WEKA: Evaluation. Knowledge flow

Lab 4

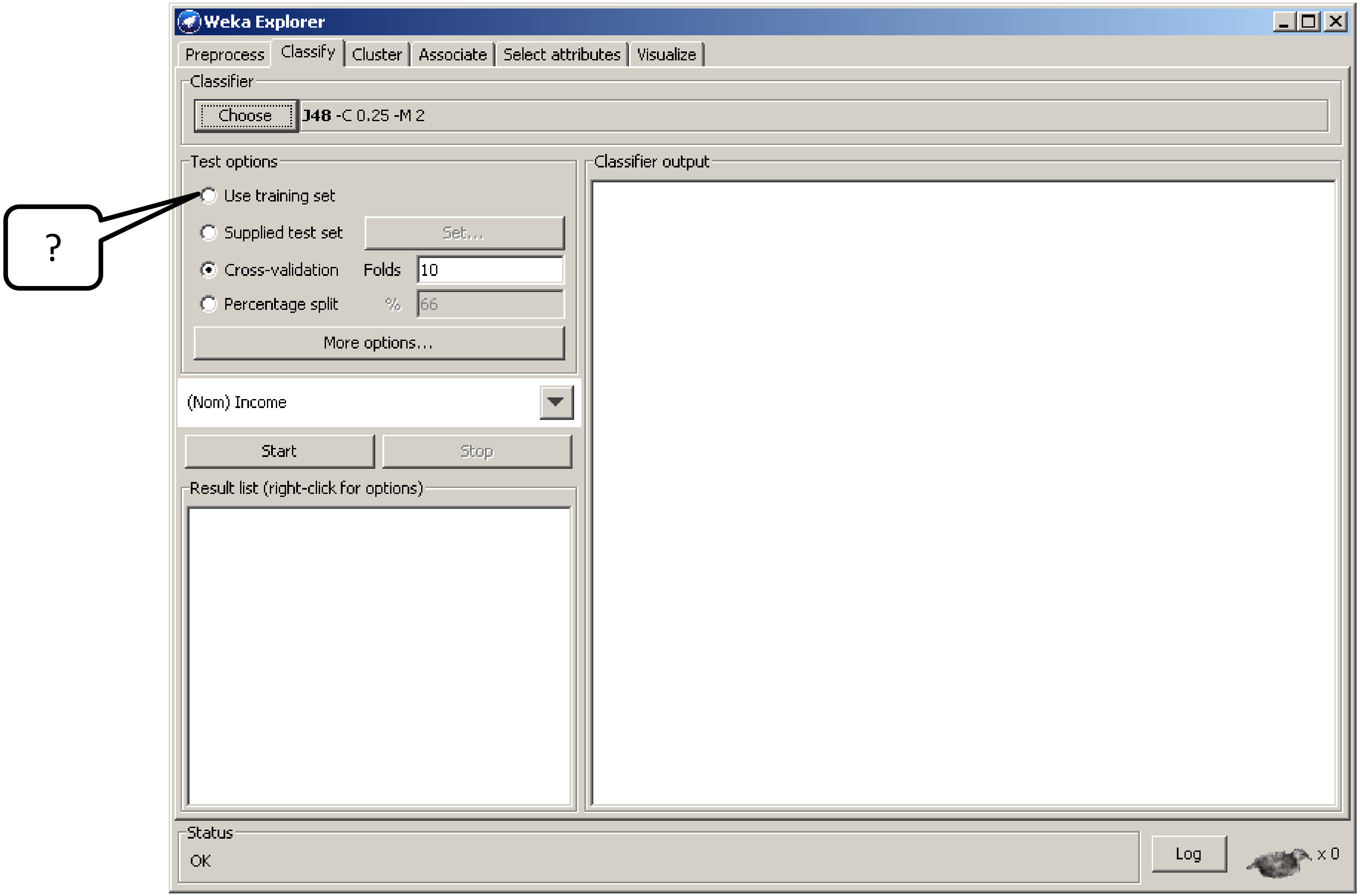
Lab outline

- Evaluation metrics in WEKA Explorer
- Knowledge flow interface
- Generating ROC curves in Knowledge flow interface

WEKA: evaluation metrics

- Open WEKA
- Open file “adult_income.arff”

Evaluation options



?

Classifier: J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set

● Cross-validation Folds 10

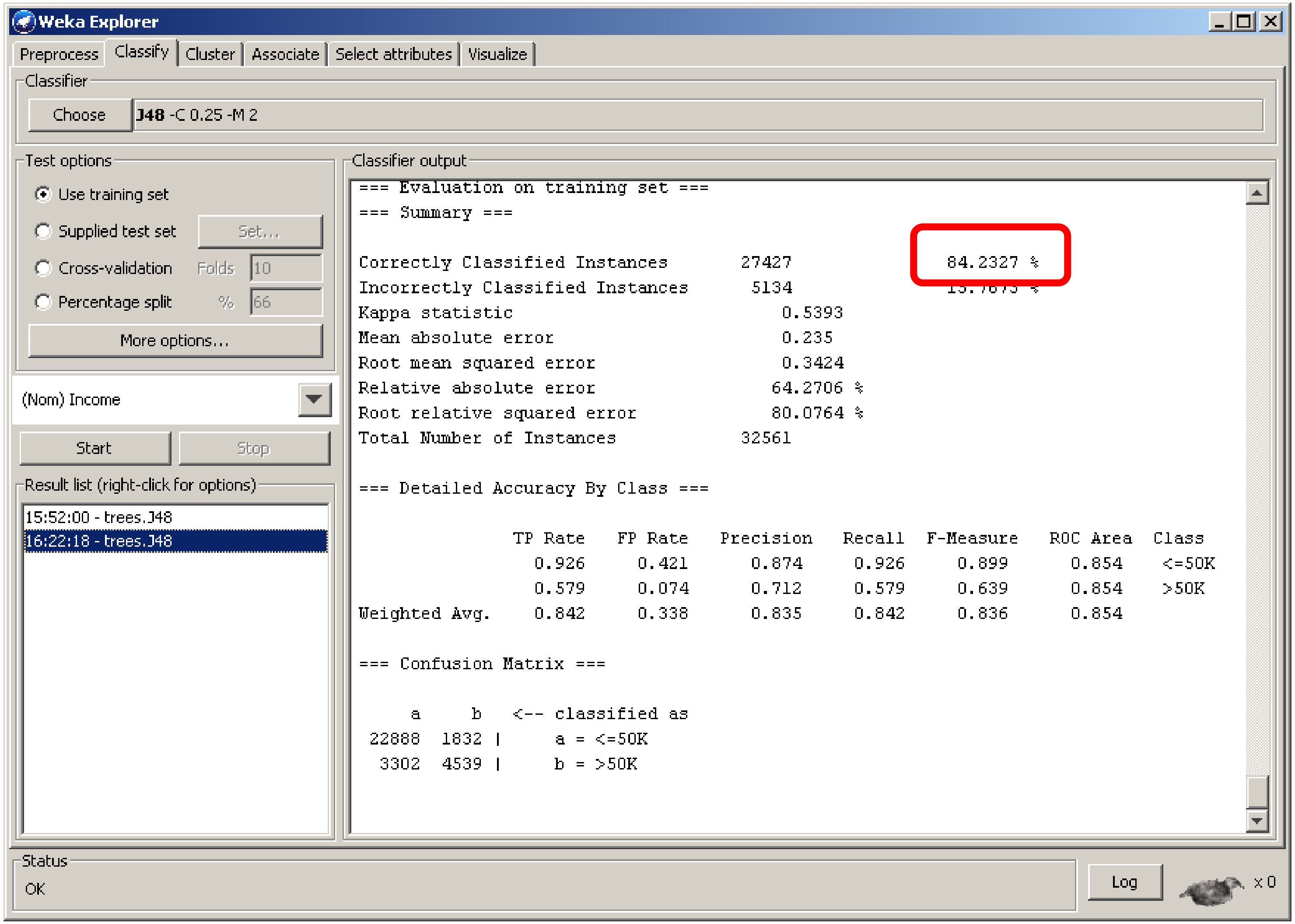
○ Percentage split % 66

(Nom) Income

Status

OK

The best possible accuracy



Classifier: J48 -C 0.25 -M 2

Test options:
- ● Use training set
- ○ Supplied test set
- ○ Cross-validation Folds 10
- ○ Percentage split % 66

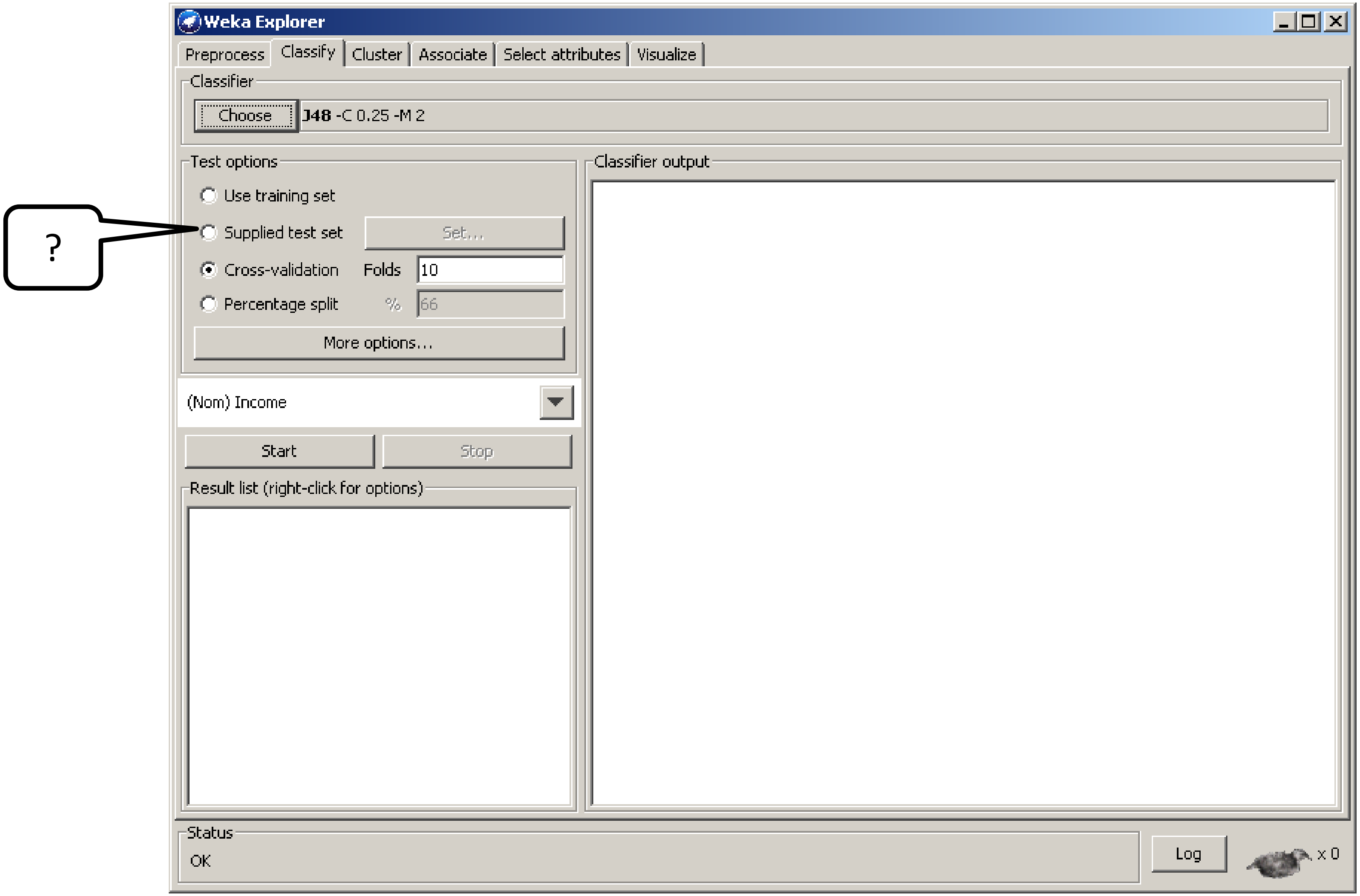
(Nom) Income

Result list:
- 15:52:00 - trees.J48
- 16:22:18 - trees.J48

```
=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances       27427               84.2327 %
Incorrectly Classified Instances      5134               15.7673 %
Kappa statistic                          0.5393
Mean absolute error                      0.235
Root mean squared error                  0.3424
Relative absolute error                 64.2706 %
Root relative squared error             80.0764 %
Total Number of Instances            32561

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.926     0.421      0.874     0.926     0.899      0.854    <=50K
                 0.579     0.074      0.712     0.579     0.639      0.854    >50K
Weighted Avg.    0.842     0.338      0.835     0.842     0.836      0.854

=== Confusion Matrix ===

     a     b   <-- classified as
 22888  1832 |     a = <=50K
  3302  4539 |     b = >50K
```

Status: OK

Evaluation options



Classifier

J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set

● Cross-validation Folds 10

○ Percentage split % 66

(Nom) Income

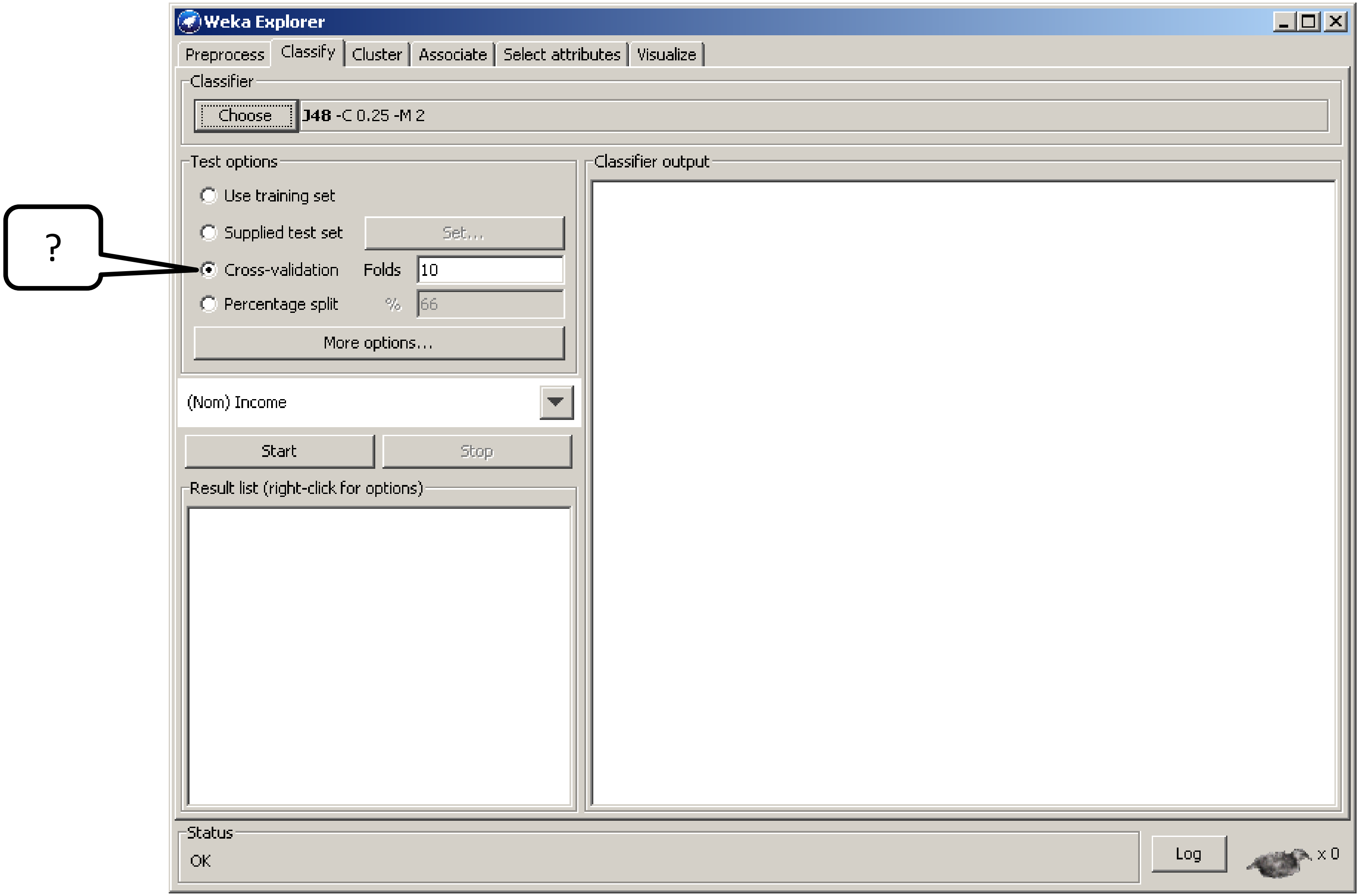
Result list (right-click for options)

Classifier output

Status

OK

Evaluation options



?

Weka Explorer

Classifier: J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set

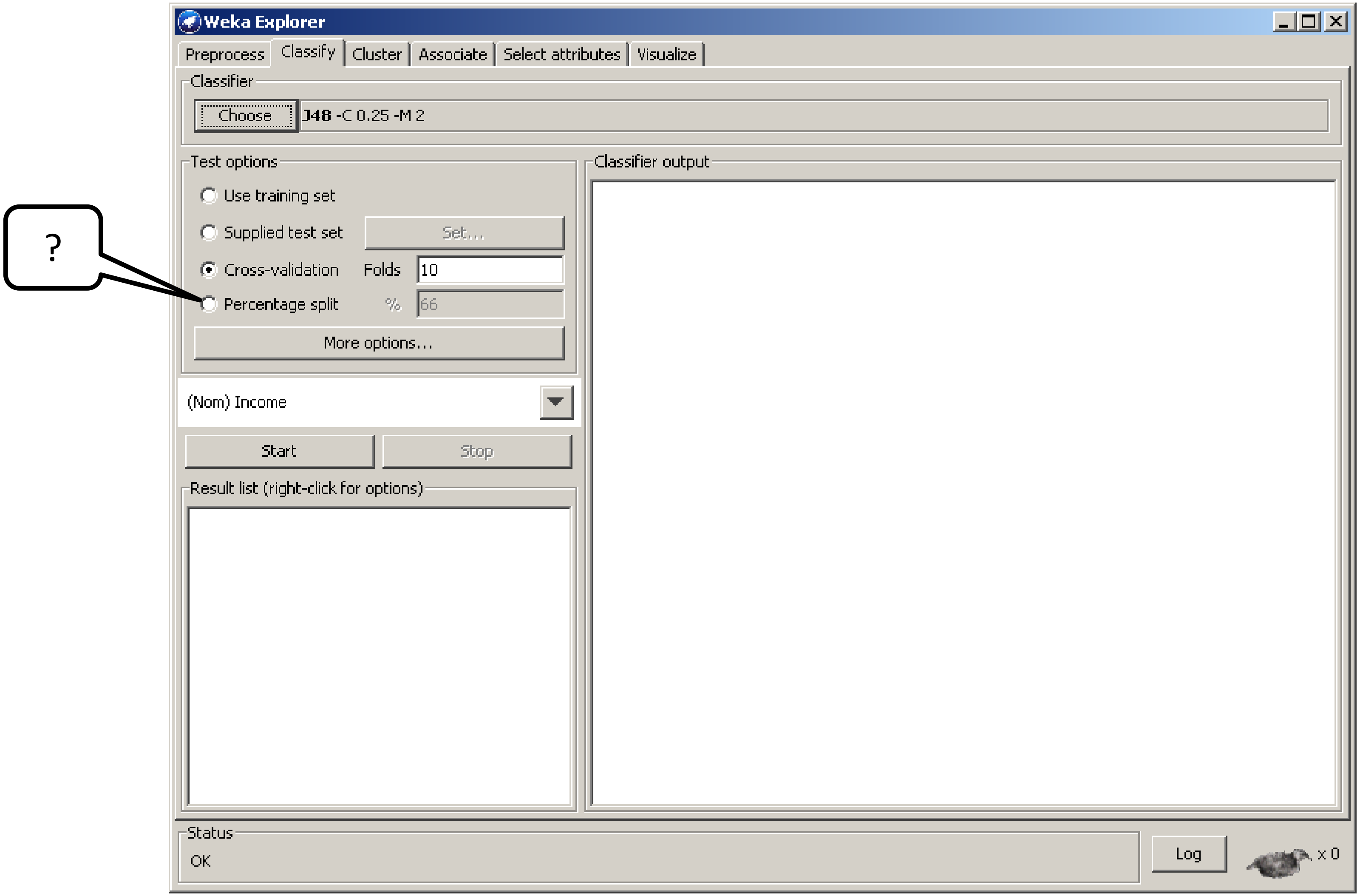
● Cross-validation Folds 10

○ Percentage split % 66

(Nom) Income

Status: OK

Evaluation options



?

Weka Explorer

Classifier

J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set

● Cross-validation Folds 10

○ Percentage split % 66

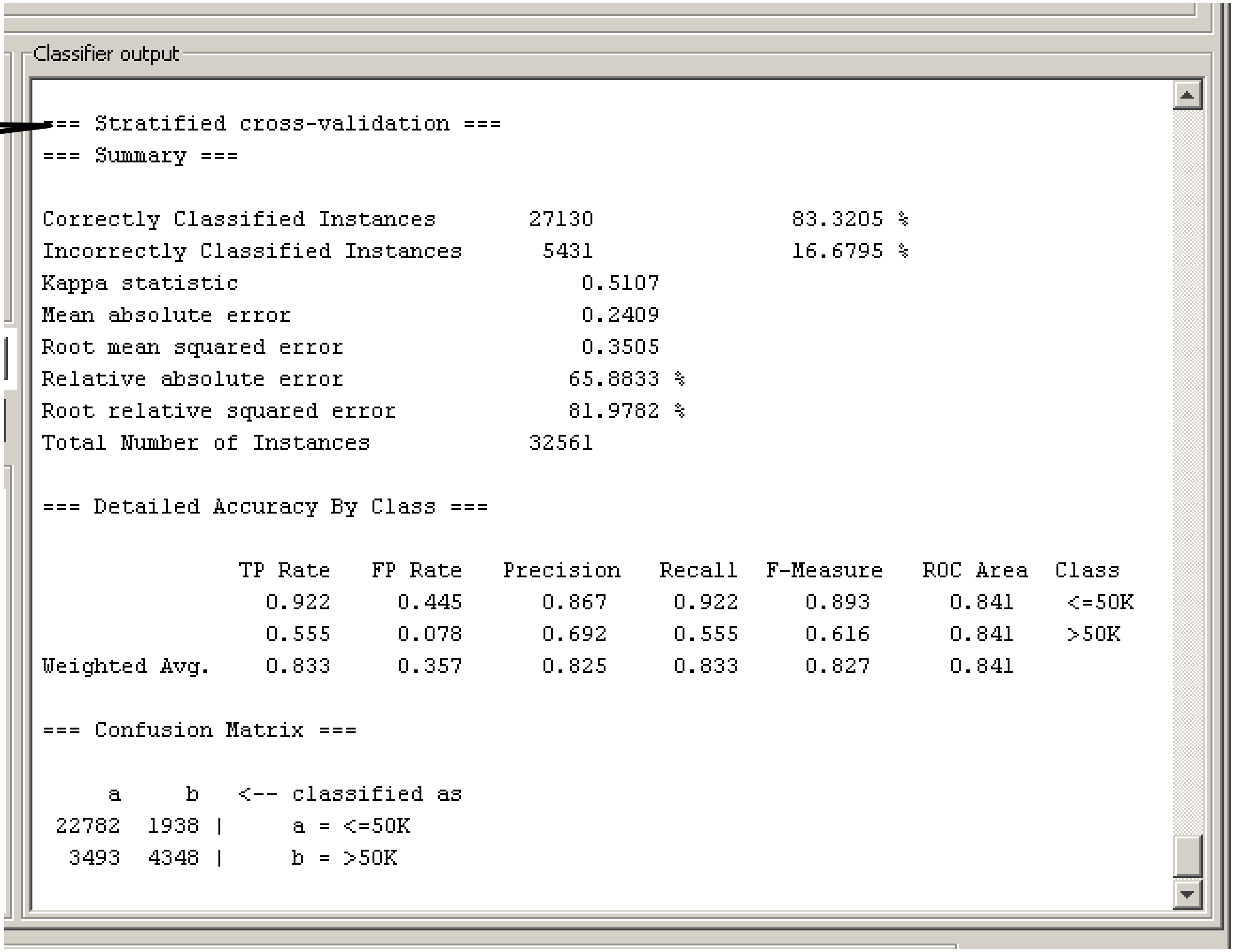
(Nom) Income

Status

OK

Build classifier: output

?

