



# Introduction to WEKA

---

Nils Schaetti

[nils.schaetti@unine.ch](mailto:nils.schaetti@unine.ch)

October 2<sup>rd</sup>, 2017



# What is WEKA?

- A flightless bird found only in New Zealand
- Collection of ML algorithms
  - Pre-processing
  - Classifiers
  - Clustering
  - Regression
  - Association rule
  - Visualization





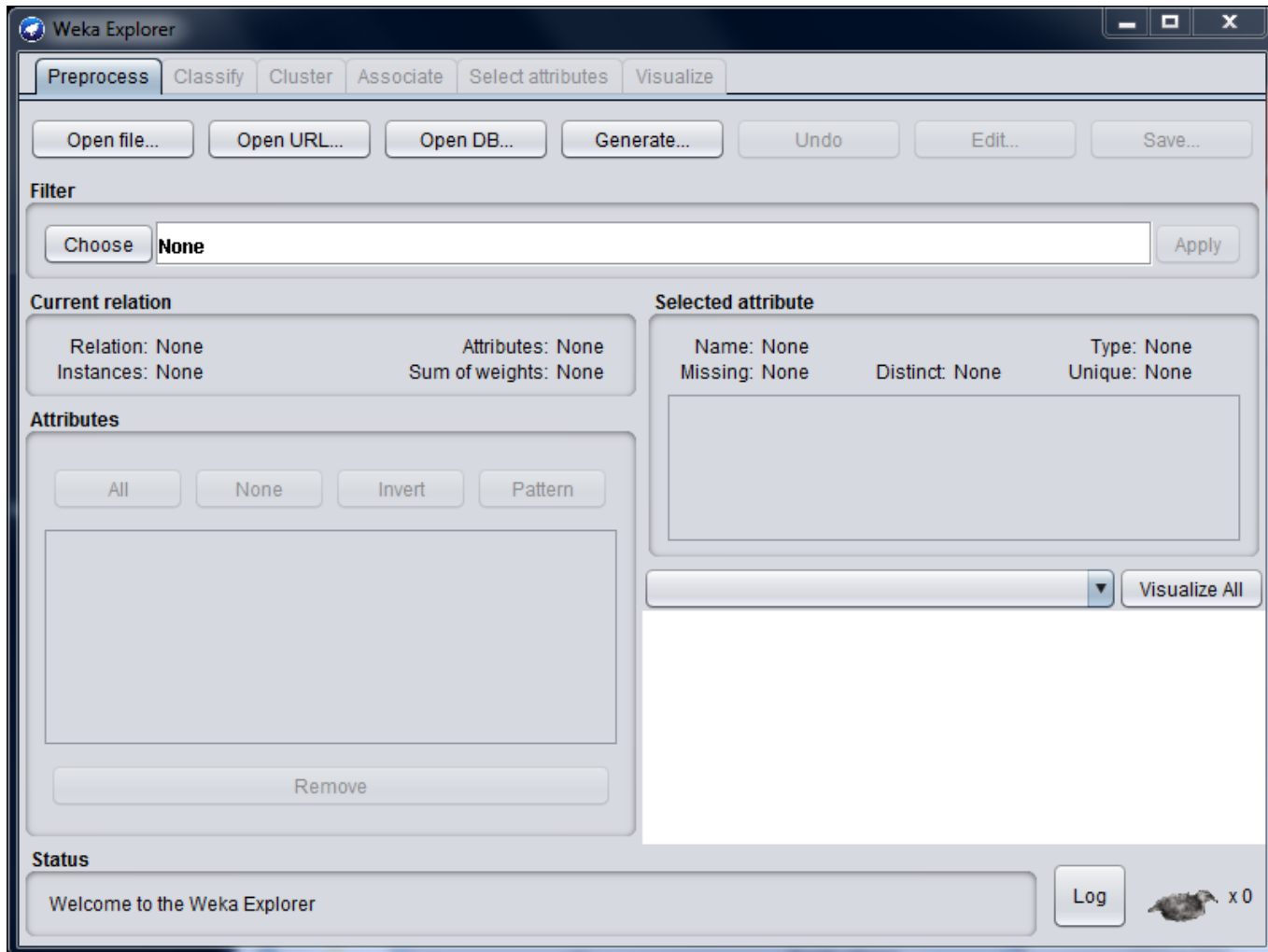
# Use WEKA

- Download:
  - <http://www.cs.waikato.ac.nz/ml/weka/downloading.html>
- GUI WEKA Chooser
  - The Explorer



