# EXERCISE -2

## 2.1 Problem Statement:

To perform the following preprocessing filters on 'Weather' dataset. (i) Add(ii) Remove (iii) Discretize (iv) Replace Missing values (v) Normalize.

## 2.2 Description:

About Dataset

The weather data is a small open data set with only 14 examples.

In RapidMiner it is named Golf Dataset, whereas Weka has two data set: weather.nominal.arff and weather.numeric.arff

The dataset contains data about weather conditions are suitable for playing a game of golf.

the original dataset that only has 5 variables:

1.outlook

2.temperature

3.humidity

4.windy

5.play

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], ...}

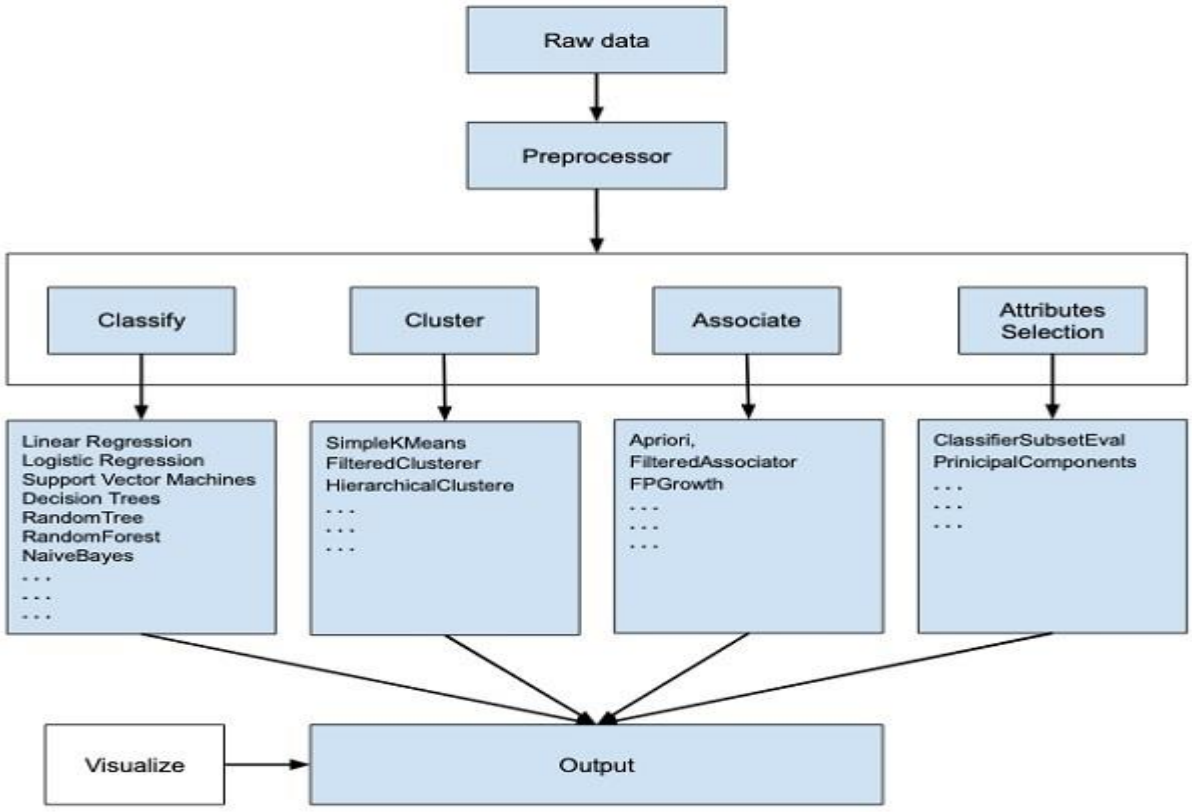
