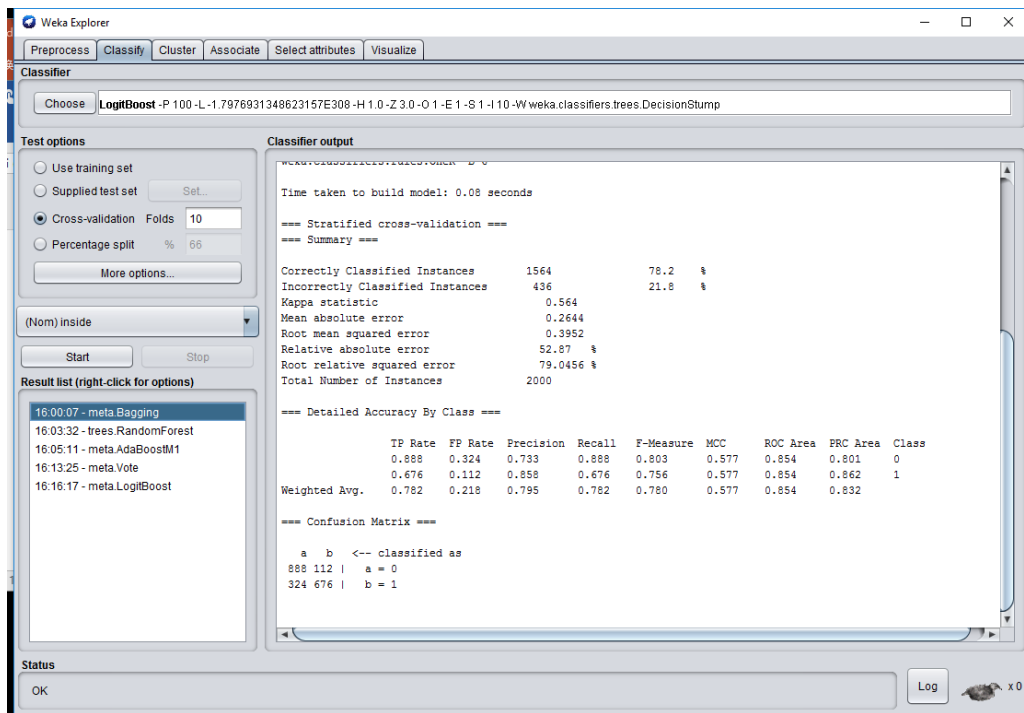
              TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
              0.892    0.199    0.818     0.892    0.853     0.696    0.907    0.866     0
              0.801    0.108    0.881     0.801    0.839     0.696    0.907    0.932     1
Weighted Avg.   0.847    0.154    0.849     0.847    0.846     0.696    0.907    0.899

=== Confusion Matrix ===

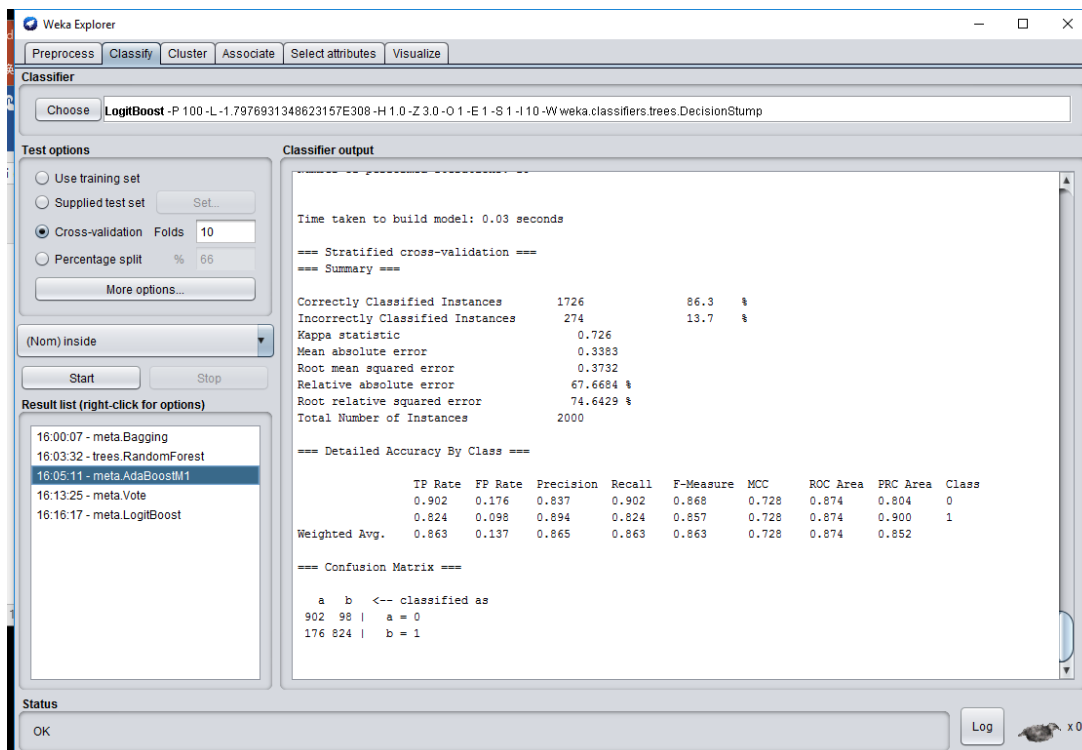
  a  b  <-- classified as
892 108 | a = 0
199 801 | b = 1
```

The 'Result list' on the left shows several entries, with '16:03:32 - trees.RandomForest' selected. The 'Status' bar at the bottom indicates 'OK' and 'Log'.

Bagging with oneR



AdaBoostM1



LogitBoost

Weka Explorer

Preprocess **Classify** Cluster Associate Select attributes Visualize

Classifier

Choose **LogitBoost** -P 100 -L -1.7976931348623157E308 -H 1.0 -Z 3.0 -O 1 -E 1 -S 1 -I 10 -W weka.classifiers.trees.DecisionStump

Test options

☐ Use training set
☐ Supplied test set Set...
☒ Cross-validation Folds **10**
☐ Percentage split % 66

More options...

(Nom) inside

Start Stop

Result list (right-click for options)

- 16:00:07 - meta.Bagging
- 16:03:32 - trees.RandomForest
- 16:05:11 - meta.AdaBoostM1
- 16:13:25 - meta.Vote
- 16:16:17 - meta.LogitBoost**

Classifier output

Time taken to build model: 0.13 seconds

=== Stratified cross-validation ===
 === Summary ===

Correctly Classified Instances	1730	86.5 %
Incorrectly Classified Instances	270	13.5 %
Kappa statistic	0.73	
Mean absolute error	0.2094	
Root mean squared error	0.3195	
Relative absolute error	41.8721 %	
Root relative squared error	63.9077 %	
Total Number of Instances	2000	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
Weighted Avg.	0.865	0.135	0.867	0.865	0.865	0.732	0.912	0.897	1

=== Confusion Matrix ===

	a	b	<-- classified as
898 102	a = 0		
168 832	b = 1		

Status

OK Log x0

Vote

Weka Explorer

Preprocess **Classify** Cluster Associate Select attributes Visualize

Classifier

Choose **Vote** -S 1 -B "weka.classifiers.bayes.BayesNet -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5" -B "weka.classifiers.functions.Logistic -R 1.0E-8 -M

Test options

☐ Use training set
☐ Supplied test set Set...
☒ Cross-validation Folds **10**
☐ Percentage split % 66

More options...

(Nom) inside

Start Stop

Result list (right-click for options)

- 22:52:22 - meta.Vote**

Classifier output

Time taken to build model: 0.17 seconds

=== Stratified cross-validation ===
 === Summary ===

Correctly Classified Instances	1729	86.45 %
Incorrectly Classified Instances	271	13.55 %
Kappa statistic	0.729	
Mean absolute error	0.4021	
Root mean squared error	0.4126	
Relative absolute error	80.43 %	
Root relative squared error	82.521 %	
Total Number of Instances	2000	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
Weighted Avg.	0.865	0.136	0.872	0.865	0.864	0.737	0.911	0.895	1

=== Confusion Matrix ===

	a	b	<-- classified as
937 63	a = 0		
208 792	b = 1		

Status

OK Log x0

Vote combines the probability distributions of these base learners:

