EXERCISE -6

(i) Implement Simple Linear Regression on an 'Employee' dataset.
(ii) Demonstrate the simple k-Means clustering algorithm on 'iris' dataset.

6.1.1 Problem Statement:

Implement Simple Linear Regression on an 'Employee' dataset.

6.1.2 Description:

About Dataset used:

HR data can be hard to come by, and HR professionals generally lag behind with respect to analytics and data visualization competency. Thus, Dr. Carla Patalano and I set out to create our own HR-related dataset, which is used in one of our graduate MSHRM courses called HR Metrics and Analytics, at New England College of Business. We created this data set ourselves. We use the data set to teach HR students how to use and analyze the data in Tableau Desktop - a data visualization tool that's easy to learn.
This version provides a variety of features that are useful for both data visualization AND creating machine learning / predictive analytics models. We are working on expanding the data set even further by generating even more records and a few additional features. We will be keeping this as one file/one data set for now. There is a possibility of creating a second file perhaps down the road where you can join the files together to practice SQL/joins, etc.
Note that this dataset isn't perfect. By design, there are some issues that are present. It is primarily designed as a teaching data set - to teach human resources professionals how to work with data and analytics.

About Arff:

An ARFF (Attribute-Relation File Format) file is an ASCII text file that describes a list of instances sharing a set of attributes.

ARFF files have two distinct sections. The first section is the **Header** information, which is followed the **Data** information.

The **Header** of the ARFF file contains the name of the relation, a list of the attributes (the columns in the data), and their types.

Lines that begin with a % are comments. The **@RELATION**, **@ATTRIBUTE** and **@DATA** declarations are case insensitive.

About CSV:

Files with .csv (Comma Separated Values) extension represent plain text files that contain records of data with comma separated values. Each line in a CSV file is a new record from the set of records contained in the file. Such files are generated when data transfer is intended from one storage system to another. Since all applications can recognize records separated by comma, import of such data files to database is done very conveniently. Almost all spreadsheet applications such as Microsoft Excel or OpenOffice Calc can import CSV without much effort. Data imported from such files is arranged in cells of a spreadsheet for representation to user.

Datatypes that are supported by Weka:

- numeric

- integer is treated as numeric
- real is treated as numeric
- [nominal-specification]
- string
- date [date-format]
- relational for multi-instance data (for future use)

where [nominal-specification] and [date-format] are defined below. The keywords **numeric**, **real**, **integer**, **string** and **date** are case insensitive.

Numeric attributes
Numeric attributes can be real or integer numbers.

Nominal attributes
Nominal values are defined by providing an [nominal-specification] listing the possible values: {[nominal-name1], [nominal-name2], [nominal-name3], ...}

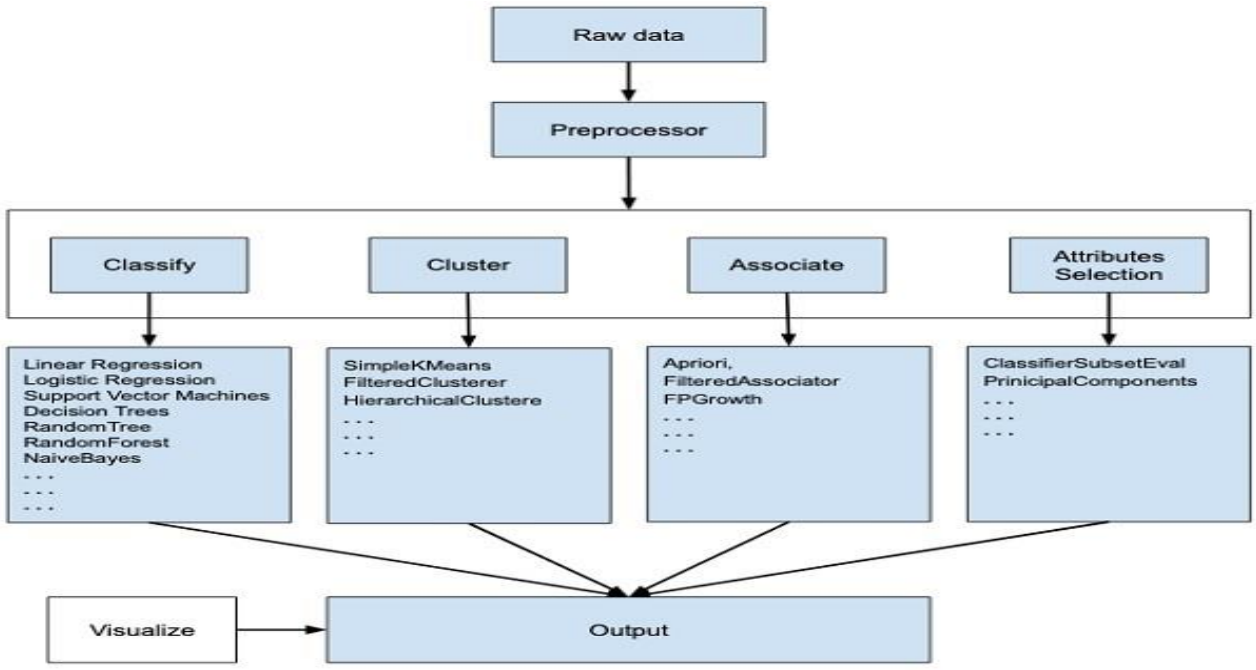
For example, the class value of the Iris dataset can be defined as follows:

```
@ATTRIBUTE class       {Iris-setosa,Iris-versicolor,Iris-virginica}
```

Values that contain spaces must be quoted.

About WEKA Software:

WEKA - an opensource software provides tools for data preprocessing, implementation of several Machine Learning algorithms, and visualization tools so that you can develop machine learning techniques and apply them to real-world data mining problems. What WEKA offers is summarized in the following diagram −



If you observe the beginning of the flow of the image, you will understand that there are many stages in dealing with Big Data to make it suitable for machine learning

First, you will start with the raw data collected from the field. This data may contain several null values and irrelevant fields. You use the data preprocessing tools provided in WEKA to cleanse the data.

Then, you would save the preprocessed data in your local storage for applying ML algorithms.

Next, depending on the kind of ML model that you are trying to develop you would select one of the options such as **Classify, Cluster**, or **Associate**. The **Attributes Selection** allows the automatic selection of features to create a reduced dataset.

Note that under each category, WEKA provides the implementation of several algorithms. You would select an algorithm of your choice, set the desired parameters and run it on the dataset.

Then, WEKA would give you the statistical output of the model processing. It provides you a visualization tool to inspect the data.

The various models can be applied on the same dataset. You can then compare the outputs of different models and select the best that meets your purpose.

