Machine Learning with Weka

Sujatha Das Gollapalli
Cornelia Caragea

August 19, 2014

Thanks to Eibe Frank for some of the slides

WEKA: the software

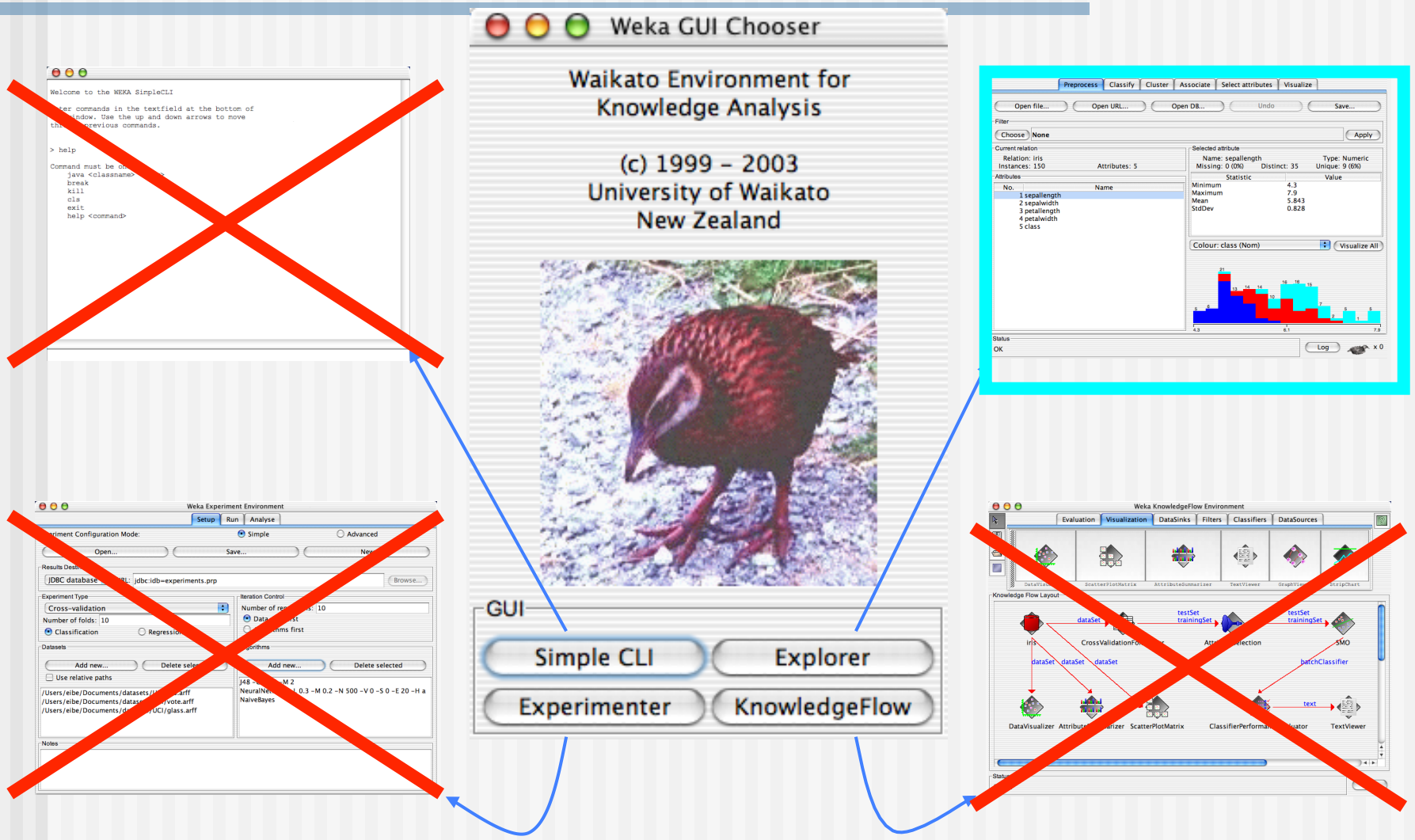
- Machine learning/data mining software written in Java (distributed under the GNU Public License)
- Used for research, education, and applications
- Main features:
  - Comprehensive set of data pre-processing tools, learning algorithms and evaluation methods
  - Graphical user interfaces (incl. data visualization)
  - Environment for comparing learning algorithms
- WEKA website:
  - http://www.cs.waikato.ac.nz/ml/weka/

WEKA: resources

- API Documentation, Tutorials, Source code.
- WEKA mailing list
- *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*
- Weka-related Projects:
  - Weka-Parallel - parallel processing for Weka
  - RWeka - linking R and Weka
  - YALE - Yet Another Learning Environment
  - Many others…

WEKA: launching

- java -jar weka.jar



Data Preparation and Loading

Data Preparation: WEKA only deals with “flat” files

```
@relation heart-disease-simplified

@attribute age numeric
@attribute sex { female, male}
@attribute chest_pain_type { typ_angina, asympt, non_anginal, atyp_angina}
@attribute cholesterol numeric
@attribute exercise_induced_angina { no, yes}
@attribute class { present, not_present}

@data
63,male,typ_angina,233,no,not_present
67,male,asympt,286,yes,present
67,male,asympt,229,yes,present
38,female,non_anginal,?,no,not_present
...
```

