

# Classification in WEKA

8.11.2007

Petra Kralj

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

# 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.

# Weka

Weka is open source software for machine learning and data mining.

<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

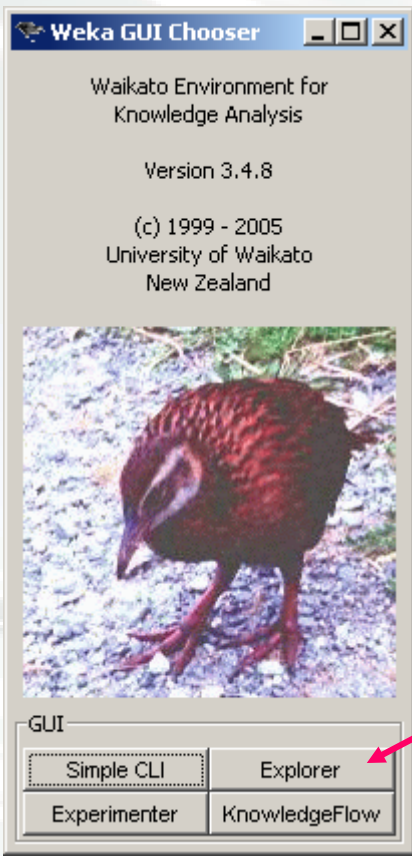
Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also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

