**Junaid A Girkar**

**60004190057**

**TE Comps A-4**

**DWM**

**LAB EXPERIMENT - 1**

**Aim**: Perform data Pre-processing task using Weka data mining tool

1. What is Weka?

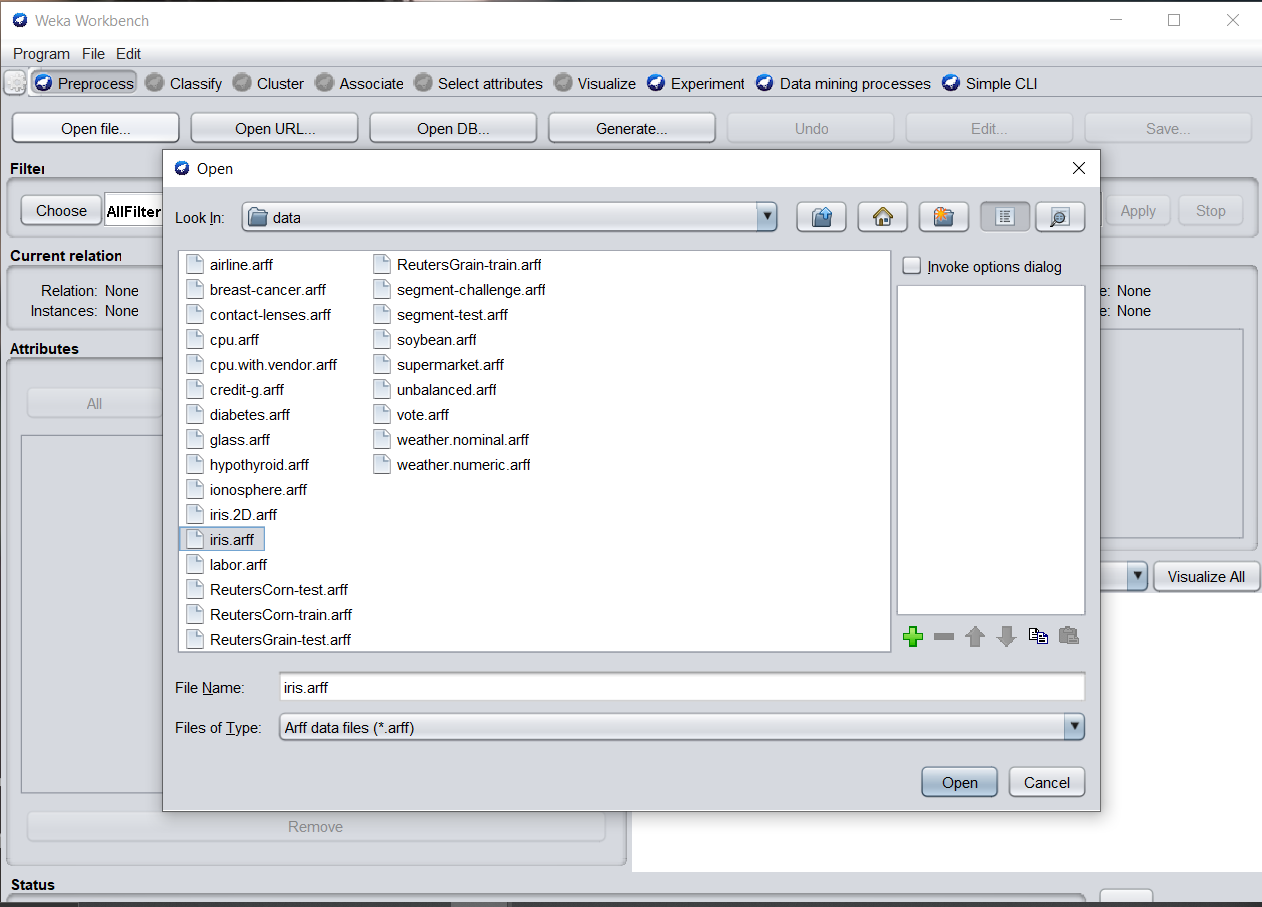
Waikato Environment for Knowledge Analysis (Weka), developed at the University of Waikato, New Zealand, is free software licensed under the GNU General Public License. Weka contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to these functions in particular for educational purposes and research. Advantages of Weka include:

* Free availability under the GNU General Public License.
* Portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform.
* A comprehensive collection of data preprocessing and modeling techniques.
* Ease of use due to its graphical user interfaces.