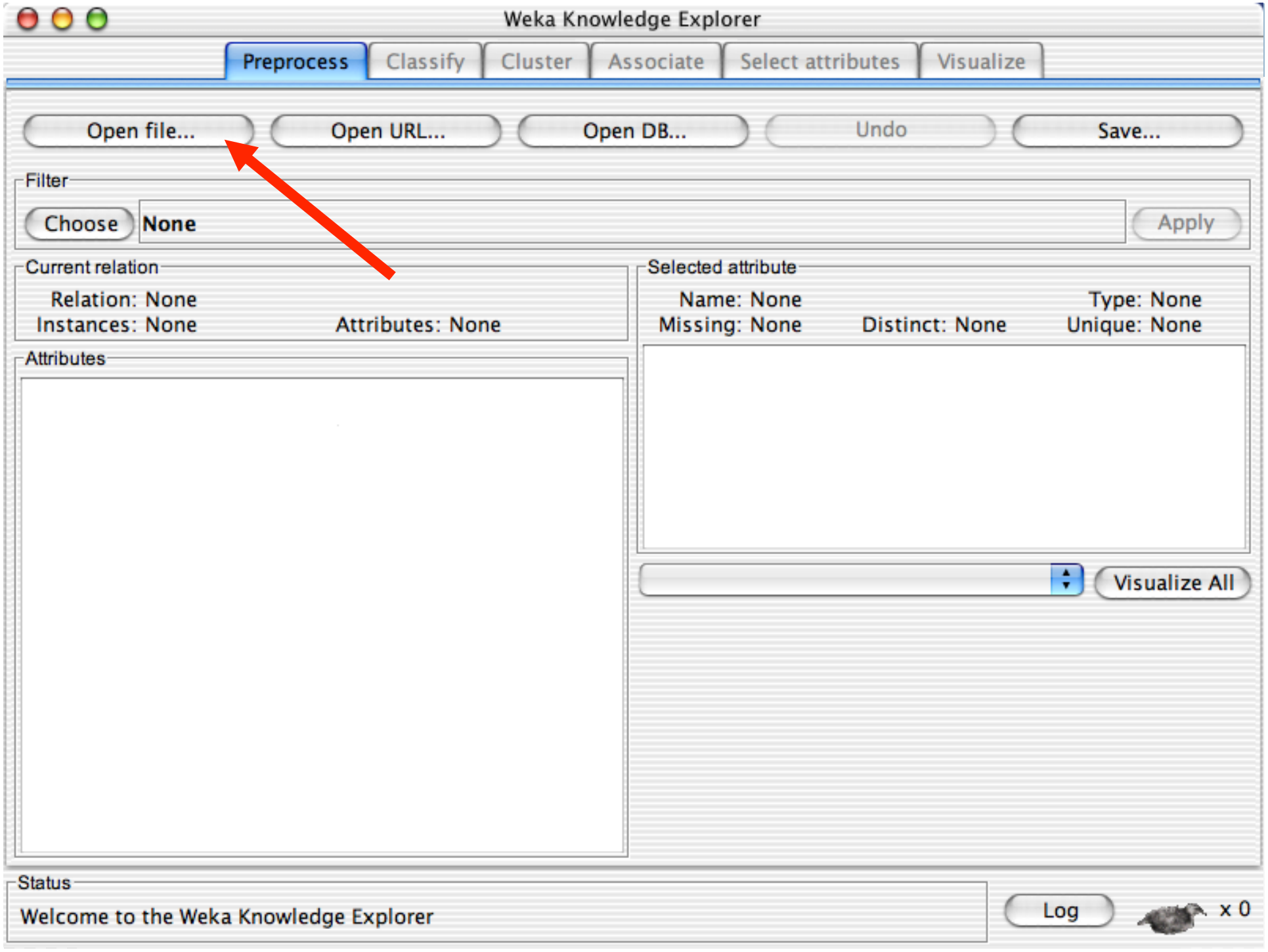
Flat file in ARFF format

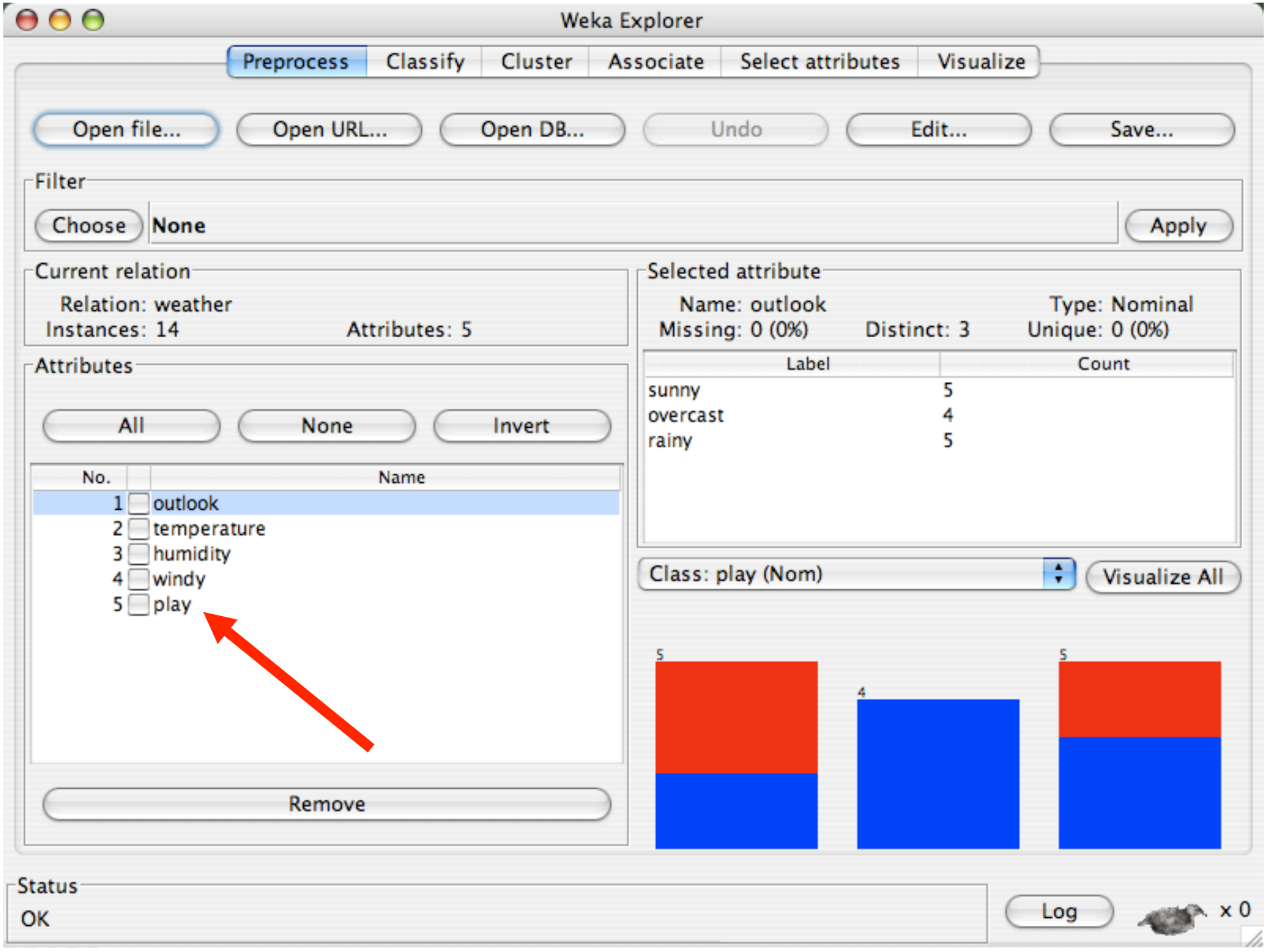
WEKA only deals with "flat" files

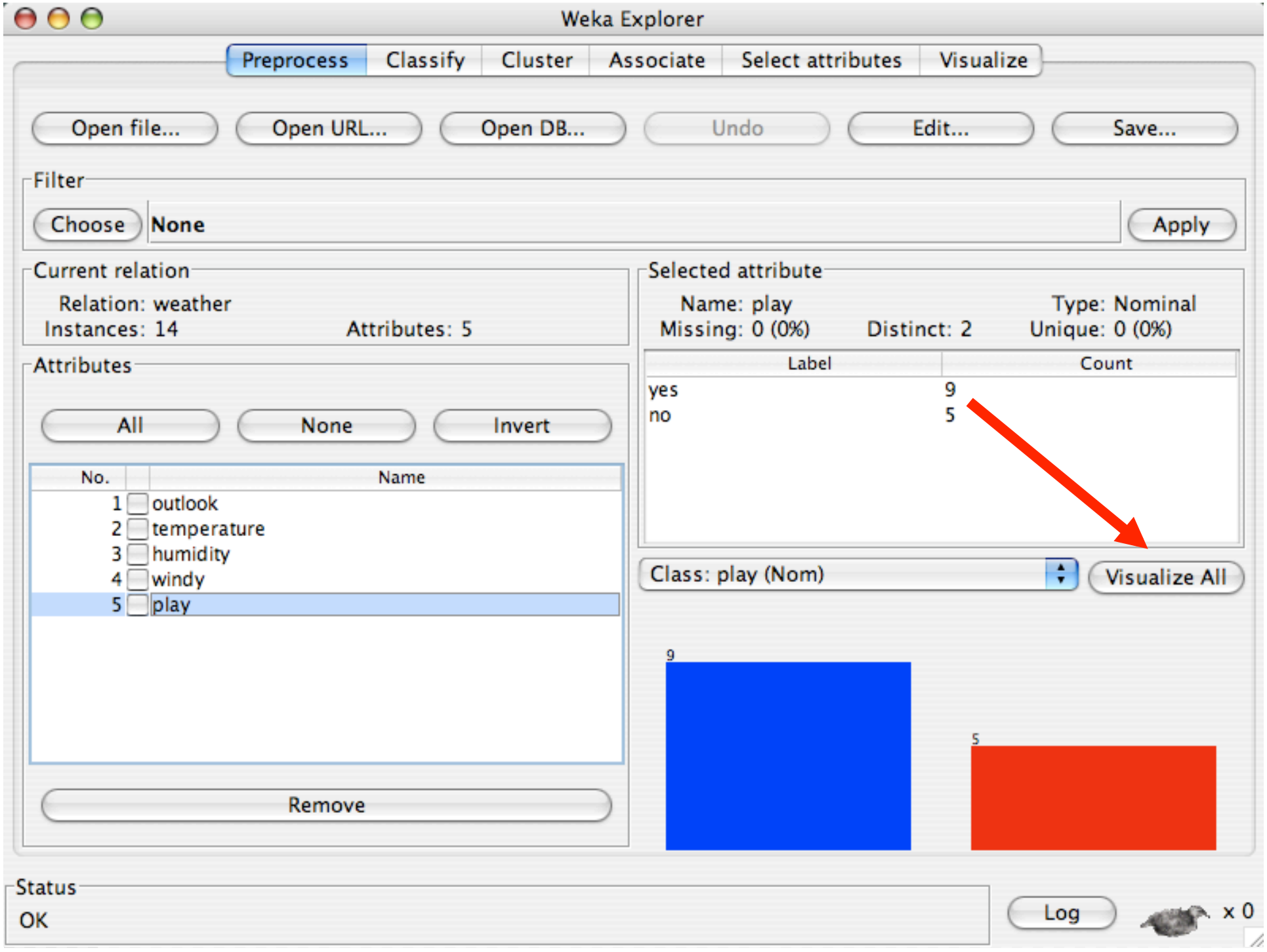

```
@relation heart-disease-simplified

@attribute age numeric
@attribute sex { female, male}
@attribute chest_pain_type { typ_angina, asympt, non_anginal, atyp_angina}
@attribute cholesterol numeric
@attribute exercise_induced_angina { no, yes}
@attribute class { present, not_present}

@data
63,male,typ_angina,233,no,not_present
67,male,asympt,286,yes,present
67,male,asympt,229,yes,present
38,female,non_anginal,?,no,not_present
...
```

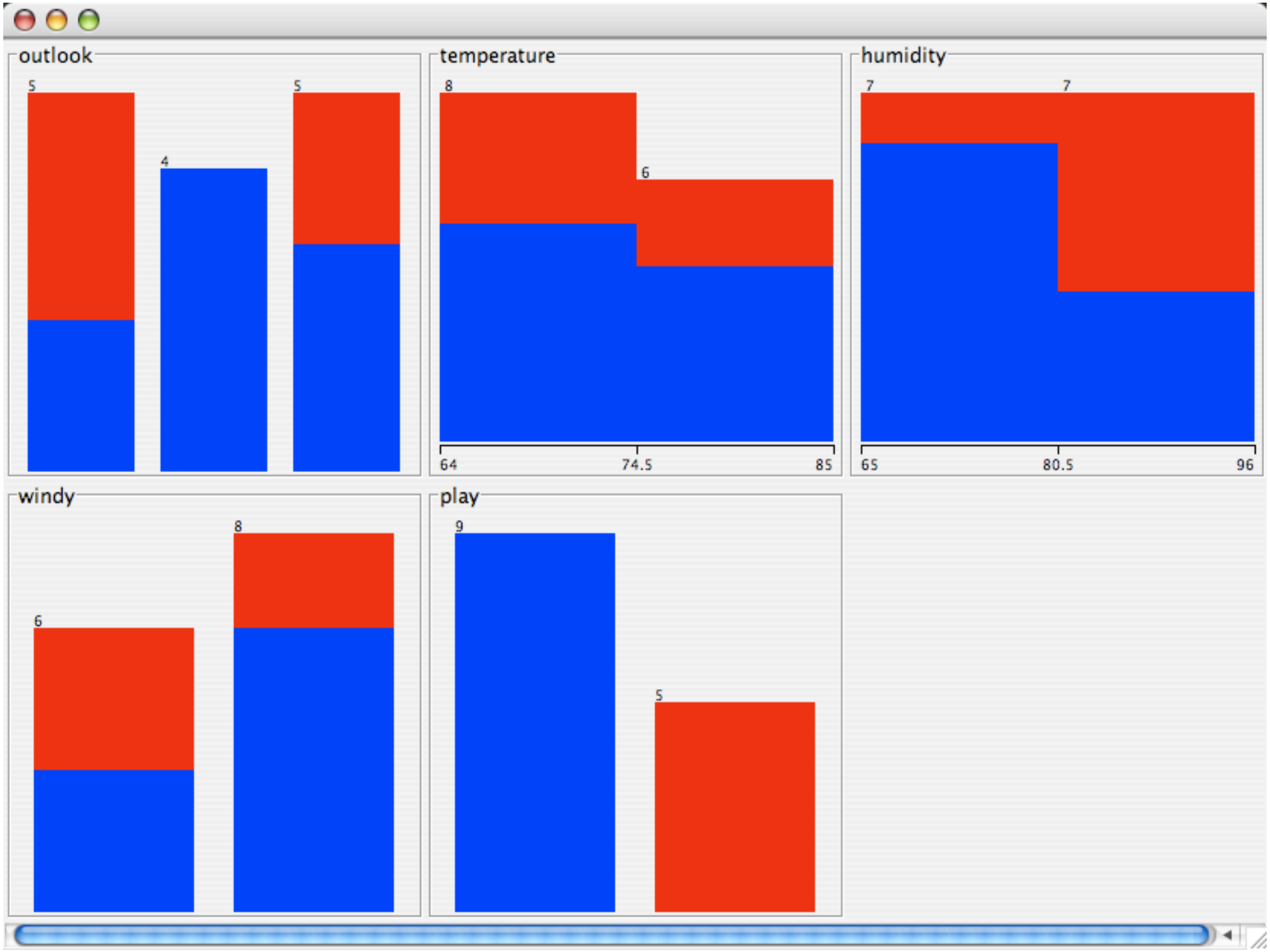
Explorer: pre-processing the data

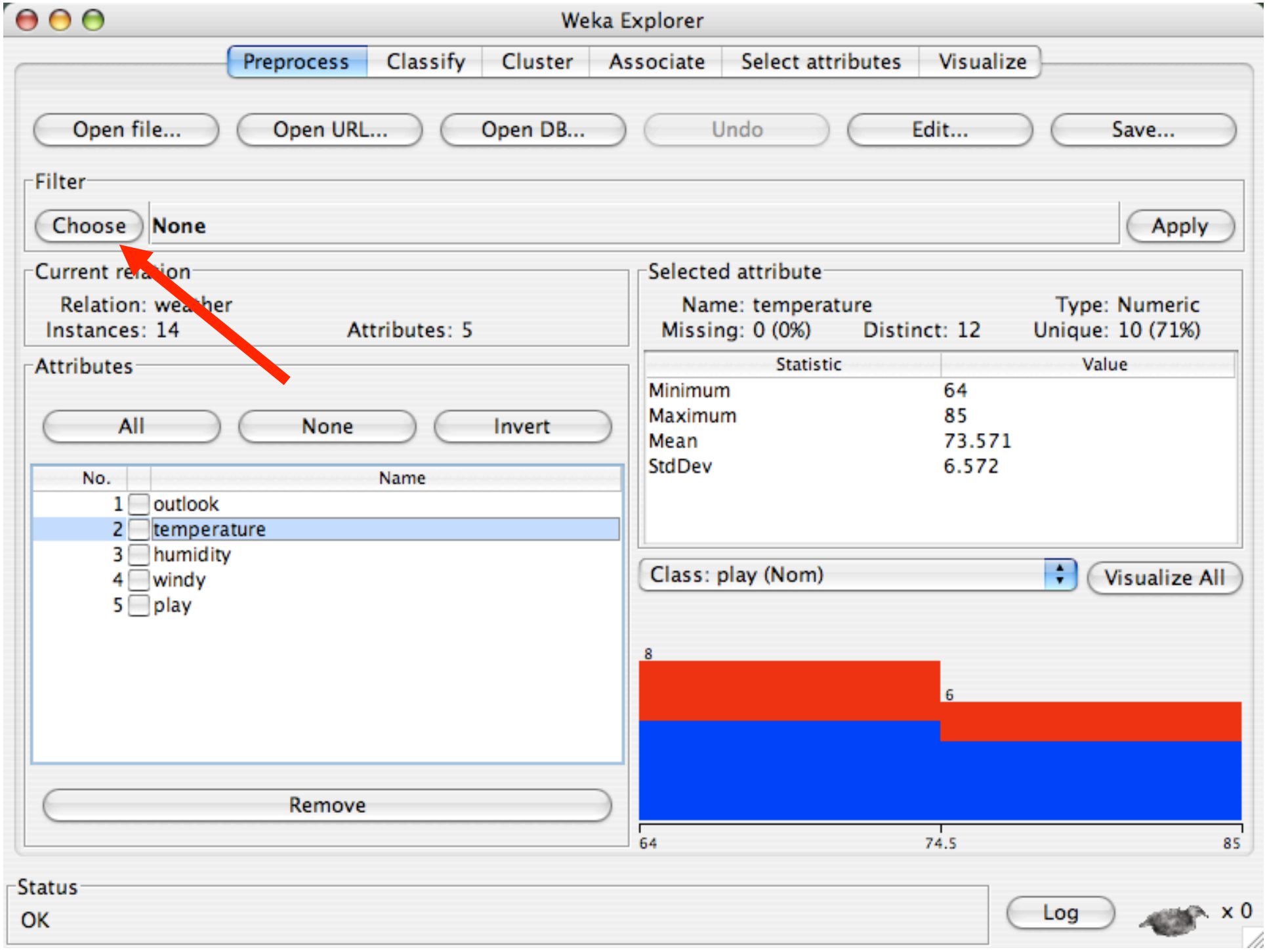
- Data can be imported from a file in various formats: ARFF, CSV, C4.5, binary
- Data can also be read from a URL or from an SQL database (using JDBC)
- Pre-processing tools in WEKA are called “filters”
- WEKA contains filters for:
  - Discretization, normalization, resampling, attribute selection, transforming and combining attributes, …

