

Petra.Kralj@ijs.si

Hands on WEKA

2006.11.15

Petra Kralj

Petra.Kralj@ijs.si

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)

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

CLASSIFICATION

CAR dataset

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

Preparing the data for WEKA - 1

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

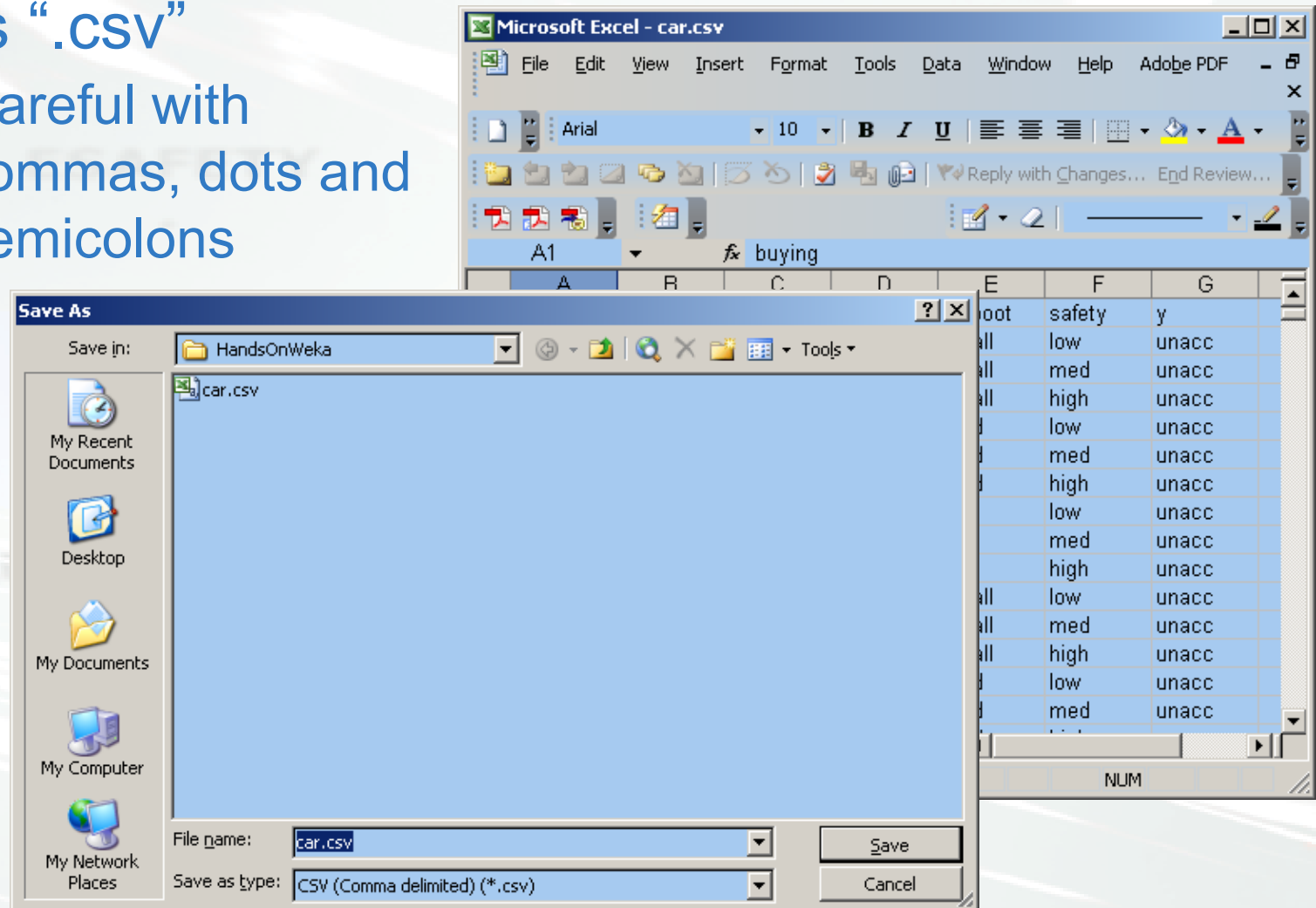
- Rows are instances
- Columns are attributes
- The last column is the target attribute

