Introduction to WEKA

Nils Schaetti

nils.schaetti@unine.ch

October 2rd, 2017

What is WEKA?

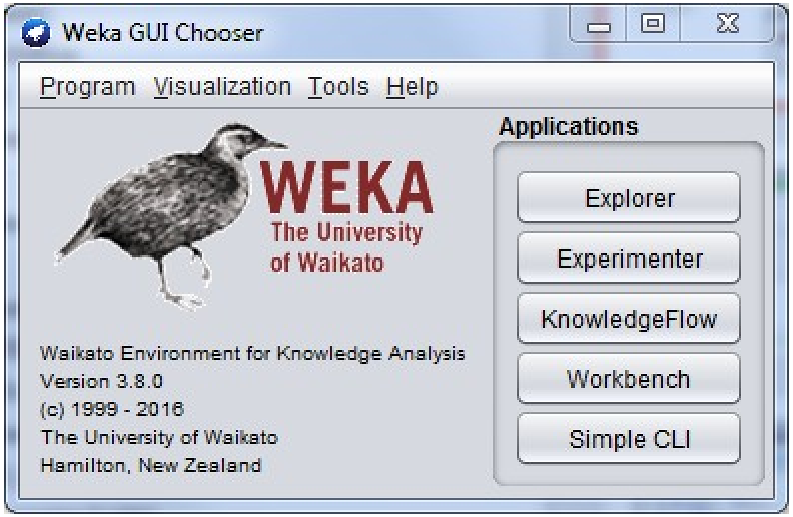
- A flightless bird found only in New Zealand

- Collection of ML algorithms
  - Pre-processing
  - Classifiers
  - Clustering
  - Regression
  - Association rule
  - Visualization

Use WEKA

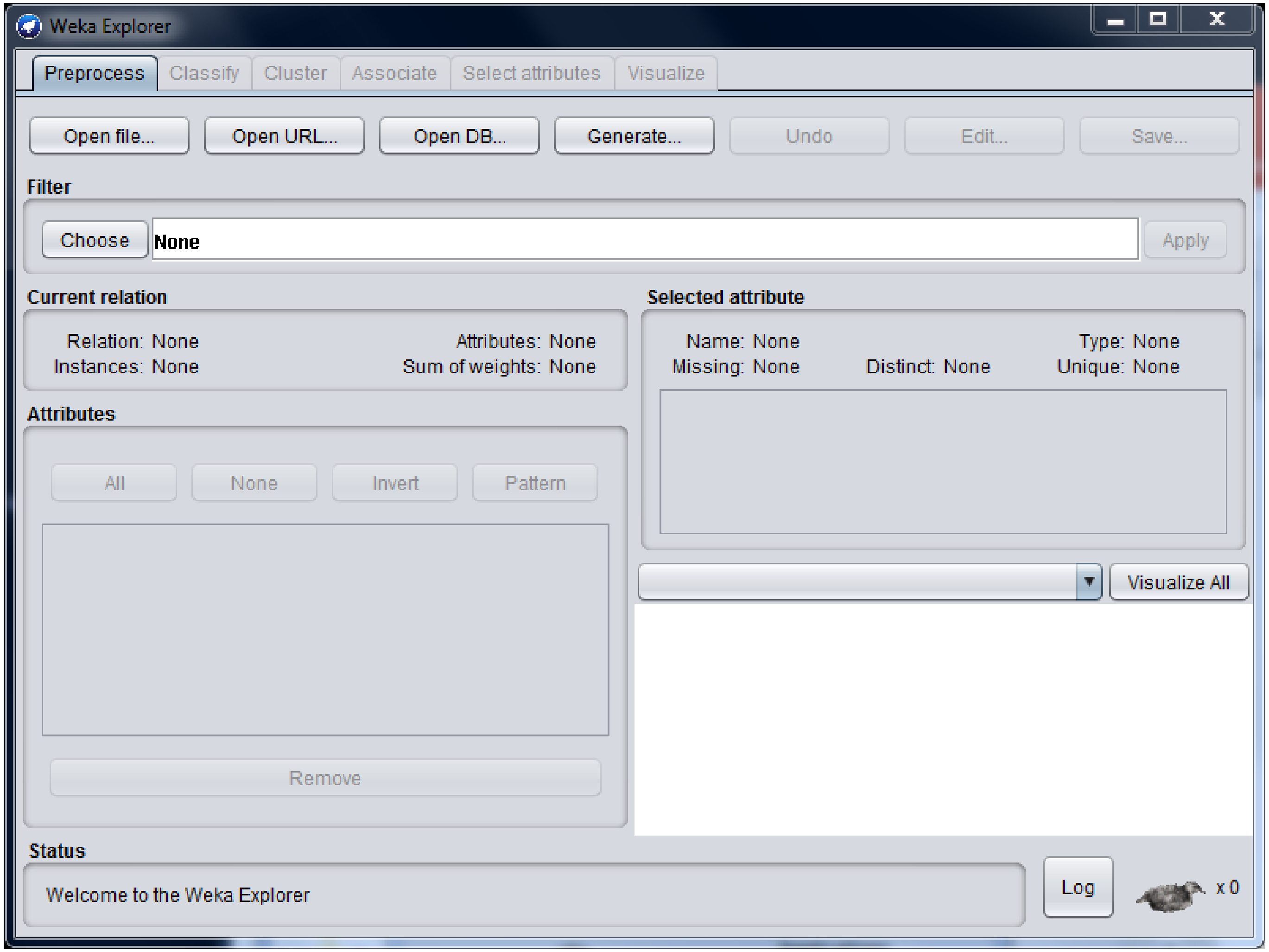
- Download:
  - http://www.cs.waikato.ac.nz/ml/weka/downloading.html

- GUI WEKA Chooser
  - The Explorer



Waikato Environment for Knowledge Analysis
Version 3.8.0
(c) 1999 - 2016
The University of Waikato
Hamilton, New Zealand

