Hand on Weka

2014/11/11

Petra Kralj Novak
Petra.Kralj.Novak@ijs.si

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Data Mining Tools

- Weka http://www.cs.waikato.ac.nz/ml/weka/
- Orange http://orange.biolab.si/
- Knime http://www.knime.org/
- Taverna http://www.taverna.org.uk/
- Rapid Miner http://rapid-i.com/content/view/181/196/
- ClowdFlows http://clowdflows.org/

Weka (Waikato Environment for Knowledge Analysis)

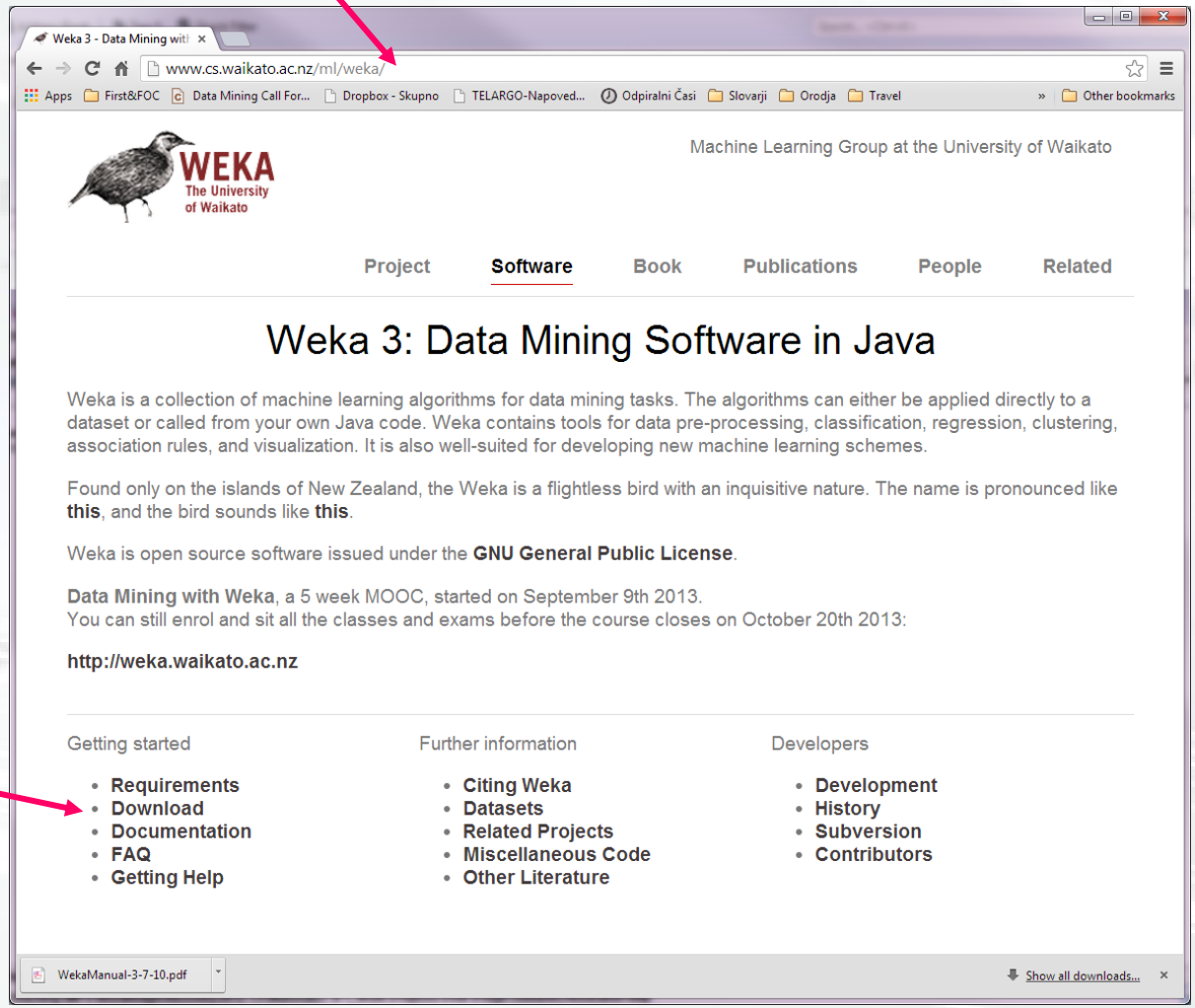
- Collection of machine learning algorithms for data mining tasks
- The algorithms
  - Can be applied directly to a dataset
  - Can be called from Java code (library)
- Weka contains tools for
  - Data pre-processing
  - Classification
  - Regression
  - Clustering
  - Association rules
  - Visualization
- Weka is open source software issued under the GNU General Public Licanse

