National Journal of Multidisciplinary Research and Development
ISSN: 2455-9040; Impact Factor: RJIF 5.22
Received: 06-02-2019; Accepted: 09-03-2019
www.nationaljournals.com
Volume 4; Issue 2; April 2019; Page No. 41-45

A comparative study of data analytics open source tools

Lakshit Sharma[1], Dr. Vijay Gupta[2]
[1] Student, MCA, International School of Informatics and Management, Jaipur, Rajasthan, India
[2] Associate Professor, International School of Informatics and Management, Jaipur, Rajasthan, India

Abstract
To prepare data for analysis, execute analytic algorithms and access the result, data analytic professionals are using a wide range of tools. With the change in time there have been increases in the depth and functionality of these tools. These tools are now having good richer user interfaces. Common tasks will be automated by these tools very easily. Due to these tools, data analytics professionals get more time to focus on analysis and get good results. Combination of new Tools and methods with the evolved scalability and process is helping for organization to tame Big Data. This paper describes a comparative study of open source tools used in data analytics in terms of interfaces, functionality and algorithms that can be used.

Keywords: data analytics, orange, Weka, R, Kmine, machine learning

1. Introduction
Data Analytics is the process of examining data sets in order to draw conclusions about the information they contain, increasingly with the aid of specialized system and software. Data Analytics techniques and technologies are mostly used in commercial industries and by scientists and researchers in order to enable organizations to make more inform business decisions and to verify or disapprove scientific models and theories. Data Analytics primarily refers to a mixture of Applications that comprises of basic business intelligence (BI), reporting and online analytical processing (OLAP) and various forms of advance analytics.

2. Open source tools for the data analytics
2.1 Orange
Orange is a component-based visual programming software package for data visualization, machine learning, data mining, and data analysis. Orange components are called widgets and they range from simple data visualization, subset selection, and preprocessing, to empirical evaluation of learning algorithms and predictive modeling. Orange is an open-source software package released under GPL. Versions up to 3.0 include core components in C++ with wrappers in Python are available on GitHub. From version 3.0 onwards, Orange uses common Python open-source libraries for scientific computing, such as Numpy, Scipy and Scikit-learn, while its graphical user interface operates within the cross-platform Qt framework. Orange3 has its own separate GitHub.

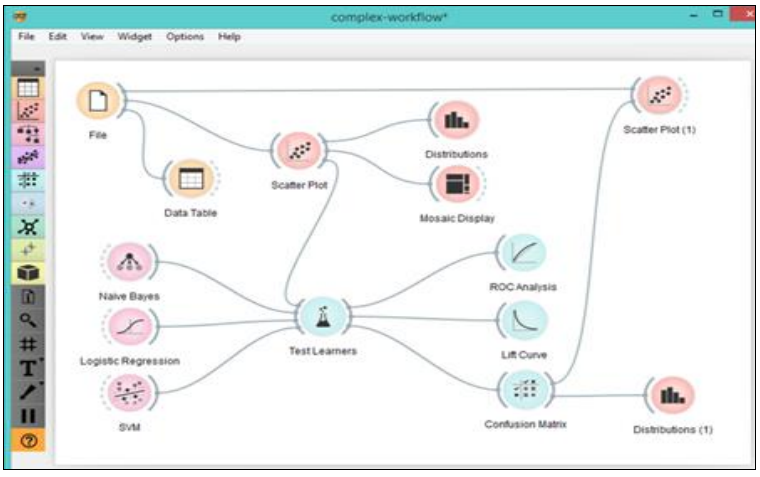

Fig 1

2.2 Rapid Miner
Rapid Miner is available in both FOSS and commercial editions and is a leading predictive analytic platform. Gartner, the US research and advisory firm, has recognized Rapid Miner and Knife as leaders in the magic quadrant for advanced analytic platforms in 2016.
Rapid Miner is helping enterprises embed predictive analysis in their business processes with its user friendly, rich library of data science and machine learning algorithms through its all-in-one programming environments like Rapid

Miner Studio. Besides the standard data mining features like data cleansing, filtering, clustering, etc., the software also features built-in templates, repeatable work flows, a professional visualization environment, and seamless integration with languages like Python and R into work flows that aid in rapid prototyping. The tool is also compatible with weak scripts.

