



Data Mining with WEKA



WEKA ?

- Waikato Environment for Knowledge Analysis
- A Collection of Machine Learning algorithms for data tasks.
- WEKA contains tools for data – pre-processing, classification, regression, clustering association rules.



Start with WEKA

1) Get the WEKA program on the web

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

2) set the CLASSPATH

system environment variables;

variable name: CLASSPATH

variable value: (e.g C:\Program Files\Weka-3-4)

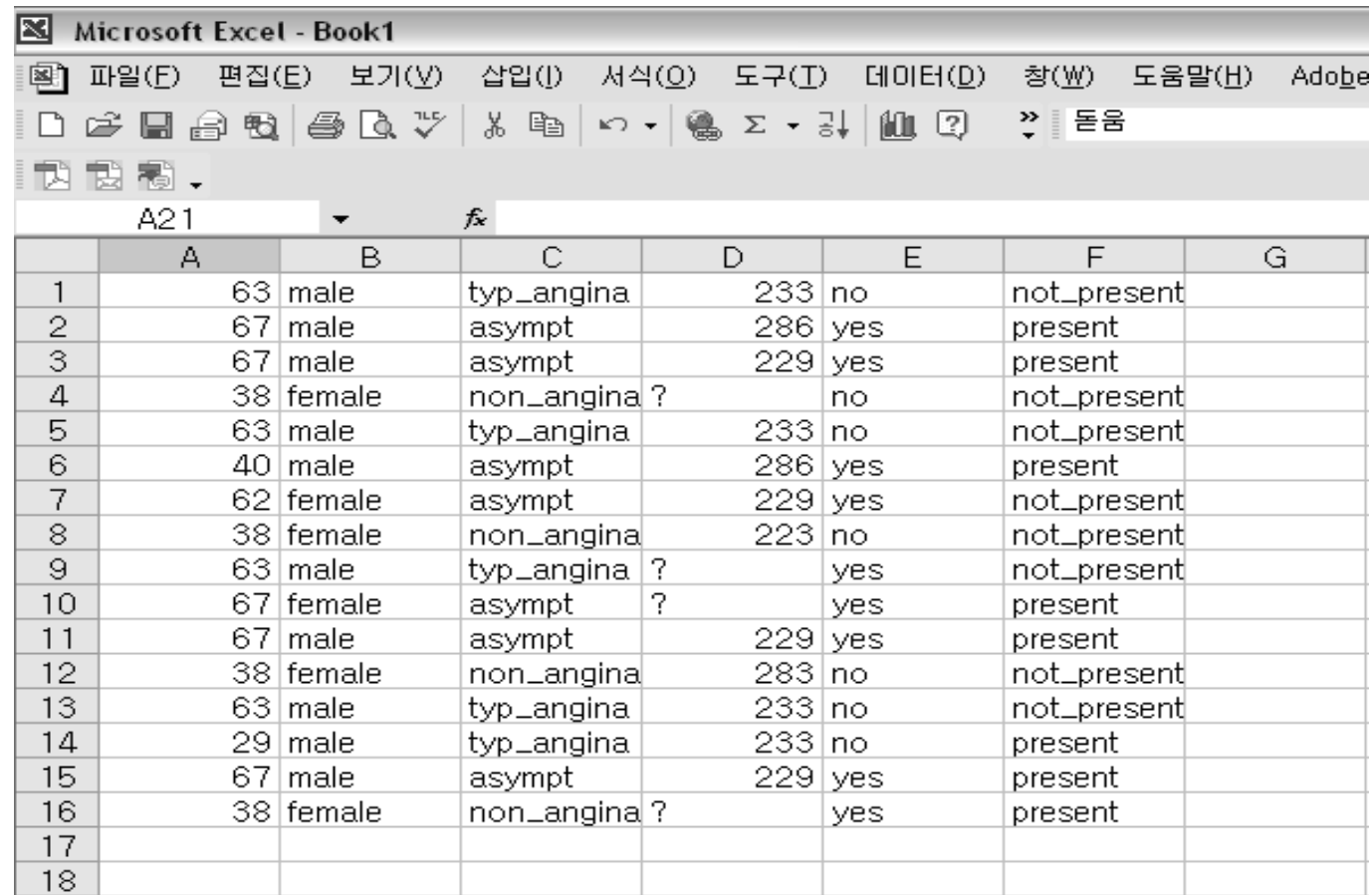


Prepare the Data Set

Need to convert ARFF format

1. Data can be load to excel spreadsheet
2. Save this data in comma-separated format (CSV)
3. Load this file into Micro Word
4. Make beginning of the ARFF file.
 - @ relation (title)
 - @ attribute (data type)
 - @ data

Load into Excel



Microsoft Excel - Book1

파일(F) 편집(E) 보기(V) 삽입(I) 서식(O) 도구(T) 데이터(D) 창(W) 도움말(H) Adobe

☰ ☱ ☲ ☳ ☴ ☵ ☶ ☷

A21 fx

	A	B	C	D	E	F	G
1	63	male	typ_angina	233	no	not_present	
2	67	male	asympt	286	yes	present	
3	67	male	asympt	229	yes	present	
4	38	female	non_angina ?		no	not_present	
5	63	male	typ_angina	233	no	not_present	
6	40	male	asympt	286	yes	present	
7	62	female	asympt	229	yes	not_present	
8	38	female	non_angina	223	no	not_present	
9	63	male	typ_angina ?		yes	not_present	
10	67	female	asympt ?		yes	present	
11	67	male	asympt	229	yes	present	
12	38	female	non_angina	283	no	not_present	
13	63	male	typ_angina	233	no	not_present	
14	29	male	typ_angina	233	no	present	
15	67	male	asympt	229	yes	present	
16	38	female	non_angina ?		yes	present	
17							
18							

Save as the CSV file format

	A	B	C	D	E	F	G	H	I
1	63	male	typ_angina	233	no	not_present			
2	67	male	asympt	286	yes	present			
3	67	male	asympt	229	yes	present			
4	38	female	non_angina ?		no	not_present			
5	63	male	typ_angina	233	no	not_present			
6	40	male	asympt	286	yes	present			
7	62	female	asympt	229	yes	not_present			
8	38	female							
9	63	male							
10	67	female							
11	67	male							
12	38	female							
13	63	male							
14	29	male							
15	67	male							
16	38	female							
17									
18									
19									
20									
21									
22									
23									
24									
25									
26									
27									
28									
29									
30									
31									

다름 이름으로 저장

저장 위치(I): My Documents

- FIFA 2003
- My eBooks
- My Music
- My Pictures
- My Received Files
- my way
- My Webs

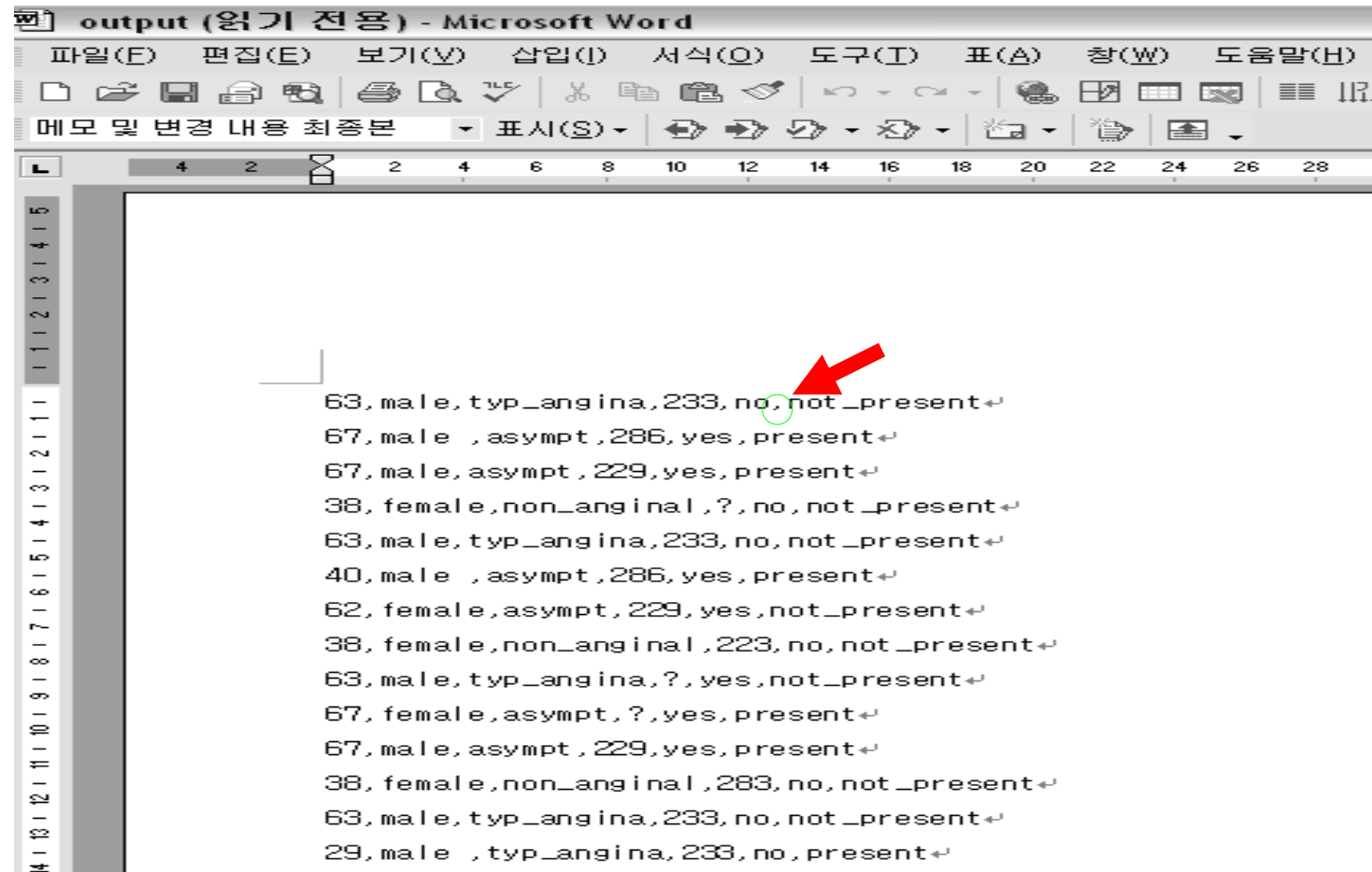
파일 이름(N): Book1

파일 형식(T): CSV (실표로 분리)

