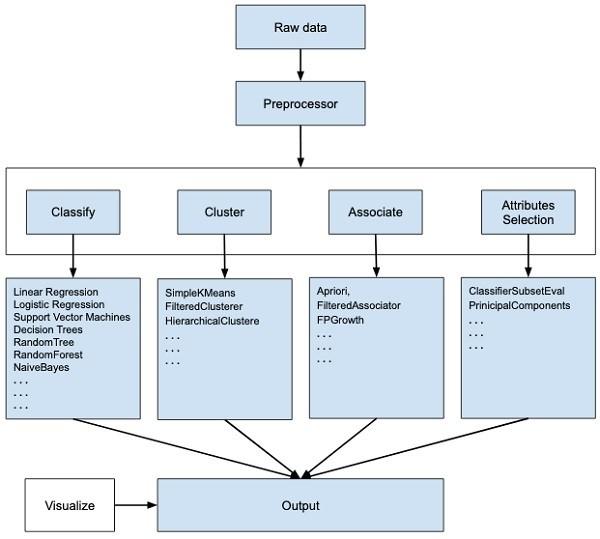
lOMoARcPSD|17241975

**Experiment 6**

**Aim:** Perform data Pre-processing task and demonstrate Classification, Clustering, Association algorithm on data sets using data mining tool (WEKA/R tool)

# Theory:

WEKA is an open-source 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.



The use of WEKA results in a quicker development of machine learning models overall.

**Data Preprocessing:** Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.

**Data Classification:** Data classification is broadly defined as the process of organizing data by relevant categories so that it may be used and protected more efficiently. On a basic level, the classification process makes data easier to locate and retrieve.

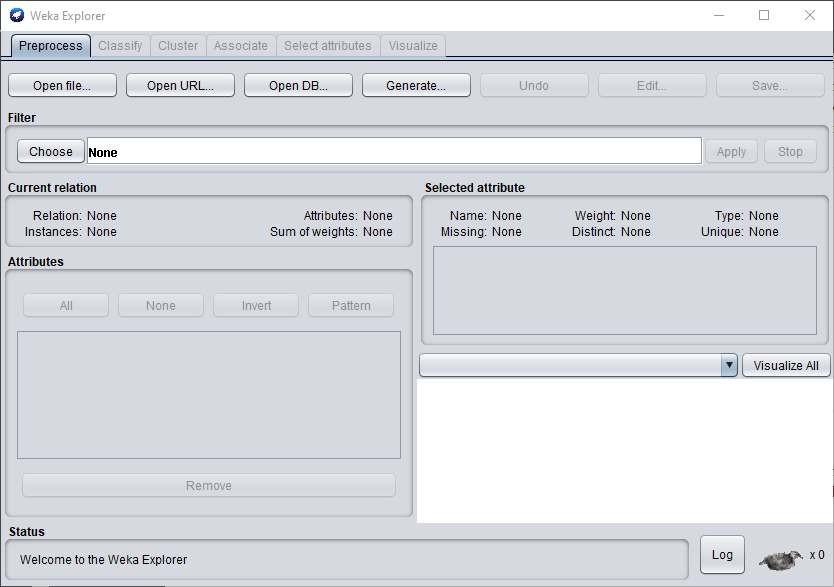
**Data Clustering:** Clustering is a classic data mining technique based on machine learning that divides groups of abstract objects into classes of similar objects. Clustering helps to split data into several subsets.

**Data Association:** A data association is a user-defined grouping of related groups and elements. It can consist of one or more groups along with some or all the elements within those groups.

Once you isolate the groups and elements, you can create data associations, which you can then use to create Rules of Visibility

# Implementation:

**Loading Dataset:**



The first four buttons at the top of the preprocess section enable you to load data into WEKA:

1. **Open file**: Brings up a dialog box allowing you to browse for the data file on the local file system.
2. **Open URL:** Asks for a Uniform Resource Locator address for where the data is stored.
3. **Open DB:** Reads data from a database. (Note that to make this work you might have to edit the file in weka/experiment/DatabaseUtils.props.)
4. **Generate:** Enables you to generate artificial data from a variety of Data Generators.

