Klasifikacija v WEKI

16.3.2007

Petra Kralj
Petra.Kralj@ijs.si

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Petra.Kralj@ijs.si

Vaje z Weko

- Ponovimo primer s kontaktnimi lečami z algoritmom ID3, testiranje:
  - Z ločeno testno množico
- Klasifikacija na CAR dataset
  - Priprava in branje podatkov
  - Gradnja odločitvenih dreves
  - Naivni Bayesov klasifikator
  - Razumevanje rezultatov

Naloga

- V Weki z algoritmom ID3 zgradi odločitveno drevo na učni množici in izračunaj njegovo klasifikacijsko točnost na testni množici
- Podatki
  - LeceBinUcna.arff
  - LeceBinTestna.arff
- Primerjaj z rezultati, ki smo jih dobili pri ročnem računanju

Program WEKA

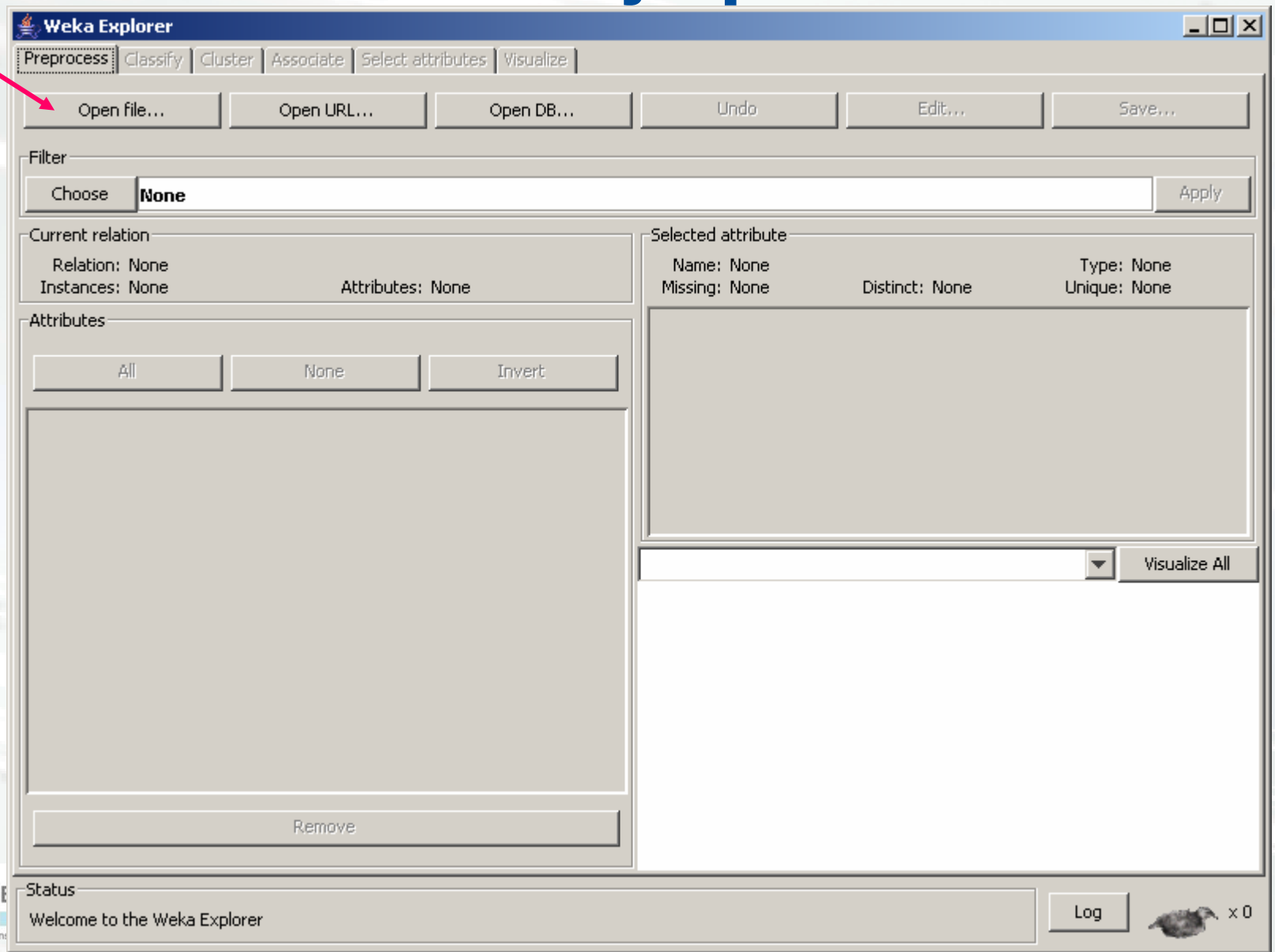
Prosto dostopen program za rudarjenje podatkov
http://www.cs.waikato.ac.nz/ml/weka/