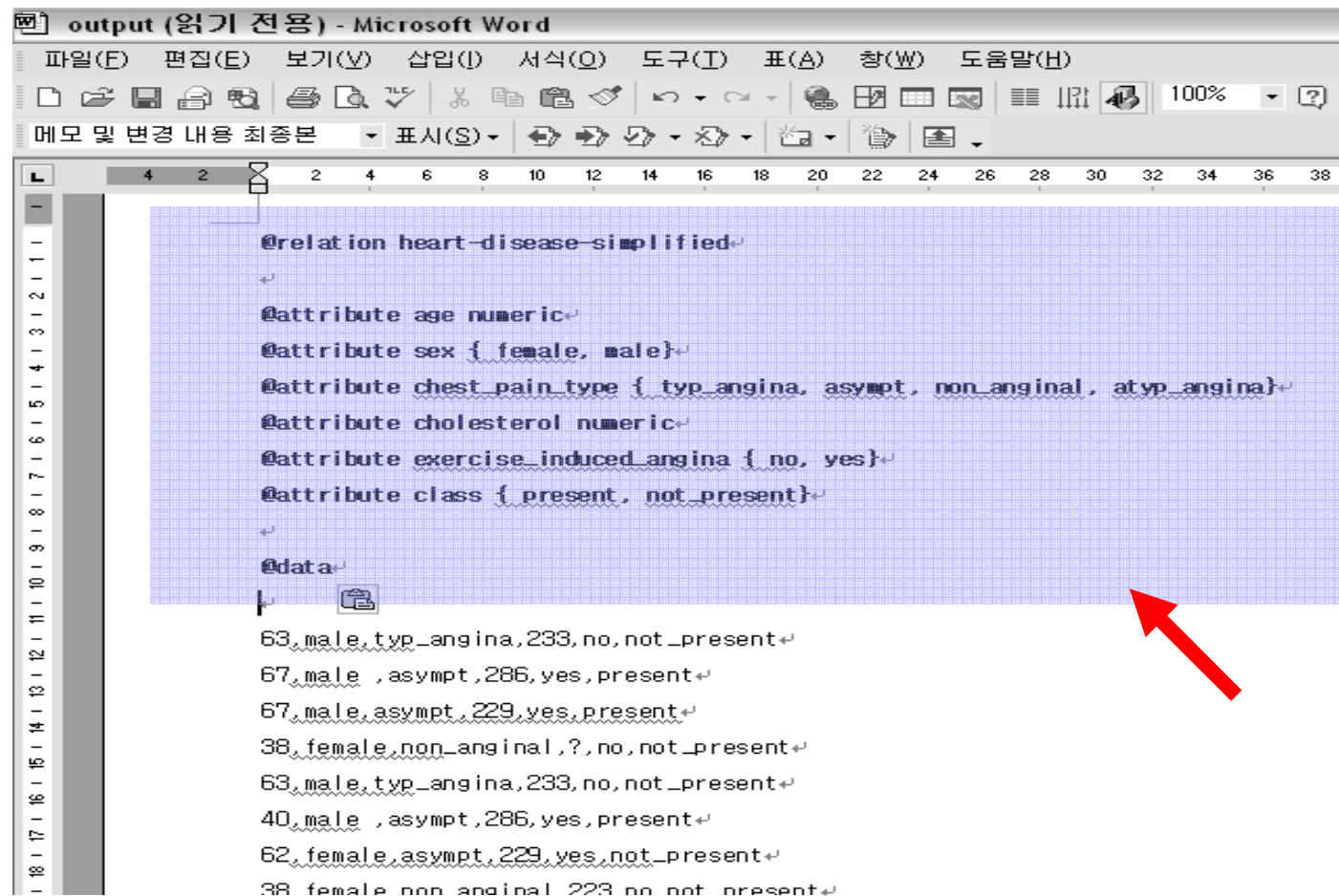
- 유니코드 텍스트
- Microsoft Excel 5.0/95 통합 문서
- Microsoft Excel 97-2002 & 5.0/95 통합 문서
- CSV (실표로 분리)
- Microsoft Excel 4.0 워크시트
- Microsoft Excel 3.0 워크시트

저장(S) 취소

Load into MS word



Make other parts..



output (읽기 전용) - Microsoft Word

파일(F) 편집(E) 보기(V) 삽입(I) 서식(O) 도구(T) 표(A) 창(W) 도움말(H)

메모 및 변경 내용 최종본 표시(S)

```
@relation heart-disease-simplified↵
↵
@attribute age numeric↵
@attribute sex { female, male}↵
@attribute chest_pain_type { typ_angina, asympt, non_anginal, atyp_angina}↵
@attribute cholesterol numeric↵
@attribute exercise_induced_angina { no, yes}↵
@attribute class { present, not_present}↵
↵
@data↵
63,male,typ_angina,233,no,not_present↵
67,male , asympt,286,yes,present↵
67,male,asympt,229,yes,present↵
38,female,non_anginal,?,no,not_present↵
63,male,typ_angina,233,no,not_present↵
40,male , asympt,286,yes,present↵
62,female,asympt,229,yes,not_present↵
38 female non anginal 223 no not present↵
```



WEKA only deals with ARFF files

@relation heart-disease-simplified

@attribute age numeric

@attribute sex { female, male }

@attribute chest_pain_type { typ_angina, asympt, non_anginal, atyp_angina }

@attribute cholesterol numeric

@attribute exercise_induced_angina { no, yes }

@attribute class { present, not_present }

@data

63,male,typ_angina,233,no,not_present

67,male,asympt,286,yes,present

67,male,asympt,229,yes,present

38,female,non_anginal,?,no,not_present



Weka GUI Chooser

Waikato Environment for Knowledge Analysis

(c) 1999 - 2003
University of Waikato
New Zealand



GUI

Simple CLI Explorer

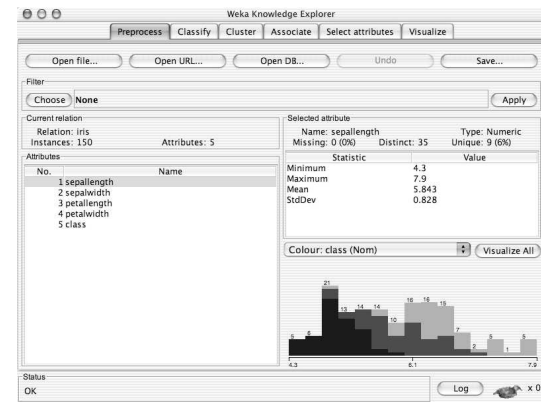
Experimenter KnowledgeFlow

```
Welcome to the WEKA SimpleCLI

Enter commands in the textfield at the bottom of
the window. Use the up and down arrows to move
through previous commands.

> help

Command must be one of:
  java <classname> <args>
  break
  kill
  cls
  exit
  help <command>
```



Weka Experiment Environment

Setup | Run | Analyse

Experiment Configuration Mode: Simple Advanced

Open... Save... New

Results Destination: JDBC database URL: jdbc:db=experiments.prp Browse...

Experiment Type: Cross-validation
Number of folds: 10
Classification Regression

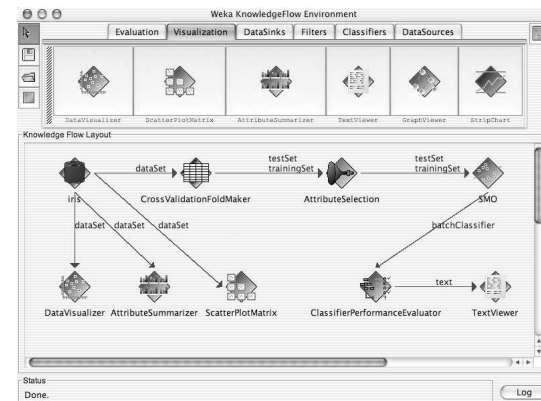
Iteration Control: Number of repetitions: 10
Data sets first Algorithms first

Datasets: Add new... Delete selected
Use relative paths

Algorithms: Add new... Delete selected

J48 -C 0.25 -M 2
NeuralNetwork -L 0.3 -M 0.2 -N 500 -V 0 -E 20 -H a
NaiveBayes

Notes





Preprocessing the data

- **Integration from different sources**
- **The Data must be assembled, integrated, and cleaned up**
- **Pre-processing tools in WEKA are called “filters”**
- **WEKA contains filters for:**
 - **Discretization, normalization, resampling, attribute selection, transforming and combining attributes, ...**

With numeric data (Iris.arff)

The screenshot shows the Weka Explorer interface with the Iris dataset loaded. The 'petallength' attribute is selected, and a histogram is displayed. A red arrow points to the 'Choose' button in the Filter section.

Weka Explorer

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

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

Filter: Choose **None** [Apply]

Current relation: iris
Relation: iris
Instances: 150
Attributes: 5

Attributes:

No.	Name
1	sepalwidth
2	sepalwidth
3	petallength
4	petalwidth
5	class

Selected attribute:

Name: petallength
Missing: 0 (0%)
Distinct: 43
Type: Numeric
Unique: 10 (7%)

Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

Colour: class (Nom) [Visualize All]

Histogram showing frequency distribution of petallength values. The x-axis ranges from 1 to 6.9, and the y-axis shows frequencies from 0 to 37.

Status: OK [Log] x 0

Select Discretize filter

The screenshot shows the Weka Explorer interface. The 'Filter' tab is active, and the 'Discretize' filter is selected in the 'Filter' list. A red arrow points to the 'Discretize' filter. Another red arrow points to the 'Apply' button in the top right corner of the filter panel.

Filter List:

- weka
 - filters
 - unsupervised
 - attribute
 - Add
 - AddCluster
 - AddExpression
 - AddNoise
 - ClusterMembership
 - Copy
 - Discretize**
 - FirstOrder
 - MakeIndicator
 - MergeTwoValues
 - NominalToBinary
 - Normalize
 - NumericToBinary
 - NumericTransform
 - Obfuscate
 - PKIDiscretize
 - RandomProjection
 - Remove
 - RemoveType
 - RemoveUseless
 - ReplaceMissingValues

Selected attribute:

Name: petalength
Missing: 0 (0%)
Distinct: 43
Type: Numeric
Unique: 10 (7%)

Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

Colour: class (Nom) [v] Visualize All

A histogram showing the distribution of petalength values. The x-axis ranges from 1 to 6.9, and the y-axis shows the frequency of values. The distribution is unimodal and slightly right-skewed.

