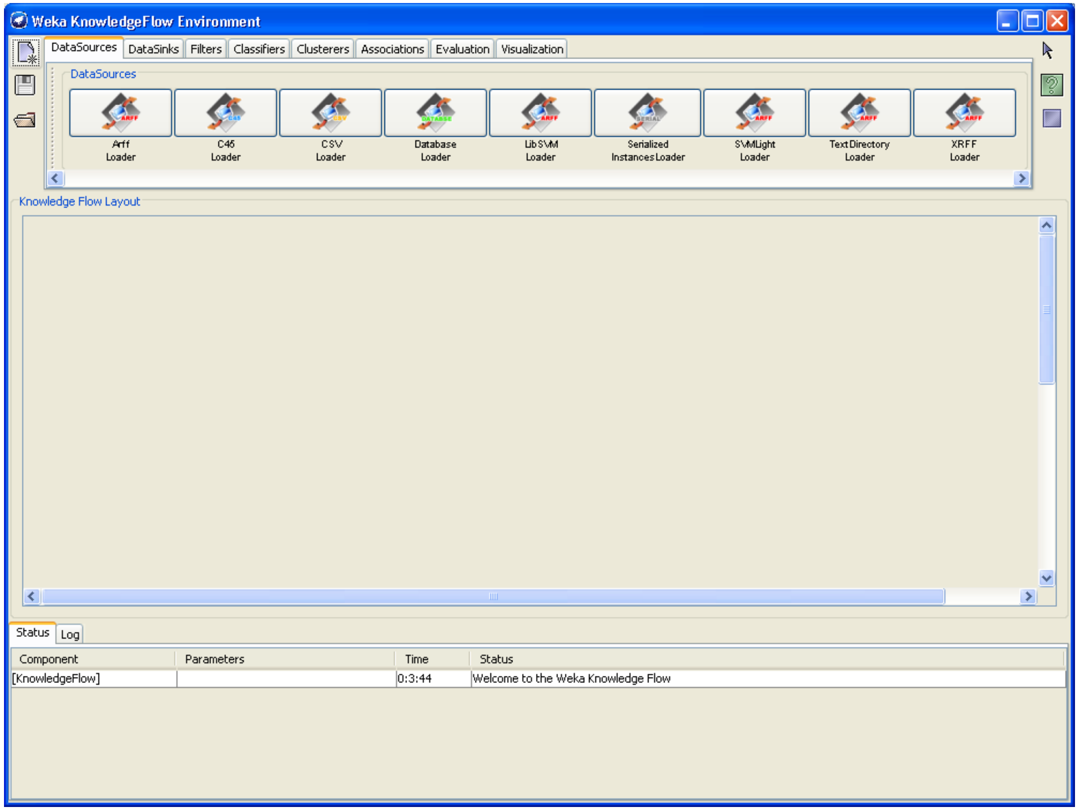
WEKA
Waikato Environment for Knowledge Analysis

Performing Classification Experiments

Prof. Pietro Ducange

The Knowledge Flow Interface

- It provides an alternative to the Explorer interface
- The user can select WEKA components from a palette, place them on a layout canvas and connect them together in order to form a knowledge flow for processing and analyzing data.



Component	Parameters	Time	Status
[KnowledgeFlow]		0:3:44	Welcome to the Weka Knowledge Flow

Knowledge Flow example (1)

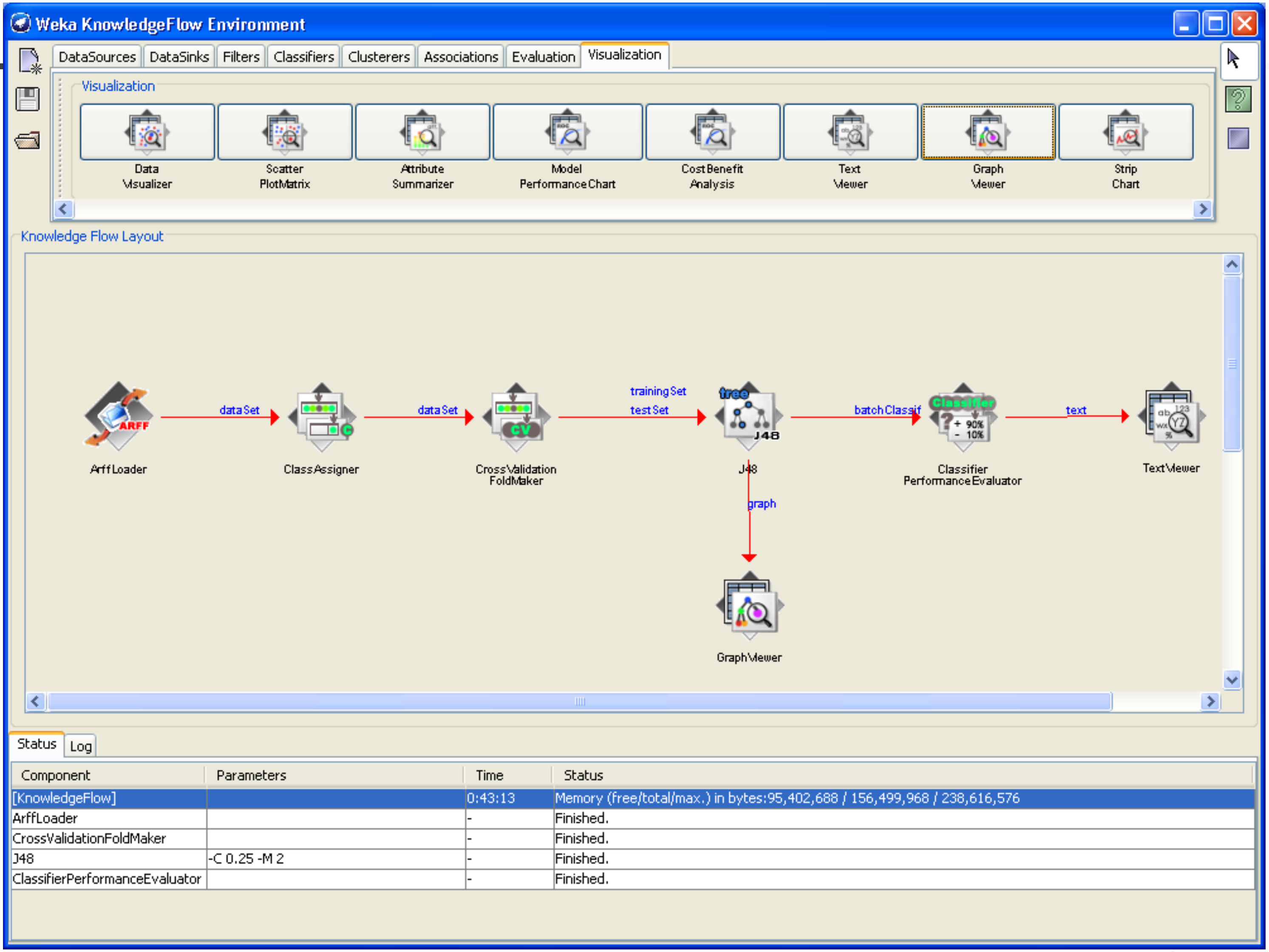
Setting up a flow to load an ARFF file and perform a cross-validation using J48