```
weka.classifiers.bayes.BayesNet -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5 -B "weka.classifiers.functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4
weka.classifiers.trees.DecisionStump
weka.classifiers.rules.ZeroR
```

using the 'Average' combination rule

With the dataset circle, LogitBoost provide the highest accuracy rate.

2. OCR

Random Forest

Weka Explorer

Preprocess | **Classify** | Cluster | Associate | Select attributes | Visualize

Classifier

Choose **RandomForest** -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

Test options

☐ Use training set
☐ Supplied test set
☒ Cross-validation Folds **10**
☐ Percentage split % 66

More options...

(Nom) lettr

Start Stop

Result list (right-click for options)

16:27:22 - trees.RandomForest

Classifier output

```
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances      19274           96.37 %
Incorrectly Classified Instances     726           3.63 %
Kappa statistic                    0.9622
Mean absolute error                 0.0131
Root mean squared error             0.0622
Relative absolute error             17.6826 %
Root relative squared error         32.3563 %
Total Number of Instances          20000

=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.986	0.001	0.975	0.986	0.981	0.980	1.000	0.995	I
	0.946	0.001	0.971	0.946	0.958	0.957	0.999	0.991	I
	0.973	0.003	0.929	0.973	0.950	0.948	0.999	0.988	D
	0.963	0.001	0.968	0.963	0.965	0.964	1.000	0.994	N
	0.964	0.002	0.950	0.964	0.957	0.955	0.999	0.987	G
	0.965	0.001	0.976	0.965	0.970	0.969	1.000	0.993	S
	0.956	0.004	0.907	0.956	0.931	0.928	0.999	0.983	B
	0.995	0.000	0.989	0.995	0.992	0.991	1.000	1.000	A
	0.949	0.002	0.961	0.949	0.955	0.953	1.000	0.991	J
	0.981	0.001	0.975	0.981	0.978	0.977	1.000	0.997	M
	0.975	0.001	0.971	0.975	0.973	0.972	1.000	0.996	X
	0.963	0.002	0.943	0.963	0.953	0.951	0.999	0.987	O
	0.945	0.002	0.943	0.945	0.944	0.942	0.999	0.984	R
	0.948	0.002	0.958	0.948	0.953	0.951	0.999	0.983	F
	0.951	0.000	0.992	0.951	0.971	0.970	1.000	0.995	C
	0.903	0.002	0.942	0.903	0.922	0.919	0.999	0.976	H

Status

OK Log x 0

Bagging+oneR

Weka Explorer

Preprocess | **Classify** | Cluster | Associate | Select attributes | Visualize

Classifier

Choose **Bagging** -P 100 -S 1 -num-slots 1 -I 10 -W weka.classifiers.rules.OneR -- -B 6

Test options

☐ Use training set
☐ Supplied test set
☒ Cross-validation Folds **10**
☐ Percentage split % 66

More options...

