

Hand on Weka

2014/11/11

Petra Kralj Novak

Petra.Kralj.Novak@ijs.si

Data Mining Tools

- Weka <http://www.cs.waikato.ac.nz/ml/weka/>
- Orange <http://orange.biolab.si/>
- Knime <http://www.knime.org/>
- Taverna <http://www.taverna.org.uk/>
- Rapid Miner <http://rapid-i.com/content/view/181/196/>
- ClowdFlows <http://clowdflows.org/>

Weka (Waikato Environment for Knowledge Analysis)

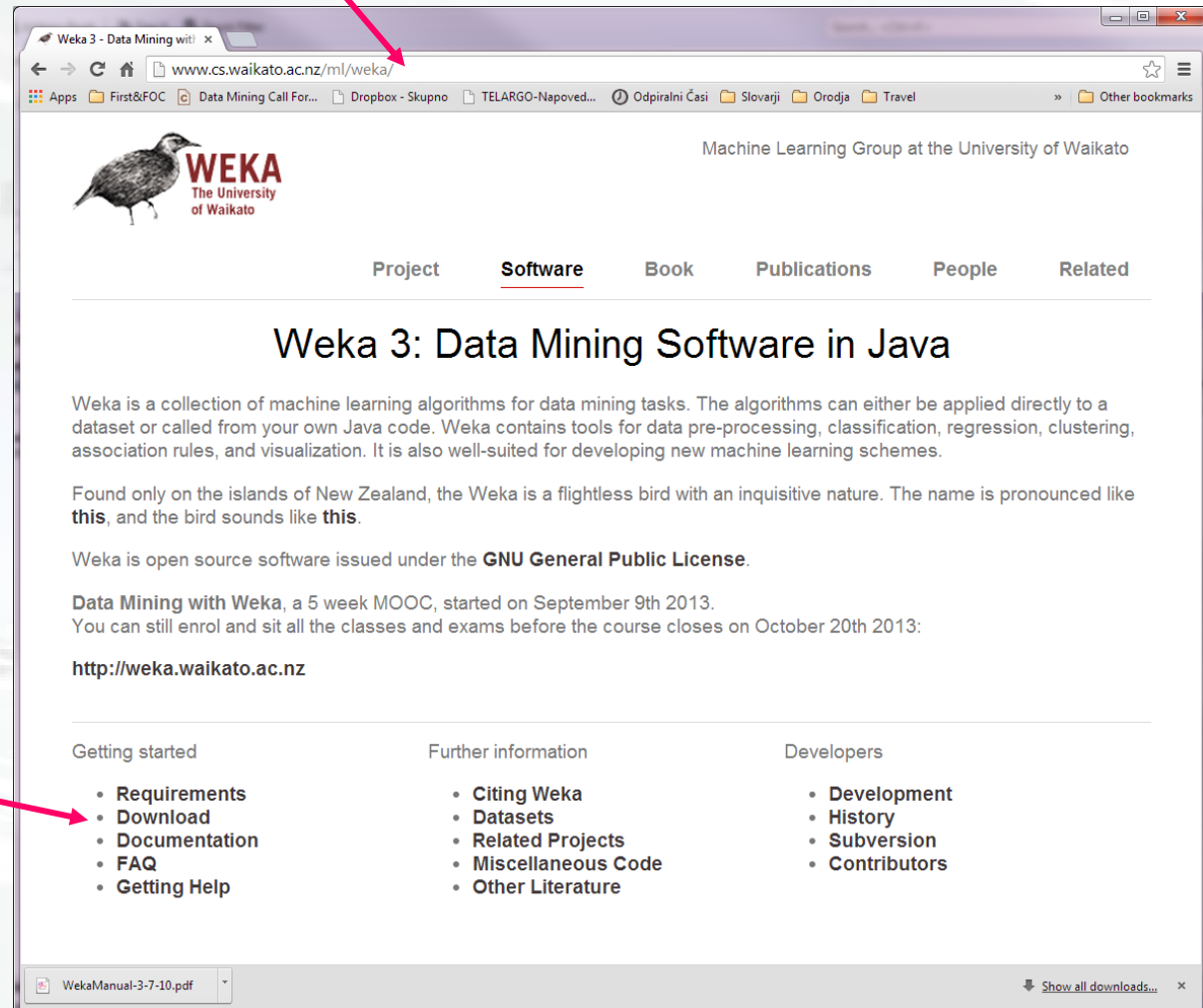
- Collection of machine learning algorithms for data mining tasks
- The algorithms
 - Can be applied directly to a dataset
 - Can be called from Java code (library)
- Weka contains tools for
 - Data pre-processing
 - Classification
 - Regression
 - Clustering
 - Association rules
 - Visualization
- Weka is open source software issued under the GNU General Public License

