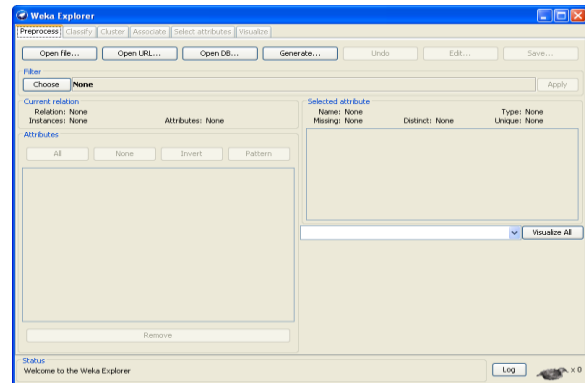
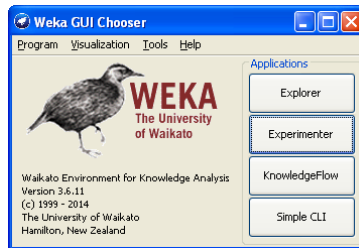


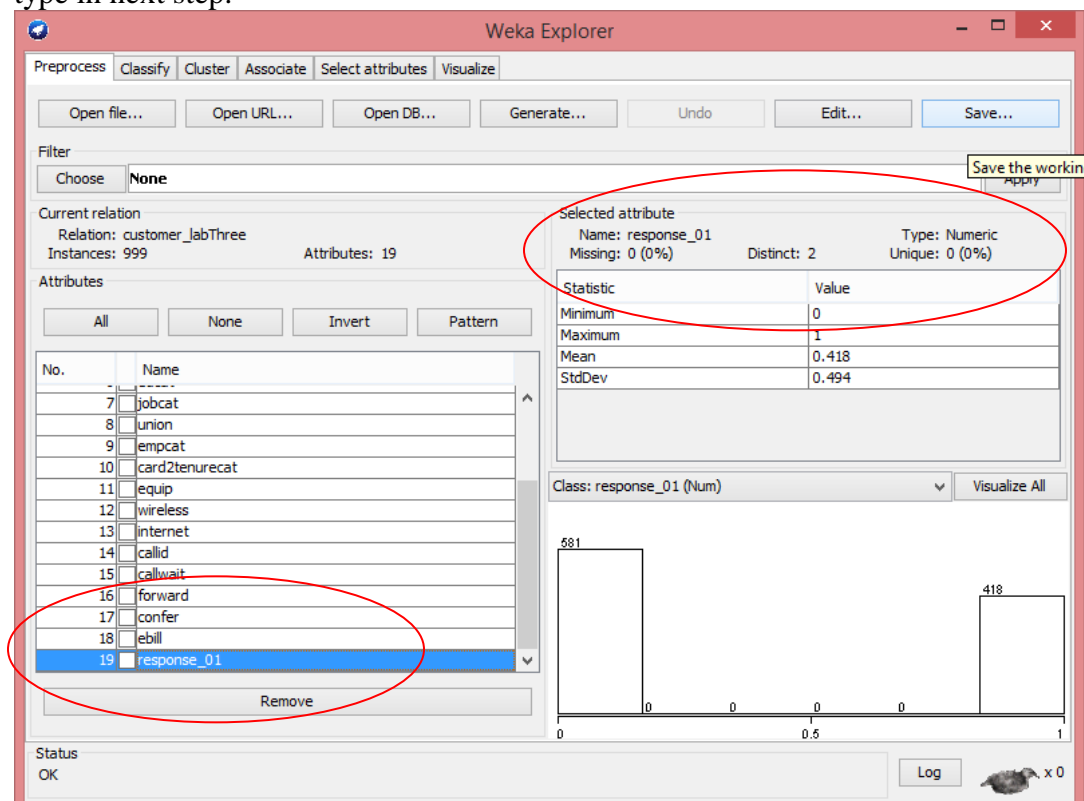
Lab Exercise Three

Classification with WEKA Explorer

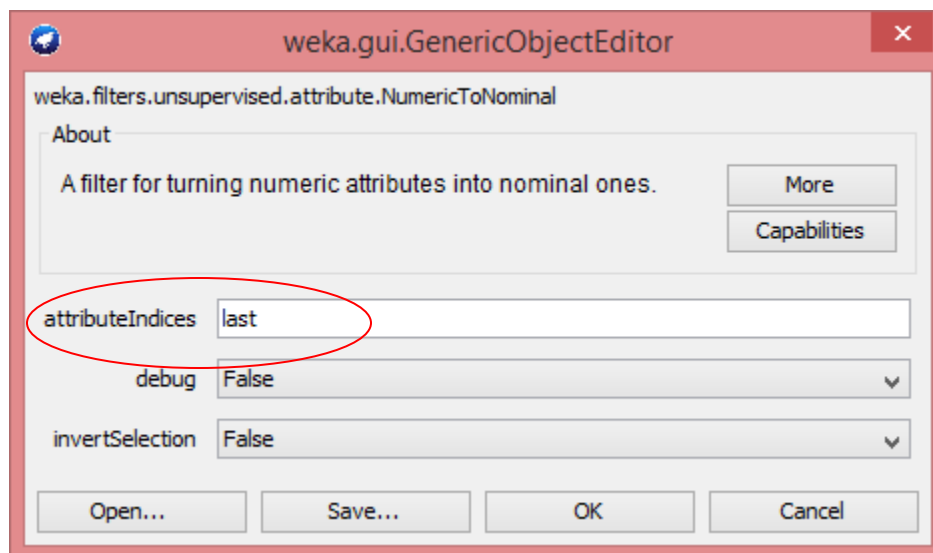
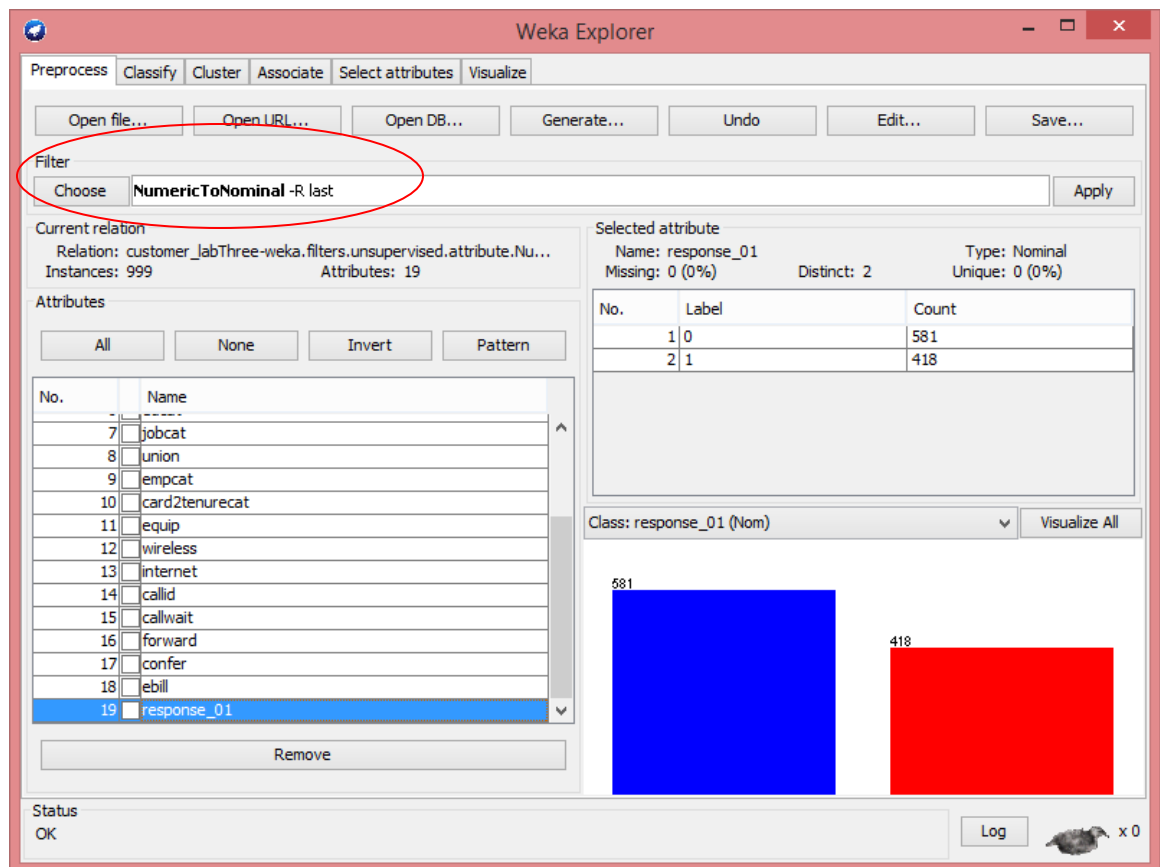
1. Fire up WEKA to get the GUI Chooser panel. Select Explorer from the four choices on the right side.



