Hands on WEKA

2006.11.15

Petra Kralj

Petra.Kralj@ijs.si

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

To assist you

- Branko Kavšek (Branko.Kavsek@ijs.si)
- Panče Panov (Pance.Panov@ijs.si)
- Dragi Kocev (Dragi.Kocev@ijs.si)
- Ivica Slavkov (Ivica.Slavkov@ijs.si)
- Matjaž Depolli (Matjaz.Depolli@ijs.si)

Petra.Kralj@ijs.si

Plan for the session

- Classification: CAR dataset
  - Preparing and loading the data
  - Building decision trees
  - Building set of rules
  - Estimating model quality
- Regression: Imports-85 dataset
  - Model trees
  - Regression trees
- Descriptive induction: Voting & Iris dataset
  - Association rules
  - Clustering

Petra.Kralj@ijs.si

CLASSIFICATION

CAR dataset

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

CAR dataset

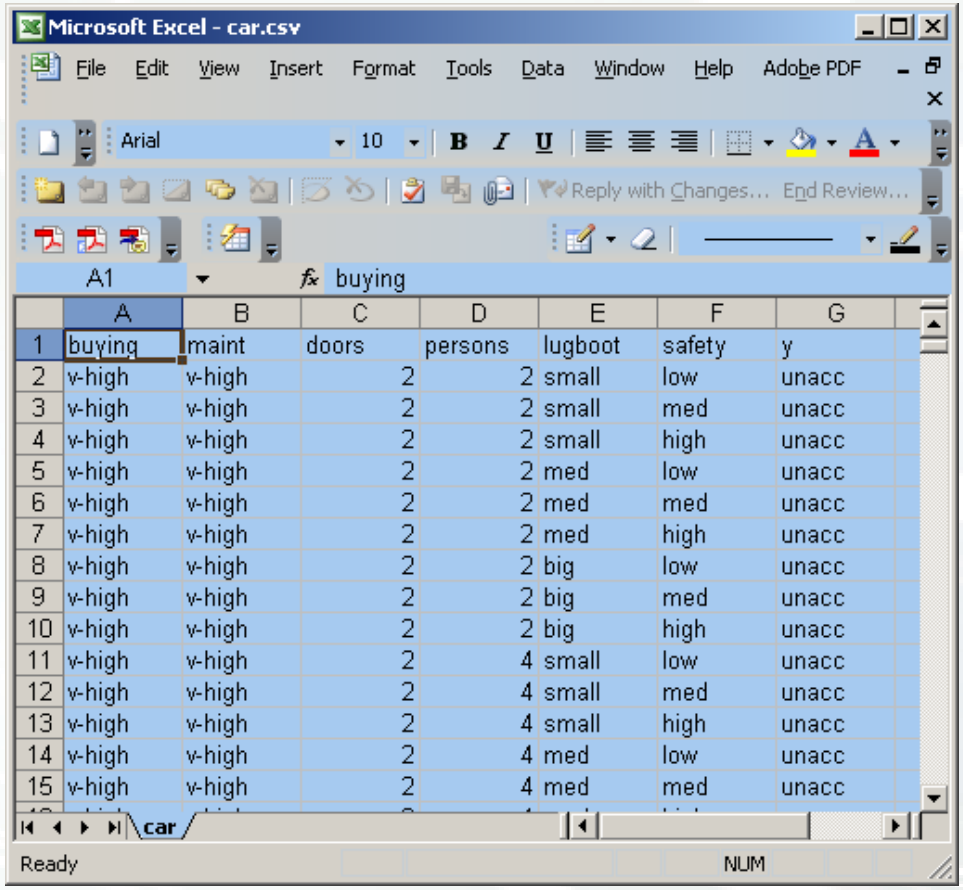
- 1728 instances
- 6 attributes
  - 6 nominal attributes
  - 0 numeric attributes
- Nominal target variable
  - 4 values: unacc, acc, good, v-good
  - Distribution of values
    - unacc (70%), acc (22%), good (4%), v-good (4%)
- Missing values
  - No missing values

Petra.Kralj@ijs.si

Preparing the data for WEKA - 1

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

- Rows are instances
- Columns are attributes
- The last column is the target attribute



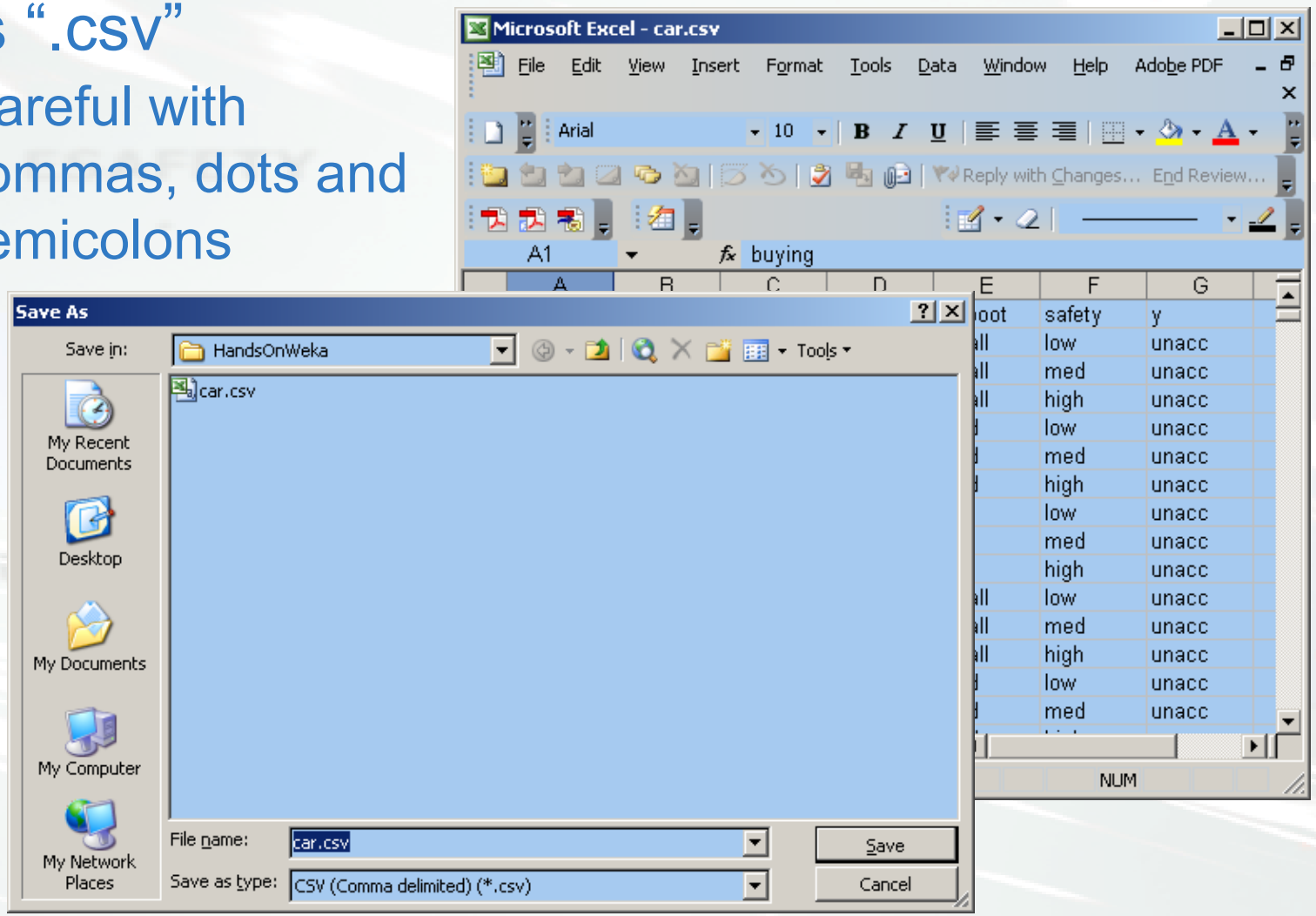
	A	B	C	D	E	F	G
1	buying	maint	doors	persons	lugboot	safety	y
2	v-high	v-high	2	2	small	low	unacc
3	v-high	v-high	2	2	small	med	unacc
4	v-high	v-high	2	2	small	high	unacc
5	v-high	v-high	2	2	med	low	unacc
6	v-high	v-high	2	2	med	med	unacc
7	v-high	v-high	2	2	med	high	unacc
8	v-high	v-high	2	2	big	low	unacc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

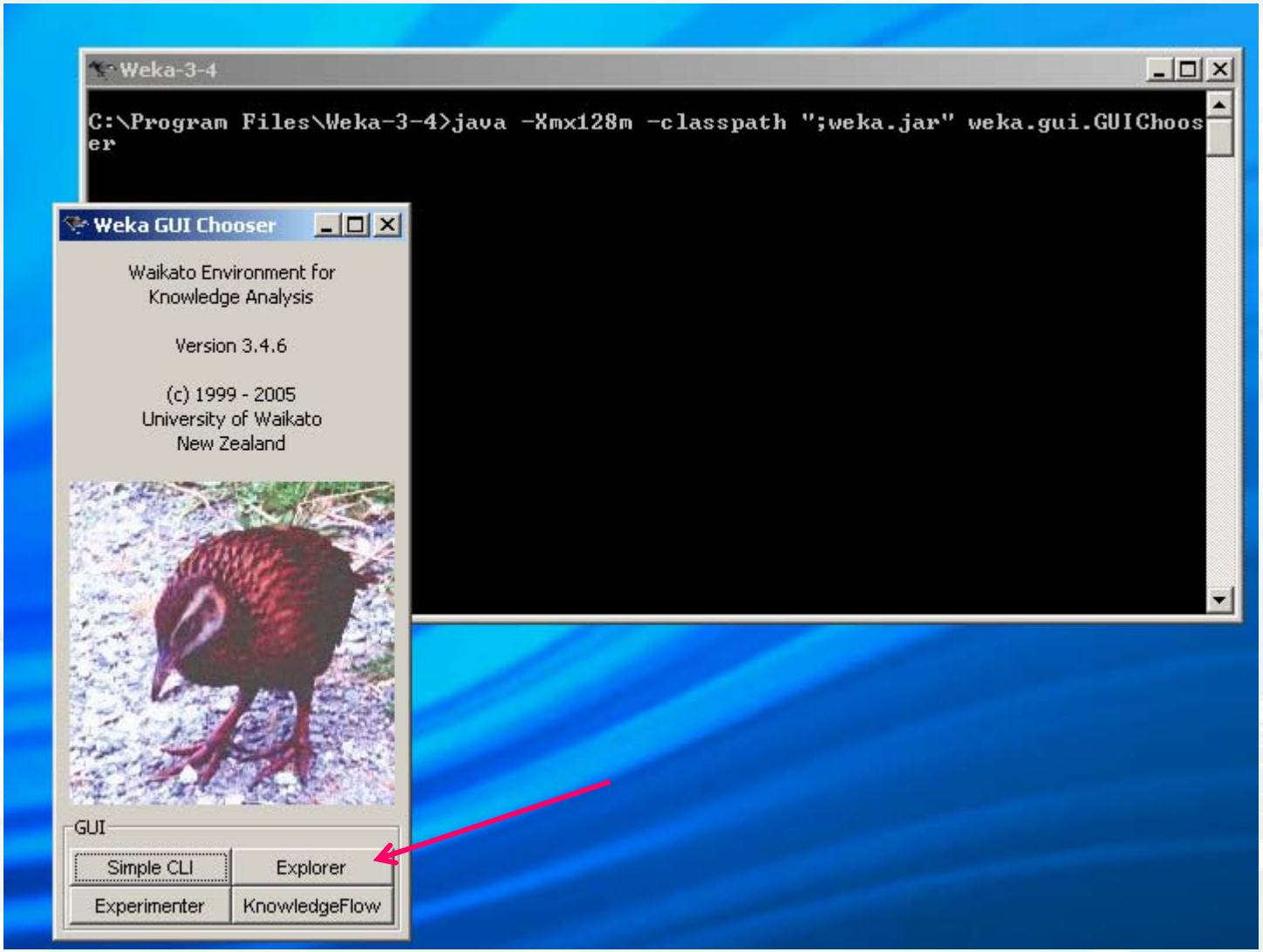
Save as ".csv"

