

Klasifikacija v WEKI

16.3.2007

Petra Kralj

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/>



Weka 3 - Data Mining with Open Source Machine Learning Software in Java - Mozilla Firefox

File Edit View History Bookmarks Tools Help

[http://www.cs.waikato.ac.nz/ml/weka/](#)



WEKA
The University
of Waikato

Software

[project](#) ▪ [software](#) ▪ [book](#) ▪ [publications](#) ▪ [people](#) ▪ [related](#)

Home

Getting started

- [Requirements](#)
- [Download](#)
- [Documentation](#)
- [FAQ](#)
- [Citing Weka](#)

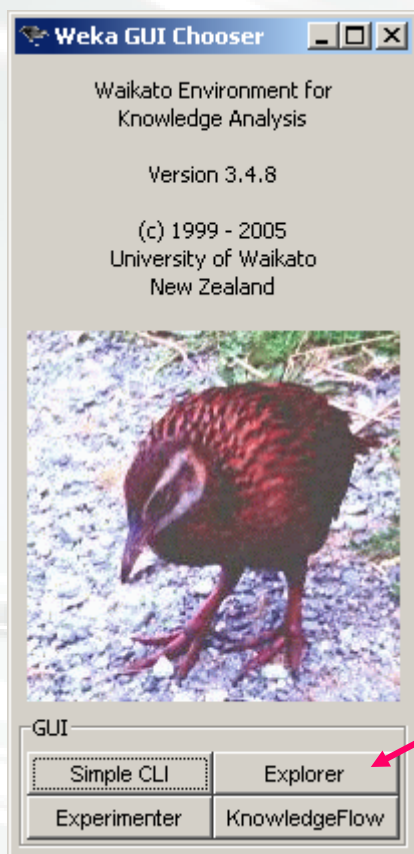
Weka 3: Data Mining Software in Java

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

Weka is open source software issued under the [GNU General Public License](#).

download

Zagon programa Weka



Izberemo Explorer

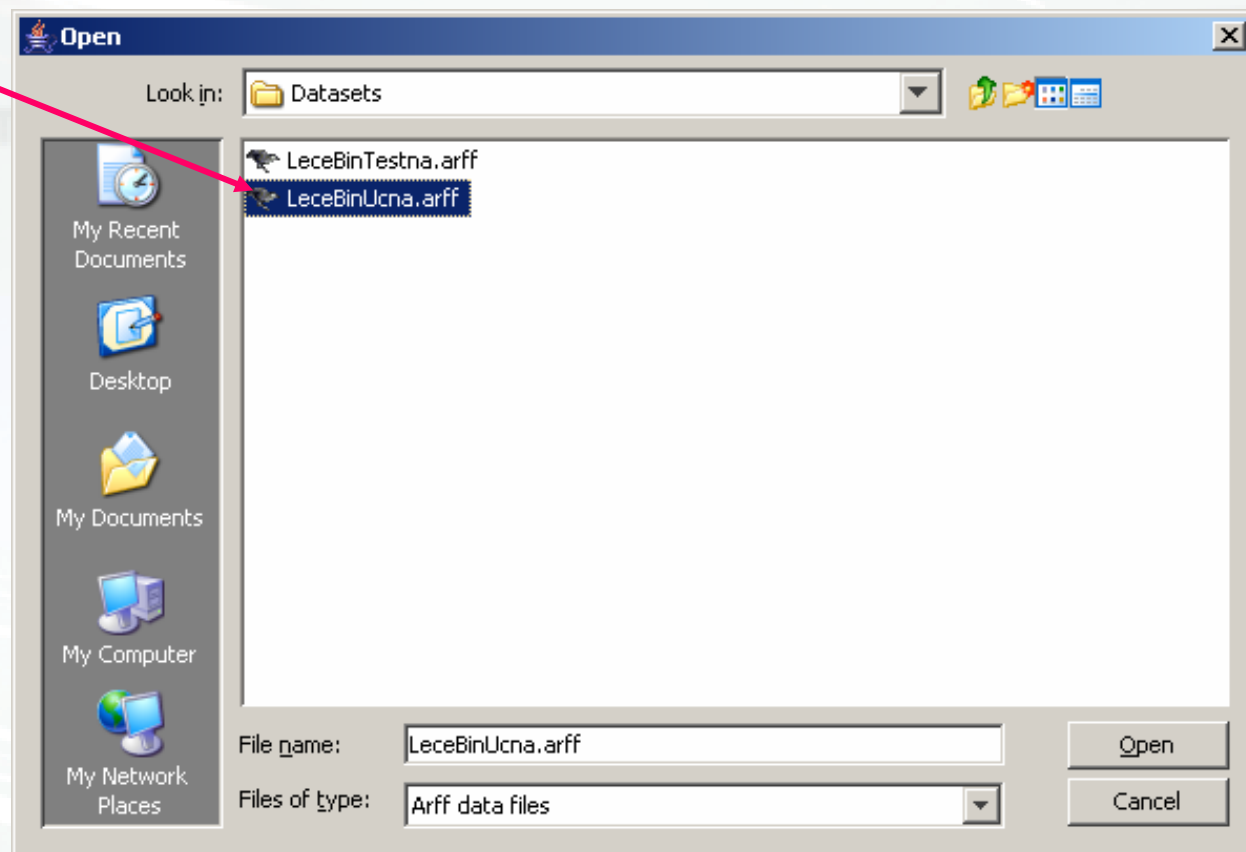
Okno za branje podatkov

The screenshot shows the Weka Explorer application window. A red arrow points to the 'Open file...' button in the 'Preprocess' tab. The interface includes several sections:

- Buttons:** Open file..., Open URL..., Open DB..., Undo, Edit..., Save...
- Filter:** Choose **None** [Apply]
- Current relation:** Relation: None, Instances: None, Attributes: None
- Selected attribute:** Name: None, Missing: None, Distinct: None, Type: None, Unique: None
- Attributes:** All, None, Invert
- Visualize:** Visualize All
- Status:** Welcome to the Weka Explorer [Log] x 0

Naložimo datoteko

LeceBinUcna.arff



Pokažejo se nam podatki

Izberemo
zavihek
"Classify"

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. A red arrow points to the 'Classify' button in the top menu. Below the menu, there are buttons for 'Open file...', 'Open URL...', 'Open DB...', 'Undo', 'Edit...', and 'Save...'. The 'Filter' section shows 'Choose None' and an 'Apply' button. The 'Current relation' section displays 'Relation: LeceBinUcna-weka.filters.unsupervised.attribute.Remove-R1' and 'Instances: 17 Attributes: 5'. The 'Attributes' section has 'All', 'None', and 'Invert' buttons, and a list of attributes: Starost, Dioptrija, Astigmatizem, Solzenje, and Lece. A red arrow points to the 'Starost' attribute in this list. The 'Selected attribute' section shows 'Name: Starost', 'Missing: 0 (0%)', 'Distinct: 3', and 'Type: Nominal Unique: 0 (0%)'. Below this is a table:

Label	Count
mlad	7
pr_st_dal	3
st_daleko	7

The 'Class: Lece (Nom)' section has a 'Visualize All' button. Below this is a bar chart with three bars representing the counts for 'mlad', 'pr_st_dal', and 'st_daleko'. The bars are stacked with blue at the bottom and red on top. The counts are 7, 3, and 7 respectively. A red arrow points to the first bar (7). The 'Status' section at the bottom shows 'OK' and a 'Log' button. The bottom left corner features the logo for the Department of Knowledge Technologies at the Jozef Stefan Institute.

Ciljna spremenljivka

Izberemo algoritem

The screenshot shows the Weka Explorer application window. The 'Classifier' tab is active, and 'ZeroR' is selected in the classifier list. A red arrow points to the 'Choose' button next to 'ZeroR'. The 'Test options' section shows 'Cross-validation' selected with 10 folds. The 'Classifier output' area is empty. The 'Result list' area is also empty. The status bar at the bottom shows 'OK' and a 'Log' button.

Weka Explorer

Preprocess | **Classifier** | Cluster | Associate | Select attributes | Visualize

Classifier

Choose ZeroR


Test options

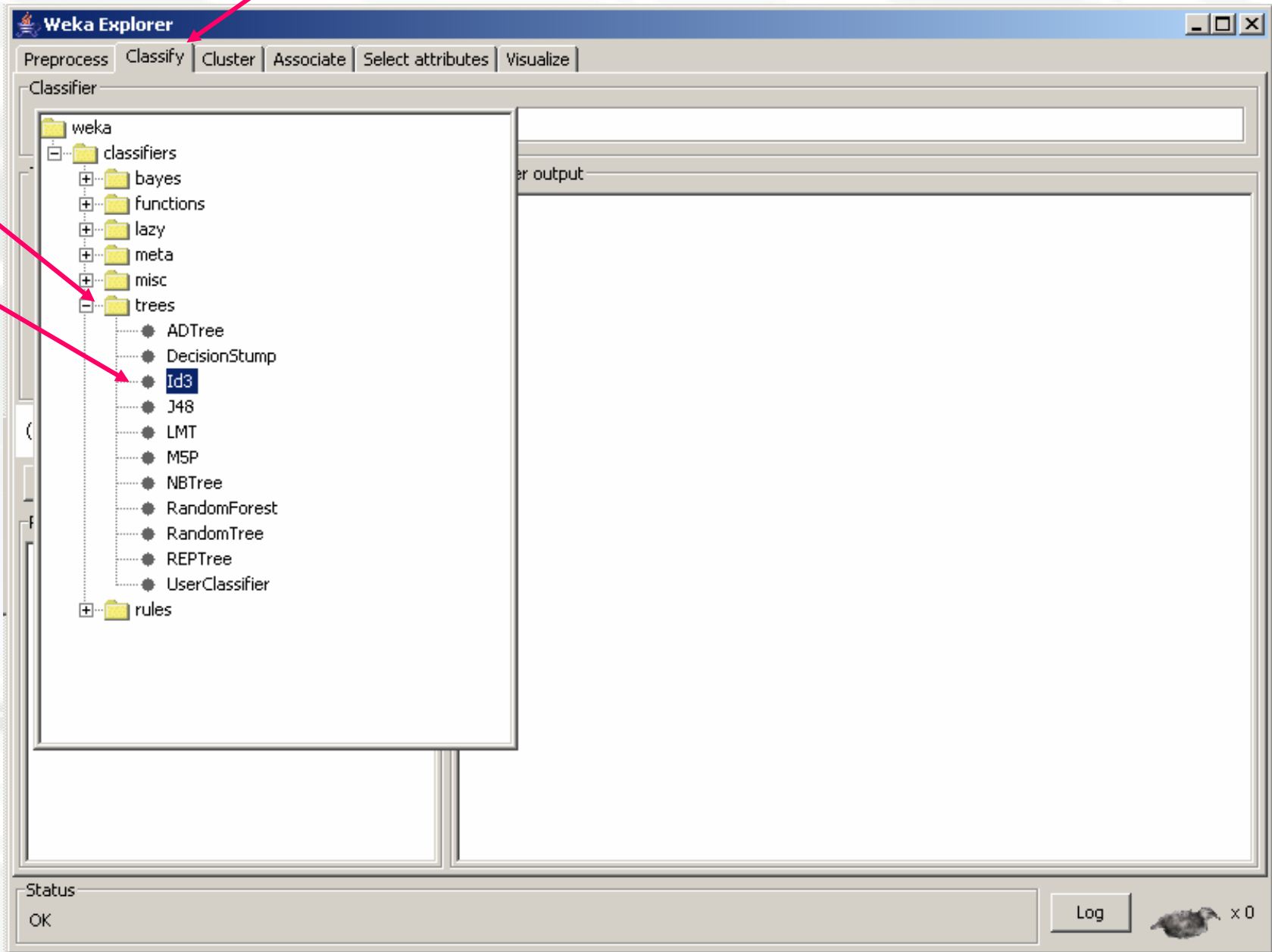
- Use training set
- Supplied test set
- Cross-validation Folds
- Percentage split %