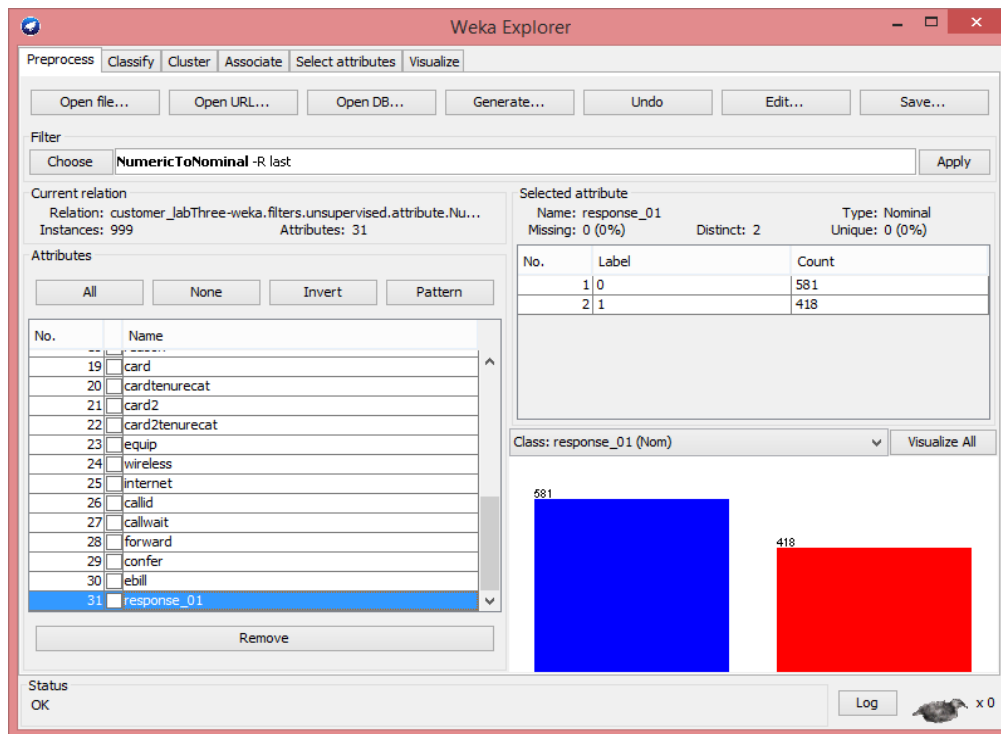
2. We are on **Preprocess** now. Click the **Open file** button to bring up a standard dialog through which you can select a file. Choose the **customer_labThree.csv** file.
3. To perform classification with Weka, the last attribute in the dataset is taken as class label and it should be *nominal*. Since the last attribute of data set **customer_labThree.csv** is *numeric* type (1/0), we should convert it to *nominal* type in next step.



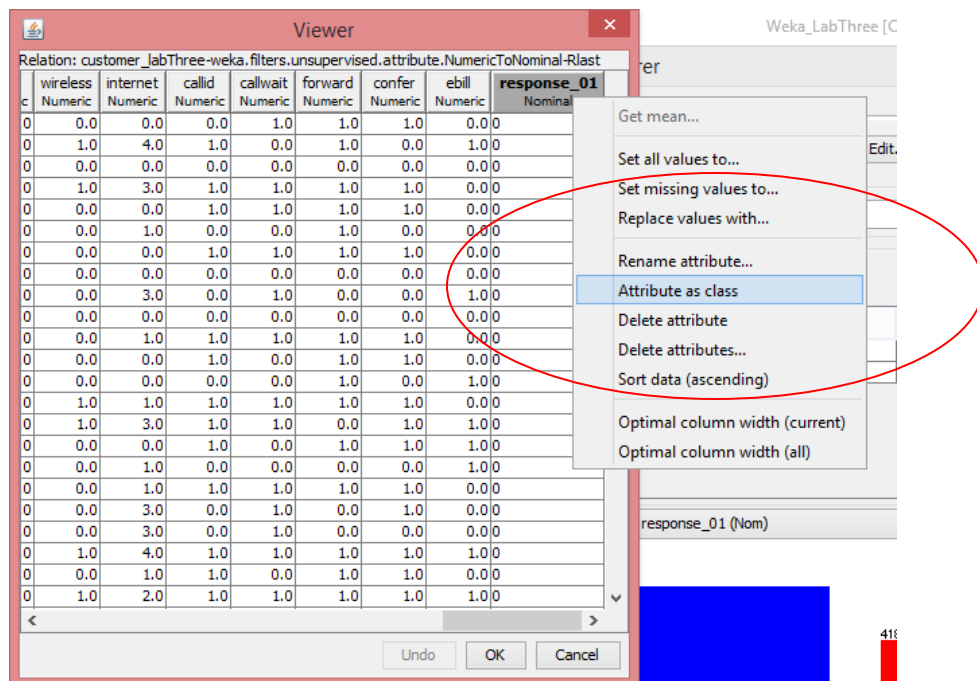
4. Unsupervised attribute filter – *NumericToNominal* is chosen to perform this conversion. Since we would like to convert the last attribute only, change the *attributeIndices* to **last**.



5. After applying the filter, the last attribute becomes nominal type and it is taken as the class label for the dataset – now the data set is visualized in two colors.



6. If the class attribute is not the last attribute, you could set it in edit window.



7. You should also convert the types of other attributes. Attributes region, townsize, agecat, jobcat, empcat, card2tenurecat, and internet are all nominal values, however, they are treated as numeric type by Weka. And attributes gender, union, equip, wireless, called, callwait, forward, confer, ebill are binary values,

they are treated as numeric types as well. **NumericToNominal** filter should be applied to convert them. You could also normalize attribute educat to [0, 1] since education categories are rankings.

Weka Explorer

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

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter: Choose **Normalize -S 1.0 -T 0.0** Apply

Current relation: Relation: customer_labThree-weka.filters.unsupervised.attribute.Nu...
Instances: 999 Attributes: 19

Attributes: All None Invert Pattern

No.	Name
1	<input checked="" type="checkbox"/> custid
2	<input checked="" type="checkbox"/> region
3	<input type="checkbox"/> townsize
4	<input type="checkbox"/> gender
5	<input type="checkbox"/> agecat
6	<input type="checkbox"/> educat
7	<input type="checkbox"/> jobcat
8	<input type="checkbox"/> union
9	<input type="checkbox"/> empcat
10	<input type="checkbox"/> card2tenurecat
11	<input type="checkbox"/> equip
12	<input type="checkbox"/> wireless
13	<input type="checkbox"/> internet

Remove

Selected attribute: Name: region Missing: 0 (0%) Distinct: 5 Type: Nominal Unique: 0 (0%)

No.	Label	Count
1	1	195
2	2	203
3	3	217
4	4	176
5	5	208

Class: response_01 (Nom) Visualize All

Status OK Log x 0

Attribute Selection - Since not all attributes are relevant to the classification job, you should perform attribute selection before training the classifier.

