

EXERCISE -1

1.1 Problem Statement:

To Create a dataset using ARFF and CSV formats and load into the Weka Explorer.

1.2 Description:

About Dataset used

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], ...}

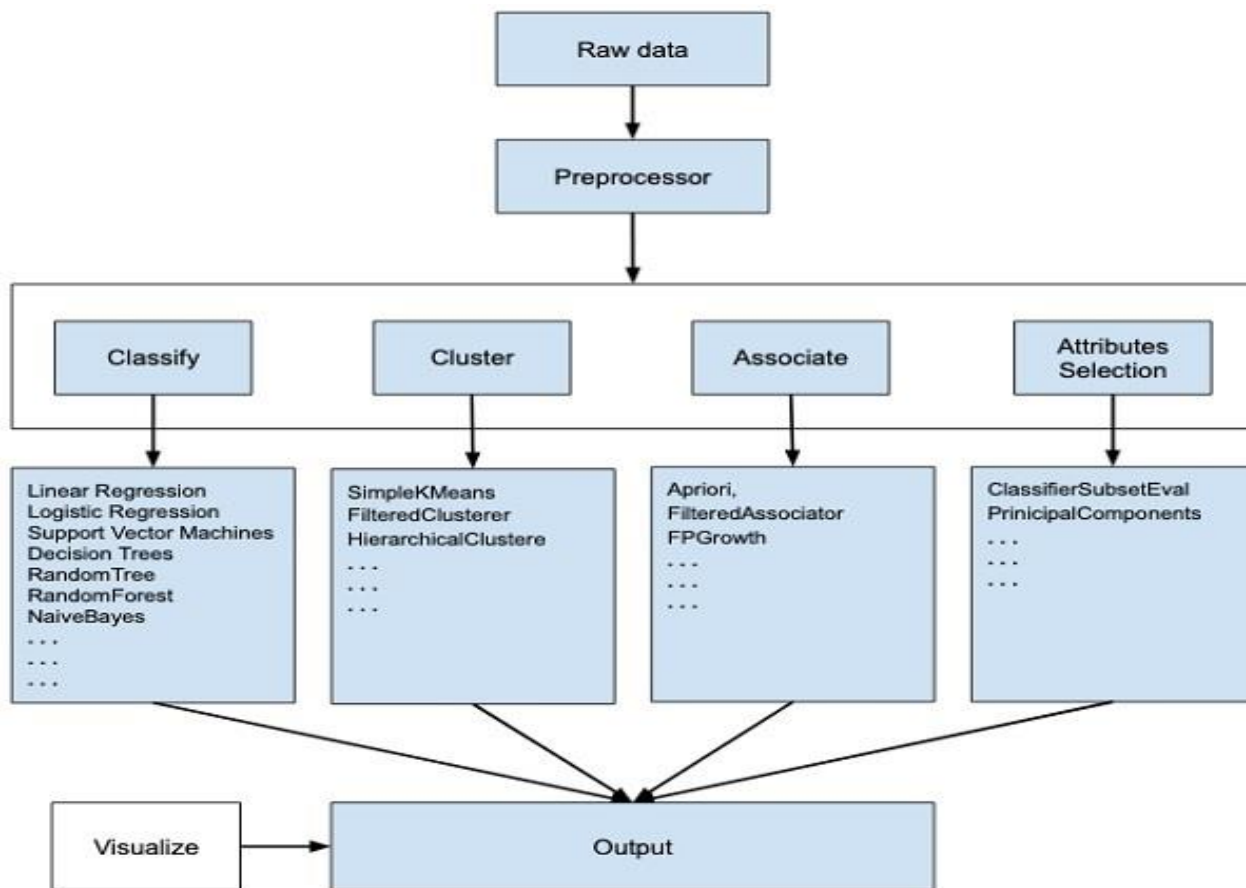
For example, the class value of the Iris dataset can be defined as follows:

```
@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.