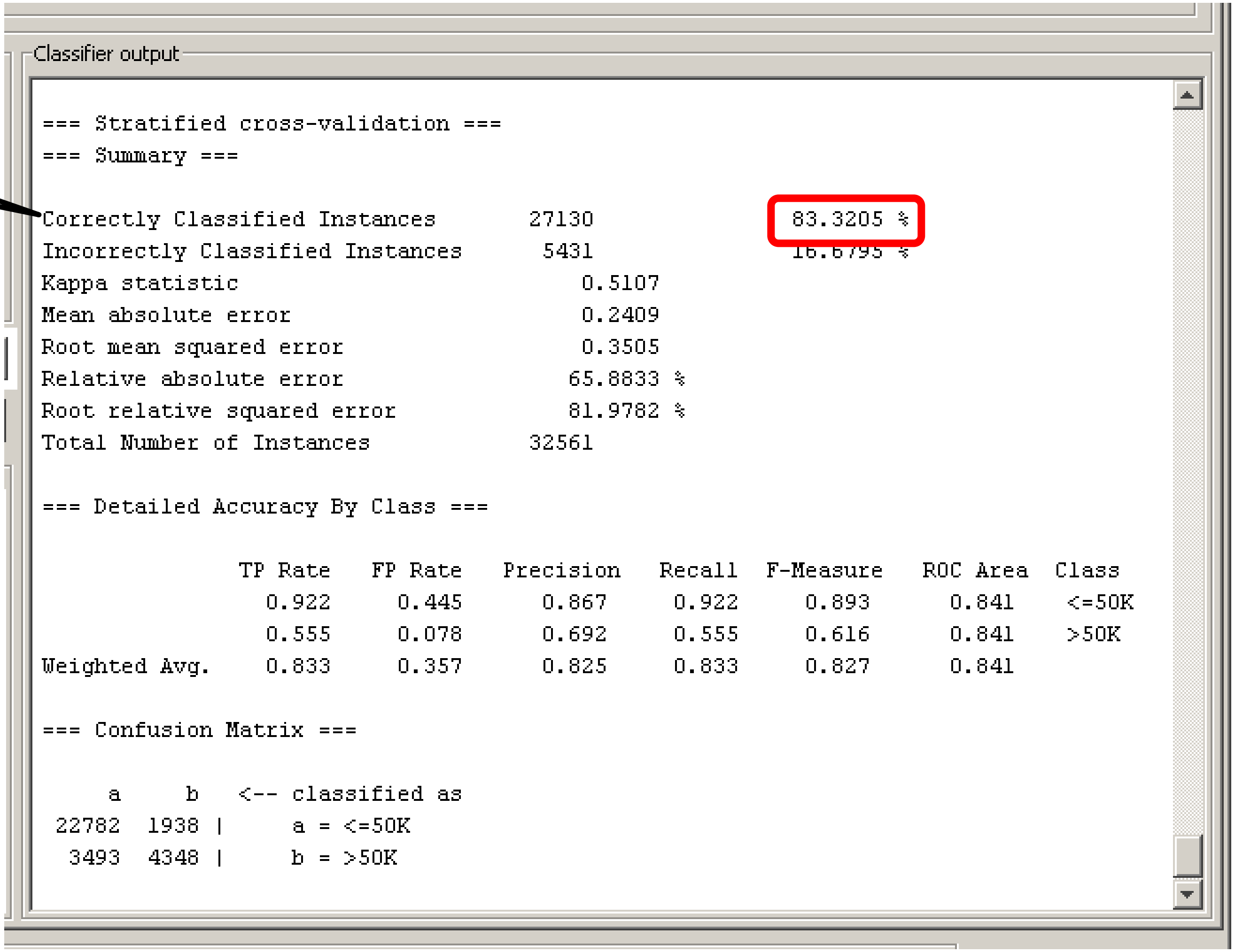


Classifier output

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       27130               83.3205 %
Incorrectly Classified Instances      5431               16.6795 %
Kappa statistic                          0.5107
Mean absolute error                      0.2409
Root mean squared error                  0.3505
Relative absolute error                 65.8833 %
Root relative squared error             81.9782 %
Total Number of Instances            32561

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.922     0.445      0.867     0.922     0.893      0.841    <=50K
                 0.555     0.078      0.692     0.555     0.616      0.841    >50K
Weighted Avg.    0.833     0.357      0.825     0.833     0.827      0.841

=== Confusion Matrix ===

     a     b   <-- classified as
 22782  1938 |     a = <=50K
  3493  4348 |     b = >50K
```

Build classifier: output

?

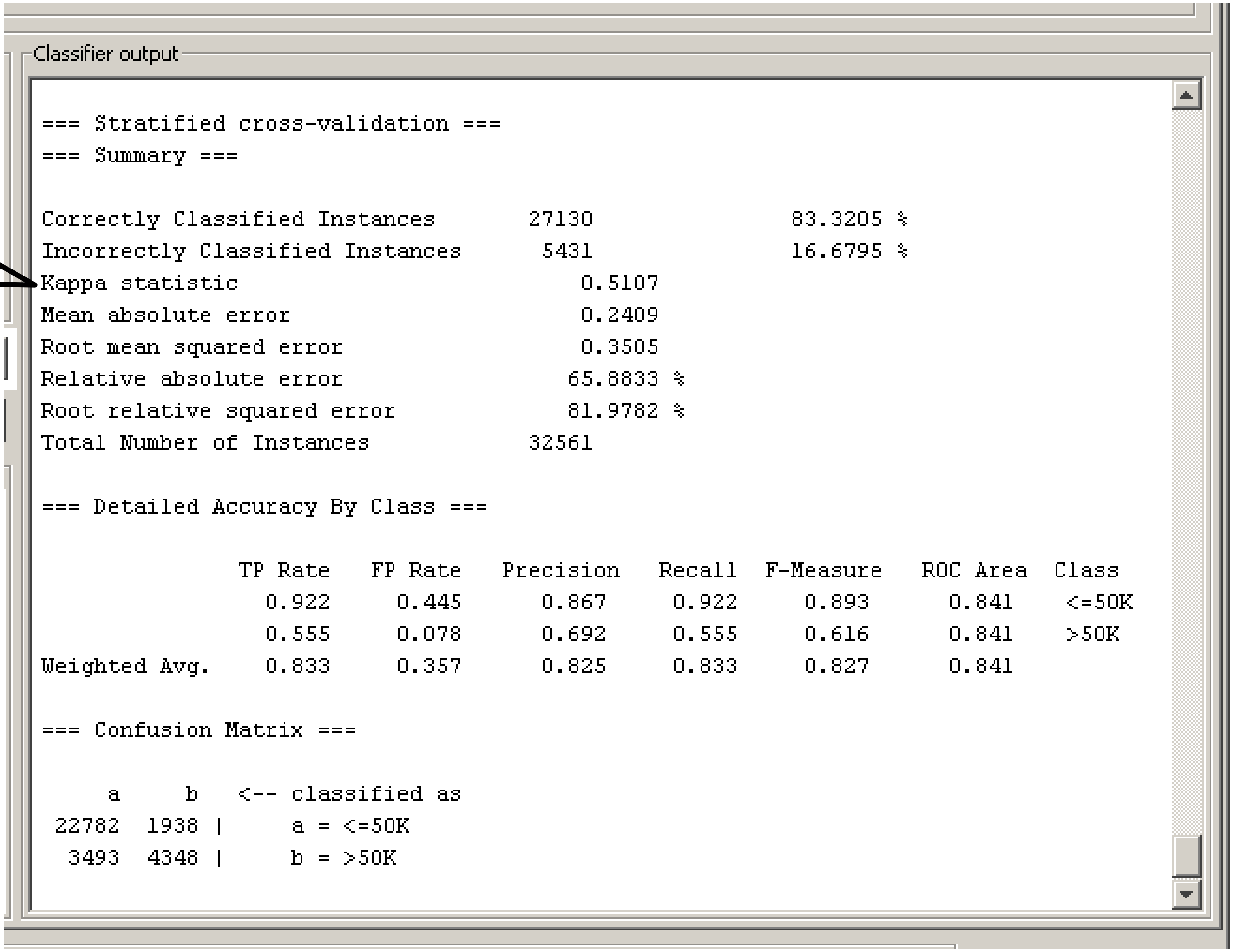


Classifier output

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       27130               83.3205 %
Incorrectly Classified Instances      5431               16.6795 %
Kappa statistic                          0.5107
Mean absolute error                      0.2409
Root mean squared error                  0.3505
Relative absolute error                 65.8833 %
Root relative squared error             81.9782 %
Total Number of Instances            32561

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.922     0.445      0.867     0.922     0.893      0.841    <=50K
                 0.555     0.078      0.692     0.555     0.616      0.841    >50K
Weighted Avg.    0.833     0.357      0.825     0.833     0.827      0.841

=== Confusion Matrix ===

     a     b   <-- classified as
 22782  1938 |     a = <=50K
  3493  4348 |     b = >50K
```

Vs. 84.23%

Build classifier: output

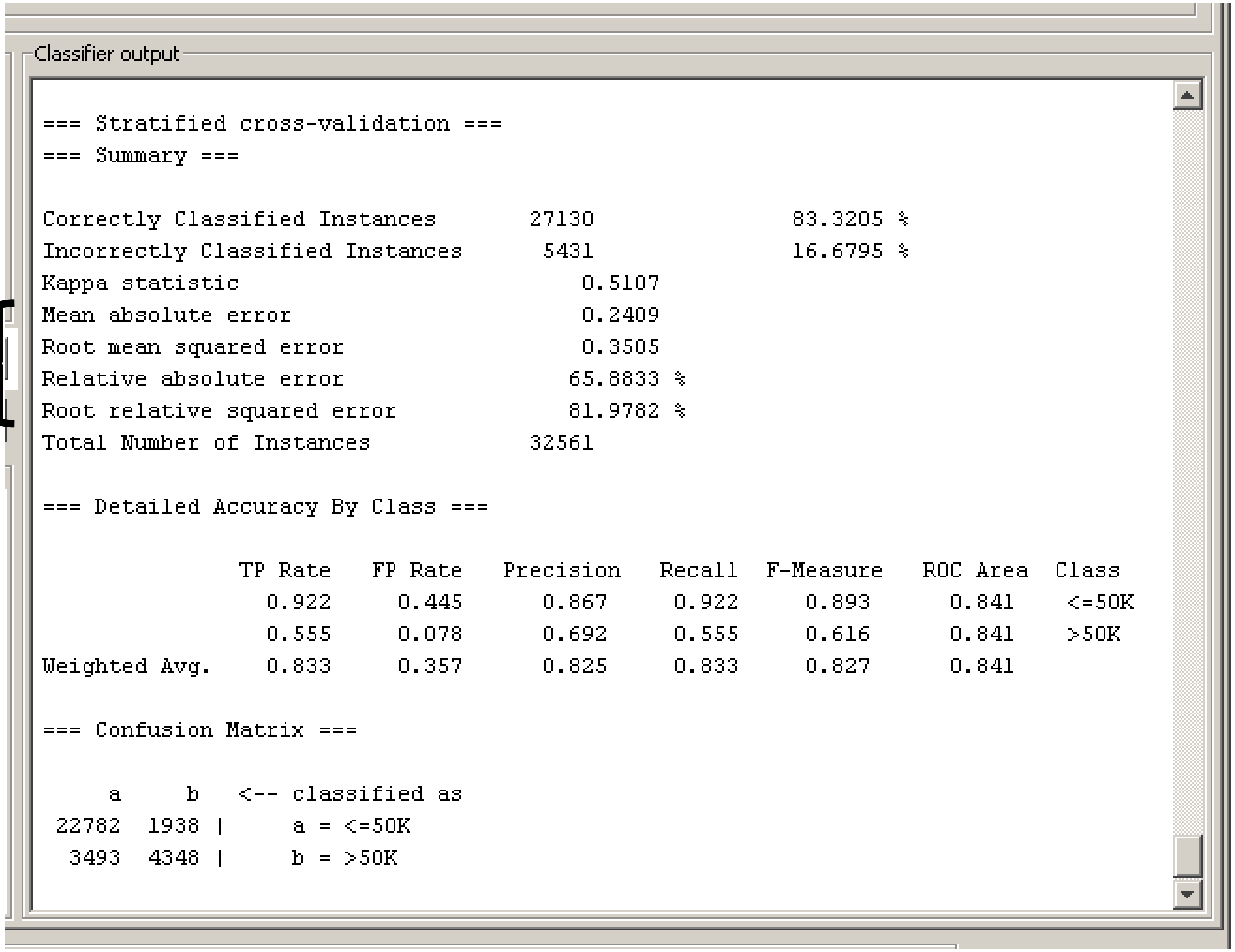


```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       27130               83.3205 %
Incorrectly Classified Instances      5431               16.6795 %
Kappa statistic                          0.5107
Mean absolute error                      0.2409
Root mean squared error                  0.3505
Relative absolute error                 65.8833 %
Root relative squared error             81.9782 %
Total Number of Instances            32561

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.922     0.445      0.867     0.922     0.893      0.841    <=50K
                 0.555     0.078      0.692     0.555     0.616      0.841    >50K
Weighted Avg.    0.833     0.357      0.825     0.833     0.827      0.841

=== Confusion Matrix ===

     a     b   <-- classified as
 22782  1938 |     a = <=50K
  3493  4348 |     b = >50K
```

Your prediction is better than random prediction by 51%

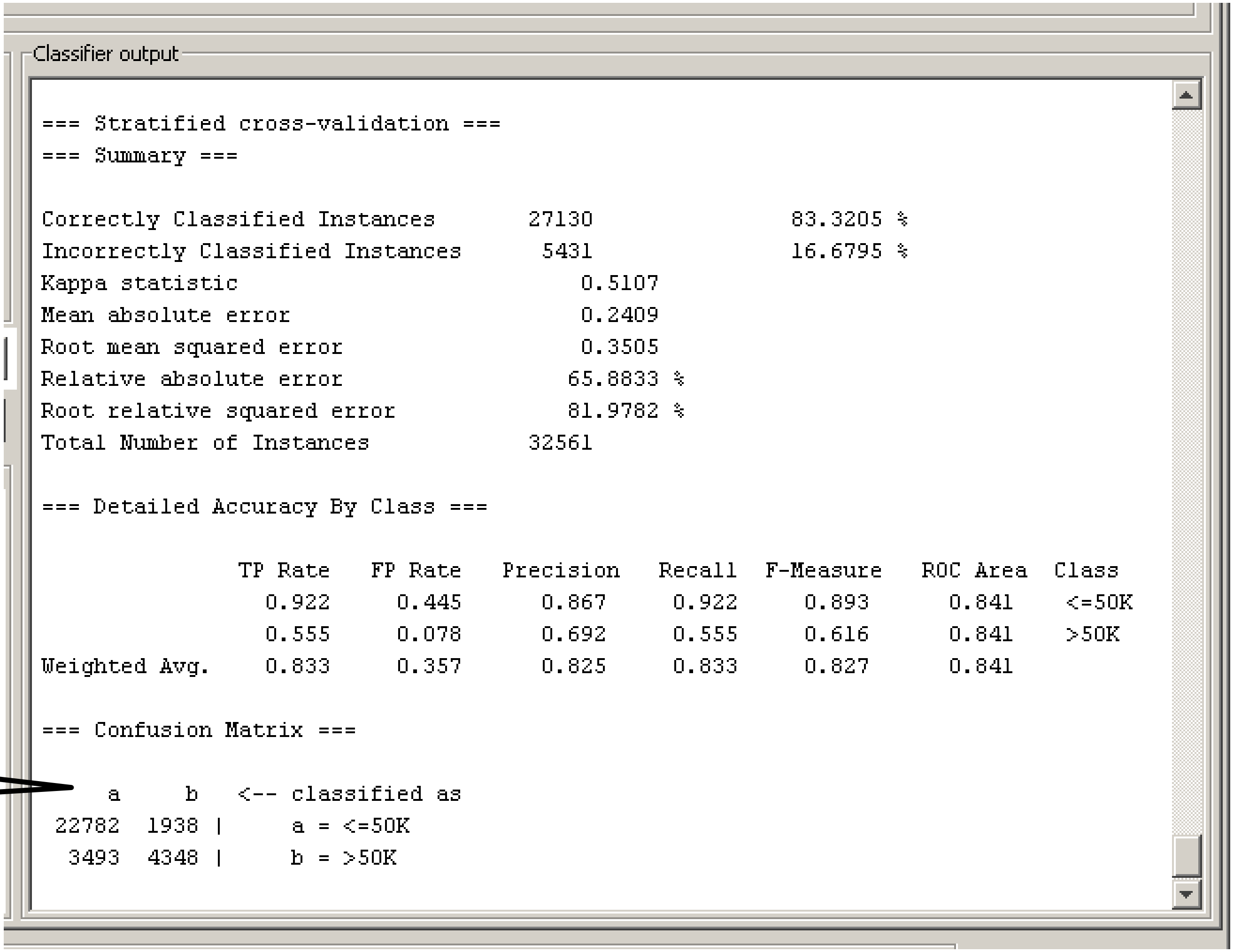
Build classifier: output

Some per/instance metrics



```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       27130               83.3205 %
Incorrectly Classified Instances      5431               16.6795 %
Kappa statistic                          0.5107
Mean absolute error                      0.2409
Root mean squared error                  0.3505
Relative absolute error                 65.8833 %
Root relative squared error             81.9782 %
Total Number of Instances            32561

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.922     0.445      0.867     0.922     0.893      0.841    <=50K
                 0.555     0.078      0.692     0.555     0.616      0.841    >50K
Weighted Avg.    0.833     0.357      0.825     0.833     0.827      0.841

=== Confusion Matrix ===

     a     b   <-- classified as
 22782  1938 |     a = <=50K
  3493  4348 |     b = >50K
```

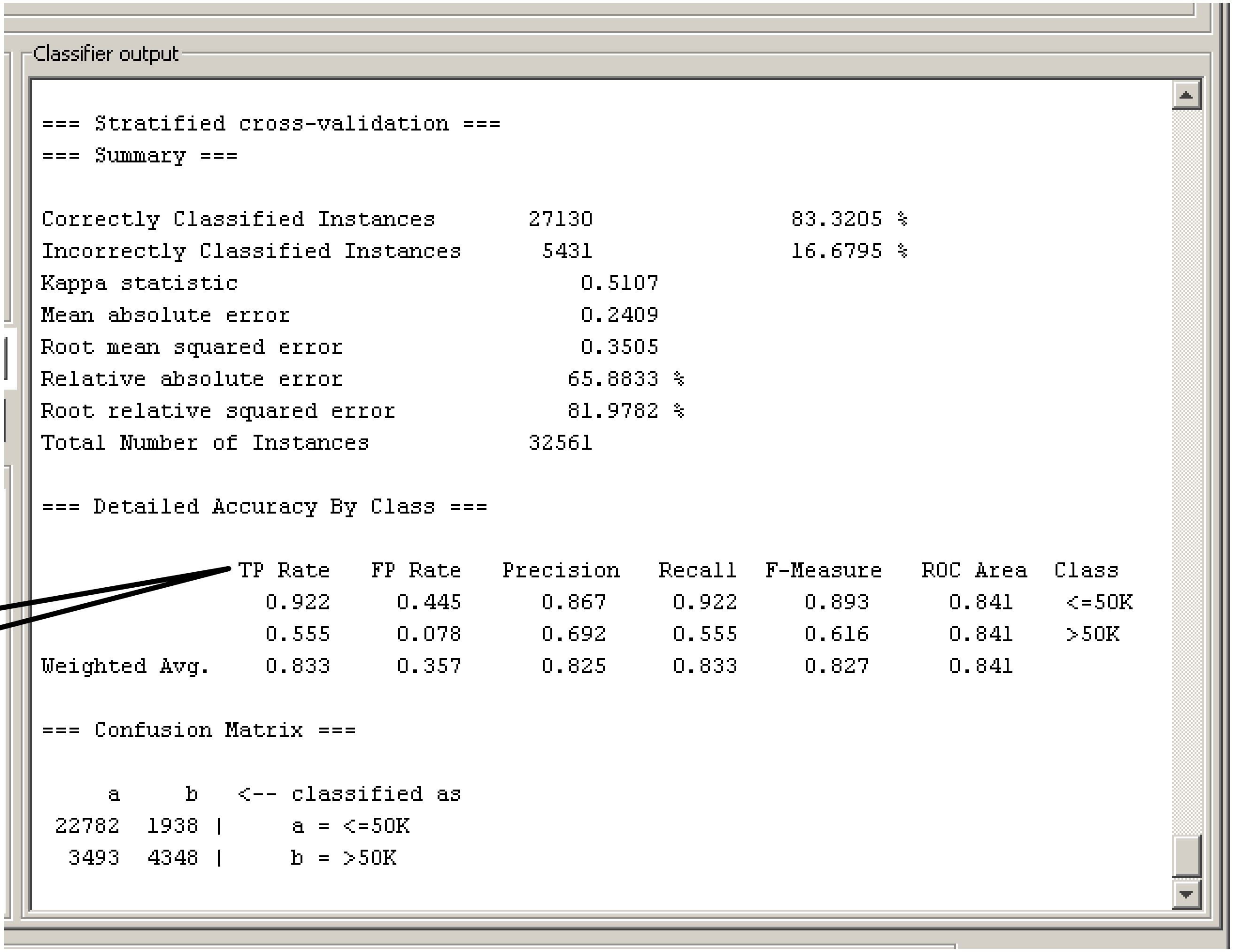
Build classifier: output



```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       27130               83.3205 %
Incorrectly Classified Instances      5431               16.6795 %
Kappa statistic                          0.5107
Mean absolute error                      0.2409
Root mean squared error                  0.3505
Relative absolute error                 65.8833 %
Root relative squared error             81.9782 %
Total Number of Instances            32561

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.922     0.445      0.867     0.922     0.893      0.841    <=50K
                 0.555     0.078      0.692     0.555     0.616      0.841    >50K
Weighted Avg.    0.833     0.357      0.825     0.833     0.827      0.841

=== Confusion Matrix ===

     a     b   <-- classified as
 22782  1938 |     a = <=50K
  3493  4348 |     b = >50K
```

?

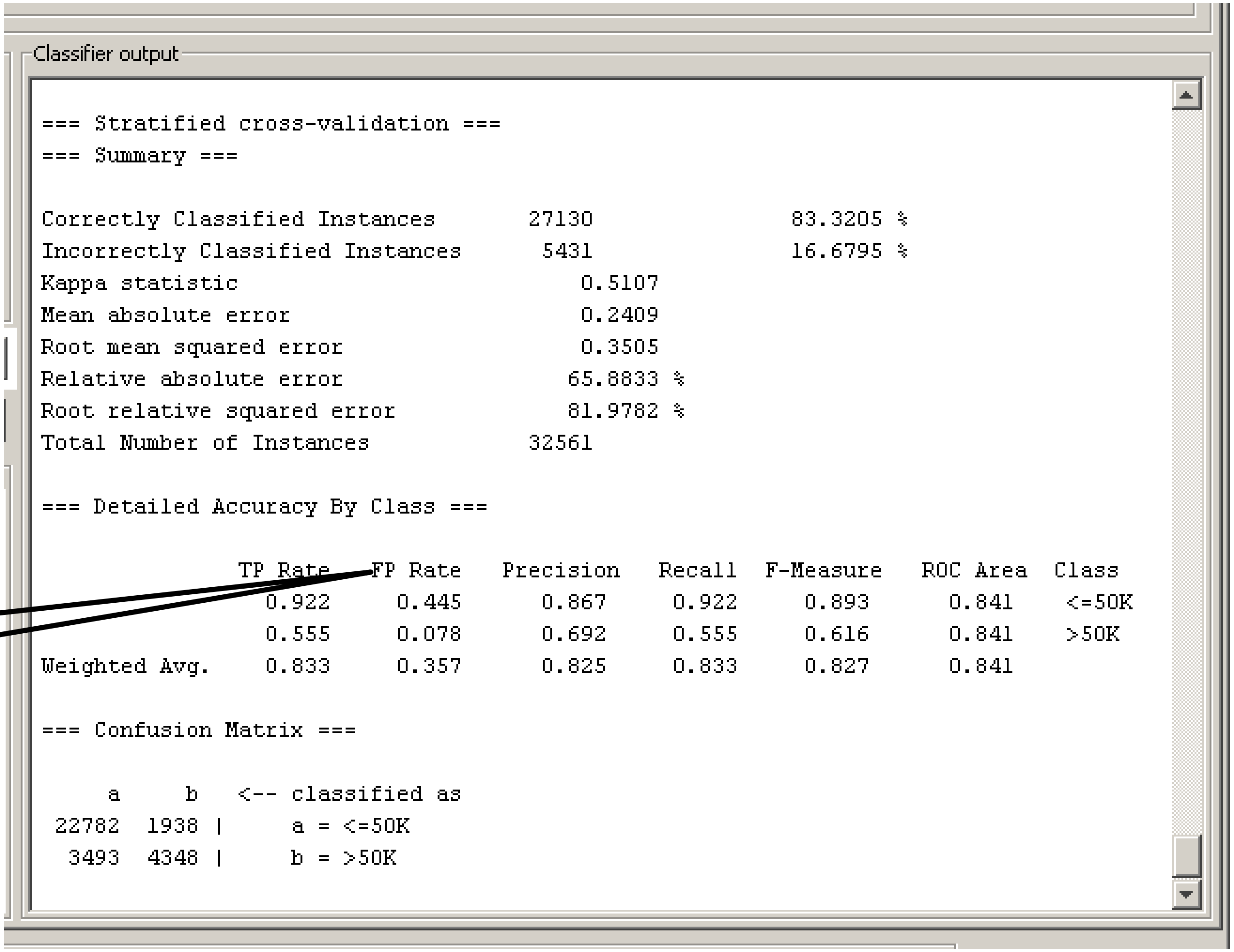
Build classifier: output



```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       27130               83.3205 %
Incorrectly Classified Instances      5431               16.6795 %
Kappa statistic                          0.5107
Mean absolute error                      0.2409
Root mean squared error                  0.3505
Relative absolute error                 65.8833 %
Root relative squared error             81.9782 %
Total Number of Instances            32561

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.922     0.445      0.867     0.922     0.893      0.841    <=50K
                 0.555     0.078      0.692     0.555     0.616      0.841    >50K
Weighted Avg.    0.833     0.357      0.825     0.833     0.827      0.841

=== Confusion Matrix ===

     a     b   <-- classified as
 22782  1938 |     a = <=50K
  3493  4348 |     b = >50K
```

TPos/Pos

Build classifier: output

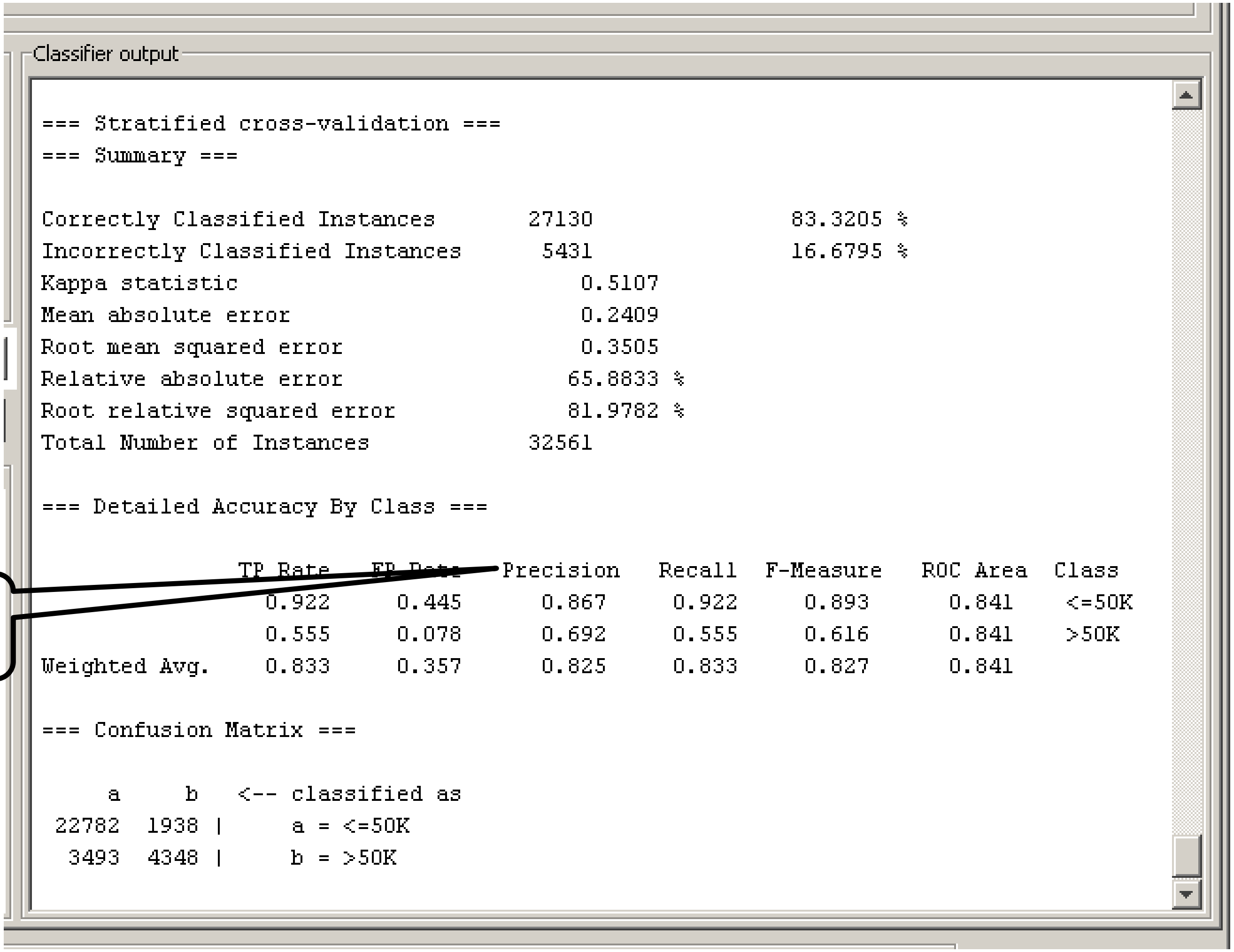


Classifier output

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       27130               83.3205 %
Incorrectly Classified Instances      5431               16.6795 %
Kappa statistic                          0.5107
Mean absolute error                      0.2409
Root mean squared error                  0.3505
Relative absolute error                 65.8833 %
Root relative squared error             81.9782 %
Total Number of Instances            32561

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.922     0.445      0.867     0.922     0.893      0.841    <=50K
                 0.555     0.078      0.692     0.555     0.616      0.841    >50K
Weighted Avg.    0.833     0.357      0.825     0.833     0.827      0.841