- You could remove irrelevant attributes by hand. For example, the first attribute ***custId*** should be removed. Select it and click **Remove** button to remove it.

Weka Explorer

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

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter: Choose **NumericToNominal -R last** Apply

Current relation: Relation: customer_labThree-weka.filters.unsupervised.attribute.Nu...
Instances: 999 Attributes: 31

Attributes: All None Invert Pattern

No.	Name
1	<input checked="" type="checkbox"/> custid
2	<input checked="" type="checkbox"/> region
3	<input type="checkbox"/> townsize
4	<input type="checkbox"/> gender
5	<input type="checkbox"/> agecat
6	<input type="checkbox"/> educat
7	<input type="checkbox"/> jobcat
8	<input type="checkbox"/> union
9	<input type="checkbox"/> empcat
10	<input type="checkbox"/> retire
11	<input type="checkbox"/> inccat
12	<input type="checkbox"/> jobcat
13	<input type="checkbox"/> reside

Remove

Selected attribute: Name: custid Missing: 0 (0%) Distinct: 999 Type: Nominal Unique: 999 (100%)

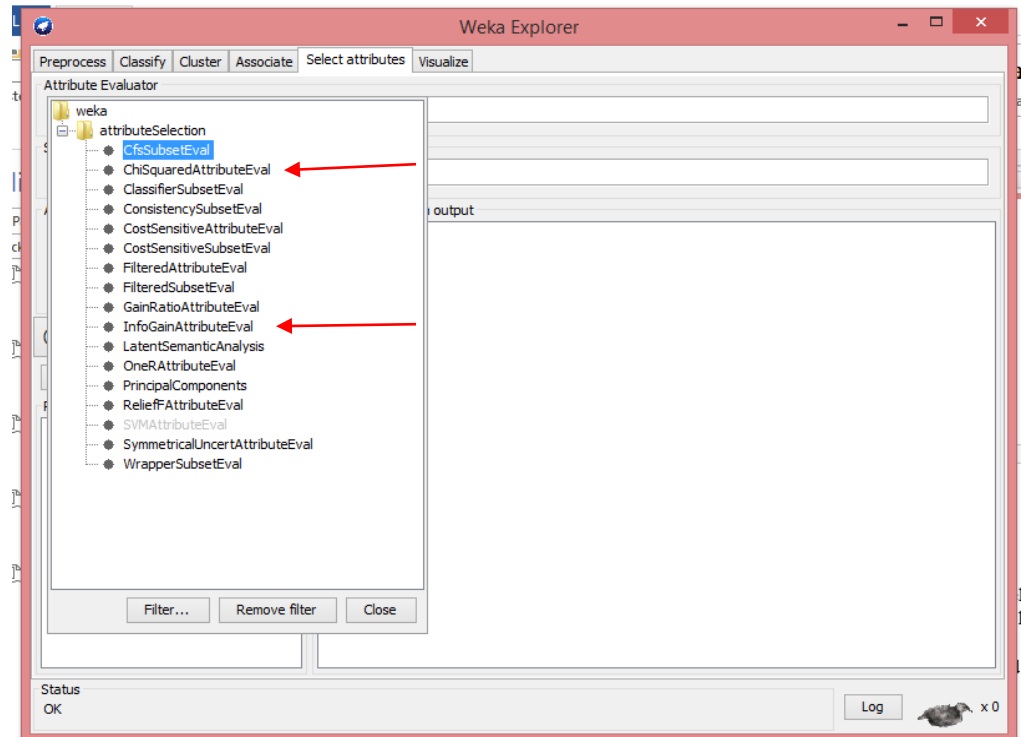
No.	Label	Count
1	4459-VLPQUH-3OL	1
2	9124-DZALHM-S6I	1
3	2228-KOLOPU-FY3	1
4	2866-TTOTRL-TA7	1
5	7217-UECHSF-PCR	1
6	4166-WEDNKN-SRK	1
7	1114-UELZXT-QT7	1

Class: response_01 (Nom) Visualize All

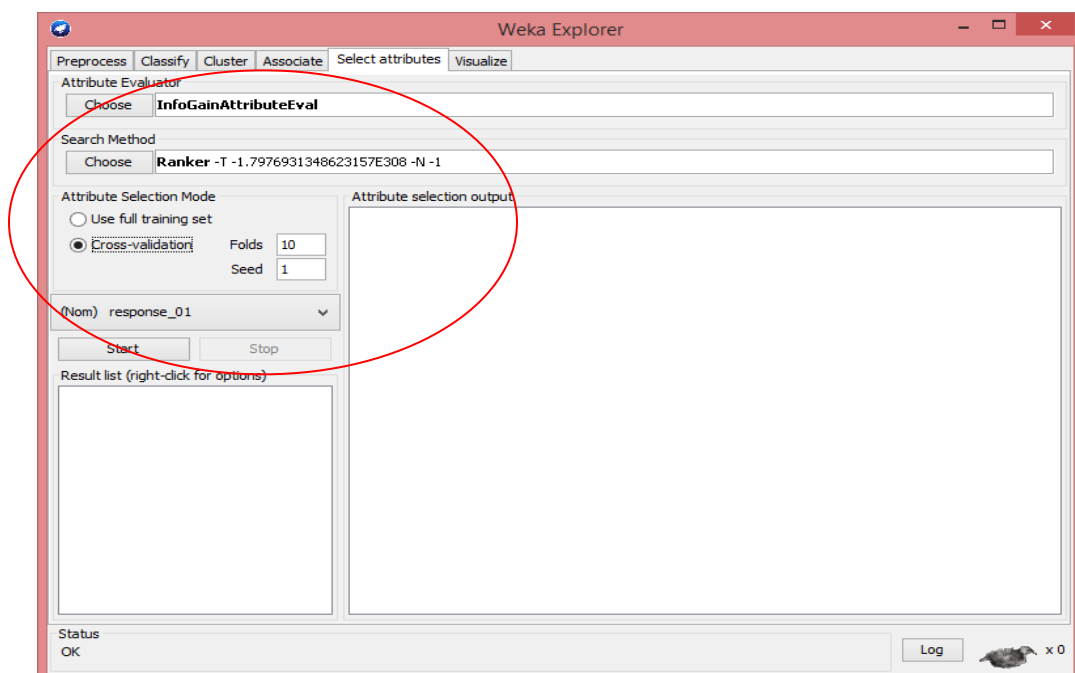
Too many values to display.

