A COMPARATIVE STUDY ON DATA MINING TOOLS

A Project

Presented to the faculty of the Department of Computer Science

California State University, Sacramento

Submitted in partial satisfaction of

the requirements for the degree of

MASTER OF SCIENCE

in

Computer Science

by

Akshay Vishwanath Bhinge

FALL

2015

© 2015

Akshay Vishwanath Bhinge

ALL RIGHTS RESERVED

A COMPARATIVE STUDY ON DATA MINING TOOLS

A Project

by

Akshay Vishwanath Bhinge

Approved by:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_, Committee Chair

Dr. Meiliu Lu, Ph.D.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_, Second Reader

Prof. Bob Buckley

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Date

Student: Akshay Vishwanath Bhinge

I certify that this student has met the requirements for format contained in the University format manual, and that this project is suitable for shelving in the Library and credit is to be awarded for the project.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_, Graduate Coordinator \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Jinsong Ouyang, Ph.D. Date

Department of Computer Science

Abstract

of

A COMPARATIVE STUDY ON DATA MINING TOOLS

by

Akshay Vishwanath Bhinge

In a data mining course such as CSc 177 (Data Warehouse and Data Mining) at California State University, Sacramento, the students need to work on a term project. The students learn to implement various classification and clustering algorithms [[1]](#footnote-1)[1]. Part of the course requirements is to implement one or more data mining tools to complete the data mining project. This project allows students to explore various data mining algorithms using different open source data mining tools. These data mining tools are very useful to predict valuable information and gain knowledge based on the input data sets. A number of popular data mining and data visualization tools like Weka [2], Rapid Miner [3], R [4] and Tableau [5] provide a free and an efficient way to implement the various data mining algorithms.

One of the challenges for students is to select one particular tool for their term project within limited time and also learn how to use it quickly. It is important to know and understand how each algorithm is implemented in tools such as Rapid Miner and Weka. It takes a lot of time to understand the features and interpret the results. For example, both Rapid Miner and Weka have unique features and their own methods of data representation.

In this project, I developed a comparative study and presented it by designing a website that contains the following:

1. The output generated for Rapid Miner and Weka for a classification and clustering algorithm using the same data set simultaneously in a single window
2. A set of illustrations on how to use Tableau to solve a real quality health care problem
3. A set of useful resources which will help users to learn R
4. Comparative Study results added as a reference for students in the CSc 177 course website at California State University, Sacramento
5. Quizzes to check student understanding of the data mining tools

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_, Committee Chair

Dr. Meiliu Lu, Ph.D.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Date

# ACKNOWLEDGEMENTS

It gives me immense pleasure for getting an opportunity to work on my master’s project. It has helped me to learn a multitude of skills for which I am thankful of. The constant inspiration to succeed with dedication and determination has helped me to achieve such an important milestone of my life.

I would like to express sincere gratitude for the guidance provided by my advisor, Dr. Meiliu Lu. She helped me to develop an interest in data mining and perform this research study. I can’t thank her enough for her continuous support.

I would also like to thank Prof. Bob Buckley for giving me constant feedback, help and a detailed review of my project report. His suggestions proved to be valuable for getting the work done according to required standards. I would also like to thank the faculty and staff of Computer Science Department at California State University, Sacramento for giving me an opportunity to pursue my Master’s Degree and guiding me throughout my master’s program.

Most importantly, I would like to thank my wonderful parents Vishwanath and Alka for their unending love and support. It would not have been possible to reach this far without their motivation. They gave me the belief of achieving good things in life without giving up. I would like to thank my beautiful sister Suchita for being there for me and bearing with me during this whole time. She has encouraged me to work hard and believed in me and has always given advice on all important decisions. I wish the best of things in life for her.

Finally, I would like to appreciate the support of my friends Kaushik, Mihir, Nikhilesh, Ameya, Akshay D., Vishal, Abhishek, Akshay M., Ankita, Kshipra, Sangram and Anant who have made this journey worthwhile and entertaining. I hope to enjoy more special moments with them in coming years.

TABLE OF CONTENTS

Page

[Acknowledgements](#h.35nkun2) .......................................................................................................... vii

[List of Tables](#h.44sinio) .................................................................................................................... xi

[List of Figures](#h.z337ya) .................................................................................................................. xii

Chapter

[1.](#h.3j2qqm3) [INTRODUCTION](#h.3j2qqm3)........................................................................................................1

[1.1 Goal of the Project](#h.1y810tw)..................................................................................................1

[1.2 Motivation](#h.4i7ojhp)...............................................................................................................2

[1.3 Research](#h.2xcytpi)..................................................................................................................3

[1.4 Related](#h.2xcytpi) Work..........................................................................................................4

[2.](#h.1ci93xb) BACKGROUND…………………………………………………………………….6

2.1 Objective and Scope…………………………………………………………..…6

2.2 Overview of the Study…………………………………………………………...6

[2.3 Data Mining](#h.2bn6wsx)……………………………………………………………...………7

[2.4 Data](#h.2p2csry) Mining Tools………………….....................................................................9

[2.4.1](#h.147n2zr) The Rapid Miner Tool.................................................................................9

[2.4.2 The W](#h.3o7alnk)eka Tool..........................................................................................12

[2.4.3](#h.23ckvvd) The Tableau Tool.......................................................................................17

2.4.4 The R Programming Tool..........................................................................18

[3. RESEARCH](#h.ihv636) AND COMPARATIVE STUDY.........................................................20

3.1 Comparative Study……………………………………………………………...21

[4.](#h.2lwamvv) DESIGN AND ARCHITECTURE…………………………………………………24

4.1 Use Case Diagram……………………………………………………………....24

4.2 Architecture…………………………………………………………………..…26

[5. IMPLEMENTATION](#h.pkwqa1)………………………………………………………………28

[6. CONCLUSION](#h.pkwqa1) ……….………………....................................................................44

[6.1](#h.39kk8xu) Learning Experience……………………………………………………………44

[6.2 Future Work](#h.1opuj5n).........................................................................................................45

Appendix A. Quiz.js Source Code...………………………………………………..……46

References………………………………………………………………………………..58

LIST OF TABLES

Tables Page

1. Comparative Features………….………………………………………..………….…23

# LIST OF FIGURES

Figures Page

1. Data Mining: A KDD Process..................................................................................…...8

2. Overview of Design Perspective: Rapid Miner Studio……………………..............…11

3. Weka GUI Chooser…………………...………………………………………….……14

4. Weka Explorer……………………………………………………………………...…16

5. Use Case Diagram……………………………………………………………………..24

6. Architecture Diagram………………………………………………………………….26

7. Comparative Study: Site Map……………………………...………………………….29

8. Screenshot of the Home Page…………………………...…………………………….31

9. Screenshot of Data Mining Tools……………………………………………………..33

10. Screenshot of Preprocessing Steps……………………………………….…..............35

11. Preprocessing: Visual Steps……………….……………………………...………….36

12. Screenshot of Classification Algorithm Implementation...…………………………..38

13. Classification Algorithm: Visual Steps………………………………………...…….39

14. Screenshot of Clustering Algorithm Implementation………………………………..40

15. Clustering Algorithm: Visual Steps……………………………………………….…41

16. Screenshot of Quiz…………………………………………………………………...43

# INTRODUCTION

## 1.1 Goal of the Project

The data mining course such as CSc 177 (Data Warehouse and Data Mining) at California State University, Sacramento requires the students to work on a term project. The students learn various classification and clustering algorithms[[2]](#footnote-2) like K-Means, K-NN, ID3, Naïve Bayesian, EM algorithm etc. during the course. These algorithms are implemented using different open source data mining tools. These data mining tools are very useful to predict valuable information and gain knowledge based on the input data sets. A number of popular data mining tools like Weka, Rapid Miner, R and data visualization tool like Tableau provide an efficient method to implement the various data mining algorithms.

