Practical No. 8

Title: Data Mining Tools

Aim: Study of Data Mining tools using WEKA / ORANGE

Software required: Orange, Weka

Theory :-

Data Mining is the set of techniques that utilize specific algorithms, statical analysis, artificial intelligence, and database systems to analyze data from different dimensions and perspectives.

Data Mining tools have the objective of discovering patterns/trends/groupings among large sets of data and transforming data into more refined information.

Orange :-

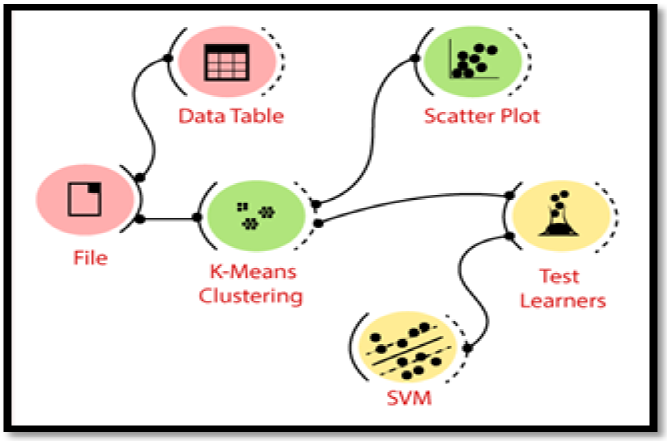
- Orange is a framework for data visualization, machine learning, and data mining with a front-end for visual programming.
- It has been around since 1996 and is free software. The analysis is achieved by connecting widgets that perform various functions, such as reading files, displaying statistics on features, constructing models, evaluating, etc.
- Moreover, if you intend to dig deeper into finer tuning, it is available as a Python library. For programmers, analysts, and data mining experts, Orange supports a versatile domain. Python, a scripting language and programming environment of the modern century, where our data mining scripts can be simple but efficient.
- For easy implementation, Orange uses a component-based method. Simply like placing the Wooden blocks, or even using an existing algorithm, we can apply our research technique.
- Orange is a great software package for machine learning and data mining

Widgets offer essential functionality, like:
Displaying data table and allowing to select features

1) Data reading

2) Training predictors and comparison of learning algorithms

3) Data element visualization, etc.



Advantages :

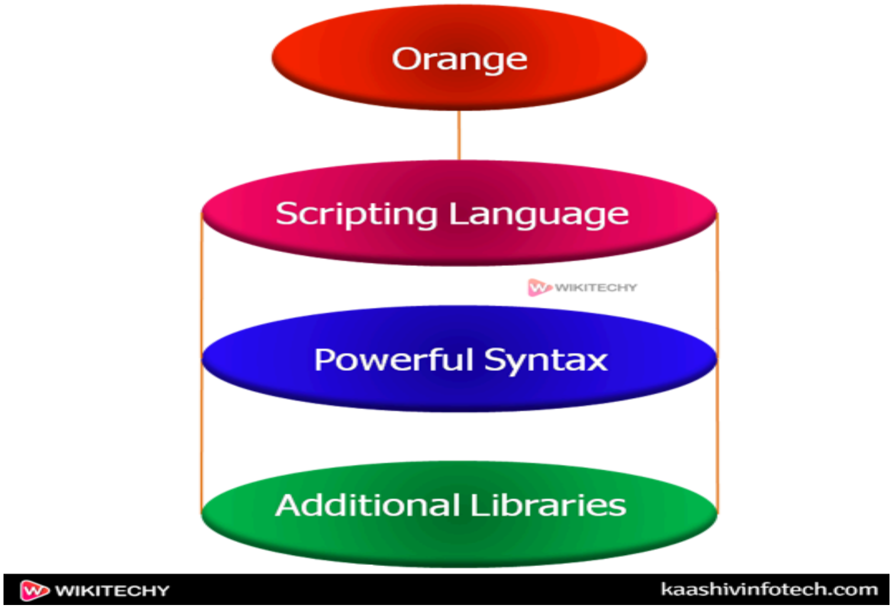
1. Open-source software is cost-effective.
2. Constant improvements are a hallmark of open-source software.
3. Visual Programming
4. Interactive Data Visualization
5. Add-ons Extended Functionality

Disadvantages :

1. Open-source software might not stick around.
2. Manual Troubleshooting
3. Advance analysis is not so easy
4. Support isn’t always reliable.
5. Security becomes a major issue.

- **Orange scripting**:

If we want to access Orange objects, then we need to write our components and design our test schemes and machine learning applications through the script. Orange interfaces to Python, a model simple to use a scripting language with clear and powerful syntax and a broad set of additional libraries. Same as any scripting language, Python can be used to test a few ideas mutually or to develop more detailed scripts and programs.