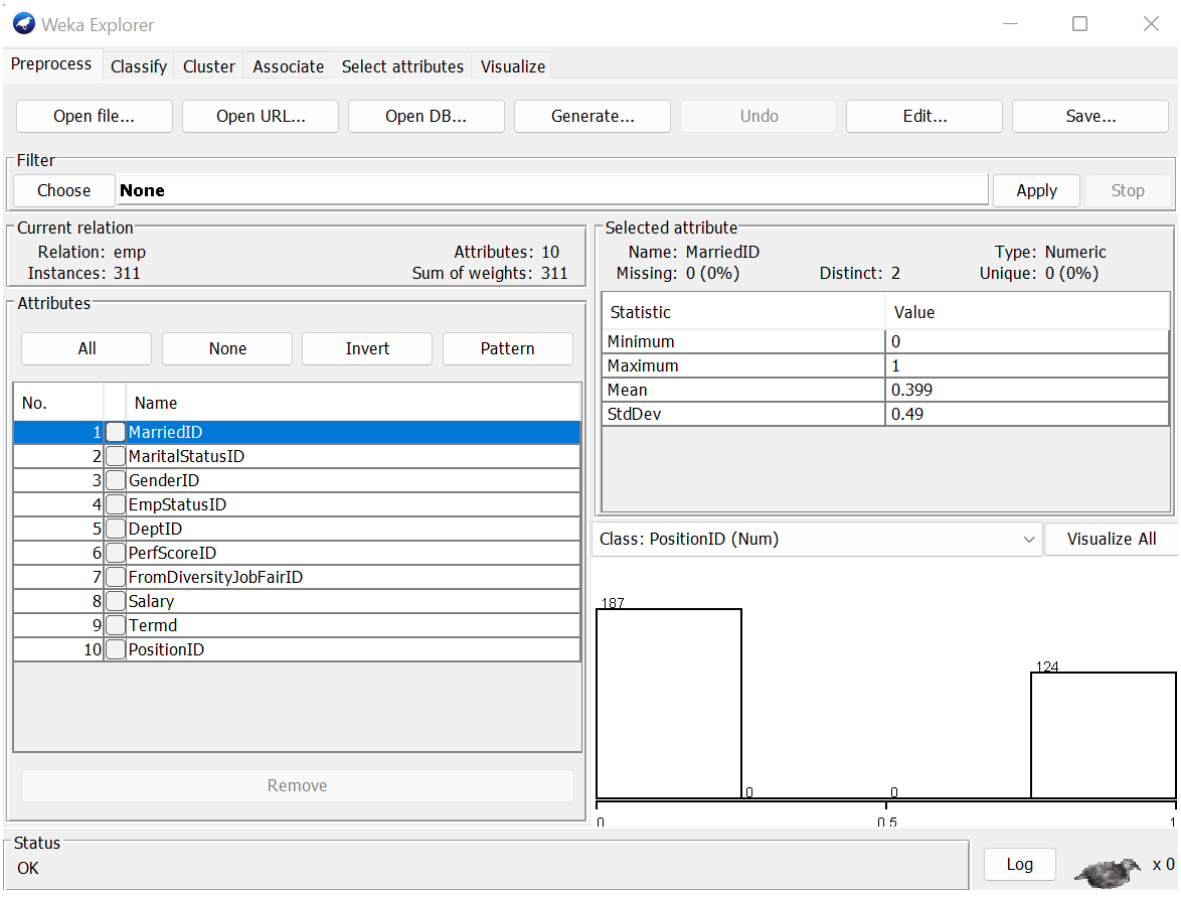
Thus, the use of WEKA results in a quicker development of machine learning models on the whole.

Now that we have seen what WEKA is and what it does, in the next chapter let us learn how to install WEKA on your local computer

6.1.3 Steps to implement the Simple Linear Regression:

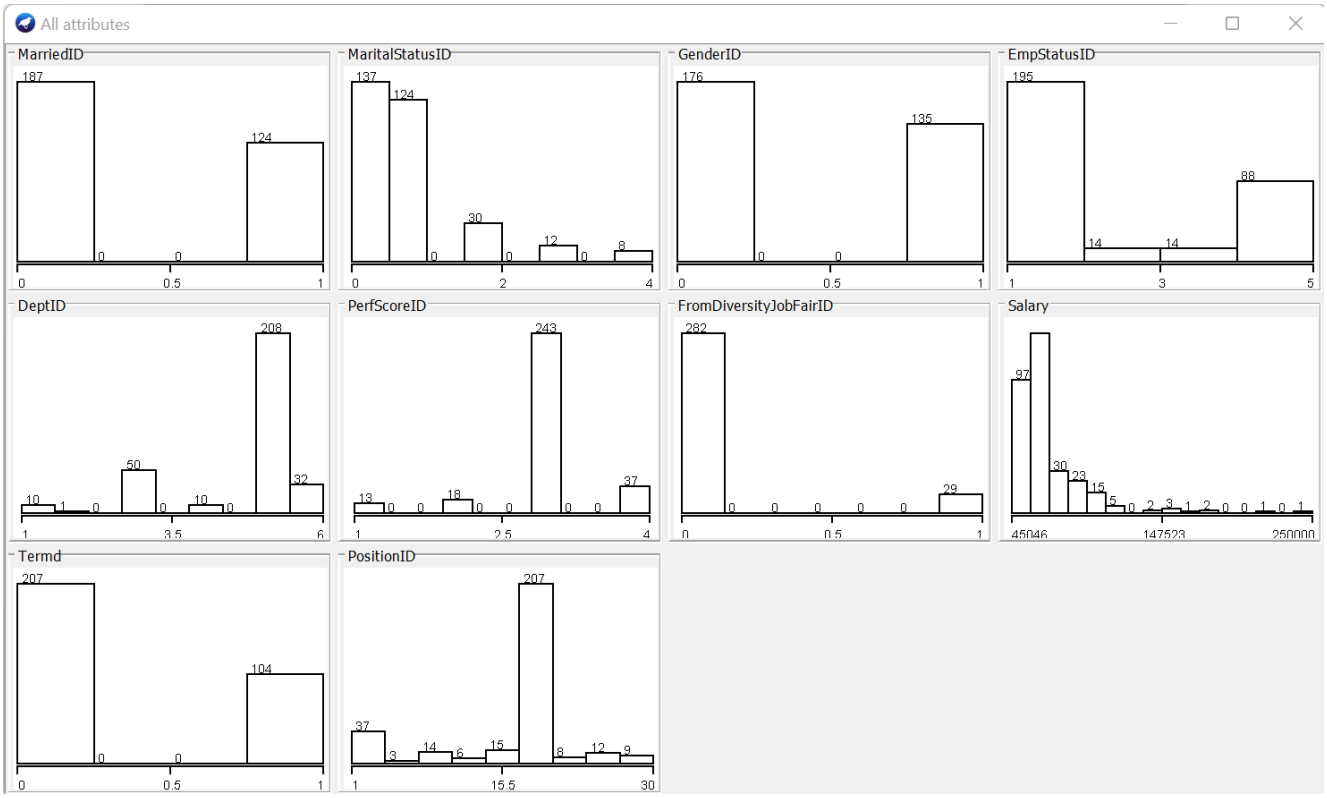
1. To implement Simple Linear Regression on an 'Employee' dataset.