	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

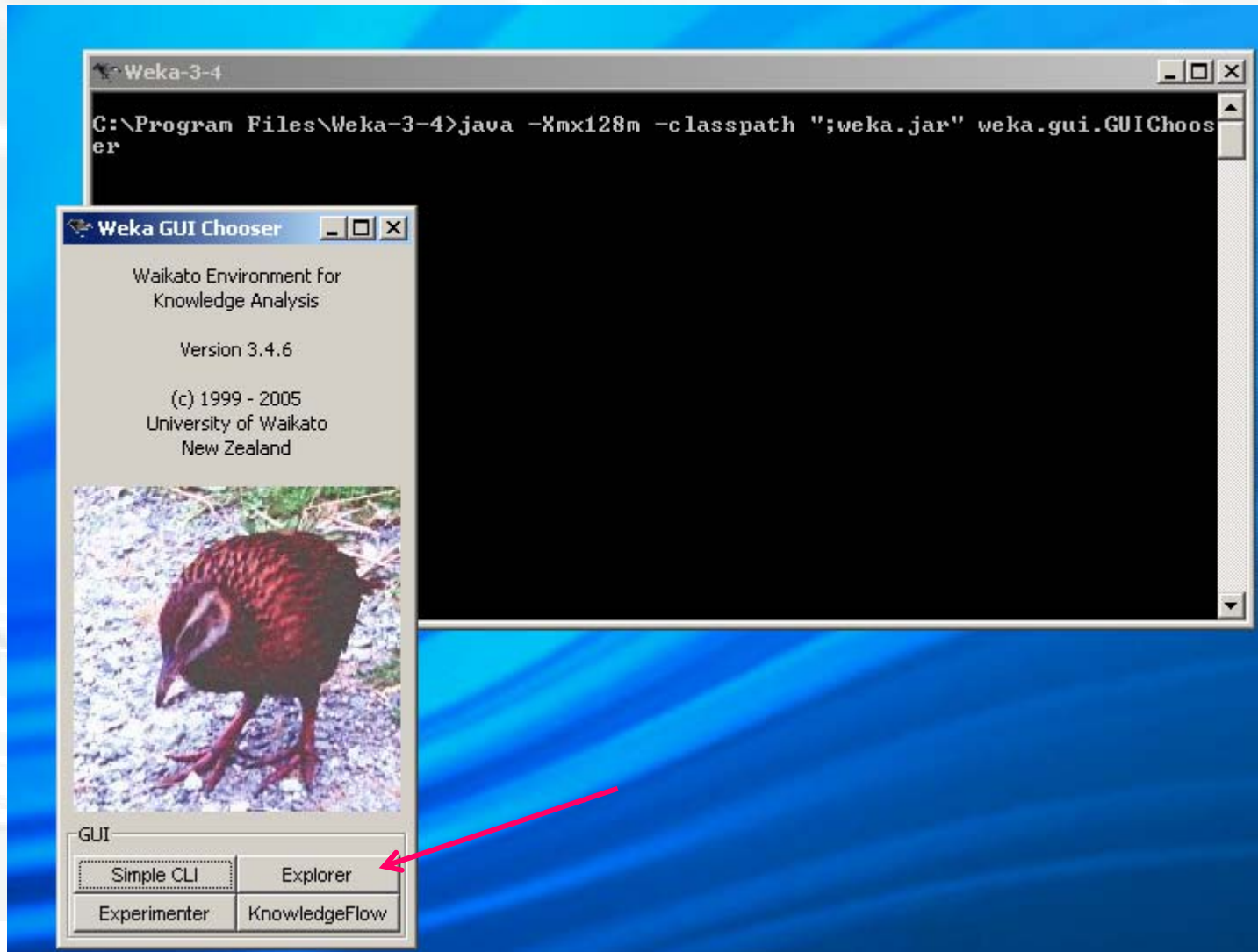
Preparing the data for WEKA - 2

Save as “.csv”

- Careful with commas, dots and semicolons



Open WEKA Explorer



1

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: car Instances: 1728 Attributes: 7

Attributes: All | None | Invert

No.	Name
1	<input type="checkbox"/> buying
2	<input type="checkbox"/> maint
3	<input type="checkbox"/> doors
4	<input type="checkbox"/> persons
5	<input type="checkbox"/> lugboot
6	<input type="checkbox"/> safety
7	<input checked="" type="checkbox"/> y

