DATA MINING LAB MANUAL

MR22-1CS0148

Index

S.No	Experiment
1	Installing Weka on Windows
2	Start working with WEKA tool kit and understand the features of WEKA tool kit. Loading Data from different sources in WEKA. Various File Formats supported by WEKA. And Study the ARFF file format.
3	Demonstration of creating a Student dataset (student.arff) using WEKA tool in Data Mining.

Experiment 1: Installing Weka on Windows

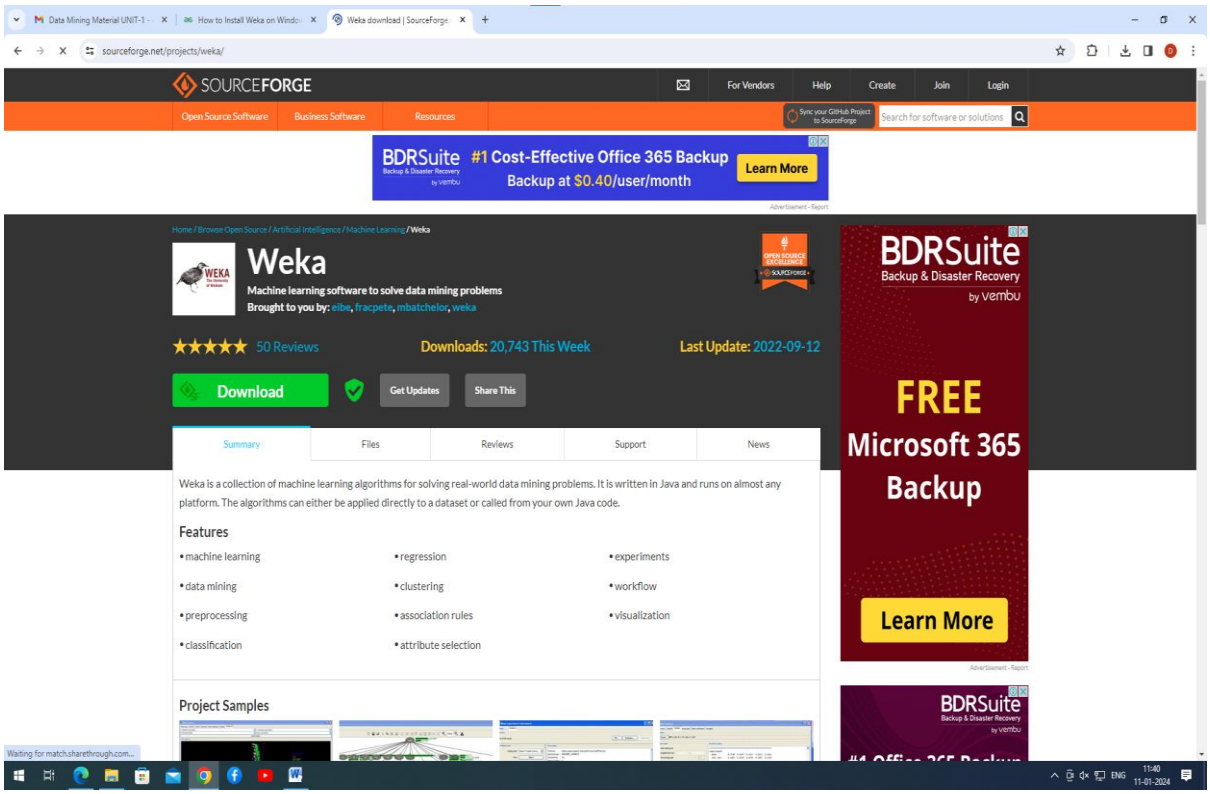
Weka stands for Waikato Environment for Knowledge Analysis, it is software that is used in the data science field for data mining. It is free software. It is written in Java hence it can be run on any system supporting Java, so weka can be run on different operating systems like Windows, Linux, Mac, etc. Weka provides a collection of visualization tools that can be used for data analysis, cleaning, and predictive modeling. Weka can perform a number of tasks like data preprocessing, clustering, classification, regression, visualization, and feature selection.

Installing Weka on Windows:

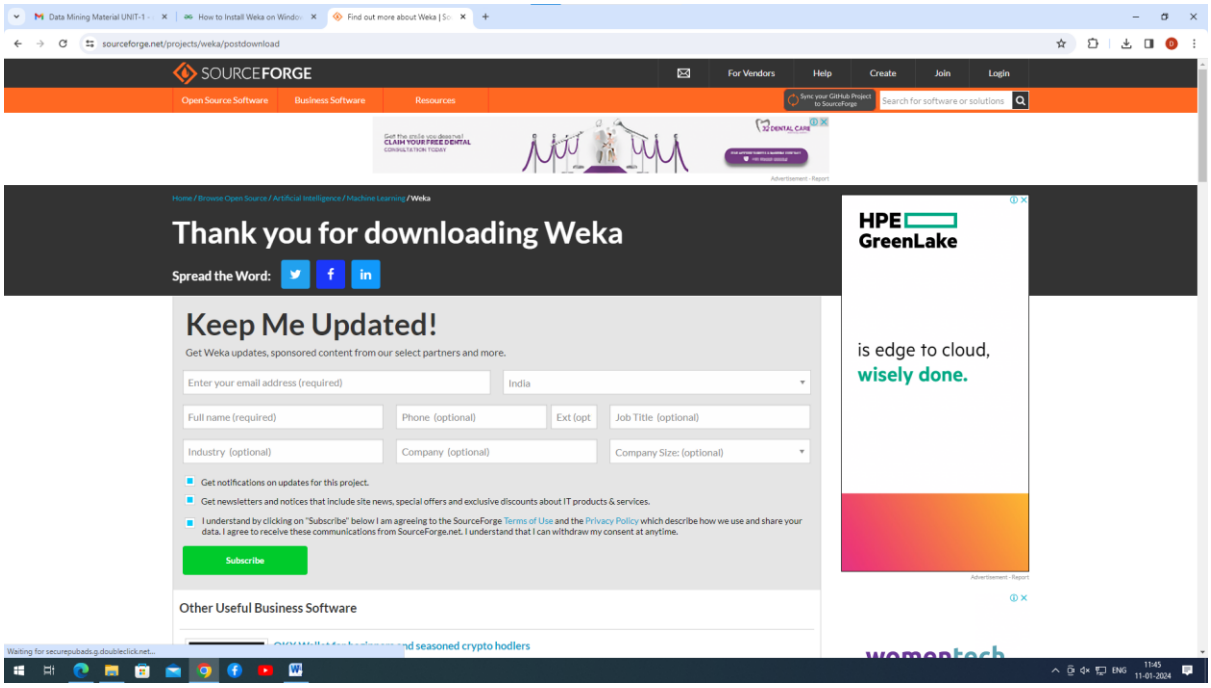
Follow the below steps to install Weka on Windows:

Step 1: Visit this website using any web browser. Click on Free Download.

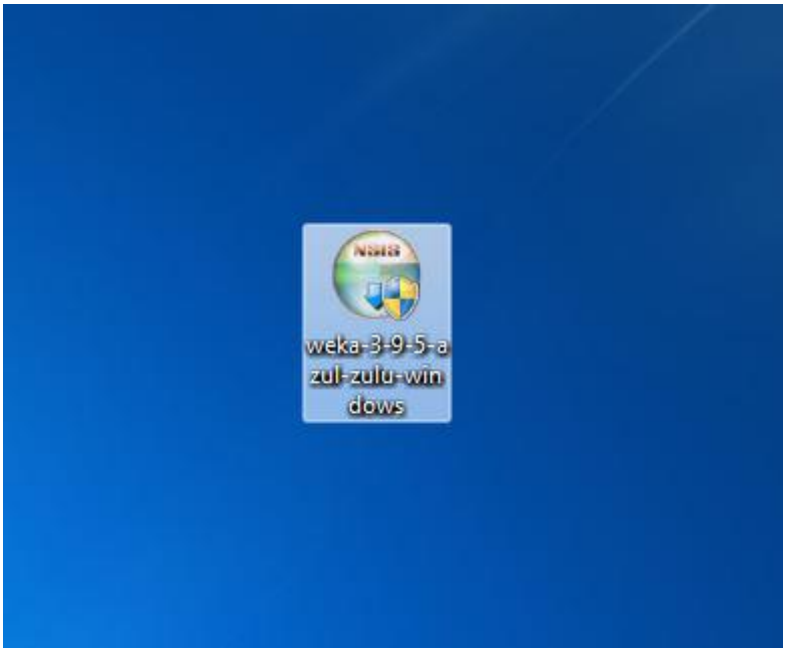
https://waikato.github.io/weka-wiki/downloading_weka/



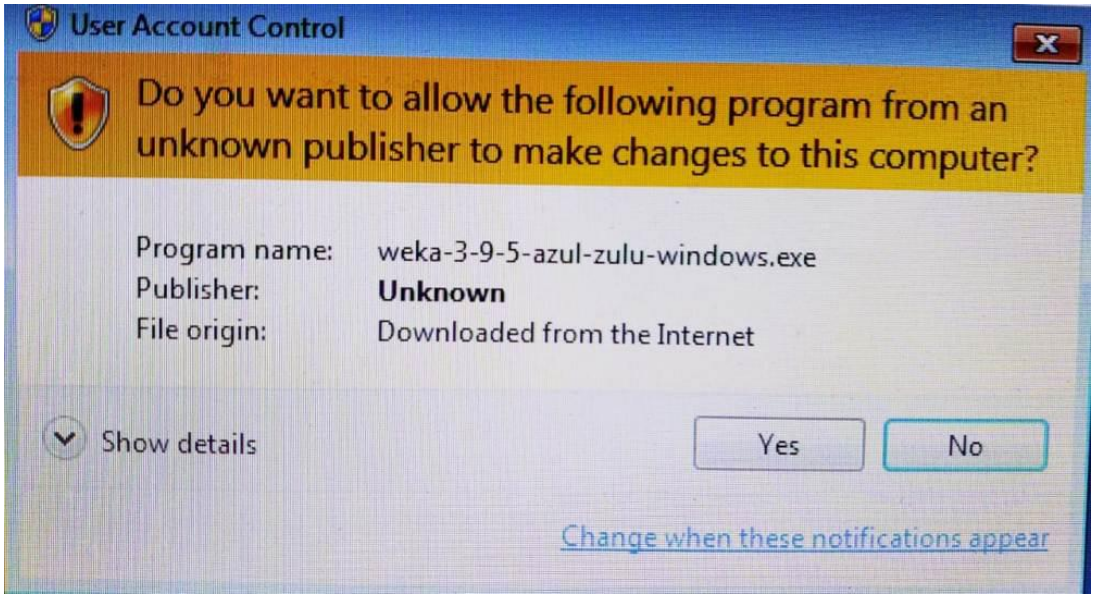
Step 2: It will redirect to a new webpage, click on Start Download. Downloading of the executable file will start shortly. It is a big 120 MB file that will take some minutes.