So first open the file by using the **Open file** ... option and select the **employee.arff** file.

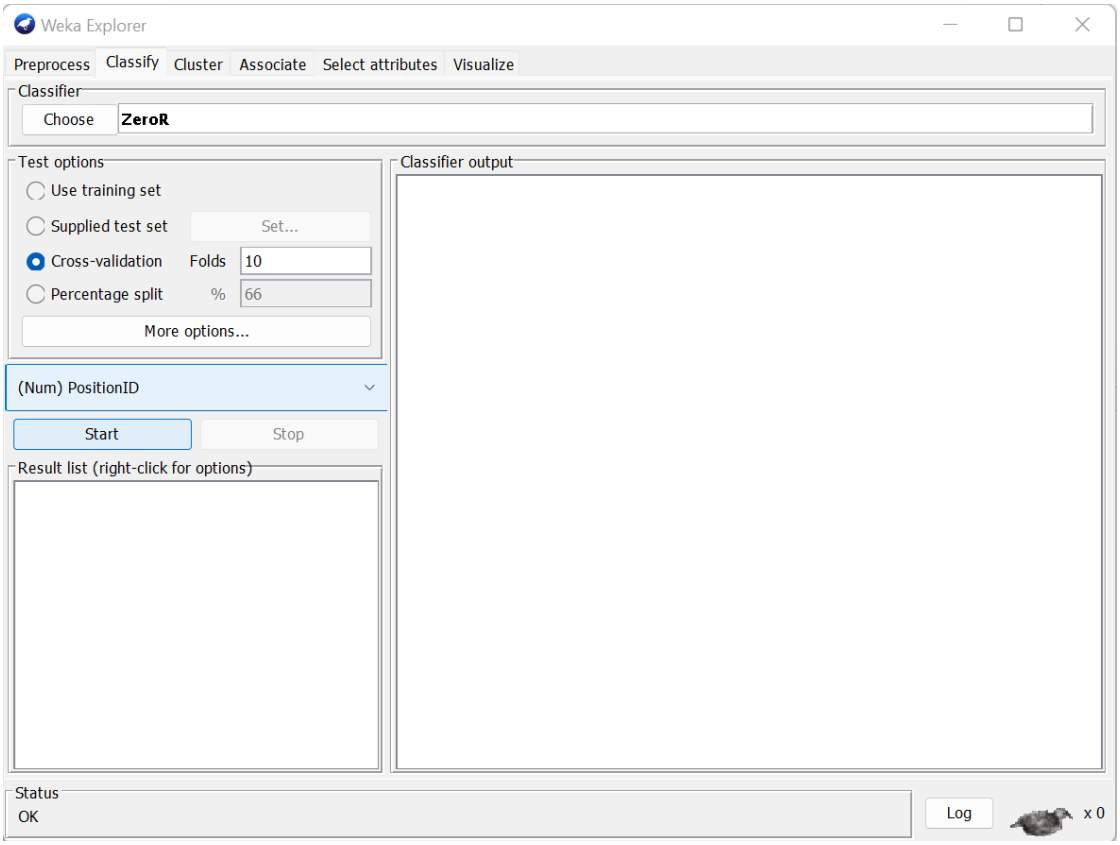


2. At the bottom of the window, you see the visual representation of the **class** values.

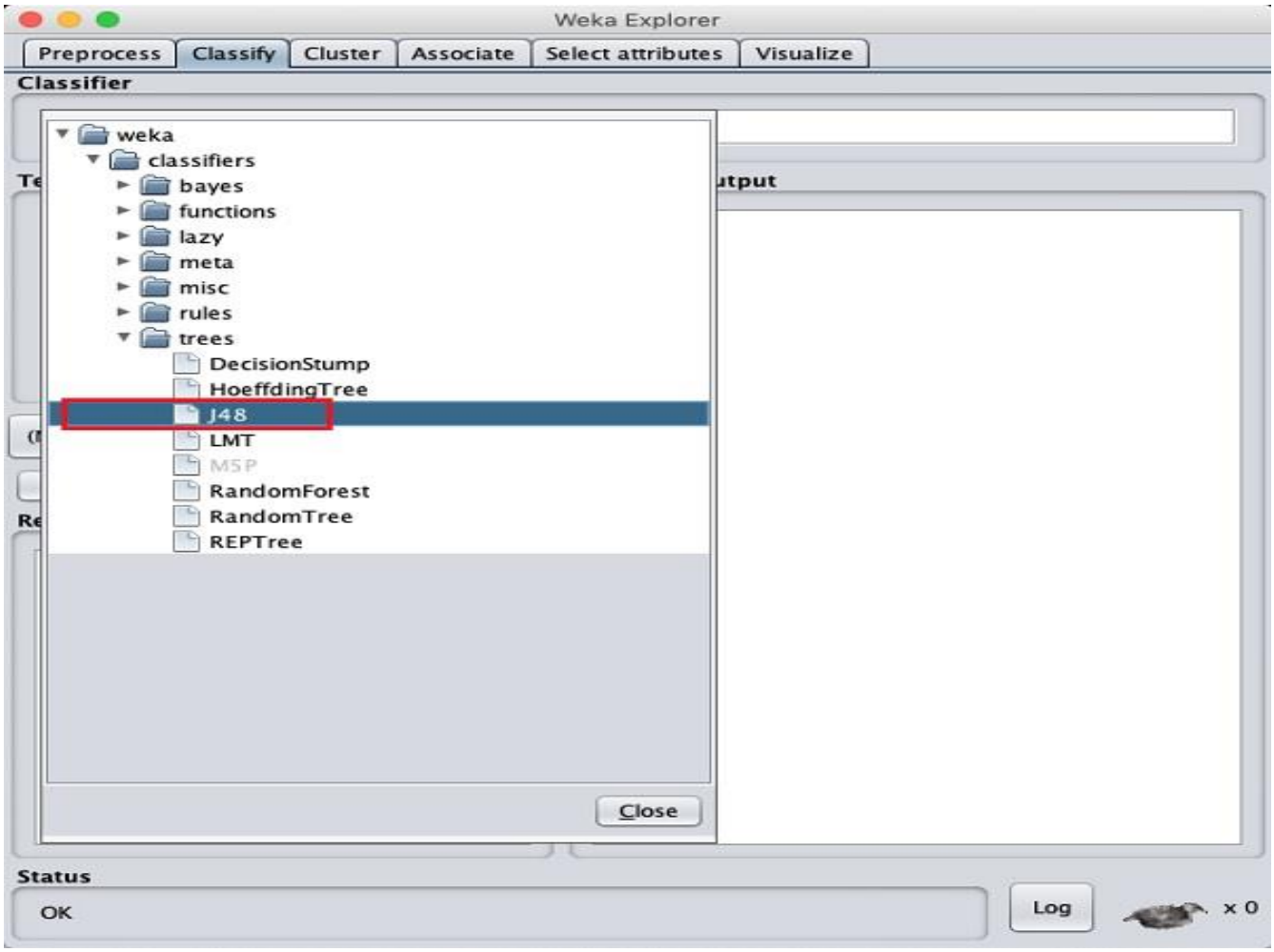
If you click on the **Visualize All** button, you will be able to see all features in one single window as shown here



