

# Hands on Weka: Part II

Petra Kralj Novak

18.11.2013

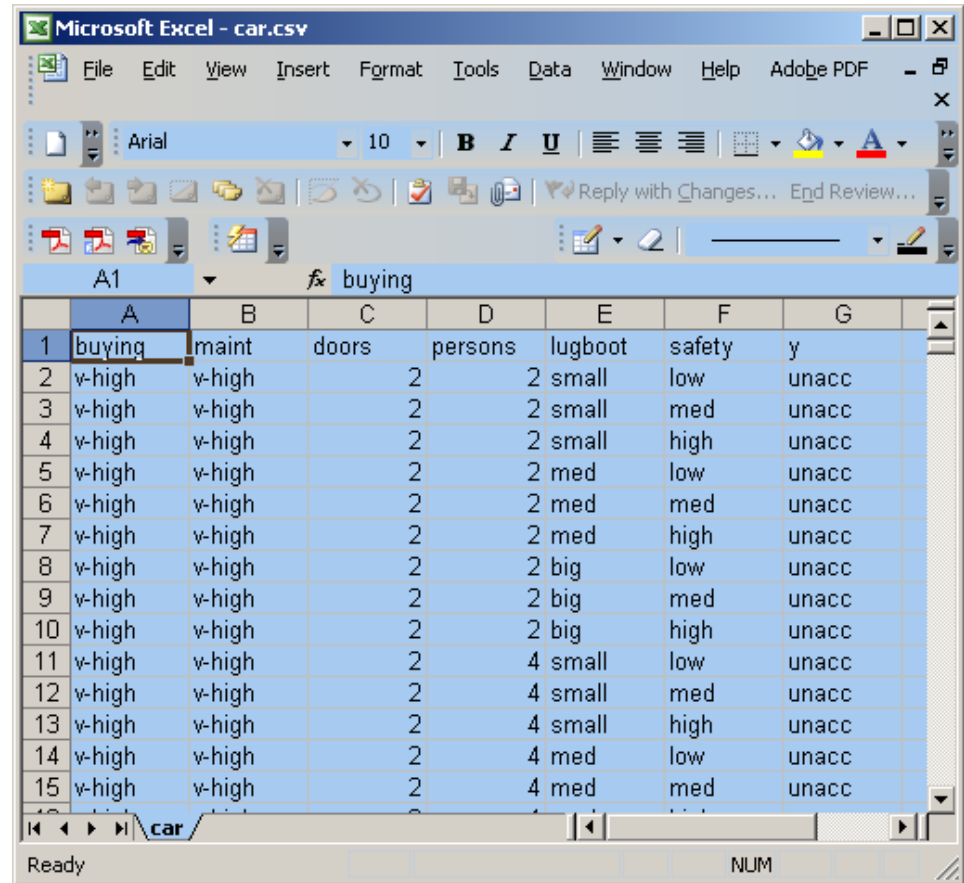
# Exercise 2: CAR dataset

- 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



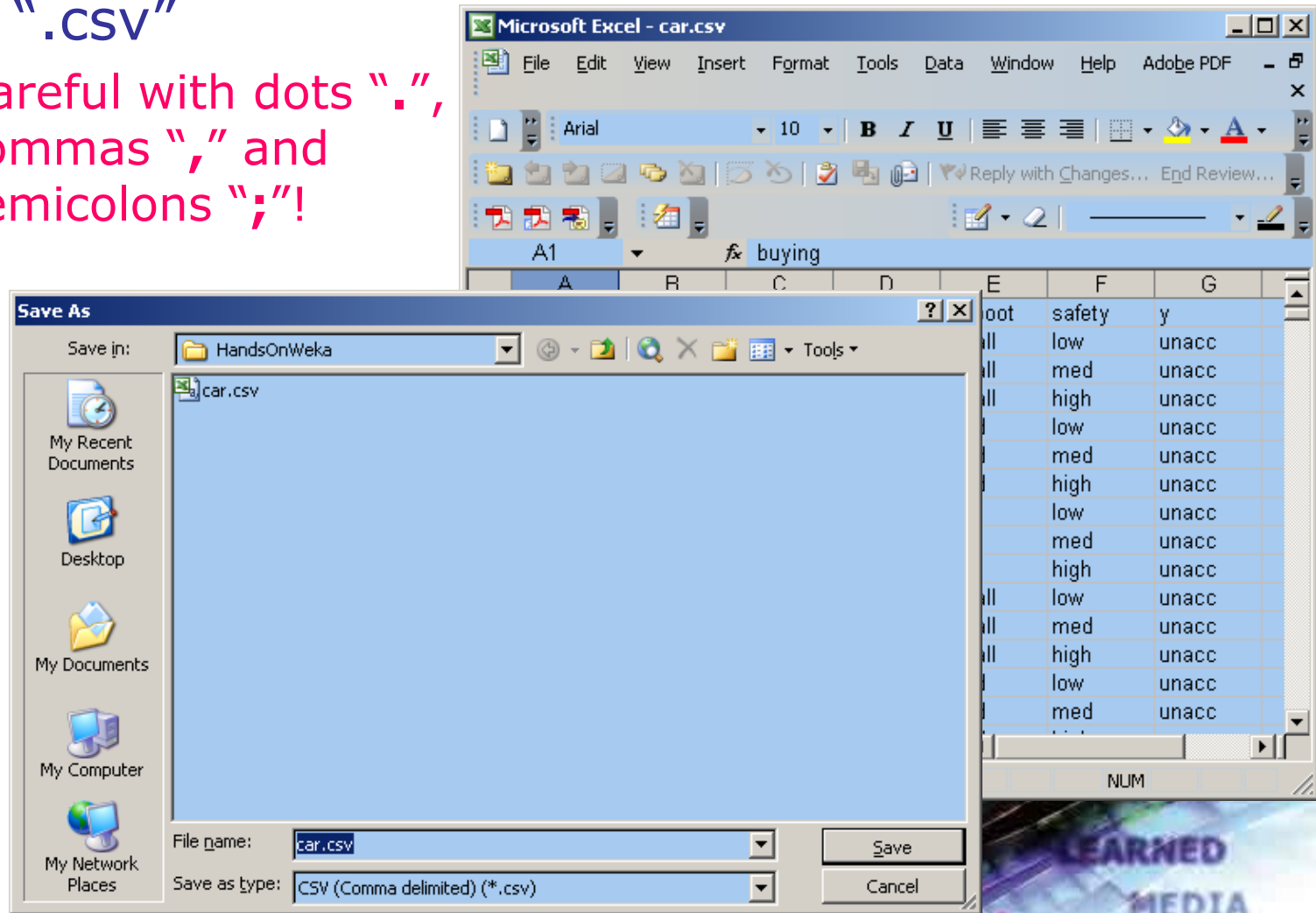
Microsoft Excel - car.csv

	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 '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 'Missing: 0 (0%)', 'Distinct: 4', and 'Type: Nominal'. A table below shows the distribution of the 'y' attribute: unacc (1210), acc (384), v-good (65), and good (69). A bar chart at the bottom right visualizes this distribution, with a blue arrow pointing to the 'y' attribute label and the text 'Target variable'.

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

Target variable

Status: OK Log x 0

# Choose algorithm J48

The screenshot shows the Weka Explorer application window. The 'Classifier' tab is active, displaying a tree view of available algorithms. The 'J48' algorithm is highlighted with a blue selection box. Three red arrows with circular numbers 1, 2, and 3 point to the 'Weka Explorer' title bar, the 'Classifier' tab, and the 'J48' algorithm respectively.

Weka Explorer

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

Classifier

- weka
  - classifiers
    - bayes
    - functions
    - lazy
    - meta
    - misc
    - trees
      - ADTree
      - DecisionStump
      - Id3
      - J48**
      - LMT
      - MSP
      - NBTree
      - RandomForest
      - RandomTree
      - REPTree
      - UserClassifier
  - rules

Status: OK

Log [Turtle icon] x 0

# Building and evaluating the tree

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

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
- 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

**Classification accuracy**

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

**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 %	
Total 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	
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 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' parameter is set to 15, which is highlighted by a red arrow from the text 'Set the minimal number of objects per leaf to 15'. Other parameters 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' table with columns for time, accuracy, and class.

Time	Accuracy	Class
14:55:00	92.3611 %	unacc
14:58:13	7.6389 %	acc
		v-good
		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
- Cross-validation Folds
- Percentage split %