WEKA Explorer



Filter

None

Current relation

Relation: None
Instances: None
Attributes: None
Sum of weights: None

Selected attribute

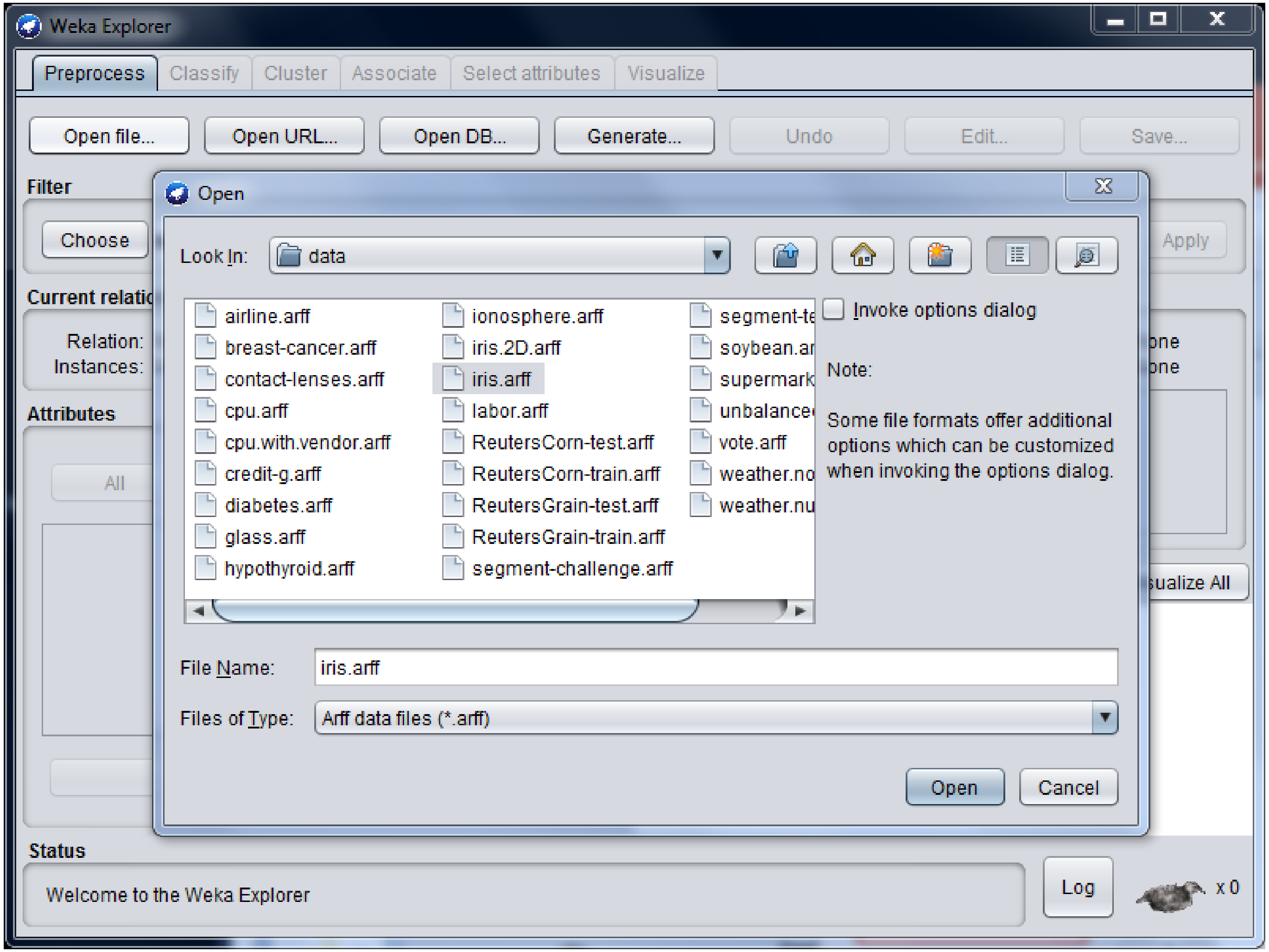
Name: None
Missing: None
Distinct: None
Type: None
Unique: None

Attributes

Status

Welcome to the Weka Explorer

WEKA Explorer



Open

Look In: data

airline.arff	ionosphere.arff	segment-te
breast-cancer.arff	iris.2D.arff	soybean.ar
contact-lenses.arff	iris.arff	supermark
cpu.arff	labor.arff	unbalance
cpu.with.vendor.arff	ReutersCorn-test.arff	vote.arff
credit-g.arff	ReutersCorn-train.arff	weather.no
diabetes.arff	ReutersGrain-test.arff	weather.nu
glass.arff	ReutersGrain-train.arff	
hypothyroid.arff	segment-challenge.arff	

□ Invoke options dialog

Note:

Some file formats offer additional options which can be customized when invoking the options dialog.

File Name: iris.arff

Files of Type: Arff data files (*.arff)

Status

Welcome to the Weka Explorer

Preparing the Data

- ARFF (Attribute-Relation File Format)
- Text file with tags and attributes
- Sections for header and data
- Comments start with %

ARFF Header

- Name of dataset
  - @relation <relation-name>

```
@RELATION iris
```

- List of attributes
  - @attribute <attribute-name> <datatype>

```
@ATTRIBUTE sepallength NUMERIC
…
@ATTRIBUTE class {Iris-setosa,Iris-versicolor,Iris-virginica}
```

Preparing the Data

- Tags (@relation, @attribute, @data) are case insensitive
- Attributes are case sensitive
- Strings with space must be quoted
- Order of attributes in header is column in data