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

Download  
version  
3.4



# Run Weka



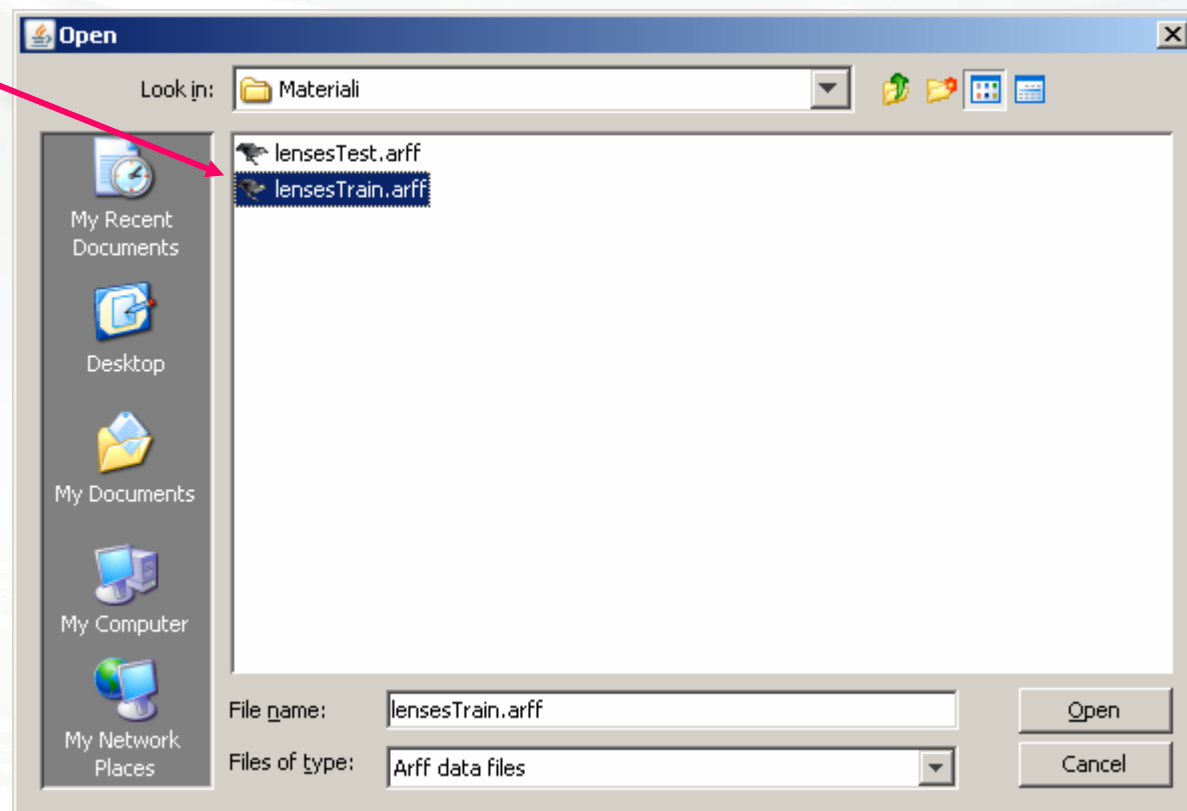
Choose Explorer

# Load the data

The screenshot shows the Weka Explorer application window. A red arrow points to the 'Open file...' button in the top toolbar. The interface includes a menu bar with 'Preprocess', 'Classify', 'Cluster', 'Associate', 'Select attributes', and 'Visualize'. Below the toolbar are buttons for 'Open file...', 'Open URL...', 'Open DB...', 'Undo', 'Edit...', and 'Save...'. The main area is divided into several sections: 'Filter' with a 'Choose' button and a text field containing 'None'; 'Current relation' showing 'Relation: None' and 'Instances: None'; 'Attributes' with 'All', 'None', and 'Invert' buttons; 'Selected attribute' showing 'Name: None', 'Missing: None', 'Distinct: None', 'Type: None', and 'Unique: None'; and a 'Visualize All' button. The status bar at the bottom displays 'Welcome to the Weka Explorer' and a 'Log' button.

# Load the data - 2

lensesTrain.arff



# The data are loaded

Choose  
"Classify"

The screenshot shows the Weka Explorer application window. The 'Classify' menu item is highlighted with a red arrow. The 'Open file...' button is also highlighted with a red arrow. The 'Current relation' section shows 'Relation: lensesTrain' and 'Instances: 17'. The 'Attributes' section lists five attributes: Age, Prescription, Astigmatic, Tear\_rate, and Lenses. The 'Selected attribute' section shows 'Name: Age' and a table of labels and counts. The 'Class: Lenses (Nom)' dropdown is highlighted with a red arrow. The bar chart shows three bars representing the distribution of the 'Lenses' variable across different age groups.

