

# A Review on Tools for Data Mining Application in the Diagnosis of Psychiatric Diseases

**Shivangi Jain**

Department of Computer Science and Engineering  
Bhabha Engineering Research Institute  
Bhopal, India [shivangijainap@gmail.com](mailto:shivangijainap@gmail.com)

**Dr. Mohit Gangwar**

Department of Computer Science and Engineering  
Bhabha Engineering Research Institute  
Bhopal, India [mohitgangwar@gmail.com](mailto:mohitgangwar@gmail.com)

**Abstract:** Data mining and processing large amount of data captured by various resources produce many outputs. Several research direction and solutions to the previously arise problems can understand by the large amount of past study data. Various resource such as UCI, different medical institutions, various medical readings and devices produce large amount of data. In order to understand the large data there are tools for mining their knowledge. Data extraction, mining knowledge from that is an important task which gives output for utilization. There are different tools such as Weka tool, Rapid miner etc are the best effort tool which provide better understanding of large data in visualize form. In this paper a study about the different data mining tools , which can be used for psychiatric disease disorder mining is studied. Different tools discuss about their features and advantages to process large medical data. Further a comparison between the tools among is presented. As per study it is described that using tools is very much effective and easy to process understanding large amount of data.

**Keywords:** Weka Tool, Rapid Miner, Orange, Knime, Data Melt, Apache Mahout, ELKI, MOA, KEEL, RATTLE

## INTRODUCTION

In the recent years, a huge amount of data is being gathered and stored in databases everywhere across the globe, which is mainly coming from information industry and social networking sites. There is a need to extract and classify useful information and knowledge from such a data collected. Data mining is an interdisciplinary field of computer science and is referred to as extracting or mining knowledge from large databases. It is the process of performing automated extraction and generating the predictive information from a large database. It is actually the process of finding the hidden information or patterns from the repositories.

The fields that use Data mining techniques include medical research, marketing, telecommunication, and stock markets, health care and so on. Data mining consists of the various technical approaches including machine learning, statistics, database system etc. The goal of the data mining process is to discover knowledge from large databases and transform into a human understandable format. The DM and knowledge discovery are essential

components to the organization due to its decision making strategy.

## TOOLS & TECHNIQUES

In this section detailing about the various data mining tools discussed, which gives their advantage in finding psychiatric disease disorders.

### WEKA

Weka is a landmark system in the history of the data mining and machine learning research communities, because it is the only toolkit that has gained such widespread adoption and survived for an extended period of time. The Weka or woodhen (*Gallirallus australis*) is an endemic bird of New Zealand. It provides many different algorithms for data mining and machine learning. Weka is open source and freely available. It is also platform-independent.

Weka is a collection with algorithms of machine learning used for data mining approaches. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

Clustering is a process of grouping a set of similar data objects within the same group based on similarity criteria (i.e. based on a set of attributes). There are many clustering algorithms. The objective of this thesis is to perform a comparative analysis of four clustering algorithms namely K-means algorithm, Hierarchical algorithm, Expectation and maximization algorithm and Density based algorithm. These algorithms are compared in terms of efficiency and accuracy, using WEKA

tool. The data for clustering is used in normalized and as well as unnormalised format. In terms of efficiency and accuracy K-means produces better results as compared to other algorithms.

### WEKA TOOL

WEKA is Waikato Environment for Knowledge Analysis. It's a data mining/machine learning tool developed by Department of Computer Science, University of Waikato, New Zealand. Weka is also a bird found only on the islands of New Zealand.

#### Features of WEKA:

- 49 data preprocessing tools
- 76 classification/regression algorithms
- 8 clustering algorithms
- 3 algorithms for finding association rules
- 15 attribute/subset evaluators + 10 search algorithms for feature selection

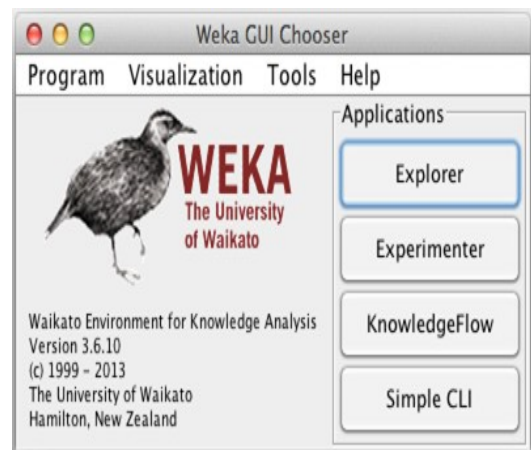


Figure: WEKA Explorer

The GUI Chooser consists of four buttons:

1. Explorer: An environment for exploring data with WEKA.

2. **Experimenter:** An environment for performing experiments and conducting statistical tests between learning schemes.
3. **Knowledge Flow:** This environment supports essentially the same functions as the Explorer but with a drag-and-drop interface. One advantage is that it supports incremental learning.
4. **Simple CLI:** Provides a simple command-line interface that allows direct execution of WEKA commands for operating systems that do not provide their own command line interface.