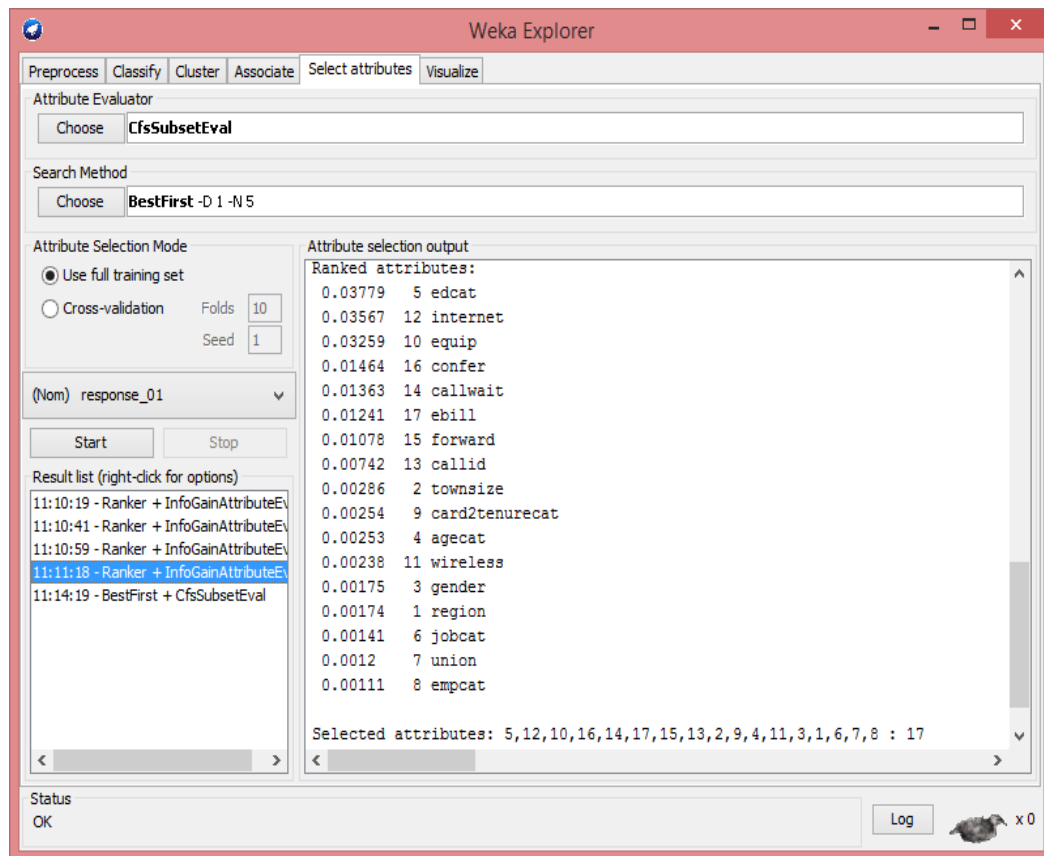
Status OK Log x 0

9. You also could run automatic attribute selection. We have introduced two methods of evaluating attributes individually – **InfoGainAttributeEval** and **ChiSquaredAttributeEval**. The default attribute selection method of Weka is **CfsSubsetEval**, which evaluates subsets of attributes.

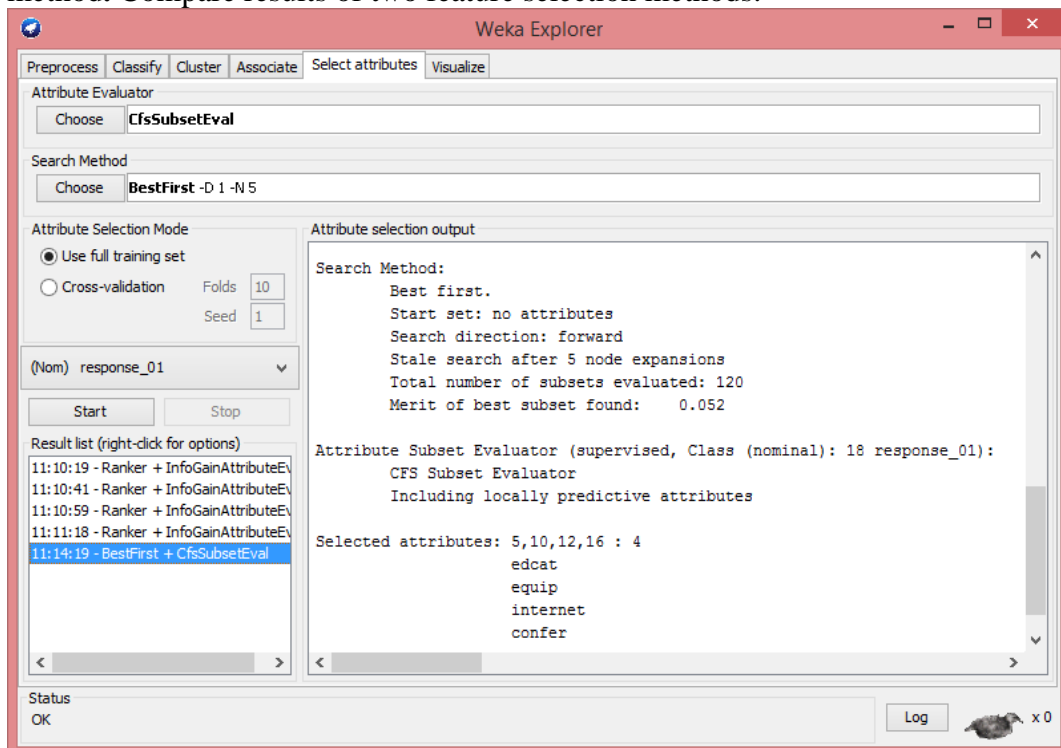


10. To use evaluator **InfoGainAttributeEval**, a search method **Ranker** is selected to rank all attributes regarding the evaluation results. We use the full dataset as training dataset. The results show that the first 8 attributes are good.

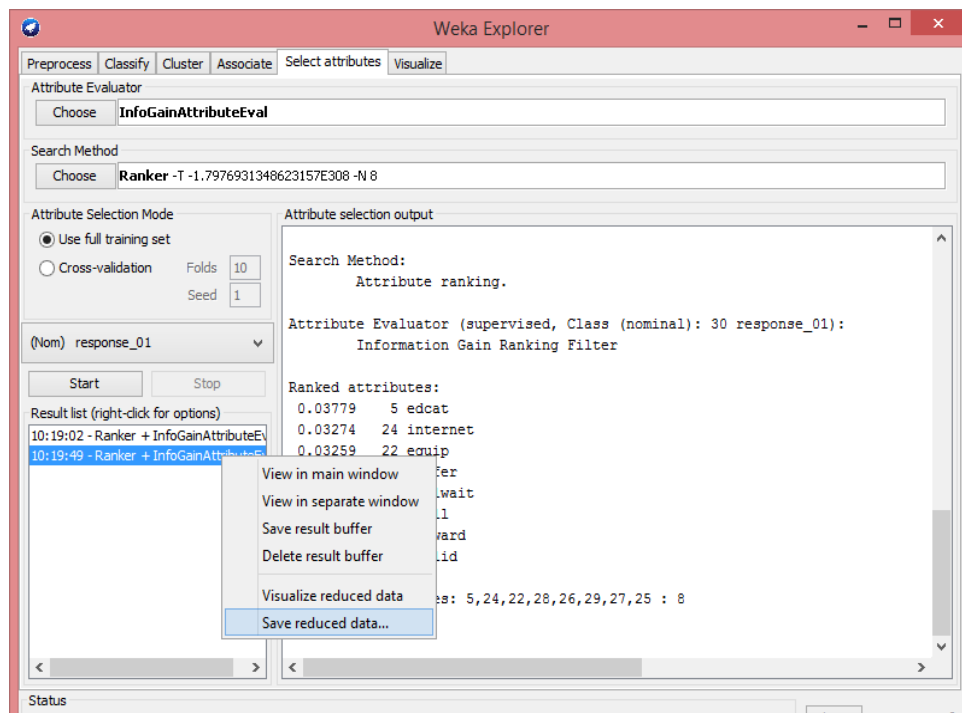




- Run feature selection the second time with **CfsSubsetEval** and **BestFirst** search method. Compare results of two feature selection methods.

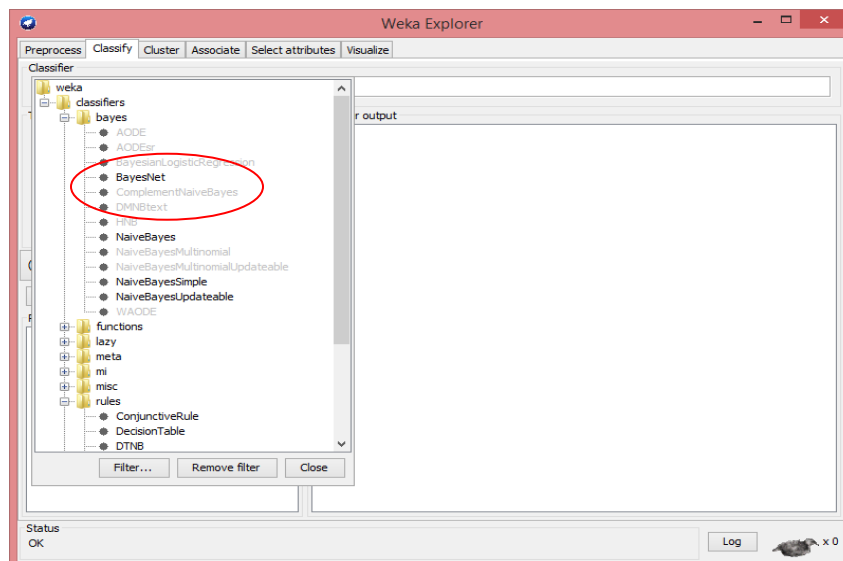


12. If you decide to reduce the dataset by removing unimportant attributes, you could choose to save the reduced dataset by right-click the Result list. Save the file name as **customer.arff**.

