Classification in WEKA

2008/10/22

Petra Kralj Novak
Petra.Kralj@ijs.si

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Petra.Kralj@ijs.si

Practice with Weka

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

2. Classification on the CAR dataset
   - Preparing the data
   - Building decision trees
   - Naive Bayes classifier
   - Understanding the Weka output

Weka

Weka is open source softwere for machine learning and data mining.
http://www.cs.waikato.ac.nz/ml/weka/

Download version 3.4

Run Weka

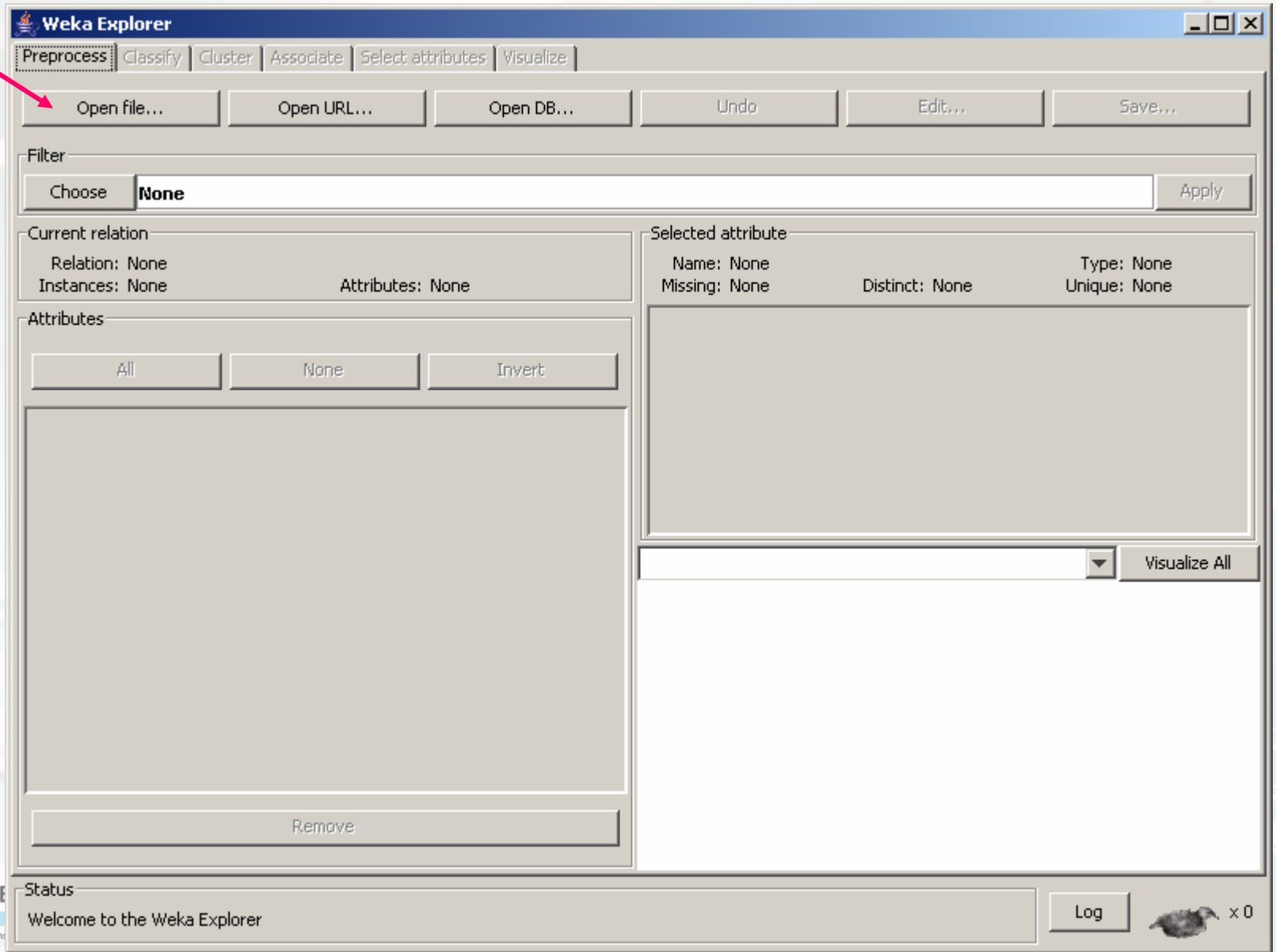
Choose Explorer

Petra.Kralj@ijs.si

Exercise 1: Lenses dataset

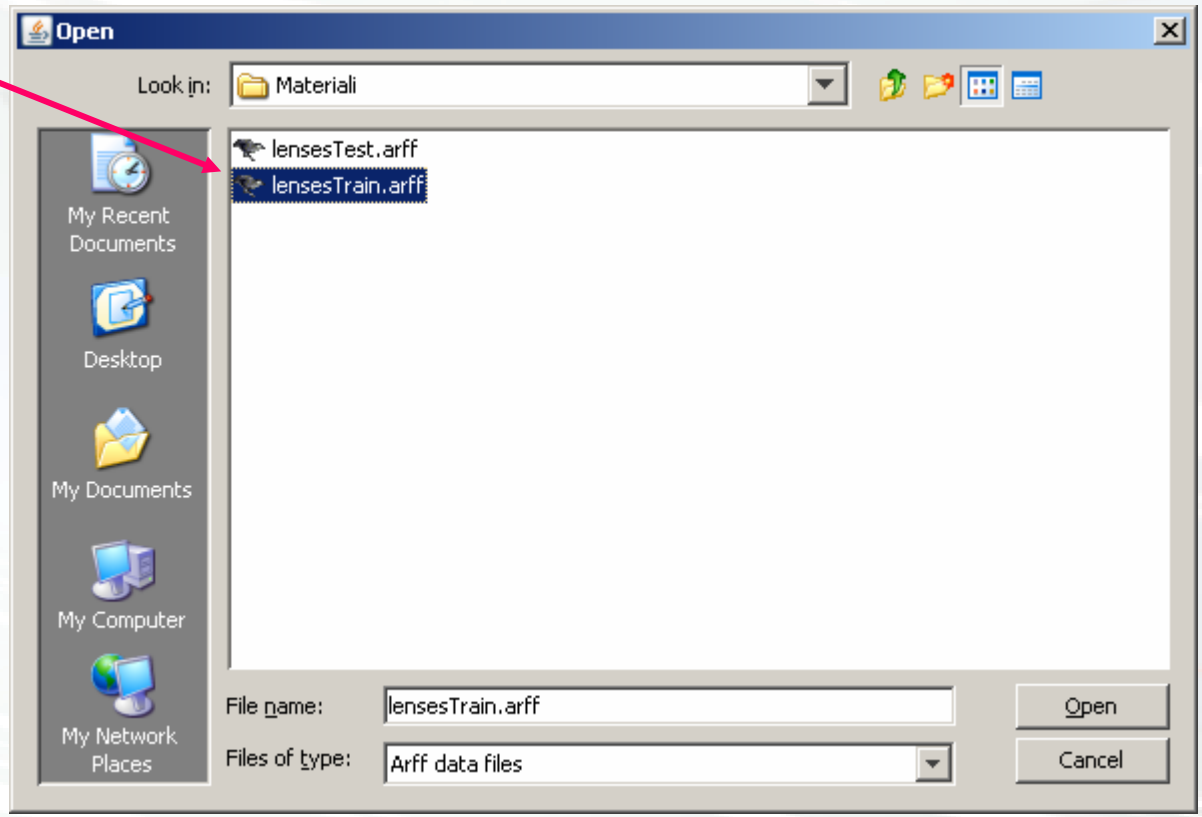
- In the Weka data mining tool induce a decision tree for the lenses dataset with the ID3 algorithm.
- Data:
  - lensesTrain.arff
  - lensesTest.arff
- Compare the outcome with the manually obtained results.