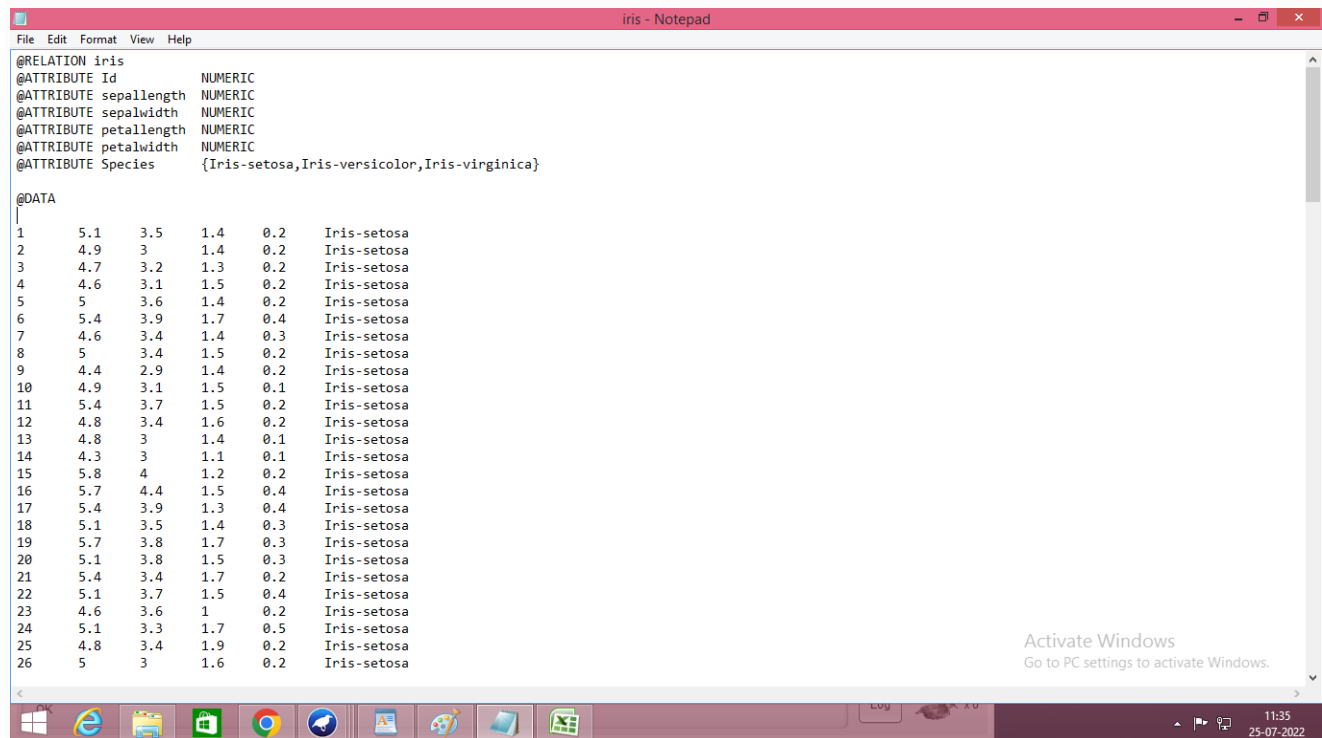
The various models can be applied on the same dataset. You can then compare the outputs of different models and select the best that meets your purpose.

