Using the Weka Machine Learning toolkit

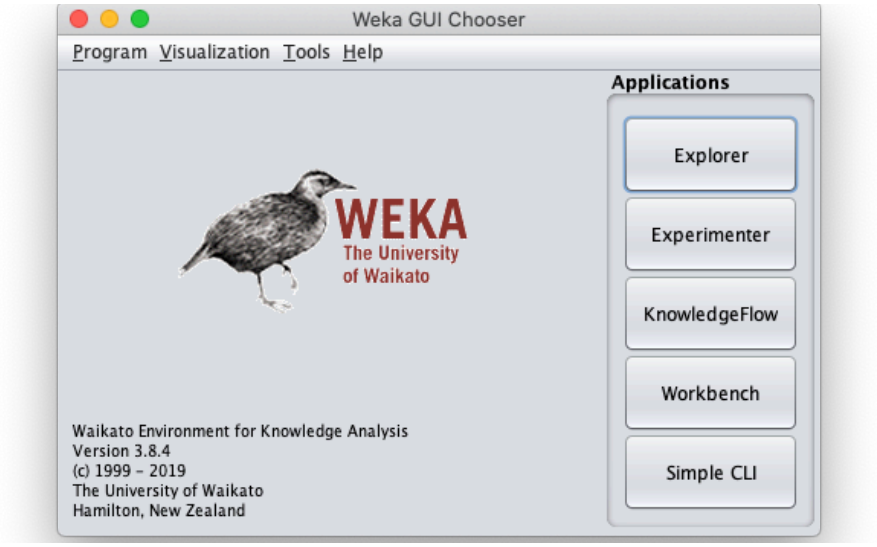
Weka is a Machine Learning toolkit written in Java. It was developed and is maintained by a team from Waikato University in New Zealand but it has packages contributed from many different sources.

Weka is reasonably easy to use as it has a GUI so you don't need to do any programming. It runs on Linux, macOS and Windows. You can download it from:

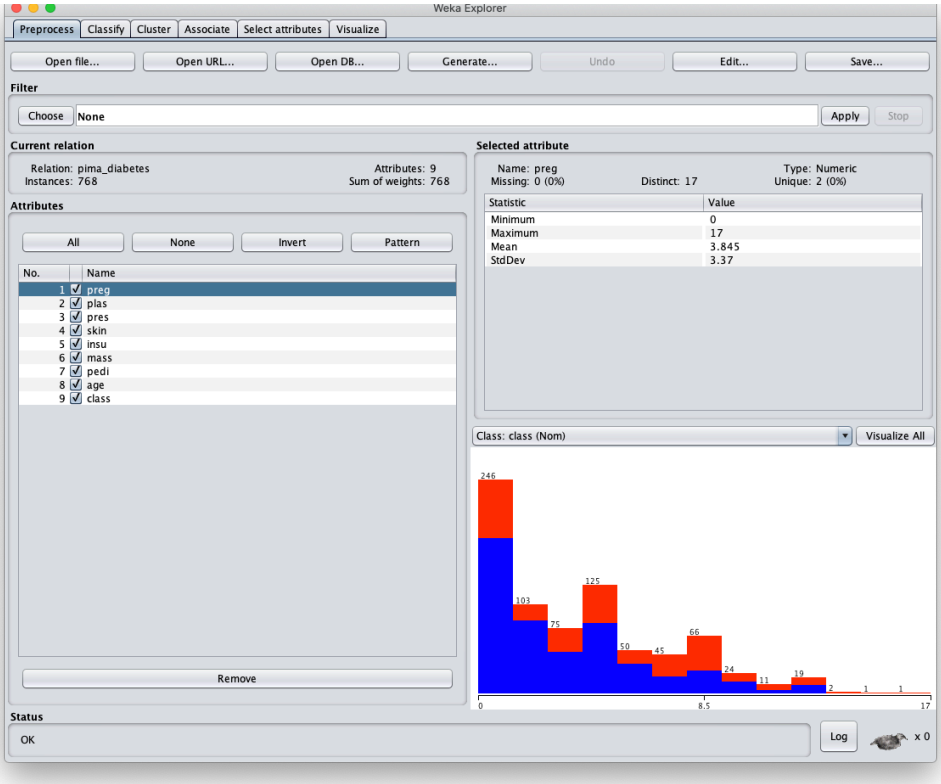
https://waikato.github.io/weka-wiki/downloading_weka/

The download includes a Java JVM, so you shouldn't need to install Java, if you haven't got it already.