Value	Frequency
1	11
2	37
3	2
4	0
5	0
6	3
7	4
8	12
9	18
10	17
11	16
12	14
13	10
14	2
15	4

Status: OK Log x 0

Changed to nominal data

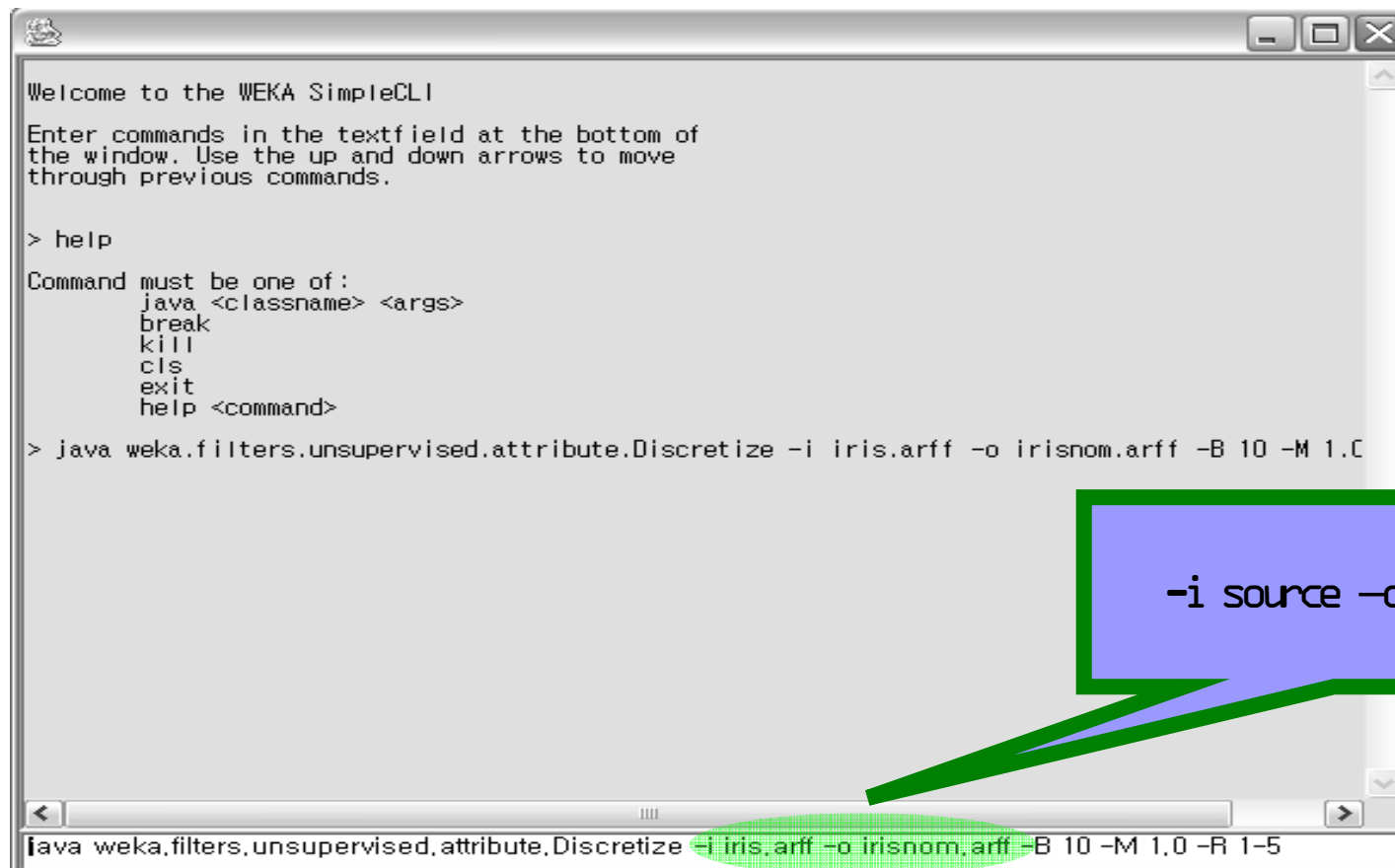
The screenshot shows the Weka Explorer interface with the 'Discretize' filter applied to the 'petal length' attribute. The filter parameters are '-B 10 -M -1.0 -R first-last'. The 'Selected attribute' section shows 'petal length' with 9 distinct values and 0 missing values. A bar chart visualizes the distribution of these values, with a red arrow pointing to the bar for the value '(3.36-3.95]'.

Label	Count
'(-inf-1.59]'	37
'(1.59-2.18]'	13
'(2.18-2.77]'	0
'(2.77-3.36]'	3
'(3.36-3.95]'	8
'(3.95-4.54]'	26
'(4.54-5.13]'	29

Colour: class (Nom) Visualize All

Status: OK Log x 0

Filtering using CLI (Iris.data)



Welcome to the WEKA SimpleCLI

Enter commands in the textfield at the bottom of the window. Use the up and down arrows to move through previous commands.

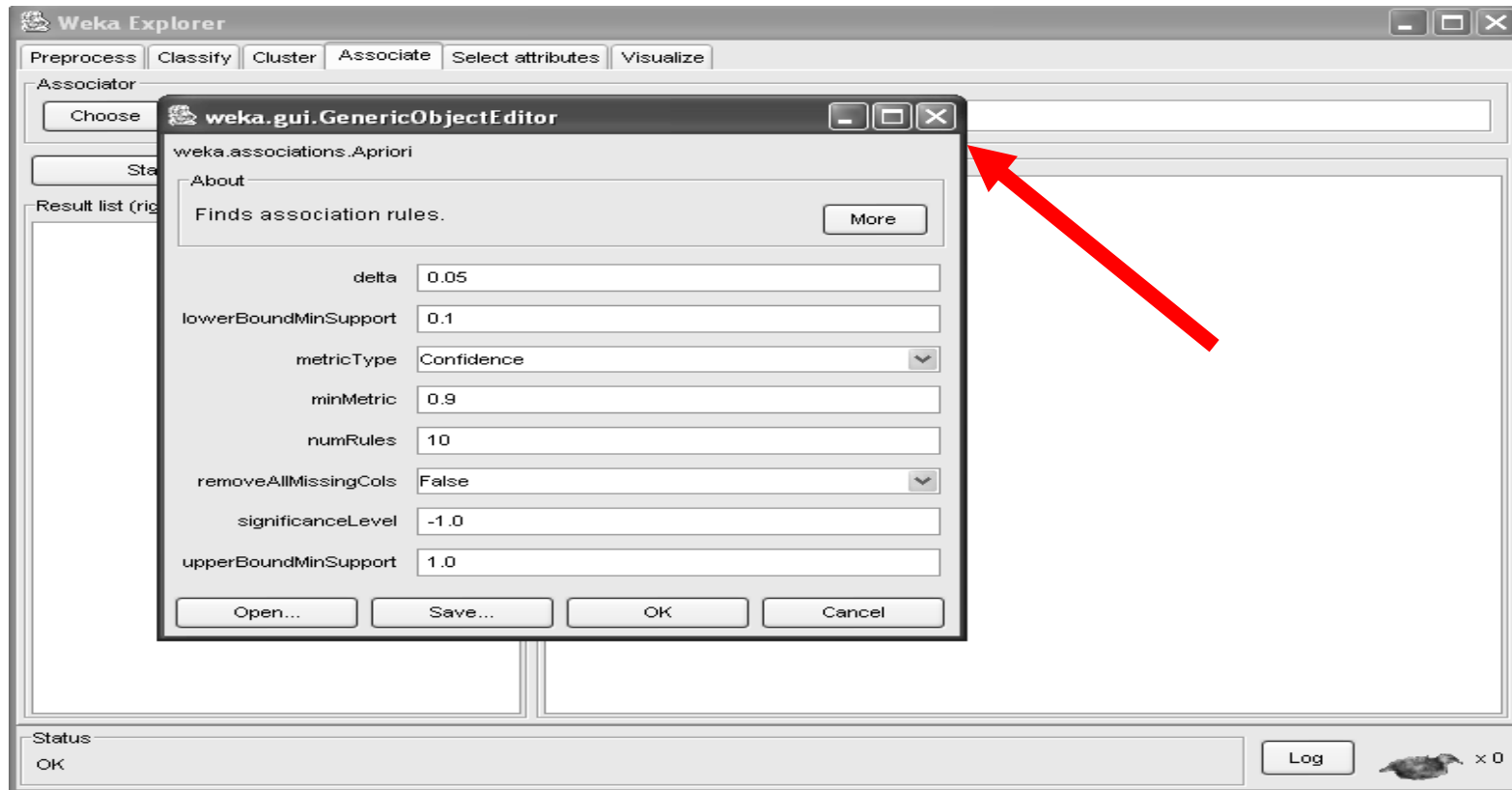
```
> help
Command must be one of:
  java <classname> <args>
  break
  kill
  cls
  exit
  help <command>

> java weka.filters.unsupervised.attribute.Discretize -i iris.arff -o irisnom.arff -B 10 -M 1.0
```

-i source -o object file

```
java weka.filters.unsupervised.attribute.Discretize -i iris.arff -o irisnom.arff -B 10 -M 1.0 -R 1-5
```

Association (weather.nominal.arff)



