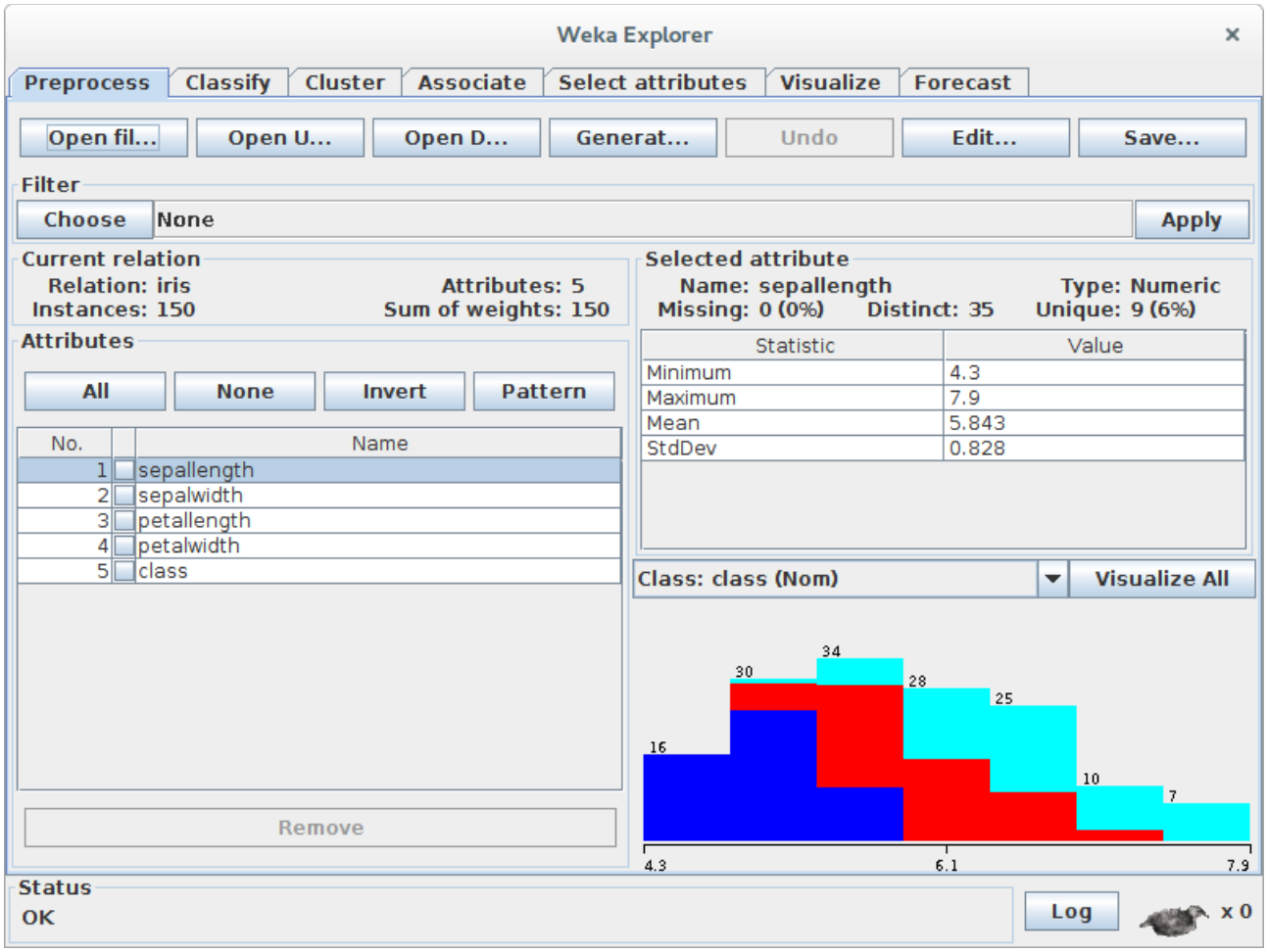
Step1. Open the data/iris.arff Dataset

Click the "***Open file…***" button to open a data set and double click on the "***data***" directory.

Weka provides a number of small common machine learning datasets that you can use to practice on.

Select the "***iris.arff***" file to load the Iris dataset.



The Iris flower dataset is a famous dataset from statistics and is heavily borrowed by researchers in machine learning. It contains 150 instances (rows) and 4 attributes (columns) and a class attribute for the species of iris flower (one of setosa, versicolor, virginica).