=== Confusion Matrix ===

     a     b   <-- classified as
 22782  1938 |     a = <=50K
  3493  4348 |     b = >50K
```

FPos/Neg

Build classifier: output

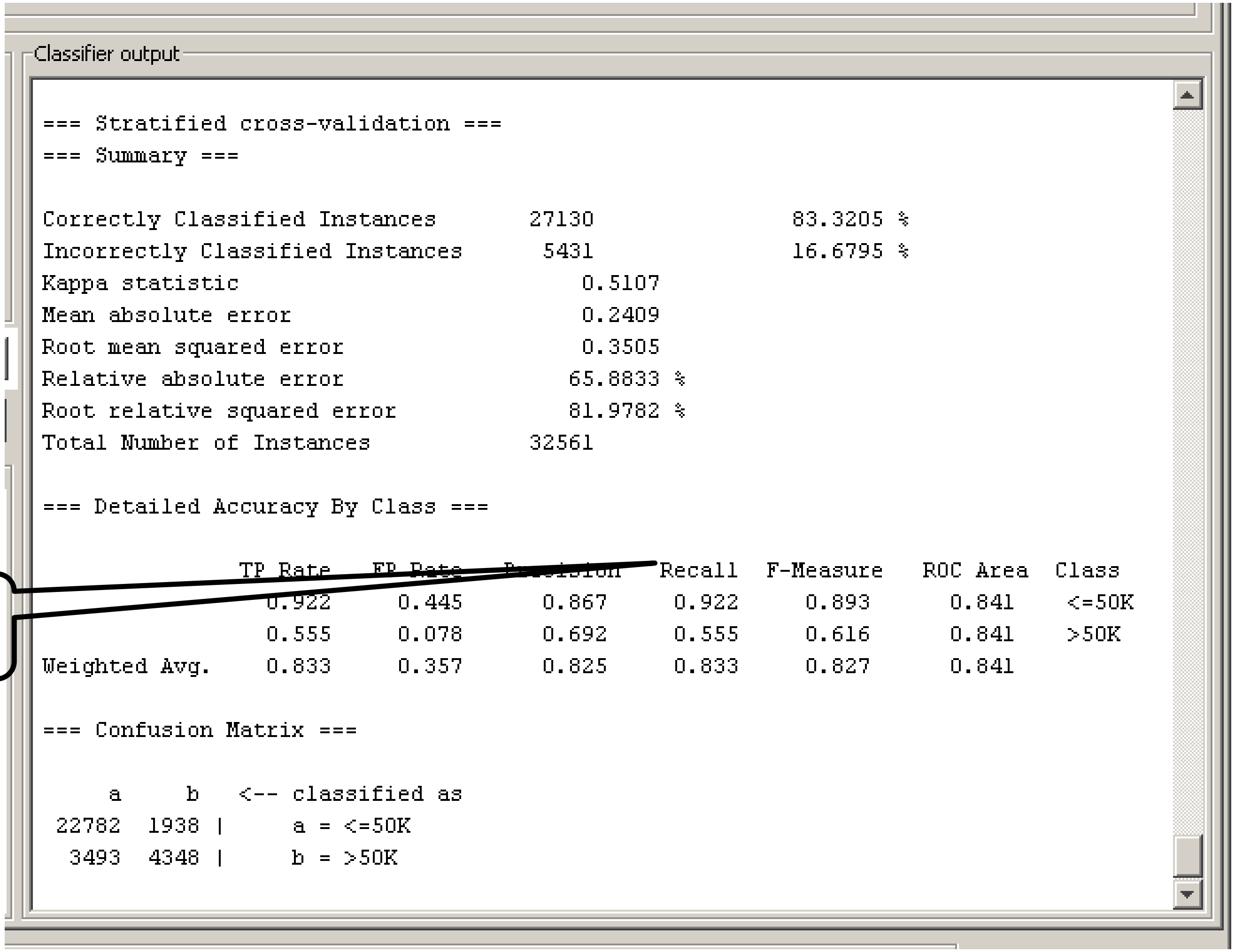


Classifier output

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       27130               83.3205 %
Incorrectly Classified Instances      5431               16.6795 %
Kappa statistic                          0.5107
Mean absolute error                      0.2409
Root mean squared error                  0.3505
Relative absolute error                 65.8833 %
Root relative squared error             81.9782 %
Total Number of Instances            32561

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.922     0.445      0.867     0.922     0.893      0.841    <=50K
                 0.555     0.078      0.692     0.555     0.616      0.841    >50K
Weighted Avg.    0.833     0.357      0.825     0.833     0.827      0.841

=== Confusion Matrix ===

     a     b   <-- classified as
 22782  1938 |     a = <=50K
  3493  4348 |     b = >50K
```

TPos/(Tpos+Fpos)

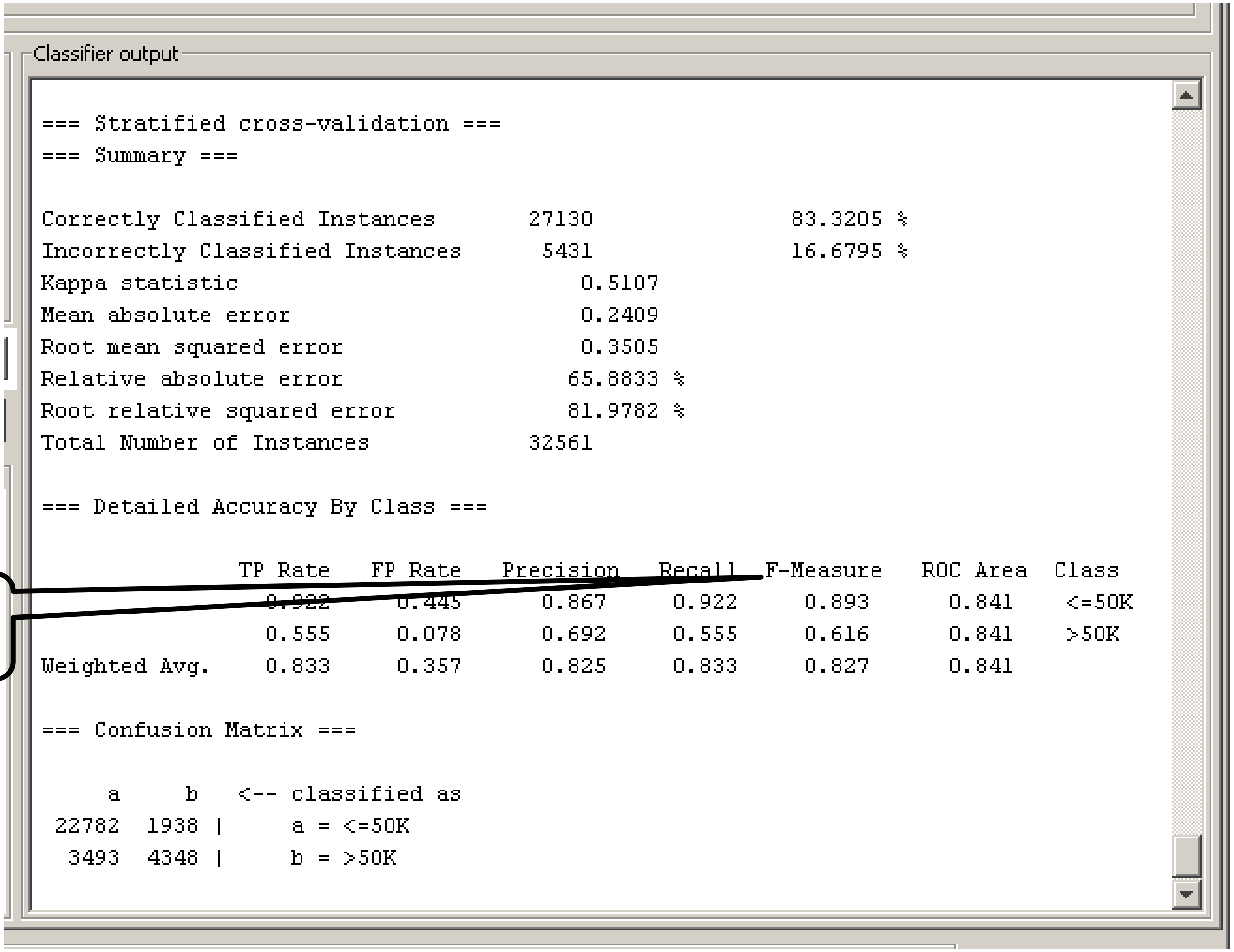
Build classifier: output



```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       27130               83.3205 %
Incorrectly Classified Instances      5431               16.6795 %
Kappa statistic                          0.5107
Mean absolute error                      0.2409
Root mean squared error                  0.3505
Relative absolute error                 65.8833 %
Root relative squared error             81.9782 %
Total Number of Instances            32561

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.922     0.445     0.867     0.922     0.893      0.841    <=50K
                 0.555     0.078     0.692     0.555     0.616      0.841    >50K
Weighted Avg.    0.833     0.357     0.825     0.833     0.827      0.841

=== Confusion Matrix ===

     a     b   <-- classified as
 22782  1938 |     a = <=50K
  3493  4348 |     b = >50K
```

TP Rate

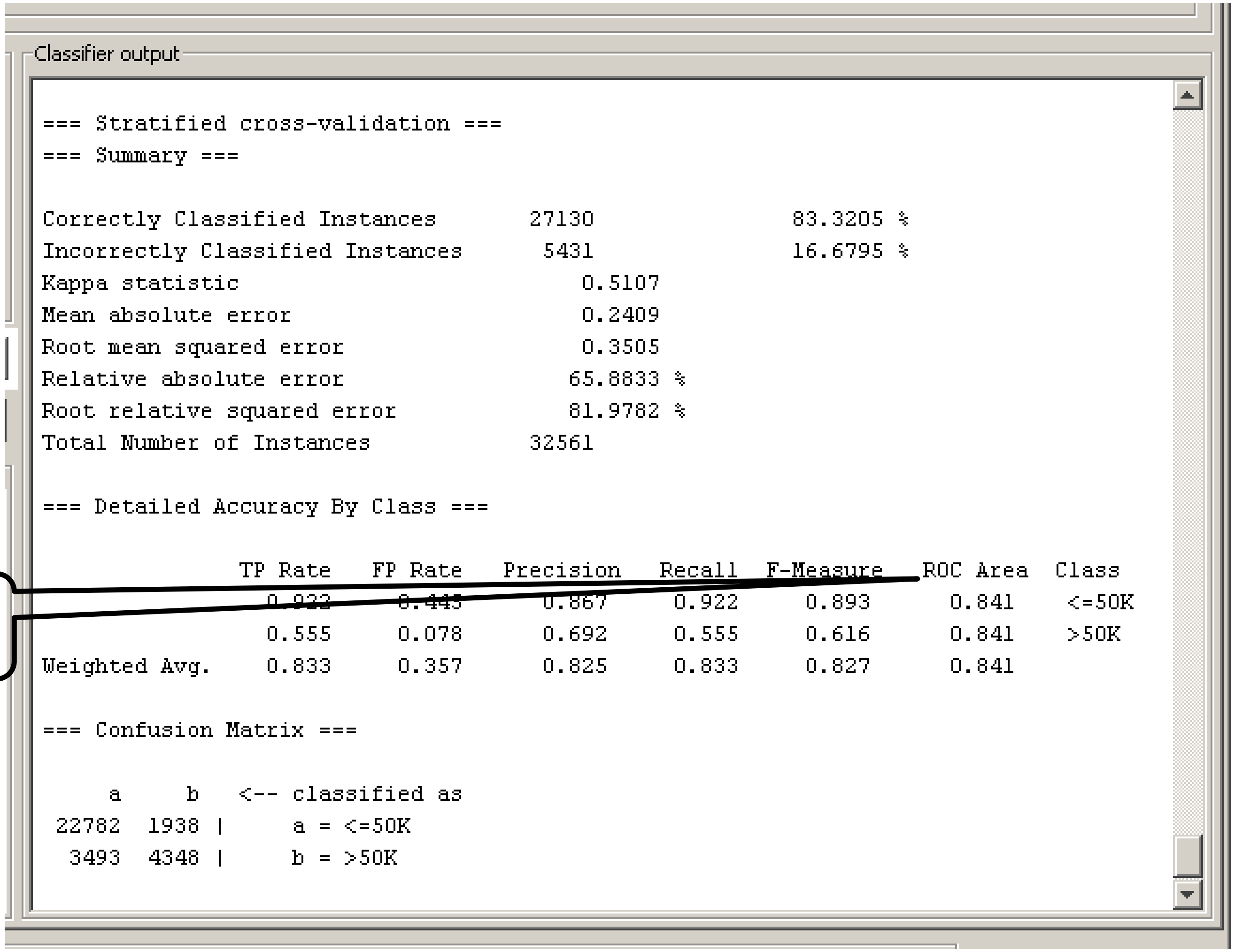
Build classifier: output



```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       27130               83.3205 %
Incorrectly Classified Instances      5431               16.6795 %
Kappa statistic                          0.5107
Mean absolute error                      0.2409
Root mean squared error                  0.3505
Relative absolute error                 65.8833 %
Root relative squared error             81.9782 %
Total Number of Instances            32561

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.922     0.445      0.867     0.922     0.893      0.841    <=50K
                 0.555     0.078      0.692     0.555     0.616      0.841    >50K
Weighted Avg.    0.833     0.357      0.825     0.833     0.827      0.841

=== Confusion Matrix ===

     a     b   <-- classified as
 22782  1938 |     a = <=50K
  3493  4348 |     b = >50K
```

$$\frac{2*precision*recall}{precision + recall}$$

Build classifier: output



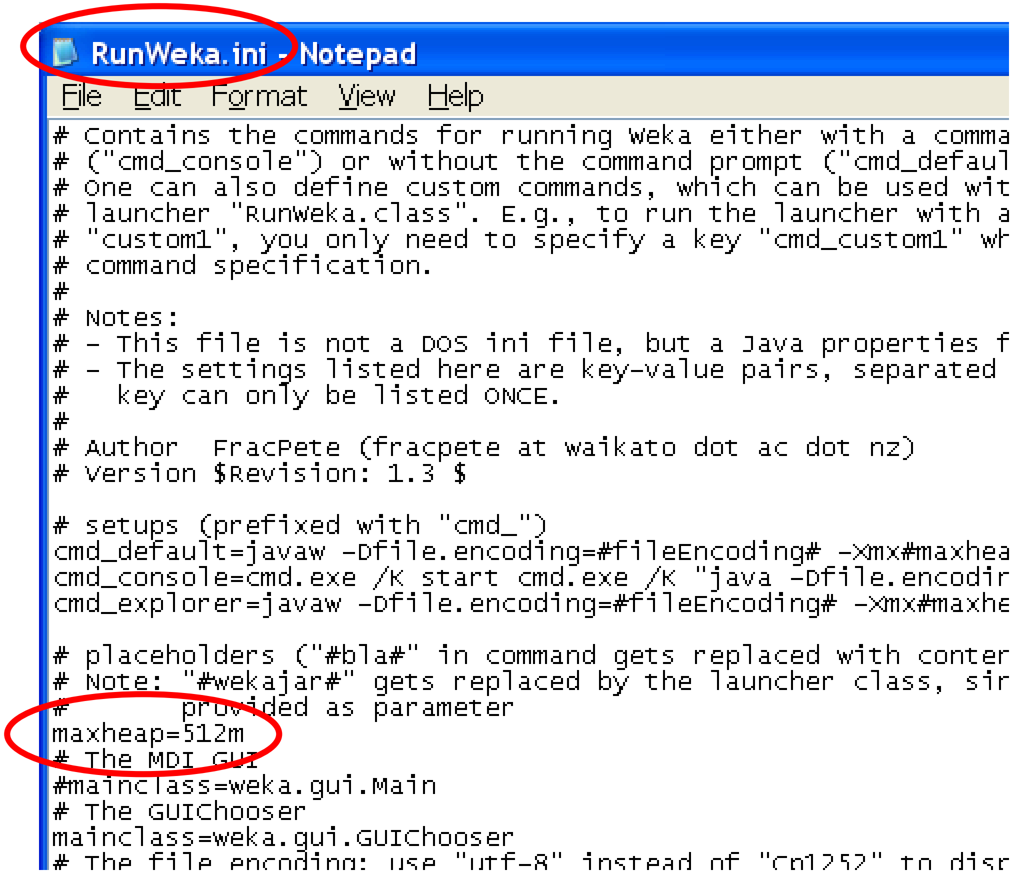
Classifier output

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       27130               83.3205 %
Incorrectly Classified Instances      5431               16.6795 %
Kappa statistic                          0.5107
Mean absolute error                      0.2409
Root mean squared error                  0.3505
Relative absolute error                 65.8833 %
Root relative squared error             81.9782 %
Total Number of Instances            32561

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.922     0.445      0.867     0.922     0.893      0.841    <=50K
                 0.555     0.078      0.692     0.555     0.616      0.841    >50K
Weighted Avg.    0.833     0.357      0.825     0.833     0.827      0.841

=== Confusion Matrix ===

     a     b   <-- classified as
 22782  1938 |     a = <=50K
  3493  4348 |     b = >50K
```

Area under the ROC curve

WEKA: dealing with large datasets

- Increase java heap space
- Still might get “Out of memory” exception



```
# Contains the commands for running Weka either with a comma
# ("cmd_console") or without the command prompt ("cmd_defaul
# One can also define custom commands, which can be used wit
# launcher "RunWeka.class". E.g., to run the launcher with a
# "custom1", you only need to specify a key "cmd_custom1" wh
# command specification.
#
# Notes:
# - This file is not a DOS ini file, but a Java properties f
# - The settings listed here are key-value pairs, separated
#   key can only be listed ONCE.
#
# Author  FracPete (fracpete at waikato dot ac dot nz)
# Version $Revision: 1.3 $

# setups (prefixed with "cmd_")
cmd_default=javaw -Dfile.encoding=#fileEncoding# -Xmx#maxhea
cmd_console=cmd.exe /K start cmd.exe /K "java -Dfile.encodir
cmd_explorer=javaw -Dfile.encoding=#fileEncoding# -Xmx#maxhe