(Nom) lettr

Start Stop

Result list (right-click for options)

16:27:22 - trees.RandomForest
16:30:27 - meta.Bagging

Classifier output

```
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances      3579           17.895 %
Incorrectly Classified Instances   16421           82.105 %
Kappa statistic                    0.1456
Mean absolute error                 0.0638
Root mean squared error             0.2225
Relative absolute error             86.2954 %
Root relative squared error        115.6819 %
Total Number of Instances          20000

=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.000	0.000	0.000	0.000	0.000	0.000	0.578	0.069	T
	0.799	0.109	0.223	0.799	0.349	0.384	0.863	0.325	I
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.040	D
	0.766	0.105	0.228	0.766	0.352	0.379	0.830	0.182	N
	0.471	0.161	0.105	0.471	0.172	0.158	0.681	0.082	G
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.037	S
	0.000	0.000	0.000	0.000	0.000	0.000	0.621	0.060	B
	0.219	0.018	0.337	0.219	0.266	0.248	0.768	0.312	A
	0.000	0.000	0.000	0.000	0.000	0.000	0.562	0.053	J
	0.617	0.032	0.443	0.617	0.516	0.500	0.877	0.340	M
	0.000	0.000	0.000	0.000	0.000	0.000	0.520	0.042	X
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.038	O
	0.000	0.000	0.000	0.000	0.000	0.000	0.731	0.098	R
	0.000	0.000	0.000	0.000	0.000	0.000	0.537	0.049	F
	0.000	0.000	0.000	0.000	0.000	0.000	0.576	0.053	C
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.037	H

Status

OK Log x 0

AdaBoost (classifier: decision stump)

The Weka Explorer interface displays the results of an AdaBoost classifier using decision stumps. The classifier is configured with 100 iterations and a learning rate of 0.1. The test options are set to cross-validation with 10 folds. The classifier output shows a summary of performance metrics and a detailed accuracy by class table.

Classifier: AdaBoostM1 - P 100 - S 1 - I 10 - W weka.classifiers.rules.JRip -- F 3 - N 2.0 - O 2 - S 1

Test options:

- ☐ Use training set
- ☐ Supplied test set
- ☒ Cross-validation Folds: 10
- ☐ Percentage split % 66

Classifier output:

```
=== Summary ===
Correctly Classified Instances      1418           7.09 %
Incorrectly Classified Instances   18582          92.91 %
Kappa statistic                    0.0329
Mean absolute error                 0.0726
Root mean squared error             0.1905
Relative absolute error             98.1504 %
Root relative squared error         99.0725 %
Total Number of Instances         20000

=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.645	0.063	T
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.639	0.058	I
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.631	0.055	D
0.815	0.330	0.091	0.815	0.164	0.197	0.197	0.732	0.081	N
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.680	0.059	G
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.675	0.057	S
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.677	0.058	B
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.554	0.046	A
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.699	0.070	J
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.697	0.074	M
0.991	0.637	0.060	0.991	0.113	0.145	0.145	0.674	0.059	X
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.627	0.051	O
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.677	0.058	R
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.553	0.045	F
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.621	0.049	C
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.511	0.037	H

Status: Building model for fold 1...

LogitBoost

The Weka Explorer interface displays the results of a LogitBoost classifier using decision stumps. The classifier is configured with 100 iterations and a learning rate of 0.1. The test options are set to cross-validation with 10 folds. The classifier output shows a summary of performance metrics and a detailed accuracy by class table.

Classifier: LogitBoost - P 100 - L -1.7976931348623157E308 - H 1.0 - Z 3.0 - O 1 - E 1 - S 1 - I 10 - W weka.classifiers.trees.DecisionStump

Test options:

- ☐ Use training set
- ☐ Supplied test set
- ☒ Cross-validation Folds: 10
- ☐ Percentage split % 66

Classifier output:

```
Class 26 (lettr=2)

Decision Stump

Classifications

x-age <= 1.5 : 0.5373831516781108
x-age > 1.5 : -0.29945303784476335
x-age is missing : 0.01061526273262466

Number of performed iterations: 10

Time taken to build model: 17.17 seconds

=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances      14725           73.625 %
Incorrectly Classified Instances    5275           26.375 %
Kappa statistic                    0.7257
Mean absolute error                 0.0336
Root mean squared error             0.1198
Relative absolute error             45.4287 %
Root relative squared error         62.2963 %
Total Number of Instances         20000

=== Detailed Accuracy By Class ===
```

Status: OK

Vote

Weka Explorer

Preprocess | **Classify** | Cluster | Associate | Select attributes | Visualize

Classifier

Choose: **Vote** - S 1 -B "weka.classifiers.functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4" -B "weka.classifiers.lazy.IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A "weka.core.EuclideanDistance -R first-last"" -E

Test options

☐ Use training set
☐ Supplied test set (Set...)
☒ Cross-validation Folds: **10**
☐ Percentage split %: **66**
 More options...