- Create a source of data (DataSources tab - ARFFLoader)
- Connect it to a ARFF file (right click over the ARFFLoader icon - Configure)
- Specify which attribute is the class (Evaluation tab – ClassAssigner)
- Connect the ArffLoader to the ClassAssigner (right click over the ArffLoader, select the dataSet under Connections and link with the ClassAssigner component with a left click
- Specify which column is the class (right click over the ClassAssigner - choose Configure)
- Add a CrossValidationFoldMaker component (Evaluation)
- Connect the ClassAssigner to the CrossValidationFoldMaker (right click over ClassAssigner, select dataSet, left click over CrossValidationFoldMaker

Knowledge Flow example (2)

- Select the J48 component (classifiers tab)
- Connect the CrossValidationFoldMaker to J48 TWICE (right click over CrossValidationFoldMaker, first choose trainingSet and then testSet)
- Select ClassifierPerformanceEvaluator component (Evaluation tab)
- Connect J48 to this component (right click over J48, select batchClassifier left click over by ClassifierPerformanceEvaluator
- Select TextViewer component (Visualization tab)
- Connect the ClassifierPerformanceEvaluator to the TextViewer (select the text entry from the pop-up menu for ClassifierPerformanceEvaluator)
- Select GraphViewer component (Vizualization tab) and link to J48 (select the graph entry from the pop-up menu for J48)
- Start the flow (select start loading from the pop-up menu for the loader)

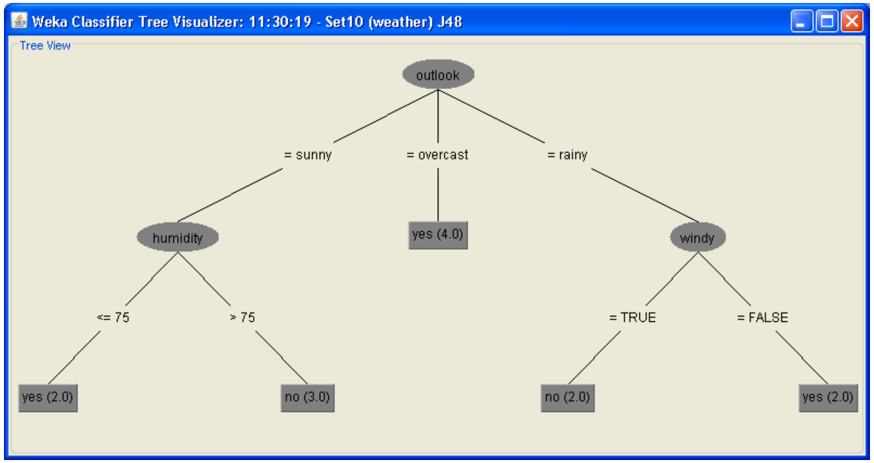
Knowledge Flow example (3)



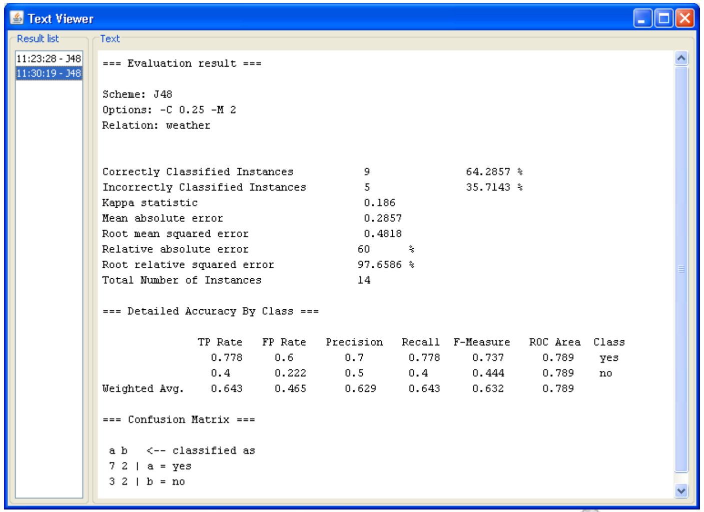
Component	Parameters	Time	Status
[KnowledgeFlow]		0:43:13	Memory (free/total/max.) in bytes:95,402,688 / 156,499,968 / 238,616,576
ArffLoader		-	Finished.
CrossValidationFoldMaker		-	Finished.
J48	-C 0.25 -M 2	-	Finished.
ClassifierPerformanceEvaluator		-	Finished.

Knowledge Flow example (4)