Remove

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

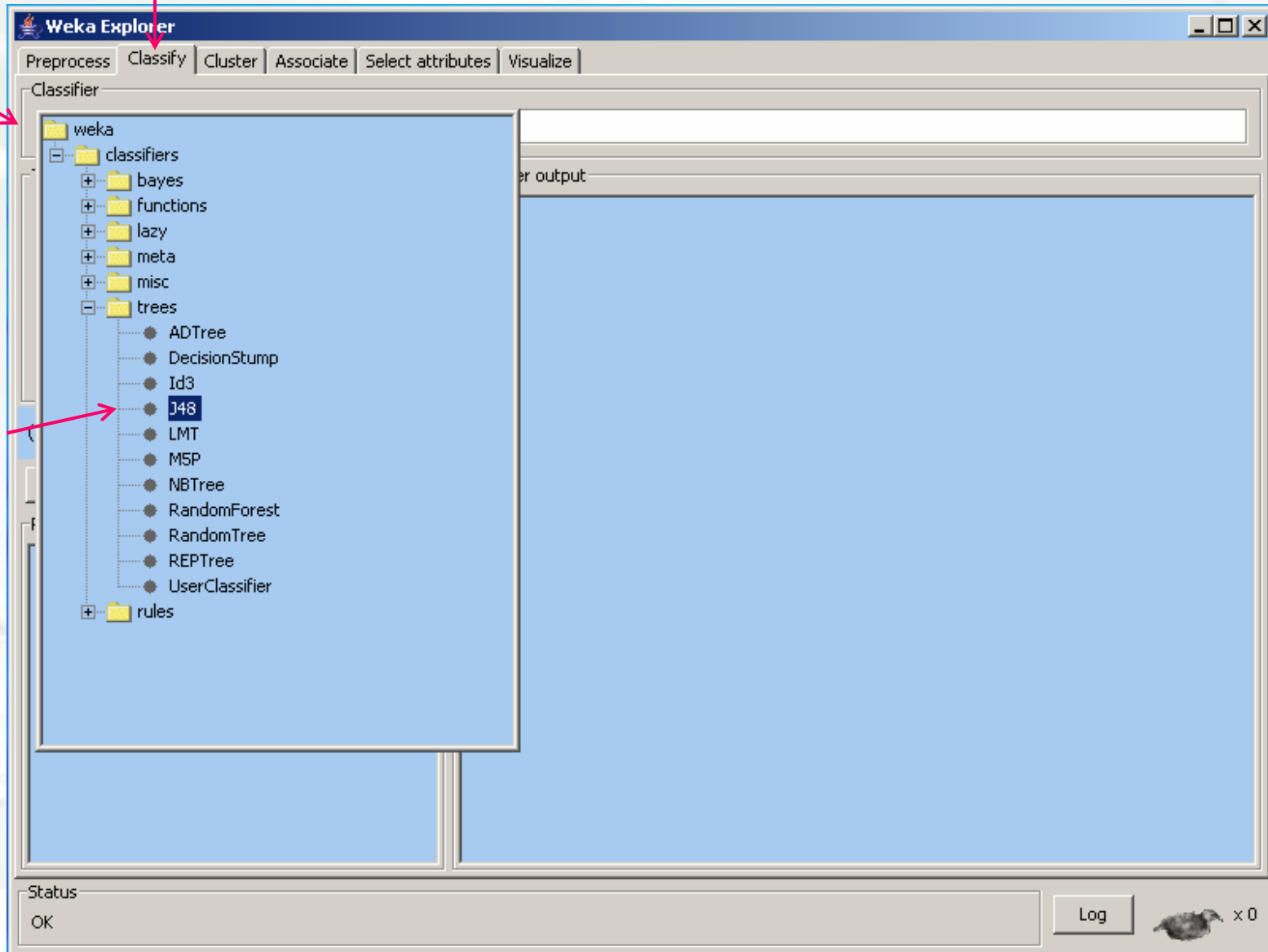
Class: y (Nom) Visualize All

Target variable

Status: OK Log x 0

Choose a tree

1



2

3

Build tree + evaluate

The screenshot shows the Weka Explorer interface with the following components:

- Classifier:** A dropdown menu showing "J48 -C 0.25 -M 2".
- Test options:** Radio buttons for "Use training set", "Supplied test set", "Cross-validation", and "Percentage split". The "Cross-validation" option is selected, with "Folds" set to 10 and "Percentage split" set to 66. A "More options..." button is also present.
- Classifier output:** A large empty text area for displaying results.
- Result list (right-click for options):** An empty list area for showing individual results.
- Buttons:** "Start" and "Stop" buttons are located below the test options.
- Status:** A status bar at the bottom left shows "OK".
- Log:** A "Log" button is located at the bottom right.

Annotations on the left side of the image:

- A red arrow labeled "1" points to the "Cross-validation" radio button.
- A red arrow labeled "2" points to the "Start" button.

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 %)

92.3611 %
7.6389 %

=== 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

Predicted class (points to 'classified as')

Actual class (points to 'a', 'b', 'c', 'd')

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
- 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
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  x 0

Right 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

Algorithm parameters

2

Set the minimum number of examples in a leaf

The screenshot shows the Weka Explorer interface with the J48 classifier selected. A dialog box titled 'weka.gui.GenericObjectEditor' is open, displaying the configuration for 'weka.classifiers.trees.J48'. The 'minNumObj' parameter is set to 15, which is highlighted by a red arrow from the text 'Set the minimum number of examples in a leaf'. Other parameters include confidenceFactor (0.25), numFolds (3), and subtreeRaising (True). The background shows the 'Classifier output' window with a table of results.

Measure	Class
92.3611 %	
7.6389 %	
843	
421	
718	
833 %	
176 %	
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

