# Practical No.1

Aim: To build Data warehouse and Explore WEKA.

Date :

# Practical No. 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 :→**

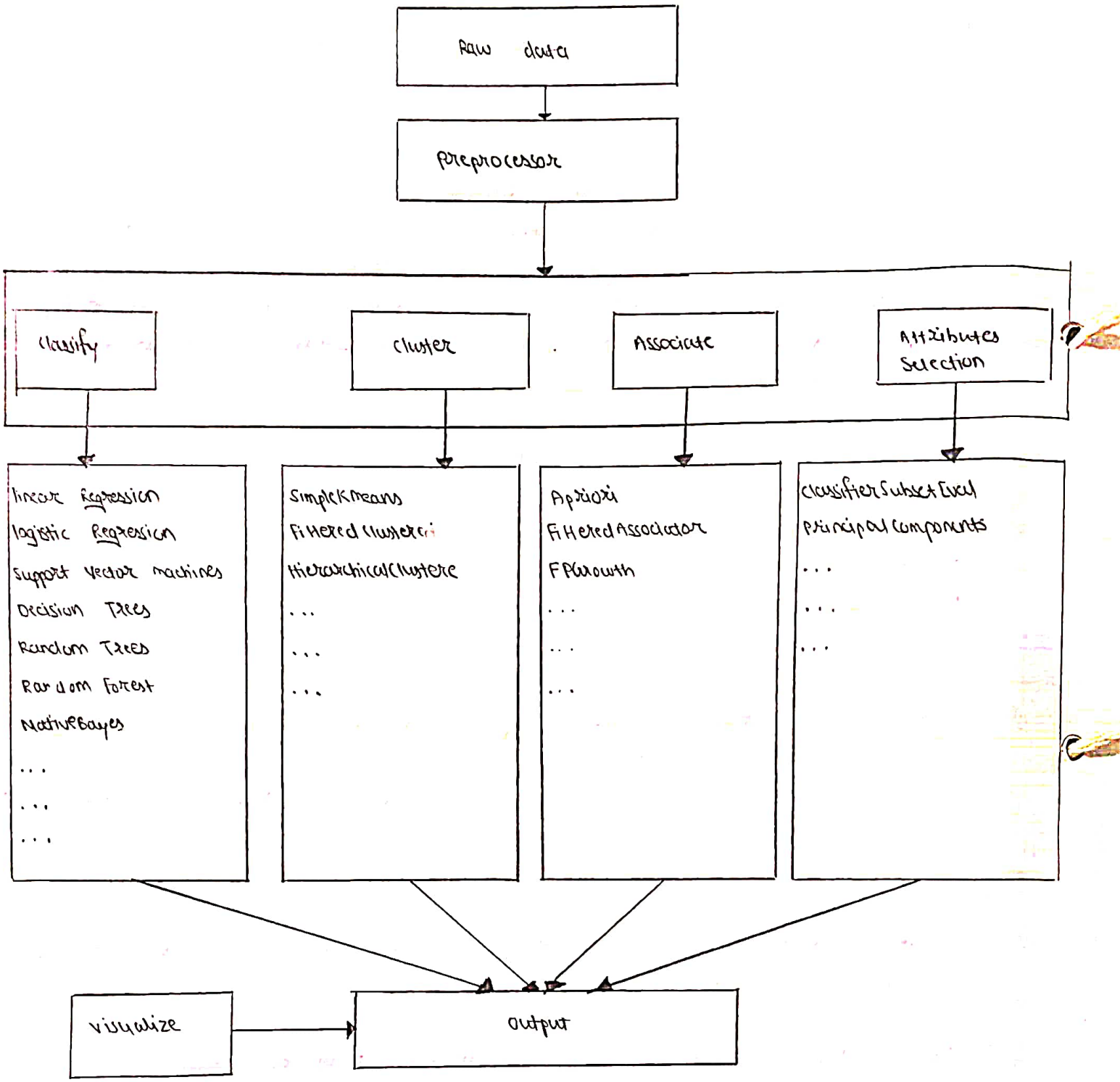
- Subject - oriented
- Integrated
- Time - variant
- Non - volatile

**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 present to them in an elementary form.

2. Store historical data : Data warehouse is required to store the time variable data from the past. This input is made to be used for various purposes.