Load the data



Load the data - 2

lensesTrain.arff

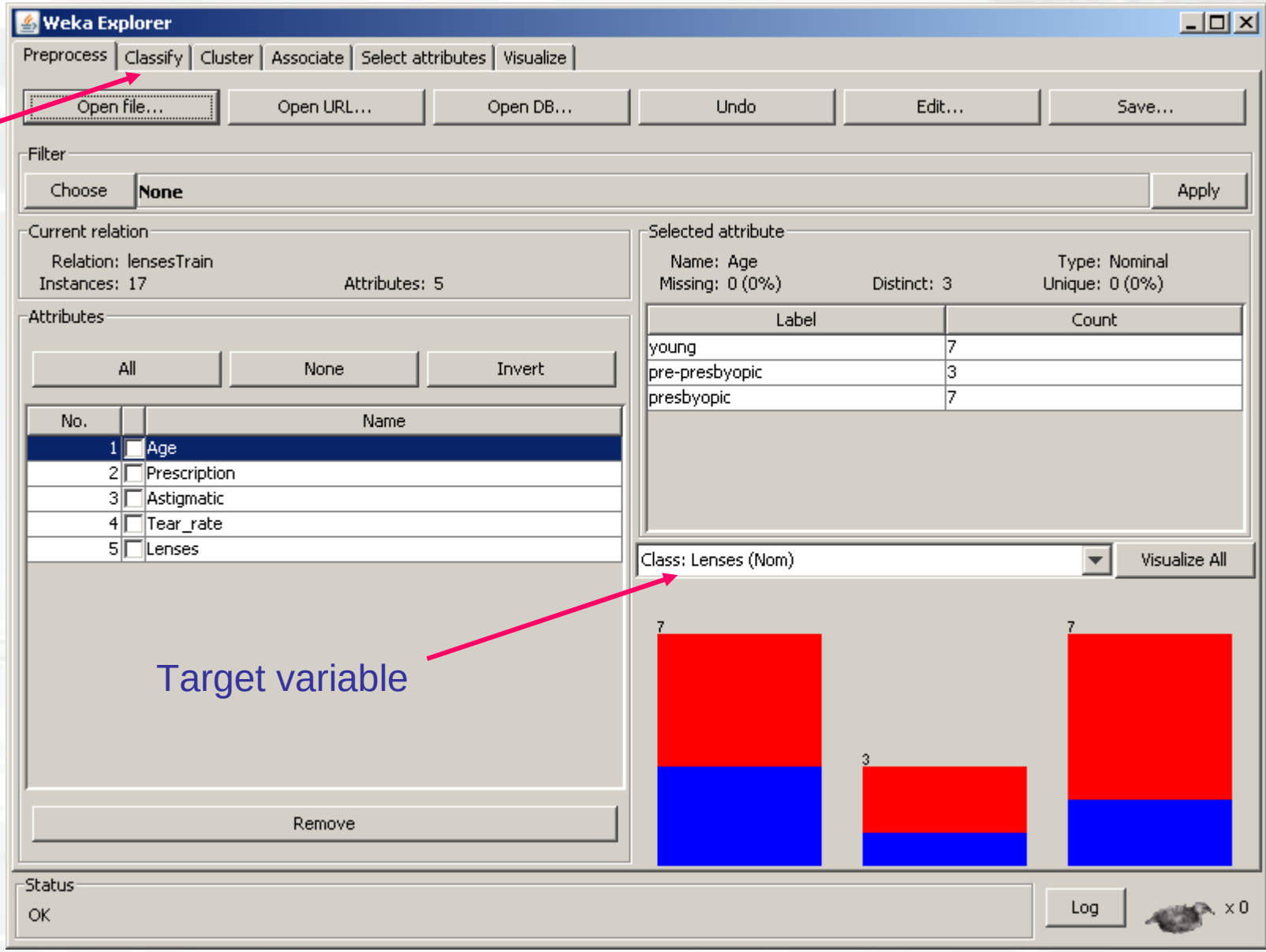


Petra.Kralj@ijs.si

The data are loaded

Choose

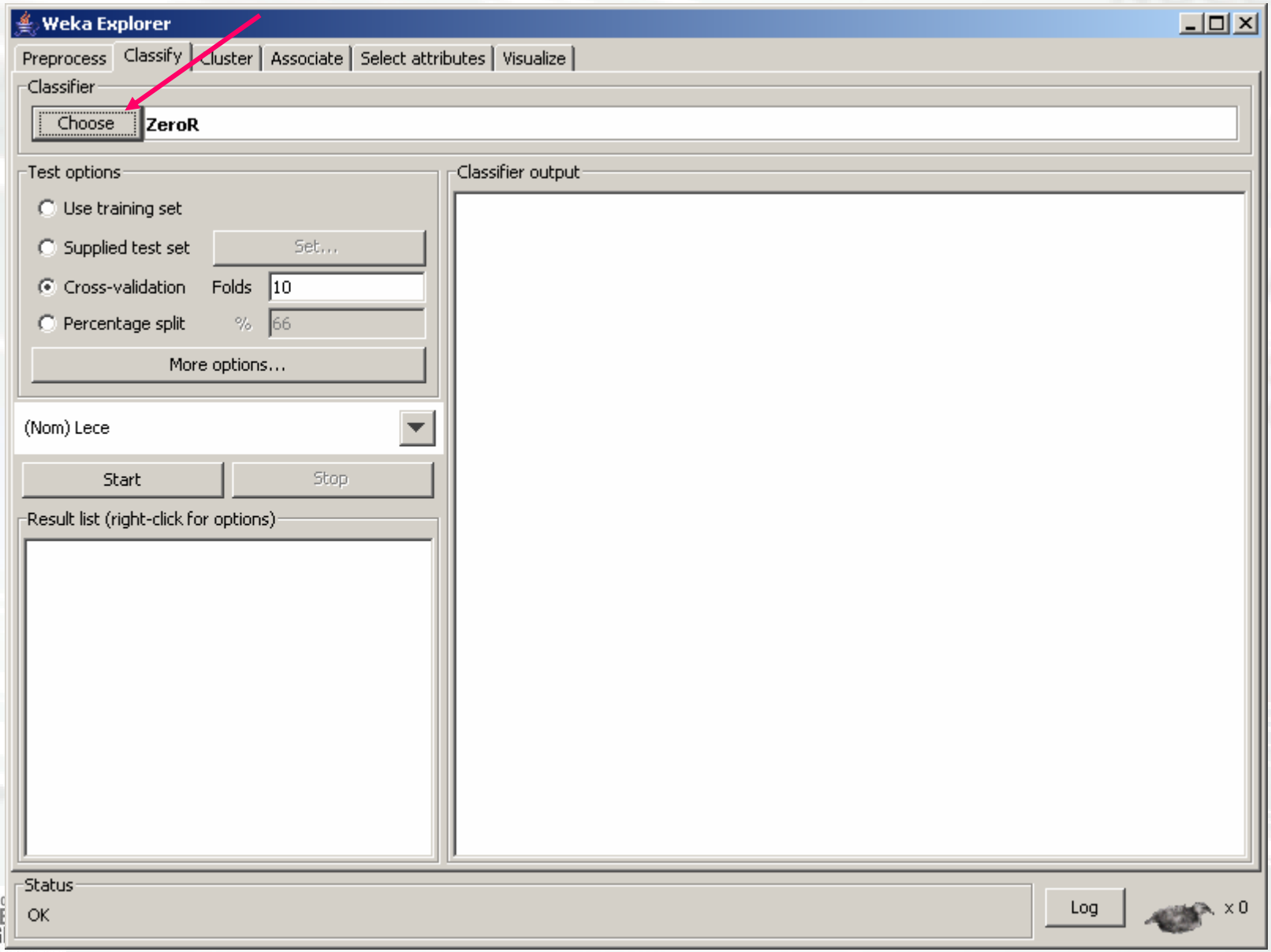
"Classify"