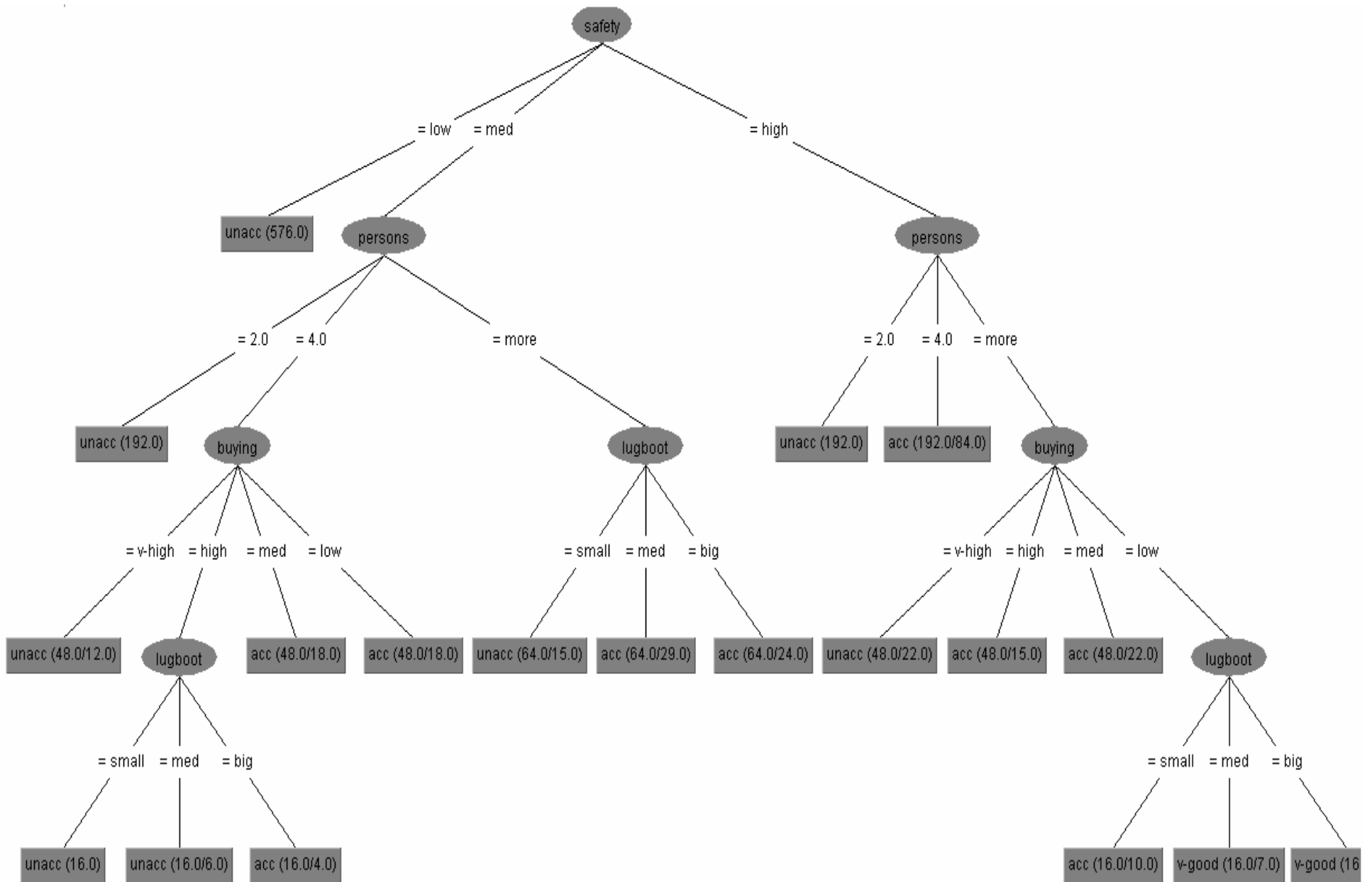
=== Confusion Matrix ===

	a	b	c	d	<-- classified as
1098	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