download

Zagon programa Weka

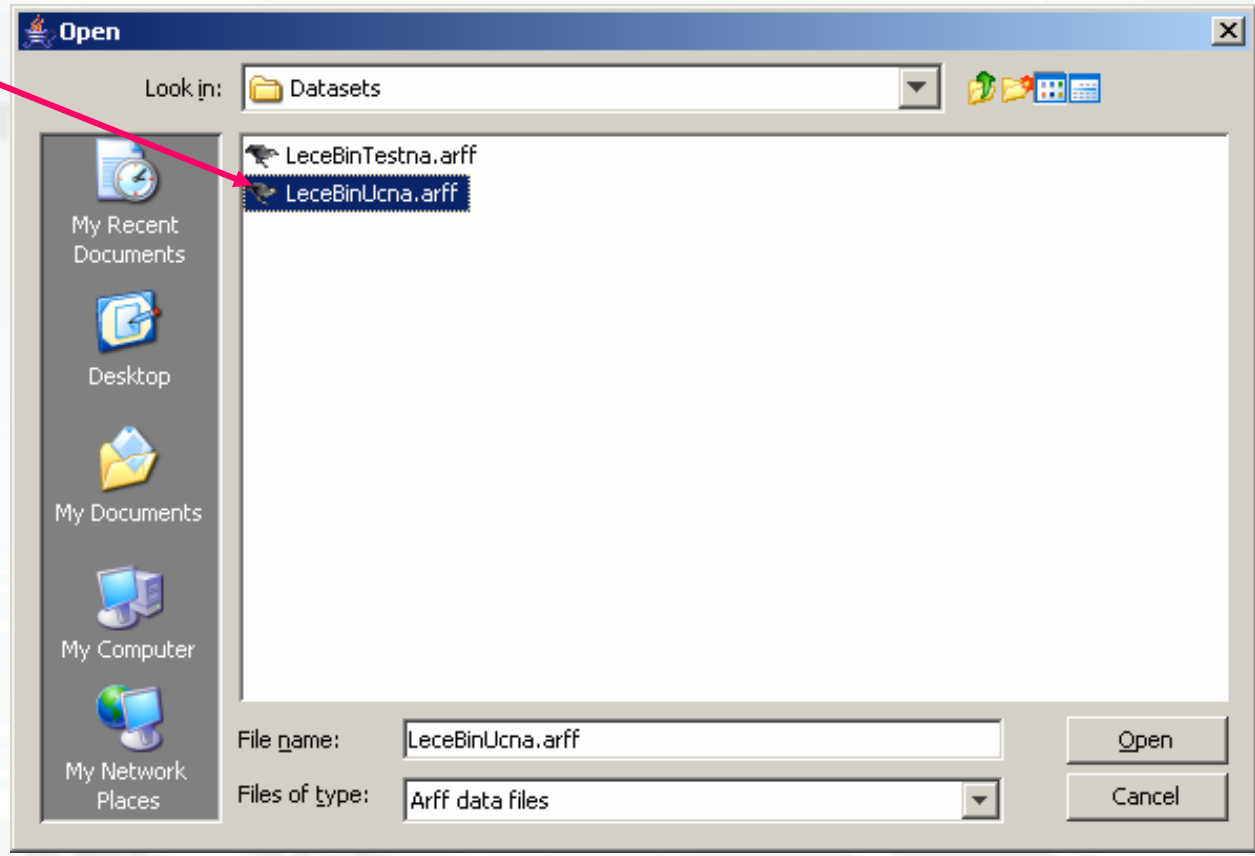
Izberemo Explorer

Okno za branje podatkov



Naložimo datoteko

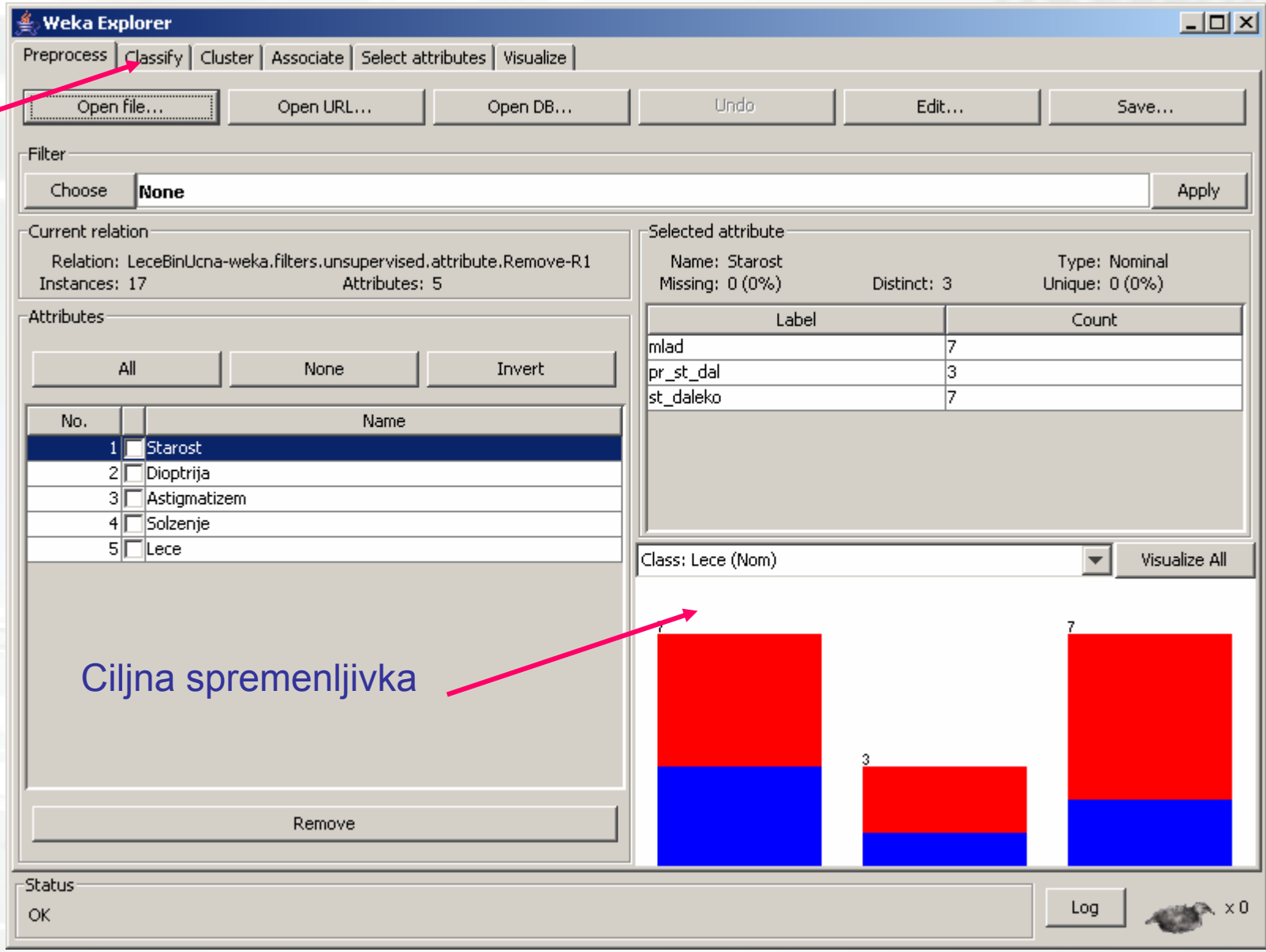
LeceBinUcna.arff



Petra.Kralj@ijs.si

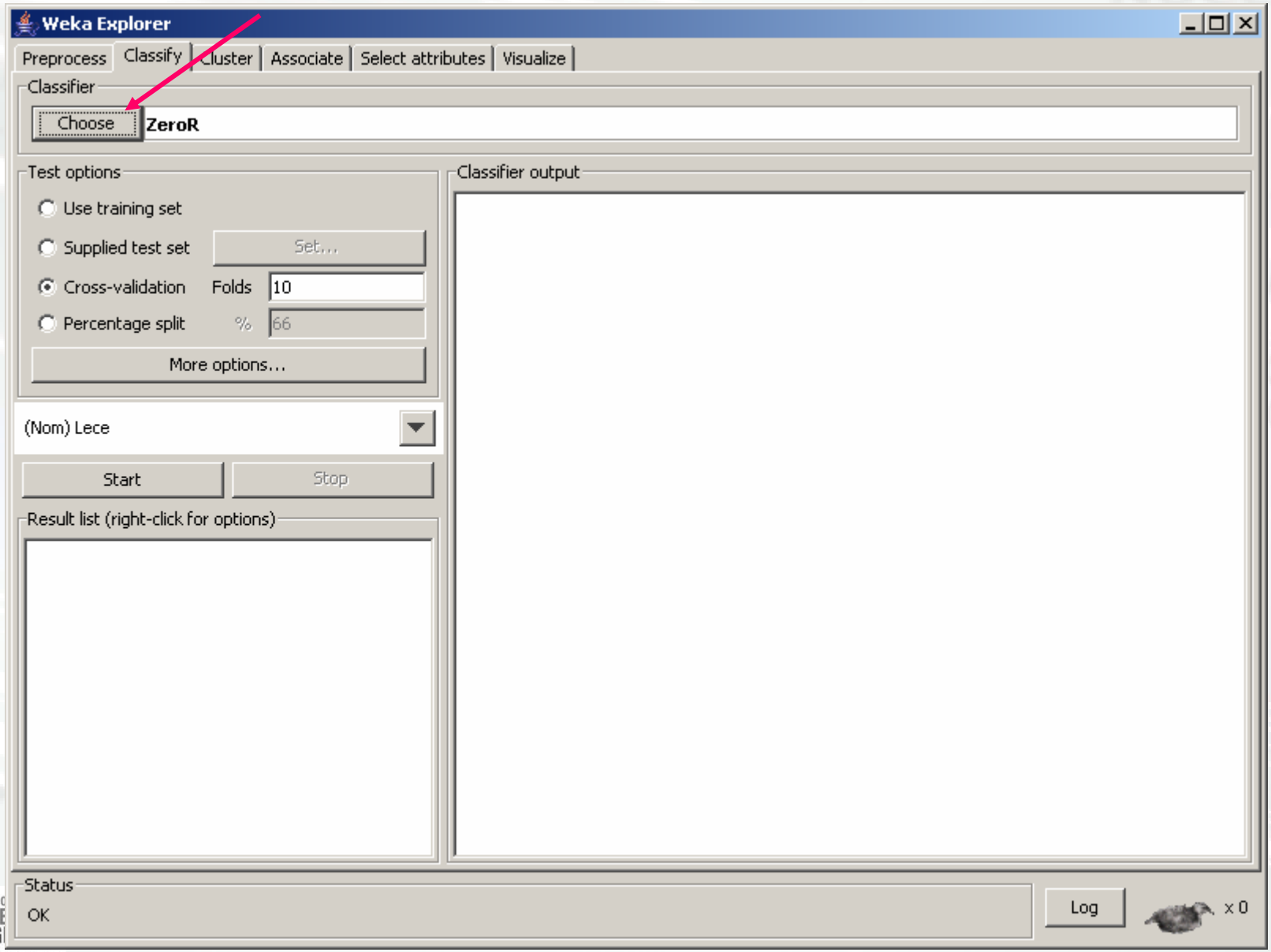
Pokažejo se nam podatki

Izberemo zavihek "Classify"