# placeholders ("#bla#" in command gets replaced with conter
# Note: "#wekajar#" gets replaced by the launcher class, sir
#       provided as parameter
maxheap=512m
# The MDI GUI
#mainclass=weka.gui.Main
# The GUIChooser
mainclass=weka.gui.GUIChooser
# The file encoding; use "utf-8" instead of "Cp1252" to disr
```

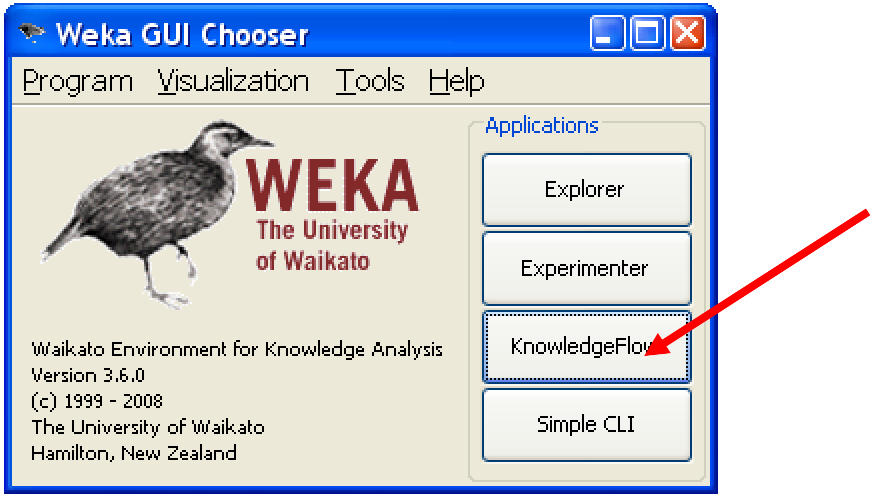
GUI I: WEKA Explorer and CLI

- Everything is in main memory: dataset, filter, model
- No large-scale data mining

GUI II. WEKA Knowledge Flow

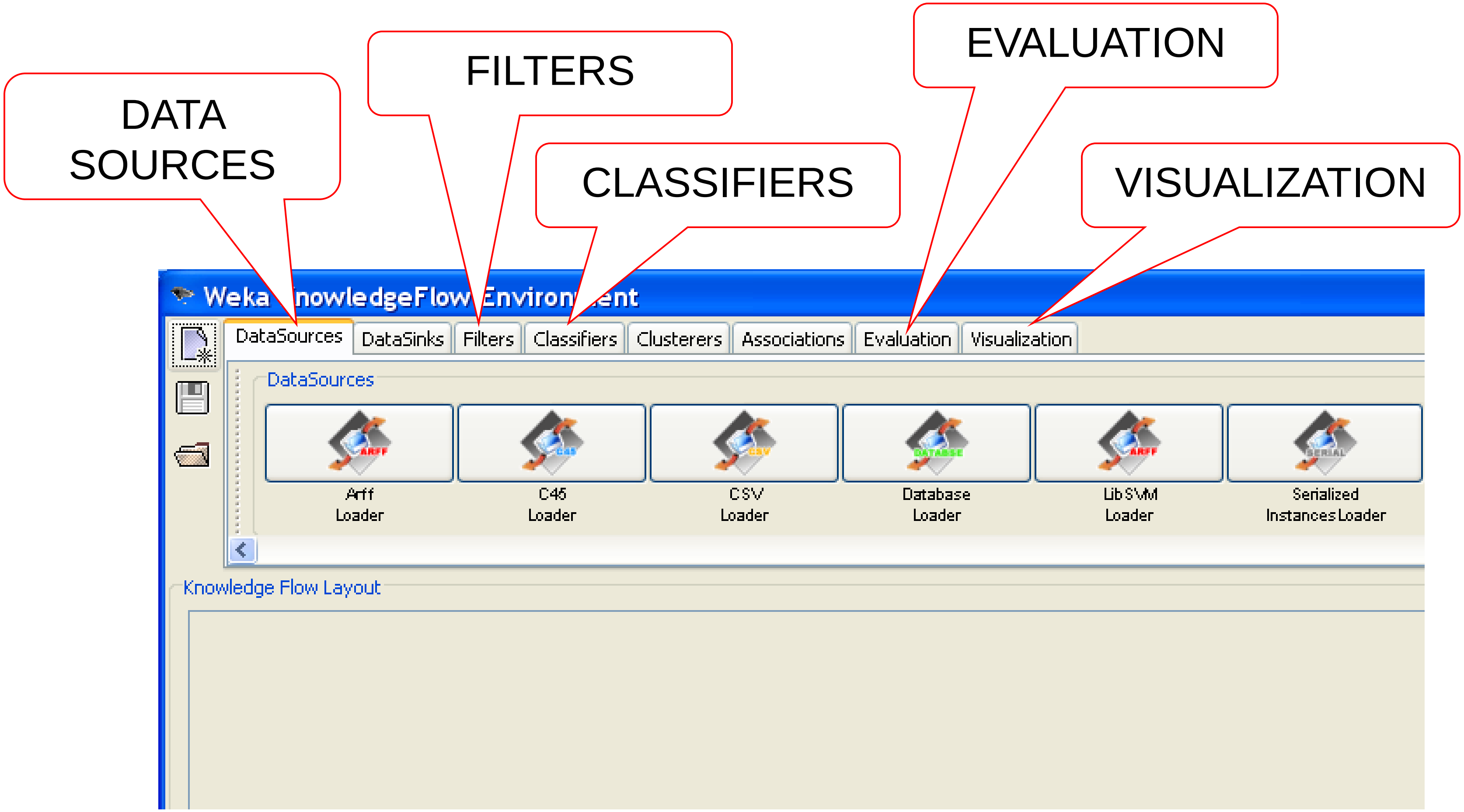
- Design configuration for streamed data processing
- Specify data stream and run algorithms which stream data from one component to another
- If the algorithm allows incremental filtering and learning, data will be loaded sequentially from disk

Comparing classifiers. Knowledge flow



Waikato Environment for Knowledge Analysis
Version 3.6.0
(c) 1999 - 2008
The University of Waikato
Hamilton, New Zealand

Knowledge flow tabs



DATA SOURCES

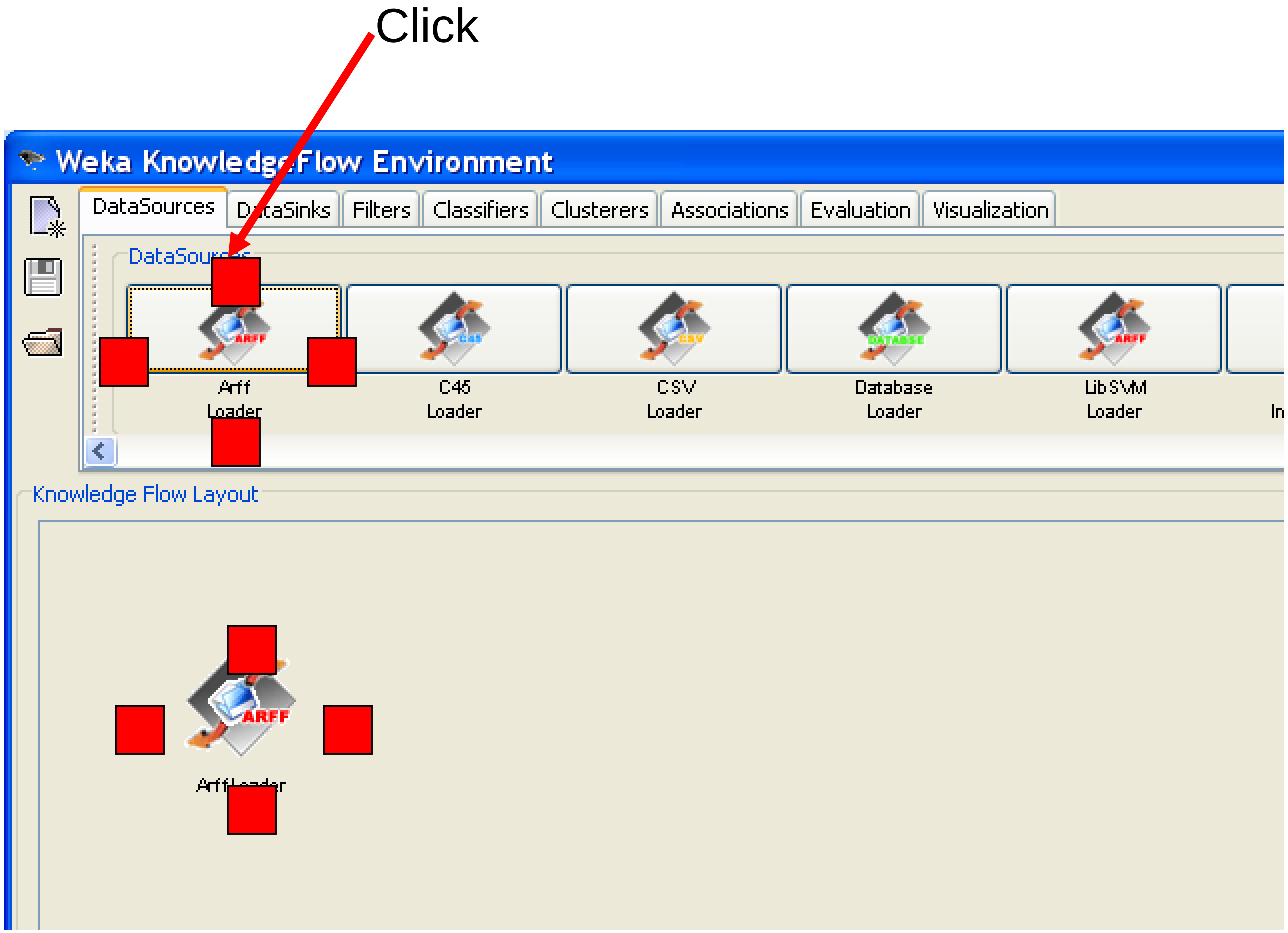
FILTERS

CLASSIFIERS

EVALUATION

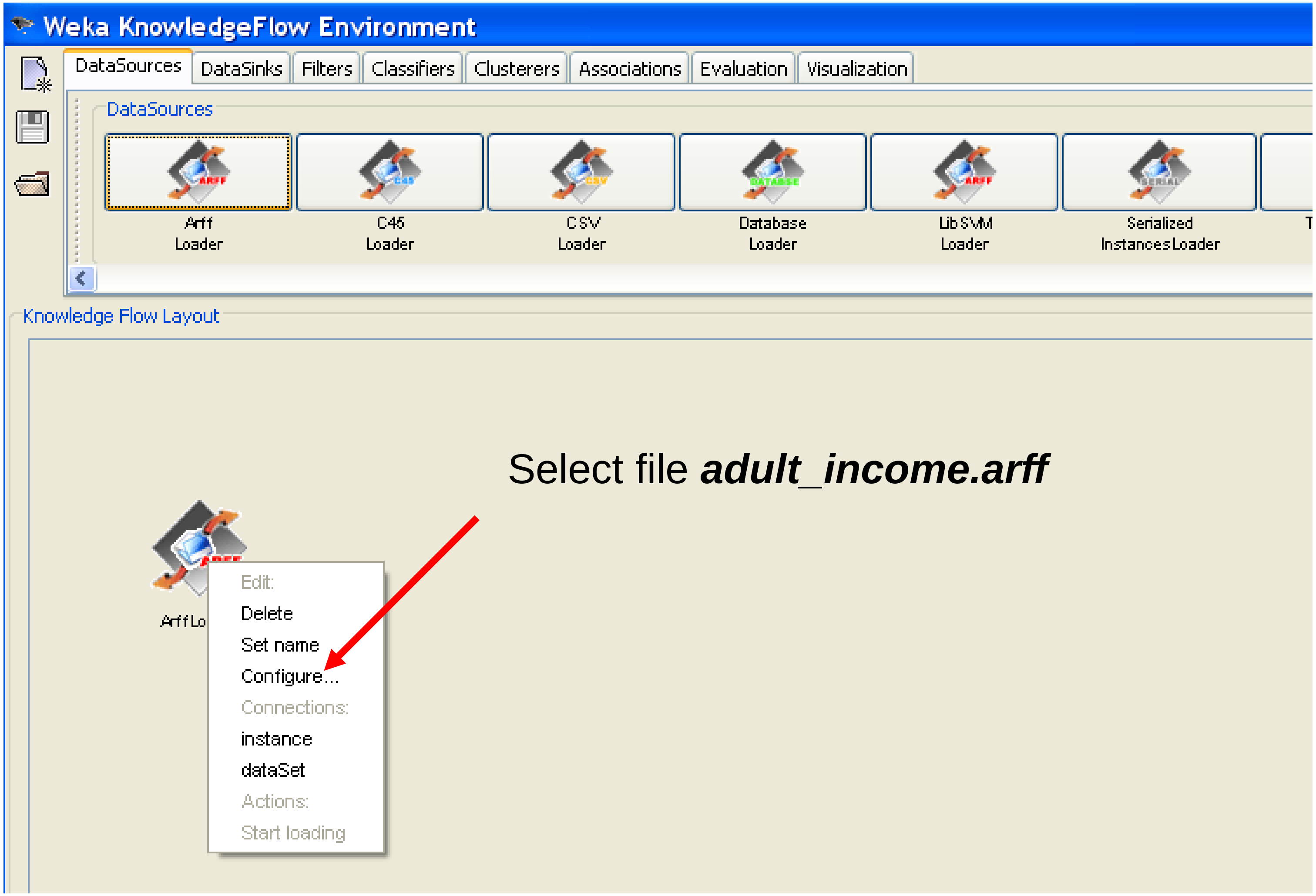
VISUALIZATION

Loading the data



Click

Loading the data



Select file ***adult_income.arff***

Attributes of interest:

age, education,

class (income >50 K: YES,NO)

1. @attribute Age numeric

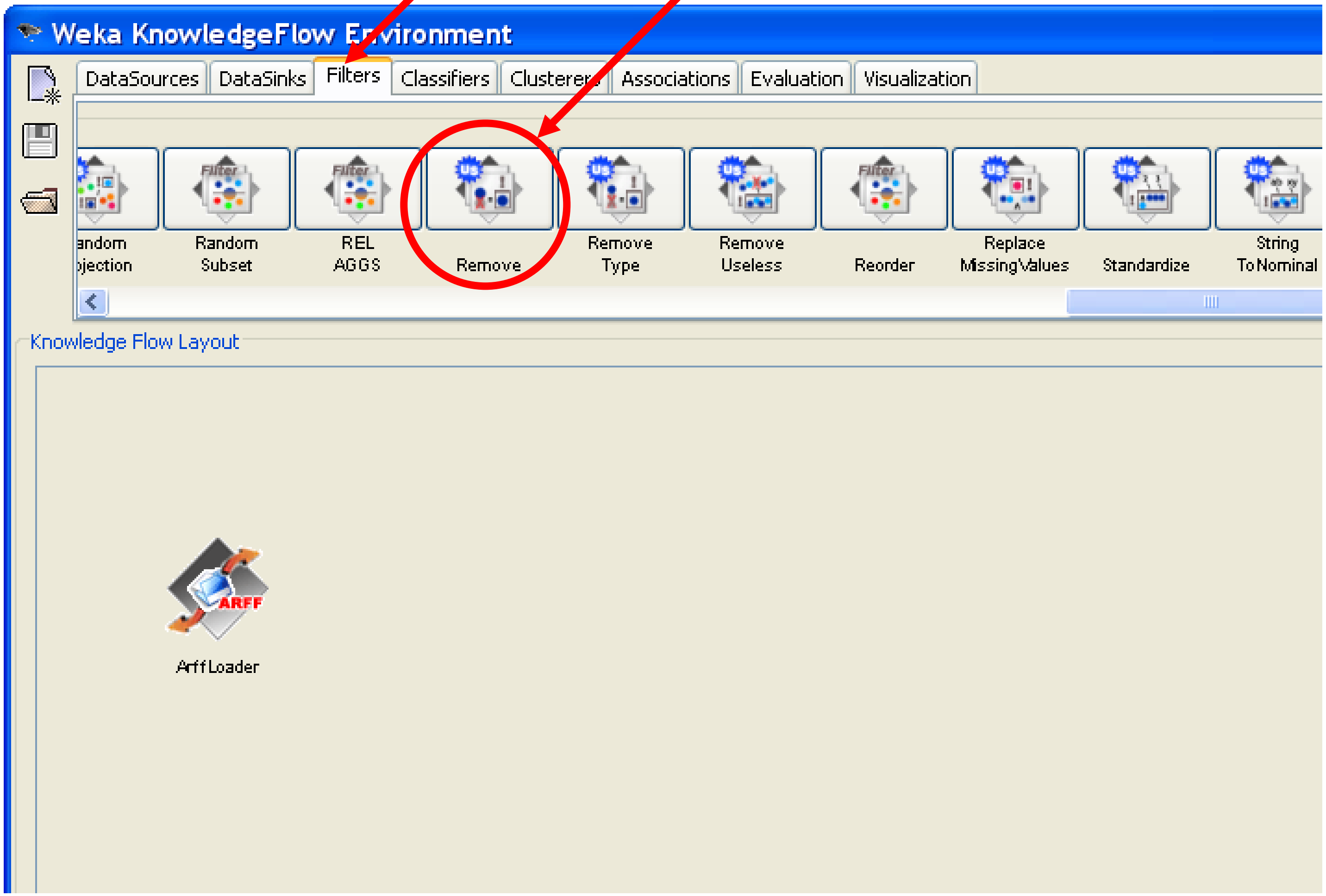
3. @attribute Education {Preschool,1st-4th,5th-6th,7th-8th,9th,10th,11th,12th,Prof-school,HS-grad,Some-college,Assoc-voc,Assoc-acdm,Bachelors,Masters,Doctorate}