3. Make strategic decision : Some strategies may be depending upon the data in the data warehouse. So, data warehouse contributes to making strategic decision.


Raw data
Preprocessor
Classify
Cluster
Associate
Attributes Selection
linear Regression
logistic Regression
Support Vector machines
Decision Trees
Random Trees
Random Forest
NativeBayes
...
SimpleKMeans
FilteredClusterer
HierarchicalClusterer
...
Apriori
FilteredAssociator
FPGrowth
...
ClassifierSubsetEval
PrincipalComponents
...
Visualize
Output

Date :

④ For data consistency and quality : Bringing the data from different sources at a commonplace, the user can effectively undertake to bring the uniformity and consistency in data.

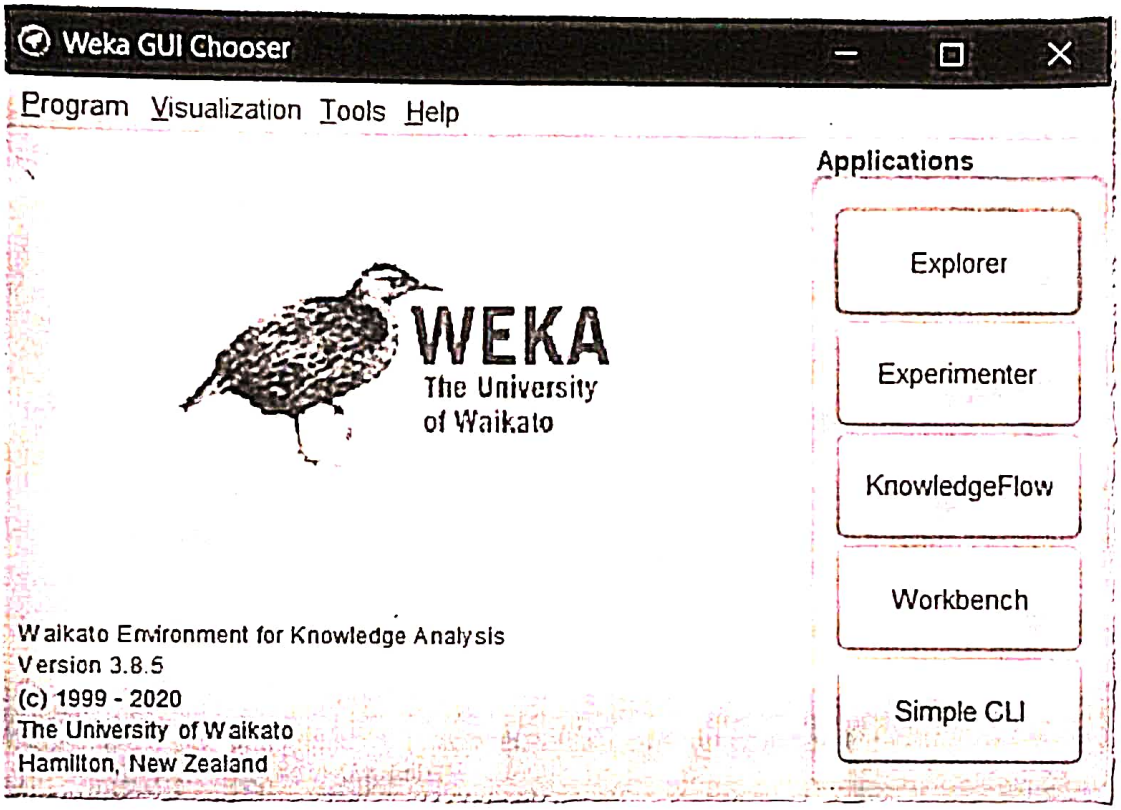
⑤ High response time : Data warehouse has to be ready for somewhat unexpected loads and types of queries, which demands a significant degree of flexibility and quick response time.