Exsercise1: ID3 in Weka

1. Build a decision tree with the ID3 algorithm on the lenses dataset, evaluate on a separate test set

Weka: Install

<http://www.cs.waikato.ac.nz/ml/weka/>



Download
version
3.6

Weka: Run Explorer



Choose Explorer

Exercise 1: ID3 in Weka

- 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.

Load the data

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: None
Instances: None
Attributes: None

Attributes
All | None | Invert

Selected attribute
Name: None
Missing: None
Distinct: None
Type: None
Unique: None

Visualize All

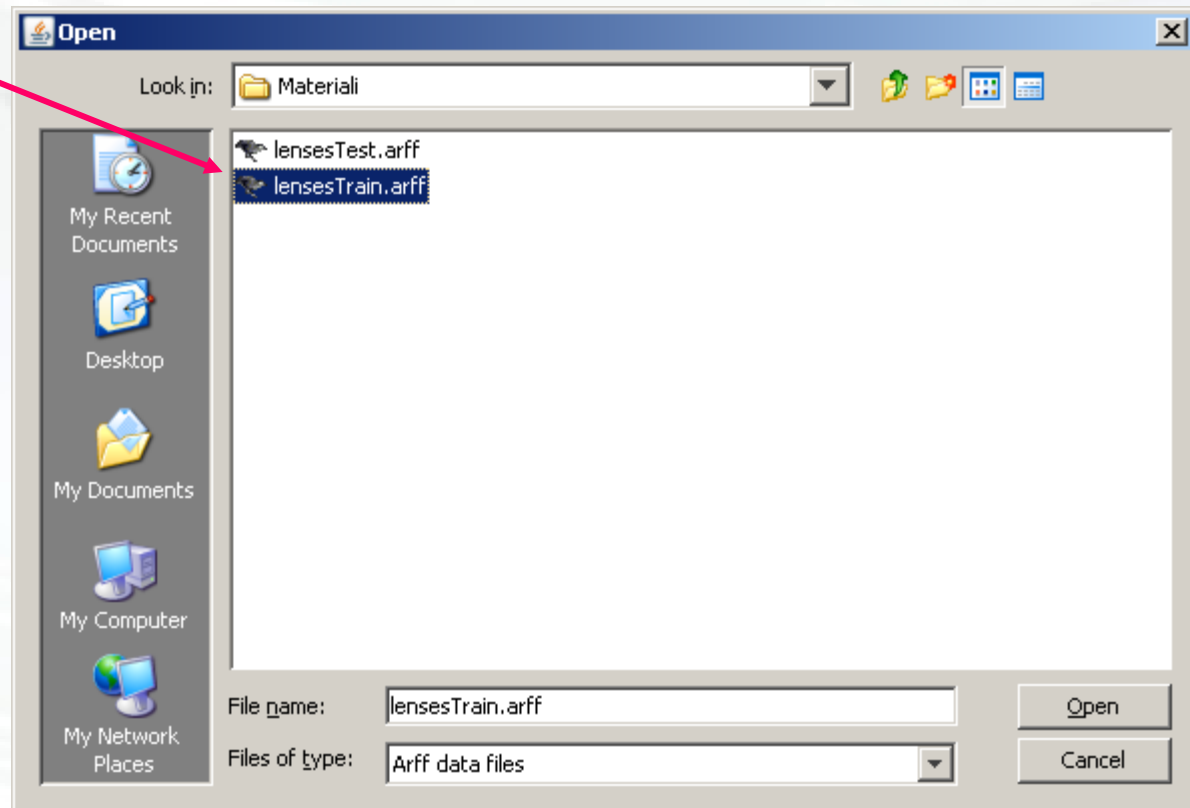
Remove

Status
Welcome to the Weka Explorer

Log x 0

Load the data - 2

lensesTrain.arff



The data are loaded

Choose
"Classify"

The screenshot shows the Weka Explorer application window. The 'Classify' tab is selected in the top menu. A red arrow points to the 'Classify' button. Below the menu, there are buttons for 'Open file...', 'Open URL...', 'Open DB...', 'Undo', 'Edit...', and 'Save...'. The 'Filter' section shows 'None' selected. The 'Current relation' section displays 'Relation: lensesTrain' and 'Instances: 17'. The 'Attributes' section lists five attributes: Age, Prescription, Astigmatic, Tear_rate, and Lenses. A red arrow points to the 'Age' attribute, which is highlighted in the list. The 'Selected attribute' section shows 'Name: Age', 'Missing: 0 (0%)', 'Distinct: 3', and 'Type: Nominal'. Below this is a table with columns 'Label' and 'Count':

Label	Count
young	7
pre-presbyopic	3
presbyopic	7

The 'Class: Lenses (Nom)' dropdown is visible, with a red arrow pointing to it. Below the dropdown are three stacked bar charts representing the distribution of the 'Lenses' class across the 'Age' categories. The first bar (young) has a total height of 7, with a blue segment at the bottom and a red segment on top. The second bar (pre-presbyopic) has a total height of 3, with a blue segment at the bottom and a red segment on top. The third bar (presbyopic) has a total height of 7, with a blue segment at the bottom and a red segment on top. The status bar at the bottom shows 'OK' and a 'Log' button.

Target variable

Choose algoritem

The screenshot shows the Weka Explorer application window. The title bar reads "Weka Explorer". The main menu includes "Preprocess", "Classify", "Cluster", "Associate", "Select attributes", and "Visualize". The "Classify" tab is active, and the "Classifier" section shows "ZeroR" selected. A red arrow points to the "Choose" button next to "ZeroR".

Test options:

- Use training set
- Supplied test set (Set...)
- Cross-validation (Folds: 10)
- Percentage split (%: 66)