Step 3: Now check for the executable file in downloads in your system and run it.

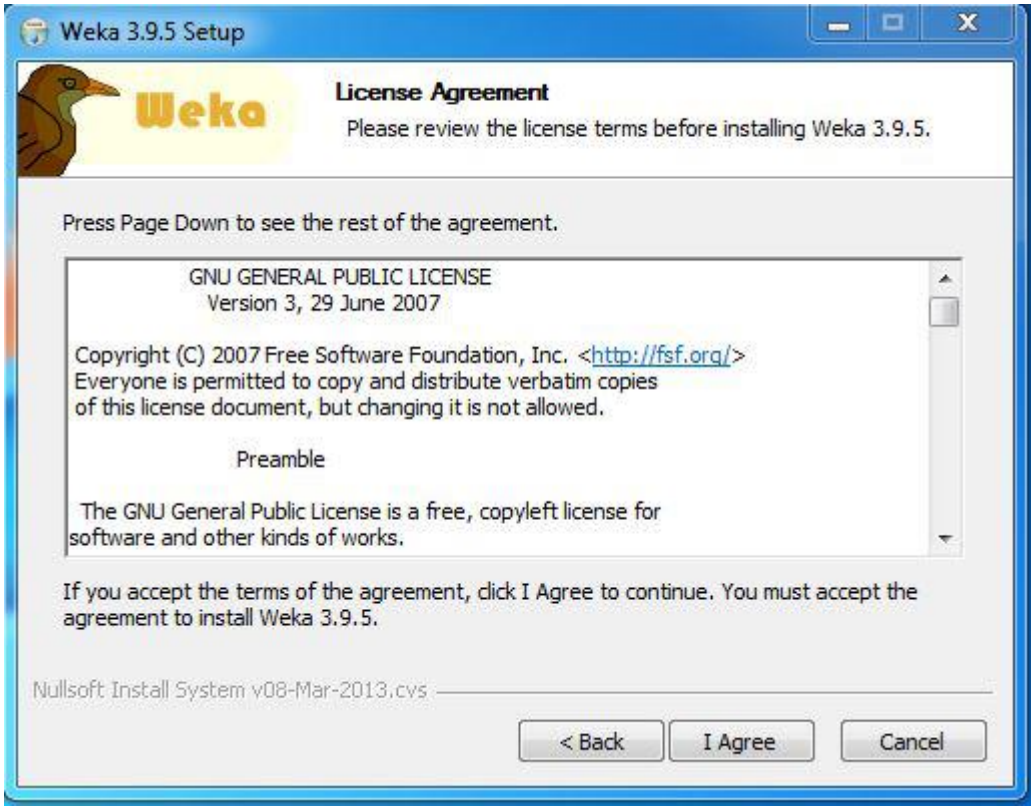


Step 4: It will prompt confirmation to make changes to your system. Click on Yes.

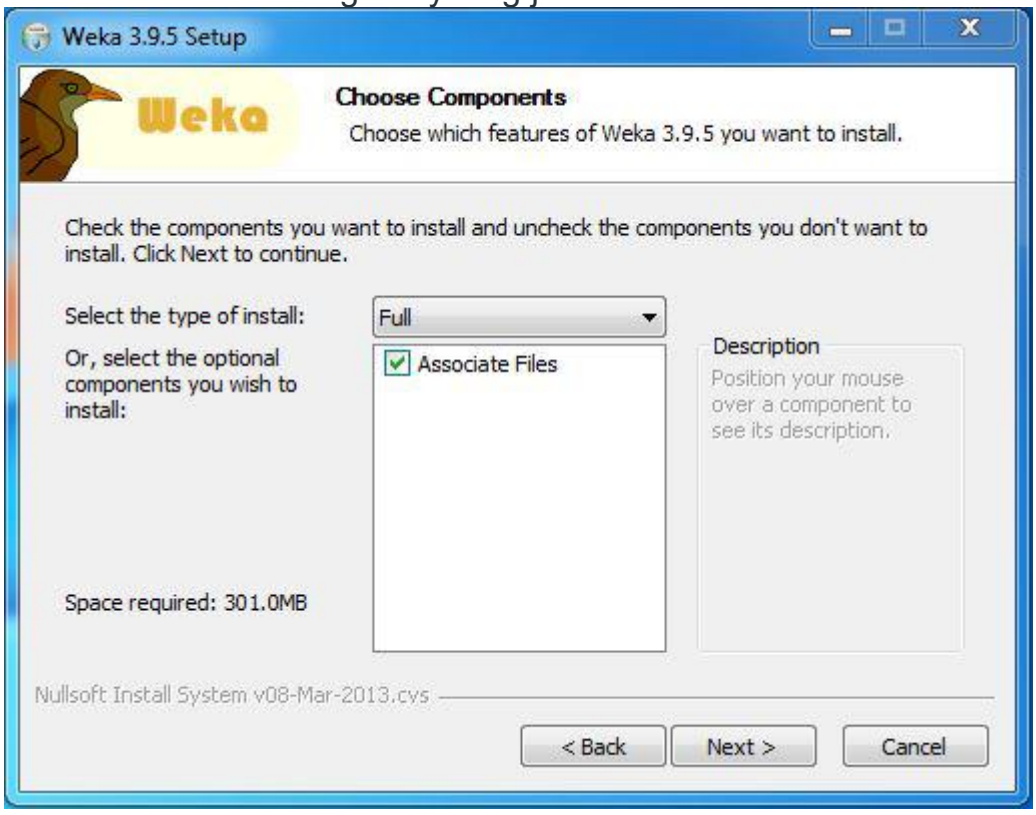


Step 5: Setup screen will appear, click on Next.

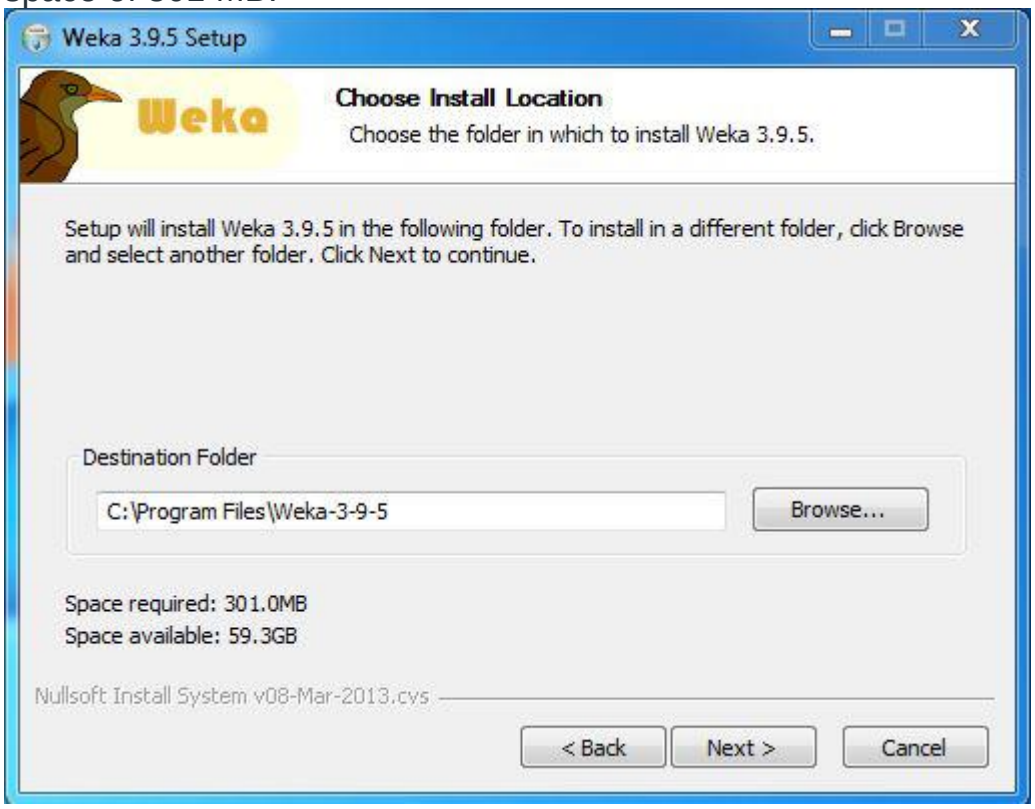
Step 6: The next screen will be of License Agreement, click on I Agree.



Step 7: Next screen is of choosing components, all components are already marked so don't change anything just click on the Install button.