Target variable

Label	Count
young	7
pre-presbyopic	3
presbyopic	7



# Choose 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 %

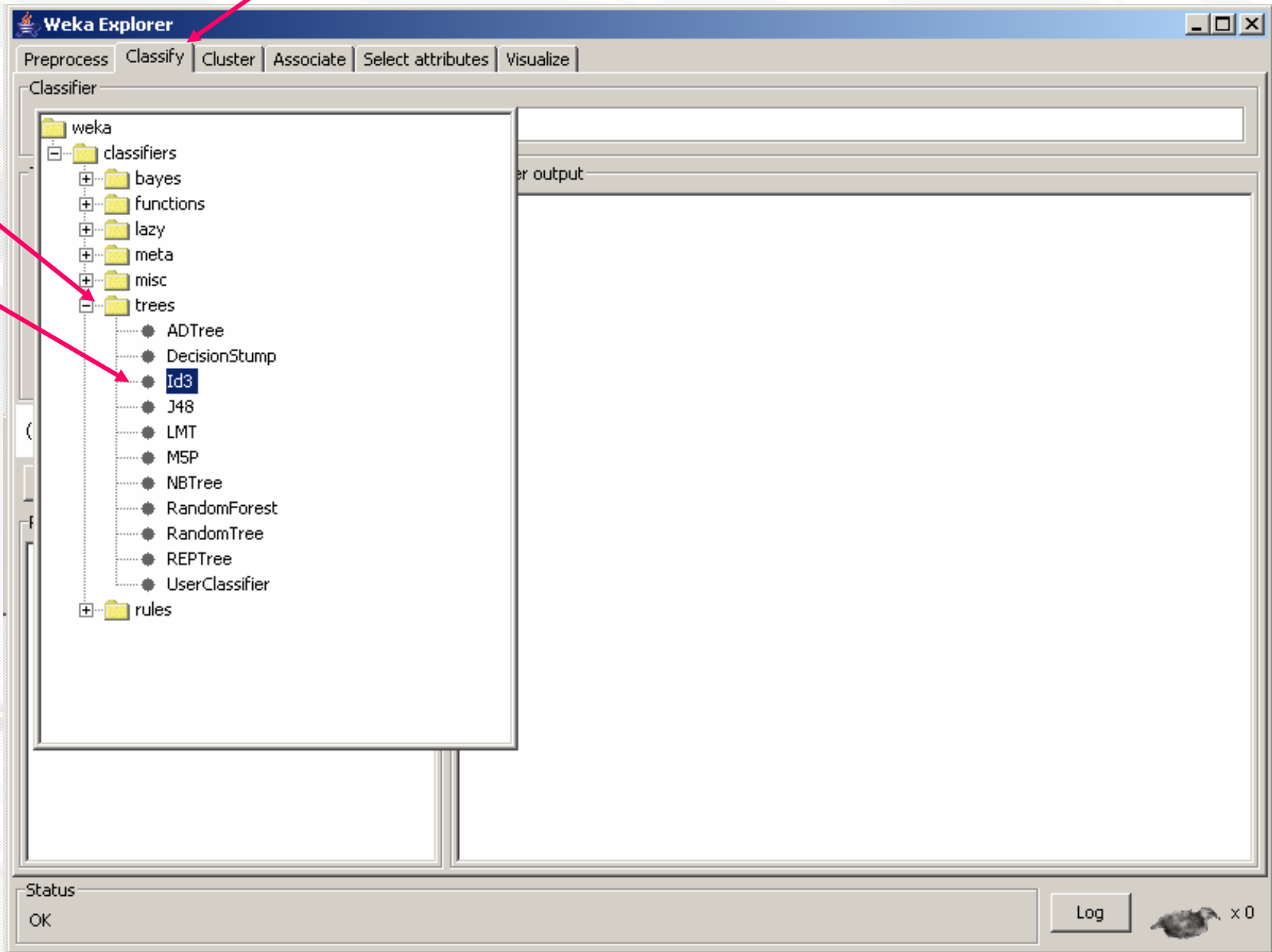
(Nom) Lece

Result list (right-click for options)

Classifier output

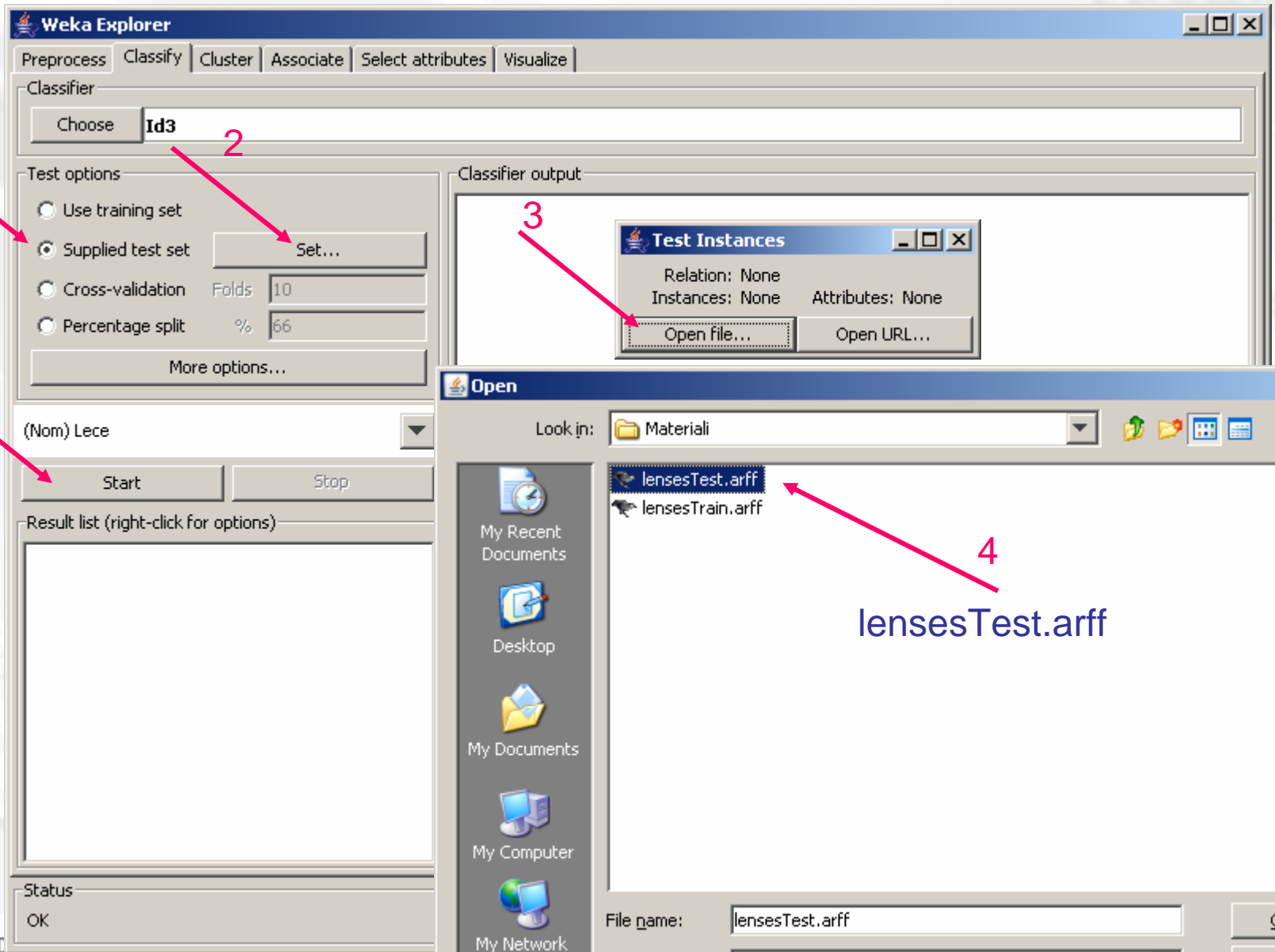
Status

OK   x 0



trees

Id3



**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) Lenses ▼

Start Stop

Result list (right-click for options)