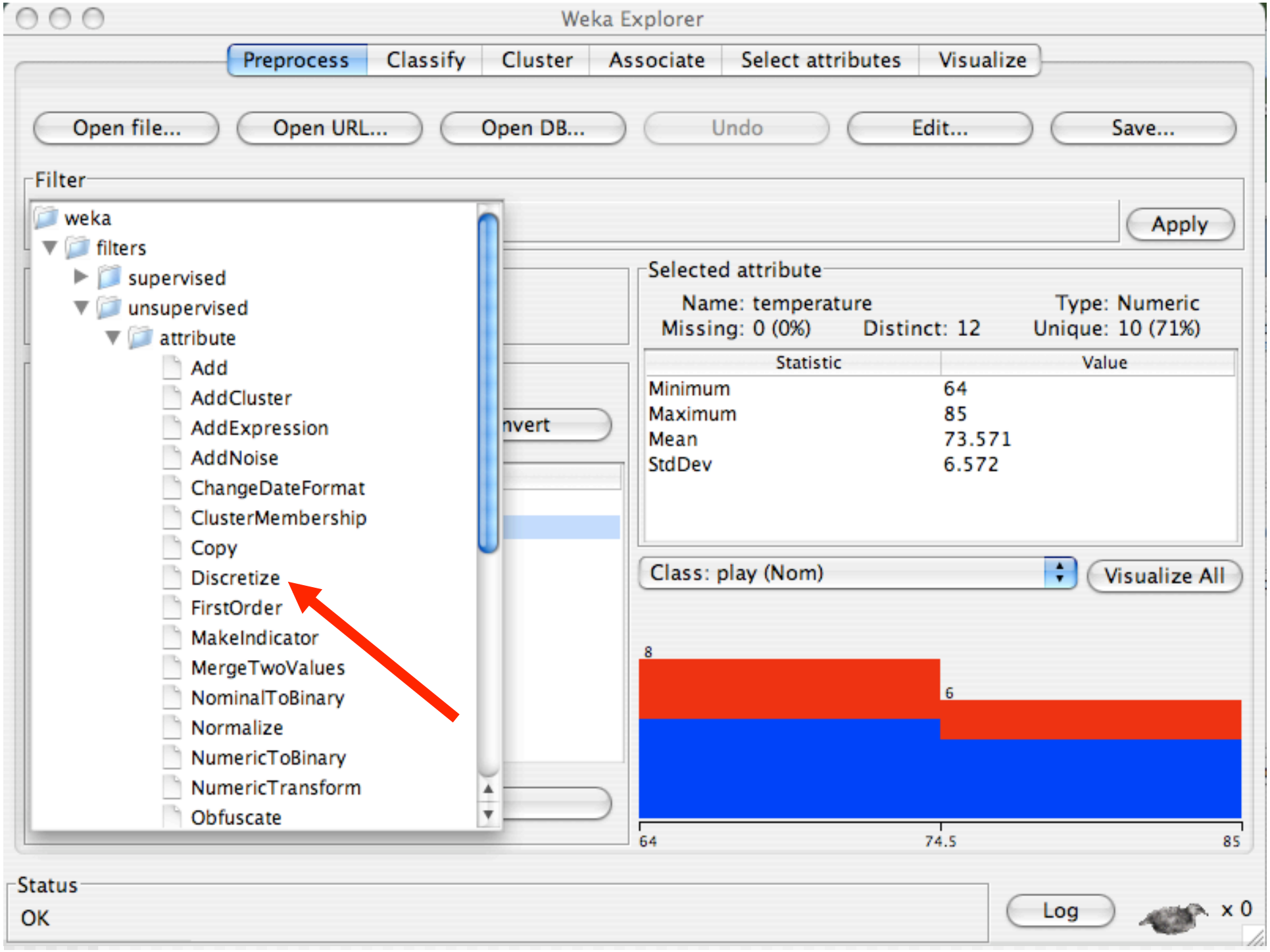

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
None
Apply
Current relation
Relation: None
Instances: None
Attributes: None
Selected attribute
Name: None
Type: None
Missing: None
Distinct: None
Unique: None
Attributes
Visualize All
Status
Welcome to the Weka Knowledge Explorer
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: weather
Instances: 14
Attributes: 5
Attributes
All
None
Invert
No.
Name
1 outlook
2 temperature
3 humidity
4 windy
5 play
Remove
Selected attribute
Name: outlook
Type: Nominal
Missing: 0 (0%)
Distinct: 3
Unique: 0 (0%)
Label
Count
sunny 5
overcast 4
rainy 5
Class: play (Nom)
Visualize All
5
4
5
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: weather
Instances: 14
Attributes: 5
Attributes
All
None
Invert
No.
Name
1 outlook
2 temperature
3 humidity
4 windy
5 play
Remove
Selected attribute
Name: play
Type: Nominal
Missing: 0 (0%)
Distinct: 2
Unique: 0 (0%)
Label
Count
yes
9
no
5
Class: play (Nom)
Visualize All
9
5
Status
OK
Log
x 0


outlook
5
4
5
temperature
8
6
64
74.5
85
humidity
7
7
65
80.5
96
windy
6
8
play
9
5

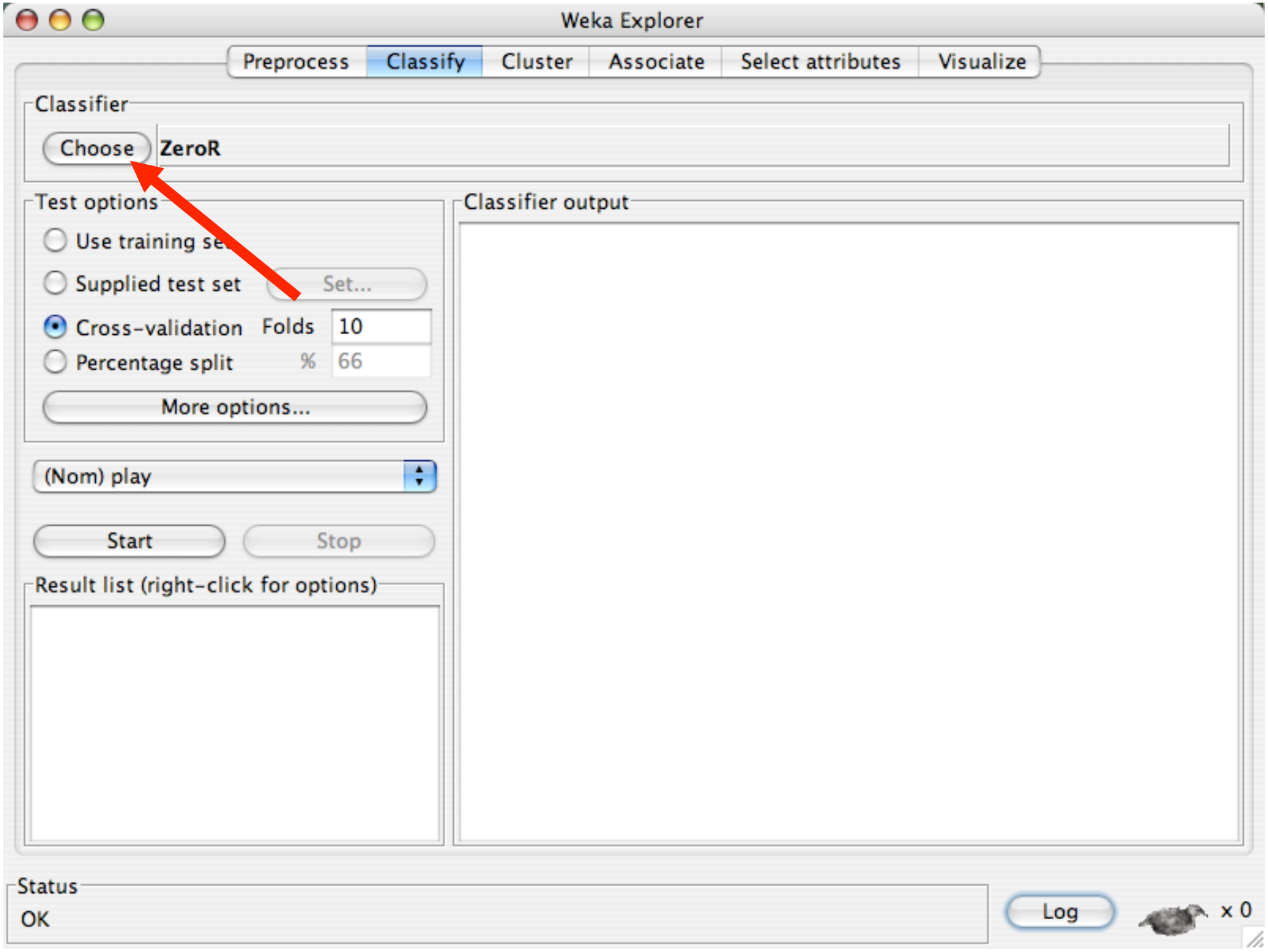

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: weather
Instances: 14
Attributes: 5
Attributes
All
None
Invert
No.
Name
1 outlook
2 temperature
3 humidity
4 windy
5 play
Remove
Selected attribute
Name: temperature
Type: Numeric
Missing: 0 (0%)
Distinct: 12
Unique: 10 (71%)
Statistic
Value
Minimum 64
Maximum 85
Mean 73.571
StdDev 6.572
Class: play (Nom)
Visualize All
8
6
64
74.5
85
Status
OK
Log
x 0