During the implementation of the project for Data Mining part of CSc 177, it was necessary to decide which tool to use and how it is to be used. For example, it was important to know and understand how each algorithm is implemented in Rapid Miner and Weka. It takes a lot of time to understand the features and interpret the results. Both Rapid Miner and Weka have unique features and their own advantages of data representation. Students need to spend more time to learn both these tools as they have different methods for implementation of algorithms. If they have to implement same algorithm in these 2 tools, it will take more decision time and therefore the tool selection process can be delayed.

The basic features of this comparative study include:

1. Information about Rapid Miner, Weka, Tableau and R at one place - Providing the introduction of data mining tools at the beginning can be very helpful to the students.
2. Pictorial representation of the data mining process using screenshots for each tool side by side- Students will get clear idea of how various processes like importing data set, preprocessing, selecting data for various algorithms are done in Rapid Miner and Weka.
3. Step by step representation of classification and clustering algorithm implementation using a concrete example.
4. Reduction time in learning of each tool
5. Quizzes to facilitate learning – Provided at the end of the comparative study to test the student knowledge and improve their understanding about the basic concepts.
6. Links for various useful resources on R, Tableau, Weka and Rapid Miner.
7. A Case Study on Quality Healthcare[[3]](#footnote-3)- Contribution by Sri Latha Buyyanapragada explaining the working of Tableau in detail.

## 1.2 Motivation

In today’s digital world, a large amount of raw data is collected every day and stored in different sources and an efficient method is required to be implemented to manage this huge data. This is the reason data warehousing and data mining has gained a lot of significance. Using data mining, we can perform data analysis on this data and extract useful information.

The main motivation for taking up this comparative study was my passion and commitment to help the students in learning the data mining tools to reduce their learning time by further exploring the theoretical concepts learnt during CSc 177 course and applying these concepts using data mining tools. Some of the other motivating factors were to develop an understanding of the following,

* The power of data mining
* The ability to extract information from a large data set and applying various data mining tools to analyze this large set of information in different tools.
* The effectiveness of comparative study to ease learning process

The reason behind choosing a comparative study on data mining tools was to make the information available to users all the time at their convenience and reduce the learning time of the various tools. After taking the CSc 177 course, I felt it would be better to have introductory information about various data mining tools and useful resources, all in one place.

## 1.3 Research

A sufficient amount of time had to be spent on research to ensure how the comparative study of the various data mining tools should be effectively presented. A very high importance is given to ease in getting information of data mining tools and available resources. The main reason is to reduce the amount of time in searching for online resources and tutorials for learning the data mining tools. Along with learning the tools, an important consideration is given to presenting information about different tools at one place. Practical exposure of implementing data mining tools during CSc 177 course helped me understand the overall design complexity and how to make the research student-oriented.

This report is structured as follows: Chapter 2 contains background information on data mining in general and the various data mining tools studied. Chapter 3 describes the research and analysis with focus on the comparative study of a set of data mining tools. Chapter 4 contains the design and architecture of the comparative study website.[[4]](#footnote-4) Chapter 5 describes the Implementation which includes creation of the user interface showing the implementation of classification and clustering algorithms and the actual working of the data mining tools. Chapter 6 includes learning experience along with some future work which could be added to the comparative study of data mining tools.

**1.4 Related Work**

A considerable amount of time in the initial stage was spent in going through various research papers which talked about comparison of various data mining tools and getting to know about the background of each tool. I went through some of the previous research work on comparative study of data mining tools done by other people. To begin with, I read the research paper “Comparative Study of Data Mining Tools” by Rangra et al., [6]. This gave me a fair idea on the basic features of Rapid Miner, Weka and R. Some of the other research papers like “Comparison of Data Mining Techniques and tools for data classification” by Luis C. Borges et al. [7], “The Comparison of Data Mining Tools” by Mesut Ozkan [8] and “A Comparison Study between Data Mining Tools over some Classification Methods” by Abdullah H. Wahbeh et al. [9] helped me to gain knowledge about the implementation of classification algorithms in these tools and gave me a fair idea on how I should work on my study. This research was very useful for me to build a platform of my comparison study. As I already had exposure to data mining concepts from CSc 177, it helped me to come up with different modules for the comparative study.

Apart from learning various data mining tools, I gained experience of web development. I learnt Bootstrap, a responsive framework for web development [10]. I had to brush up my JavaScript and Jquery skills to present this comparative study using a web interface.

# BACKGROUND

**2.1 Objective and Scope**

The main objective for undertaking this comparative study was to gain more knowledge about data mining and data mining tools in general. After taking the Data Mining and data Warehousing Course, I felt that it would be beneficial to have a some sort of resource where future students would be able to learn about data mining tools without having to spend a large amount of time in deciding which tool to use and where to find resources to learn them. This is mainly because students are required to work on a data mining project but have sufficiently less amount of time to learn various data mining tools. Another objective of this comparative study is also to provide relevant and quality resources.

One of the objectives of this comparative study is to provide a web interface for learning data mining tools and making it available to students for their help. In this project, we are going to describe the implementation of some of the data mining tools like Weka, Rapid Miner, R and Tableau. Also, we will see sample implementations of classification and clustering algorithms in Rapid Miner and Weka so that it helps to save students decision time.

**2.2 Overview of the Study**

The development process of the comparative study focuses on three phases. In the first phase, it concentrates on the research on various data mining tools. The second phase includes the implementation of classification and clustering algorithms in Rapid Miner and Weka by providing information about preprocessing the data in both of these tools. This phase also includes providing useful resources for R which is a statistical analysis tool and Tableau which is a powerful data visualization tool. The third phase focuses on designing a web interface for displaying the results of the comparative study on data mining tools and some useful resources.

**2.3 Data Mining**

Data Mining is a step by step process of “mining” or extracting valuable information from a large data set, i.e. big data to perform concrete analysis and predict future trends [11]. Data mining field has evolved tremendously in recent years with growing quality of research. Data mining is known as one of the important tasks of **KDD (Knowledge Discovery Process)** which involves various steps such as data selection, data preprocessing, data transformation, data integration, pattern assessment, etc. These processes are applied to the data stored in data warehouses or databases.

As we can see from figure 1 below, data mining is implemented at a later stage and we cannot simply apply data mining algorithms on raw data. For the simplicity of data mining process, preprocessing is one of the first important steps to make the data task-relevant and useful for evaluations. Preprocessing involves data cleaning and data integration. The data mining tools primarily focus on pattern evaluations based on historic data which is collected from the data warehouse to gain knowledge which acts as a basis for future analysis.

DATABASES

DATA CLEANING

DATA WAREHOUSE

DATA SELECTION

TASK RELEVANT DATA

DATA MINING

PATTERN EVALUATIONS

KNOWLEDGE

**Fig. 1 Data Mining: A KDD Process [12]**

**2.4 Data Mining Tools**

A number of open source and proprietary based data mining and data visualization tools exist which are used for information extraction from large data repositories and for data analysis. Some of the data mining tools[[5]](#footnote-5) which exist in the market are Weka, Rapid Miner, Orange, R, KNIME, ELKI, GNU Octave, Apache Mahout, SCaViS, Natural Language Toolkit, Tableau, etc.

In this comparative study, I have concentrated on 4 of the commonly used tools: Rapid Miner, Weka, R and the popular data visualization tool, Tableau.