- Select show results from the pop-up menu for the graph viewer

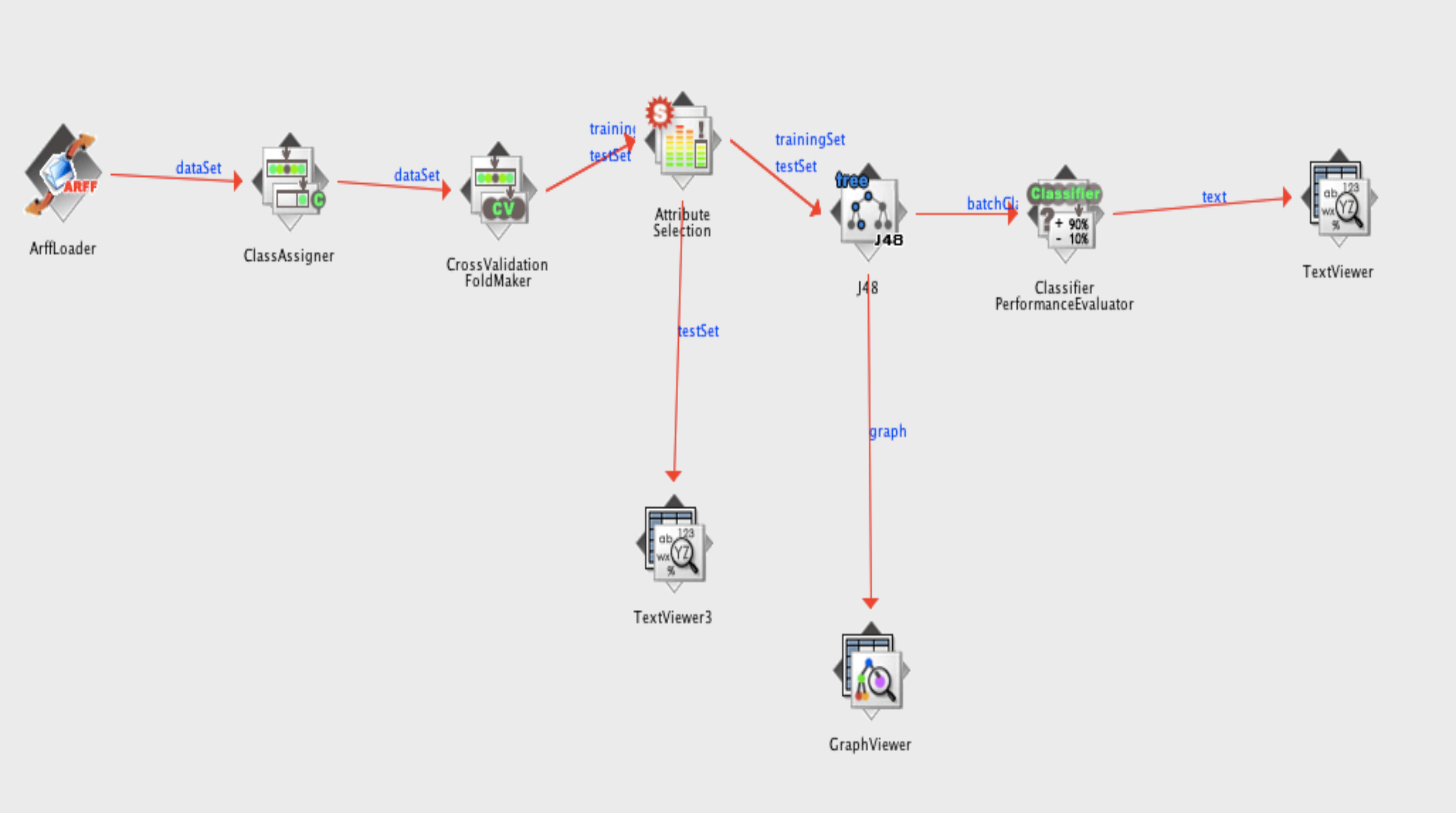


- Select show results from the pop-up menu for the text viewer



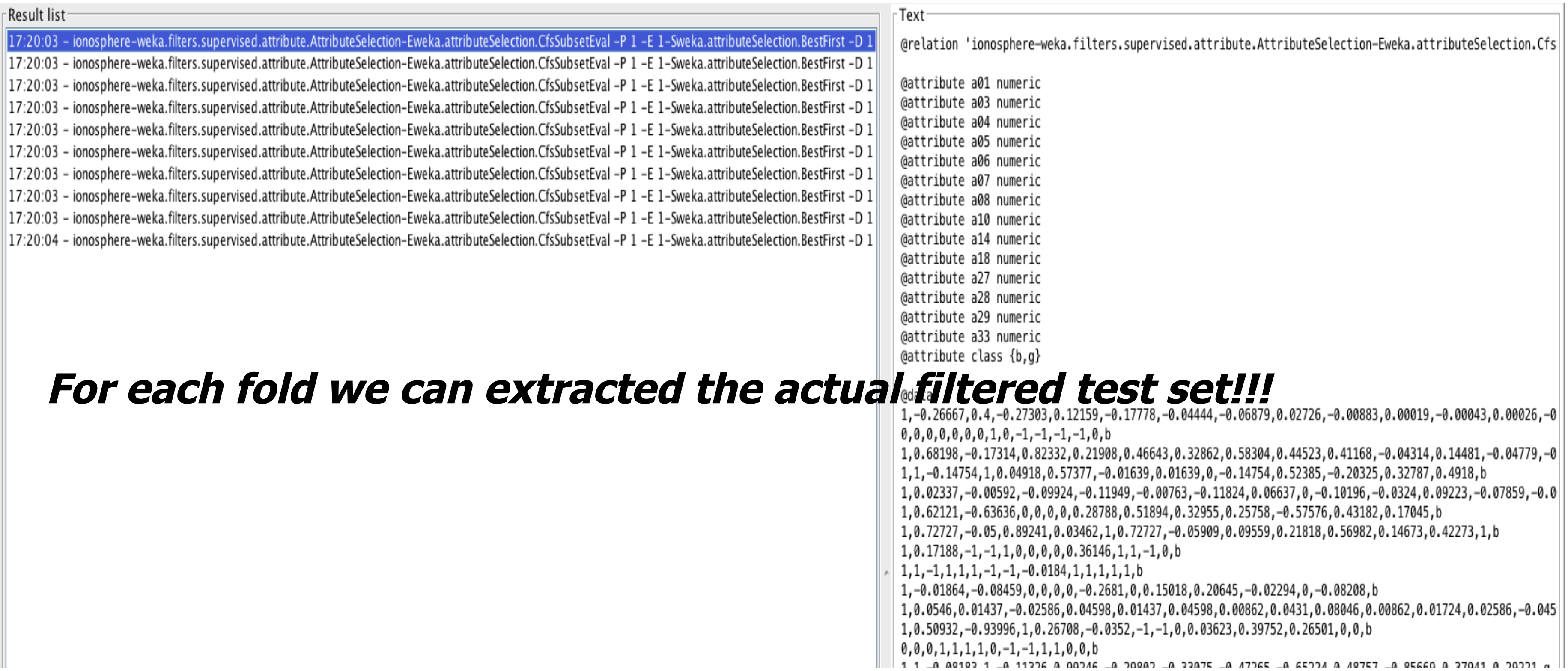
```
=== Evaluation result ===

Scheme: J48
Options: -C 0.25 -M 2
Relation: weather


Correctly Classified Instances           9               64.2857 %
Incorrectly Classified Instances         5               35.7143 %
Kappa statistic                          0.186
Mean absolute error                      0.2857
Root mean squared error                  0.4818
Relative absolute error                 60      %
Root relative squared error             97.6586 %
Total Number of Instances               14

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.778     0.6        0.7        0.778     0.737      0.789    yes
                 0.4       0.222      0.5        0.4       0.444      0.789    no
Weighted Avg.    0.643     0.465      0.629      0.643     0.632      0.789

=== Confusion Matrix ===

 a b   <-- classified as
 7 2 | a = yes
 3 2 | b = no
```

Knowledge Flow: attribute selection



Knowledge Flow: attribute selection

Select show results from the pop-up menu for the text viewer connected to the Attribute Selection Block