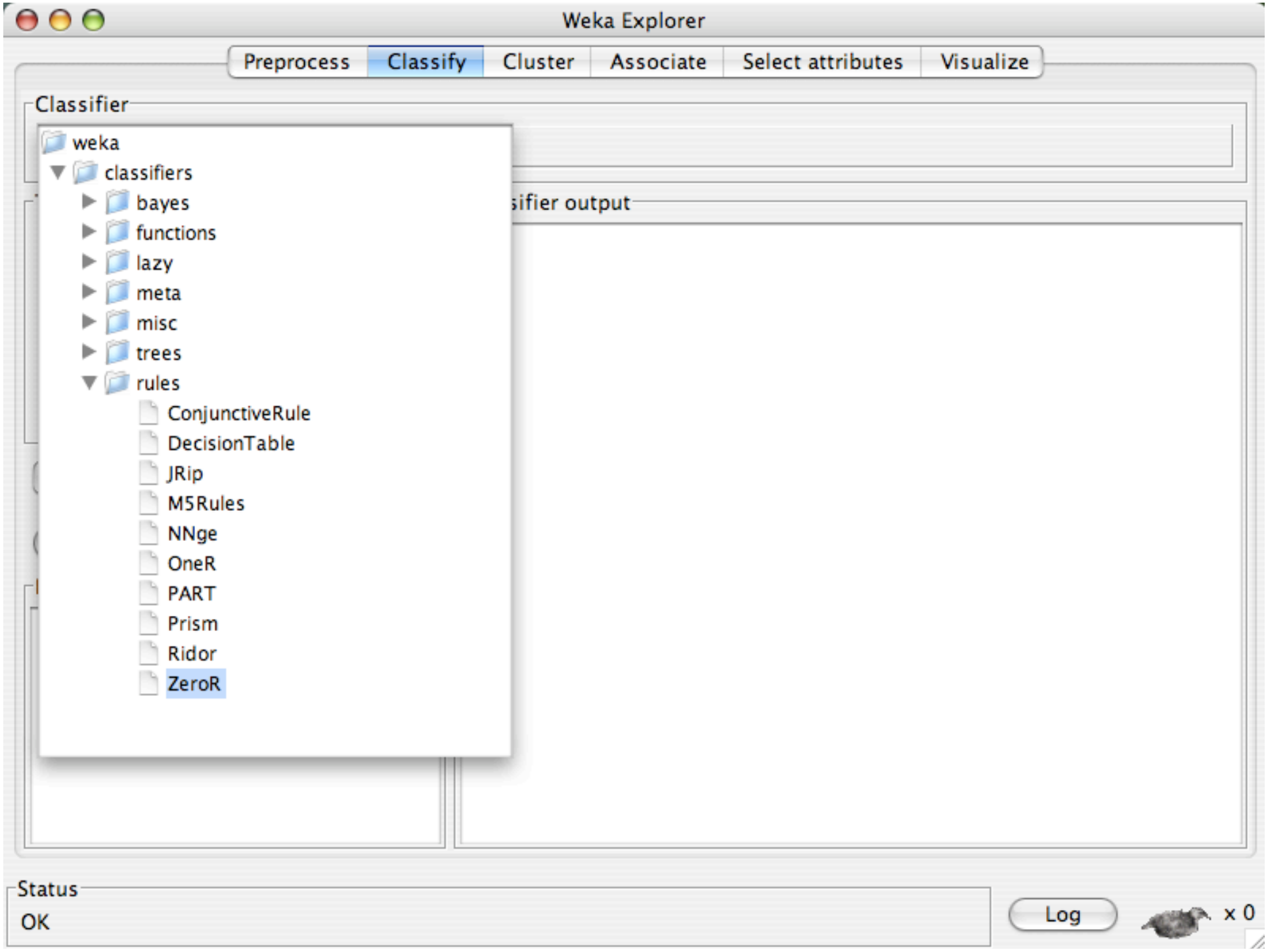

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Edit...
Save...
Filter
weka
filters
supervised
unsupervised
attribute
Add
AddCluster
AddExpression
AddNoise
ChangeDateFormat
ClusterMembership
Copy
Discretize
FirstOrder
MakeIndicator
MergeTwoValues
NominalToBinary
Normalize
NumericToBinary
NumericTransform
Obfuscate
Apply
Selected attribute
Name: temperature
Type: Numeric
Missing: 0 (0%)
Distinct: 12
Unique: 10 (71%)
Statistic
Value
Minimum
64
Maximum
85
Mean
73.571
StdDev
6.572
Class: play (Nom)
Visualize All
8
6
64
74.5
85
Status
OK
Log
x 0

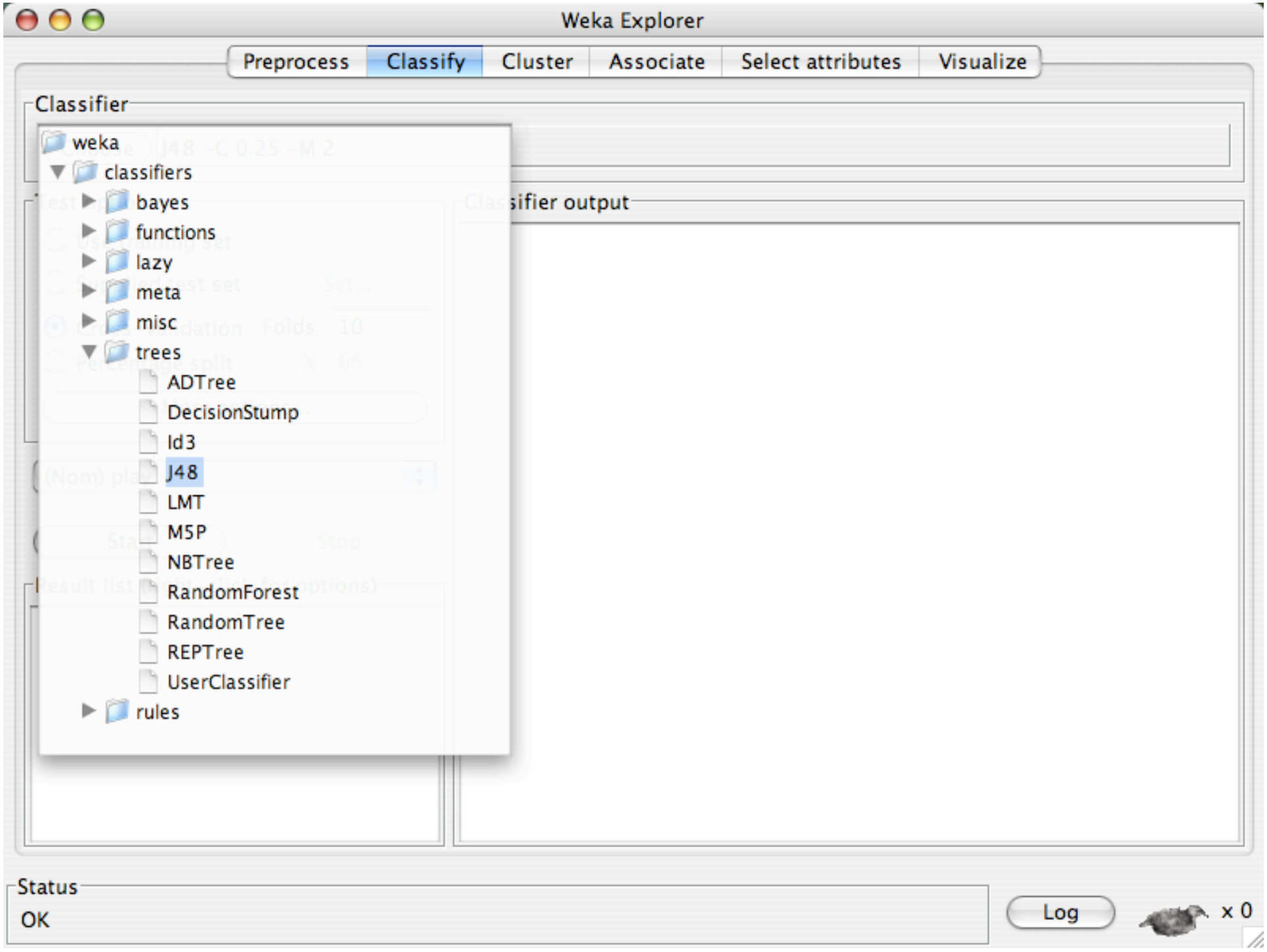
Building Classifiers

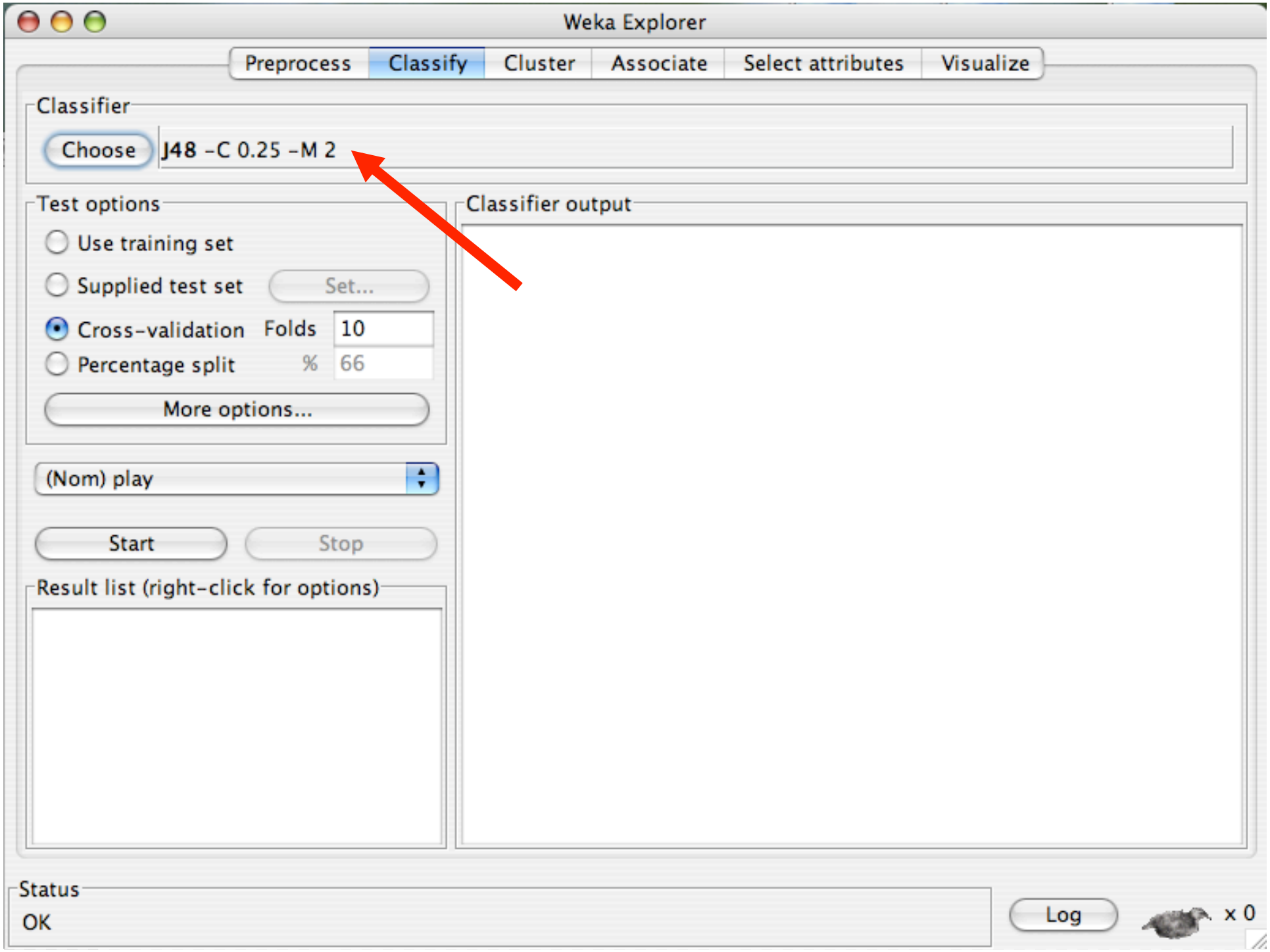
Explorer: building "classifiers"