- Careful with commas, dots and semicolons



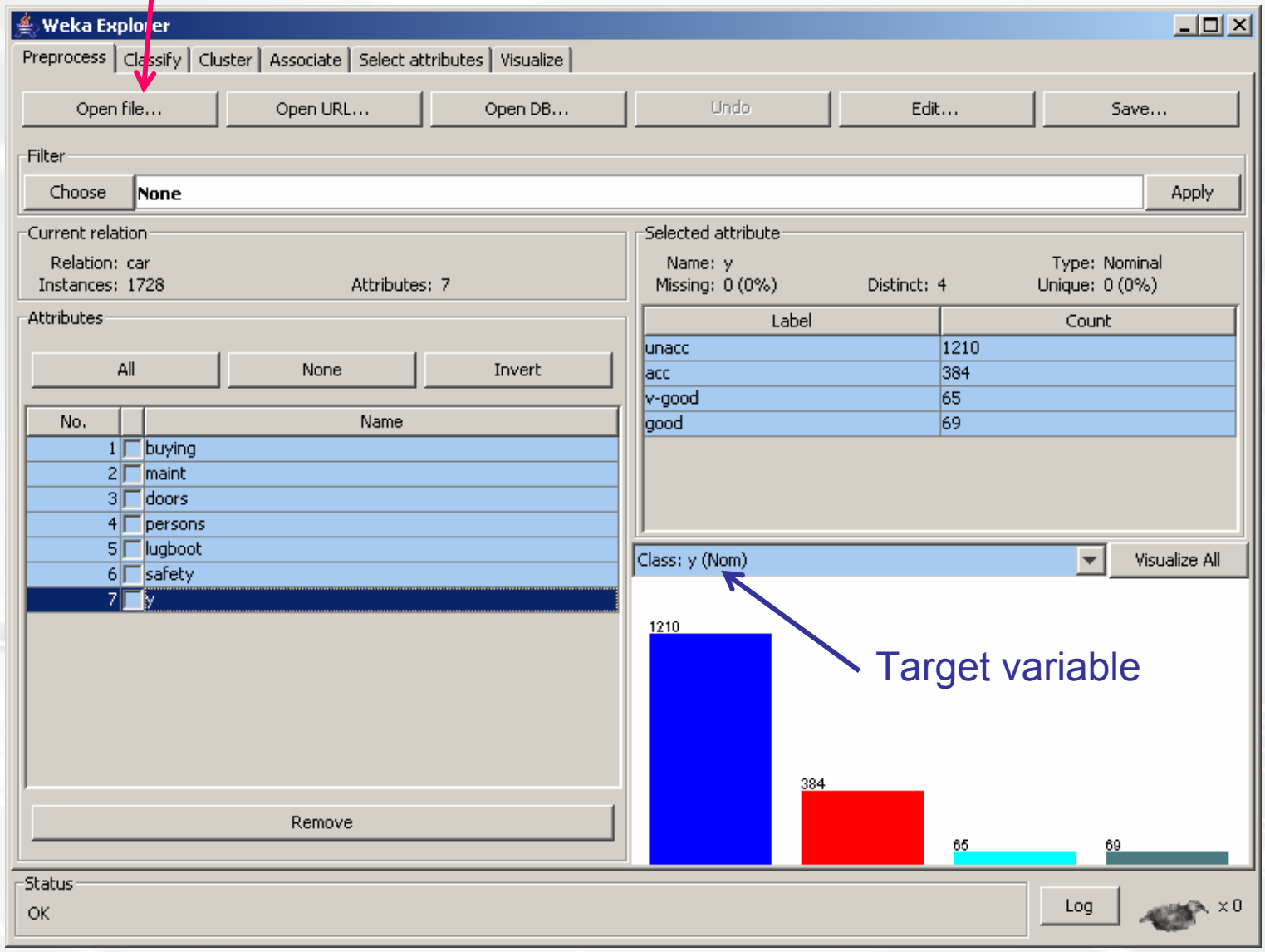
Petra.Kralj@ijs.si

Open WEKA Explorer



Petra.Kralj@ijs.si

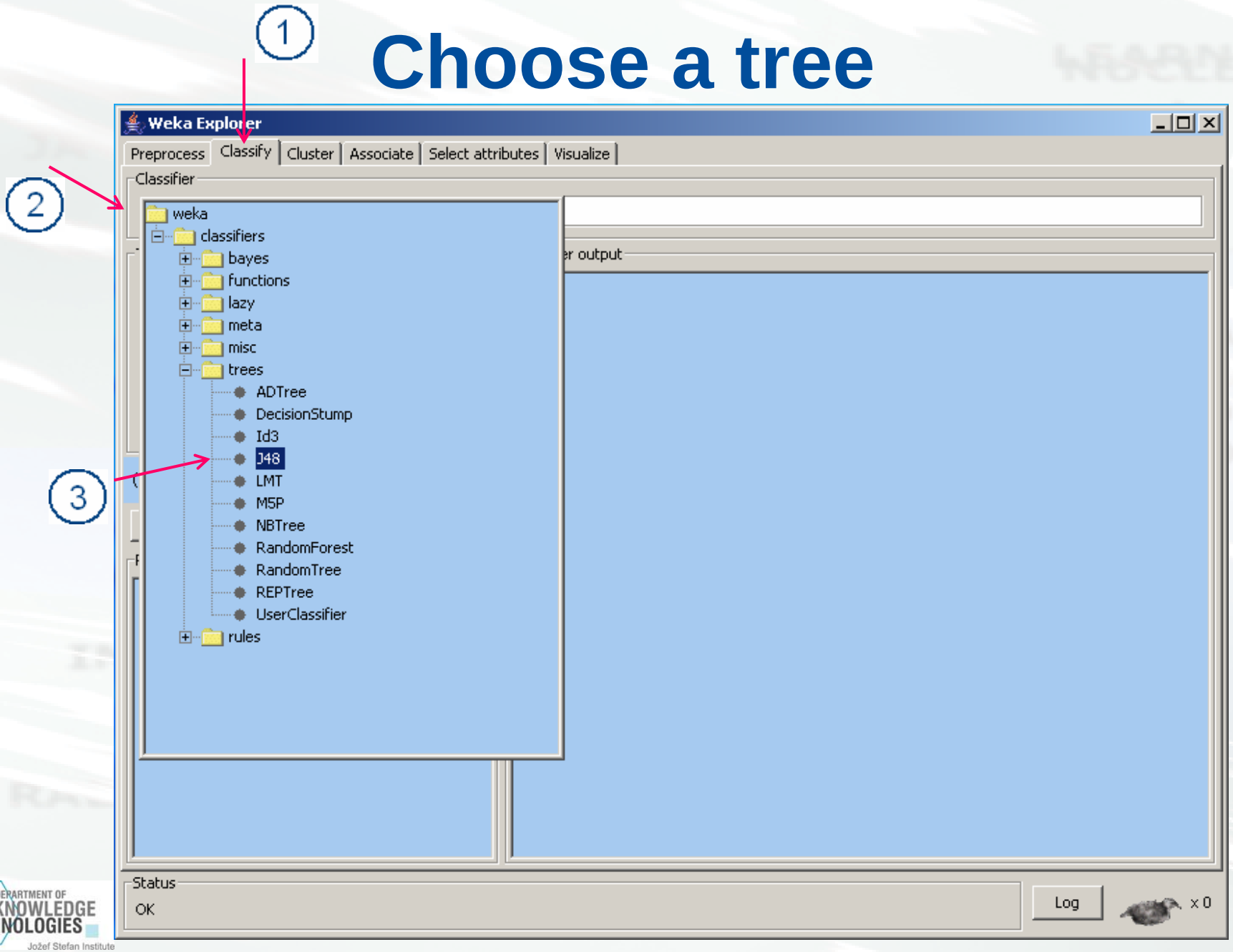
Load the data



DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

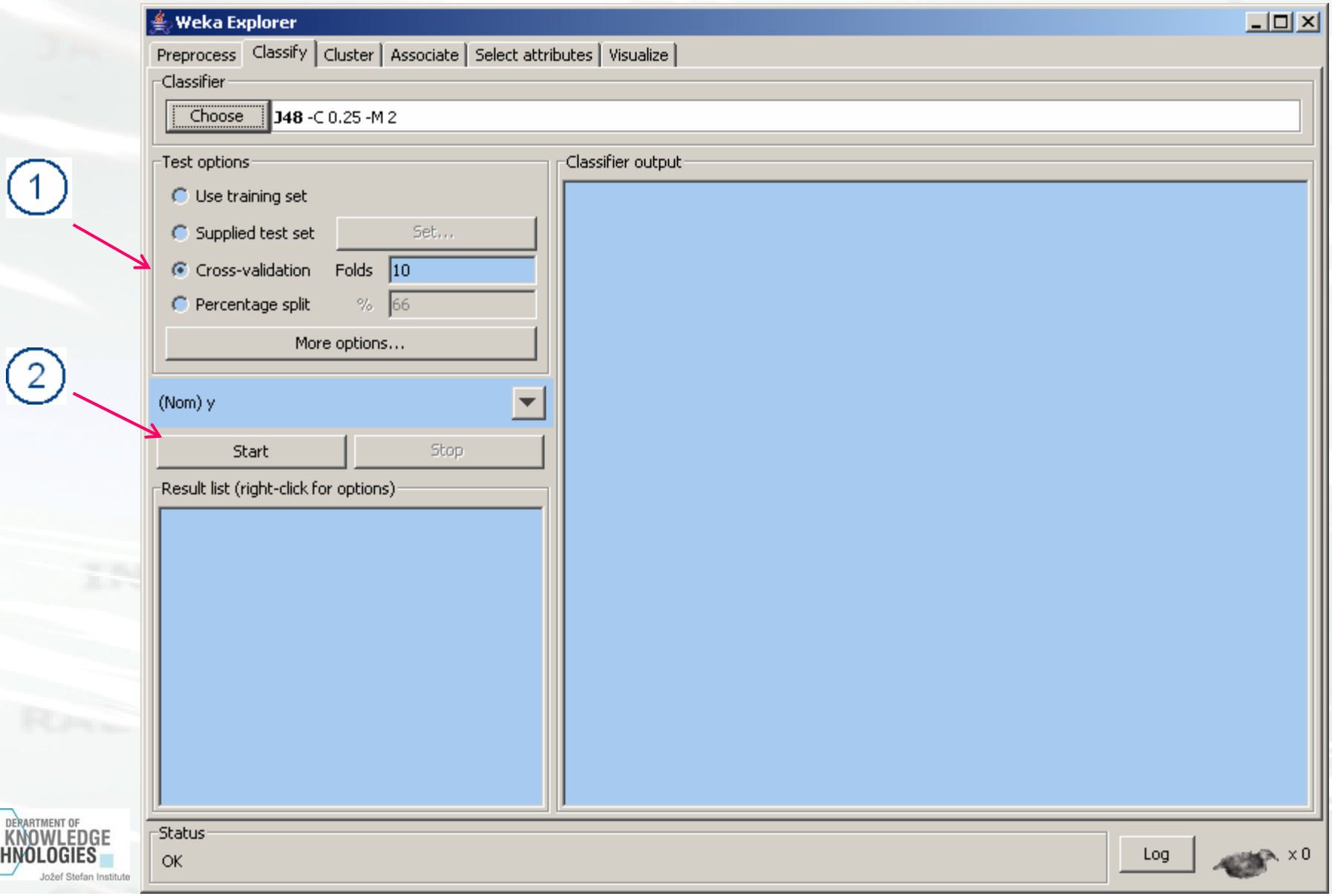
Petra.Kralj@ijs.si