3. Now to classify the **employee** database select and open classify option

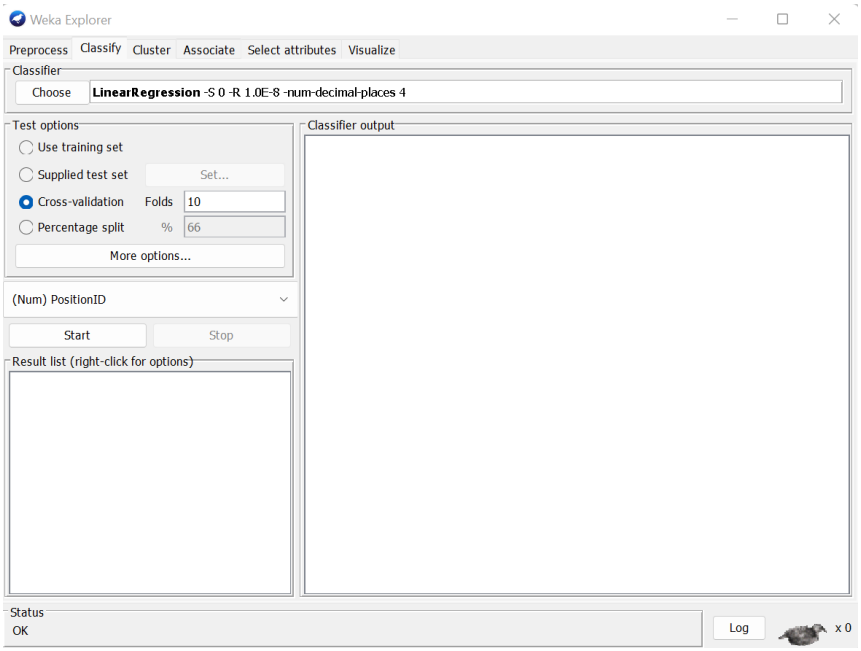


There are many classification methods are available in weka software some of them are bayes, meta, trees, lazy, rules ect..

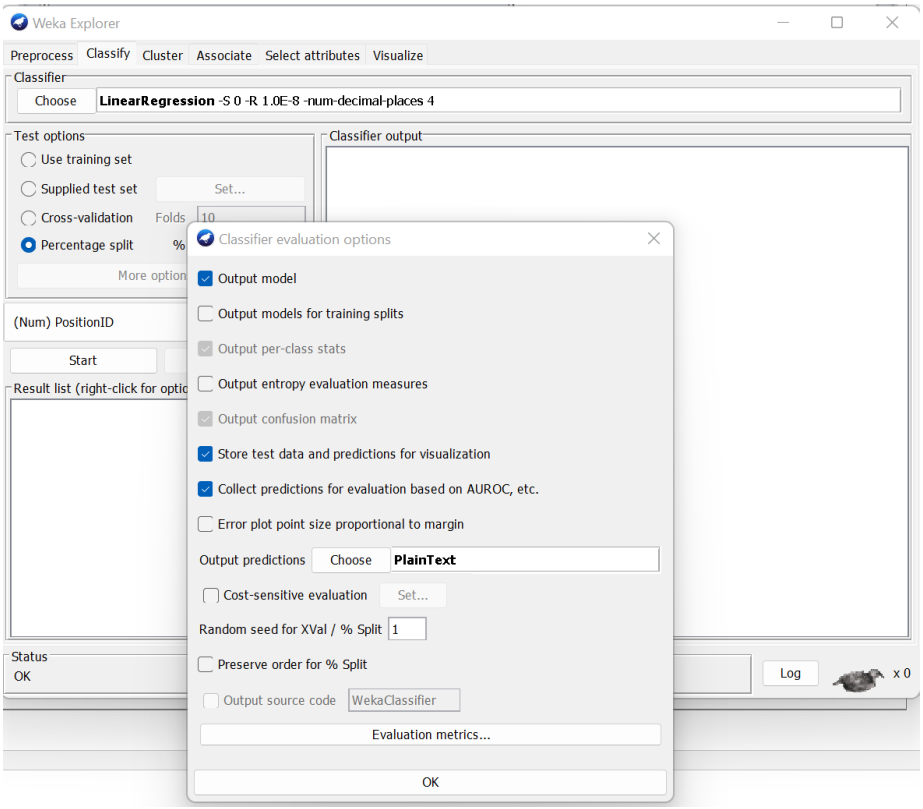


4. According to the problem we are implementing the simple linear regression.

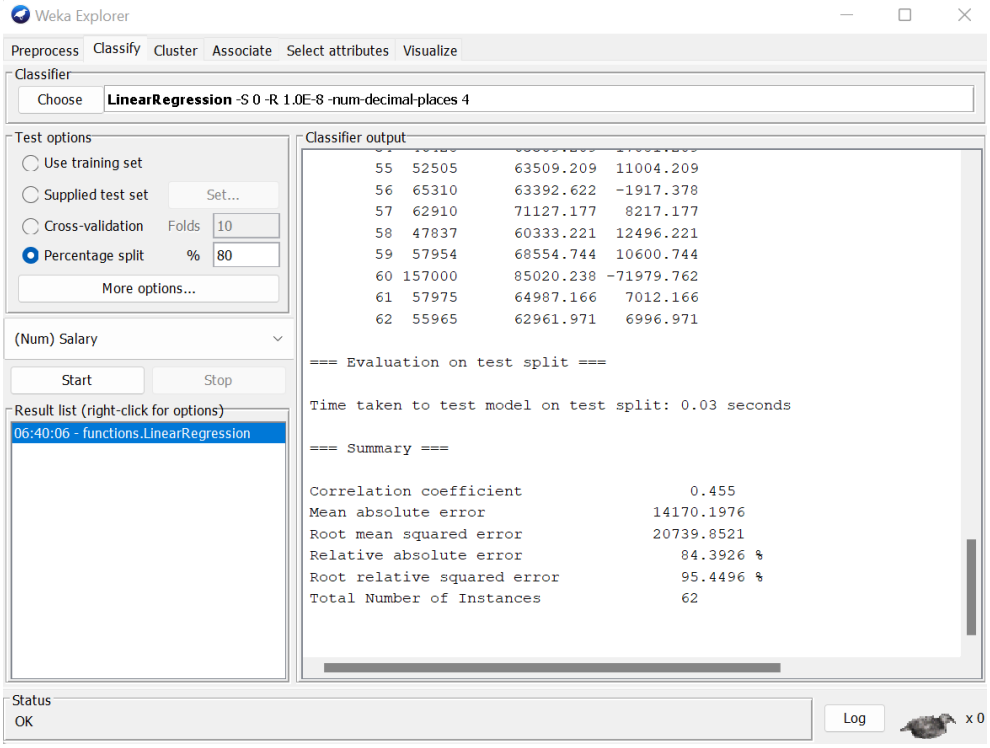
So, select choose > classifiers > functions > LinearRegression



5. In the More options.. Select output predictions as plain text and set percentage split are 80% and set the target parameter as salary.