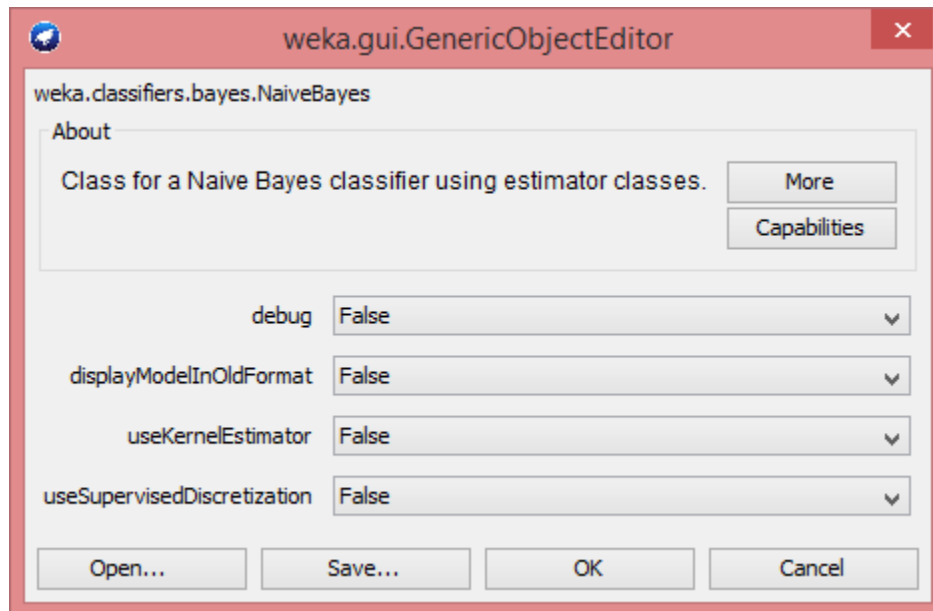


Naïve Bayes Classifier: bayes/NaiveBayes

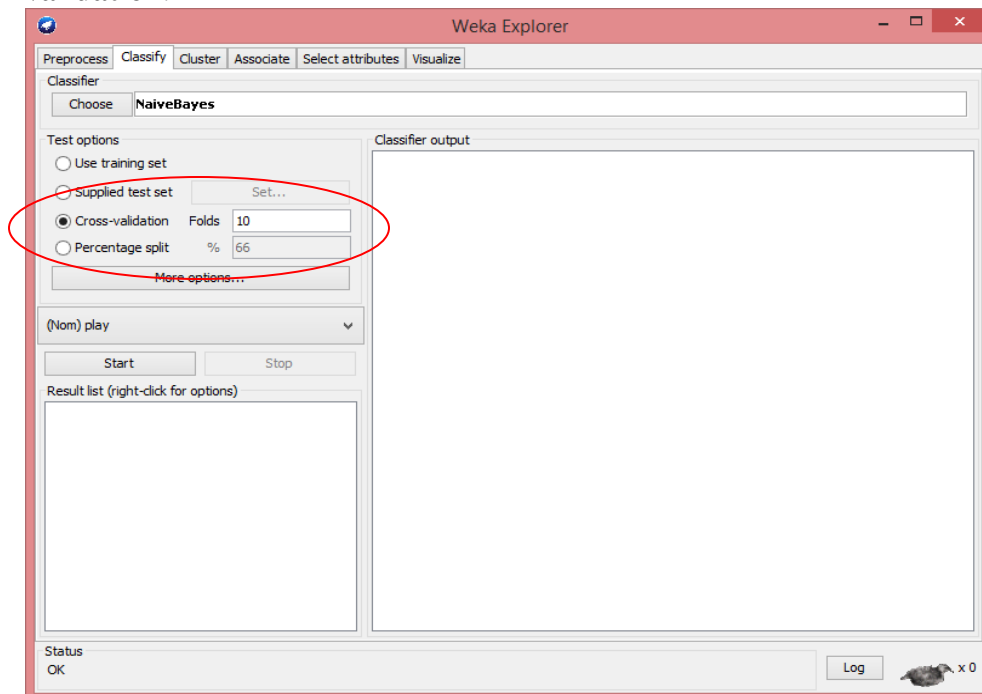
13. Open the saved processed data file **customer.arff** and then click **Classify** Tab on top of the window. Click Choose button under *Classifier*. The drop down list of all classifiers show. Choose **NaiveBayes** from **bayes** folder.



14. Left click the field of **Classifier**, choose Show Property from the drop down list. The property window of **NaiveBayes** opens, if you do not want to use Normal Distribution for numeric data, set *useKernelEstimator* to **true**; You also could perform supervised discretization on numeric data by setting *useSupervisedDiscretization* to **true**. Click OK button to save all the settings.



15. To partition the training data set and test data set, choose **10-fold cross-validation**.



16. Click **Start** button on the left of the window, the algorithm begins to run. The output is showing in the right window.

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier: Choose **NaiveBayes**

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

(Nom) response_01

Start Stop

Result list (right-click for options):
11:23:37 - bayes.NaiveBayes

Classifier output:

Attribute	Class 0	Class 1
=====		
edcat		
mean	0.4355	0.305
std. dev.	0.2892	0.2761
weight sum	581	418
precision	0.25	0.25
equip		
0	361.0	341.0
1	222.0	79.0
[total]	583.0	420.0
internet		
0	269.0	283.0
1	100.0	52.0
2	71.0	32.0
3	72.0	34.0
4	74.0	22.0
[total]	586.0	423.0
confer		
0	304.0	159.0
1	279.0	261.0
[total]	583.0	420.0

parameters of normal distributions for numeric

frequency counts of nominal values

NaiveBayes avoids zero frequencies by applying the Laplace correction.

Status: OK Log x 0

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier: Choose **NaiveBayes**

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

(Nom) response_01

Start Stop

Result list (right-click for options):
11:23:37 - bayes.NaiveBayes

Classifier output:

Time taken to build model: 0.03 seconds

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

Metric	Value	Percentage
Correctly Classified Instances	616	61.6617 %
Incorrectly Classified Instances	383	38.3383 %
Kappa statistic	0.2235	
Mean absolute error	0.4267	
Root mean squared error	0.4831	
Relative absolute error	87.6667 %	
Root relative squared error	97.9272 %	
Total Number of Instances	999	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
	0.633	0.407	0.684	0.633	0.658	0.663
	0.593	0.367	0.538	0.593	0.564	0.663
Weighted Avg.	0.617	0.39	0.623	0.617	0.619	0.663