15:42:23 - trees.Id3

Classifier output

```

=== Run information ===

Scheme:      weka.classifiers.trees.Id3
Relation:    lensesTrain
Instances:   17
Attributes:  5
              Age
              Prescription
              Astigmatic
              Tear_rate
              Lenses

Test mode:   user supplied test set: 7 instances

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

Id3

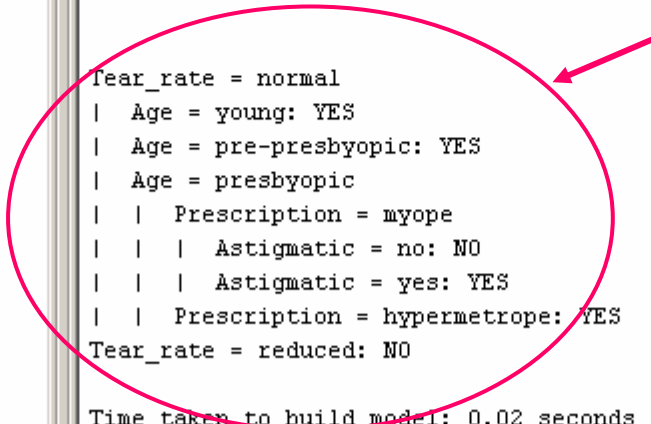
Tear_rate = normal
| Age = young: YES
| Age = pre-presbyopic: YES
| Age = presbyopic
| | Prescription = myope
| | | Astigmatic = no: NO
| | | Astigmatic = yes: YES
| | Prescription = hypermetrope: YES
Tear_rate = reduced: NO

Time taken to build model: 0.02 seconds
  
```

Decision tree

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):  
 15:42:23 - trees.Id3  
 15:45:48 - trees.Id3

Classifier output:

Time taken to build model: 0.02 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	YES
0.5	0	1	0.5	0.667	NO

=== Confusion Matrix ===

```

a b <-- classified as
3 0 | a = YES
2 2 | b = NO
    
```