Using the Open file button, you can read files in a variety of formats: WEKA’s ARFF format, CSV format, C4.5 format, or serialized Instances format. ARFF files typically have a .arff extension, CSV files a .csv extension, C4.5 files a .data and .names extension, and serialized Instances objects a .bsi extension

# Data Preprocessing:

1. Open Weka.
2. Click the “Explorer” button to open the Weka Explorer.
3. Select **weather-numeric.arff** file from the “choose file” under the preprocess tab option.
4. Apply Filters

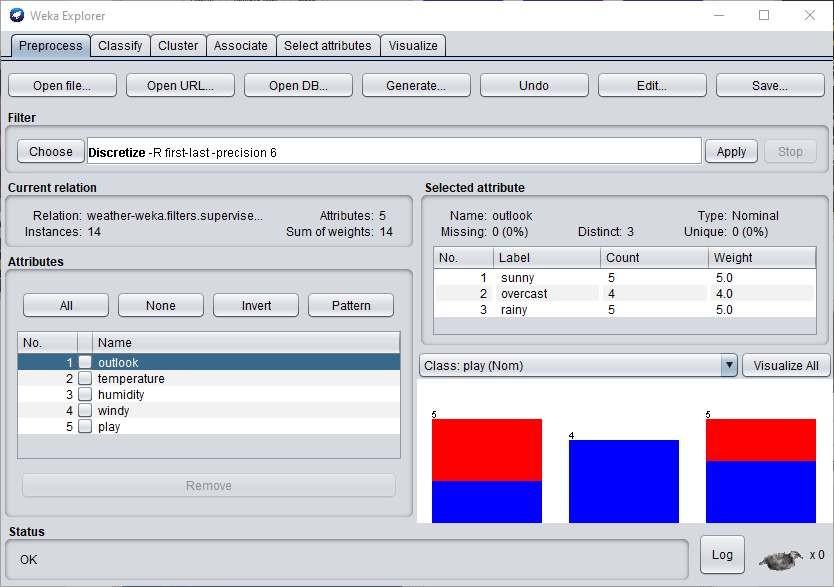
# Applying Filters:

To illustrate the use of filters, we will use **weather-numeric.arff** database that contains two **numeric** attributes - **temperature** and **humidity**.

We will convert these to **nominal** by applying a filter on our raw data. Click on the **Choose** button in the **Filter** section and select the following filter -

# weka→filters→supervised→attribute→Discretize

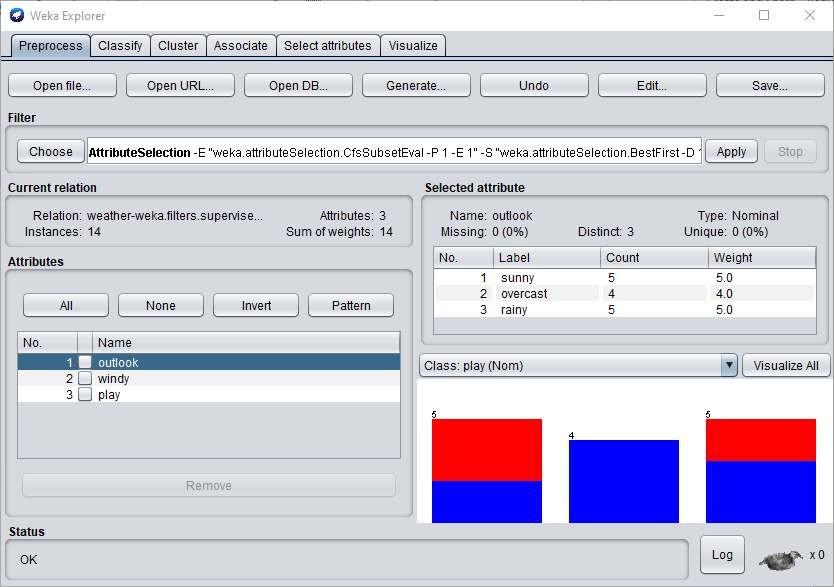
Click on the **Apply** button and examine the **temperature** and/or **humidity** attribute. You will notice that these have changed from numeric to nominal types.



Let us try another filter now. Suppose you want to select the best attributes for deciding the **play**. Select and apply the following filter −

# weka→filters→supervised→attribute→AttributeSelection

You will notice that it removes the temperature and humidity attributes from the database.



After you are satisfied with the preprocessing of your data, save the data by clicking the **Save**

button. You will use this saved file for model building.