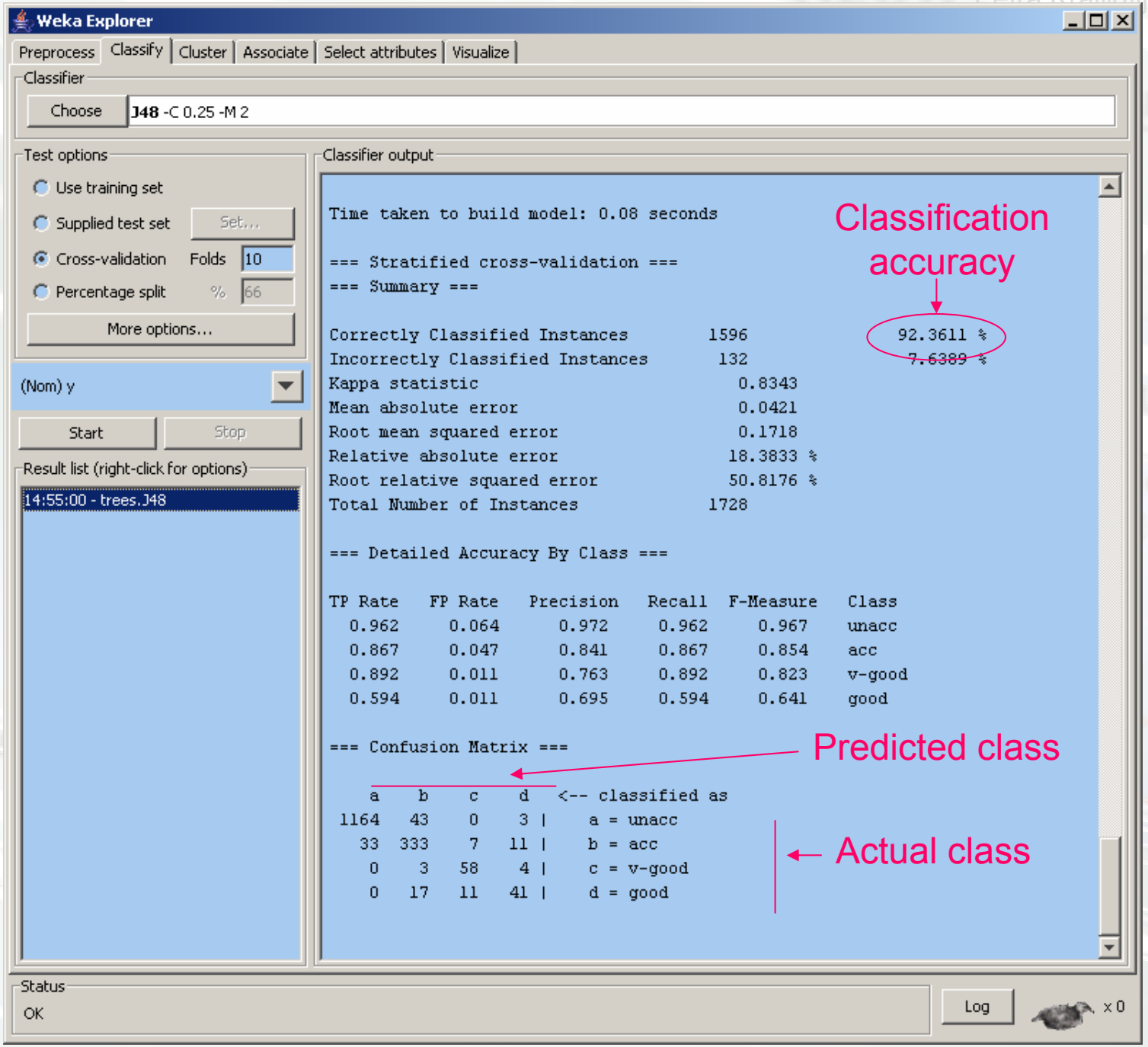
Choose a tree

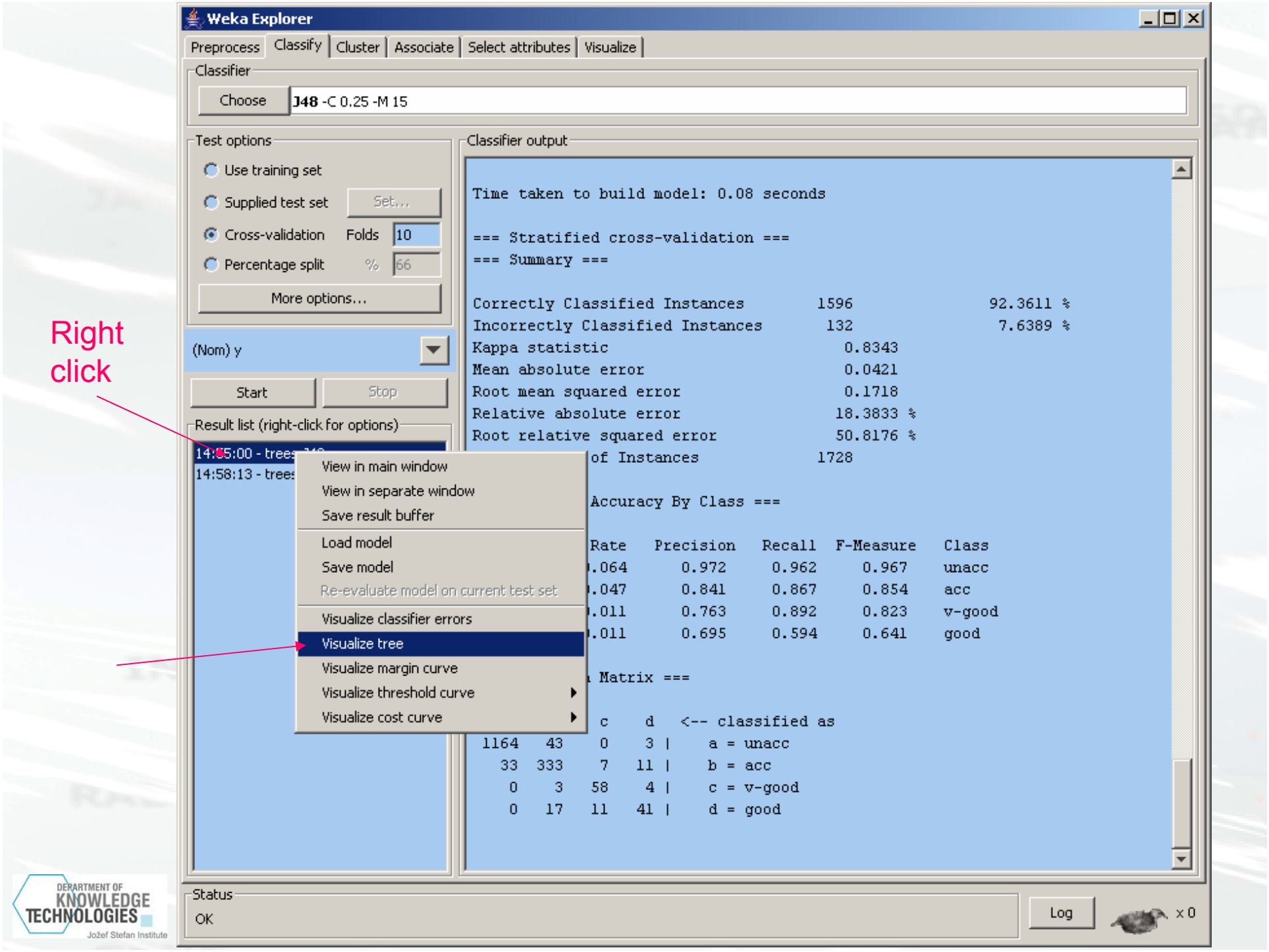


Petra.Kralj@ijs.si

Build tree + evaluate







Right click

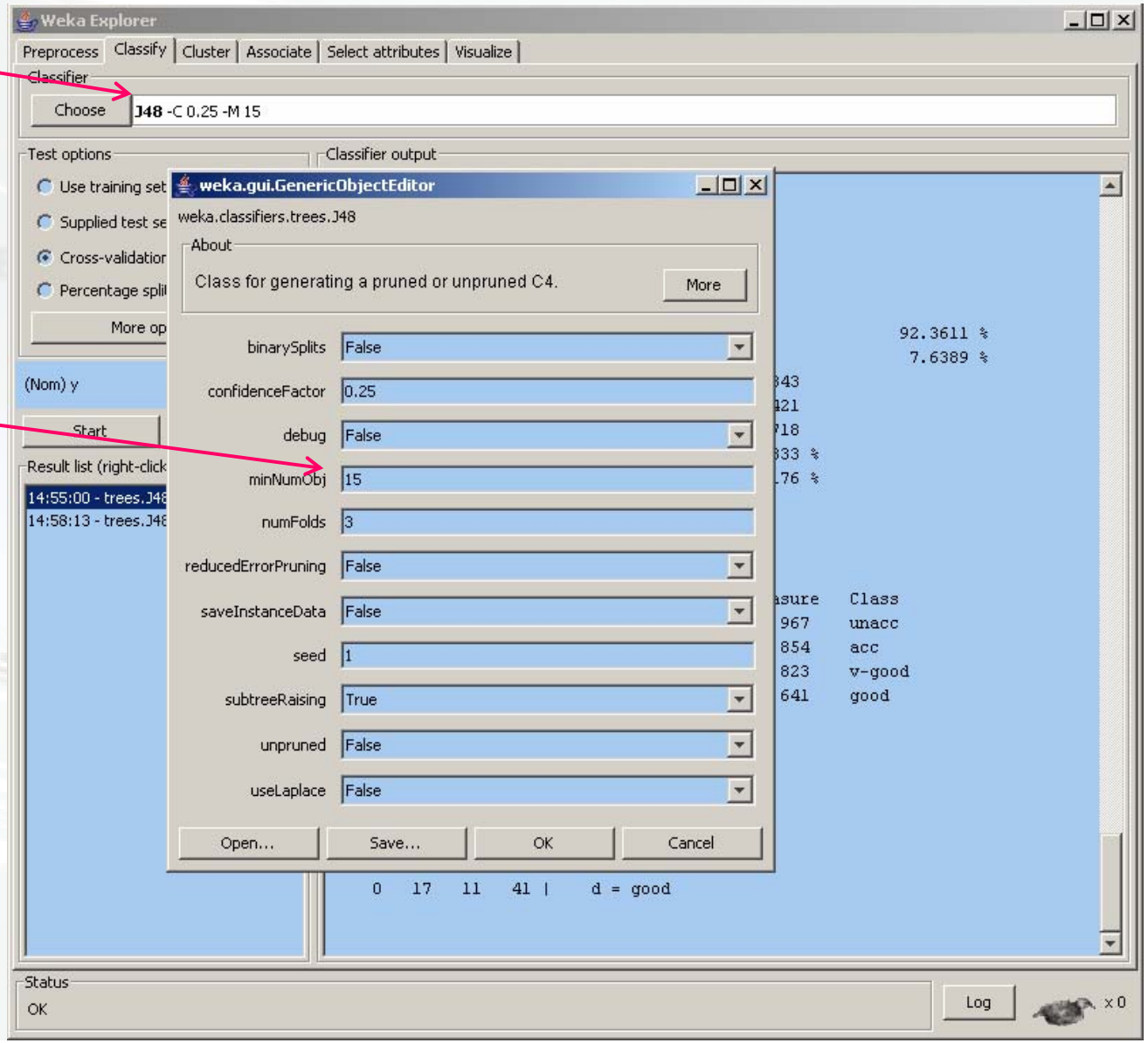
Tree pruning

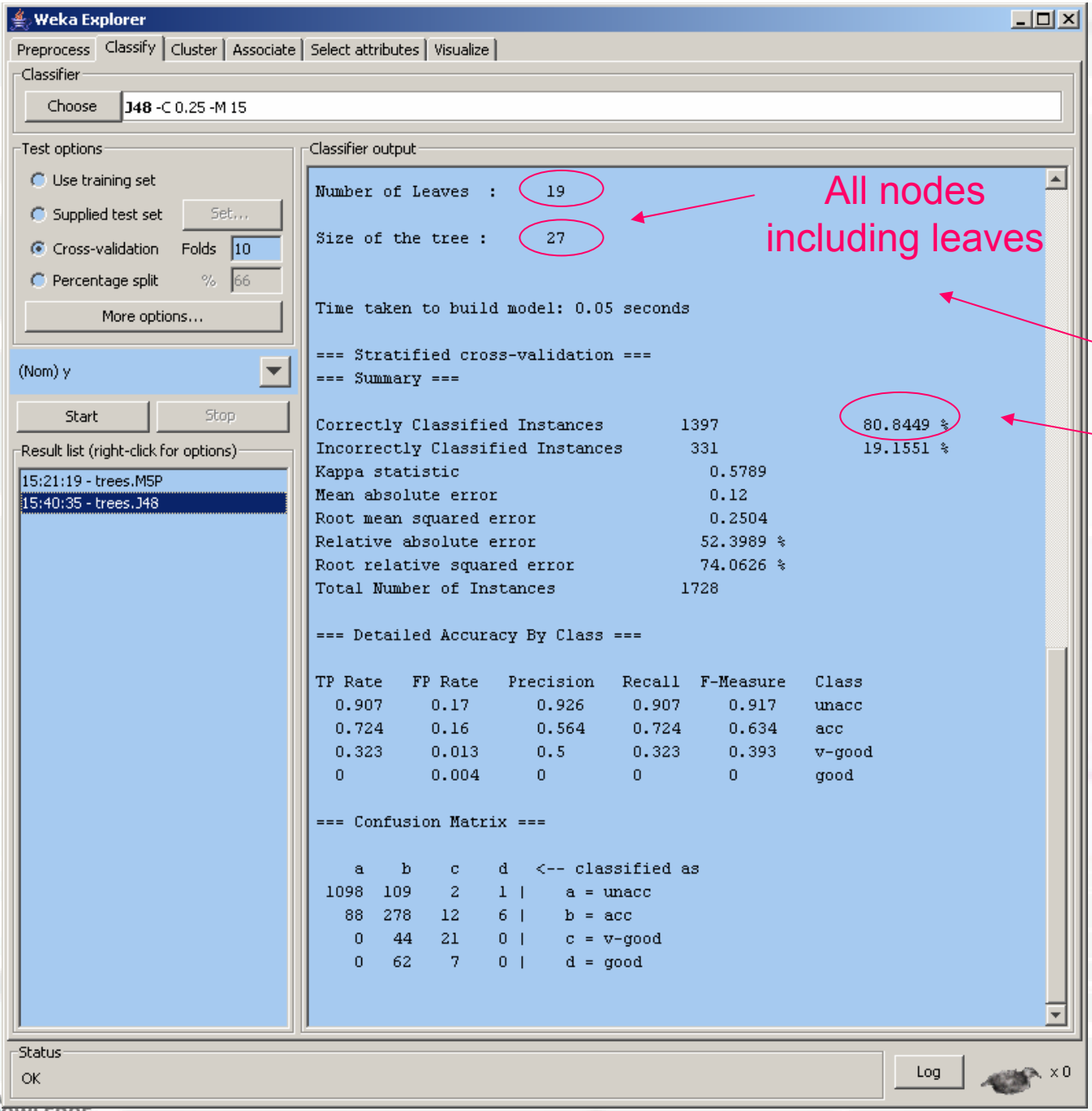
1

Algorithm parameters

2

Set the minimum number of examples in a leaf