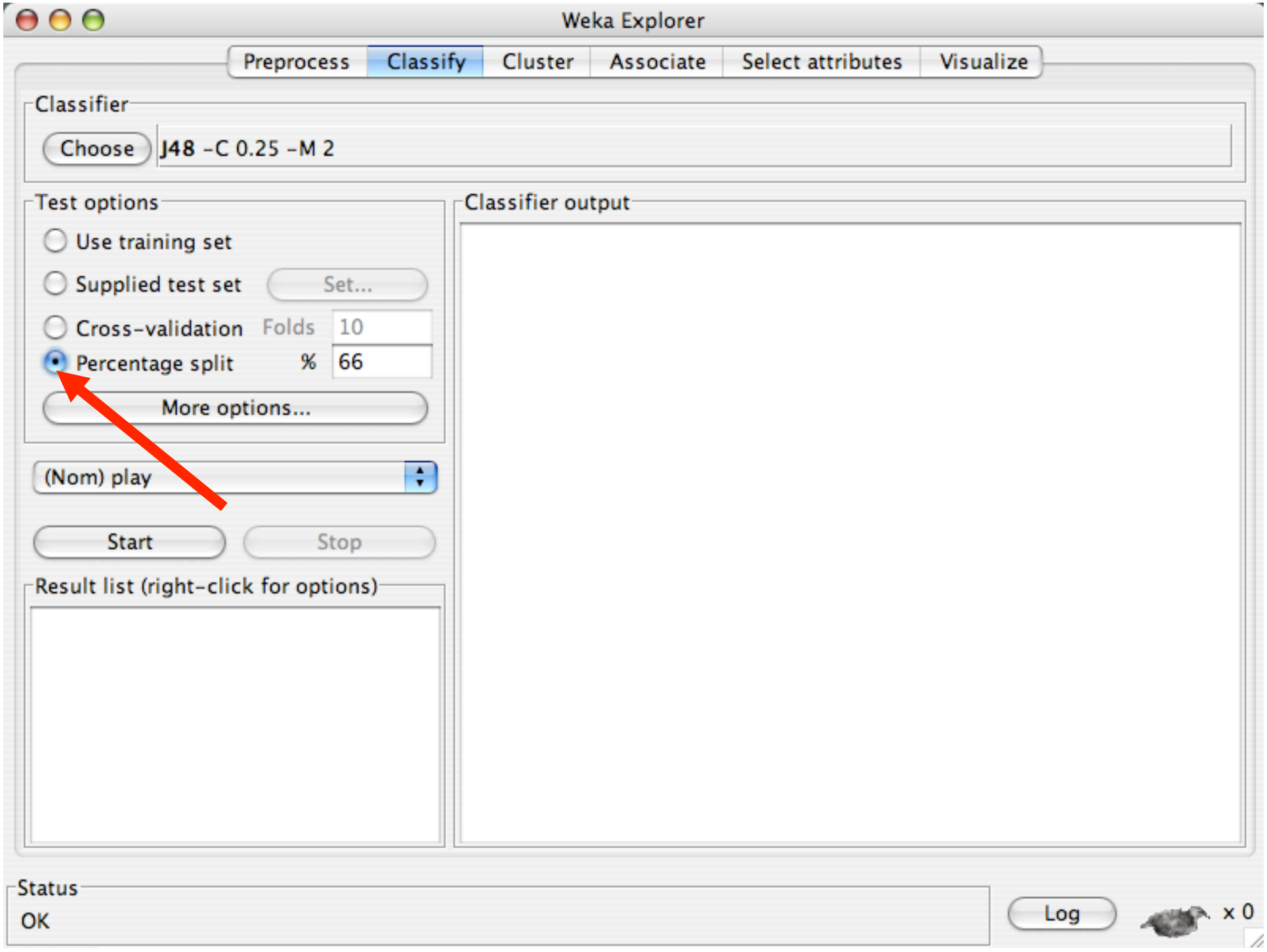
- Classifiers in WEKA are models for predicting nominal or numeric quantities
- Implemented learning schemes include:
  - Decision trees and lists, instance-based classifiers, support vector machines, multi-layer perceptrons, logistic regression, Bayes' nets, …
- "Meta"-classifiers include:
  - Bagging, boosting, stacking, etc.

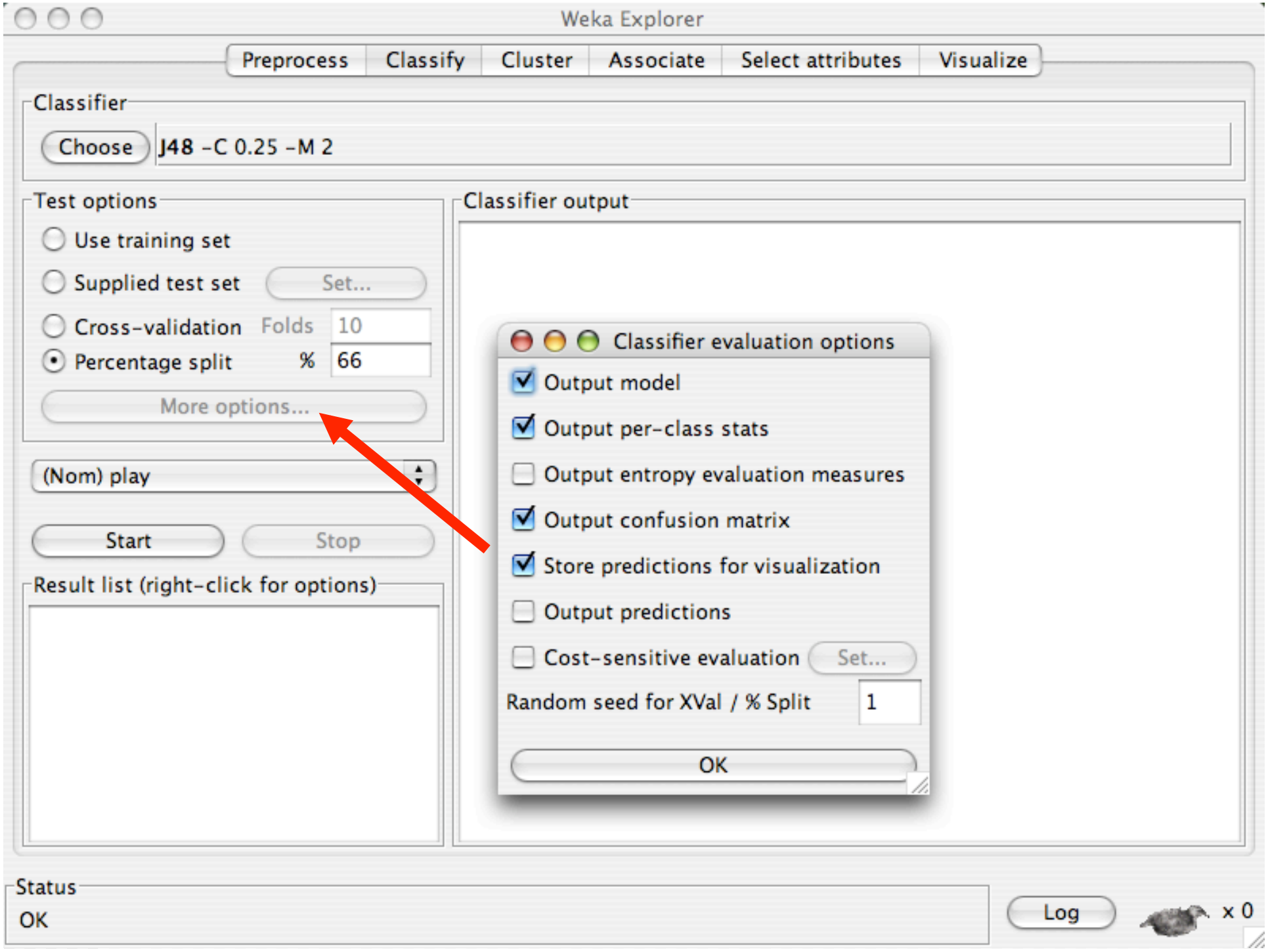

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
ZeroR
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
Classifier output
(Nom) play
Start
Stop
Result list (right-click for options)
Status
OK
Log
x 0

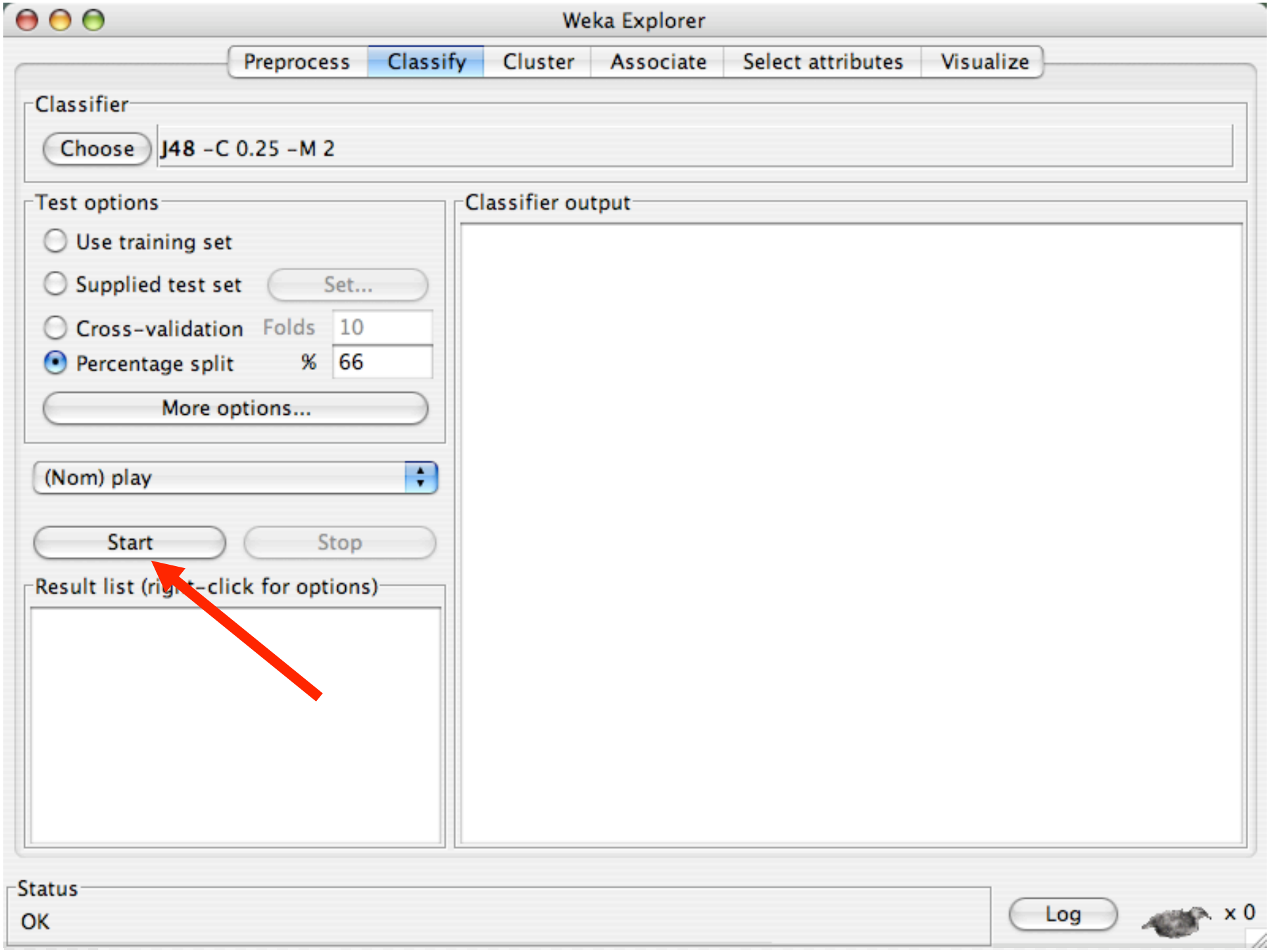

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
weka
classifiers
bayes
functions
lazy
meta
misc
trees
rules
ConjunctiveRule
DecisionTable
JRip
M5Rules
NNge
OneR
PART
Prism
Ridor
ZeroR
sifier output
Status
OK
Log
x 0