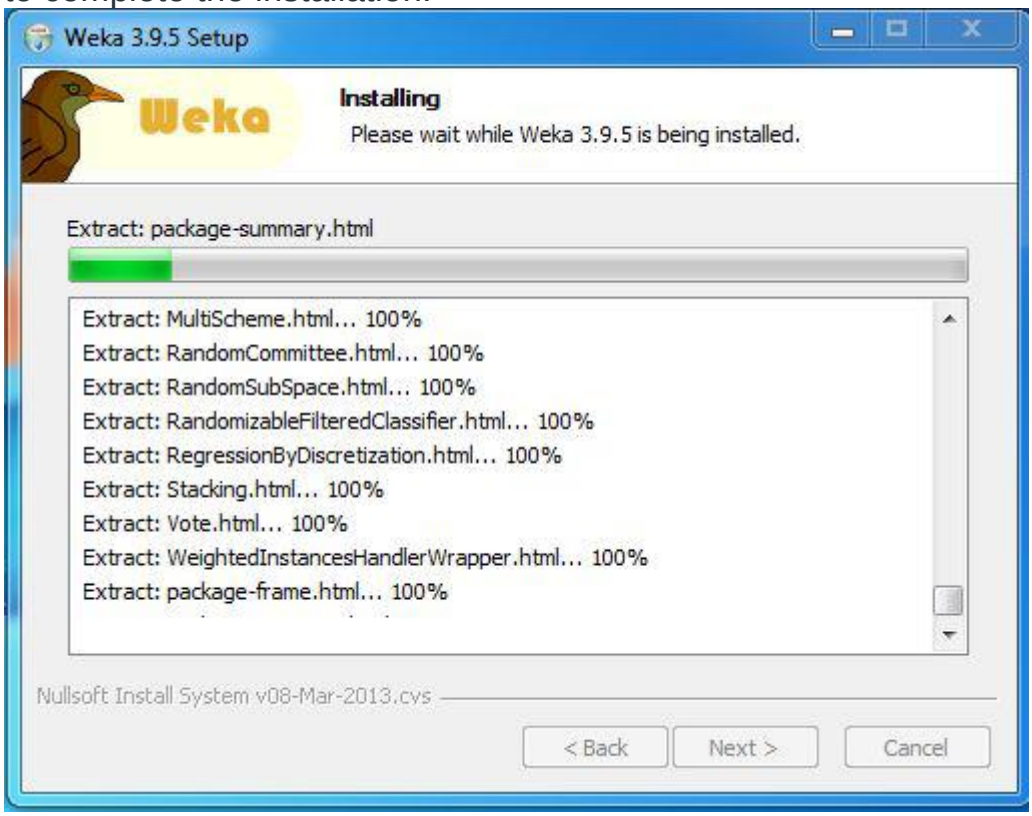


Step 8: The next screen will be of installing location so choose the drive which will have sufficient memory space for installation. It needed a memory space of 301 MB.

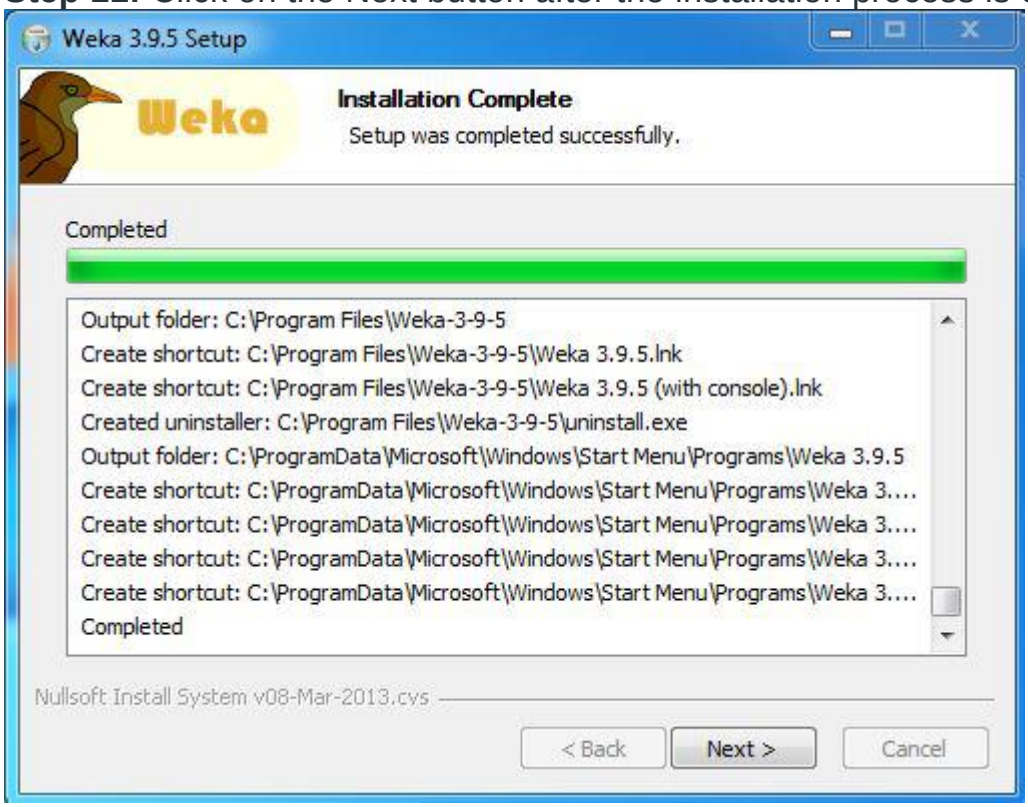


Step 9: Next screen will be of choosing the Start menu folder so don't do anything just click on Install Button.

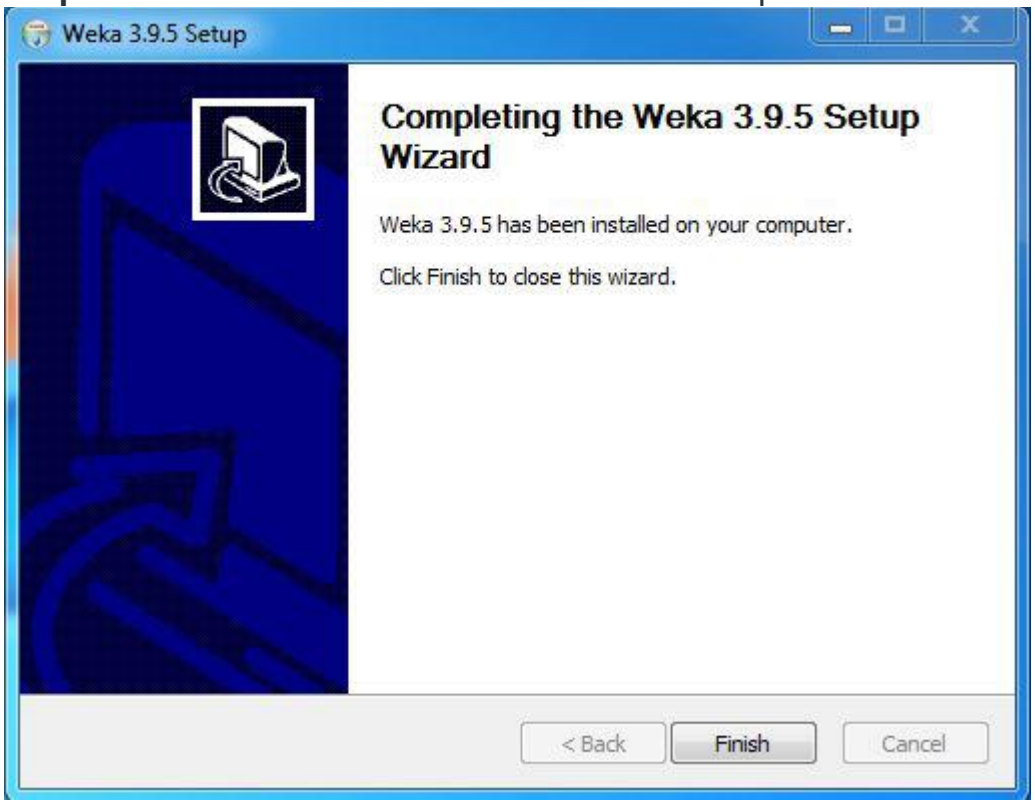
Step 10: After this installation process will start and will hardly take a minute to complete the installation.



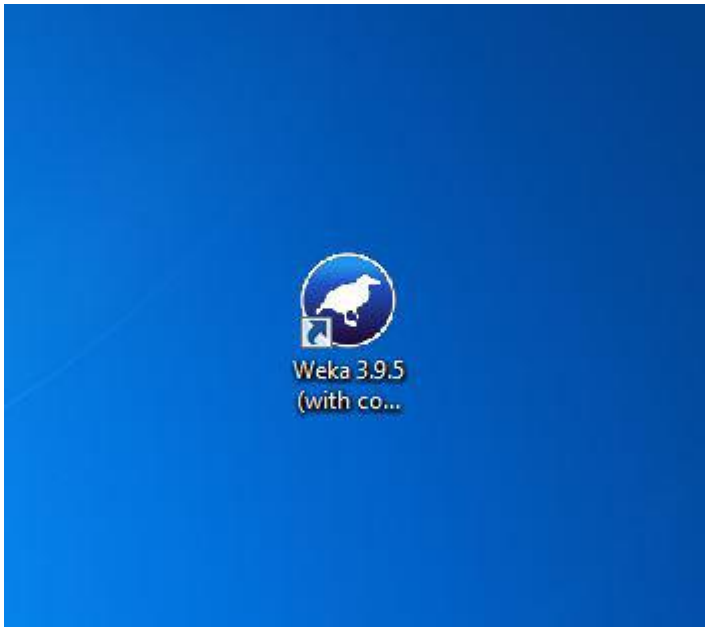
Step 11: Click on the Next button after the installation process is complete.