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

1.3 Sample Arff and CSV files of the Dataset:

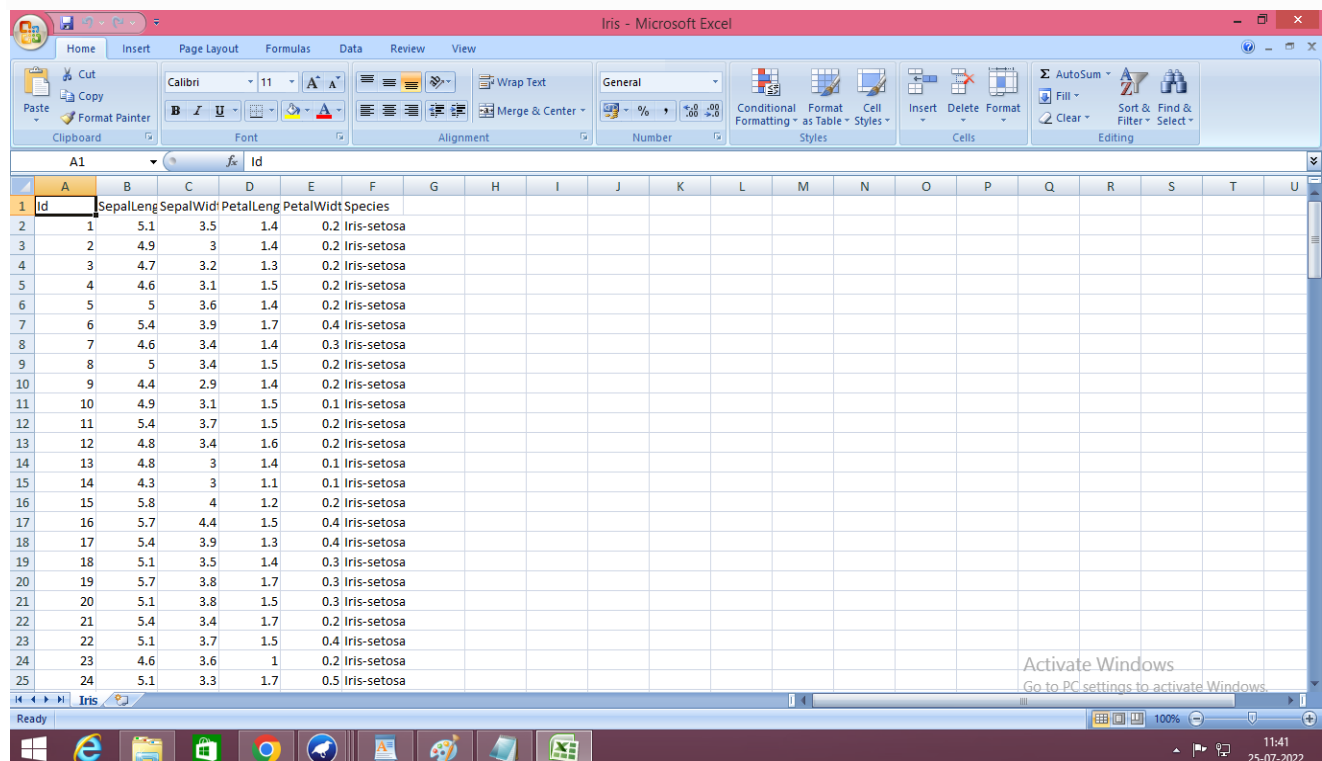
Arff file for Iris Dataset:



```
@RELATION iris
@ATTRIBUTE Id NUMERIC
@ATTRIBUTE SepalLength NUMERIC
@ATTRIBUTE SepalWidth NUMERIC
@ATTRIBUTE PetalLength NUMERIC
@ATTRIBUTE PetalWidth NUMERIC
@ATTRIBUTE Species {Iris-setosa,Iris-versicolor,Iris-virginica}

@DATA
1 5.1 3.5 1.4 0.2 Iris-setosa
2 4.9 3 1.4 0.2 Iris-setosa
3 4.7 3.2 1.3 0.2 Iris-setosa
4 4.6 3.1 1.5 0.2 Iris-setosa
5 5 3.6 1.4 0.2 Iris-setosa
6 5.4 3.9 1.7 0.4 Iris-setosa
7 4.6 3.4 1.4 0.3 Iris-setosa
8 5 3.4 1.5 0.2 Iris-setosa
9 4.4 2.9 1.4 0.2 Iris-setosa
10 4.9 3.1 1.5 0.1 Iris-setosa
11 5.4 3.7 1.5 0.2 Iris-setosa
12 4.8 3.4 1.6 0.2 Iris-setosa
13 4.8 3 1.4 0.1 Iris-setosa