For example, the class value of the Iris dataset can be defined as follows:

```
@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

About preprocessing filters:

Data preprocessing is the process of transforming raw data into a useful, understandable format. Real-world or raw data usually has inconsistent formatting, human errors, and can also be incomplete. Data preprocessing resolves such issues and makes datasets more complete and efficient to perform data analysis

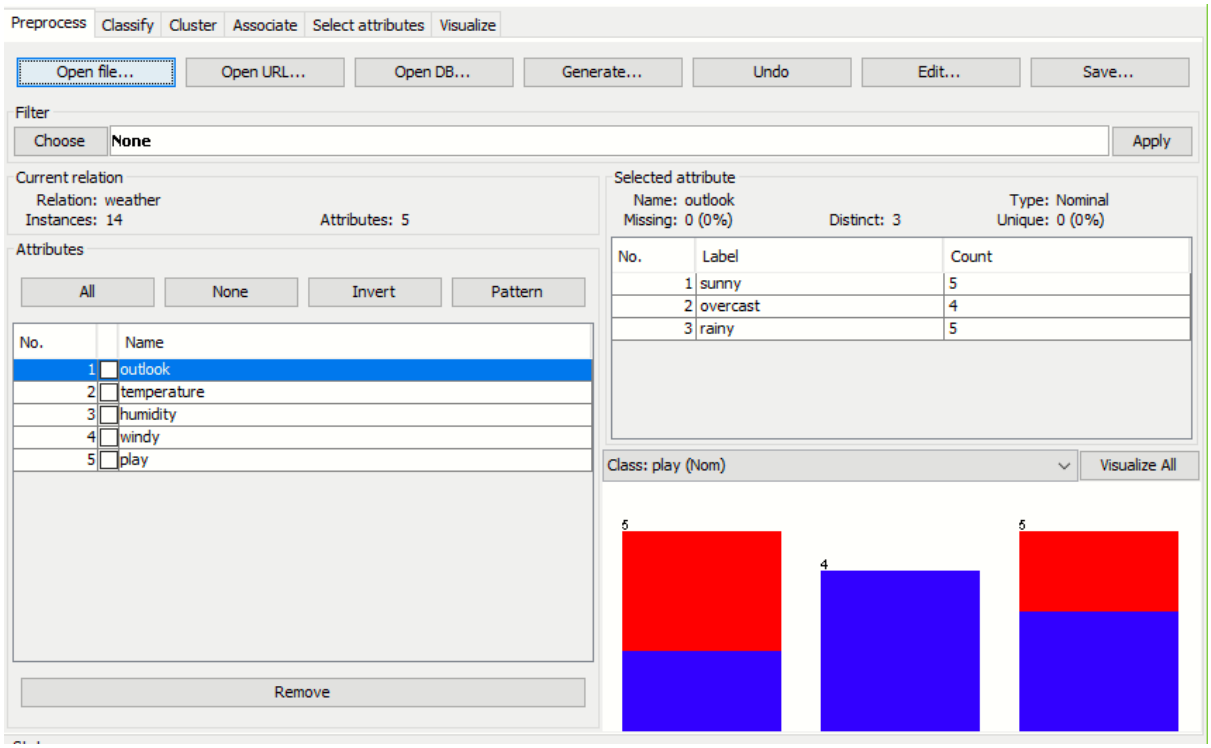
The weka.filters package contains Java classes that transform datasets -- by removing or adding attributes, resampling the dataset, removing examples and so on. This package offers useful support for data preprocessing, which is an important step in machine learning.