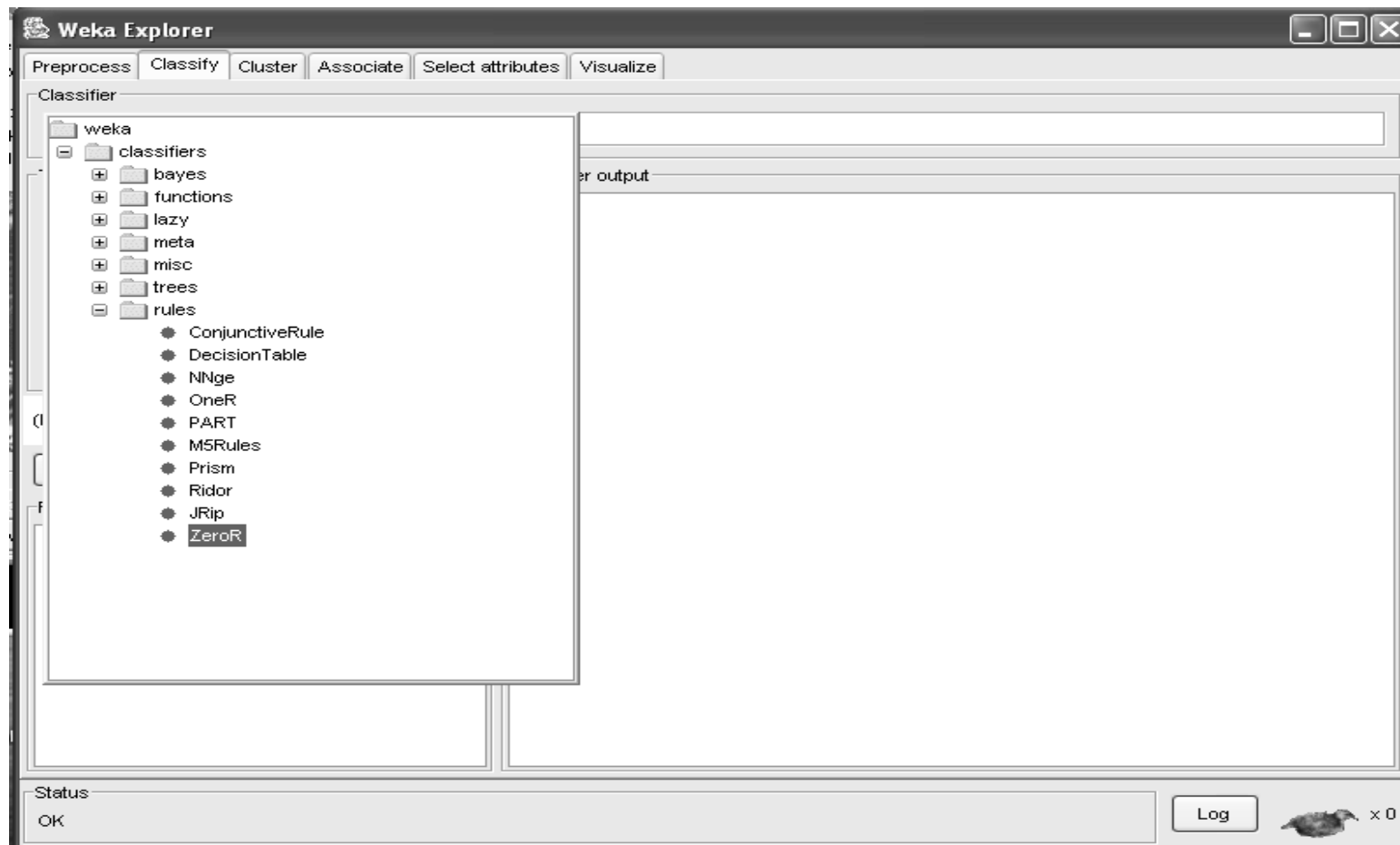


Association -result

- **Best rules found:**

1. **humidity=normal windy=FALSE 4 ==> play=yes 4 conf:(1)**
2. **temperature=cool 4 ==> humidity=normal 4 conf:(1)**
3. **outlook=overcast 4 ==> play=yes 4 conf:(1)**
4. **humidity=normal 7 ==> play=yes 6 conf:(0.86)**
5. **play=no 5 ==> humidity=high 4 conf:(0.8)**
6. **windy=FALSE 8 ==> play=yes 6 conf:(0.75)**
7. **play=yes 9 ==> windy=FALSE 6 conf:(0.67)**
8. **play=yes 9 ==> humidity=normal 6 conf:(0.67)**
9. **humidity=normal play=yes 6 ==> windy=FALSE 4 conf:(0.67)**
10. **windy=FALSE play=yes 6 ==> humidity=normal 4 conf:(0.67)**

Classification – voting records



Classification - zeroR

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose ZeroR

Test options

Use training set

Supplied test set Set...

Cross-validation Folds 10

Percentage split % 66

More options...

(Nom) Class_Name

Start Stop

Result list (right-click for options)

16:29:29 - rules.ZeroR

16:29:39 - rules.ZeroR

Classifier output

Correctly Classified Instances 267 61.3793 %

Incorrectly Classified Instances 168 38.6207 %

Kappa statistic 0

Mean absolute error 0.4742

Root mean squared error 0.4869

Relative absolute error 100 %

Root relative squared error 100 %

Total Number of Instances 435

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
1	1	0.614	1	0.761	democrat
0	0	0	0	0	republican

=== Confusion Matrix ===

```
a b <-- classified as
267 0 | a = democrat
168 0 | b = republican
```

Status

OK Log x 0

Classification -oneR

The screenshot shows the Weka Explorer interface with the OneR classifier selected. The 'Classifier output' pane displays the following summary statistics:

Correctly Classified Instances	416	95.6322 %
Incorrectly Classified Instances	19	4.3678 %
Kappa statistic	0.9088	
Mean absolute error	0.0437	
Root mean squared error	0.209	
Relative absolute error	9.21 %	
Root relative squared error	42.9237 %	
Total Number of Instances	435	

Below the summary, the 'Detailed Accuracy By Class' is shown:

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.948	0.03	0.981	0.948	0.964	democrat
0.97	0.052	0.921	0.97	0.945	republican

The 'Confusion Matrix' is also displayed:

```
=== Confusion Matrix ===
  a  b  <-- classified as
253 14 | a = democrat
  5 163 | b = republican
```

Two red arrows in the image point to the overall accuracy (95.6322%) and the confusion matrix.

Classification –J48

The screenshot shows the Weka Explorer interface with the J48 classifier selected. The classifier output window displays the following performance metrics:

Metric	Value	Percentage
Correctly Classified Instances	419	96.3218 %
Incorrectly Classified Instances	16	3.6782 %
Kappa statistic	0.9224	
Mean absolute error	0.0611	
Root mean squared error	0.1748	
Relative absolute error	12.887 %	
Root relative squared error	35.9085 %	
Total Number of Instances	435	

Below the main metrics, the 'Detailed Accuracy By Class' table is shown:

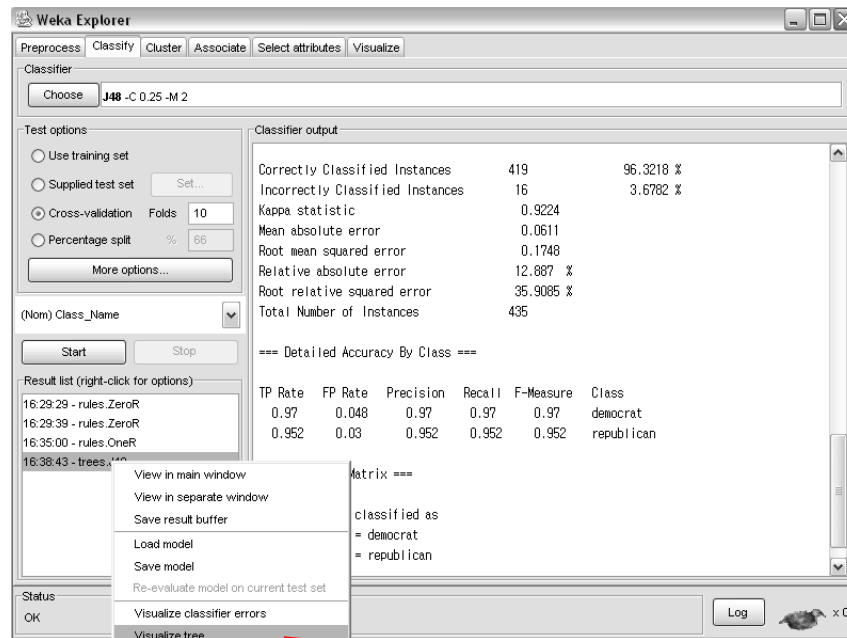
TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.97	0.048	0.97	0.97	0.97	democrat
0.952	0.03	0.952	0.952	0.952	republican

The 'Confusion Matrix' section shows the following data:

```
=== Confusion Matrix ===
 a  b  <-- classified as
259  8 | a = democrat
 8 160 | b = republican
```

The status bar at the bottom left shows 'Status OK'.

Decision Tree from J48 result



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...

Classifier output:

Correctly Classified Instances	419	96.3218 %
Incorrectly Classified Instances	16	3.6782 %
Kappa statistic	0.9224	
Mean absolute error	0.0611	
Root mean squared error	0.1748	
Relative absolute error	12.887 %	
Root relative squared error	35.9065 %	
Total Number of Instances	435	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.97	0.048	0.97	0.97	0.97	democrat
0.952	0.03	0.952	0.952	0.952	republican

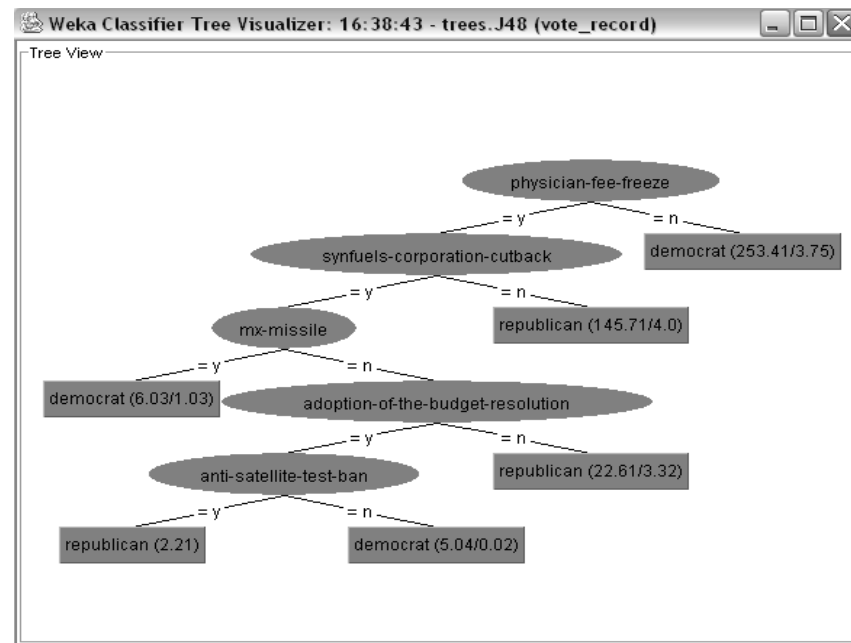
Result list (right-click for options):

