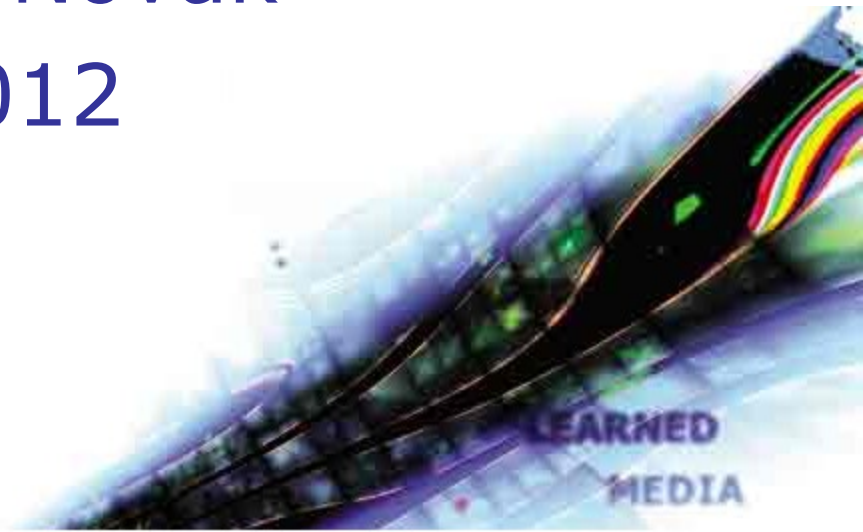
Hands on Weka: Part II

Petra Kralj Novak

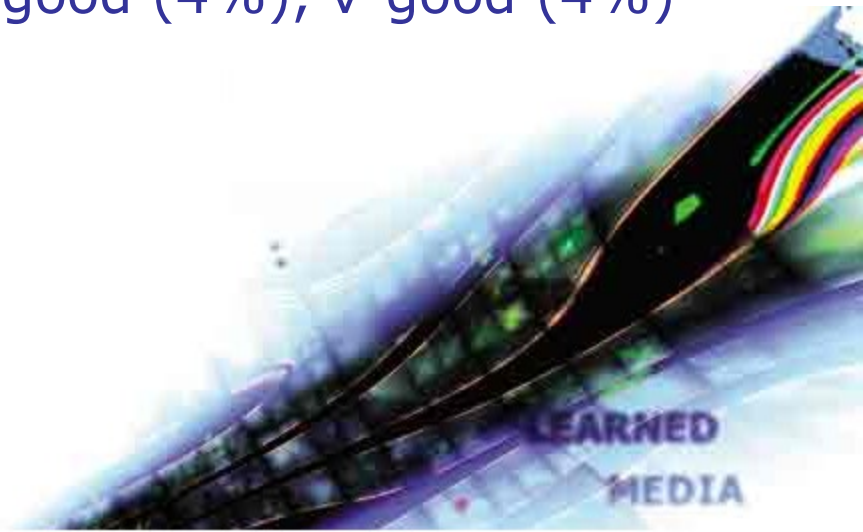
4.12.2012



DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Exercise 2: CAR dataset

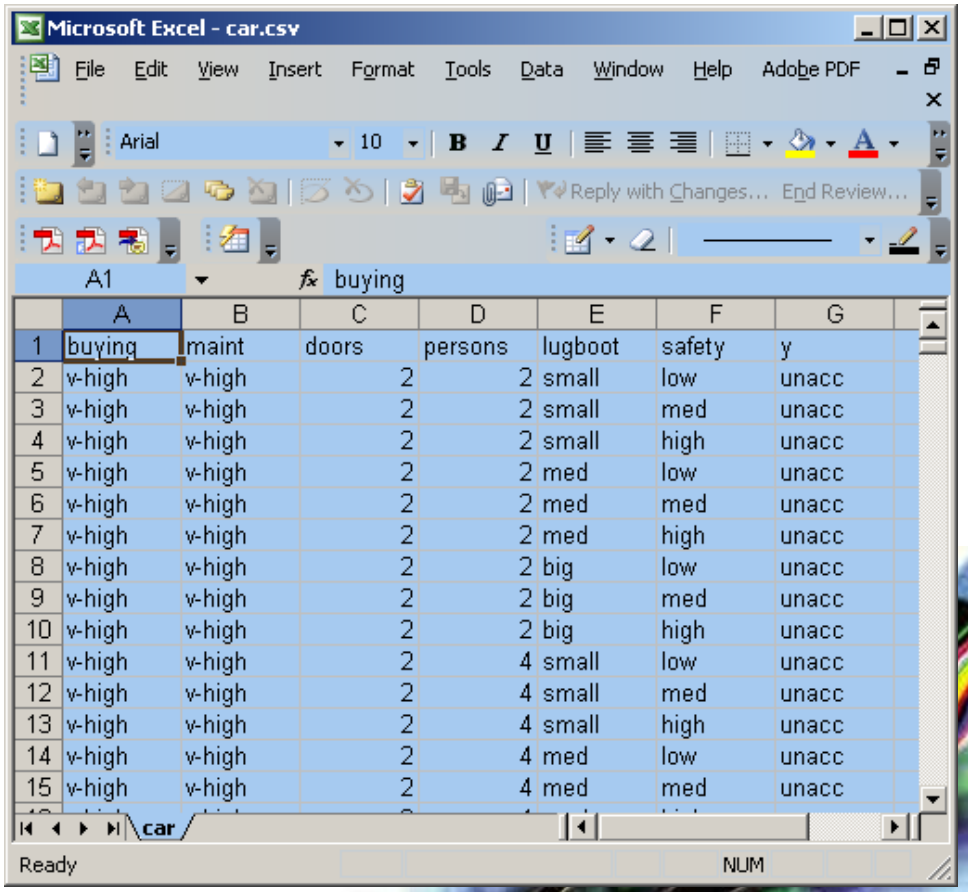
- 1728 examples
- 6 attributes
  - 6 nominal
  - 0 numeric
- Nominal target variable
  - 4 classes: unacc, acc, good, v-good
  - Distribution of classes
    - unacc (70%), acc (22%), good (4%), v-good (4%)
- No missing values



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

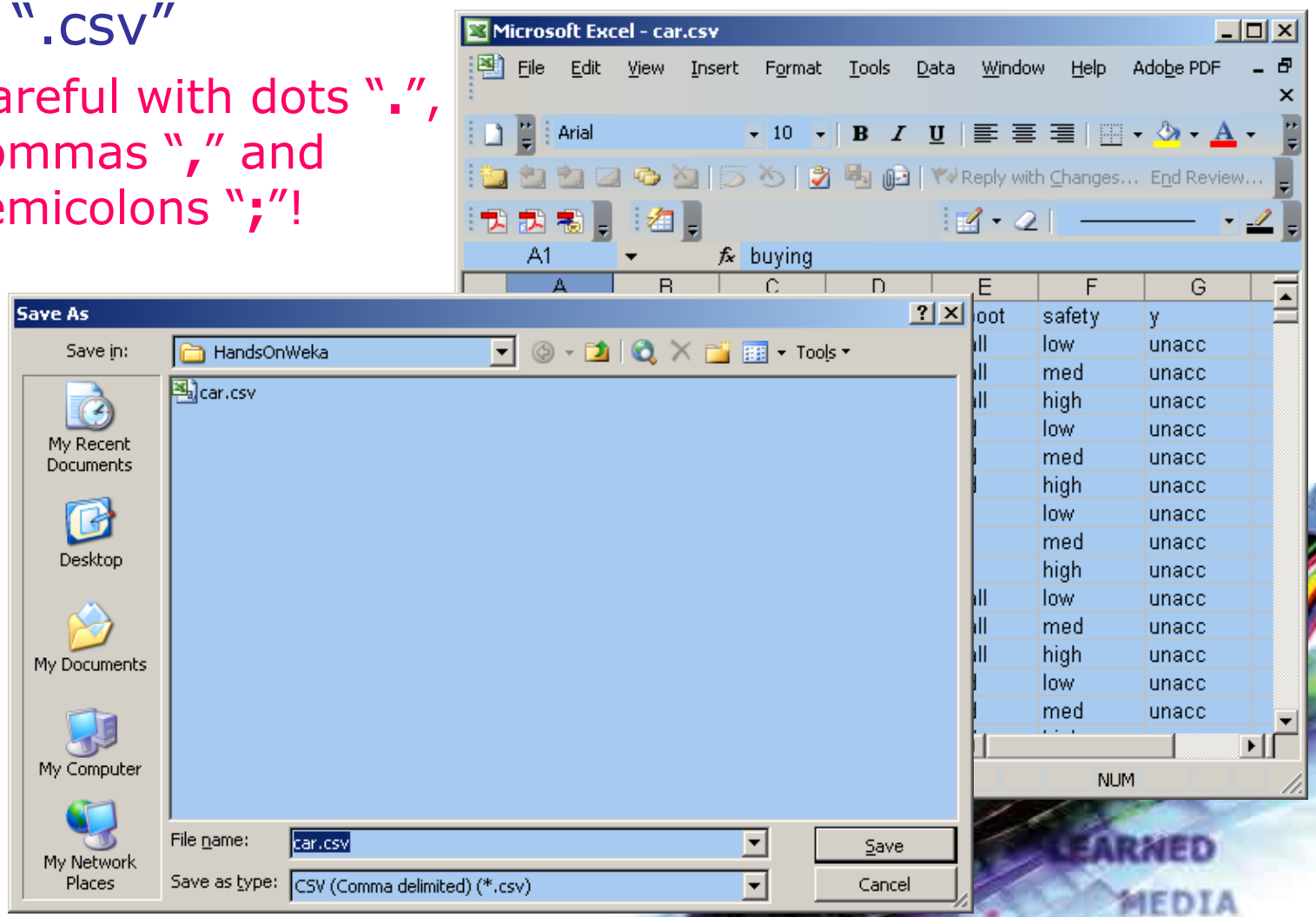


	A	B	C	D	E	F	G
1	buying	maint	doors	persons	lugboot	safety	y
2	v-high	v-high	2	2	small	low	unacc
3	v-high	v-high	2	2	small	med	unacc
4	v-high	v-high	2	2	small	high	unacc
5	v-high	v-high	2	2	med	low	unacc
6	v-high	v-high	2	2	med	med	unacc
7	v-high	v-high	2	2	med	high	unacc
8	v-high	v-high	2	2	big	low	unacc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