(Nom) Lece

Result list (right-click for options)

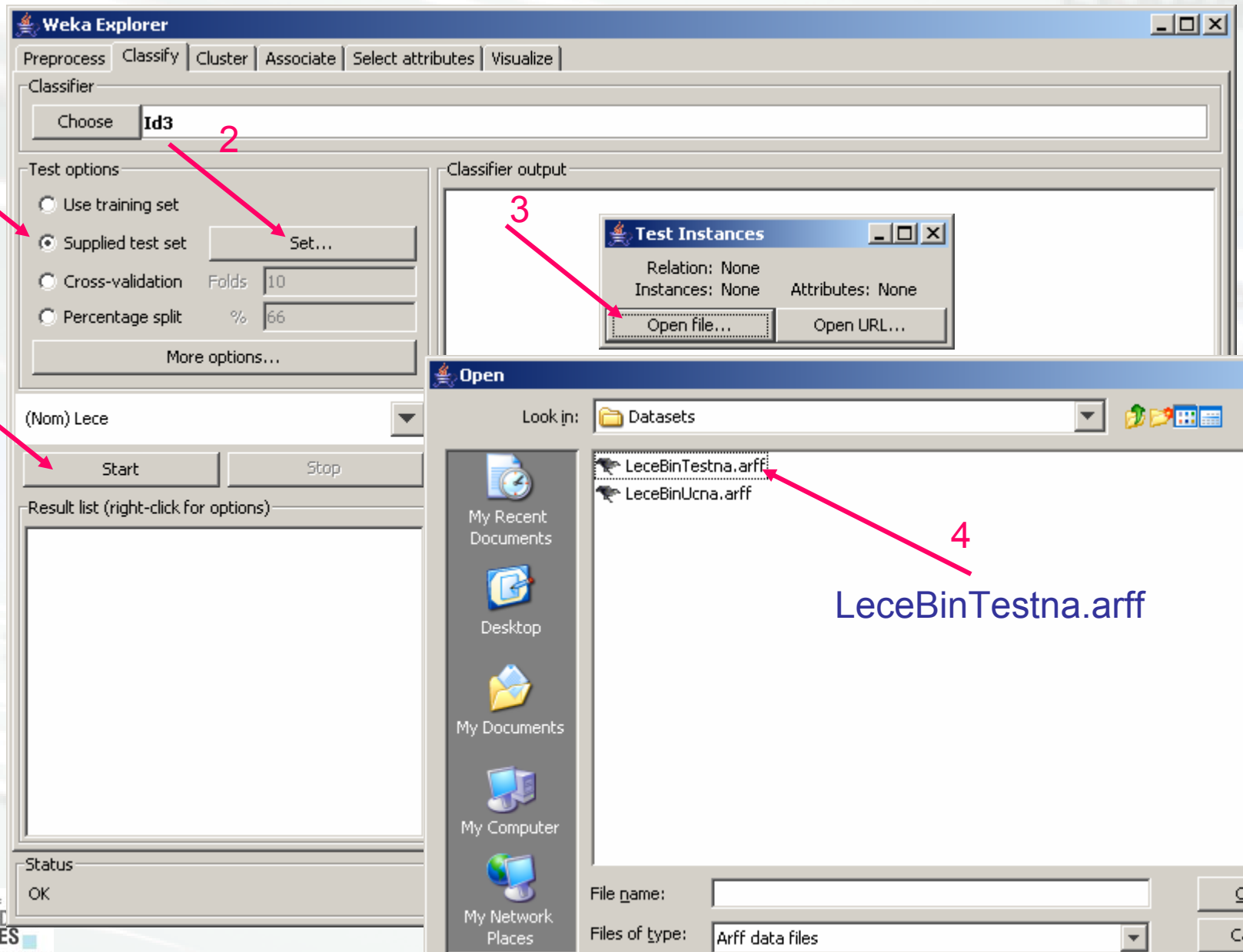
Status

OK  x 0



trees

Id3



LeceBinTestna.arff

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

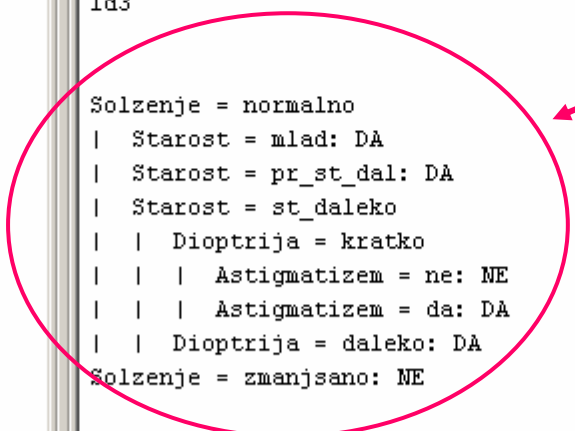
Solzenje = normalno
| Starost = mlad: DA
| Starost = pr_st_dal: DA
| Starost = st_daleko
| | Dioptrija = kratko
| | | Astigmatizem = ne: NE
| | | Astigmatizem = da: DA
| | Dioptrija = daleko: DA
Solzenje = zmanjsano: NE

Time taken to build model: 0 seconds
  
```

Odločitveno drevo

Status: OK

Log x 0



Odločitveno drevo

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