Target variable

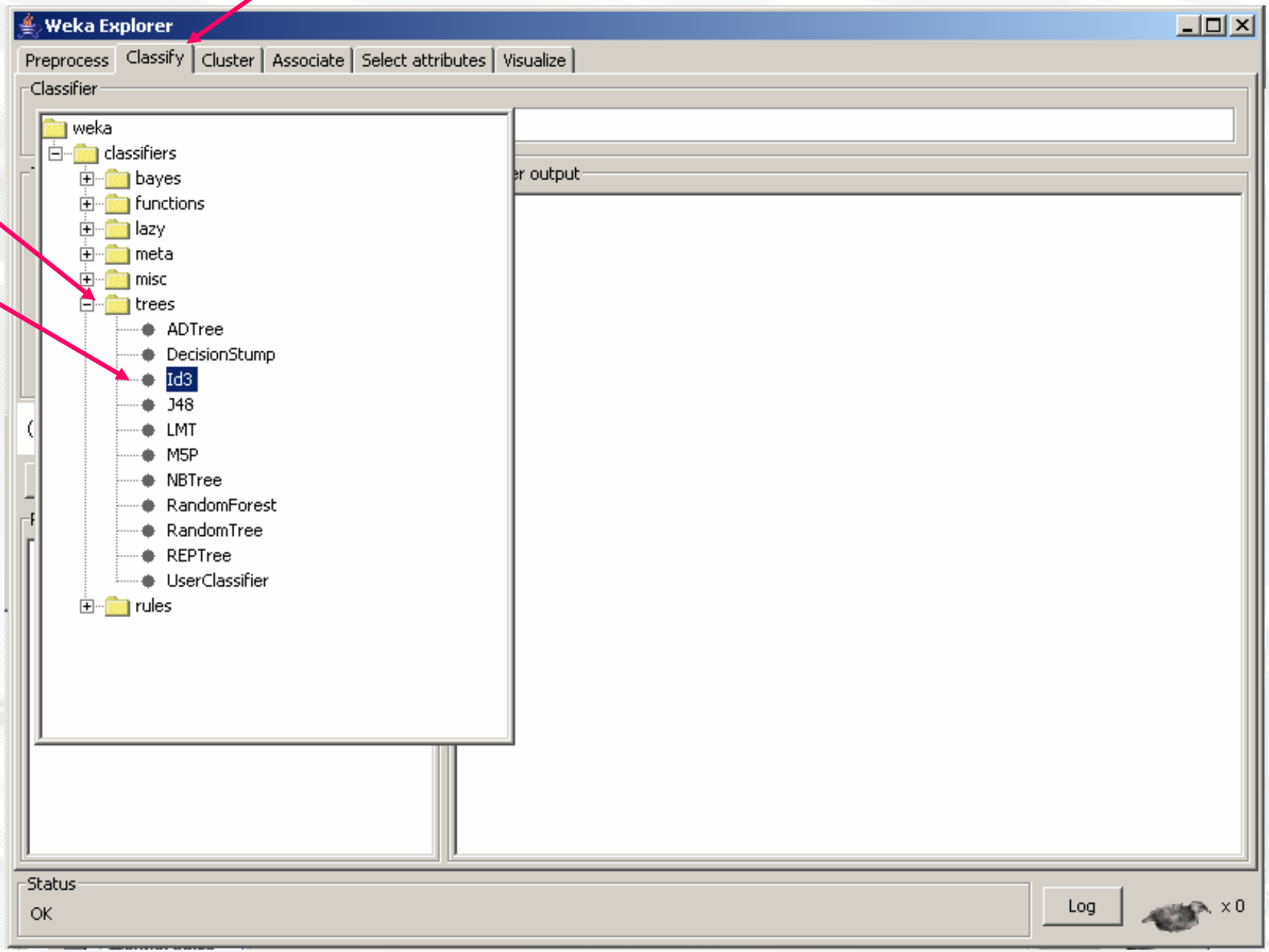
Petra.Kralj@ijs.si

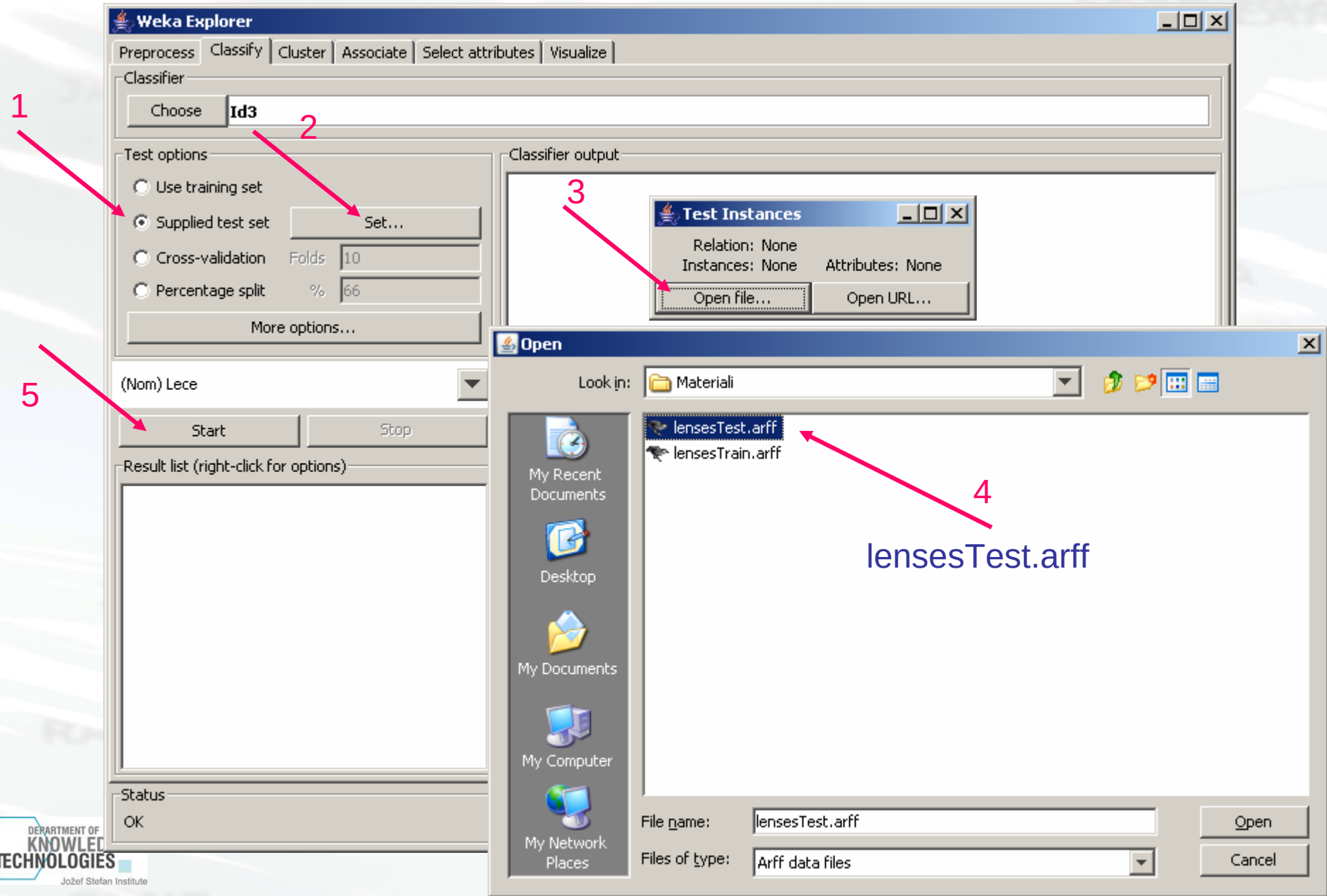
Choose algoritem

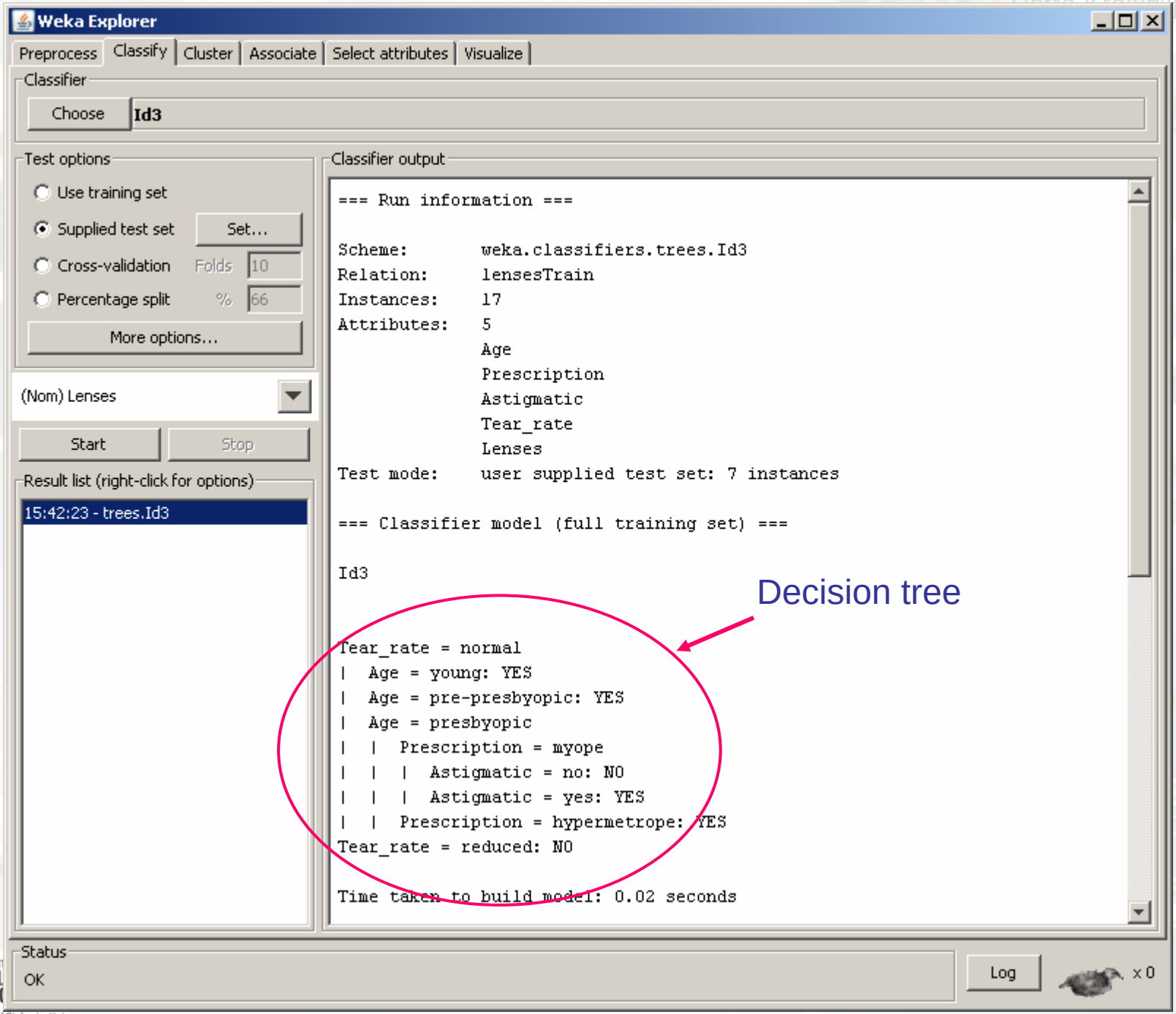


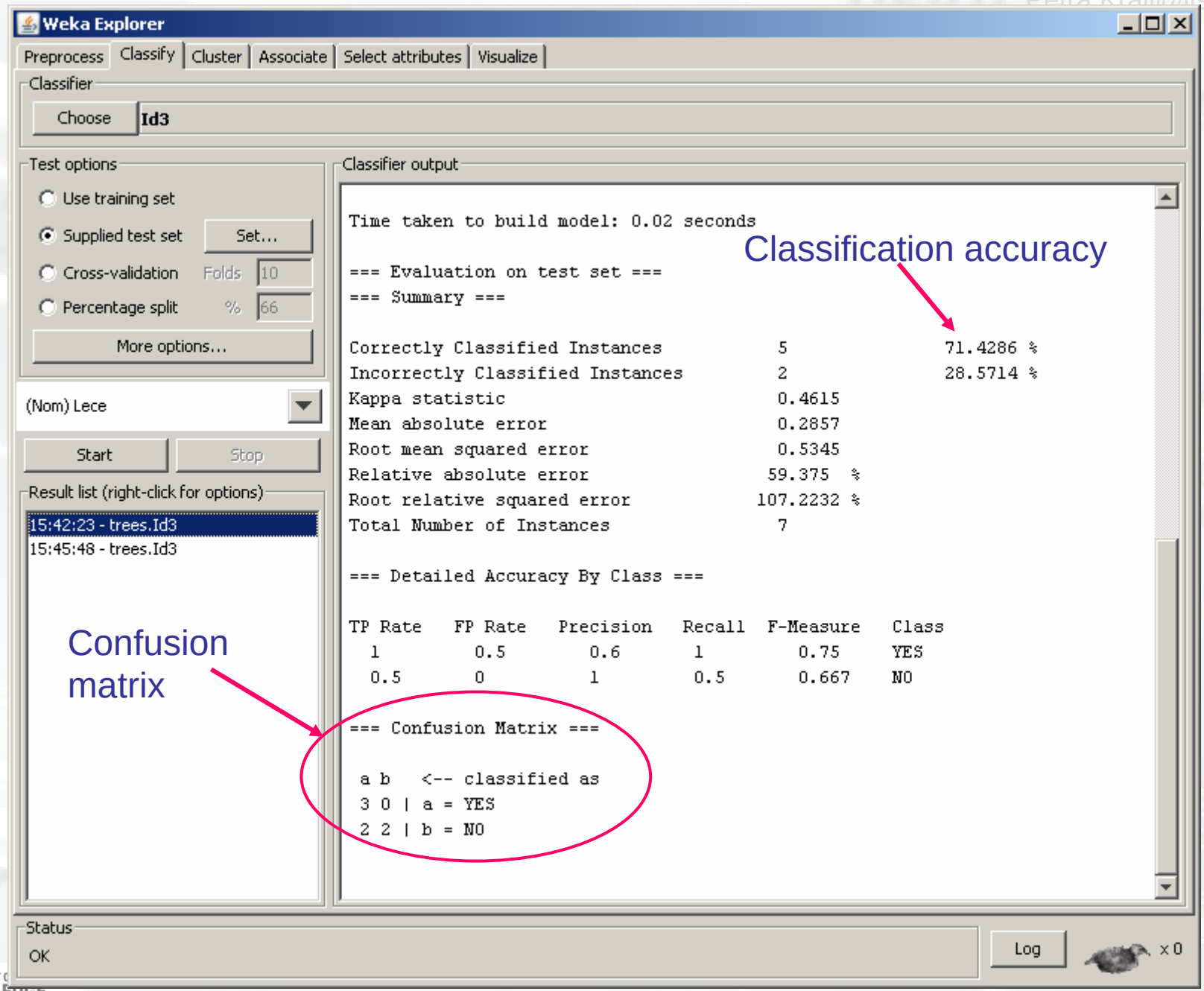
trees

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...
(Nom) Lece
5
Start
Stop
Result list (right-click for options)
Status
OK
Open
Look in:
Materiali
lensesTest.arff
lensesTrain.arff
My Recent Documents
Desktop
My Documents
My Computer
My Network Places
4
lensesTest.arff
File name:
lensesTest.arff
Files of type:
Arff data files
Open
Cancel





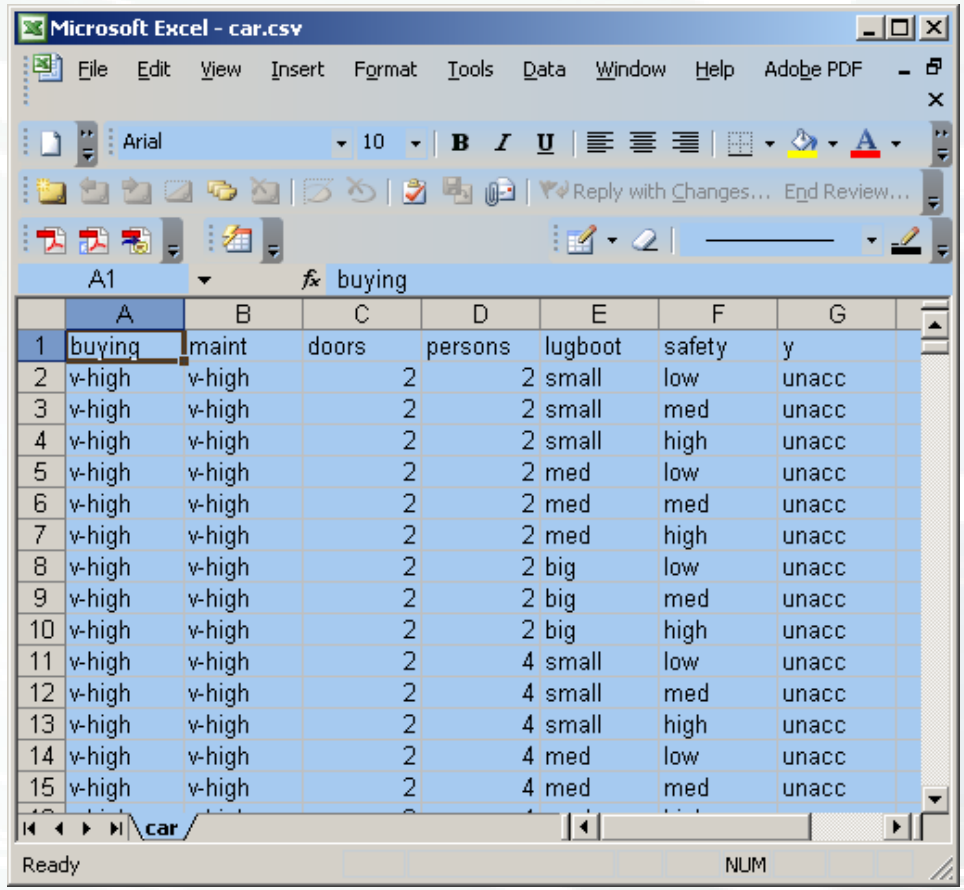
Exercise 2: CAR dataset

- 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



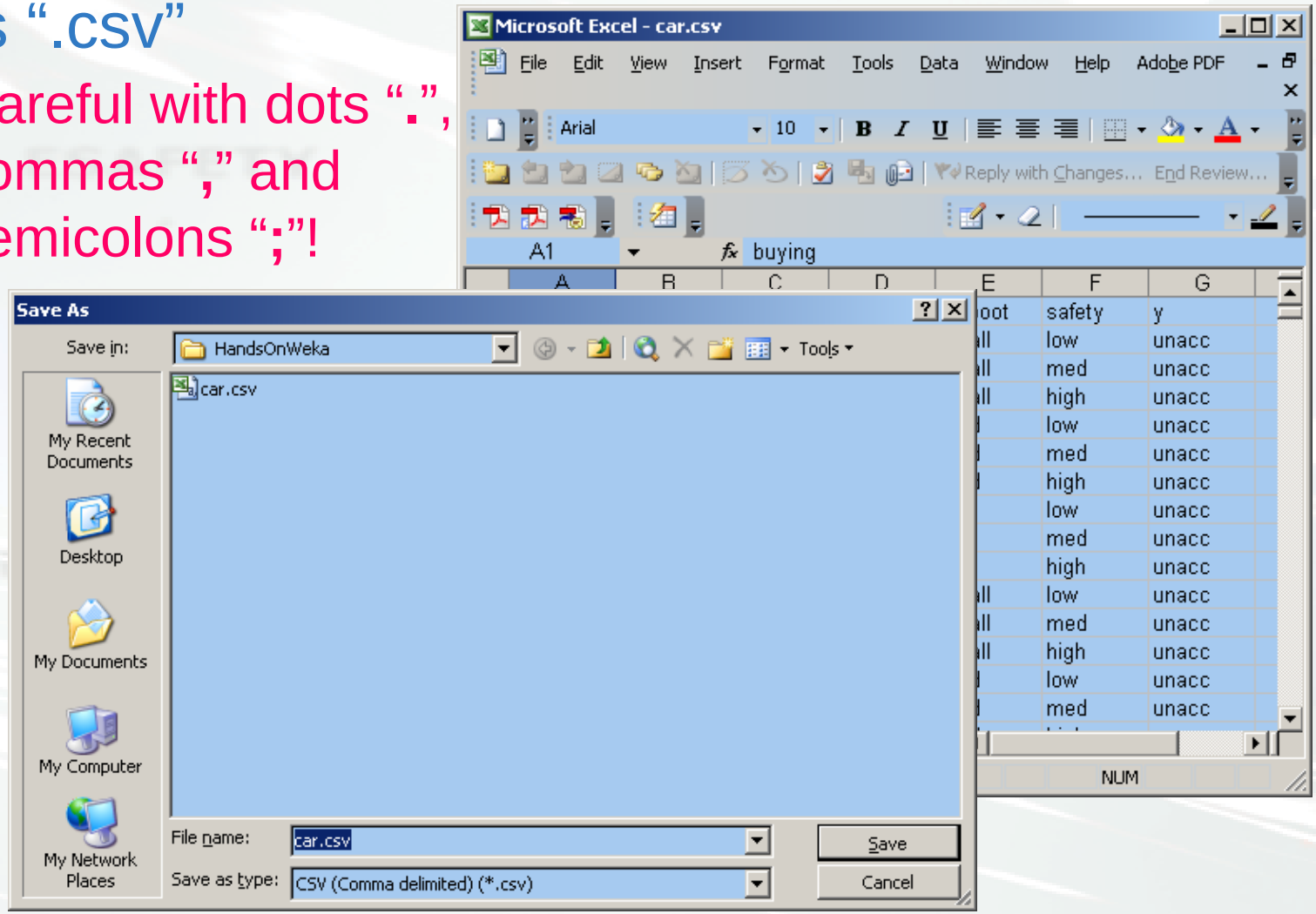
	A	B	C	D	E	F	G
1	buying	maint	doors	persons	lugboot	safety	y
2	v-high	v-high	2	2	small	low	unacc
3	v-high	v-high	2	2	small	med	unacc
4	v-high	v-high	2	2	small	high	unacc
5	v-high	v-high	2	2	med	low	unacc
6	v-high	v-high	2	2	med	med	unacc
7	v-high	v-high	2	2	med	high	unacc
8	v-high	v-high	2	2	big	low	unacc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

