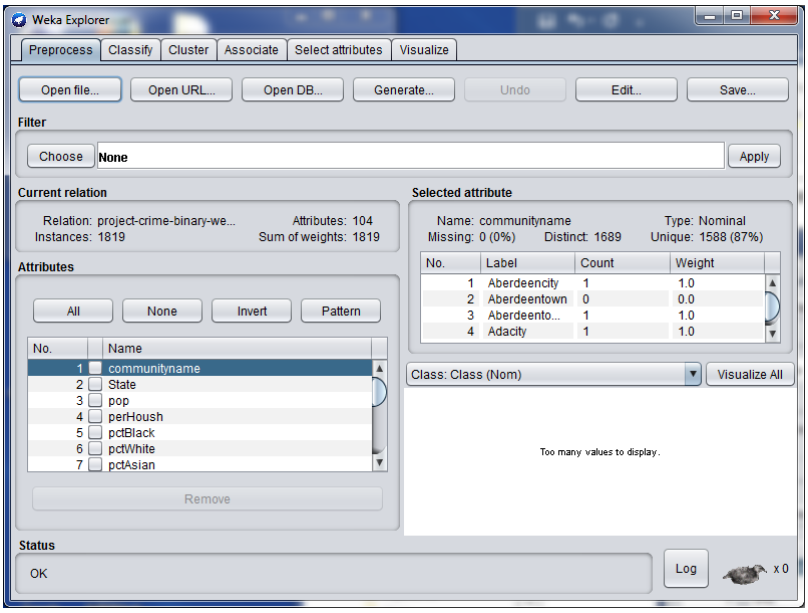
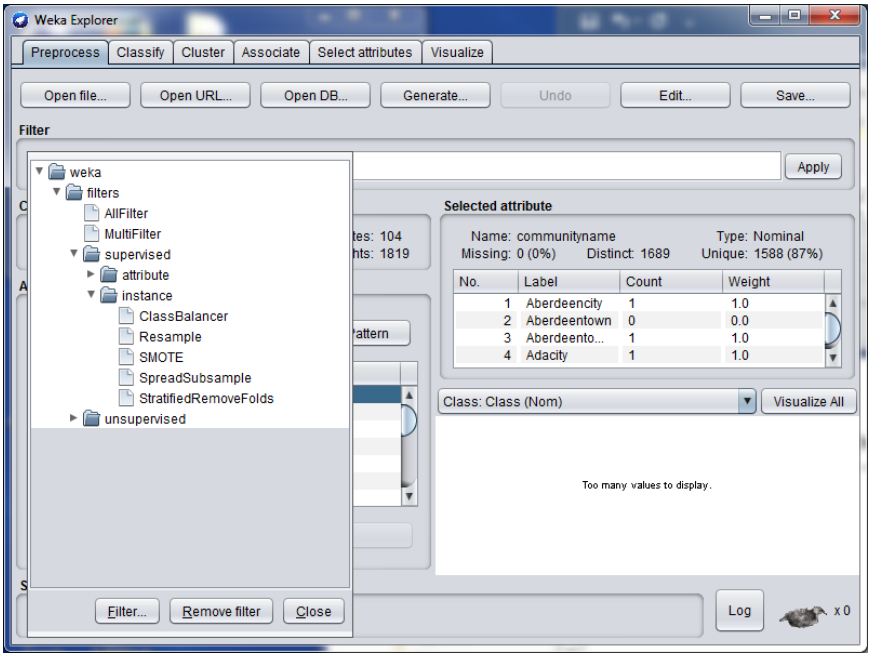
This document shows how to split a given dataset into a training dataset and a test dataset, while preserving class distribution, using Weka. Note that the screenshots shown here may not be exactly the same as those you will see on your computer.

