# Assignment 03

Review of Data Mining Software Packages

**Name: Dean D'souza**

**H.U. ID: 168424**

## Statistica Data Miner:

Statistica Data Miner (or simply Statistica) is an advanced analytical software package for data mining and analysis. It is one of the oldest data mining software packages initially developed by StatSoft in the 1980's. It has a proprietary license (with trial versions available) and is easily available for Windows.

Statistica comes packaged with a number of algorithms for performing data analysis and provides an interactive GUI for performing analysis through a number of menus, tabs and also provides a number of visualizations. It can work with a number of file formats such as flat files, .csv and can also access databases.

Statistica allows for univariate and bivariate analysis through simple statistics as well as visualizations such as histograms, bar charts, scatterplots, etc. with options for creating interactive plots for better representation of data.

Statistica also has one of the largest range of available algorithms for building predictive models through linear regression, Support Vector Machines, etc. as well as for clustering through k-means, X-means, K-medoids etc. It also has a selection of algorithms for building decision trees as well as Artificial Neural Networks and also has a number of options for performing Association analysis in order to obtain and apply Association rules.

Statistica also allows for an open architecture which can be used for custom extensions written in a variety of languages (such as C, C++, Java, R) for pre-processing data as well as for deploying models in.

Overall, Statistica has one of the most extensive selection of algorithms for use in data mining with significantly better performance over a variety of hardware. Additionally, it has some of the best customer support and online tutorials available.

## Weka:

WEKA (Waikato Environment for Knowledge Analysis) is a suite of machine learning algorithms and software used for data mining. It is written in java and was developed at the University of Waikato, New Zealand. Weka has an open-source GNU license and is readily available for download on Windows, Mac OS and on Linux distributions.

Weka allows for the use of packaged algorithms as well as algorithms written by the user in java. The source of data to be analyzed should usually be in the form of flat files (usually in the form of .arff files), as well as SQL database files through the use of the Java Database Connector (JDBC). Weka can also work on .csv format files but the use of such files is not advised due to complications that could occur during input and processing.

Weka provides an easy to use interface through its GUI Explorer which provides a number of tabs, buttons, drop down menus, radio buttons, etc. to simplify its use. It also has a number of pre-processing tools (known as “filters”) which allows for Discretization, normalization, resampling, attribute selection, etc.

Weka allows for univariate analysis through a number of basic statistics (minimum, maximum, standard deviation, etc.) and visual tools such as histograms and bar charts. It also allows for bivariate and multivariate analysis through a number of scatterplots and mosaic plots. It also allows for a level of interactivity of plots through the information on the side.

In terms of predictive models, Weka provides a number of pre-packaged algorithms which allows for classification through regression algorithms (such as linear regression), as well as clustering algorithms such as k-means clustering, X-means, etc. Weka also comes with Apriori algorithm implementation as well as algorithms for building Decision Trees and lists, Support Vector machines and Bayesian Networks. Weka also provides utilities for building Artificial Neural Networks through Multilayer perceptron algorithms.

Many of these algorithms can be easily run through the various tabs of classify, associate, cluster, etc. and selection of algorithms is done through the drop-down menus which have a hierarchical directory structure. For those more comfortable with command line tools, Weka also provides a simple command line interface for building models.

Overall, Weka is easy to use, has a number of features which provide a lot of functionality (which also includes advance features for machine learning) and has good performance for relatively large data sets with a good amount of documentation and customer support.

## RapidMiner:

RapidMiner is a software platform for machine learning, data mining, text mining, predictive analysis and business analytics. It is written in java and was developed by the company of the same name. RapidMiner has an AGPL which allows for free use (allowing it be called open-source) but prevents further distribution through third parties and also had a paid pro version which allows for better performance and more data to be loaded. It can be readily downloaded for Windows, Mac OS and Linux Distributions.

RapidMiner is based on a client/server architecture and so is supported by a cloud infrastructure for analyzing data and for better performance. It also has a marketplace for a number of plug-ins for more varied use and can also be used as an API to facilitate development of customized algorithms. It also provides support for using Weka and R scripts to extend functionality. RapidMiner mainly uses flat files of the .arff format but can also use .csv files and database files.

RapidMiner uses an extensive GUI for designing and executing analytical workflows, with the main focus being on a GUI that allows for drag and drop and point and click tools. It also provides a number of visualization tools such as bar charts, histograms, etc. Most of the useful information for data analysis can be found through various views of the model and data such as the meta-data view, data-view, plot-view and annotations.

RapidMiner provides univariate analysis of individual variables through a statistics column with options to obtain bar charts and histograms. It allows for bivariate analysis through scatterplots with sufficient amount of customization options such as jitter.

In terms of predictive models, RapidMiner can be easily used to create linear regression models, polynomial regression models and also supports clustering algorithms such as k-means, k-medoids, etc. It also supports various forms of Decision Trees (weighted, multi-way, etc.) as well as Bayesian networks and Support vector models.

RapidMiner also allows for easily creating and applying association rules as well as generalized sequential patterns and FP-growth algorithm. RapidMiner also provides a number of algorithms for building Artificial Neural Networks with ease. It also provides a number of utilities for cleaning data and validating models. For those who prefer a command line utility, RapidMiner provides an easy to use command line interface as well.

Overall, RapidMiner has a good selection of algorithms (both basic and advanced) which can be easily applied to the data with minimal code for the user. It also has better performance due to the cloud infrastructure and has better customer support and easily accessible developer documentation.