(Nom) y

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 ===

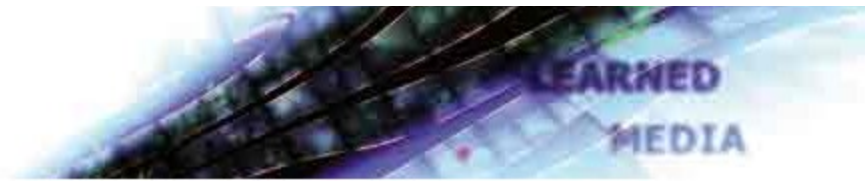
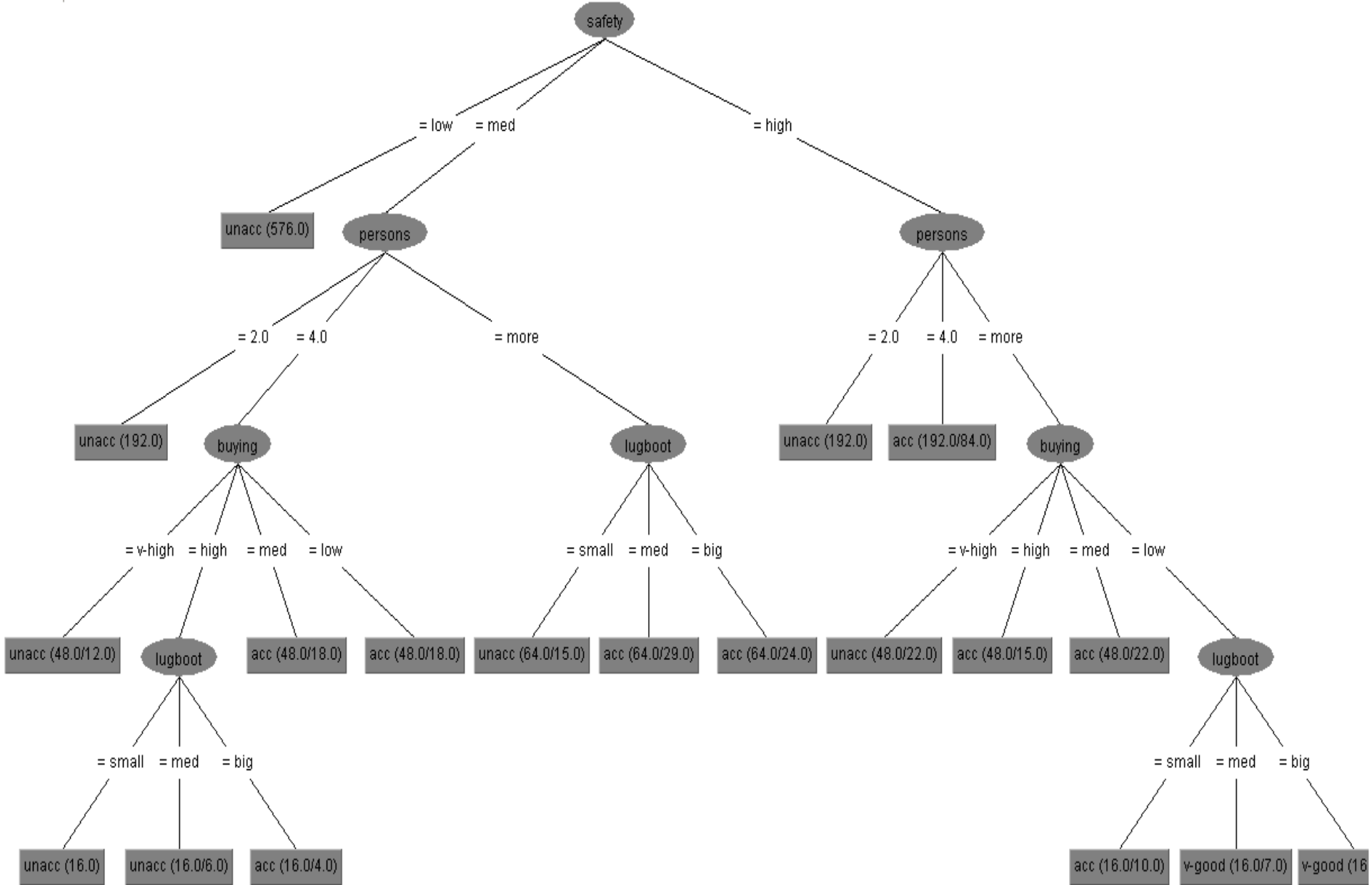
a	b	c	d	<-- classified as
1098	109	2	1	a = unacc
88	278	12	6	b = acc
0	44	21	0	c = v-good
0	62	7	0	d = good