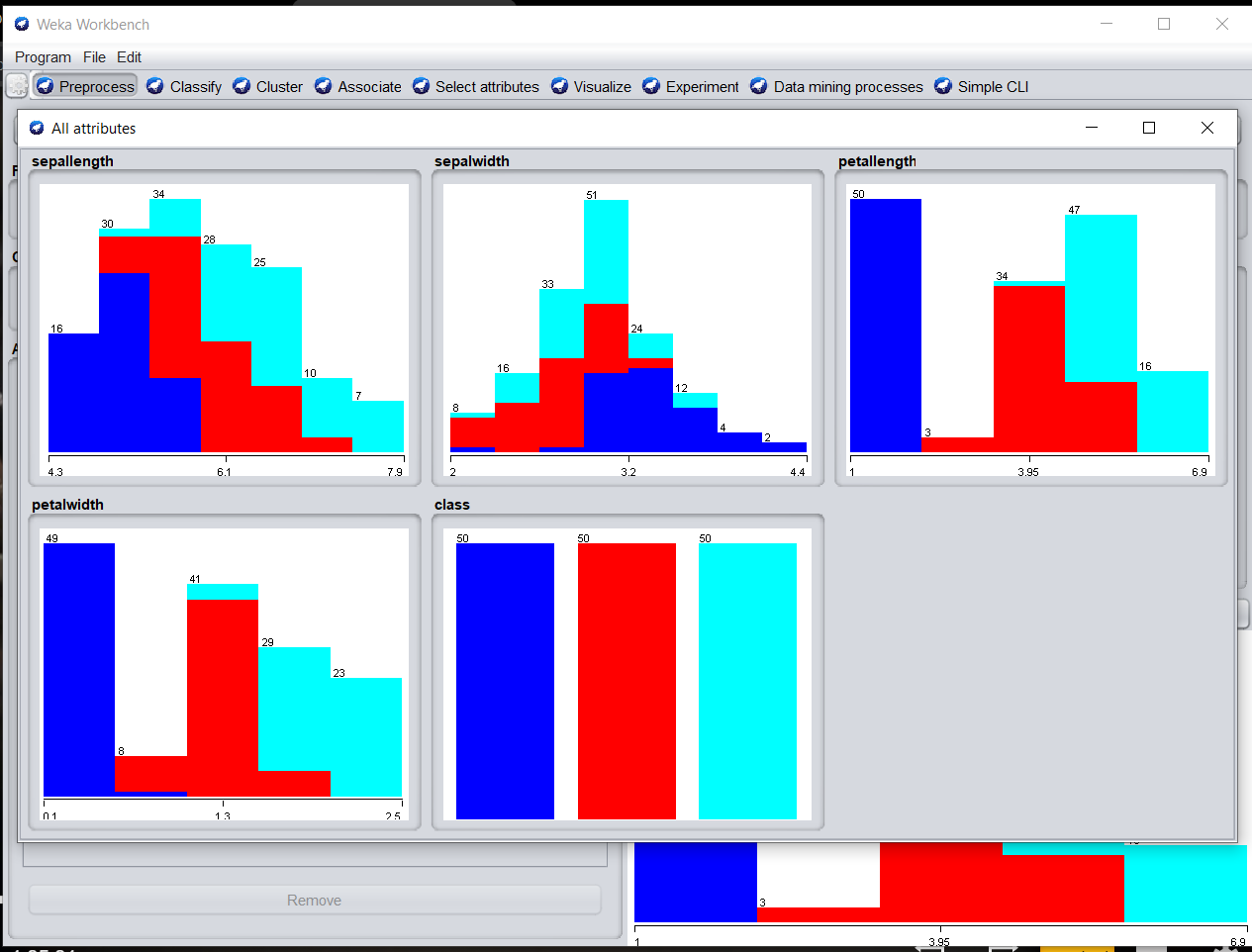
1. Using Weka

The interface is divided into 6 tabs, each with a specific function:

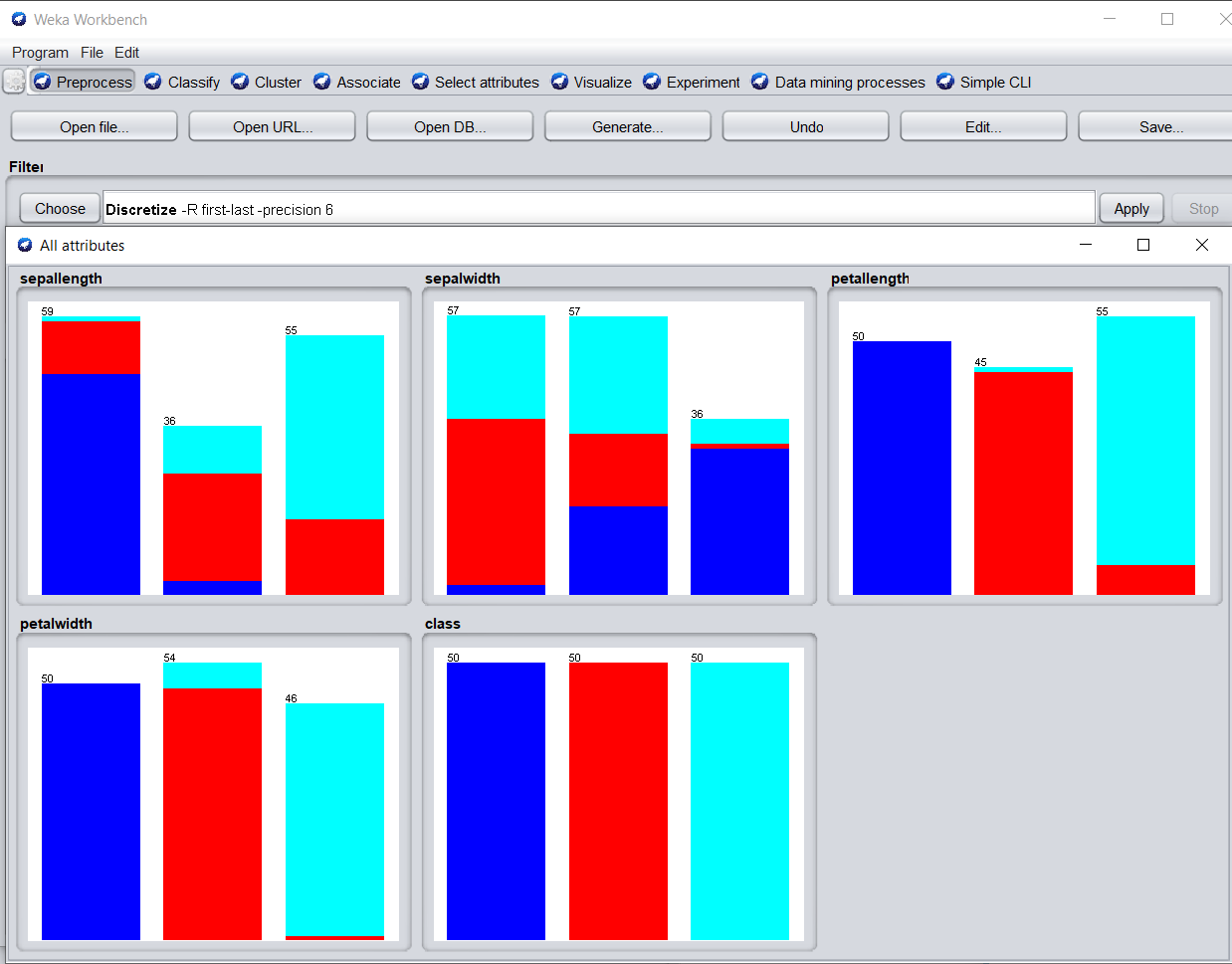
The preprocess tab is for loading your dataset and applying filters to transform the data into a form that better exposes the structure of the problem to the modeling processes. Also provides some summary statistics about loaded data.