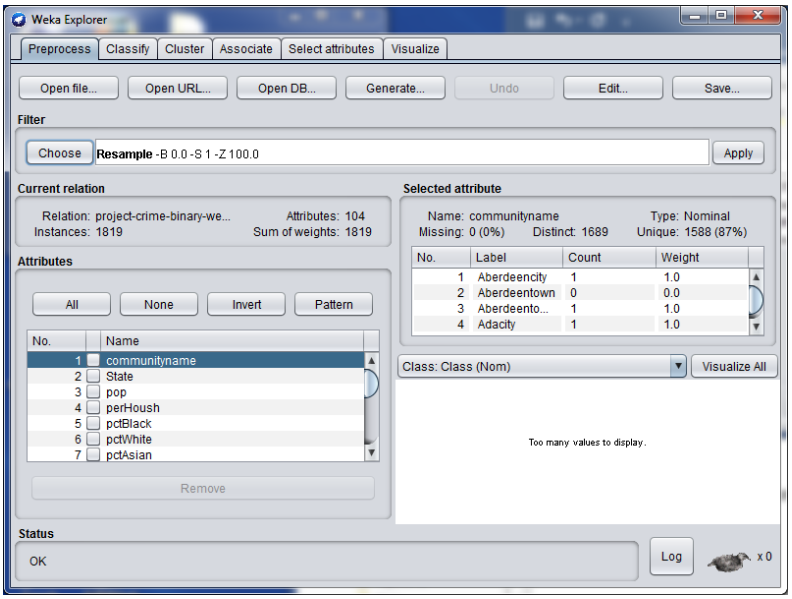
Open the initial dataset.



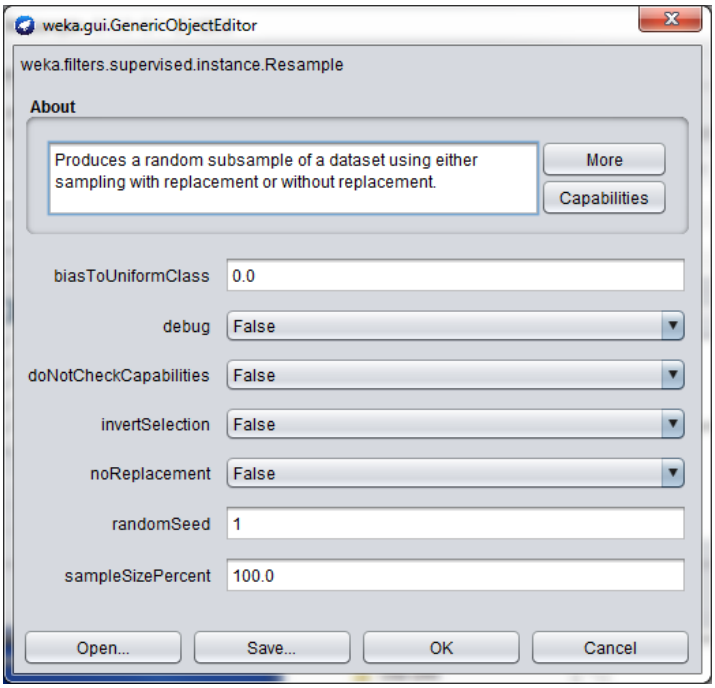
Click the *Choose* button below *Filter*.
Then, click *filters – supervised – instance*.



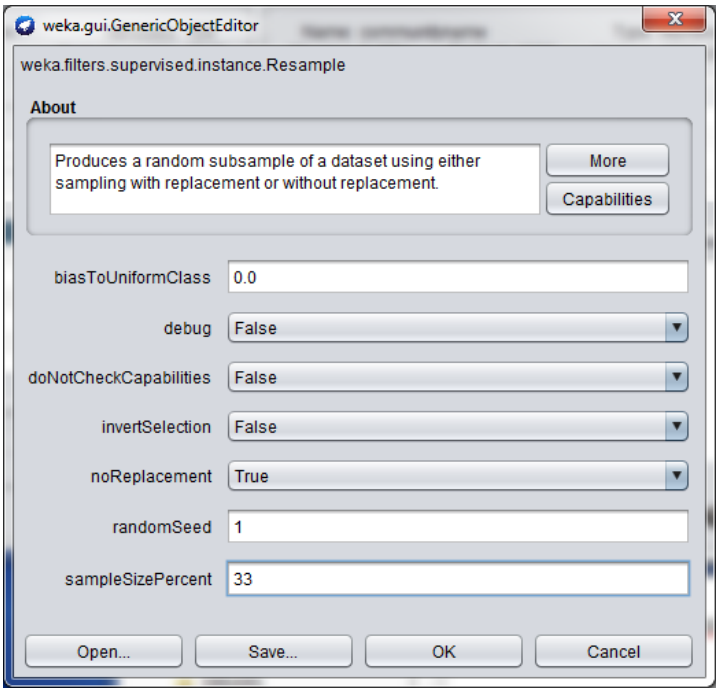
Then, choose *Resample*.