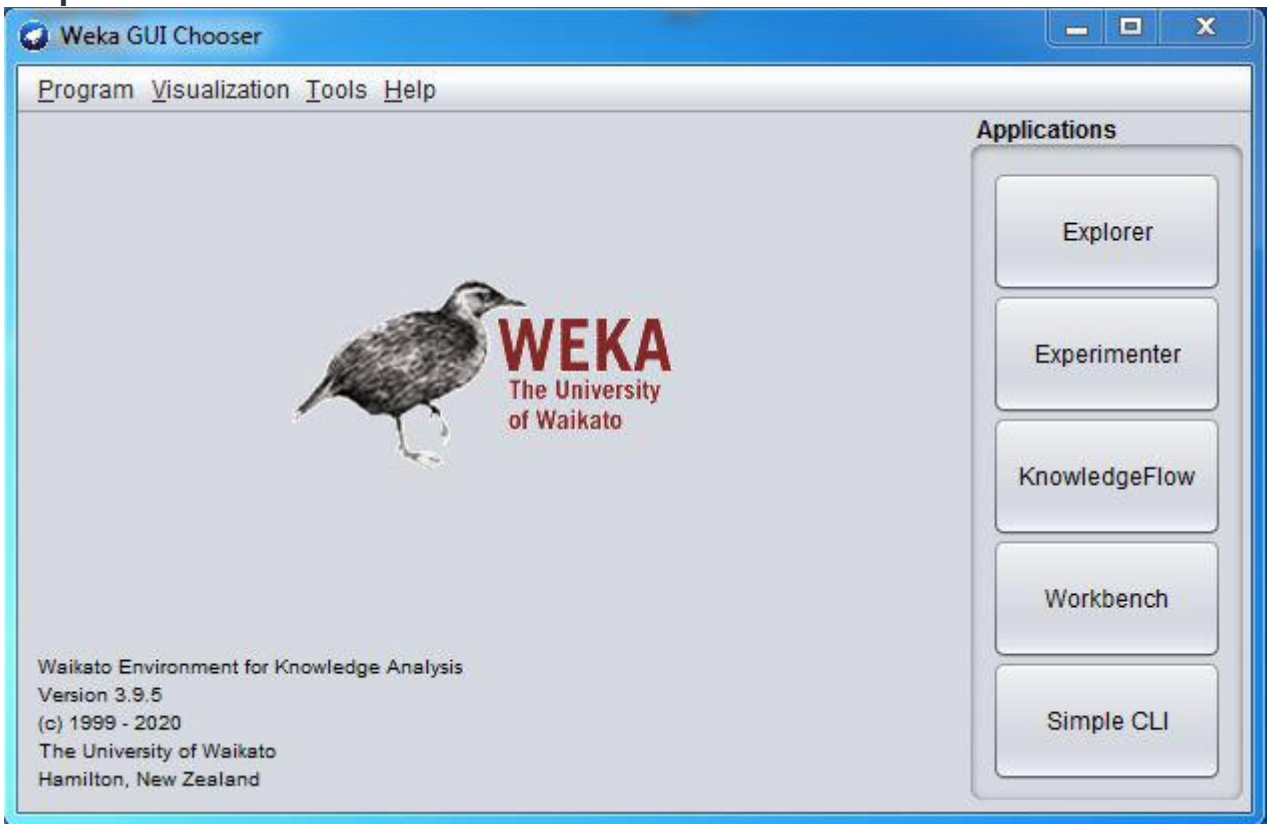
Step 12: Click on Finish to finish the installation process.



Step 13: Weka is successfully installed on the system and an icon is created on the desktop.



Step 14: Run the software and see the interface.



Congratulations!! At this point, you have successfully installed Weka on your windows system.

Experiment 2: Start working with WEKA tool kit and understand the features of WEKA tool kit. Loading Data from different sources in WEKA. Various File Formats supported by WEKA. And Study the ARFF file format.

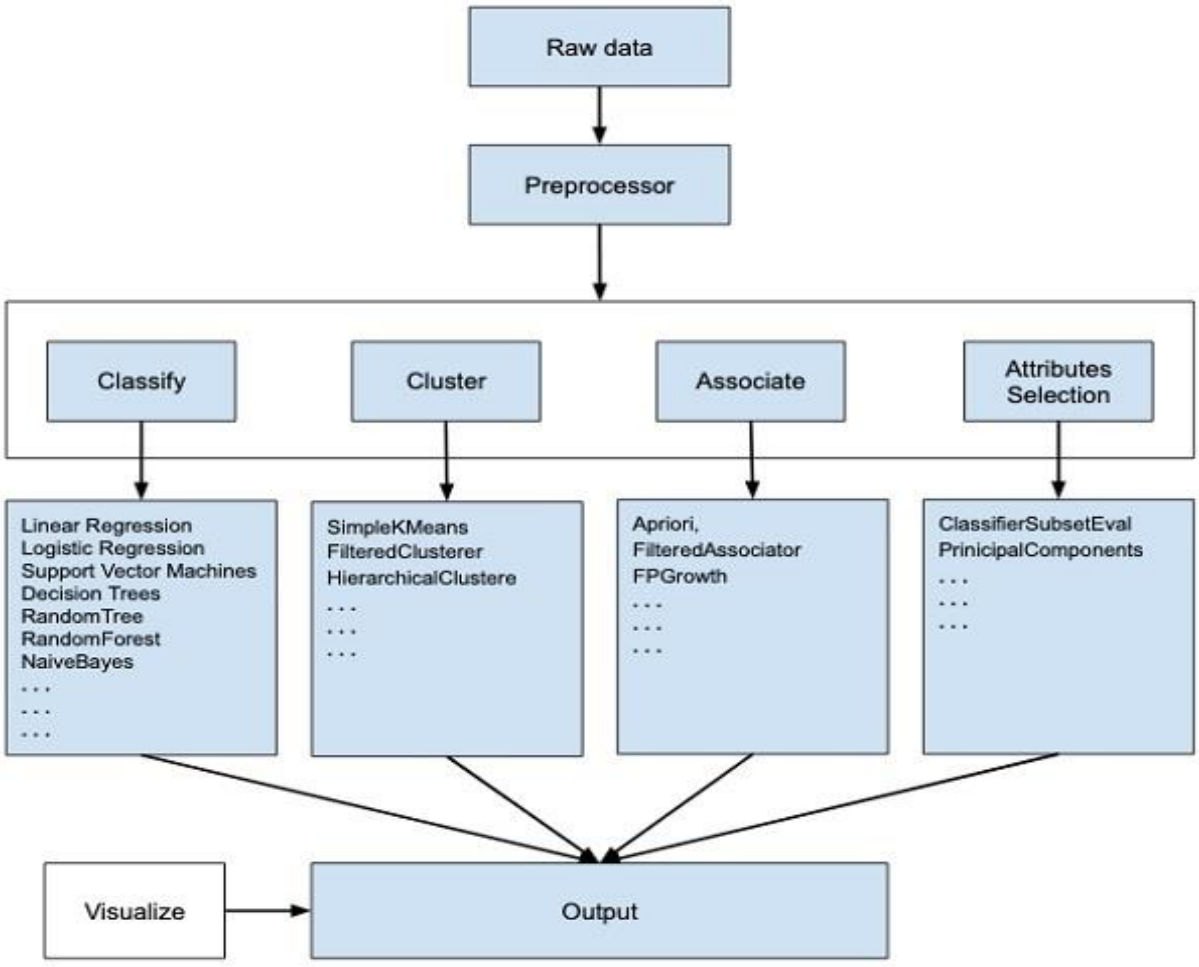
Start working with WEKA tool kit and understand the features of WEKA tool kit.

Solution :

What is WEKA:

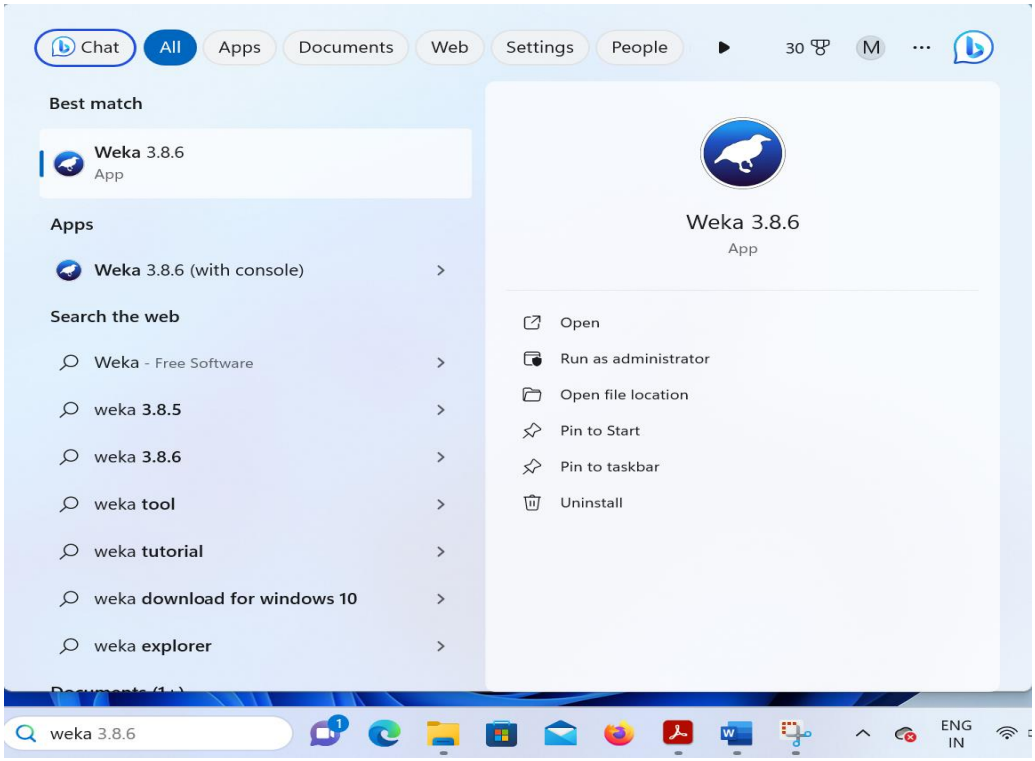
WEKA, an open-source software, offers a range of tools for data preprocessing, implementation of various Data Mining algorithms, and visualization tools. These resources enable users to develop data mining techniques and effectively apply them to real-world data mining problems.

The diagram presented below provides a concise summary of the offerings provided by WEKA.



To start Weka:

Search for Weka 3.8.6 and click on **Weka 3.8.6 app.**

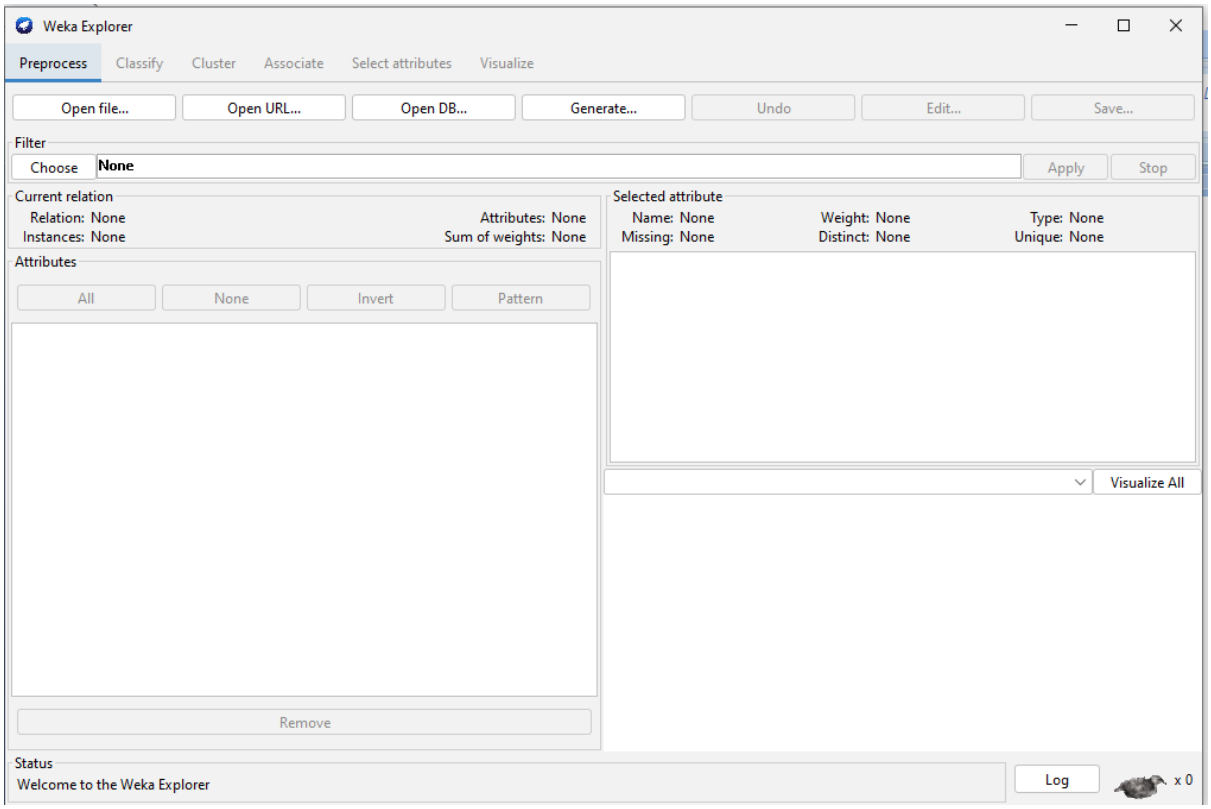


The following **Graphical User Interface** Of WEKA you get when you click on **Weka 3.8.6 app**.

The GUI of WEKA gives five options: **Explorer, Experimenter, Knowledge flow, Workbench,** and **Simple CLI.** Let us understand each of these individually.

1. Explorer

It is an environment for exploring data with WEKA. And it apply the various data mining algorithms. When you click on the **Explorer** button in the **Applications** selector, it displays the following window.



Located at the uppermost section of the window, positioned just below the title bar, is a series of tabs. Upon launching the Explorer, only the first tab is enabled, while the remaining tabs are displayed in an unresponsive manner. This is due to the prerequisite of opening and pre-processing a data set before data exploration.

The **tabs** are as follows:

Preprocess:

The first step in Data Mining is to preprocess the data. You will select the data file in the Preprocess option. Then, you will process the data and make it suitable for applying the different Data Mining algorithms.

Classify:

The Classify tab offers a range of Data Mining algorithms for the classification of your data. Some of the algorithms that can be applied include Linear Regression, Logistic Regression, Support Vector Machines, Decision Trees, Random Tree, Random Forest, Naive Bayes, and others.

Cluster:

The Cluster tab contains a variety of clustering algorithms, including Simple K-Means, Filtered Clusterer, Hierarchical Clusterer, and many more.

Associate:

The Associate tab contains Apriori, Filtered Associator and FPGrowth. These are used to learn / discover association rules in the data.

Select attributes:

This tab contains various methods to select the most relevant attributes in the data.

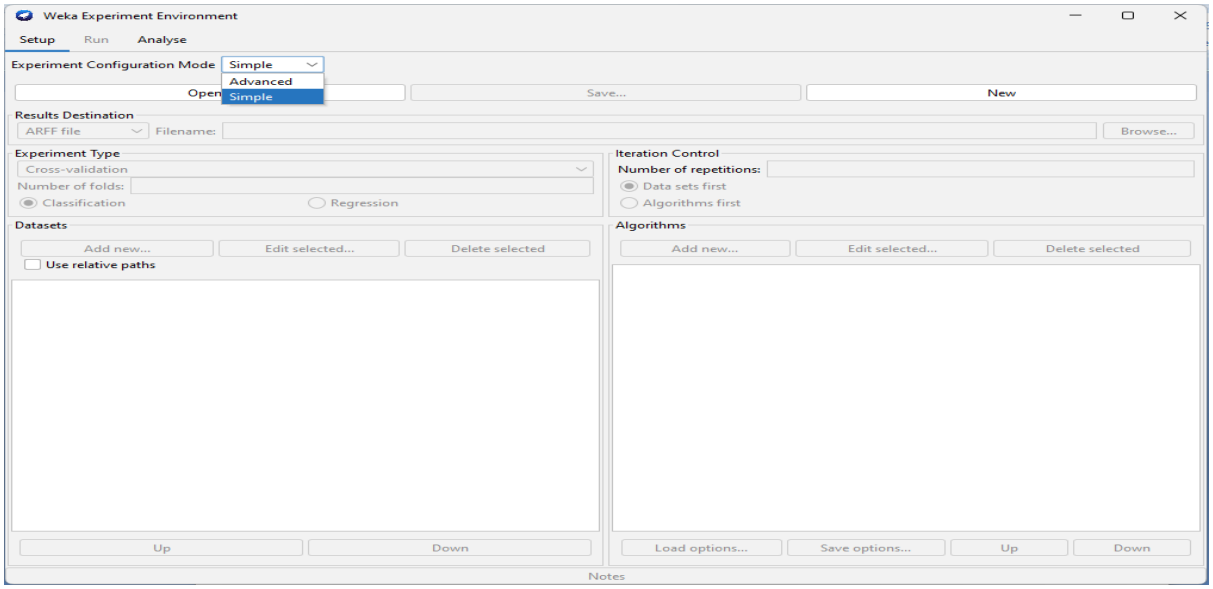
Visualize:

In this tab, various plots and graphs are available to show the trends identified by the model. I.e. it displays an interactive 2D plot of the data.

2. Experimenter

The Experimenter Environment allows users to easily create, run, modify, and analyze experiments. Users can create experiments that test multiple schemes on different datasets and analyze the results to determine statistical differences between the schemes.

When you click on the **Experimenter** button in the **Applications selector**, it displays the following window.



The Experimenter is available in two variants, those are

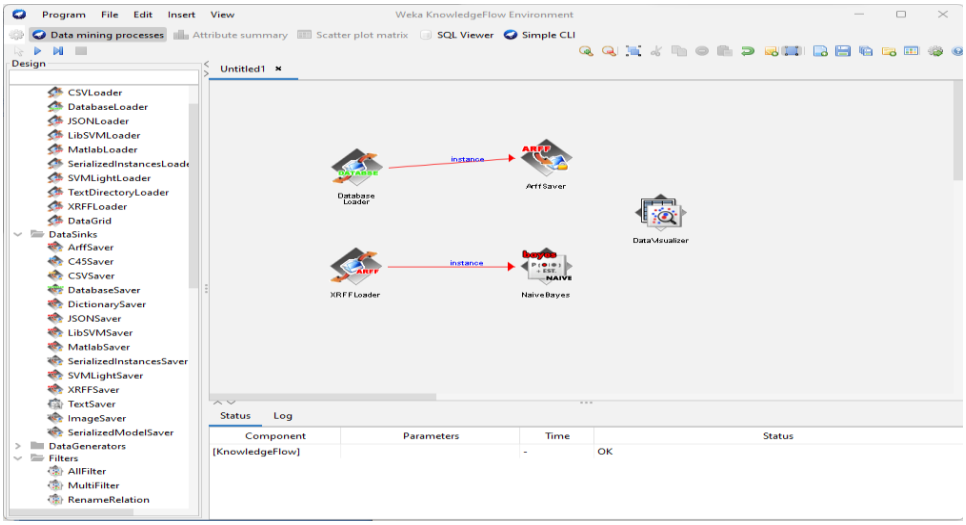
- **Simple**
  This variant provides most of the functionality one needs for experiments
- **Advanced**
  This is an interface with full access to the Experimenter's capabilities.

3. Knowledge flow

The Knowledge Flow offers an alternative to the Explorer as a graphical user interface for accessing the core algorithms of WEKA.

The Knowledge Flow platform offers an interface that draws inspiration from data-flow principles, specifically designed for WEKA. Users are able to choose components from a selection of WEKA tools, position them on a layout canvas, and establish connections between them. This facilitates the creation of a knowledge flow, enabling efficient processing and analysis of data.

When you click on the **Knowledge flow** button in the **Applications selector**, it displays the following window.



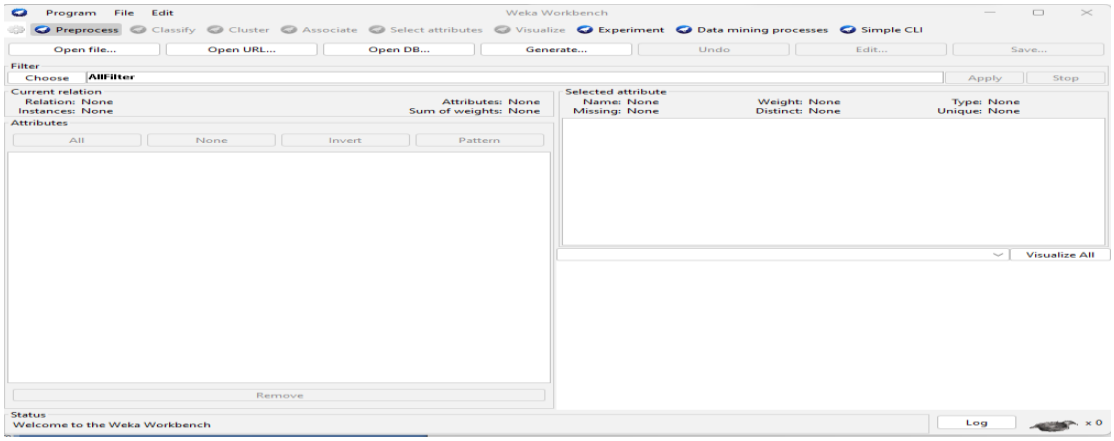
Currently, all classifiers, filters, clusterers, associators, loaders, and savers provided by WEKA are accessible within the Knowledge Flow platform, along with extra tools.

4. Workbench

The Workbench is an integrated environment that combines all graphical user interfaces into a unified or single interface.

If you frequently switch between multiple interfaces, such as the Explorer and the Experiment Environment, it can be beneficial. This is often the case when testing various scenarios in the Explorer and promptly implementing acquired knowledge into controlled experiments.