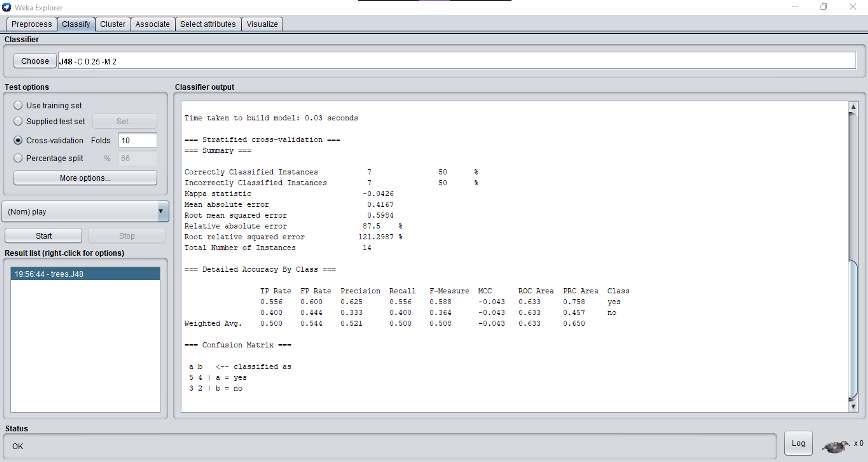
# Data Classification:

1. Open Weka.
2. Click the “Explorer” button to open the Weka Explorer.
3. Select **weather-nominal.arff** file from the “choose file” under the preprocess tab option.
4. Go to the “Classify” tab for classifying the unclassified data. Click on the “Choose” button. From this, select “trees -> J48”.

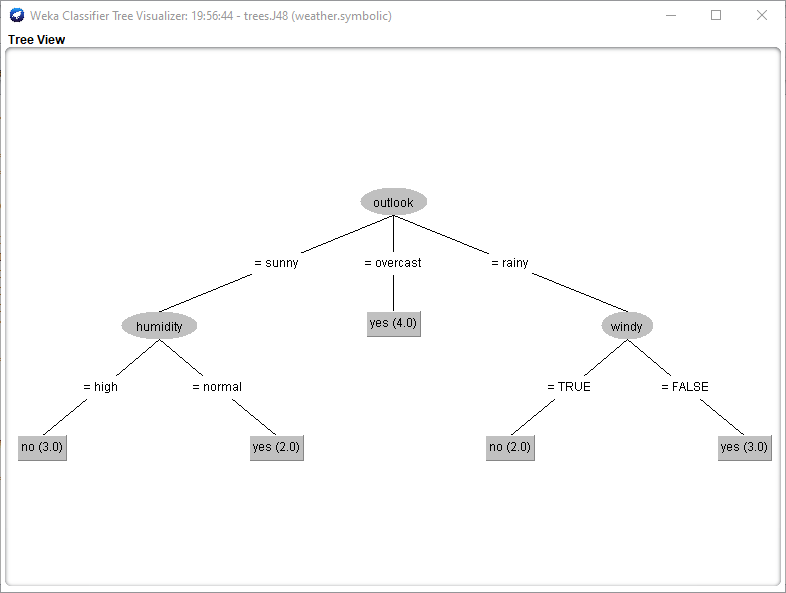
# Let us also have a quick look at other options in the Choose button:

* + **Bayes:** It is a density estimation for numerical attributes.
  + **Meta:** It is a multi-response linear regression.
  + **Functions**: It is logistic regression.
  + **Lazy:** It sets the blend entropy automatically.
  + **Rule:** It is a rule learner.
  + **Trees:** Trees classifies the data.

1. Click on Start Button. The classifier output will be seen on the Right-hand panel. **It shows the run information in the panel as:**
   * **Scheme:** The classification algorithm used.
   * **Instances:** Number of data rows in the dataset.
   * **Attributes:** The dataset has 5 attributes. The number of leaves and the size of the tree describes the decision tree.
   * **Time taken to build the model:** Time for the output. Full classification of the J48 pruned with the attributes and number of instances.