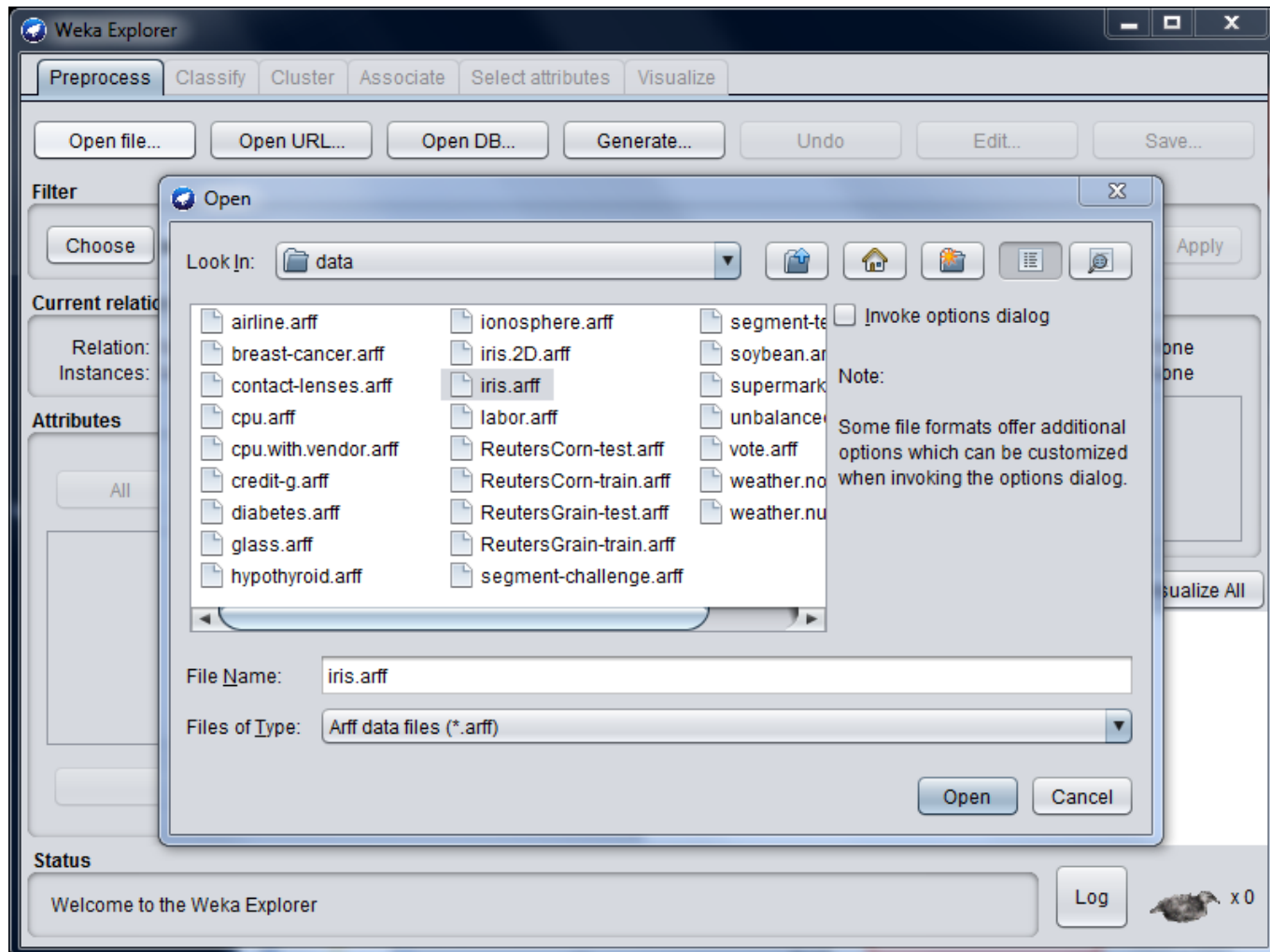


# WEKA Explorer





# 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*

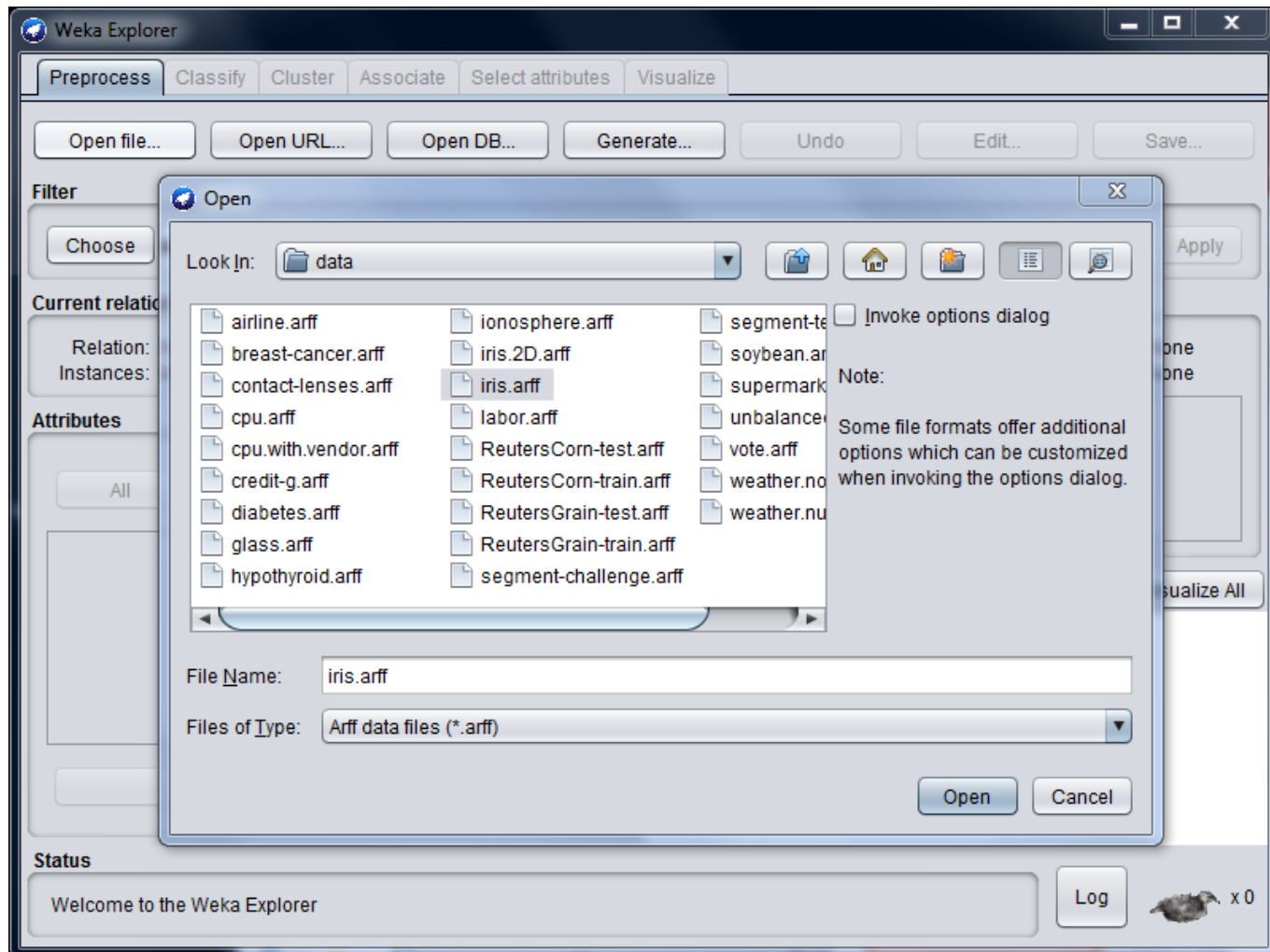
*4.9, 3.0, 1.4, 0.2, Iris-setosa*

*4.7, 3.2, 1.3, 0.2, Iris-setosa*

*...*



# WEKA Explorer





# WEKA Explorer - Preprocess

**Weka Explorer**

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

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

**Filter**

Choose  Apply

**Current relation**

Relation: iris  
Instances: 150

Attributes: 5  
Sum of weights: 150

**Attributes**

All | None | Invert | Pattern

No.	Name
1	<input checked="" type="checkbox"/> sepallength
2	<input type="checkbox"/> sepalwidth
3	<input type="checkbox"/> petallength
4	<input type="checkbox"/> petalwidth
5	<input type="checkbox"/> class

Remove

**Selected attribute**

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

Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

Class: class (Nom) Visualize All

The histogram shows the distribution of the 'sepallength' attribute across three classes: blue (class 0), red (class 1), and cyan (class 2). The x-axis represents the value of 'sepallength' from 4.3 to 7.9. The y-axis represents the frequency of instances. The blue bars represent class 0, red bars represent class 1, and cyan bars represent class 2.