=== Confusion Matrix ===

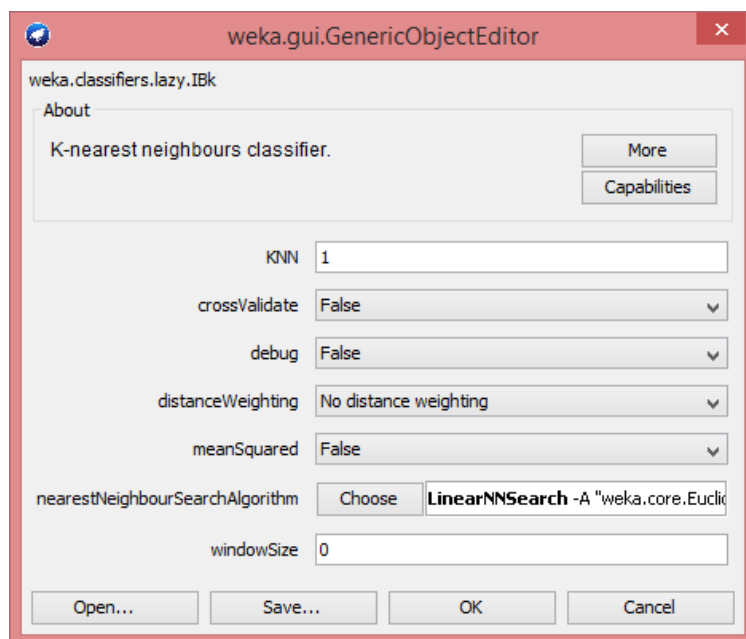
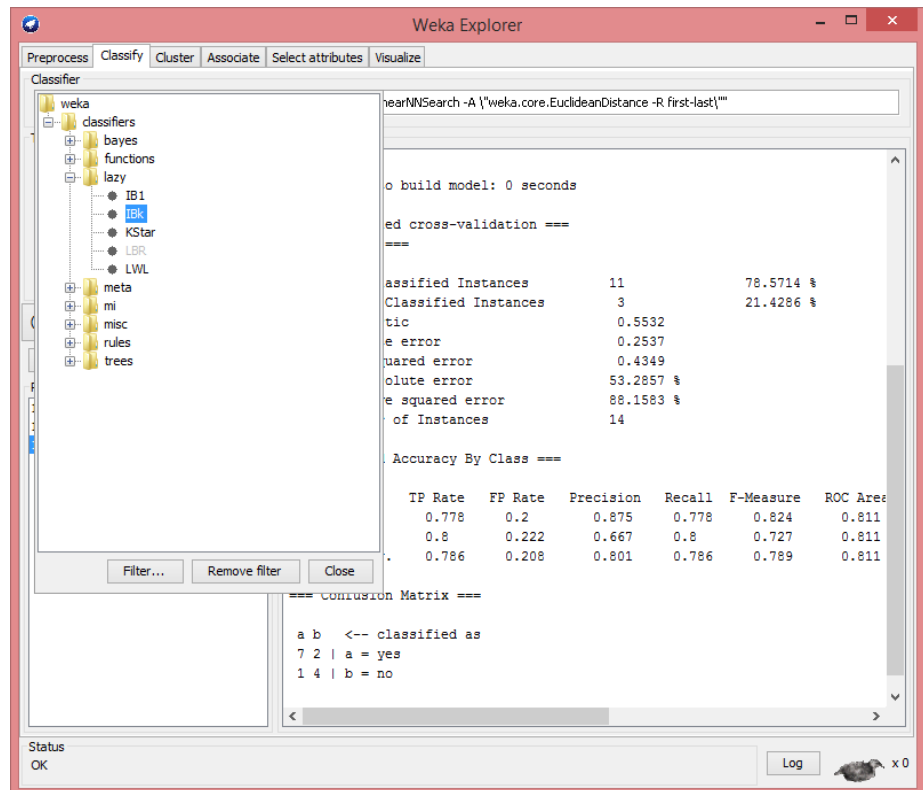
a	b	-- classified as
368	213	a = 0
170	248	b = 1

Accuracy

Status: OK Log x 0

K-Nearest-Neighbor: lazy/IBK

17. We would like to perform K-Nearest-Neighbor classification on the same dataset. You could try different K and see what value gives a better result. Compare the results with Naïve Bayes classifier.



Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose **IBk -K 3 -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) response_01

Start Stop

Result list (right-click for options)

12:27:06 - bayes.NaiveBayes

12:27:48 - bayes.NaiveBayes

12:32:16 - lazy.IBk

Classifier output

Correctly Classified Instances 596 59.6597 %

Incorrectly Classified Instances 403 40.3403 %

Kappa statistic 0.1521

Mean absolute error 0.4414

Root mean squared error 0.4891

Relative absolute error 90.6828 %

Root relative squared error 99.1419 %

Total Number of Instances 999

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
	0.711	0.562	0.637	0.711	0.672	0.628
	0.438	0.289	0.521	0.438	0.476	0.628
Weighted Avg.	0.597	0.448	0.589	0.597	0.59	0.628

=== Confusion Matrix ===

a b <-- classified as

413 168 | a = 0

235 183 | b = 1

Status OK

Log x0

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose **IBk -K 20 -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) response_01

Start Stop

Result list (right-click for options)

12:27:06 - bayes.NaiveBayes

12:27:48 - bayes.NaiveBayes

12:32:16 - lazy.IBk

12:40:47 - lazy.IBk

12:41:02 - lazy.IBk

12:41:19 - lazy.IBk

Classifier output

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 607 60.7608 %

Incorrectly Classified Instances 392 39.2392 %

Kappa statistic 0.1599

Mean absolute error 0.4423

Root mean squared error 0.4761

Relative absolute error 90.8728 %

Root relative squared error 96.5204 %

Total Number of Instances 999

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
	0.766	0.612	0.635	0.766	0.694	0.654
	0.388	0.234	0.544	0.388	0.453	0.654
Weighted Avg.	0.608	0.454	0.597	0.608	0.593	0.654

=== Confusion Matrix ===

a b <-- classified as

445 136 | a = 0

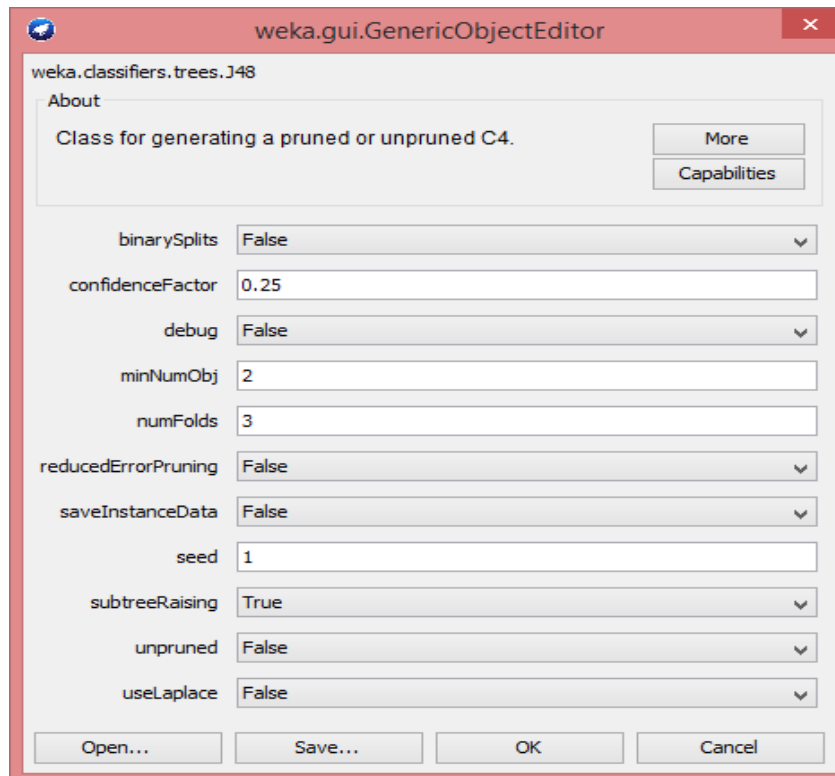
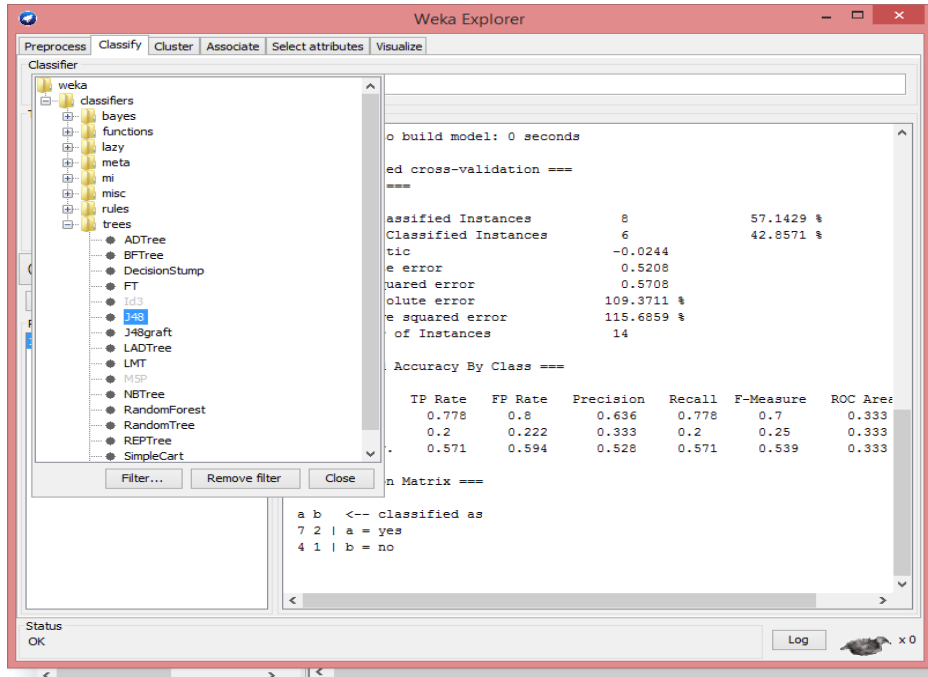
256 162 | b = 1