last @attribute class {>50K, <=50K}

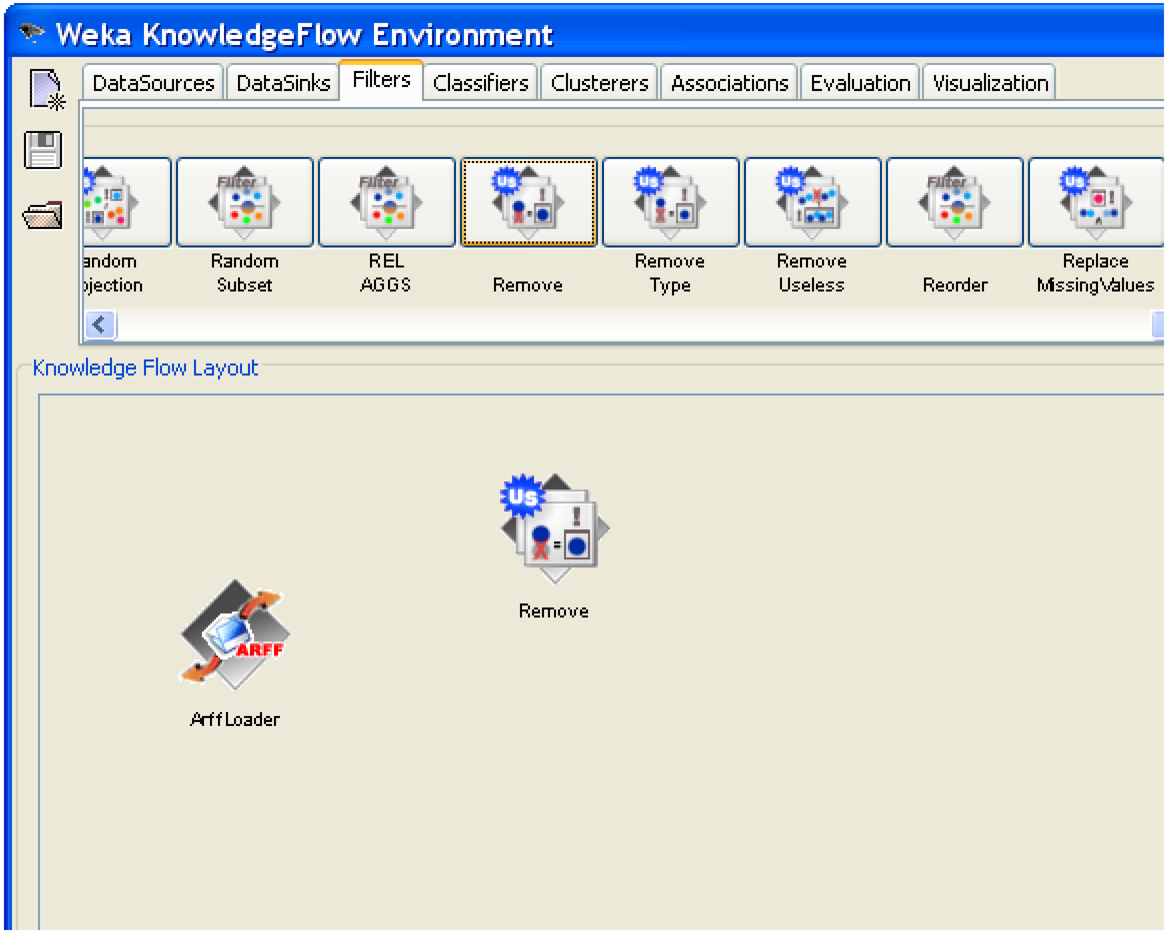
We remove all other attributes and leave only attributes 1,3, last – for simplicity

We build a classifier, which predicts income based on age and education

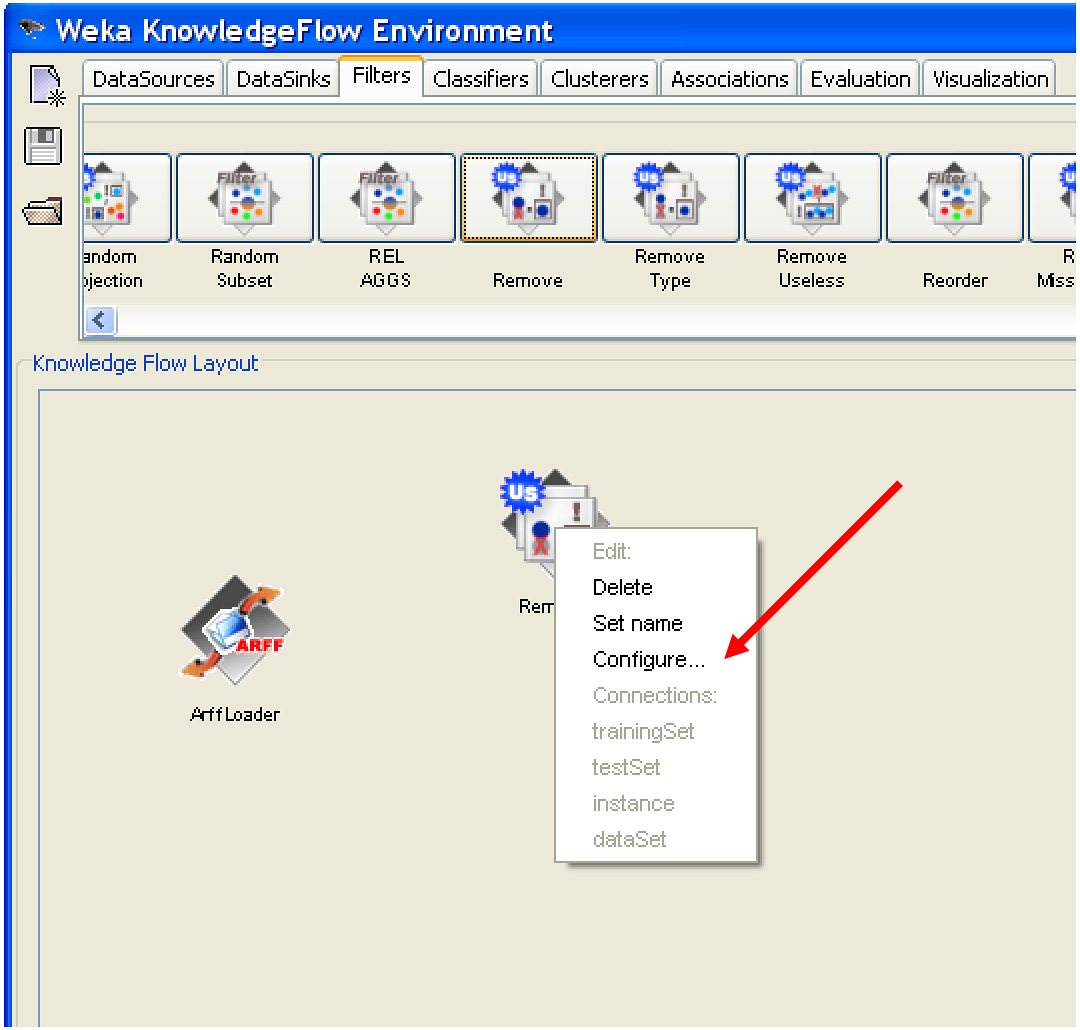
Removing attributes



Removing attributes

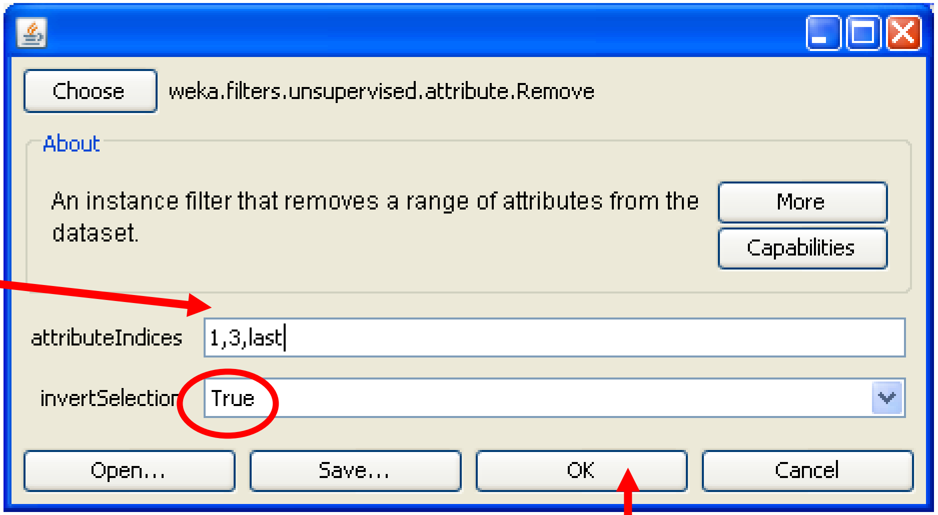


Removing attributes



Removing attributes

What not to remove



weka.filters.unsupervised.attribute.Remove

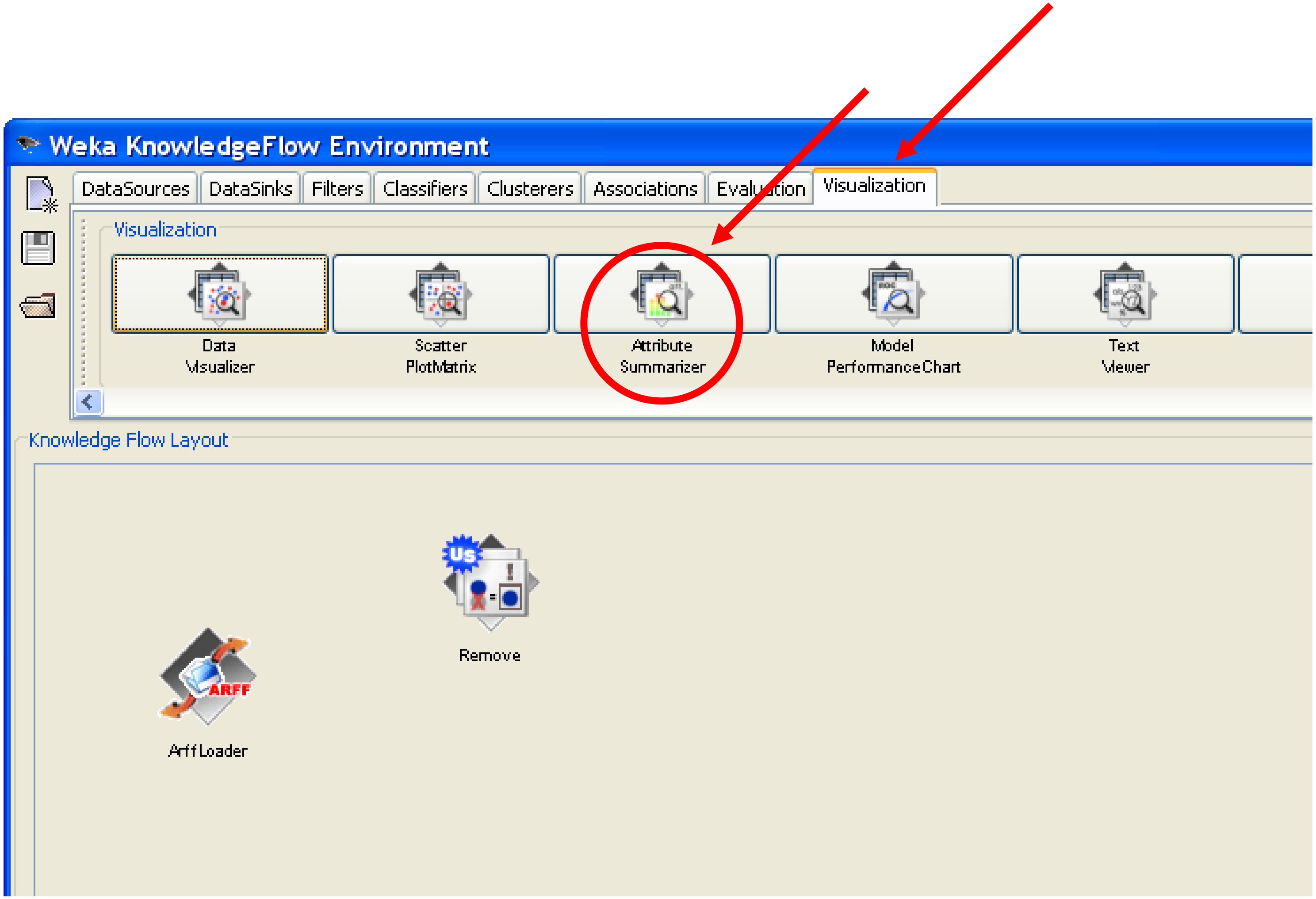
An instance filter that removes a range of attributes from the dataset.

attributeIndices 1,3,last

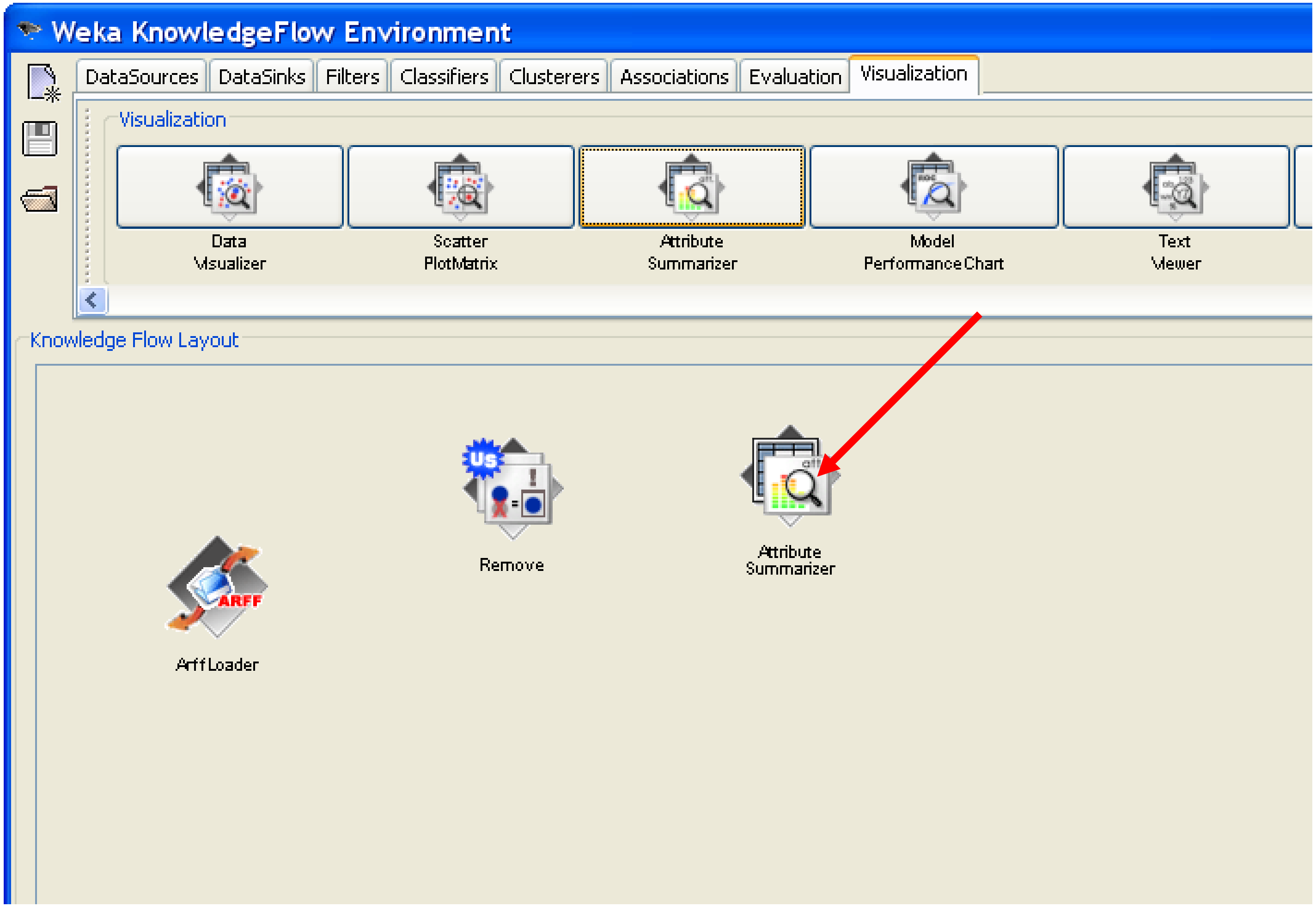
invertSelection True

It means: remove all except attributes 1,3,last

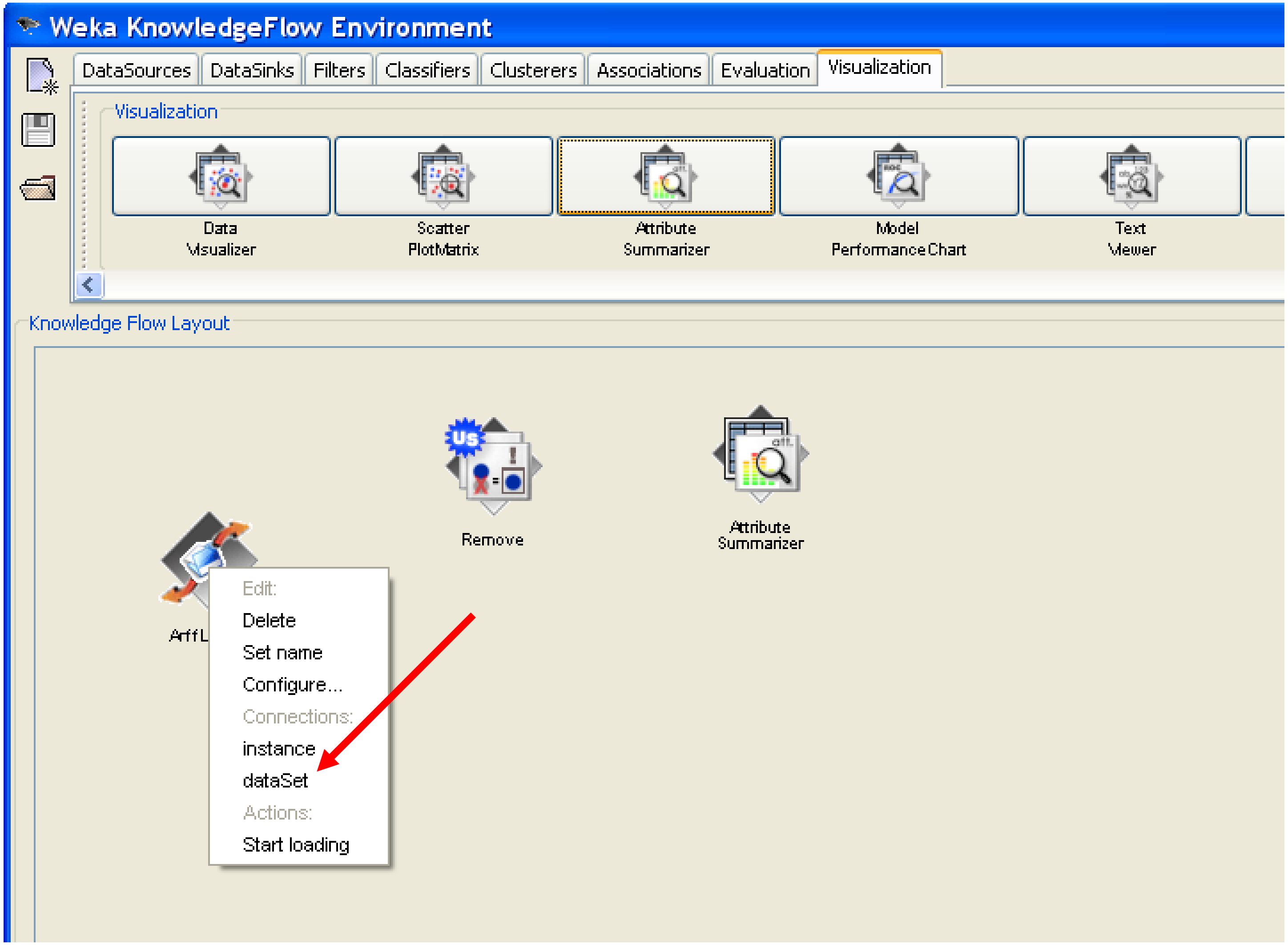
Visualize data



Visualize data

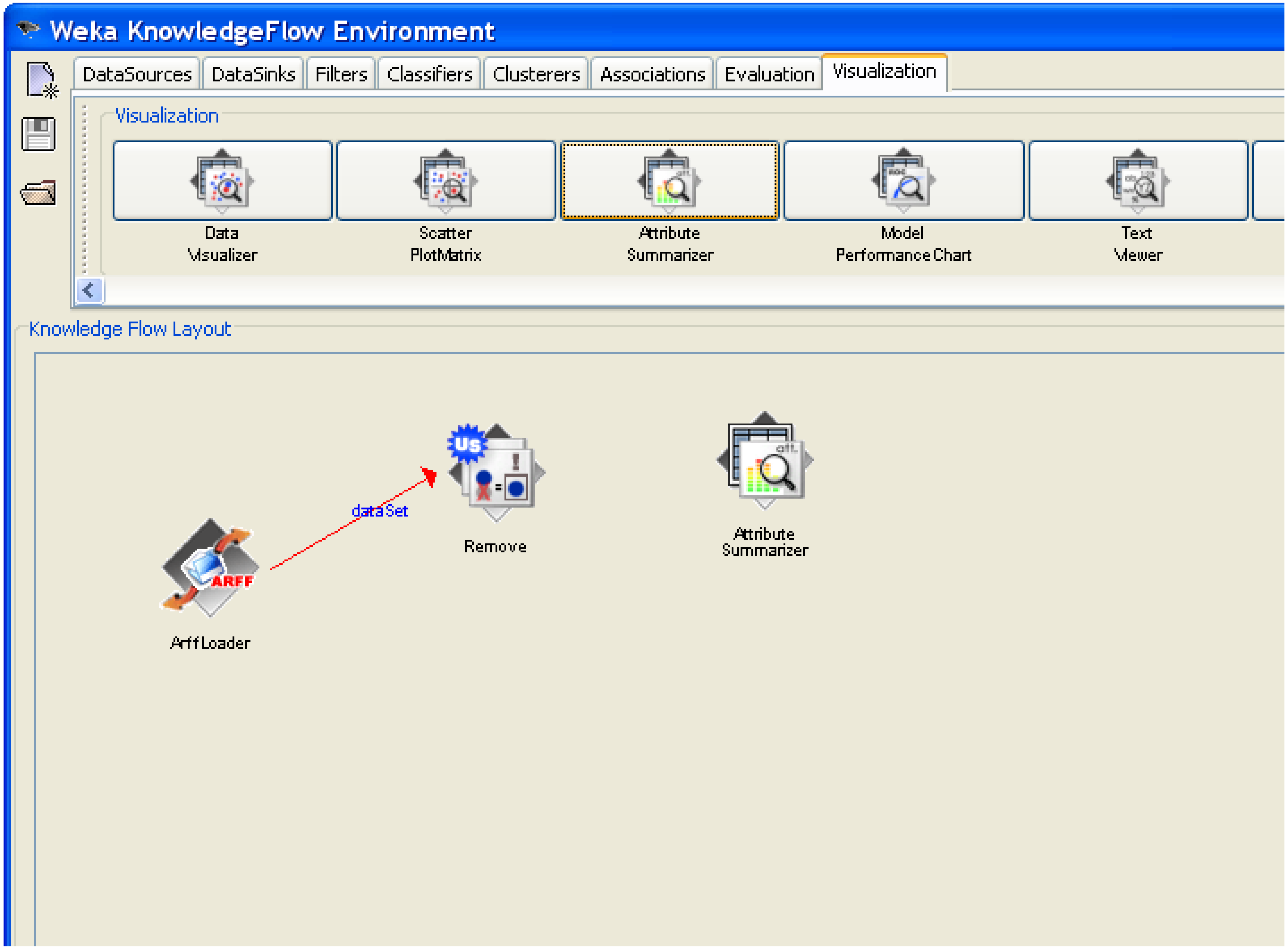


Connect the flow



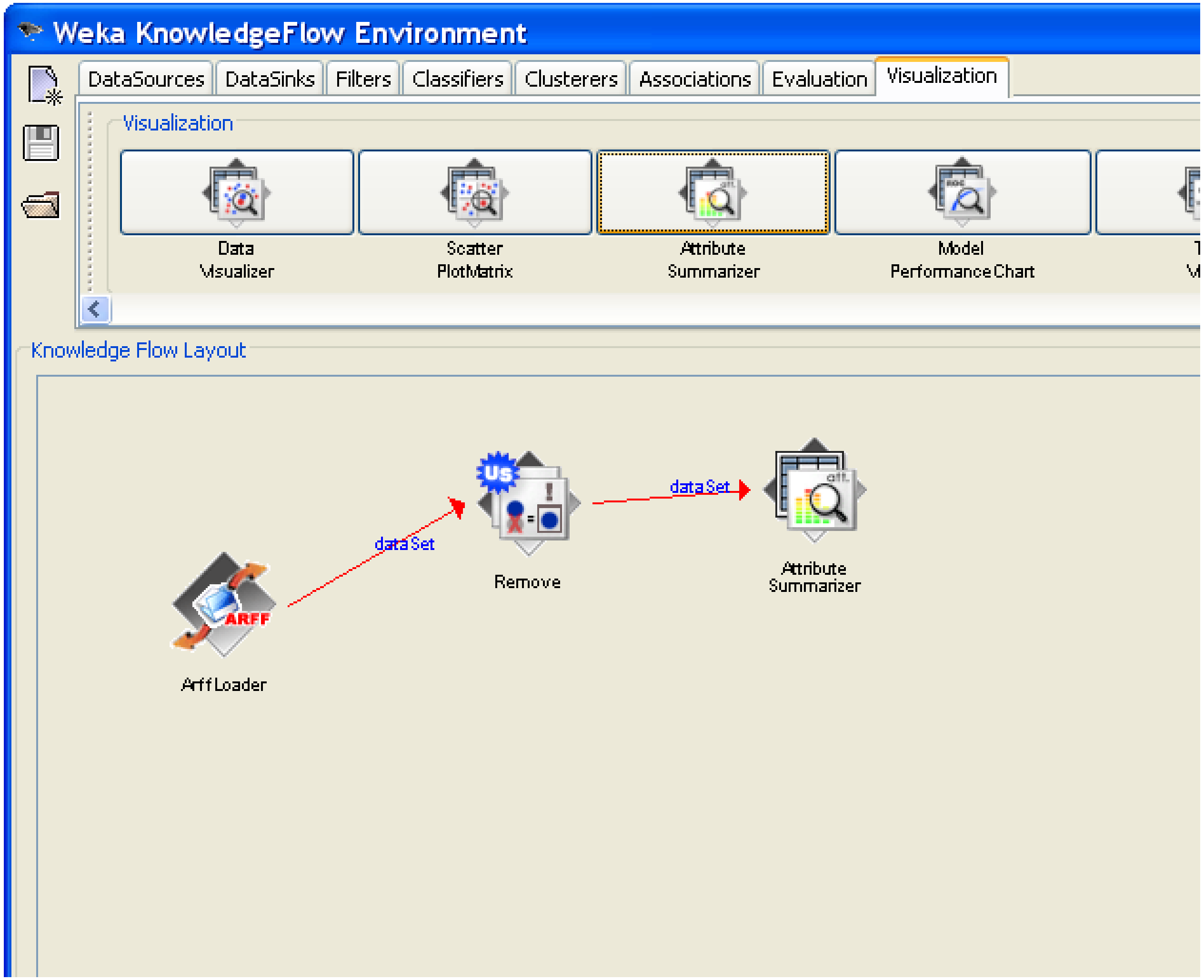
Connect the flow:

from data loader to attribute remover

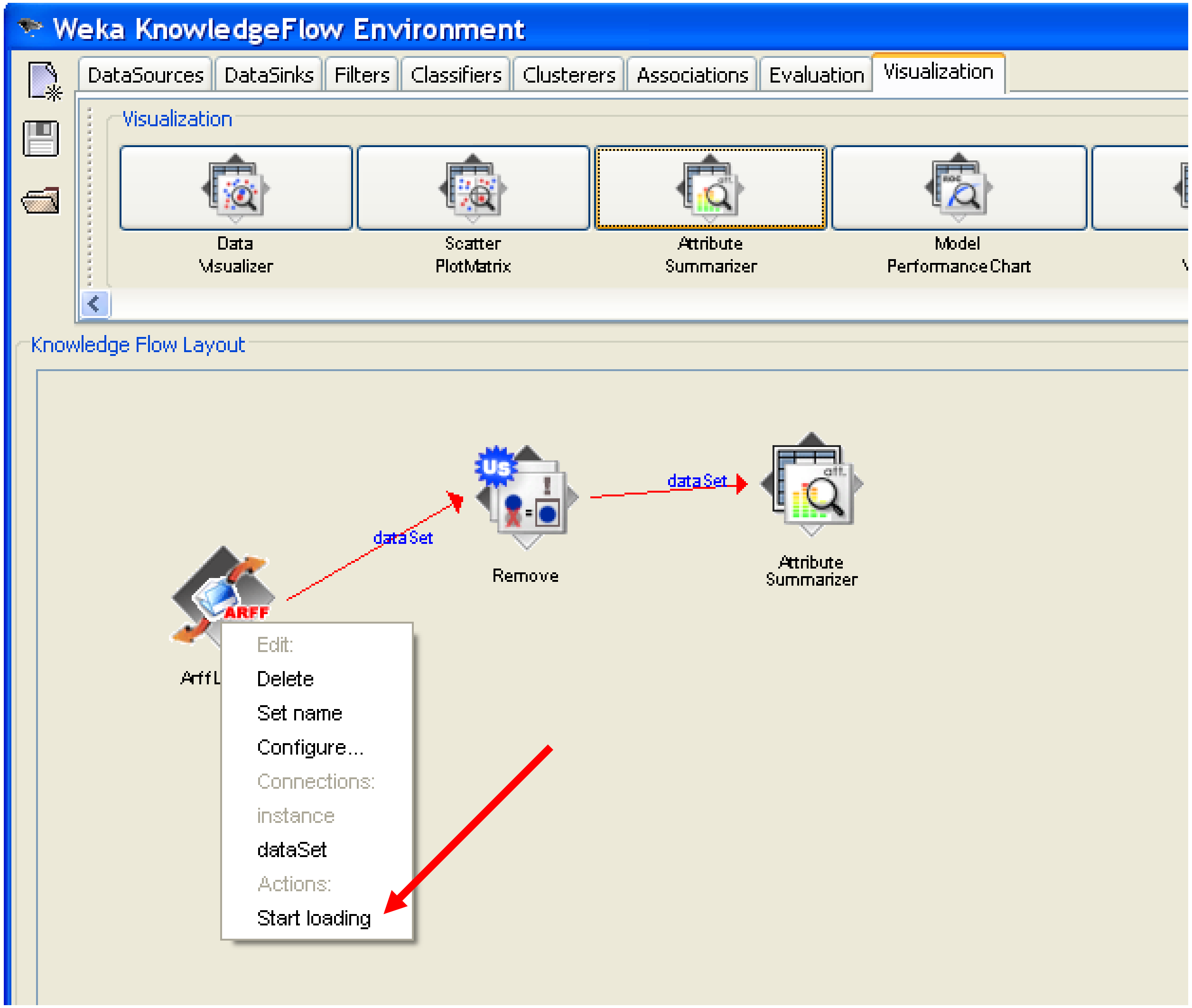


Connect the flow:

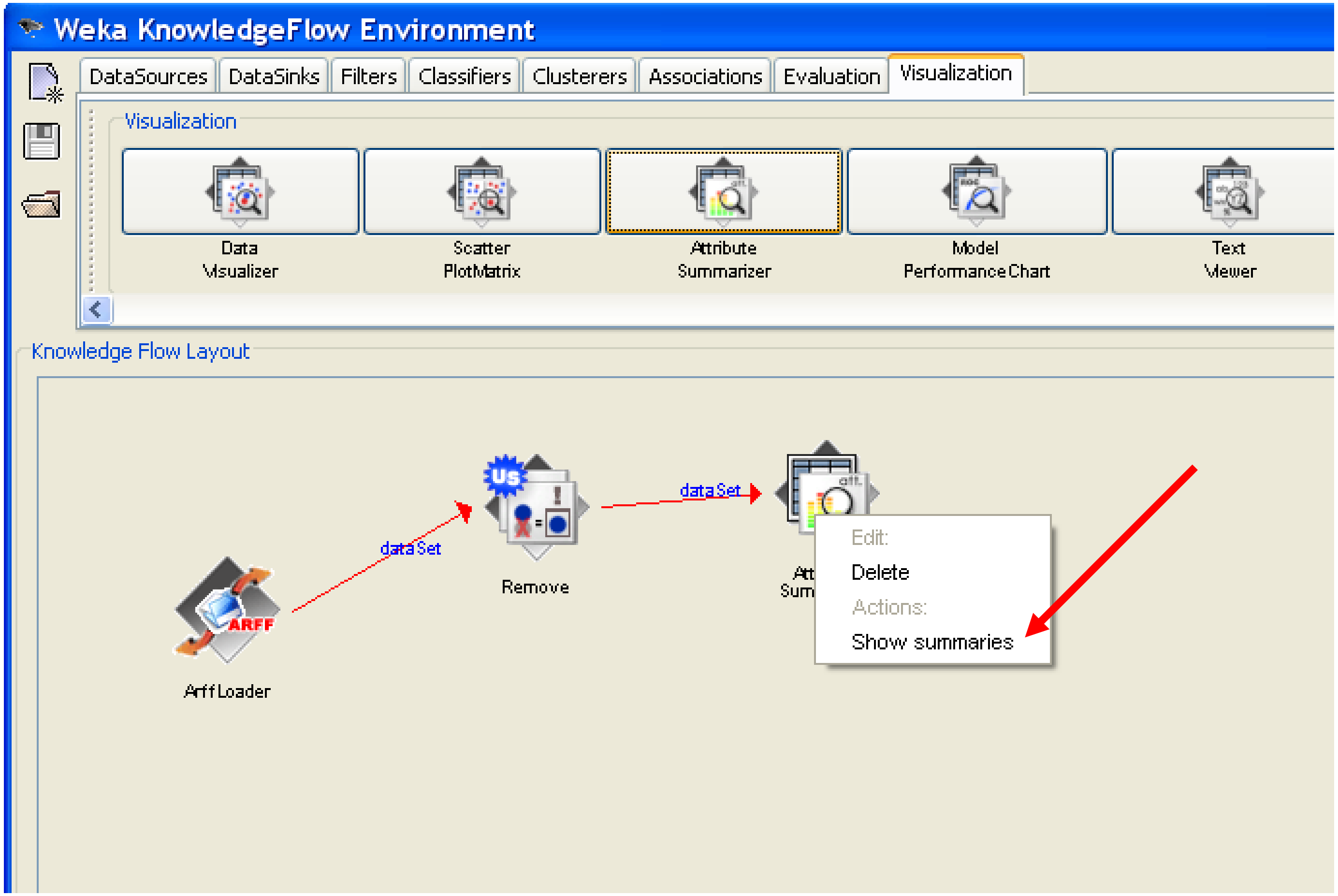
from attribute remover to summarizer



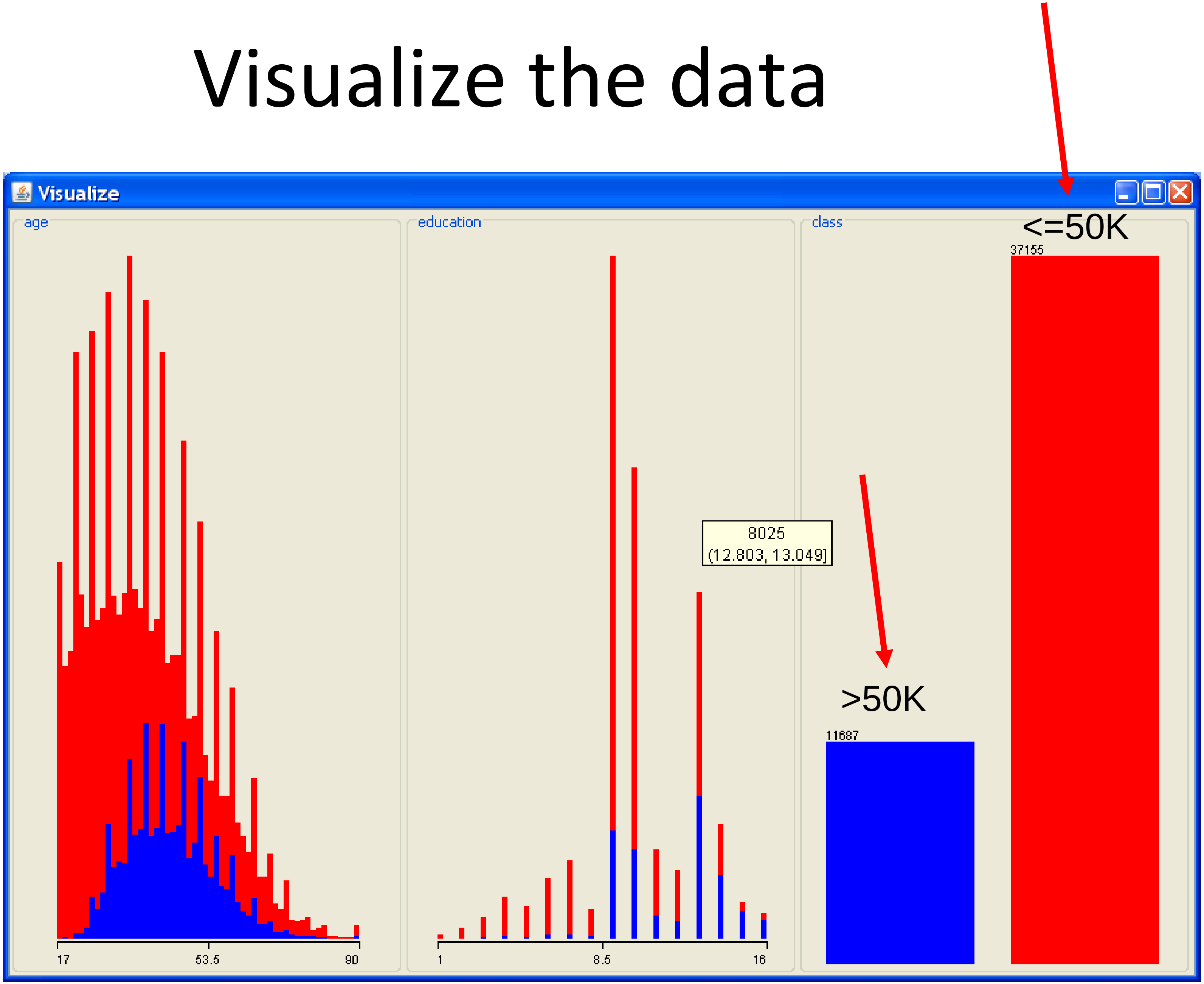
Start data flow



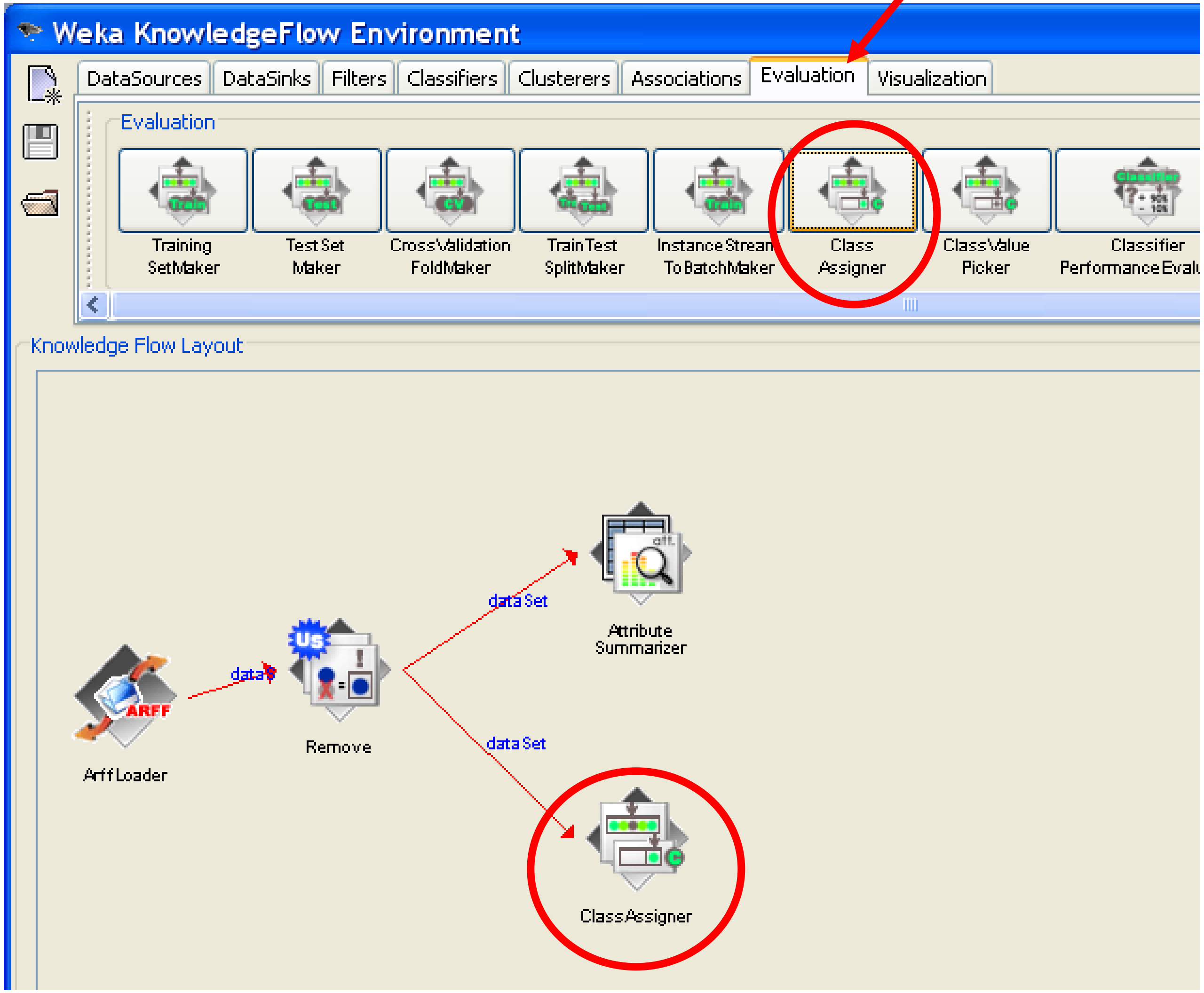
Visualize the data



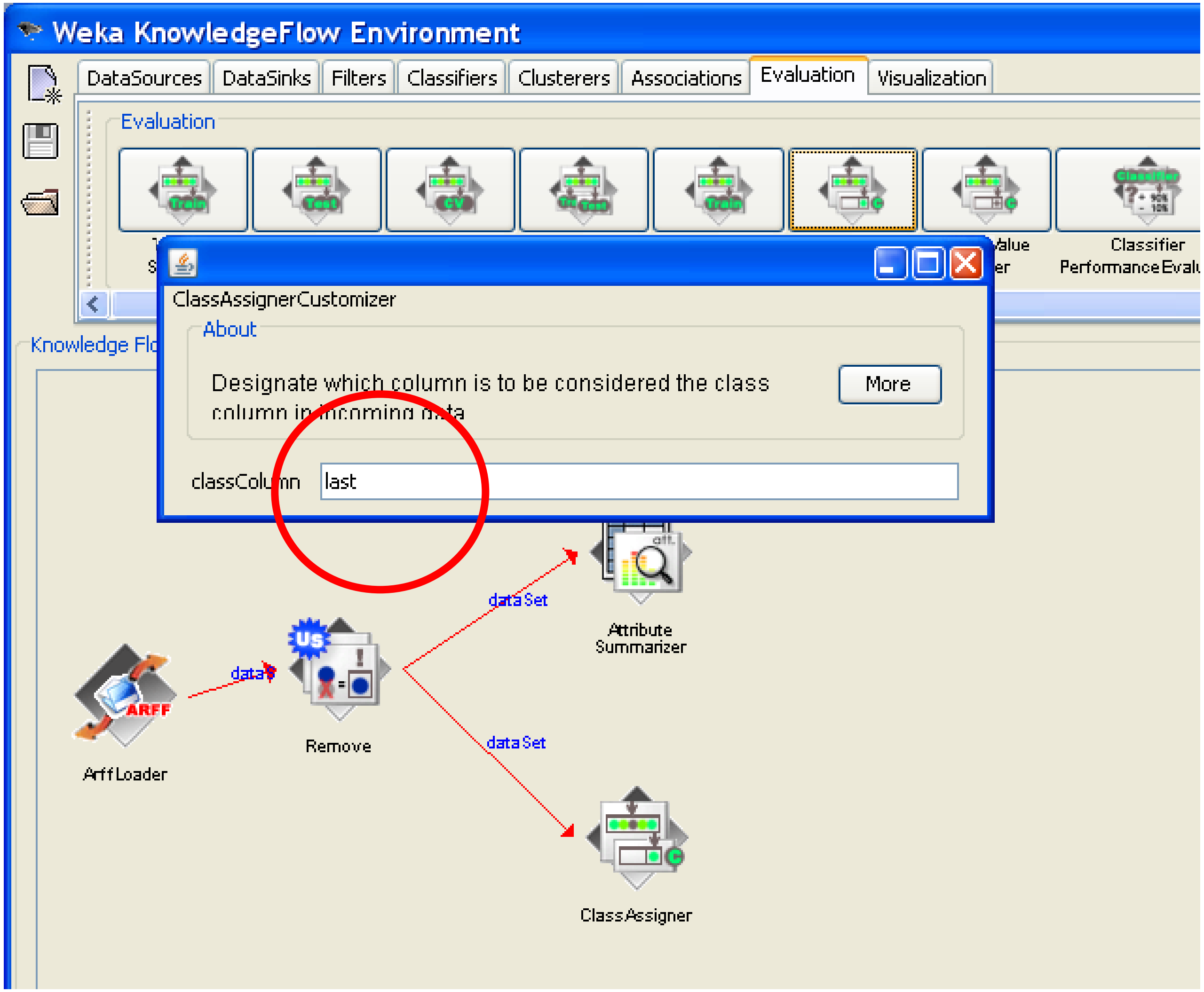
Visualize the data



Assigning the class



Configuring class assigner



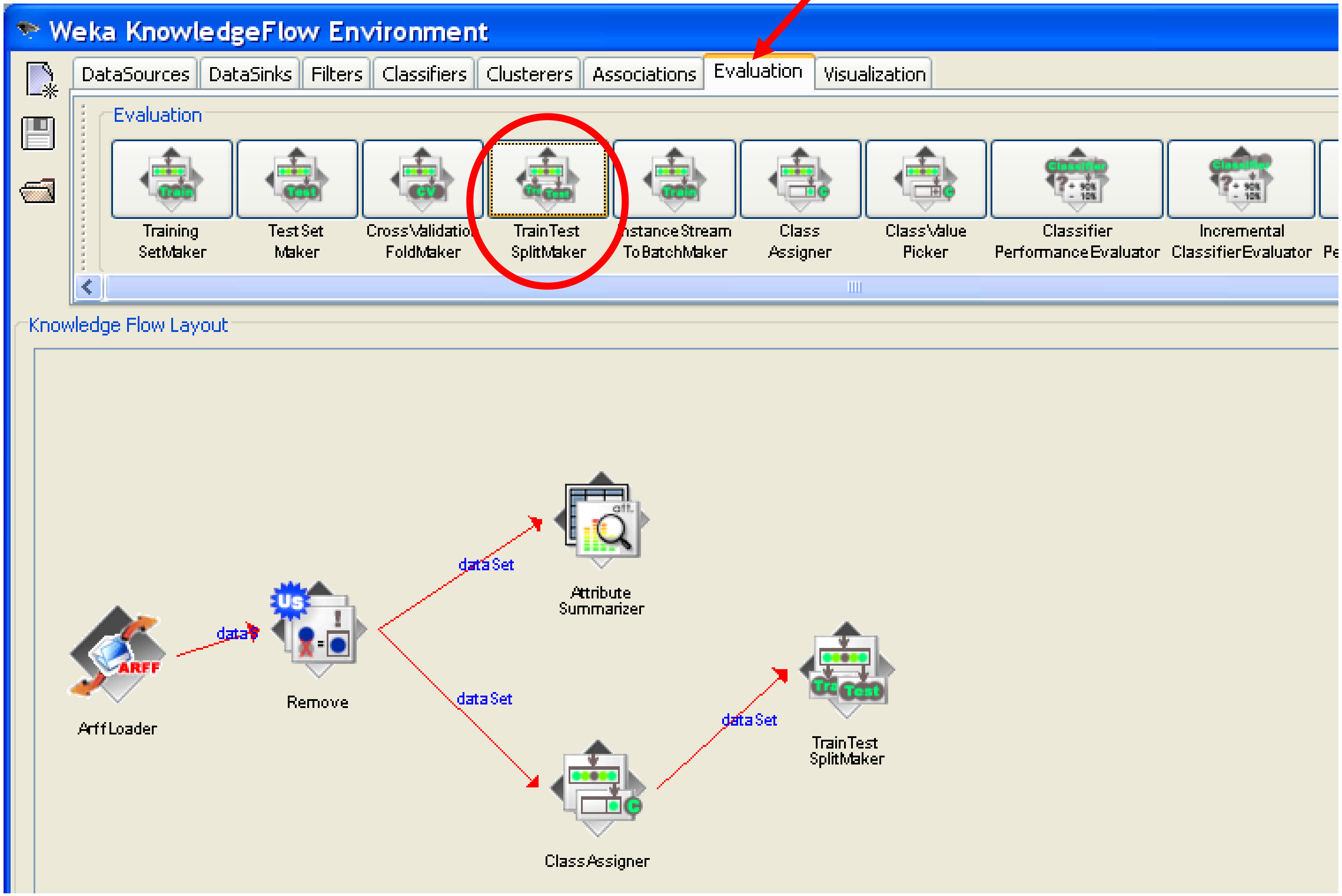
ClassAssignerCustomizer

About

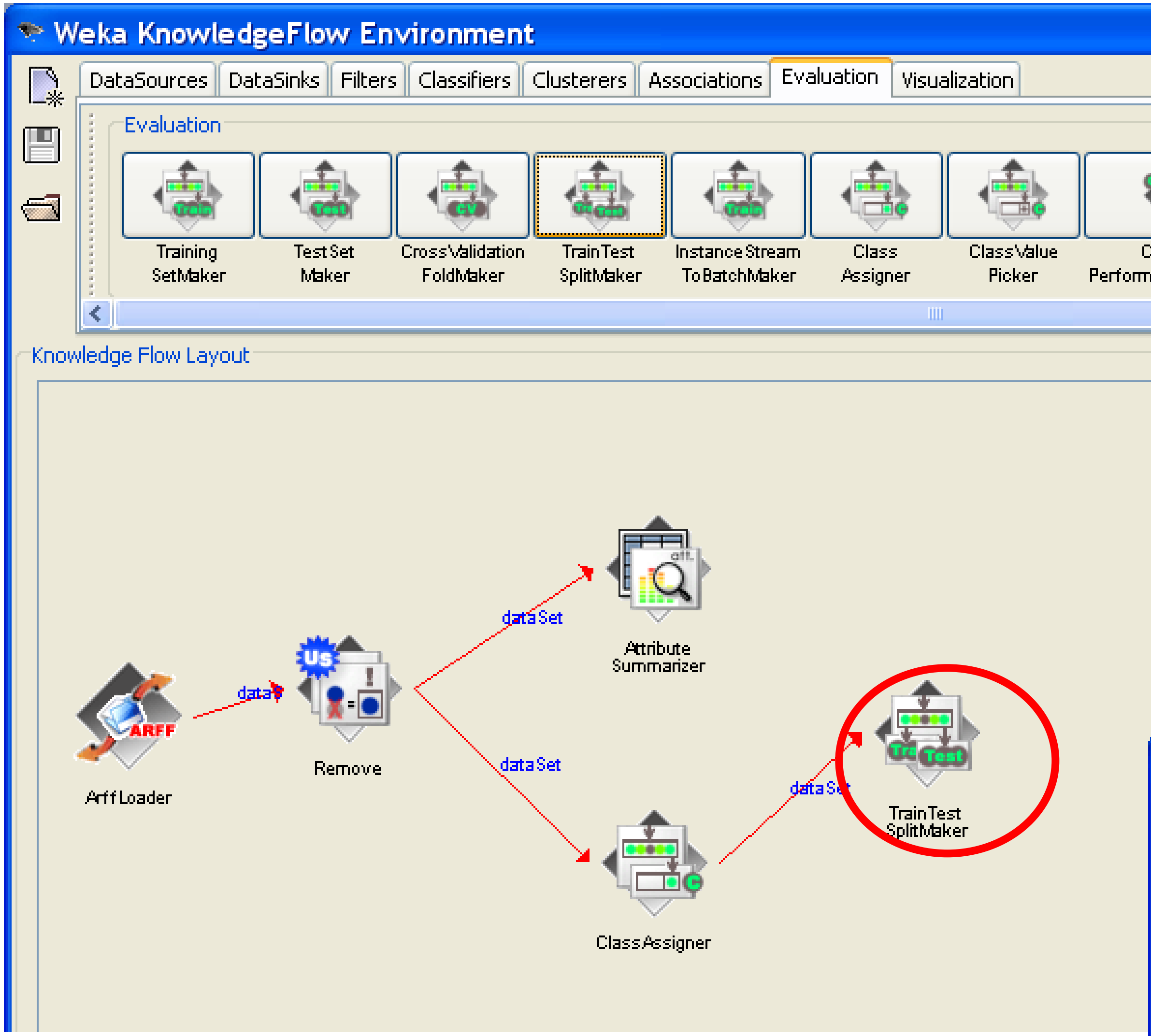
Designate which column is to be considered the class column in incoming data