Status: OK   x 0

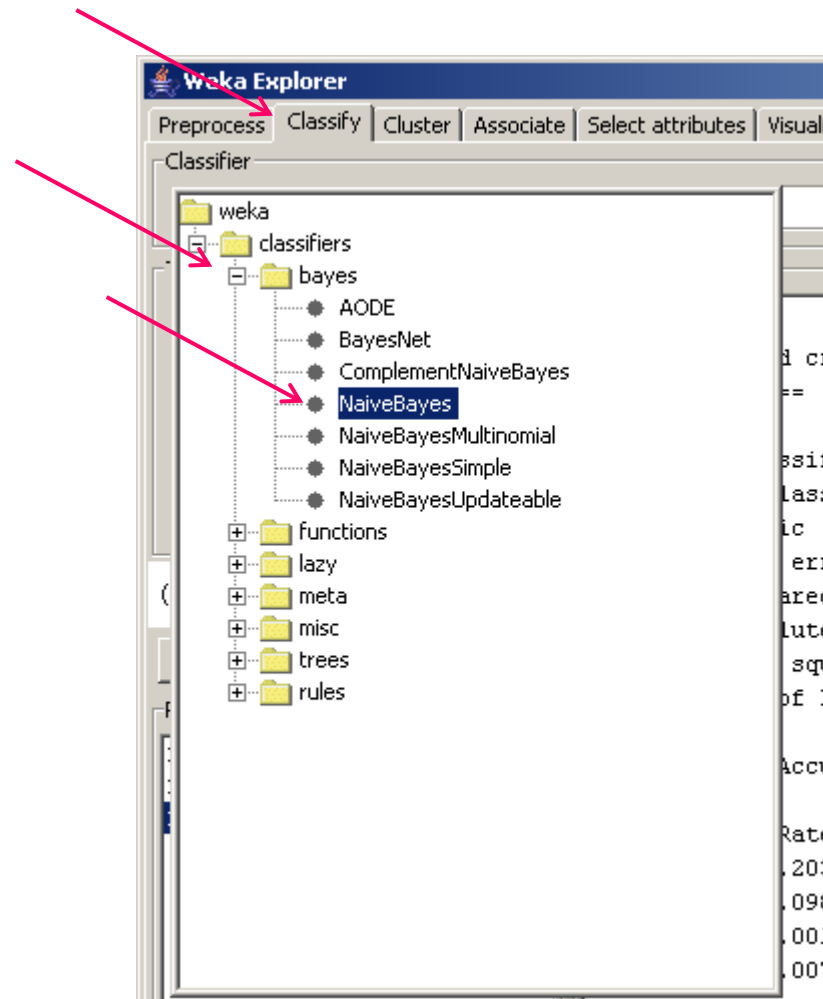
Reduced number of leaves and nodes

Easier to interpret

Lower classification accuracy



# 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

**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
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
  
```

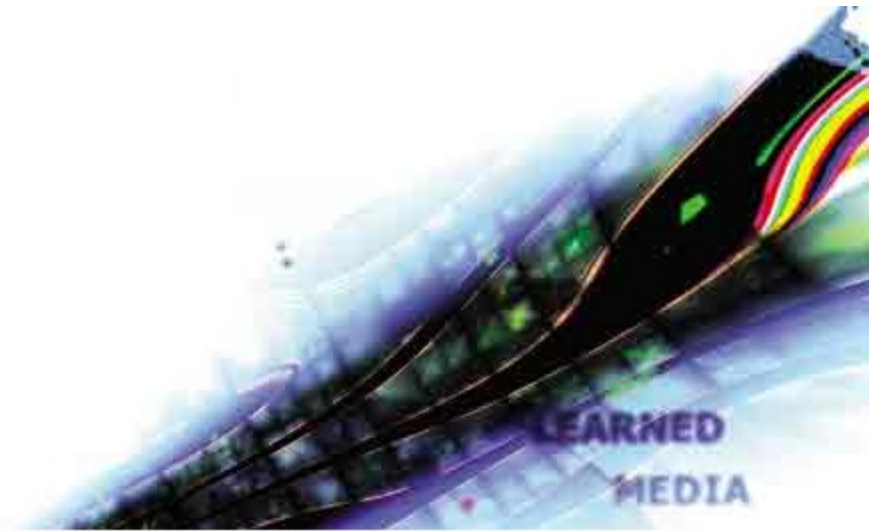
85.5324 %

# Numeric prediction in Weka



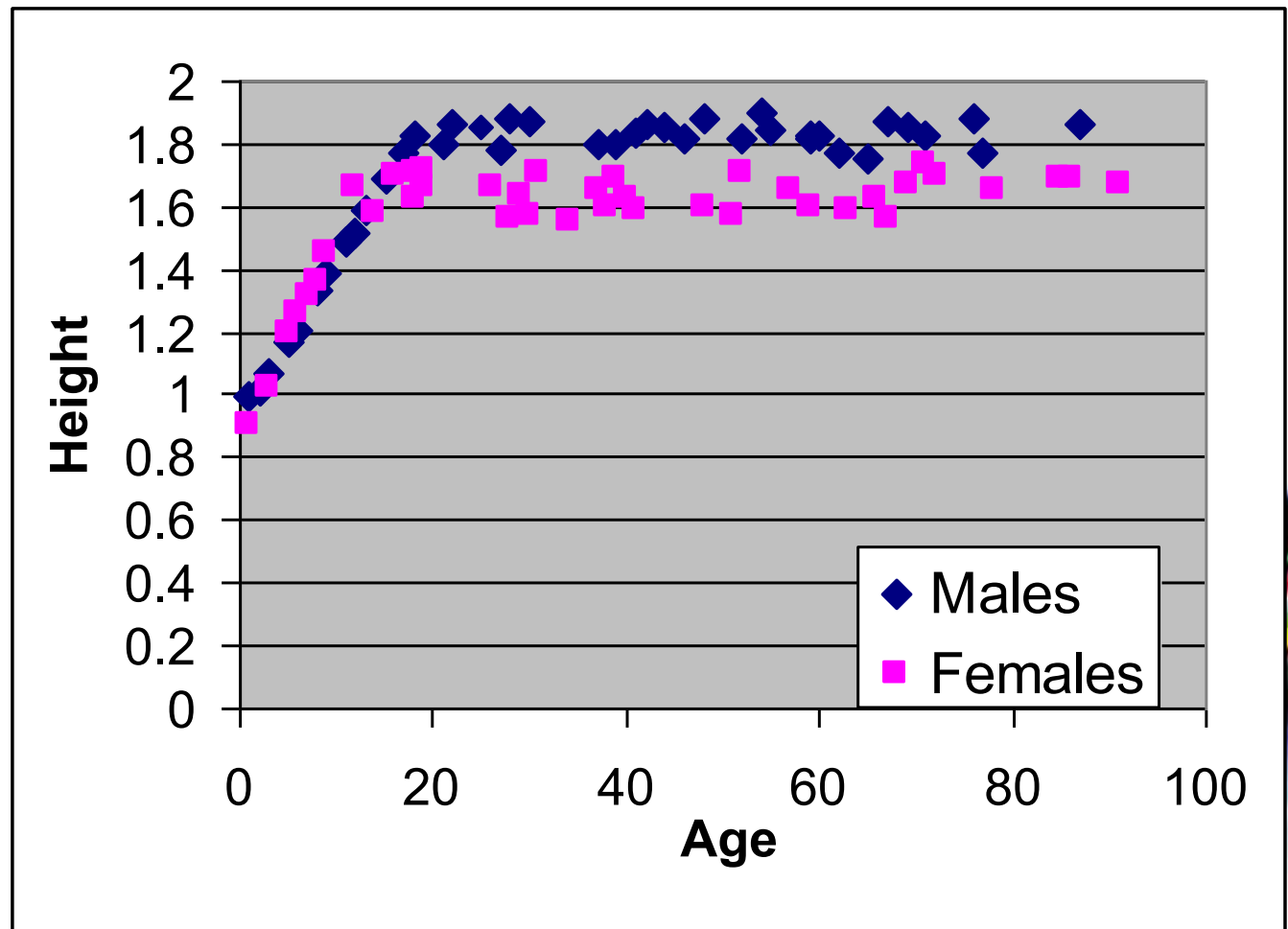
# Numeric prediction models

- LinearRegression
- M5P Regression and model trees
- KNN
- Baseline predictor



# regressionAgeHeight.csv

- Data about 80 people:
- Age
- Gender
- Height



# Filename:

The screenshot shows the Weka Explorer application window. The 'Classify' tab is selected. The 'Open file...' button is highlighted with a red arrow. The 'Current relation' is 