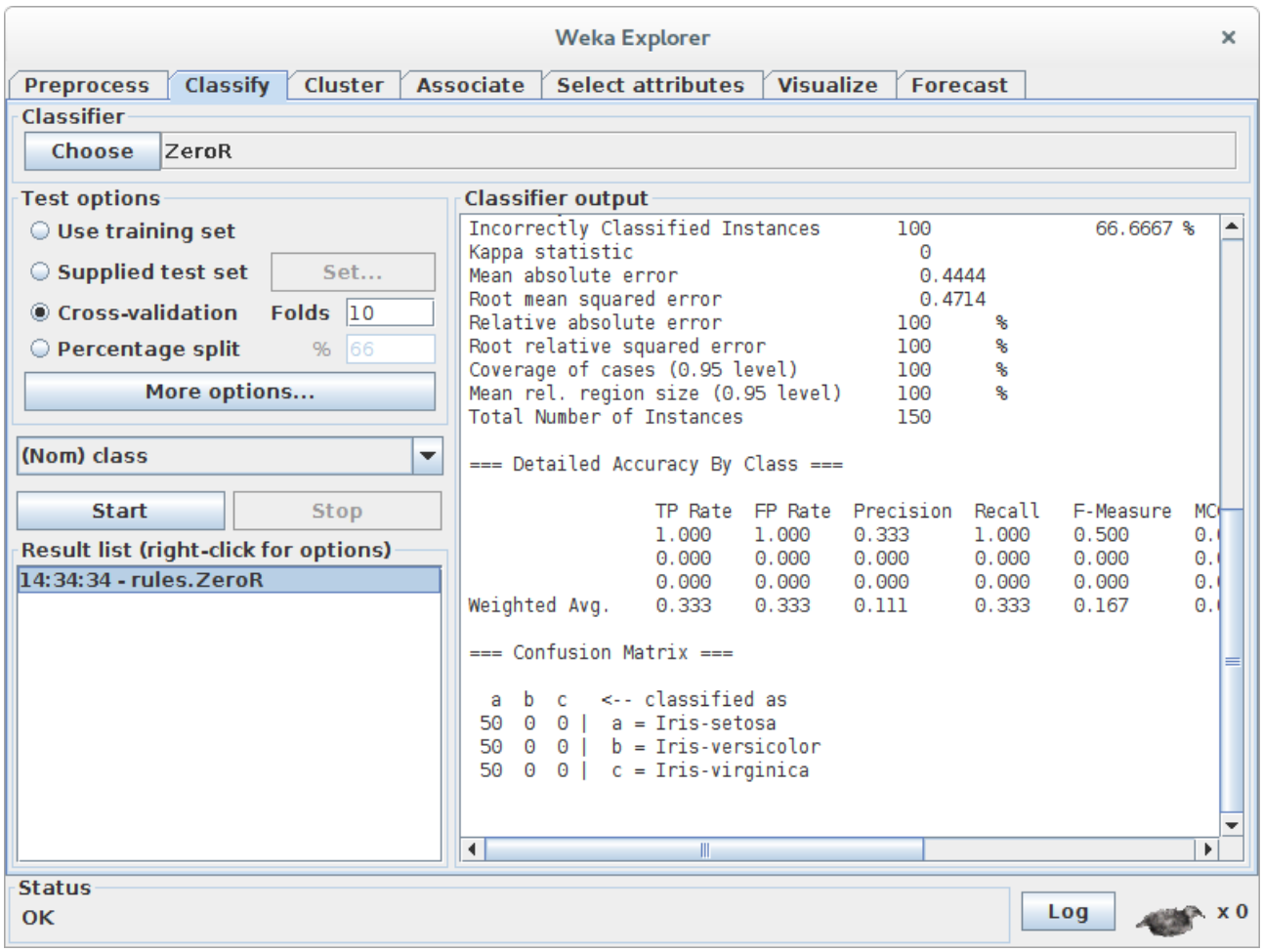
Step2. Select and Run an Algorithm

Now that you have loaded a dataset, it's time to choose a machine learning algorithm to model the problem and make predictions.

Click the "***Classify***" tab. This is the area for running algorithms against a loaded dataset in Weka.

You will note that the "***ZeroR***" algorithm is selected by default.

Click the "***Start***" button to run this algorithm.



The ZeroR algorithm selects the majority class in the dataset (all three species of iris are equally present in the data, so it picks the first one: setosa) and uses that to make all predictions. This is the baseline for the dataset and the measure by which all algorithms can be compared. The result is **33%**, as expected (3 classes, each equally represented, assigning one of the three to each prediction results in 33% classification accuracy).