All filters offer the command-line option -*i* for specifying the input dataset, and the option -*o* for specifying the output dataset. If any of these parameters is not given, this specifies standard input resp. output for use within pipes. Other parameters are specific to each filter and can be found out via -*h*, as with any other class. The weka.filters package is organized into supervised and unsupervised filtering, both of which are again subdivided into instance and attribute filtering. We will discuss each of the four subsection separately.

### 2.3 Steps for **preprocessing filters:**

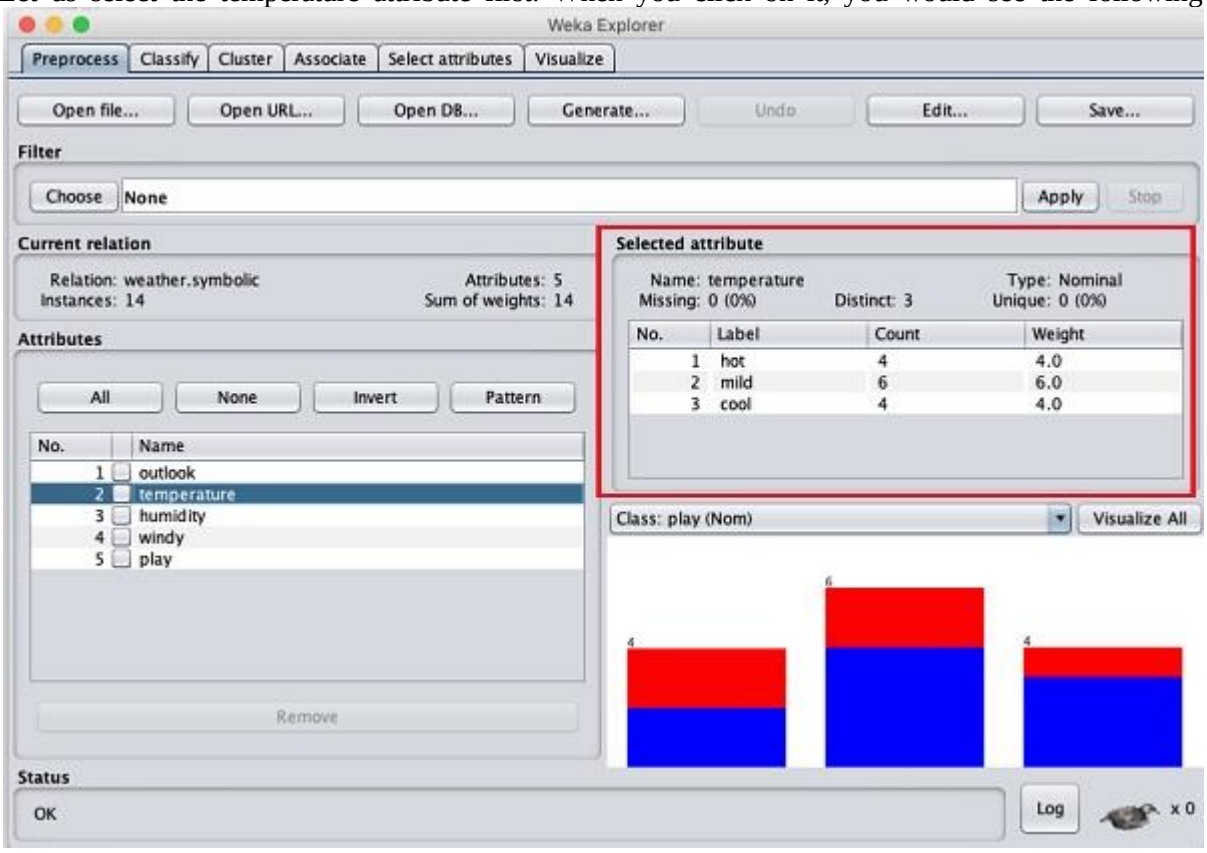