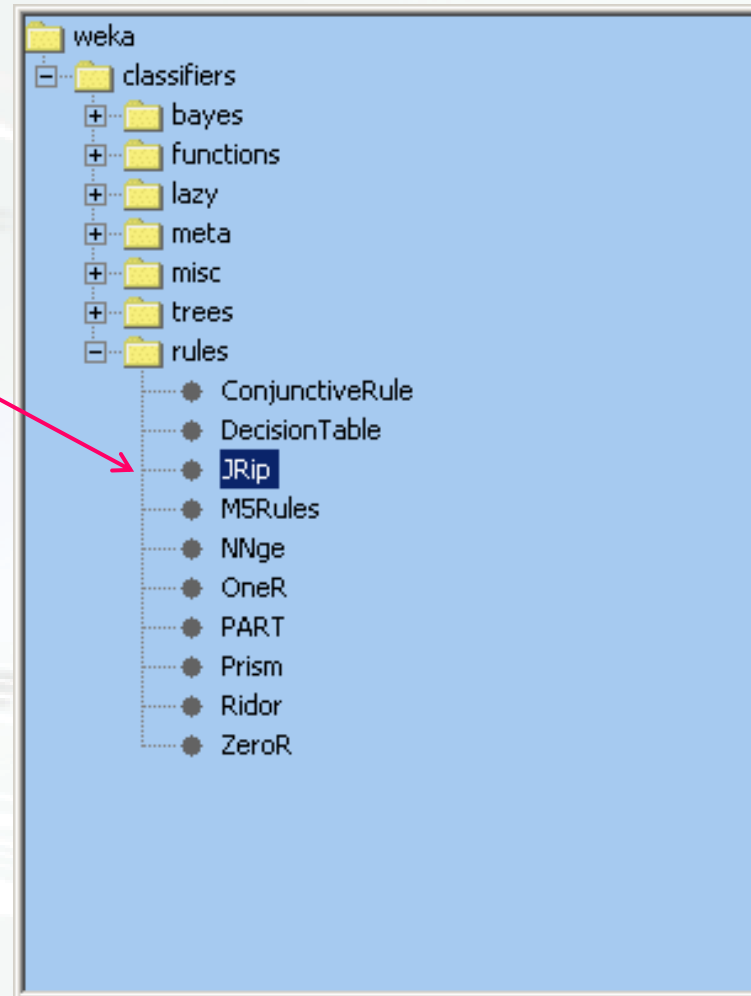
All nodes including leaves

higher understandability, lower accuracy



LANGUAGE

Building classification rules

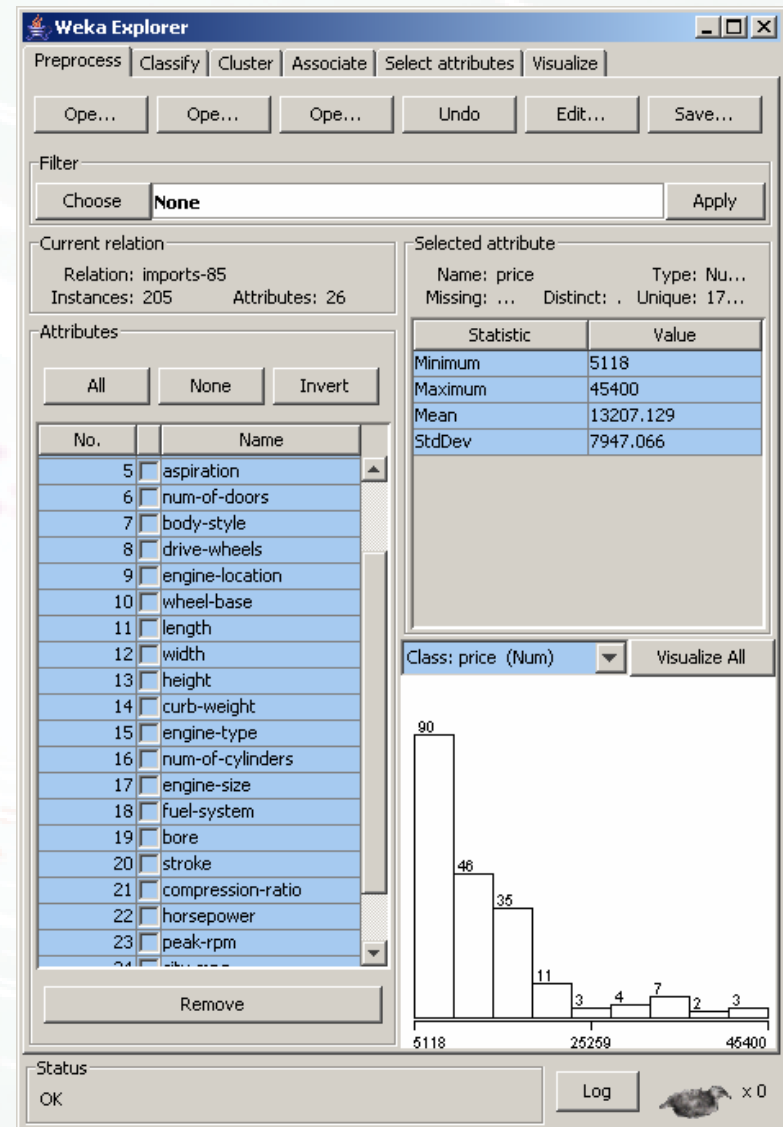


REGRESSION

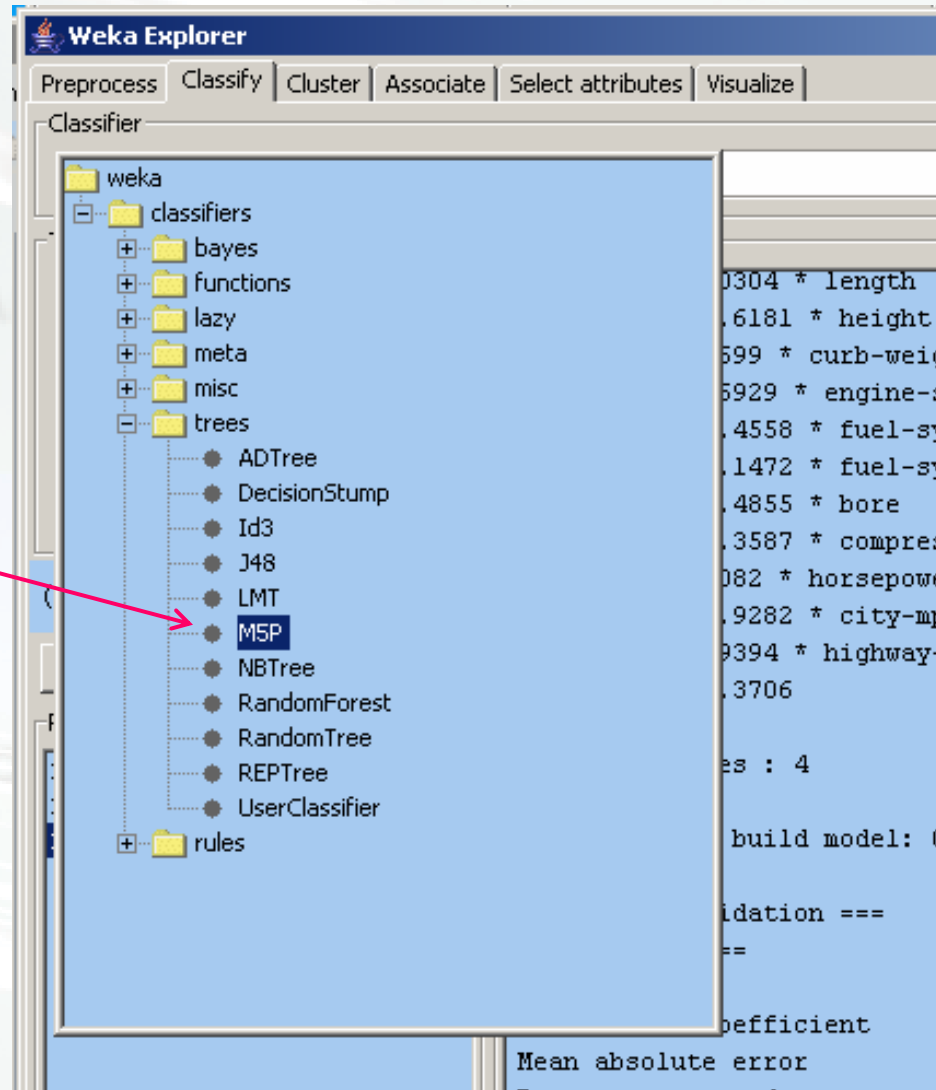
Imports-85 dataset

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%)



Model & regression tree



Evaluating regression model

Weka Explorer

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

Classifier: Choose MSP -M 4.0

Test options:

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

More options...

(Num) price

Start Stop