'regressionAheHeight' with 80 instances and 3 attributes. The 'Attributes' list shows 'Age', 'Gender', and 'Height' (selected). The 'Selected attribute' section shows statistics for 'Height': Minimum (0.902), Maximum (1.895), Mean (1.628), and StdDev (0.236). A histogram for 'Class: Height (Num)' is displayed at the bottom right, showing the distribution of height values.

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: regressionAheHeight  
Instances: 80 Attributes: 3

Attributes: All | None | Invert

No.	Name
1	<input type="checkbox"/> Age
2	<input type="checkbox"/> Gender
3	<input checked="" type="checkbox"/> Height

Remove

Selected attribute: Name: Height  
Missing: 0 (0%) Distinct: 80 Type: Numeric  
Unique: 80 (100%)

Statistic	Value
Minimum	0.902
Maximum	1.895
Mean	1.628
StdDev	0.236

Class: Height (Num) Visualize All

Bin Range	Frequency
0.9 - 1.0	5
1.0 - 1.1	4
1.1 - 1.2	6
1.2 - 1.3	30
1.3 - 1.4	35

Status: OK Log x 0

# Visualization in Weka



Weka Explorer

Preprocess

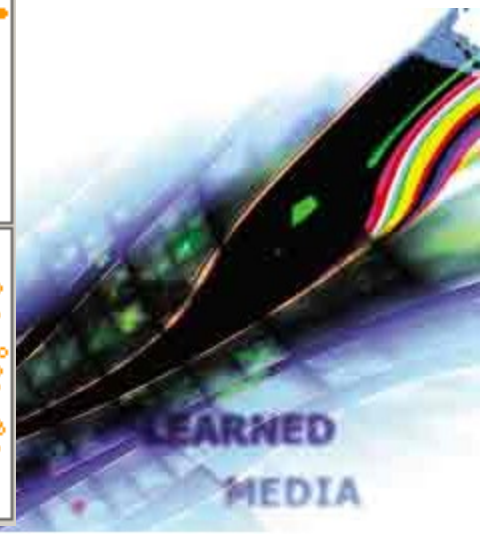
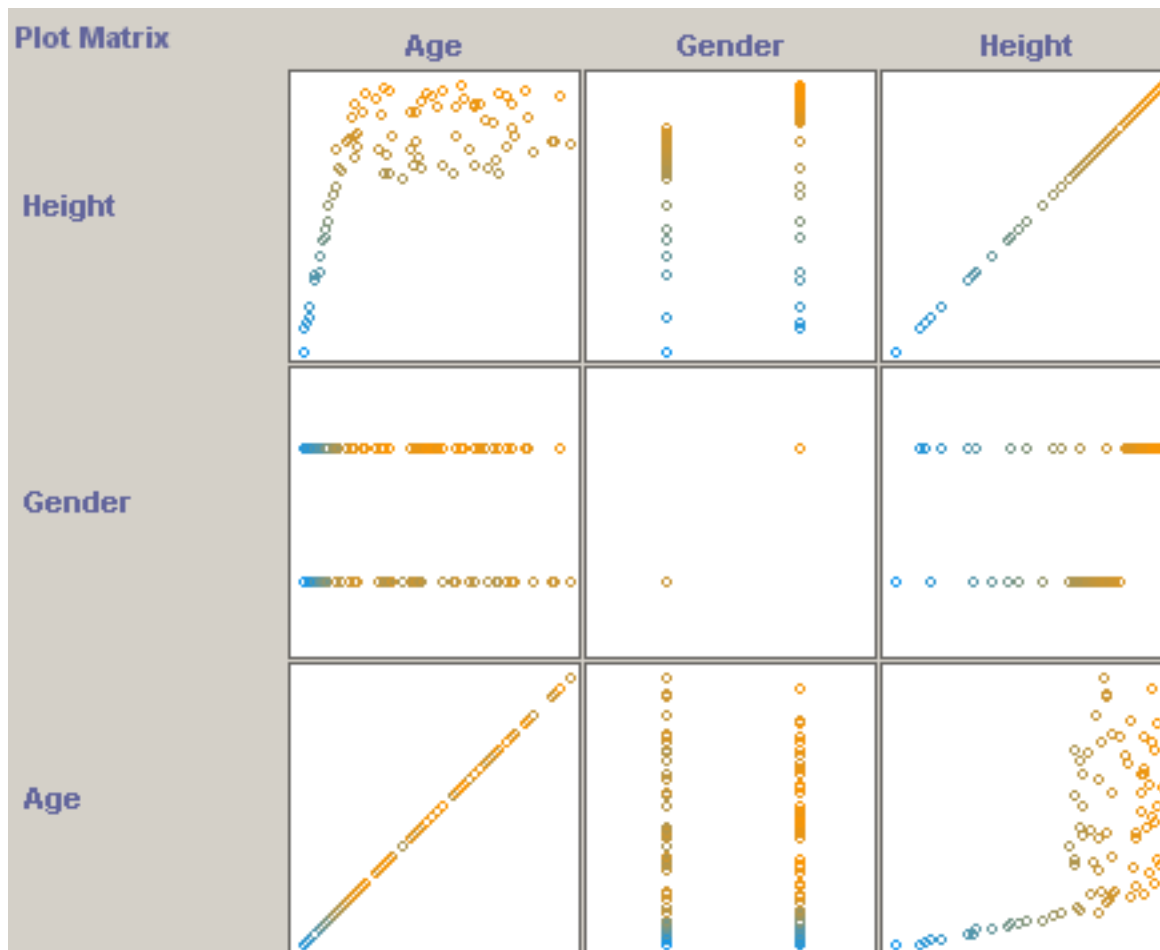
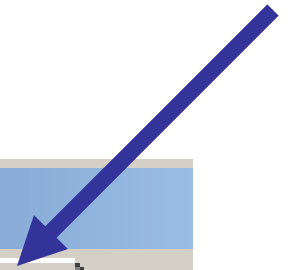
Classify

Cluster

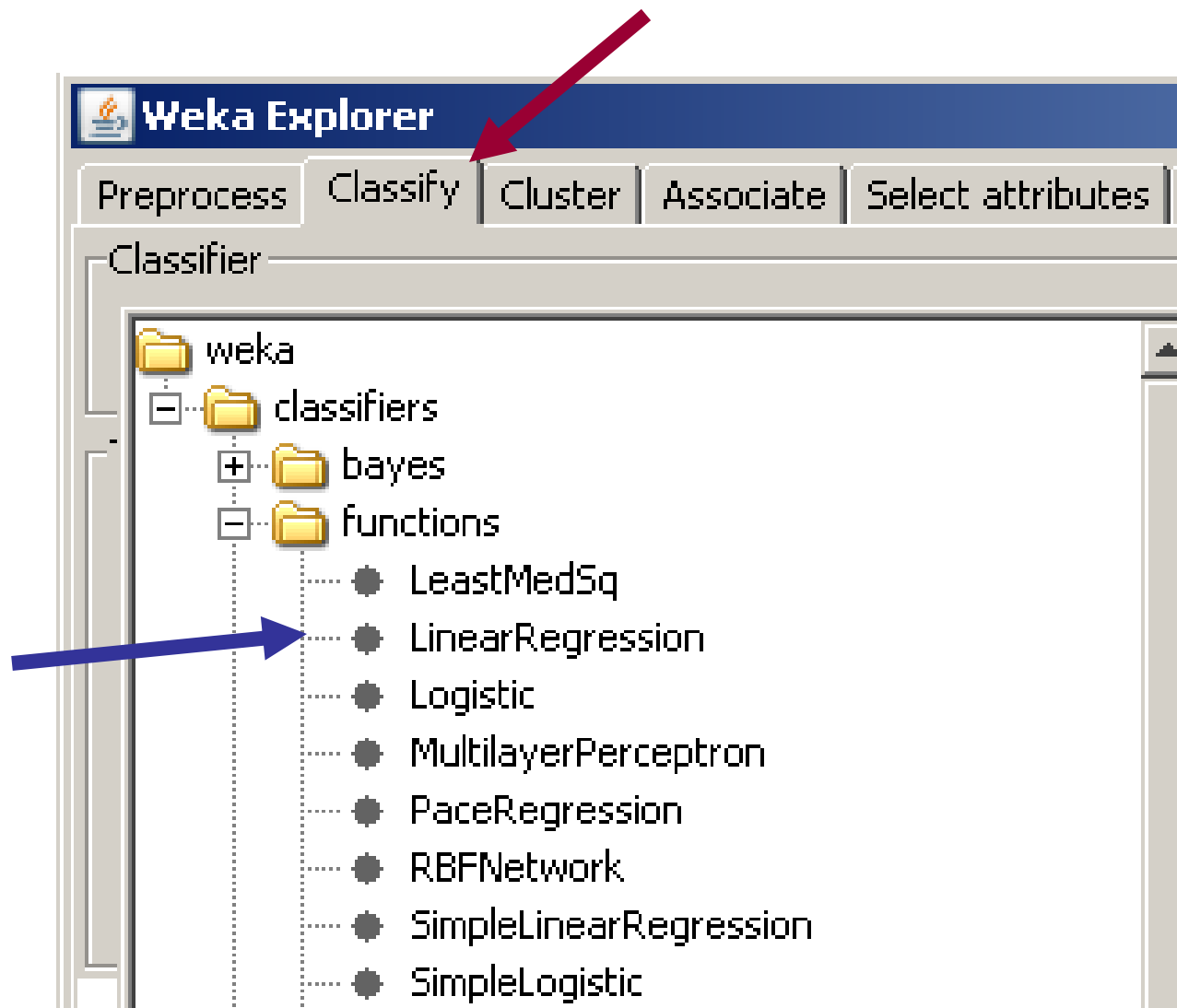
Associate

Select attributes

Visualize



# Weka → classifiers → functions → LinearRegression



**Weka Explorer**

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

Classifier: Choose **MSP -M 4.0**

Test options:

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

(Num) Height

Result list (right-click for options):

- 15:24:52 - trees.M5P
- 15:58:13 - functions.LinearRegression**
- 16:03:32 - functions.LinearRegression

Classifier output:

```

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

Linear Regression Model


Height =

      0.0056 * Age +
      0.1292 * Gender=M +
      1.3506

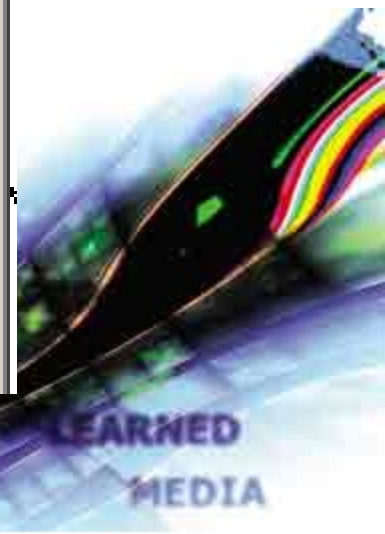
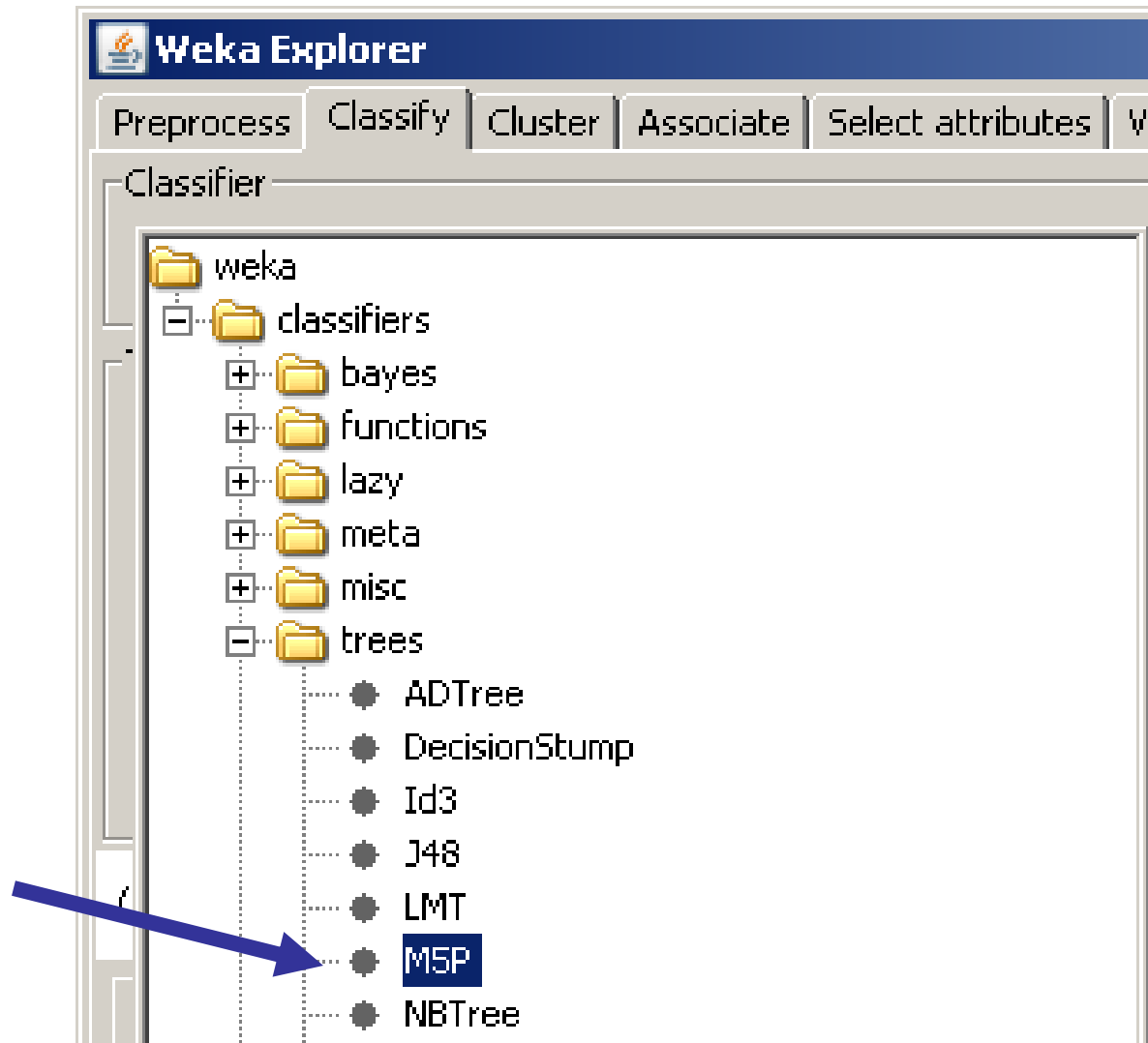
Time taken to build model: 0 seconds