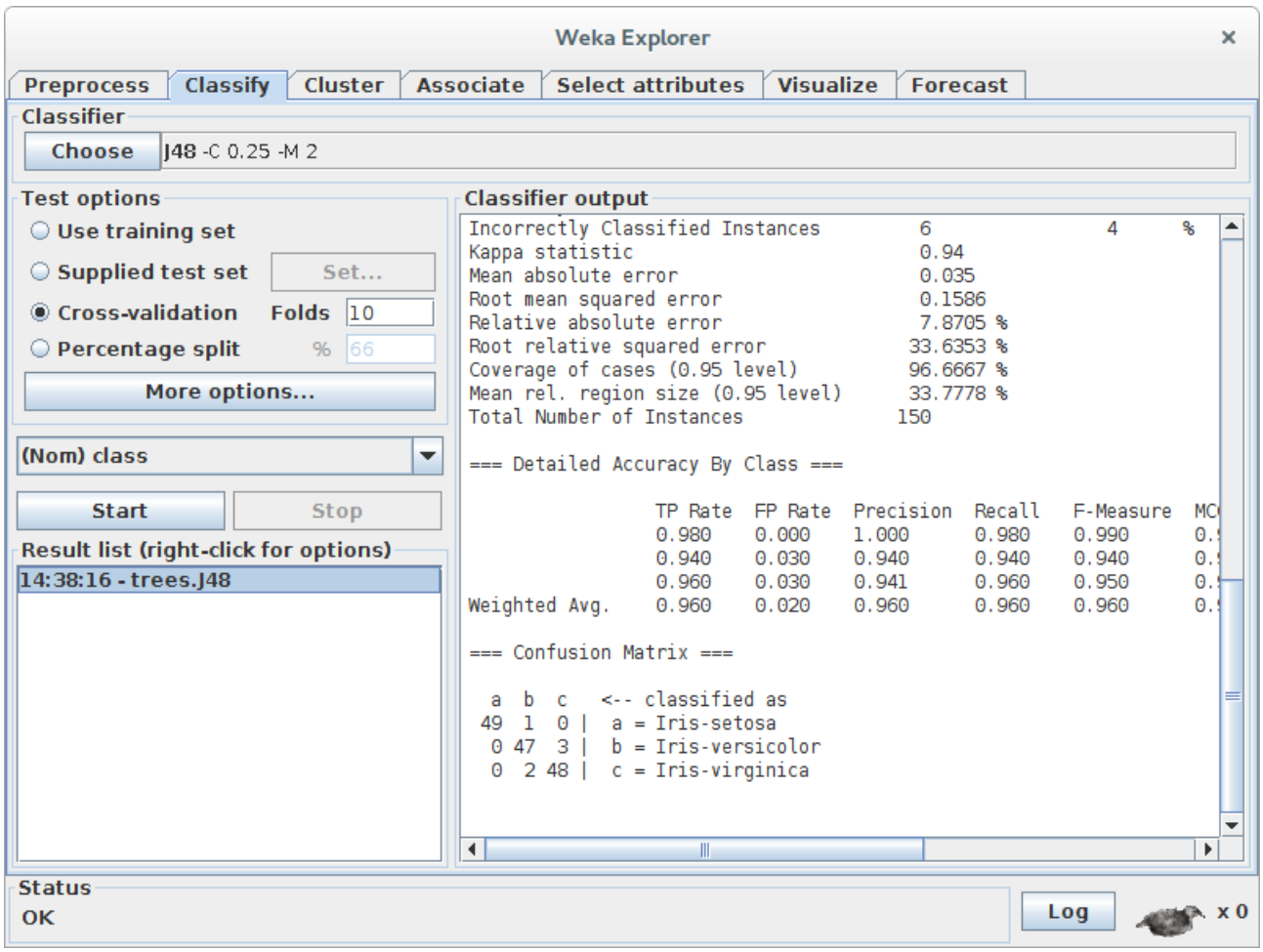
You will also note that the test options uses Cross Validation by default with 10 folds. This means that the dataset is split into 10 parts, the first 9 are used to train the algorithm, and the 10th is used to assess the algorithm. This process is repeated allowing each of the 10 parts of the split dataset a chance to be the held out test set.

Step3: Choose other Algorithm

Click the "***Choose***" button in the "Classifier" section and click on "***trees***" and click on the "***J48***" algorithm.

This is an implementation of the C4.8 algorithm in Java ("J" for Java, 48 for C4.8, hence the J48 name) and is a minor extension to the famous C4.5 algorithm.

Click the "***Start***" button to run the algorithm.



Step4: Review Results

After running the J48 algorithm, you can note the results in the "Classifier output" section.

The algorithm was run with 10 fold cross validation, this means it was given an opportunity to make a prediction for each instance of the dataset (with different training folds) and the presented result is a summary of those predictions.

Firstly, note the Classification Accuracy. You can see that the model achieved a result of **144/150** correct or **96%**, which seems a lot better than the baseline of **33%**.

Secondly, look at the Confusion Matrix. You can see a table of actual classes compared predicted classes and you can see that was 1 error where a Iris-setosa was classified as a Iris-versicolor, 2 cases where Iris-virginica was classified as a Iris-versicolor and 3 cases where a Iris-versicolor was classified as a Iris-setosa (a total of 6 errors). This table can help to explain the accuracy achieved by the algorithm.

You can see the output given by J48 algorithm in a tree fashion by right click on the Results List and by choosing the option Visualize Tree.