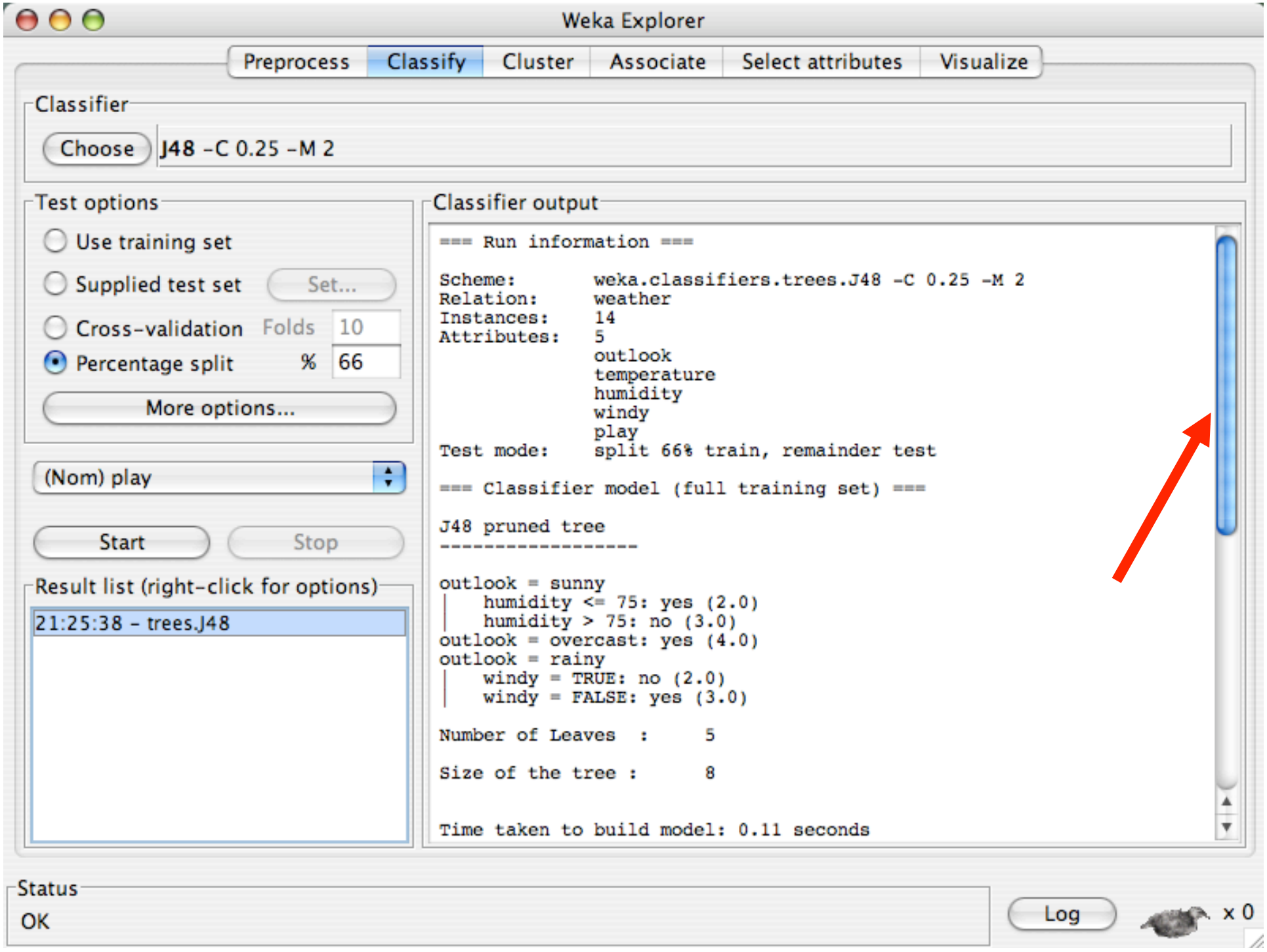

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
weka
classifiers
bayes
functions
lazy
meta
misc
trees
ADTree
DecisionStump
Id3
J48
LMT
M5P
NBTree
RandomForest
RandomTree
REPTree
UserClassifier
rules
Classifier output
Status
OK
Log
x 0

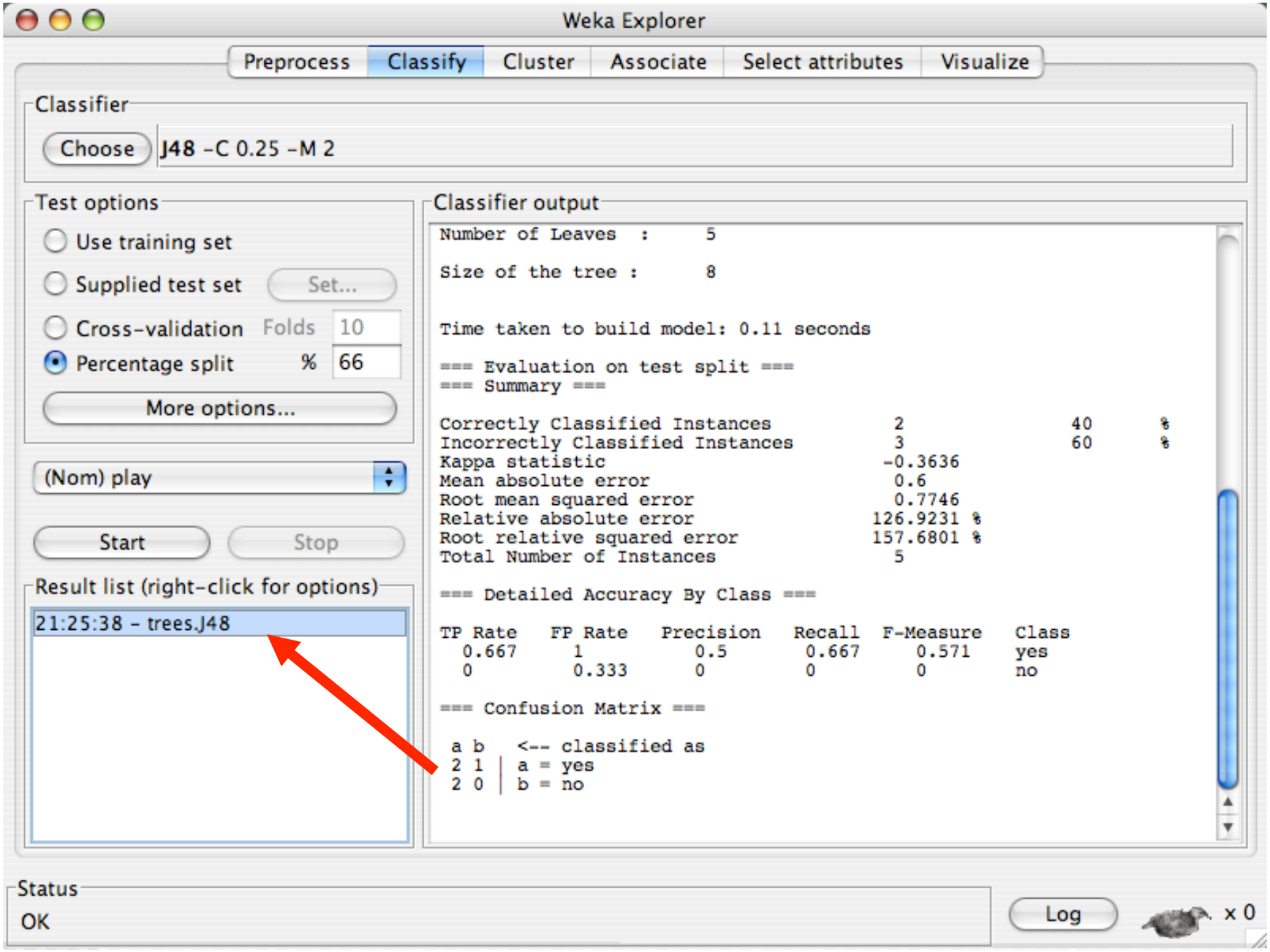

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
Classifier output
(Nom) play
Start
Stop
Result list (right-click for options)
Status
OK
Log
x 0

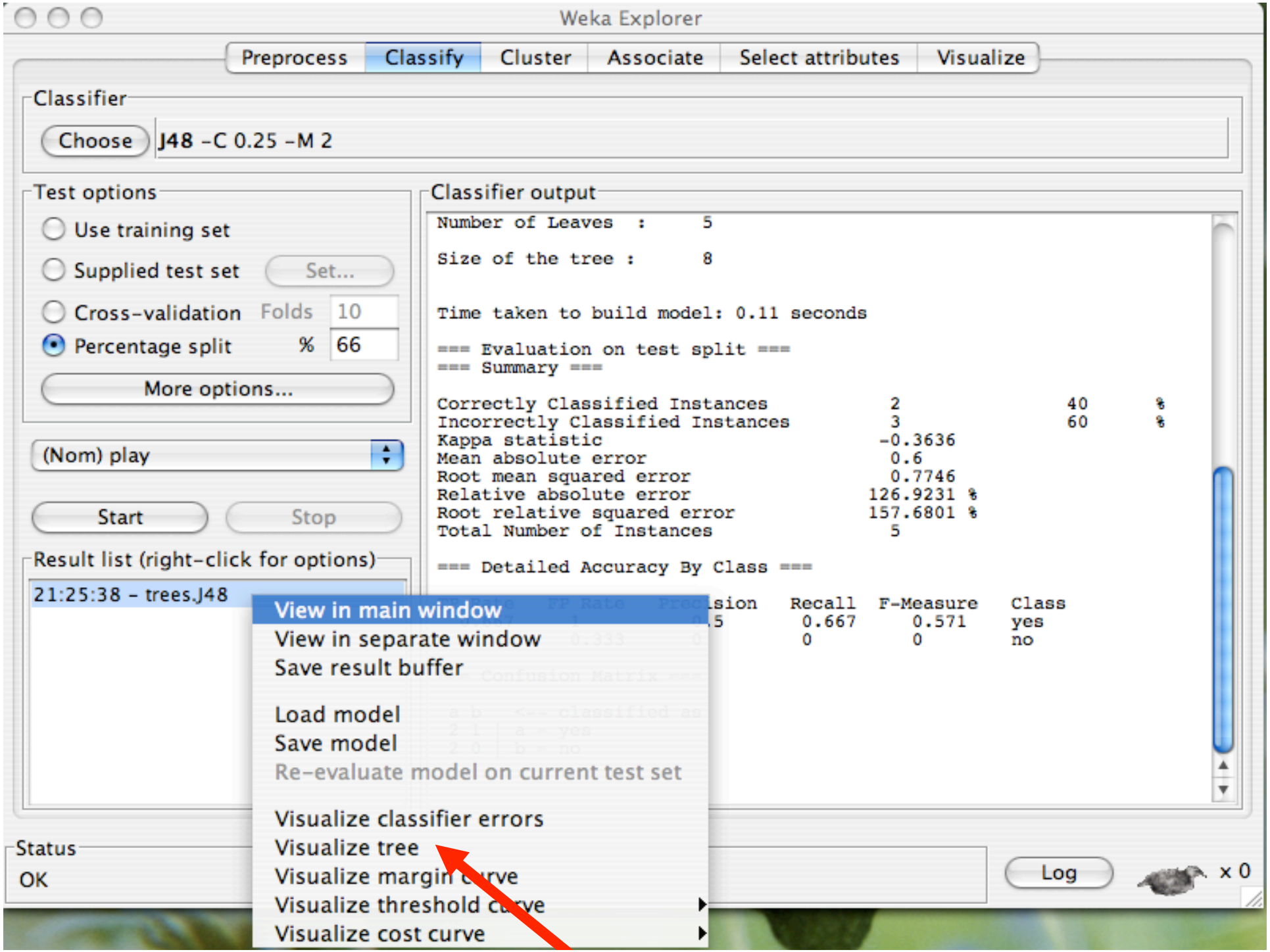

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
Classifier output
(Nom) play
Start
Stop
Result list (right-click for options)
Status
OK
Log
x 0