ARFF Data

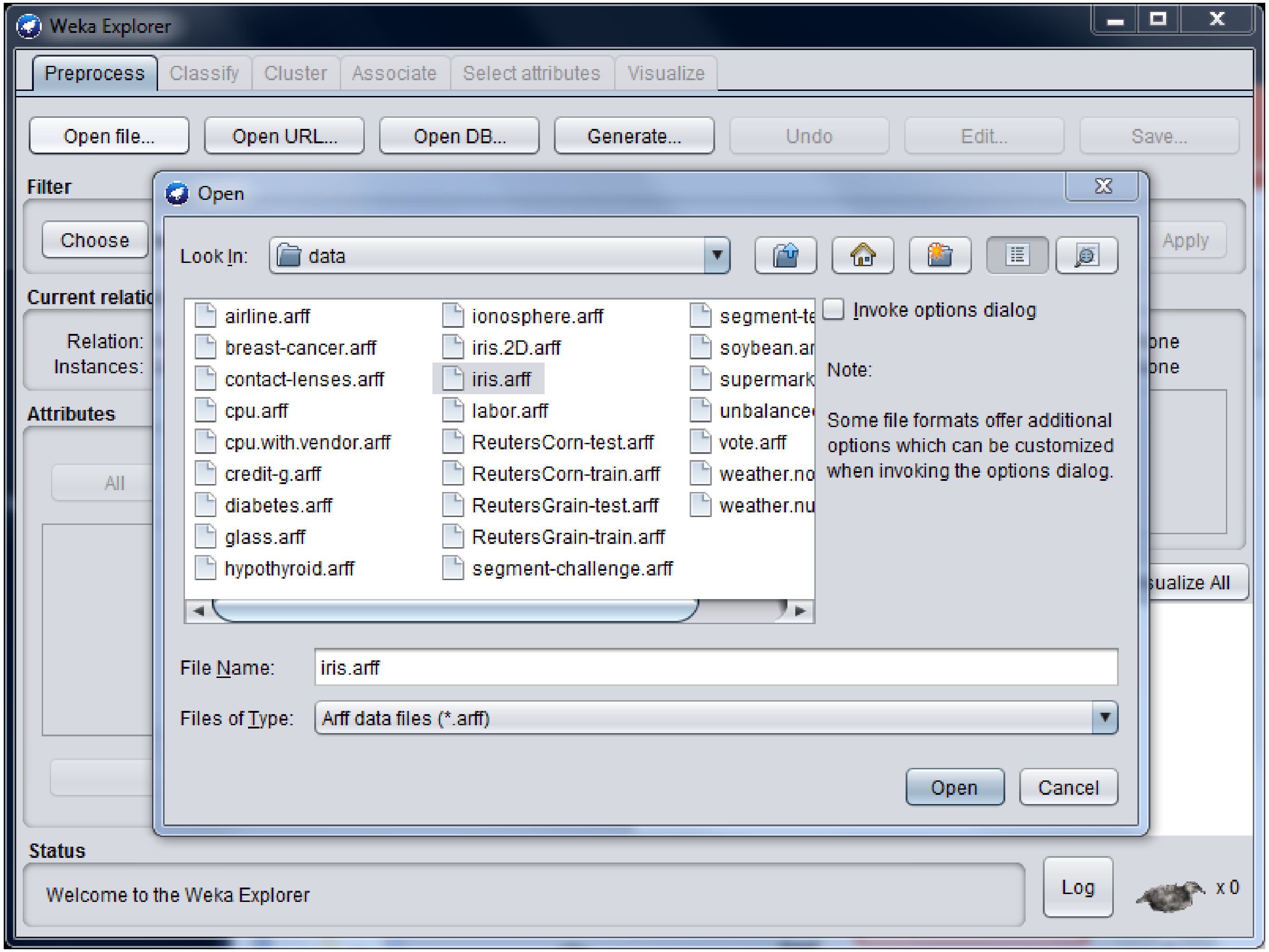
- @data: start of the data segment in the file
- One line for each sample
- Values separated by commas

```
@DATA
5.0, 3.3, 1.4, 0.2, Iris-setosa
5.4, 3.9, 1.3, 0.4, Iris-setosa
7.0, 3.2, 4.7, 1.4, Iris-versicolor
5.5, 2.6, 4.4, 1.2, Iris-versicolor
```

ARFF Example

```
% Iris Plants Database %
@RELATION iris

@ATTRIBUTE sepallength NUMERIC
@ATTRIBUTE sepalwidth NUMERIC
@ATTRIBUTE petallength NUMERIC
@ATTRIBUTE petalwidth NUMERIC
@ATTRIBUTE class {Iris-setosa, Iris-versicolor, Iris-virginica}

@DATA
5.1, 3.5, 1.4, 0.2, Iris-setosa
4.9, 3.0, 1.4, 0.2, Iris-setosa
4.7, 3.2, 1.3, 0.2, Iris-setosa
…
```

WEKA Explorer



Look In: data

□ Invoke options dialog

Note:

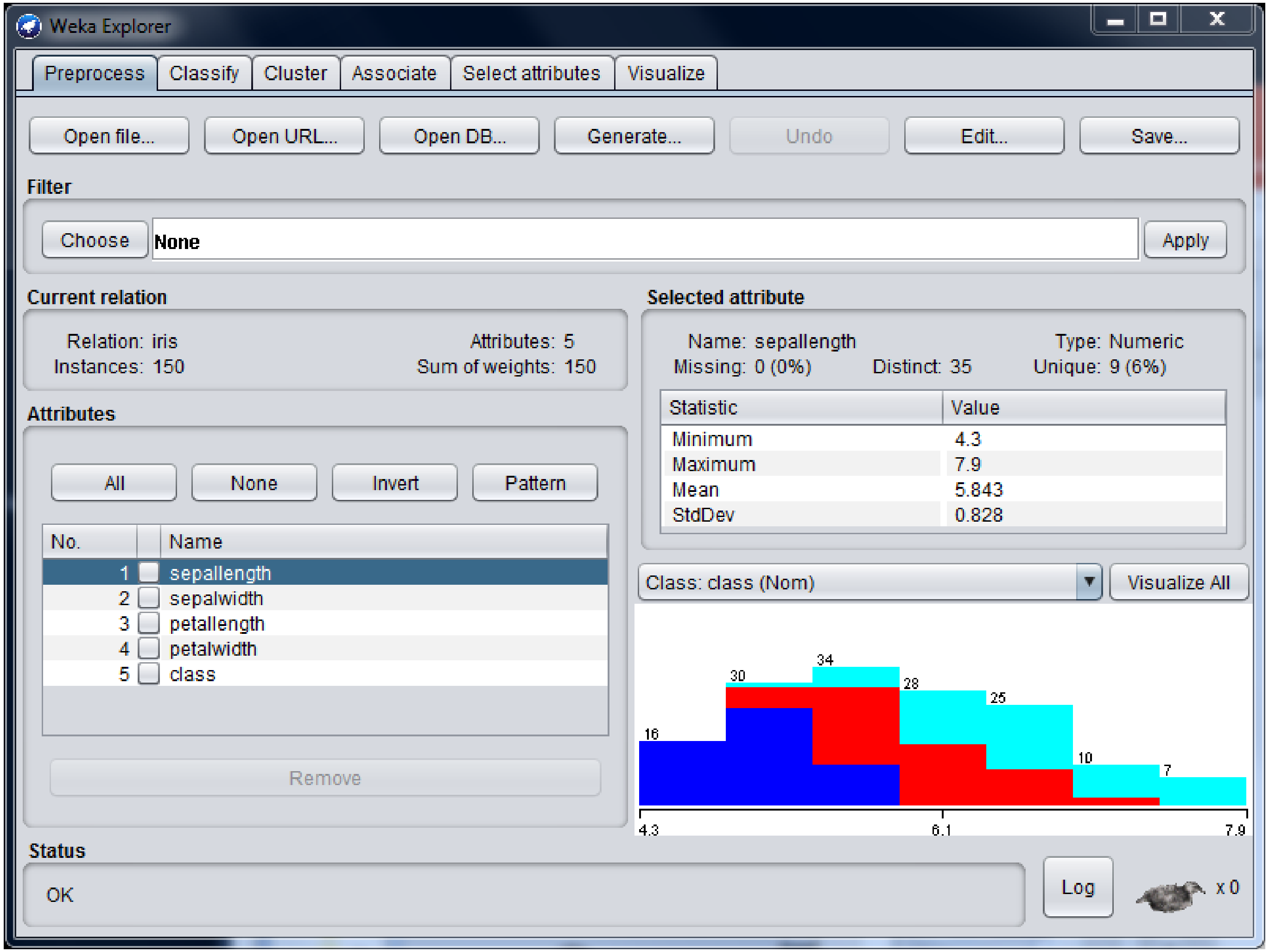
Some file formats offer additional options which can be customized when invoking the options dialog.

File Name: iris.arff

Files of Type: Arff data files (*.arff)

Welcome to the Weka Explorer

WEKA Explorer - Preprocess



Filter

None

Current relation

Relation: iris
Attributes: 5
Instances: 150
Sum of weights: 150

Selected attribute

Name: sepallength
Type: Numeric
Missing: 0 (0%)
Distinct: 35
Unique: 9 (6%)

Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

Attributes

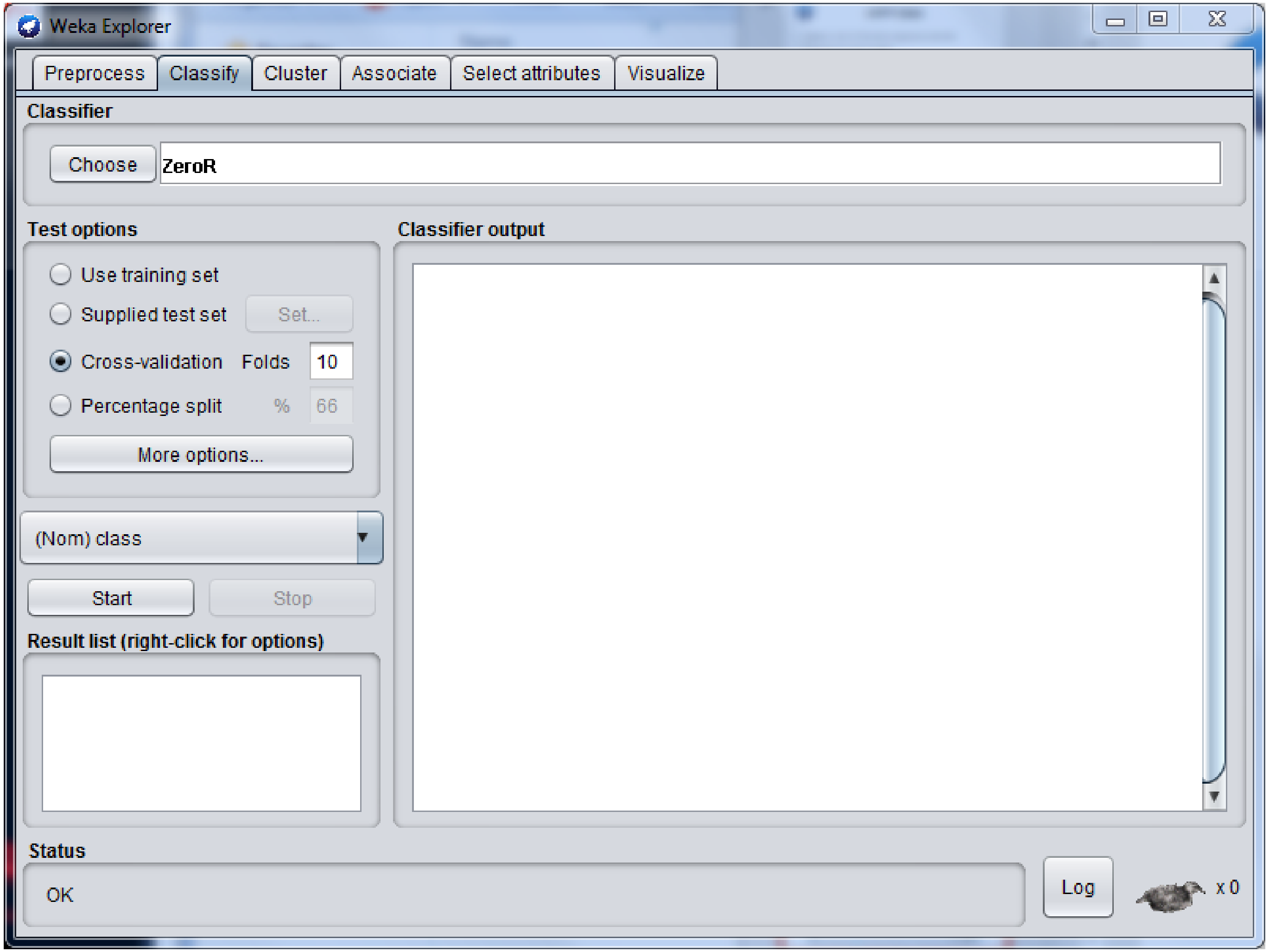
No.		Name
1	□	sepallength
2	□	sepalwidth
3	□	petallength
4	□	petalwidth
5	□	class