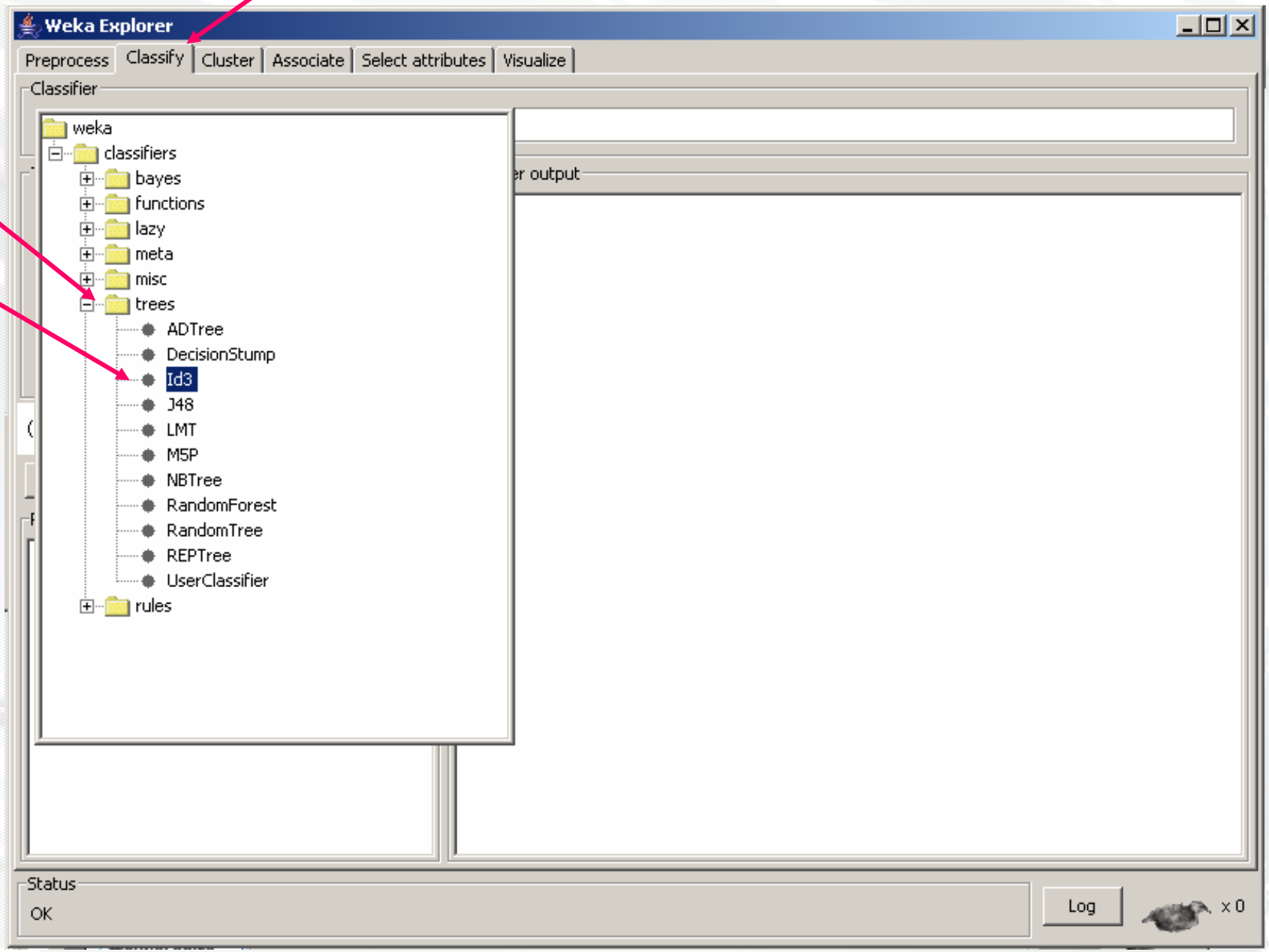
Petra.Kralj@ijs.si

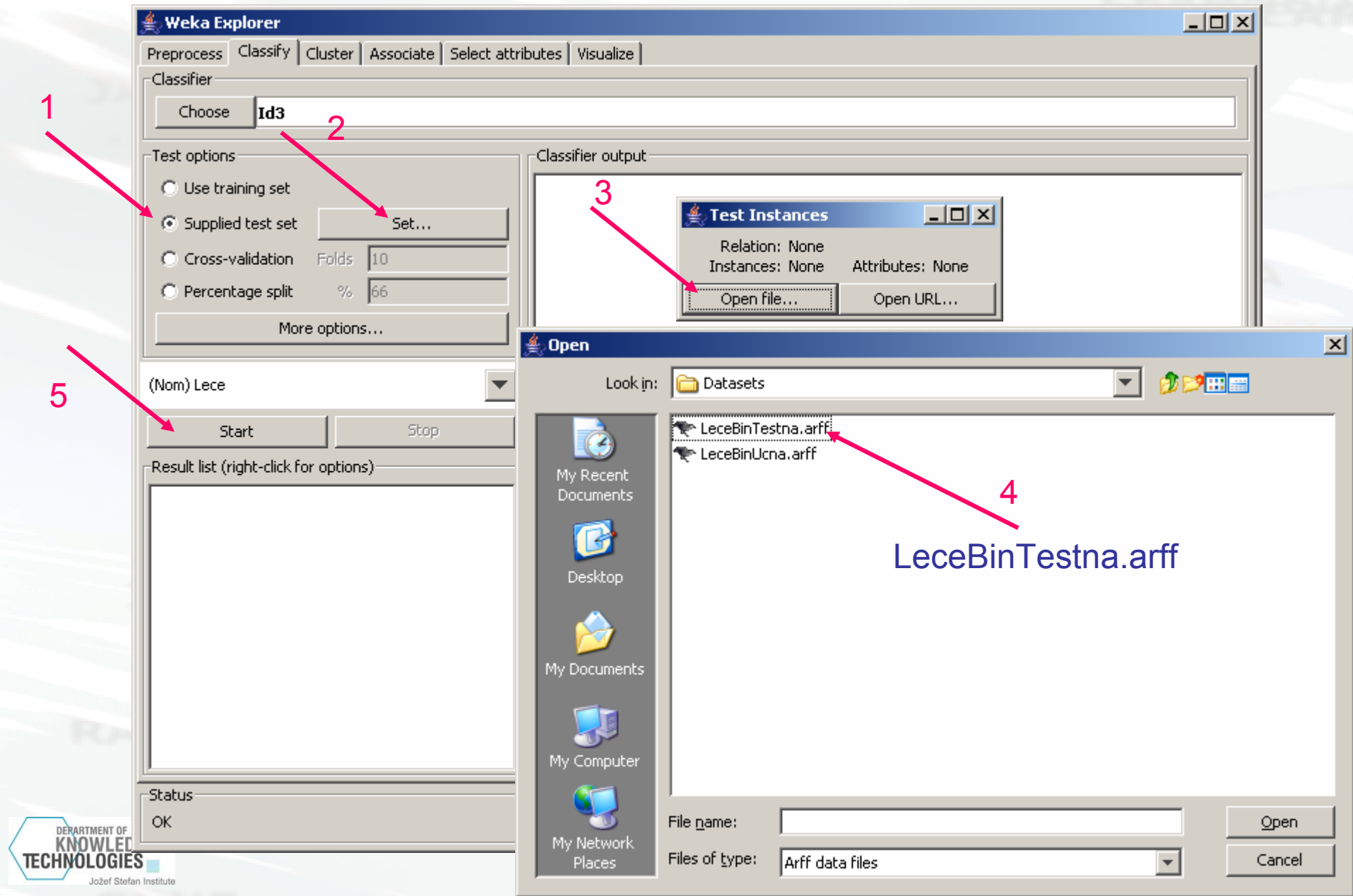
Izberemo algoritem

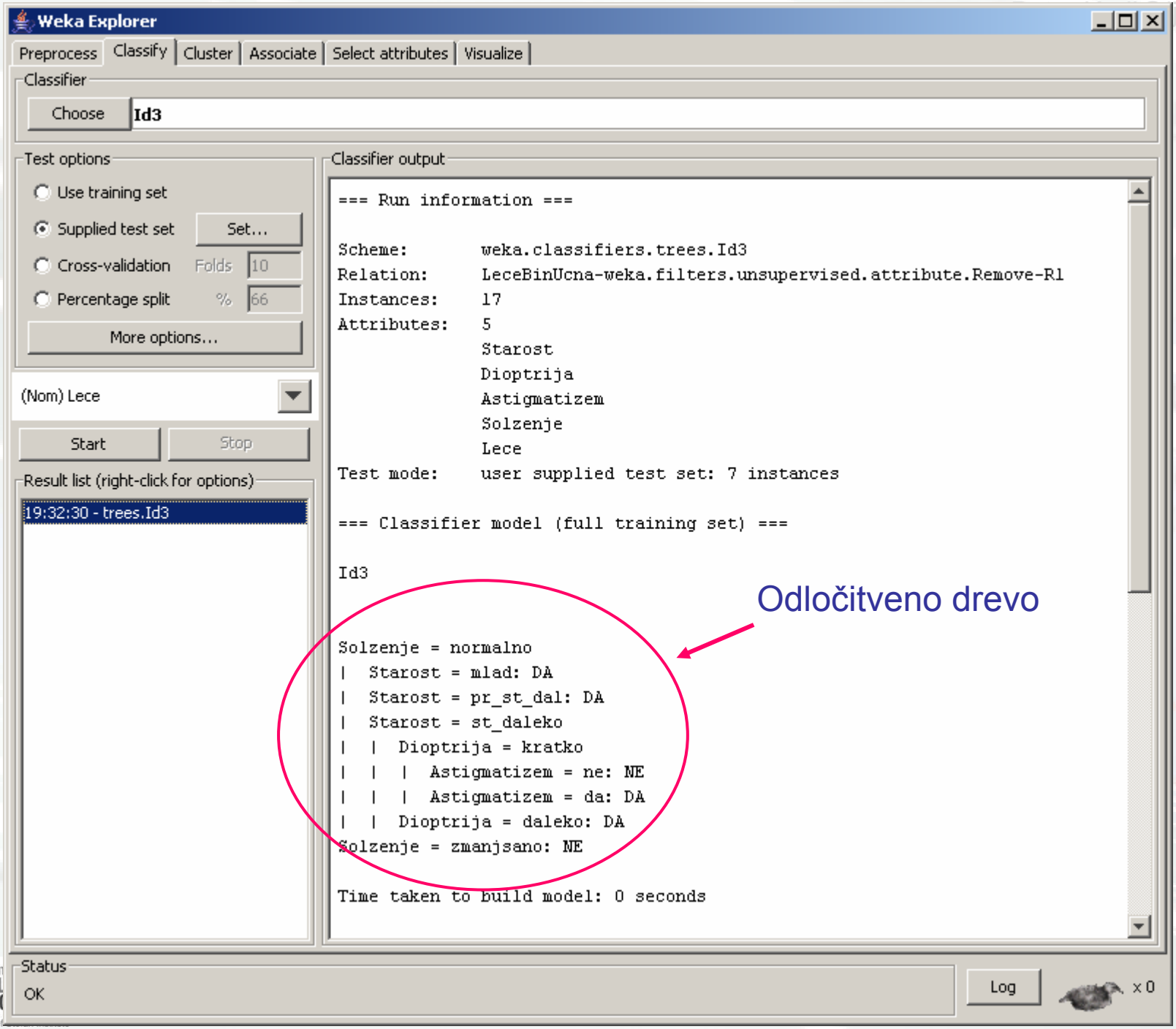


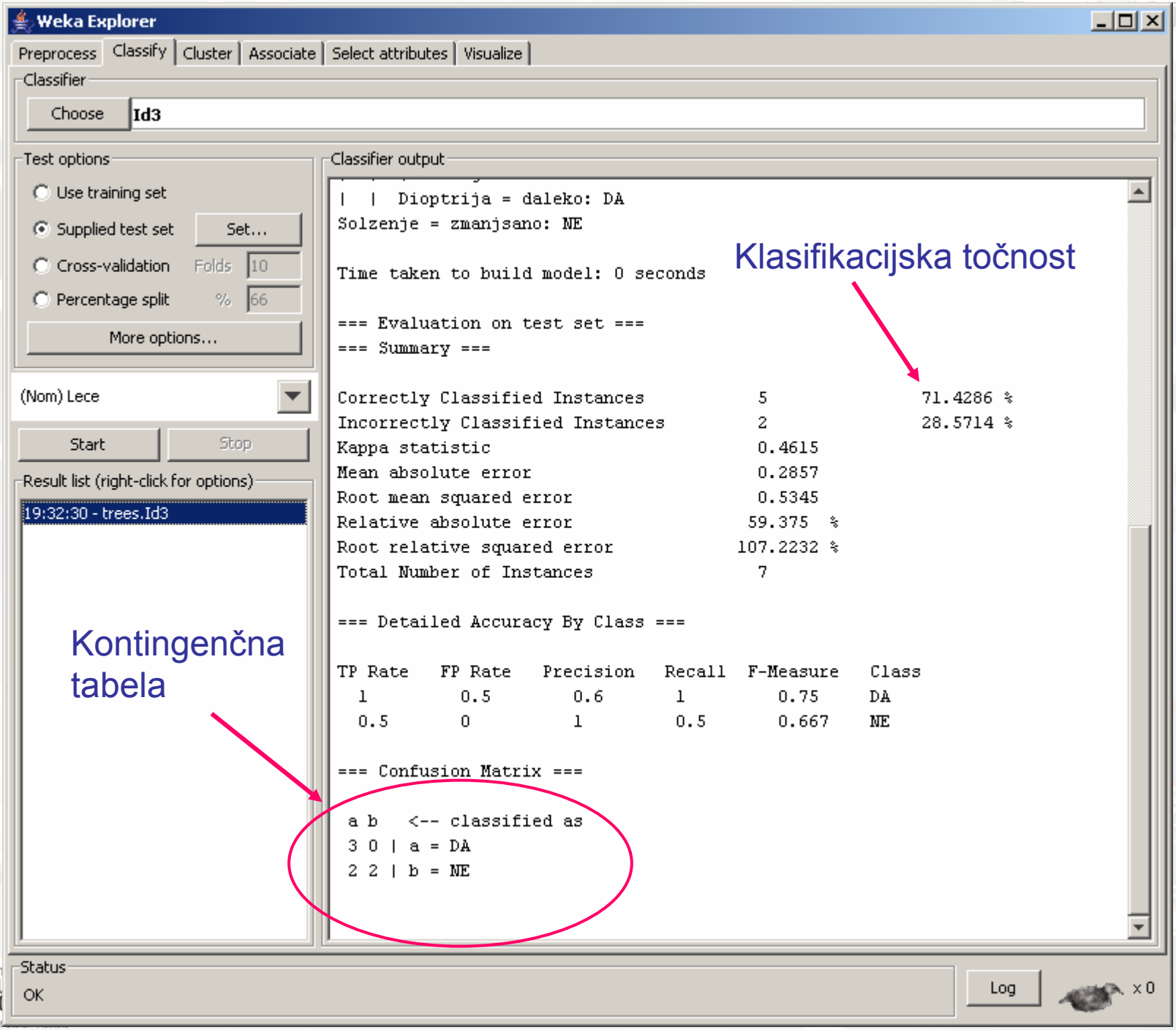
trees

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:
Datasets
My Recent Documents
Desktop
My Documents
My Computer
My Network Places
LeceBinTestna.arff
LeceBinUcna.arff
4
LeceBinTestna.arff
File name:
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)
19:32:30 - trees.Id3
Classifier output
=== Run information ===
Scheme: weka.classifiers.trees.Id3
Relation: LeceBinUcna-weka.filters.unsupervised.attribute.Remove-R1
Instances: 17
Attributes: 5
Starost
Dioptrija
Astigmatizem
Solzenje
Lece
Test mode: user supplied test set: 7 instances
=== Classifier model (full training set) ===
Id3
Solzenje = normalno
| Starost = mlad: DA
| Starost = pr_st_dal: DA
| Starost = st_daleko
| | Dioptrija = kratko
| | | Astigmatizem = ne: NE
| | | Astigmatizem = da: DA
| | Dioptrija = daleko: DA
Solzenje = zmanjsano: NE
Time taken to build model: 0 seconds
Odločitveno drevo
Status
OK
Log
x 0