6. Click on start and we get the classifier output as shown below.



6.1.4 Results and Discussion:

Implementing Simple Linear Regression on an ‘Employee’ dataset is successfully completed.\

6.2.1 Problem Statement:

Demonstrate the simple k-Means clustering algorithm on 'iris' dataset.

6.2.2 Description:

About Dataset used

The attributes are:
1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class:
6. Number of samples of each species of iris flowers:
7. Predicted attribute: class of iris plant.
8. Missing Attribute Values: None

The Iris Dataset contains information of three species of Iris flowers (Iris setosa, Iris virginica and Iris versicolor). The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other.
Data Set Characteristics: Multivariate
Area: Life Sciences
Number of samples (or instances) in the dataset: 150
Number of attributes (or features): 05 Attribute Information:
-- Iris Setosa
-- Iris Versicolour
-- Iris Virginica
Class Distribution: 33.3% for each of 3 classes.

About Arff:

An ARFF (Attribute-Relation File Format) file is an ASCII text file that describes a list of instances sharing a set of attributes.

ARFF files have two distinct sections. The first section is the **Header** information, which is followed the **Data** information.

The **Header** of the ARFF file contains the name of the relation, a list of the attributes (the columns in the data), and their types.

Lines that begin with a % are comments. The **@RELATION**, **@ATTRIBUTE** and **@DATA** declarations are case insensitive.

About CSV:

Files with .csv (Comma Separated Values) extension represent plain text files that contain records of data with comma separated values. Each line in a CSV file is a new record from the set of records contained in the file. Such files are generated when data transfer is intended from one storage system to another. Since all applications can recognize records separated by comma, import of such data files to database is done very conveniently. Almost all spreadsheet applications such as Microsoft Excel or OpenOffice Calc can import CSV without much effort. Data imported from such files is arranged in cells of a spreadsheet for representation to user.

Datatypes that are supported by Weka:

- numeric
- integer is treated as numeric
- real is treated as numeric
- [nominal-specification]
- string
- date [date-format]
- relational for multi-instance data (for future use)

where [nominal-specification] and [date-format] are defined below. The keywords **numeric**, **real**, **integer**, **string** and **date** are case insensitive.

Numeric attributes
Numeric attributes can be real or integer numbers.

Nominal attributes
Nominal values are defined by providing an [nominal-specification] listing the possible values: {[nominal-name1], [nominal-name2], [nominal-name3], ...}

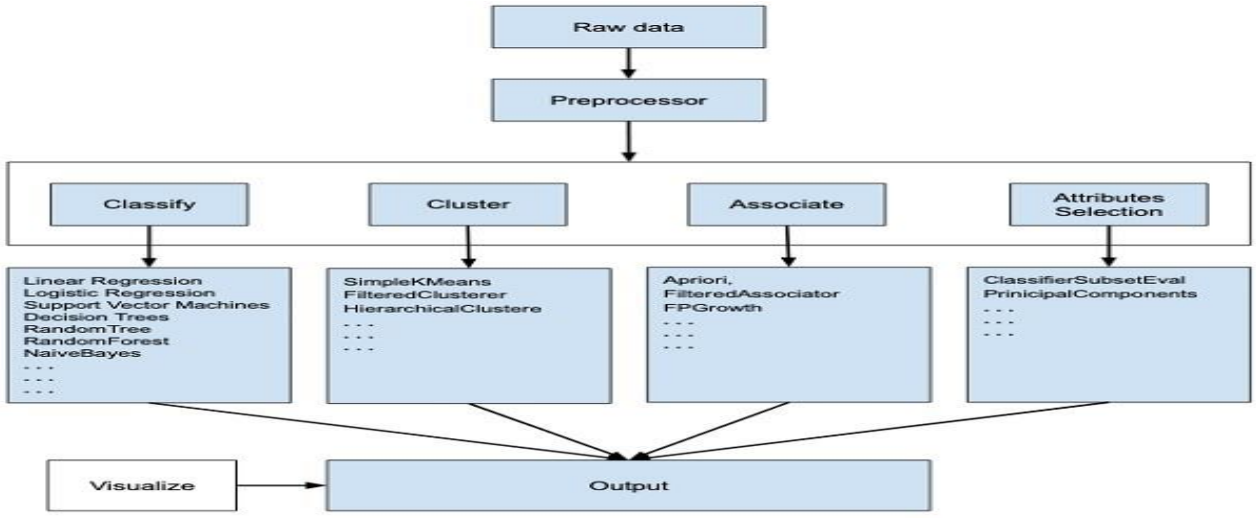
For example, the class value of the Iris dataset can be defined as follows:

```
@ATTRIBUTE class       {Iris-setosa,Iris-versicolor,Iris-virginica}
```

Values that contain spaces must be quoted.

About WEKA Software:

WEKA - an opensource software provides tools for data preprocessing, implementation of several Machine Learning algorithms, and visualization tools so that you can develop machine learning techniques and apply them to real-world data mining problems. What WEKA offers is summarized in the following diagram −



If you observe the beginning of the flow of the image, you will understand that there are many stages in dealing with Big Data to make it suitable for machine learning −

First, you will start with the raw data collected from the field. This data may contain several null values and irrelevant fields. You use the data preprocessing tools provided in WEKA to cleanse the data.

Then, you would save the preprocessed data in your local storage for applying ML algorithms.

Next, depending on the kind of ML model that you are trying to develop you would select one of the options such as **Classify, Cluster**, or **Associate**. The **Attributes Selection** allows the automatic selection of features to create a reduced dataset.

Note that under each category, WEKA provides the implementation of several algorithms. You would select an algorithm of your choice, set the desired parameters and run it on the dataset.

Then, WEKA would give you the statistical output of the model processing. It provides you a visualization tool to inspect the data.

The various models can be applied on the same dataset. You can then compare the outputs of different models and select the best that meets your purpose.