Result list

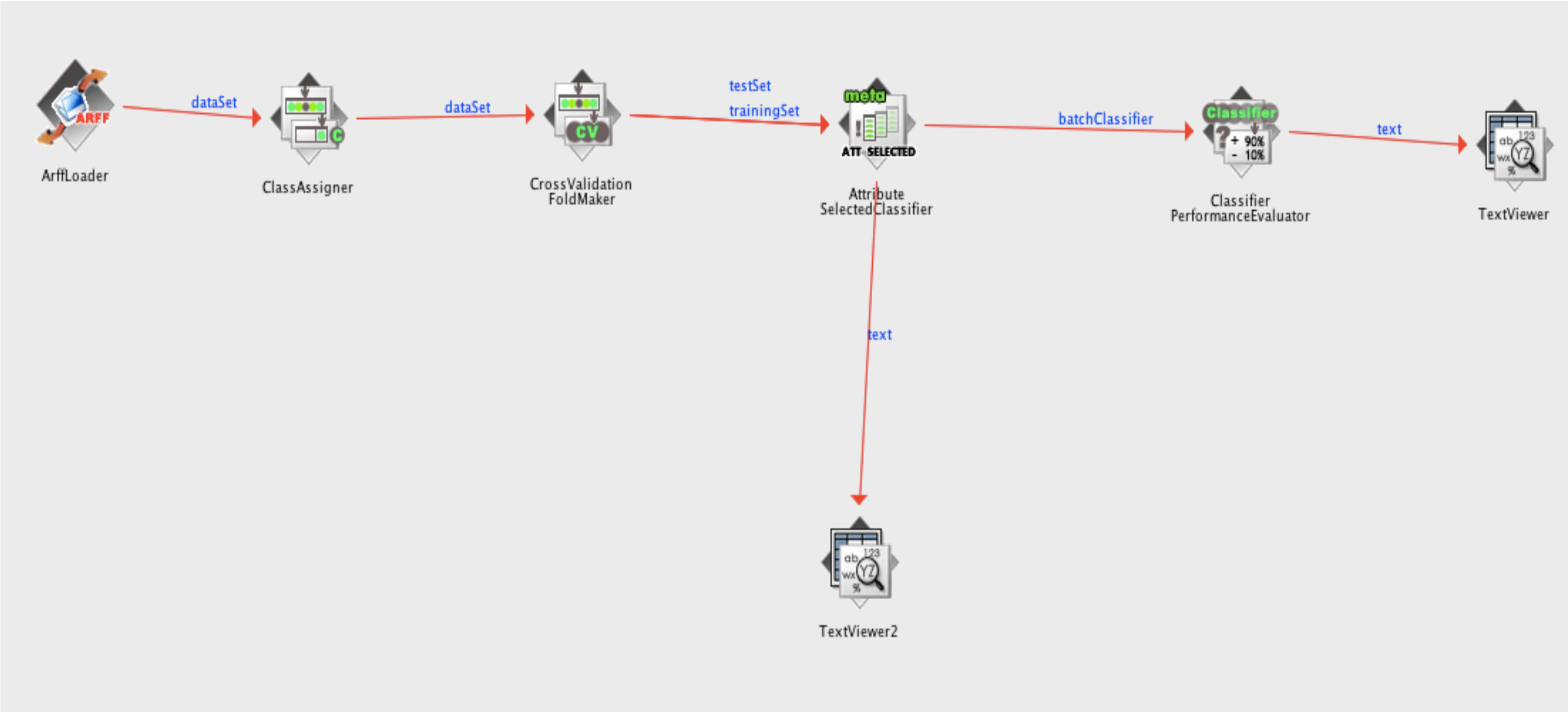
```
17:20:03 - ionosphere-weka.filters.supervised.attribute.AttributeSelection-Eweka.attributeSelection.CfsSubsetEval -P 1 -E 1-Sweka.attributeSelection.BestFirst -D 1
17:20:03 - ionosphere-weka.filters.supervised.attribute.AttributeSelection-Eweka.attributeSelection.CfsSubsetEval -P 1 -E 1-Sweka.attributeSelection.BestFirst -D 1
17:20:03 - ionosphere-weka.filters.supervised.attribute.AttributeSelection-Eweka.attributeSelection.CfsSubsetEval -P 1 -E 1-Sweka.attributeSelection.BestFirst -D 1
17:20:03 - ionosphere-weka.filters.supervised.attribute.AttributeSelection-Eweka.attributeSelection.CfsSubsetEval -P 1 -E 1-Sweka.attributeSelection.BestFirst -D 1
17:20:03 - ionosphere-weka.filters.supervised.attribute.AttributeSelection-Eweka.attributeSelection.CfsSubsetEval -P 1 -E 1-Sweka.attributeSelection.BestFirst -D 1
17:20:03 - ionosphere-weka.filters.supervised.attribute.AttributeSelection-Eweka.attributeSelection.CfsSubsetEval -P 1 -E 1-Sweka.attributeSelection.BestFirst -D 1
17:20:03 - ionosphere-weka.filters.supervised.attribute.AttributeSelection-Eweka.attributeSelection.CfsSubsetEval -P 1 -E 1-Sweka.attributeSelection.BestFirst -D 1
17:20:03 - ionosphere-weka.filters.supervised.attribute.AttributeSelection-Eweka.attributeSelection.CfsSubsetEval -P 1 -E 1-Sweka.attributeSelection.BestFirst -D 1
17:20:03 - ionosphere-weka.filters.supervised.attribute.AttributeSelection-Eweka.attributeSelection.CfsSubsetEval -P 1 -E 1-Sweka.attributeSelection.BestFirst -D 1
17:20:04 - ionosphere-weka.filters.supervised.attribute.AttributeSelection-Eweka.attributeSelection.CfsSubsetEval -P 1 -E 1-Sweka.attributeSelection.BestFirst -D 1
```

Text

```
@relation 'ionosphere-weka.filters.supervised.attribute.AttributeSelection-Eweka.attributeSelection.Cfs

@attribute a01 numeric
@attribute a03 numeric
@attribute a04 numeric
@attribute a05 numeric
@attribute a06 numeric
@attribute a07 numeric
@attribute a08 numeric
@attribute a10 numeric
@attribute a14 numeric
@attribute a18 numeric
@attribute a27 numeric
@attribute a28 numeric
@attribute a29 numeric
@attribute a33 numeric
@attribute class {b,g}

@data
1,-0.26667,0.4,-0.27303,0.12159,-0.17778,-0.04444,-0.06879,0.02726,-0.00883,0.00019,-0.00043,0.00026,-0
0,0,0,0,0,0,0,1,0,-1,-1,-1,-1,0,b
1,0.68198,-0.17314,0.82332,0.21908,0.46643,0.32862,0.58304,0.44523,0.41168,-0.04314,0.14481,-0.04779,-0
1,1,-0.14754,1,0.04918,0.57377,-0.01639,0.01639,0,-0.14754,0.52385,-0.20325,0.32787,0.4918,b
1,0.02337,-0.00592,-0.09924,-0.11949,-0.00763,-0.11824,0.06637,0,-0.10196,-0.0324,0.09223,-0.07859,-0.0
1,0.62121,-0.63636,0,0,0,0,0.28788,0.51894,0.32955,0.25758,-0.57576,0.43182,0.17045,b
1,0.72727,-0.05,0.89241,0.03462,1,0.72727,-0.05909,0.09559,0.21818,0.56982,0.14673,0.42273,1,b
1,0.17188,-1,-1,1,0,0,0,0,0.36146,1,1,-1,0,b
1,1,-1,1,1,1,-1,-1,-0.0184,1,1,1,1,1,b
1,-0.01864,-0.08459,0,0,0,0,-0.2681,0,0.15018,0.20645,-0.02294,0,-0.08208,b
1,0.0546,0.01437,-0.02586,0.04598,0.01437,0.04598,0.00862,0.0431,0.08046,0.00862,0.01724,0.02586,-0.045
1,0.50932,-0.93996,1,0.26708,-0.0352,-1,-1,0,0.03623,0.39752,0.26501,0,0,b
0,0,0,1,1,1,1,0,-1,-1,1,1,0,0,b
1,1,-0.08183,1,-0.11326,0.99246,-0.29802,-0.33075,-0.47265,-0.65224,0.48757,-0.85669,0.37941,0.29221,g
```