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)
19:32:30 - trees.Id3
Classifier output
| | Dioptrija = daleko: DA
Solzenje = zmanjsano: NE
Time taken to build model: 0 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 DA
0.5 0 1 0.5 0.667 NE
=== Confusion Matrix ===
a b <-- classified as
3 0 | a = DA
2 2 | b = NE
Klasifikacijska točnost
Kontingenčna tabela
Status
OK
Log
x 0

Petra.Kralj@ijs.si

CAR dataset

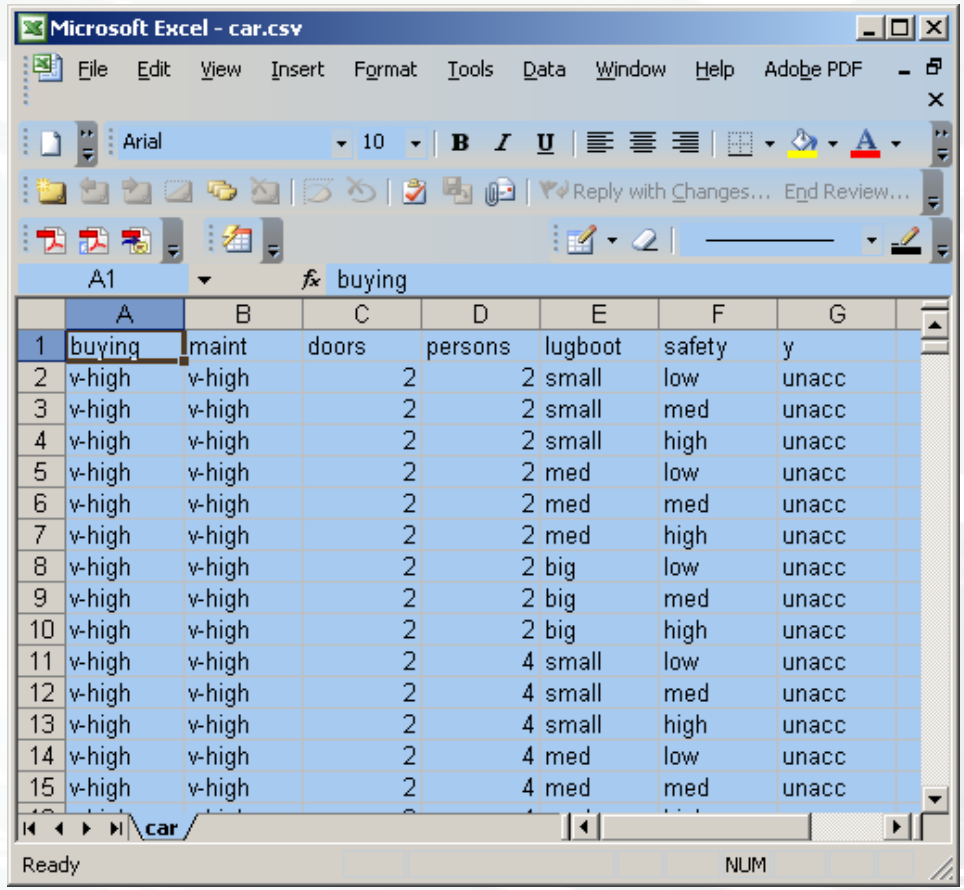
- 1728 primerov
- 6 atributov
  - 6 nominalnih
  - 0 numeričnih
- Nominalna ciljna spremenljivka
  - 4 vrednosti: unacc, acc, good, v-good
  - Distribucija vrednosti
    - unacc (70%), acc (22%), good (4%), v-good (4%)
- Brez manjkajočih vrednosti

Petra.Kralj@ijs.si

Priprava podatkov za WEKO - 1

Podatki v tabeli
(npr. MS Excel)

- Vrstice so primeri
- Stolpci so atributi
- Zadnji stolpec je ciljna spremenljivka



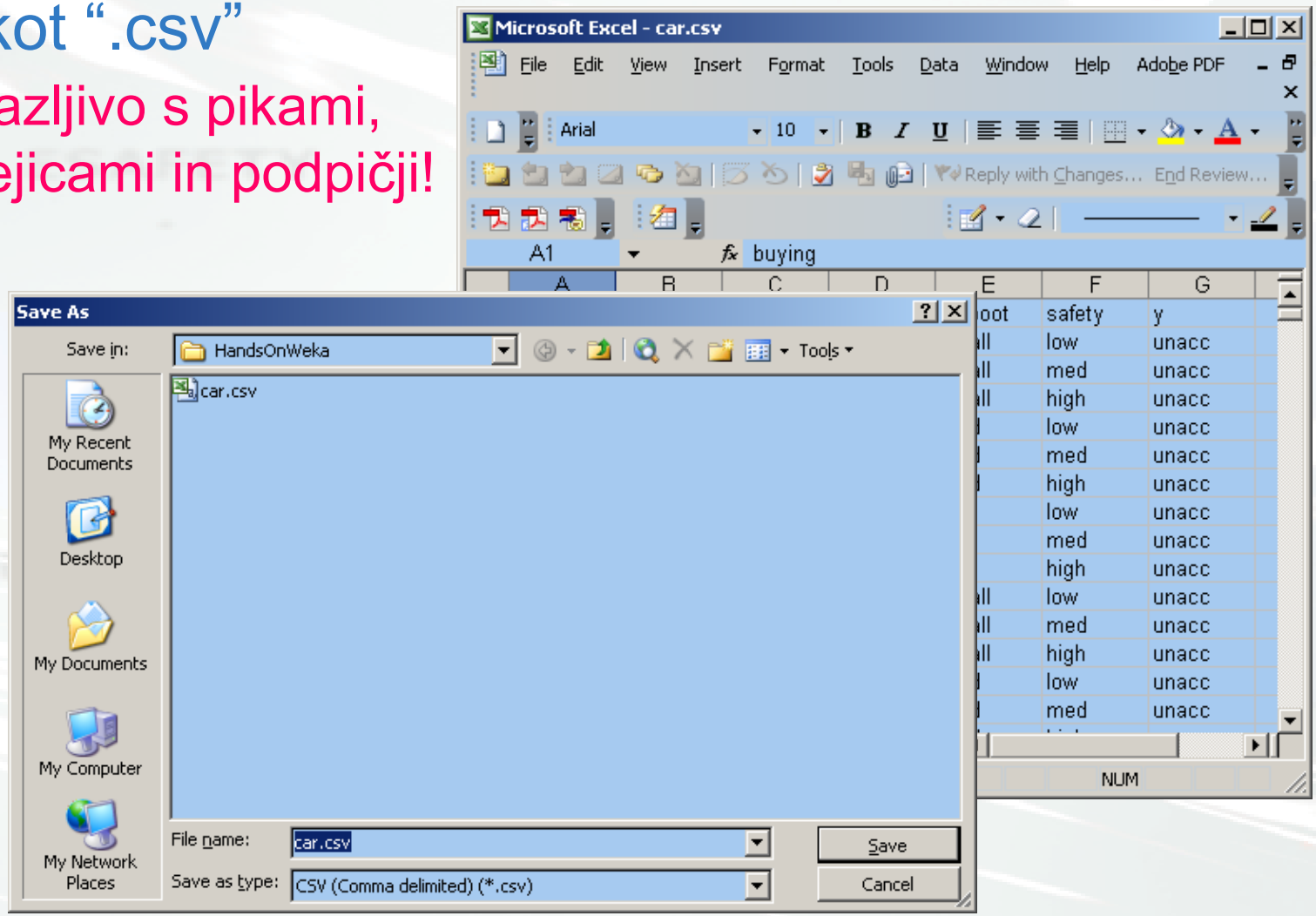
	A	B	C	D	E	F	G
1	buying	maint	doors	persons	lugboot	safety	y
2	v-high	v-high	2	2	small	low	unacc
3	v-high	v-high	2	2	small	med	unacc
4	v-high	v-high	2	2	small	high	unacc
5	v-high	v-high	2	2	med	low	unacc
6	v-high	v-high	2	2	med	med	unacc
7	v-high	v-high	2	2	med	high	unacc
8	v-high	v-high	2	2	big	low	unacc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