**2.4.1 The Rapid Miner Tool**

Rapid Miner is one of the most widely used open source data mining tool developed in 2001 by Ingo Mierswa and Ralf Klinkenberg. Prior to 2006, it was known as YALE (Yet another Learning Tool) [3]. Rapid Miner is a XML based data mining tool used to implement various machine learning and data mining processes. It is a popular tool to implement classification and clustering algorithms[[6]](#footnote-6). An important feature of Rapid Miner is its ability to display results visually. “Rapid Miner provides learning schemes and models and algorithms from [Weka](https://en.wikipedia.org/wiki/Weka) and [R](https://en.wikipedia.org/wiki/R_(programming_language)) scripts that can be used through extensions.” [13]

Rapid Miner also provides an integrated environment for machine learning, predictive analytics, data mining, text mining and business analytics. It is used for business and industrial applications as well as for research, education, training, rapid prototyping, and application development and supports all steps of the data mining process. The Rapid Miner Extensions marketplace provides a platform for developers to create data analysis algorithms and publish them to the community [13]. Rapid Miner Studio provides 3 perspectives for users to work with,

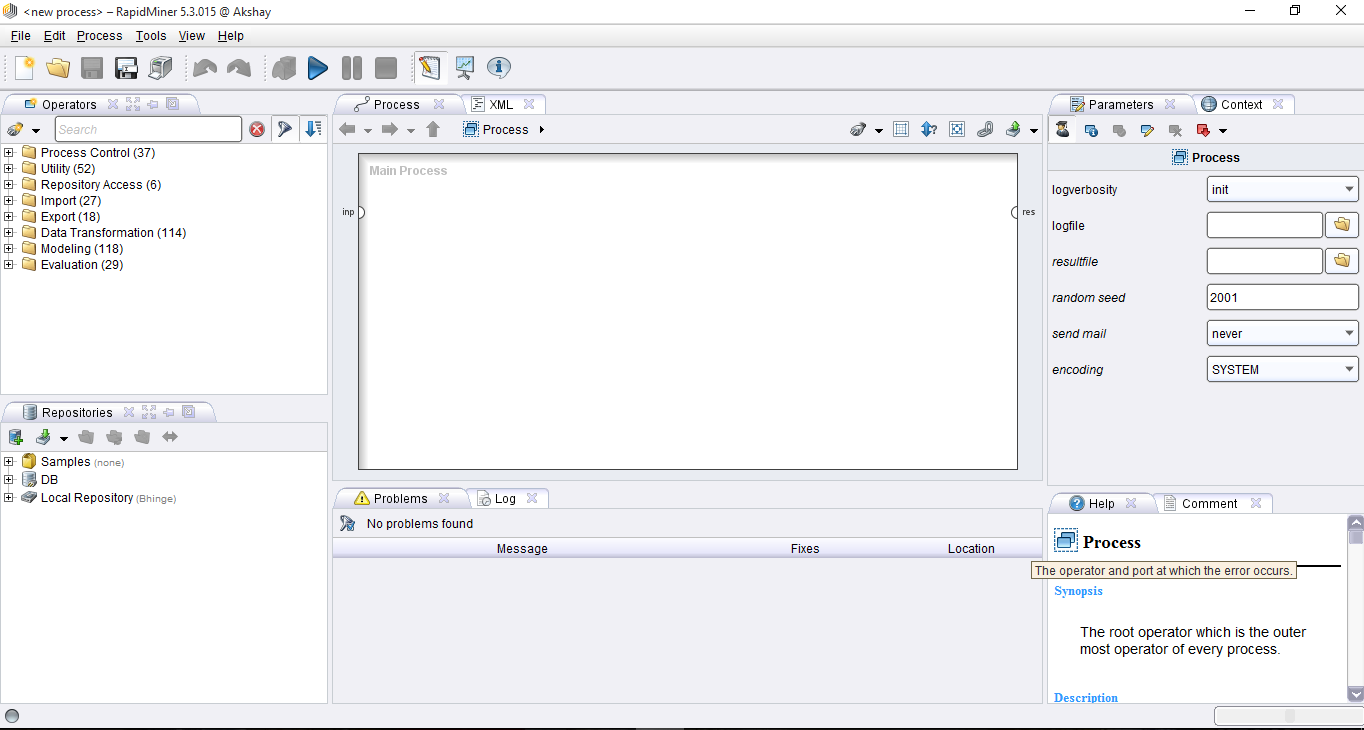
**The first perspective is Home perspective-**

This is the default perspective when you start Rapid miner for the first time. Rapid Miner provides information about new releases or current version which the user is running in this perspective.

**The second is Design perspective-**

It is the main Rapid Miner Studio perspective where in all the processes are implemented and worked with. The design perspective is displayed in figure 2 below. It displays various modules of Rapid Miner Studio.

The “Operator” View is used to select various operators for performing any data mining task. “Process” View shows the actual implementation of data mining algorithms and display the results. “Repository” View consists of all the data repositories stored in Rapid Miner Studio. “Problems” View to display any errors during execution of process and also provides suggestions to remove those errors. “Parameters” View to set any parameters for a particular operator. For example, set size of clusters for the K-means operator.

****

**Fig. 2 Overview of Design Perspective: Rapid Miner Studio [3]**

**The third perspective is Wizard perspective-**

This perspective is used to apply data mining algorithms on the user’s data set. It is a very useful feature provided by Rapid Miner Studio.

Rapid Miner Studio provides a platform for existing data mining techniques. A great feature of Rapid Miner is its drag-and-drop methodology. Users can select operators and combine them by using this approach [14]. This makes it easier to perform the desired tasks of implementing various data mining algorithms. This whole procedure is completed using combinations of different processes.

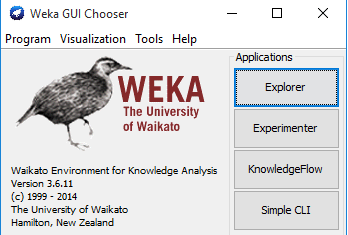
Another important feature of Rapid Miner is that it provides very good visualizations of the data like 3-D graphs and matrices [14]. Rapid Miner focuses a lot on providing more visual impact of the data, as we can understand visual images more quickly as compared to plain text. Rapid Miner consists of more than 100 learning schemes 0classification and clustering analysis. Another important feature of Rapid Miner is its responsive and intuitive GUI.

**2.4.2 The Weka Tool**

Weka is one of the very popular open source data mining tools developed at the University of Waikato in New Zealand in 1992. It is a Java based tool and can be used to implement various machine learning and data mining algorithms written in Java [9]. The simplicity of using Weka has made it a landmark for machine learning and data mining implementation [15]. Weka supports reading of files from several different databases and also allows importing the data from the internet, from web pages or from a remotely located SQL database server by entering the URL of resource. Among all the available data mining tools, Weka is the most commonly used of all due to its fast performance and support for major classification and clustering algorithm[[7]](#footnote-7). Weka can be easily downloaded and deployed.

Weka provides both, a GUI and CLI for performing data mining and does a good job of providing support for all the data mining tasks [16]. Weka supports a variety of data formats like CSV (Comma-separated Value), ARFF and Binary. Weka focuses more on textual representation of the data rather than visualization although it does provide support to display some visualization but those are very generic. Also, Weka does not provide visual representation of results of processing in an effective and understanding manner like Rapid Miner.

Weka performs accurately when the size of the data set is not large. If the size is large, then Weka does experience some performance issues. Weka provides support for filtering out data or attributes [16]. Figure 3 on next page shows the available choices for user interface. Weka supports the following three graphical user interfaces-

****

**Fig. 3 Weka GUI Chooser [4]**

**1. The Explorer**

It is the most commonly used graphical user interface in Weka to implement data mining algorithms[[8]](#footnote-8). It supports exploratory data analysis to perform preprocessing, attribute selection, learning and visualization. This interface consists of different tabs to access various components for performing data mining which can be seen in Figure 4 below. The different tabs are-

A) Preprocessing

Using this tab, we can load input data files and perform preprocessing on this data using filters.

B) Classify

This tab is used to implement different classification and regression algorithms. We can do this by selecting a particular classifier from this tab. For example, the K-NN or Naïve Bayesian algorithm can be implemented by using this tab.

C) Associate

This tab is used to find out all association rules between different attributes of the data and which can be used for further mining. For example, Association rule mining, etc. [17].