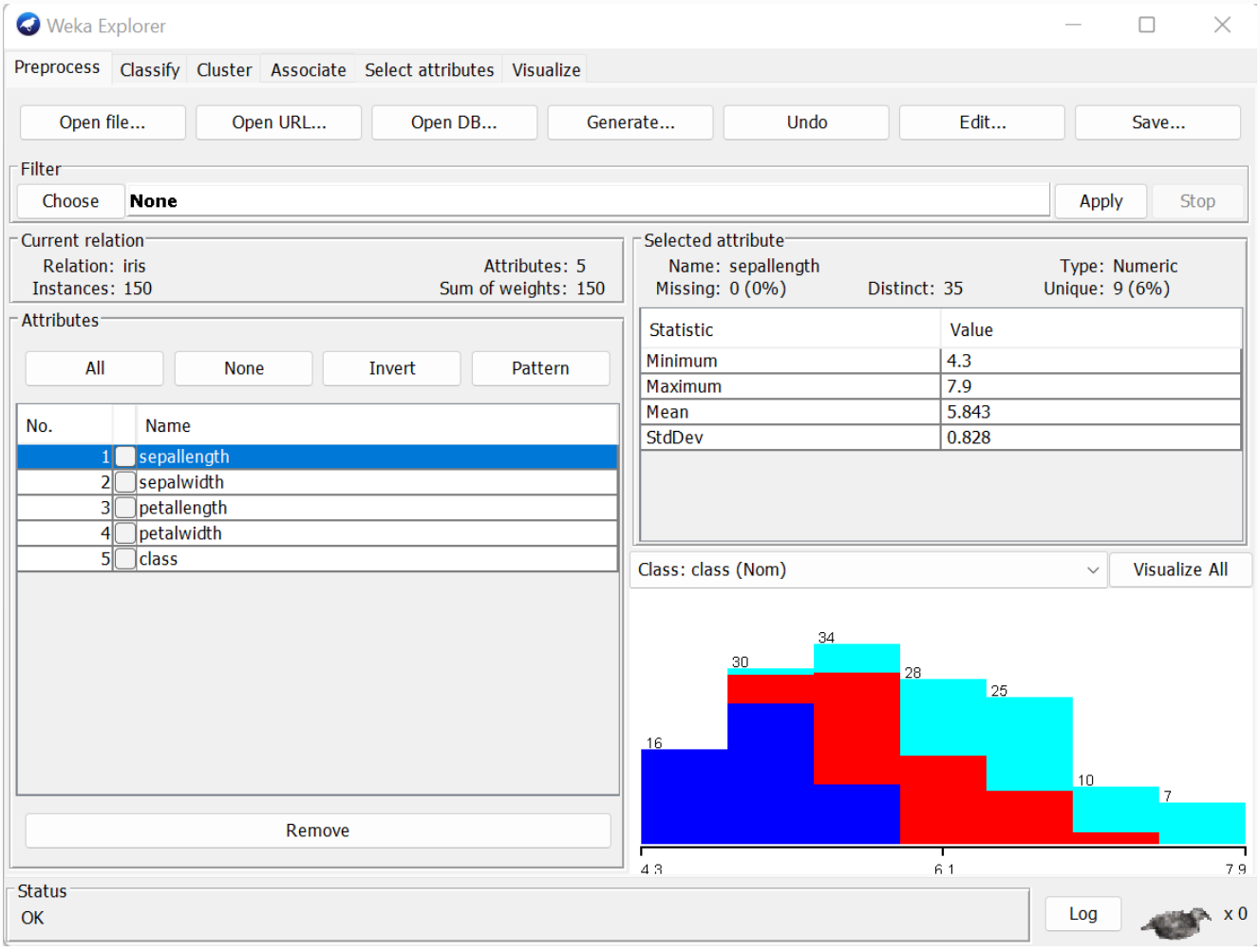
Thus, the use of WEKA results in a quicker development of machine learning models on the whole.

Now that we have seen what WEKA is and what it does, in the next chapter let us learn how to install WEKA on your local computer.

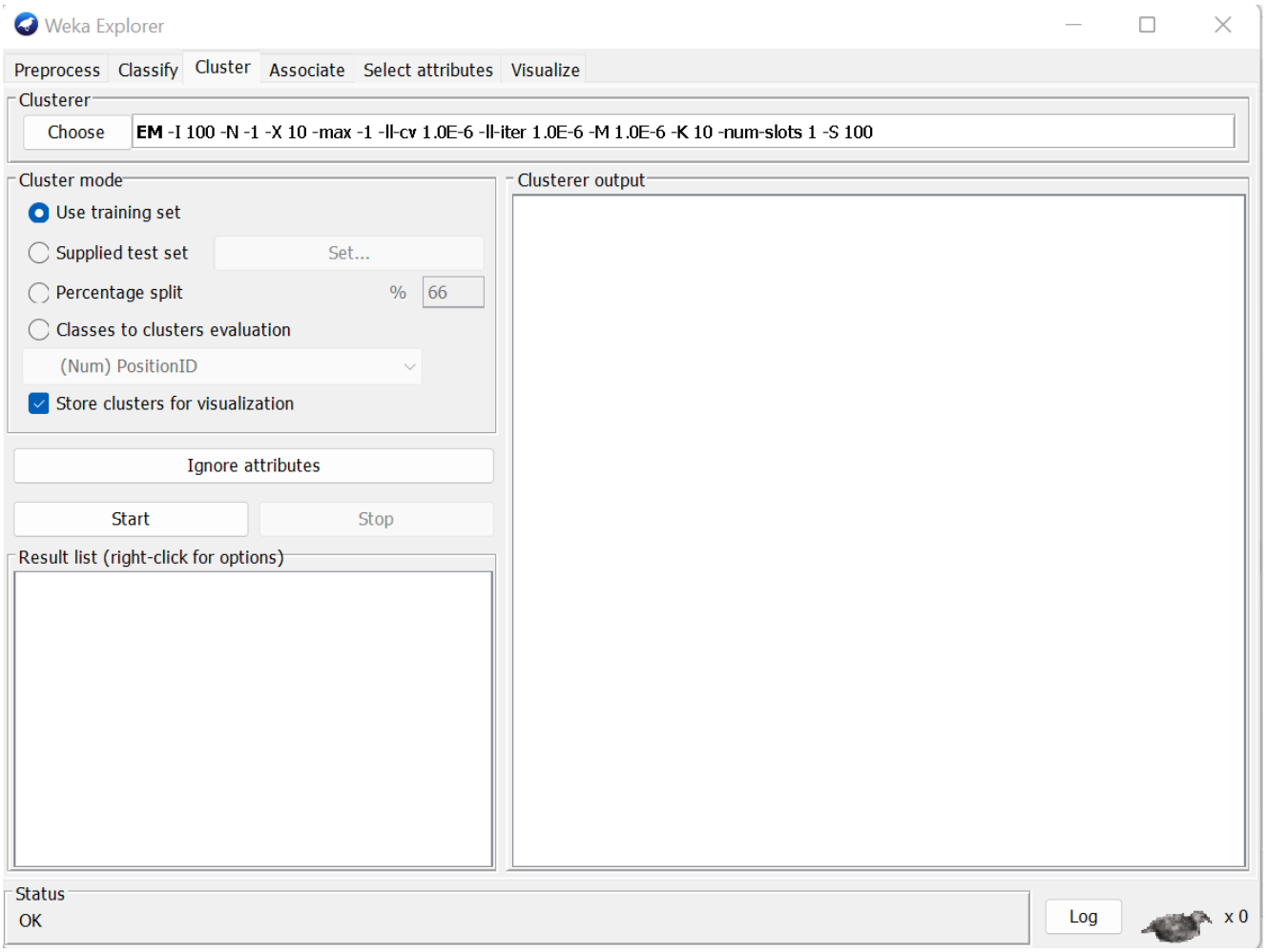
6.2.3 Steps to implement the simple K-Means clustering algorithm:

1. To implement the simple K-Means clustering algorithm on the Iris database.

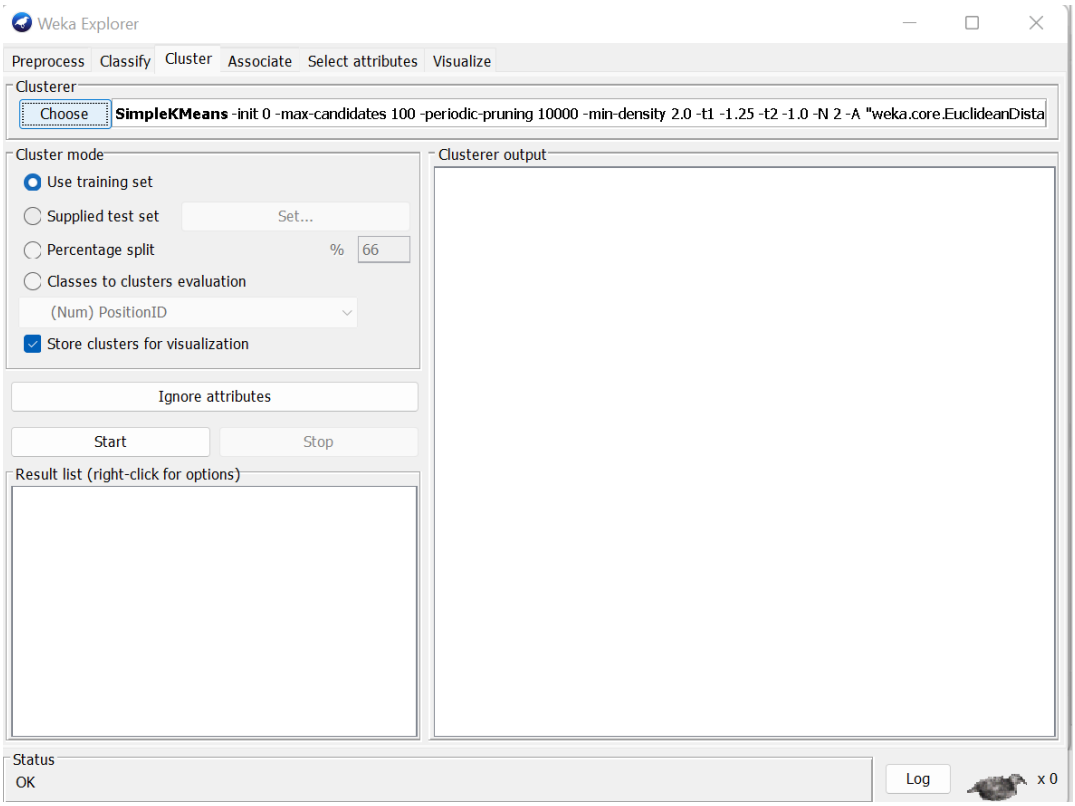
So first open the file by using the **Open file** ... option and select the **Iris.arff** file.



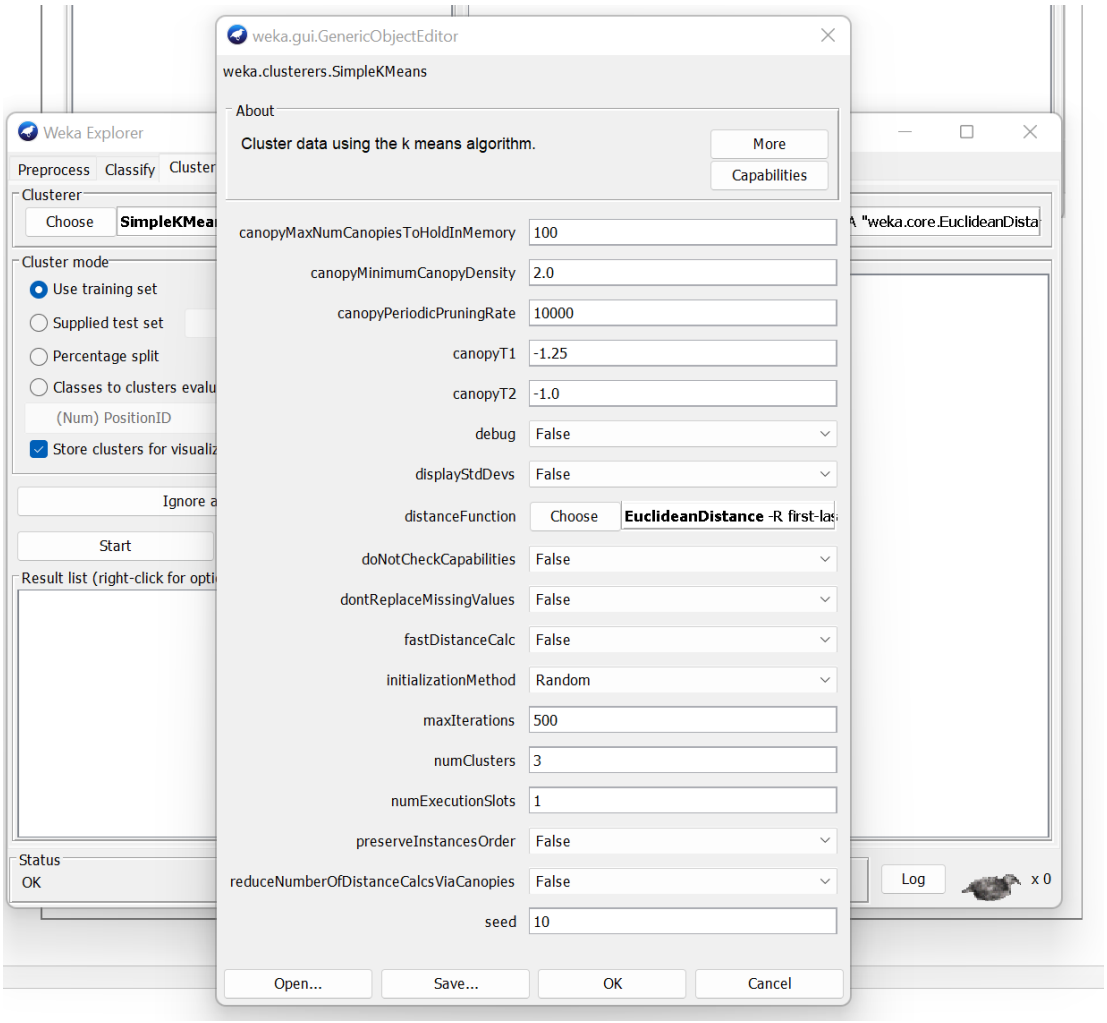
2. Now to apply the clustering on the **Iris** database select and open cluster tab.