Priprava podatkov za WEKO - 2

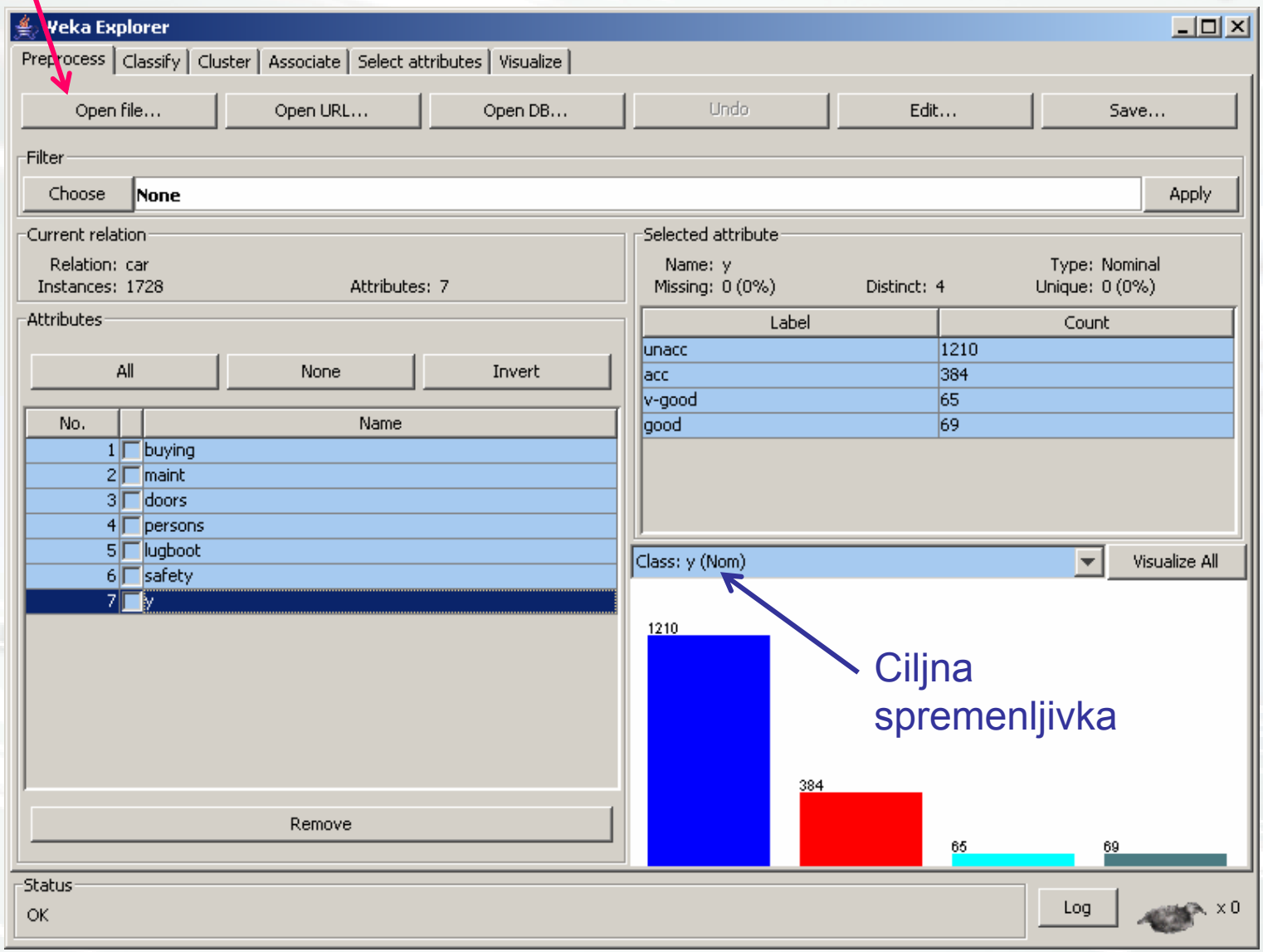
Shrani kot ".csv"

- Pazljivo s pikami, vejicami in podpičji!