- 16:29:29 - rules.ZeroR
- 16:29:39 - rules.ZeroR
- 16:35:00 - rules.OneR
- 16:38:43 - trees.J48

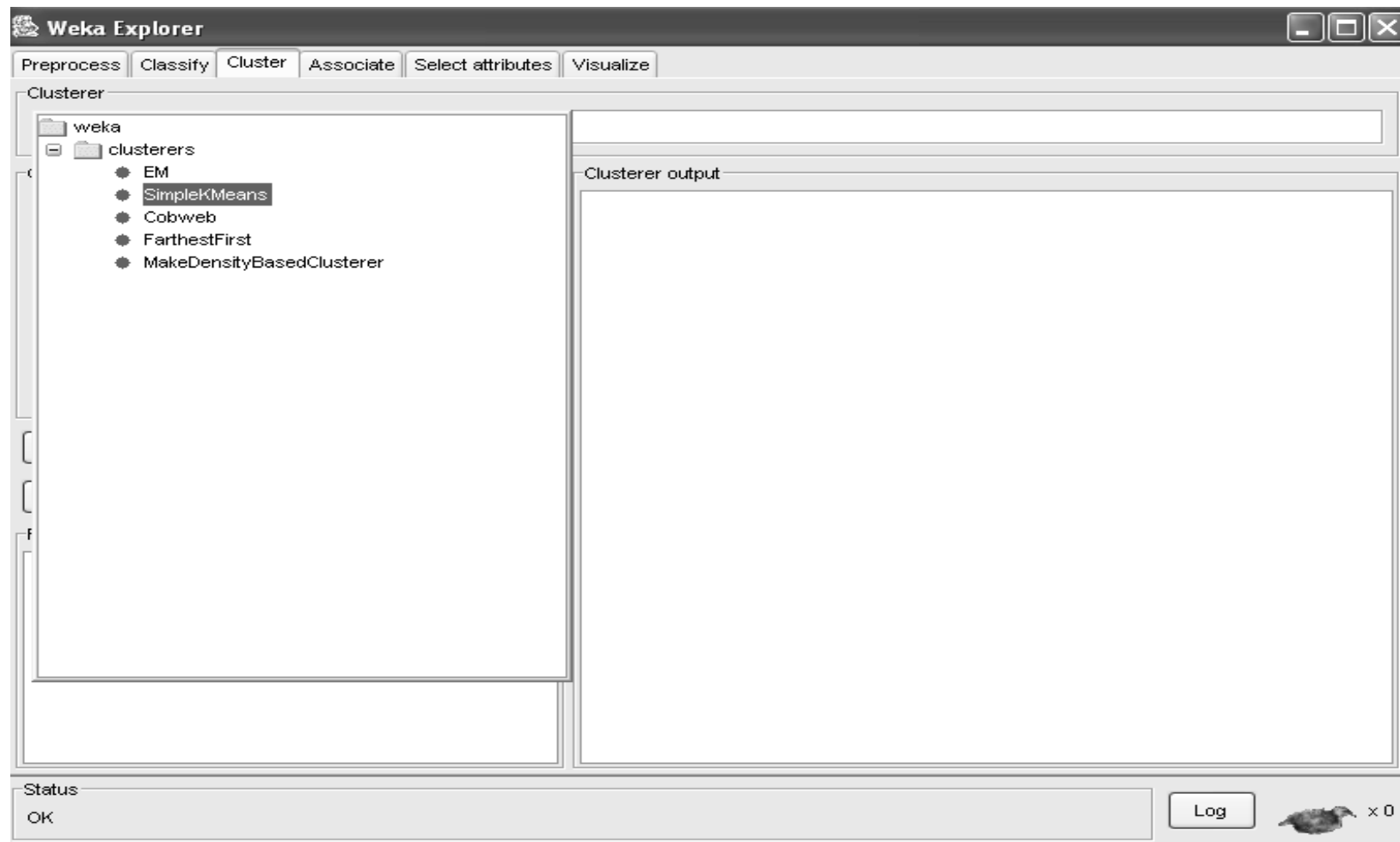
Context menu for 16:38:43 - trees.J48:

- 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

Status: OK Log x 0



Cluster (Iris.ARFF data)



Cluster – k-means

The screenshot shows the Weka Explorer interface with the 'Cluster' tab selected. The 'Clusterer' dropdown is set to 'SimpleKMeans -N 2 -S 10'. The 'Cluster mode' section has 'Use training set' selected, and 'Store clusters for visualization' is checked. The 'Start' button is highlighted with a red arrow. The 'Clusterer output' pane displays the following results:

```
kMeans
*****
Number of iterations: 2

Cluster centroids:

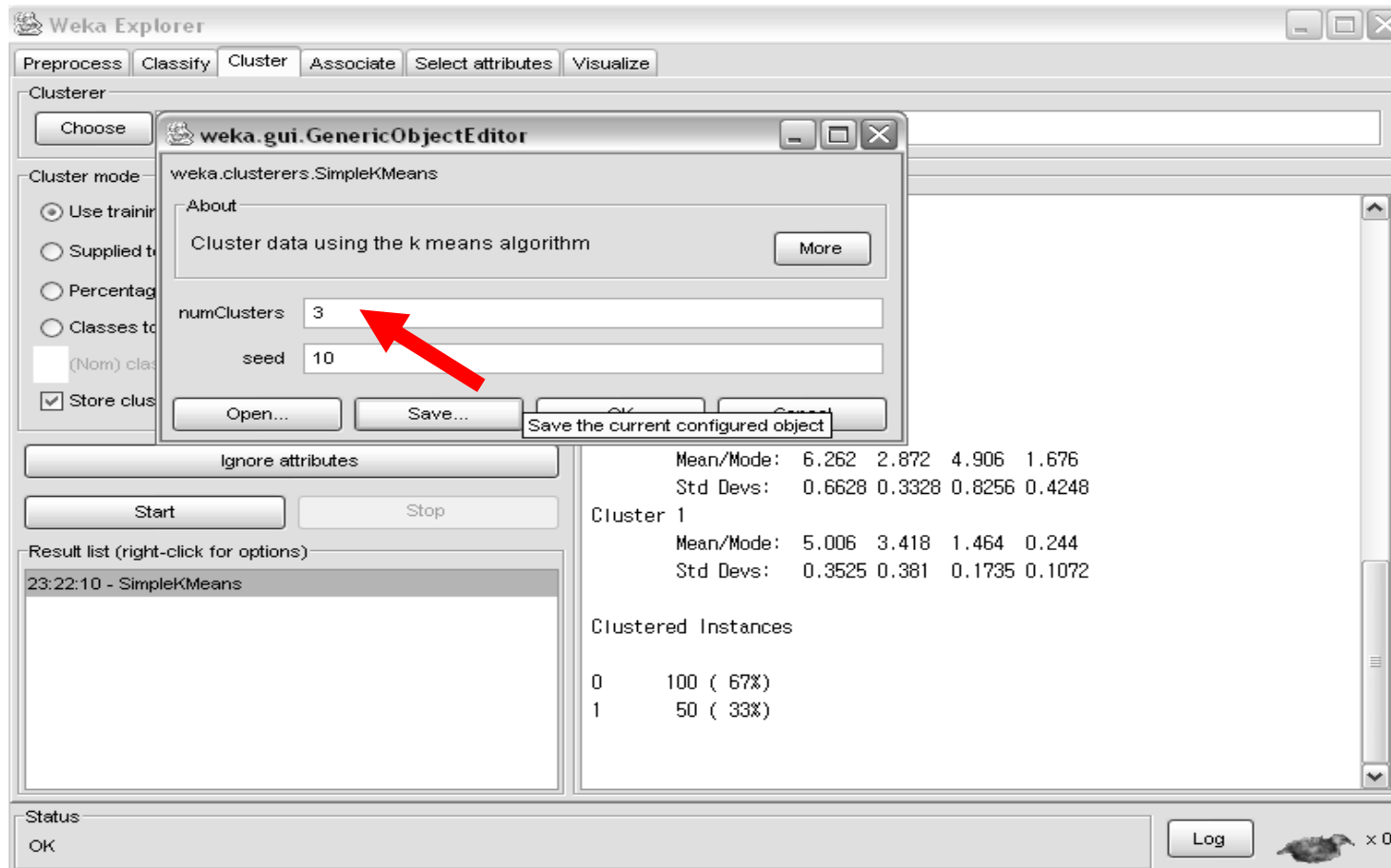
Cluster 0
  Mean/Mode:  6.262  2.872  4.906  1.676
  Std Devs:   0.6628 0.3328 0.8256 0.4248
Cluster 1
  Mean/Mode:  5.006  3.418  1.464  0.244
  Std Devs:   0.3525 0.381  0.1735 0.1072

Clustered Instances

0    100 ( 67%)
1     50 ( 33%)
```

The 'Result list' shows a single entry: '23:22:10 - SimpleKMeans'. The status bar at the bottom indicates 'OK' and 'Log'.

K- means: numClusters to 3



The screenshot shows the Weka Explorer interface with the 'Cluster' tab selected. A dialog box titled 'weka.gui.GenericObjectEditor' is open, showing the configuration for 'weka.clusterers.SimpleKMeans'. The 'numClusters' field is set to 3, highlighted by a red arrow. The 'seed' field is set to 10. The 'Save...' button is highlighted with a tooltip that reads 'Save the current configured object'.