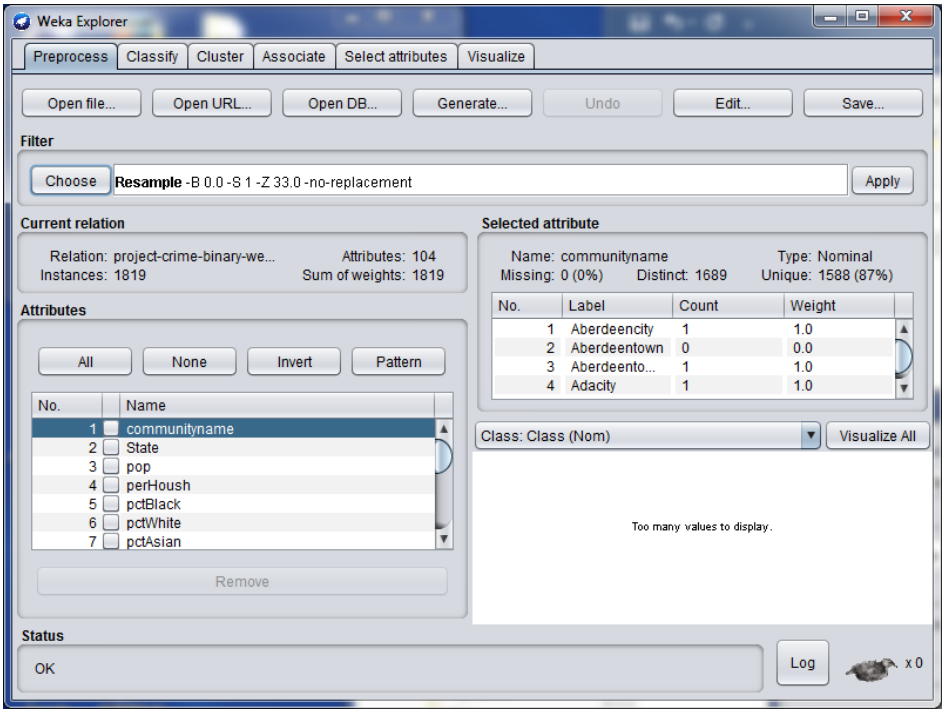
Click the horizontal space next to *Choose* where *Resample* is. The following window appears.