1. To demonstrate the available features in preprocessing, we will use the **Weather** database that is provided in the installation.

Using the **Open file** ... option and select the **weather-numeric.arff** file.



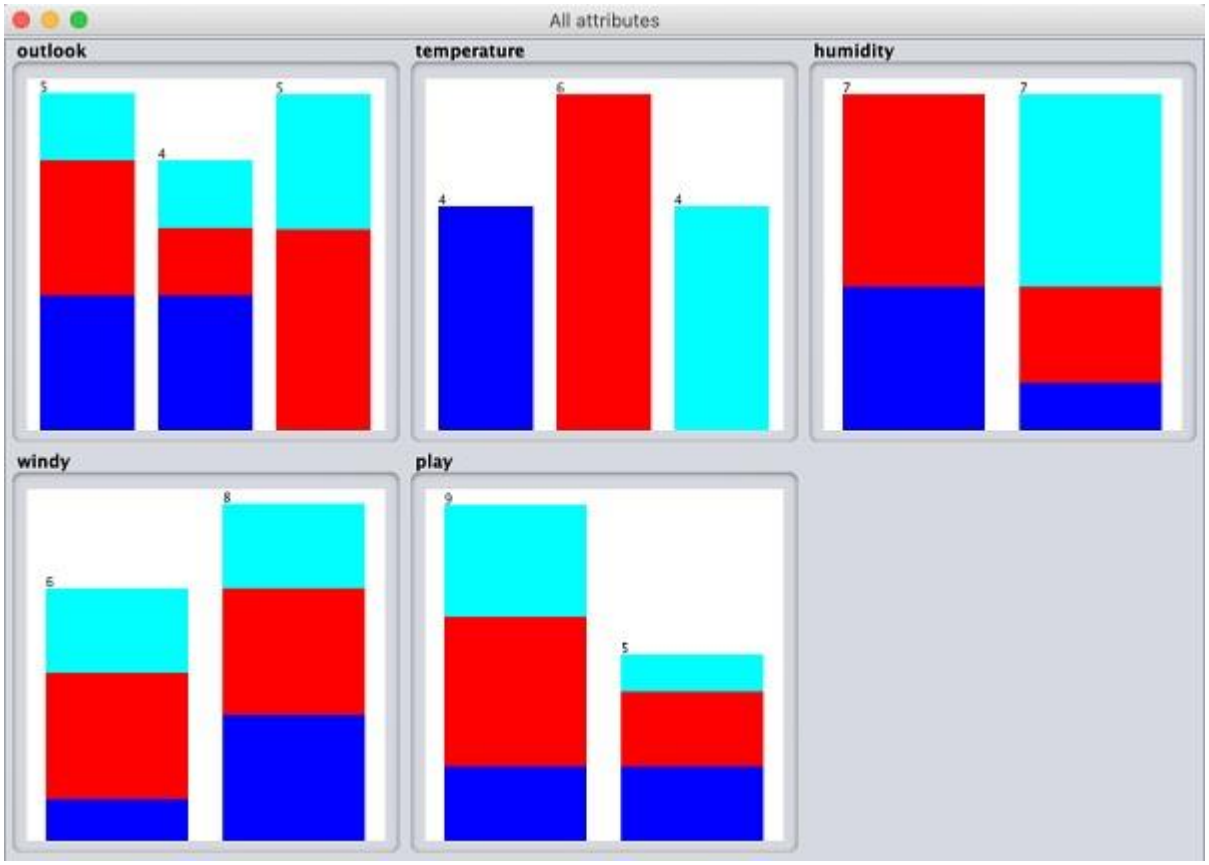

2. The **weather** database contains five fields - outlook, temperature, humidity, windy and play. When you select an attribute from this list by clicking on it, further details on the attribute itself are displayed on the right hand side.

Let us select the temperature attribute first. When you click on it, you would see the following screen –




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 –