14 4.3 3 1.1 0.1 Iris-setosa
15 5.8 4 1.2 0.2 Iris-setosa
16 5.7 4.4 1.5 0.4 Iris-setosa
17 5.4 3.9 1.3 0.4 Iris-setosa
18 5.1 3.5 1.4 0.3 Iris-setosa
19 5.7 3.8 1.7 0.3 Iris-setosa
20 5.1 3.8 1.5 0.3 Iris-setosa
21 5.4 3.4 1.7 0.2 Iris-setosa
22 5.1 3.7 1.5 0.4 Iris-setosa
23 4.6 3.6 1 0.2 Iris-setosa
24 5.1 3.3 1.7 0.5 Iris-setosa
25 4.8 3.4 1.9 0.2 Iris-setosa
26 5 3 1.6 0.2 Iris-setosa
```

CSV file for Iris Dataset:



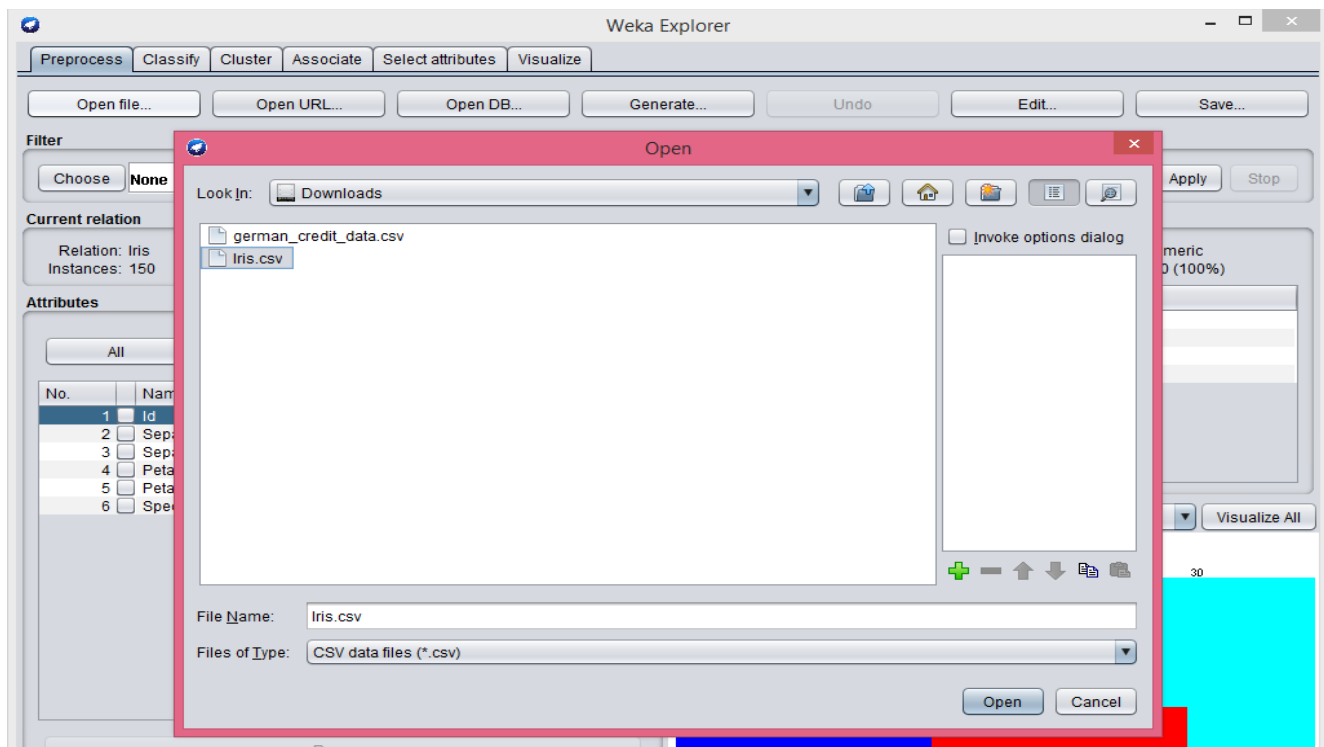
Id	SepalLeng	SepalWidt	PetalLeng	PetalWidt	Species
1	5.1	3.5	1.4	0.2	Iris-setosa
2	4.9	3	1.4	0.2	Iris-setosa
3	4.7	3.2	1.3	0.2	Iris-setosa
4	4.6	3.1	1.5	0.2	Iris-setosa
5	5	3.6	1.4	0.2	Iris-setosa
6	5.4	3.9	1.7	0.4	Iris-setosa
7	4.6	3.4	1.4	0.3	Iris-setosa
8	5	3.4	1.5	0.2	Iris-setosa
9	4.4	2.9	1.4	0.2	Iris-setosa