Exsercise1: ID3 in Weka

1. Build a decision tree with the ID3 algorithm on the lenses dataset, evaluate on a separate test set

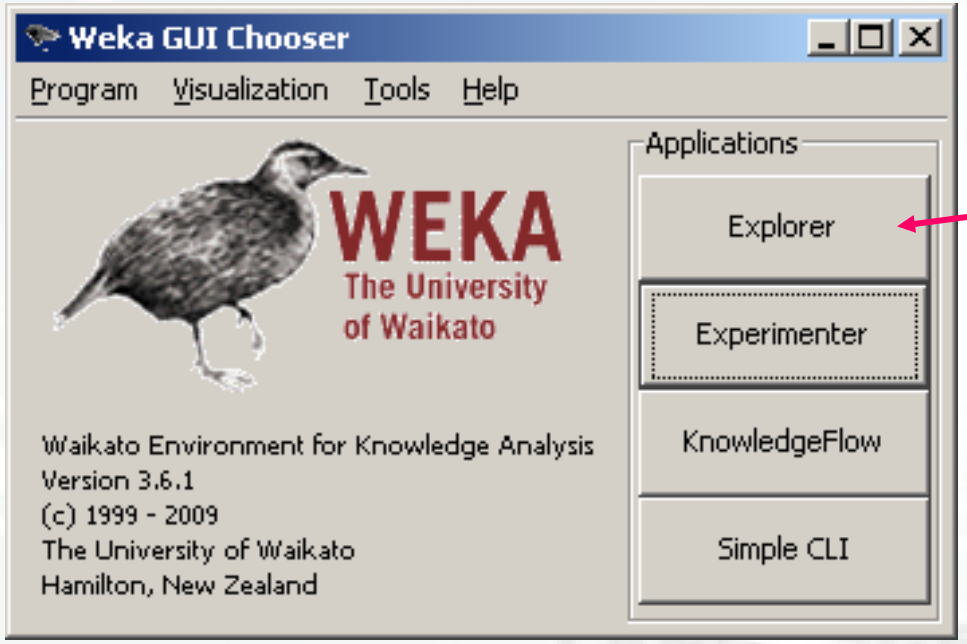
Weka: Install

http://www.cs.waikato.ac.nz/ml/weka/



Download version 3.6

Weka: Run Explorer

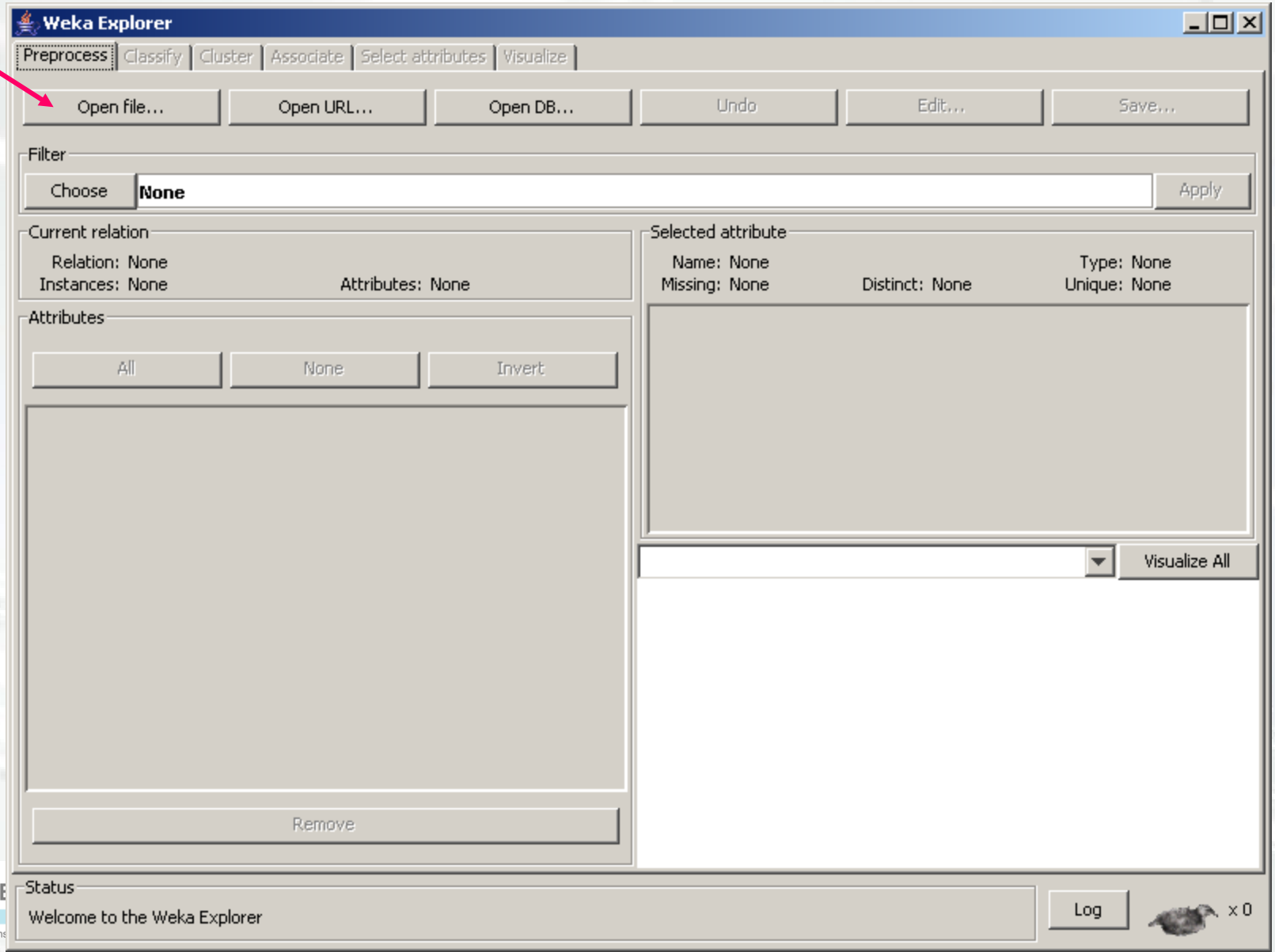


Choose Explorer

Exercise 1: ID3 in Weka

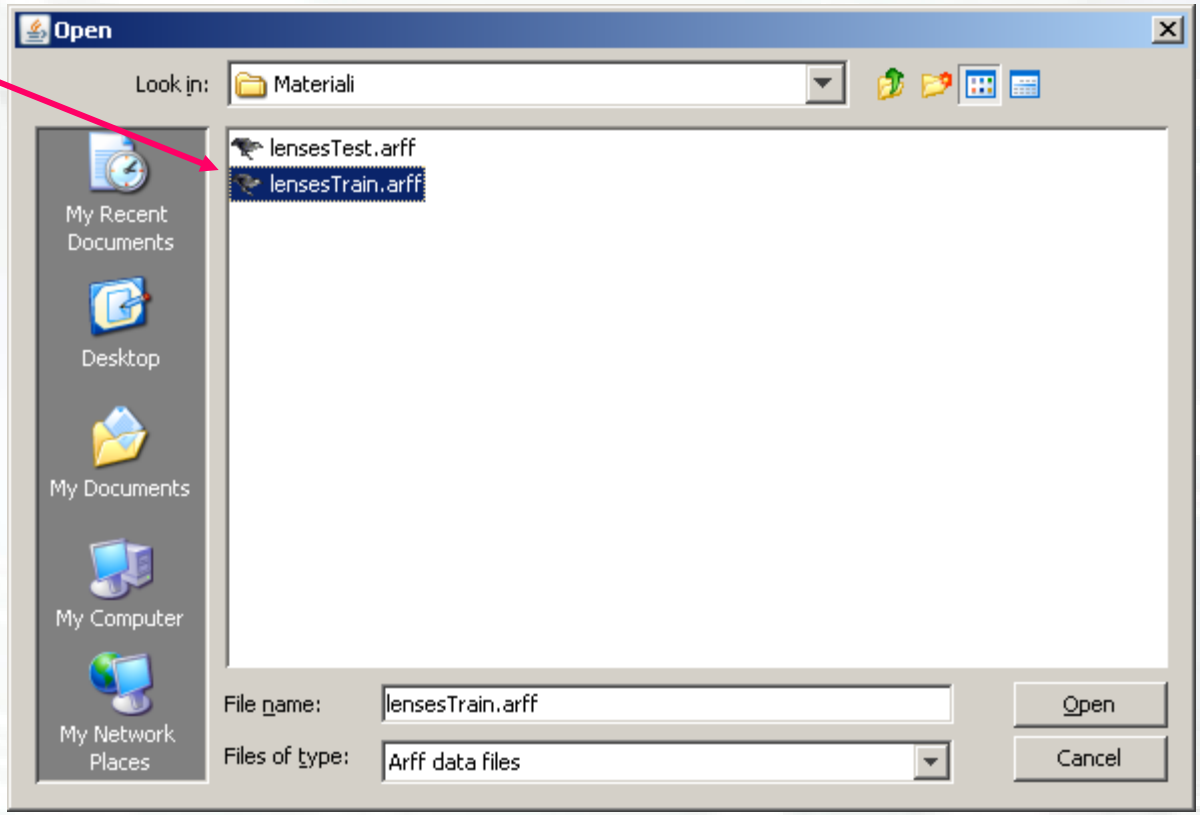
- In the Weka data mining tool, induce a decision tree for the **lenses** dataset with the ID3 algorithm.
- Data:
  - lensesTrain.arff
  - lensesTest.arff
- Compare the outcome with the manually obtained results.

Load the data



Load the data - 2

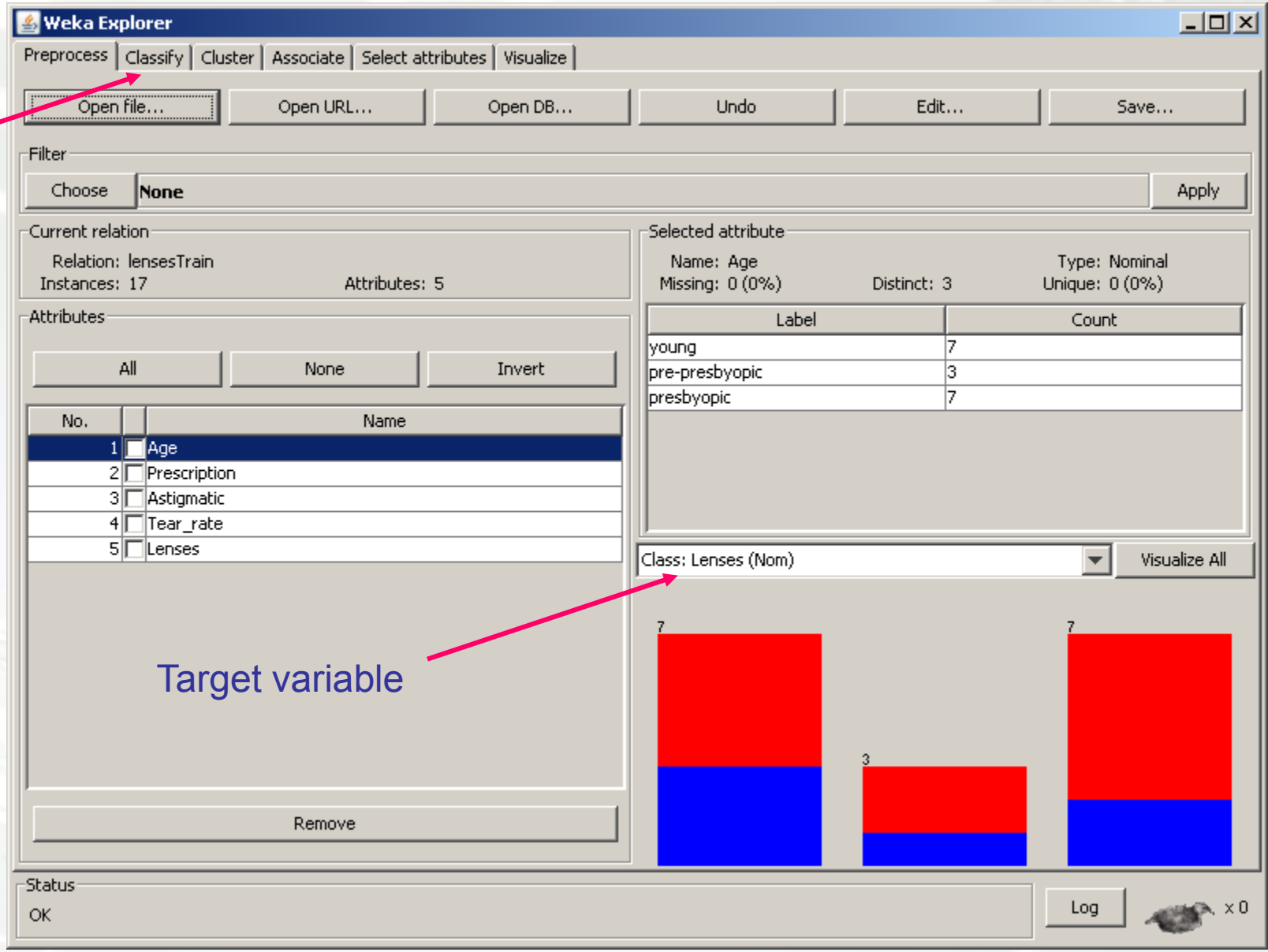
lensesTrain.arff



The data are loaded

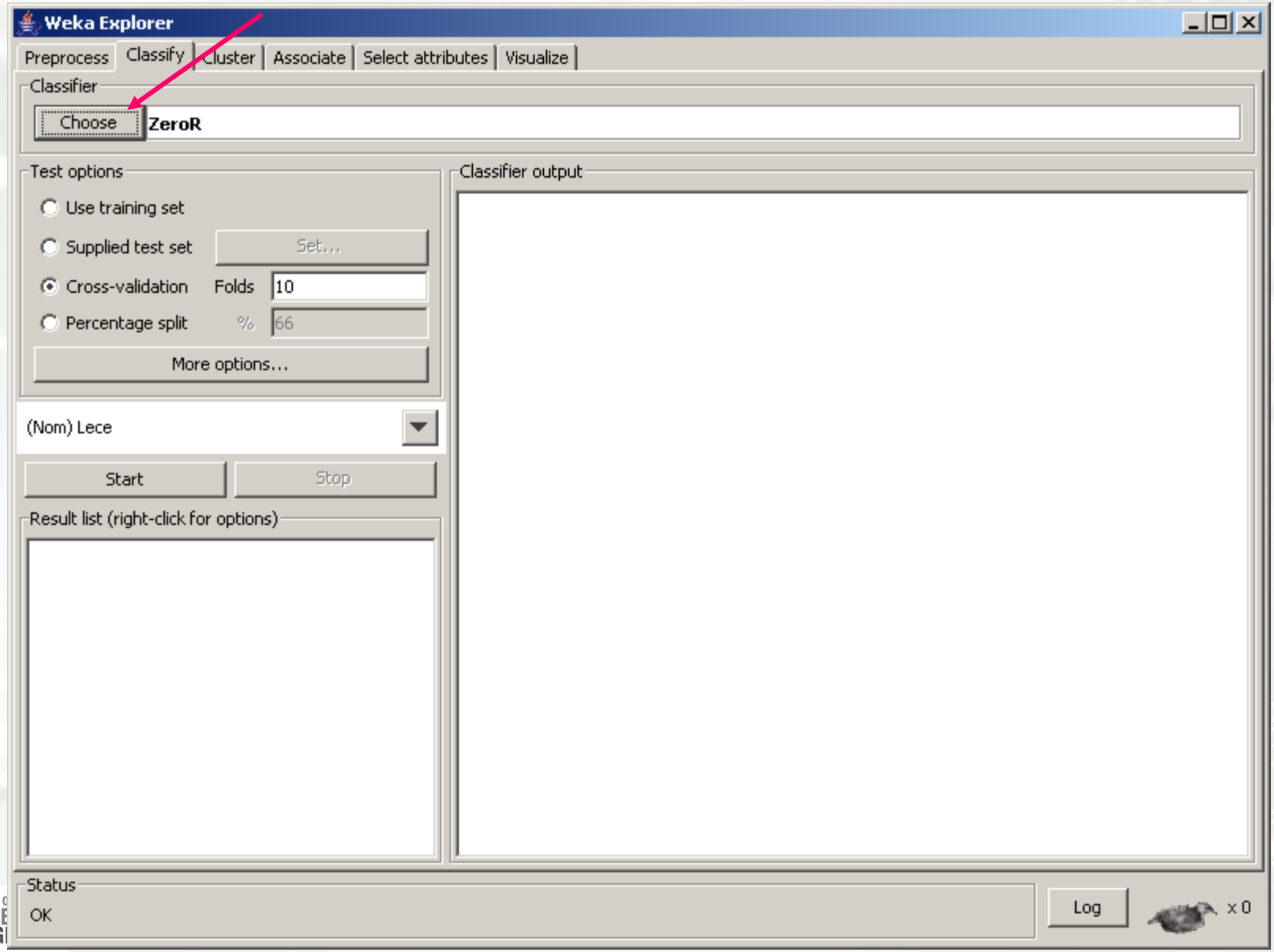
Choose

"Classify"