D) Cluster

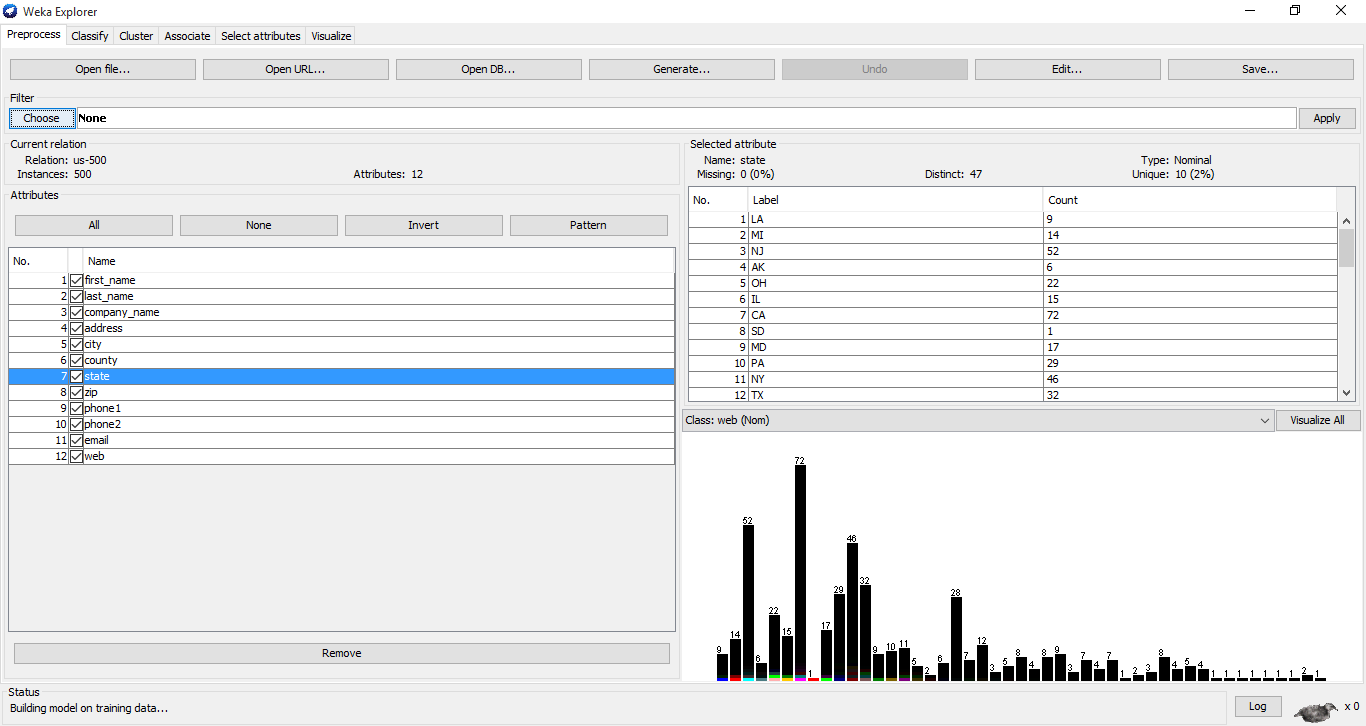
Using this tab, we can select a particular clustering algorithm to implement for our data set. Clustering algorithms like K-means can be implemented using this tab.

E) Select attributes

This tab is used to select particular attributes from the data set useful for implementing the algorithm.

F) Visualize

This tab is used to visualize the data whenever available or supported by a particular algorithm in the form of scatter plot matrix [17].



**Fig. 4 Weka Explorer [4]**

**2. The Experimenter**

This user interface provides experimental environment for testing and evaluating machine learning algorithms.

**3. The Knowledge Flow**

Knowledge flow is basically a component based interface similar to explorer. This interface is used for new process evaluations.

**2.4.3 The Tableau Tool**

Tableau is a powerful data visualization tool used in business intelligence and data analysis. Tableau Software was invented by Chris Stolte, Christian Chabot and Pat Hanrahan in January, 2003 [18]. The visualization provided by Tableau has completely enhanced the ability to gain more knowledge about the data we are working on and can be used to provide more accurate predictions. “The product queries [relational databases](https://en.wikipedia.org/wiki/Relational_databases), [cubes](https://en.wikipedia.org/wiki/OLAP_cube), [cloud databases](https://en.wikipedia.org/wiki/Cloud_database), and [spreadsheets](https://en.wikipedia.org/wiki/Spreadsheets) and then generates a number of graph types that can be combined into dashboards which can be securely shared over a [computer network](https://en.wikipedia.org/wiki/Corporate_network) or the [internet](https://en.wikipedia.org/wiki/Internet)” [18].

Unlike Rapid Miner and Weka, Tableau does not implement data mining algorithms provides visualizations of the data. For this, Tableau provides integration with another popular statistical analysis tool R[[9]](#footnote-9), to provide support for data mining. “Tableau offers five main products namely Tableau Desktop, Tableau Server, Tableau Online, Tableau Reader and Tableau Public. Tableau Public and Tableau Reader are available freely, whereas Tableau Server and Tableau Desktop come with a free trial period afterwards which the user has to pay” [18]. Tableau has made it possible to explore and present the data in a much simpler and beautiful manner. Working on projects using Tableau is less time consuming and easy to handle. Tableau uses a feature called Dashboard which is a collection of worksheets which can be easily imported from anywhere.

In my project, I have added contribution by Sri Latha Buyyanapragada, a fellow student as a reference to learn Tableau implementation using a case study. The case study is based on Quality Healthcare Analytical Reports and explains working of Tableau systematically. Students can use this case study to create Tableau dashboards and visualize their own data which could be used as a guide to perform data analysis.

**2.4.4 The R Programming Tool**

R is an open source statistical analysis tool based on C and FORTRAN programming language developed by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand [19]. R was released in 1997 and it is currently licensed using GNU General Public License. Using R, well-designed publication-quality plots can be produced, including mathematical symbols and formulae wherever required [4]. R uses a number of different packages to support data mining or statistics and provides a well-integrated collection of intermediate tools for data analysis

R-Integration can be used in combination with Tableau to create visual representations of various data mining algorithms due to the clear and interactive visualizations which can be created using Tableau. Although R provides less support to data mining algorithms as compared to Rapid Miner and Weka, it does implement a few data mining algorithms. R supports implementation of Naïve Bayes algorithm, confusion matrix and summary of data which is used for classification of data. R uses a code-driven methodology which involves use of a number of in-built functions and commands to perform statistical analysis and data mining.

Apart from support for data mining, R has a large collection of statistical library [6]. R-project can be said to be analogous to programming in MATLAB due to its similar functionality. There is no inbuilt preprocessing function available in R [8]. As mentioned above, to implement Naïve Bayes, first the package has to be loaded before beginning with further coding.

# RESEARCH AND COMPARATIVE STUDY

While working on any project, the primary goal is to gather all the requirements to have a clear vision on how to go about completing the desired tasks in a specific amount of time. Performing research and analysis helps to accomplish our goal. This directly leads to increase in the effectiveness of our work and save time and energy during project work.

In this project, I had to first perform requirements gathering and understand the scope of the study in detail. Another import aspect of any project is to get familiar with the technologies used in implementation of the project. Apart from learning various data mining tools, as I had to develop a web interface, I learned Bootstrap, which is a JavaScript framework for responsive web development. The decision to use Bootstrap was based on extensive research of how the website could be made to support different device-types. The primary reason was that as the study was going to be an aid for the students, it should be easy to view on mobile devices and tablets too. I read many tutorials and articles to get some hands-on experience of this framework before effectively implementing it.