10	4.9	3.1	1.5	0.1	Iris-setosa
11	5.4	3.7	1.5	0.2	Iris-setosa
12	4.8	3.4	1.6	0.2	Iris-setosa
13	4.8	3	1.4	0.1	Iris-setosa
14	4.3	3	1.1	0.1	Iris-setosa
15	5.8	4	1.2	0.2	Iris-setosa
16	5.7	4.4	1.5	0.4	Iris-setosa
17	5.4	3.9	1.3	0.4	Iris-setosa
18	5.1	3.5	1.4	0.3	Iris-setosa
19	5.7	3.8	1.7	0.3	Iris-setosa
20	5.1	3.8	1.5	0.3	Iris-setosa
21	5.4	3.4	1.7	0.2	Iris-setosa
22	5.1	3.7	1.5	0.4	Iris-setosa
23	4.6	3.6	1	0.2	Iris-setosa
24	5.1	3.3	1.7	0.5	Iris-setosa
25	4.8	3.4	1.9	0.2	Iris-setosa
26	5	3	1.6	0.2	Iris-setosa

1.4 Steps to upload dataset:

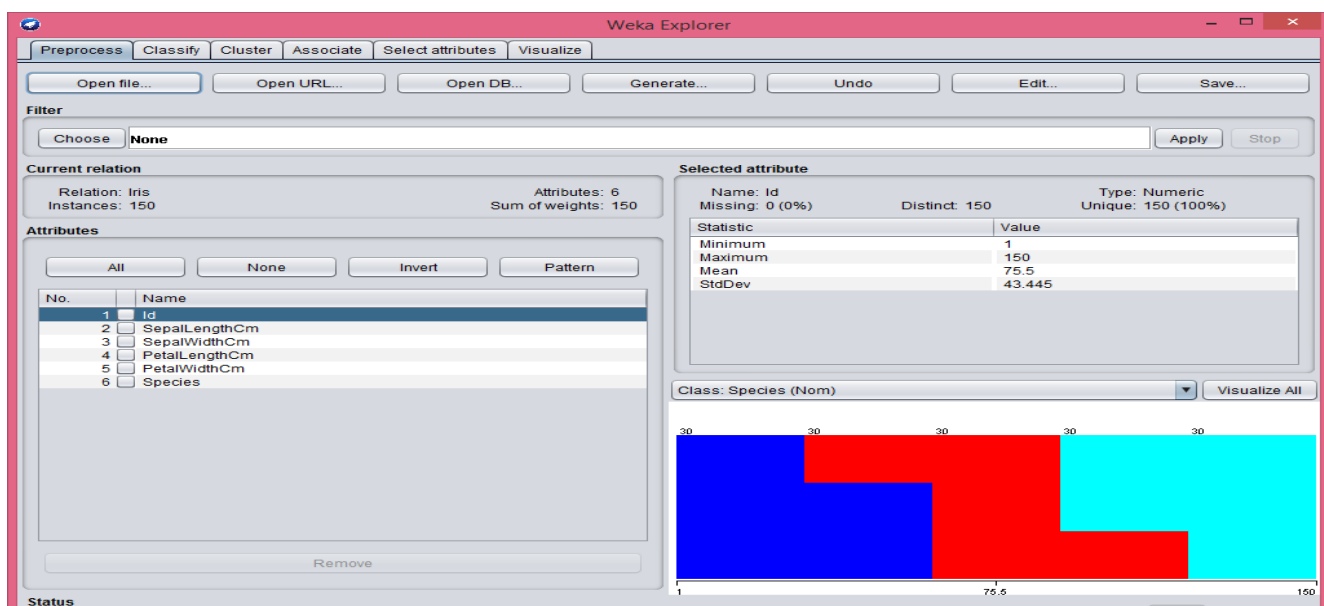
1.4.1 To upload CSV Dataset:

1. Open weka software and click on Explorer, then it will open the weka explorer in that choose open file option. There you can select the file which is in that system.