- Orange interfaces to Python, model simple to use a scripting language with clear and powerful syntax and broad set of additional libraries.

```
import orange

data1 = orange.ExampleTable('voting.tab')

print('Instance:', len(data1))

print(Attributes:', 1len(data.domain.attributes))
```

If we store this script in script.py and run it by shell command "python script.py" ensure that the data file is in the same directory then we get

```
 Instances: 543

 Attributes: 16
```

Let us proceed with our script that uses the same data created by a naïve Bayesian classifier and print the classification of the first five instances:

```
model = orange.BayesLearner(data1)

for i in range(5):

print(model(data1[i]))
```

It is easy to produce the classification model; we have called Oranges object (Bayes Learner) and gave it the data set. It returned another object (naïve Bayesian classifier) when given an instance returns the label of the possible class.

Output

```
inc

inc

inc

bjp

bjp
```

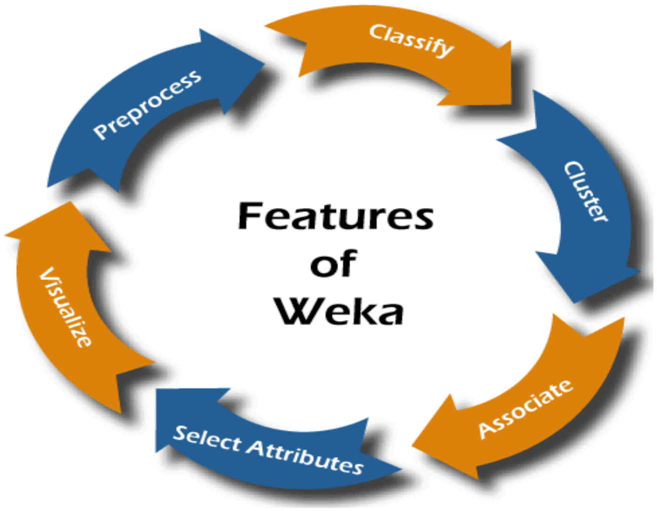
- **WEKA :-**

  Weka contains a collection of visualization tools and algorithms for data analysis and predictive modelling, together with graphical user interfaces for easy access to these functions. The original non-Java version of Weka was a Tcl/Tk front-end to (mostly third-party) modelling algorithms implemented in other programming languages, plus data preprocessing utilities in C and a makefile-based system for running machine learning experiments.

  Weka supports several standard data mining tasks, specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection. Input to Weka is expected to be formatted according to the Attribute-Relational File Format and filename with the .arff extension.

- **Features of Weka**

  Weka has the following features, such as:



1. Preprocess

The preprocessing of data is a crucial task in data mining. Because most of the data is raw, there are chances that it may contain empty or duplicate values, have garbage values, outliers, extra columns, or have a different naming convention. All these things degrade the results.

2. Classify

Classification is one of the essential functions in machine learning, where we assign classes or categories to items. The classic examples of classification are: declaring a brain tumour as "malignant" or "benign" or assigning an email to a "spam" or "not_spam" class.

After selecting the desired classifier, we select test options for the training set. Some of the options are:

- **Use training set:** the classifier will be tested on the same training set.
- **A supplied test set:** evaluates the classifier based on a separate test set.
- **Cross-validation Folds:** assessment of the classifier based on cross-validation using the number of provided folds.
- **Percentage split:** the classifier will be judged on a specific percentage of data.

3. Cluster

In clustering, a dataset is arranged in different groups/clusters based on some similarities. In this case, the items within the same cluster are identical but different from other clusters. Examples of clustering include identifying customers with similar behaviours and organizing the regions according to homogenous land use.

4. Associate

Association rules highlight all the associations and correlations between items of a dataset. In short, it is an if-then statement that depicts the probability of relationships between data items. A classic example of association refers to a connection between the sale of milk and bread.