Benefits of data warehouse :→

① Understand business trends and make better forecasting decisions.
② Data warehouses are designed to perform well enormous amount of data.
③ The structure of data warehouses is more accessible for end-users to navigate, understand and query.
④ Queries that would be complex in many normalized databases could be easier to build and maintain in data warehouses
⑤ Data warehousing is an efficient method to manage demand for lots of information from lots of users.
⑥ Data warehousing provides the capabilities to analyze a large amount of historical data.

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 apply them to real-world data mining problems. What weka offers is summarized in the diagram -


Weka GUI Chooser
Program Visualization Tools Help
WEKA
The University of Waikato
Waikato Environment for Knowledge Analysis
Version 3.8.5
(c) 1999 - 2020
The University of Waikato
Hamilton, New Zealand
Applications
Explorer
Experimenter
KnowledgeFlow
Workbench
Simple CLI

Date :

If you observe the beginning of the flow of the image, you will understand that there are many stages 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 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 creates a reduced dataset.

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 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.

Weka is a data mining sw that uses a collection of machine learning algorithms. These algorithm can be applied directly to the data or called from the Java code

Date :

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 interfaces Explorer preprocessing, attribute selection, learning, visualization Experimenter Testing and evaluating machine learning algorithms knowledge Flow visual design of KDD process simple 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 below are some formats that weka supports, from where data can be imported :

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

Date :

Arff Formate :

An Arff file contains two sections - headers and data

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

An ARFF file requires the declarations of the relations.

@relation :-

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

@attribute :-

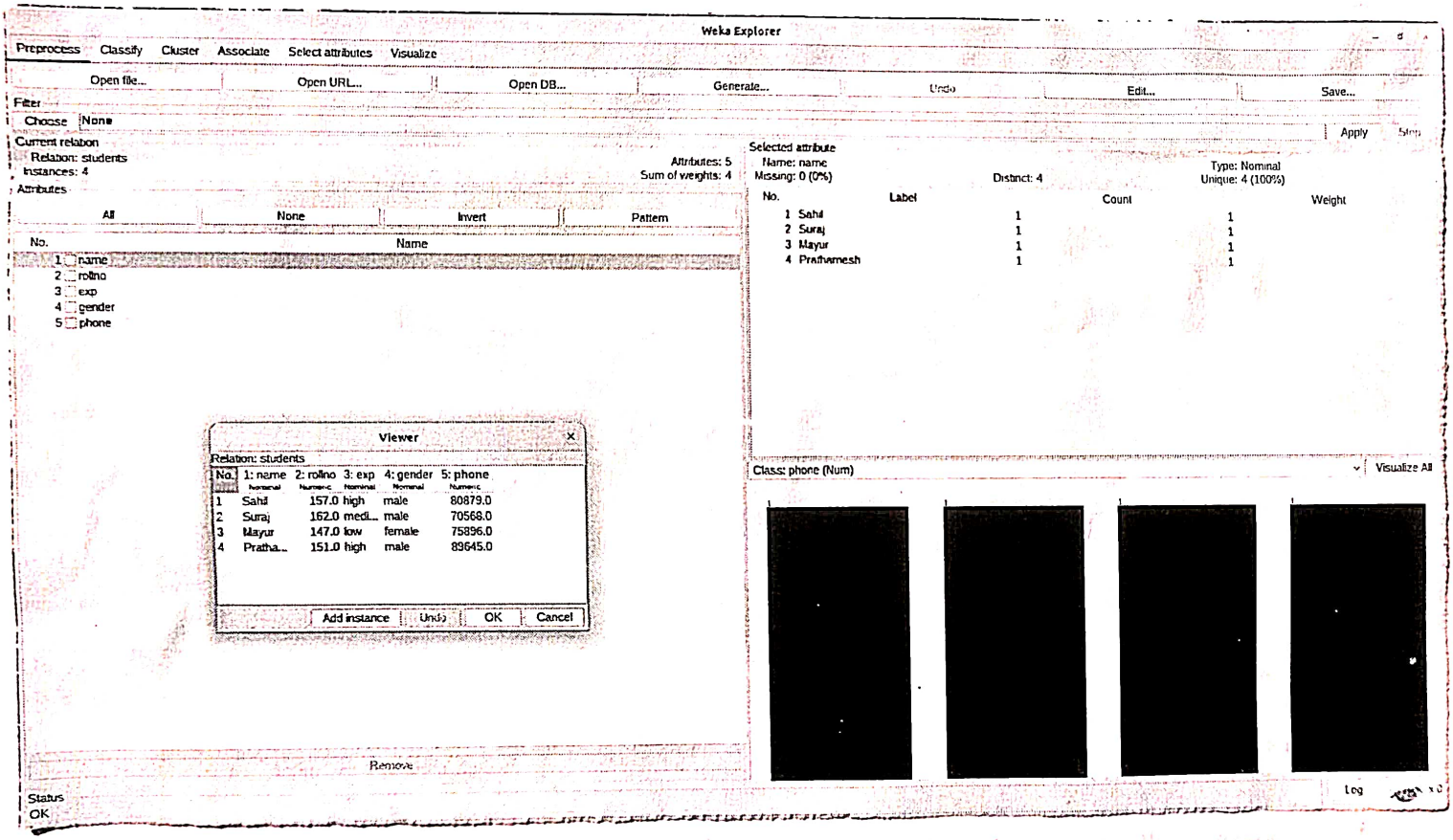
These are declared with their names and the type or range in the headersections. Weka supports the following data types for attributes :