Rapid Miner is used for business/commercial applications, research and education.

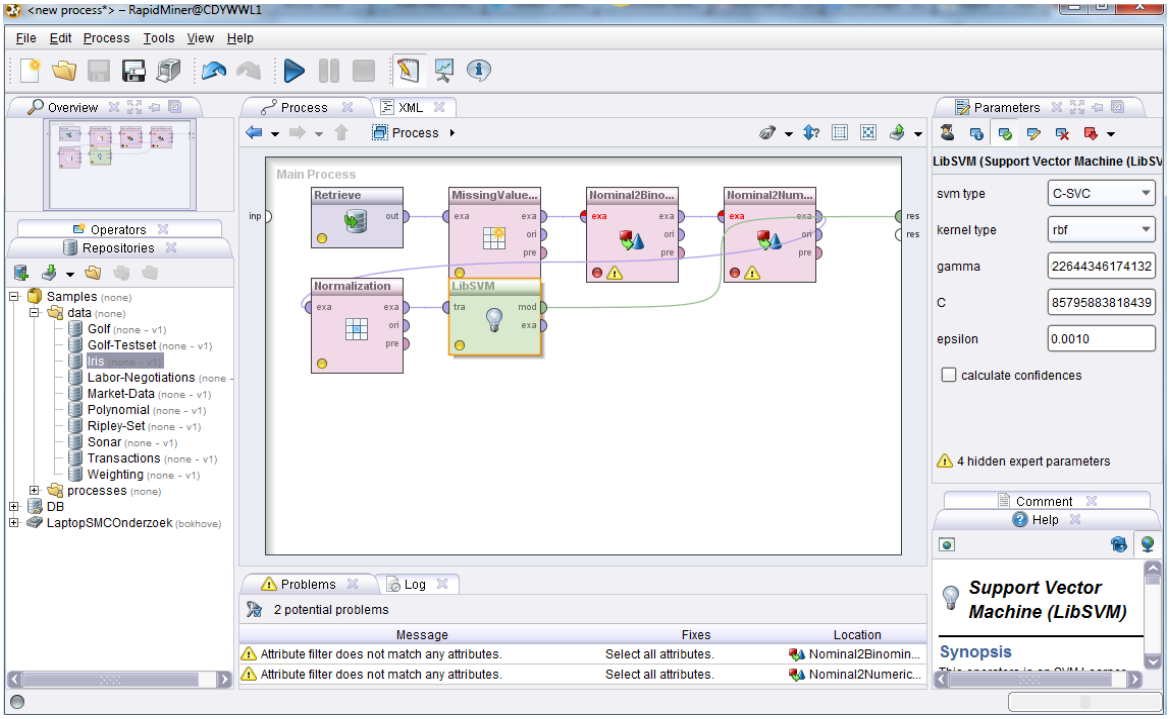

Fig 2

2.3 Knime

Knime is one of the leading open source analytic, integration and reporting platforms that comes as free software and as well as a commercial version. Written in Java and built upon Eclipse, its access is through a GUI that provides options to create the data flow and conduct data pre-processing, collection, analysis, modeling and reporting. A Gartner survey reveals that customers are happy with the platform's flexibility, openness and smooth integration with other software like Weka and R. Given the small size of the company, Knime has a large user base and an active community. It makes use of Eclipse's extension mechanism capability to add plugins for the required functionalities like text and image mining. This software is ideal for enterprise use.