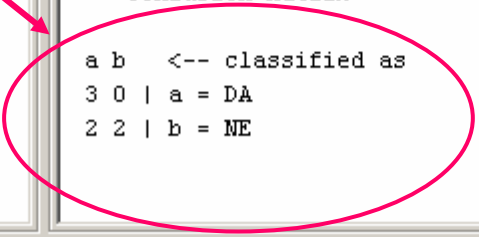
=== Confusion Matrix ===
 a b  <-- classified as
3 0 | a = DA
2 2 | b = NE
```

====

Klasifikacijska točnost



Kontingenčna tabela



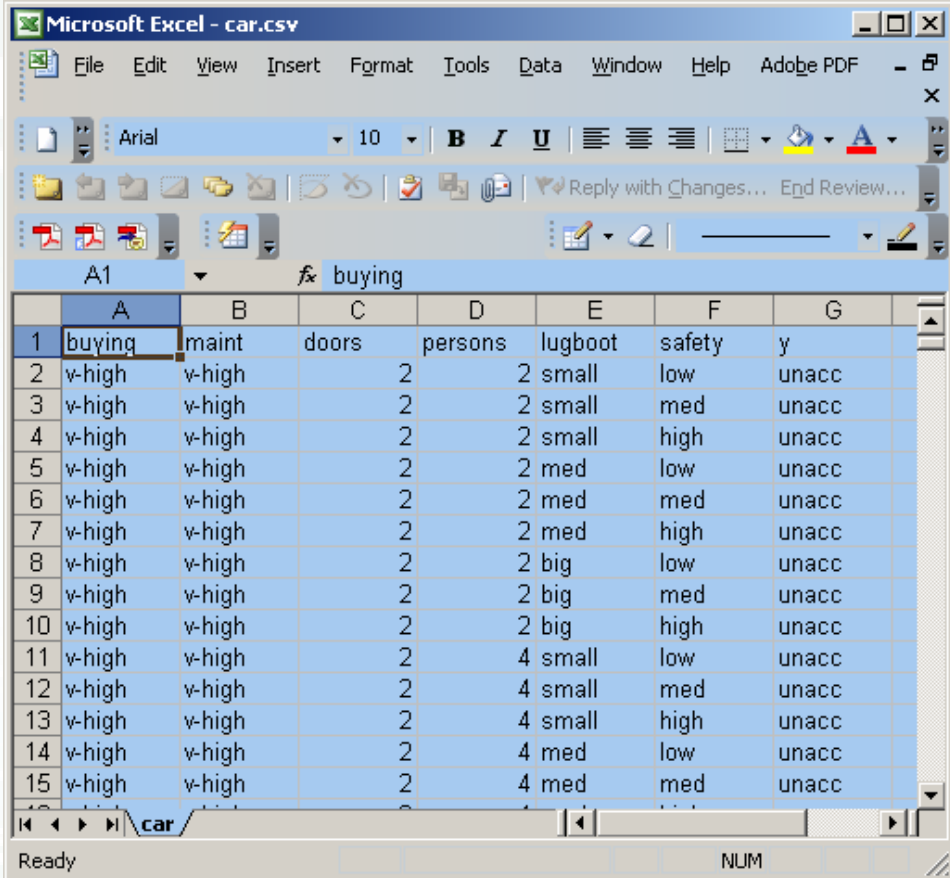
CAR dataset

- 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

Priprava podatkov za WEKO - 1

Podatki v tabeli
(npr. MS Excel)

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



Microsoft Excel - car.csv

	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

Priprava podatkov za WEKO - 2

Shrani kot “.csv”

- Pazljivo s pikami, vejicami in podpičji!