A large amount of time had to be spent in studying about data mining tools. For information on the data mining tools, I went through the CSc 177 course website [1]. As the website consists of various useful resources to learn data mining, this was the first resource I referred to carry out my analysis. I learnt about classification and clustering from the website which helped me to practically implement it using Rapid Miner and Weka. I used a sample tennis data set obtained online from <http://archive.ics.uci.edu/ml/datasets/Tennis+Major+Tournament+Match+Statistics> to illustrate the implementation of Preprocessing, Classification and Clustering in Rapid Miner and Weka. I also reviewed articles and watched video tutorials[[10]](#footnote-10) to understand the working of various data mining tools and finally got fair idea of how I should conduct my comparative study.

**3.1 Comparative Study**

As Comparative Study is the primary focus of my project, I had to become proficient in the use of Rapid Miner, Weka, Tableau and R in detail. Each of these data mining and data visualization tools has their own special features and also drawbacks. The differences are what make them unique and popular in their own way.

After studying the features of each tool, I have compiled the following list of comparable features provided by each of these data mining tools. These are defined in the list and there after introduced in Table 1.

**1. Usability**

This feature determines the usability of each tool. This describes which user interface is comparatively easier to use.

**2. Speed**

Speed is an important distinguishing factor among different data mining tools. This feature helps to understand how system configuration impacts the working of a particular data mining tool.

**3. Visualization**

Visualization is the most important feature of a data mining tool. This comparative feature distinguishes each data mining tool based on different visualization options provided.

**4. Algorithms supported**

This feature classifies data mining tools based on the algorithm implementation supported by them and the descriptor selection choice available. [20]

**5. Data Set Size**

Smaller or larger data set support is another comparable feature among different data mining tools.

**6. Memory Usage**

As memory usage affects performance, memory usage is another important feature for comparing data mining tools.

**7. Primary Usage**

Each data mining tool has a particular usage which is one of the comparable features. For example, both R and Weka can be used to implement data mining algorithms but the primary usage of R is in statistical computing.

**8. Interface Type Supported**

The type of interface provided for algorithm implementation is one of the comparative features of this study. From the context of this comparative study, use of Graphical User Interface (GUI) or Command line Interface (CLI) differentiates each tool.

**Table 1: Comparative Features [6]**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Rapid Miner** | **Weka** | **Tableau** | **R** |
| **Usability** | Easy to use | Most easiest to use | Simple to use | Complicated as coding required |
| **Speed** | Requires more memory to operate | Works faster on any machine. | Works fast on any machine | Works fast on any machine |
| **Visualization** | More options but less than Tableau | Less options | Many visualization options | Less options as compared Rapid Miner |
| **Algorithms supported** | Classification and Clustering | Classification and Clustering | Not used to implement Classification and Clustering | Very few Classification and Clustering algorithms |
| **Data Set Size** | Supports large and small data set | Supports only small data sets | Supports any data set | Supports large and small data set |
| **Memory Usage** | Requires more memory | Less Memory hence works faster | Less Memory | More Memory |
| **Primary Usage** | Data Mining, Predictive Analysis | Machine Learning | Business Intelligence | Statistical Computing |
| **Interface Type Supported** | GUI | GUI / CLI | GUI | CLI |

# DESIGN AND ARCHITECTURE

In this section, the design and architecture details of the comparative study are provided. Design and Architecture is one of the important steps of a project. It helps us to decide on how to proceed with the actual implementation part and gather information about the different modules of a project.

**4.1 Use Case Diagram**

A Use Case Diagram depicts the users of a project. In this case, users are the people who will go through the comparative study for their reference. Use cases help to clearly identify system requirements.

Users

**Fig. 5 Use Case Diagram**

As can be seen from the above use case diagram, the comparative study consists of only one type of use case. The users / “Actors” would be the students of CSc 177 who would access this comparative study for their projects or learning. Other users would be professor who would review the content of the comparative study and suggest modifications to the content. These users would choose one of the following tabs after accessing the comparative study website-

**A) Data Mining Tools**

This tab provides description of the four data mining tools namely Rapid Miner, Weka, Tableau and R.

**B) Comparative Study**

This tab describes the working of classification and clustering algorithms in Rapid Miner and Weka. It also describes how preprocessing is done in each of these tools as a basis of comparison.

**C) Quiz**

This tab deals with the quiz based on the understanding of each data mining tool. Users would take the quiz and able to test their knowledge on Rapid Miner, Weka, Tableau and R.

**D) Resources**

This tab provides useful and meaningful resources for further learning of the data mining tools.

**E) Site Map**

Site Map provides the general flow of the comparative study which can be useful for the users to easily navigate through the different tabs and access a tab of their choice.

**4.2 Architecture**

The Architecture of a project comprises of the major entities involved in the development. In this comparative study, the architecture can be defined to be consisting of the data mining tools, the comparison between each tool and the web interface to present the working of the data mining tools. The figure below best describes the architecture of my comparative study.

Data Mining Tools

Comparison Study

Web Interface

**Fig. 6 Architecture Diagram**

To explain the above architectural diagram in brief, the comparative study can be divided into three main entities,

**1. Data mining tools**

This is one of the main entities which involve the data mining tools to be studied and base the comparison study on them.

**2. Comparative Study**

This entity deals with the actual comparison study of the data mining tools. I have described implementation of classification and clustering algorithms in Rapid Miner and Weka in this section. For Tableau and R, I have provided useful resources.

**3. Web Interface**

Web interface is used to present the comparative study and data mining tools information.

1. IMPLEMENTATION

The main technologies used to design the web interface for the comparative study are Bootstrap, HTML5, JavaScript, jQuery and CSS3. The comparative study includes description of the Rapid Miner, Weka, Tableau and R, implementation of classification and clustering algorithms in Rapid Miner and Weka with their description and visual steps. The comparative study also involves a sample quiz based on the knowledge of these data mining tools.

As the main focus of this comparison study is on data mining tools, I have initially provided information about Rapid Miner, Weka, Tableau and R. Figure 7 below shows the total flow of the web interface for the comparative study. This is the work flow of the web interface.