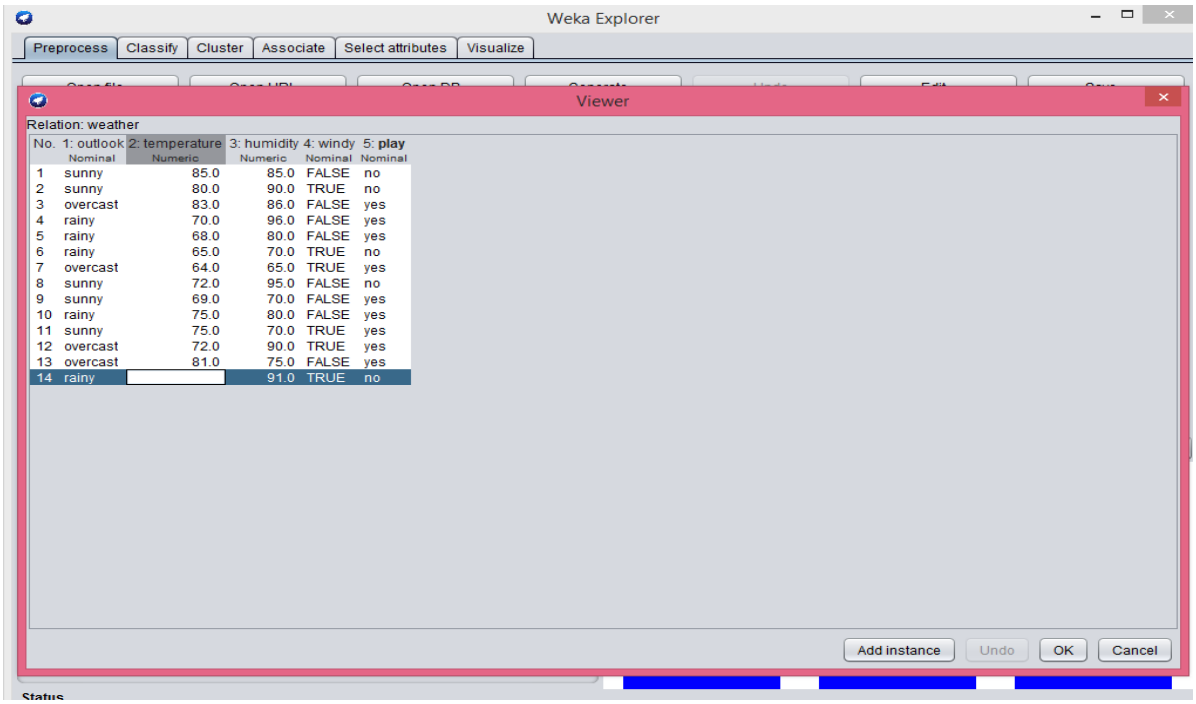

3.you can see the EDIT option on the screen.Click it and there we can add, remove and fill the missing values in the rows

Relation: weather

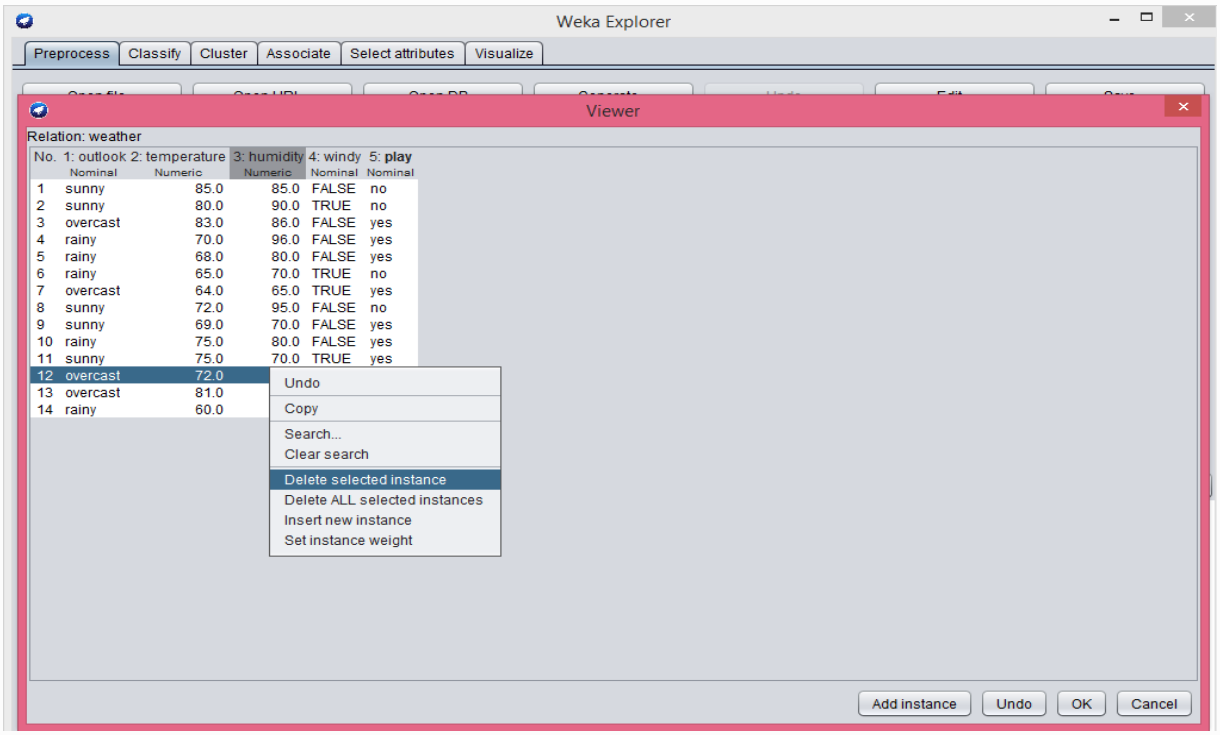
| No. | outlook Nominal | temperature Numeric | humidity Numeric | windy Nominal | play Nominal |
|---|---|---|---|---|---|
| 1 | sunny | 85.0 | 85.0 | FALSE | no |
| 2 | sunny | 80.0 | 90.0 | TRUE | no |
| 3 | overcast | 83.0 | 86.0 | FALSE | yes |
| 4 | rainy | 70.0 | 96.0 | FALSE | yes |
| 5 | rainy | 68.0 | 80.0 | FALSE | yes |
| 6 | rainy | 65.0 | 70.0 | TRUE | no |
| 7 | overcast | 64.0 | 65.0 | TRUE | yes |
| 8 | sunny | 72.0 | 95.0 | FALSE | no |