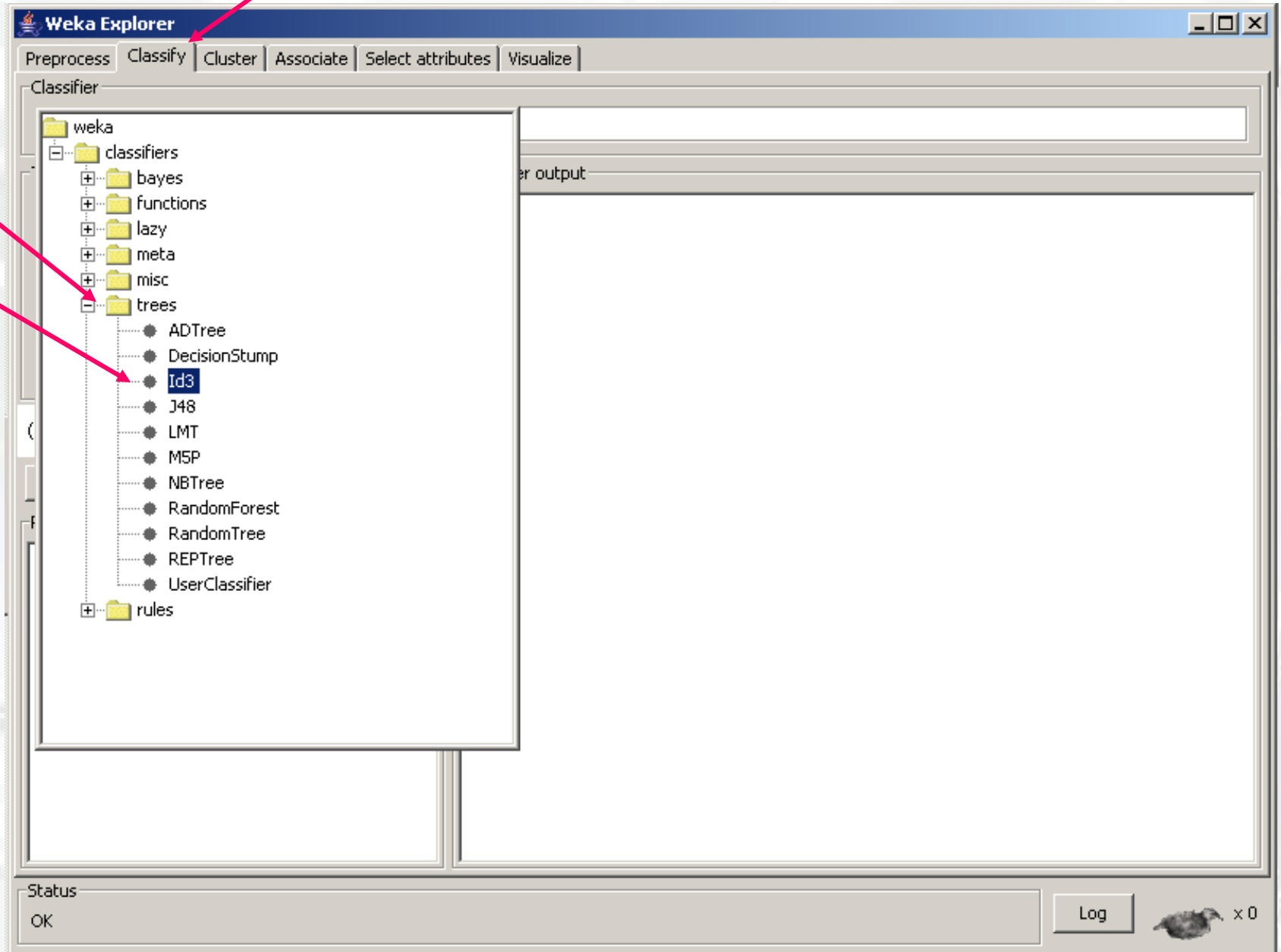
Classifier output: (Empty text area)

Result list (right-click for options): (Empty list area)

Buttons: Start, Stop, More options...

Status: OK

Log: Log button with a small icon and "x 0" next to it.



trees

Id3

The image shows the Weka Explorer software interface with an 'Open' dialog box overlaid. The Weka Explorer window has tabs for 'Preprocess', 'Classify', 'Cluster', 'Associate', 'Select attributes', and 'Visualize'. The 'Classify' tab is active, showing the 'Classifier' dropdown set to 'Id3'. Below this, the 'Test options' section has 'Supplied test set' selected, with a 'Set...' button. The 'Classifier output' area is empty. The 'Open' dialog box shows the 'Look in:' directory as 'Materiali'. Two files are listed: 'lensesTest.arff' and 'lensesTrain.arff'. The 'lensesTest.arff' file is selected and highlighted. The 'File name:' field at the bottom of the dialog contains 'lensesTest.arff' and the 'Files of type:' dropdown is set to 'Arff data files'. The 'Open' button is visible at the bottom right of the dialog.