classColumn: last

Subdivision of the dataset into “training” and “test” set



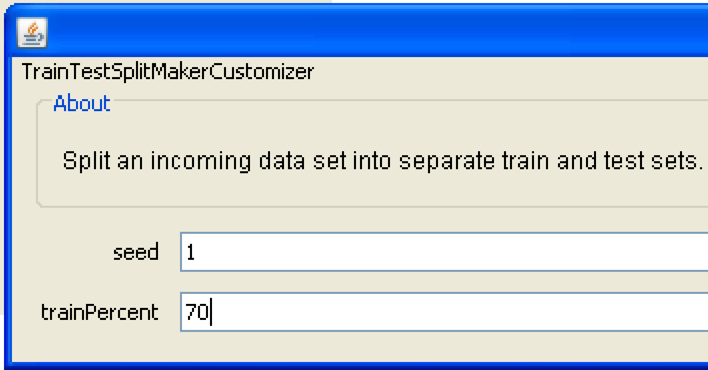
Subdivision of the dataset into “learning” and “test” set



We want to build our prediction model on the 70% of the whole dataset,

and test on the remaining

So, we set the TRAINTEST SPLIT MAKER (EVALUATION) in
the diagram and configure its parameters.



TrainTestSplitMakerCustomizer

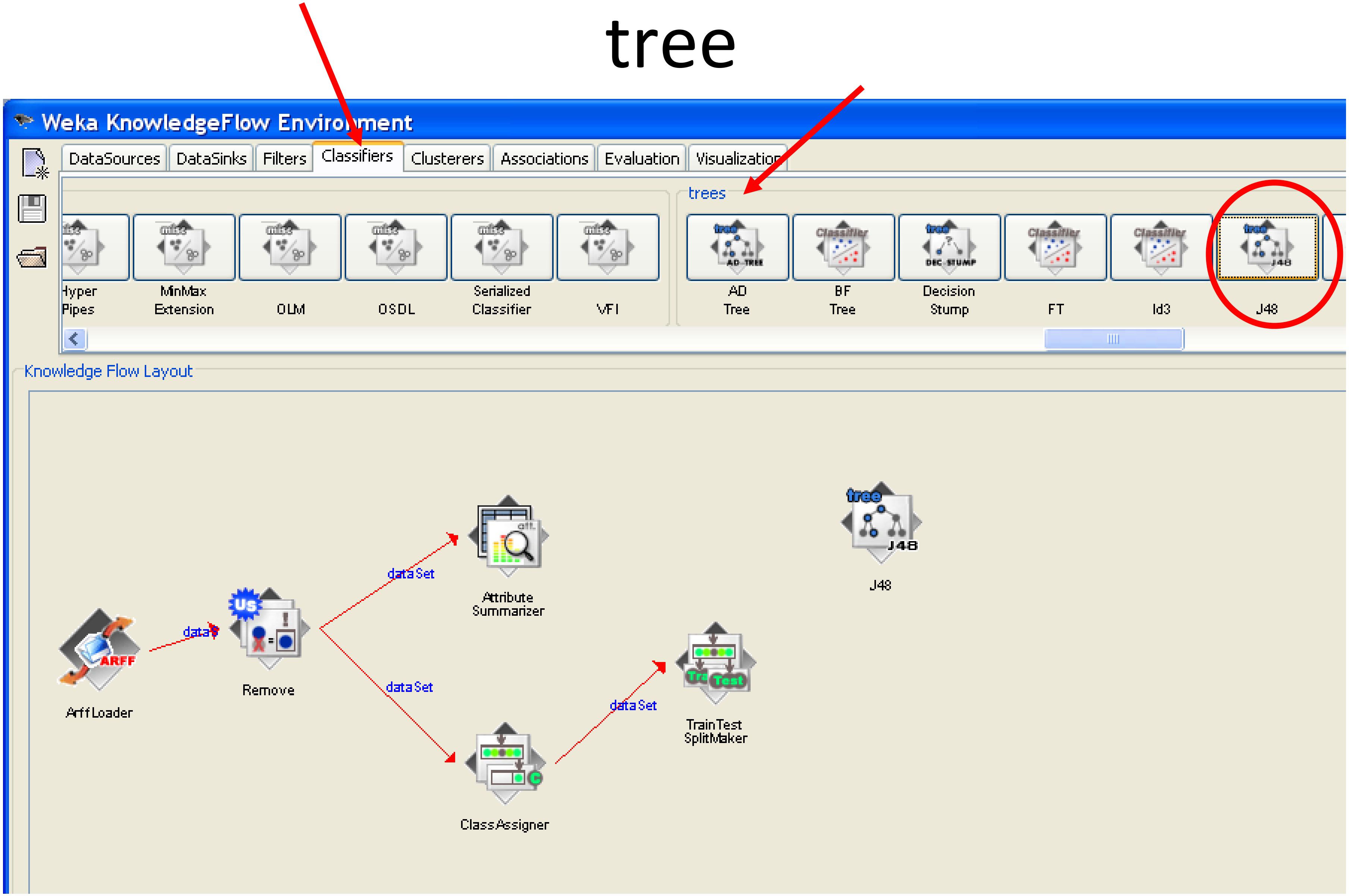
About

Split an incoming data set into separate train and test sets.

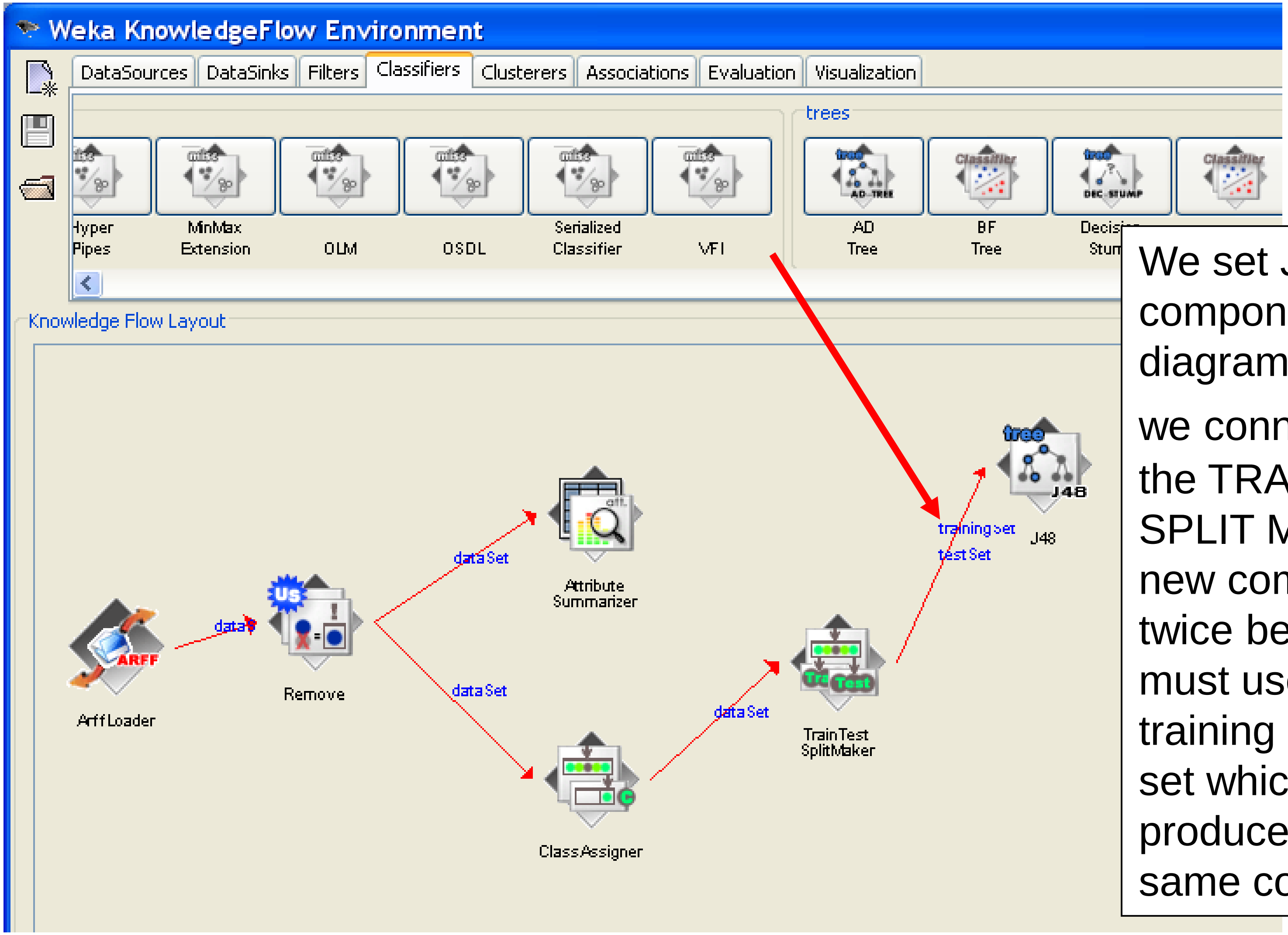
seed 1

trainPercent 70

Choosing discrete classifier – decision tree



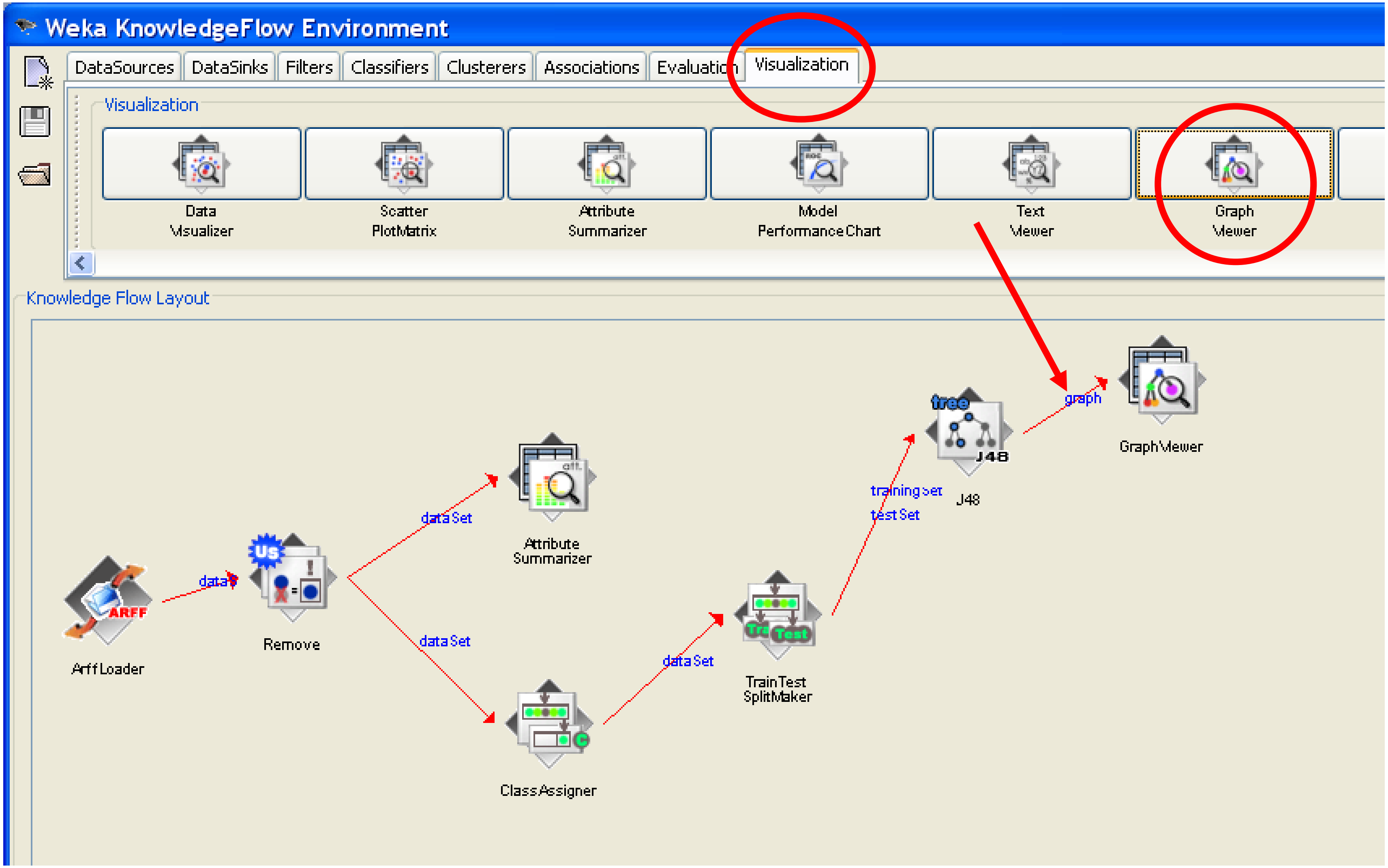
Connecting classifier to the data



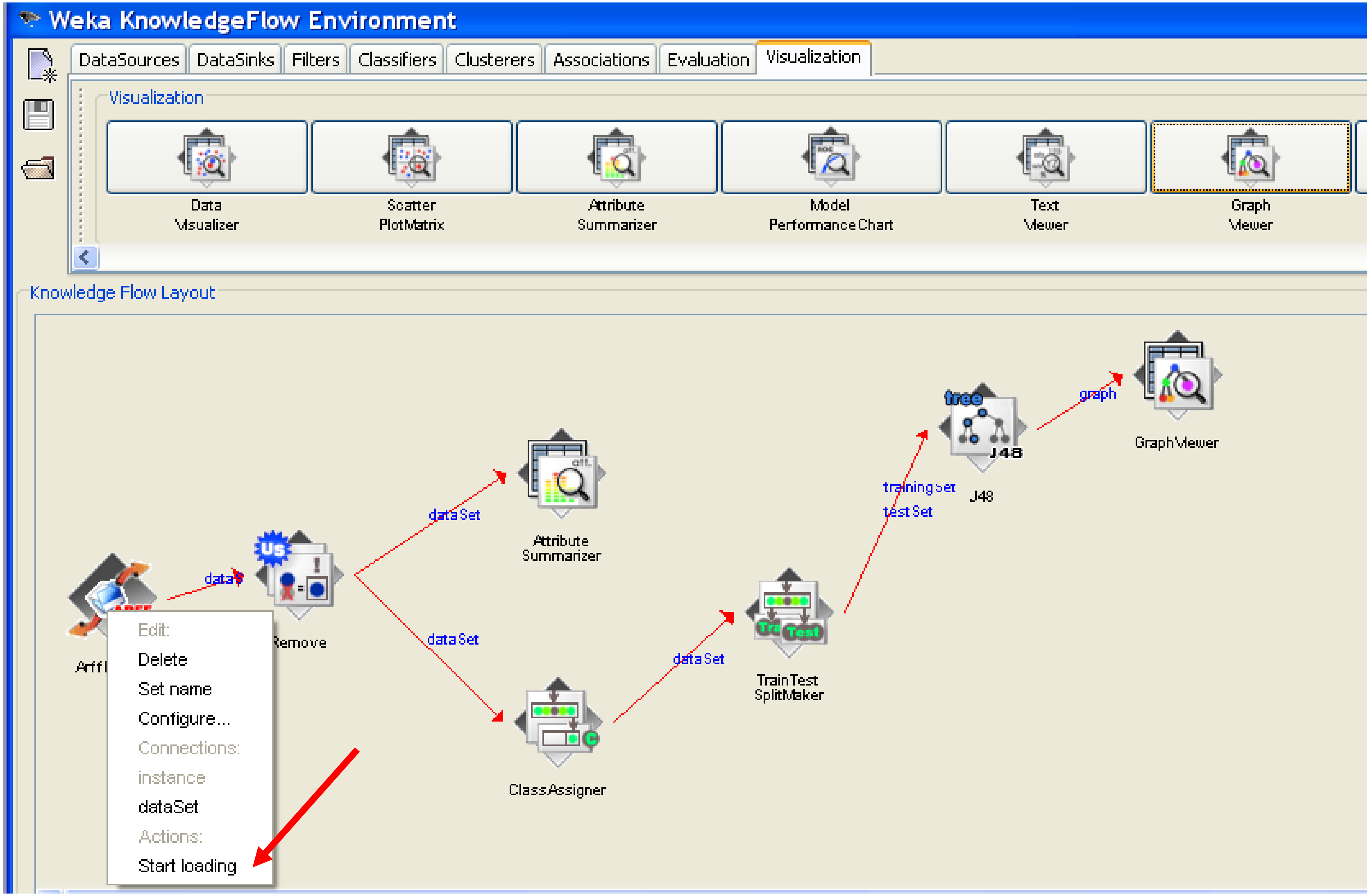
We set J48 component in the diagram,

we connect **twice** the TRAIN TEST SPLIT MAKER to this new component: twice because we must use together the training and the test set which are produced by the same component.

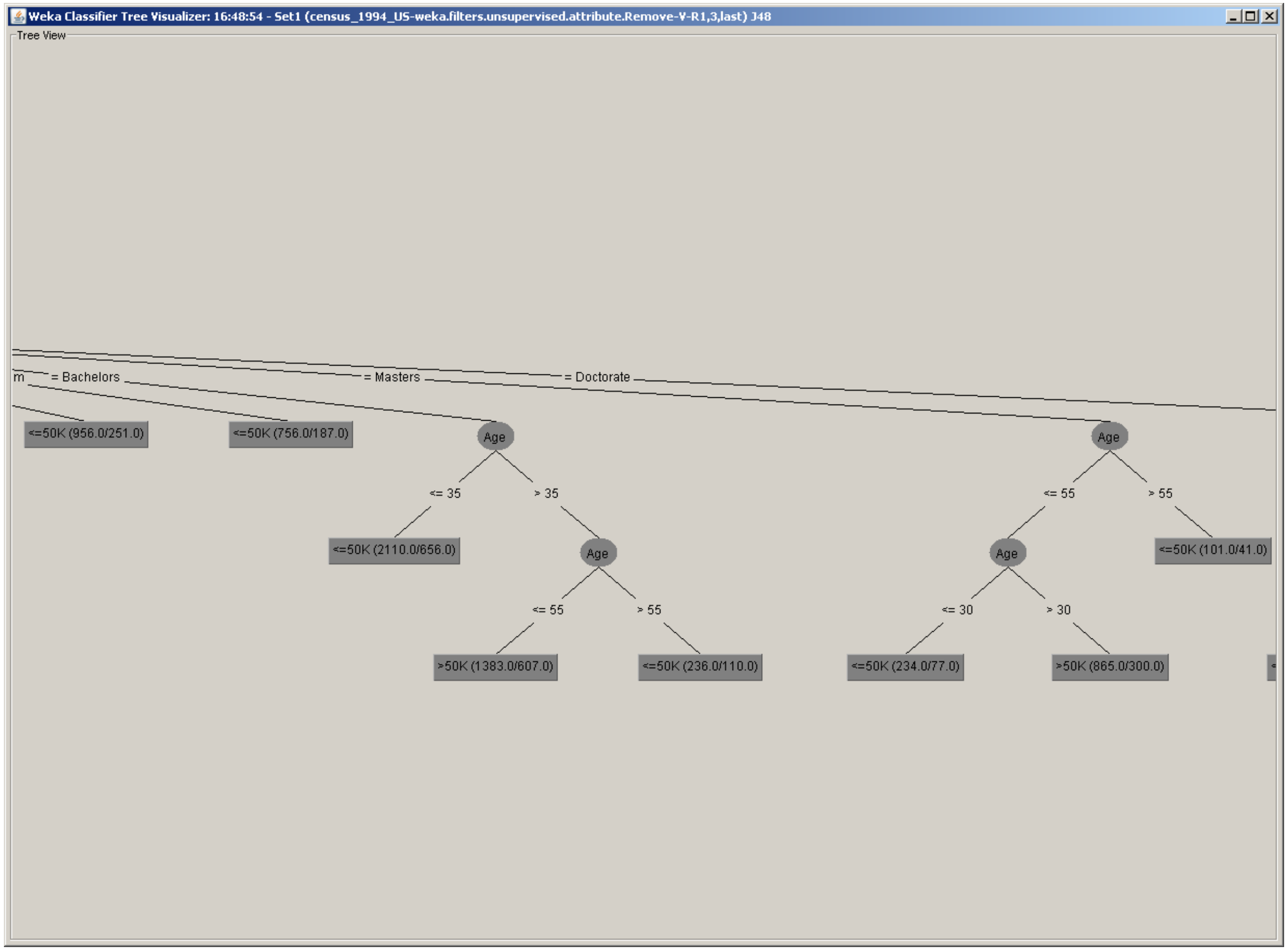
Adding visualizer to see the classification results



Perform classification

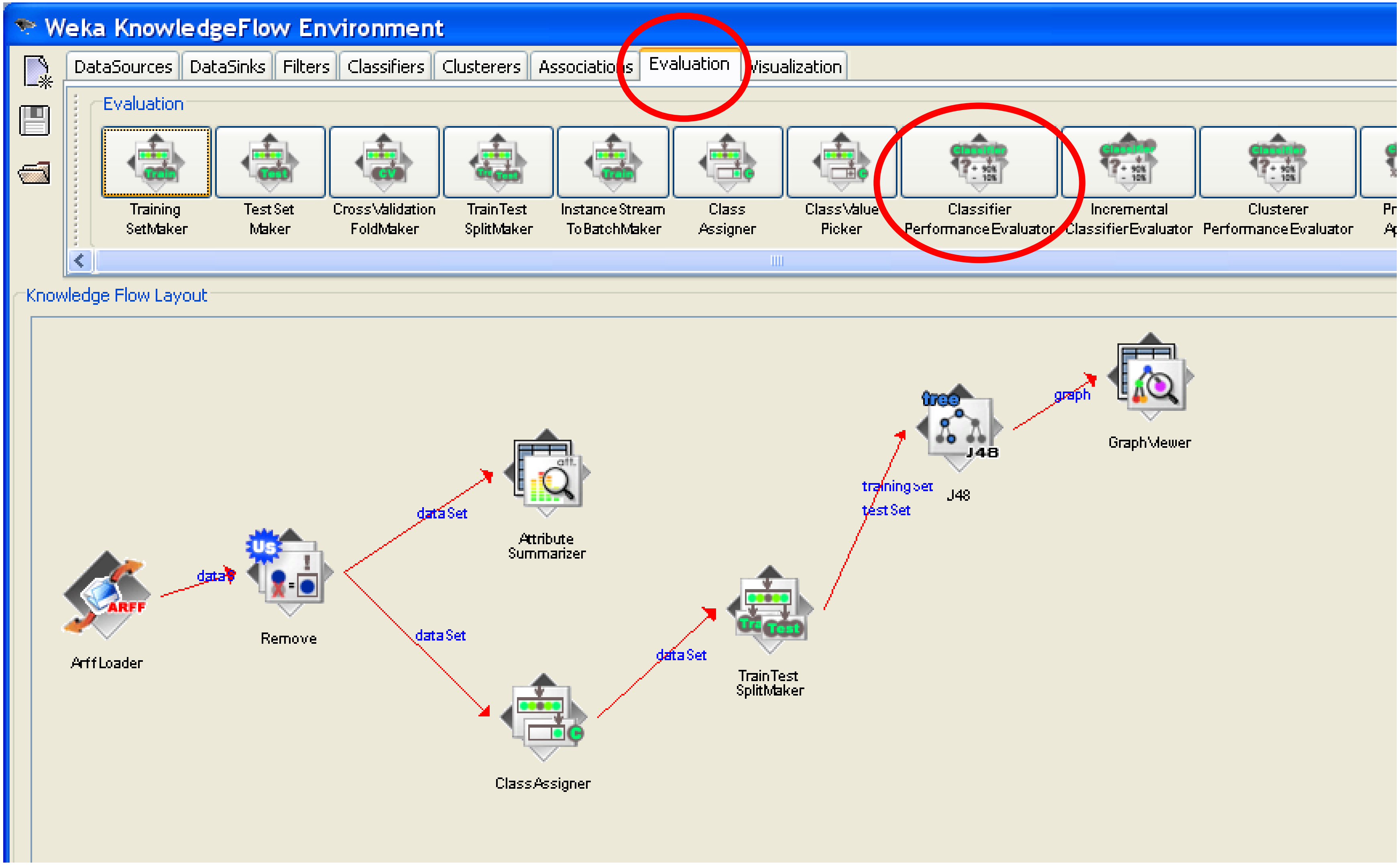


Show classification results (decision tree)

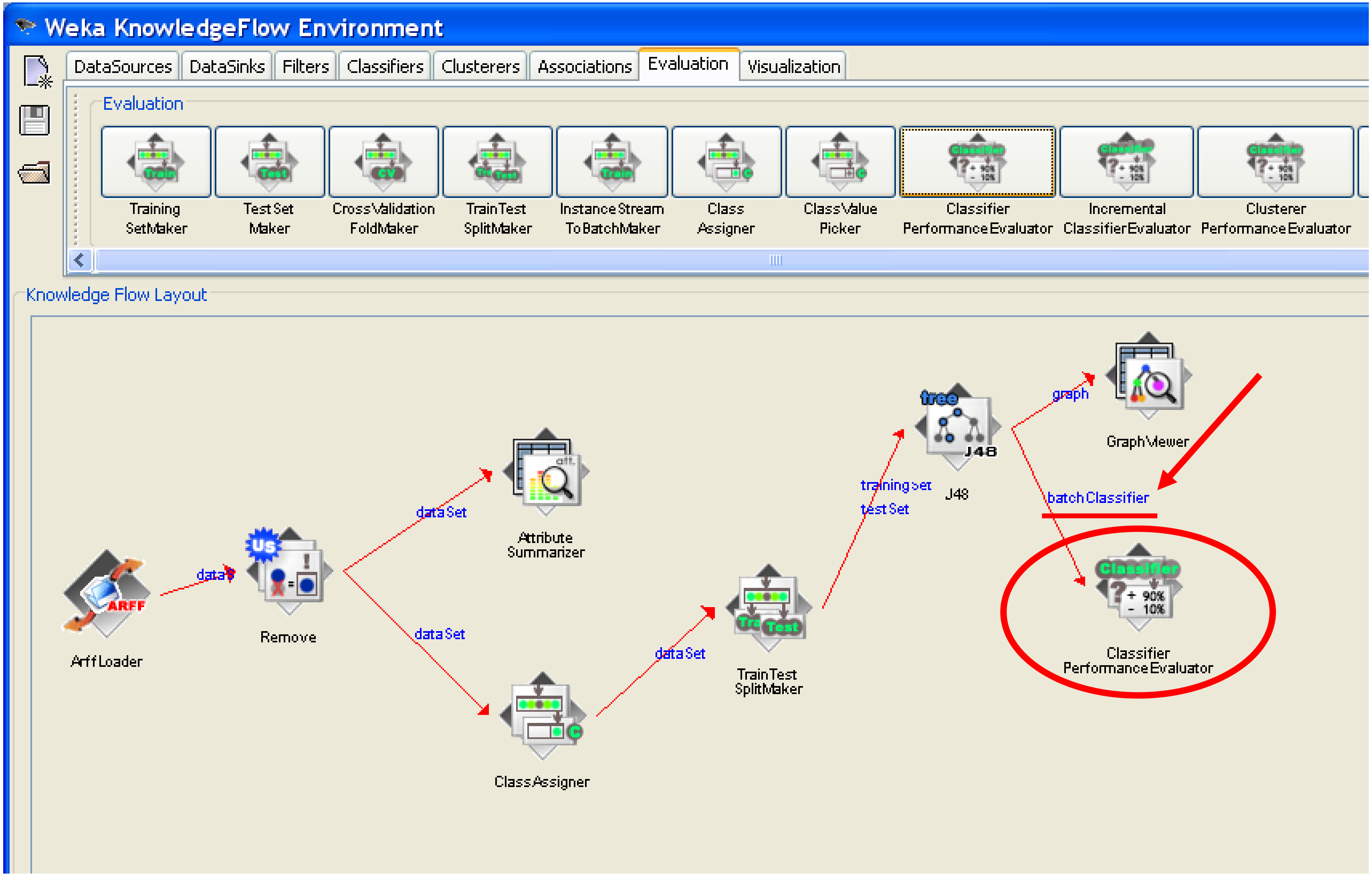


Weka Classifier Tree Visualizer: 16:48:54 - Set1 (census_1994_US-weka.filters.unsupervised.attribute.Remove-V-R1,3,last) J48

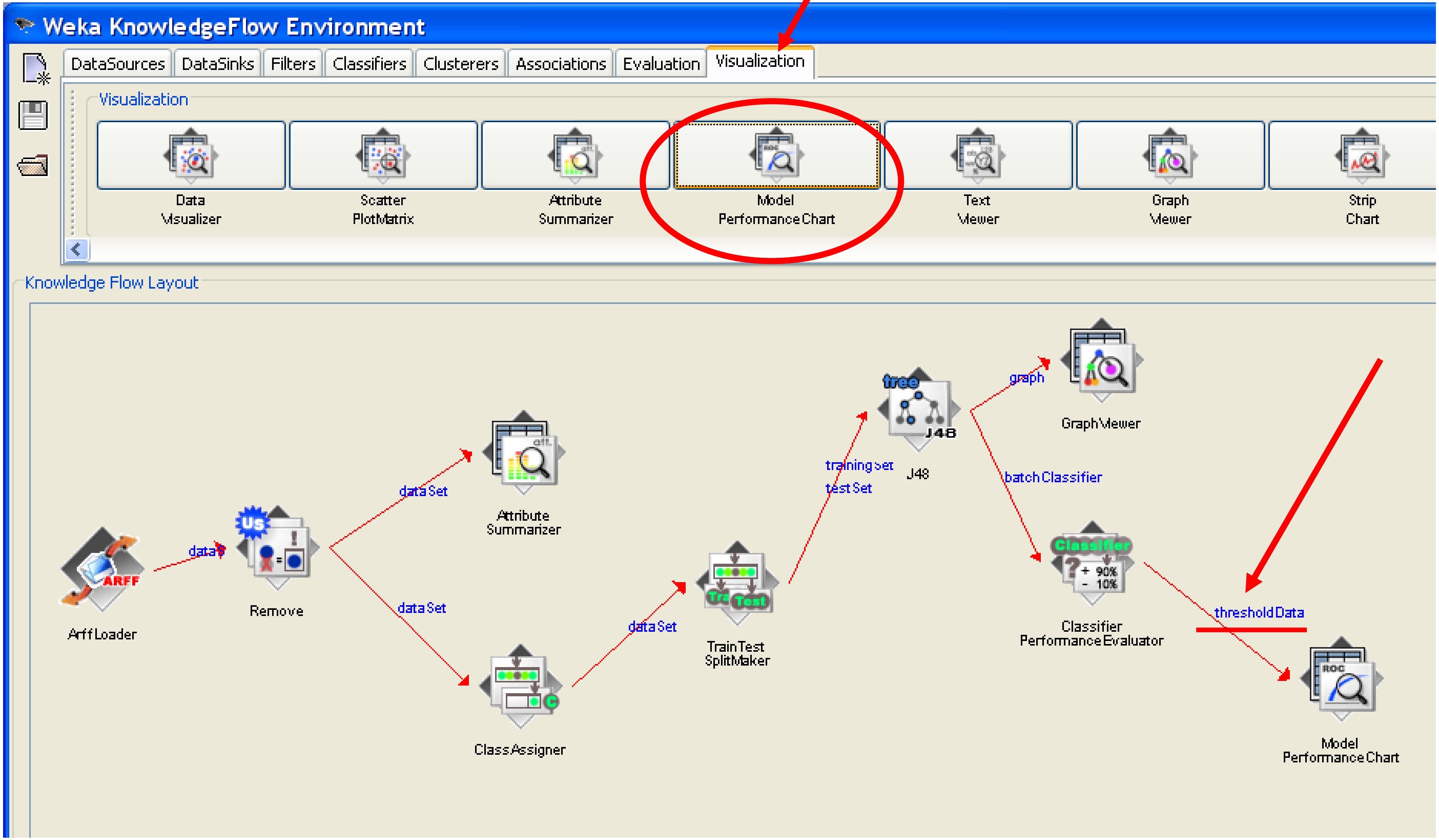
Classifier evaluation



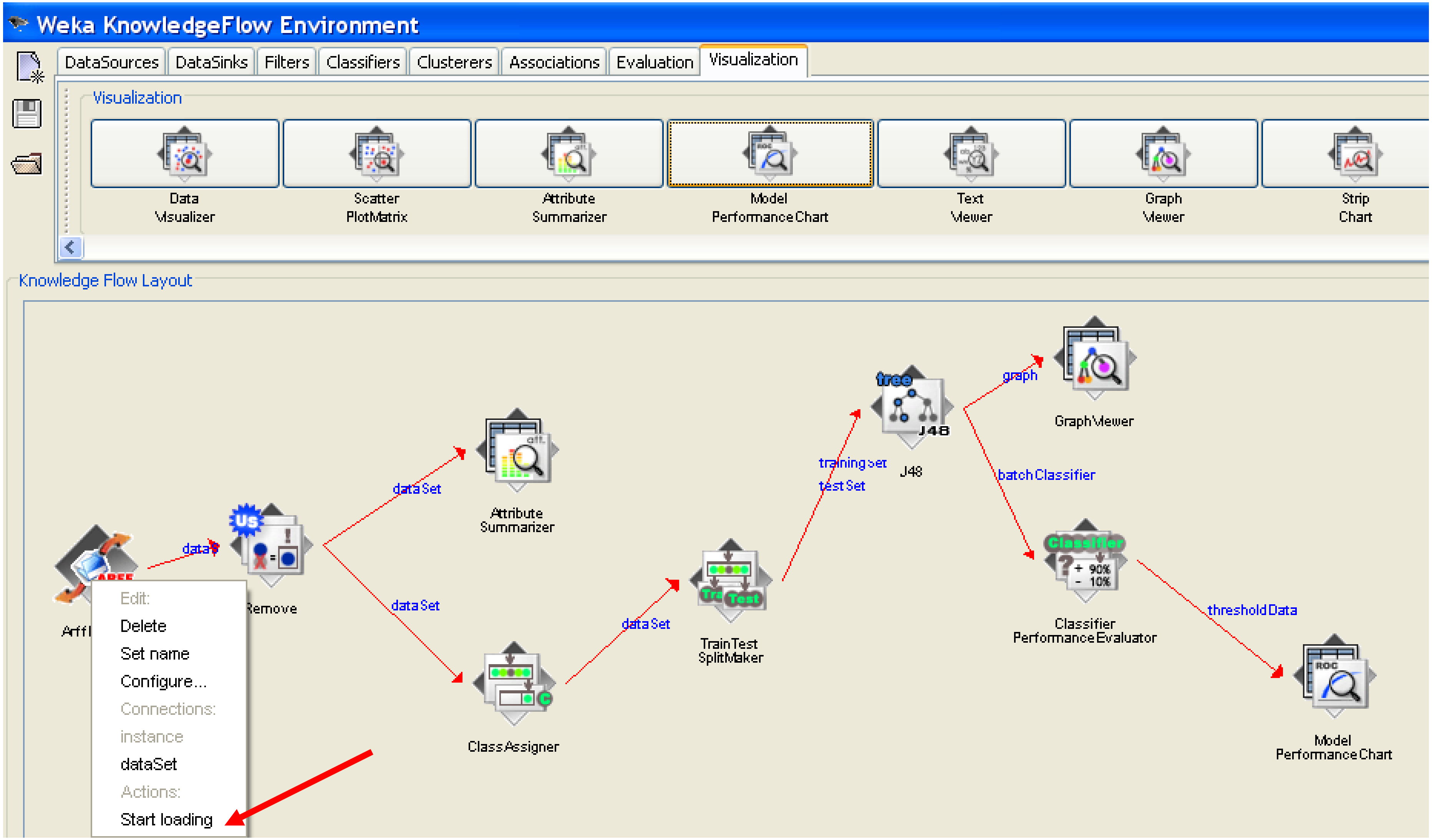
Connecting classifier to the evaluator



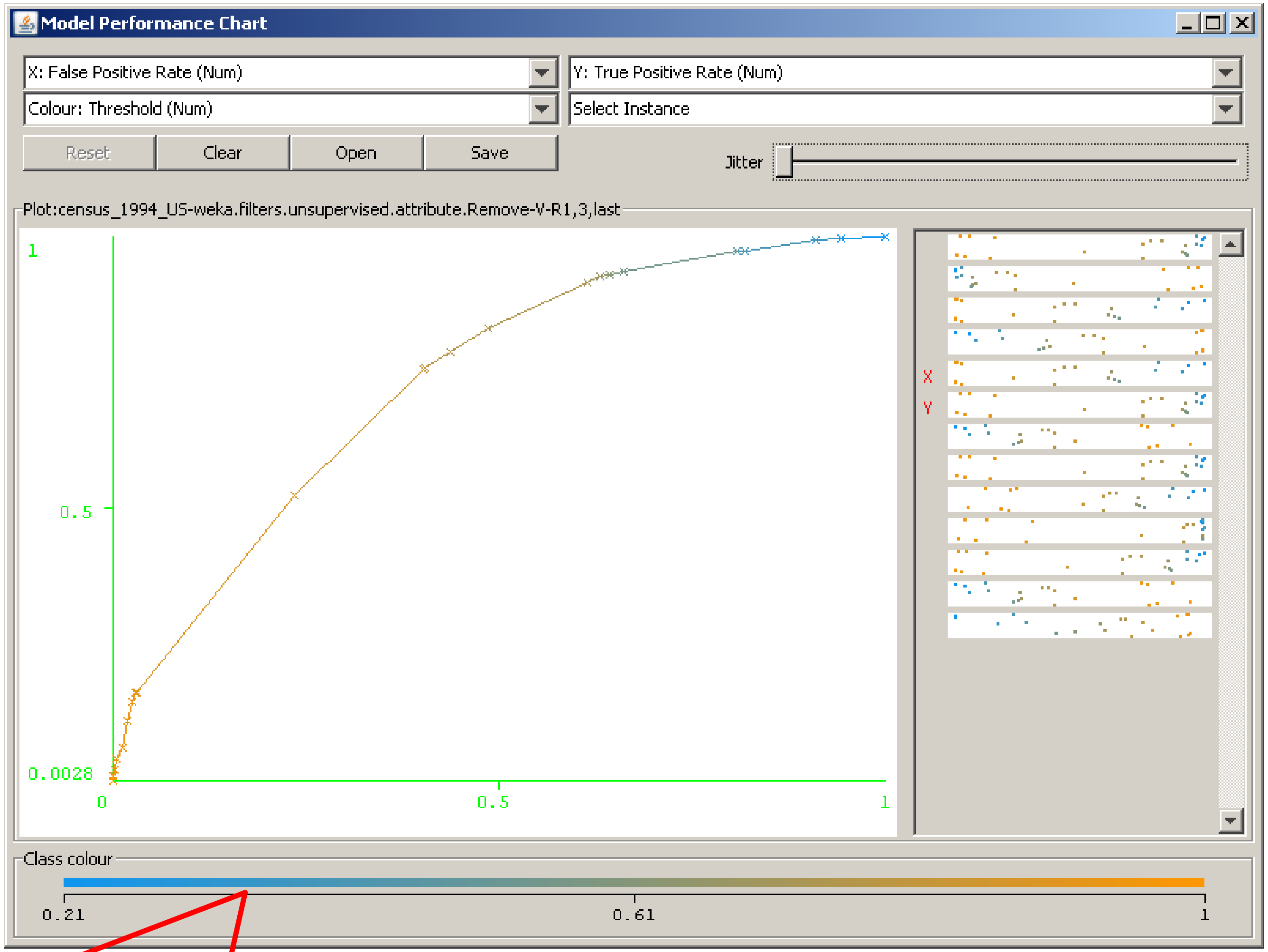