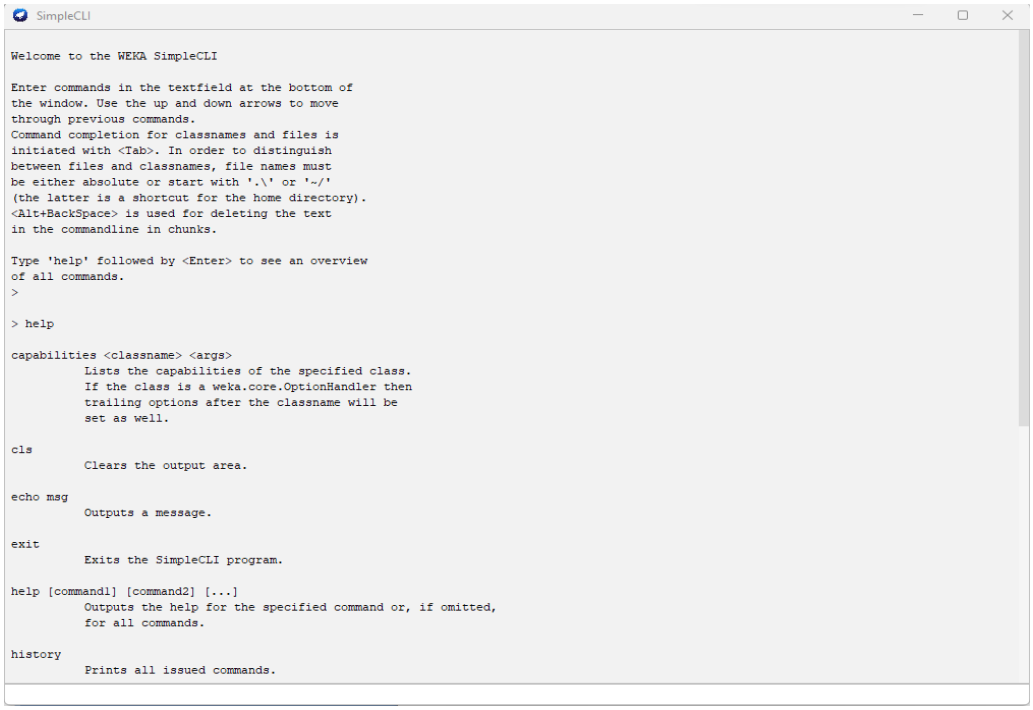
When you click on the Workbench button in the Applications selector, it displays the following window.



5. Simple CLI

The Simple Command Line Interface (CLI) grants comprehensive access to all Weka classes, including classifiers, filters, clusterers, and more, while eliminating the inconvenience of managing the CLASSPATH (it simplifies the one used during Weka's initialization). It presents a straightforward Weka shell with distinct command line and output sections.

When you click on the **Simple CLI** button in the **Applications selector**, it displays the following window.



Loading Data from different sources in WEKA.

Solution :

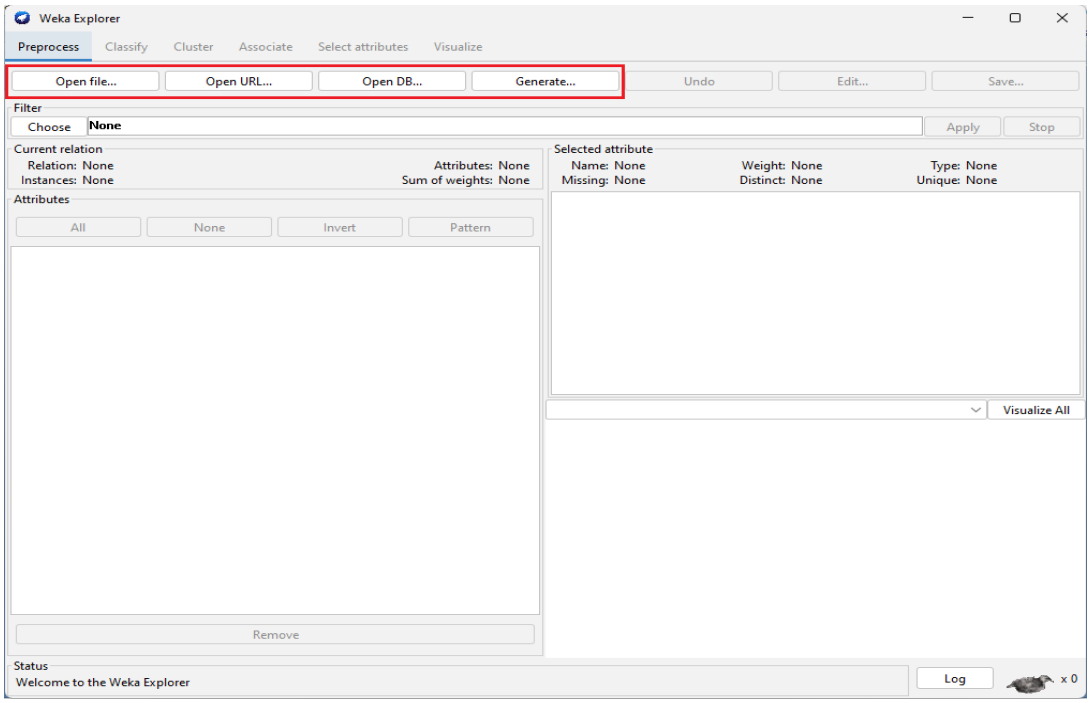
Loading Data from different sources in WEKA :

In order to utilize Weka Explorer, it is essential to start by loading the data into the application.

Multiple sources are available for data loading within Weka Explorer.Those are,

1. **Local file system**
2. **Web**
3. **Database**
4. **Generate Artificial Data**

The diagram presented below provides a concise summary of the offerings provided by WEKA.

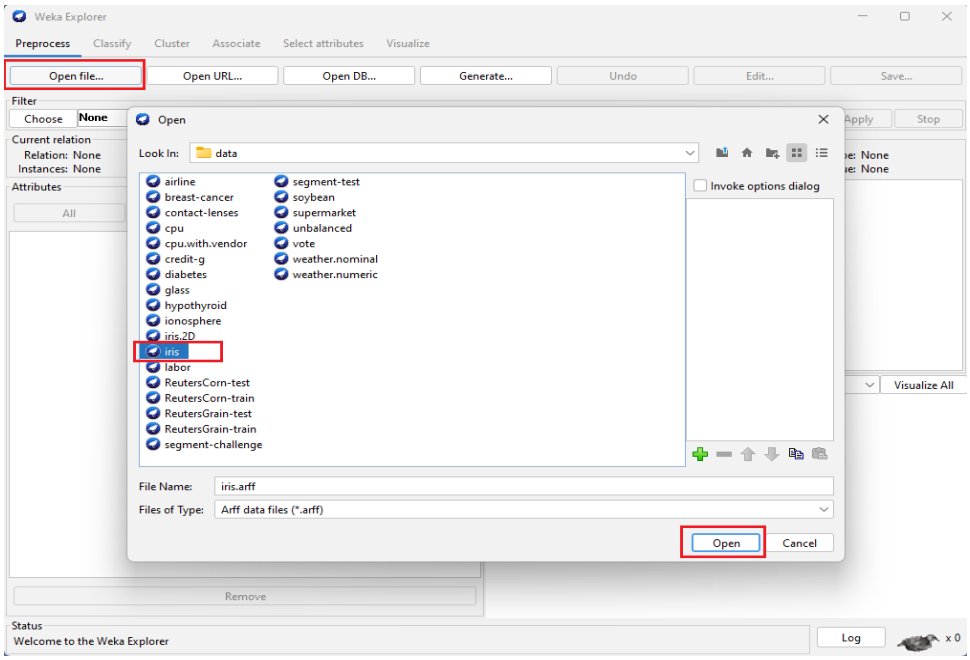


1. Loading data from Local file System:

In this, we are going to load data from the Local File System by clicking on the **Open file...** button.

Steps:

- Click on **Open file...** button.
- **Navigate the folder,** where the data files are stored.
- Select the **required data file.**
- Click on **Open** button.

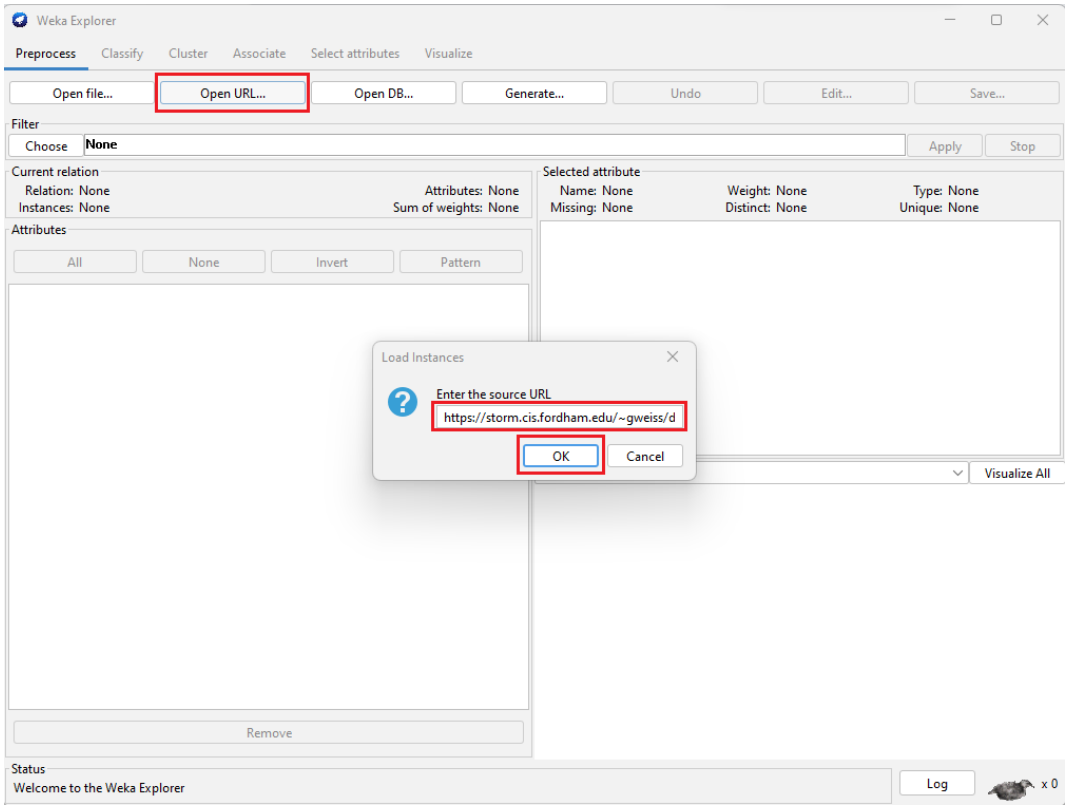


2. Loading data from Web:

In this, we are going to load data from the Web by clicking on the **Open URL...** button.