1. Classifier dropdown (Id3)

2. Set... button

3. Open file... button

4. Selected file (lensesTest.arff)

5. Start button

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

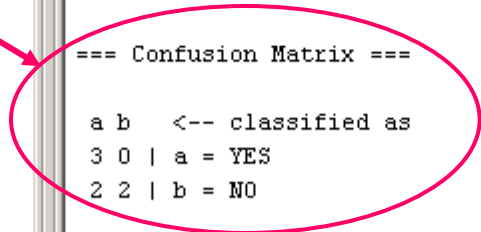
Status: OK

Log x 0

Classification accuracy



Confusion matrix



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

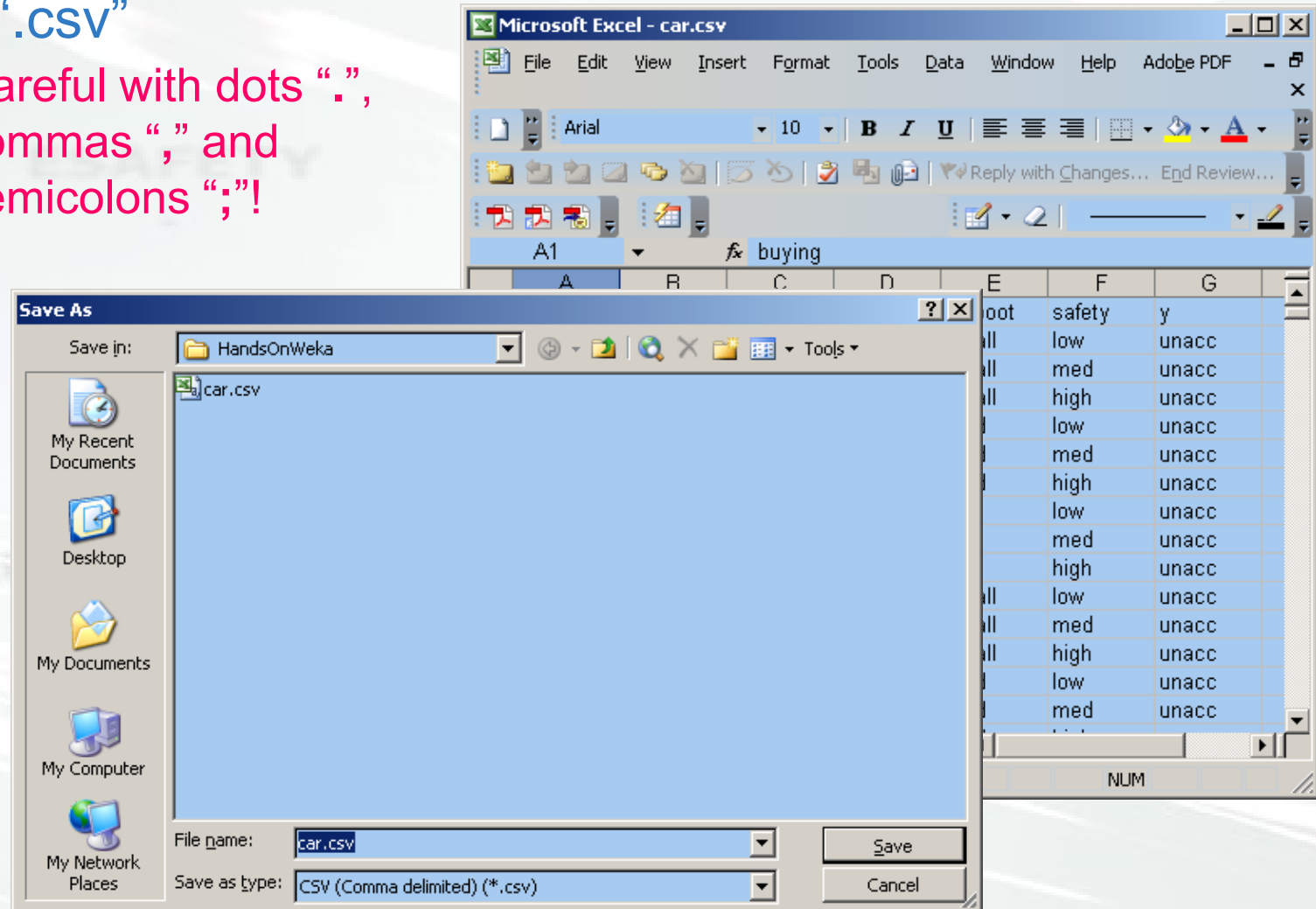
Microsoft Excel - car.csv

	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 “;”!



Load the data

Car.csv



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 interface displays the 'Current relation' as 'car' with 1728 instances and 7 attributes. The 'Attributes' list includes 'buying', 'maint', 'doors', 'persons', 'lugboot', 'safety', and 'y'. The 'Selected attribute' section shows 'y' with 4 distinct values and 0 missing values. A table below shows the distribution of 'y': unacc (1210), acc (384), v-good (65), and good (69). A bar chart at the bottom visualizes this distribution, with a blue arrow pointing to the 'y (Nom)' dropdown menu and the text 'Target variable'.

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

Choose algorithm J48

The image shows a screenshot of the Weka Explorer software interface. The window title is "Weka Explorer". At the top, there are several tabs: "Preprocess", "Classify", "Cluster", "Associate", "Select attributes", and "Visualize". The "Classify" tab is currently selected. Below the tabs, there is a "Classifier" section. On the left side of this section, there is a tree view showing the directory structure of classifiers. The tree starts with "weka", which contains a sub-directory "classifiers". Under "classifiers", there are several sub-directories: "bayes", "functions", "lazy", "meta", "misc", "trees", and "rules". The "trees" sub-directory is expanded, showing a list of classifiers: "ADTree", "DecisionStump", "Id3", "J48", "LMT", "MSP", "NBTree", "RandomForest", "RandomTree", "REPTree", and "UserClassifier". The "J48" classifier is highlighted with a blue selection