The screenshot shows a Microsoft Excel window titled 'Microsoft Excel - car.csv'. The spreadsheet contains data in columns E, F, and G. The 'Save As' dialog box is open, showing the file name 'car.csv' and the save type 'CSV (Comma delimited) (*.csv)'. The 'Save in' location is 'HandsOnWeka'.

root	safety	y
all	low	unacc
all	med	unacc
all	high	unacc
	low	unacc
	med	unacc
	high	unacc
	low	unacc
	med	unacc
	high	unacc
all	low	unacc
all	med	unacc
all	high	unacc
	low	unacc
	med	unacc

Car.csv

Load the data

The screenshot shows the Weka Explorer application window. The 'Preprocess' tab is active, and the 'Open file...' button is highlighted with a red arrow. The 'Current relation' section shows 'Relation: car' and 'Instances: 1728'. The 'Attributes' section lists 7 attributes, with 'y' selected. The 'Selected attribute' section shows 'Name: y', 'Type: Nominal', and a table of counts for labels: unacc (1210), acc (384), v-good (65), and good (69). A bar chart below shows these counts, with a blue arrow pointing to the 'unacc' bar and the text 'Ciljna spremenljivka'.

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... | Open URL... | Open DB... | Undo | Edit... | Save...

Filter: Choose None Apply

Current relation: Relation: car, Instances: 1728, Attributes: 7

Attributes: All None Invert

No.	Name
1	buying
2	maint
3	doors
4	persons
5	lugboot
6	safety
7	y

Selected attribute: Name: y, Missing: 0 (0%), Distinct: 4, Type: Nominal, Unique: 0 (0%)