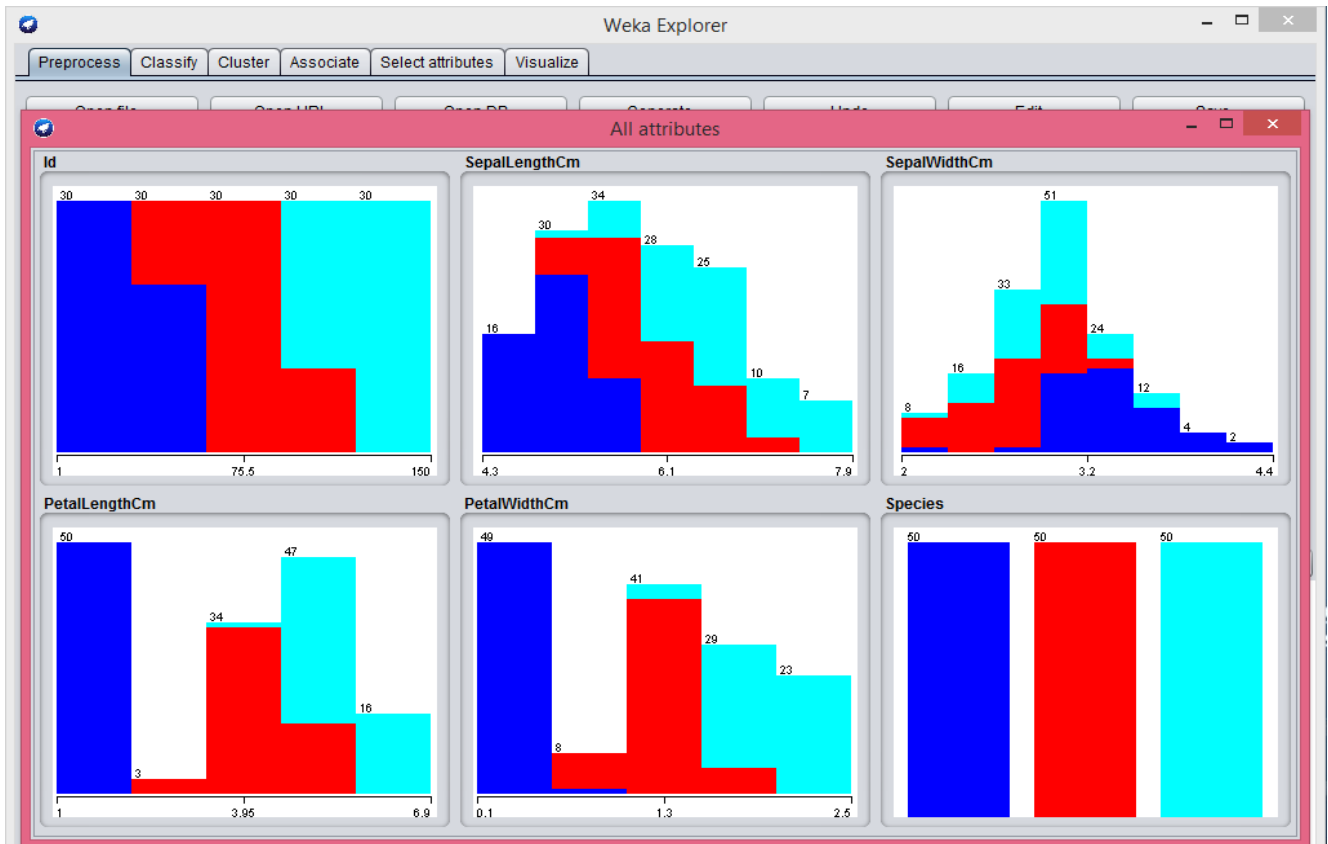
Now I am selecting Iris CSV file to upload the file in weka explorer



2. After the iris CSV file is uploaded in weka the it will display like below.



3. we can visualize all the attributes in the Iris dataset, by clicking on Visualize All option.