box. Three red arrows with numbered circles (1, 2, 3) point to specific elements: arrow 1 points to the "Classify" tab, arrow 2 points to the "classifiers" directory, and arrow 3 points to the "J48" classifier. The right side of the "Classifier" section is a large empty area. At the bottom of the window, there is a "Status" bar showing "OK", a "Log" button, and a small icon of a dog.

Building and evaluating the tree

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

Cross-validation Folds

Percentage split %

(Nom) y

Classifier output

Result list (right-click for options)

Status

OK 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 (Set...)
 Cross-validation Folds **10**
 Percentage split % **66**
More options...

(Nom) y

Start Stop

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

Classification accuracy (92.3611 %)

=== 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

Classified as (points to 'classified as' column)

Actual values (points to 'a', 'b', 'c', 'd' columns)

Status: OK Log 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 (Set...)
- Cross-validation Folds: **10**
- Percentage split %: **66**

 More options...

(Nom) y

Start Stop

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	
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

Status: OK

Log  x 0

Right mouse click

- 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

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 configuration for the J48 classifier. The 'minNumObj' setting is set to 15, which is highlighted by a red arrow. Other settings include 'binarySplits' (False), 'confidenceFactor' (0.25), 'debug' (False), 'numFolds' (3), 'reducedErrorPruning' (False), 'saveInstanceData' (False), 'seed' (1), 'subtreeRaising' (True), 'unpruned' (False), and 'useLaplace' (False). The background shows the 'Result list' with the following data:

Time	Classifier	Accuracy
14:55:00	- trees.J48	92.3611 %
14:58:13	- trees.J48	7.6389 %

The status bar at the bottom shows 'OK' and a 'Log' button.