Classification accuracy

Confusion matrix

Status: OK

Log x 0

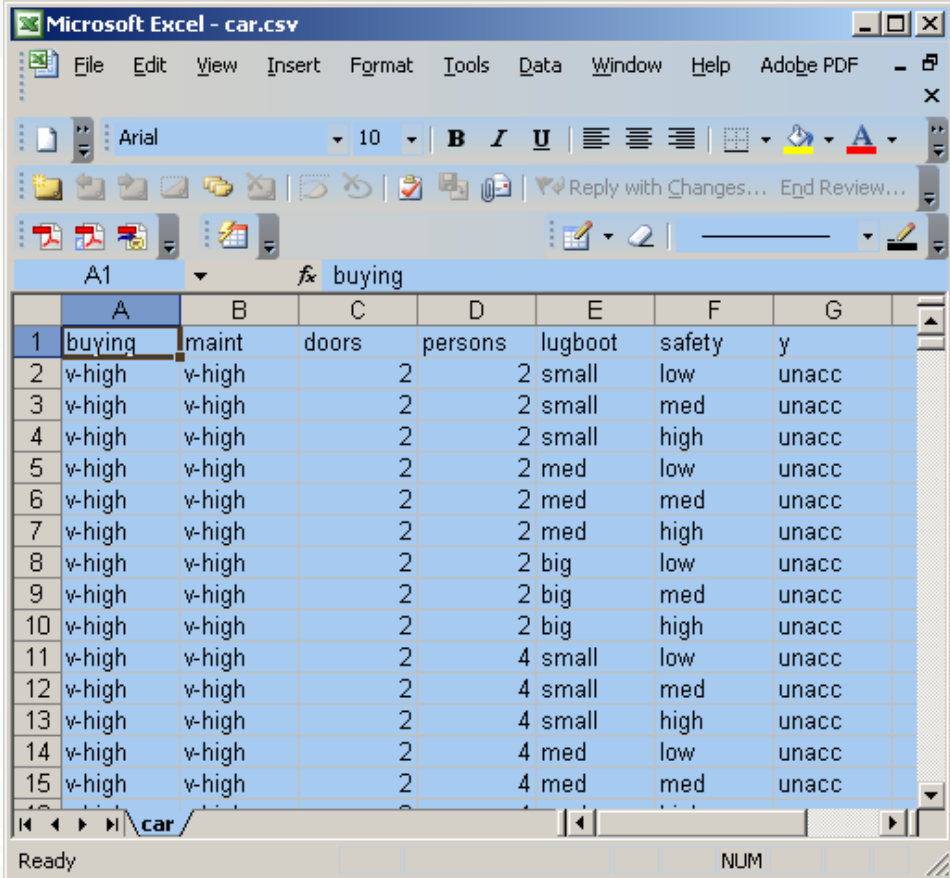
# Exercise 2: CAR dataset

- 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



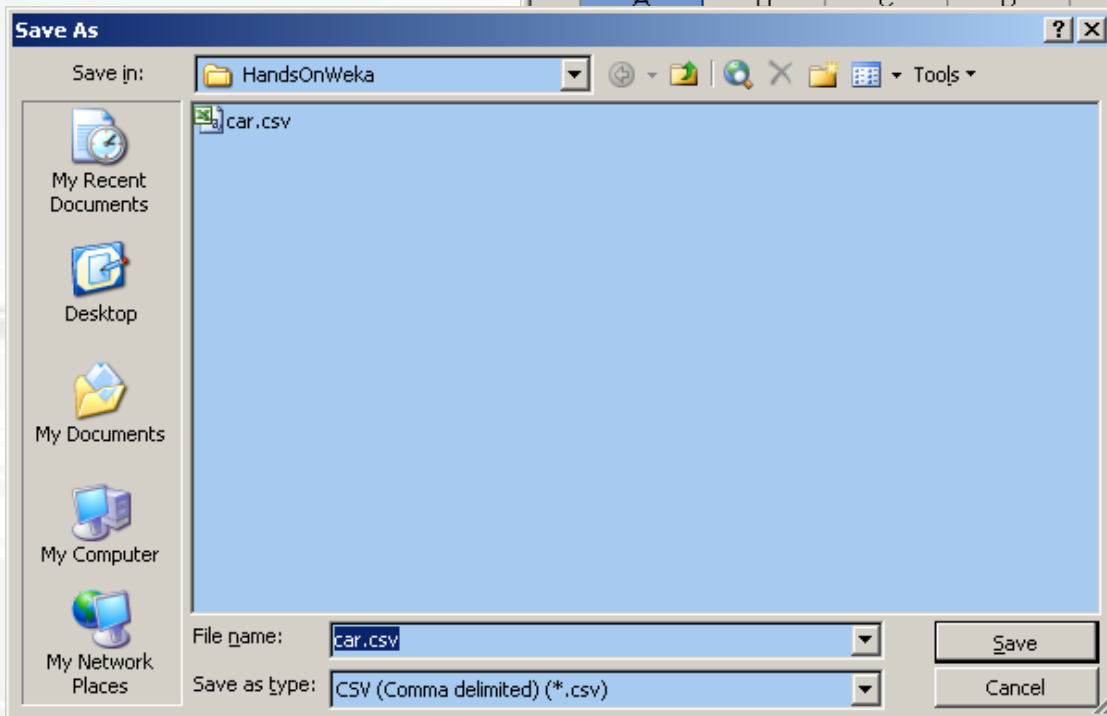
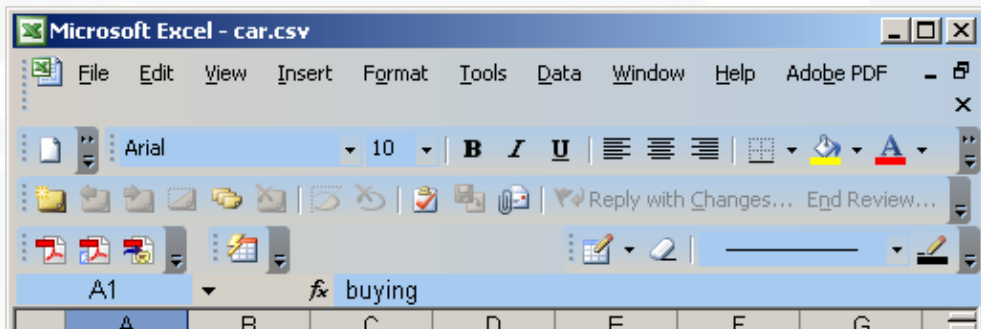
Microsoft Excel - car.csv

	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

# Preparing the data for WEKA - 2

Save as “.csv”

- Careful with dots “.”, commas “,” and semicolons “;”!