| 9 | sunny | 69.0 | 70.0 | FALSE | yes |
| 10 | rainy | 75.0 | 80.0 | FALSE | yes |
| 11 | sunny | 75.0 | 70.0 | TRUE | yes |
| 12 | overcast | 72.0 | 90.0 | TRUE | yes |
| 13 | overcast | 81.0 | 75.0 | FALSE | yes |
| 14 | rainy | 71.0 | 91.0 | TRUE | no |

Undo OK Cancel

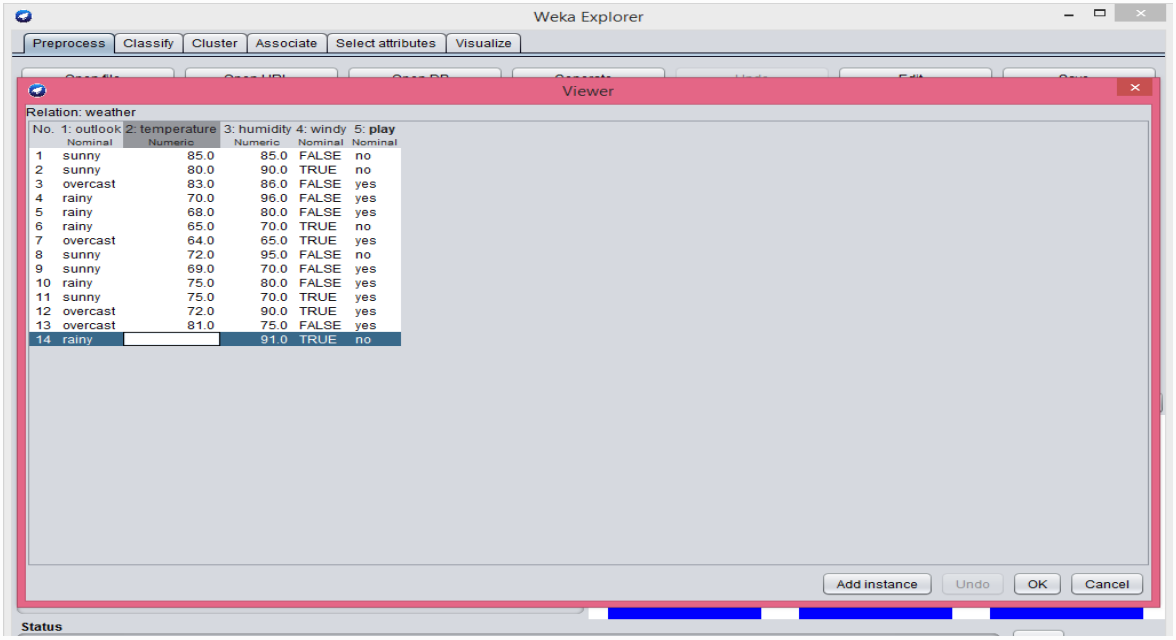
4. Now You can perform ADD preprocessing on the 'Weather' dataset, at the bottom right side we have add instance by clicking on that button automatically new instances are added.

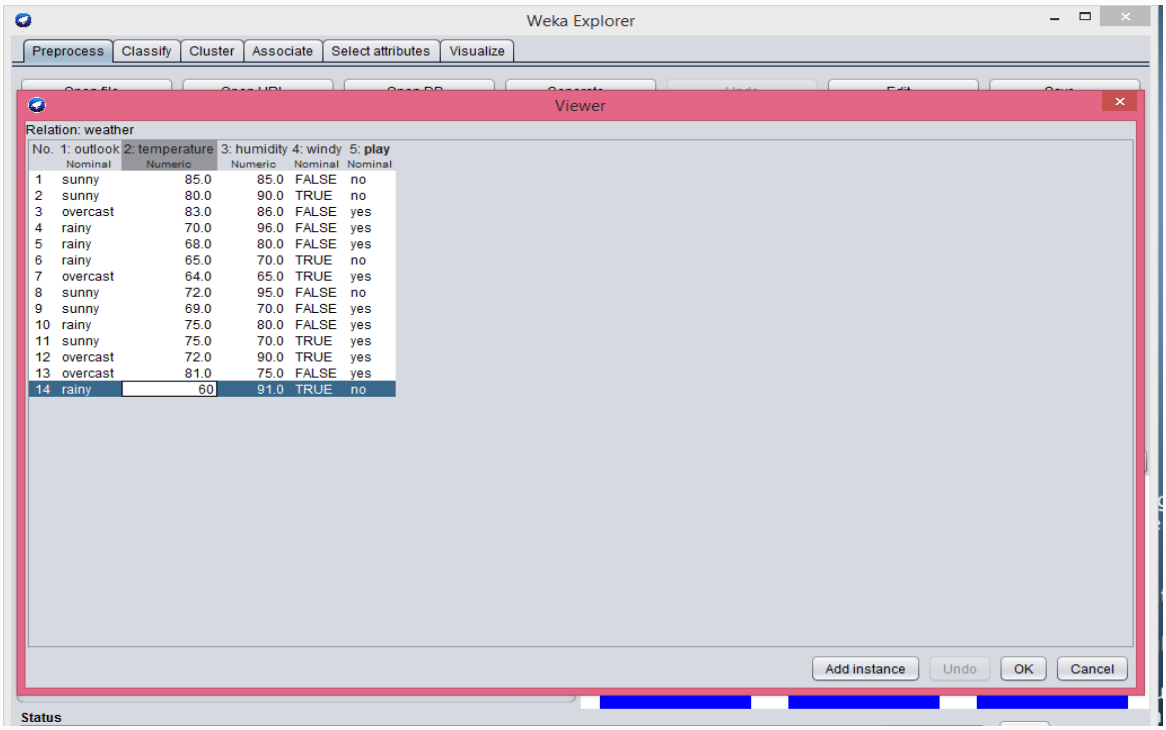



5. Now to perform remove preprocessing select the instances which you have to remove, right click on it than we will get an option like Delete selected instances. By clicking on that the instances which we are selected will be deleted.