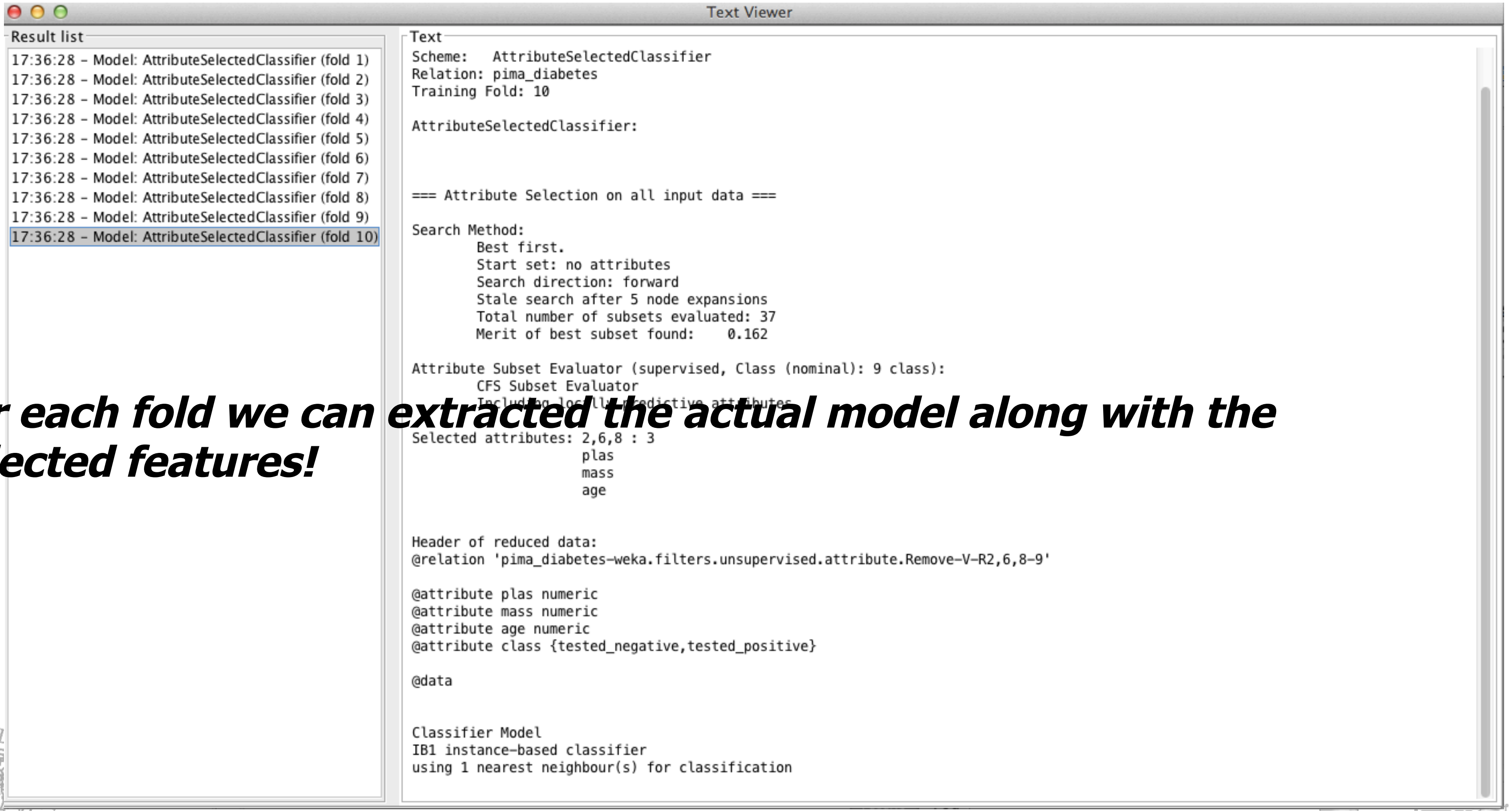
For each fold we can extracted the actual filtered test set!!!

Knowledge Flow: metaclassification



Knowledge Flow: attribute selection

Select show results from the pop-up menu for the text viewer connected to the Meta Classifier Block



```
17:36:28 - Model: AttributeSelectedClassifier (fold 1)
17:36:28 - Model: AttributeSelectedClassifier (fold 2)
17:36:28 - Model: AttributeSelectedClassifier (fold 3)
17:36:28 - Model: AttributeSelectedClassifier (fold 4)
17:36:28 - Model: AttributeSelectedClassifier (fold 5)
17:36:28 - Model: AttributeSelectedClassifier (fold 6)
17:36:28 - Model: AttributeSelectedClassifier (fold 7)
17:36:28 - Model: AttributeSelectedClassifier (fold 8)
17:36:28 - Model: AttributeSelectedClassifier (fold 9)
17:36:28 - Model: AttributeSelectedClassifier (fold 10)
```

```
Scheme:   AttributeSelectedClassifier
Relation: pima_diabetes
Training Fold: 10

AttributeSelectedClassifier:



=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 37
        Merit of best subset found:    0.162

Attribute Subset Evaluator (supervised, Class (nominal): 9 class):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 2,6,8 : 3
                     plas
                     mass
                     age


Header of reduced data:
@relation 'pima_diabetes-weka.filters.unsupervised.attribute.Remove-V-R2,6,8-9'

@attribute plas numeric
@attribute mass numeric
@attribute age numeric
@attribute class {tested_negative,tested_positive}

@data


Classifier Model
IB1 instance-based classifier
using 1 nearest neighbour(s) for classification
```

For each fold we can extracted the actual model along with the selected features!