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 pointing to the 'Car.csv' text above. The 'Current relation' section shows 'Relation: car' and 'Instances: 1728'. The 'Attributes' section lists 7 attributes: buying, maint, doors, persons, lugboot, safety, and y. The 'Selected attribute' section shows 'Name: y' with a 'Type: Nominal' and a 'Count' table. The 'Count' table is as follows:

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

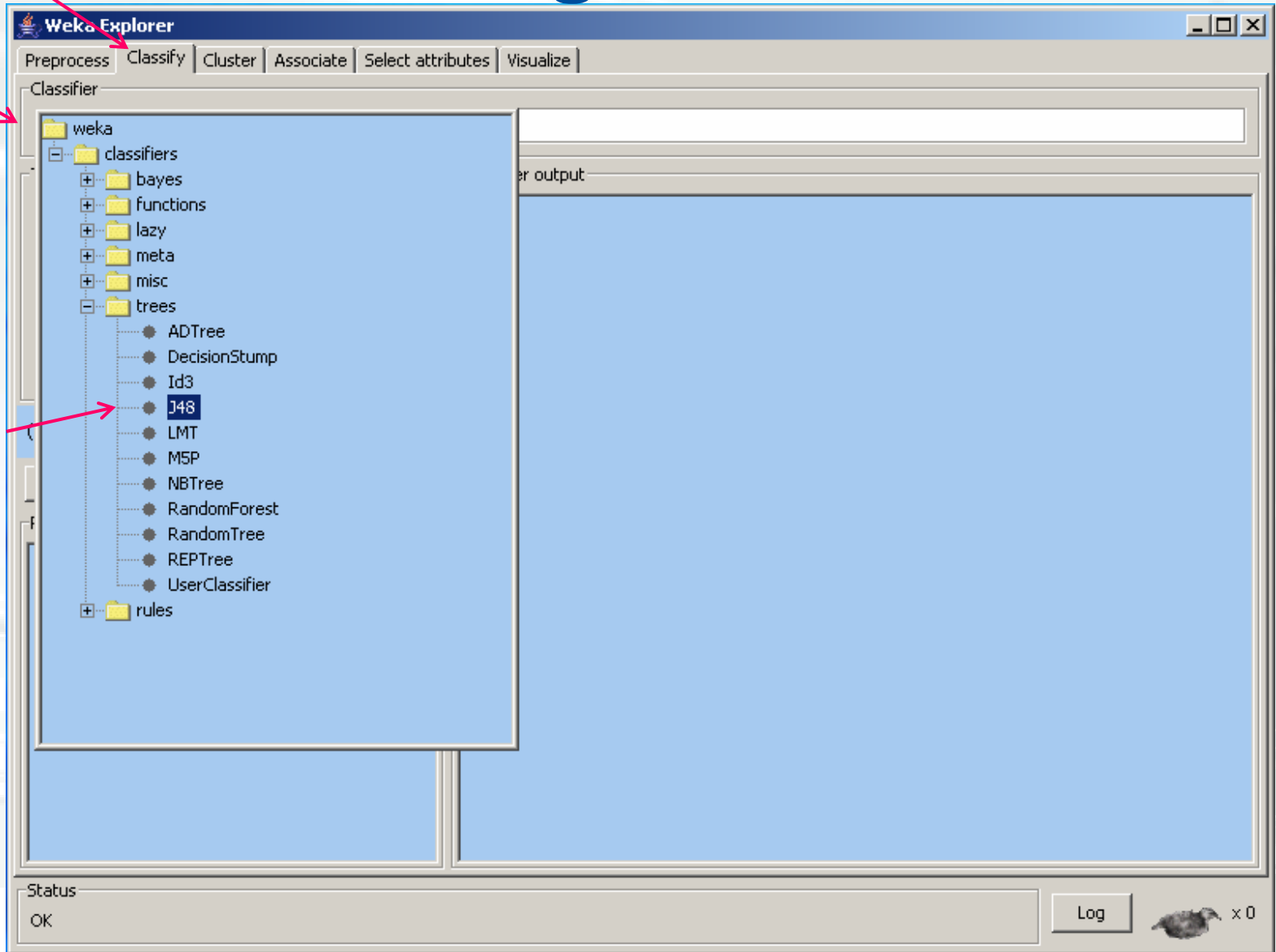
Below the table, a bar chart visualizes the counts for the 'y' variable. The bars are colored blue (1210), red (384), cyan (65), and dark green (69). A blue arrow points to the 'Class: y (Nom)' dropdown menu and the bar chart, with the text 'Target variable' next to it. The status bar at the bottom shows 'OK' and 'Log' buttons.