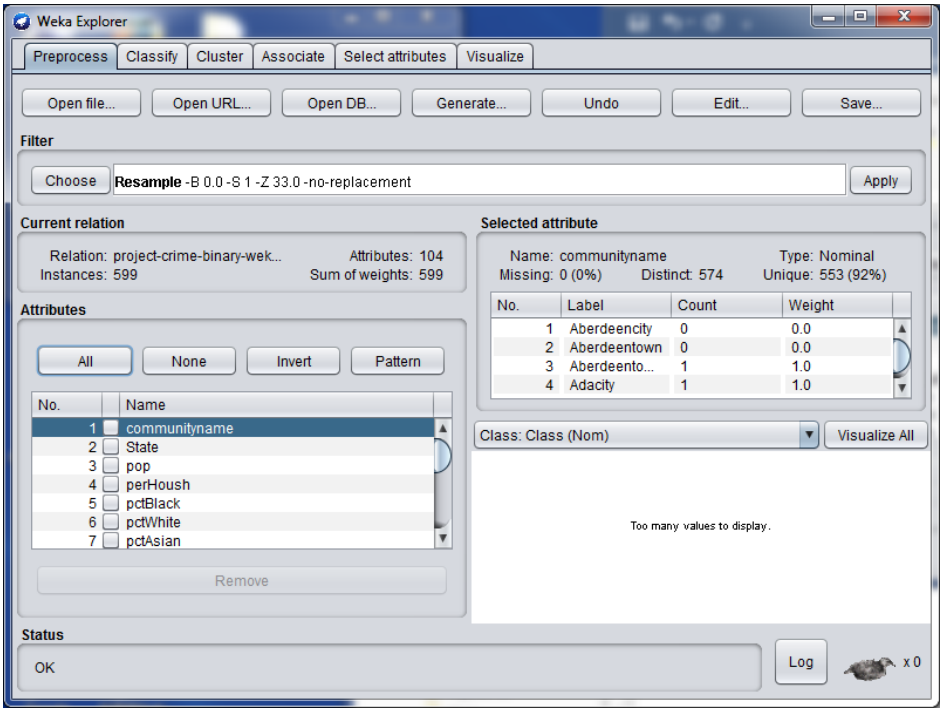
Change *noReplacement* to *True* and change *sampleSizePercent* to 33.



Click OK.

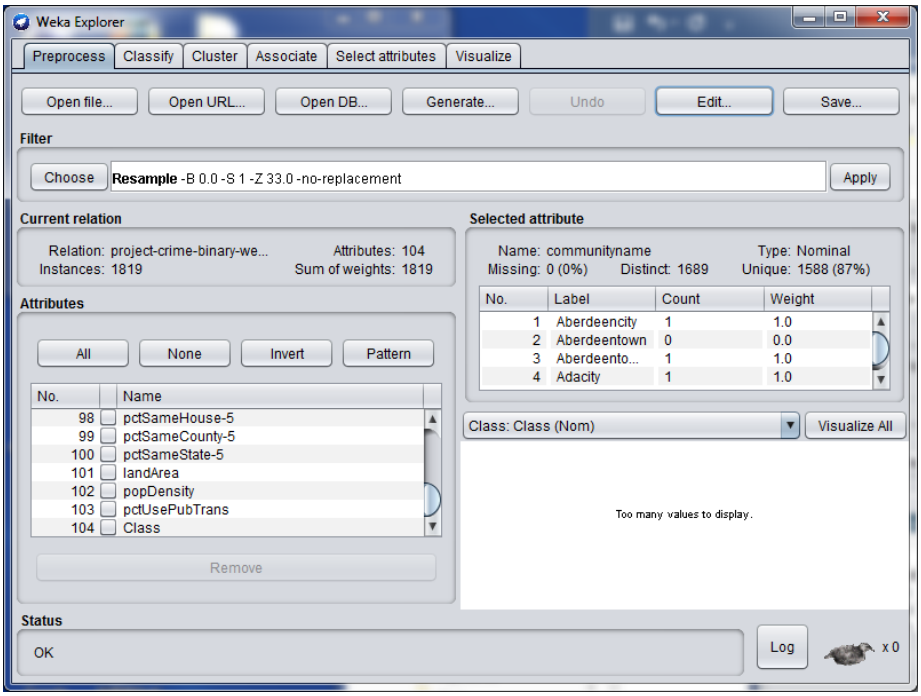


Click *Apply*.