(Nom) lettr

Start Stop

Result list (right-click for options)

- 16:27:22 - trees RandomForest
- 16:30:27 - meta Bagging
- 16:31:12 - meta AdaBoostM1
- 16:45:12 - meta AdaBoostM1
- 16:46:32 - meta AdaBoostM1
- 17:06:49 - meta LogitBoost
- 17:10:44 - meta Vote**

Classifier output

```

y-ge <= 2.5
T I D N G S B A J M X O R F C H W L P
0.07542579075425791 0.06898525833691141 0.009446114212108201 0.09131243738371261 0.0 7.156147130385E-4 4.2936882782310007E-4
y-ge > 2.5
T I D N G S B A J M X O R F C H W L P
0.02067163605625144 0.02097902097902098 0.056789364481672176 0.011142703450395757 0.059402136325213246 0.05709674940444171 0.058
y-ge is missing
T I D N G S B A J M X O R F C H W L P
0.0398 0.03775 0.04025 0.03915 0.03865 0.0374 0.0383 0.03945 0.03735 0.0396 0.03935 0.03765 0.0379 0.03875 0.0368 0.0367 0.0376 0.03805 0.040

ZeroR predicts class value: U

Time taken to build model: 134.98 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances 19205 96.025 %
Incorrectly Classified Instances 795 3.975 %
Kappa statistic 0.9587
Mean absolute error 0.0442
Root mean squared error 0.1186
Relative absolute error 59.7939 %
Root relative squared error 61.6778 %
Total Number of Instances 20000

=== Detailed Accuracy By Class ===

```

Status: OK Log x 0

Vote combines the probability distributions of these base learners:

```

weka.classifiers.functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4
weka.classifiers.lazy.IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A "weka.core.EuclideanDistance -R first-last""
weka.classifiers.trees.DecisionStump
weka.classifiers.rules.ZeroR

```

using the 'Average' combination rule

Random Forest provide the highest accuracy. Some methods has really low accuracies and long running times...

3. Eclipse

Random Forest

Weka Explorer

Preprocess **Classify** Cluster Associate Select attributes Visualize

Classifier

Choose **RandomForest** -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

Test options

☐ Use training set
☐ Supplied test set
☒ Cross-validation Folds
☐ Percentage split %

(Nom) y

Result list (right-click for options)

21:40:23 - trees.RandomForest

Classifier output

```
weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 1.08 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      3760           94 %
Incorrectly Classified Instances    240            6 %
Kappa statistic                    0.8354
Mean absolute error                 0.0652
Root mean squared error             0.2247
Relative absolute error             22.7038 %
Root relative squared error         51.8828 %
Total Number of Instances          4000

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure  MDC     ROC Area  PRC Area  Class
          0.838    0.026    0.915     0.838    0.875     0.837    0.949    0.924     0
          0.974    0.162    0.947     0.974    0.961     0.837    0.949    0.971     1
Weighted Avg.   0.940    0.128    0.939     0.940    0.939     0.837    0.949    0.959

=== Confusion Matrix ===

      a  b  <-- classified as
838 162 |  a = 0
 78 2922 | b = 1
```

Status

OK x 0

Bagging+oneR

Weka Explorer

Preprocess **Classify** Cluster Associate Select attributes Visualize

Classifier

Choose **Bagging** -P 100 -S 1 -num-slots 1 -I 10 -W weka.classifiers.rules.OneR -- -B 6

Test options

☐ Use training set
☐ Supplied test set
☒ Cross-validation Folds
☐ Percentage split %

(Nom) y

Result list (right-click for options)

21:40:23 - trees.RandomForest
21:45:15 - meta.Bagging

Classifier output

```
weka.classifiers.rules.OneR -B 6

Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      3570           89.25 %
Incorrectly Classified Instances    430           10.75 %
Kappa statistic                    0.6864
Mean absolute error                 0.1163
Root mean squared error             0.3099
Relative absolute error             30.9943 %
Root relative squared error         71.5635 %
Total Number of Instances          4000

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure  MDC     ROC Area  PRC Area  Class
          0.656    0.029    0.884     0.656    0.753    0.699    0.845    0.758     0
          0.971    0.344    0.894     0.971    0.931    0.699    0.845    0.908     1
Weighted Avg.   0.893    0.265    0.892     0.893    0.887    0.699    0.845    0.870

=== Confusion Matrix ===

      a  b  <-- classified as
656 344 |  a = 0
 86 2914 | b = 1
```

Status

OK x 0

AdaBoost

Weka Explorer

Preprocess | **Classify** | Cluster | Associate | Select attributes | Visualize

Classifier

Choose: **AdaBoostM1** - P 100 - S 1 - I 10 - W weka.classifiers.trees.DecisionStump

Test options

☐ Use training set
☐ Supplied test set
☒ Cross-validation Folds:
☐ Percentage split %

(Nom) y

Result list (right-click for options)

- 21:40:23 - trees.RandomForest
- 21:45:15 - meta.Bagging
- 21:56:00 - meta.AdaBoostM1**

Classifier output

```

Number of performed Iterations: 10

Time taken to build model: 0.06 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      3525           88.125 %
Incorrectly Classified Instances    475           11.875 %
Kappa statistic                    0.6302
Mean absolute error                 0.1702
Root mean squared error             0.2955
Relative absolute error             45.3782 %
Root relative squared error        68.2409 %
Total Number of Instances          4000


=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure  MCC   ROC Area  PRC Area  Class
          0.547    0.007    0.961    0.547    0.697    0.669  0.912    0.873    0
          0.993    0.453    0.868    0.993    0.926    0.669  0.912    0.947    1
Weighted Avg.   0.881    0.342    0.891    0.881    0.869    0.669  0.912    0.929

=== Confusion Matrix ===

      a    b  <-- classified as
547  453 |   a = 0
 22 2978 |   b = 1
  
```

Status

OK  x 0

LogitBoost

Weka Explorer

Preprocess | **Classify** | Cluster | Associate | Select attributes | Visualize

Classifier

Choose: **LogitBoost** - P 100 - L -1.7976931348623157E308 - H 1.0 - Z 3.0 - O 1 - E 1 - S 1 - I 10 - W weka.classifiers.trees.DecisionStump

Test options

☐ Use training set
☐ Supplied test set
☒ Cross-validation Folds:
☐ Percentage split %

(Nom) y

Result list (right-click for options)