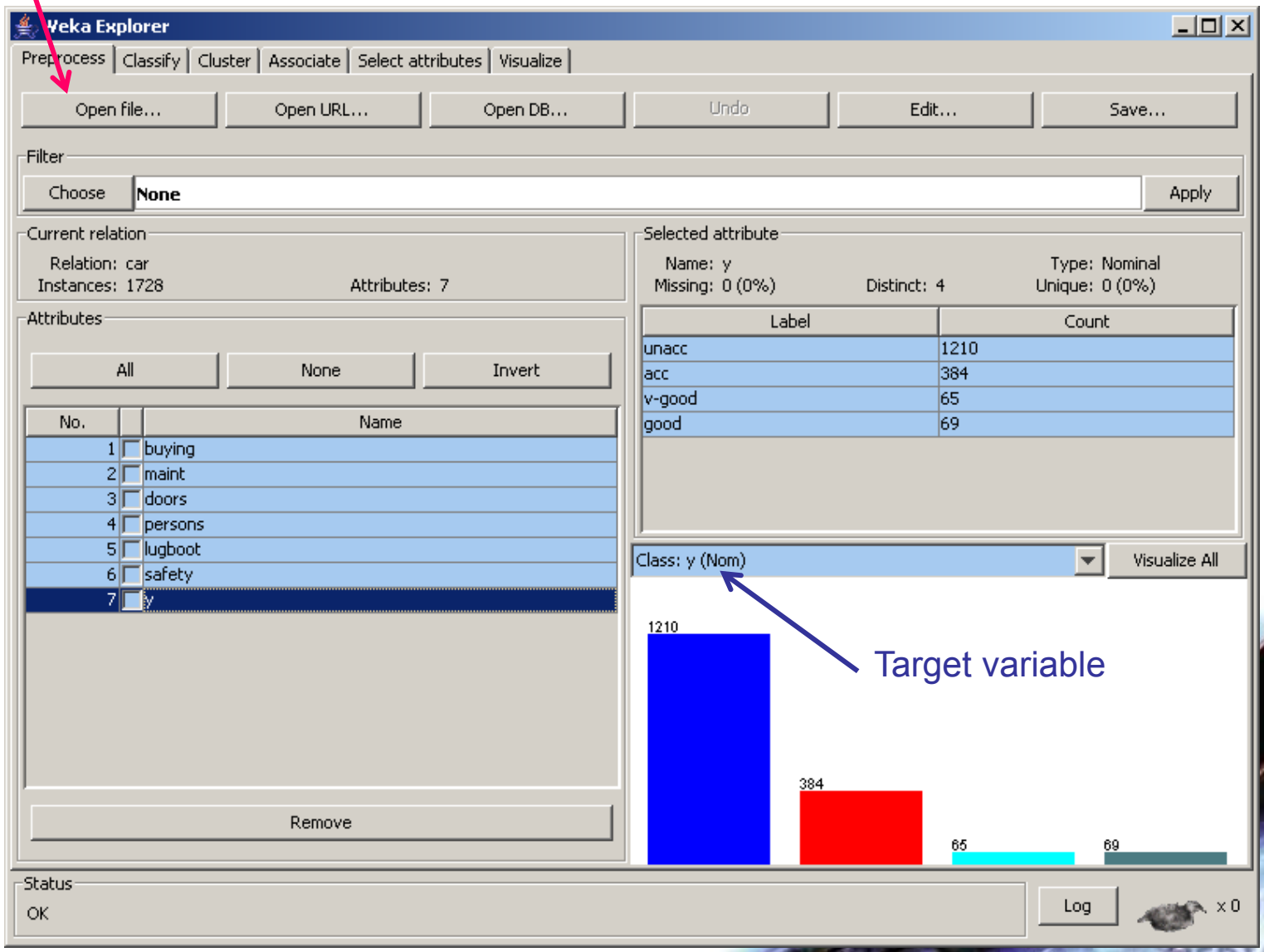
Save as ".csv"