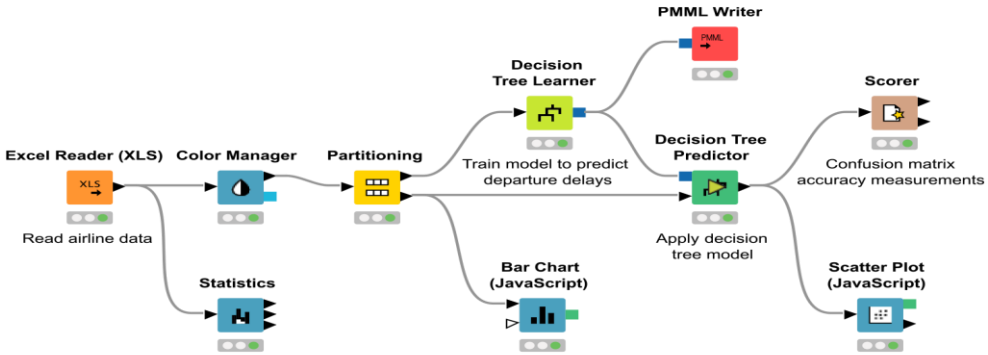

Fig 3

2.4 Weka

Weka is a Java based free and open source software licensed under the GNU GPL and available for use on Linux, Mac OS X and Windows. It comprises a collection of machine learning algorithms for data mining. It packages tools for data pre-processing, classification, regression, clustering, association rules and visualization. The various ways of accessing it are – Weka Knowledge Explorer, Experimenter, Knowledge Flow and a simple CL. This software also provides a Java Appetizer for use in applications and can connect to databases using CJD. Weka has proved to be an ideal choice for educational and research purposes, as well as for rapid prototyping.

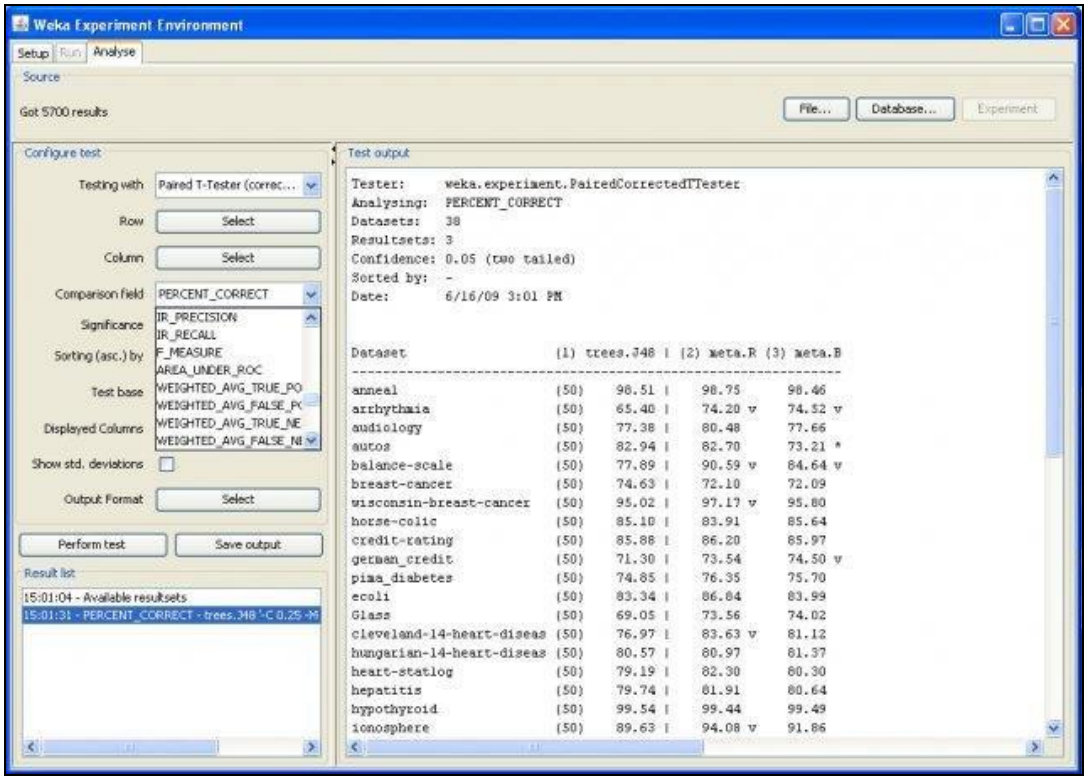

Fig 4: 2.5 R

R is a free and open source package for statistics and graphing. R is traditionally command line; however, there are many feely available open source tools that integrate into R. One such example is R Studio which provides a graphical user interface for R. R can be employed for a variety of statistical and analytics tasks including but not limited to clustering, regression, time series analysis, text mining, and statistical modeling.