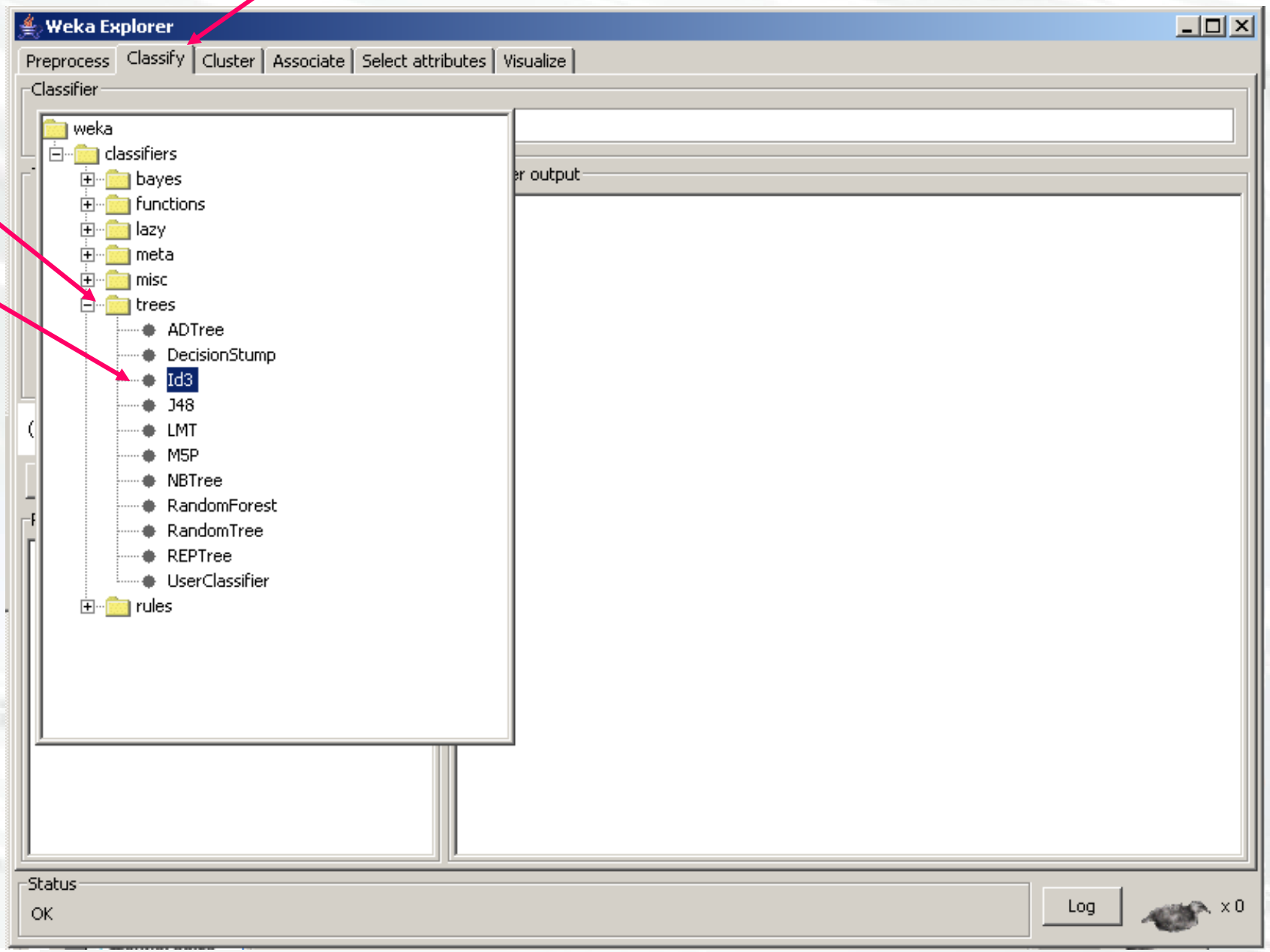
Target variable

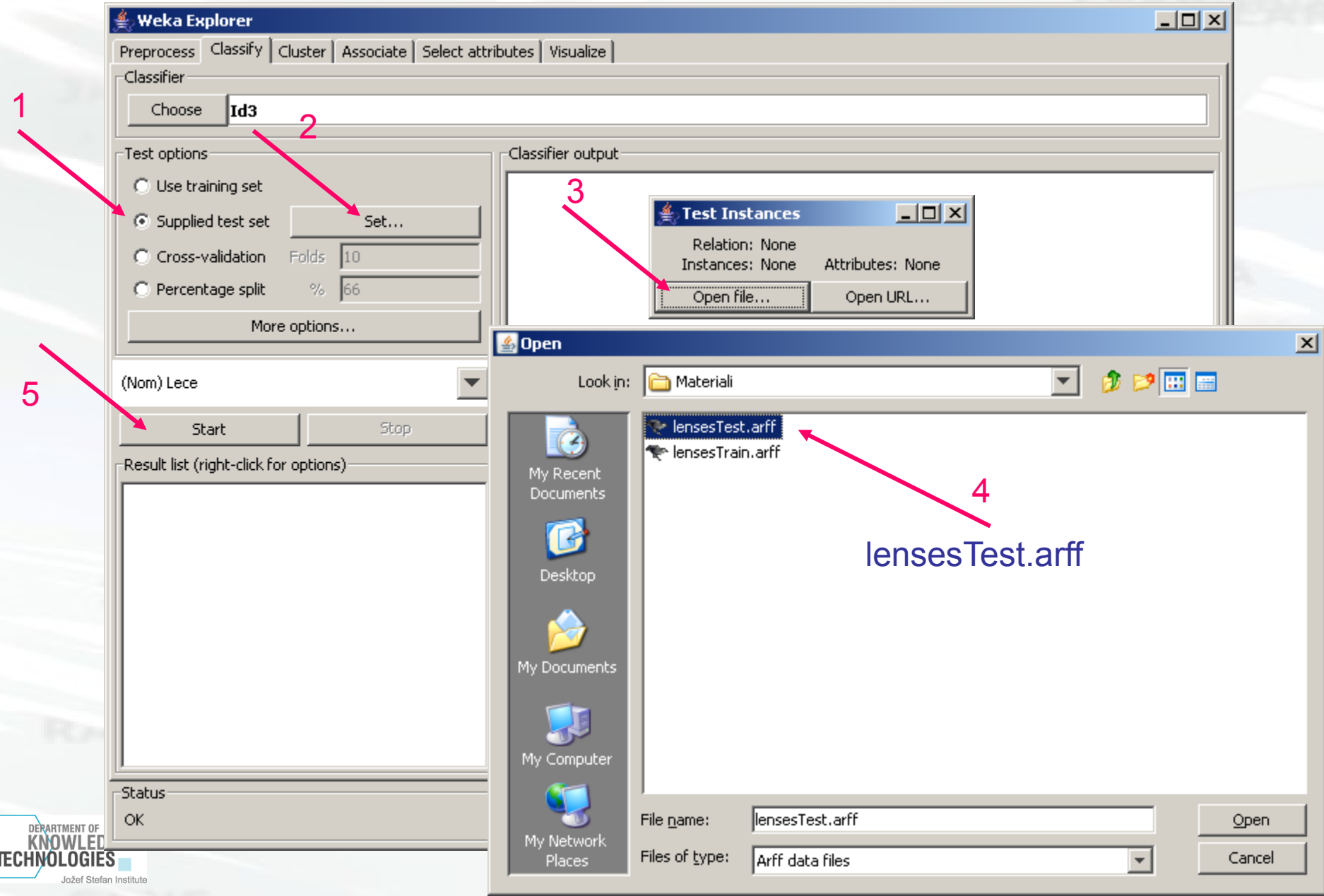
Choose algoritem

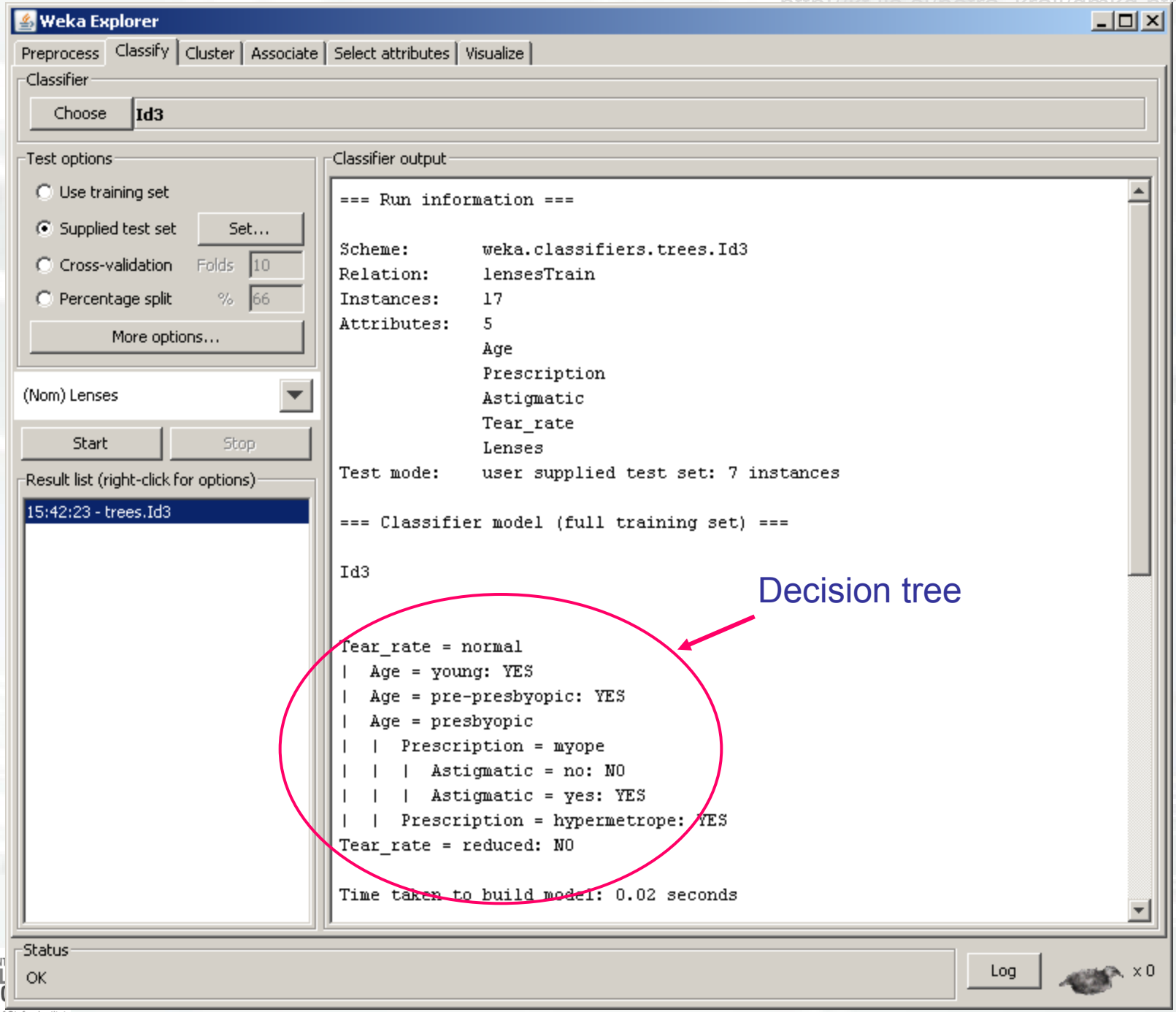


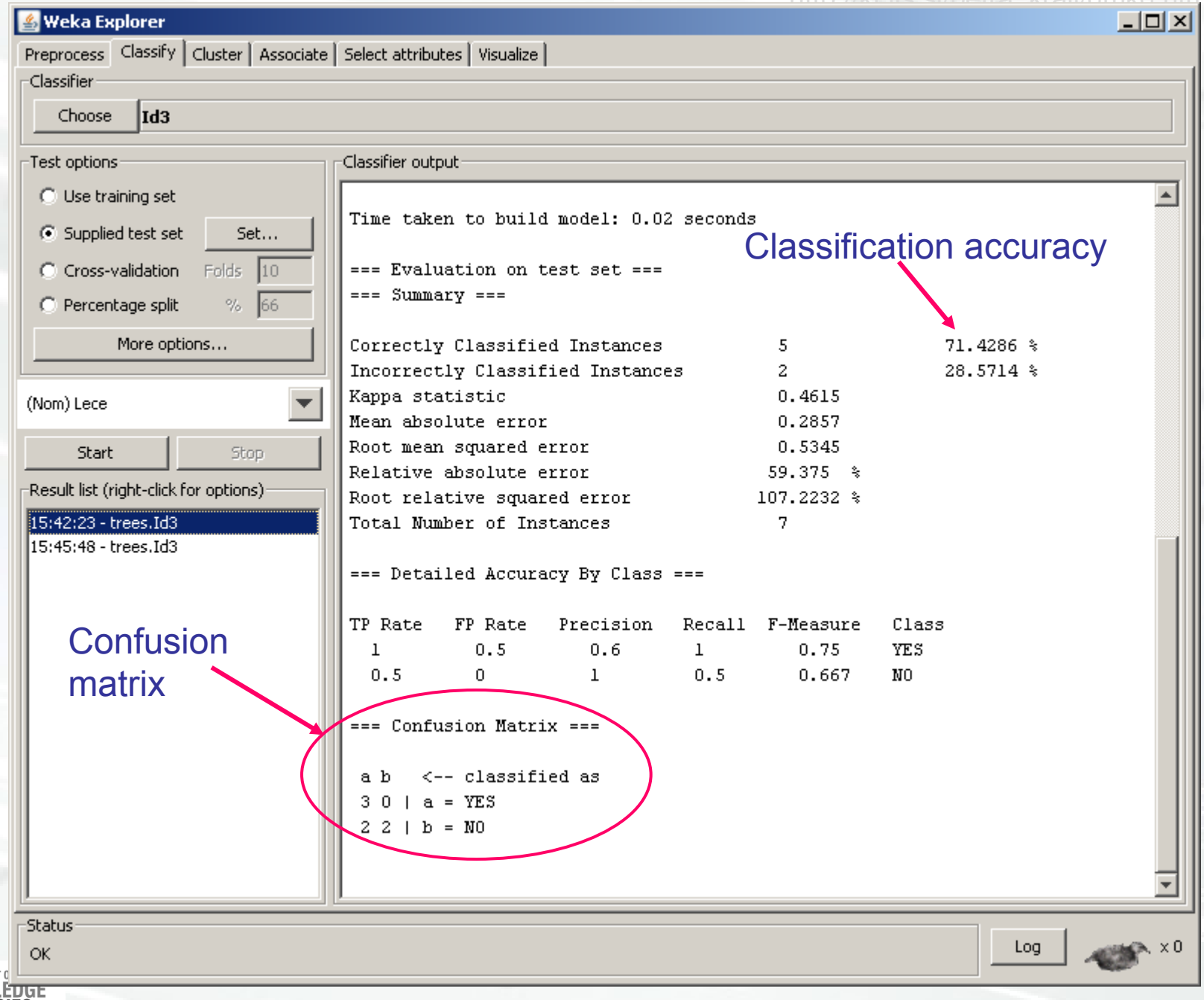
trees

Id3


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
output
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
Id3
1
2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
Classifier output
3
Test Instances
Relation: None
Instances: None
Attributes: None
Open file...
Open URL...
5
(Nom) Lece
Start
Stop
Result list (right-click for options)
Status
OK
Open
Look in:
Materiali
My Recent Documents
Desktop
My Documents
My Computer
My Network Places
lensesTest.arff
lensesTrain.arff
4
lensesTest.arff
File name:
lensesTest.arff
Files of type:
Arff data files
Open
Cancel




Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
Id3
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) Lece
Start
Stop
Result list (right-click for options)
15:42:23 - trees.Id3
15:45:48 - trees.Id3
Classifier output
Time taken to build model: 0.02 seconds
=== Evaluation on test set ===
=== Summary ===
Correctly Classified Instances 5 71.4286 %
Incorrectly Classified Instances 2 28.5714 %
Kappa statistic 0.4615
Mean absolute error 0.2857
Root mean squared error 0.5345
Relative absolute error 59.375 %
Root relative squared error 107.2232 %