Once you have it installed, when you start Weka, you will see window like this:



1. Click on the explorer button and you will get a new window:

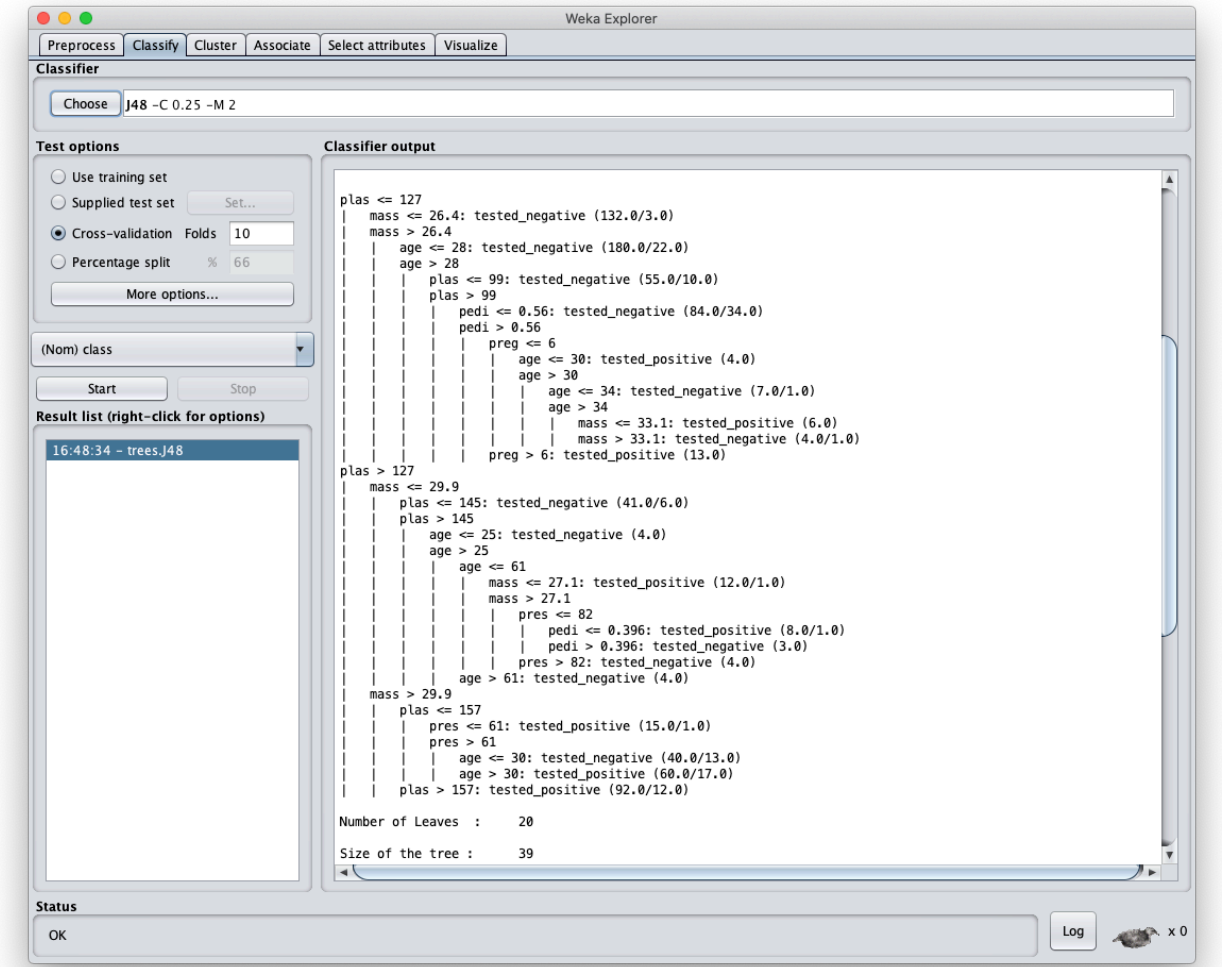


This shows you the data panel.