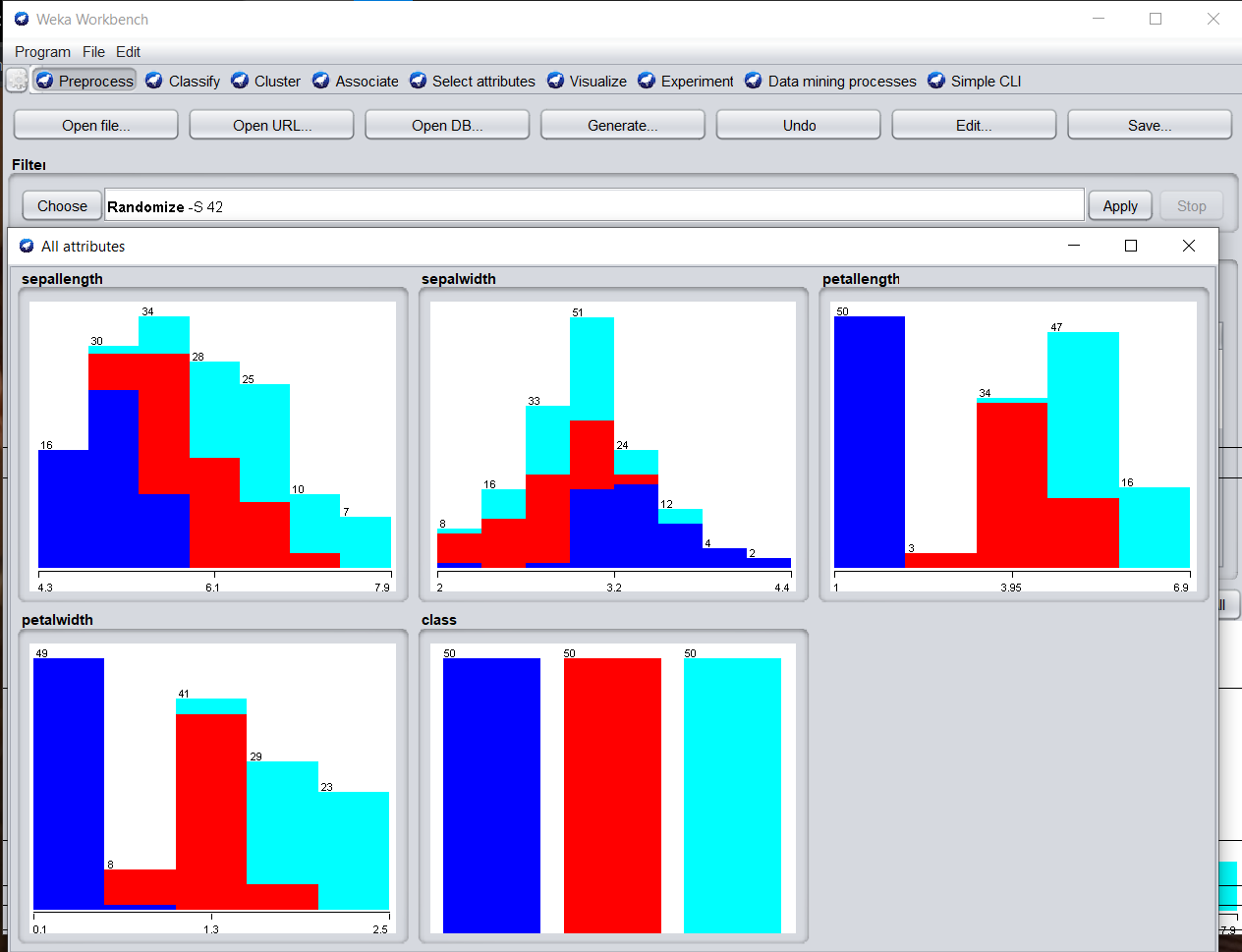
Visualise before filtering



SUPERVISED DISCRETE FILTERING

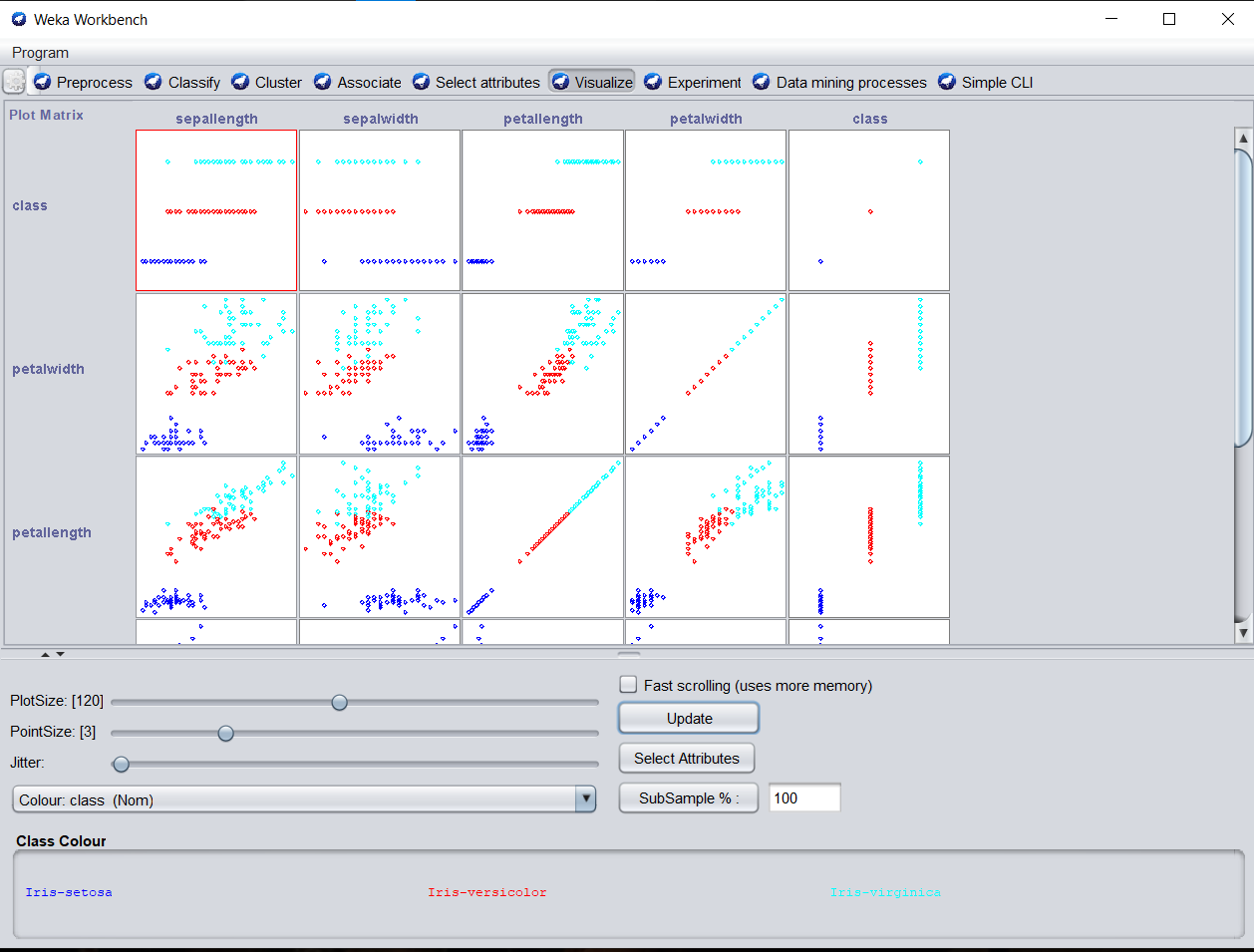


UNSUPERVISED RANDOMIZE FILTERING

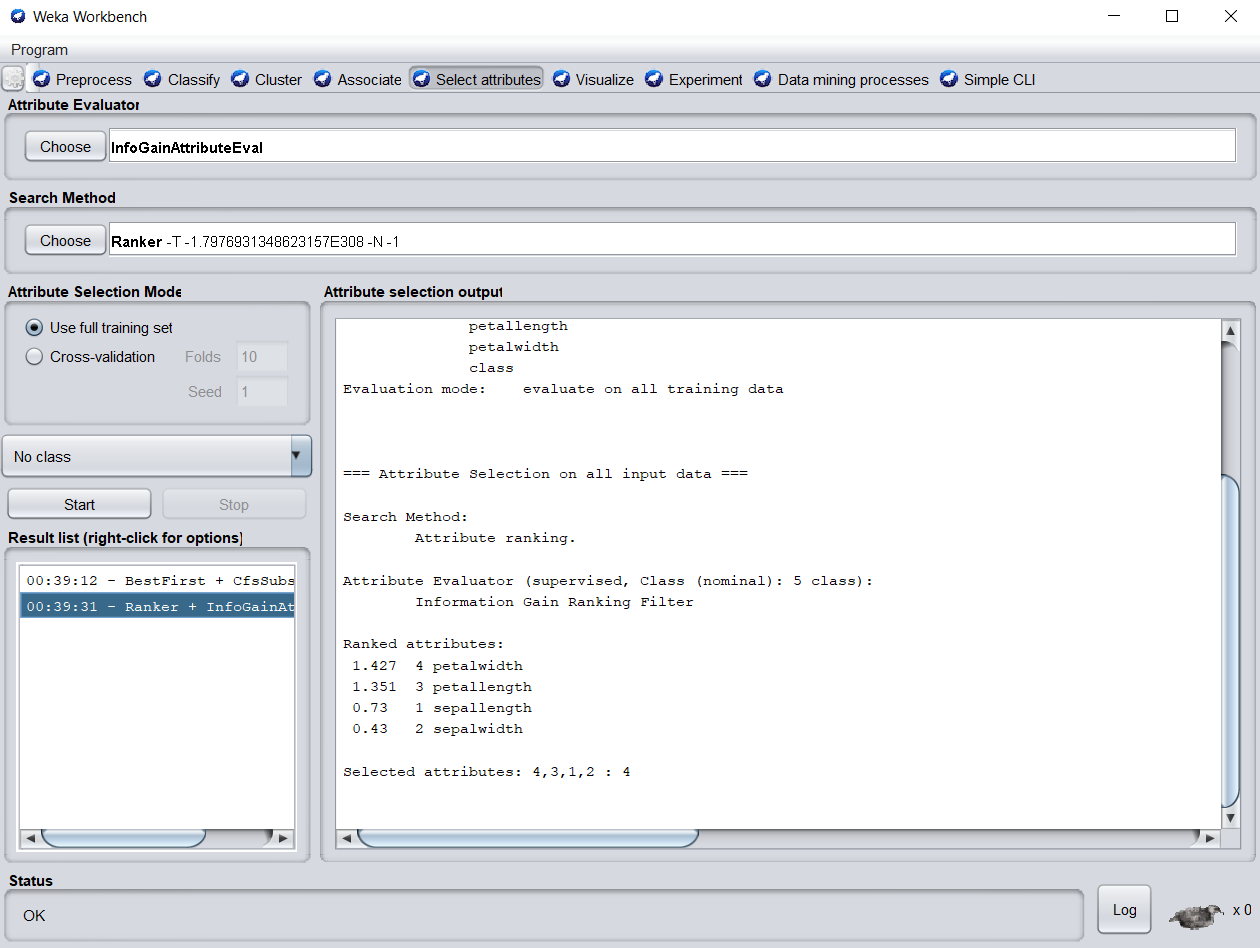