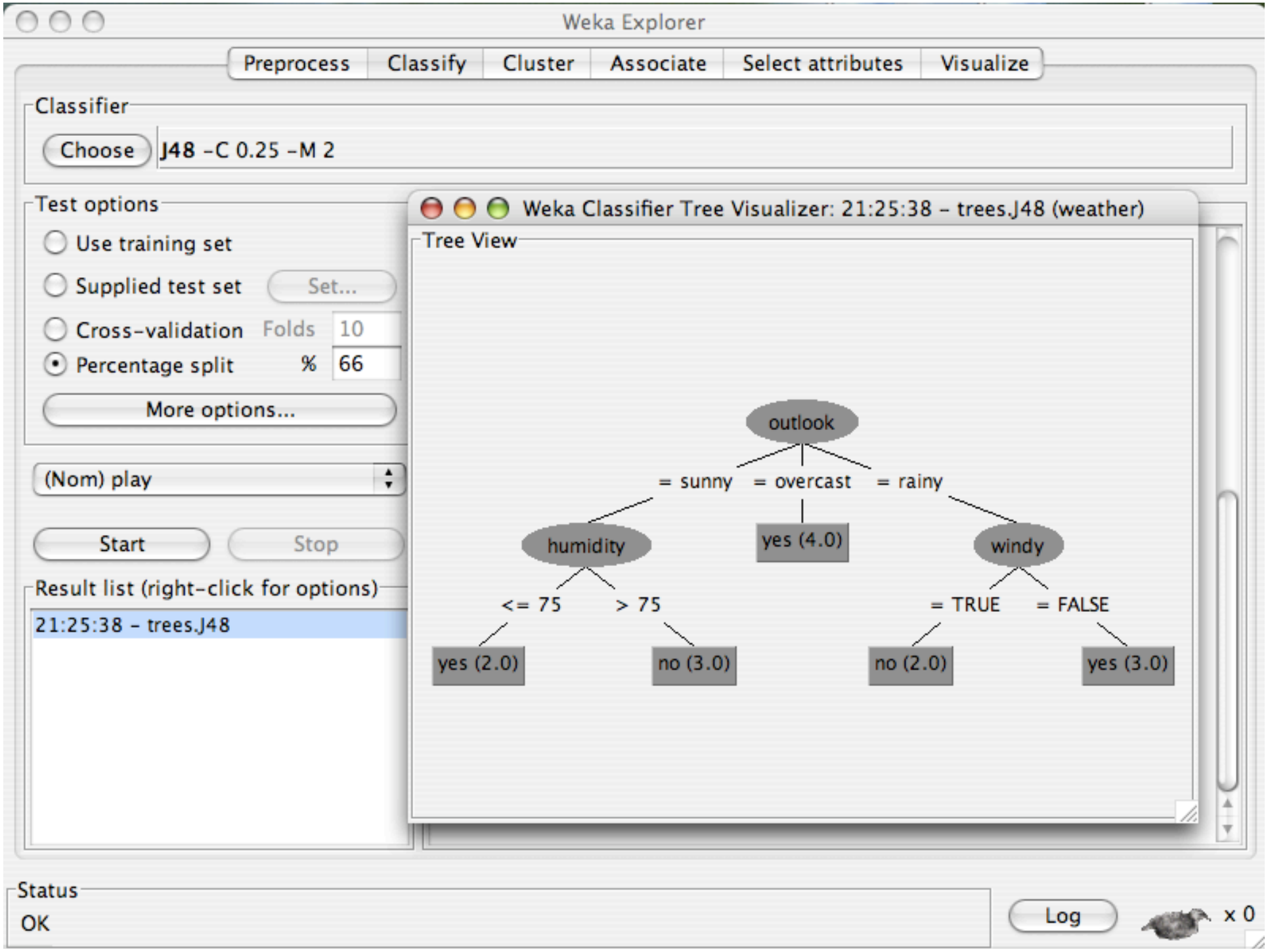

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
Classifier output
(Nom) play
Start
Stop
Result list (right-click for options)
Classifier evaluation options
Output model
Output per-class stats
Output entropy evaluation measures
Output confusion matrix
Store predictions for visualization
Output predictions
Cost-sensitive evaluation
Set...
Random seed for XVal / % Split
1
OK
Status
OK
Log
x 0

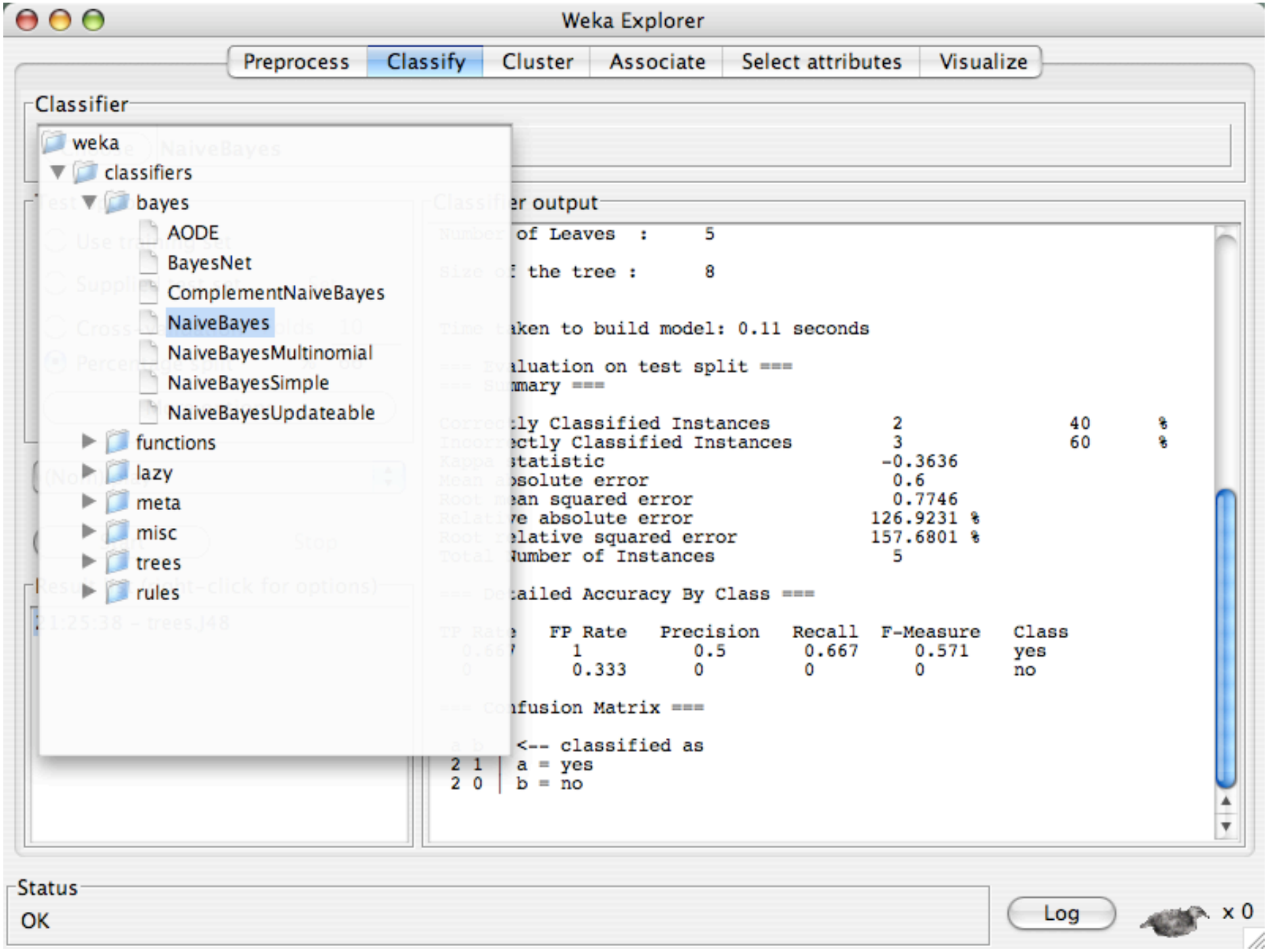

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
Classifier output
(Nom) play
Start
Stop
Result list (right-click for options)
Status
OK
Log
x 0

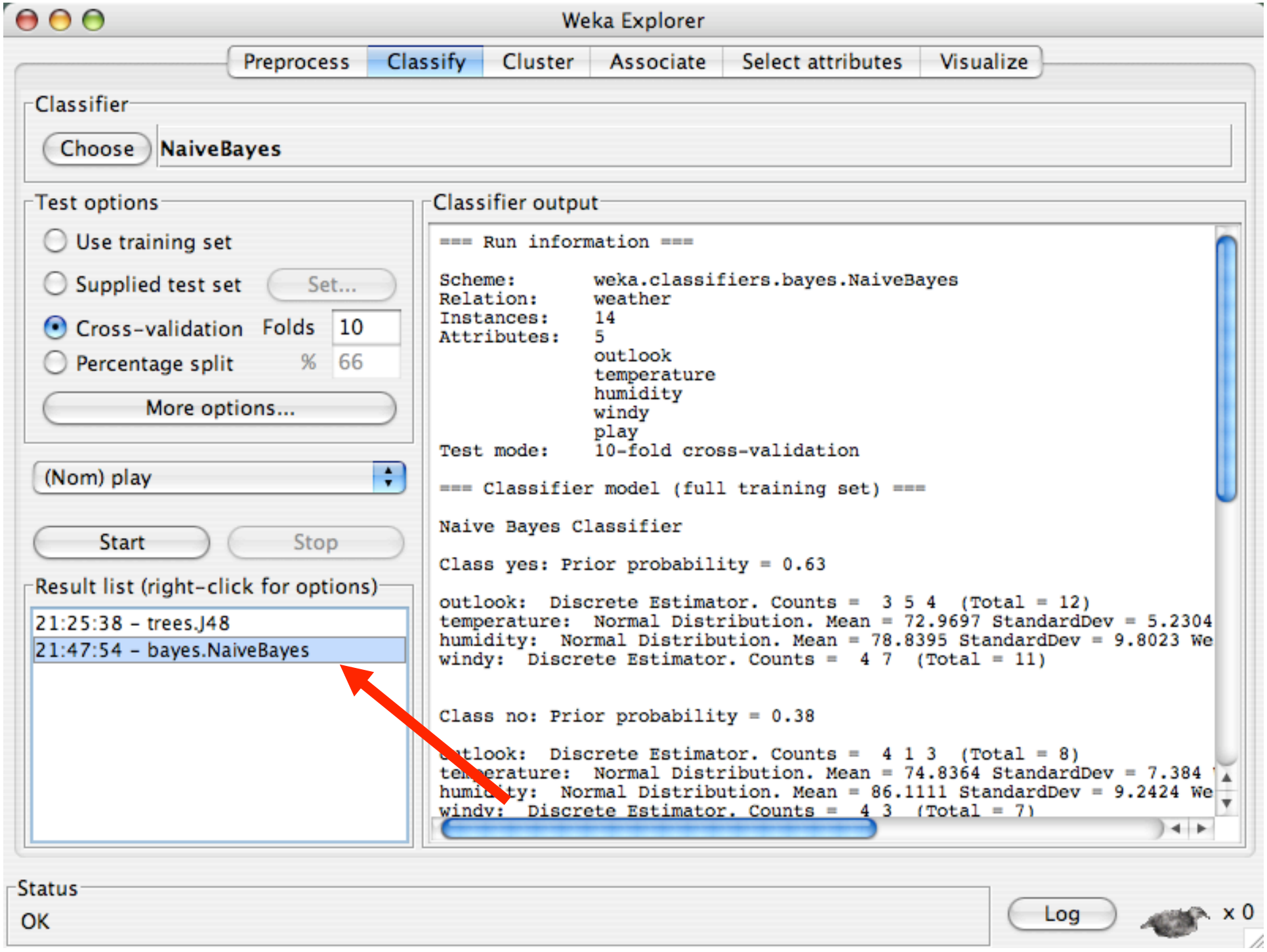

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) play
Start
Stop
Result list (right-click for options)
21:25:38 - trees.J48
Classifier output
=== Run information ===
Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: weather
Instances: 14
Attributes: 5
outlook
temperature
humidity
windy
play
Test mode: split 66% train, remainder test
=== Classifier model (full training set) ===
J48 pruned tree
outlook = sunny
| humidity <= 75: yes (2.0)
| humidity > 75: no (3.0)
outlook = overcast: yes (4.0)
outlook = rainy
| windy = TRUE: no (2.0)
| windy = FALSE: yes (3.0)
Number of Leaves : 5
Size of the tree : 8
Time taken to build model: 0.11 seconds
Status
OK
Log
x 0