Steps:

- Click on **Open URL...** button.
- **Enter the URL** of data source in the popup box.
- Click on **Ok** button.



Example public **Data source URLs** are:

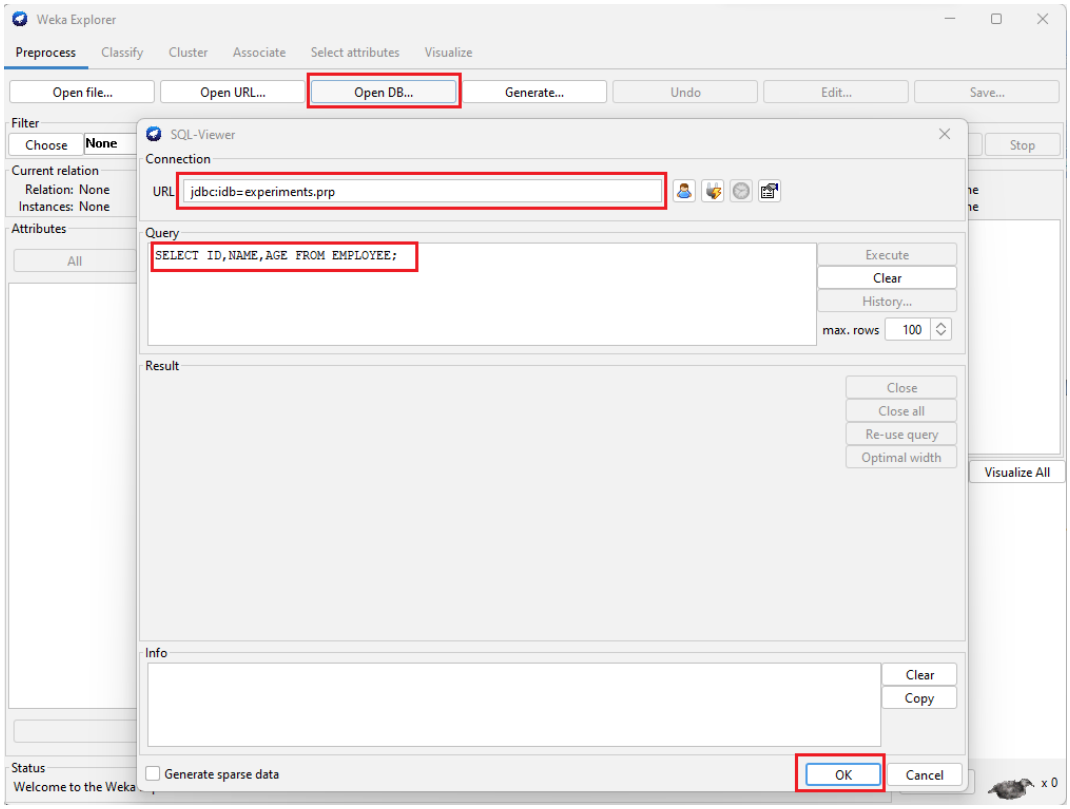
- https://storm.cis.fordham.edu/~gweiss/data-mining/weka-data/weather.nominal.arff
- https://storm.cis.fordham.edu/~gweiss/data-mining/weka-data/iris.arff
- https://storm.cis.fordham.edu/~gweiss/data-mining/weka-data/vote.arff

3. Loading data from Database:

In this, we are going to load data from the Database by clicking on the **Open DB...** button.

Steps:

- Click on **Open DB...** button.
- Enter the **Connection URL** of database.
- Type **SQL Query** to get data from Database table.
- Click on **Ok** button.

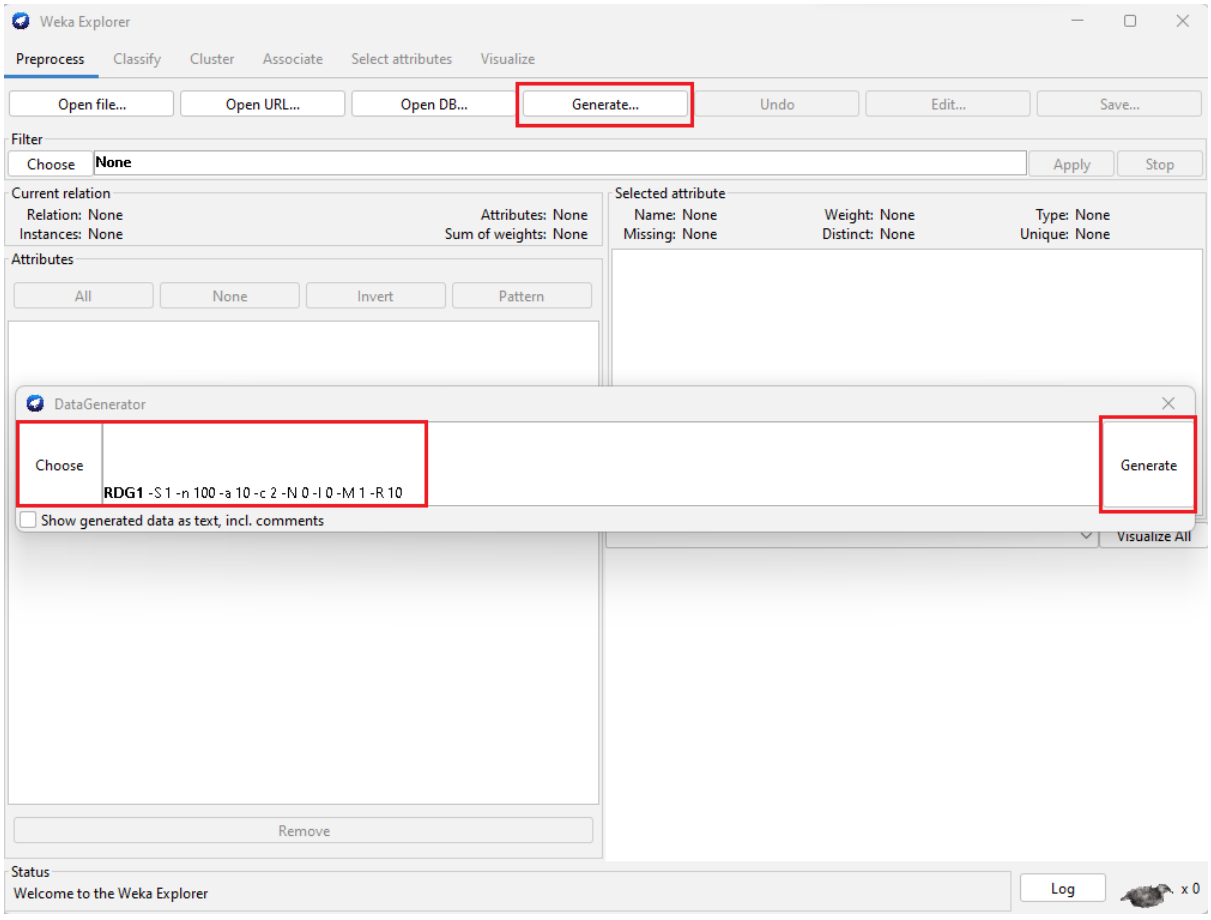


4. Generate Artificial Data:

In this, the artificial data (random data) will be generated by clicking on the **Generate...** button.

Steps:

- Click on **Generate...** button.
- **Choose Data Generator**.
- Click on **Generate** button.



Various File Formats supported by WEKA. And Study the ARFF file format.

Solution :

Various File Formats supported by WEKA :

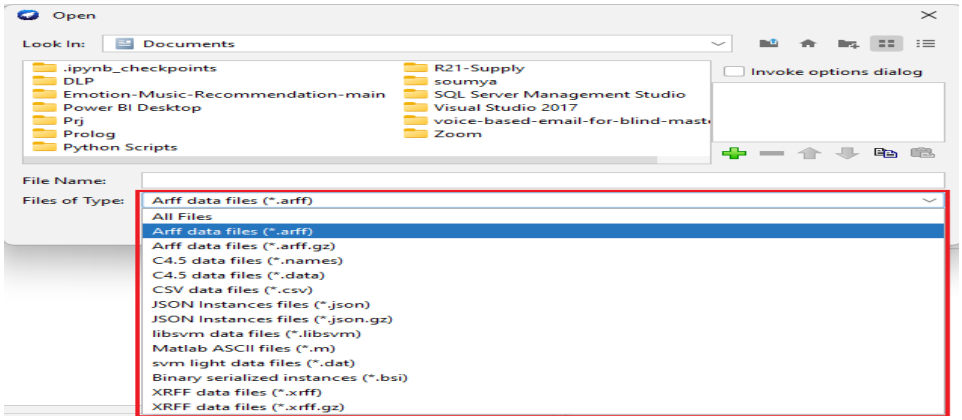
In order to work with Weka Explorer, it is essential to **load the data** into the application. The data may in different formats such as CSV, Text, JSON and so on.

WEKA supports a **wide range of file formats** to load the data.

The following is the complete list of file formats

- ☛ Arff data files(*.arff)
- ☛ Arff data files(*arff.gz)
- ☛ C4.5 data files(*.names)
- ☛ C4.5 data files(*.data)
- ☛ CSV data files(*.csv)
- ☛ JSON instance files(*.json)
- ☛ JSON instance files(*.json.gz)
- ☛ libsvm data files(*.libsvm)
- ☛ Matlab ASCII files(*.m)
- ☛ svm light data files(*.dat)
- ☛ Binary Serialized instances(*.bsi)
- ☛ XRFF data files(*.xrff)
- ☛ XRFF data files(*.xrff.gz)

The following screen displays all supported file formats in dropdown list at the bottom of window.