The visualize tab is for reviewing the pairwise scatterplot matrix of each attribute plotted against every other attribute in the loaded dataset. It is useful to get an idea of the shape and relationship of attributes that may aid in data filtering, transformation and modeling.

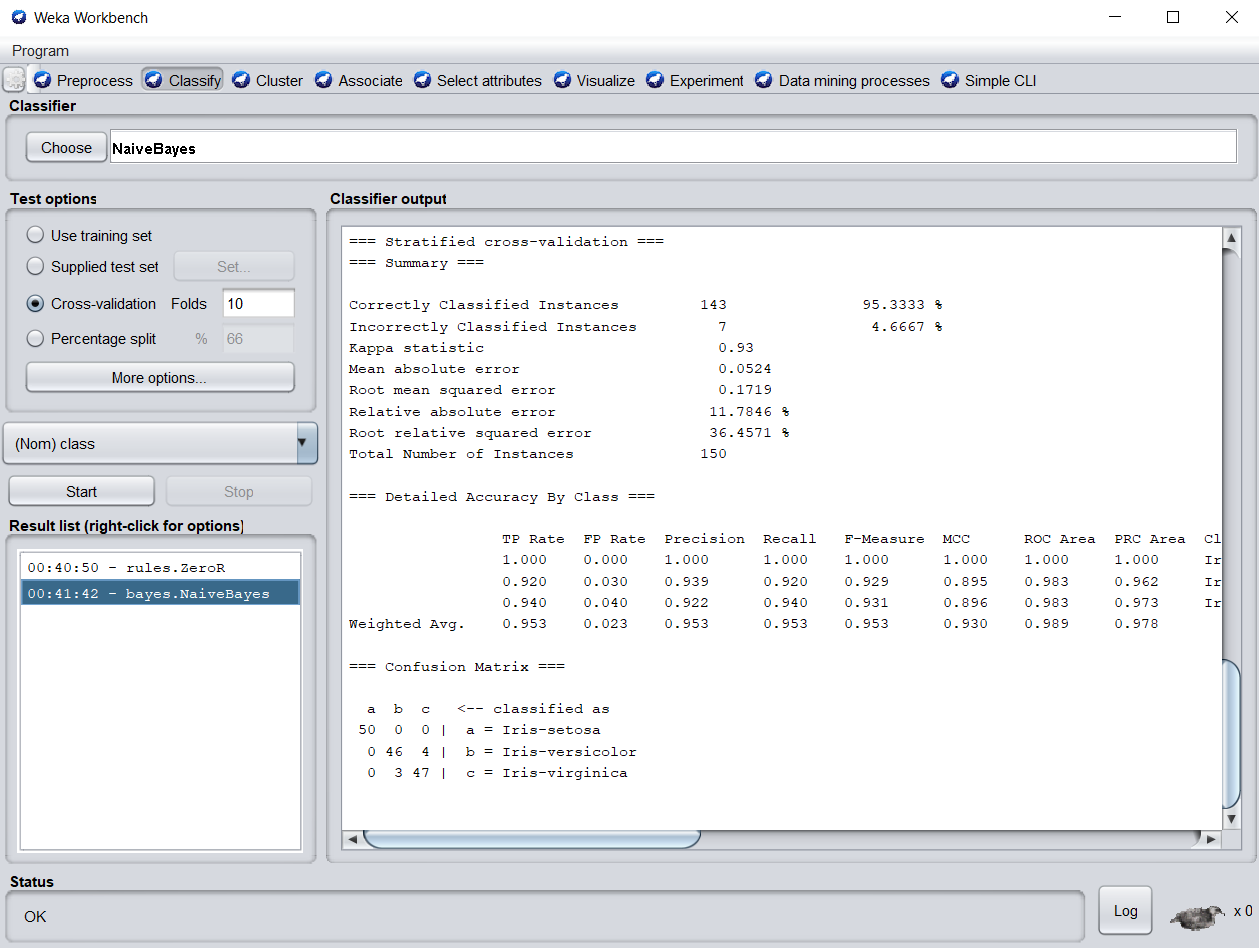
Increase the point size and the jitter and click the “Update” button to set an improved plot of the categorical attributes of the loaded dataset.