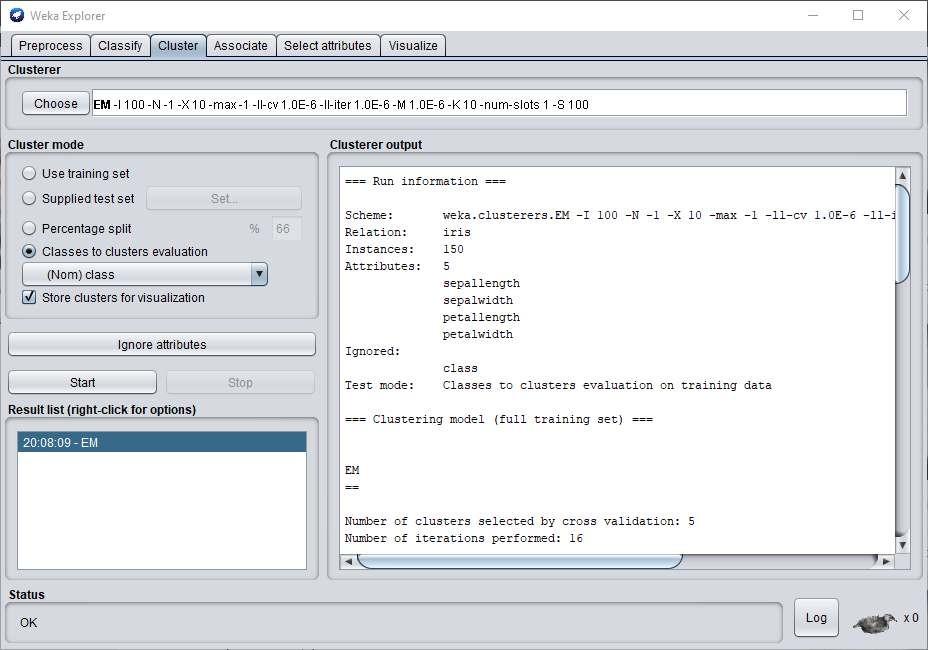


1. To visualize the tree, right-click on the result and select visualize the tree.



# Data Clustering:

1. Open Weka.
2. Click the “Explorer” button to open the Weka Explorer.
3. Select **iris.arff** file from the “choose file” under the preprocess tab option.
4. Go to the “Cluster” to apply the clustering algorithms to our loaded data. Click on the “Choose” button. From this select **EM** as the clustering algorithm.
5. In the **Cluster mode** section, select the **Classes to clusters evaluation**
6. Click on the **Start** button to process the data. After a while, the results will be presented on the screen.



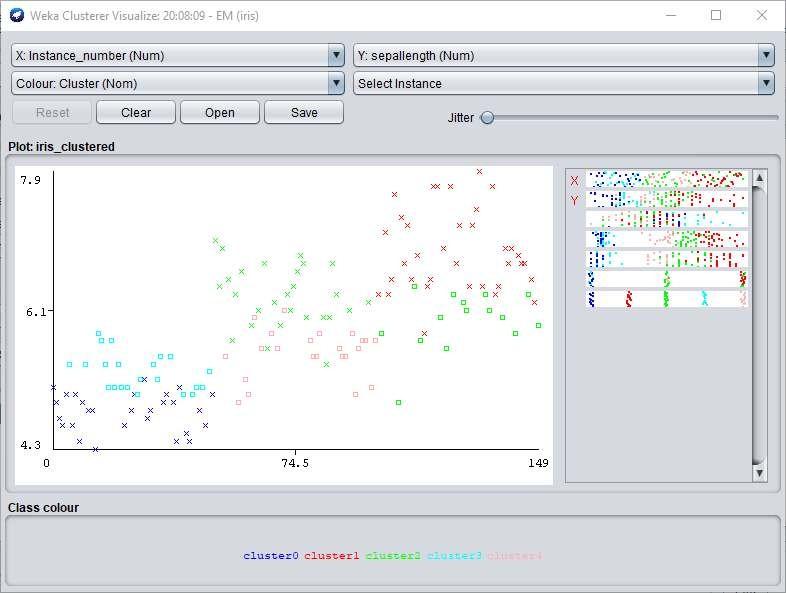
From the output screen, you can observe that:

* + There are 5 clustered instances detected in the database.
  + The **Cluster 0** represents setosa, **Cluster 1** represents virginica, **Cluster**

