02/09/2020 **Mohd. Aftab Alam**
02013302717

EXPERIMENT 1

Aim: Introduction to WEKA.

Introduction: Named after a flightless New Zealand bird, Weka is a set of machine learning algorithms that can be applied to a data set directly, or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

Machine learning is nothing but a type of artificial intelligence which enables computers to learn the data without help of any explicit programs. Machine learning systems crawl through the data to find the patterns and, when these are found, adjust the program's actions accordingly.

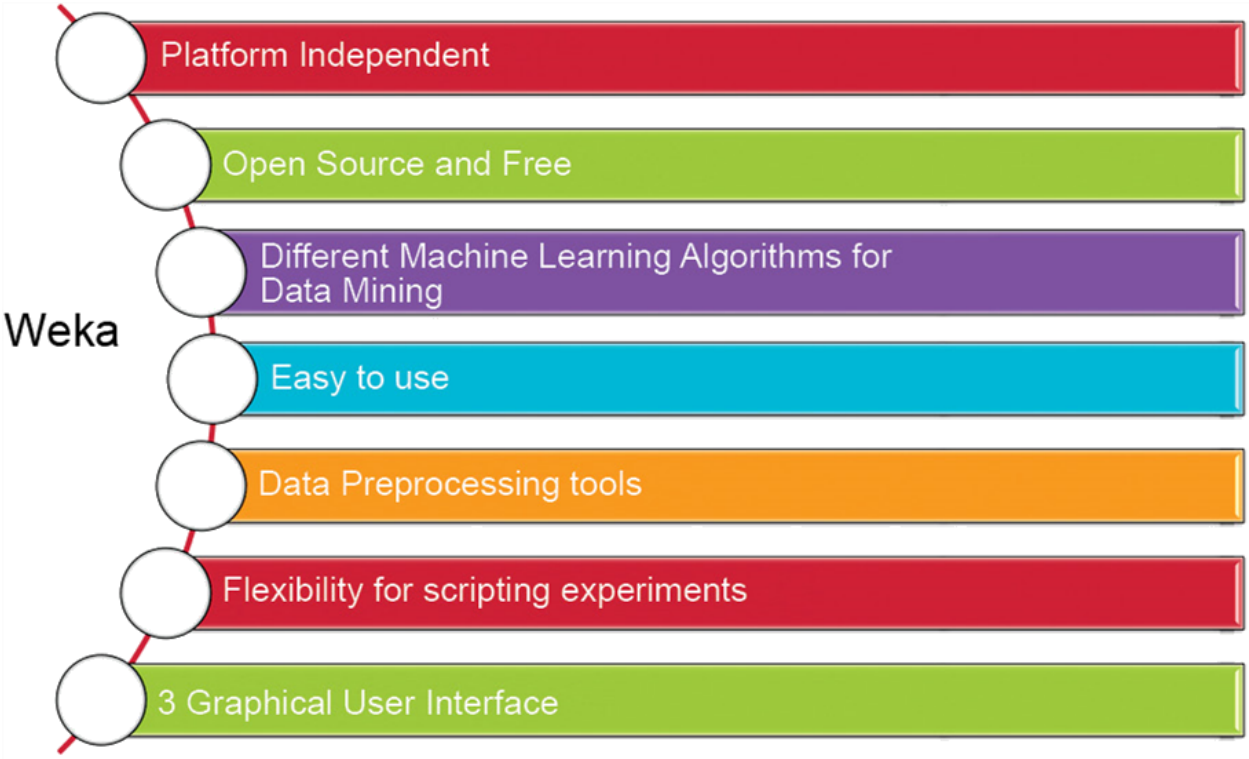
Data mining analyses the data from different perspectives and summarizes it into parcels of useful information. The machine learning method is similar to data mining. The difference is that data mining systems extract the data for human comprehension.