Total Number of Instances 7
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure Class
1 0.5 0.6 1 0.75 YES
0.5 0 1 0.5 0.667 NO
=== Confusion Matrix ===
a b <-- classified as
3 0 | a = YES
2 2 | b = NO
Status
OK
Log
x 0
Classification accuracy
Confusion matrix

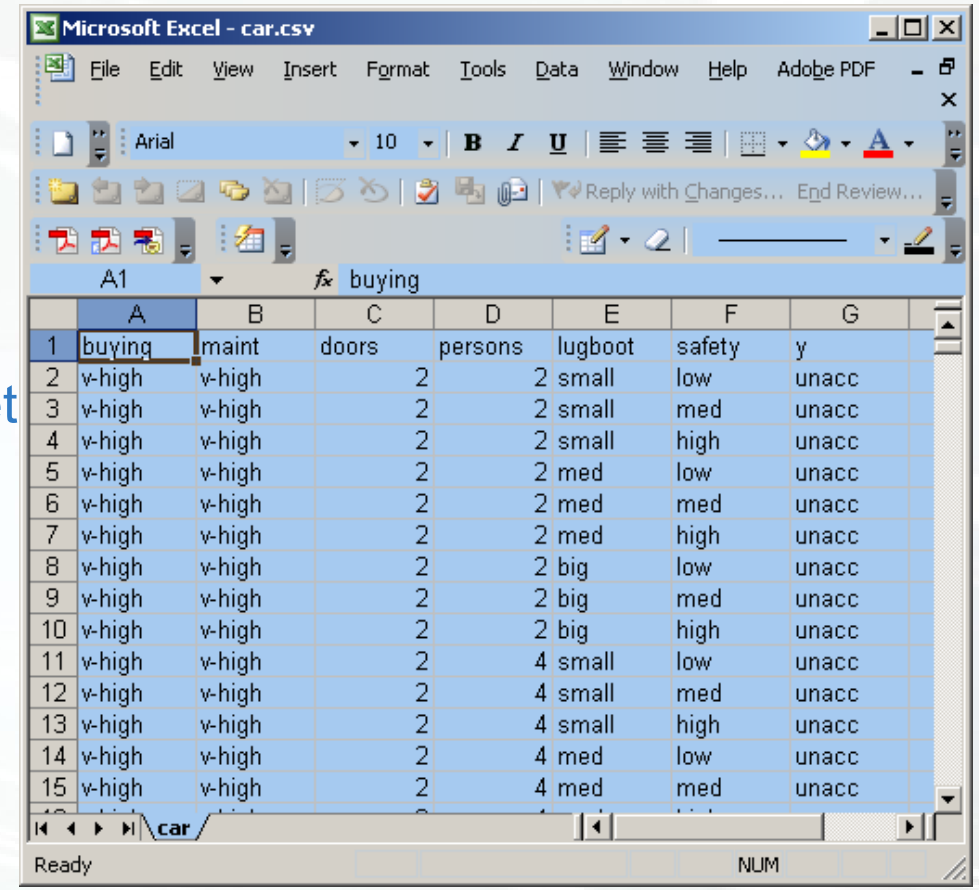
Exercise 2: CAR dataset

- 1728 examples
- 6 attributes
  - 6 nominal
  - 0 numeric
- Nominal target variable
  - 4 classes: unacc, acc, good, v-good
  - Distribution of classes
    - unacc (70%), acc (22%), good (4%), v-good (4%)
- No missing values

Preparing the data for WEKA - 1

Data in a spreadsheet
(e.g. MS Excel)

- Rows are examples
- Columns are attributes
- The last column is the target variable

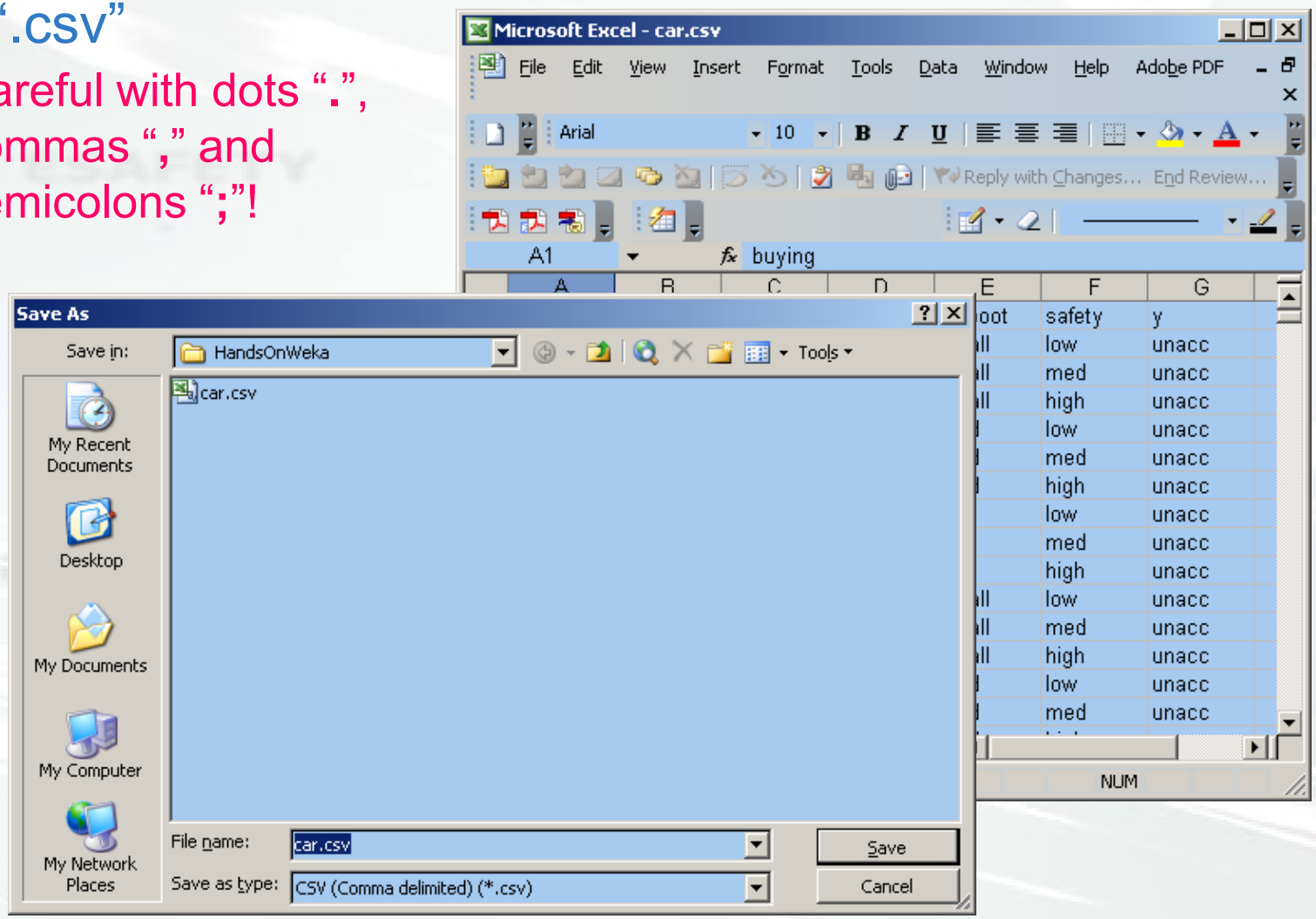


	A	B	C	D	E	F	G
1	buying	maint	doors	persons	lugboot	safety	y
2	v-high	v-high	2	2	small	low	unacc
3	v-high	v-high	2	2	small	med	unacc
4	v-high	v-high	2	2	small	high	unacc
5	v-high	v-high	2	2	med	low	unacc
6	v-high	v-high	2	2	med	med	unacc
7	v-high	v-high	2	2	med	high	unacc
8	v-high	v-high	2	2	big	low	unacc
9	v-high	v-high	2	2	big	med	unacc
10	v-high	v-high	2	2	big	high	unacc
11	v-high	v-high	2	4	small	low	unacc
12	v-high	v-high	2	4	small	med	unacc
13	v-high	v-high	2	4	small	high	unacc
14	v-high	v-high	2	4	med	low	unacc
15	v-high	v-high	2	4	med	med	unacc