Comparative Study on Data Mining Tools

Home

Data Mining Tools

Rapid Miner

Weka

Tableau

Case Study

R

Comparative Study

Preprocessing

Classification Algorithm

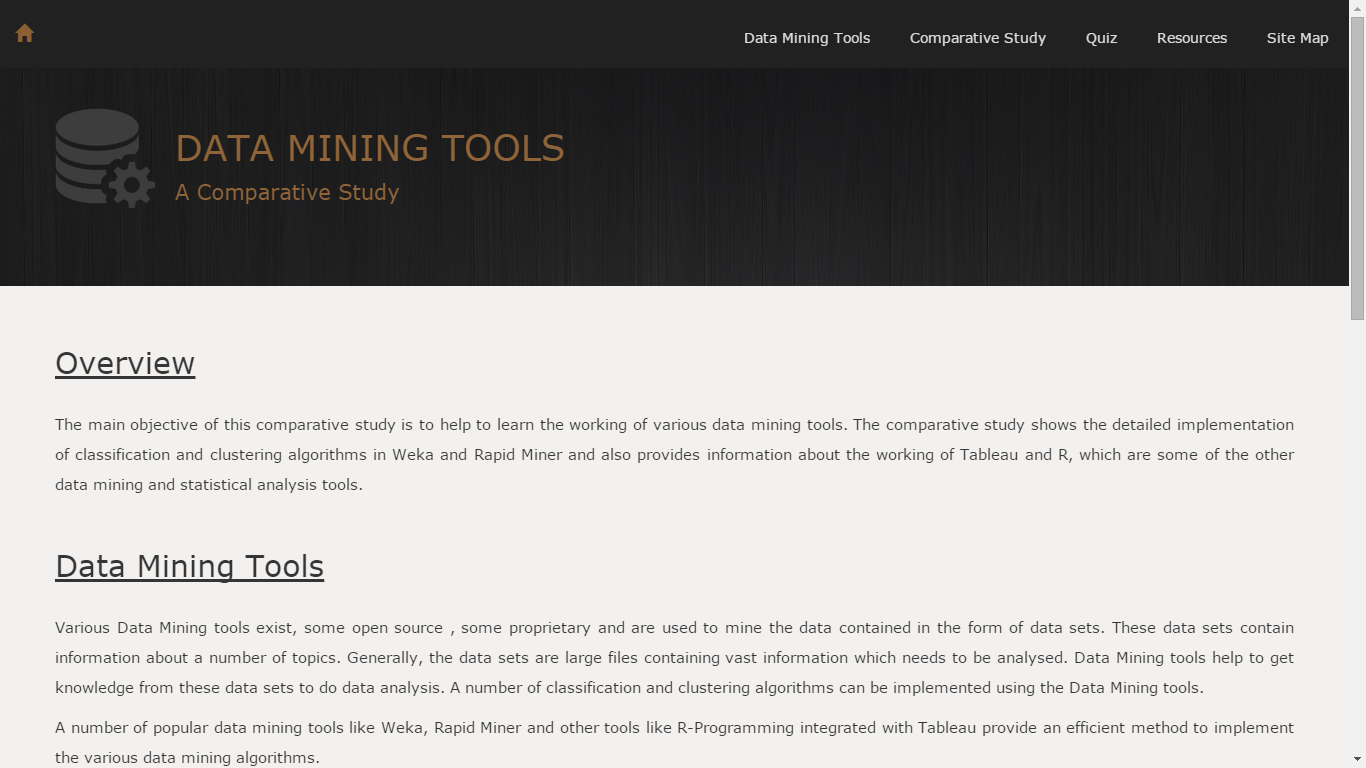
Resources

Quiz

Clustering Algorithm

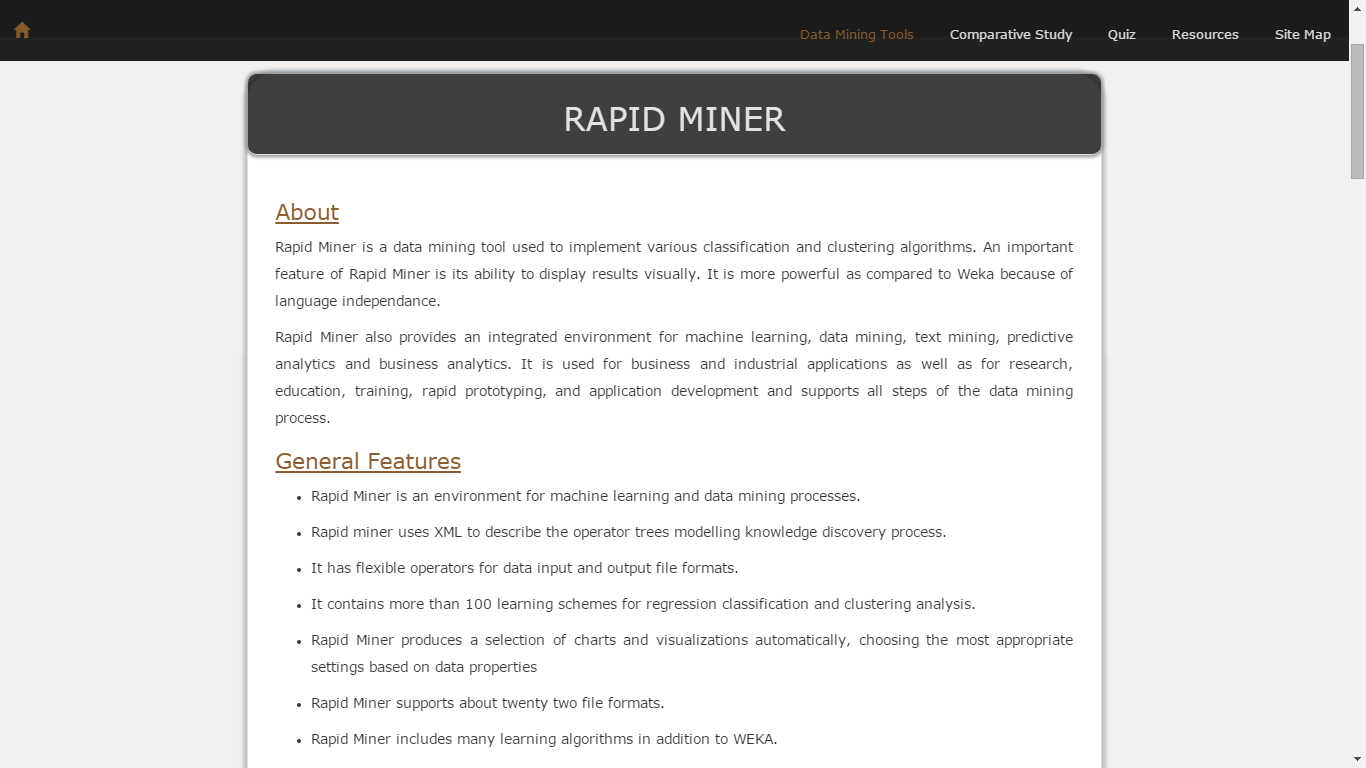
**Fig. 7 Comparative Study: Site Map**

Figure 8 on next page shows the overall content of the web interface which is basically the home page. The home page contains overview of the comparative study and information about the tools on which I have based my comparative study. From the home page, students can directly navigate to any other web page of their choice. For example, students can directly go to the Comparative Study page to learn about classification and clustering algorithm implementation or check for online resources using the Resource page. Students can answer the Quizzes by directly going to that page.