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

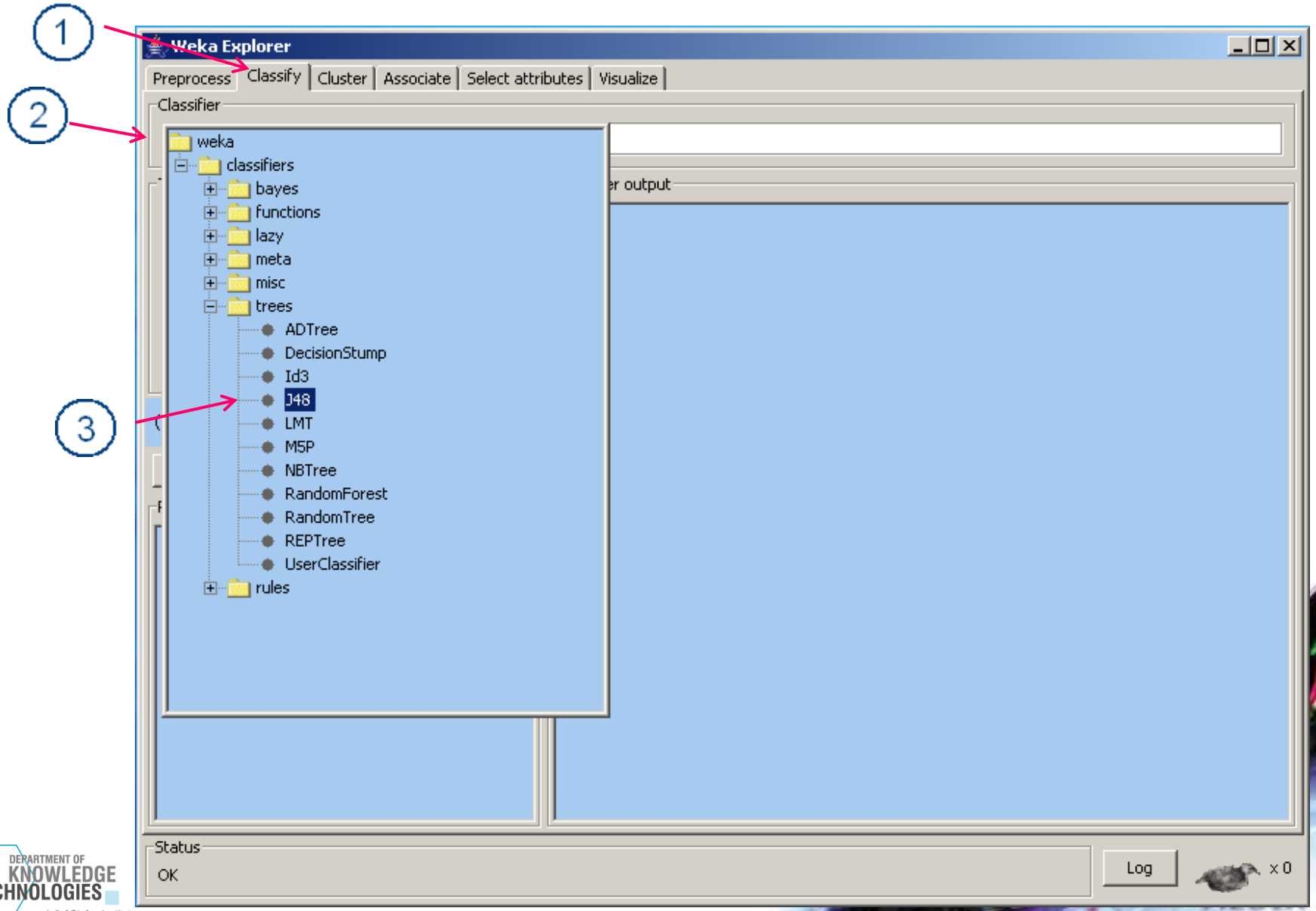


Load the data

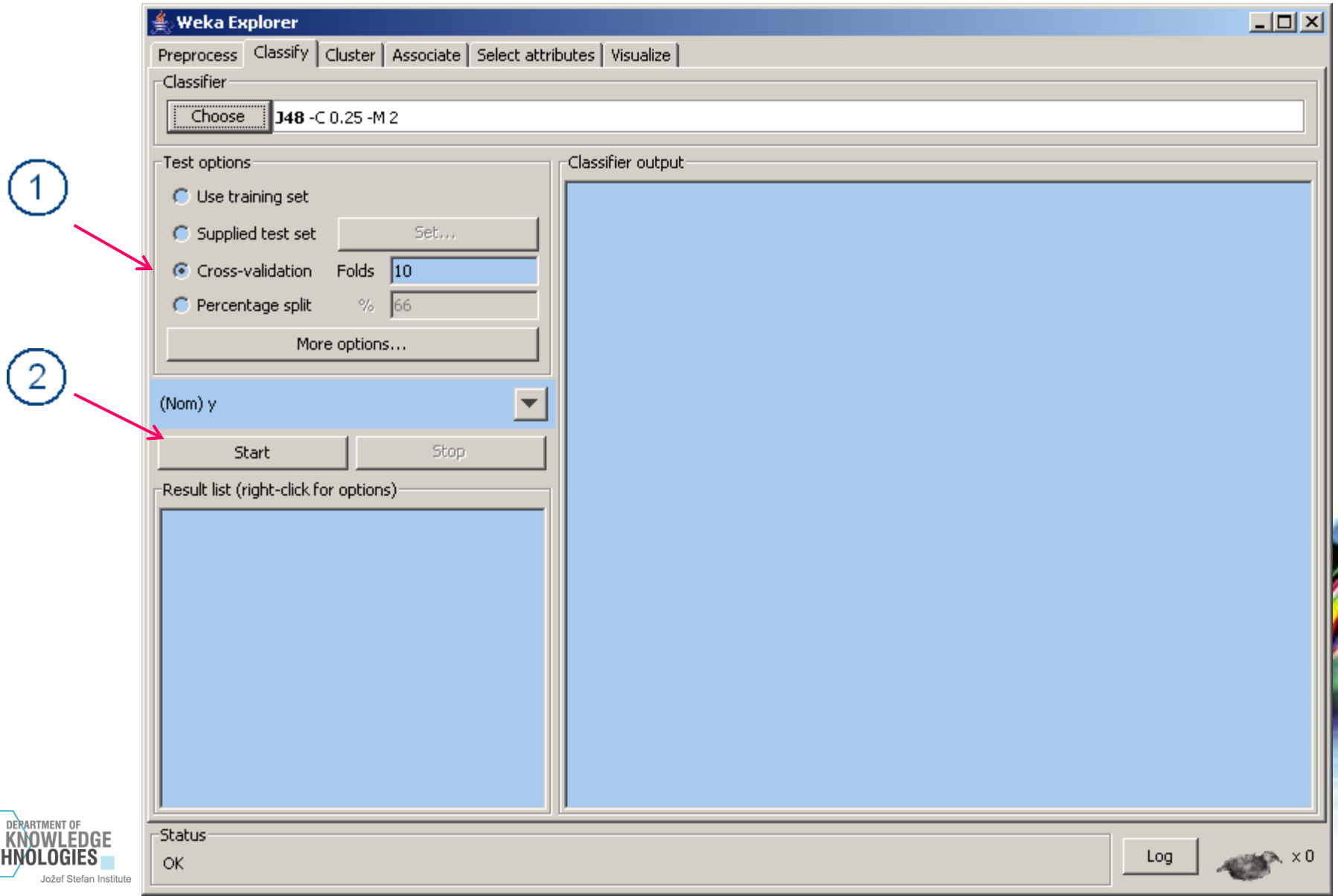
Car.csv

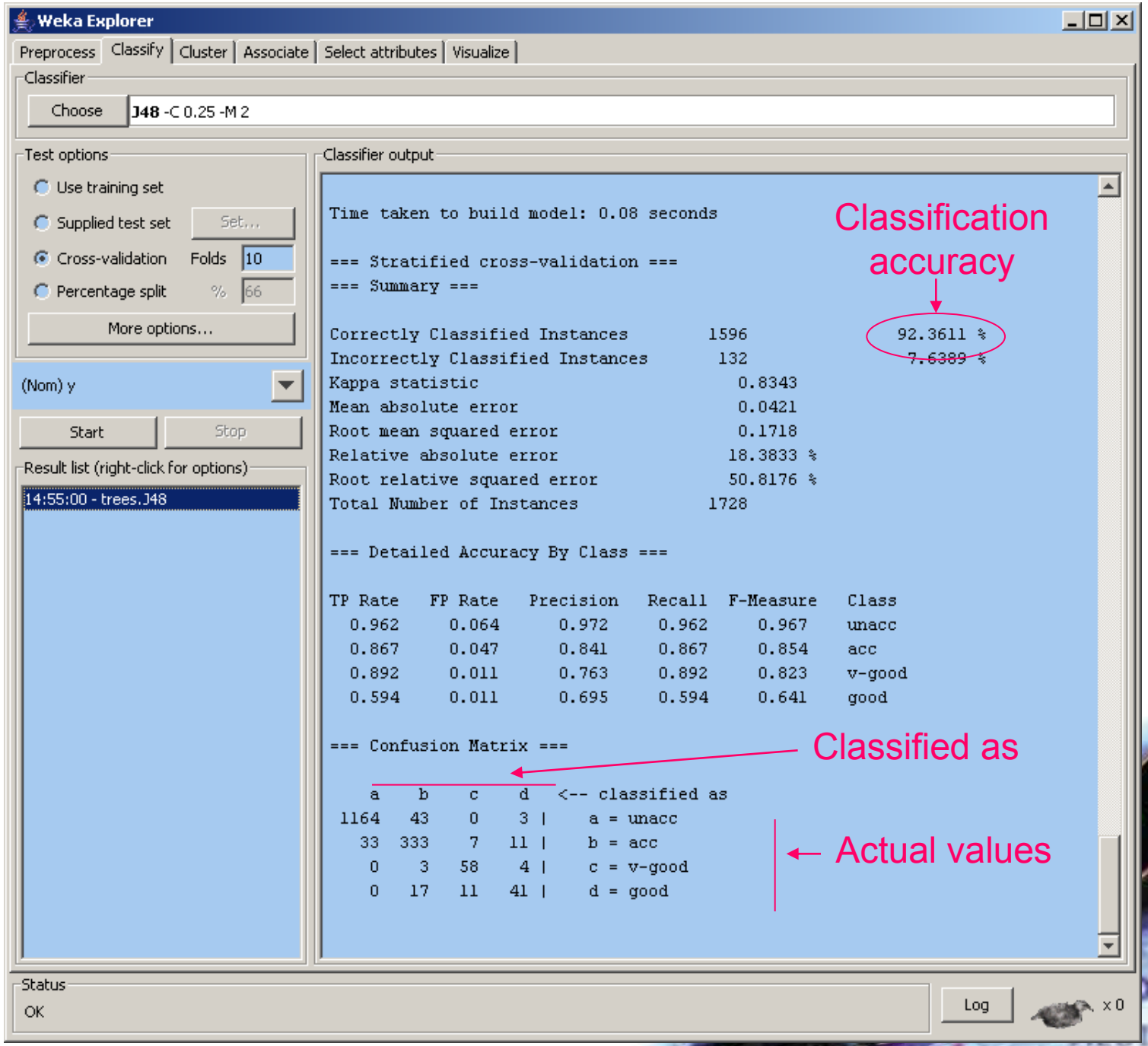


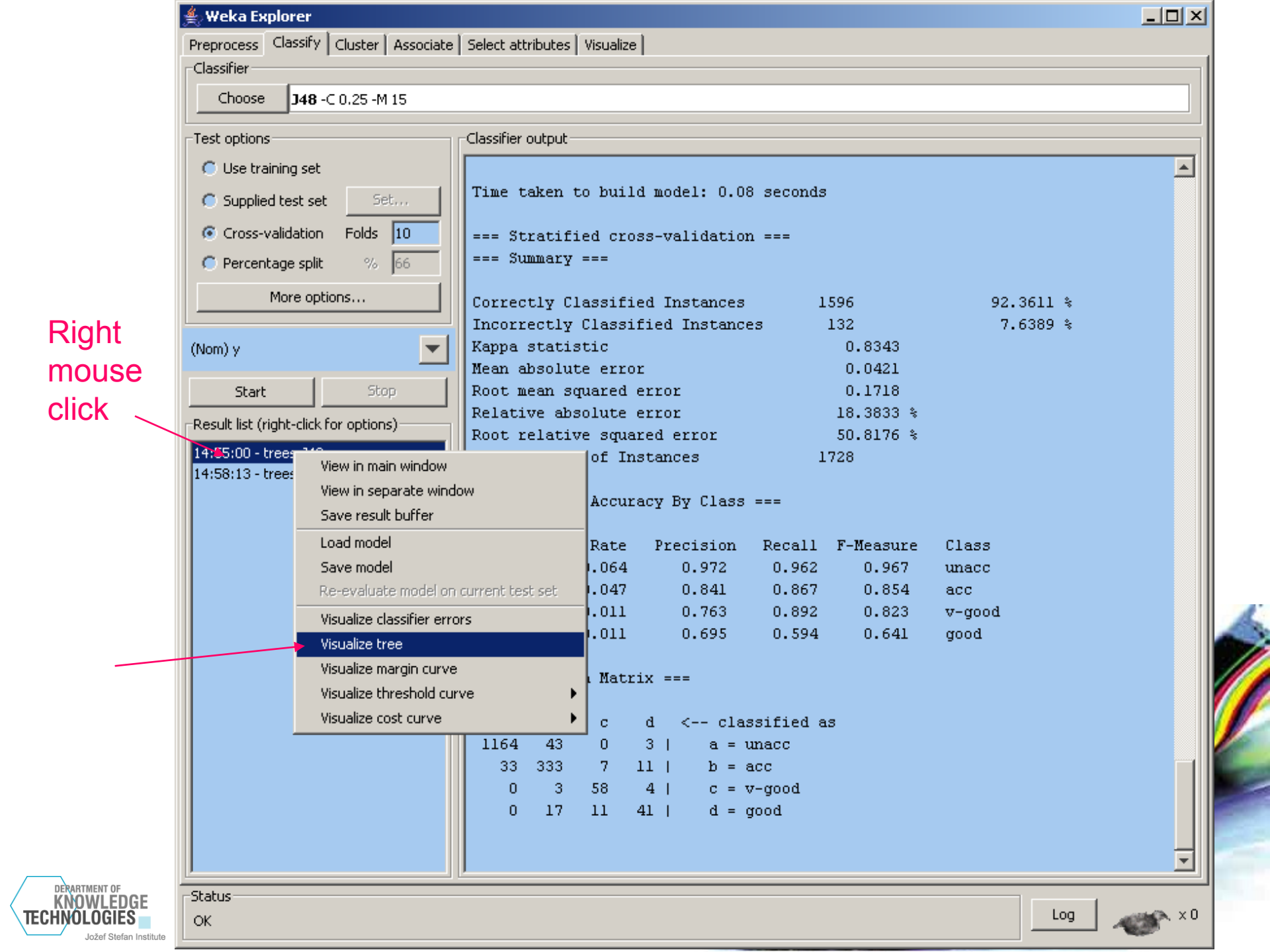
Choose algorithm J48