R is considered an interpreted language more so than an environment. R supports big data processing with RHadoop. RHadoop connects R to Hadoop environments and runs R programs across Hadoop nodes and clusters. Natively, visual features are not available making creating workflows challenging, especially for a novice; still, its broad community provides many graphical utilities such as R Studio

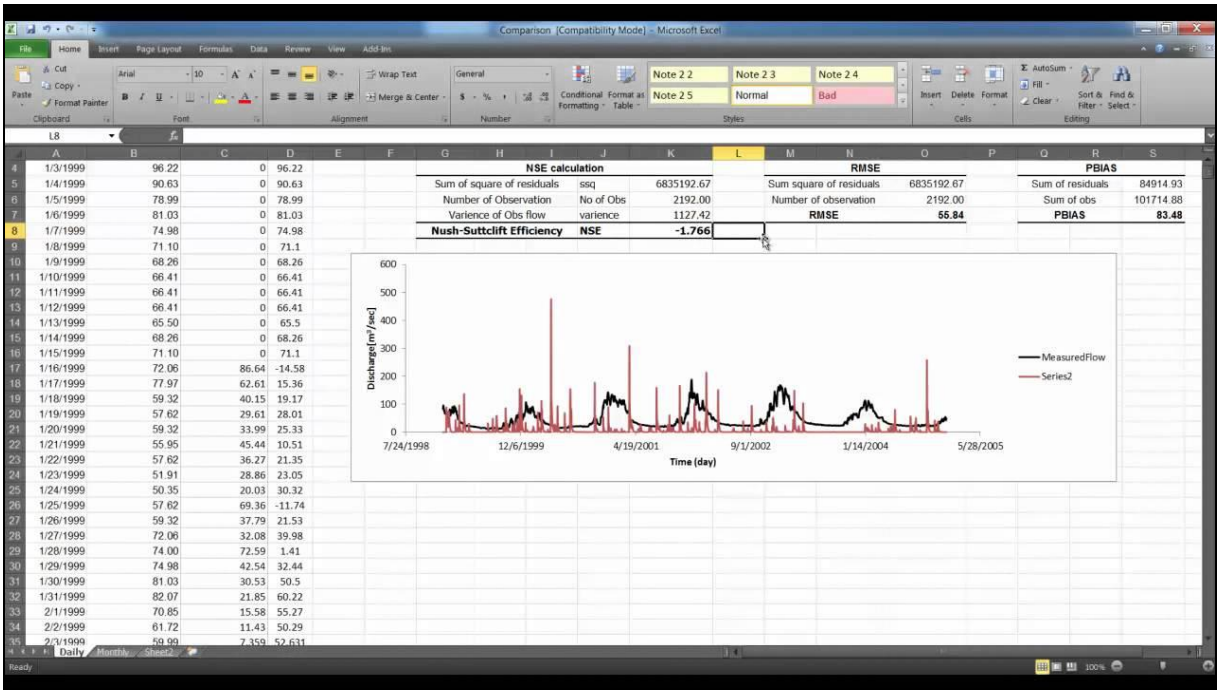

Fig 5

Comparison

1. Technical Overview of best data mining open source tools

Table 1: Technical Overview of best six data minim open source tools

S. No.	Tool Name	Release Date	Release date/ Latest version	License	Operating System	Language	Website
1	RAPID MINER	2006	21November, 2013 /6.0	AGPL Proprietary	Cross platform	Language Independent	www.rapidmine r.com
2	ORANGE	2009	6 May,2013/2.7	GNU General Public License	Cross Platform	Python C++,C	www.orange.bi olab.si
3	KNIME	2004	6December,2013/2.9	GNU General Public License	Linux, OS X, Windows	Java	www.knime.org
4	WEKA	1993	24 Apr11.2014/3.7.11	GNU General Public License	Cross Platform	Java	www.cs.waikat o.ac.nz/~ml/we ka
5	KEEL	2004	5 June,2010/2.0	GNU GPL v3	Cross Platform	Java	www.sci2s.ugr.es/keel
6	R.	1997	10 Apri1,2014/3.1.0	GNU General Public License	Cross Platform	C, Fortran and R	www.rproject.org

2. Analytics of feature of open source data mining tools

The given table describes the basic features and functionality of data mining tools i.e. Rapid miner, R, Weka, R, Keel and Orange.