As we see the WEKA supports various formats of data to load, among those formats the most commonly used data formats are **Arff data files(*.arff)** and **CSV data files(*.csv)**.

NOTE : The default data format of WEKA is **Arff data files (*.arff)**.

Study the ARFF file format:

An ARFF (**Attribute-Relation File Format**) file is an ASCII text file that describes a list of instances sharing a set of attributes.

The ARFF file format has mainly **two sections**, those are

• **Header** section

• **Data** section

Header section:

The Header section of the ARFF file contains the **name of the relation**, a **list of the attributes** and their **types**.

@RELATION Declaration

The relation name is defined as the first line in the ARFF file.

format:

@RELATION <relation-name>

- where <relation-name> is a string. The relation name must be quoted if the name includes spaces.

@ATTRIBUTE Declaration

The attribute specifies name of the attribute along with type.

format:

@ATTRIBUTE <attribute-name> <datatype>

- where the <attribute-name> must start with an alphabet. The attribute name must be quoted if the name includes spaces.

Weka supports the following four datatypes:

1. Numeric attributes:

Numeric attributes can be real or integer numbers.

2. Nominal attributes:

Nominal values are defined by providing the possible values: { nominal-value1, nominal-value2, nominal-value3,… }

3. String attributes:

String attributes allow us to define attributes holding textual values.

4. Date attributes:

Date attribute defined as follows

@ATTRIBUTE <name> date [<date-format>]

- where <name> is the name for the attribute and <date-format> is an optional string. The default date-format string is yyyy-MM-dd'T'HH:mm:ss.

Example of Header Section:

```
% Title: Student Database
%
% Sources:
% (a) Creator: Mr.T.M
% (b) Date: Oct, 2023
%

@RELATION student

@ATTRIBUTE sid NUMERIC
@ATTRIBUTE age NUMERIC
@ATTRIBUTE gender {male, female}
```

In the above example,

The lines which start with **%** are treated as comments.

@**RELATION** specifies the name of the relation.

@**ATTRIBUTE** specifies name of the attribute along with type and possible values.

Data section:

The Data section of the ARFF file contains the **list of data values (instance data)** separated by comma.

Example of Body Section:

```
@DATA
101,20,male
102,19,female
103,?,male
```

In the above, there are 3 instances with numeric and nominal values. And the symbol ? indicates missing values.

✍**NOTE :** The **@RELATION, @ATTRIBUTE** and **@DATA** declarations are **case insensitive**. i.e @RELATION and @relation are treated as same in ARFF file format.

The following is the **complete ARFF file**

Filename: student. arff

```
% Title: student Dataset
%
% Sources:
%      (a) Creator: Mr.T.M
%      (b) Date: Oct, 2023
%

@RELATION student

@ATTRIBUTE sid NUMERIC
@ATTRIBUTE age NUMERIC
@ATTRIBUTE gender {male, female}

@DATA
101,20,male
102,19,female
103,21,male
```

Comments

Relation Name

Attributes

Data Instances

Experiment 3: Demonstration of creating a Student dataset (student.arff) using WEKA tool in Data Mining.

Aim:

Demonstration of creating a Student dataset (student.arff) using WEKA tool in Data Mining.

Solution :

Creating a student dataset (student.arff):

Description:

We need to create a Student Table with training data set which includes attributes like sid, name, various subject marks, total and result.

Procedure:

Steps:

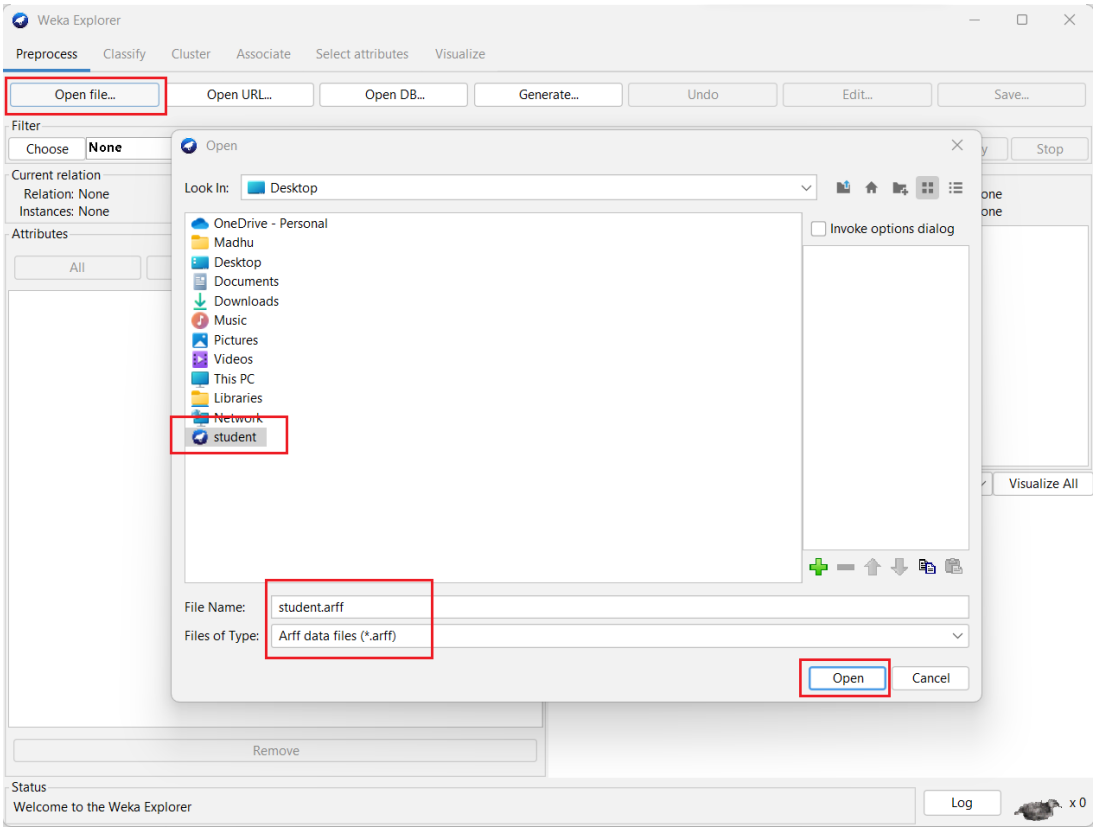
1) Open any **text editor** (e.g. Notepad)

2) Type the following training **data set** in the Notepad.

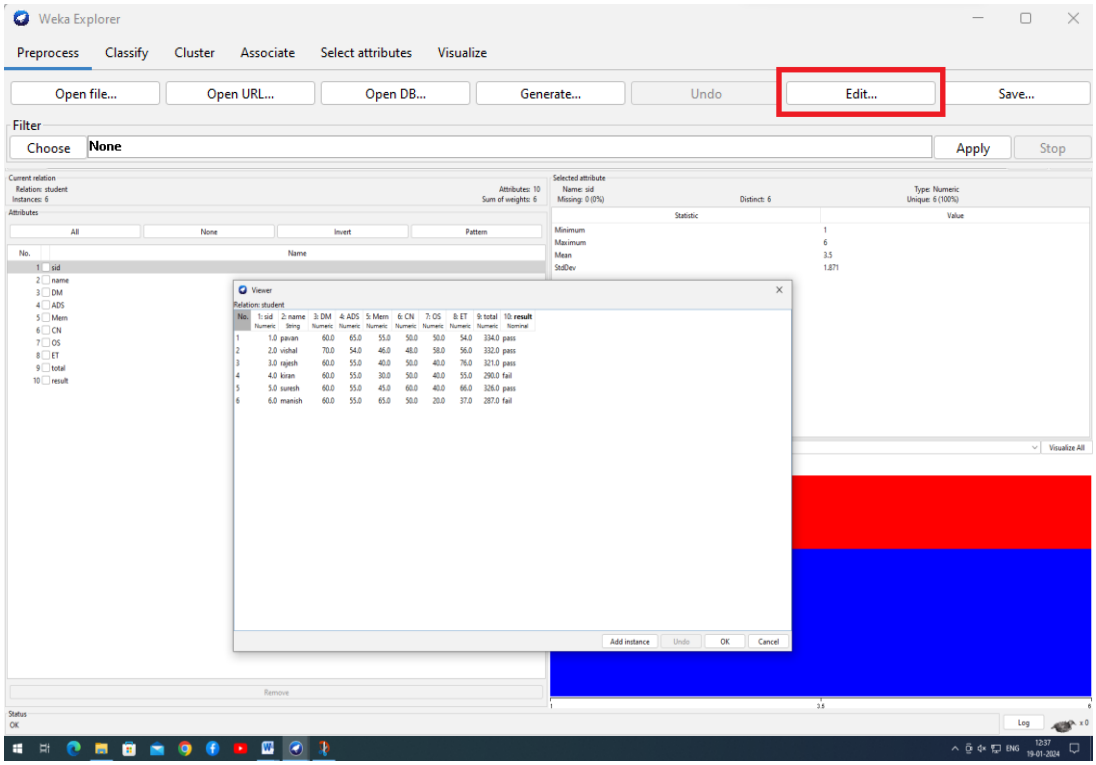
```
@relation student

@attribute sid numeric
@attribute name string
@attribute DM numeric
@attribute ADS numeric
@attribute MERN numeric
@attribute CN numeric
@attribute OS numeric
@attribute ET numeric
@attribute total numeric
@attribute result {pass, fail}

@data
1, pavan, 60, 65, 55, 50, 50, 54, 334, pass
2, vishal, 70, 54, 46, 48, 58, 56, 332, pass
3, rajesh, 60, 55, 40, 50, 40, 76, 321, pass
4, kiran, 60, 55, 30, 50, 40, 55, 290, fail
5, suresh, 60, 55, 45, 60, 40, 66, 326, pass
6, manish, 60, 55, 65, 50, 20, 37, 287, fail
```

3) After that save the file with **.arff file** format.
4) Minimize the arff file and then open **Start → Programs → weka-3.8.6.**
5) Click on **weka-3.8.6**, then **Weka GUI chooser** is displayed on the screen.
6) In that **Weka GUI chooser** there are five applications, click on **Explorer**.

7) Explorer shows many options. In that click on **'Open file...'** button and select the **.arff file** (e.g. student.arff).



8) Click on **edit** button which shows student table on weka.



Result:

The **student dataset** (student.arff) was created successfully using WEKA tool kit.