- 21:40:23 - trees.RandomForest
- 21:45:15 - meta.Bagging
- 21:56:00 - meta.AdaBoostM1
- 21:56:28 - meta.LogitBoost**

Classifier output

```

Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      3714           92.85 %
Incorrectly Classified Instances    286           7.15 %
Kappa statistic                    0.8025
Mean absolute error                 0.1232
Root mean squared error             0.2452
Relative absolute error             32.8371 %
Root relative squared error        56.6371 %
Total Number of Instances          4000


=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure  MCC   ROC Area  PRC Area  Class
          0.805    0.030    0.898    0.805    0.849    0.805  0.930    0.897    0
          0.970    0.195    0.937    0.970    0.953    0.805  0.930    0.957    1
Weighted Avg.   0.929    0.154    0.927    0.929    0.927    0.805  0.930    0.942

=== Confusion Matrix ===

      a    b  <-- classified as
805  195 |   a = 0
 91 2909 |   b = 1
  
```

Status

OK  x 0

Vote

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose **Vote** -S 1 -B "weka.classifiers.rules.OneR -B 6" -B "weka.classifiers.trees.DecisionStump" -B "weka.classifiers.lazy.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) y

Start Stop

Result list (right-click for options)

- 21:40:23 - trees RandomForest
- 21:45:15 - meta Bagging
- 21:56:00 - meta AdaBoostM1
- 21:56:23 - meta LogitBoost
- 21:59:14 - meta Vote**

Classifier output

Time taken to build model: 0.05 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	3591	89.775 %
Incorrectly Classified Instances	409	10.225 %
Kappa statistic	0.6933	
Mean absolute error	0.1718	
Root mean squared error	0.2757	
Relative absolute error	45.8086 %	
Root relative squared error	63.6797 %	
Total Number of Instances	4000	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
Weighted Avg.	0.629	0.013	0.943	0.629	0.755	0.716	0.922	0.887	0
	0.987	0.371	0.889	0.987	0.935	0.716	0.922	0.945	1

=== Confusion Matrix ===

a	b	<-- classified as
629	371	a = 0
38	2962	b = 1

Status

OK Log x 0

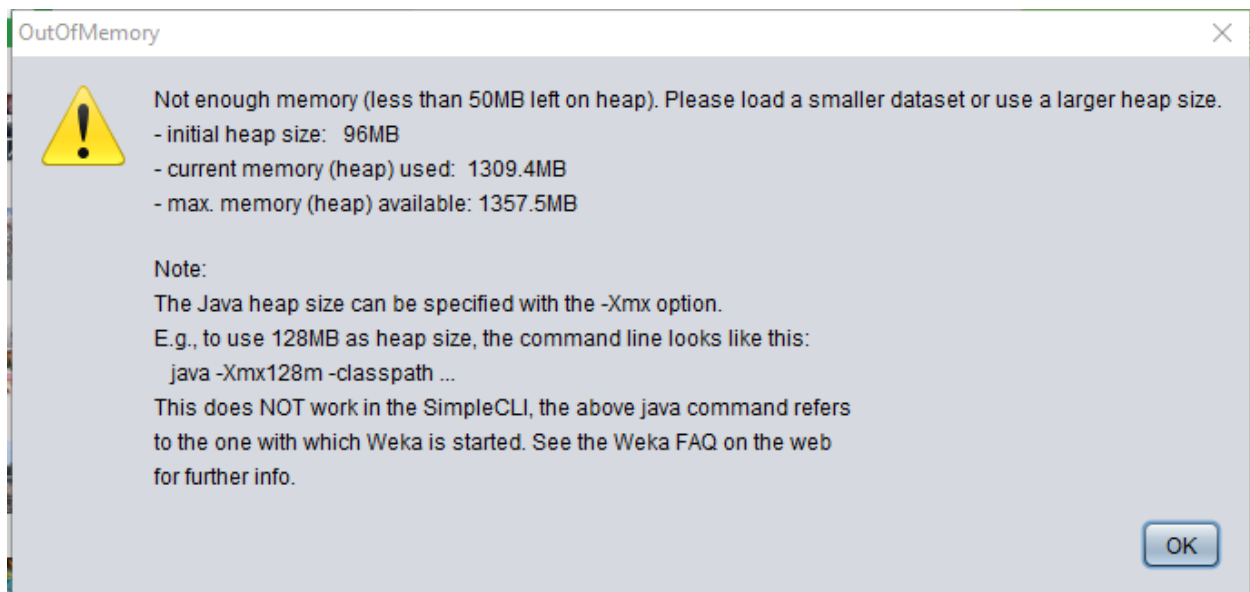
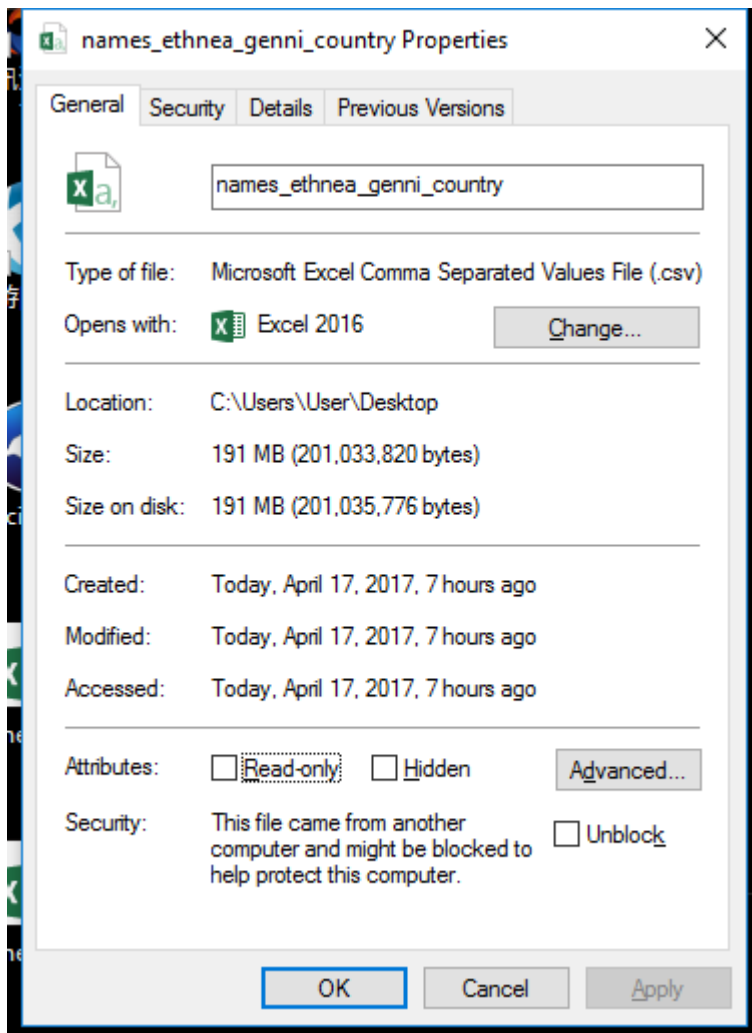
Vote combines the probability distributions of these base learners:

```
weka.classifiers.rules.OneR -B 6
weka.classifiers.trees.DecisionStump
weka.classifiers.lazy.IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""
weka.classifiers.functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4
```