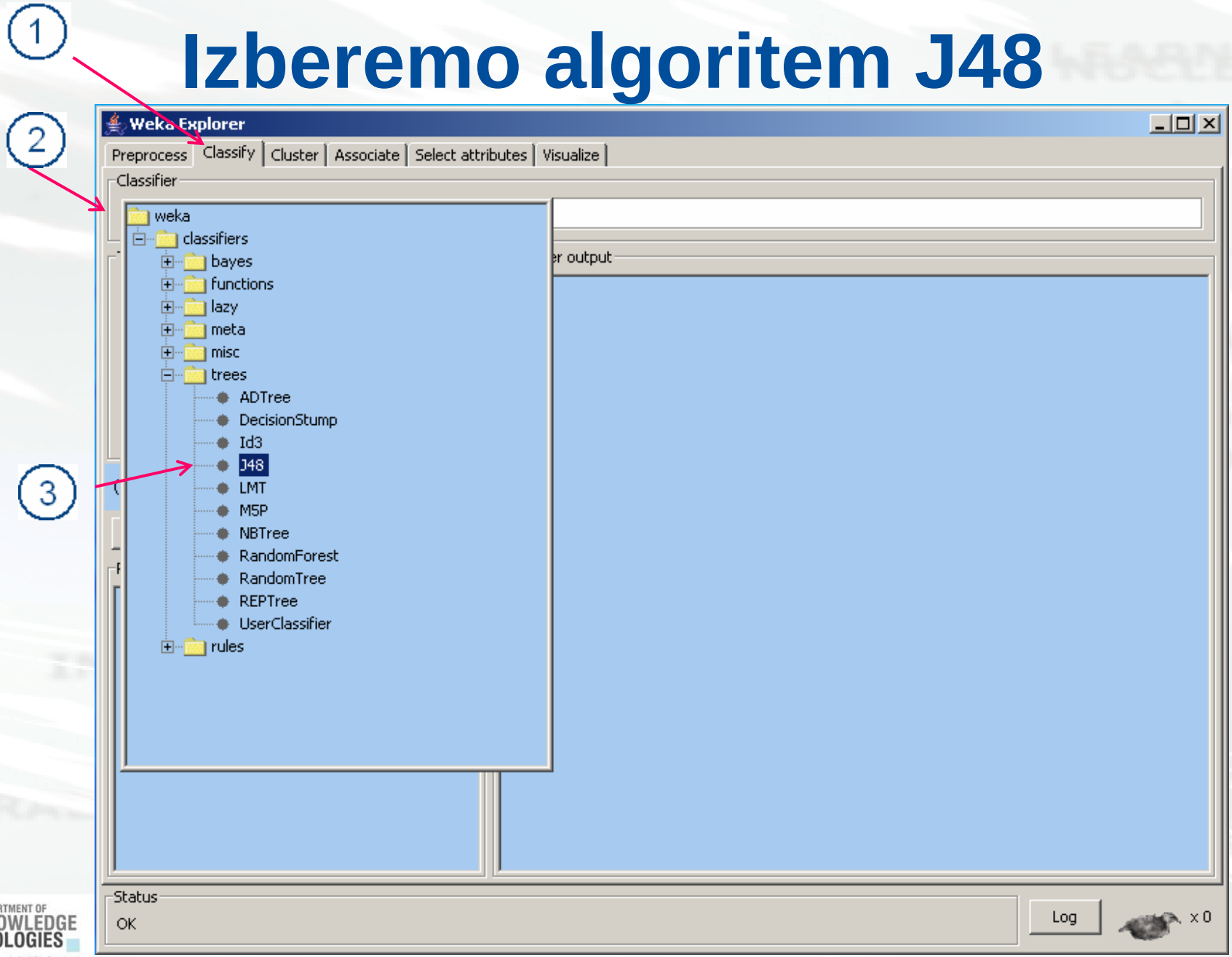


Car.csv

Load the data

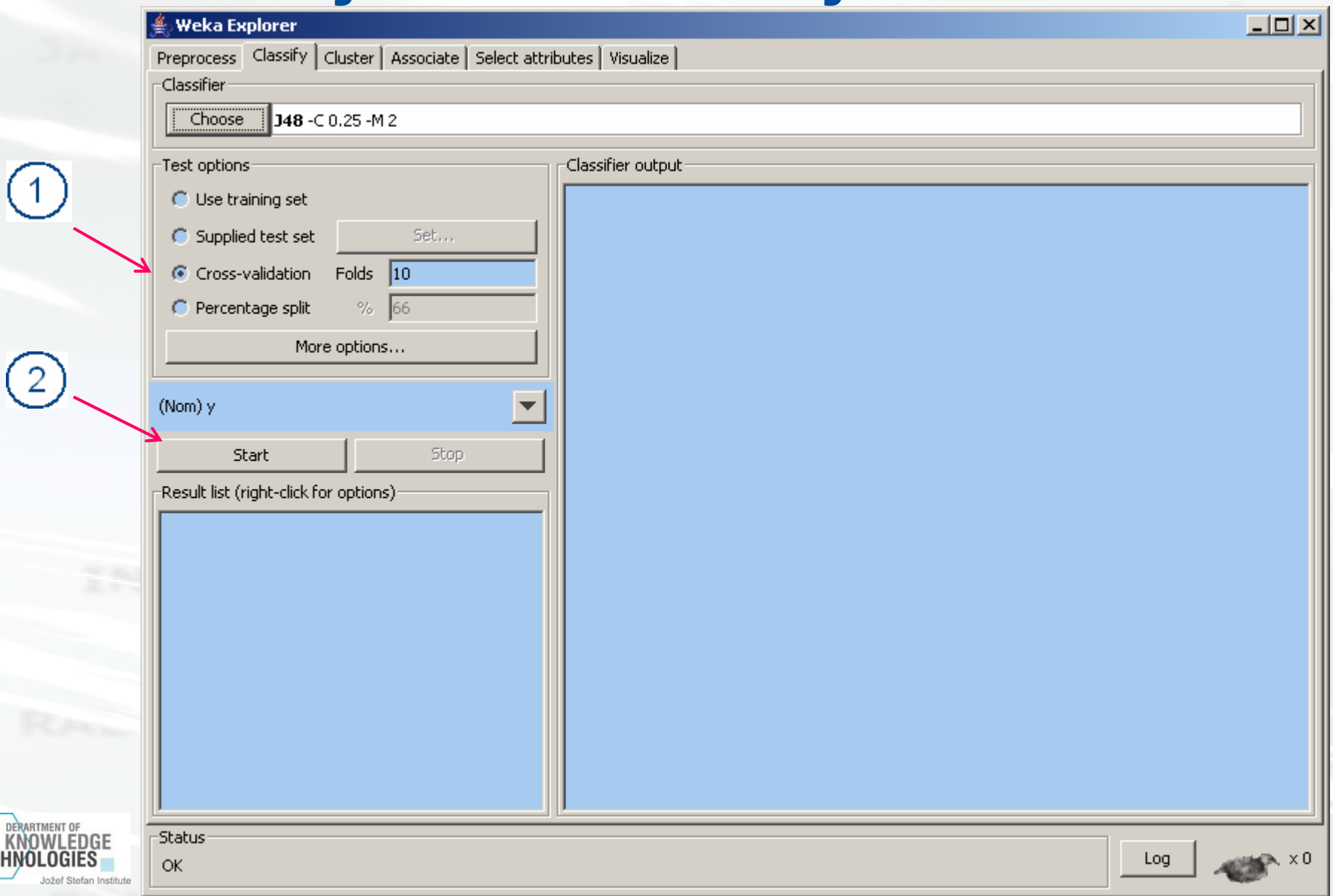


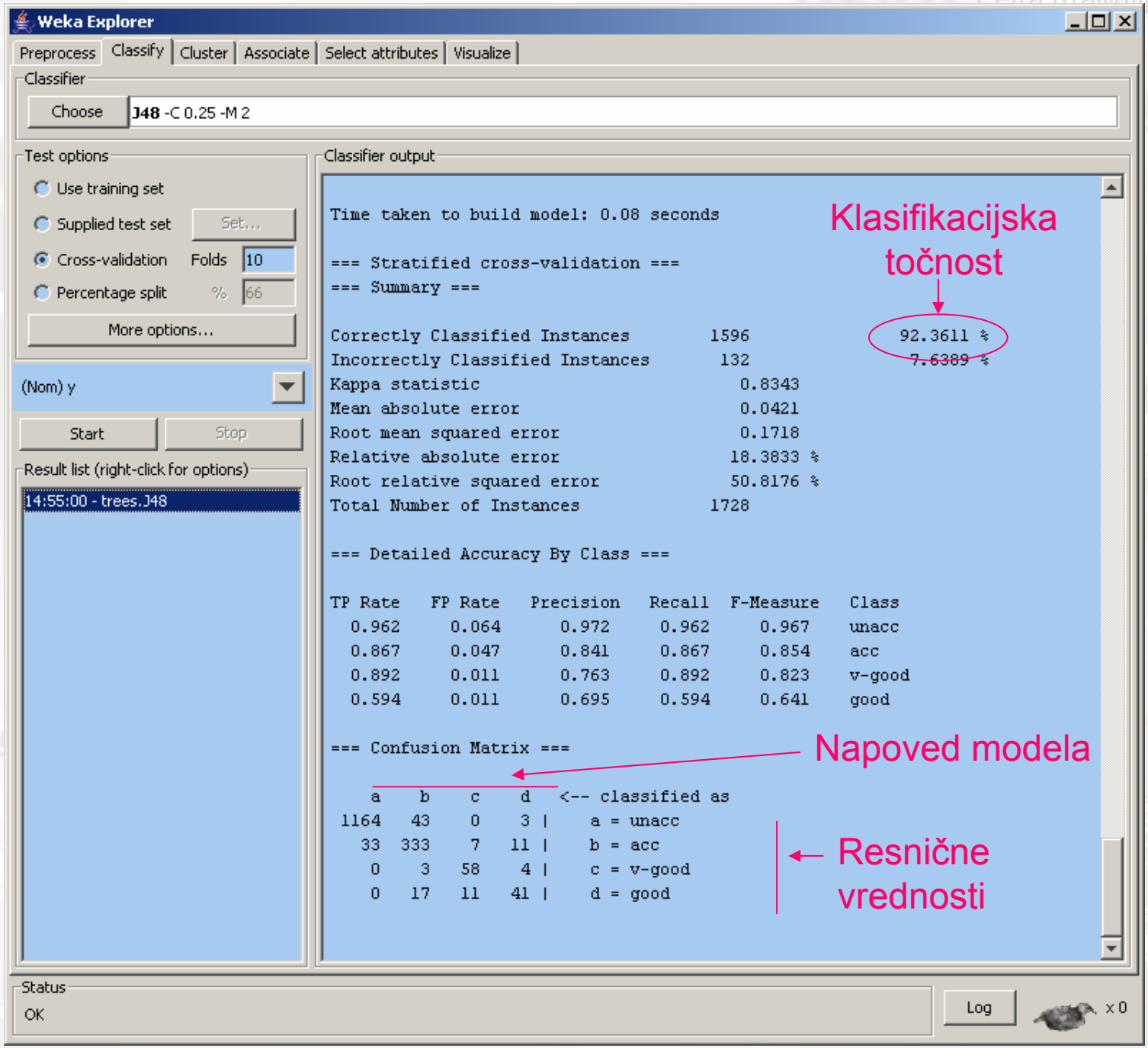
Izberemo algoritem J48

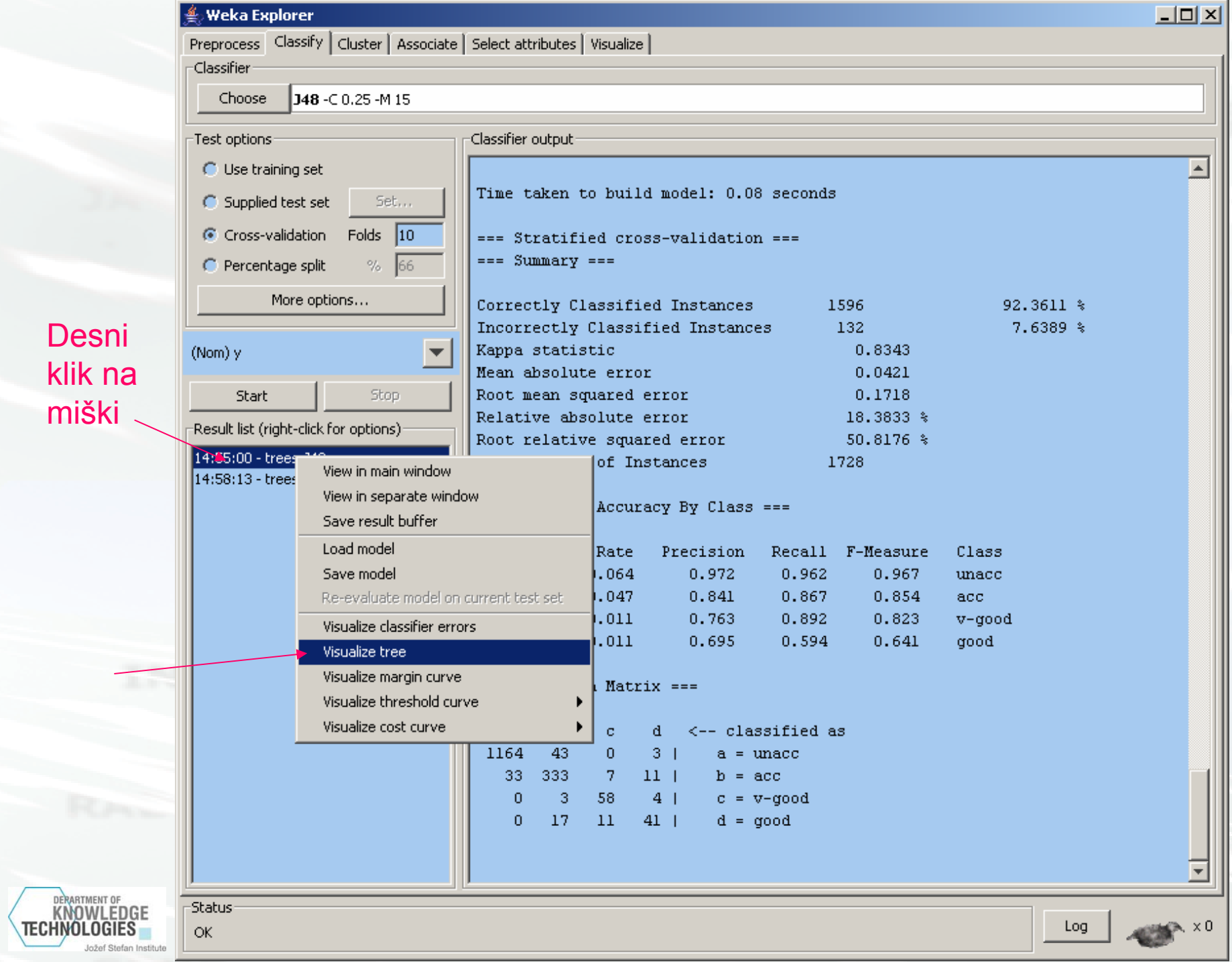


Petra.Kralj@ijs.si

Gradnja in evalvacija drevesa






Desni klik na miški
Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 15
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
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 %
of Instances 1728
Accuracy By Class ===
Rate Precision Recall F-Measure Class
.064 0.972 0.962 0.967 unacc
.047 0.841 0.867 0.854 acc
.011 0.763 0.892 0.823 v-good
.011 0.695 0.594 0.641 good
Matrix ===
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
(Nom) y
Start
Stop
Result list (right-click for options)
14:55:00 - trees
14:58:13 - trees
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