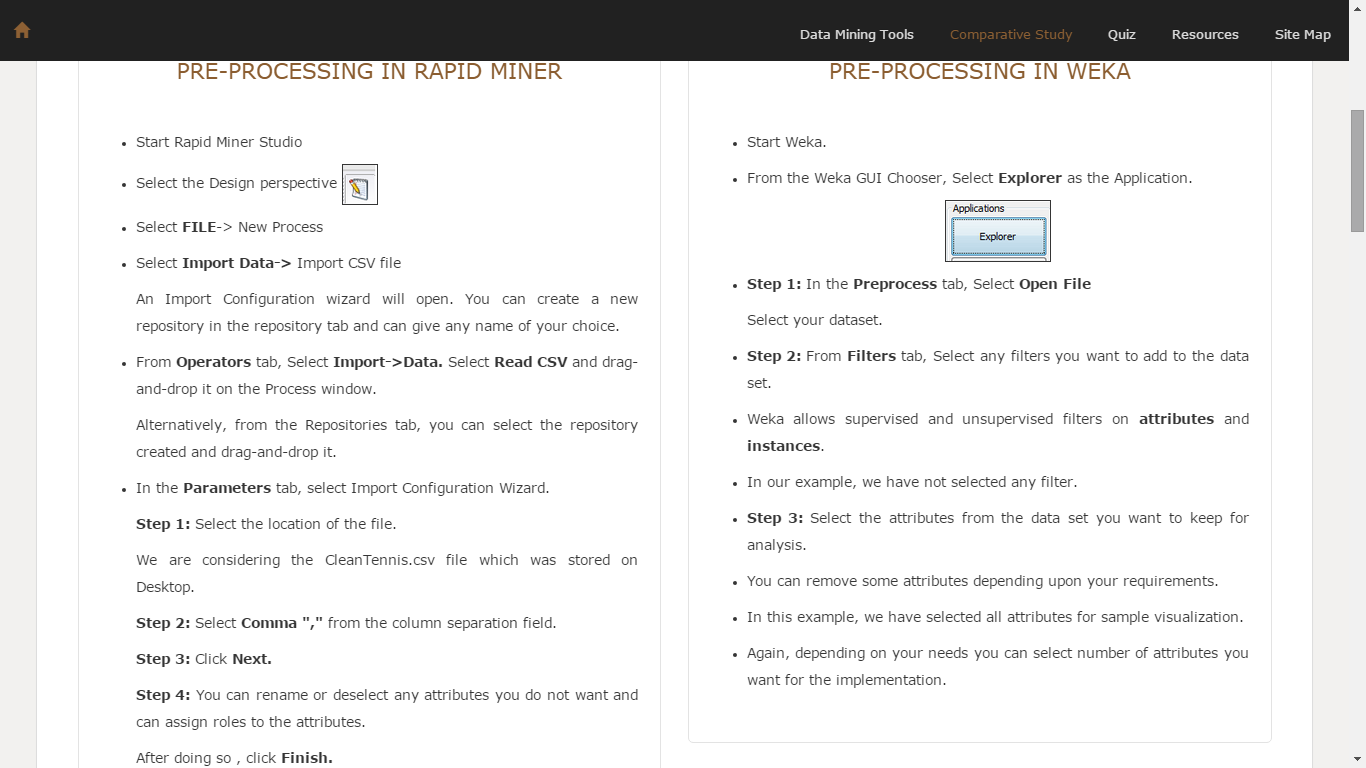
**Fig. 8 Screenshot of the Home Page**

Figure 9 on next page shows the next page of the web interface which is the description of various data mining tools. I have provided brief description and general features of Rapid Miner, Weka, Tableau and R in this page. In the description of Tableau, I have added a link to a Case Study on Quality Health Care which would explain the working of Tableau in detail to students. I have provided various useful and meaningful resources along with some articles to learn about R language in the description of R. This would be highly useful for the students to learn about R data mining.

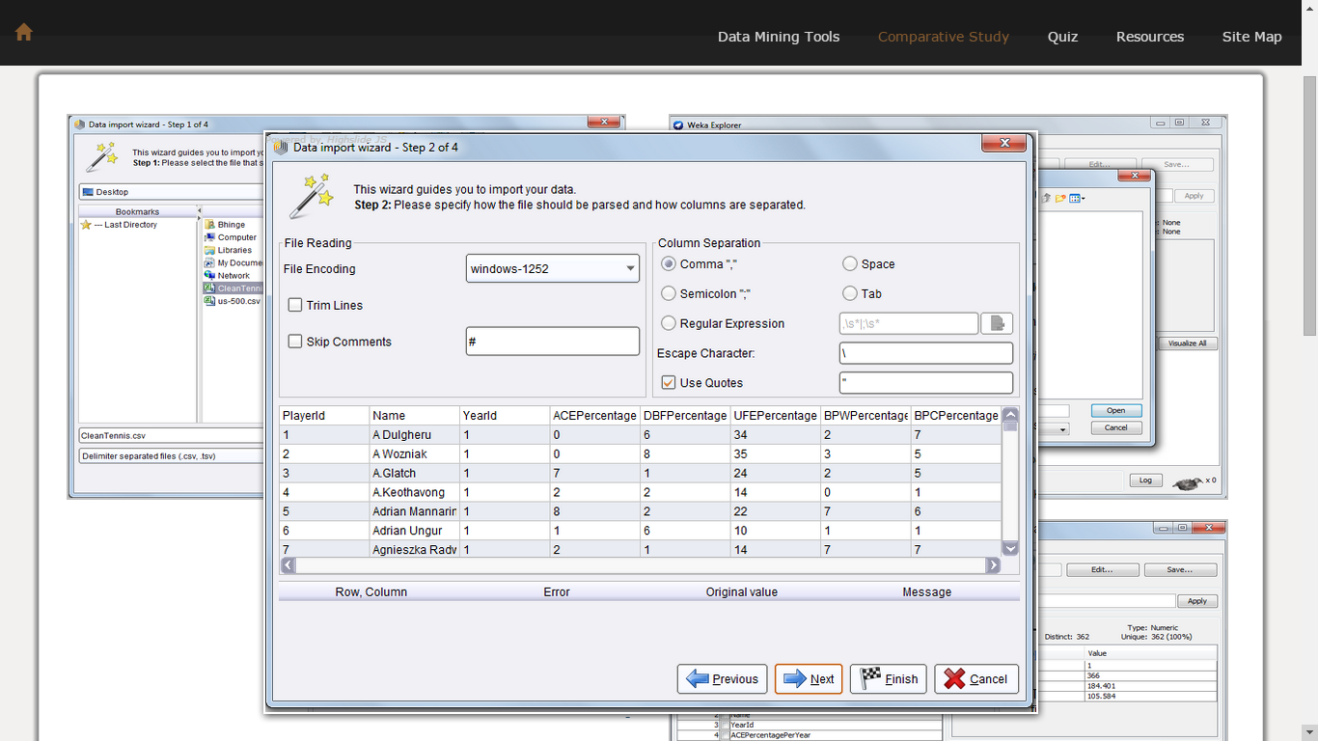


**Fig. 9 Screenshot of Data Mining Tools**

Figure 10 describes the Comparative Study. It involves description of preprocessing steps in Rapid Miner and Weka using sample data set. I have also provided visual steps so that they can view each step visually if they want. For this, I took screenshots of each step in Rapid Miner and Weka. To put together these screenshots on a webpage, I have used Highslide JavaScript which makes it convenient to view images on web pages. Highslide JS is used to create image gallery [21].