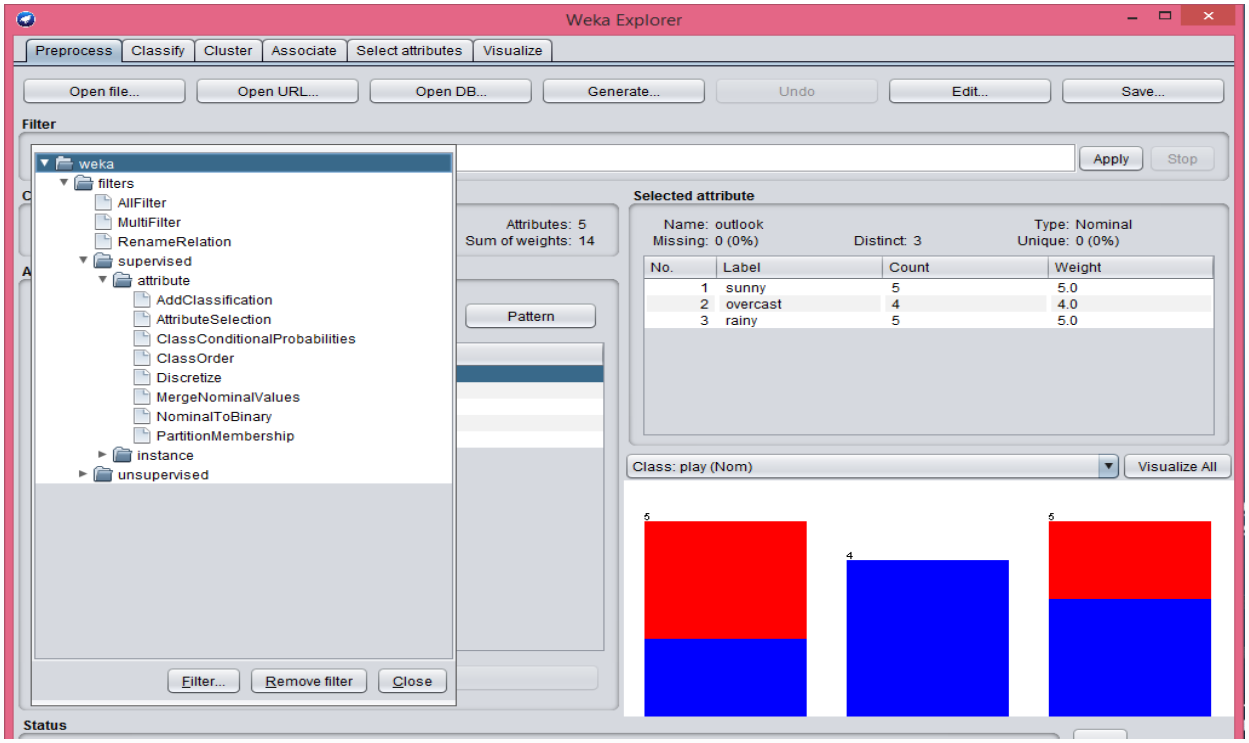
6. Perform the preprocessing filter of replacing missing values.



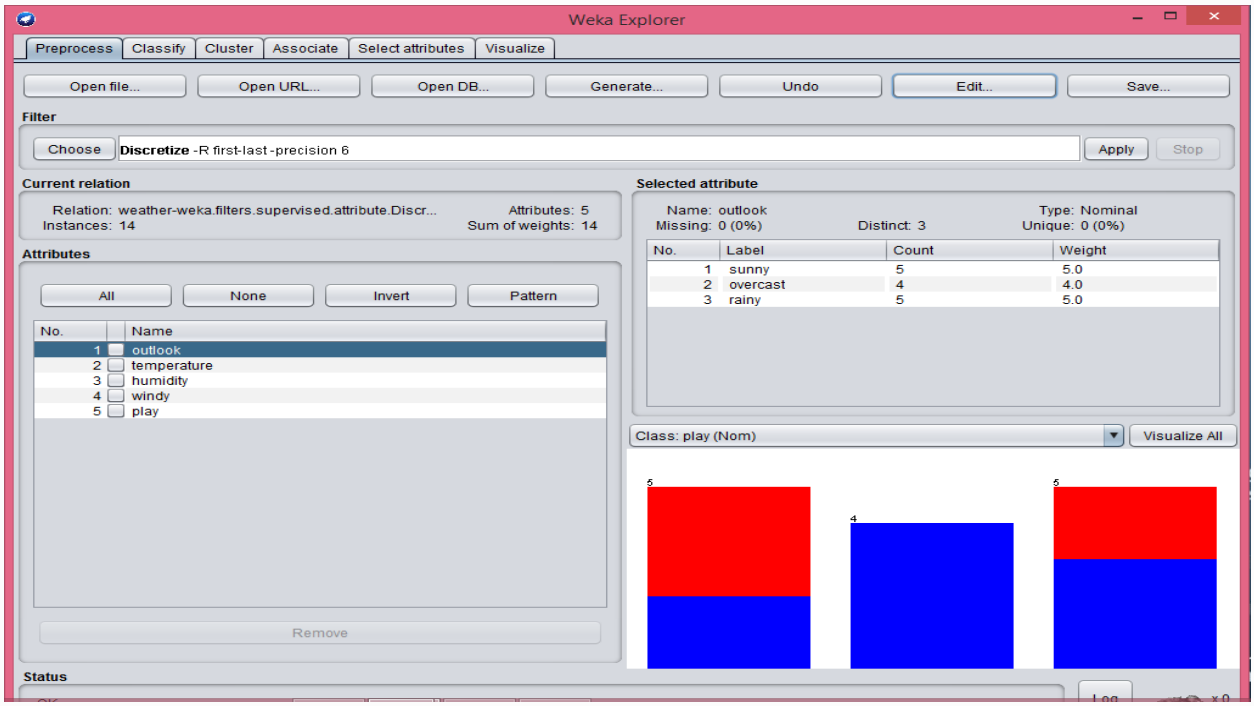




7. Now to perform Discretize preprocessing filter, go to choose option now select the filter option under filter we have different modes of filters to perform preprocessing.

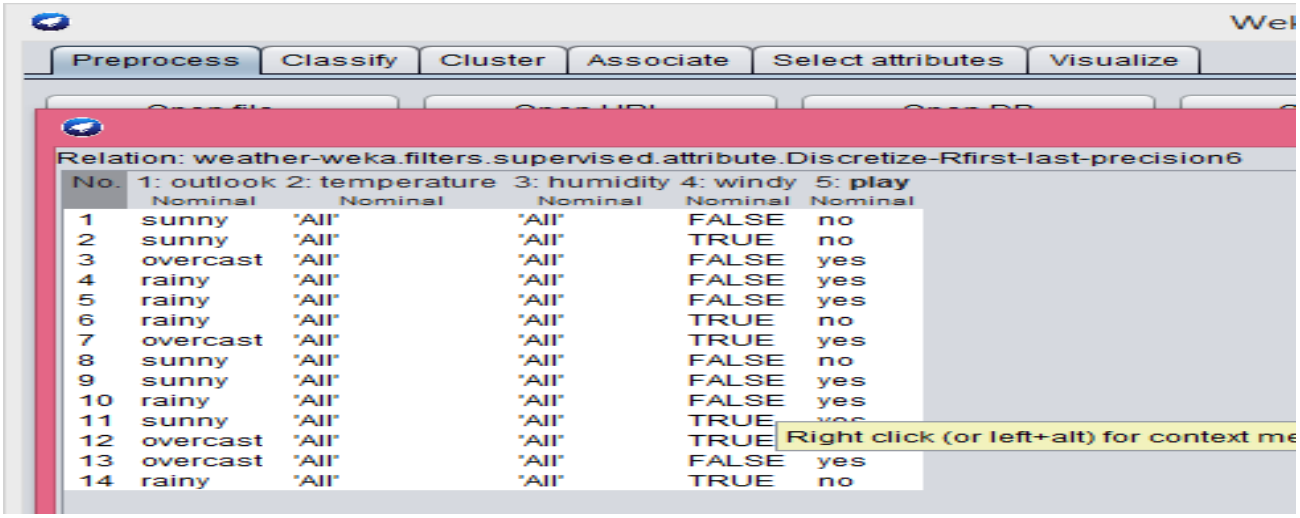
Select the attribute under supervised there we have a Discretize option choose that.



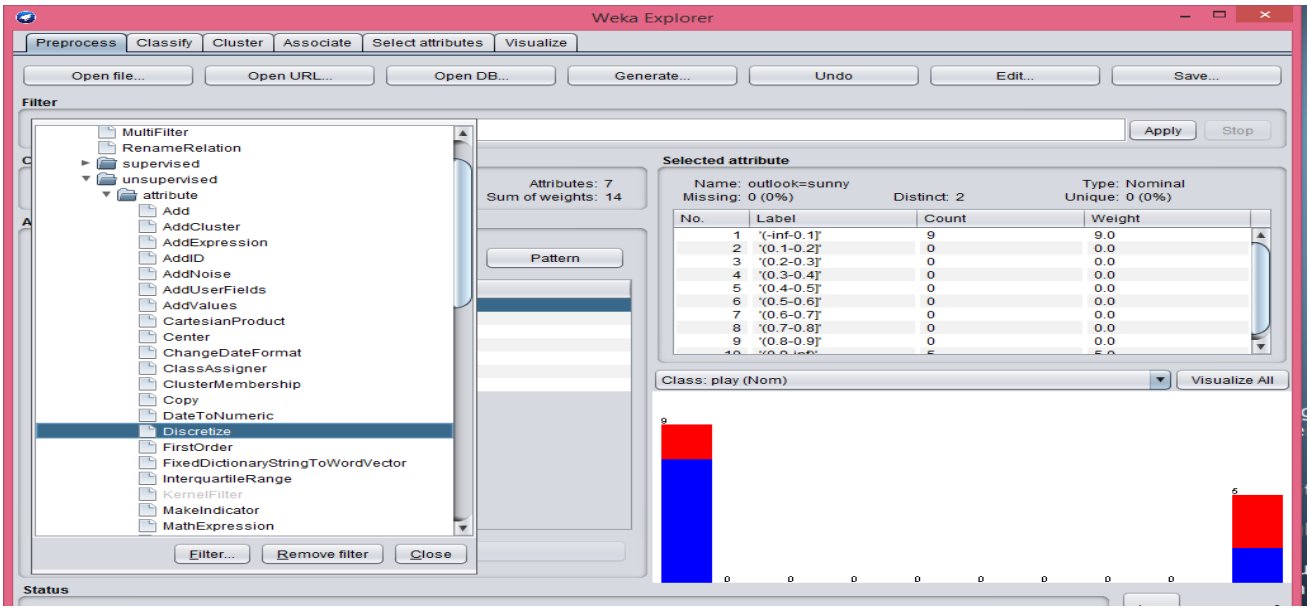

8.After choosing Discretize option click on Apply option to apply the filter to the dataset.

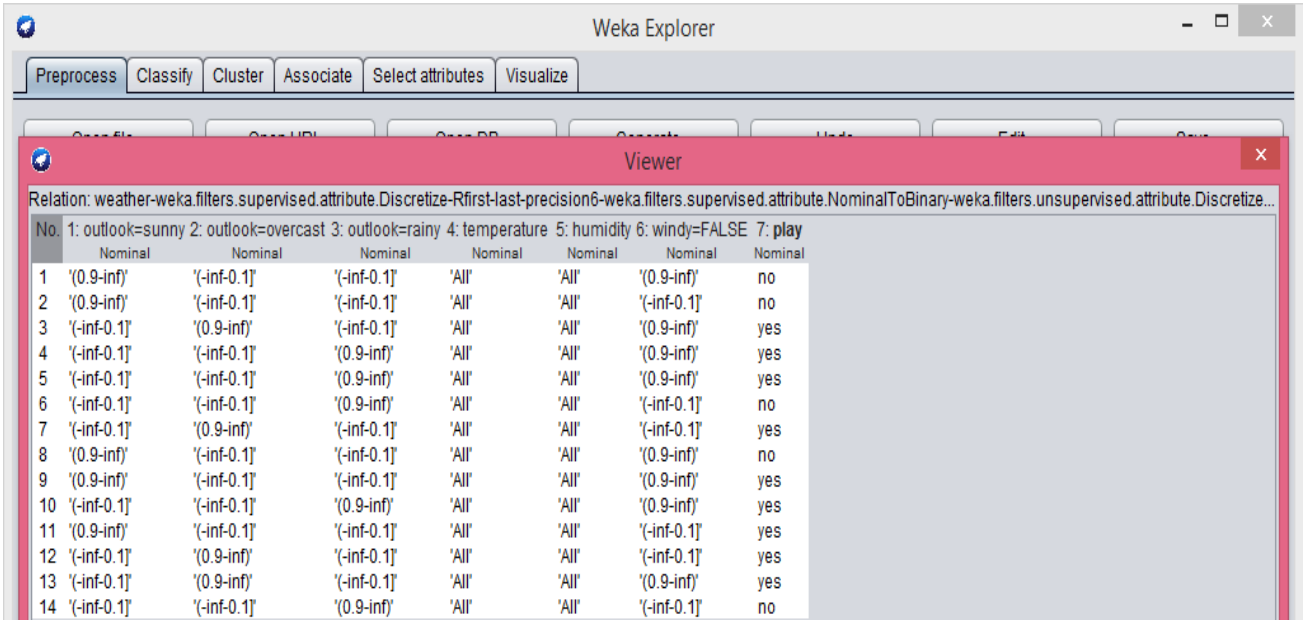