Building and evaluating the tree







Right mouse click

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

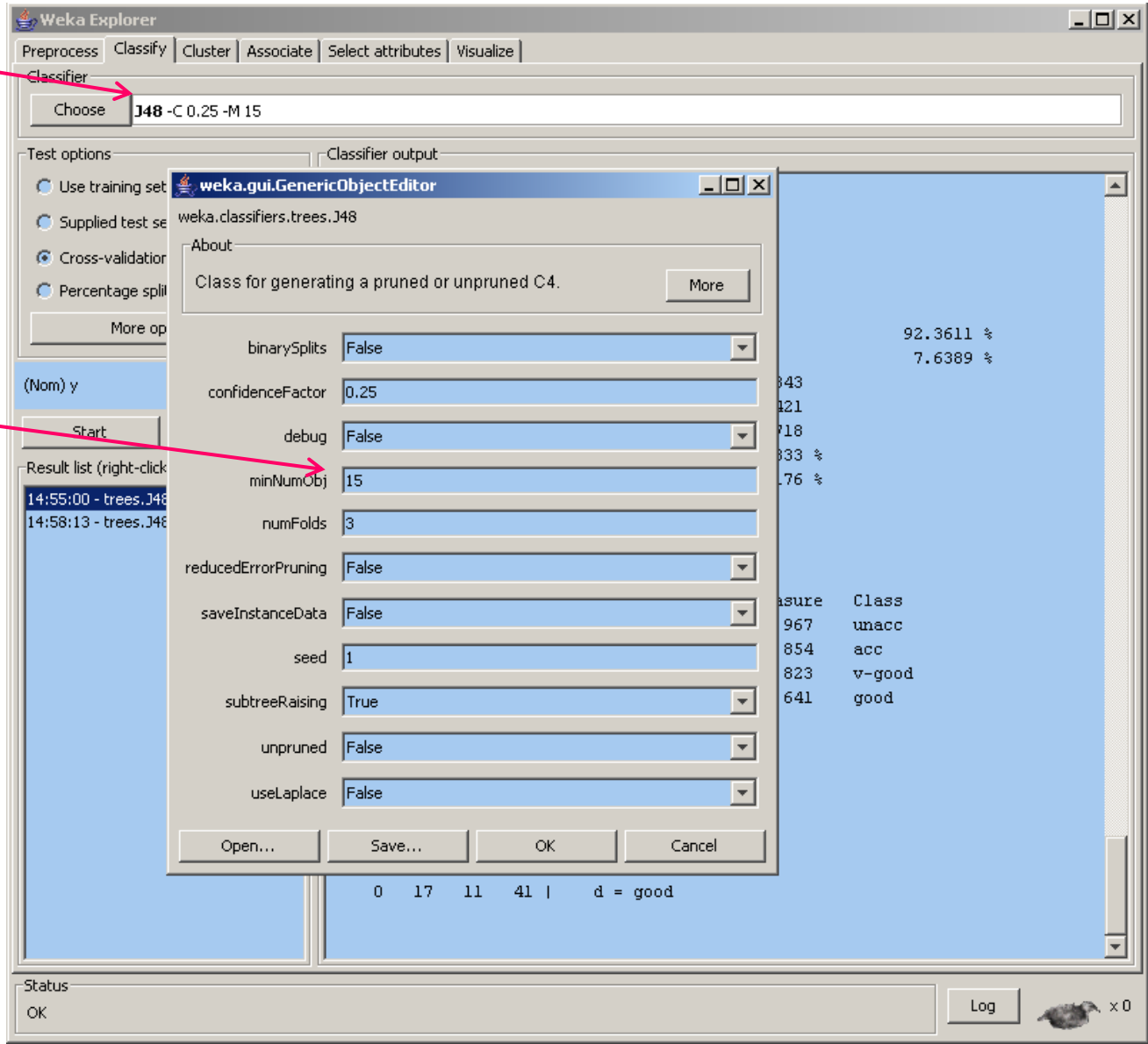
Tree pruning

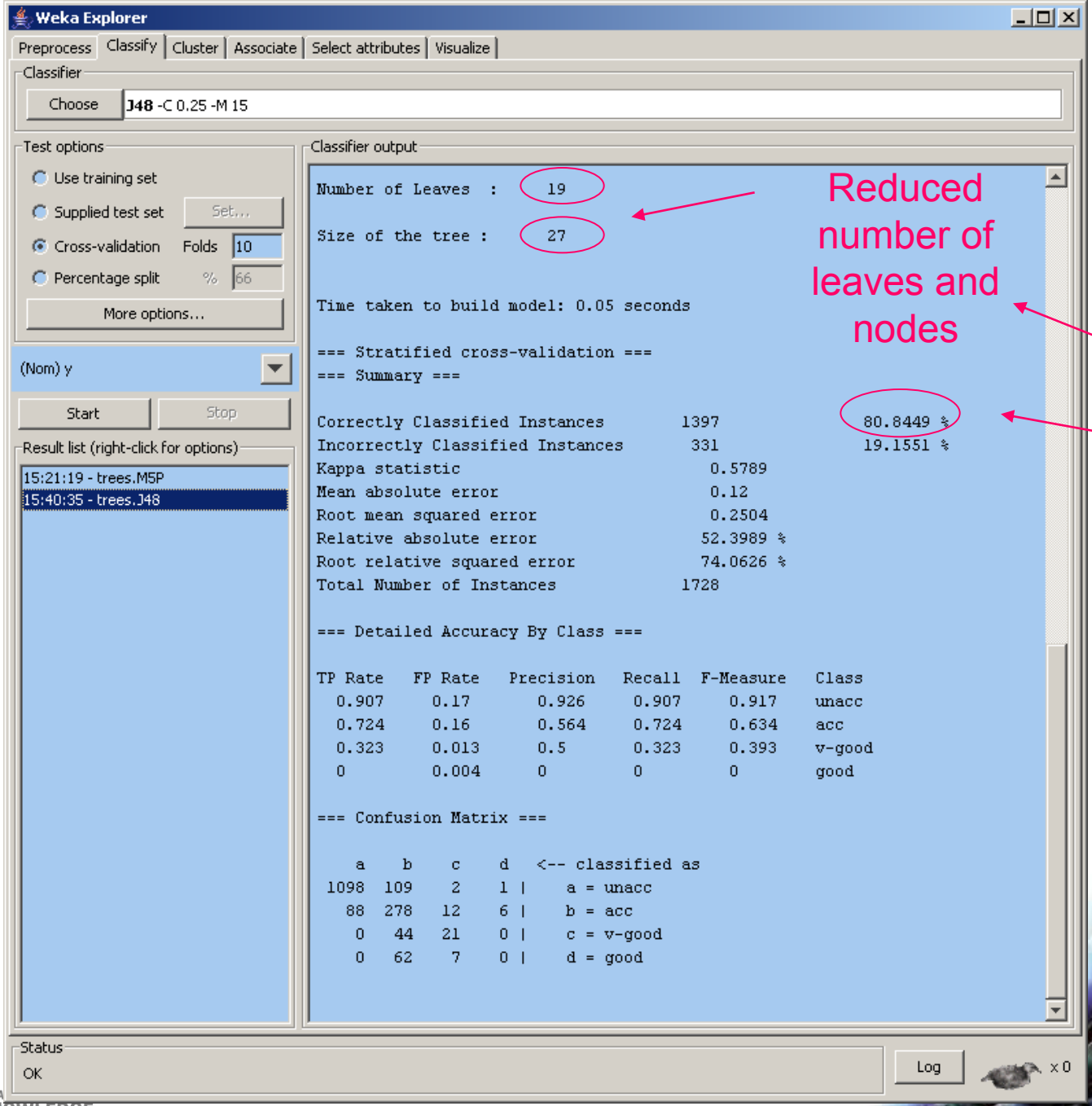
1

Parameters of the algorithm (right mouse click)