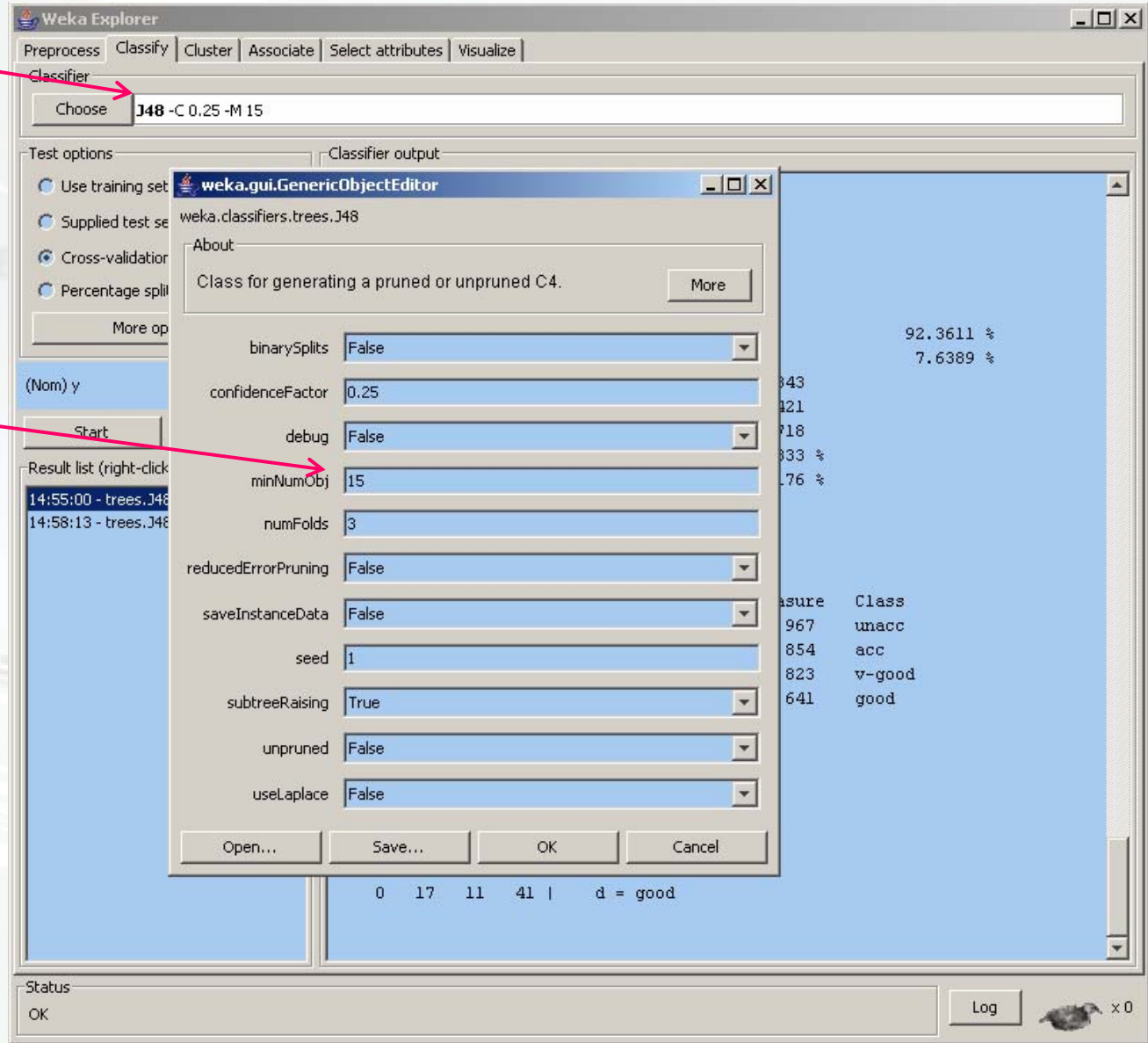
Rezanje dreves

1

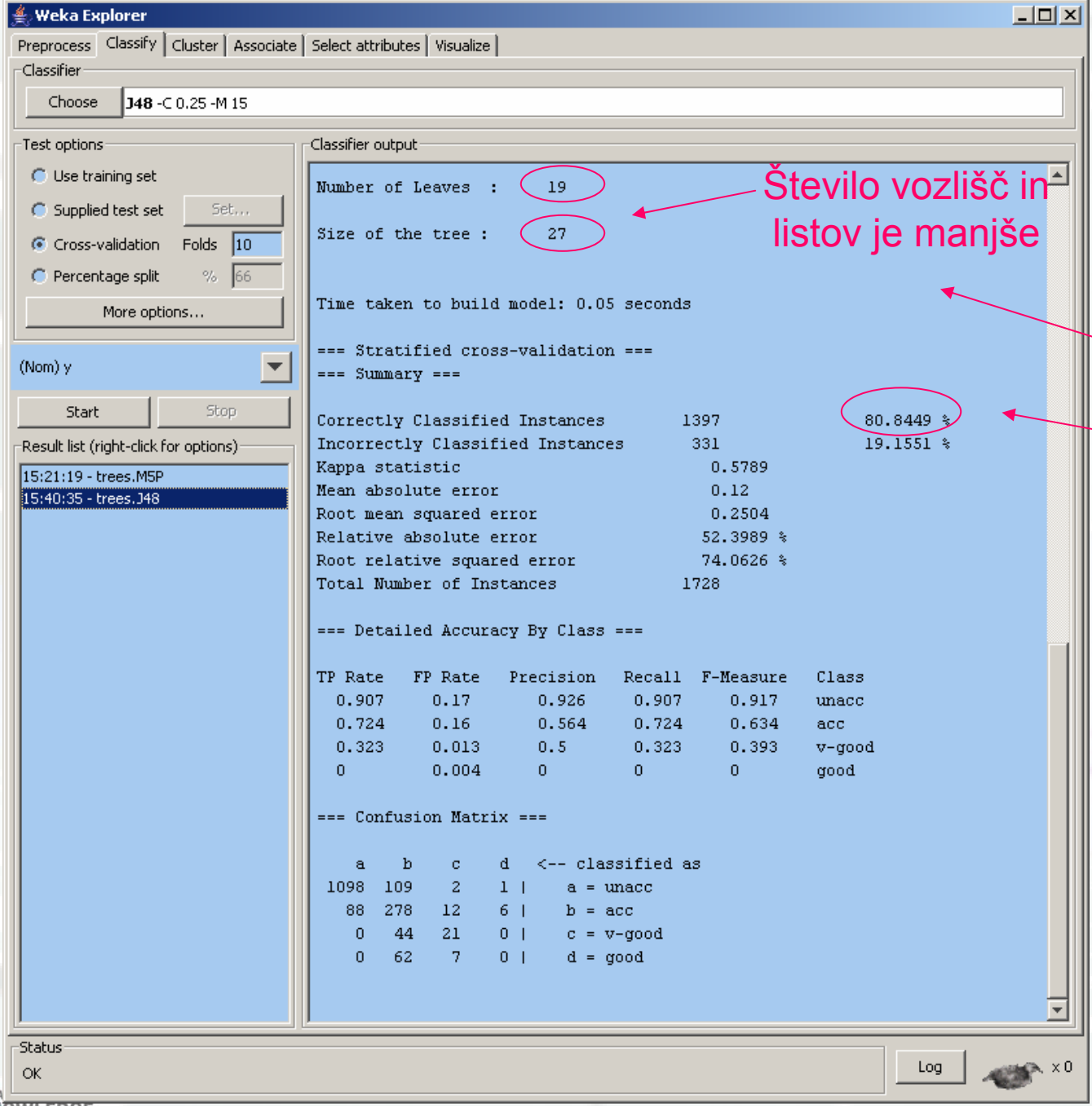
Parametri algoritma (desni klik na miški)