Result list (right-click for options)

- 15:21:19 - trees.MSP
- 15:40:35 - trees.J48
- 15:55:13 - trees.MSP

Status: OK

Classifier output:

```

+ 18.9394 * highway-mpg
- 517.3706

Number of Rules : 4

Time taken to build model: 0.39 seconds

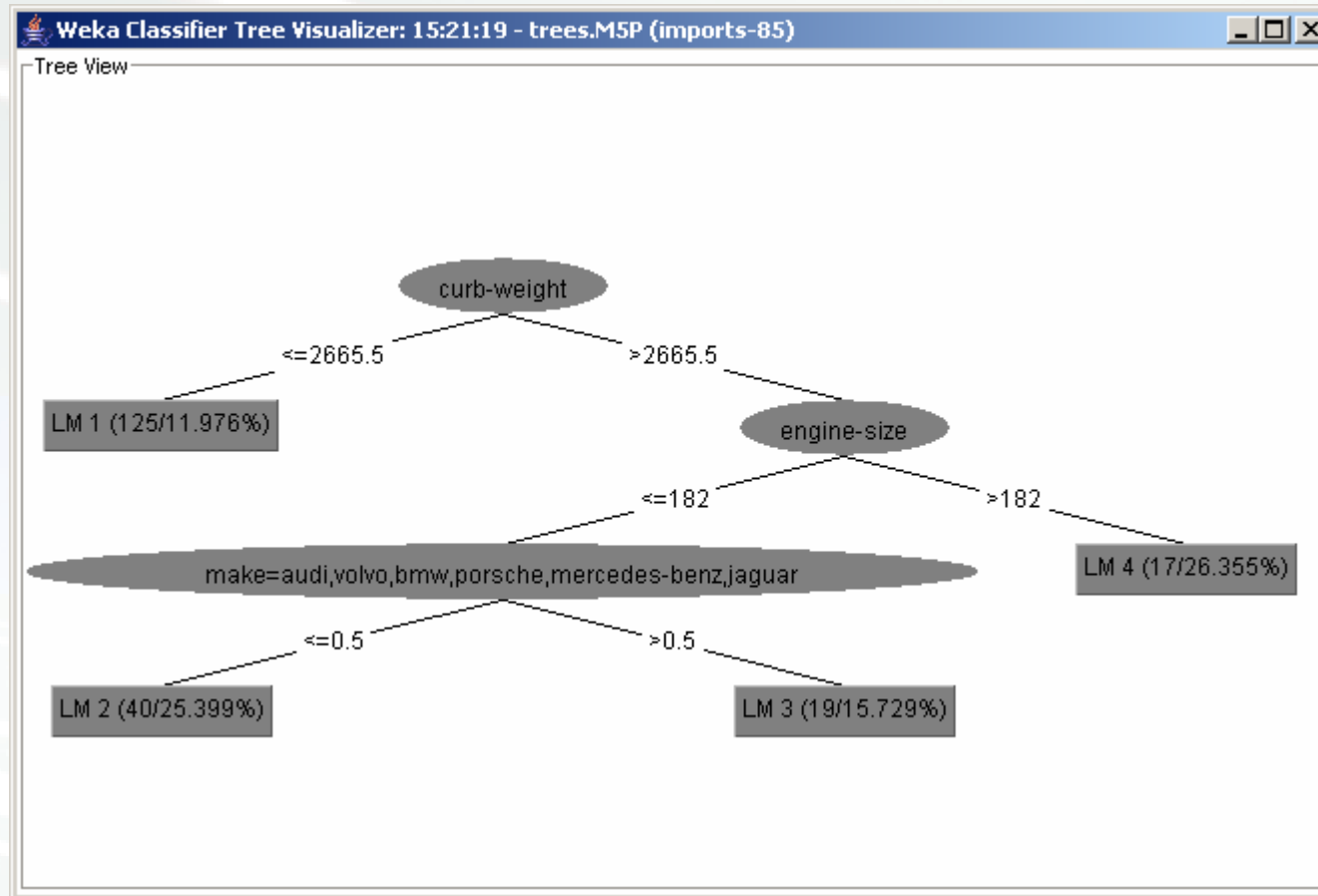
=== Cross-validation ===
=== Summary ===

Correlation coefficient          0.9599
Mean absolute error             1524.3286
Root mean squared error         2224.9209
Relative absolute error         25.8998 %
Root relative squared error     27.8279 %
Total Number of Instances      201
Ignored Class Unknown Instances 4
  