2

Set the minimal number of objects per leaf to 15

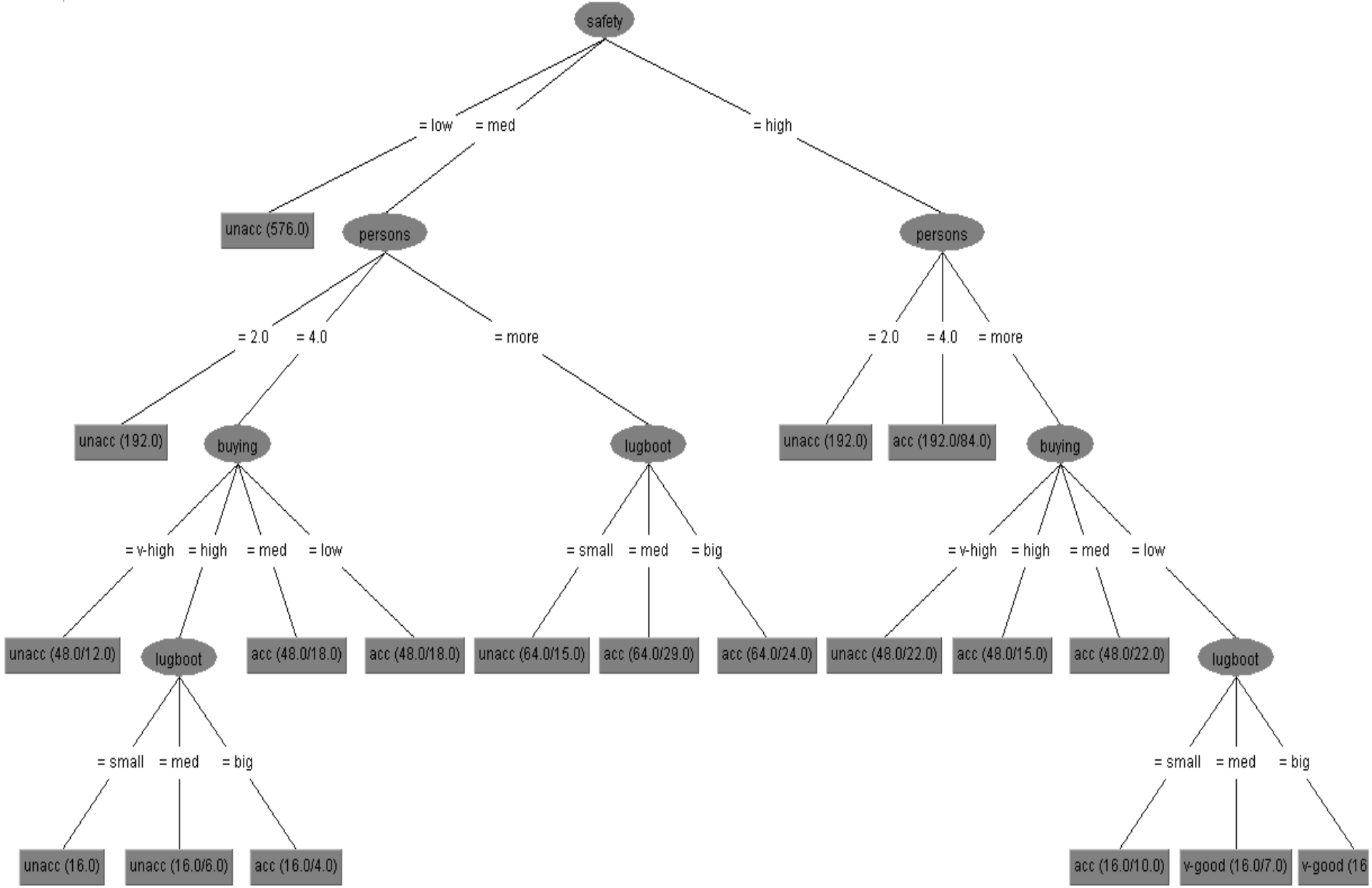




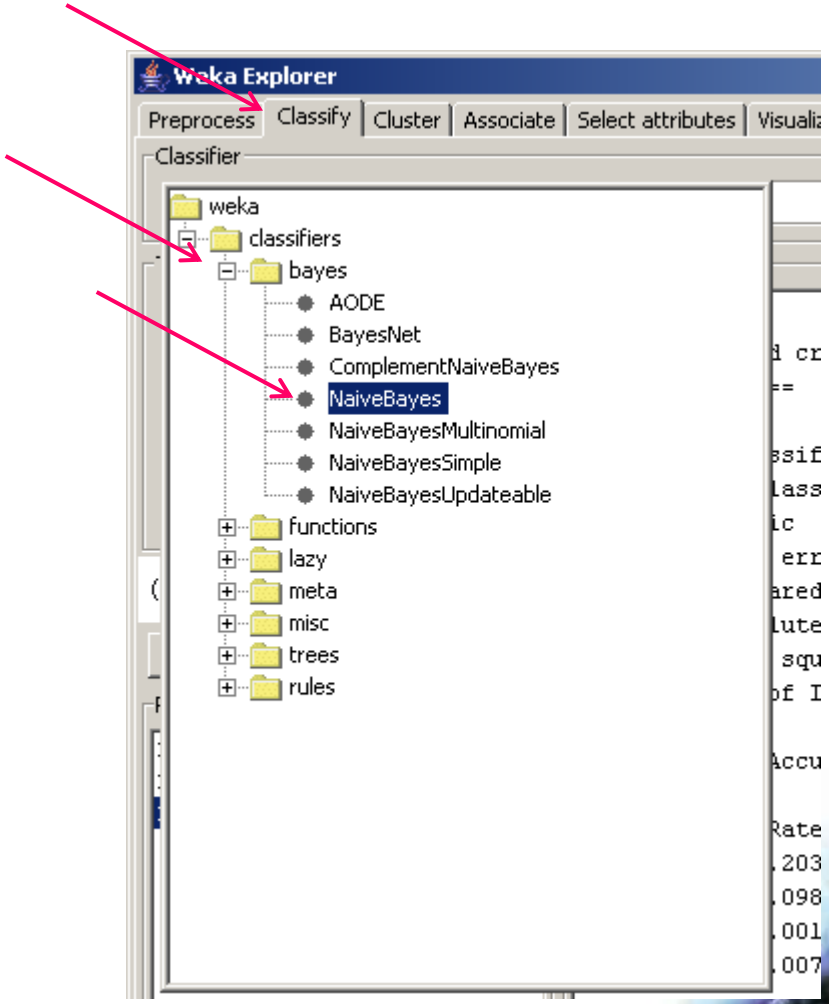
Reduced number of leaves and nodes

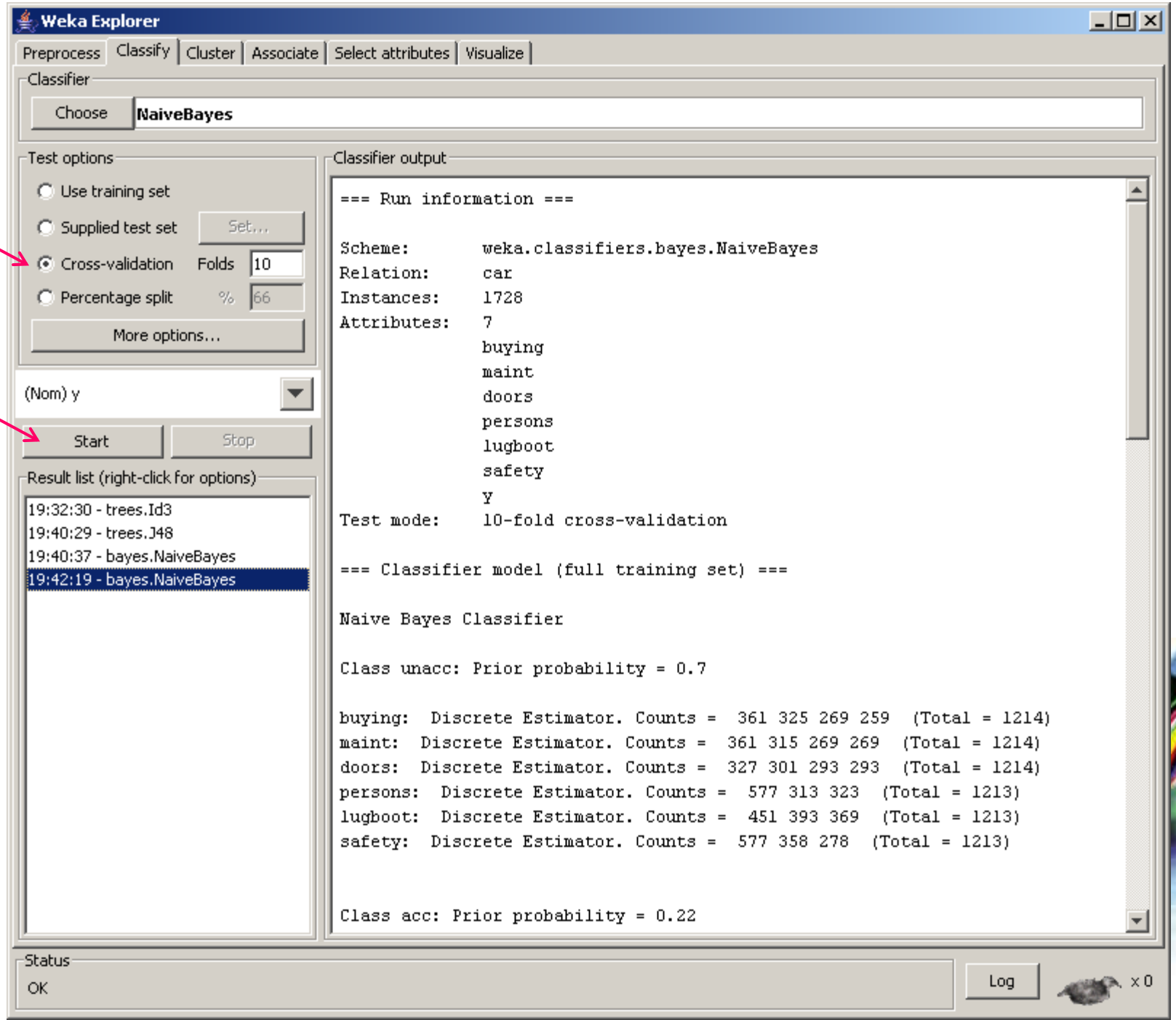
Easier to interpret

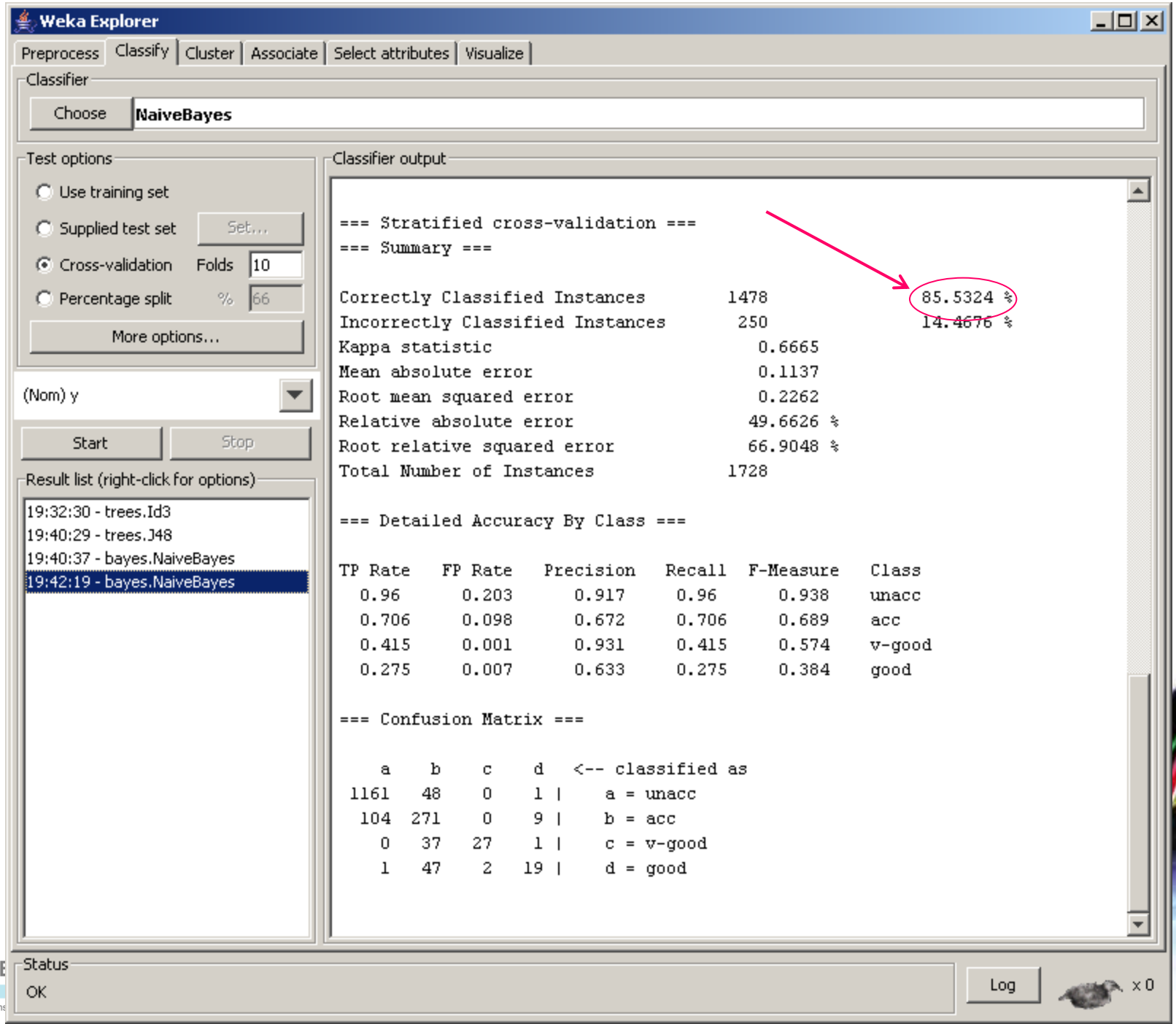
Lower classification accuracy



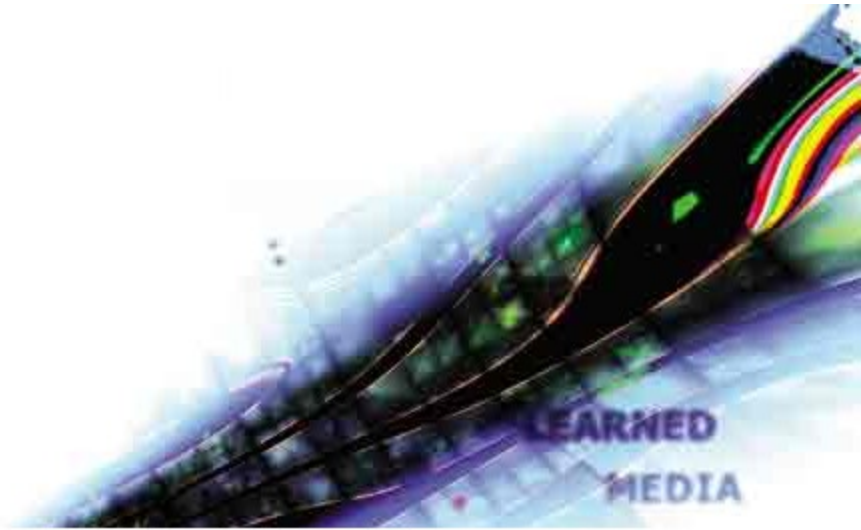
Naïve Bayes classifier





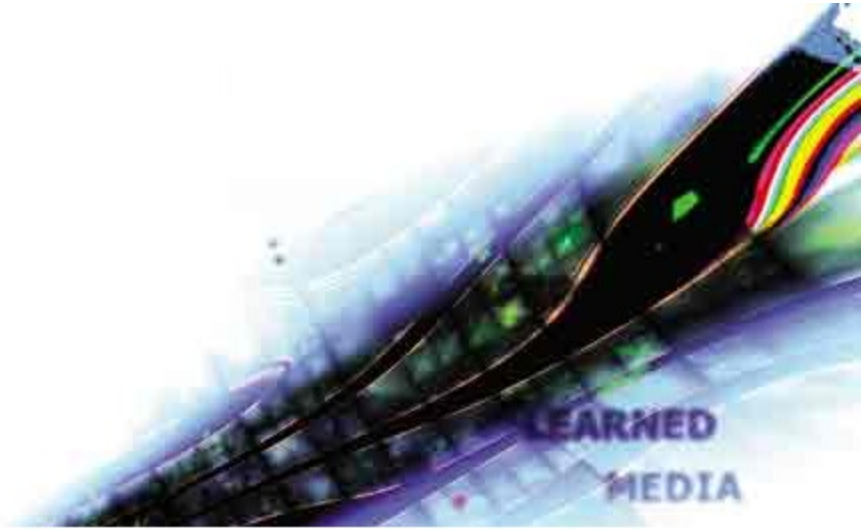

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) y
Start
Stop
Result list (right-click for options)
19:32:30 - trees.Id3
19:40:29 - trees.J48
19:40:37 - bayes.NaiveBayes
19:42:19 - bayes.NaiveBayes
Classifier output
=== 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