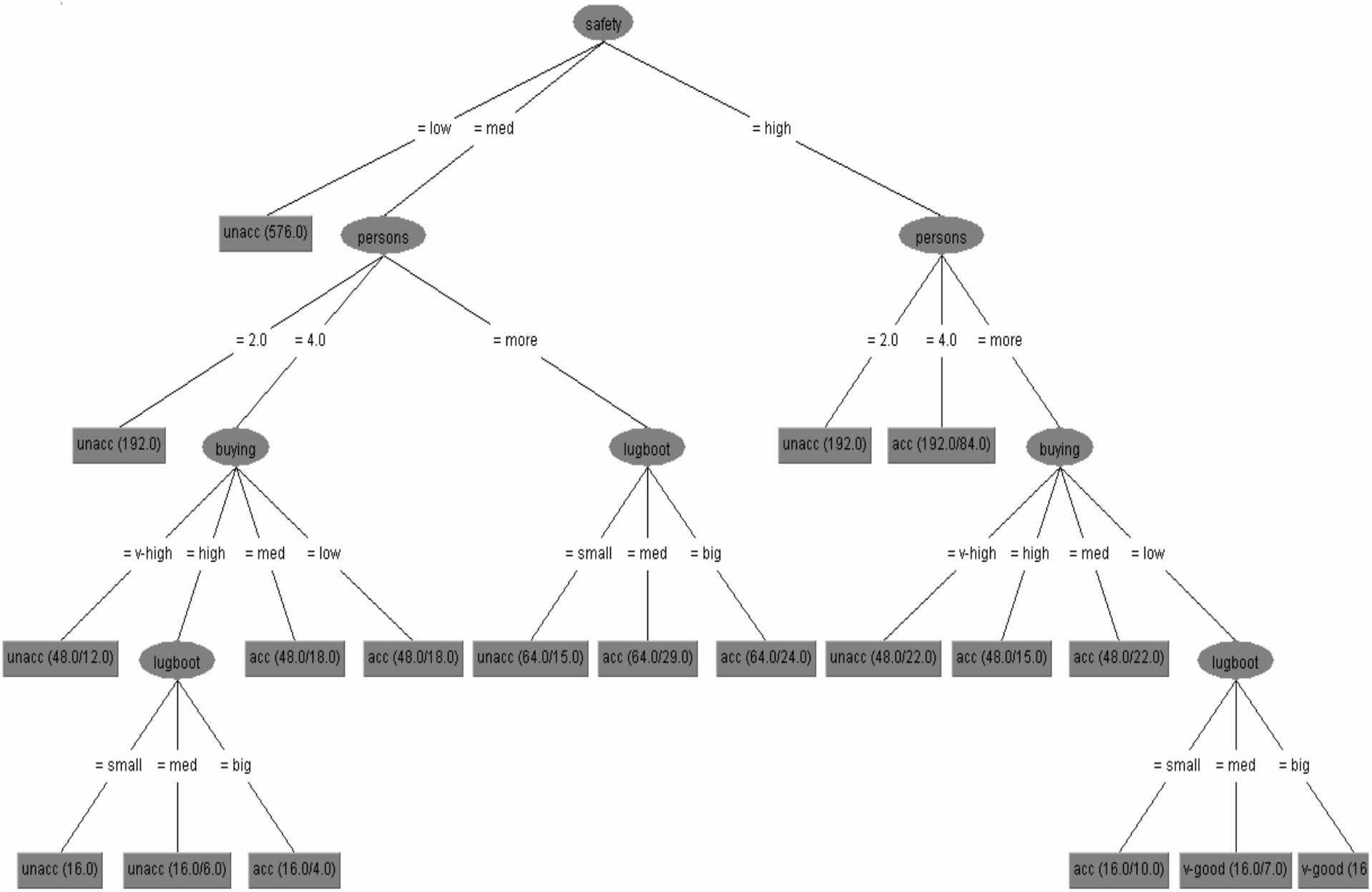




All nodes including leaves

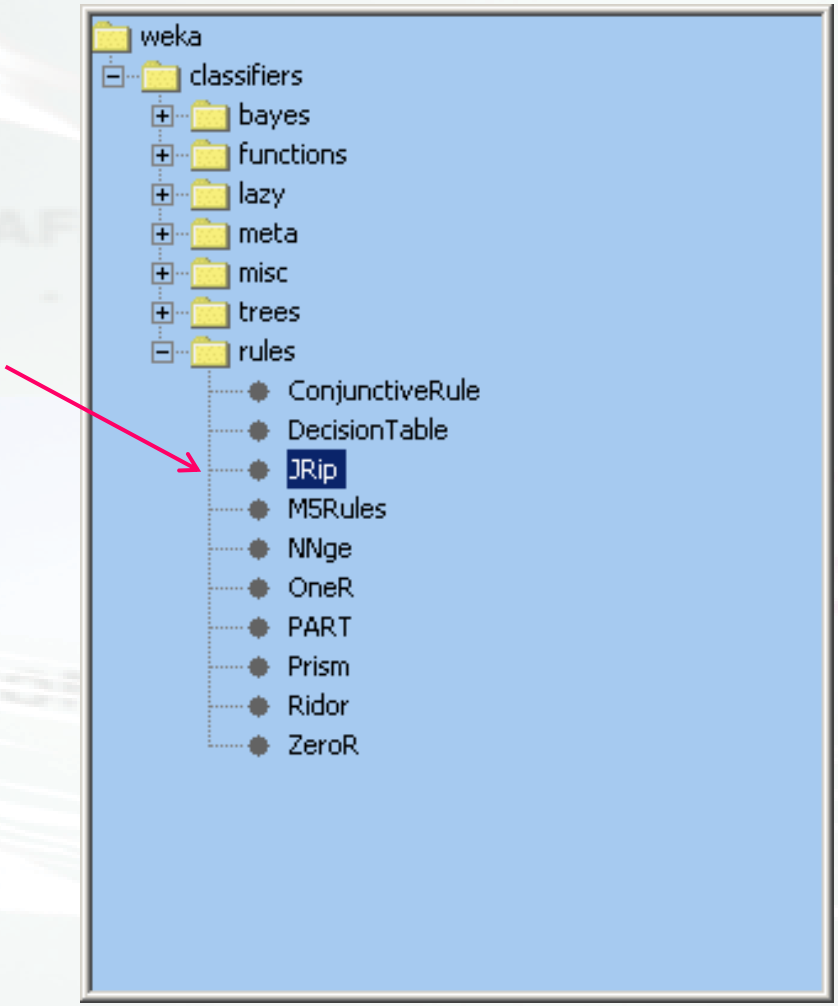
higher understandability,

lower accuracy



Petra.Kralj@ijs.si

Building classification rules



Petra.Kralj@ijs.si

REGRESSION

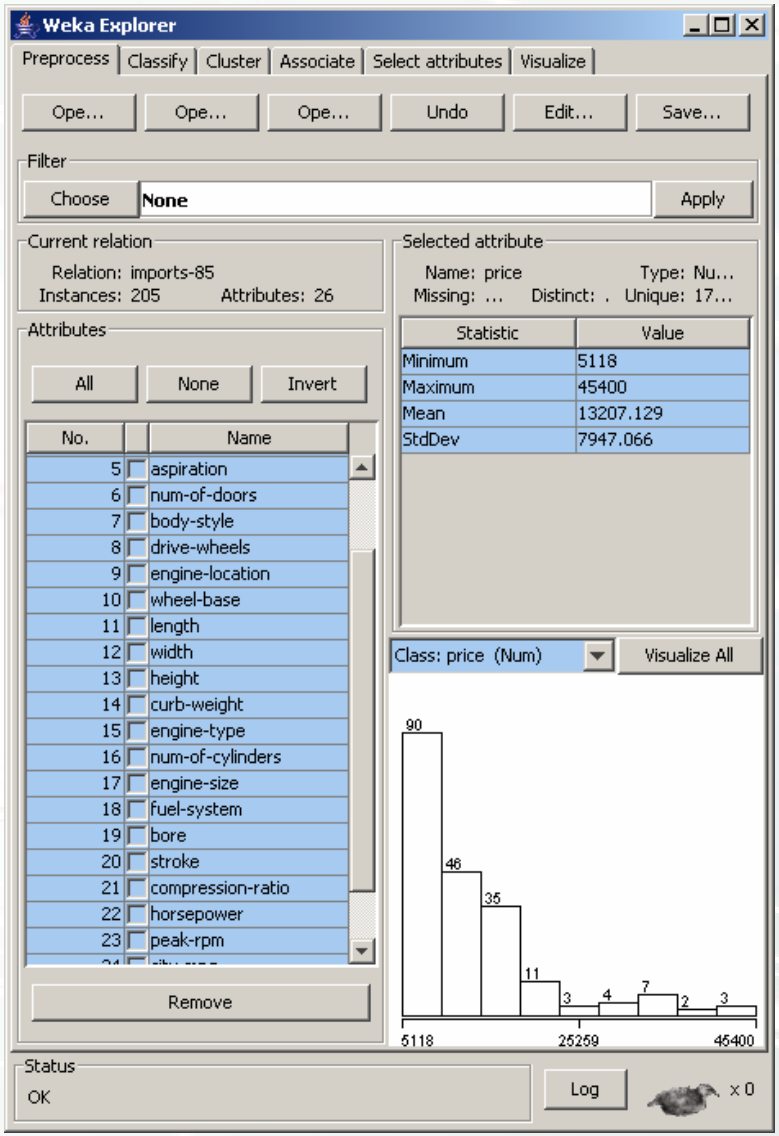
Imports-85 dataset

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Petra.Kralj@ijs.si