using the 'Average' combination rule

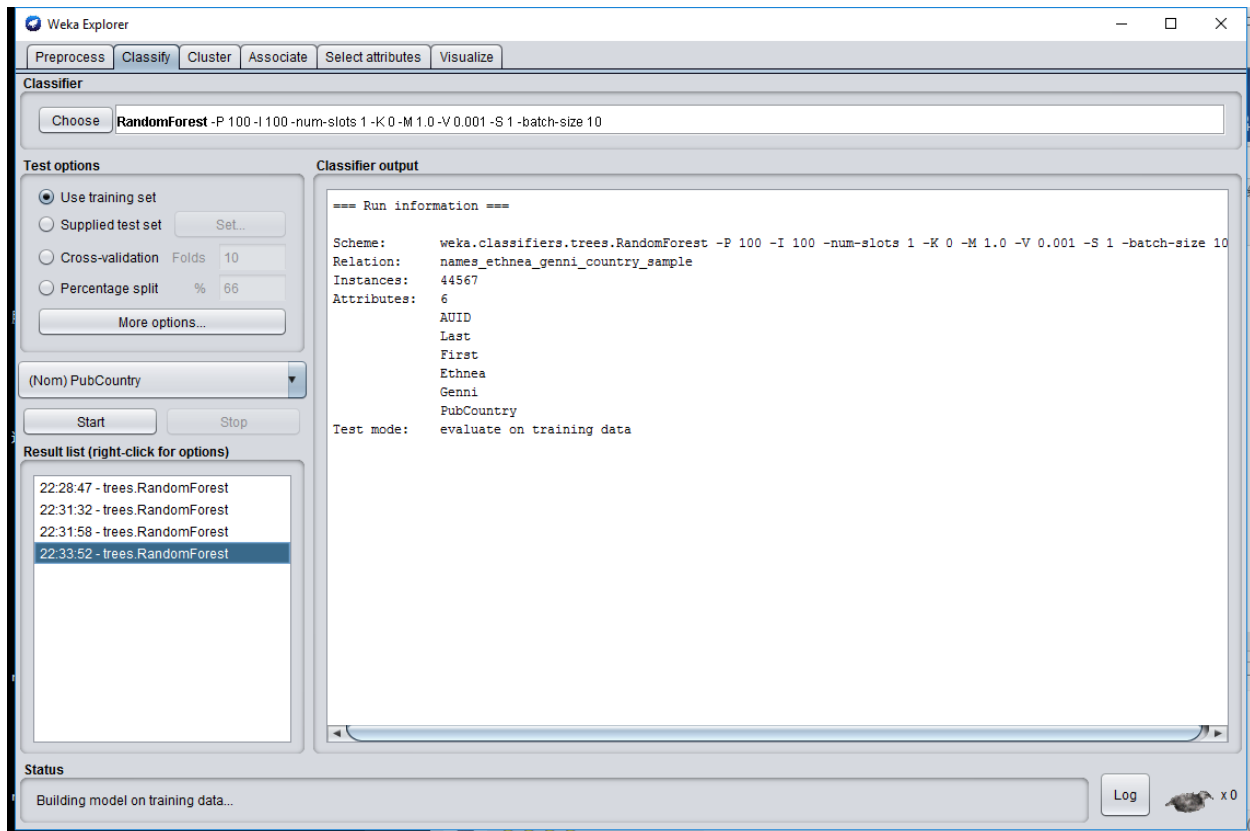
Random Forest still has the highest accuracy...