Preparing the data for WEKA - 2

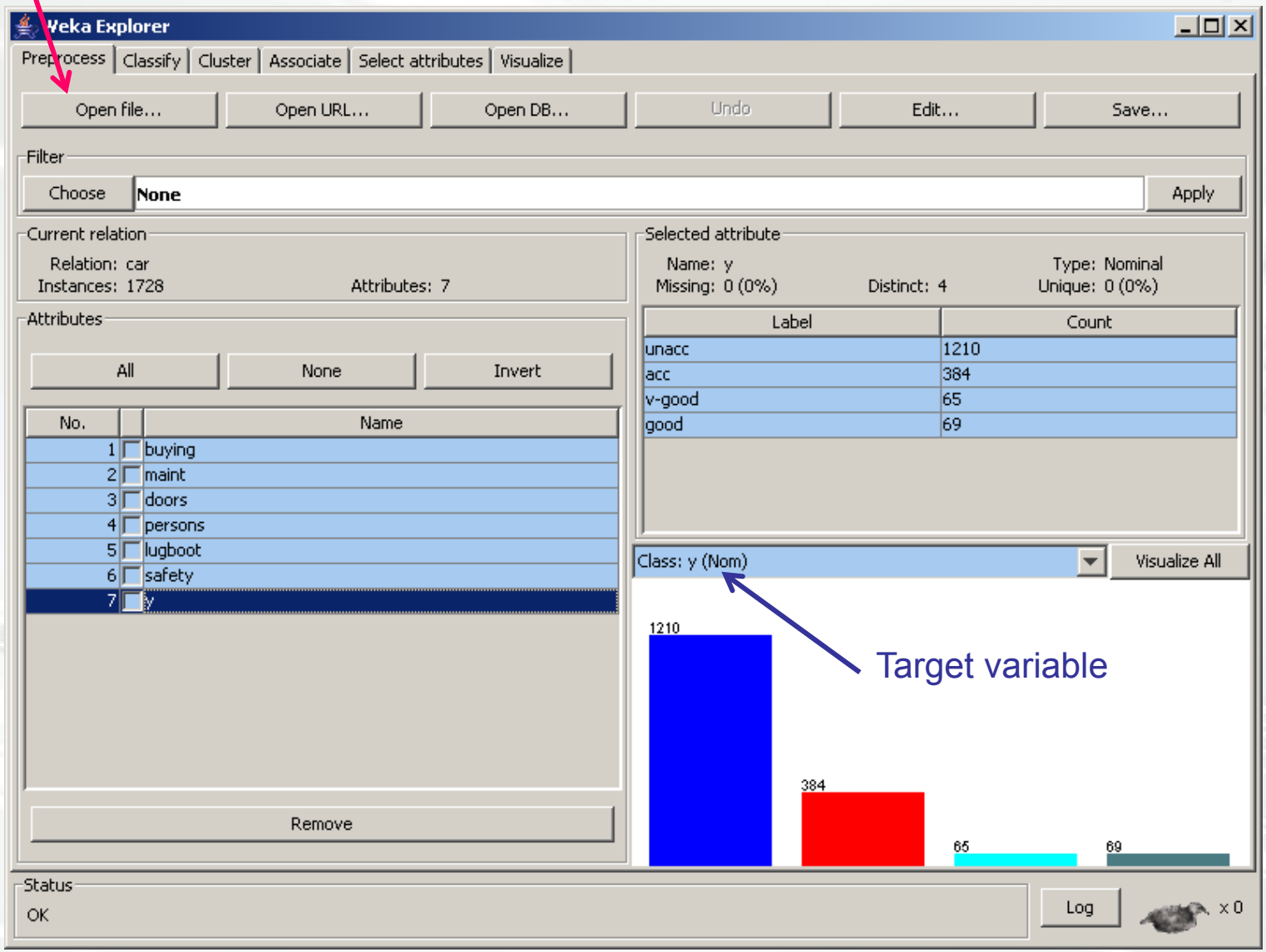
Save as ".csv"

- Careful with dots ".", commas "," and semicolons ";"!