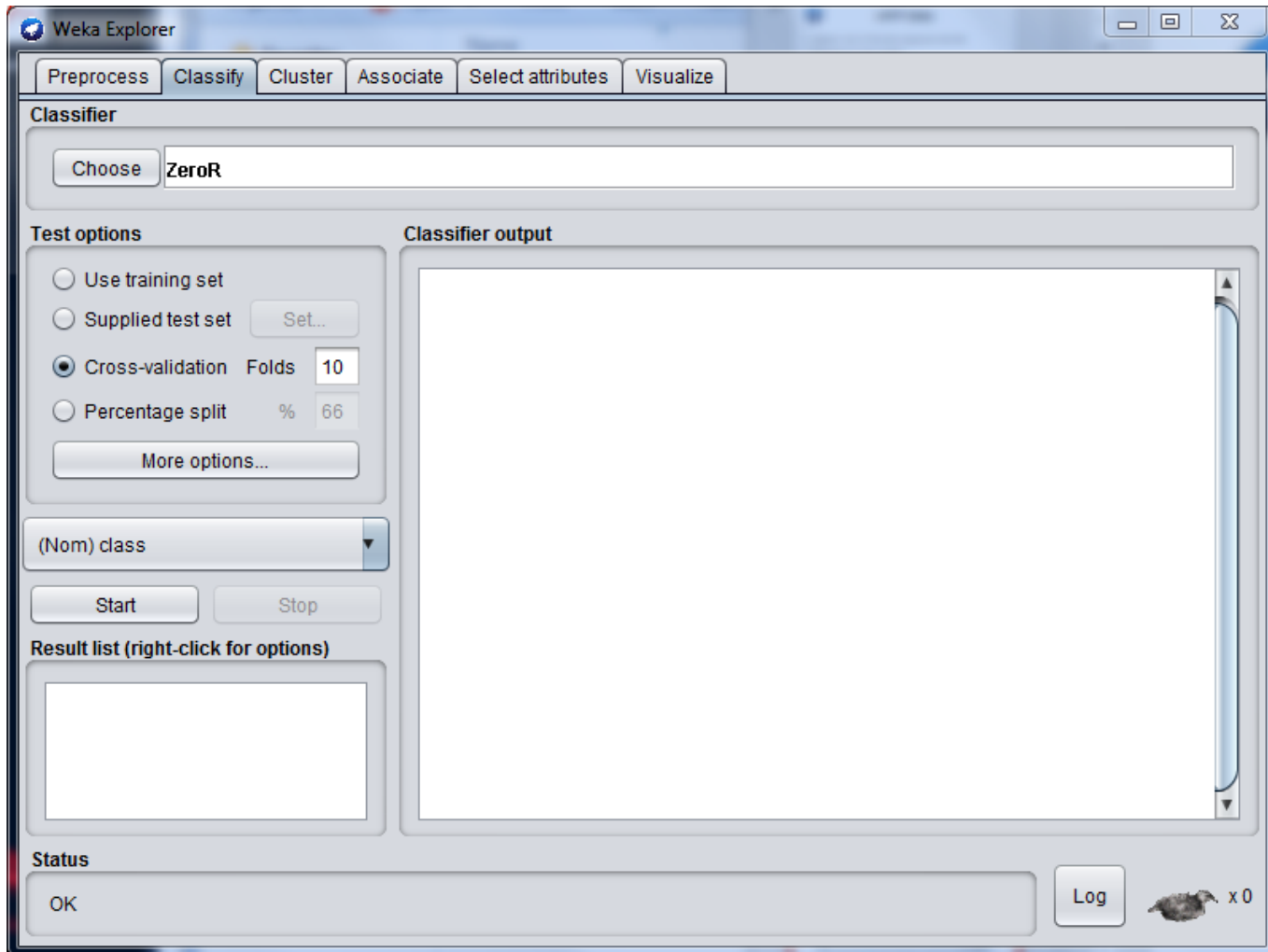
Class	Frequency
0 (Blue)	16
1 (Red)	30
2 (Cyan)	34