The Experimenter

- A robust experimental part involves running several learning schemes on different datasets.

- The Experimenter interface enables us to set-up large scale experiments.
- The user can create an experiment that runs several schemes against a series of datasets and then analyze the results to determine if one of the schemes is (statistically) better than the other schemes.

Simple setup

Experiment type:

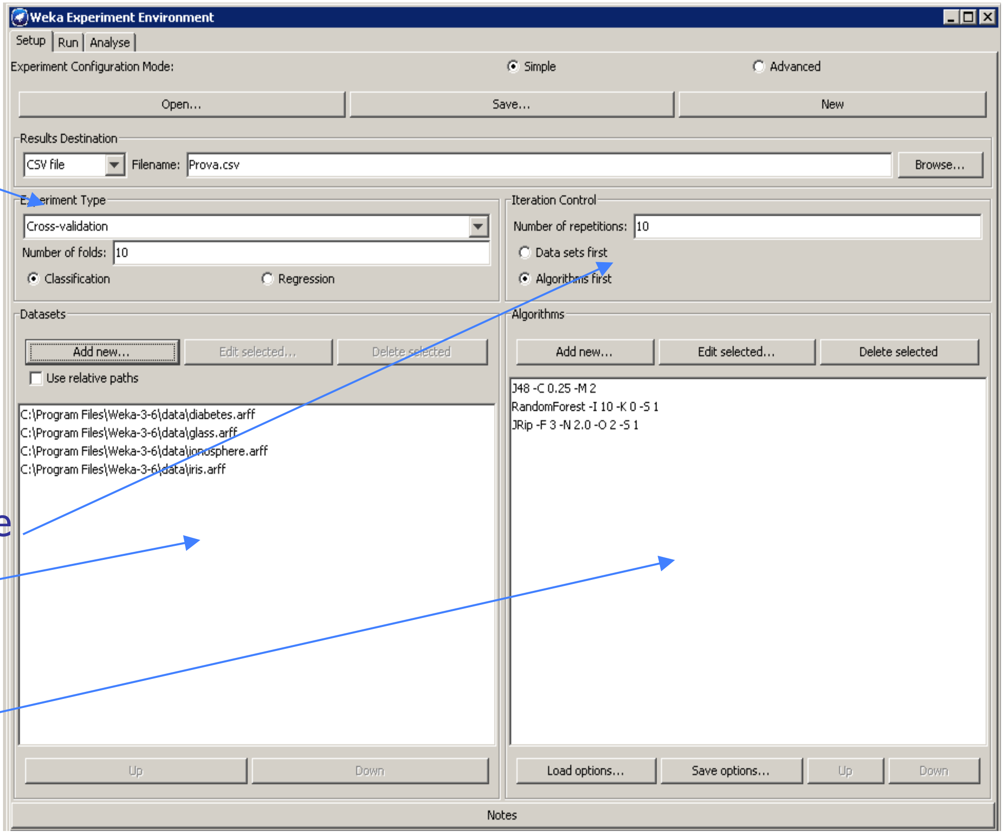
- Cross-validation (default), Train/Test Percentage Split (data randomized or order preserved)
- Number of folds
- Classification/Regression

Iteration control

- Set the number of repetition and change the order of iterations

Datasets

Algorithms



Weka Experiment Environment

Experiment Configuration Mode: ● Simple ○ Advanced

Results Destination

CSV file Filename: Prova.csv

Experiment Type

Cross-validation

Number of folds: 10

● Classification ○ Regression

Iteration Control

Number of repetitions: 10

○ Data sets first

● Algorithms first

Datasets

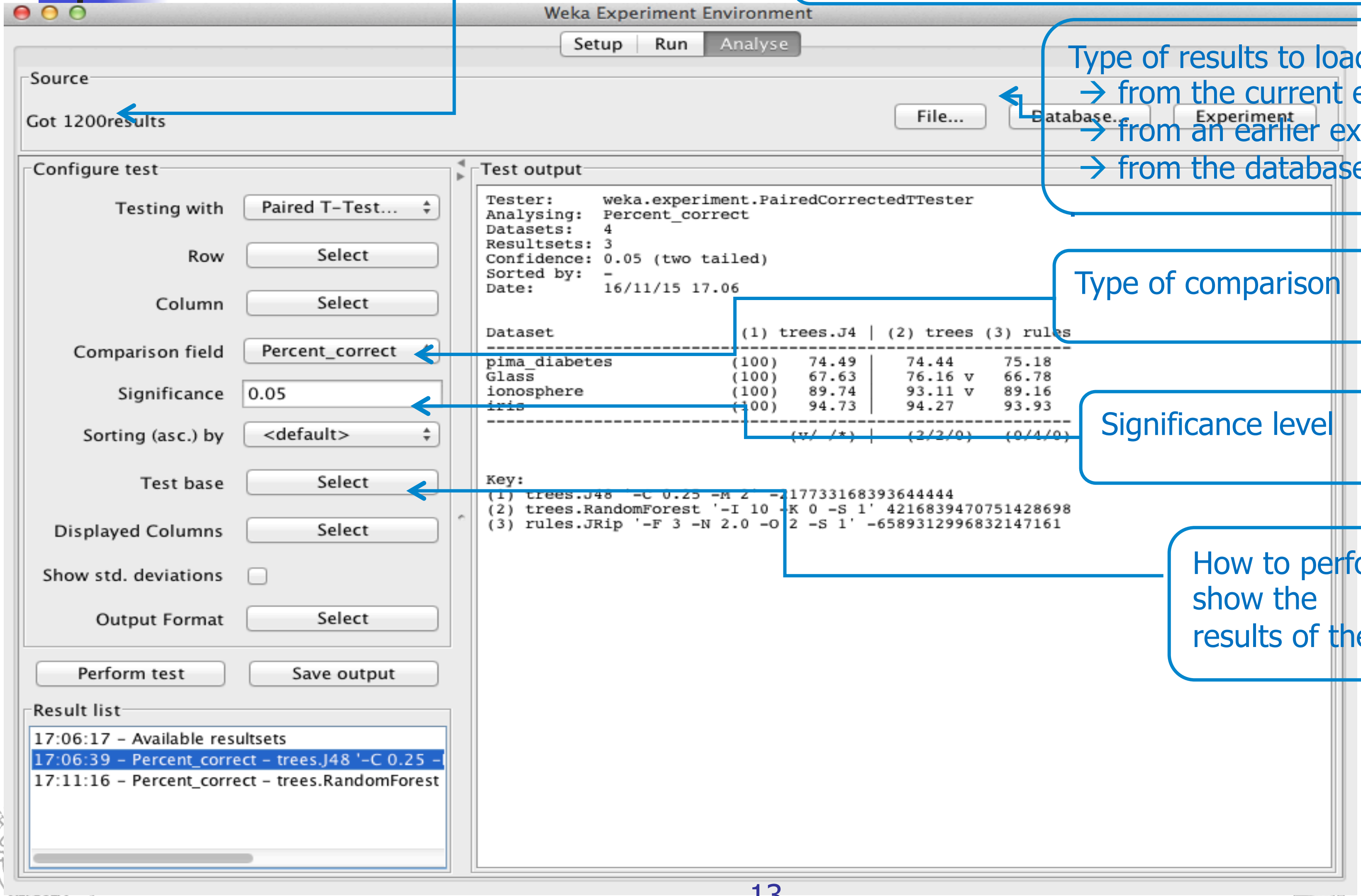
□ Use relative paths

```
C:\Program Files\Weka-3-6\data\diabetes.arff
C:\Program Files\Weka-3-6\data\glass.arff
C:\Program Files\Weka-3-6\data\ionosphere.arff
C:\Program Files\Weka-3-6\data\iris.arff
```