## RAPID MINER

Rapid Miner is a data science software platform developed by the company of the same name that provides an integrated environment for data preparation, machine learning, deep learning, text mining, and predictive analytics. ... Rapid Miner is developed on an open core

model. Rapid Miner is a centralized solution that features a very powerful and robust graphical user interface that enables users to create, deliver, and maintain predictive analytics. Aside from allowing users to create very advanced workflows, Rapid Miner features scripting support in several languages. With Rapid Miner, the whole process of modelling to implementation is unhindered. The system also boasts of rich technology that is useful when working in the various stages of an advanced analytic project.

## HISTORY

Rapid Miner, formerly known as YALE (Yet Another Learning Environment), was developed starting in 2001 by Ralf Klinkenberg, Ingo Mierswa, and Simon Fischer at the Artificial Intelligence Unit of the [Technical University of](http://www.tu-bs.de)

[Dortmund](http://www.tu-bs.de).<sup>[3]</sup> Starting in 2006, its development was driven by Rapid-I, a company founded by Ingo Mierswa and Ralf Klinkenberg in the same year.<sup>[4]</sup> In 2007, the name of the software was changed from YALE to Rapid Miner. In 2013, the company rebranded from Rapid-I to Rapid Miner.

## FEATURES

1. Graphical user interface
2. Analysis processes design
3. Multiple data management methods
4. Data from file, database, web, and cloud services
5. In-memory, in-database and in-Hadoop analytics
6. Application templates
7. -D graphs, scatter matrices, self-organizing map
8. GUI or batch processing
9. Integrates with in-house databases
10. Interactive, sharable dashboards
11. Big Data predictive analytics
12. Remote analysis processing
13. RapidMiner Radoop add-on
14. RapidMiner Streams add-on
15. CConnect to Apache Kafka and Amazon SQS
16. Data filtering, merging, joining and aggregating
17. Build, train and validate predictive models
18. Run more than 1500 operations
19. Store streaming data to numerous databases
20. Reports and triggered notifications

## ADVANTEGES

Multiple data management methods: data loading, data transformation, data modeling, and data visualization methods  
Works with multiple data sources: Excel, Access, Oracle, IBM DB2, Microsoft SQL, Sybase, Ingres, MySQL, Postgres, SPSS, dBase, Text files, and more  
Brand-new templates: Including churn

reduction, sentiment analysis, predictive maintenance and direct marketing Runs on every major platform and operating system Run more than 1500 operations. From data partitioning, to market-based analysis, to attribute generation. RapidMiner Radoop can connect to many different Hadoop clusters: Cloudera Distribution including Apache Hadoop (CDH), the Hortonworks Data Platform (HDP), Apache Hadoop with Hive, Amazon Elastic Map Reduce, MapR Hadoop, and DataStax Enterprise Data Storage: Store streaming data and the results of your analytics to numerous databases including Cassandra, MongoDB, Redis, Apache Solr and others.

### Limitations

1. Rapid Miner Studio is a visual data science workflow designer that accelerates the building of complete analytic workflows. It's code-optional with guided analytics, predefined connections, built-in templates, and repeatable workflows.
2. It contains a rich library of algorithms and functions to build the strongest possible model for any use case. Over 1500+ built-in operations (predefined functions).
3. Rapid Miner Studio is open and extensible with the industry's leading user community marketplace. Leverage expertise & best practices of 150,000+ users, and easily integrate existing R and Python code.

### CONCLUSION

Data mining is an important research topic which deals with real time scenario and examples in the world. Every aspect of dealing with data such as medical, social or people around the world, they deal with

data mining and its application. Medical science gets a large data from different resource such as machinery, doctors report etc. There are tools which help in understanding of the large produced data. This paper help in understanding various tools and their advance feature which help in recognition pattern. Detecting psychiatric disorder from the dataset production. This paper presented large number of popular avail tool in open source and paid platform. A study can determine that the visualization and mining of knowledge extraction gives better understanding of disease and its diagnosis. Data mining technique make big advantage on data analysis. A further work to study psychiatric disorder analysis with the selected and well features weka tool, which is open source and more graphical usage. A study analysis using weka with psychiatric data is left for the future work analysis.

### REFERENCES

1. Chauhan R, kaur H Alam MA, "Data Clustering Method Discovering Clusters in Spatial Cancer Database," International Journal of Computer Application, (0975-8887) Vol.10-no.6, November 2010.
2. Data Pre-processing in WEKA, Available at: <http://facweb.cs.depaul.edu/mobasher/classes/ect584/weka/preprocess.html>.
3. <https://rapidminer.com/products/studio/feature-list/>.
4. Introduction to WEKA available at : <https://rapidminer.com/products/studio/feature-list/>.
5. Introduction to Rapid Miner available at: <https://reviews.financesonline.com/p/rapidminer/>



6. History:  
[https://en.wikipedia.org/wiki/Rapid Miner](https://en.wikipedia.org/wiki/Rapid_Miner)
7. Features and advantages available at:  
<https://www.getapp.com/business-intelligence-analytics-software/a/rapidminer/>
8. Limitations:  
<https://rapidminer.com/resource/rapidminer-pros-cons/>