Petra.Kralj@ijs.si

Preparing the data for WEKA - 2

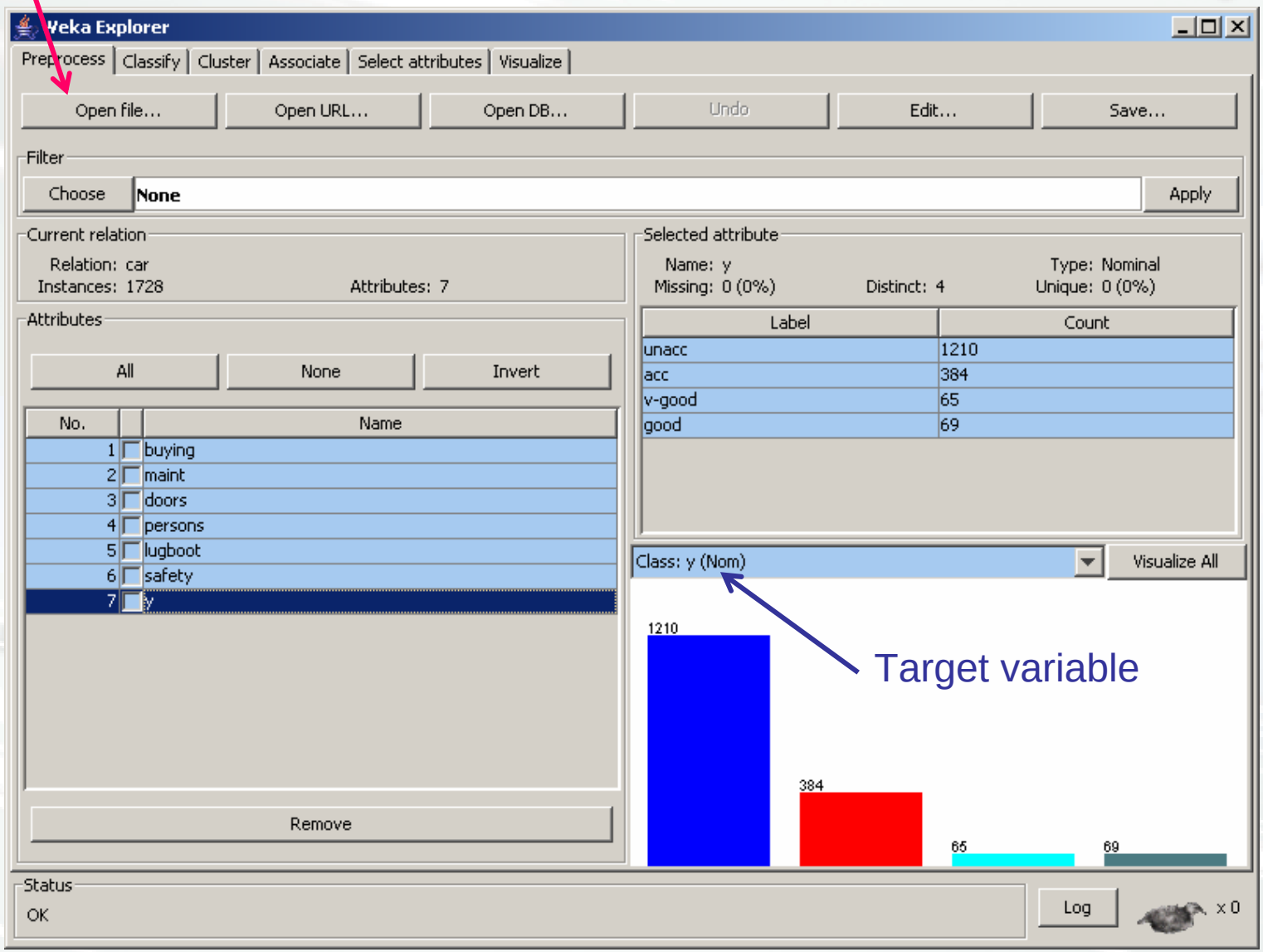
Save as ".csv"