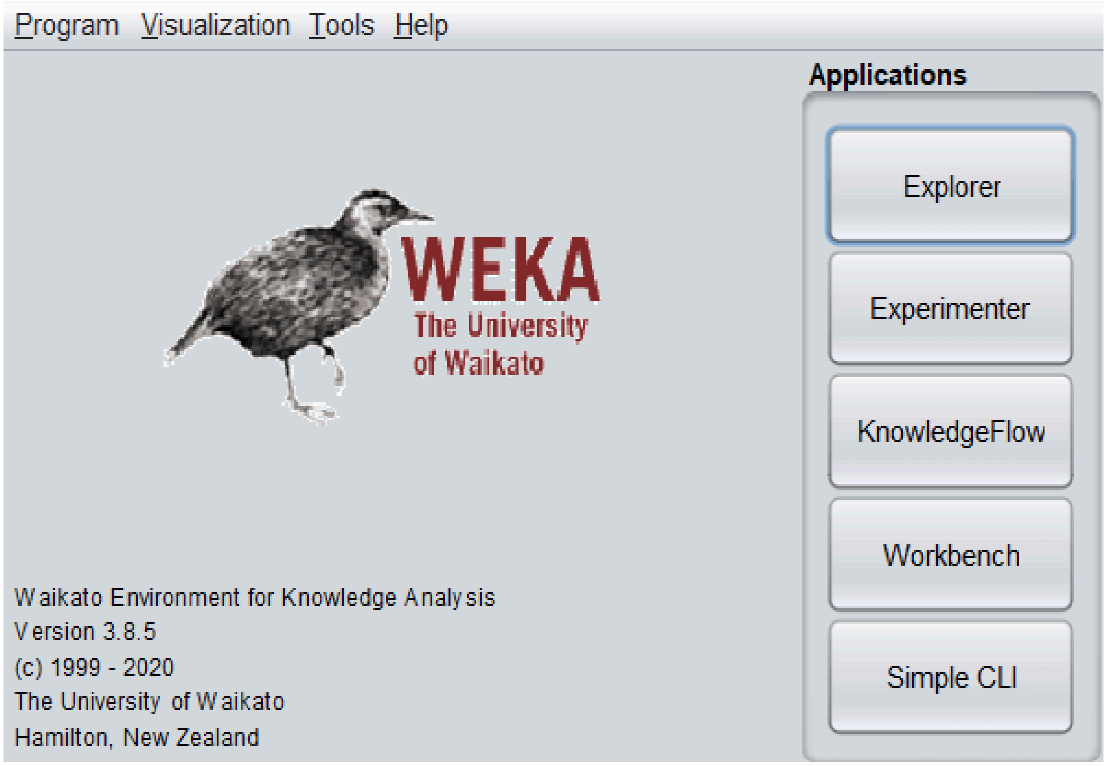
5.Select Attributes

Every dataset contains a lot of attributes, but several of them may not be significantly valuable. Therefore, removing the unnecessary and keeping the relevant details are very important for building a good model.

Many attribute evaluators and search methods include **BestFirst, GreedyStepwise**, and **Ranker**.

6. **Visualize**

In the visualize tab, different plot matrices and graphs are available to show the trends and errors identified by the model.



Waikato Environment for Knowledge Analysis
Version 3.8.5
(c) 1999 - 2020
The University of Waikato
Hamilton, New Zealand

As shown in the above screenshot, five options are available in the Applications category.

- The Exploreris the central panel where most data mining tasks are performed. We will further explore this panel in upcoming sections.
- The tool provides an Experimenter In this panel, we can run experiments and also design them.
- WEKA provides the KnowledgeFlow panel. It provides an interface to drag and drop components, connect them to form a knowledge flow and analyze the data and results.
- The Simple CLIpanel provides the command line powers to run WEKA. For example, to fire up the ZeroR classifier on the arff data, we'll run from the command line:

```
java weka.classifiers.trees.ZeroR -t iris.arff
```

- **Weka Datatypes and Format of Data -**

Numeric (Integer and Real), String, Date, and Relational are the only four datatypes provided by WEKA. By default, WEKA supports the ARFF format. The ARFF, attribute-relation file format, is an ASCII format that describes a list of instances sharing a set of attributes. Every ARFF file has two sections: header and data.

- The header section consists of attribute types,
- And the data section contains a comma-separated list of data for that attributes.

- **Types of Algorithms by Weka**

WEKA provides many algorithms for machine learning tasks. Because of their core nature, all the algorithms are divided into several groups. These are available under the Explorer tab of the WEKA. Let's look at those groups and their core nature:

- Bayes: consists of algorithms based on Bayes theorem like Naive Bayes
- functions: comprises the algorithms that estimate a function, including Linear Regression
- lazy: covers all algorithms that use lazy learning similar to KStar, LWL
- meta: consists of those algorithms that use or integrate multiple algorithms for their work like Stacking, Bagging
- misc: miscellaneous algorithms that do not fit any of the given categories
- rules: combines algorithms that use rules such as OneR, ZeroR
- trees: contains algorithms that use decision trees, such as J48, RandomForest

Conclusion :-

Data mining tools Orange and Weka are studied.

FAQ -

1)What are the types of data in Weka ?

2) What are the different types of classifiers in WEKA?

3)Can WEKA be used to make two predictions on a given set of data

4)What are the different evaluation methods available in Orange.

5)What are the key features of Orange

6)How does Orange handle missing values in data?