Status OK

Log x0

Decision Tree: trees/J48 (Implementing C4.5)

18. We would like to build a Decision Tree model on the same given training data set. Take all default values of the parameters.



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) response_01

Start Stop

Result list (right-click for options)

- 12:27:06 - bayes.NaiveBayes
- 12:27:48 - bayes.NaiveBayes
- 12:32:16 - lazy.IBk
- 12:40:47 - lazy.IBk
- 12:41:02 - lazy.IBk
- 12:41:19 - lazy.IBk
- 12:46:25 - trees.J48**

Classifier output

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: customer_labThree-weka.filters.unsupervised.attribute.Numeric

Instances: 999

Attributes: 5

edcat

equip

internet

confer

response_01

Test mode:10-fold cross-validation

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

J48 pruned tree

equip = 0

| edcat <= 0.25

| | confer = 0: 0 (213.0/101.0)

| | confer = 1: 1 (254.0/98.0)

| edcat > 0.25: 0 (233.0/83.0)

equip = 1: 0 (299.0/78.0)

Number of Leaves : 4

Size of the tree : 7

Status OK Log x 0

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) response_01

Start Stop

Result list (right-click for options)

- 12:27:06 - bayes.NaiveBayes
- 12:27:48 - bayes.NaiveBayes
- 12:32:16 - lazy.IBk
- 12:40:47 - lazy.IBk
- 12:41:02 - lazy.IBk
- 12:41:19 - lazy.IBk
- 12:46:25 - trees.J48**

Classifier output

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	613	61.3614 %
Incorrectly Classified Instances	386	38.6386 %
Kappa statistic	0.1693	
Mean absolute error	0.4571	
Root mean squared error	0.4826	
Relative absolute error	93.9202 %	
Root relative squared error	97.832 %	
Total Number of Instances	999	

=== Detailed Accuracy By Class ===

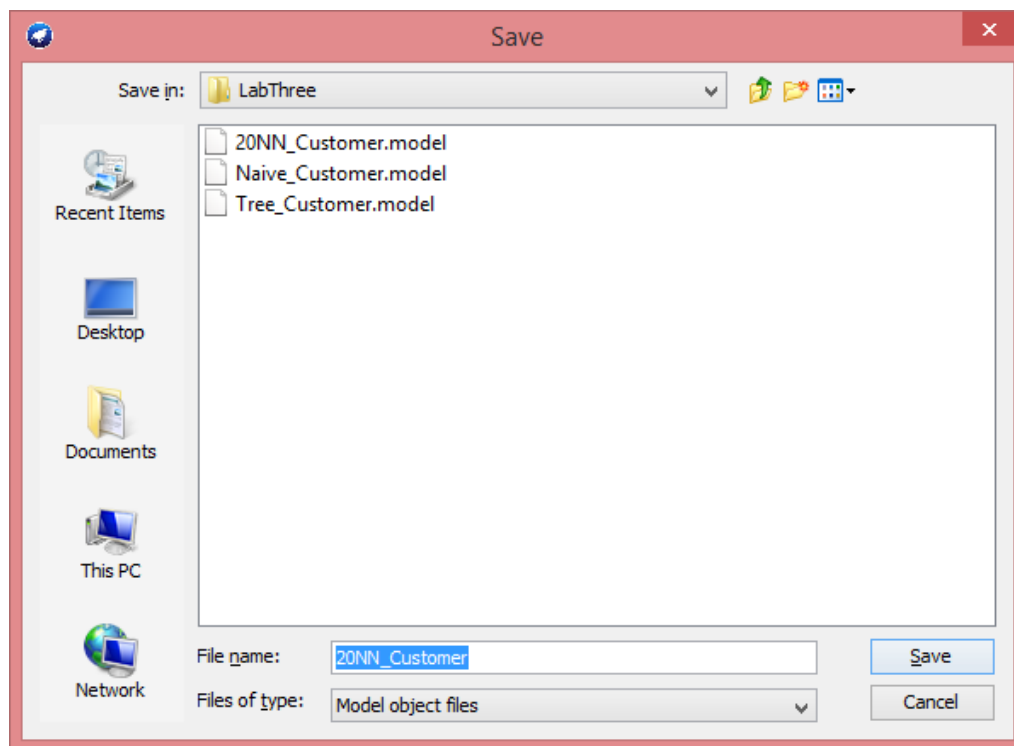
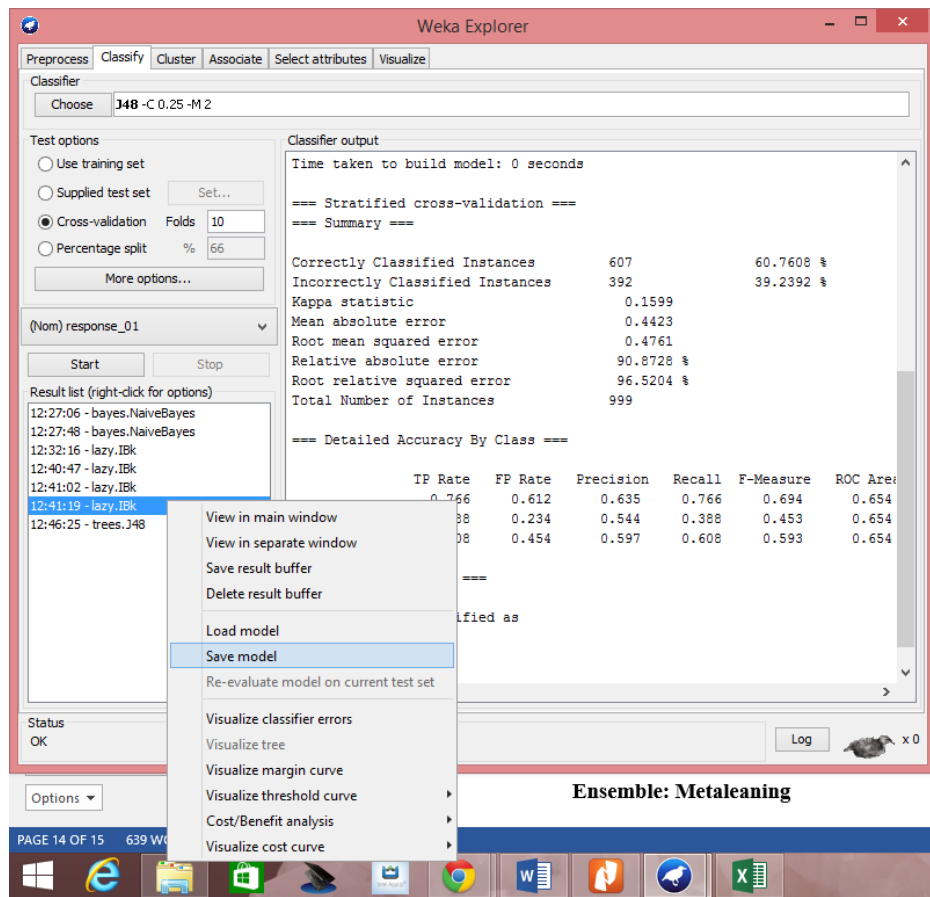
	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
	0.781	0.62	0.637	0.781	0.702	0.619
	0.38	0.219	0.556	0.38	0.452	0.619
Weighted Avg.	0.614	0.452	0.603	0.614	0.597	0.619