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

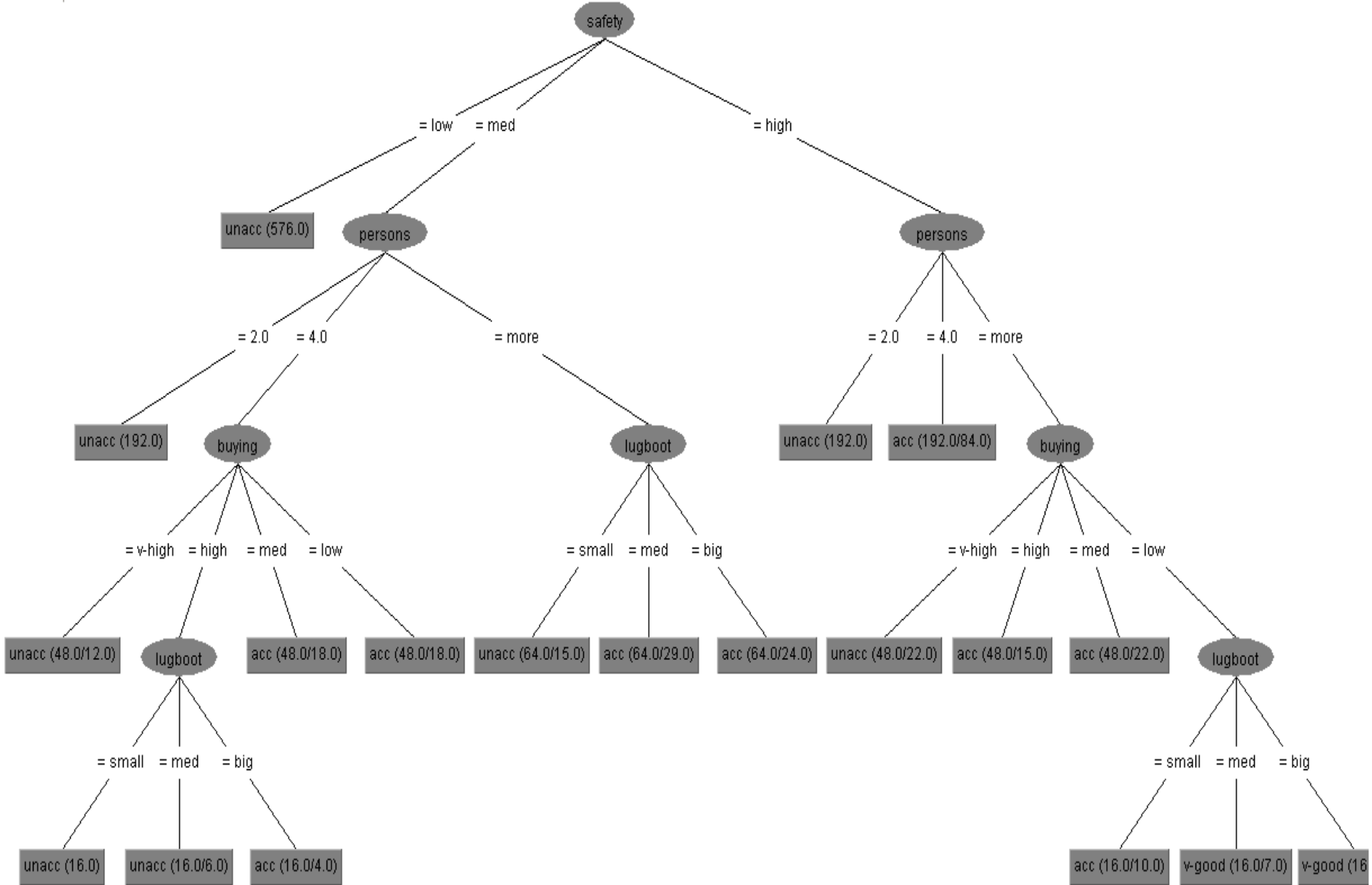
Status: OK

Log x 0

Reduced number of leaves and nodes

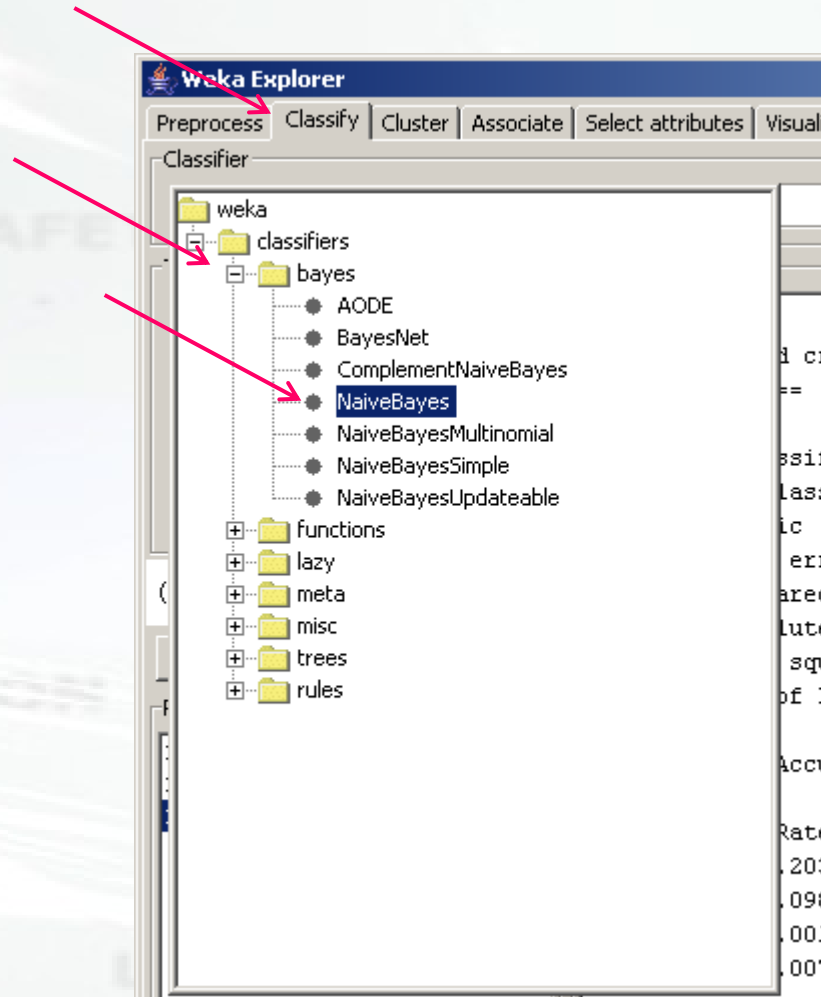
Easier to interpret

Lower classification accuracy



LANGUAGE

Naïve Bayes classifier



Weka Explorer

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

Classifier: **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



Weka Explorer

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

Classifier: Choose **NaiveBayes**

Test options:
 Use training set
 Supplied test set (Set...)
 Cross-validation Folds:
 Percentage split %:
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:

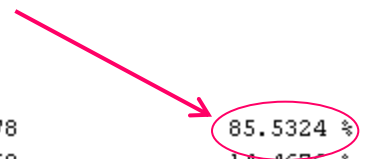
```
=== Stratified cross-validation ===  
=== Summary ===  
Correctly Classified Instances      1478      85.5324 %  
Incorrectly Classified Instances    250       14.4676 %  
Kappa statistic                     0.6665  
Mean absolute error                  0.1137  
Root mean squared error              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          |


```



Summary

- Weka
- ID3, separate test set
- Data preparation
- J48 (C4.5), cross validation, tree pruning
- Naïve Bayes