Selecting performance model: chart



Running the model



Show chart: View ROC curve



Model Performance Chart

X: False Positive Rate (Num)

Y: True Positive Rate (Num)

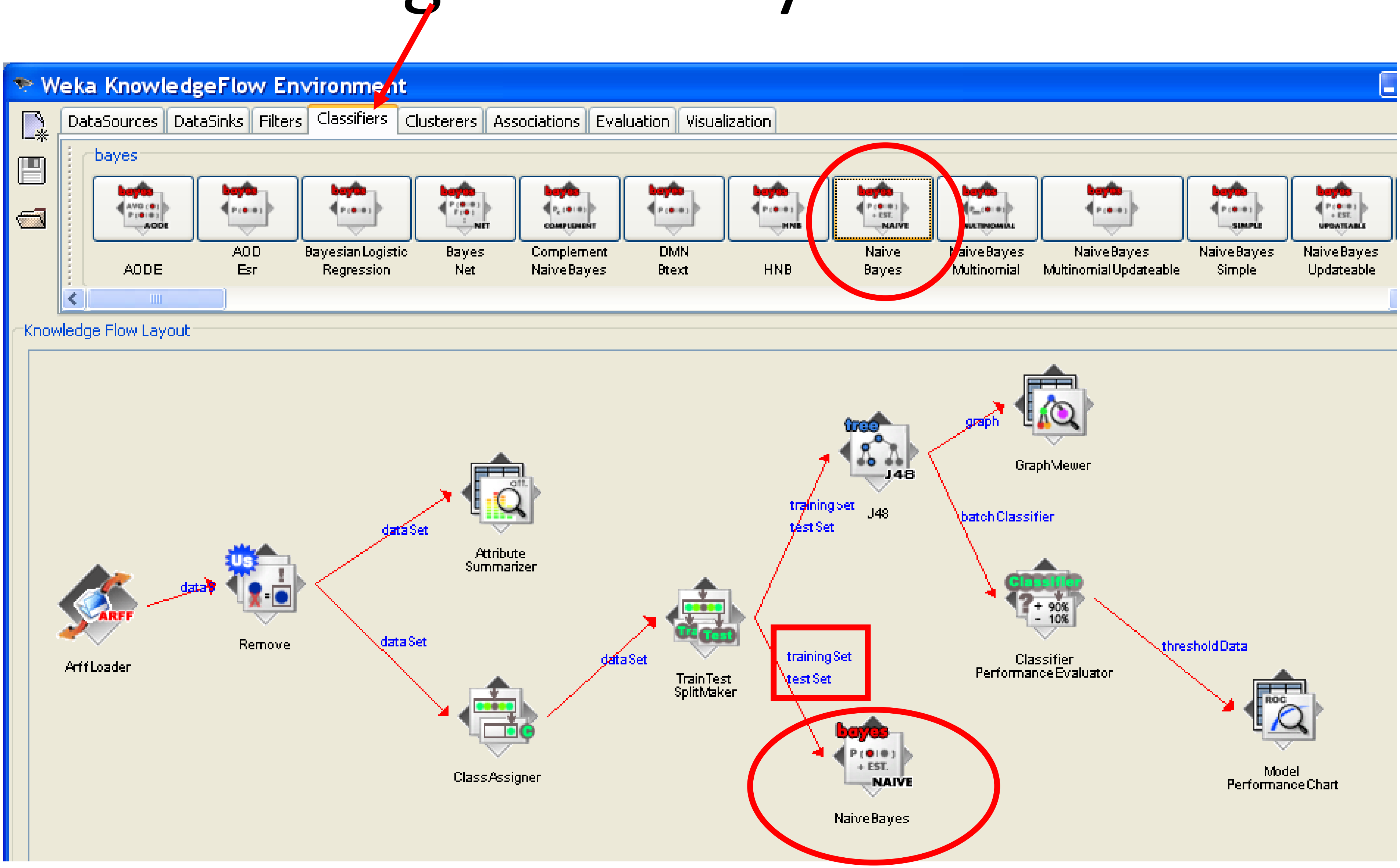
Colour: Threshold (Num)

Select Instance

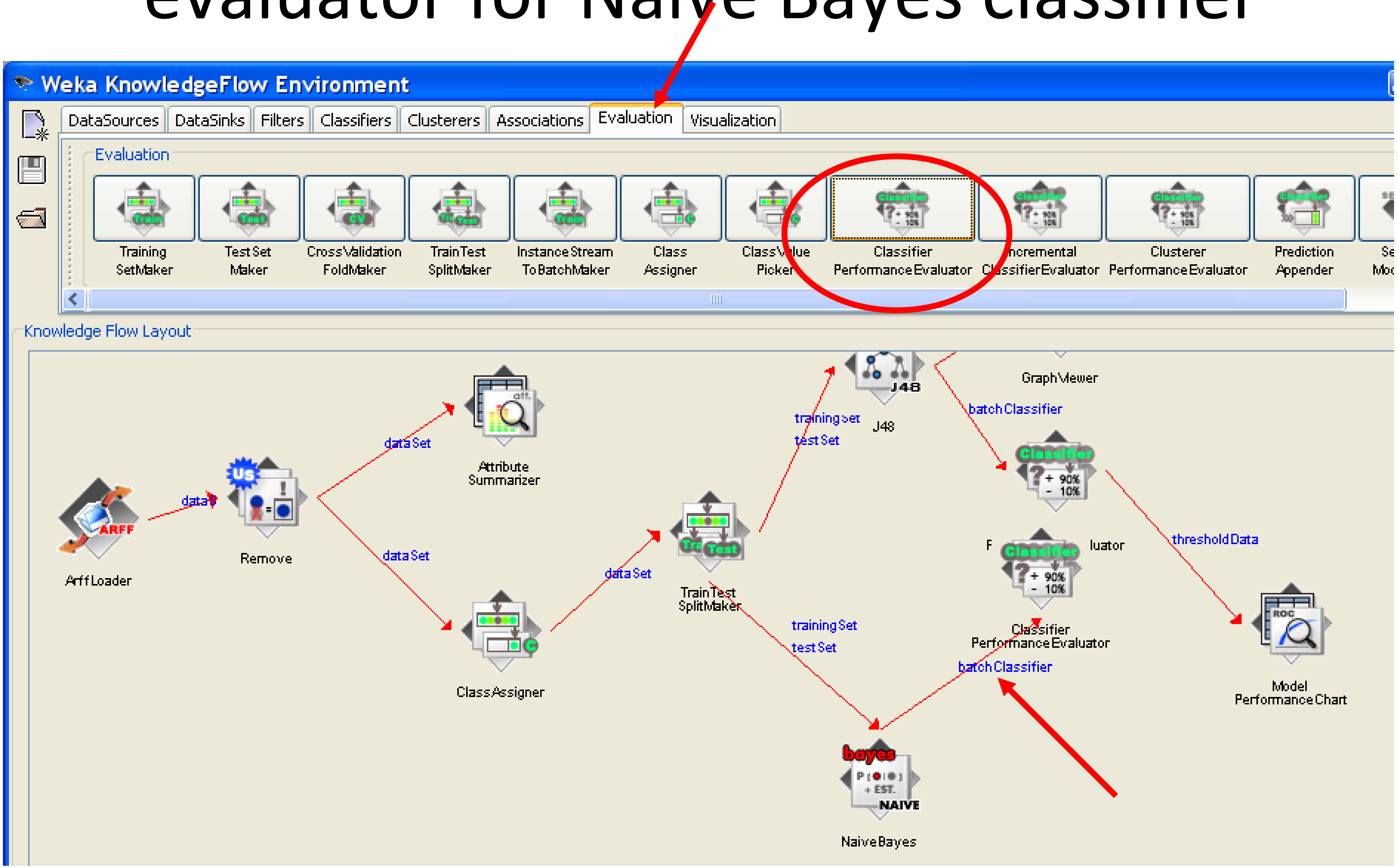
Plot:census_1994_US-weka.filters.unsupervised.attribute.Remove-V-R1,3,last

Threshold value for dividing positives from negatives

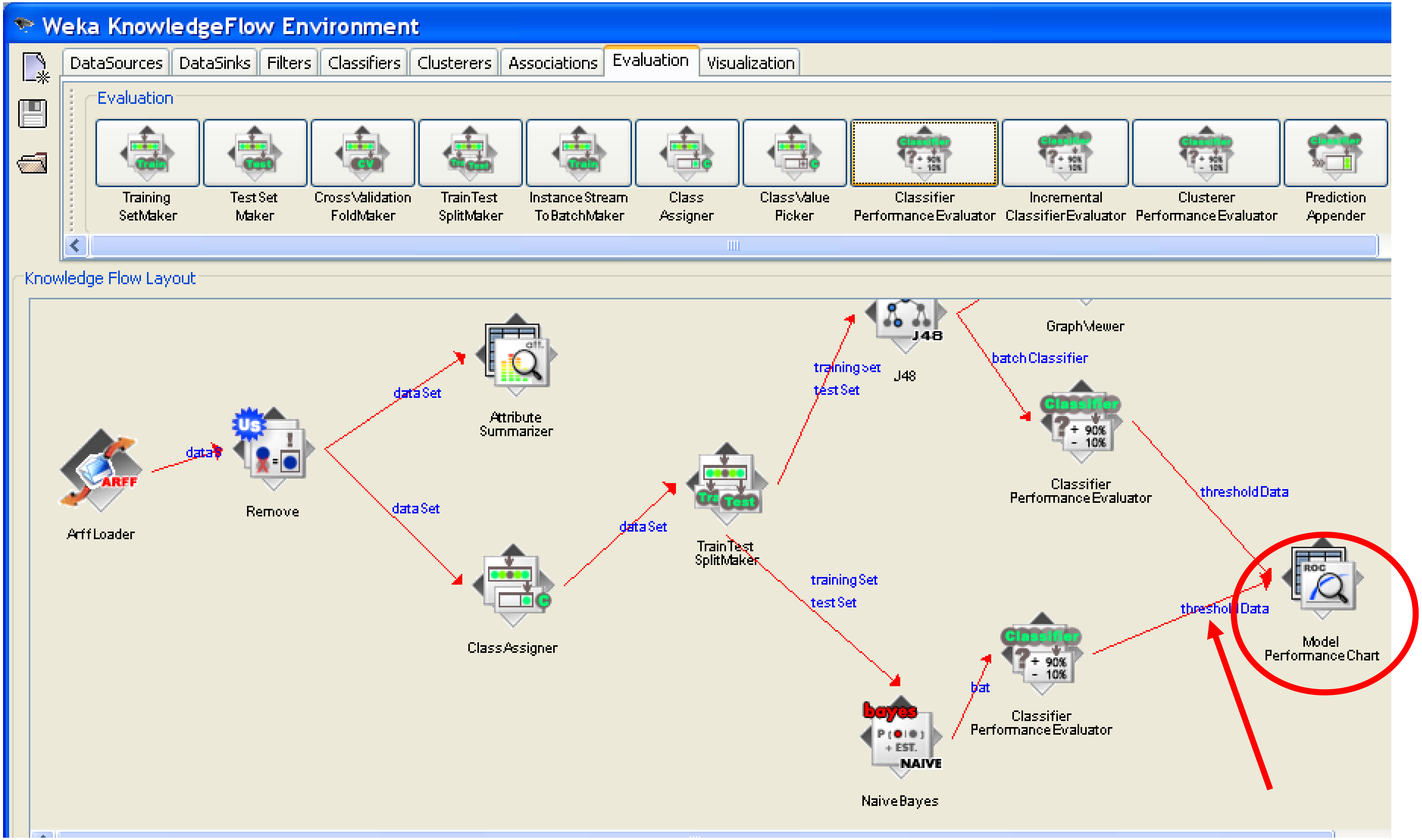
Adding Naïve Bayes classifier



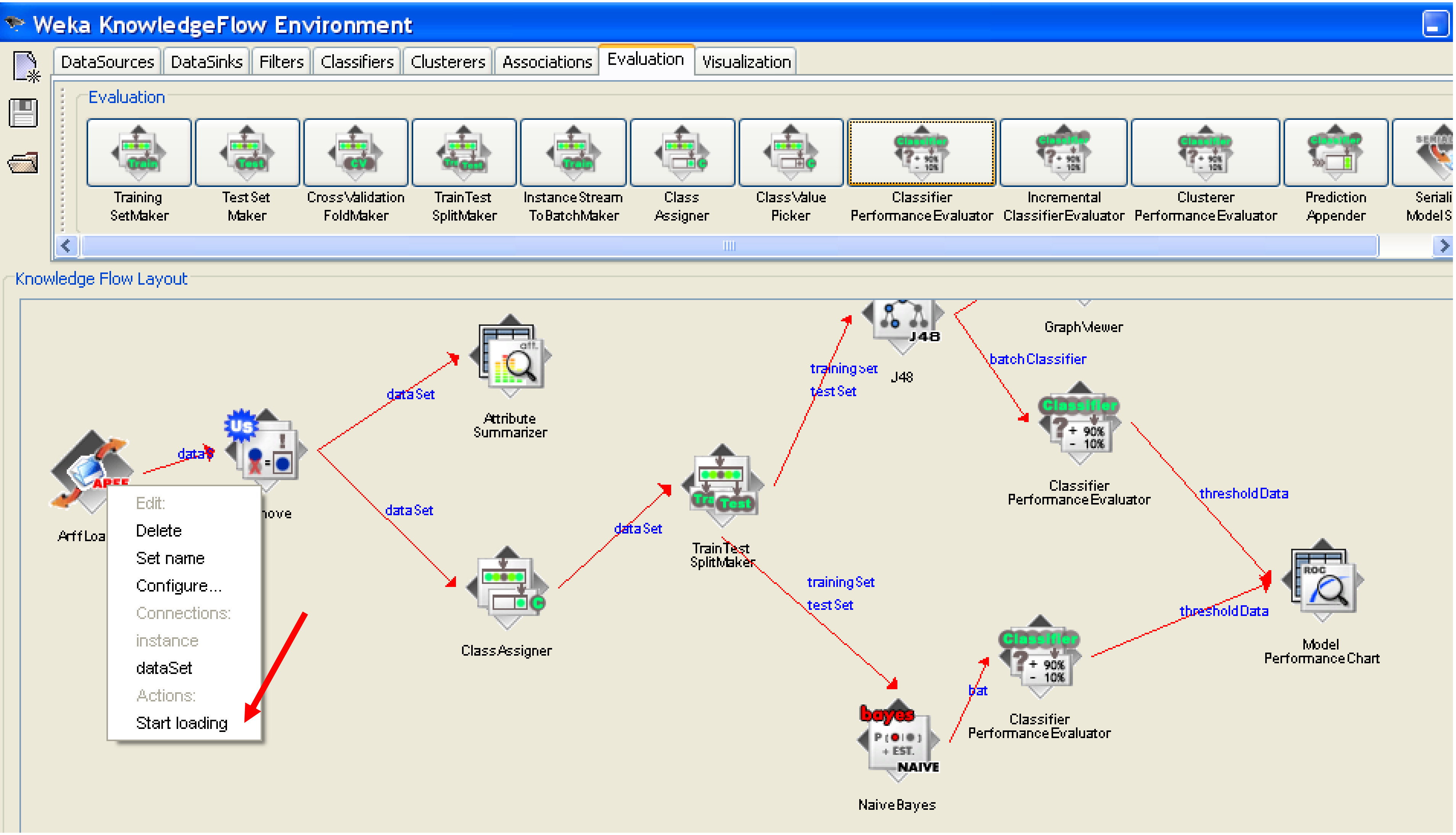
Adding separate performance evaluator for Naïve Bayes classifier



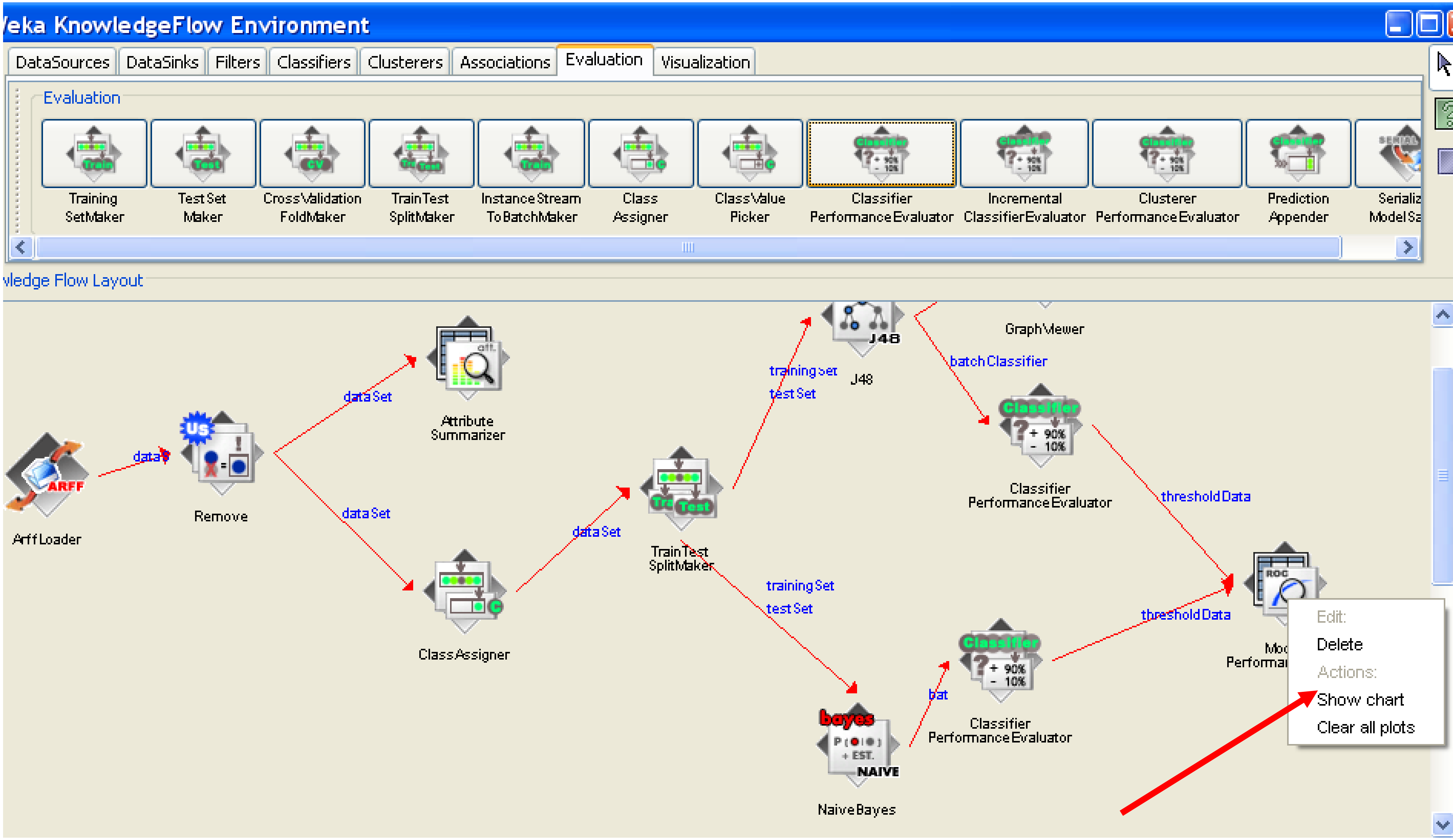
Connecting second performance evaluator to the same Model Performance Chart



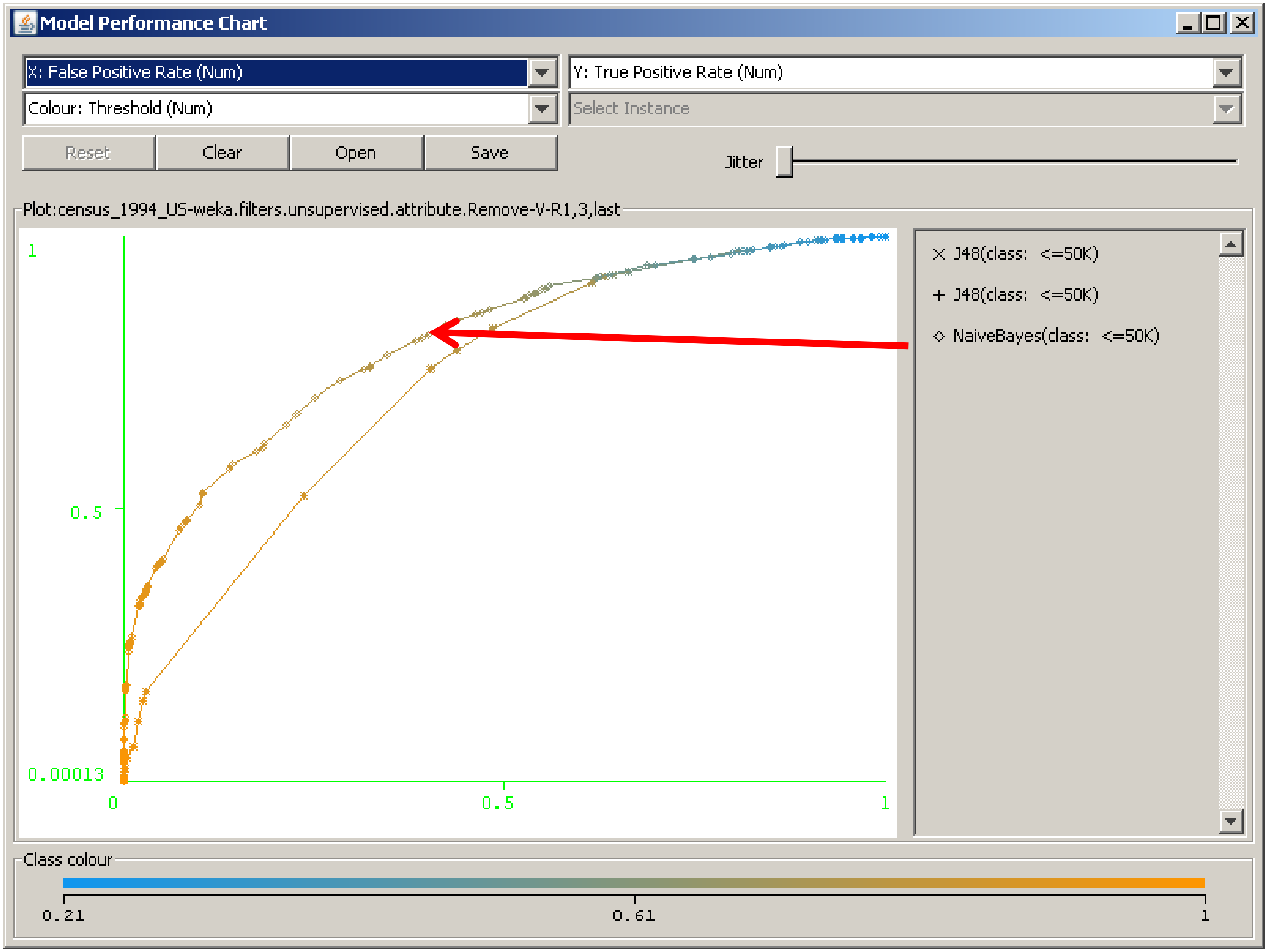
Run both classifiers



View ROC curves for both classifiers



Compare classifiers using their ROC curves



Model Performance Chart

X: False Positive Rate (Num)

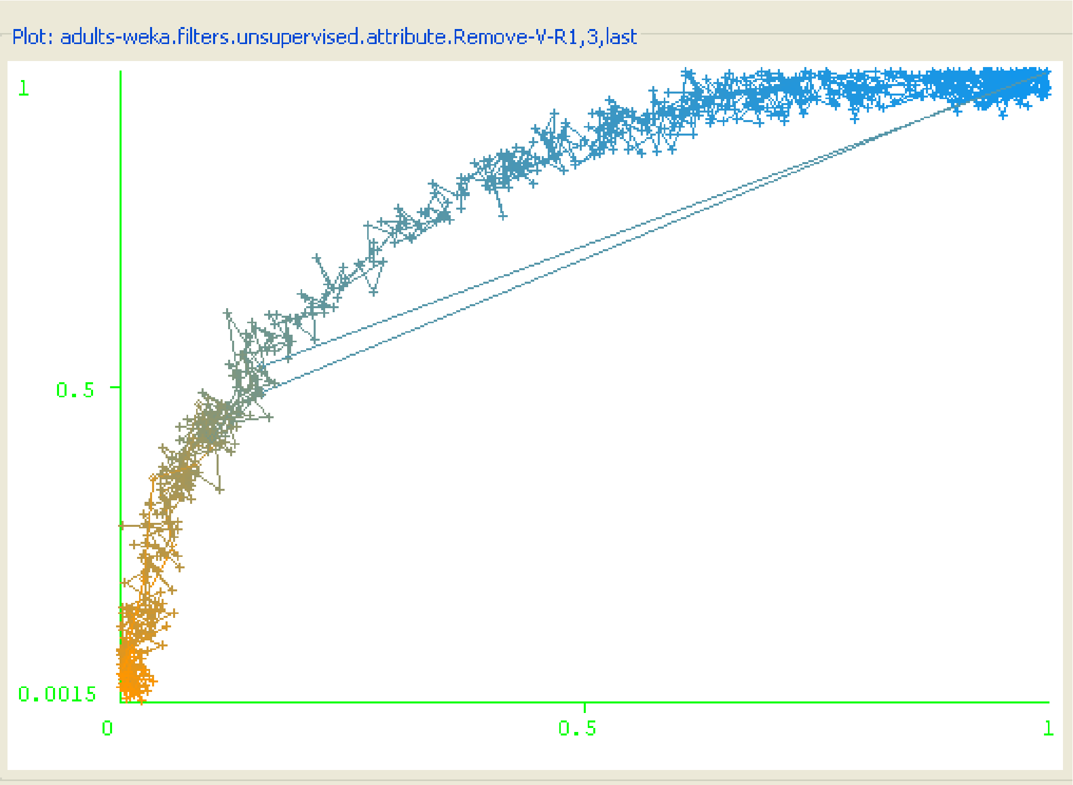
Y: True Positive Rate (Num)

Colour: Threshold (Num)

Select Instance

Plot:census_1994_US-weka.filters.unsupervised.attribute.Remove-V-R1,3,last

How good is the classifier

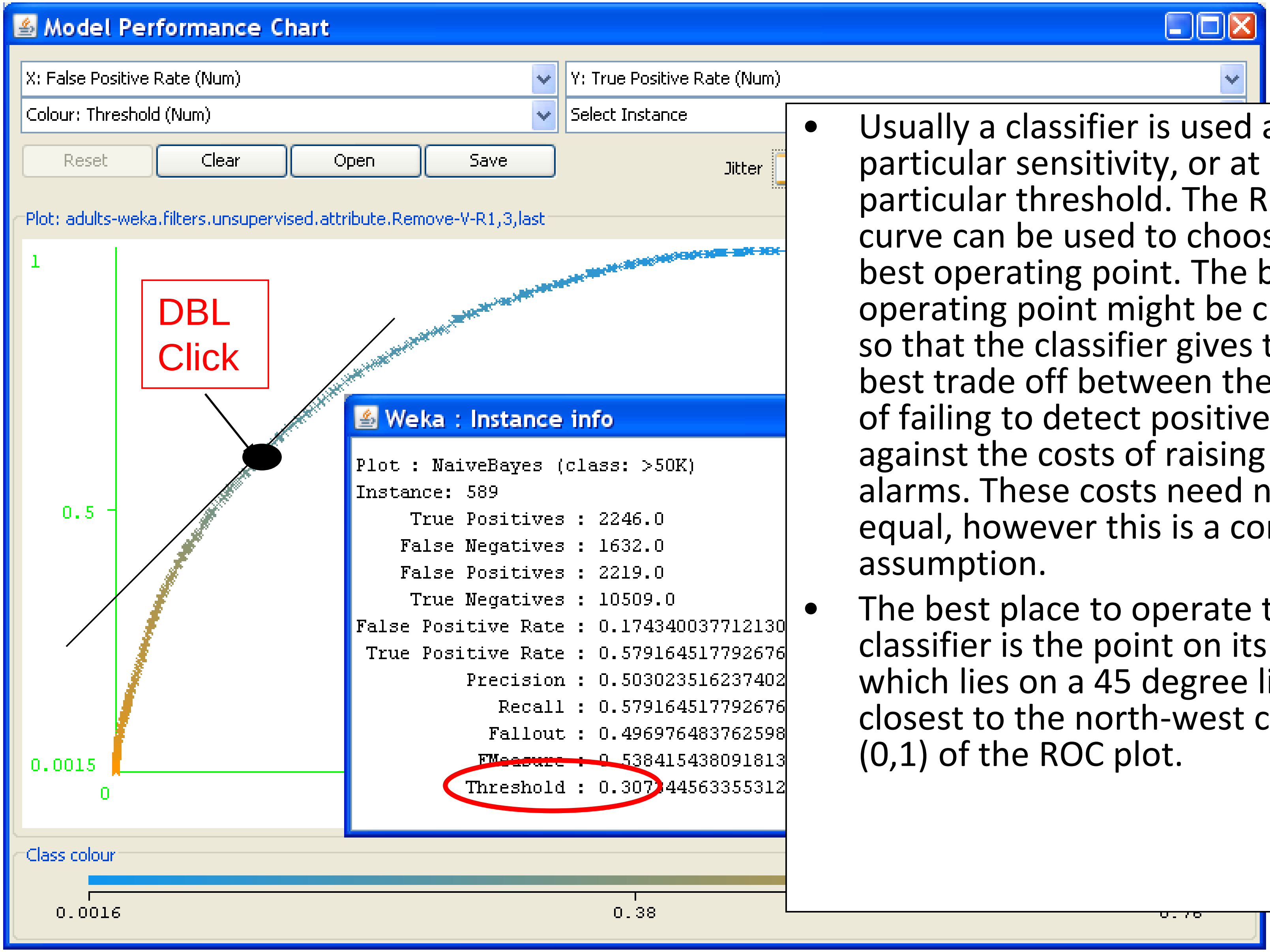


Plot: adults-weka.filters.unsupervised.attribute.Remove-V-R1,3,last

The area under the ROC curve shows the quality of a classifier – not accuracy, but the ability to separate between positive and negative instances.

What classifier is better?

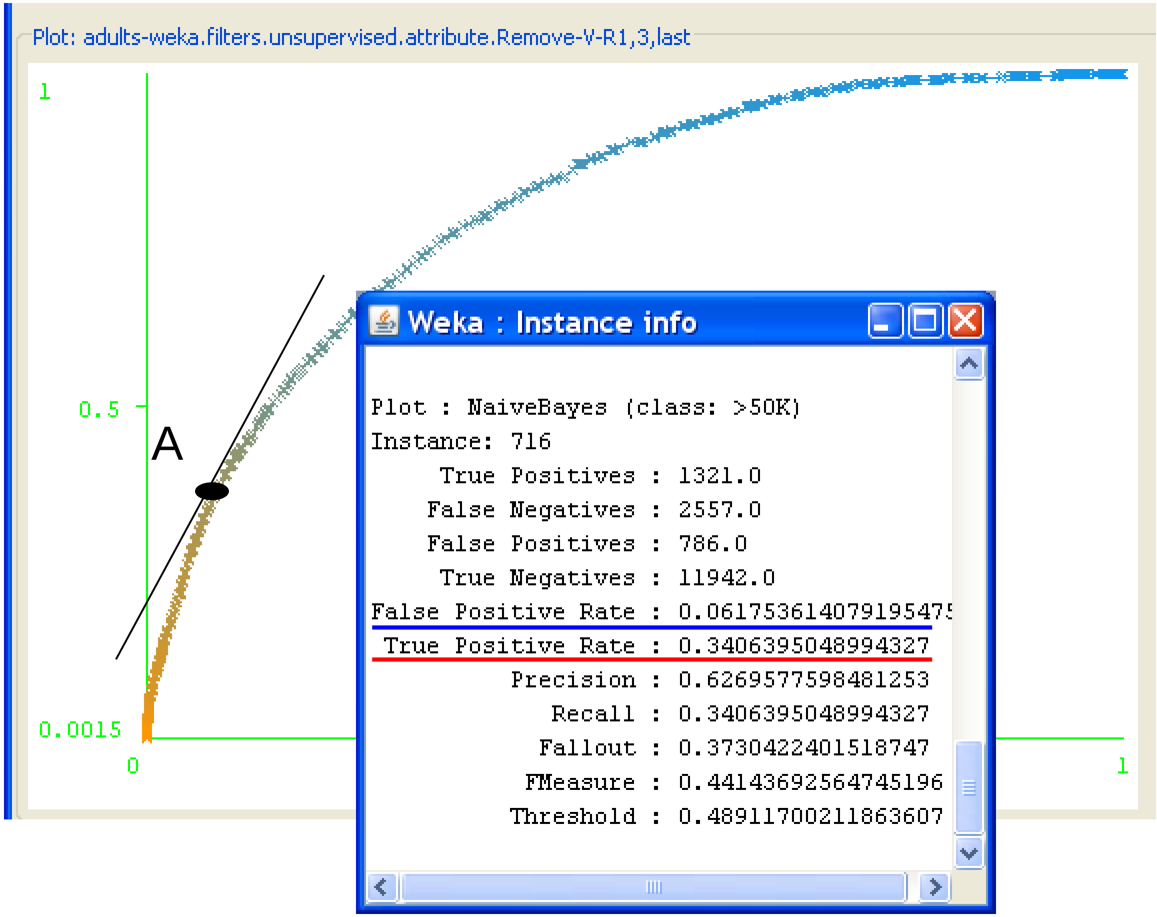
Choosing the Operating Point