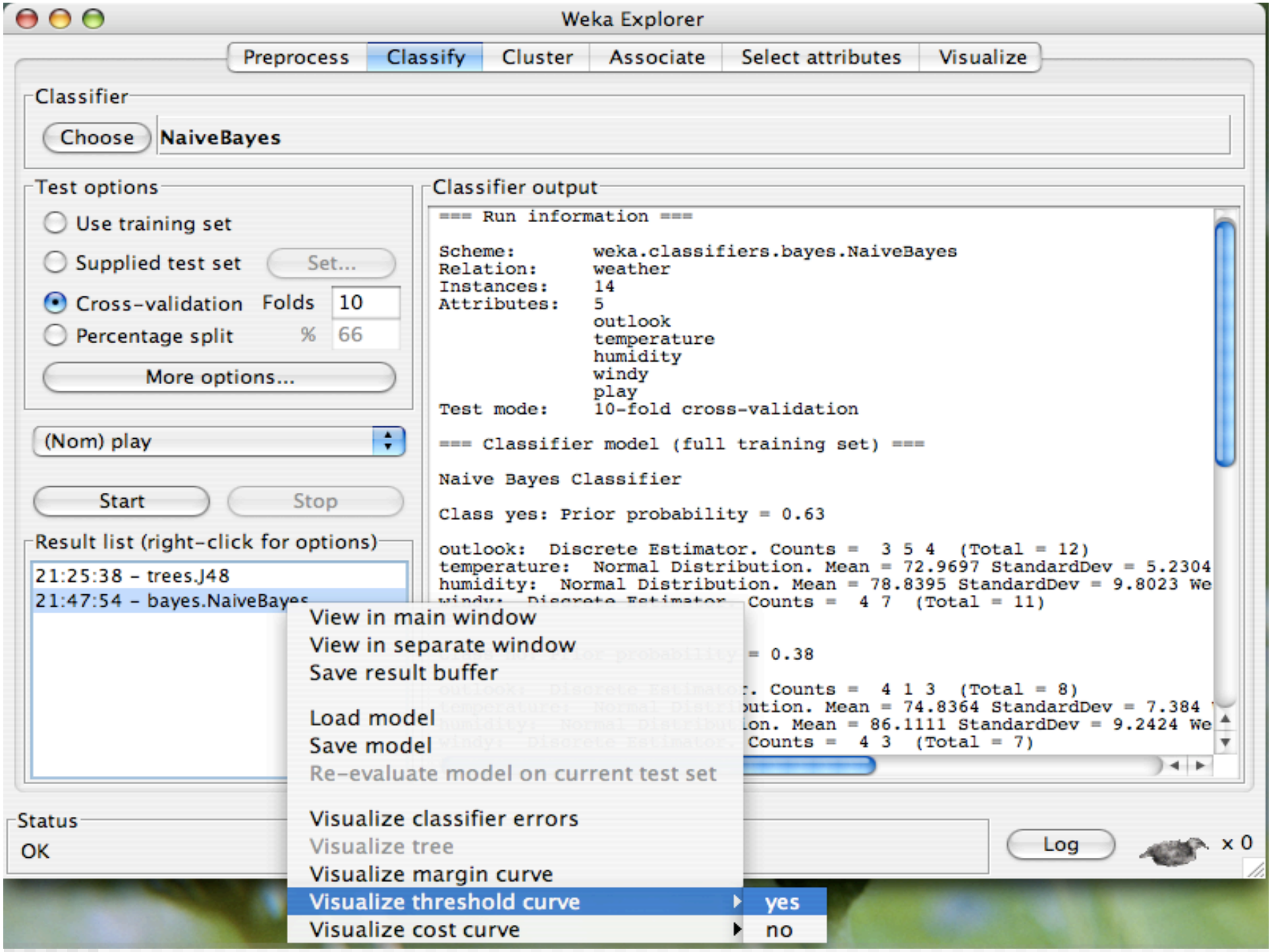

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) play
Start
Stop
Result list (right-click for options)
21:25:38 - trees.J48
Classifier output
Number of Leaves : 5
Size of the tree : 8
Time taken to build model: 0.11 seconds
=== Evaluation on test split ===
=== Summary ===
Correctly Classified Instances 2 40 %
Incorrectly Classified Instances 3 60 %
Kappa statistic -0.3636
Mean absolute error 0.6
Root mean squared error 0.7746
Relative absolute error 126.9231 %
Root relative squared error 157.6801 %
Total Number of Instances 5
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure Class
0.667 1 0.5 0.667 0.571 yes
0 0.333 0 0 0 no
=== Confusion Matrix ===
a b <-- classified as
2 1 | a = yes
2 0 | b = no
Status
OK
Log
x 0

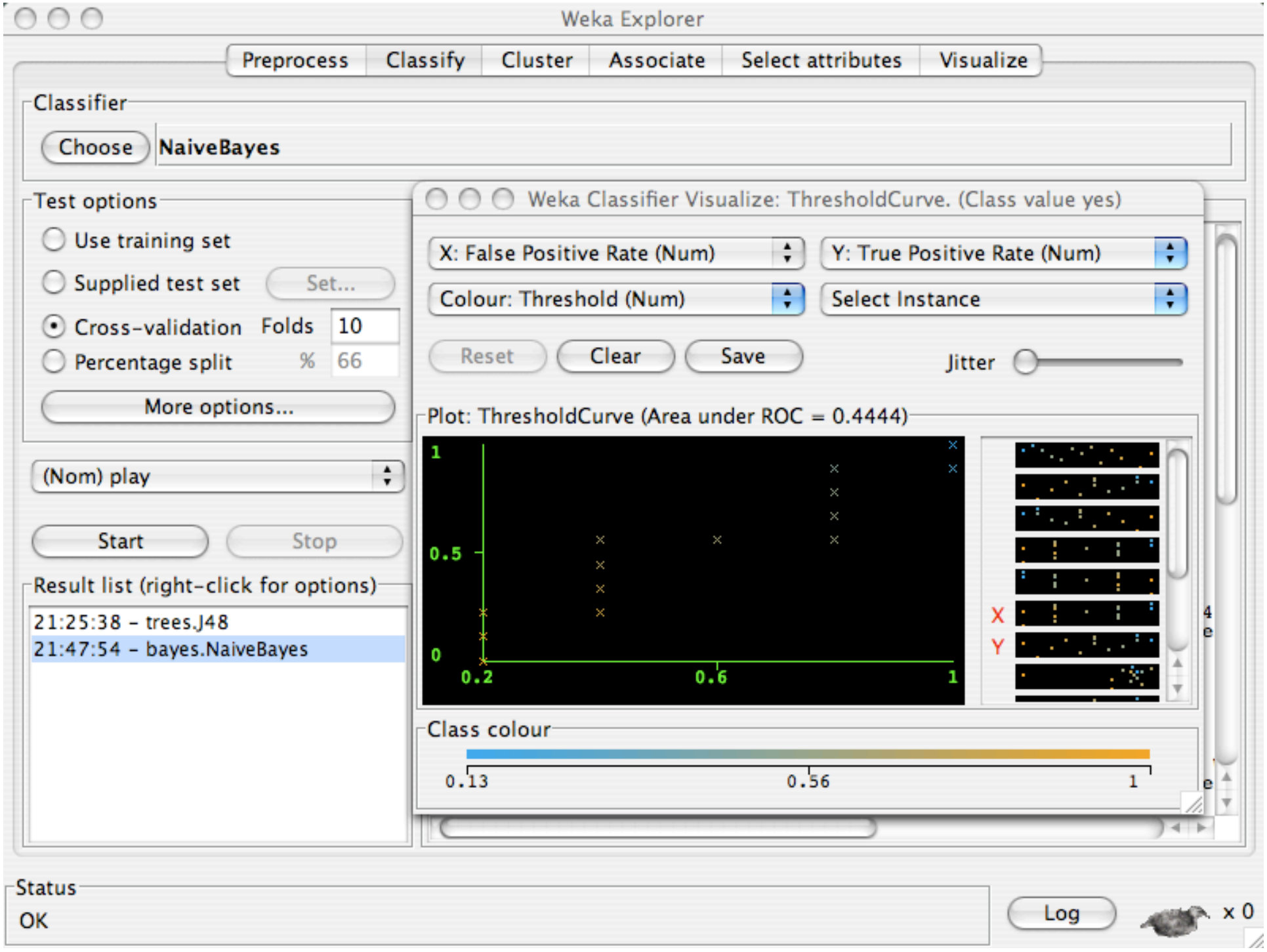

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) play
Start
Stop
Result list (right-click for options)
21:25:38 - trees.J48
Classifier output
Number of Leaves : 5
Size of the tree : 8
Time taken to build model: 0.11 seconds
=== Evaluation on test split ===
=== Summary ===
Correctly Classified Instances 2 40 %
Incorrectly Classified Instances 3 60 %
Kappa statistic -0.3636
Mean absolute error 0.6
Root mean squared error 0.7746
Relative absolute error 126.9231 %
Root relative squared error 157.6801 %
Total Number of Instances 5
=== Detailed Accuracy By Class ===
Recall F-Measure Class
0.667 0.571 yes
0 0 no
View in main window
View in separate window
Save result buffer
Load model
Save model
Re-evaluate model on current test set
Visualize classifier errors
Visualize tree
Visualize margin curve
Visualize threshold curve
Visualize cost curve
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) play
Start
Stop
Result list (right-click for options)
21:25:38 - trees.J48
Weka Classifier Tree Visualizer: 21:25:38 - trees.J48 (weather)
Tree View
outlook
= sunny
= overcast
= rainy
humidity
yes (4.0)
windy
<= 75
> 75
= TRUE
= FALSE
yes (2.0)
no (3.0)
no (2.0)
yes (3.0)
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
weka
classifiers
bayes
AODE
BayesNet
ComplementNaiveBayes
NaiveBayes
NaiveBayesMultinomial
NaiveBayesSimple
NaiveBayesUpdateable
functions
lazy
meta
misc
trees
rules
er output
of Leaves : 5
the tree : 8
aken to build model: 0.11 seconds
aluation on test split ===
mmary ===
ly Classified Instances 2 40 %
ctly Classified Instances 3 60 %
statistic -0.3636
solute error 0.6
an squared error 0.7746
e absolute error 126.9231 %
elative squared error 157.6801 %
Number of Instances 5
ailed Accuracy By Class ===
FP Rate Precision Recall F-Measure Class
1 0.5 0.667 0.571 yes
0.333 0 0 0 no
fusion Matrix ===
<-- classified as
a = yes
b = no
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) play
Start
Stop
Result list (right-click for options)
21:25:38 - trees.J48
21:47:54 - bayes.NaiveBayes
Classifier output
=== Run information ===
Scheme: weka.classifiers.bayes.NaiveBayes
Relation: weather
Instances: 14
Attributes: 5
outlook
temperature
humidity
windy
play
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
Naive Bayes Classifier
Class yes: Prior probability = 0.63
outlook: Discrete Estimator. Counts = 3 5 4 (Total = 12)
temperature: Normal Distribution. Mean = 72.9697 StandardDev = 5.2304
humidity: Normal Distribution. Mean = 78.8395 StandardDev = 9.8023 We
windy: Discrete Estimator. Counts = 4 7 (Total = 11)
Class no: Prior probability = 0.38
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...
Classifier output
=== Run information ===
Scheme: weka.classifiers.bayes.NaiveBayes
Relation: weather
Instances: 14
Attributes: 5
outlook
temperature
humidity
windy
play
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
Naive Bayes Classifier
Class yes: Prior probability = 0.63
outlook: Discrete Estimator. Counts = 3 5 4 (Total = 12)
temperature: Normal Distribution. Mean = 72.9697 StandardDev = 5.2304
humidity: Normal Distribution. Mean = 78.8395 StandardDev = 9.8023 We
Counts = 4 7 (Total = 11)
= 0.38
. Counts = 4 1 3 (Total = 8)
ution. Mean = 74.8364 StandardDev = 7.384
on. Mean = 86.1111 StandardDev = 9.2424 We
Counts = 4 3 (Total = 7)
(Nom) play
Start
Stop
Result list (right-click for options)
21:25:38 - trees.J48
21:47:54 - bayes.NaiveBayes
View in main window
View in separate window
Save result buffer
Load model
Save model
Re-evaluate model on current test set
Visualize classifier errors
Visualize tree
Visualize margin curve
Visualize threshold curve
Visualize cost curve
yes
no
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) play
Start
Stop
Result list (right-click for options)
21:25:38 - trees.J48
21:47:54 - bayes.NaiveBayes
Weka Classifier Visualize: ThresholdCurve. (Class value yes)
X: False Positive Rate (Num)
Y: True Positive Rate (Num)
Colour: Threshold (Num)
Select Instance
Reset
Clear
Save
Jitter
Plot: ThresholdCurve (Area under ROC = 0.4444)
1
0.5
0
0.2
0.6
1
X
Y
Class colour
0.13
0.56
1
Status
OK
Log
x 0

To Do List

- Try Decision Tree, Naïve Bayes, and Logistic Regression and Support Vector Machines classifiers on a CiteSeerX dataset
  - The dataset contains titles and abstracts of papers from Computer Science that are available in the CiteSeer digital library;
  - The class for each example in the dataset is the topic of the paper. There are six possible classes.
  - The dataset is available in arff format.
- Use various model parameters