1.4.2 To upload Arff Dataset:

1. To upload Arff file in Weka Explorer first we have to convert CSV file of the dataset into Arff file.
2. To convert CSV file into Arff we have use Arff format which is displayed like below.

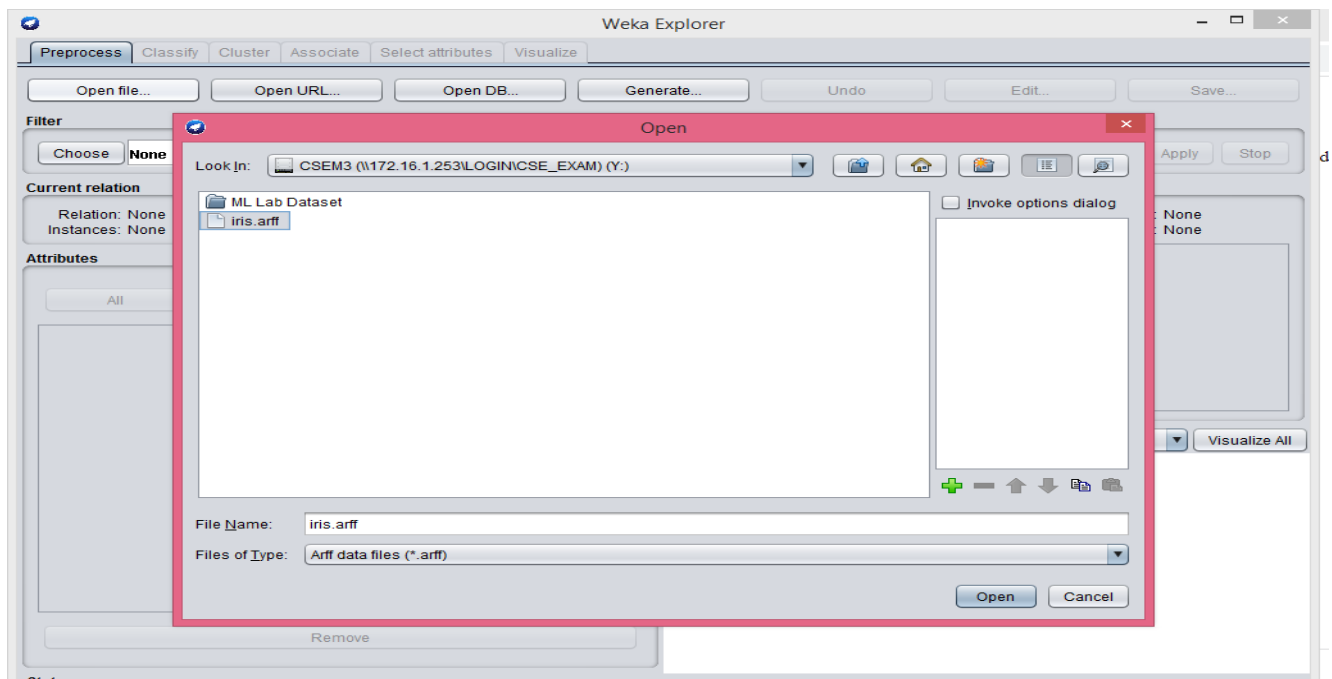
```

@RELATION iris
@ATTRIBUTE Id NUMERIC
@ATTRIBUTE sepallength NUMERIC
@ATTRIBUTE sepalwidth NUMERIC
@ATTRIBUTE petallength NUMERIC
@ATTRIBUTE petalwidth NUMERIC
@ATTRIBUTE Species {Iris-setosa,Iris-versicolor,Iris-virginica}

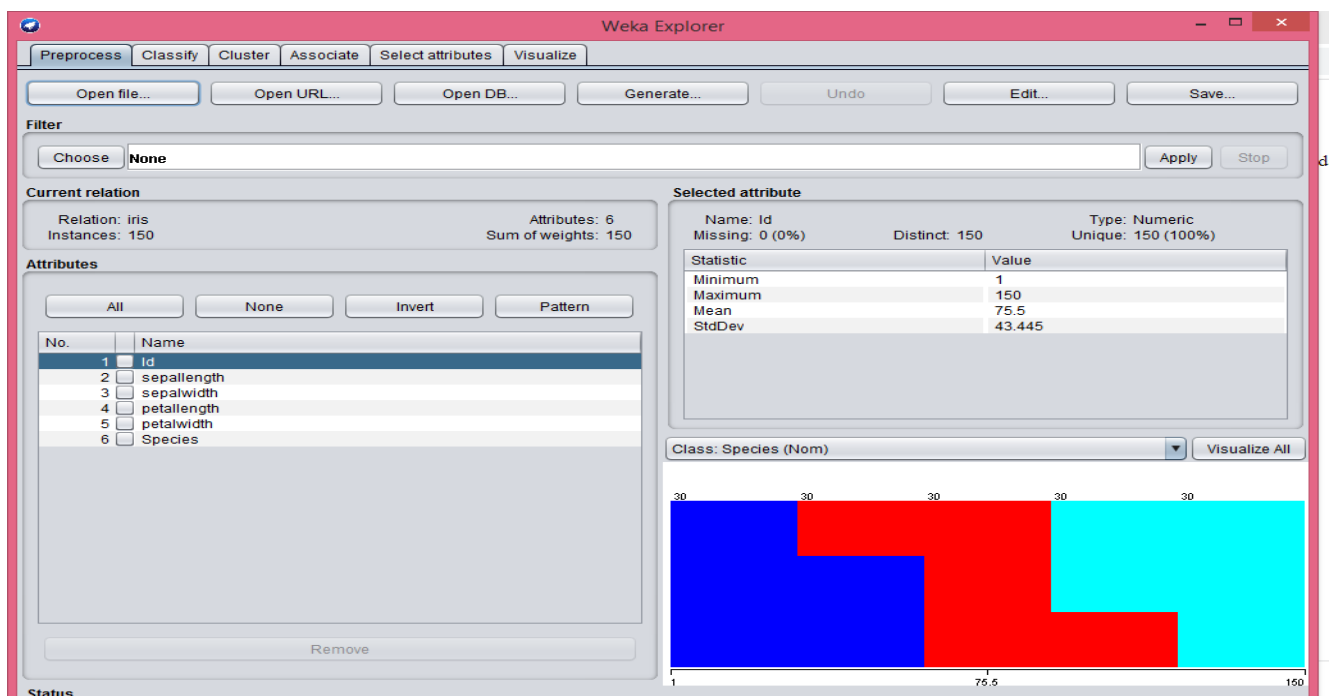
@DATA
1 5.1 3.5 1.4 0.2 Iris-setosa
2 4.9 3 1.4 0.2 Iris-setosa
3 4.7 3.2 1.3 0.2 Iris-setosa
4 4.6 3.1 1.5 0.2 Iris-setosa
5 5 3.6 1.4 0.2 Iris-setosa
6 5.4 3.9 1.7 0.4 Iris-setosa
7 4.6 3.4 1.4 0.3 Iris-setosa
8 5 3.4 1.5 0.2 Iris-setosa
9 4.4 2.9 1.4 0.2 Iris-setosa
10 4.9 3.1 1.5 0.1 Iris-setosa
11 5.4 3.7 1.5 0.2 Iris-setosa
12 4.8 3.4 1.6 0.2 Iris-setosa
13 4.8 3 1.4 0.1 Iris-setosa
14 4.3 3 1.1 0.1 Iris-setosa
15 5.8 4 1.2 0.2 Iris-setosa
16 5.7 4.4 1.5 0.4 Iris-setosa
17 5.4 3.9 1.3 0.4 Iris-setosa
18 5.1 3.5 1.4 0.3 Iris-setosa
19 5.7 3.8 1.7 0.3 Iris-setosa
20 5.1 3.8 1.5 0.3 Iris-setosa
21 5.4 3.4 1.7 0.2 Iris-setosa
22 5.1 3.7 1.5 0.4 Iris-setosa
23 4.6 3.6 1 0.2 Iris-setosa
24 5.1 3.3 1.7 0.5 Iris-setosa
25 4.8 3.4 1.9 0.2 Iris-setosa
26 5 3 1.6 0.2 Iris-setosa

```

- Now we have to save the above file by using FILE NAME.Arff form, then only the file is saved as Arff file.
- Now we have to upload the this Arff file by using same steps like uploading csv file.
- Open the Weka Explorer and click on open file option, select the Arff file which we saved already.



- Now you can open and upload the Arff file into weka Explorer.



1.5 Results and Discussion:

To Create a dataset using ARFF and CSV formats and load into the Weka Explorer is executed Successfully by using CSV file format of dataset and converting that file into Arff file with the help of Arff Data format.