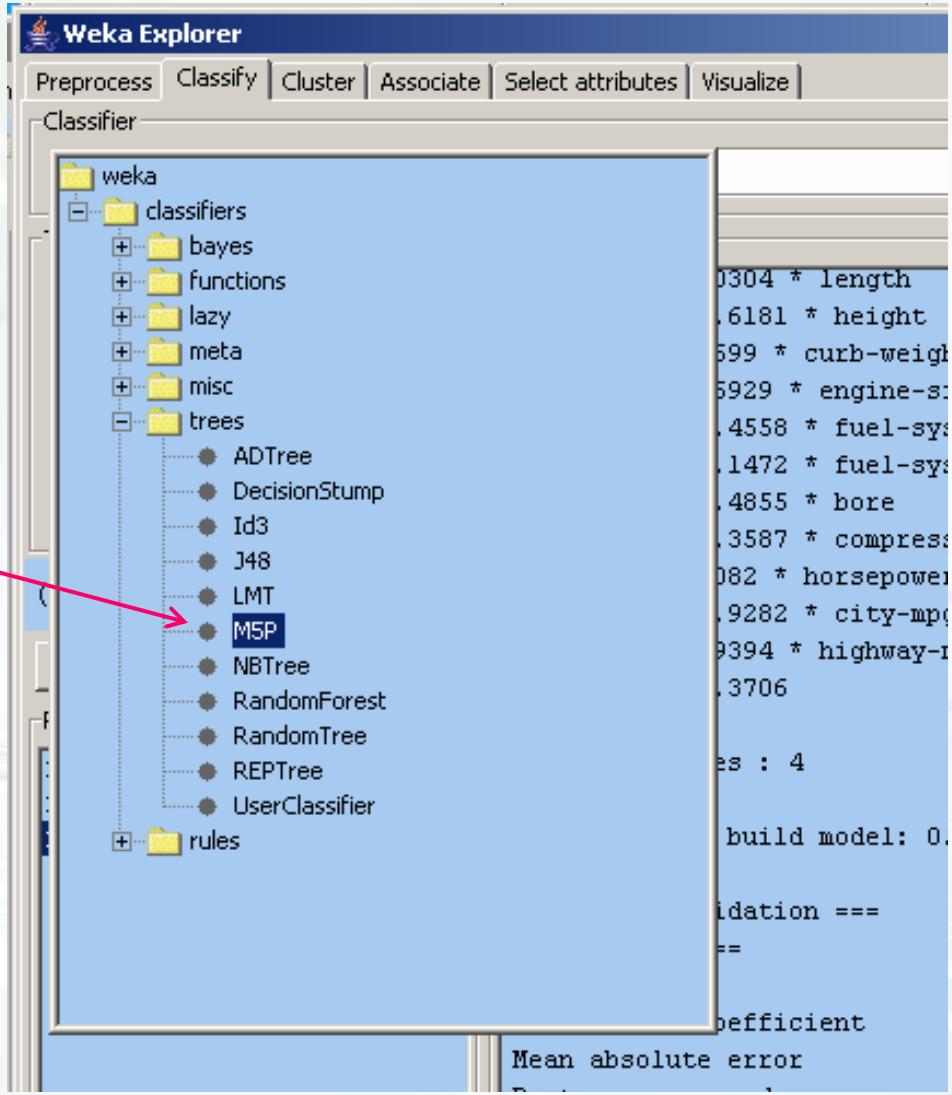
Imports-85 dataset

- 205 instances
- 25 attributes
  - 11 nominal attributes
  - 14 numeric attributes
- Continuous target variable
  - Distribution:
    - Minimum 5118
    - Maximum 45400
    - Mean 13207.129
    - StdDev 7947
- Missing values
  - 4 (2%)