Table 2: Analytics of feature of best six open source data mining tools

S. No.	Tool Name	Type	Features
1.	RAPID MINER	Statistical analysis, data mining. predictive analytics.	▪ More than 20 new functions for analysis and data handling, including multiple new aggregation functions ▪ File operators to operate directly from Rapid Miner ▪ A macro viewer that shows macros and their values in real time during process execution ▪ Intutive GUI
2.	ORANGE	Machine learning. Data mining, Data visualization	▪ Visual Programming. Visualization, ▪ Interaction And Data Analytics ▪ Large toolbox, Scripting interface ▪ Extendable Documentation
3.	KNIME	Enterprise Reporting, Business Intelligence, Data mining	▪ Scalability, Intuitive user interface, High extensibility ▪ well-defined API for phigin extensions ▪ sophisticated dab handling, intelligent automatic caching of data, Data visualization ▪ Import/export of workflows. Parallel execution on multi-core systems ▪ Command line version for "headless", "batch executions", Hilting,
4.	WEKA	Machine Learning.	▪ Forty nine data preprocessing tools, seventy six classification/regression algorithms, eight clustering algorithms. fifteen attribute/subset evaluators. ten search algorithms for feature selection. ▪ three algorithms for finding association rules ▪ three graphical user interfaces — "The Explorer" (exploratory data analysis) — "The Experimenter" (experimental environment) — "The Knowledge Flow" (new process model inspired. ▪ Poor documentation
5.	KEEL	Machine Learning	▪ Classification Discovery, Cluster Discovery, Regression Discovery, Association Discovery, Data Visualization, Discovery Visualization, a user-friendly graphical interface,evolutionary learning
6.	K	Statistical Computing	▪ Data Exploration, Outlier detection, Clustering, Text Mining. Time Series Analysis, Social Network Analysis, Parallel Computing. Graphics, Visualization of geo spatial data. Web Application Big data ▪ Data and error handlingrequires array lingua ge.poor mining.

Table 3

	Orange	Tanagra	Rapid Miner	KNIME	R	Weka
K-means Clustering	Yes	Yes	Yes	Yes	Yes	Yes
Association Rule Mining	Yes	Yes	Yes	Yes	Yes	Yes
Linear Regression	Yes	Yes	Yes	Yes	Yes	Yes
Logistic Regression	Yes	Yes	Yes	Yes	Yes	Yes
Naïve Bayesian Classifiers	Yes	Yes	Yes	Yes	Yes	Yes
Decision Tree	Yes	Yes	Yes	Yes	Yes	Yes
Time Series Analysis	No	No	Some	Yes	Yes	Yes
Text Analytics	Yes	No	Yes	Yes	Yes	Yes
Big Data Processing	No	No	No	No	Yes	Yes
Visual Work Flows	Yes	Yes	Yes	Yes	No	Yes

3. Conclusion

Data analytics plays a very important role in data mining techniques. This paper gives the comparative study of different open source tools which are used in data analytics. It also describes the functionality as well technical overview of these tools.

4. References

1. "Data Mining Curriculum". ACM SIGKDD, 2006.
2. Clifton, Christopher. "Encyclopedia Britannica: Definition of Data Mining". Retrieved, 2010.
3. Hastie Trevor, Tibshirani Robert, Friedman Jerome. "The Elements of Statistical Learning: Data Mining, Inference, and Prediction". Archived from the original, 2009.
4. Han Kamber, Pei Jaiwei, Micheline Jian. Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann, 2011.
5. Azevedo A, Santos MF, KDD SEMMA, CRISP-DM. A parallel overview Archived 2013-01-09 at the Wayback Machine. In Proceedings of the IADIS European Conference on Data Mining, 2008, pp182-185.
6. Hawkins, Douglas M. "The problem of overfitting". Journal of Chemical Information and Computer Sciences. 2004; 44(1):1-12. doi:10.1021/ci0342472. PMID 14741005.
7. Microsoft Academic Search: Top conferences in data mining". Microsoft Academic Search.
8. http://opensourceforu.com/2017/03/top-10-open-source-data-mining-tools/
9. https://thenewstack.io/six-of-the-best-open-source-data-mining-tools/
10. http://blog.galvanize.com/four-data-mining-techniques-for-businesses-that-everyone-should-know/
11. http://www.rdatamining.com/resources/tools