2. Click “Open file …” to select the data file. This is in “ARFF” format, which is easy to construct from a CSV file. An example is appended at the end of this document.
3. By ticking the check boxes, you can see the distribution of objects with the ticked attribute values.

Note that you should combine the training and test sets in the ARFF file.

4. Next, click on the “classify” button (at the top of the window) to bring up the classification panel:



5. Clicking the “Choose” button will give you a list of different classifier learning algorithms. Choose “J48” under the “Trees” subcategory. This is Weka’s implementation of Quinlan’s C4.5 algorithm.
6. The default setting in for testing is 10-fold cross-validation. You can also click on “Percentage split” to get a default 66% training data 34% test data.
7. Clicking “Start” will run J48 on the data set and produce a representation of the decision tree on the right.

It will be acceptable to submit that in your assignment. The panel will also contain error statistics for the decision tree.

Example ARFF File

The file needs an **@relation** line to tell it the name of the data table.

Each attribute is defined in an **@attribute** line, which also give the type of the attribute. A real valued attribute is ‘real’; discrete values attributes are given a list of the valid values. A question mark means missing value.

The example below is from the UCI repository. It is not the complete data file, as it is too big to list here. If you are interested, more example data sets are here:

https://waikato.github.io/weka-wiki/datasets/

```
@relation labor
@attribute 'duration' real
@attribute 'wage-increase-first-year' real
@attribute 'wage-increase-second-year' real
@attribute 'wage-increase-third-year' real
@attribute 'cost-of-living-adjustment' {'none','tcf','tc'}
@attribute 'working-hours' real
@attribute 'pension' {'none','ret_allw','empl_contr'}
@attribute 'standby-pay' real
@attribute 'shift-differential' real
@attribute 'education-allowance' {'yes','no'}
@attribute 'statutory-holidays' real
@attribute 'vacation' {'below_average','average','generous'}
@attribute 'longterm-disability-assistance' {'yes','no'}
@attribute 'contribution-to-dental-plan' {'none','half','full'}
@attribute 'bereavement-assistance' {'yes','no'}
@attribute 'contribution-to-health-plan' {'none','half','full'}
@attribute 'class' {'bad','good'}
@data
1,5,?,?,?,40,?,?,2,?,11,'average',?,?,'yes',?,'good'
2,4.5,5.8,?,?,35,'ret_allw',?,?,'yes',11,'below_average',?,'full',?,'full','good'
?,?,?,?,?,38,'empl_contr',?,5,?,11,'generous','yes','half','yes','half','good'
3,3.7,4,5,'tc',?,?,?,?,'yes',?,?,?,?,'yes',?,'good'
3,4.5,4.5,5,?,40,?,?,?,?,12,'average',?,'half','yes','half','good'
2,2,2.5,?,?,35,?,?,6,'yes',12,'average',?,?,?,?,'good'
3,4,5,5,'tc',?,'empl_contr',?,?,?,12,'generous','yes','none','yes','half','good'
3,6.9,4.8,2.3,?,40,?,?,3,?,12,'below_average',?,?,?,?,'good'
2,3,7,?,?,38,?,12,25,'yes',11,'below_average','yes','half','yes',?,'good'
1,5.7,?,?,'none',40,'empl_contr',?,4,?,11,'generous','yes','full',?,?,'good'
3,3.5,4,4.6,'none',36,?,?,3,?,13,'generous',?,?,'yes','full','good'
2,6.4,6.4,?,?,38,?,?,4,?,15,?,?,'full',?,?,'good'
2,3.5,4,?,'none',40,?,?,2,'no',10,'below_average','no','half',?,'half','bad'
3,3.5,4,5.1,'tcf',37,?,?,4,?,13,'generous',?,'full','yes','full','good'
1,3,?,?,'none',36,?,?,10,'no',11,'generous',?,?,?,?,'good'
2,4.5,4,?,'none',37,'empl_contr',?,?,?,11,'average',?,'full','yes',?,'good'
1,2.8,?,?,?,35,?,?,2,?,12,'below_average',?,?,?,?,'good'
1,2.1,?,?,'tc',40,'ret_allw',2,3,'no',9,'below_average','yes','half',?,'none','bad'
```