Label	Count
unacc	1210
acc	384
v-good	65
good	69

Class: y (Nom) Visualize All

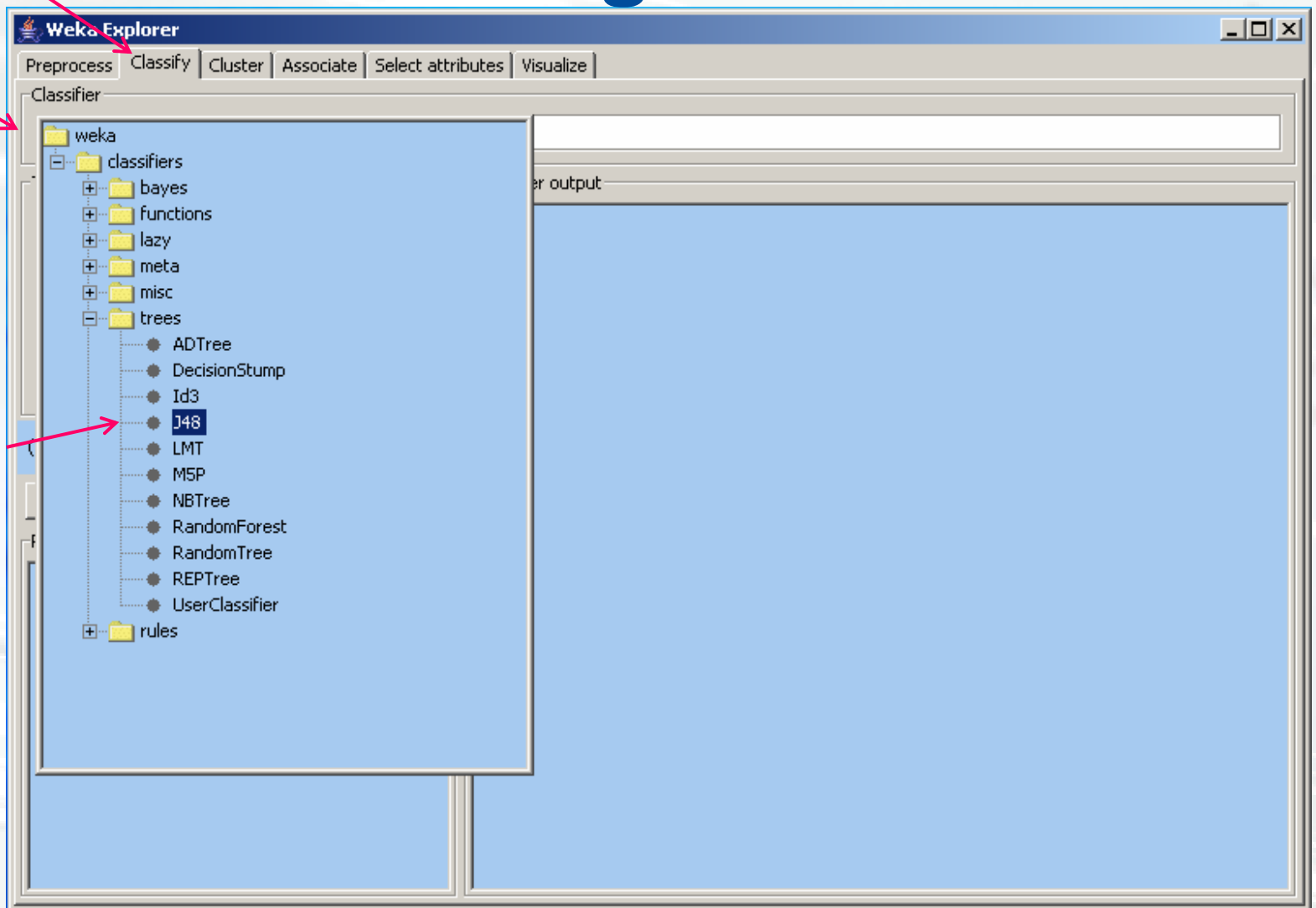
Ciljna spremenljivka

Status: OK Log x 0

1

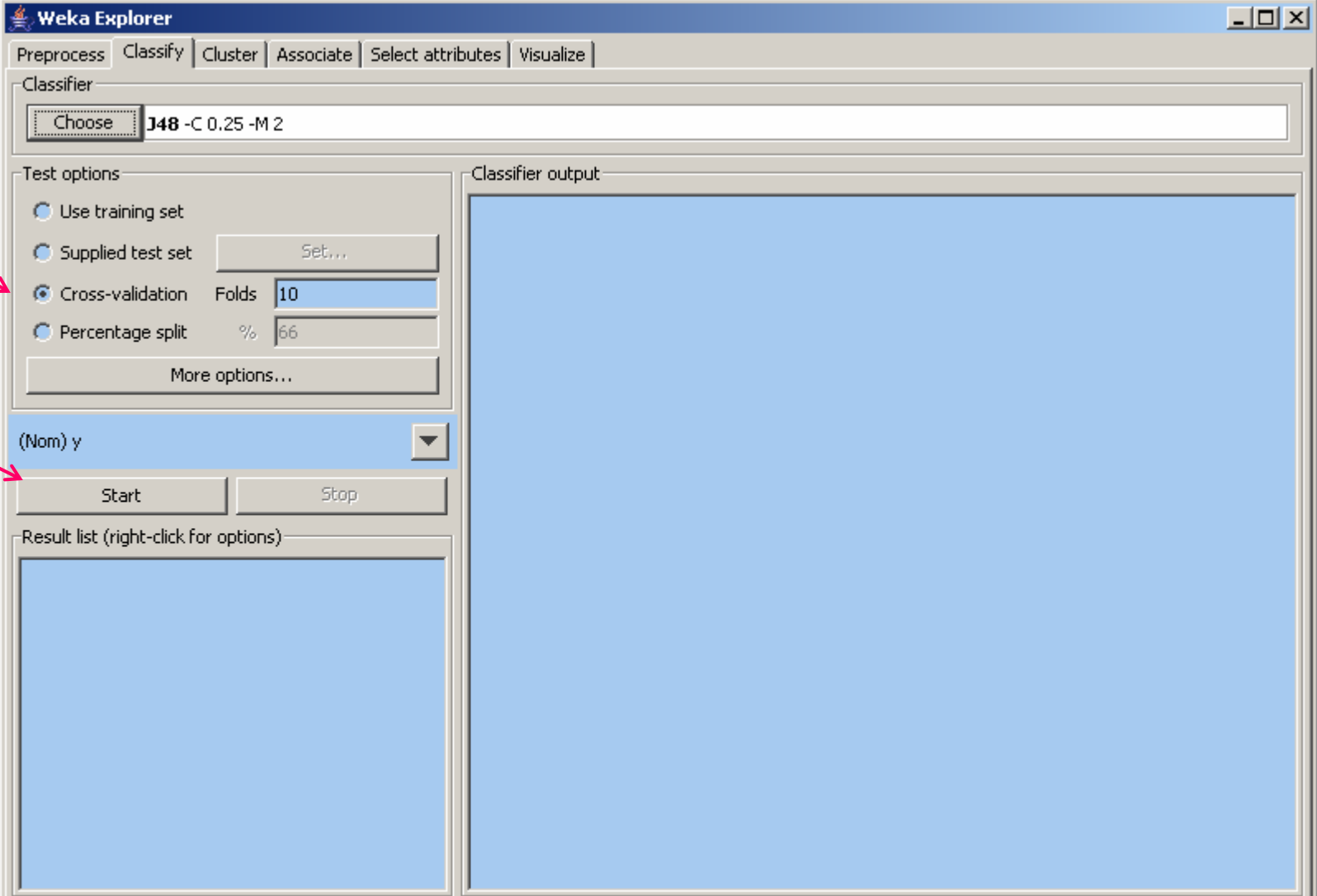
Izberemo algoritem J48

2



3

Gradnja in evalvacija drevesa



The screenshot displays the Weka Explorer application window. The 'Classifier' tab is active, showing the 'J48 -C 0.25 -M 2' classifier selected. The 'Test options' section is highlighted with a red arrow and a circled '1', indicating the configuration of the test set. The 'Cross-validation' option is selected, with 'Folds' set to 10. Below this, the 'Start' button is highlighted with a red arrow and a circled '2', indicating the execution of the classifier. The 'Classifier output' and 'Result list' areas are currently empty.