Clusterer:

Cluster mode:

- Use training data
- Supplied training data
- Percentage of training data
- Classes to cluster
- (Nom) class
- Store clusters

Ignore attributes:

Start: Stop:

Result list (right-click for options):

- 23:22:10 - SimpleKMeans

Mean/Mode: 6.262 2.872 4.906 1.676
Std Devs: 0.6628 0.3328 0.8256 0.4248

Cluster 1
Mean/Mode: 5.006 3.418 1.464 0.244
Std Devs: 0.3525 0.381 0.1735 0.1072

Clustered Instances

0	100 (67%)
1	50 (33%)

Status: OK x 0

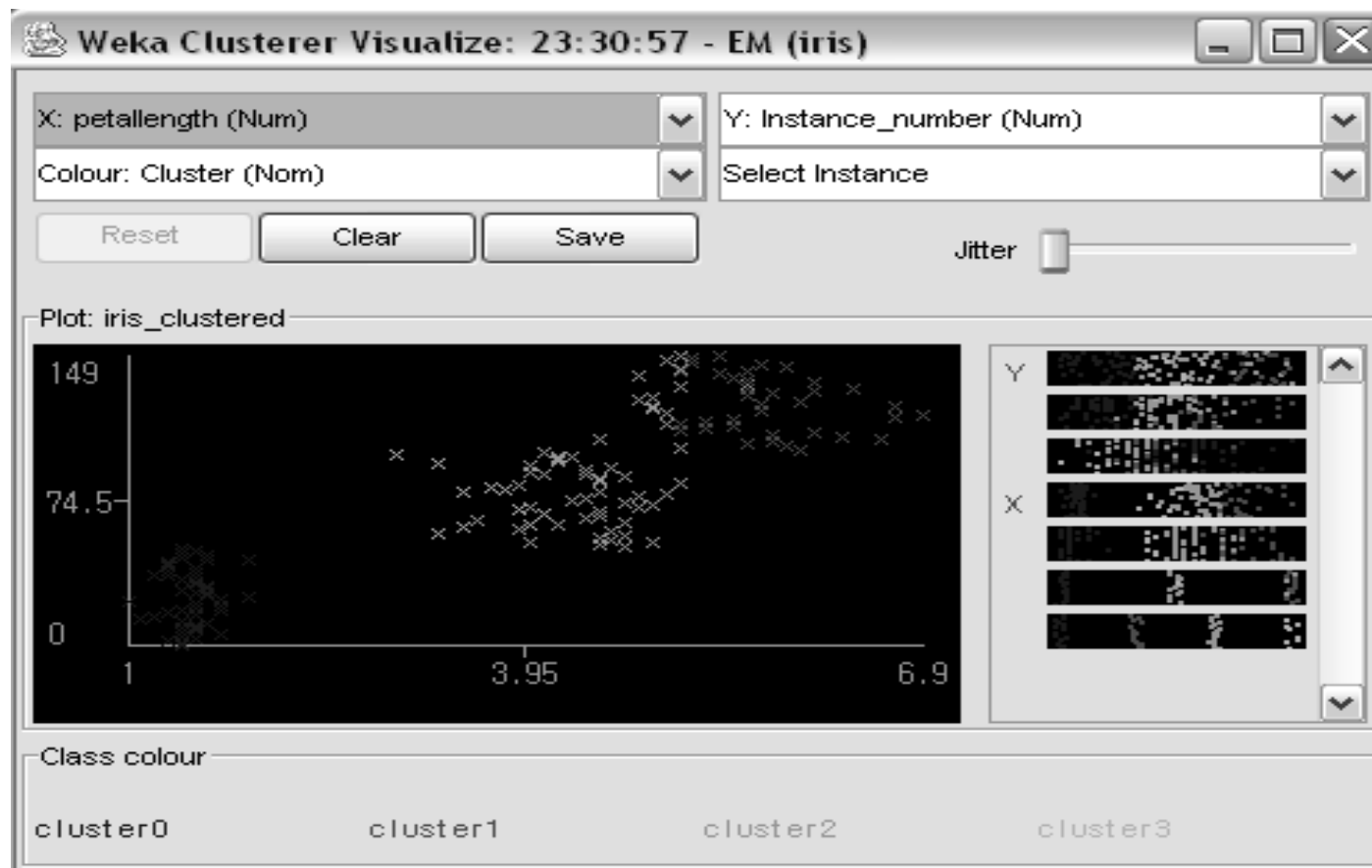
K – means clustered to 3 group

The screenshot shows the Weka Explorer interface with the 'Cluster' tab selected. The 'SimpleKMeans' algorithm is configured with 3 clusters and 10 iterations. The 'Cluster mode' section is set to 'Use training set'. The 'Clustered Instances' section shows the following distribution:

Cluster	Count	Percentage
0	52	35%
1	50	33%
2	48	32%

A red arrow points to the 'Clustered Instances' section. The 'Status' bar at the bottom shows 'OK'.

Visualization of clustering



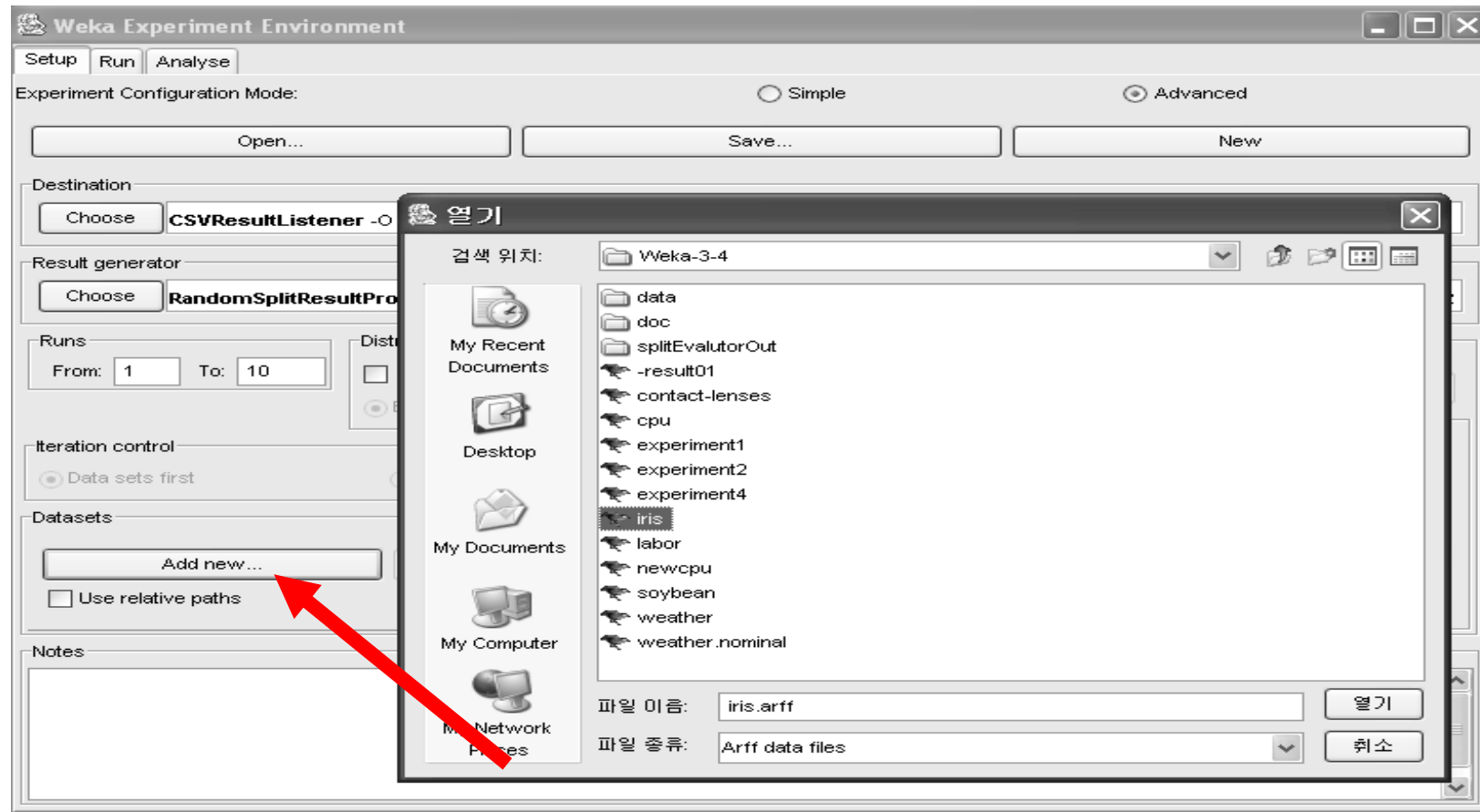
Cluseter – CobWeb

The screenshot displays the Weka Explorer application interface. The 'Clusterer' tab is active, showing the 'Cobweb' algorithm selected. The 'Cluster mode' section includes options for 'Use training set', 'Supplied test set', 'Percentage split' (set to 66%), and 'Classes to clusters evaluation'. A 'Start' button is visible. The 'Clusterer output' pane shows the following text:

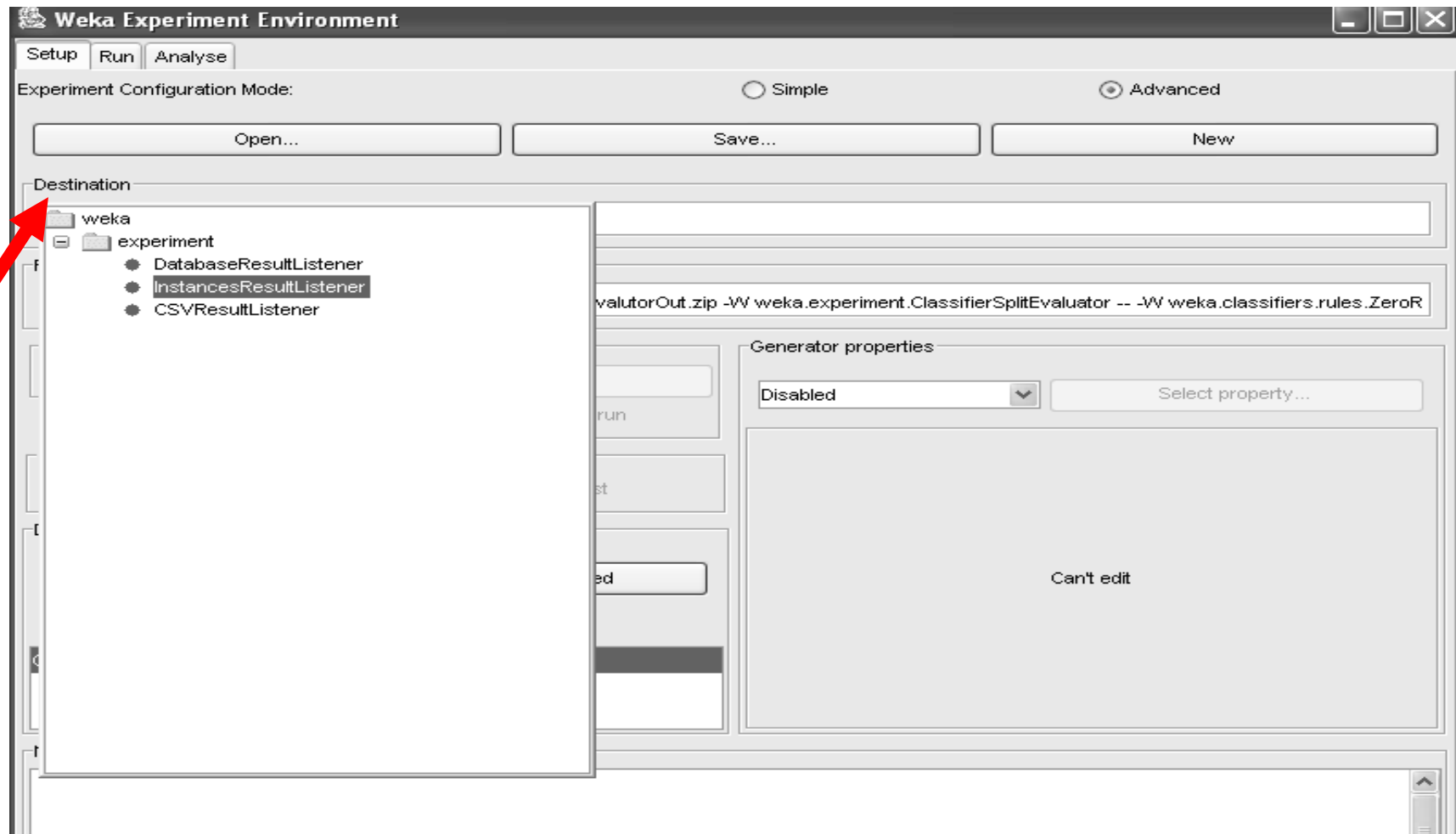
```
=== Evaluation on training set ===  
Number of merges: 2  
Number of splits: 2  
Number of clusters: 4  
  
node 0 [150]  
| leaf 1 [50]  
node 0 [150]  
| leaf 2 [50]  
node 0 [150]  
| leaf 3 [50]
```

A 'Weka Classifier Tree Visualiz...' window is open, displaying a tree view with a root node 'node 0 (150)' and three leaf nodes: 'leaf 1 (50)', 'leaf 2 (50)', and 'leaf 3 (50)'. A context menu is open over the result list, with 'Visualize tree' selected. The status bar at the bottom shows 'OK' and a 'Log' button.

