Data Mining Practical - Weka

This practical requires you to build a model from a set of data and then use that model to classify new examples from a different file. The example is the same one your saw in the first lecture - the problem of identifying fruit from its weight, colour and shape.

You will download two files - one containing the data that you will use to build a model, and the other with a list of examples of fruits that have not been identified. You must identify the new mystery fruits!

You will be asked to write down certain things as you work through this page. Take care to write them down in order and number them to match this page, as you will be assessed on what you create from this exercise!!

The new fruits will, of course, be of the types that appear in the training data, they are:

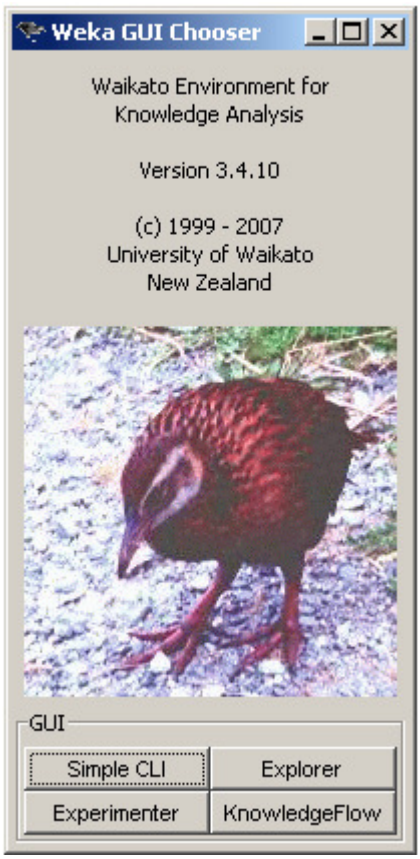
- Melon
- Apple
- Pineapple
- Courgette
- Banana
- Tomato
- No_fruit

Note the last value: No_fruit - there are examples in the data of objects that are not a fruit at all. For these, the correct answer is 'No_fruit'.

You will be using the data mining software package, 'Weka'. Weka is a bit fiddly to use, but it is very powerful, and free! Follow the steps below and you shouldn't have any problems.

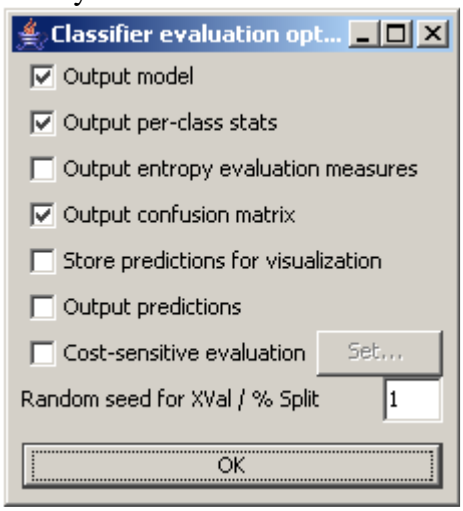
1. **Run Weka**: Go to Start, Program files and choose Weka. Run Weka 3.4

2. You should see this:



   Click the **Explorer** button
3. You should now see this the Weka Explorer with the Preprocess tab selected
4. You will now need two files. Right click the two links below and save them to your disk space:
   Training File
   Test File
5. Now, return to Weka and click the **Open file ...** button.
6. Select the file you just downloaded called *fruit.arff* this is your training file.
7. Look at the attributes listed in the Attributes window, click on each to select it and see its details
   Question 1 Write down the attributes that are in the file. Which will you use as your output?
   Question 2 Two attributes are numeric - write down their names.
   Question 3 Look at the charts - which are the two most expensive fruits?
8. Now select the **Classify** tab. You will see a drop-list containing a list of the attributes in the file. Select the one you want as the output (you are classifying the 'fruit' variable).
9. In the 'Test options' box, select 'Percentage Split' and decide how much of the training data you want to use for building the model. What is left will be used to test if the model works.
10. Next you must choose a technique. Weka has lots of them, but you should have learned about *Multi-layer Perceptrons*, *Decision trees* and *rules*. Click the **choose** button in the 'Classifier' box and pick a technique. Multi-layer perceptron is under the 'functions' folder. If you want to use a tree, try 'Decision Stump'. That is a good one to start with.

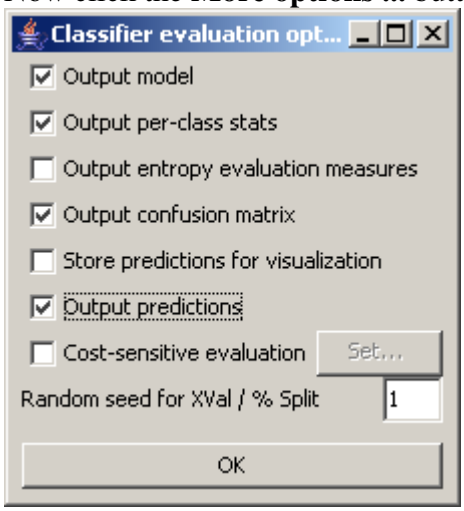
11. Nearly there ... Now click the **More options ...** button and check options thus:



    Make sure you don't check Output predictions!!!
12. Click OK to close that dialog and you are ready to go!
13. Click the **Start** button to build your model
14. When the model has been built, the results will be shown in the large window.
15. Go to the very bottom of this window and look at the confusion matrix.
    Question 4 How many test cases did the model mis-classify?
16. You can save your model by right clicking on its name in the 'result list' and selecting **Save Model**.
17. Go back and try some other techniques - do some perform better or worse? Techniques to try are:

- Trees->J48. When you have a model, right click on its name and select Visualize Tree to see what the tree looks like
- Functions->Multi layer perceptron.

18. Finally, you can classify the new fruits in the test file:
    - In the 'Test options' box, select **Supplied test set**
    - Click the **Set...** button and then the **Open file...** button
    - Choose the file *testfruit.arff*
    - Now click the **More options ...** button again and check options thus:



    Notice that you should have checked the **Output predictions** box

- Now, right click on the model name in the Result list and select **Re-evaluate model on current test set**.
- The results should go to the window to the right.
- Another right click on the same model, this time selecting **Save result buffer** will allow you to save your predictions to a file.
  Give the file a name using your full name with no spaces, then a number so you can save more than one, and end it with '.txt'
  e.g. johnsmith1.txt then, johnsmith2.txt if you make another

19. Go to the practicals page on the course web site and use the upload box at the foot of the page to upload your results file and see how well you did.