```
Plot : NaiveBayes (class: >50K)
Instance: 589
     True Positives : 2246.0
    False Negatives : 1632.0
    False Positives : 2219.0
     True Negatives : 10509.0
False Positive Rate : 0.174340037712130
 True Positive Rate : 0.579164517792676
          Precision : 0.503023516237402
             Recall : 0.579164517792676
            Fallout : 0.496976483762598
           FMeasure : 0.538415438091813
          Threshold : 0.307344563355312
```

- Usually a classifier is used at a particular sensitivity, or at a particular threshold. The ROC curve can be used to choose the best operating point. The best operating point might be chosen so that the classifier gives the best trade off between the costs of failing to detect positives against the costs of raising false alarms. These costs need not be equal, however this is a common assumption.
- The best place to operate the classifier is the point on its ROC which lies on a 45 degree line closest to the north-west corner (0,1) of the ROC plot.

Cost sensitive operating points


Plot: adults-weka.filters.unsupervised.attribute.Remove-V-R1,3,last
1
0.5
0.0015
0
1
A
Weka : Instance info

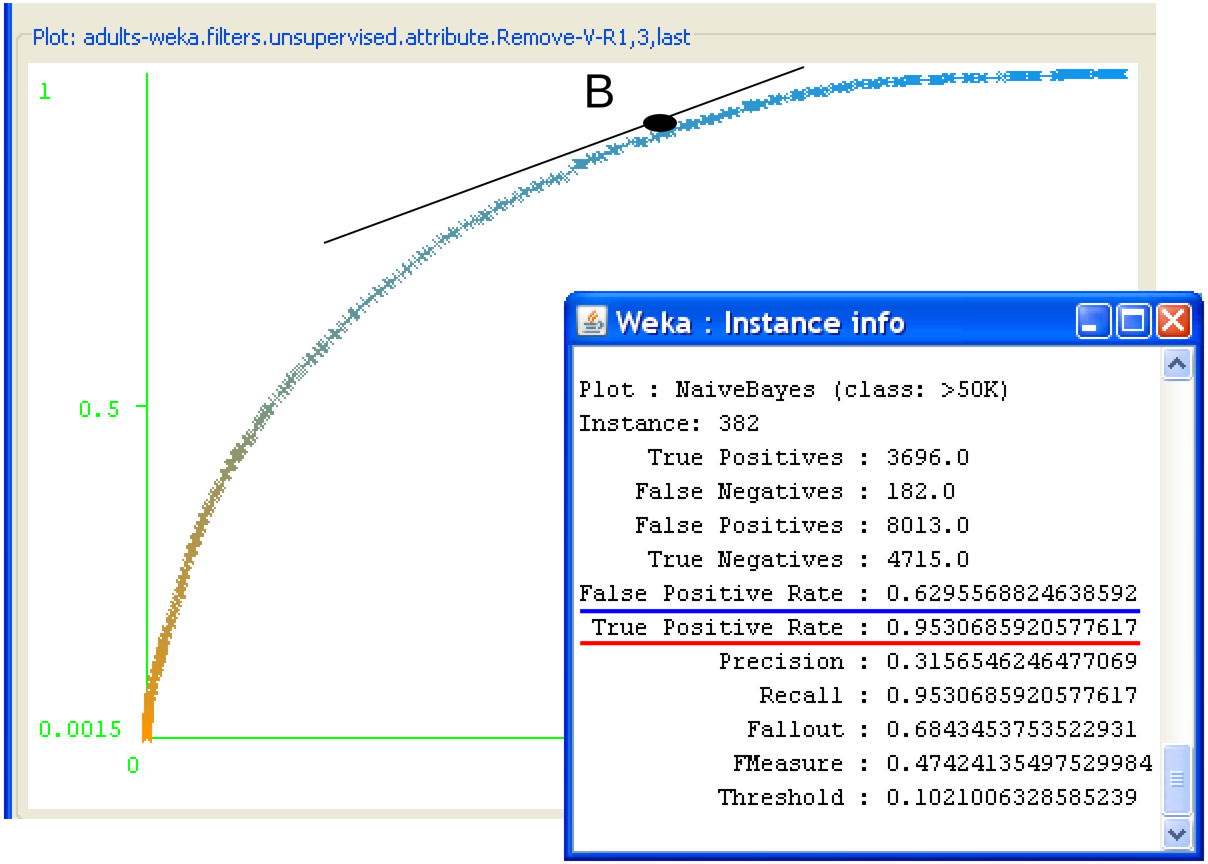
```
Plot : NaiveBayes (class: >50K)
Instance: 716
     True Positives : 1321.0
    False Negatives : 2557.0
    False Positives : 786.0
     True Negatives : 11942.0
False Positive Rate : 0.06175361407919547
 True Positive Rate : 0.3406395048994327
          Precision : 0.6269577598481253
             Recall : 0.3406395048994327
            Fallout : 0.3730422401518747
           FMeasure : 0.44143692564745196
          Threshold : 0.48911700211863607
```

Is this threshold good :

for cancer detection?

for targeting potential customers?

Cost sensitive operating points


Plot: adults-weka.filters.unsupervised.attribute.Remove-V-R1,3,last
B
1
0.5
0.0015
0
Weka : Instance info

```
Plot : NaiveBayes (class: >50K)
Instance: 382
     True Positives : 3696.0
    False Negatives : 182.0
    False Positives : 8013.0
     True Negatives : 4715.0
False Positive Rate : 0.6295568824638592
 True Positive Rate : 0.9530685920577617
          Precision : 0.3156546246477069
             Recall : 0.9530685920577617
            Fallout : 0.6843453753522931
           FMeasure : 0.47424135497529984
          Threshold : 0.1021006328585239
```

Is this threshold good :

for cancer detection?

for targeting potential customers?

Conclusions

- WEKA is a powerful datamining tool with the state-of-the art GUI, but is not very easy to use

- There are other open source data mining tools:
  - Orange:
    - http://www.ailab.si/orange
  - Tanagra:
    - http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html