=== Confusion Matrix ===

a b <-- classified as

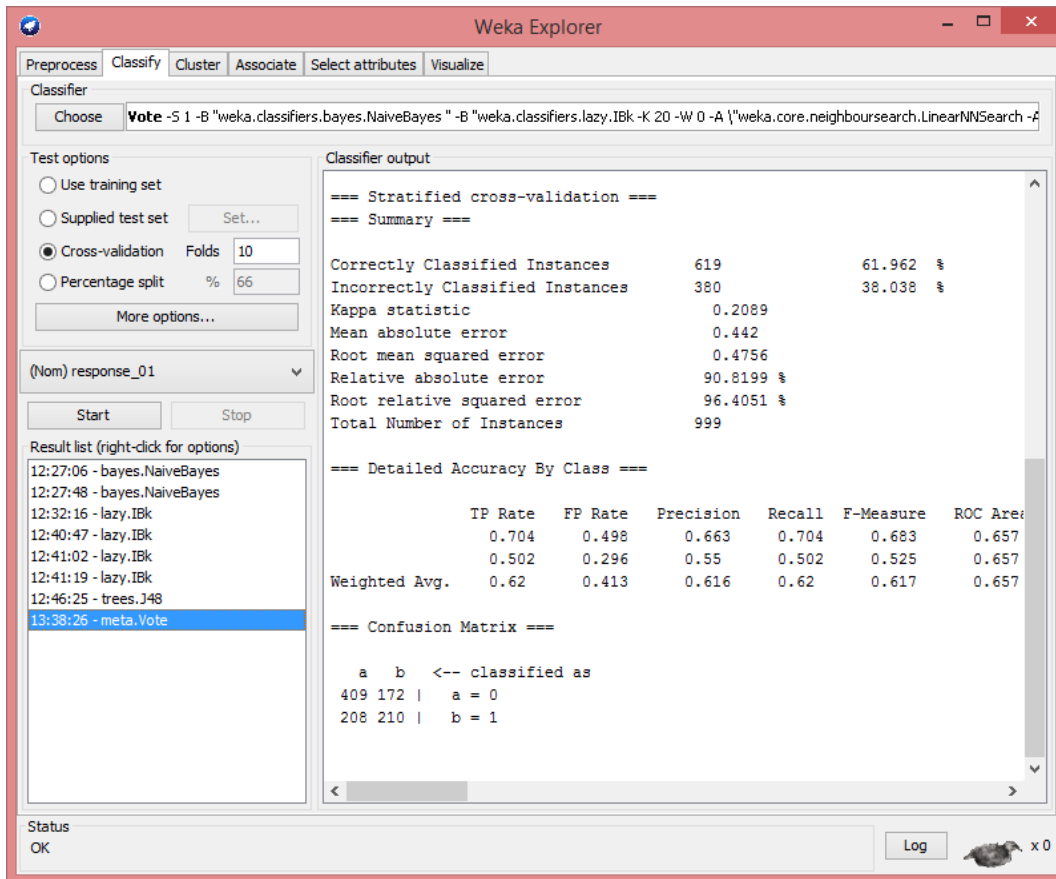
454	127	a = 0
259	159	b = 1

Status OK Log x 0



Ensemble (Metalearning) classifier.meta.Voting

21. You could combine multiple classifiers to perform an ensemble method.



The screenshot shows the Weka Explorer interface with the Classifier tab selected. The classifier chosen is `Vote -S 1 -B "weka.classifiers.bayes.NaiveBayes" -B "weka.classifiers.lazy.IBk -K 20 -W 0 -A \"weka.core.neighboursearch.LinearNNSearch -A`. The test options are set to Cross-validation with 10 folds. The classifier output is displayed in the right pane.

Classifier output

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

Correctly Classified Instances      619      61.962 %
Incorrectly Classified Instances    380      38.038 %
Kappa statistic                    0.2089
Mean absolute error                 0.442
Root mean squared error             0.4756
Relative absolute error             90.8199 %
Root relative squared error         96.4051 %
Total Number of Instances          999

=== Detailed Accuracy By Class ===

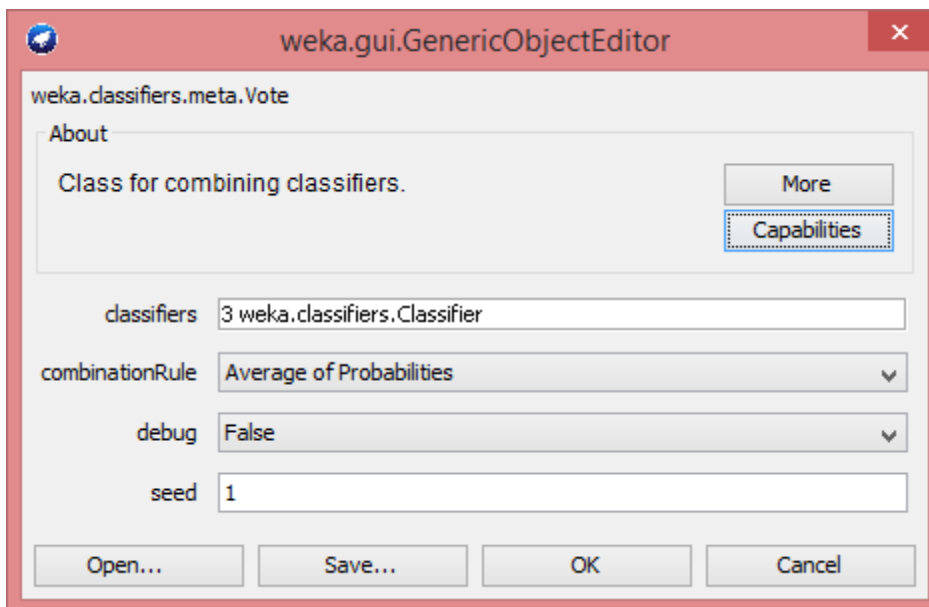
          TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area
          0.704    0.498    0.663     0.704    0.683     0.657
          0.502    0.296    0.55      0.502    0.525     0.657
Weighted Avg.   0.62     0.413    0.616     0.62     0.617     0.657

=== Confusion Matrix ===

  a   b   <-- classified as
409 172 |   a = 0
208 210 |   b = 1
```

The result list on the left shows the following entries:

- 12:27:06 - bayes.NaiveBayes
- 12:27:48 - bayes.NaiveBayes
- 12:32:16 - lazy.IBk
- 12:40:47 - lazy.IBk
- 12:41:02 - lazy.IBk
- 12:41:19 - lazy.IBk
- 12:46:25 - trees.J48
- 13:38:26 - meta.Vote



The screenshot shows the `weka.gui.GenericObjectEditor` window for the `weka.classifiers.meta.Vote` classifier. The window displays the following configuration:

- About:** Class for combining classifiers. (Buttons: More, Capabilities)
- classifiers:** 3 weka.classifiers.Classifier
- combinationRule:** Average of Probabilities
- debug:** False
- seed:** 1

Buttons at the bottom: Open..., Save..., OK, Cancel.