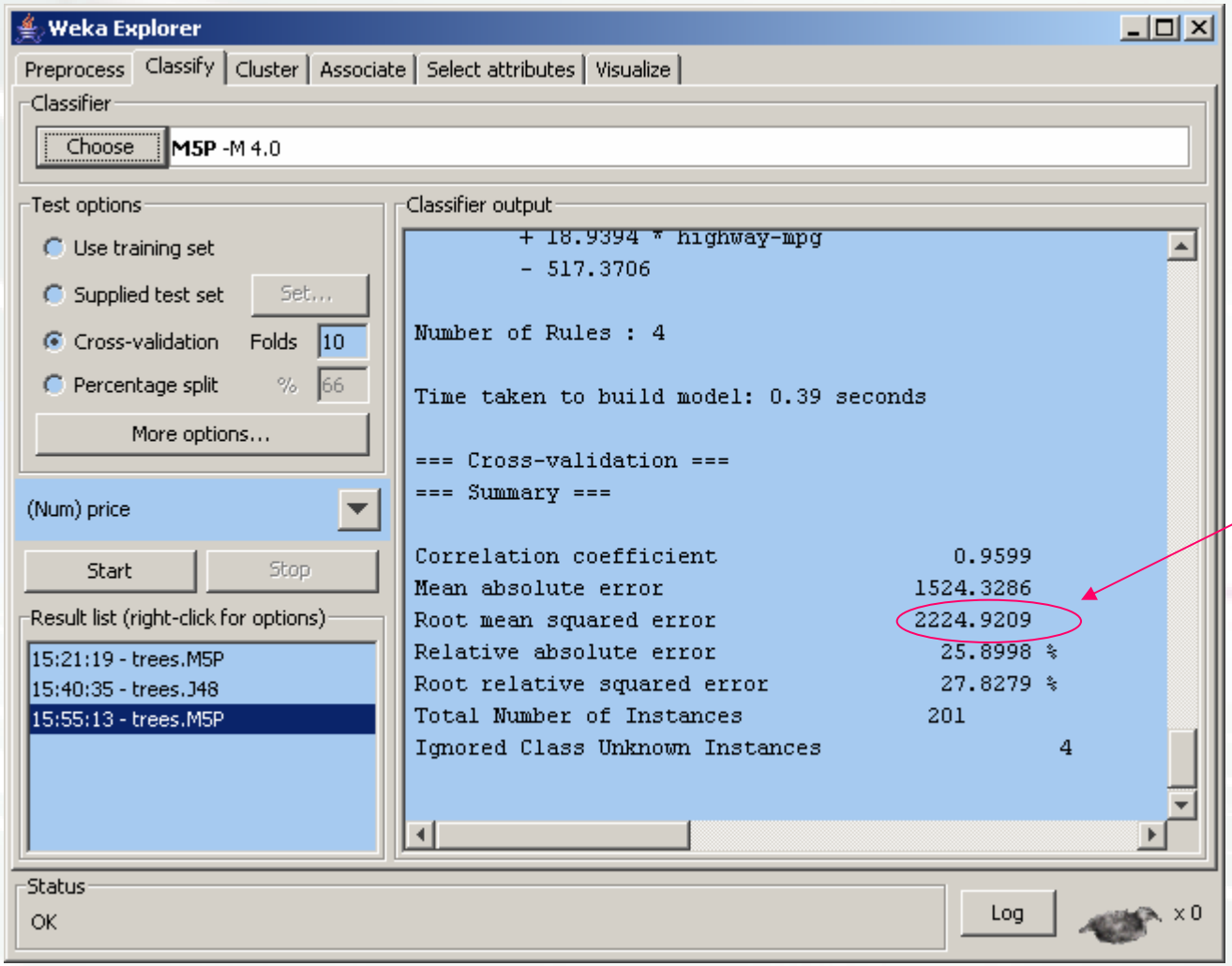


Petra.Kralj@ijs.si

Model & regression tree



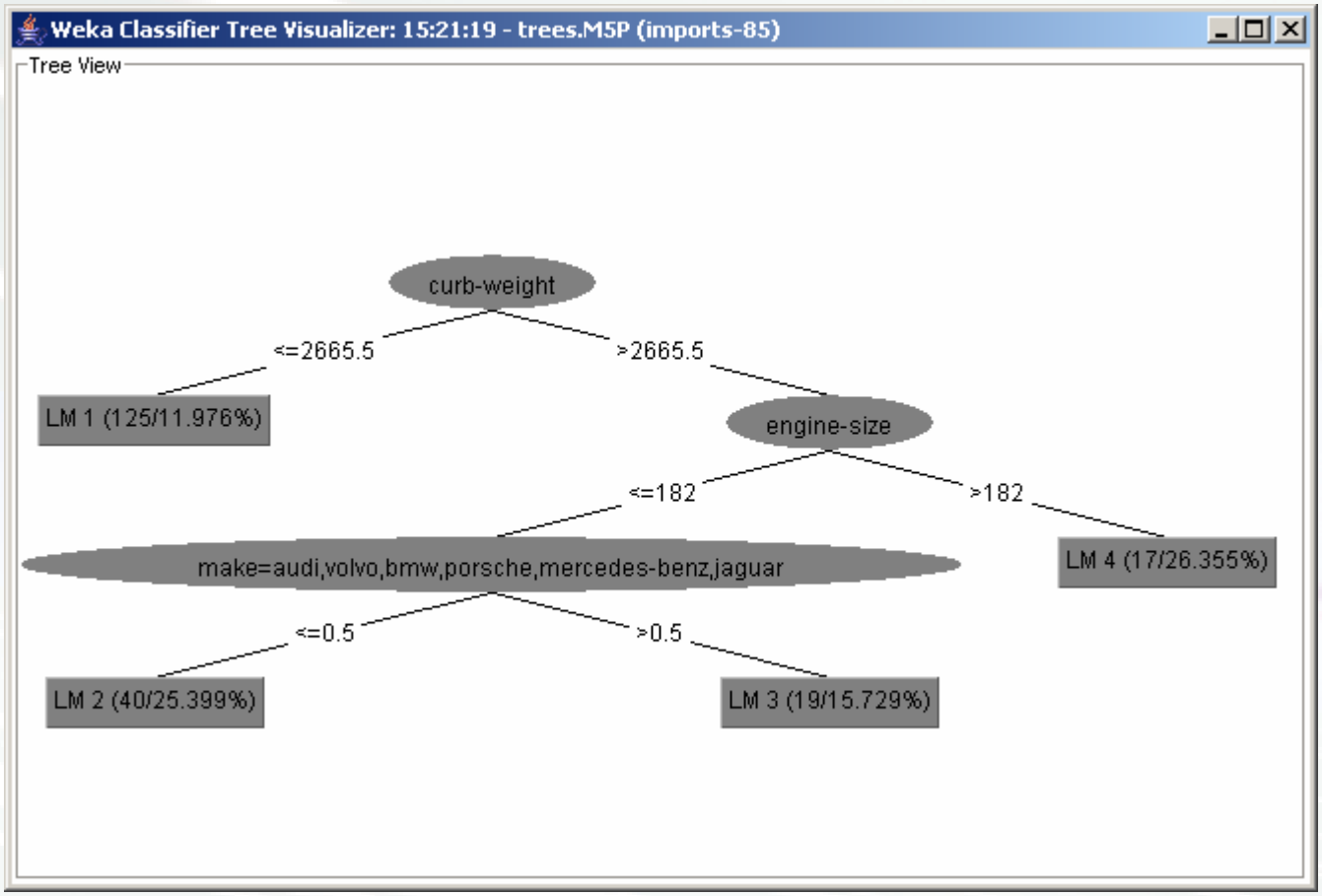
Evaluating regression model



Root Mean Squared Error

$$RMS = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2}$$

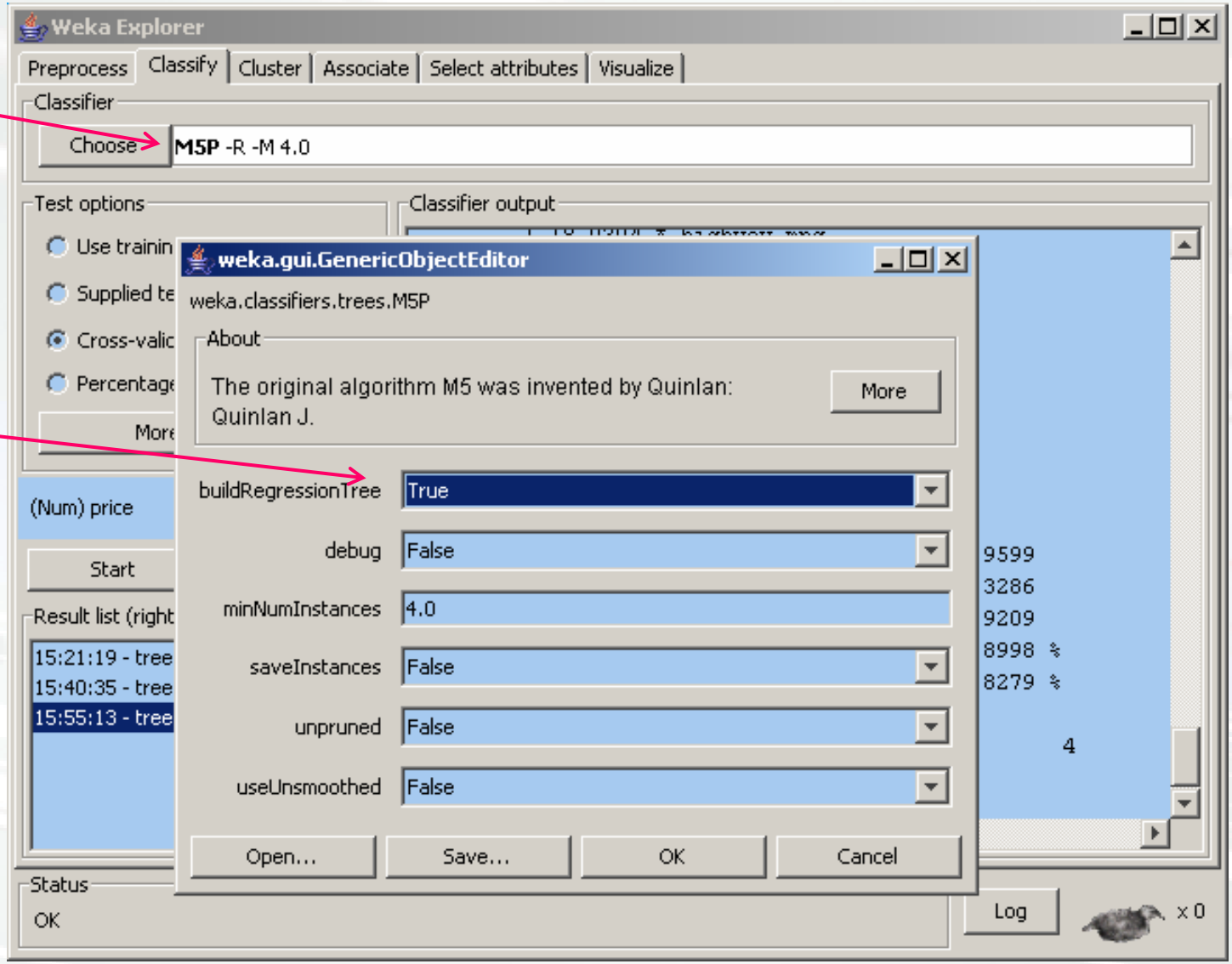
Model tree visualization



LM = linear model

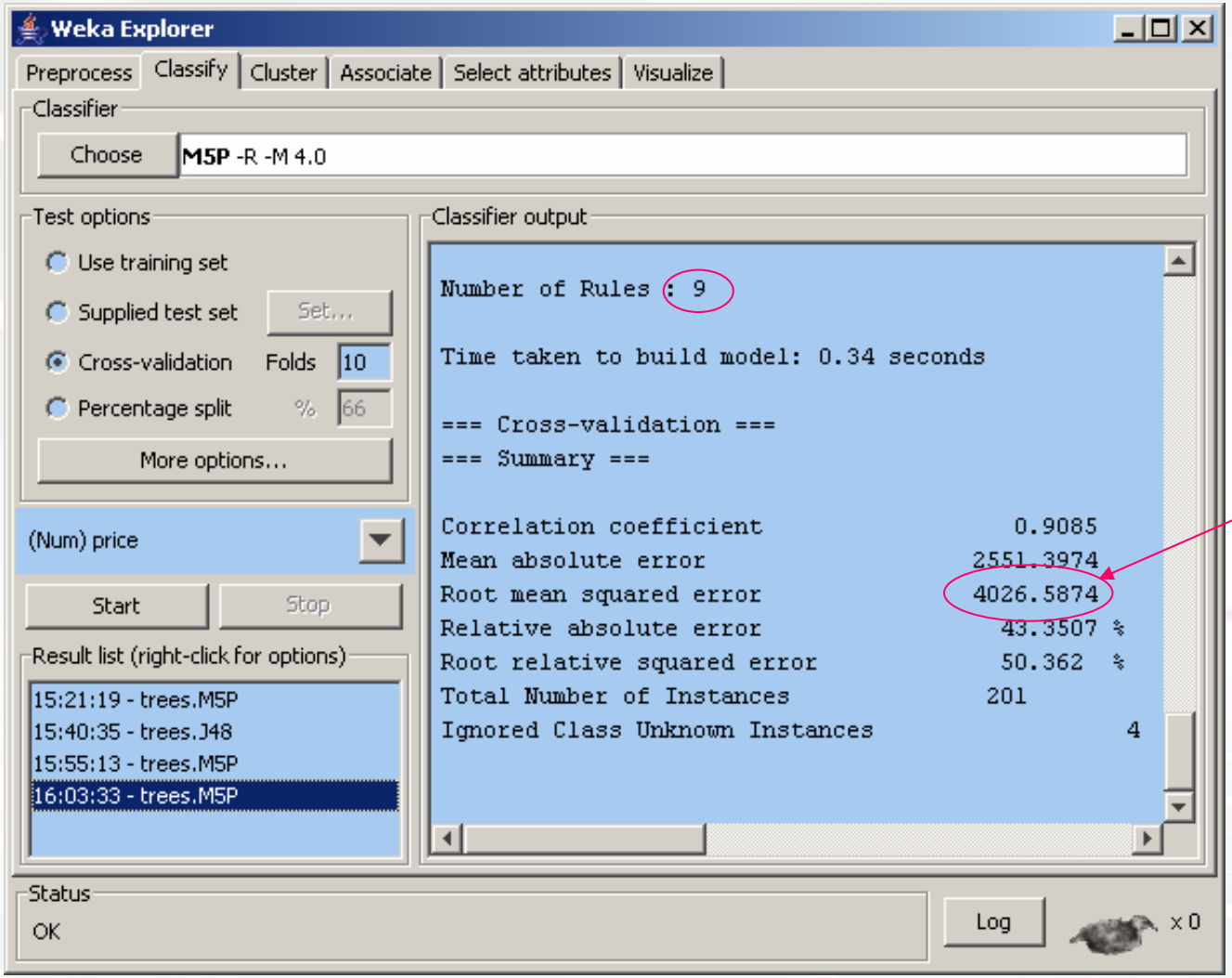
Petra.Kralj@ijs.si

Regression trees



Petra.Kralj@ijs.si

Evaluation of the regression tree



Larger error compared to the model tree

Petra.Kralj@ijs.si

Understandability of regression and model trees

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

M5 pruned regression tree:
(using smoothed linear models)