1

# Choose algorithm J48

2

3



# Building and evaluating the tree

1



2



The screenshot shows the Weka Explorer application window. At the top, there are tabs for 'Preprocess', 'Classify', 'Cluster', 'Associate', 'Select attributes', and 'Visualize'. The 'Classify' tab is active, and the classifier is set to 'J48 -C 0.25 -M 2'. Below this, the 'Test options' section is visible, with 'Cross-validation' selected and 'Folds' set to 10. A 'Start' button is located below the test options. To the right of the test options is a large 'Classifier output' area. At the bottom, there is a 'Result list' and a 'Status' bar showing 'OK'. A 'Log' button and a small icon are also present in the bottom right corner.

**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
- Percentage split %

(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

**Classification accuracy** (92.3611 %)

**Classified as**

**Actual values**

Status: OK   x 0

**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
- 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 %	
Number 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	1164	43	0	3   a = unacc
b	33	333	7	11   b = acc
c	0	3	58	4   c = v-good
d	0	17	11	41   d = good

Context menu for result list:

- 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

Right mouse click

# Tree pruning

1

Parameters of the algorithm (right mouse click)

2

Set the minimal number of objects per leaf to 15

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. The 'Classifier' dropdown is set to 'J48 -C 0.25 -M 15'. A dialog box titled 'weka.gui.GenericObjectEditor' is open, showing the settings for the 'weka.classifiers.trees.J48' classifier. The 'minNumObj' parameter is highlighted with a red arrow and set to 15. Other parameters include 'confidenceFactor' (0.25), 'numFolds' (3), and 'subtreeRaising' (True). The background shows the 'Result list' with classification results for 'trees.J48'.

Measure	Class
92.3611 %	unacc
7.6389 %	acc
843	v-good
421	good
718	
833 %	
1.76 %	
967	unacc
854	acc
823	v-good
641	good