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

Correlation coefficient           0.6204
Mean absolute error              0.142
Root mean squared error         0.1844
Relative absolute error         80.1623 %
Root relative squared error     77.2023 %
Total Number of Instances      80
  
```

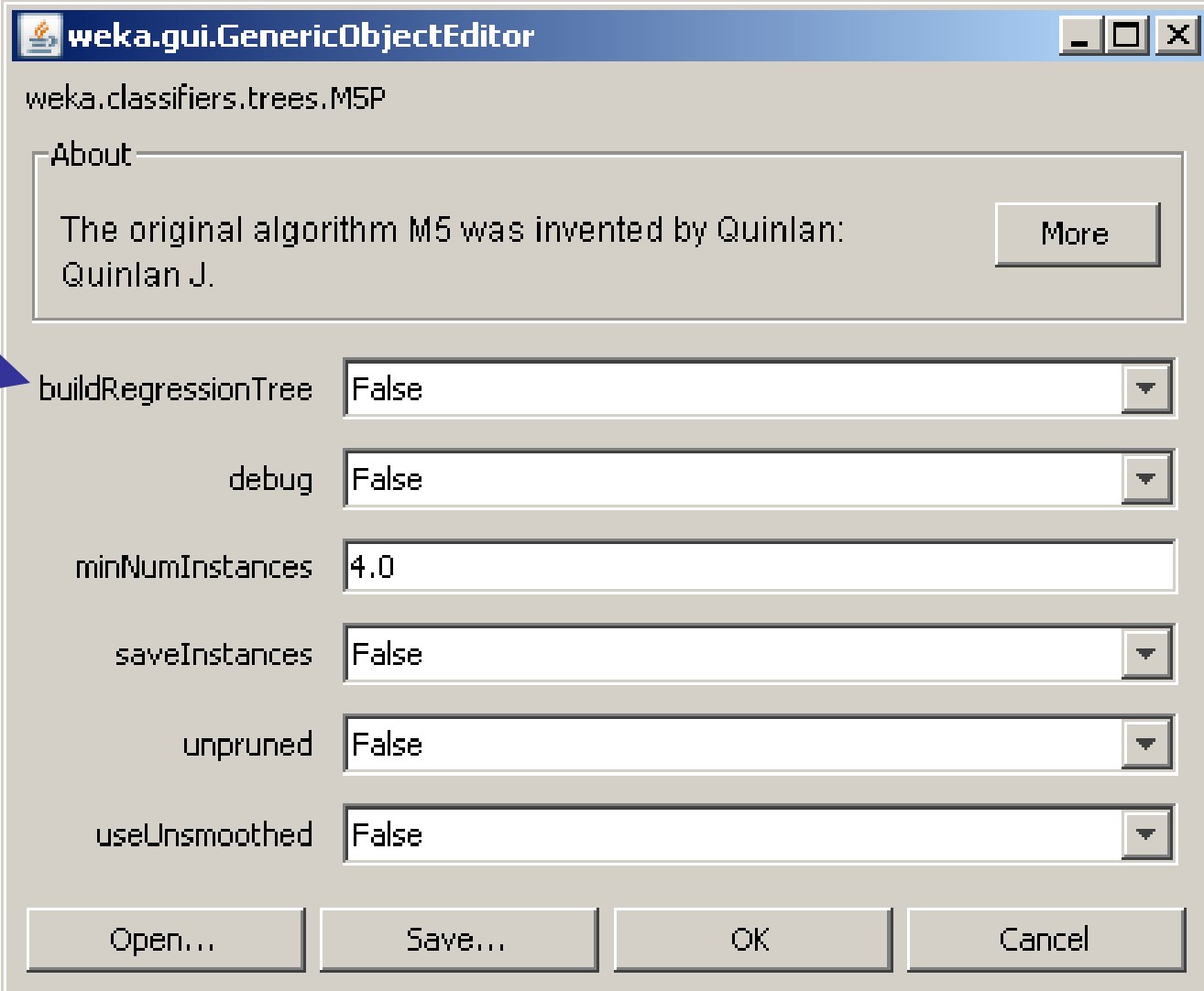
Status: OK   x 0

# Weka → classifiers → trees → M5P



LEARNED  
MEDIA

builtRegressionTree = True → regression tree  
builtRegressionTree = False → model tree



The screenshot shows the 'weka.gui.GenericObjectEditor' window for the 'weka.classifiers.trees.M5P' classifier. The window has a title bar with standard window controls. Below the title bar, the text 'weka.classifiers.trees.M5P' is displayed. There is an 'About' section with the text 'The original algorithm M5 was invented by Quinlan: Quinlan J.' and a 'More' button. Below this, there are several configuration options, each with a dropdown menu:

- builtRegressionTree: False
- debug: False
- minNumInstances: 4.0
- saveInstances: False
- unpruned: False
- useUnsmoothed: False

At the bottom of the window, there are four buttons: 'Open...', 'Save...', 'OK', and 'Cancel'. A blue arrow points to the 'builtRegressionTree' dropdown menu.



# kNN: weka → classifiers → lazy → IBk

The image shows two windows from the Weka software. The left window, 'Weka Explorer', displays a tree view of the classifier hierarchy. A blue arrow points to the 'IBk' classifier under the 'lazy' folder. The right window, 'weka.gui.GenericObjectEditor', shows the configuration for the selected classifier. A blue arrow points to the 'KNN' field, which is set to '3'. Other configuration options include 'crossValidate' (False), 'debug' (False), 'distanceWeighting' (No distance weighting), 'meanSquared' (False), 'noNormalization' (False), and 'windowSize' (0). The 'About' section describes it as a 'K-nearest neighbours classifier'.

**Weka Explorer**

- weka
  - classifiers
    - bayes
    - functions
    - lazy
      - IB1
      - IBk**
      - KStar
      - LBR
      - LWL
    - meta
    - misc
    - trees
    - rules

**weka.gui.GenericObjectEditor**

weka.classifiers.lazy.IBk

About: K-nearest neighbours classifier. [More](#)

KNN: 3

crossValidate: False

debug: False

distanceWeighting: No distance weighting

meanSquared: False

noNormalization: False

windowSize: 0

Buttons: Open..., Save..., OK, Cancel

# Weka → classifiers → rules → ZeroR