- names_ethnea_genni_country.csv (cannot reading the file, not enough memory)

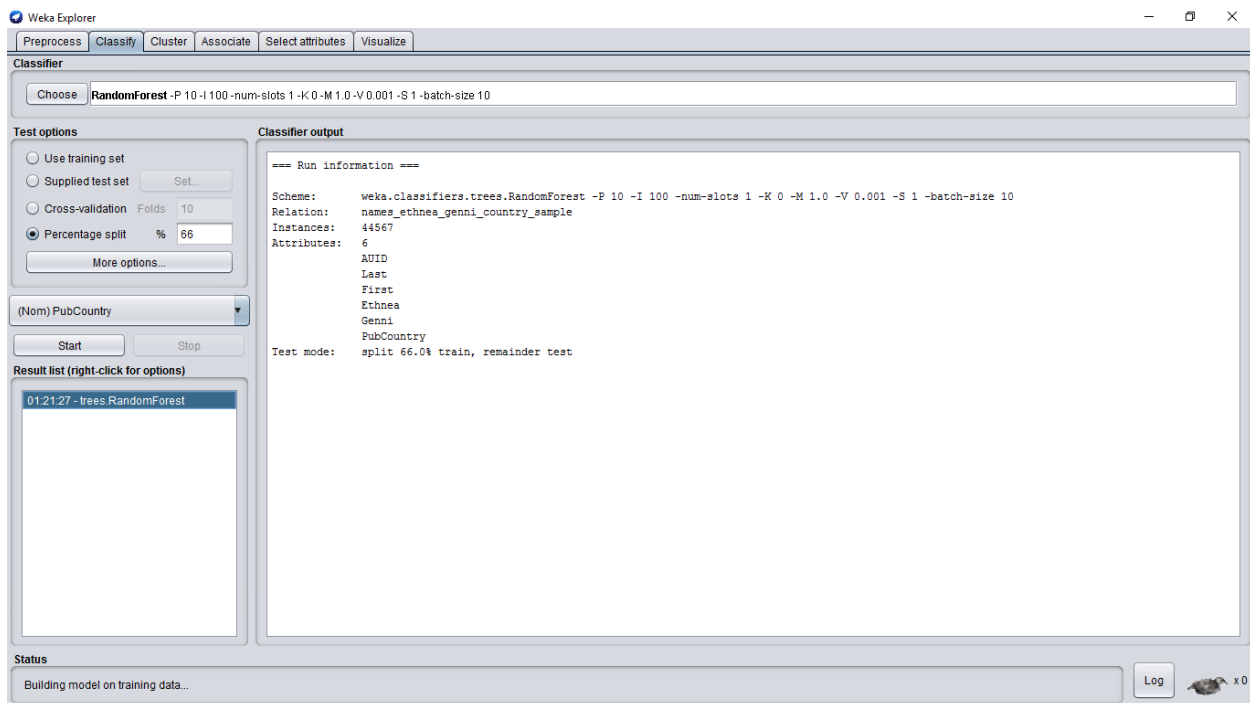


5. names_ethnea_genni_country_sample.csv

my computer can read the file but cannot run any classifiers...



I adjust the batch size to 10 and it still cannot run



I change the iteration to 1 and it ends like this...

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. The classifier chosen is 'RandomForest -P 10 -I 1 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1 -batch-size 5'. The 'Test options' section shows 'Percentage split' at 66%. The 'Classifier output' pane displays the following results:

```
Test mode: split 66.0% train, remainder test

=== Classifier model (full training set) ===

RandomForest

Bagging with 1 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities -batch-size 5

Time taken to build model: 1.69 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.29 seconds

=== Summary ===

Correctly Classified Instances      4896      32.3104 %
Incorrectly Classified Instances    10257      67.6896 %
Kappa statistic                     0
Mean absolute error                 0.0113
Root mean squared error            0.0751
Relative absolute error            100.2991 %
Root relative squared error        100.0154 %
Total Number of Instances         15153

=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	1.000	0.024	1.000	0.046	0.000	0.500	0.024	Brazil
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.033	France
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.322	USA
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.001	Tunisia
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.002	Egypt
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.000	Ghana
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.032	Spain
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.004	Iran
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.001	UnitedArabEmirates
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.010	Turkey
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.098	Japan
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.002	SaudiArabia
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.000	Qatar
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.015	India
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.070	UK
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.033	Taiwan

The accuracy becomes low...

Bagging+1R

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. The classifier chosen is 'Bagging -P 100 -S 1 -num-slots 1 -I 10 -W weka.classifiers.rules.OneR ---B 6'. The 'Test options' section shows 'Cross-validation' with 'Folds' set to 10. The 'Classifier output' pane displays the following results:

```
=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances      1054      2.365 %
Incorrectly Classified Instances    43513      97.635 %
Kappa statistic                     0
Mean absolute error                 0.0127
Root mean squared error            0.1126
Relative absolute error            112.3325 %
Root relative squared error        149.9293 %
Total Number of Instances         44567

=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	1.000	0.024	1.000	0.046	0.000	0.500	0.024	Brazil
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.033	France
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.322	USA
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.001	Tunisia
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.002	Egypt
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.000	Ghana
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.032	Spain
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.004	Iran
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.001	UnitedArabEmirates
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.010	Turkey
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.098	Japan
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.002	SaudiArabia
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.000	Qatar
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.015	India
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.070	UK
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.033	Taiwan

The accuracy is low...