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

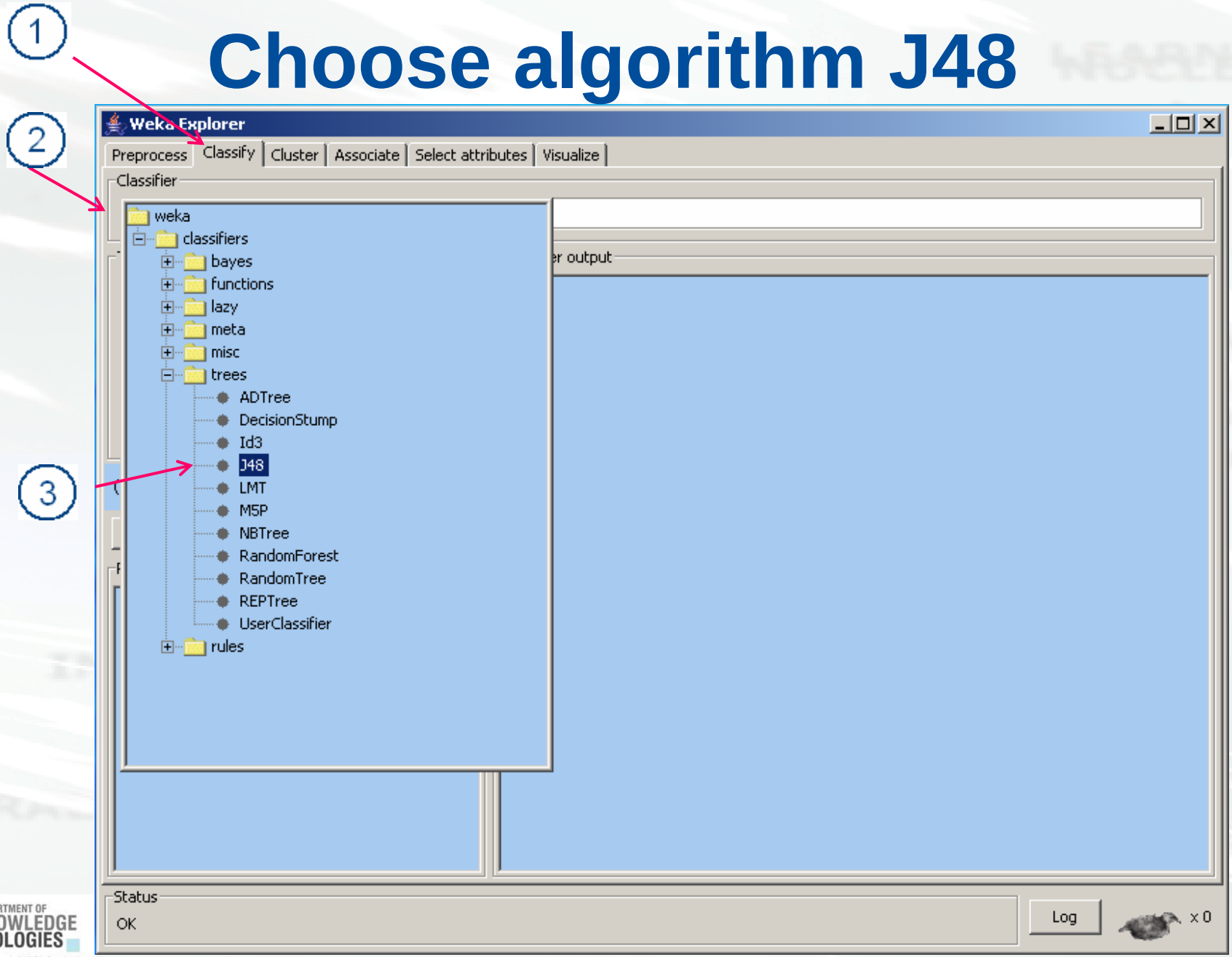


Car.csv

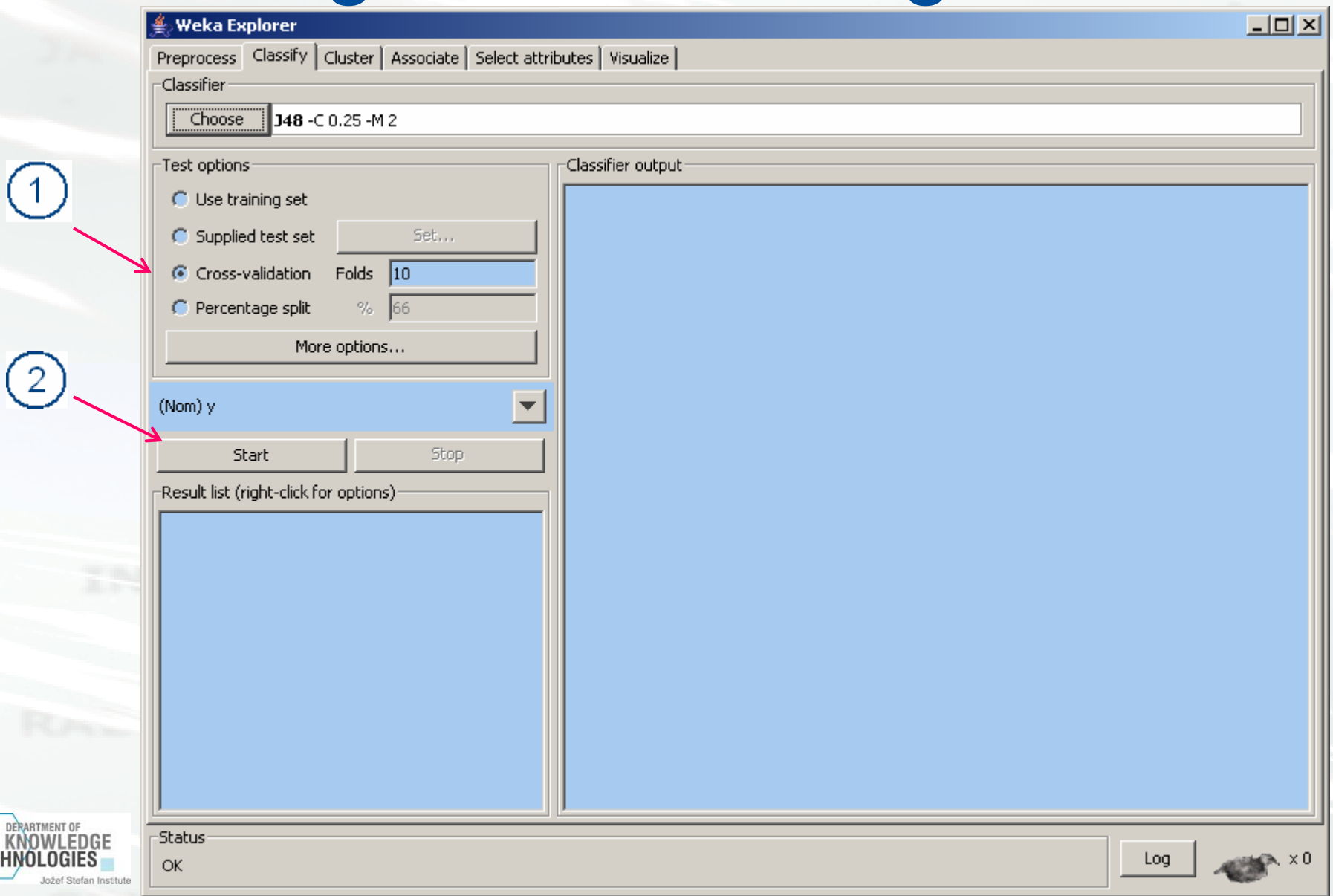
Load the data

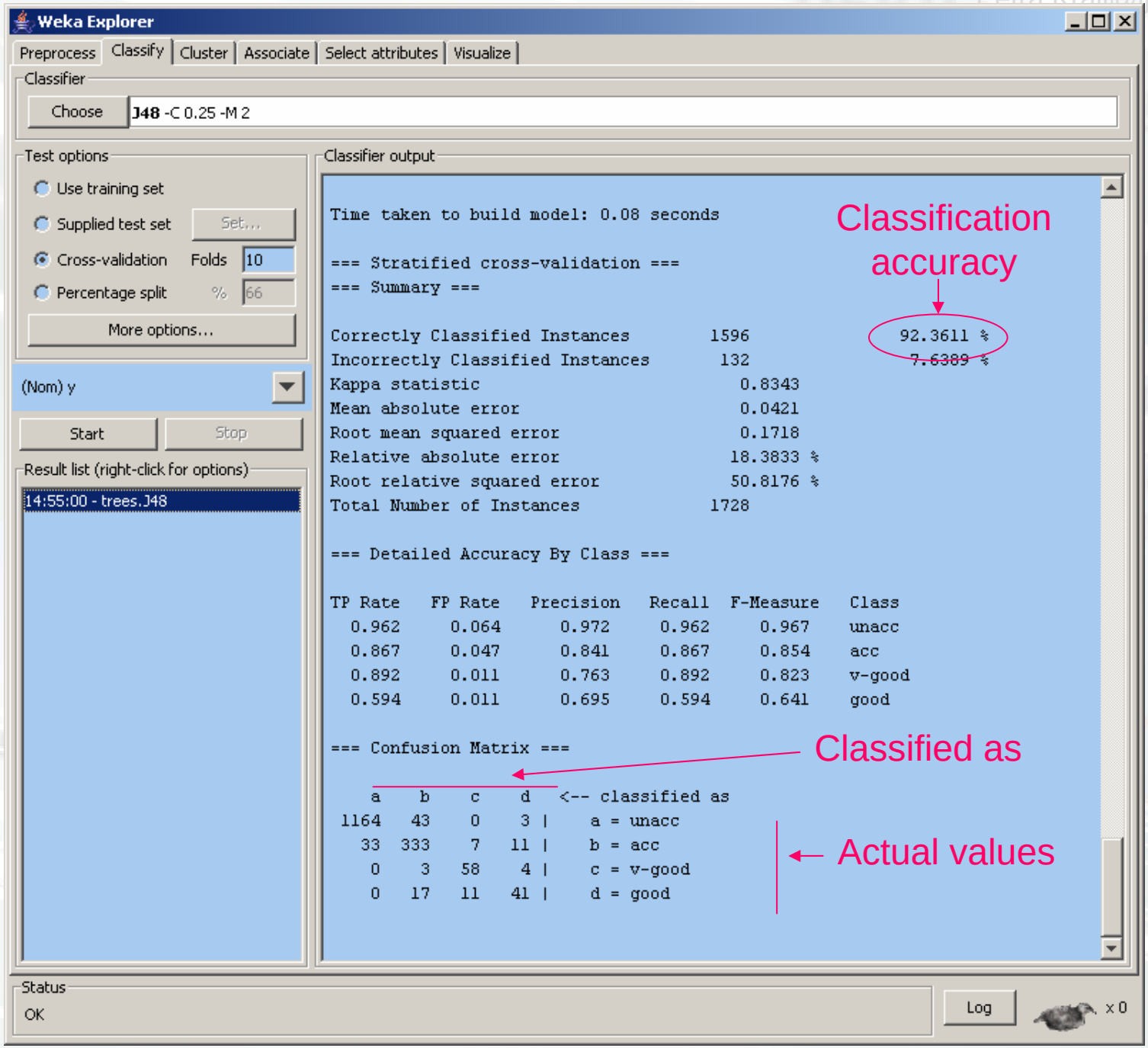


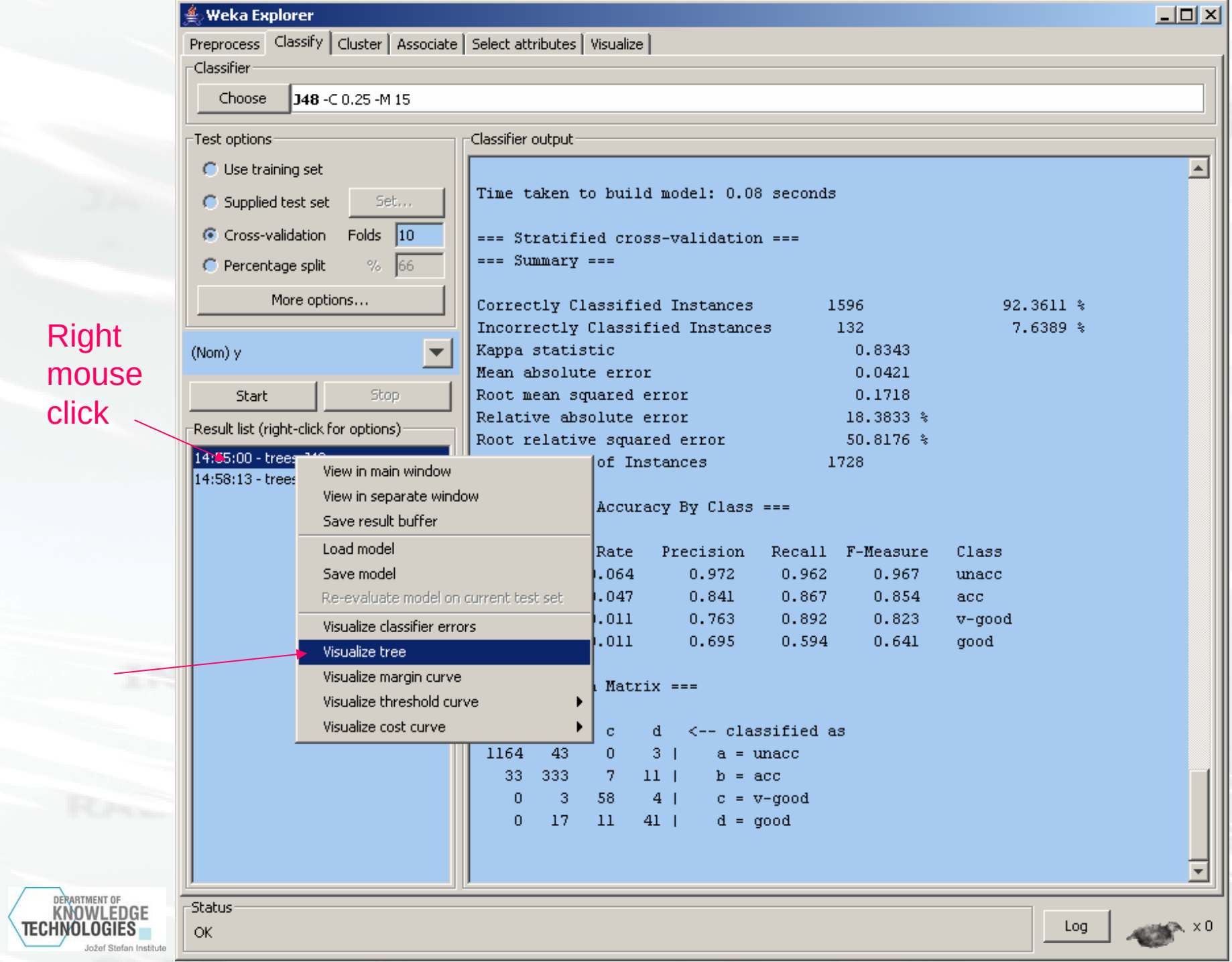