**2** represents versicolor, while the last two clusters do not have any class associated with them.

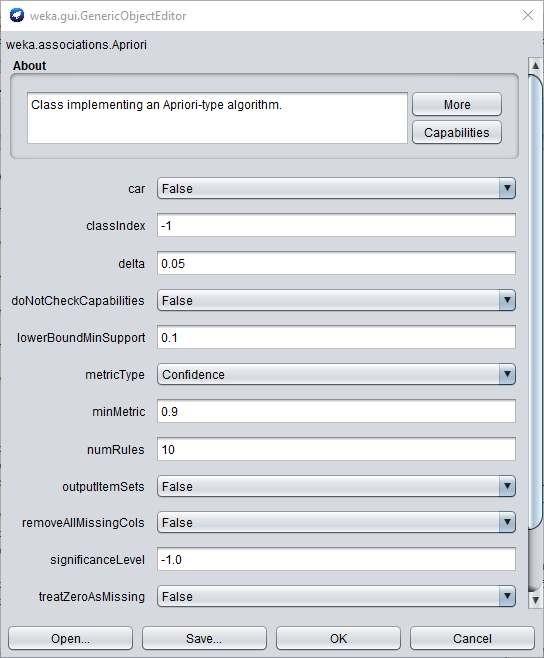
If you scroll up the output window, you will also see some statistics that gives the mean and standard deviation for each of the attributes in the various detected clusters.

1. To visualize the clusters, right click on the **EM** and select **Visualize cluster assignments**.

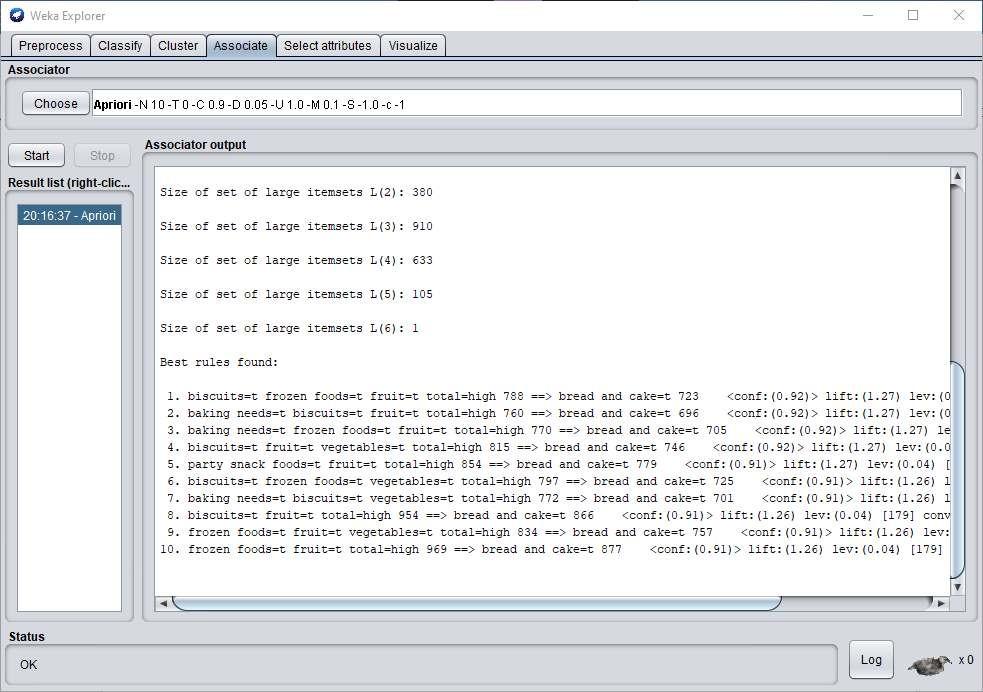


# Data Association:

1. Open Weka.
2. Click the “Explorer” button to open the Weka Explorer.
3. Select **supermarket.arff** file from the “choose file” under the preprocess tab option.
4. Click on the **Associate** tab and click on the **Choose** button. Select the **Apriori** association**.**
5. To set the parameters for the Apriori algorithm, click on its name, a window will pop up as shown below that allows you to set the parameters



1. After you set the parameters, click the **Start** button. After a while you will see the results.



At the bottom, you will find the detected best rules of associations. This will help the supermarket in stocking their products in appropriate shelves.

**Conclusion:** Thus, all data pre-processing tasks and demonstration of Classification, Clustering, Association algorithm on data sets using data mining tool were implemented.