Class: class (Nom)

Status

OK

WEKA Explorer - Classify



Classifier

ZeroR

Test options

○ Use training set

○ Supplied test set

● Cross-validation Folds 10

○ Percentage split % 66

(Nom) class

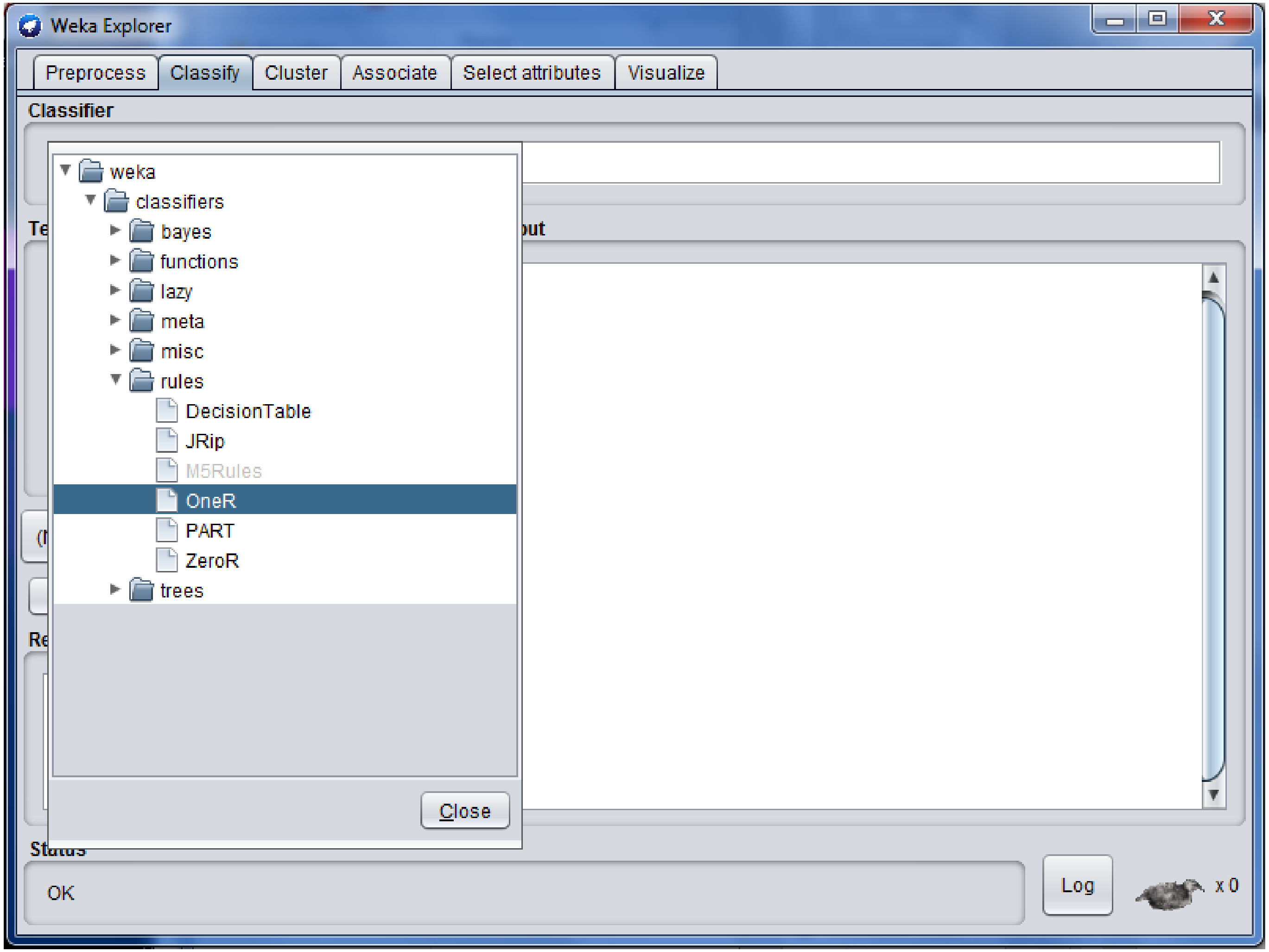
Result list (right-click for options)

