Practical 1

Aim: To build Data Warehouse and Explore WEKA.

Theory:

Data Warehouse : Data warehousing is the process of constructing and using a data warehouse. A data warehouse is constructed by integrating data from multiple heterogeneous sources that support analytical reporting,structured and/or ad hoc queries, and decision making. Data warehousing involves data cleaning, data integration, and data consolidations.

Data Warehouse Characteristics -

- Subject-Oriented
- Integrated
- Time-variant
- Non-volatile

Need of Data Warehouse:

1. Business User: Business users require a data warehouse to view summarized data from the past. Since these people are non-technical, the data may be presentedto them in an elementary form.
2. Store historical data: Data Warehouse is required to store the time variable datafrom the past. Thisinput is made to be used for various purposes.
3. Make strategic decisions: Some strategies may be depending upon the data in thedata warehouse. So,data warehouse contributes to making strategic decisions.
4. For data consistency and quality: Bringing the data from different sources at acommonplace, the user can effectively undertake to bring the uniformity and consistency in data.
5. High response time: Data warehouse has to be ready for somewhat unexpectedloads and types of queries, which demands a significant degree of flexibility and quick response time.

Benefits of Data Warehouse:

1. Understand business trends and make better forecasting decisions.
2. Data Warehouses are designed to perform well enormous amounts of data.
3. The structure of data warehouses is more accessible for end-users to navigate,understand, andquery.
4. Queries that would be complex in many normalized databases could be easier tobuild and maintainin data warehouses.

5. Data warehousing is an efficient method to manage demand for lots of informationfrom lots of users.
6. Data warehousing provides the capabilities to analyze a large amount of historicaldata.

Creating an Employee Table with the help of Data Mining Tool WEKA

Procedure:

Steps:

1) Open Start Programs Accessories Notepad
2) Type the following training data set with the help of Notepad for Employee Table.

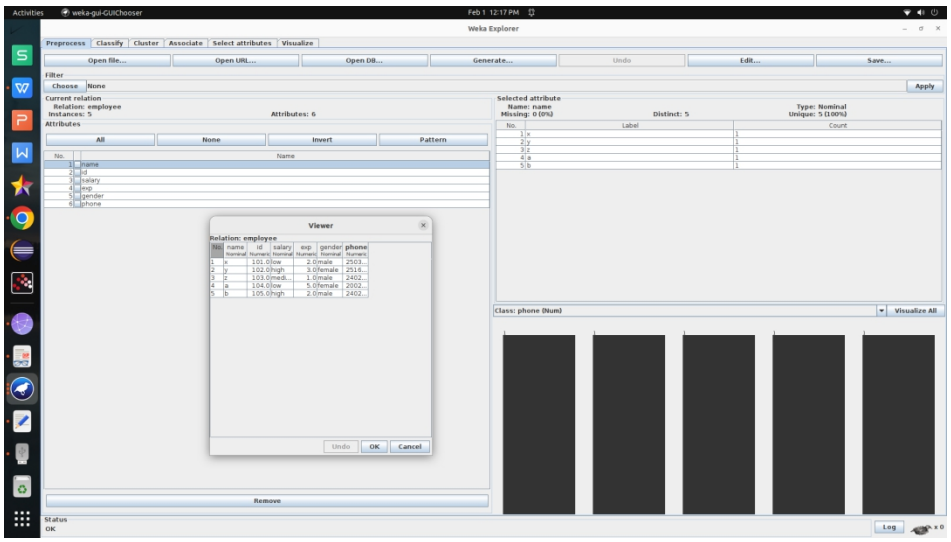
```
@relation
employee
@attribute
name
{x,y,z,a,b}
@attribute
id numeric
@attribute salary
{low,medium,high}
@attribute exp
numeric
@attribute
gender
{male,female}
@attribute
phone numeric

@data
x,101,low,2,
male,250311
y,102,high,3,
female,2516
65
z,103,mediu
m,1,male,24
```