1

2

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

Classifier output

Result list (right-click for options)

Status

OK Log x 0

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
- Cross-validation Folds **10**
- Percentage split % **66**

(Nom) y

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
 Incorrectly Classified Instances 132

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

Klasifikacijska točnost → 92.3611 %

Napoved modela →

Resnične vrednosti →

Status: OK  x 0

Weka Explorer

Preprocess | Classifier | Cluster | Associate | Select attributes | Visualize

Classifier: Choose J48 -C 0.25 -M 15

Test options

- Use training set
- Supplied test set
- Cross-validation Folds
- Percentage split %

(Nom) y

Result list (right-click for options)

- 14:05:00 - trees 148
- 14:58:13 - trees 148

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
0.064	0.972	0.962	0.967	unacc
0.047	0.841	0.867	0.854	acc
0.011	0.763	0.892	0.823	v-good
0.011	0.695	0.594	0.641	good

Confusion Matrix ===

	c	d	<-- classified as
a = unacc	1164	43	0
b = acc	33	333	7
c = v-good	0	3	58
d = good	0	17	11

Context menu options:

- 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

x 0

Desni klik na miški

Rezanje dreves

1

Parametri algoritma (desni klik na miški)

2

Nastavimo minimalno število primerov v listu na 15

The screenshot shows the Weka Explorer interface with the J48 classifier selected. A configuration dialog box is open, showing various parameters. The 'minNumObj' parameter is highlighted with a red arrow and set to 15. The background shows the classifier's performance metrics and a partial confusion matrix.

Classifier: J48 -C 0.25 -M 15

Test options: Use training set, Supplied test set, Cross-validation (selected), Percentage split

Classifier output: weka.gui.GenericObjectEditor

weka.gui.GenericObjectEditor Parameters:

- binarySplits: False
- confidenceFactor: 0.25
- debug: False
- minNumObj: 15
- numFolds: 3
- reducedErrorPruning: False
- saveInstanceData: False
- seed: 1
- subtreeRaising: True
- unpruned: False
- useLaplace: False

Classifier Performance:

- 92.3611 %
- 7.6389 %
- 843
- 121
- 718
- 833 %
- 176 %

Confusion Matrix (partial):

Actual \ Predicted	Class	unacc	acc	v-good	good
967	unacc				
854	acc				
823	v-good				
641	good				

Status: OK

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...

(Nom) y

Start Stop

Result list (right-click for options):

- 15:21:19 - trees.M5P
- 15:40:35 - trees.J48**

Classifier output:

Number of Leaves : **19**

Size of the tree : **27**

Time taken to build model: 0.05 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	1397	80.8449 %
Incorrectly Classified Instances	331	19.1551 %

Kappa statistic 0.5789

Mean absolute error 0.12

Root mean squared error 0.2504

Relative absolute error 52.3989 %

Root relative squared error 74.0626 %

Total Number of Instances 1728

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.907	0.17	0.926	0.907	0.917	unacc
0.724	0.16	0.564	0.724	0.634	acc
0.323	0.013	0.5	0.323	0.393	v-good
0	0.004	0	0	0	good

=== Confusion Matrix ===

a	b	c	d	<-- classified as
1098	109	2	1	a = unacc
88	278	12	6	b = acc
0	44	21	0	c = v-good
0	62	7	0	d = good