Classifier output

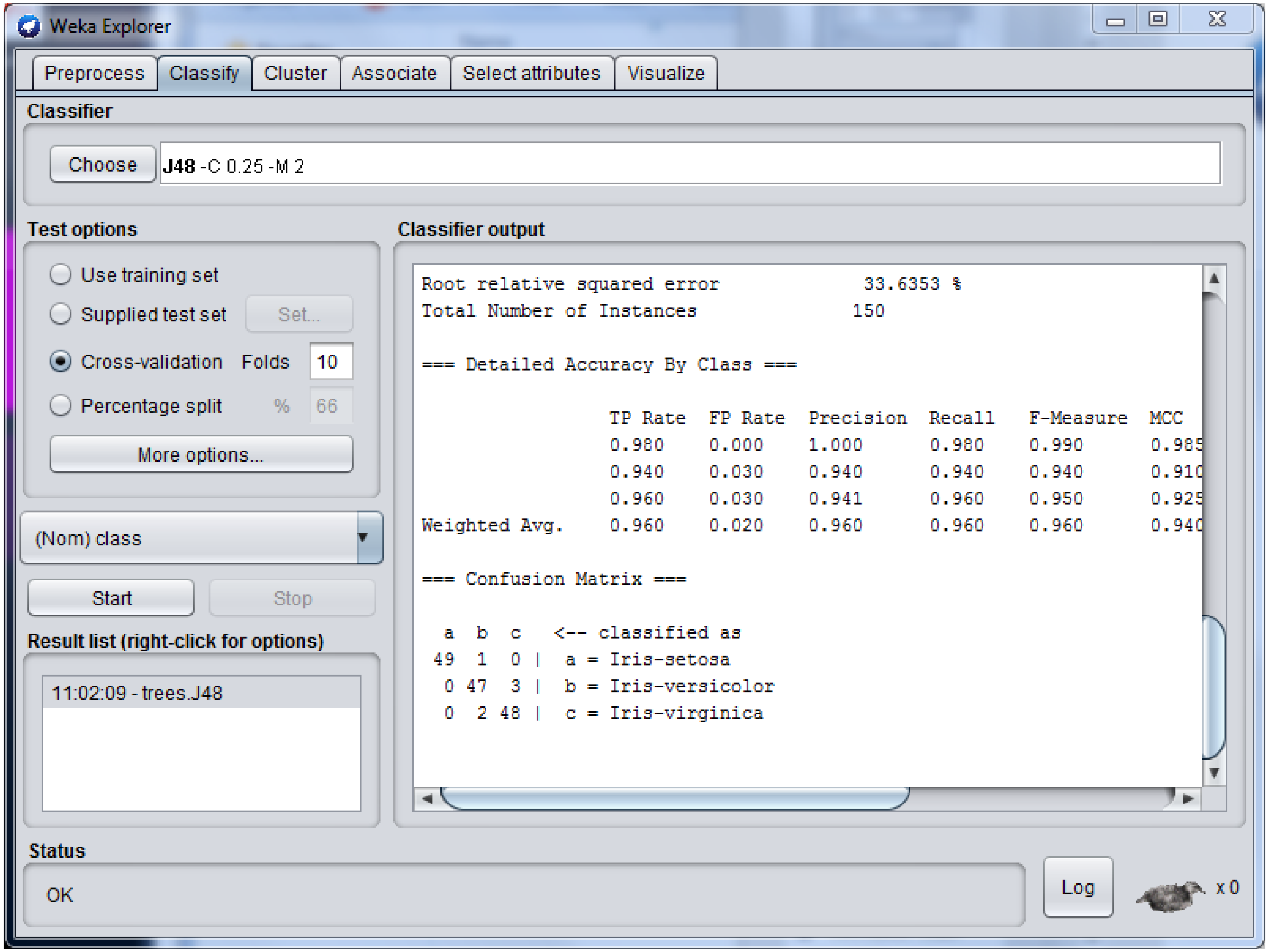
Status

OK

WEKA Explorer - Classify



WEKA Explorer - Classify



J48 -C 0.25 -M 2

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

(Nom) class

11:02:09 - trees.J48

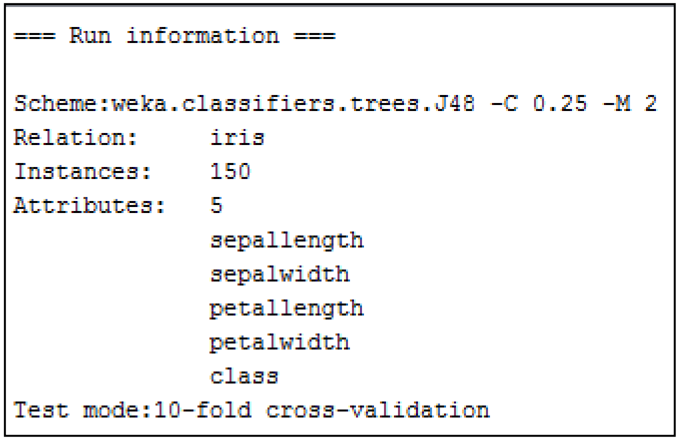
OK

```
Root relative squared error             33.6353 %
Total Number of Instances              150

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC
                 0.980    0.000    1.000      0.980    0.990      0.985
                 0.940    0.030    0.940      0.940    0.940      0.910
                 0.960    0.030    0.941      0.960    0.950      0.925
Weighted Avg.    0.960    0.020    0.960      0.960    0.960      0.940

=== Confusion Matrix ===

  a  b  c   <-- classified as
 49  1  0 |  a = Iris-setosa
  0 47  3 |  b = Iris-versicolor
  0  2 48 |  c = Iris-virginica
```