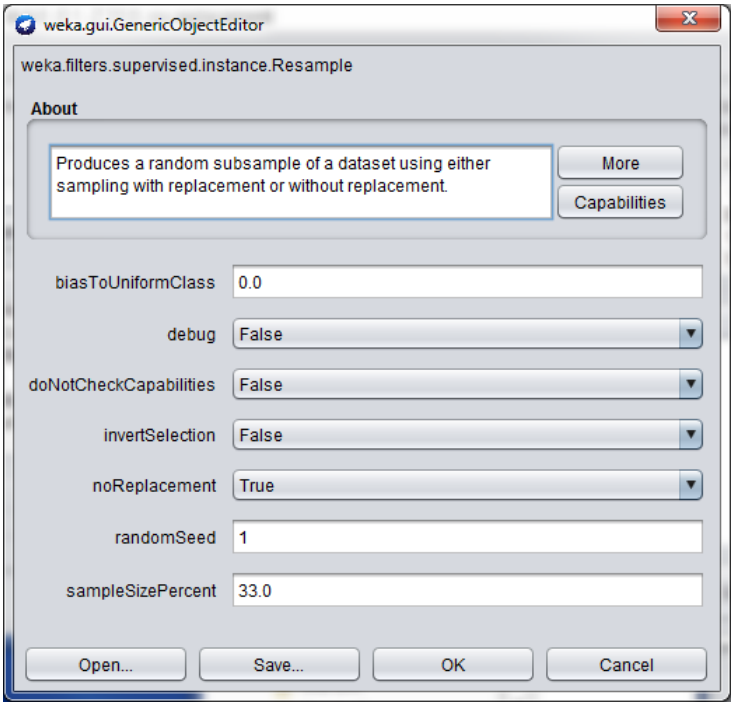
It shows that 599 tuples were sampled (assignment says to select 600 tuples but 599 or any number of tuples close to 600 is OK). Click *Save* button and save it as project-test.arff (or with other appropriate name).

Right after saving the test dataset, click *Undo* button.

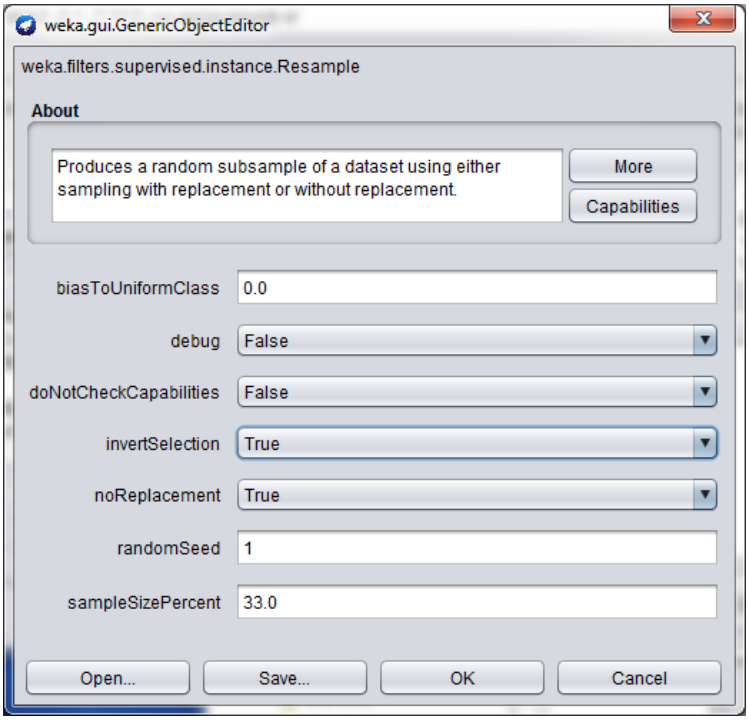


As shown above, you are returned to the initial dataset (with 1819 tuples).