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

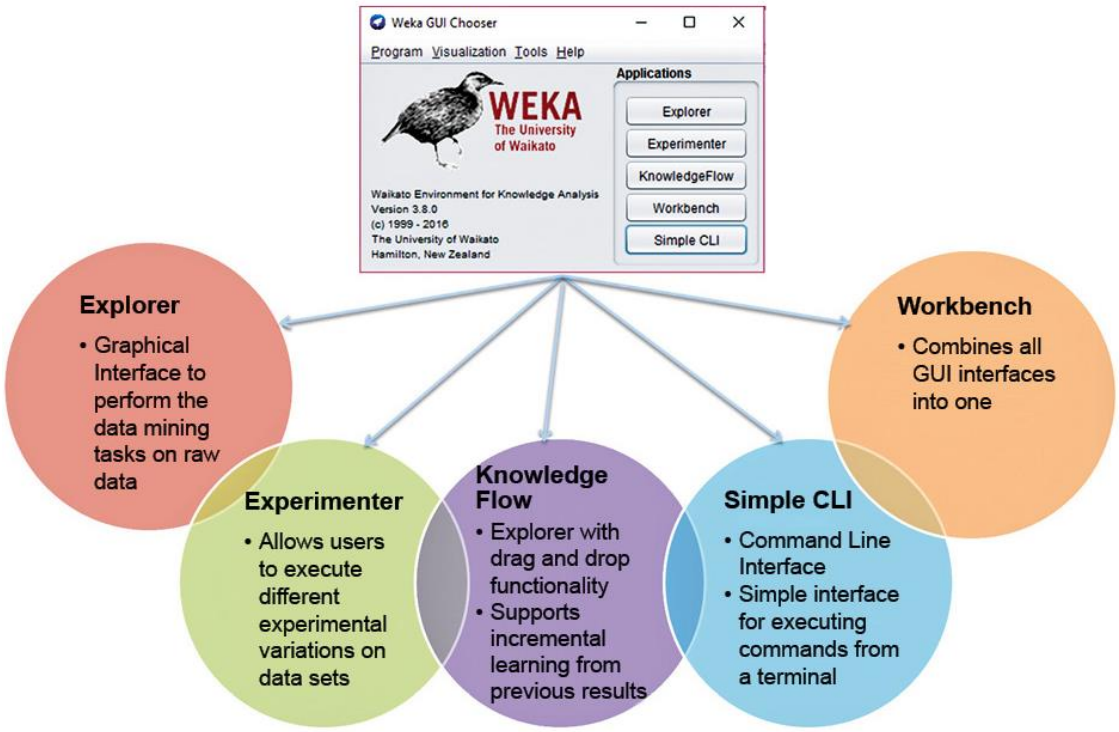
Weka is a collection of tools for:

1. Regression
2. Clustering
3. Association
4. Data pre-processing
5. Classification
6. Visualization

Features of Weka:



Weka's application interfaces:



Installation of Weka:

You can download Weka from the official website http://www.cs.waikato.ac.nz/ml/weka/.

Execute the following commands at the command prompt to set the Weka environment variable for Java, as follows:

setenv WEKAHOME /usr/local/weka/weka-3-0-2

setenv CLASSPATH $WEKAHOME/weka.jar:$CLASSPATH

Once the download is completed, run the exe file and choose the default set-up.

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:

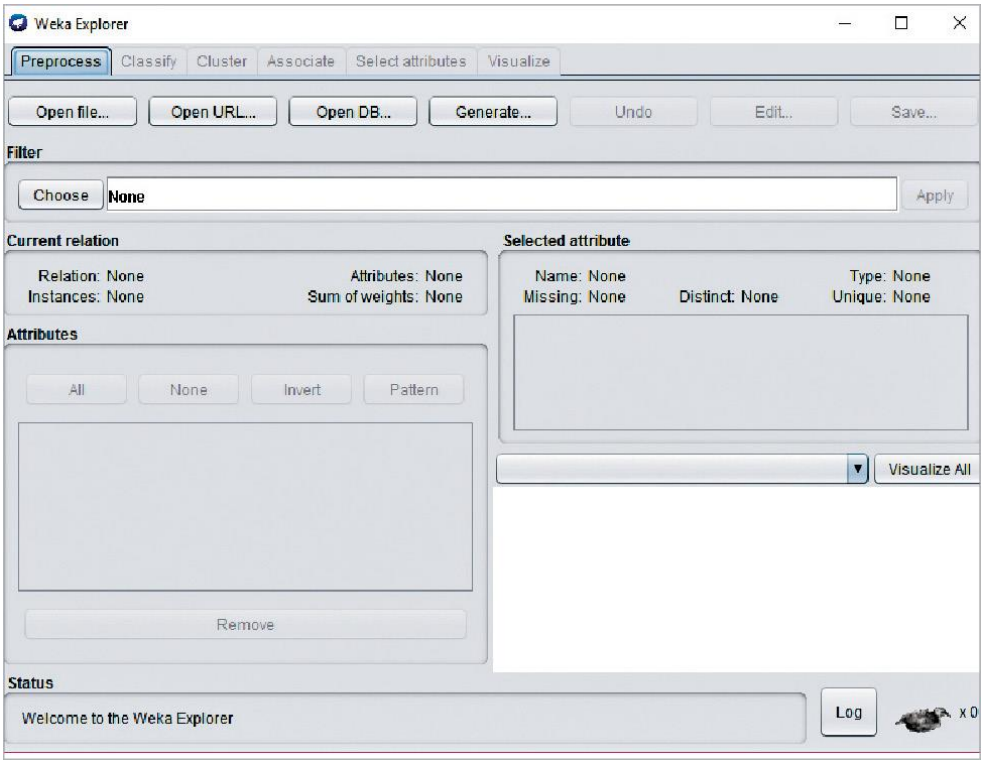
- CSV (Comma Separated Values)
- ARFF
- Database using ODBC

02/09/2020 **Mohd. Aftab Alam**
02013302717

Weka Explorer:

The Weka Explorer contains a total of six tabs.

1. Preprocess: This allows us to choose the data file.
2. Classify: This allows us to apply and experiment with different algorithms on preprocessed data files.
3. Cluster: This allows us to apply different clustering tools, which identify clusters within the data file.
4. Association: This allows us to apply association rules, which identify the association within the data.
5. Select attributes: These allow us to see the changes on the inclusion and exclusion of attributes from the experiment.
6. Visualize: This allows us to see the possible visualization produced on the data set in a 2D format, in scatter plot and bar graph output.



Pre-processing:

Data pre-processing is a must. There are three ways to inject the data for pre-processing:

- Open File – enables the user to select the file from the local machine
- Open URL – enables the user to select the data file from different locations
- pen Database – enables users to retrieve a data file from a database source

Classification:

To predict nominal or numeric quantities, we have classifiers in Weka. Available learning schemes are decision-trees and lists, support vector machines, instance-based classifiers, logistic regression and Bayes' nets.

02/09/2020 **Mohd. Aftab Alam**
02013302717

Clustering:

The cluster tab enables the user to identify similarities or groups of occurrences within the data set. Clustering can provide data for the user to analyse.

Association:

The only available scheme for association in Weka is the Apriori algorithm. It identifies statistical dependencies between clusters of attributes, and only works with discrete data. The Apriori algorithm computes all the rules having minimum support and exceeding a given confidence level.

Attribute selection:

Attribute selection crawls through all possible combinations of attributes in the data to decide which of these will best fit the desired calculation which subset of attributes works best for prediction. The attribute selection method contains two parts.

- Search method: Best-first, forward selection, random, exhaustive, genetic algorithm, ranking algorithm
- Evaluation method: Correlation-based, wrapper, information gain, chi-squared

Visualisation:

The user can see the final piece of the puzzle, derived throughout the process. It allows users to visualise a 2D representation of data, and is used to determine the difficulty of the learning problem. We can visualise single attributes (1D) and pairs of attributes (2D), and rotate 3D visualisations in Weka.