0238
a,104,low,5,f
emale,20020
0
b,105,high,2,
male,240240

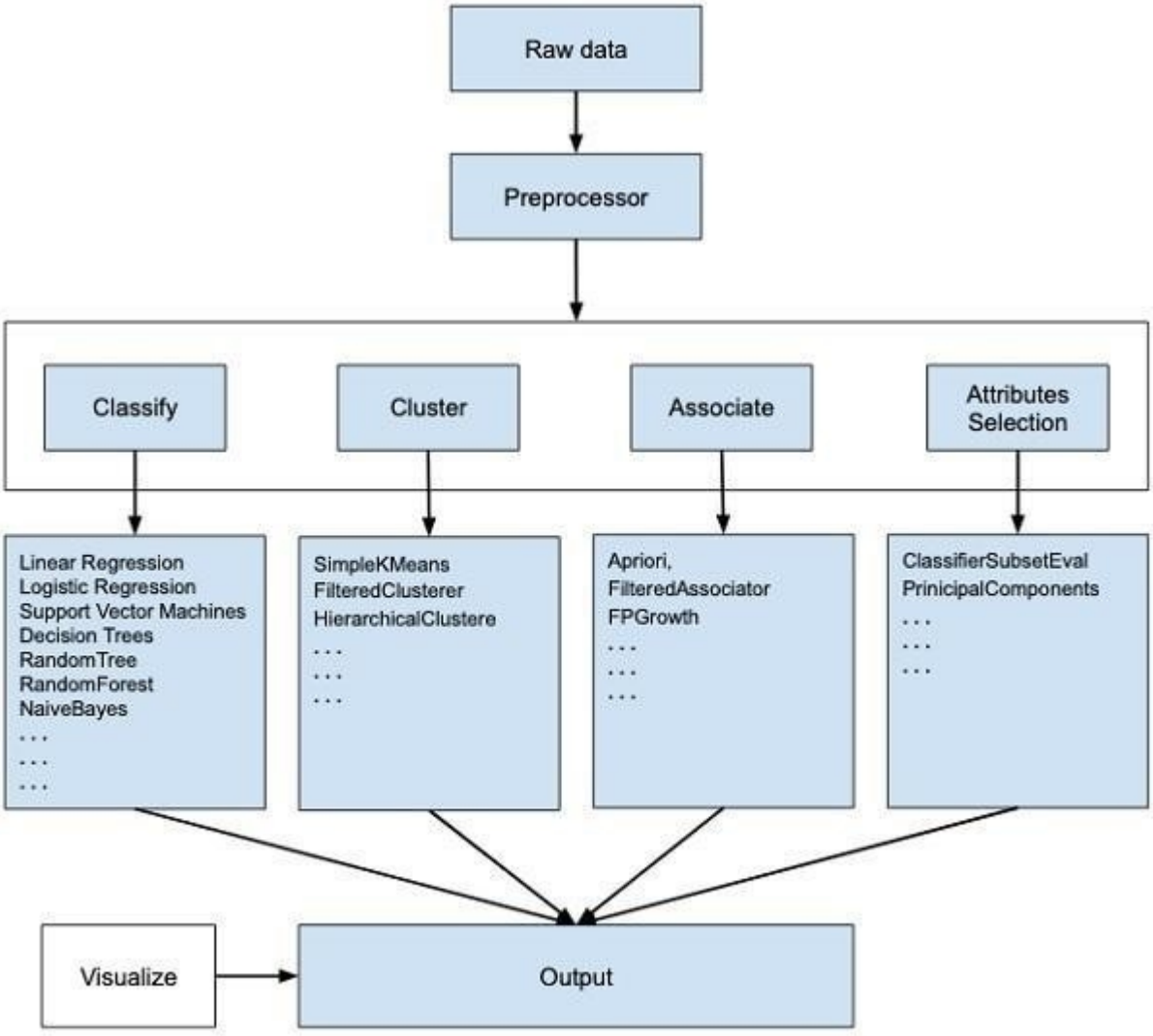
3) After that the file is saved with **.arff** file format.
4) Minimize the arff file and then open Start Programs weka-3-4.
5) Click on **weka-3-4**, then Weka dialog box is displayed on the screen.
6) In that dialog box there are four modes, click on **explorer**.
7) Explorer shows many options. In that click on **'open file'** and select the arff file
8) Click on **edit button** which shows employee table on weka.

Training Data Set Employee Table



Exploring Weka Tool:

WEKA 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 applythem 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 manystages indealing 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 andirrelevant 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.

WEKA provides the implementation of several algorithms. You wouldselect an algorithm of your choice, setthe desired parameters and run it on the dataset.

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

The various models can be applied on the same dataset. You can then compare the outputs ofdifferentmodels 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 toinstallWEKA on your local computer.

Weka is data mining software that uses a collection of machine learning algorithms. Thesealgorithms can beapplied directly to the data or called from the Java code.
Weka is a collection of tools for:

- Regression
- Clustering
- Association
- Data pre-processing
- Classification
- Visualisation Weka application interface

There are totally five application interfaces available for Weka. When we open Weka, it will start the *Weka GUIChooser* screen from where we can open the Weka application interface Explorer Preprocessing, attribute selection, learning, visualization Experimenter Testing and evaluating machine learning algorithms KnowledgeFlow Visual design of KDD processSimple Command Line A simple interface for typing commands.

Weka data formats

Weka uses the Attribute Relation File Format for data analysis, by default. But listed beloware some formatsthat Weka supports, from where data can be imported:

- arff
- arff.gz
- bsi

- csv
- dat
- data
- json
- json.gz
- libsvm
- m
- names
- xrff
- xrff.gz

Arff Format

An Arff file contains two sections - header and data.

- The header describes the attribute types.
- The data section contains a comma separated list of data.

An ARFF file requires the declaration of the relation, attribute and data. Figure 3 is anexample of an ARFFfile.

· *@relation:* This is the first line in any ARFF file, written in the header section, followed bythe relation/dataset name. The relation name must be a string and if it contains spaces, then it should be enclosed between quotes.

· *@attribute:* These are declared with their names and the type or range in the headersection. Wekasupports the following data types for attributes:

- Numeric
- <nominal-specification>
- String
- date
- @data – Defined in the Data section followed by the list of all data segments

Conclusion: Thus the training data table is reated and Weka tool is explored.

Viva Questions

1. Define Data Warehouse.
2. What are the characteristics of Data Warehousing?
3. What is Weka Tool and what are the significance of Weka ?