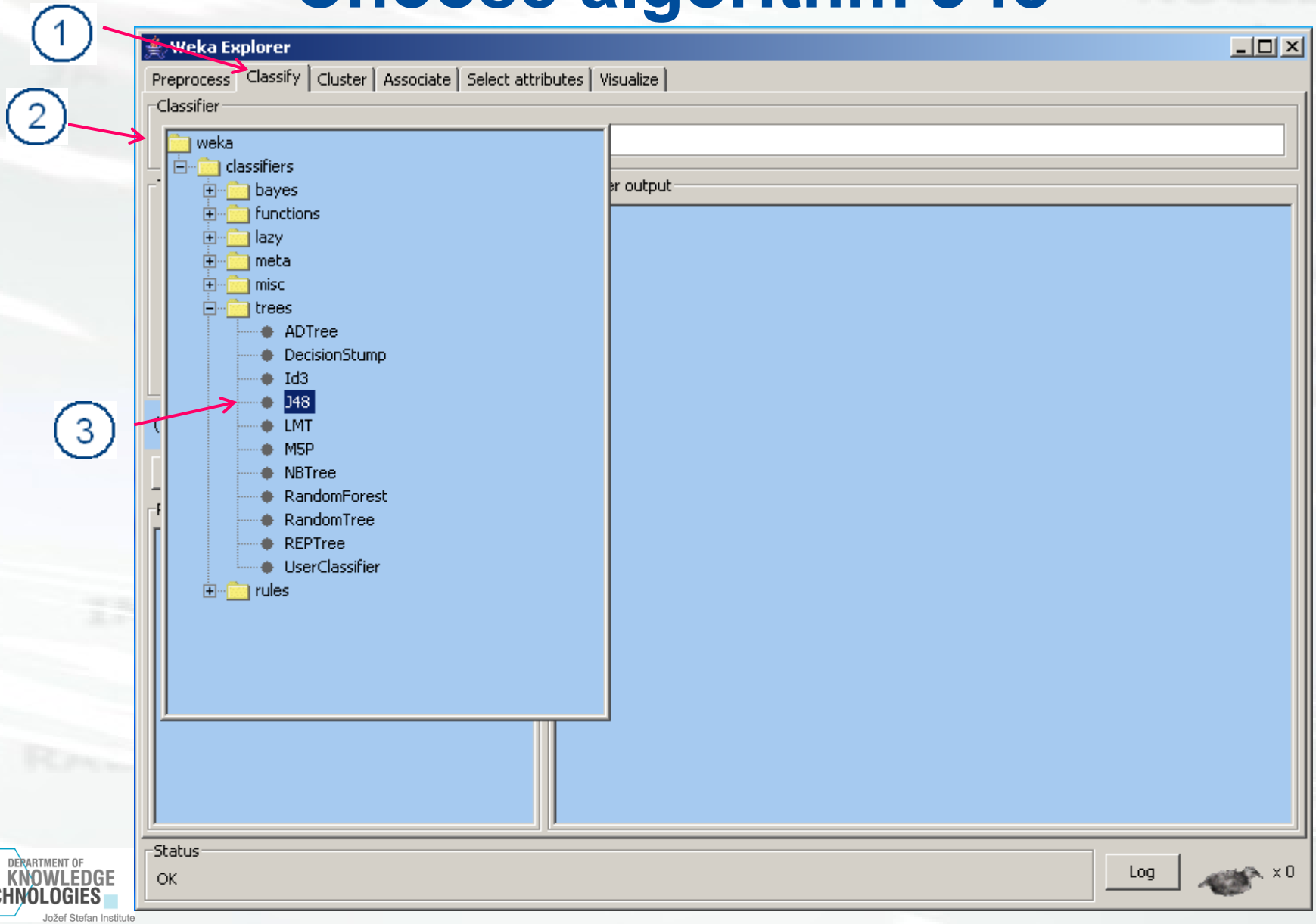


Load the data

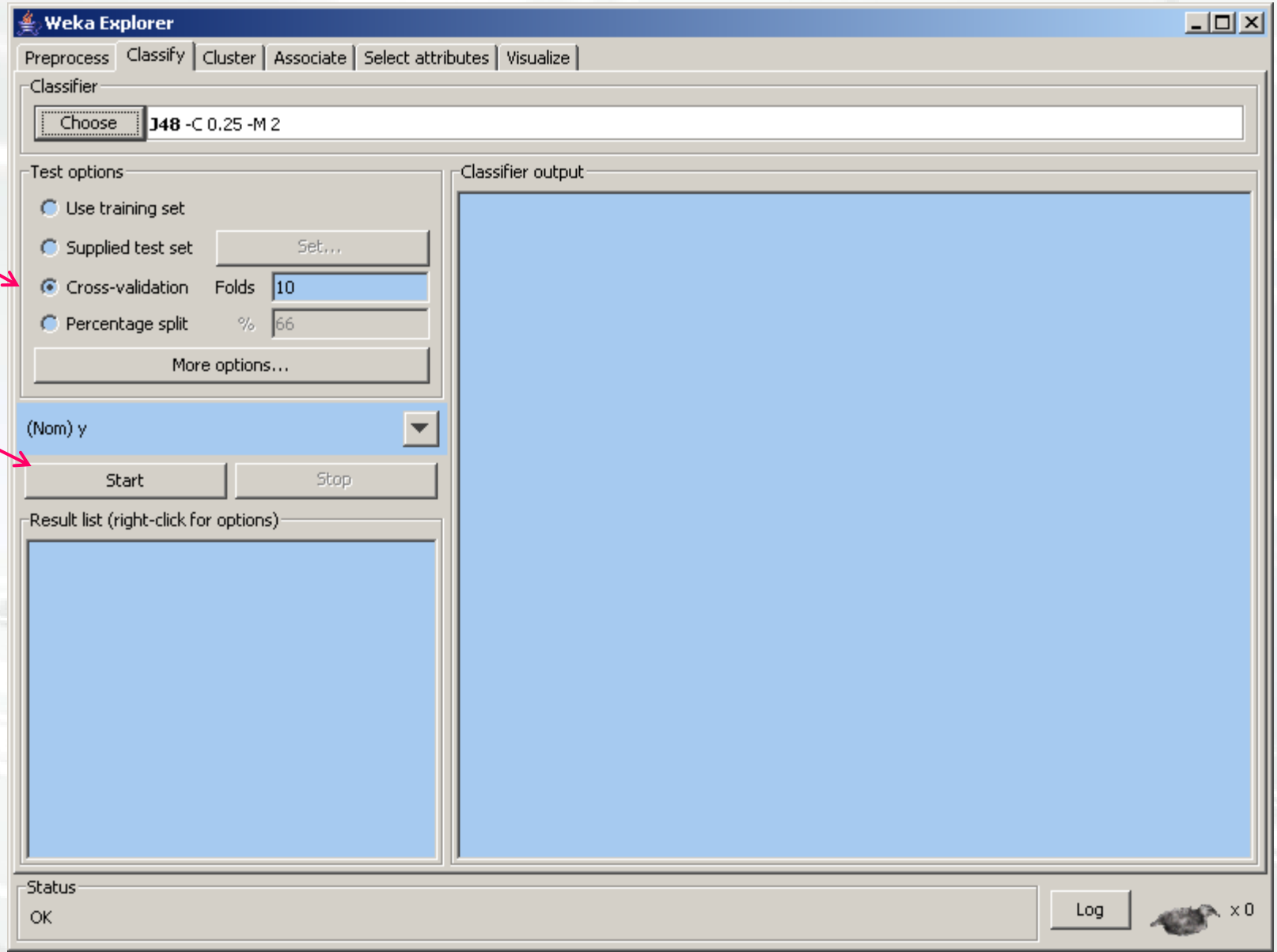
Car.csv

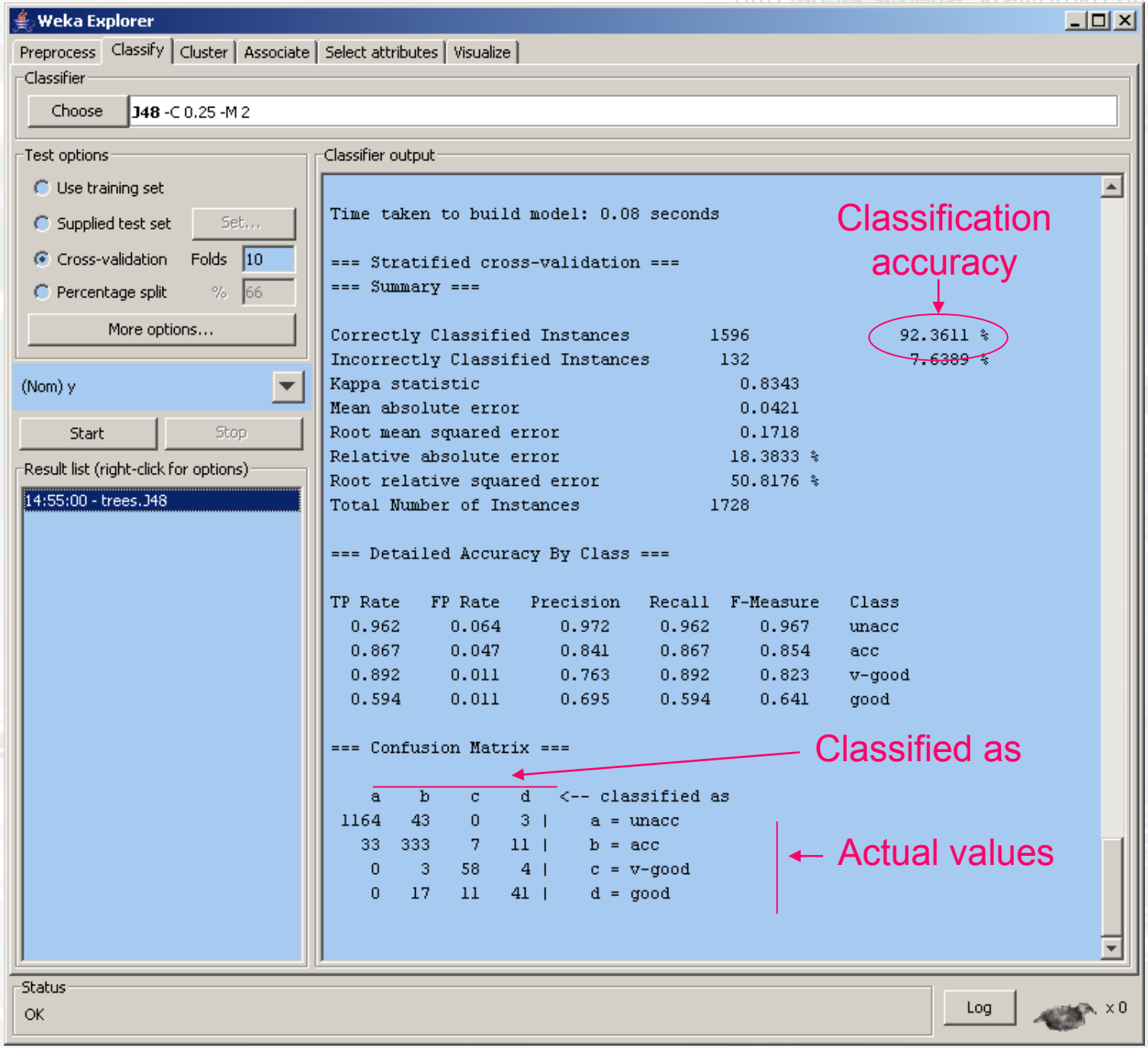


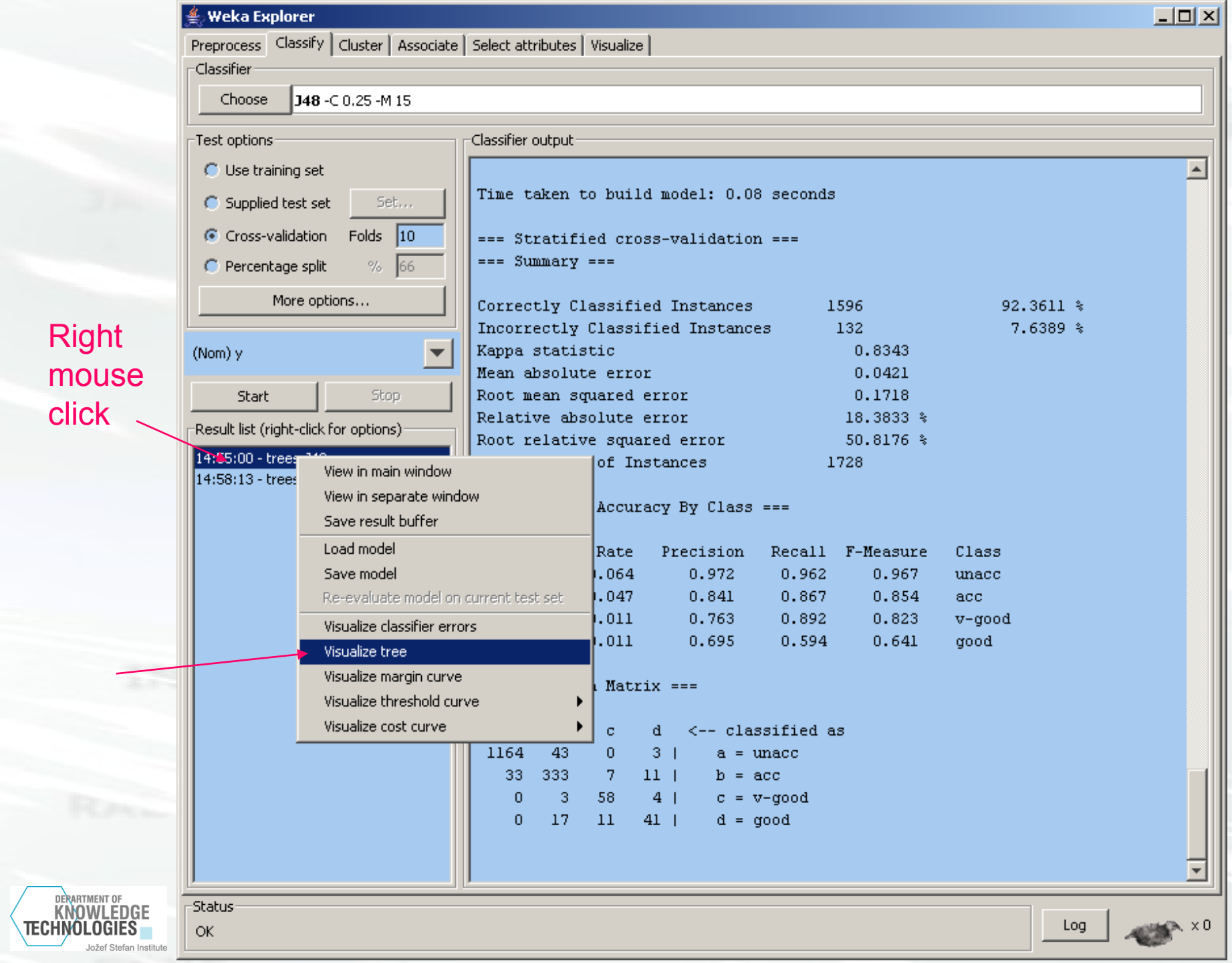
Choose algorithm J48



Building and evaluating the tree




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) y
Start
Stop
Result list (right-click for options)
14:55:00 - trees.J48
Classifier output
Time taken to build model: 0.08 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 1596 92.3611 %
Incorrectly Classified Instances 132 7.6389 %
Kappa statistic 0.8343
Mean absolute error 0.0421
Root mean squared error 0.1718
Relative absolute error 18.3833 %
Root relative squared error 50.8176 %
Total Number of Instances 1728
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure Class
0.962 0.064 0.972 0.962 0.967 unacc
0.867 0.047 0.841 0.867 0.854 acc
0.892 0.011 0.763 0.892 0.823 v-good
0.594 0.011 0.695 0.594 0.641 good
=== Confusion Matrix ===
a b c d <-- classified as
1164 43 0 3 | a = unacc
33 333 7 11 | b = acc
0 3 58 4 | c = v-good
0 17 11 41 | d = good
Classification accuracy
Classified as
Actual values
Status
OK
Log
x 0