2

Nastavimo minimalno število primerov v listu na 15



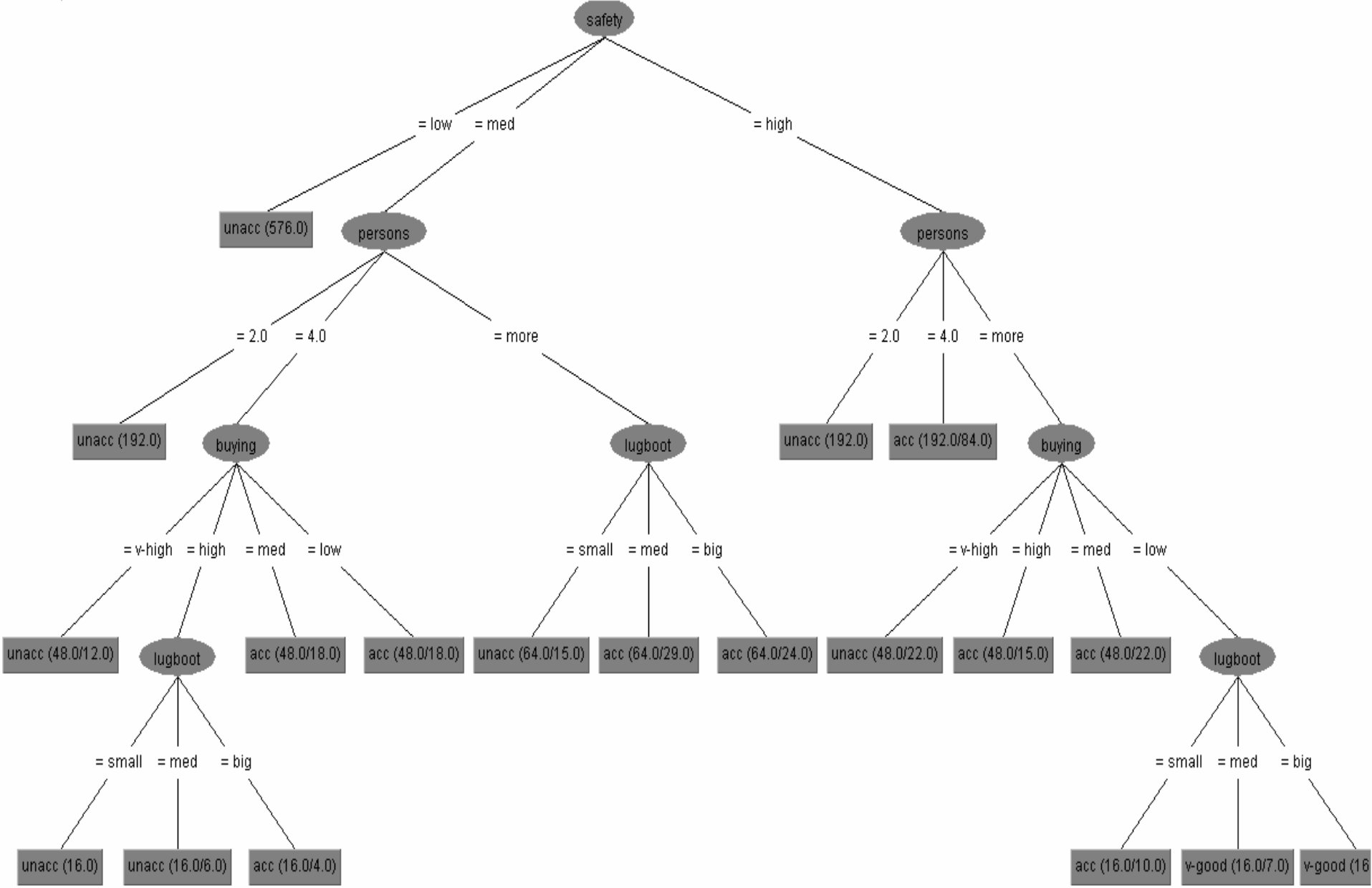
Petra.Kralj@ijs.si



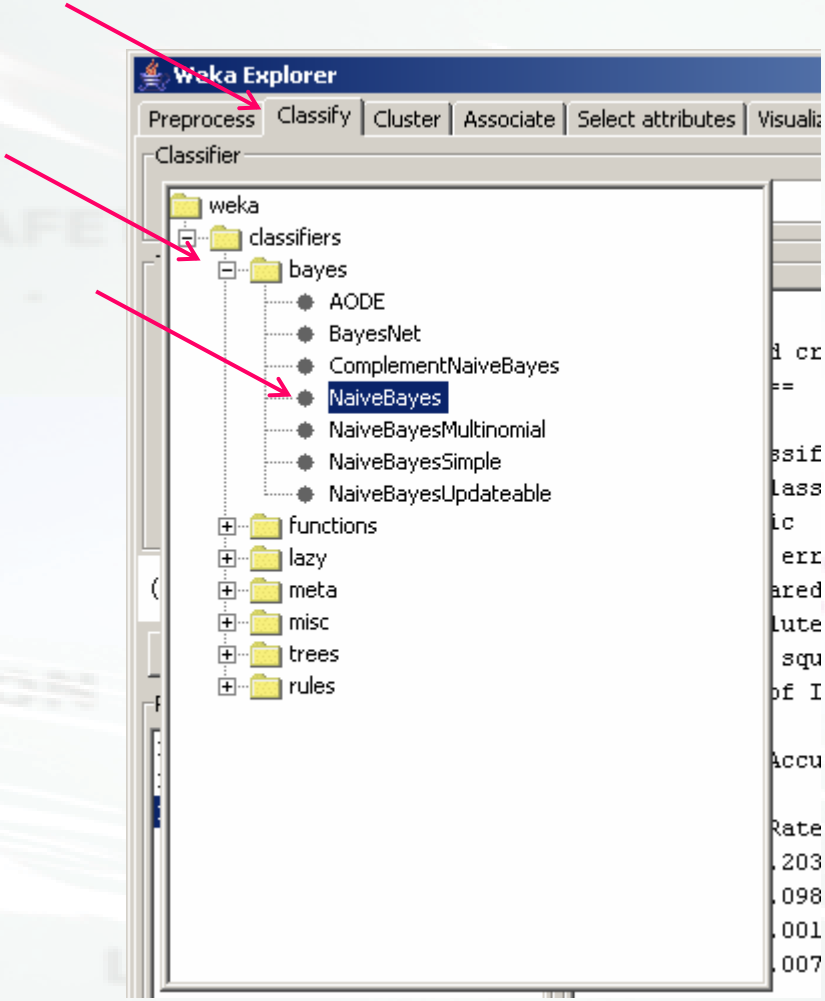
Število vozlišč in listov je manjše

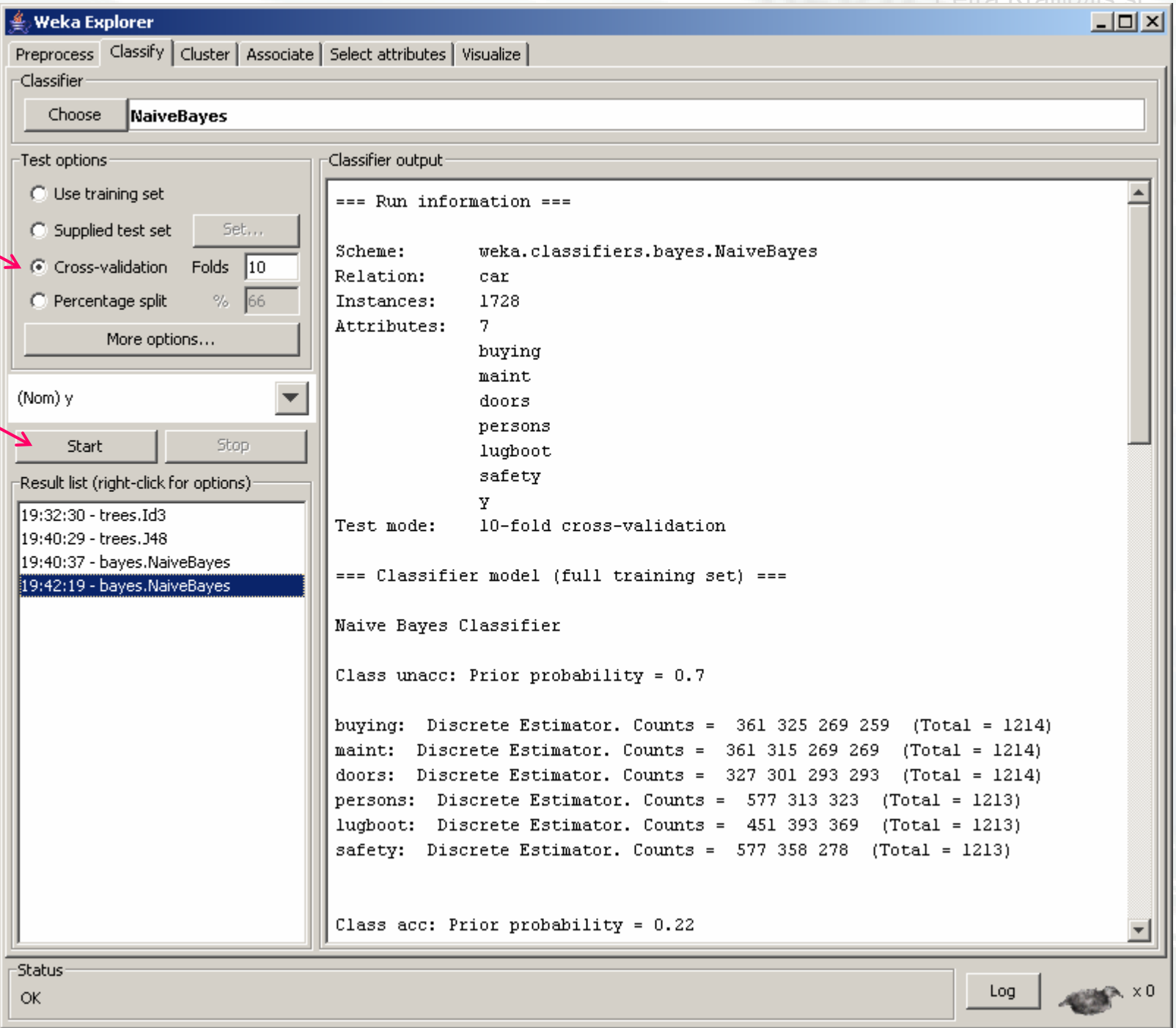
Lažja interpretacija,

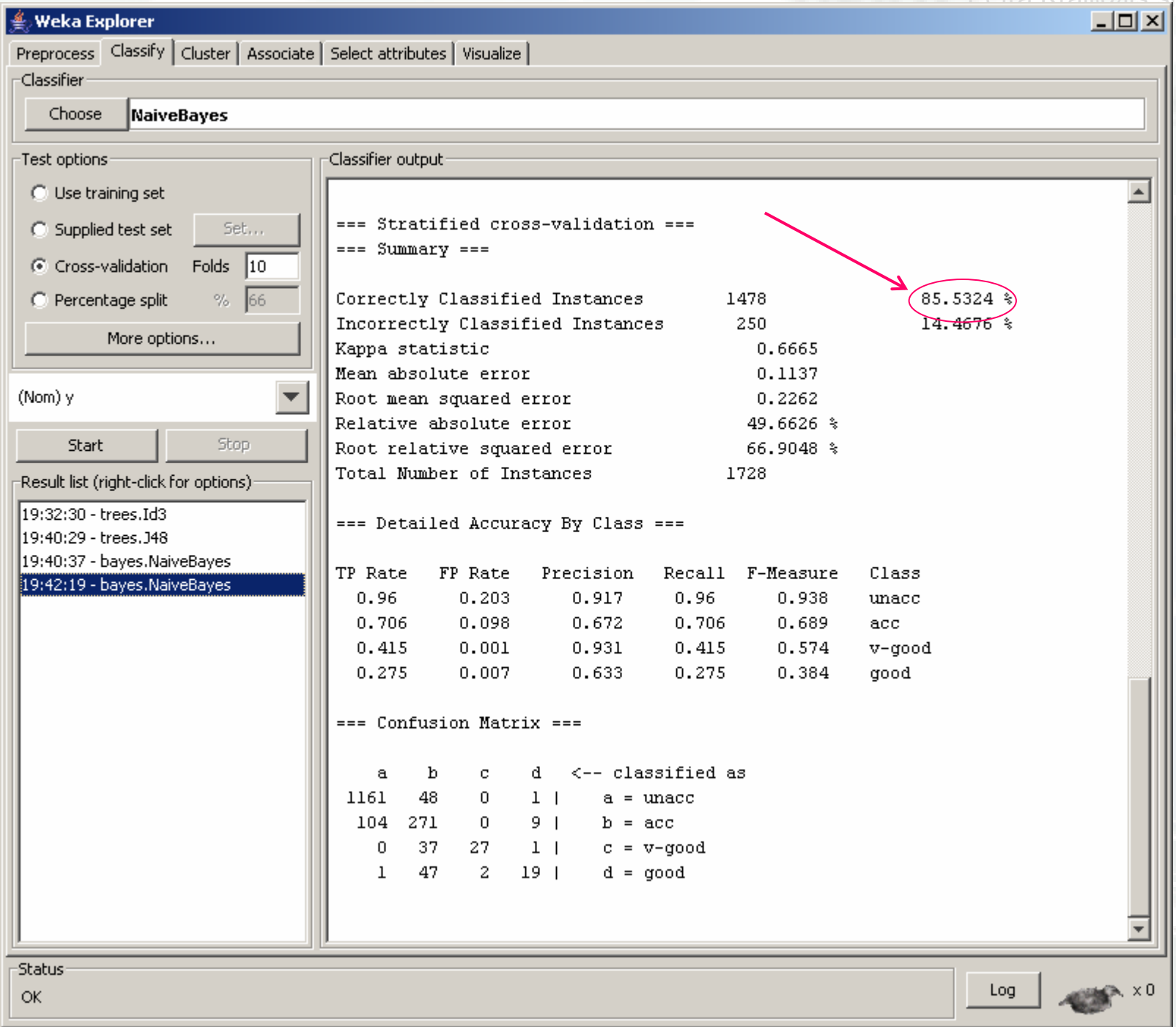
manjša klasifikacijska točnost



Naivni Bayesov klasifikator




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



Petra.Kralj@ijs.si

Za doma

Na datoteki

- LeceBin.csv

Zgradite odločitveno drevo z ID3 algoritmom in J48 algoritmom, za evalvacijo uporabite prečno preverjanje.