## R:

R is software environment and a programming language for statistical computing which provides a number of utilities for basic and advanced data mining and analysis. R is an implementation of the S programming language which is combined with lexical scoping semantics and was developed by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand. It is an open-source programming language which is readily available for Windows, Mac OS and Linux distributions.

Base R comes with a number of basic statistical algorithms pre-packaged but it can be easily extended with the help of a number of packages available freely. R can accept a number of file formats as input such as flat files, .csv files, as well as data base files and connections through appropriate functions and libraries.

While R does not come packaged with a GUI, one can easily download a number of options to complement base R based on their own preferences. One such popular IDE is RStudio which provides a good platform for working with R. R can also be extended with point and click type GUIs such as Rattle GUI for R.

Base R provides a number of functions to easily obtain statistical data for univariate and bivariate analysis, and visualization tools such as histograms, boxplots, bar charts, scatterplots, etc. Visualizations can also be further extended with the help of libraries such as ggplot2 and plotly which allows the user to build colorful and interactive graphics with a greater amount of control on the details of interaction. A number of libraries and templates are also present for easily preparing reports, storyboards, etc.

When it comes to predictive models, R comes with a number of options for building linear regression models and also has a number of sources for performing cluster analysis through k-means, k-medoids, etc. Association analysis can also be performed quite easily with the arules library to obtain association rules.

R also has a number of other libraries which can be easily used to build other types of models such as support vector machines, decision trees, artificial neural networks, etc. A number of these tasks can also be accomplished through GUIs for R such as building decision trees through Rattle GUI for R.

For the most part, however, R is more command line based and is easier to use for those used to command line utilities. Additionally, performance is limited by the RAM size on the computer on which the analysis is being performed, though it is mostly optimized and can easily work with relatively large data sets.

Overall, R provides an extensive list of algorithms and utilities not only for data analysis but also for data cleaning and visualization. While performance can be limited based on the size of the data, it can be easily optimized. Additionally, most of the documentation for R and its various libraries can be easily accessed and a number of solutions for various problems exist on various websites such as stackoverflow.

## Mlpy:

Mlpy is a machine learning library (a python module) which provides a wide range of machine learning methods for supervised and unsupervised machine learning. It is written in Python,C and C++, over the NumPy/SciPy and the GNUScientific libraries. It has an open source GNU version 3 license and can easily be installed and deployed on Windows, Mac OS and Linux distributions.

Unlike most of the software packages in this assignment, mlpy does not have an exact GUI and is mostly used the functions it provides in its libraries. Those familiar with python can easily use this library to perform machine learning tasks with ease. Due to it essentially being a python module, it can easily work with flat files, .csv format files and database files and connections. It also allows for easier modifications in python.

We can easily perform univariate and bivariate analysis using this library of functions and plots for the same can be obtained as histograms, bar charts, scatterplots, etc. provided that the matplotlib package has been installed. It also come with ready to use algorithms (in the form of functions) such as Ordinary Least squares, Ridge regression, Kernel Ridge Regression, etc. and also support algorithms for clustering such as hierarchical clustering and k-means clustering. It also has a number of algorithms for building support vector machines and for cross-validation.

However, it does not appear to have clearly defined functions for performing association analysis and building Artificial Neural Networks.

Overall, mlpy provides an extensive toolset of easy to use functions for python programmers (as well as those familiar with programming), to perform a variety of basic data mining tasks. Performance is limited by the machine on which this library is used. A good amount of documentation is also provided through the hosting site at sourceforge.net

## Conclusion:

The software packages/environments we have seen above have been extensively used by different kinds of data analysts (those used to point and click tools as well as those used to programming). However, they all come with their own quirks and limitations. Though most the above software

packages/environments differ only by a few degrees, these differences play an important role in deciding which of them should be used for a particular data set and situation. While selecting between these, I personally feel that R provides the most amount of intuitive use and flexibility for data mining as the analyst is forced to consider all aspects of the options available to him/her. Additionally, performing analysis as well as generating reports and presentations can be done through the same application, with a good amount of control on how these reports look aesthetically. R also gives a great number of options for creating some of the best interactive graphics and non-interactive graphics as well.

## References:


[1] "Statistica", Wikipedia, (https://en.wikipedia.org/wiki/Statistica)

[2] "STATISTICA Data Miner", statsoft, (http://www.statsoft.com/products/statistica/data-miner)

[3] "Weka Data Mining", Predictive Analytics Today, (http://www.predictiveanalyticstoday.com/weka-data-mining/)

[4] "Weka (machine learning)", Wikipedia, (https://en.wikipedia.org/wiki/Weka_(machine_learning))

[5] "RapidMiner", Predictive Analytics Today, (http://www.predictiveanalyticstoday.com/rapidminer/)

[6] "RapidMiner", Wikipedia, (https://en.wikipedia.org/wiki/RapidMiner)

[7] "R Software Environment", Predictive Analytics Today, (http://www.predictiveanalyticstoday.com/r-software-environment/)

[8] "R (programming language)", Wikipedia, (https://wn.wikipedia.org/wiki/R_(programming_language))

[9] "mlpy- Machine Learning Python", David Albanese, (http://mlpy.sourceforge.net)

[10] "Mlpy", Wikipedia, (https://en.wikipedia.org/wiki/Mlpy)