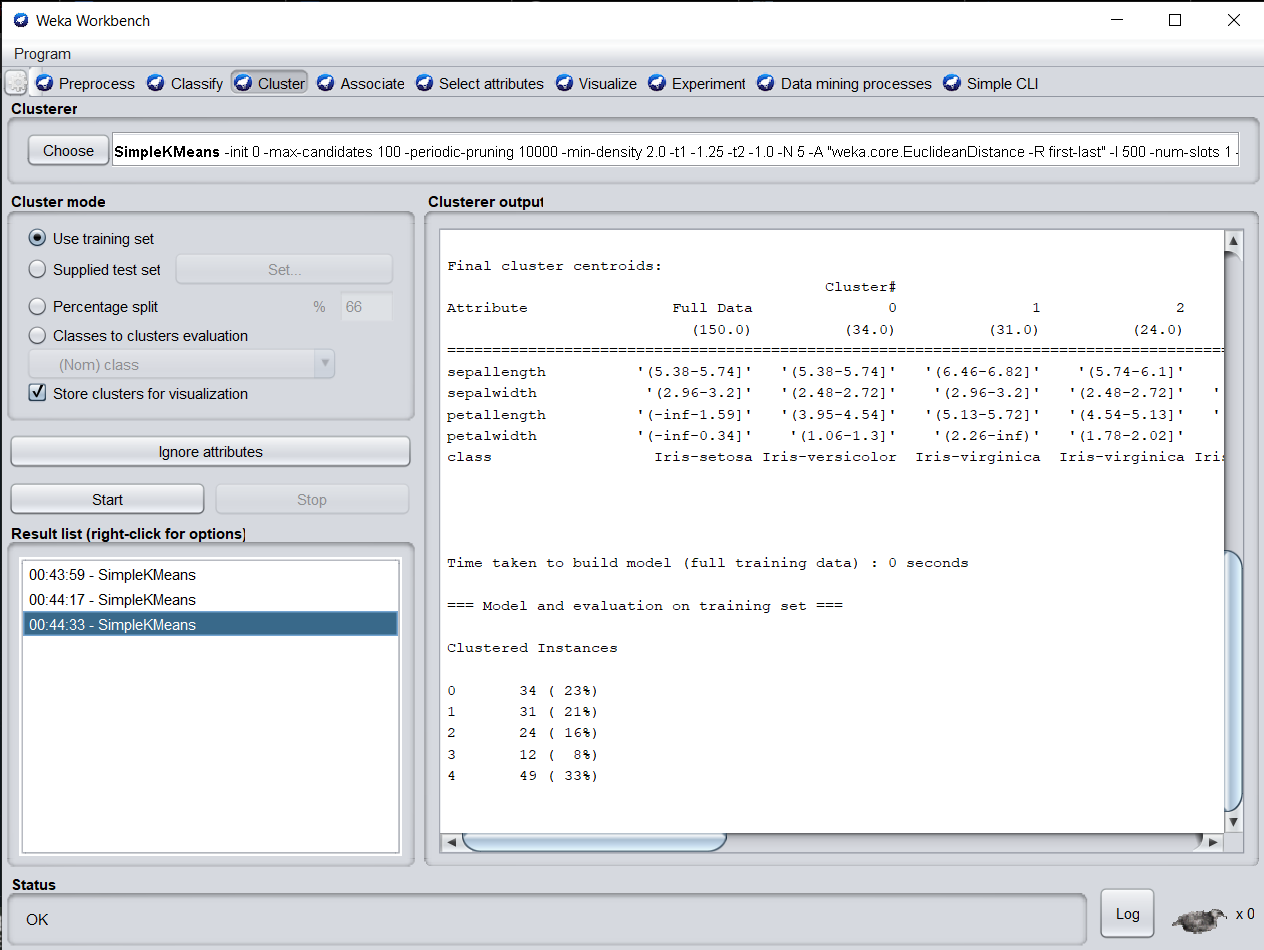
The **select attributes** tab is for performing feature selection on the loaded dataset and identifying those features that are most likely to be relevant in developing a predictive model.