Numeric prediction in Weka



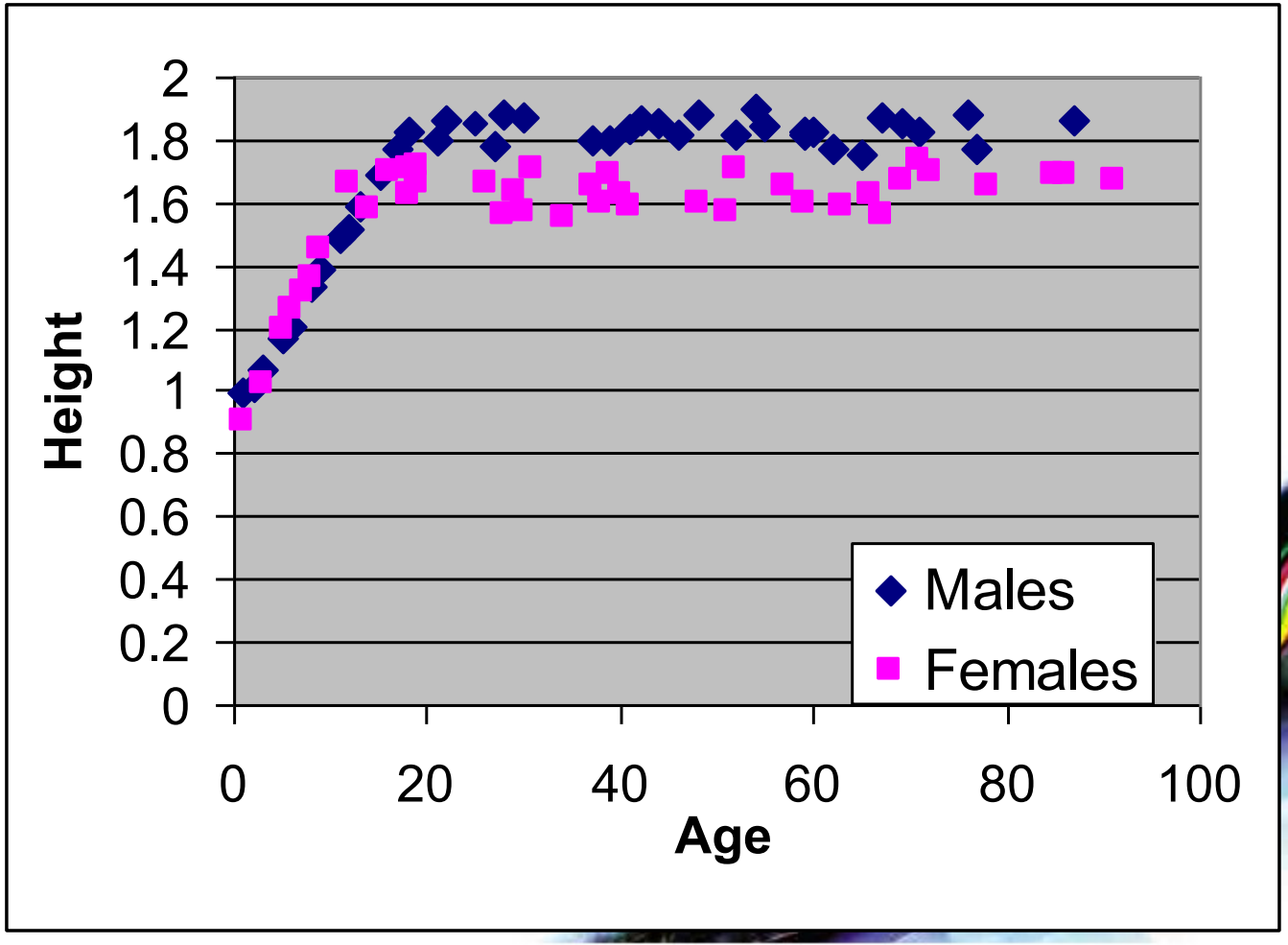
Numeric prediction models

- LinearRegression
- M5P Regression and model trees
- KNN
- Baseline predictor

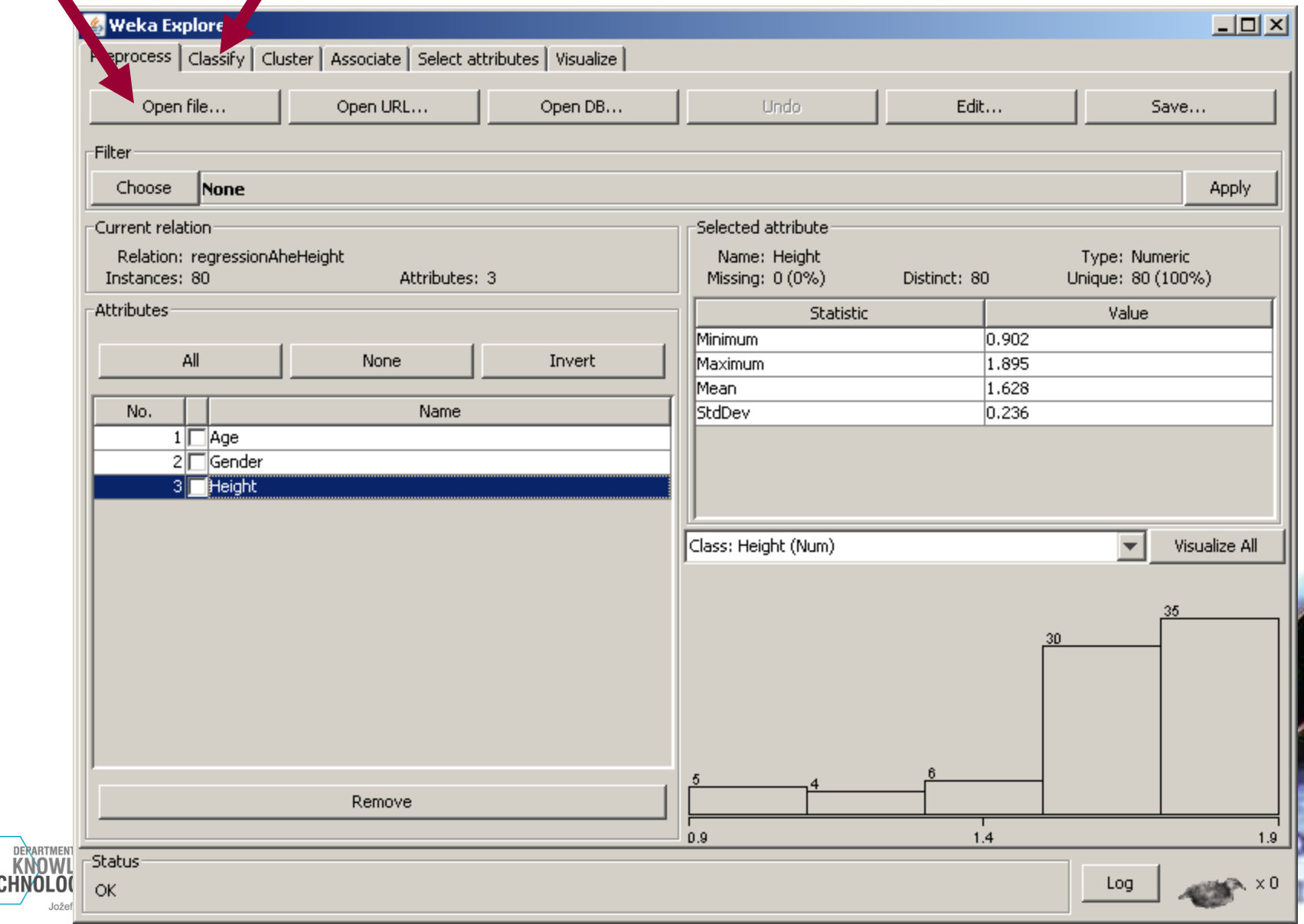


regressionAgeHeight.csv

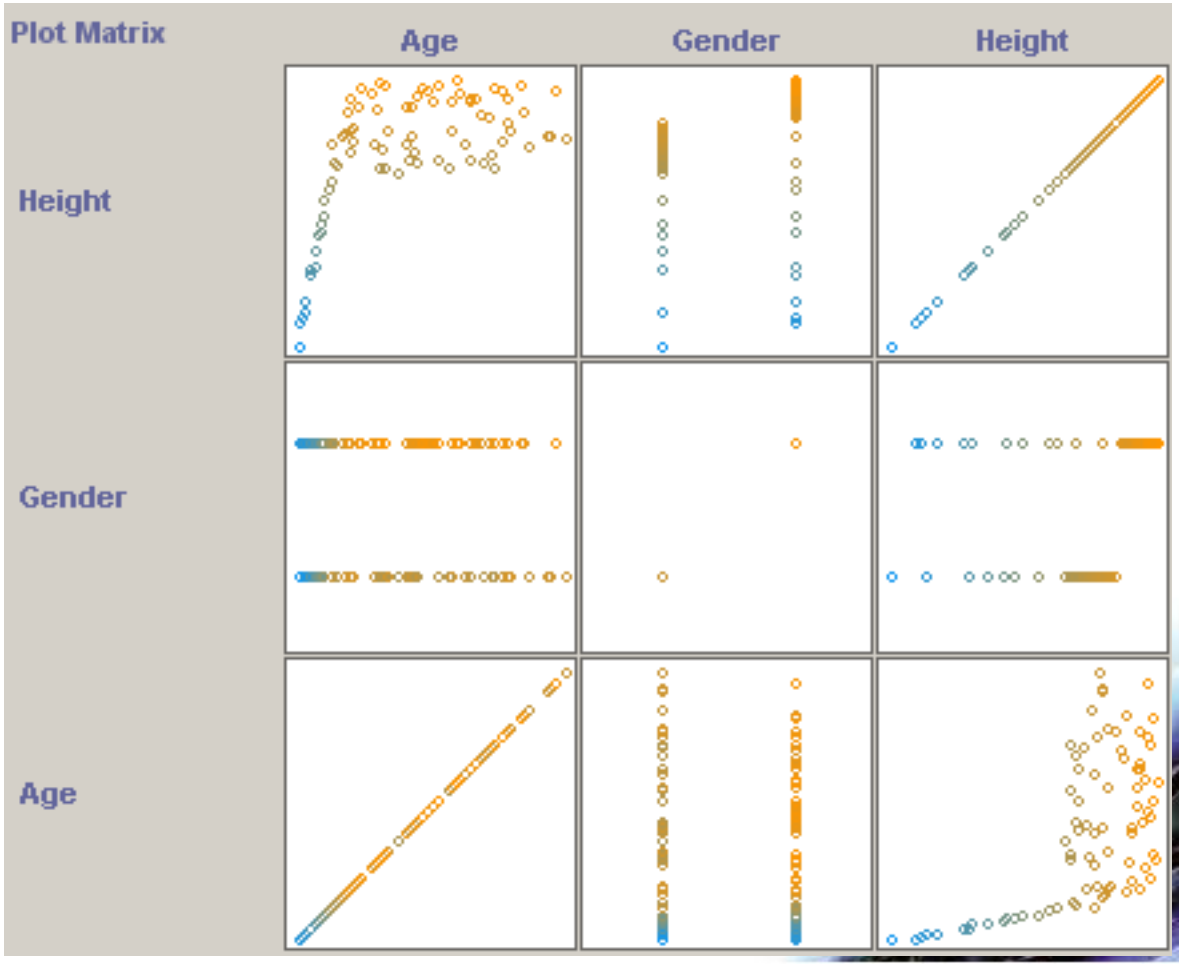
- Data about 80 people:
- Age
- Gender
- Height