Choose algorithm J48



Building and evaluating the tree







Right mouse click

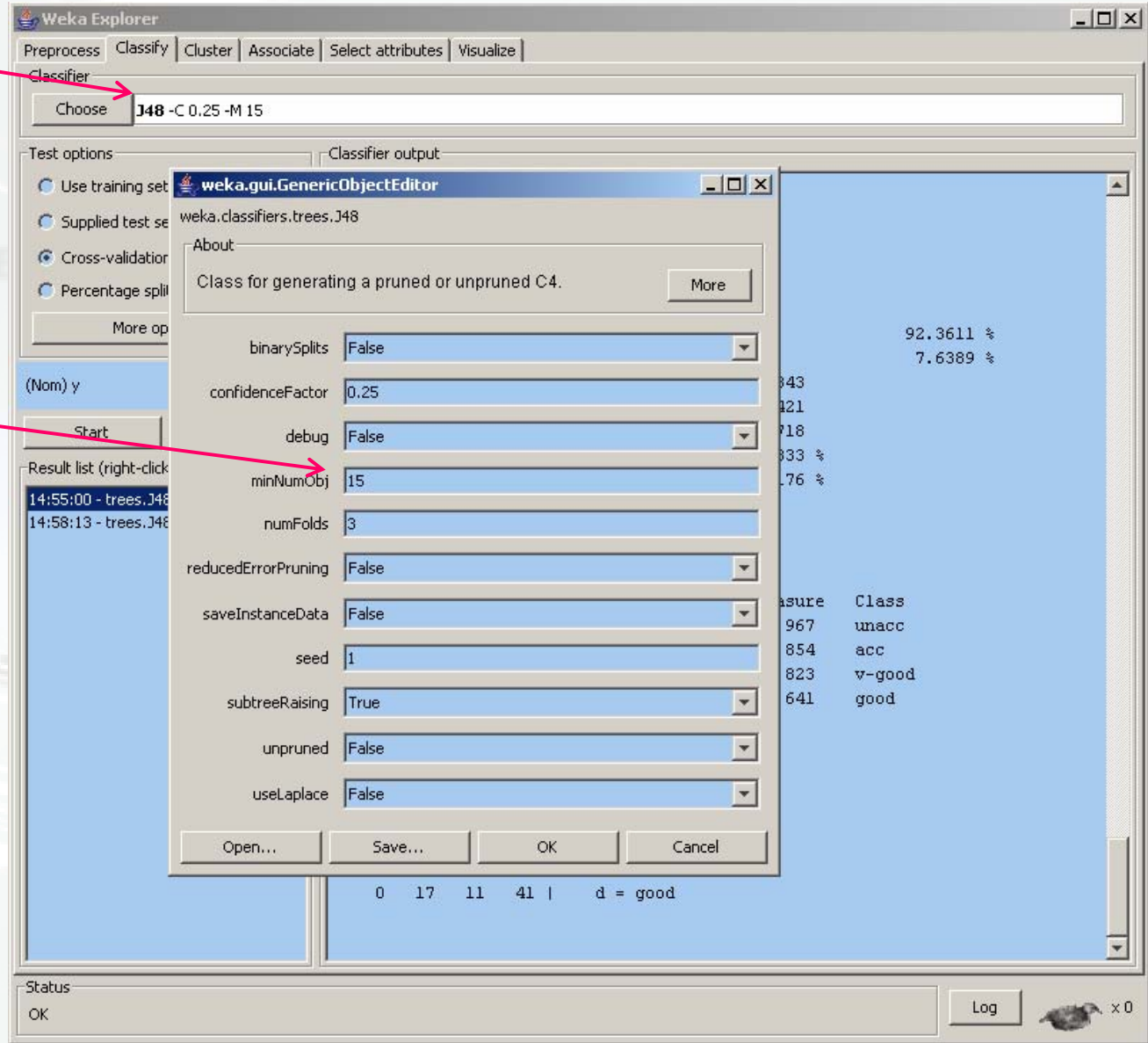
Tree pruning

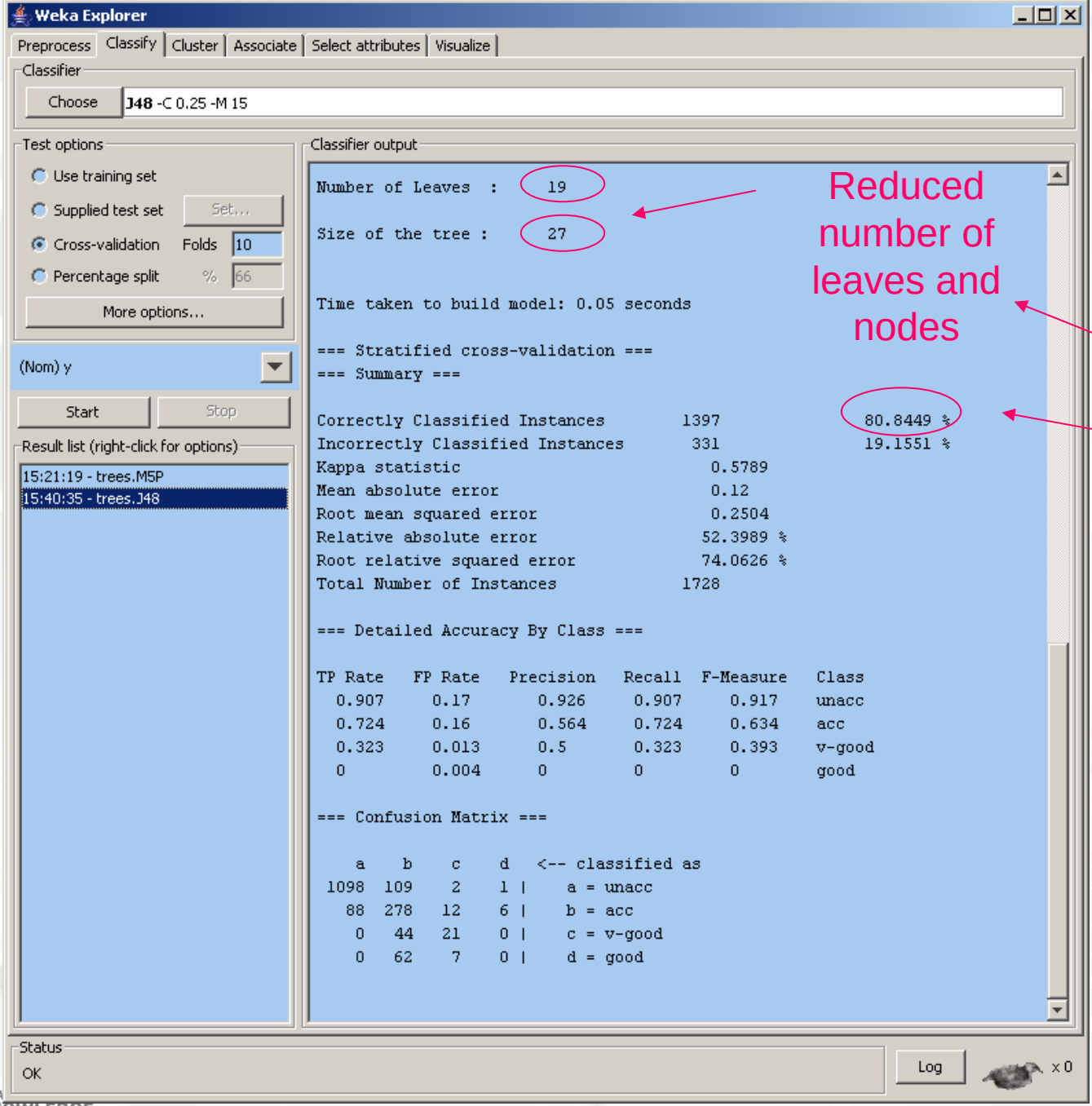
1

Parameters of the algorithm (right mouse click)