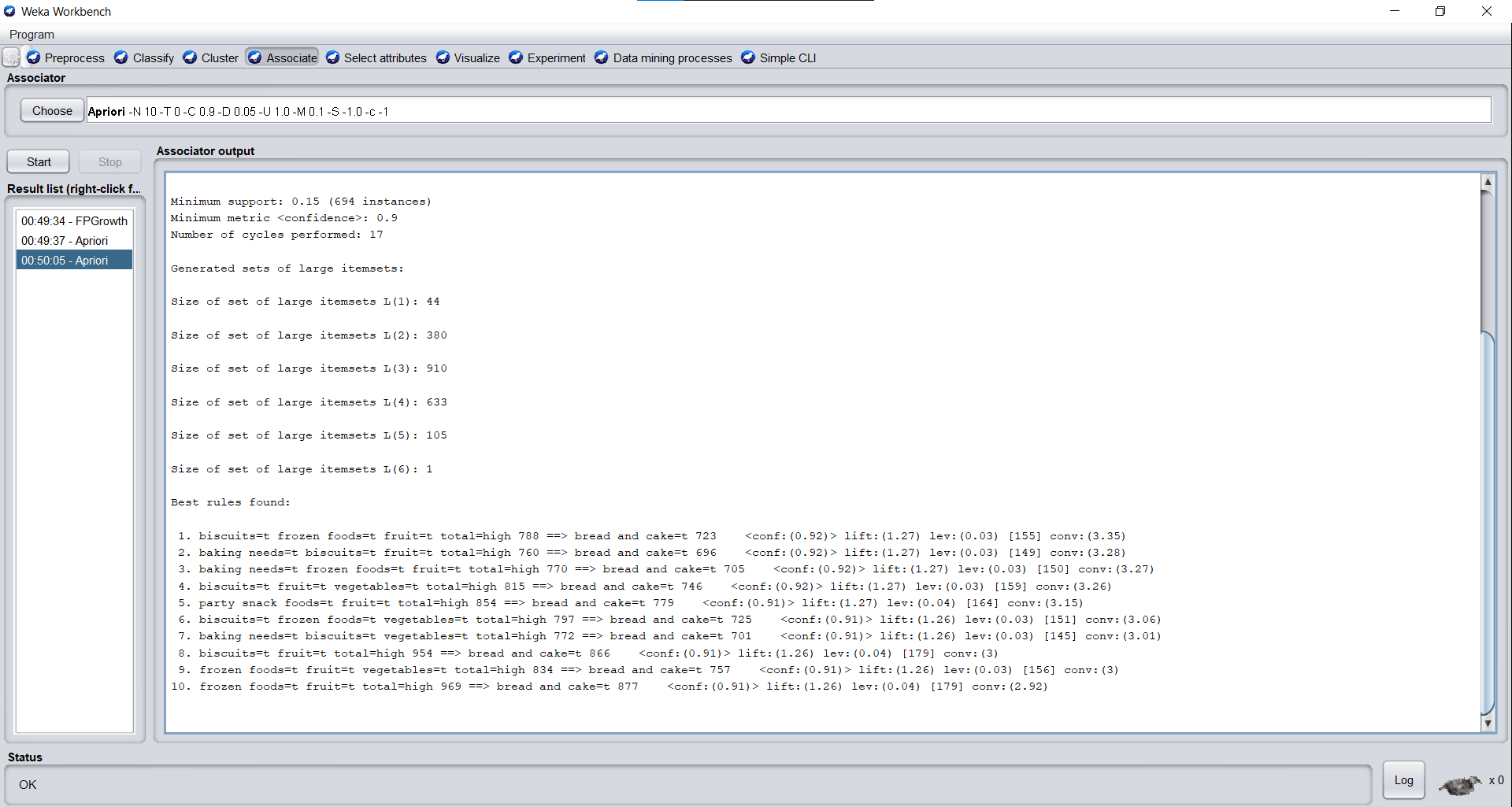


The **classify** tab is for training and evaluating the performance of different machine learning algorithms on your classification or regression problem. Algorithms are divided up into groups, results are kept in a result list and summarized in the main Classifier output.



The **cluster** tab is for training and evaluating the performance of different unsupervised clustering algorithms on your unlabeled dataset. Like the Classify tab, algorithms are divided into groups, results are kept in a result list and summarized in the main Clusterer output.



The **associate** tab is for automatically finding associations in a dataset. The techniques are often used for market basket analysis type data mining problems and require data where all attributes are categorical.

**CONCLUSION**: We learnt about the Weka tool and how to do data analysis with it. We used 2 different databases : Iris petals and Supermarket.

We tried both the supervised and unsupervised learning algorithms. We can easily visualize with charts how the data transforms when we filter it using different algorithms.

We also used the select attribute to find out which attribute is ranked best for classification. We implemented different clustering and classification algorithms.

In the second database i.e. the supermarket one, we implemented the associate function where we found the different associations in a dataset.