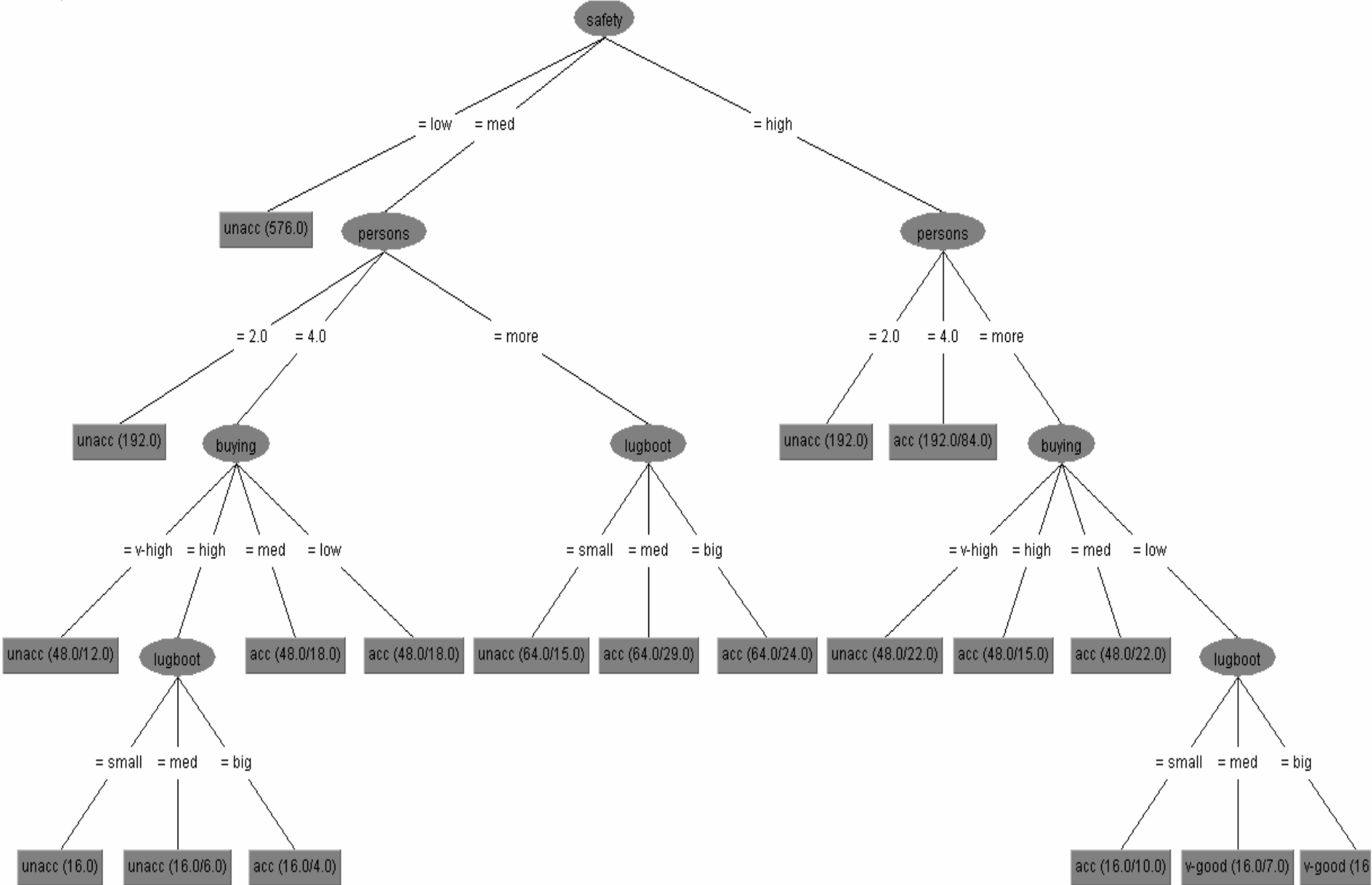
Status: OK

Log x 0

Število vozlišč in listov je manjše

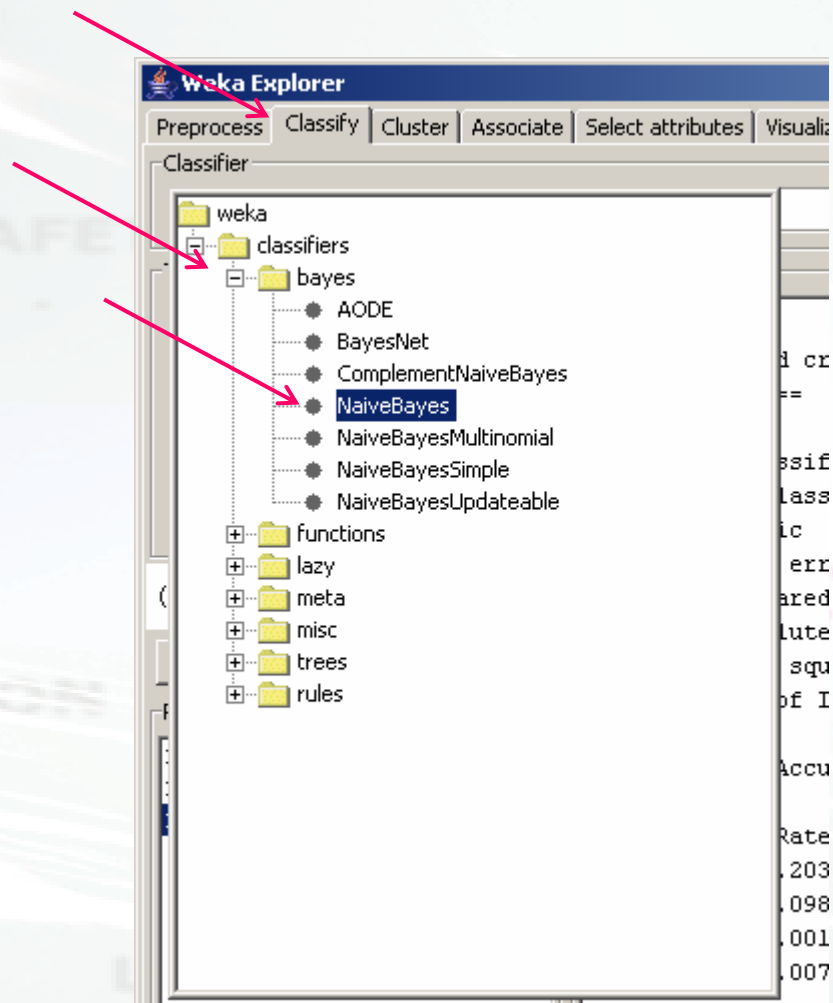
Lažja interpretacija,

manjša
klasifikacijska
točnost



LANGUAGE

Naivni Bayesov klasifikator



Weka Explorer

Preprocess | **Classify** | Cluster | Associate | Select attributes | Visualize

Classifier: Choose **NaiveBayes**

Test options

- Use training set
- Supplied test set
- Cross-validation Folds:
- Percentage split %:

(Nom) y

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  x 0

Classifier

Choose NaiveBayes

Test options

Use training set
 Supplied test set
 Cross-validation Folds
 Percentage split %

(Nom) y

Start

Stop

Result list (right-click for options)

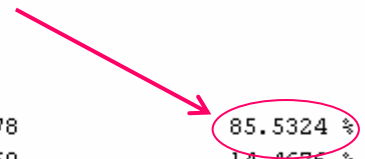
- 19:32:30 - trees.Id3
- 19:40:29 - trees.J48
- 19:40:37 - bayes.NaiveBayes
- 19:42:19 - bayes.NaiveBayes

Classifier output

```

=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances      1478
Incorrectly Classified Instances    250
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



Za doma

Na datoteki

- LeceBin.csv

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