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

Creating a Student Table with the help of data mining Tool weka :

@relation students

@attribute name {Sahil, Suraj, mayur, Prathamesh}
@attribute rollno numeric
@attribute exp {low, medium, high}
@attribute gender {male, female}
@attribute phone numeric


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: students
Instances: 4
Attributes: 5
Sum of weights: 4
Attributes
All
None
Invert
Pattern
No.
Name
1 name
2 rollno
3 exp
4 gender
5 phone
Remove
Selected attribute
Name: name
Missing: 0 (0%)
Distinct: 4
Type: Nominal
Unique: 4 (100%)
No.
Label
Count
Weight
1 Sahil 1 1
2 Suraj 1 1
3 Mayur 1 1
4 Prathamesh 1 1
Class: phone (Num)
Visualize All
Viewer
Relation: students
No. 1: name 2: rollno 3: exp 4: gender 5: phone
1 Sahil 157.0 high male 80879.0
2 Suraj 162.0 medi... male 70568.0
3 Mayur 147.0 low female 75896.0
4 Pratha... 151.0 high male 89645.0
Add instance
Undo
OK
Cancel
Status
OK
Log

Date :

@data

Sahil , 157 , high , male , 80879
Suraj , 162 , medium , male , 70568
Mayur , 147 , low , female , 75896
Prathamesh , 151 , high , male , 89645

Conclusion :-

Thus the training data table is created and WEKA tool is explored.

Viva Questions :

① Define Data warehouse.

→ It is a system used for reporting and data analysis and is considered a core component of business intelligence. A datawarehouse is a subject oriented, integrated, non-volatile and time variant collection of data in support of managements decision making process.

② what are the characteristics of Data warehousing ?

→ There are four characteristic of data warehousing and they are :

a) Subject oriented
b) Integrated
c) Time - variant
d) Non - Volatile.

③ what is weka Tool and what are the significance of weka ?

→ ① weka is an open source software software.
② It is collection of machine learning algorithms for data mining tasks.
③ weka contains tools for data pre-processing , classification , regression , clustering , association rules and visualization.

* Significance of weka :

① Free availability under GNU General public License

Date :

② Portability, since it is fully implemented in Java programming language
③ A comprehensive collection of data preprocessing and modelling techniques.
④ Ease of use due to its graphical user interfaces.
⑤ It provides you a visualization tool to inspect the data.
⑥ The use of WEKA results in quicker development of ML models on the whole.