9. Now to observe the dataset in Discretization click on Edit option.

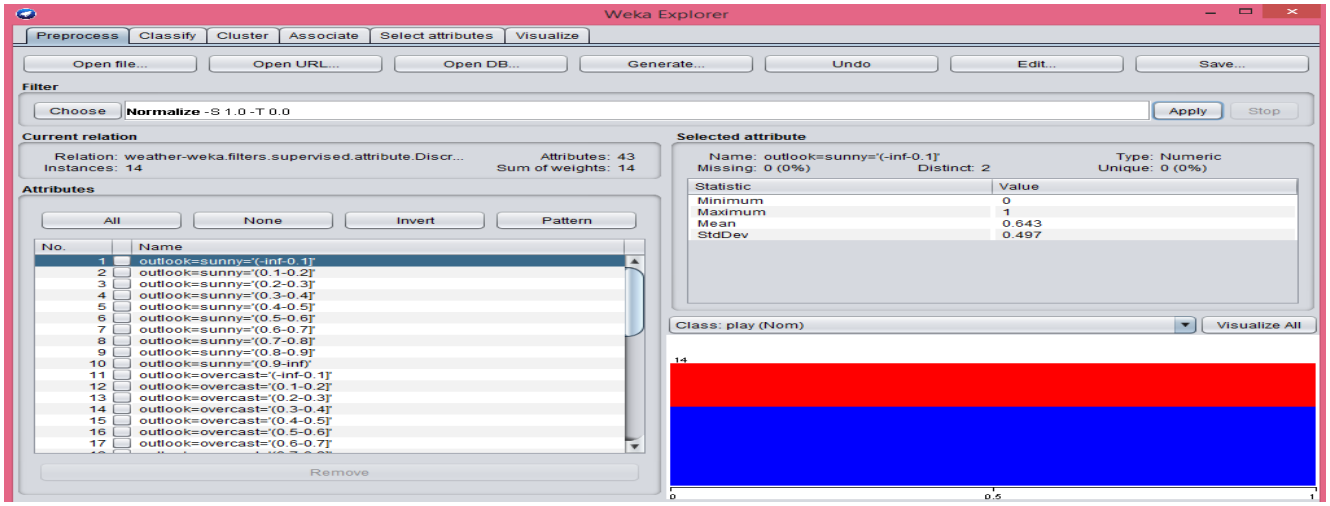



10. Now to perform Discretization under unsupervised mode select Discretize option under unsupervised filter.

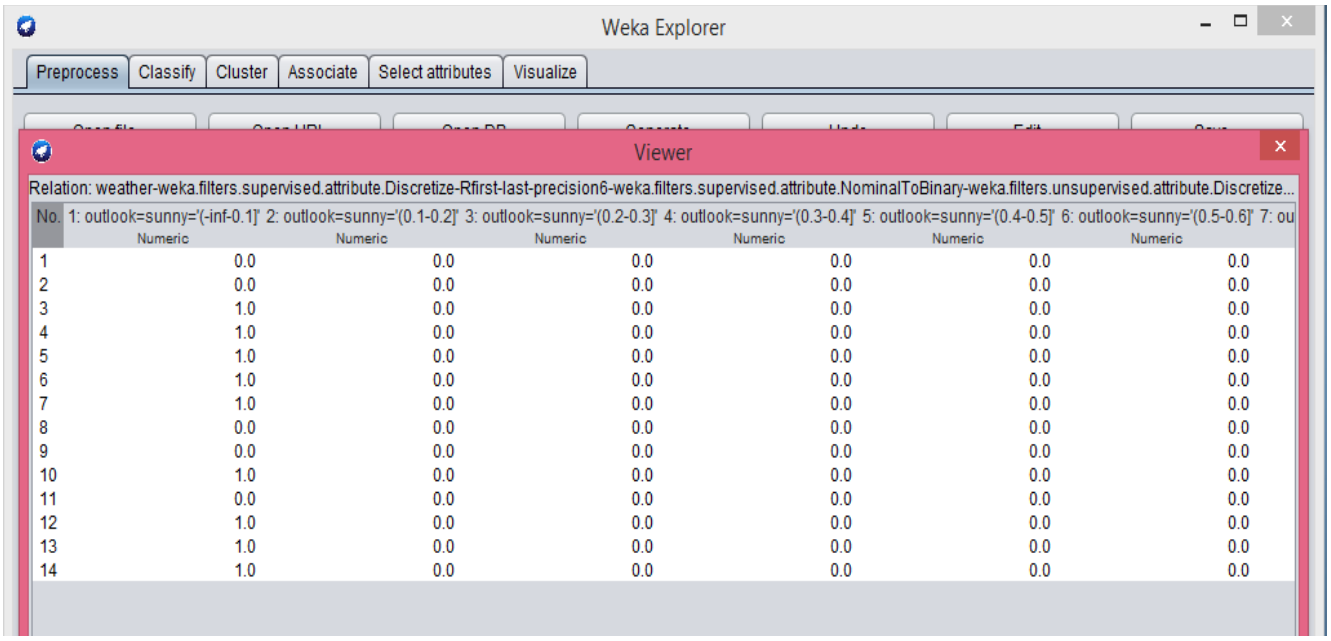


11.Now perform Normalization by selecting Normalize filter.




Apply the Normalize to dataset and observe the each and every instances.




## 2.4 Results and Discussion:

Therefore to perform the following preprocessing filters on 'Weather' dataset. Add, Remove, Discretize, Replace Missing values, Normalize, is executed successfully in the Weka software by observing and understanding each preprocessing filter in different modes.