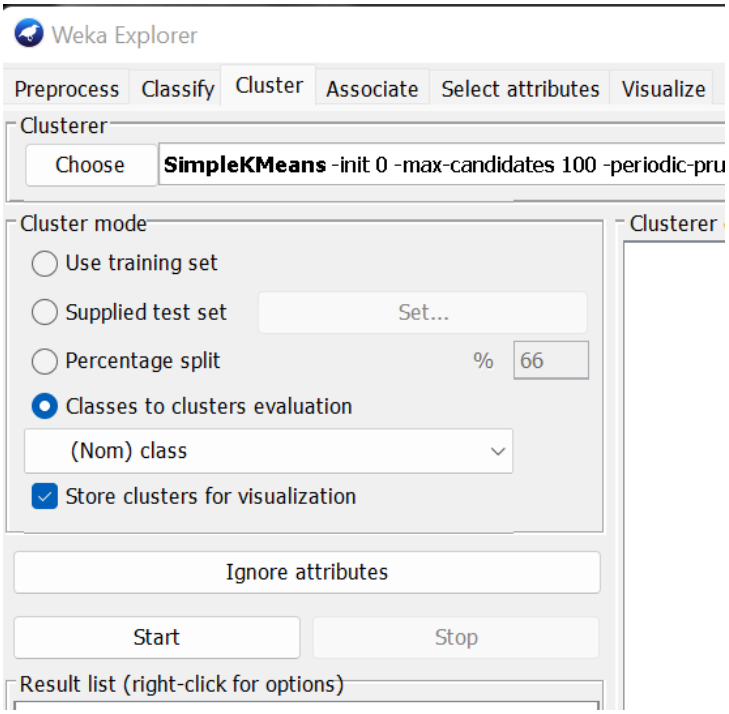
3. Select choose and select SimpleKMeans



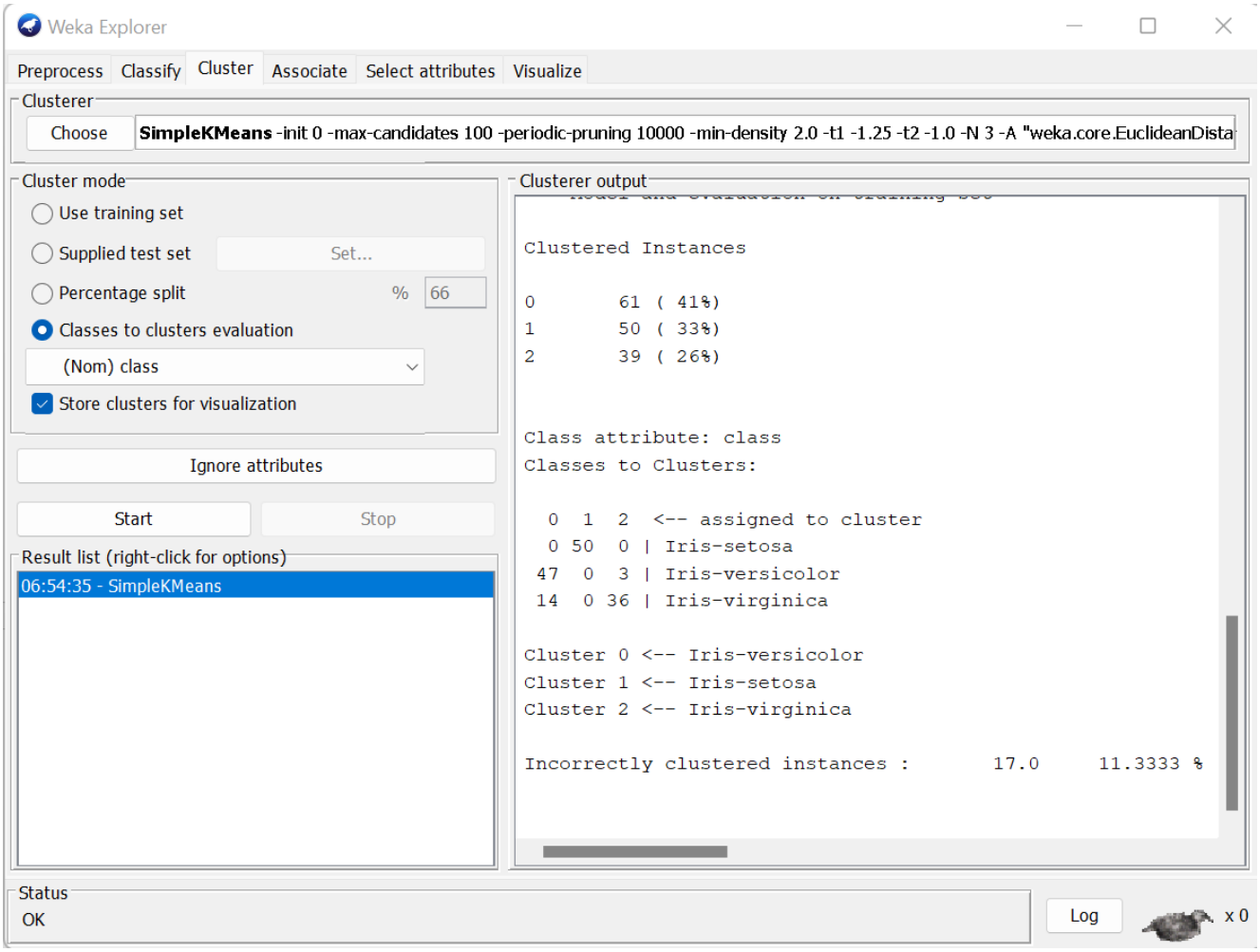
4. Now click on SimpleKMeans and set the numClusters to 3 as we have 3 classes.



5. Select classes to cluster evaluation as (Nom) class and click on start.



6. We get the classifier output as shown below



6.2.4 Results and Discussion:

Demonstration of the simple k-Means clustering algorithm on ‘iris’ dataset is successfully completed.

We observed that it classified into 3 clusters and 11.33% of the instances are incorrectly classified.