Right mouse click

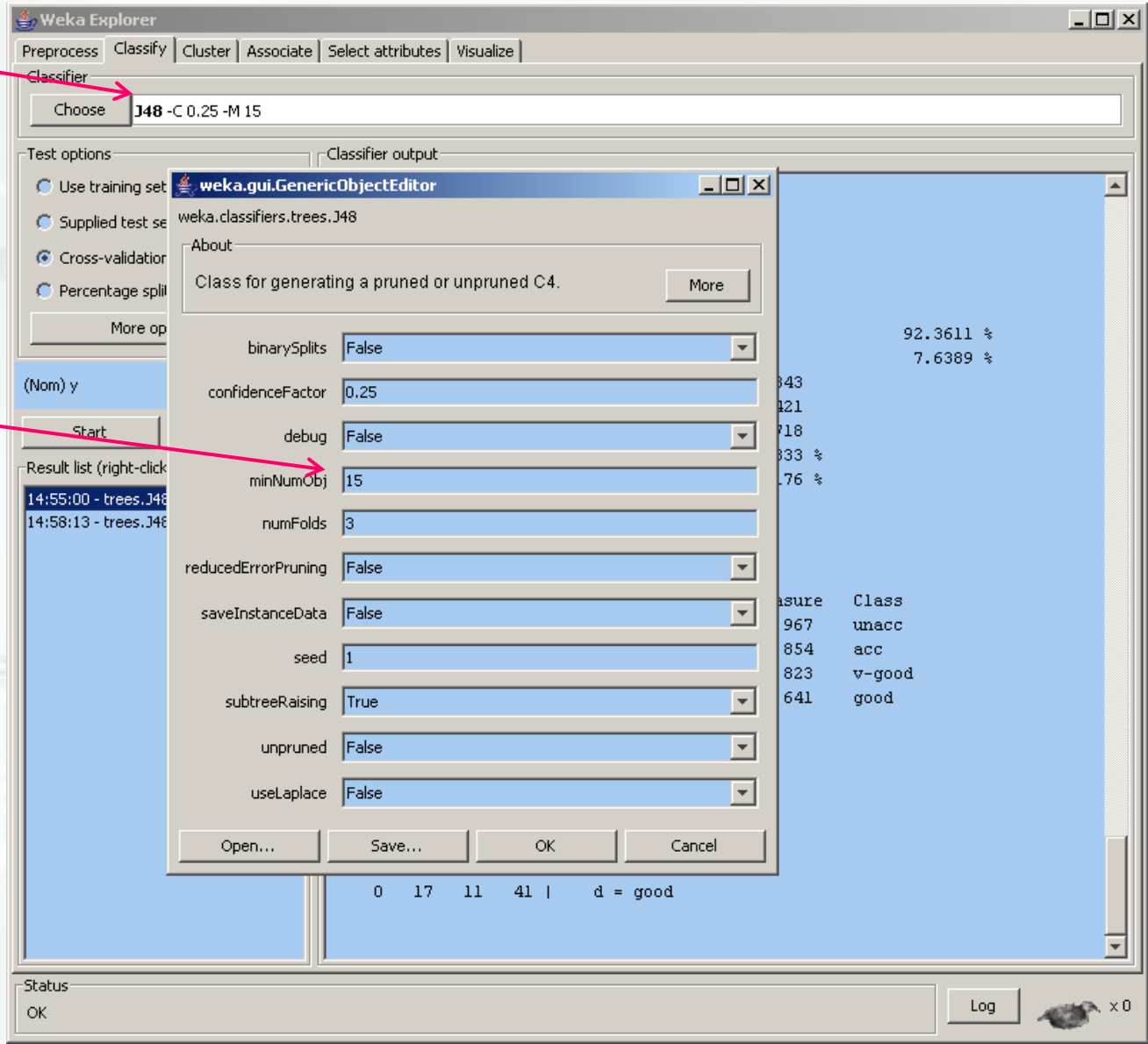
Tree pruning

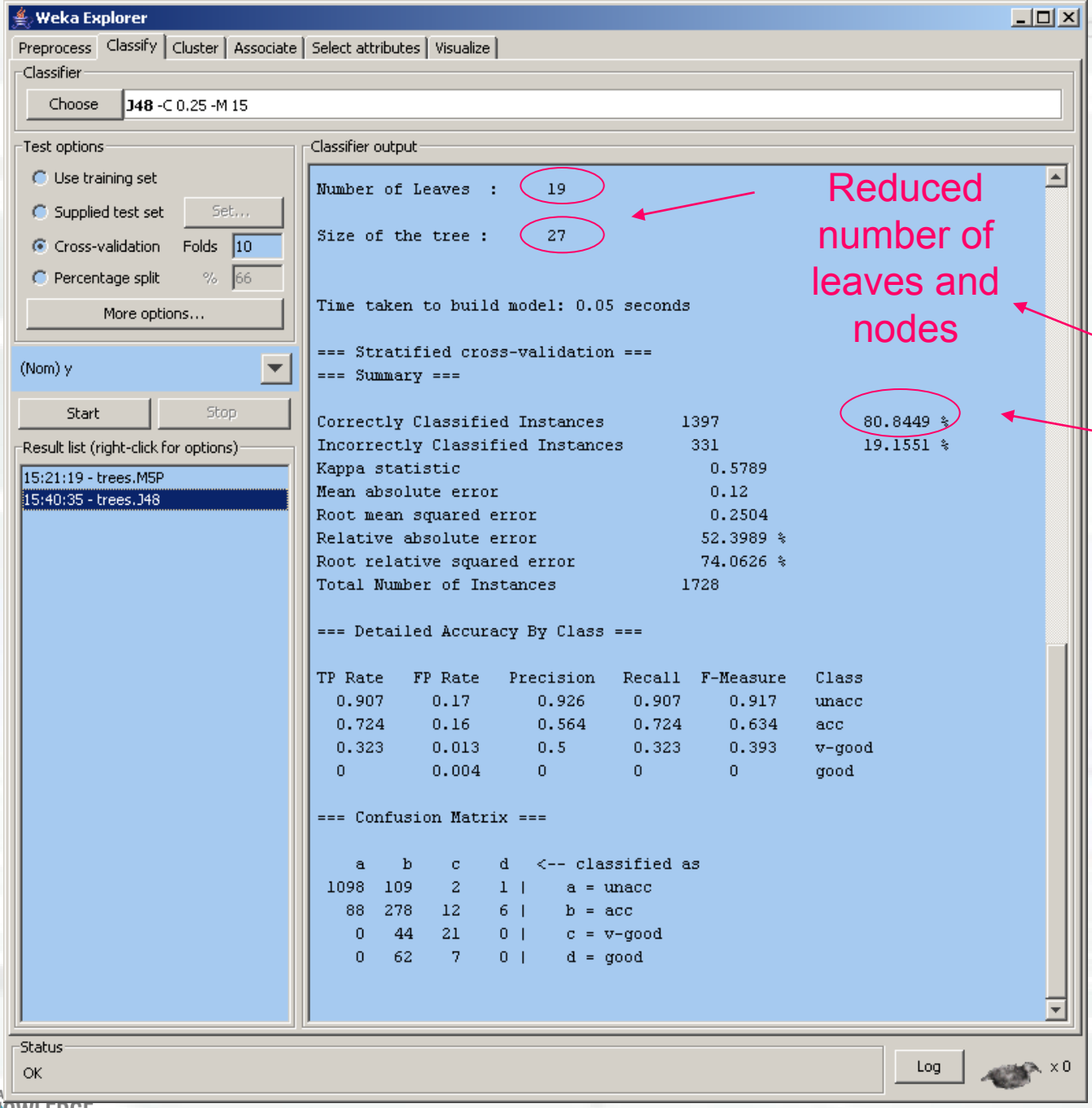
1

Parameters of the algorithm (right mouse click)