```

Root Mean Squared Error

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

Model tree visualization



LM = linear model

Regression trees

Weka Explorer

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

Classifier: Choose → **M5P -R -M 4.0**

Test options: Use training set, Supplied test set, Cross-validation, Percentage

Classifier output: `weka.classifiers.trees.M5P`

weka.gui.GenericObjectEditor

About: The original algorithm M5 was invented by Quinlan: Quinlan J. More

buildRegressionTree: **True**

debug: False

minNumInstances: 4.0

saveInstances: False

unpruned: False

useUnsmoothed: False

Start

Result list (right): 15:21:19 - tree, 15:40:35 - tree, 15:55:13 - tree

Status: OK

Log x 0

9599
3286
9209
8998 %
8279 %
4

1

2

Evaluation of the regression tree

Weka Explorer

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

Classifier: Choose **MSP -R -M 4.0**

Test options:

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

(Num) price

Result list (right-click for options):

- 15:21:19 - trees.MSP
- 15:40:35 - trees.J48
- 15:55:13 - trees.MSP
- 16:03:33 - trees.MSP

Classifier output:

```

Number of Rules : 9
Time taken to build model: 0.34 seconds
=== Cross-validation ===
=== Summary ===
Correlation coefficient           0.9085
Mean absolute error              2551.3974
Root mean squared error          4026.5874
Relative absolute error          43.3507 %
Root relative squared error      50.362 %
Total Number of Instances       201
Ignored Class Unknown Instances 4
  
```

Status: OK x 0

Larger error compared to the model tree

Understandability of regression and model trees

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

```
M5 pruned regression tree:  
(using smoothed linear models)
```

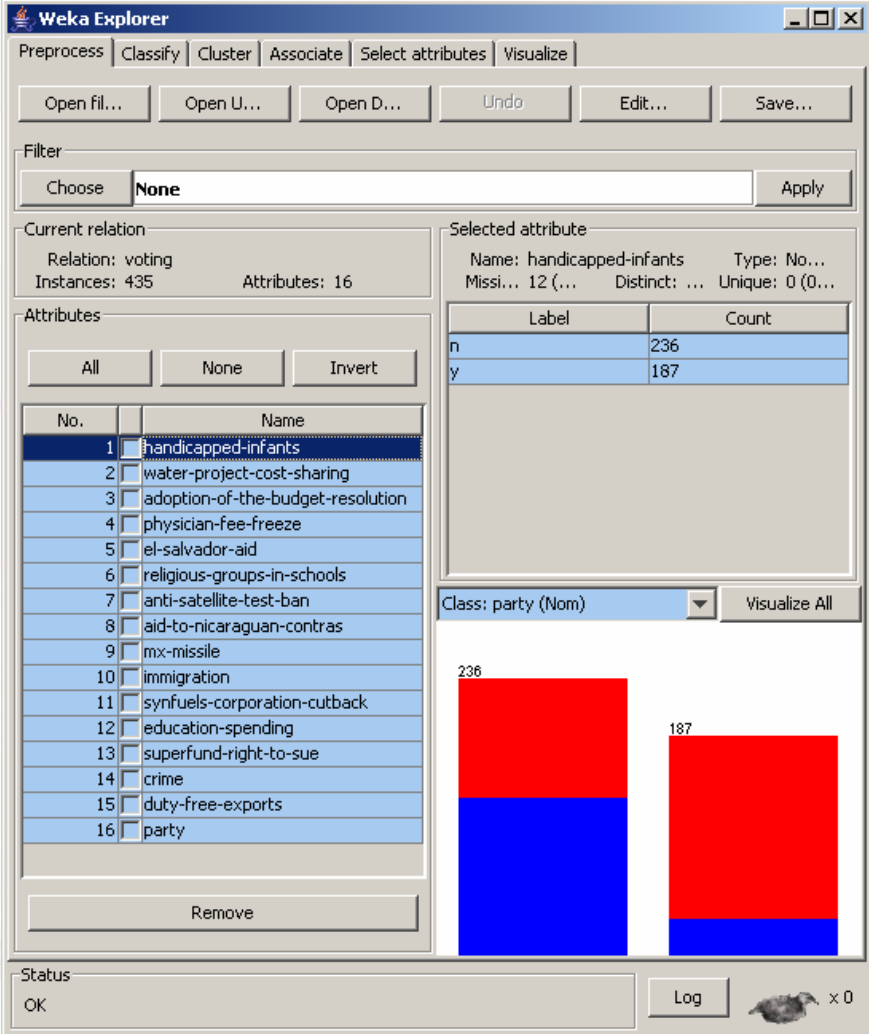
```
curb-weight <= 2665.5 :  
|   curb-weight <= 2291.5 :  
|   |   curb-weight <= 2121 :  
|   |   |   length <= 160.75 : LM1 (27/8.596%)  
|   |   |   length > 160.75 : LM2 (17/7.78%)  
|   |   |   curb-weight > 2121 : LM3 (27/11.735%)  
|   |   curb-weight > 2291.5 :  
|   |   |   fuel-system=spfi,4bbl,mfi,idi,mpfi <= 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%)
```

Descriptive induction

Voting dataset
Iris dataset

Voting dataset

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



Weka Explorer

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

Open fil... | Open U... | Open D... | Undo | Edit... | Save...