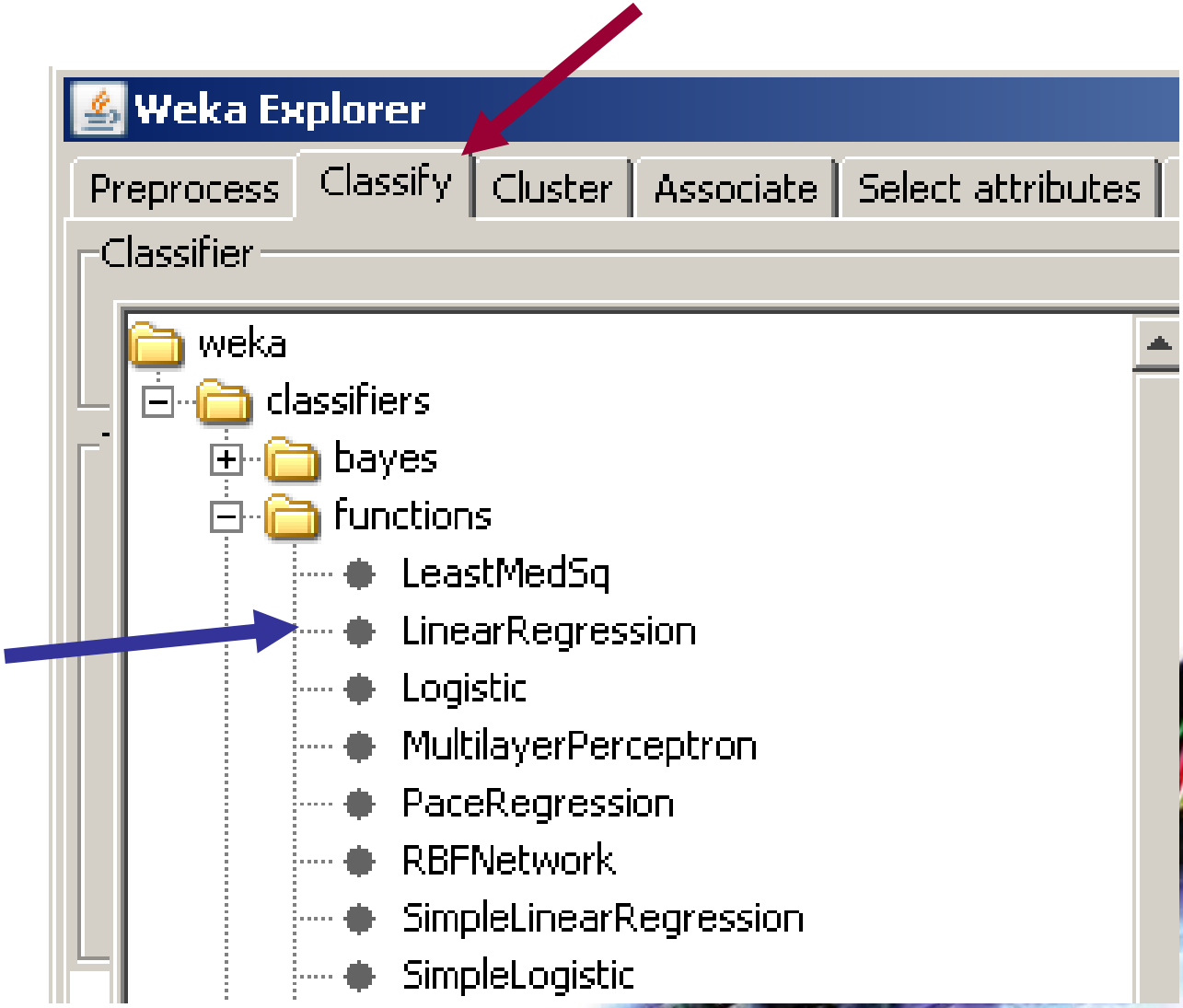
Filename:

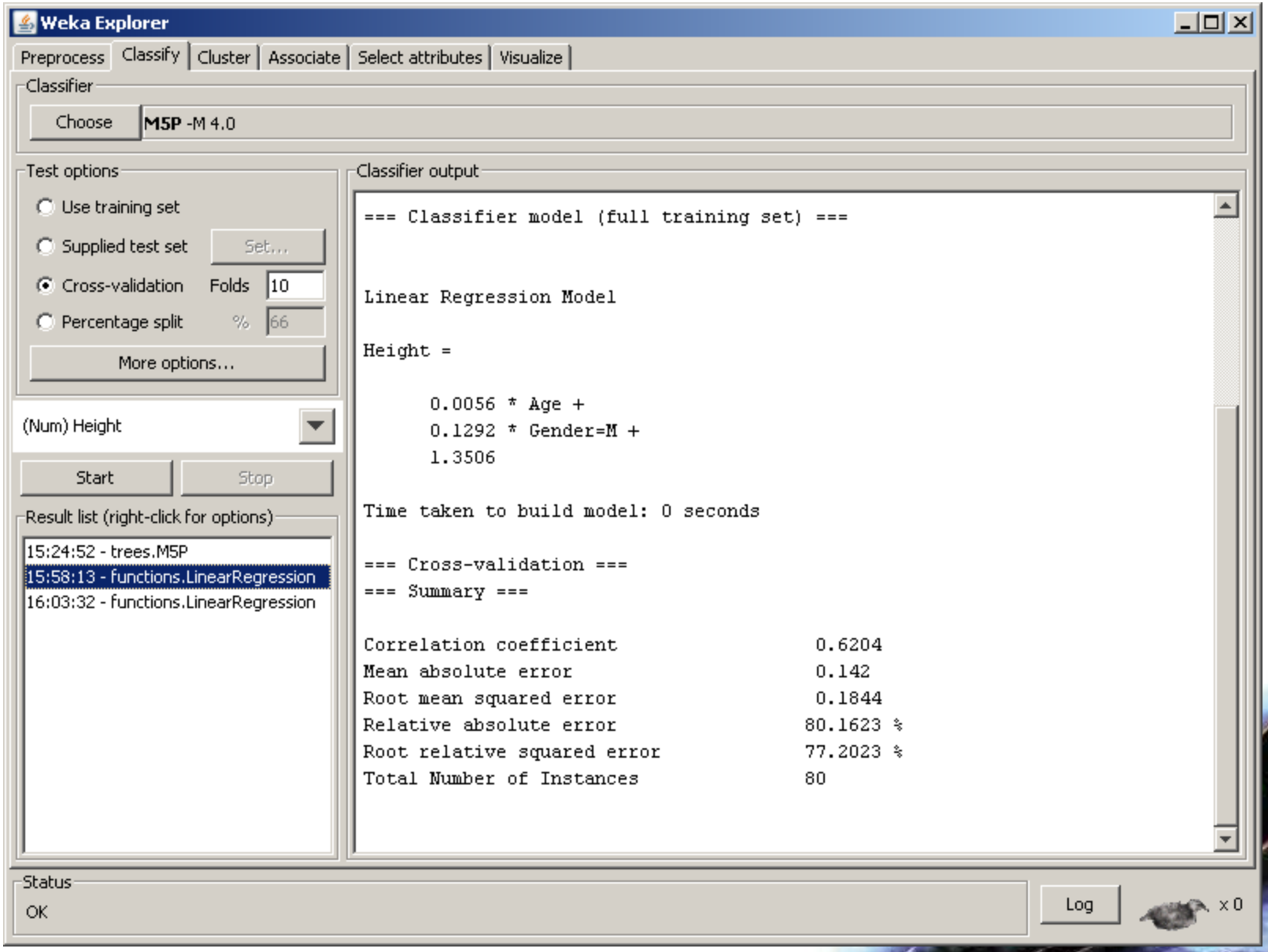

Weka Explore
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: regressionAheHeight
Instances: 80
Attributes: 3
Attributes
All
None
Invert
No.
Name
1 Age
2 Gender
3 Height
Remove
Selected attribute
Name: Height
Type: Numeric
Missing: 0 (0%)
Distinct: 80
Unique: 80 (100%)
Statistic
Value
Minimum 0.902
Maximum 1.895
Mean 1.628
StdDev 0.236
Class: Height (Num)
Visualize All
35
30
6
5
4
0.9
1.4
1.9
Status
OK
Log
x 0