2

Set the minimal number of objects per leaf to 15

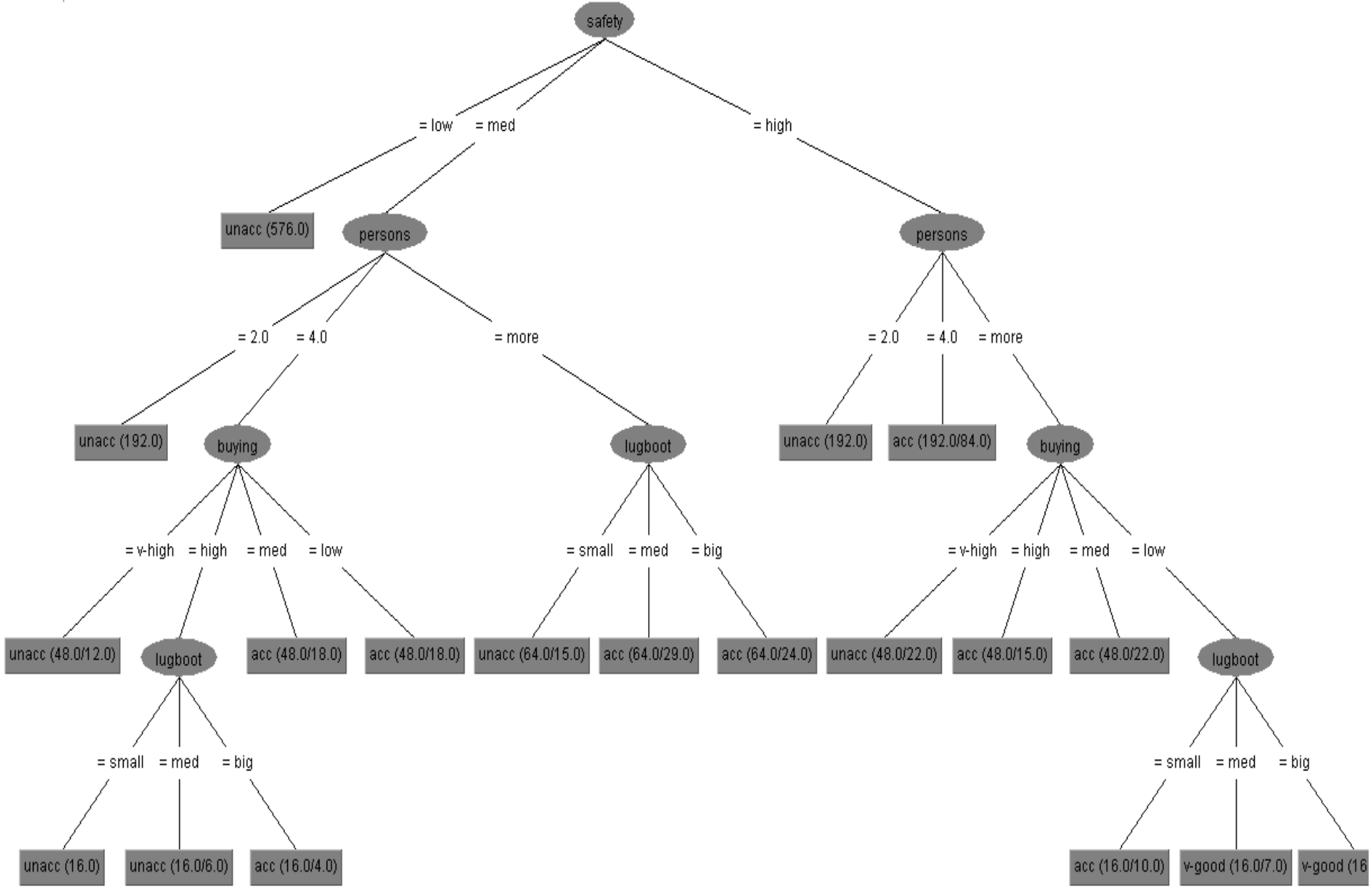




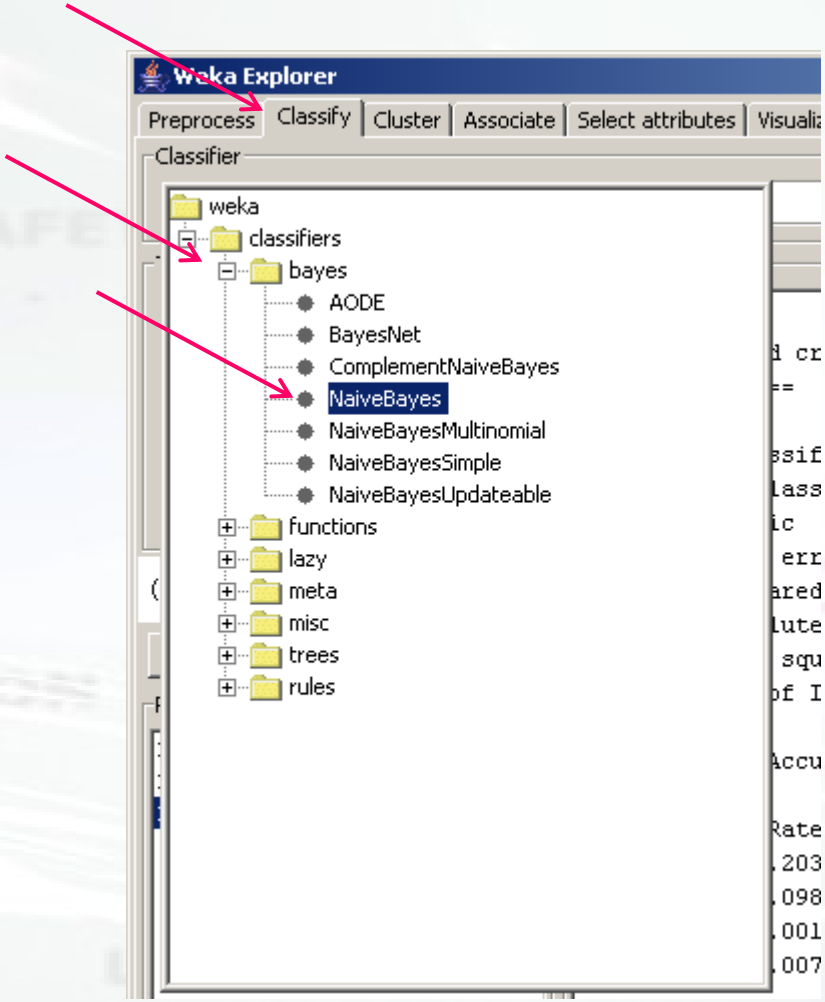
Reduced number of leaves and nodes

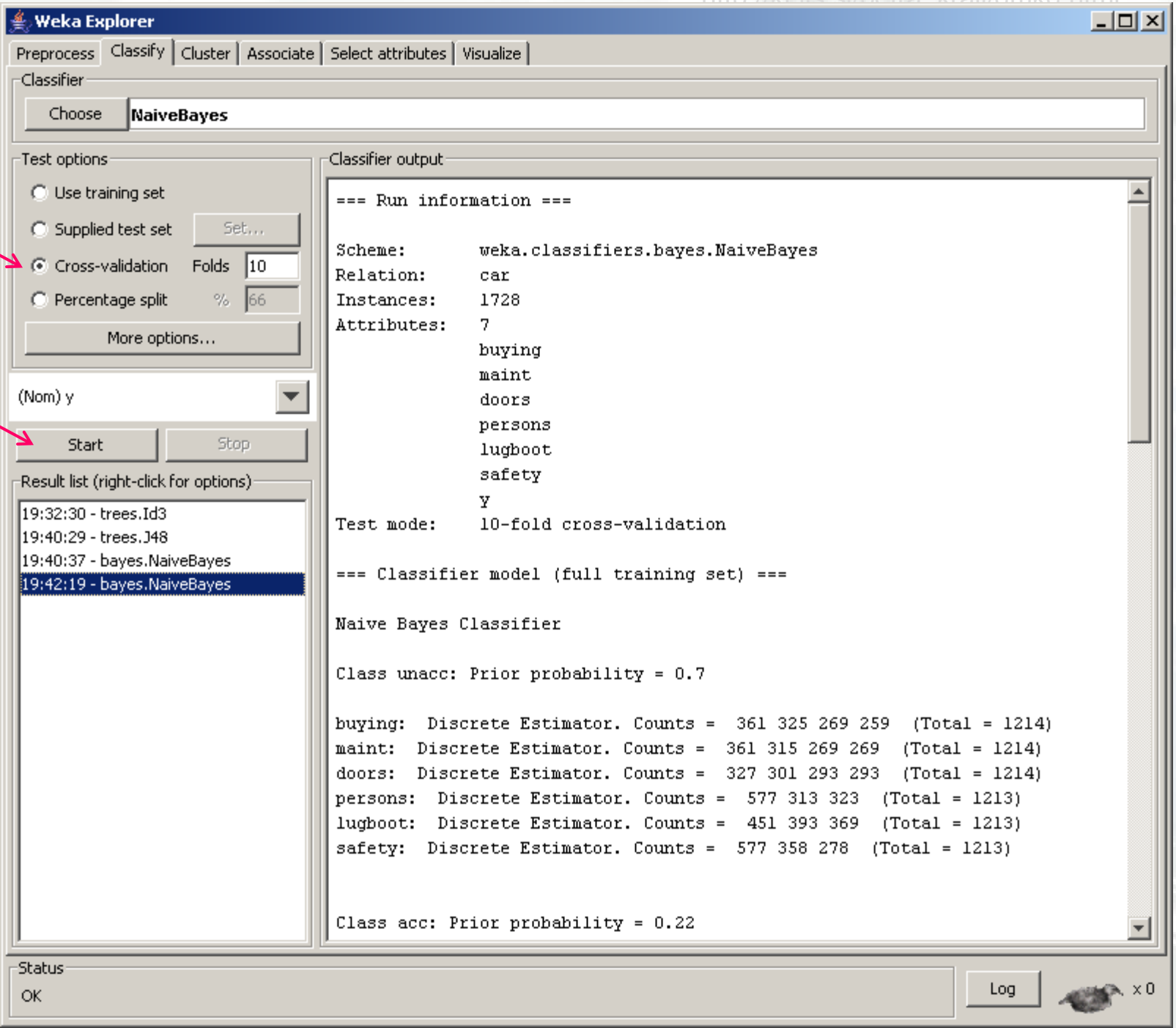
Easier to interpret

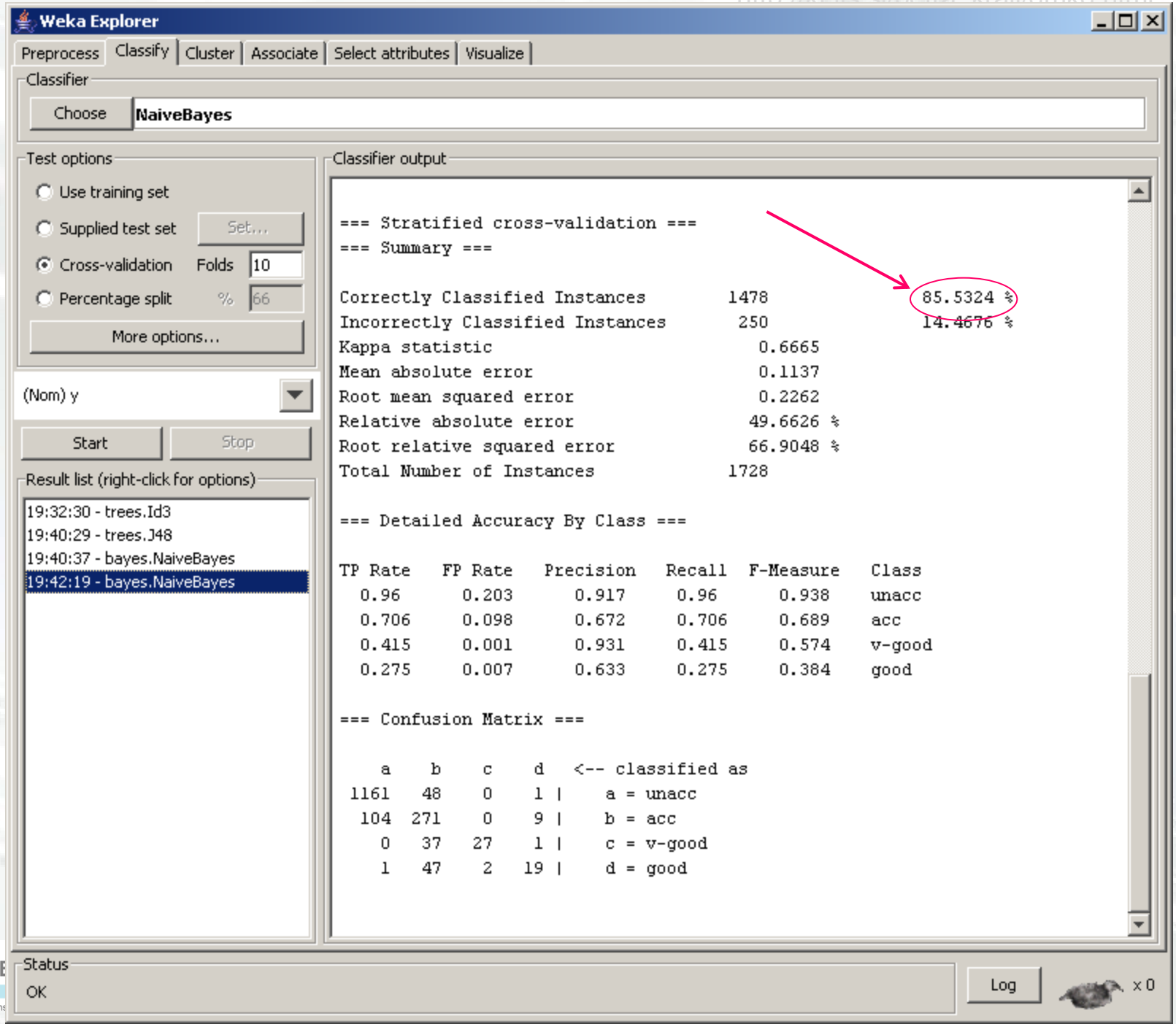
Lower classification accuracy



Naïve Bayes classifier




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) y
Start
Stop
Result list (right-click for options)
19:32:30 - trees.Id3
19:40:29 - trees.J48
19:40:37 - bayes.NaiveBayes
19:42:19 - bayes.NaiveBayes
Classifier output
=== Run information ===
Scheme: weka.classifiers.bayes.NaiveBayes
Relation: car
Instances: 1728
Attributes: 7
buying
maint
doors
persons
lugboot
safety
y
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
Naive Bayes Classifier
Class unacc: Prior probability = 0.7
buying: Discrete Estimator. Counts = 361 325 269 259 (Total = 1214)
maint: Discrete Estimator. Counts = 361 315 269 269 (Total = 1214)
doors: Discrete Estimator. Counts = 327 301 293 293 (Total = 1214)
persons: Discrete Estimator. Counts = 577 313 323 (Total = 1213)
lugboot: Discrete Estimator. Counts = 451 393 369 (Total = 1213)
safety: Discrete Estimator. Counts = 577 358 278 (Total = 1213)
Class acc: Prior probability = 0.22
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) y
Start
Stop
Result list (right-click for options)
19:32:30 - trees.Id3
19:40:29 - trees.J48
19:40:37 - bayes.NaiveBayes
19:42:19 - bayes.NaiveBayes
Classifier output
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 1478 85.5324 %
Incorrectly Classified Instances 250 14.4676 %
Kappa statistic 0.6665
Mean absolute error 0.1137
Root mean squared error 0.2262
Relative absolute error 49.6626 %
Root relative squared error 66.9048 %
Total Number of Instances 1728
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure Class
0.96 0.203 0.917 0.96 0.938 unacc
0.706 0.098 0.672 0.706 0.689 acc
0.415 0.001 0.931 0.415 0.574 v-good
0.275 0.007 0.633 0.275 0.384 good
=== Confusion Matrix ===
a b c d <-- classified as
1161 48 0 1 | a = unacc
104 271 0 9 | b = acc
0 37 27 1 | c = v-good
1 47 2 19 | d = good
Status
OK
Log
x 0

Summary

- Weka
- ID3, separate test set
- Data preparation
- J48 (C4.5), cross validation, tree prunning
- Naïve Bayes