Filter: Choose **None** Apply

Current relation
Relation: voting
Instances: 435 Attributes: 16

Selected attribute
Name: handicapped-infants Type: No...
Missi... 12 (... Distinct: ... Unique: 0 (0...)

Label	Count
n	236
y	187

Class: party (Nom) Visualize All

Attributes: All | None | Invert

No.	Name
<input checked="" type="checkbox"/>	1 handicapped-infants
<input type="checkbox"/>	2 water-project-cost-sharing
<input type="checkbox"/>	3 adoption-of-the-budget-resolution
<input type="checkbox"/>	4 physician-fee-freeze
<input type="checkbox"/>	5 el-salvador-aid
<input type="checkbox"/>	6 religious-groups-in-schools
<input type="checkbox"/>	7 anti-satellite-test-ban
<input type="checkbox"/>	8 aid-to-nicaraguan-contras
<input type="checkbox"/>	9 mx-missile
<input type="checkbox"/>	10 immigration
<input type="checkbox"/>	11 synfuels-corporation-cutback
<input type="checkbox"/>	12 education-spending
<input type="checkbox"/>	13 superfund-right-to-sue
<input type="checkbox"/>	14 crime
<input type="checkbox"/>	15 duty-free-exports
<input type="checkbox"/>	16 party

Remove

Status: OK Log x 0

Association rules

1

2

Weka Explorer

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

Associator

Choose **Apriori** -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0

Start Stop

Associator output

Result list (right-click for o
16:25:34 - Apriori

Size of set of large itemsets L(4): 1

Best rules found:

1. adoption-of-the-budget-resolution=y physician-fee-freeze=n 219 ==> party=democrat 219 conf: (1)
2. adoption-of-the-budget-resolution=y physician-fee-freeze=n aid-to-nicaraguan-contras=y 198 ==> party=democrat 198 conf: (1)
3. physician-fee-freeze=n aid-to-nicaraguan-contras=y 211 ==> party=democrat 210 conf: (1)
4. physician-fee-freeze=n education-spending=n 202 ==> party=democrat 201 conf: (1)
5. physician-fee-freeze=n 247 ==> party=democrat 245 conf: (0.99)
6. el-salvador-aid=n party=democrat 200 ==> aid-to-nicaraguan-contras=y 197 conf: (0.99)
7. el-salvador-aid=n 208 ==> aid-to-nicaraguan-contras=y 204 conf: (0.98)
8. adoption-of-the-budget-resolution=y aid-to-nicaraguan-contras=y party=democrat 203 ==> physician-fee-freeze=n 203 conf: (0.97)
9. el-salvador-aid=n aid-to-nicaraguan-contras=y 204 ==> party=democrat 197 conf: (0.97)
10. aid-to-nicaraguan-contras=y party=democrat 218 ==> physician-fee-freeze=n 210 conf: (0.96)

Status
OK

Log x 0

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

Weka Explorer

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

Ope... | Ope... | Ope... | Undo | Edit... | Sav...

Filter: Choose **None** Apply

Current relation
Relation: iris
Instances: 150 Attributes: 5

Selected attribute
Name: iris Type: N...
Missing: ... Distinct: Unique: 0...

Label	Count
Iris-setosa	50
Iris-versicolor	50
Iris-virginica	50

Class: iris (Nom) Visualize All

50 50 50

Attributes: All None Inv...

No.	Name
1	sepal length
2	sepal width
3	petal length
4	petal width
5	iris

Remove

Status: OK Log x 0

Clustering

1

The screenshot shows the Weka Explorer interface with the 'Cluster' tab selected. In the 'Clusterer' list, 'SimpleKMeans' is highlighted. A dialog box titled 'weka.gui.GenericObjectEditor' is open, showing the configuration for 'weka.clusterers.SimpleKMeans'. The 'About' section states 'Cluster data using the k means algorithm'. The 'numClusters' field is set to 3, and the 'seed' field is set to 10. Below the configuration, a table shows the resulting cluster distribution:

0	61	(41%)
1	50	(33%)
2	39	(26%)

2

3

Clustering visualization

The screenshot displays the Weka Clusterer Visualize window for a SimpleKMeans model. The main window on the left shows the 'Cluster:' tab with a 'Visualize cluster assignments' option highlighted by a red arrow. The visualization window on the right shows a scatter plot of 'sepal length (Num)' on the X-axis and 'petal width (Num)' on the Y-axis. The plot contains three clusters of data points, represented by 'x' markers in red, blue, and green. A 'Jitter' slider is visible to the right of the plot. Below the plot, a 'Class colour' section shows the mapping: cluster0 (red), cluster1 (blue), and cluster2 (green).

Weka Clusterer Visualize: 17:12:22 - SimpleKMeans (iris-weka.filters.uns...)

X: sepal length (Num) Y: petal width (Num)

Colour: Cluster (Nom) Select Instance

Reset Clear Save Jitter

Plot: iris-weka.filters.unsupervised.attribute.Remove-R5_clustered

Class colour

cluster0 cluster1 cluster2