Visualization in Weka

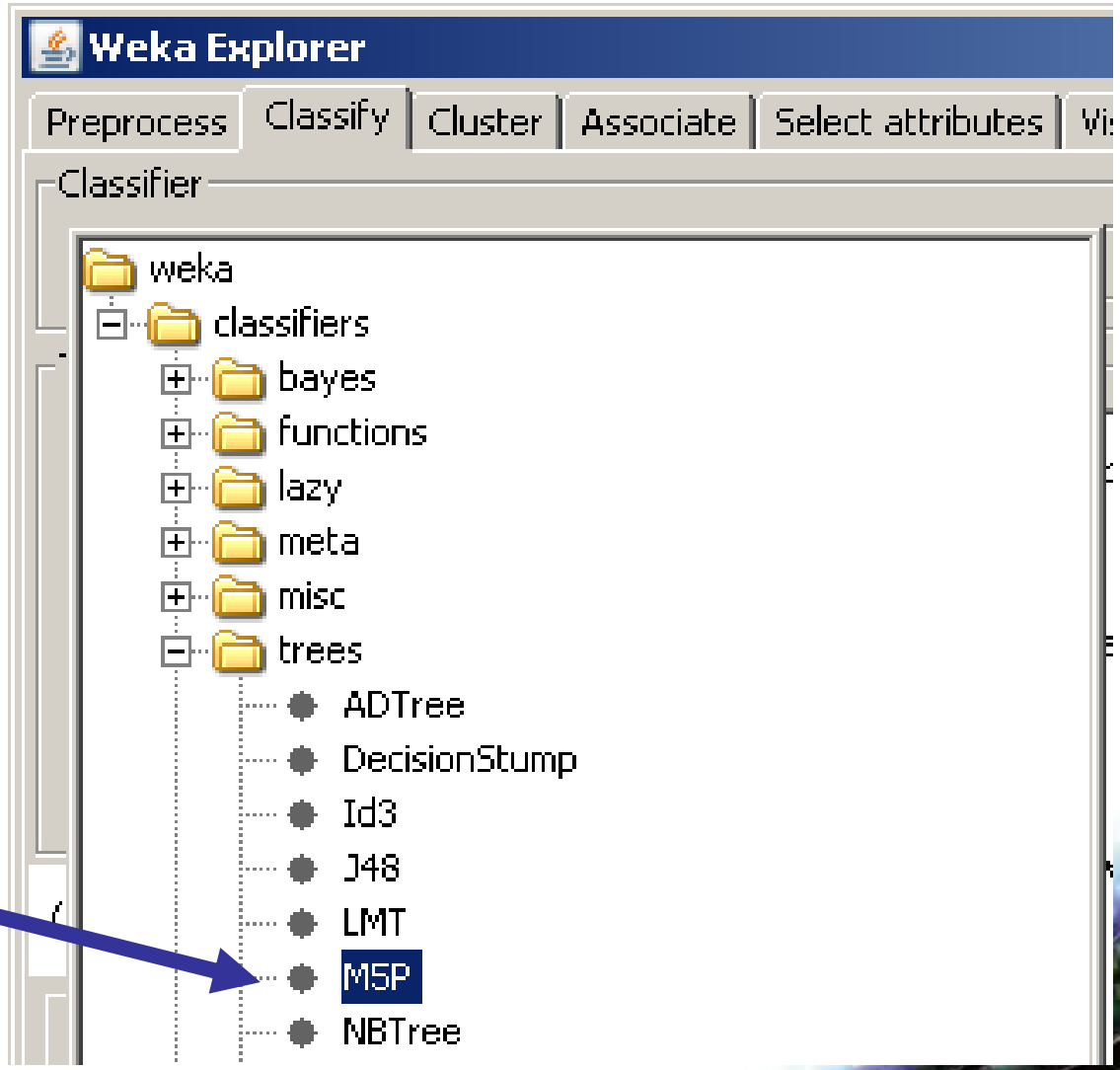


Weka → classifiers → functions → LinearRegression

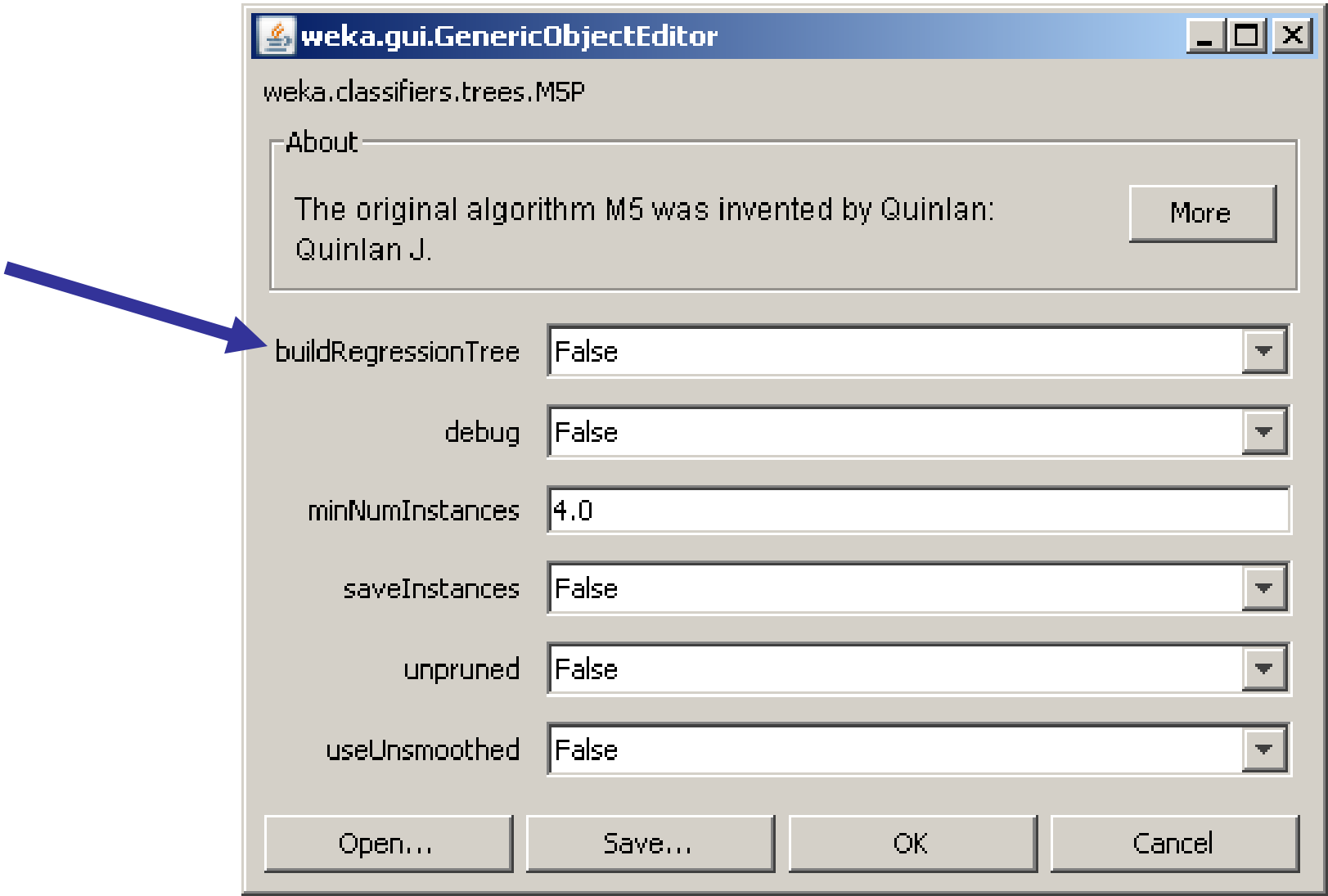




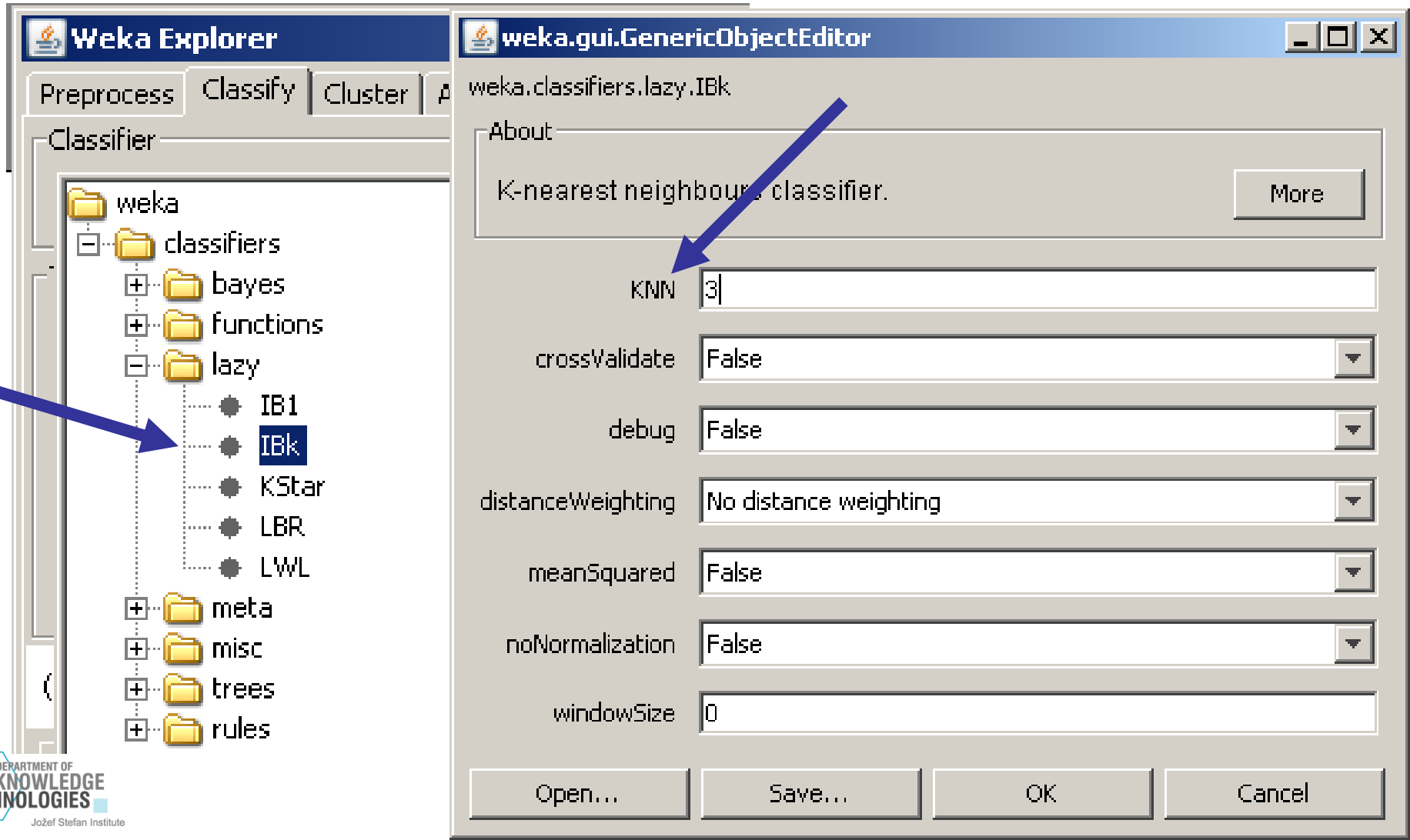
Weka → classifiers → trees → M5P



builtRegressionTree = True → regression tree
builtRegressionTree = False → model tree



kNN: weka → classifiers → lazy →IBk



Weka → classifiers → rules → ZeroR