curb-weight <= 2665.5 :
|   curb-weight <= 2291.5 :
|   |   curb-weight <= 2121 :
|   |   |   length <= 160.75 : LM1 (27/8.596%)
|   |   |   length >  160.75 : LM2 (17/7.78%)
|   |   curb-weight >  2121 : LM3 (27/11.735%)
|   curb-weight >  2291.5 :
|   |   fuel-system=spfi,4bbl,mfi,idi,mpfi <= 0.5 : LM4 (21/13.739%)
|   |   fuel-system=spfi,4bbl,mfi,idi,mpfi >  0.5 :
|   |   |   make=volkswagen,nissan,mazda,saab,peugot,alfa-romero,mercury,audi,volvo,bmw,porsche,mercedes-benz,jaguar <= 0.5 : LM5 (16/13.
|   |   |   make=volkswagen,nissan,mazda,saab,peugot,alfa-romero,mercury,audi,volvo,bmw,porsche,mercedes-benz,jaguar >  0.5 : LM6 (17/28.
curb-weight >  2665.5 :
|   engine-size <= 182 :
|   |   make=audi,volvo,bmw,porsche,mercedes-benz,jaguar <= 0.5 : LM7 (40/32.152%)
|   |   make=audi,volvo,bmw,porsche,mercedes-benz,jaguar >  0.5 : LM8 (19/40.389%)
|   engine-size >  182 : LM9 (17/62.212%)
```

Petra.Kralj@ijs.si