2

Set the minimal number of objects per leaf to 15

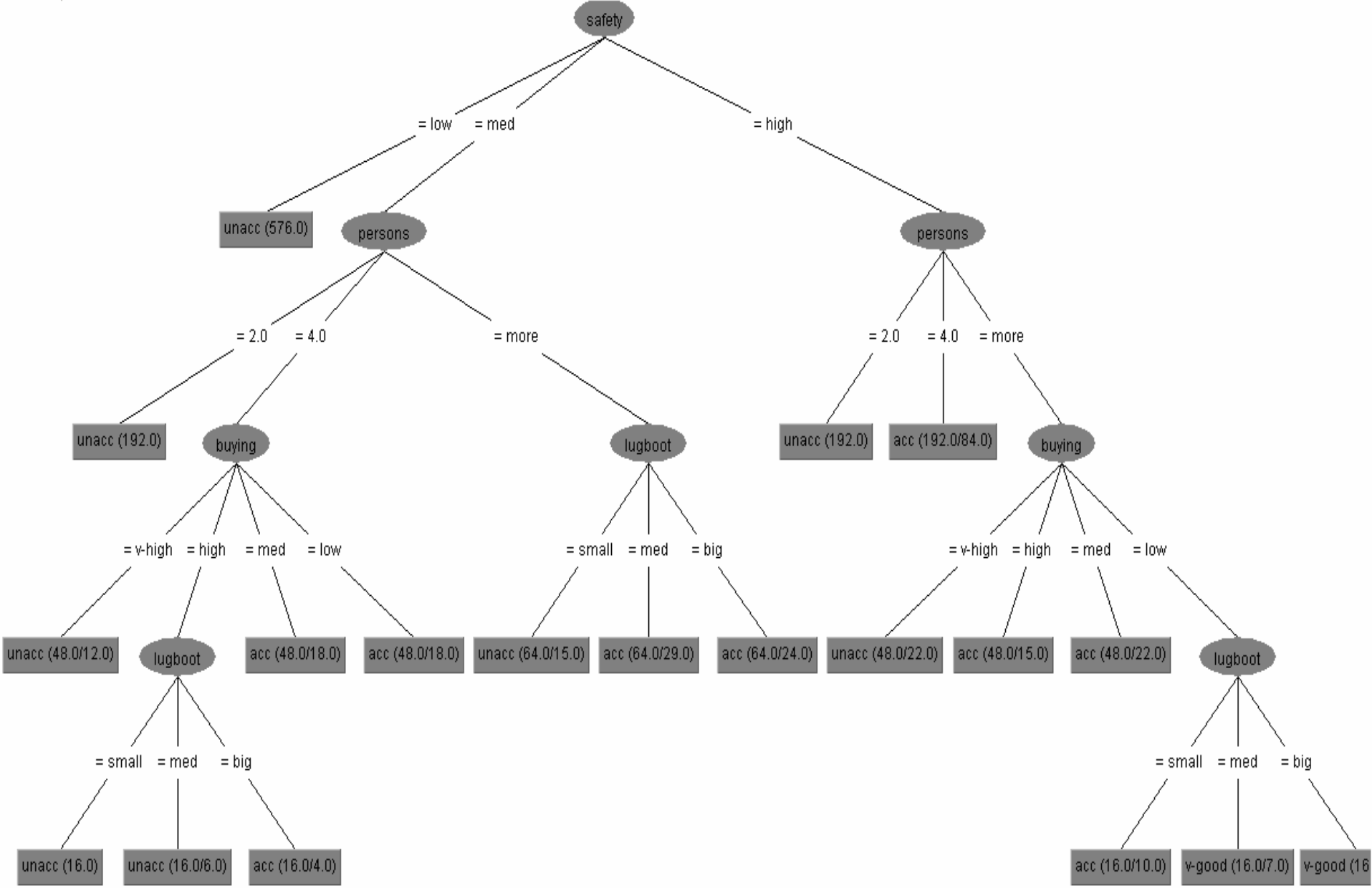




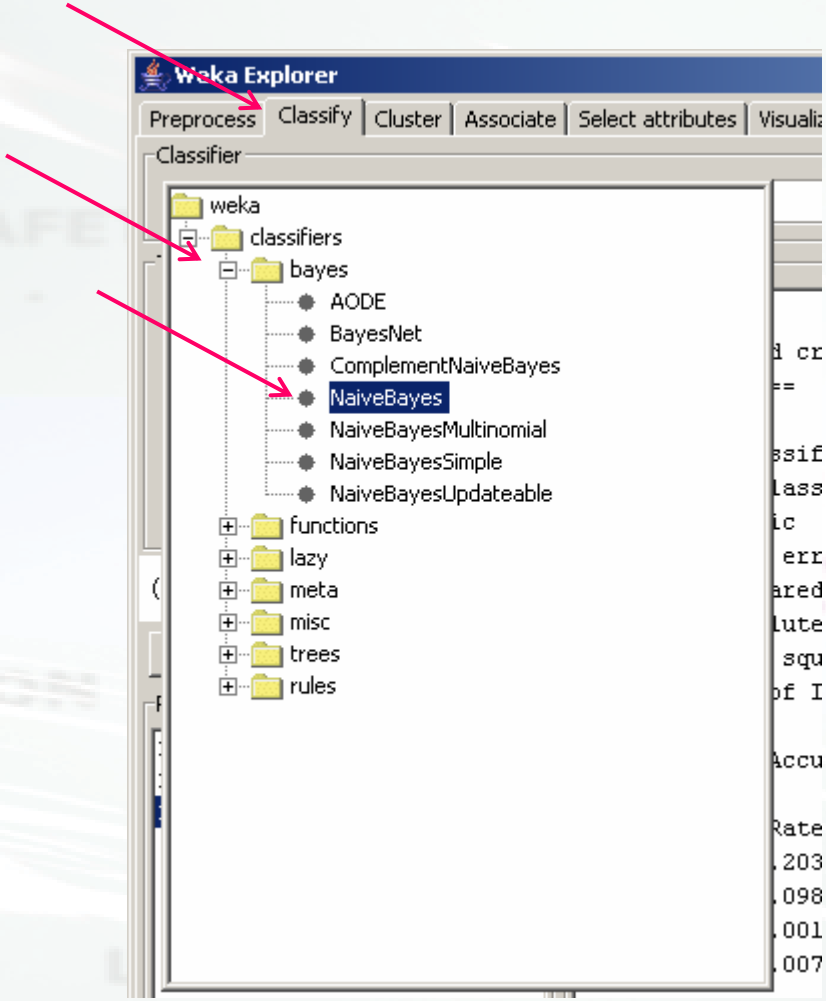
Reduced number of leaves and nodes

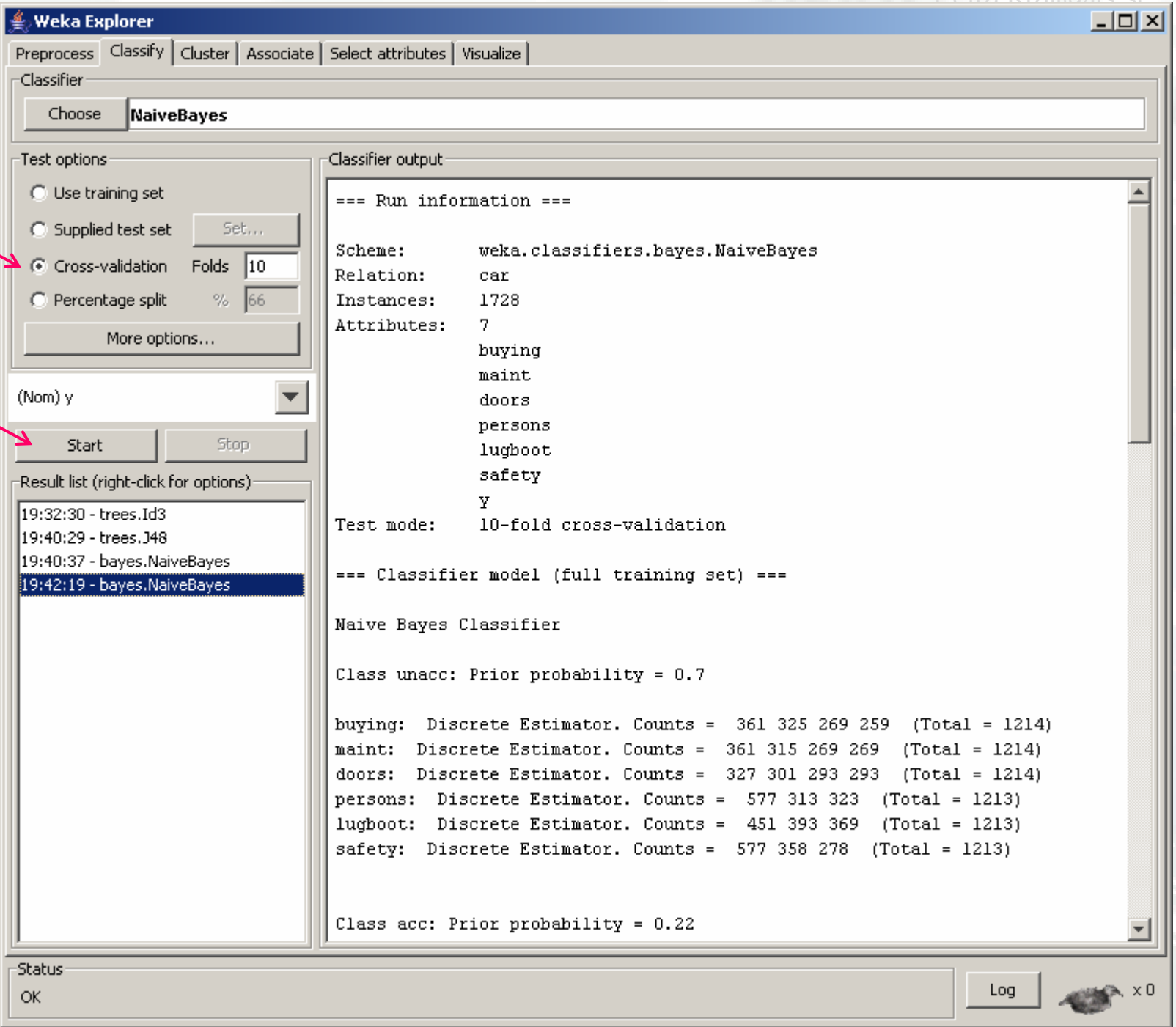
Easier to interpret

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