Algorithms

```
J48 -C 0.25 -M 2
RandomForest -I 10 -K 0 -S 1
JRip -F 3 -N 2.0 -O 2 -S 1
```

The Analyze panel



Got 1200results

Testing with: Paired T-Test...

Row: Select

Column: Select

Comparison field: Percent_correct

Significance: 0.05

Sorting (asc.) by: <default>

Test base: Select

Displayed Columns: Select

Show std. deviations: □

Output Format: Select

```
Tester:     weka.experiment.PairedCorrectedTTester
Analysing:  Percent_correct
Datasets:   4
Resultsets: 3
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       16/11/15 17.06


Dataset                    (1) trees.J4 | (2) trees (3) rules
------------------------------------------------------------
pima_diabetes             (100)   74.49 |   74.44     75.18
Glass                     (100)   67.63 |   76.16 v   66.78
ionosphere                (100)   89.74 |   93.11 v   89.16
iris                      (100)   94.73 |   94.27     93.93
------------------------------------------------------------
                              (v/ /*) |   (2/2/0)   (0/4/0)


Key:
(1) trees.J48 '-C 0.25 -M 2' -217733168393644444
(2) trees.RandomForest '-I 10 -K 0 -S 1' 4216839470751428698
(3) rules.JRip '-F 3 -N 2.0 -O 2 -S 1' -6589312996832147161
```

```
17:06:17 - Available resultsets
17:06:39 - Percent_correct - trees.J48 '-C 0.25 -
17:11:16 - Percent_correct - trees.RandomForest
```

The number of result lines available

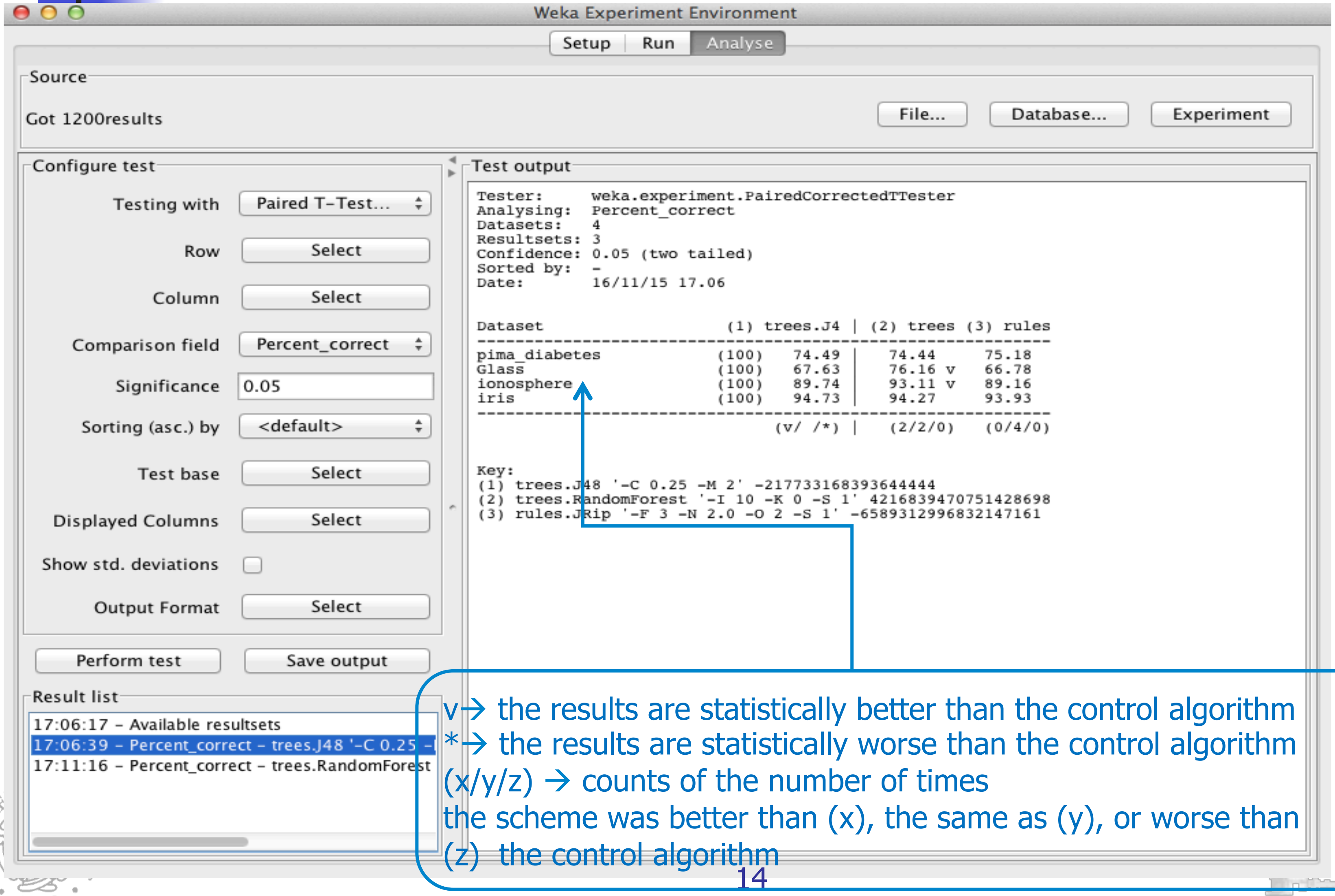
Type of results to load:
→ from the current experiment
→ from an earlier experiment file
→ from the database

Type of comparison

Significance level

How to perform and show the results of the test

The paired T-test results respect to a control algorithm (C4.5)



Source

Got 1200results

Configure test

Testing with: Paired T-Test...