Descriptive induction

Voting dataset

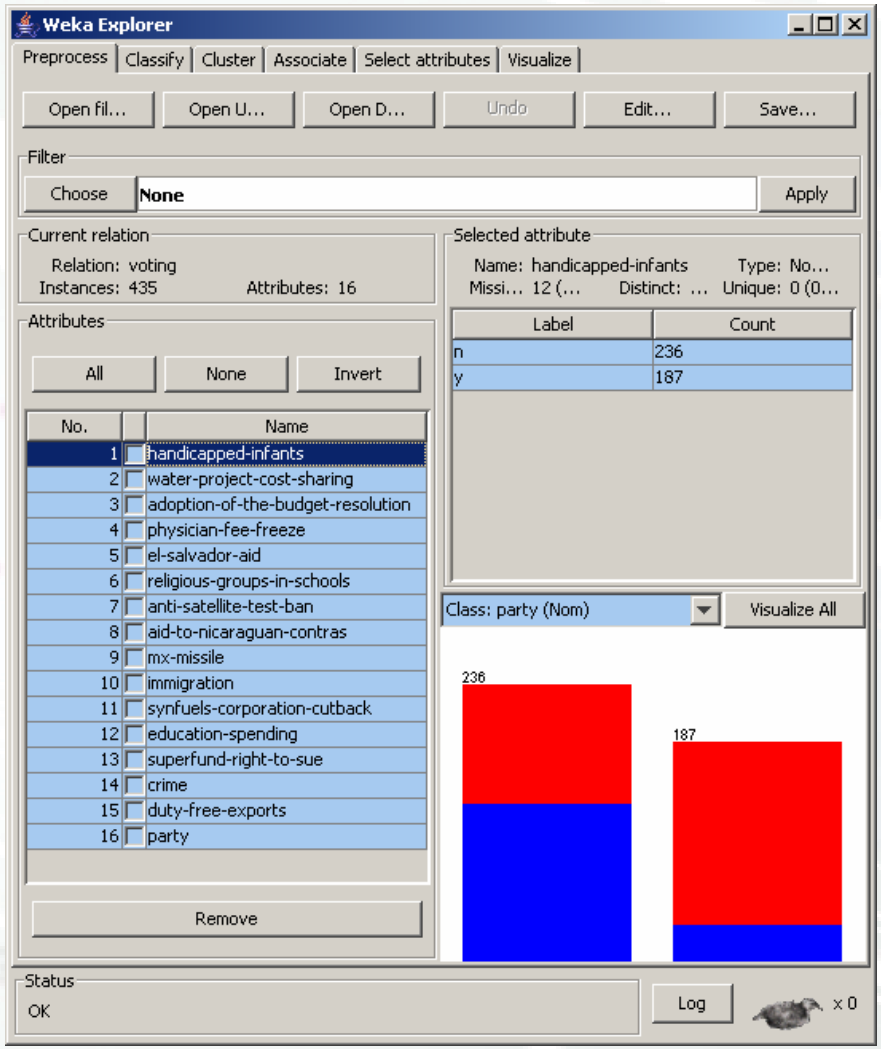
Iris dataset

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Petra.Kralj@ijs.si

Voting dataset

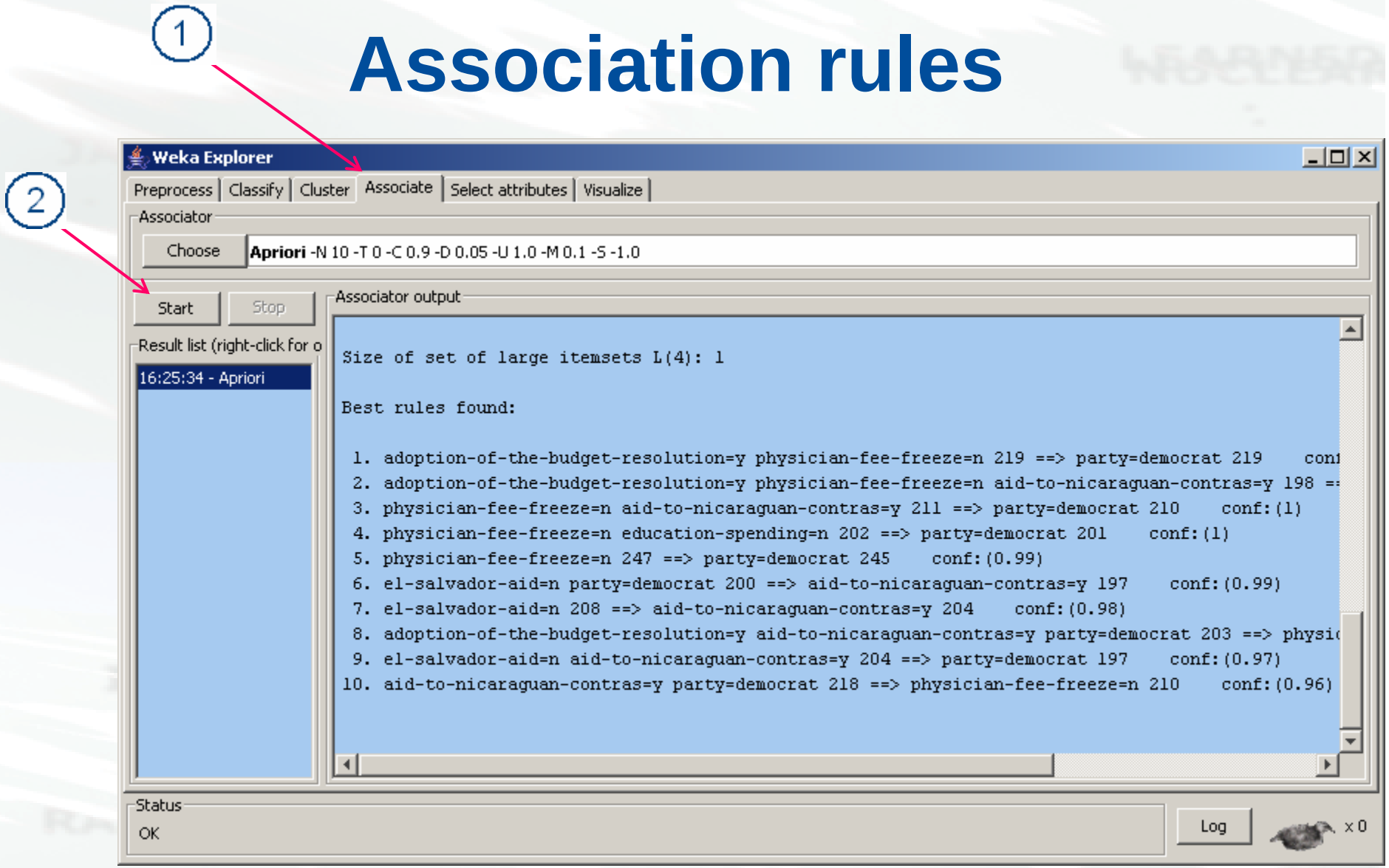
- 435 instances
- 16 attributes
  - 16 nominal attributes
  - 0 numeric attributes
- No target variable
- No missing values



DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Petra.Kralj@ijs.si

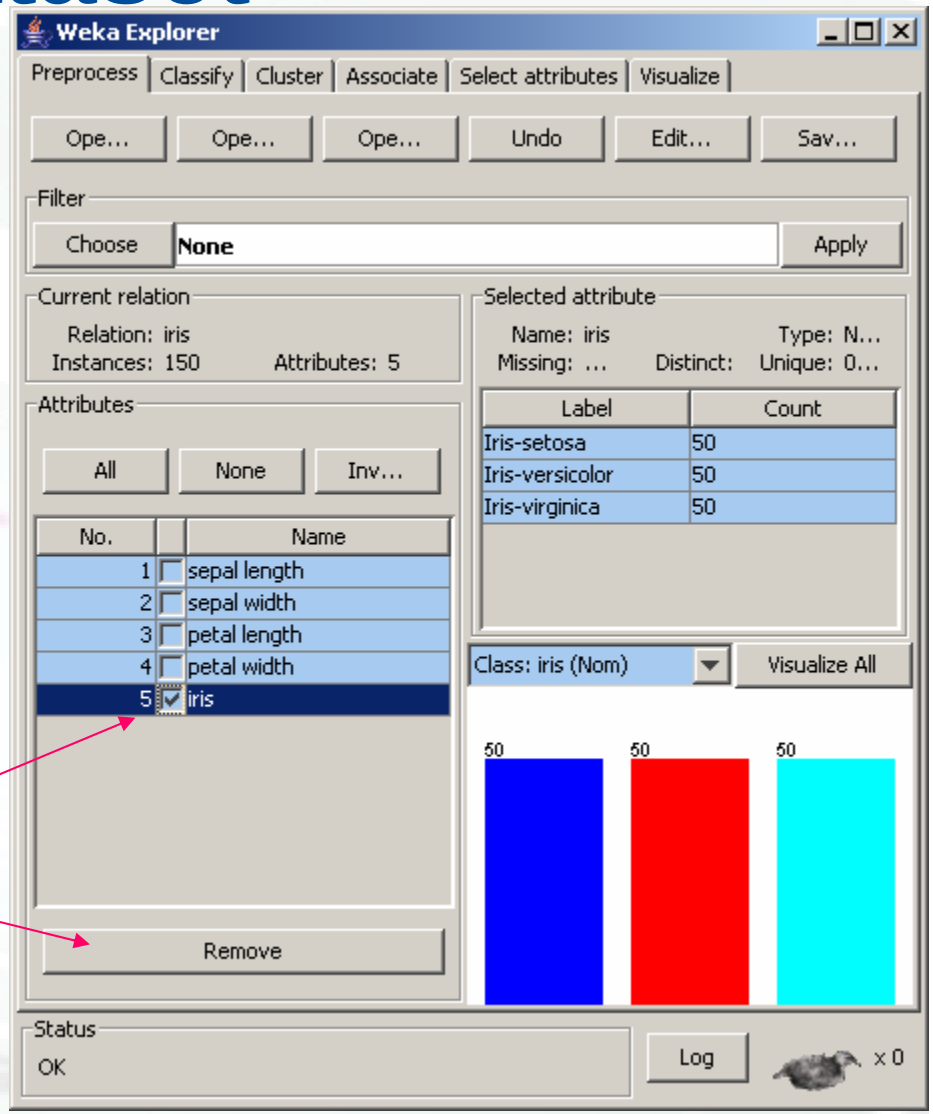
Association rules



Petra.Kralj@ijs.si

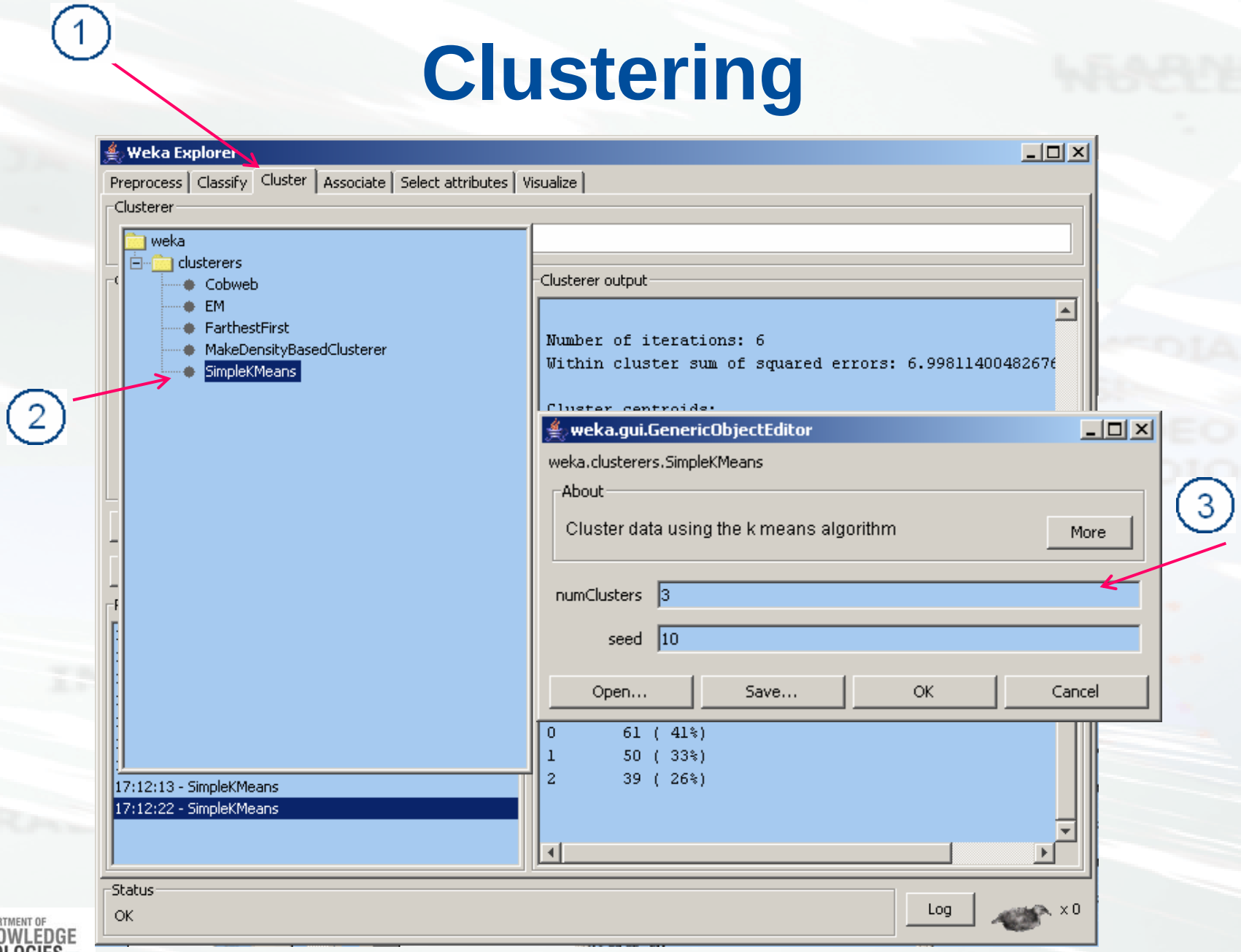
Iris dataset

- 150 instances
- 4 attributes
  - 0 nominal attributes
  - 4 numeric attributes
- Nominal target variable
  - 3 values:
    - Iris-setosa (30%)
    - Iris-versicolor (30%)
    - Iris-virginica (30%)
- No missing values



Petra.Kralj@ijs.si

Clustering



DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Petra.Kralj@ijs.si

Clustering visualization