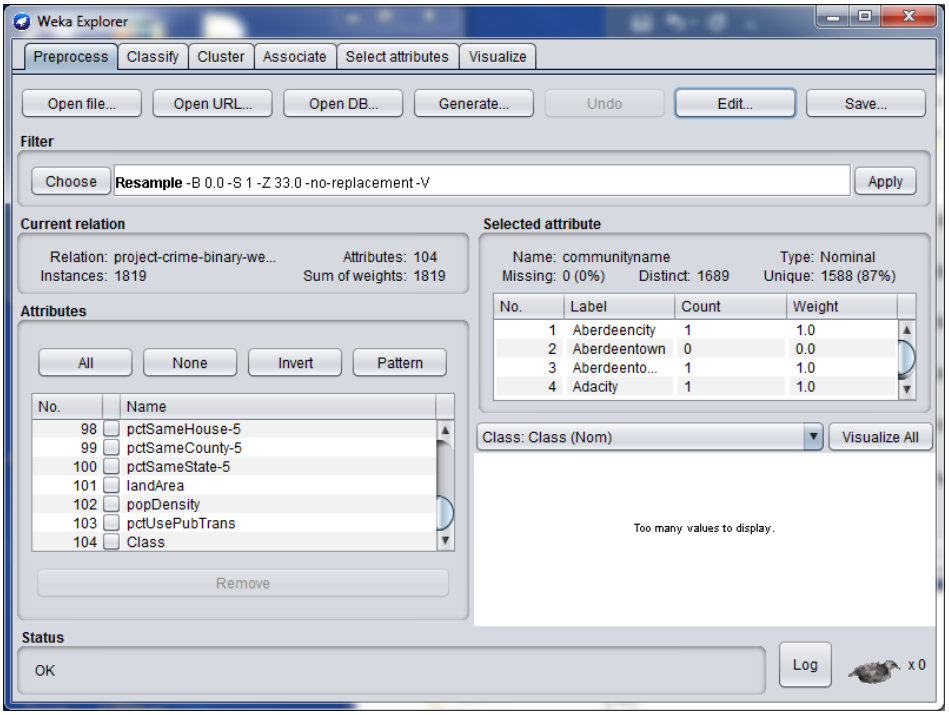
Click the horizontal space next to *Choose*.
The following window pops up. Note that your previous selections (*noReplacement* is *True* and *sampleSizePercent* is 33.0) are still there. Don't change them.



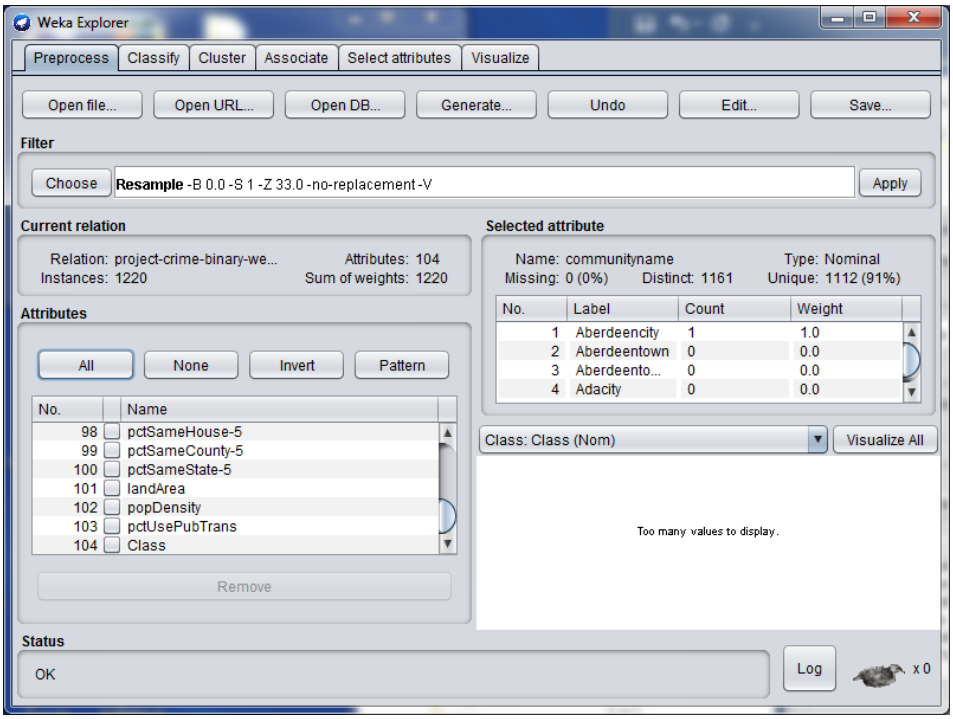
Change *InvertSelection* to *True*.



Click *OK*.



Click *Apply*.



You can see that 1220 tuples were sampled. This is your training dataset. Save it as project-training.arff (or with other appropriate name).

These two datasets are disjoint and each of them has the same class distribution as that of the initial dataset (because you chose *supervised*).