AdaBoostM1

Weka Explorer

Preprocess **Classify** Cluster Associate Select attributes Visualize

Classifier

Choose **AdaBoostM1** -P 100 -S 1 -I 10 -W weka.classifiers.trees.DecisionStump

Test options

☐ Use training set
☐ Supplied test set
☒ Cross-validation Folds **10**
☐ Percentage split % **66**

(Nom) PubCountry

Result list (right-click for options)

- 23:20:20 - meta Bagging
- 23:21:21 - meta AdaBoostM1**

Classifier output

```

Ethnea is missing : USA

Class distributions

Ethnea = ENGLISH
Brazil France USA Tunisia Egypt Ghana Spain Iran UnitedArabEmirates Turkey Japan SaudiArabia Qatar India UK Italy
9.01275651691625E-4 0.003951747088186356 0.6446894065446478 0.0 1.3865779256794233E-4 6.932889628397117E-5 6.239600665557404E-4
Ethnea != ENGLISH
Brazil France USA Tunisia Egypt Ghana Spain Iran UnitedArabEmirates Turkey Japan SaudiArabia Qatar India UK Italy
0.03453538134890356 0.04674385429452941 0.1680655419168628 0.0016587599110904688 0.0033175198221809376 1.658759911090468E-4 0.046
Ethnea is missing
Brazil France USA Tunisia Egypt Ghana Spain Iran UnitedArabEmirates Turkey Japan SaudiArabia Qatar India UK Italy
0.02364978571588844 0.032894293984338185 0.32232369241815695 0.0011219063432584648 0.00228868940247268 1.3462876119101576E-4 0.031


Time taken to build model: 0.81 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      14365      32.2324 %
Incorrectly Classified Instances    30202      67.7676 %
Kappa statistic                    0
Mean absolute error                 0.0104
Root mean squared error             0.0723
Relative absolute error             92.513 %
Root relative squared error         96.2104 %
Total Number of Instances          44567

```

Status

OK  x 0

Accuracy higher but still not high enough

LogitBoost

Weka Explorer

Preprocess **Classify** Cluster Associate Select attributes Visualize

Classifier

Choose **Vote** -S 1 -B "weka.classifiers.functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4" -B "weka.classifiers.lazy.IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A W"weka.core.EuclideanDistance -R first-last" -E

Test options

☐ Use training set
☐ Supplied test set
☒ Cross-validation Folds **10**
☐ Percentage split % **66**

(Nom) PubCountry

Result list (right-click for options)

- 23:20:20 - meta Bagging
- 23:21:21 - meta AdaBoostM1
- 23:22:48 - meta LogitBoost**

Classifier output

```

Class 154 (PubCountry=Grenada)

Decision Stump

Classifications

AUID = 12199756_2 : 1.000260878194961
AUID != 12199756_2 : -1.0000008027850509
AUID is missing : -0.9830175317645159

Number of performed iterations: 10

Time taken to build model: 28.13 seconds


=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      26082      58.5231 %
Incorrectly Classified Instances    18485      41.4769 %
Kappa statistic                    0.5117
Mean absolute error                 0.0075
Root mean squared error             0.0611
Relative absolute error             66.0742 %
Root relative squared error         81.4058 %
Total Number of Instances          44567

=== Detailed Accuracy By Class ===

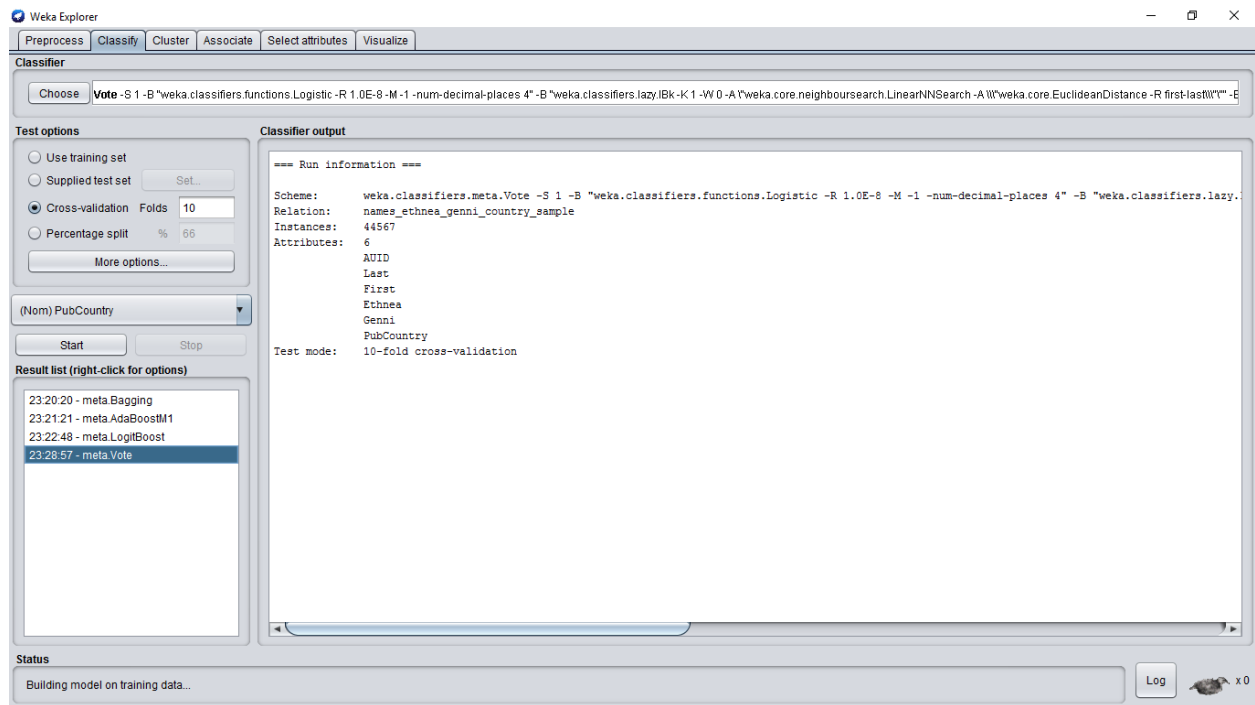
```

Status

OK  x 0

Still not accurate.

Vote



The bird go die again...

Not all classifiers can run on the sample csv, even if they can, the accurate rate is not high as expected.