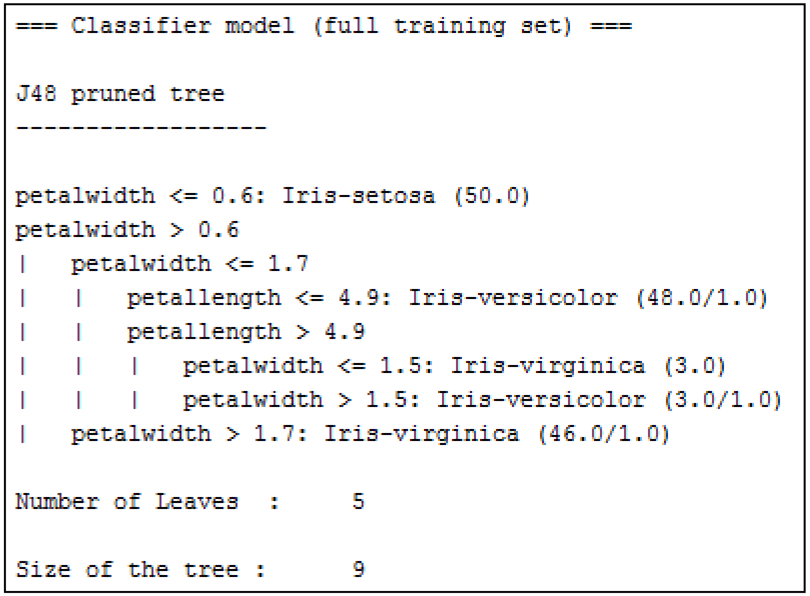
WEKA Explorer - Output

- Summary of the dataset

- Decision tree in textual form (if tree classifier)



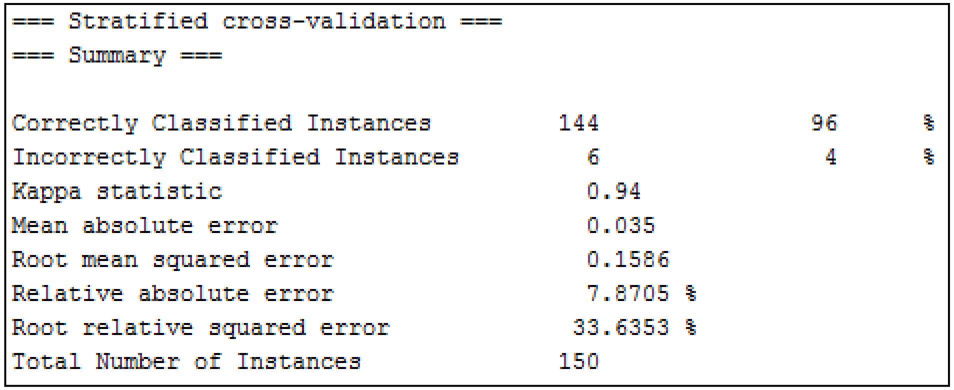
```
=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     iris
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
Test mode:10-fold cross-validation
```



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

J48 pruned tree
------------------

petalwidth <= 0.6: Iris-setosa (50.0)
petalwidth > 0.6
|   petalwidth <= 1.7
|   |   petallength <= 4.9: Iris-versicolor (48.0/1.0)
|   |   petallength > 4.9
|   |   |   petalwidth <= 1.5: Iris-virginica (3.0)
|   |   |   petalwidth > 1.5: Iris-versicolor (3.0/1.0)
|   petalwidth > 1.7: Iris-virginica (46.0/1.0)

Number of Leaves  :     5

Size of the tree :      9
```

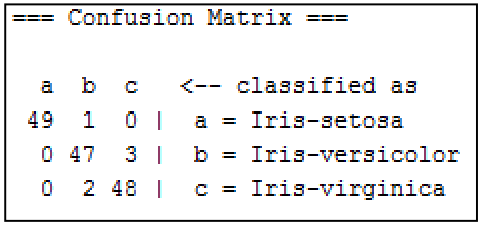
WEKA Explorer - Output

- Estimation of performance



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

Correctly Classified Instances         144               96      %
Incorrectly Classified Instances         6                4      %
Kappa statistic                          0.94
Mean absolute error                      0.035
Root mean squared error                  0.1586
Relative absolute error                  7.8705 %
Root relative squared error             33.6353 %
Total Number of Instances              150
```

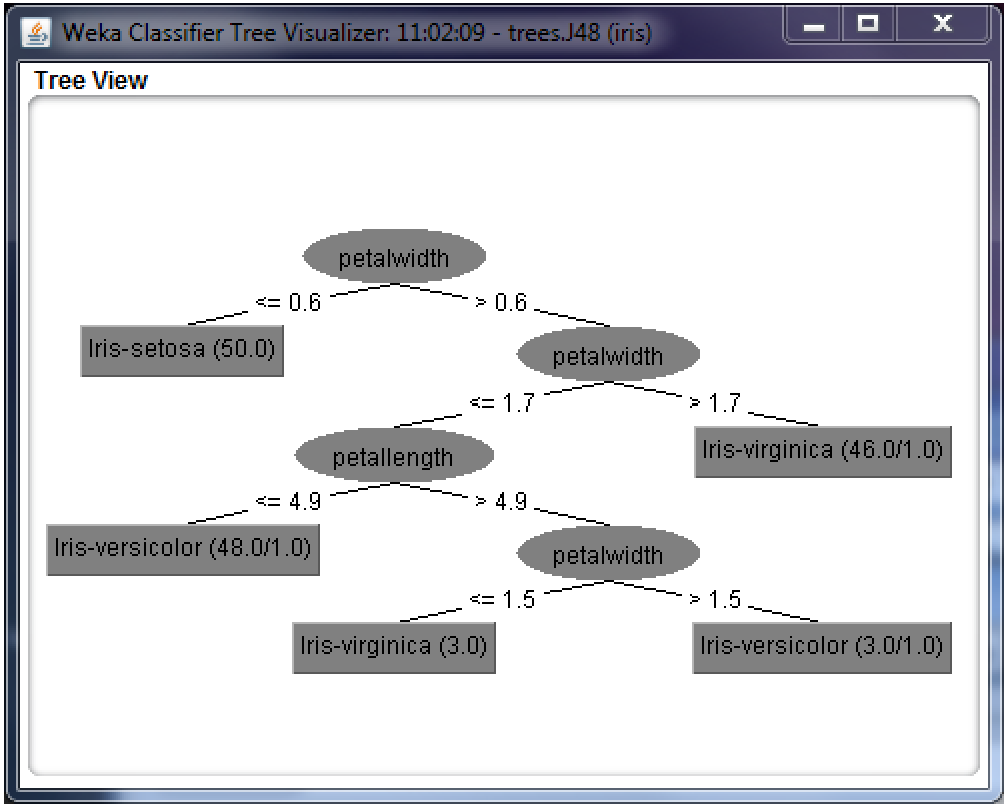
- Confusion matrix
  - Actual class in the row, predicted class in column



```
=== Confusion Matrix ===

  a  b  c   <-- classified as
 49  1  0 |  a = Iris-setosa
  0 47  3 |  b = Iris-versicolor
  0  2 48 |  c = Iris-virginica
```