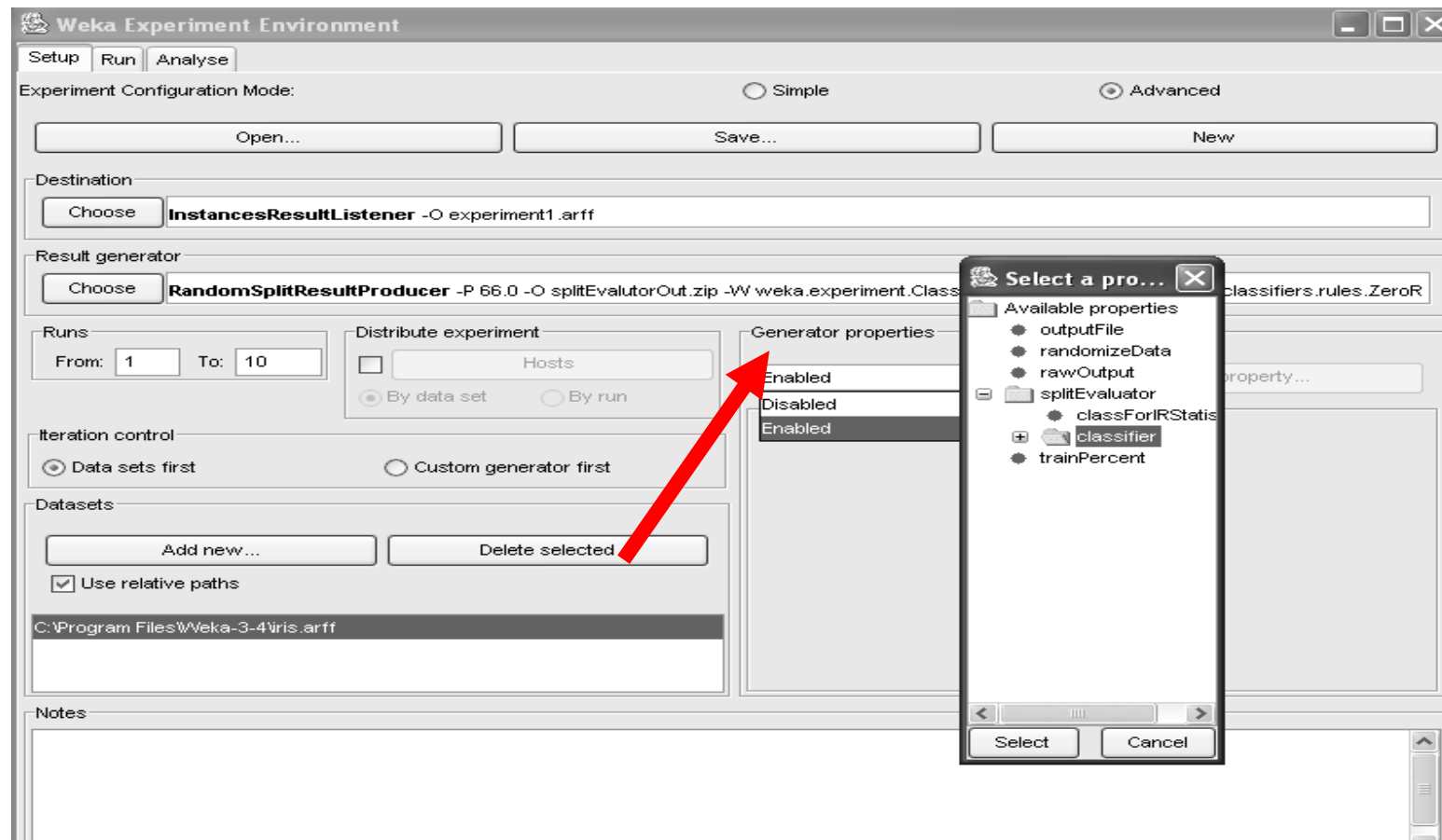
Experiment – add DataSet



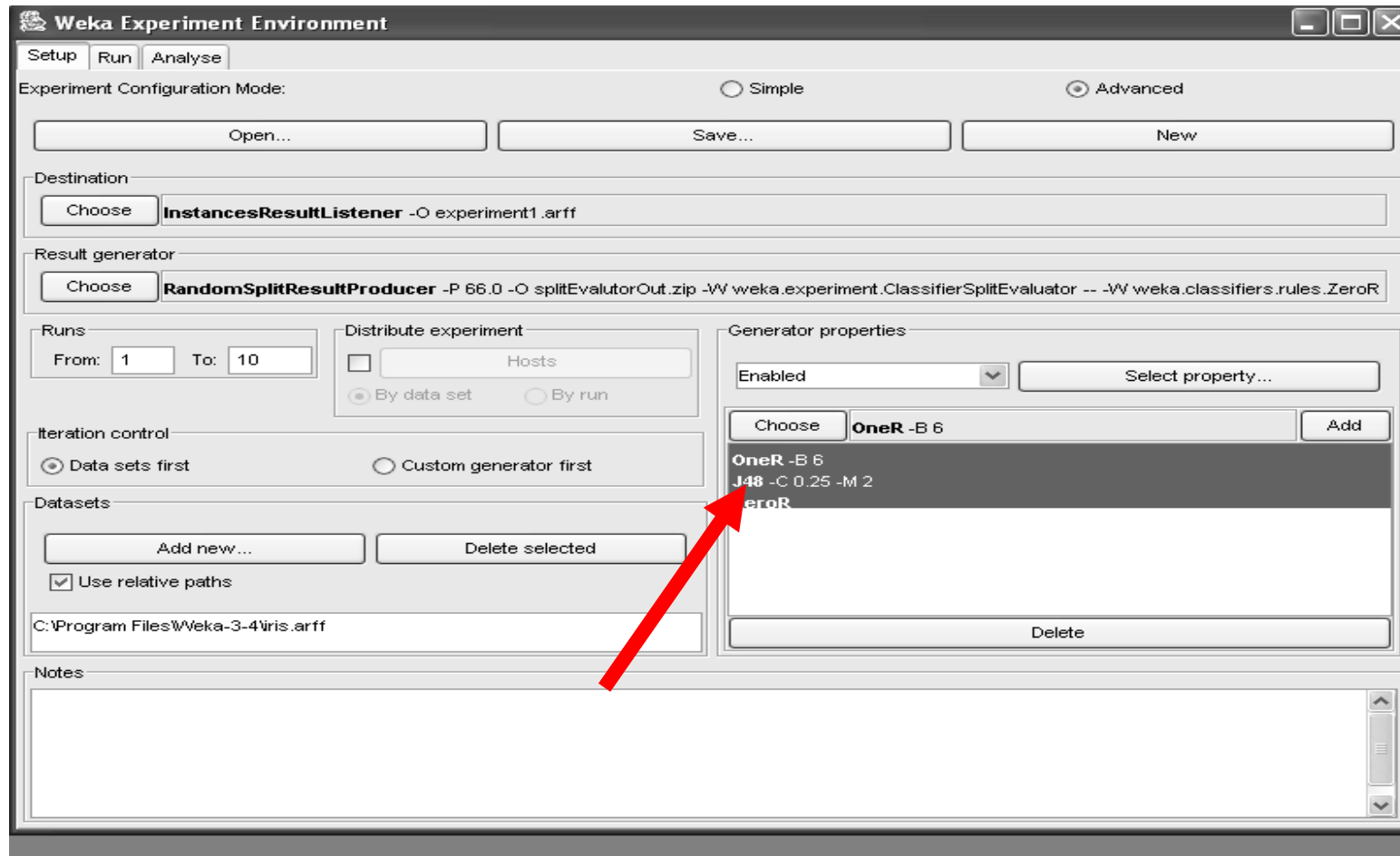
Experiment - destination



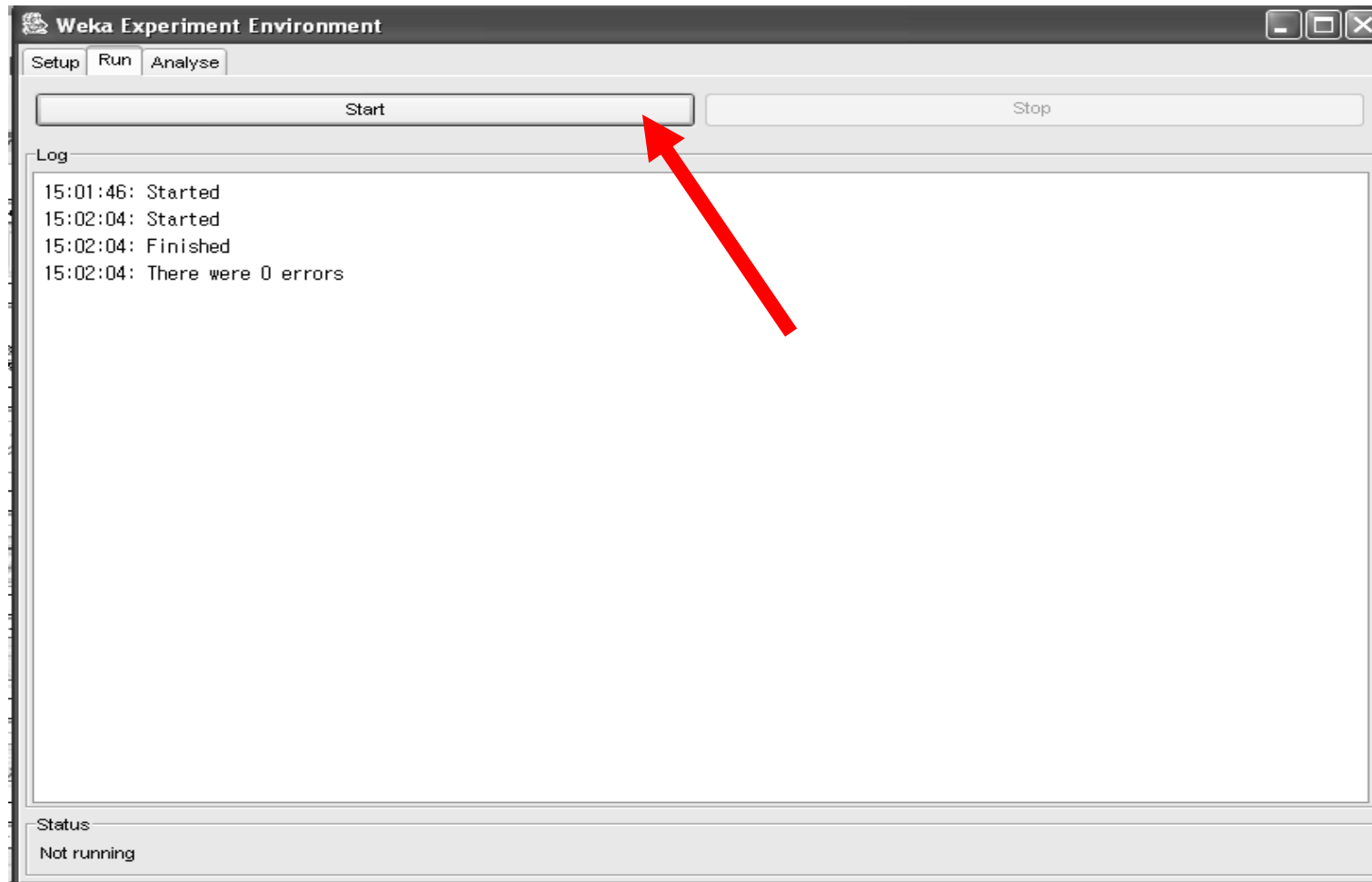
Experiment – classifying algorithm



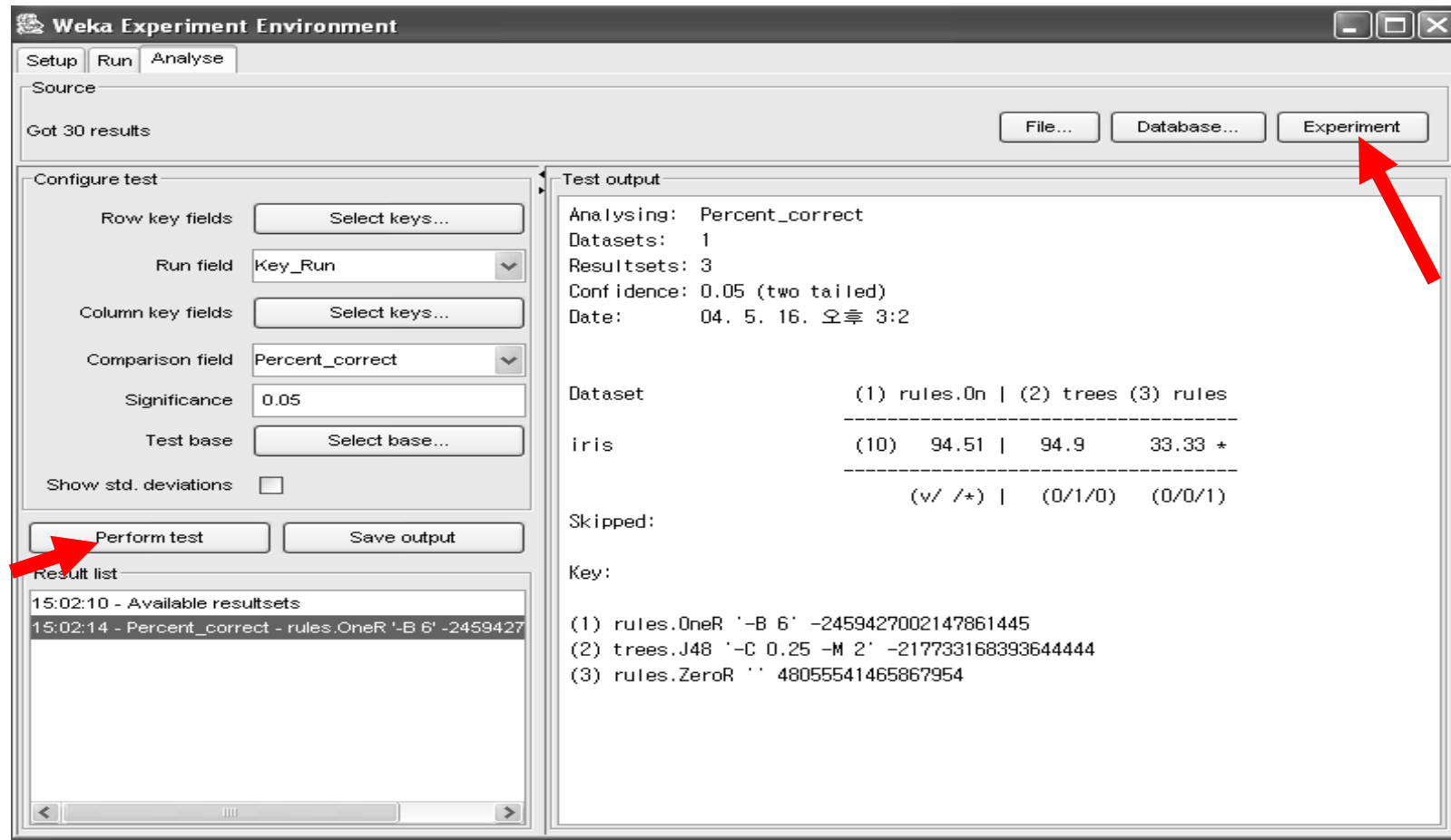
Experiment- multiple scheme



Experiment -run



Experiment - analysis



Weka Experiment Environment

Setup | Run | Analyse

Source

Got 30 results

File... Database... Experiment

Configure test

Row key fields: Select keys...

Run field: Key_Run

Column key fields: Select keys...

Comparison field: Percent_correct

Significance: 0.05

Test base: Select base...

Show std. deviations:

Perform test Save output

Result list

- 15:02:10 - Available resultsets
- 15:02:14 - Percent_correct - rules_OneR '-B 6' -2459427

Test output

Analysing: Percent_correct
Datasets: 1
Resultsets: 3
Confidence: 0.05 (two tailed)
Date: 04. 5. 16. 오후 3:2

Dataset	(1) rules.OneR	(2) trees	(3) rules
iris	(10) 94.51	94.9	33.33 +
	(v/ /+)	(0/1/0)	(0/0/1)

Skipped:

Key:

- (1) rules.OneR '-B 6' -2459427002147861445
- (2) trees.J48 '-C 0.25 -M 2' -217733168393644444
- (3) rules.ZeroR '' 48055541465867954

Experiment – better or worse

- Analysing: Percent_correct
- Datasets: 1
- Resultsets: 3
- Confidence: 0.05 (two tailed)
- Date: 04. 5. 16. 11:2

- Dataset (1) rules.On | (2) trees (3) rules

■	-----
■ iris	(10) 94.51 94.9 33.33 *

■	-----
■	(v/ /*) (0/1/0) (0/0/1)


- Skipped:

- Key:

- (1) rules.OneR '-B 6' -2459427002147861445
- (2) trees.J48 '-C 0.25 -M 2' -217733168393644444
- (3) rules.ZeroR " 48055541465867954



Experiment - summary

- Analysing: Percent_correct
 - Datasets: 1
 - Resultsets: 3
 - Confidence: 0.05 (two tailed)
 - Date: 04. 5. 16. 13:20
-
- a b c (No. of datasets where [col] >> [row])
 - - 0 0 | a = rules.OneR '-B 6' -2459427002147861445
 - 0 - 0 | b = trees.J48 '-C 0.25 -M 2' -217733168393644444
 - 1 1 - | c = rules.ZeroR " 48055541465867954
- 



Experiment - ranking

- Analysing: Percent_correct
 - Datasets: 1
 - Resultsets: 3
 - Confidence: 0.05 (two tailed)
 - Date: 04. 5. 16. 13:23
-
- >-< > < Resultset
 - 1 1 0 trees.J48 '-C 0.25 -M 2' -217733168393644444
 - 1 1 0 rules.OneR '-B 6' -2459427002147861445
 - -2 0 2 rules.ZeroR " 48055541465867954