Row: Select

Column: Select

Comparison field: Percent_correct

Significance: 0.05

Sorting (asc.) by: <default>

Test base: Select

Displayed Columns: Select

Show std. deviations: □

Output Format: Select

Result list

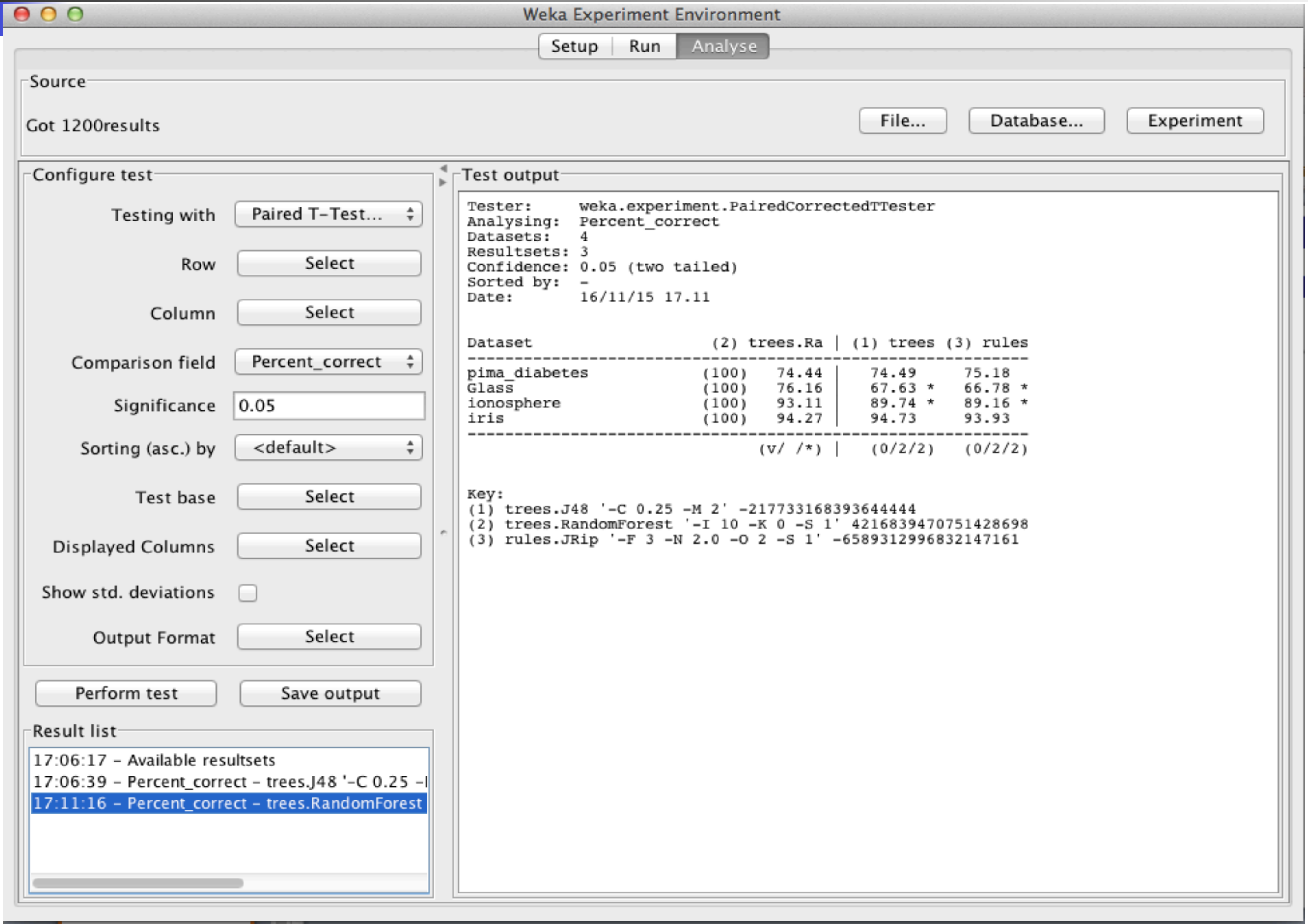
17:06:17 - Available resultsets
17:06:39 - Percent_correct - trees.J48 '-C 0.25 -
17:11:16 - Percent_correct - trees.RandomForest

Test output

```
Tester:     weka.experiment.PairedCorrectedTTester
Analysing:  Percent_correct
Datasets:   4
Resultsets: 3
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       16/11/15 17.06


Dataset                   (1) trees.J4 | (2) trees (3) rules
------------------------------------------------------------
pima_diabetes            (100)   74.49 |   74.44     75.18
Glass                    (100)   67.63 |   76.16 v   66.78
ionosphere               (100)   89.74 |   93.11 v   89.16
iris                     (100)   94.73 |   94.27     93.93
------------------------------------------------------------
                               (v/ /*) |   (2/2/0)   (0/4/0)


Key:
(1) trees.J48 '-C 0.25 -M 2' -217733168393644444
(2) trees.RandomForest '-I 10 -K 0 -S 1' 4216839470751428698
(3) rules.JRip '-F 3 -N 2.0 -O 2 -S 1' -6589312996832147161
```

v→ the results are statistically better than the control algorithm
*→ the results are statistically worse than the control algorithm
(x/y/z) → counts of the number of times
the scheme was better than (x), the same as (y), or worse than
(z) the control algorithm

The paired T-test results respect to a control algorithm (Random Forest)



Got 1200results

Testing with: Paired T-Test...

Comparison field: Percent_correct

Significance: 0.05

Sorting (asc.) by: <default>

Show std. deviations: □

17:06:17 - Available resultsets
17:06:39 - Percent_correct - trees.J48 '-C 0.25 -I
17:11:16 - Percent_correct - trees.RandomForest

```
Tester:     weka.experiment.PairedCorrectedTTester
Analysing:  Percent_correct
Datasets:   4
Resultsets: 3
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       16/11/15 17.11


Dataset                   (2) trees.Ra | (1) trees (3) rules
------------------------------------------------------------
pima_diabetes            (100)   74.44 |   74.49     75.18  
Glass                    (100)   76.16 |   67.63 *   66.78 *
ionosphere               (100)   93.11 |   89.74 *   89.16 *
iris                     (100)   94.27 |   94.73     93.93  
------------------------------------------------------------
                               (v/ /*) |   (0/2/2)   (0/2/2)


Key:
(1) trees.J48 '-C 0.25 -M 2' -217733168393644444
(2) trees.RandomForest '-I 10 -K 0 -S 1' 4216839470751428698
(3) rules.JRip '-F 3 -N 2.0 -O 2 -S 1' -6589312996832147161
```

Exercise

- Load the ionosphere dataset and prepare a 5 fold cross validation
- Perform the classification by using the three different classifiers and identify the most performing one
- Once selected the best classifier, perform the classification by using a metaclassifier with three different attribute selection methods
- Which is the best attribute selection method?
- Which are the most relevant selected attributes?