0 17 11 41 | d = good

**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

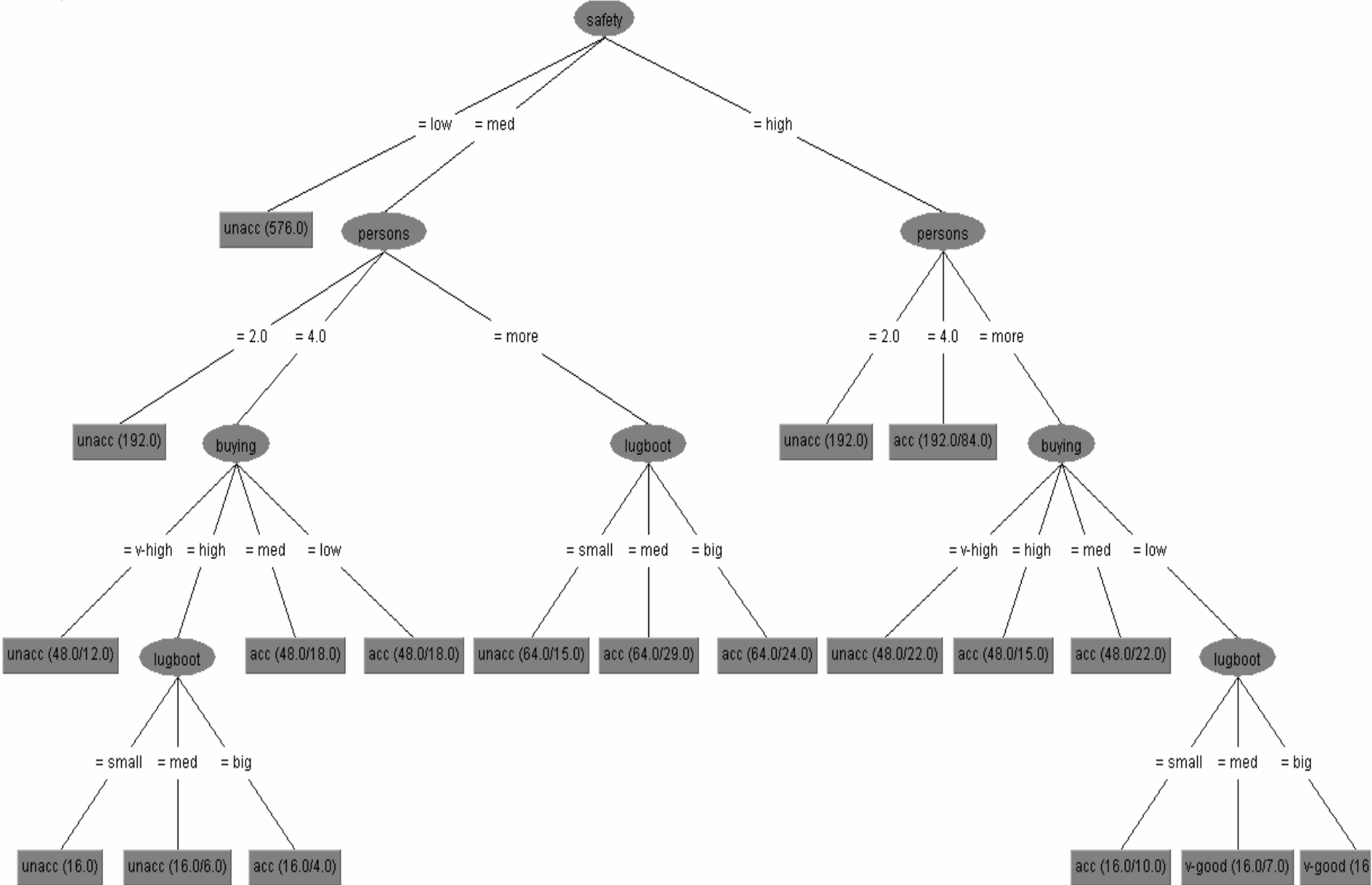
Status: OK

Log x 0

Reduced number of leaves and nodes

Easier to interpret

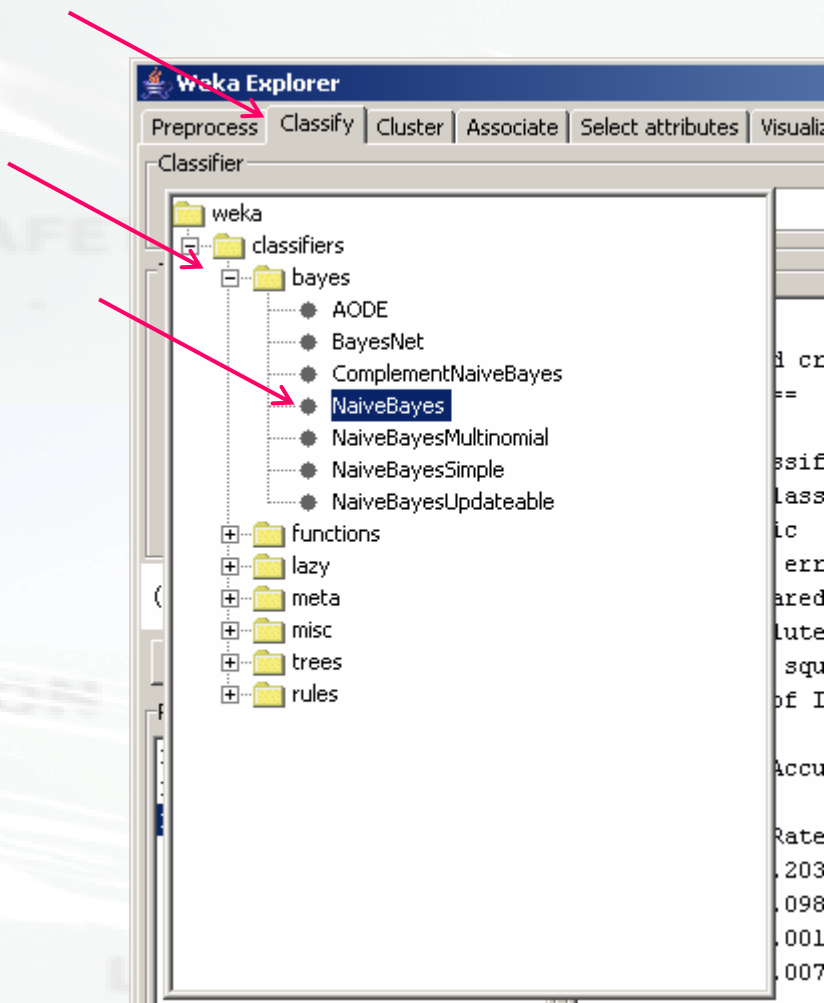
Lower classification accuracy



LANGUAGE



# Naïve Bayes classifier



**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