WEKA Explorer - Output

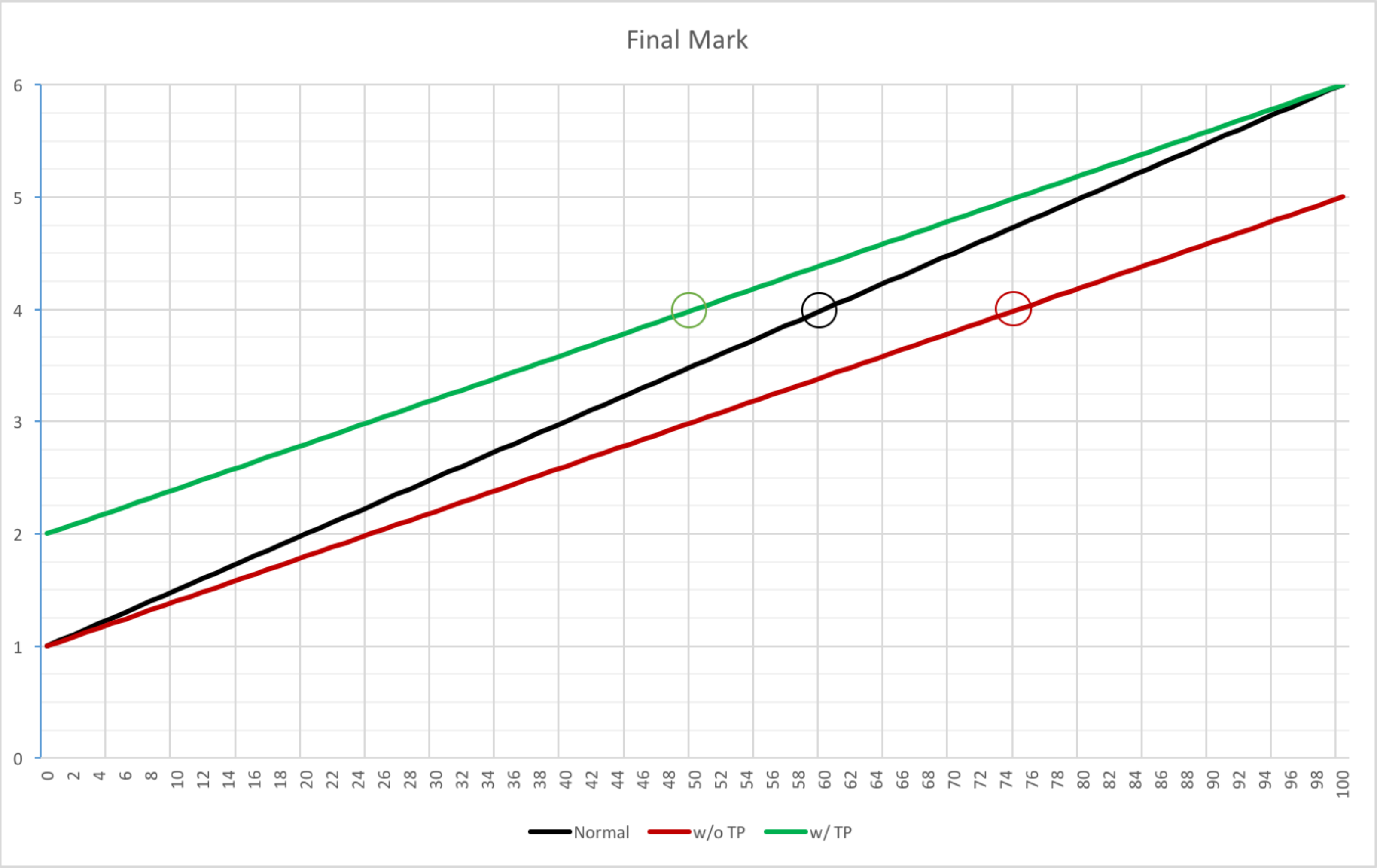
- Right click on entry in result list
  - Visualize tree (if tree classifier)
  - Save result buffer


Weka Classifier Tree Visualizer: 11:02:09 - trees.J48 (iris)
Tree View
petalwidth
<= 0.6
> 0.6
Iris-setosa (50.0)
petalwidth
<= 1.7
> 1.7
petallength
Iris-virginica (46.0/1.0)
<= 4.9
> 4.9
Iris-versicolor (48.0/1.0)
petalwidth
<= 1.5
> 1.5
Iris-virginica (3.0)
Iris-versicolor (3.0/1.0)

Problem Sets

- Each problem set gives between 0 and 10 points
- You need 80 points for the final exercise point
- Final exercise point is binary!
- Read the exercises carefully and answer to all parts
- You have one week to solve them
- Questions or problems? Write an email
- Time extend? Ask in advance
- Solve the exercises individually, not in groups

Problem Sets



Problem Set 01

- Solve a decision problem by hand with 1R
- Decide for a specific sample
- Transform the data to ARFF
- Use WEKA to check

- Deadline: October 8th, 2017 at 23:59

Questions?