**Status**

OK Log x 0

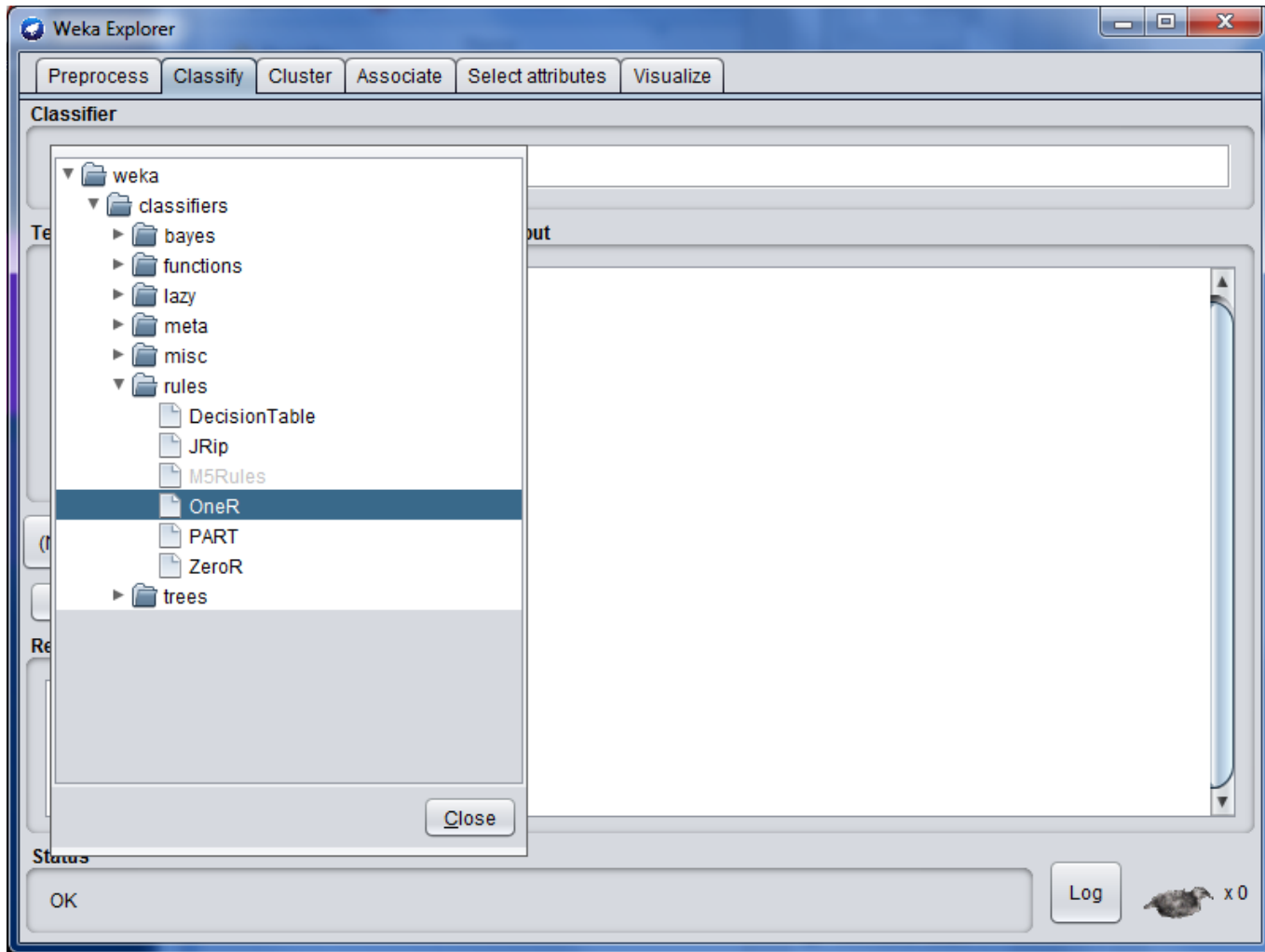


# WEKA Explorer - Classify





# WEKA Explorer - Classify





# WEKA Explorer - Classify

**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) class

Start Stop

**Result list (right-click for options)**

11:02:09 - trees.J48

**Classifier output**

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	0.988
	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
```

**Status**

OK Log x0



# 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