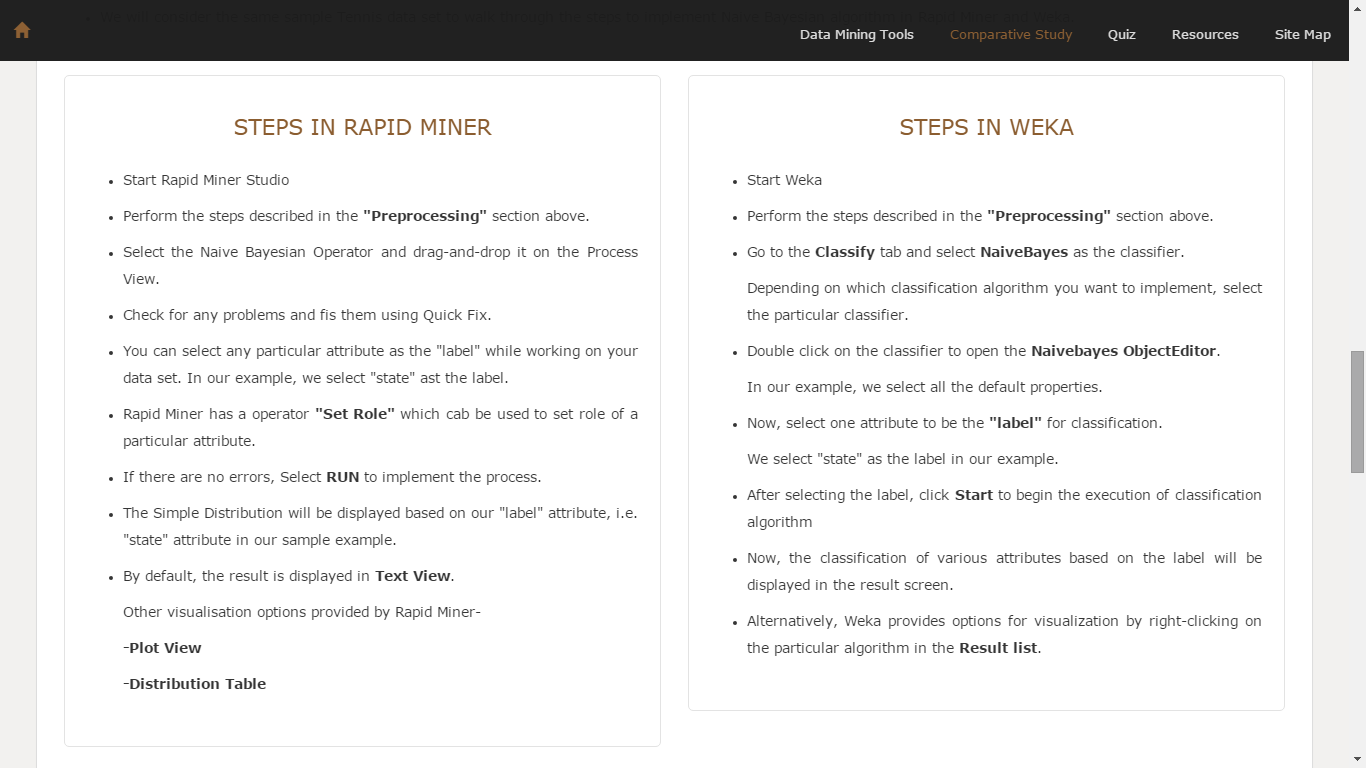


**Fig. 10 Screenshot of Preprocessing Steps**

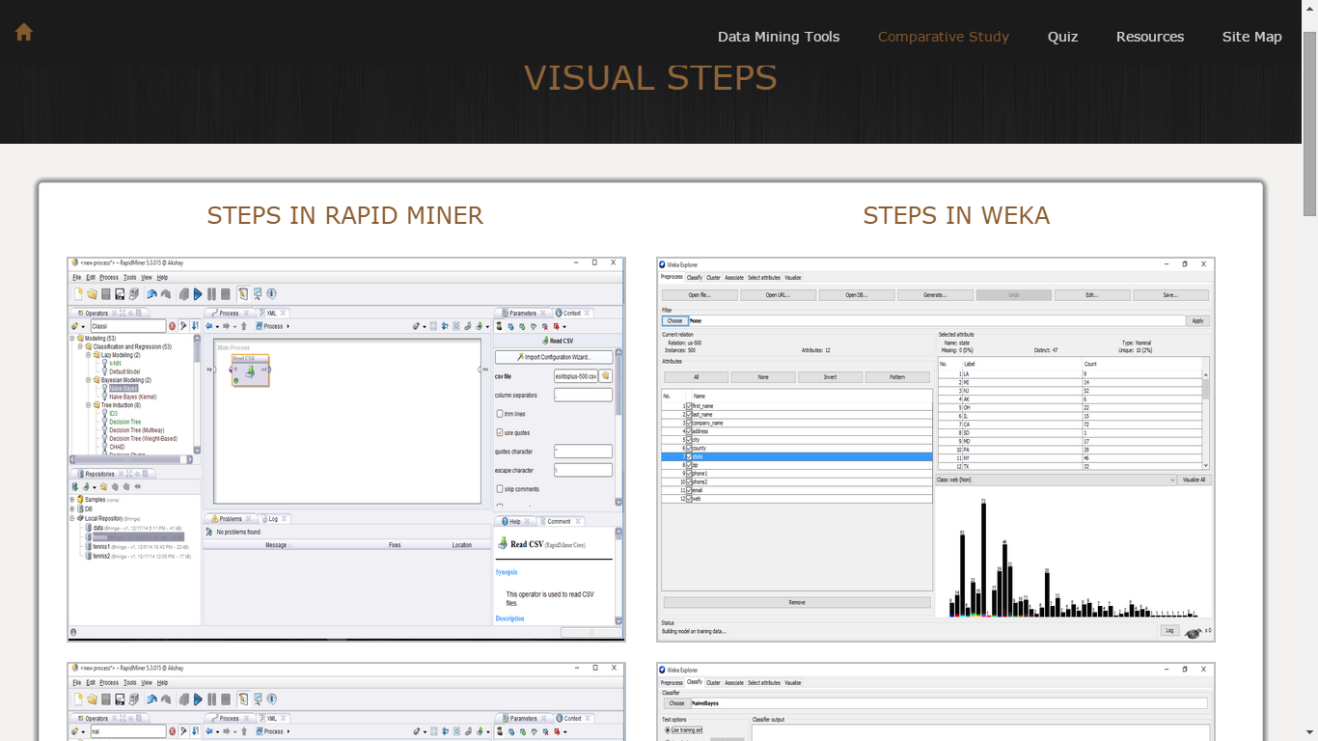
****

**Fig. 11 Preprocessing: Visual Steps**

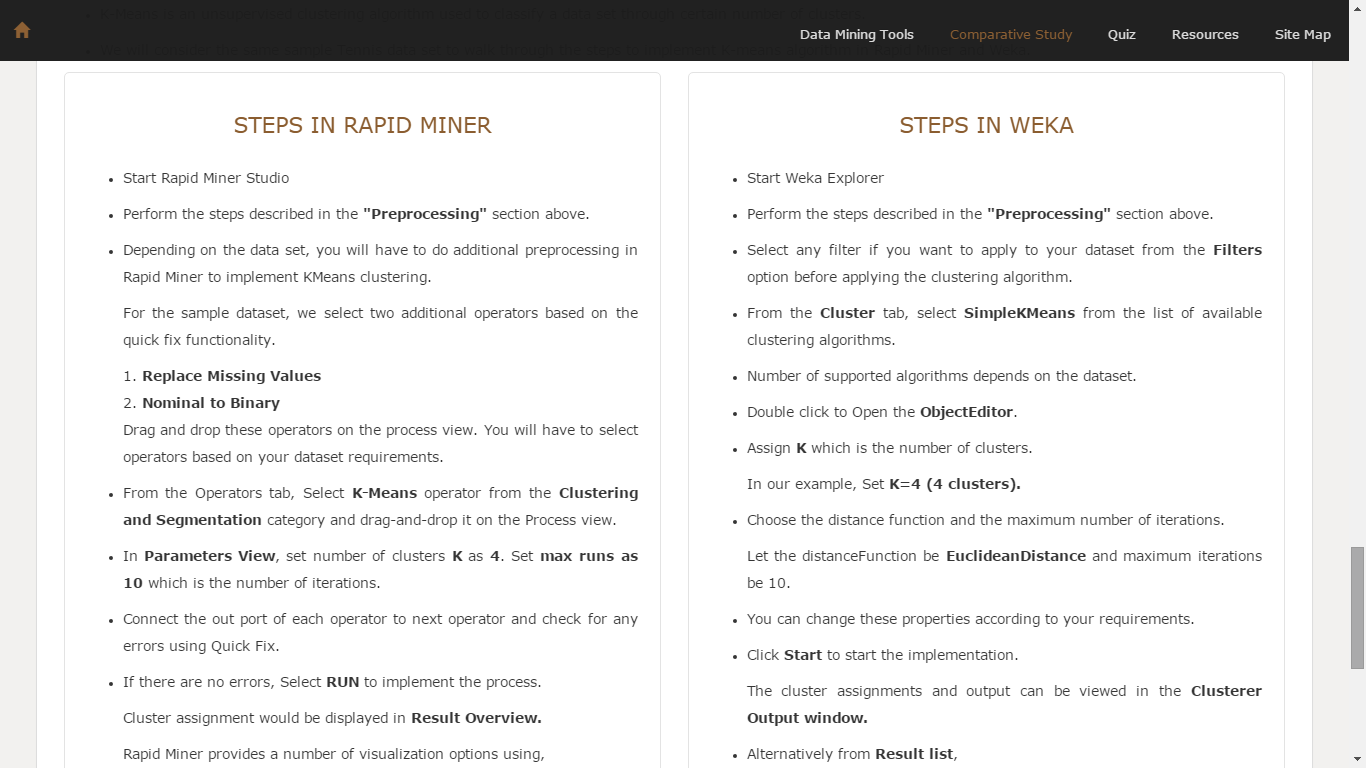
Finally, I have implemented one classification and clustering algorithm in Rapid Miner and Weka to have a fair idea of the difference in steps needed to perform in these two tools. I have implemented the K-means clustering algorithm and Naïve Bayesian classification algorithm as an example by explaining each step systematically in such a way that student’s understand the working of both the tools. Figures 11, 12, 13, 14 and 15 describe the implementation of classification and clustering algorithms.



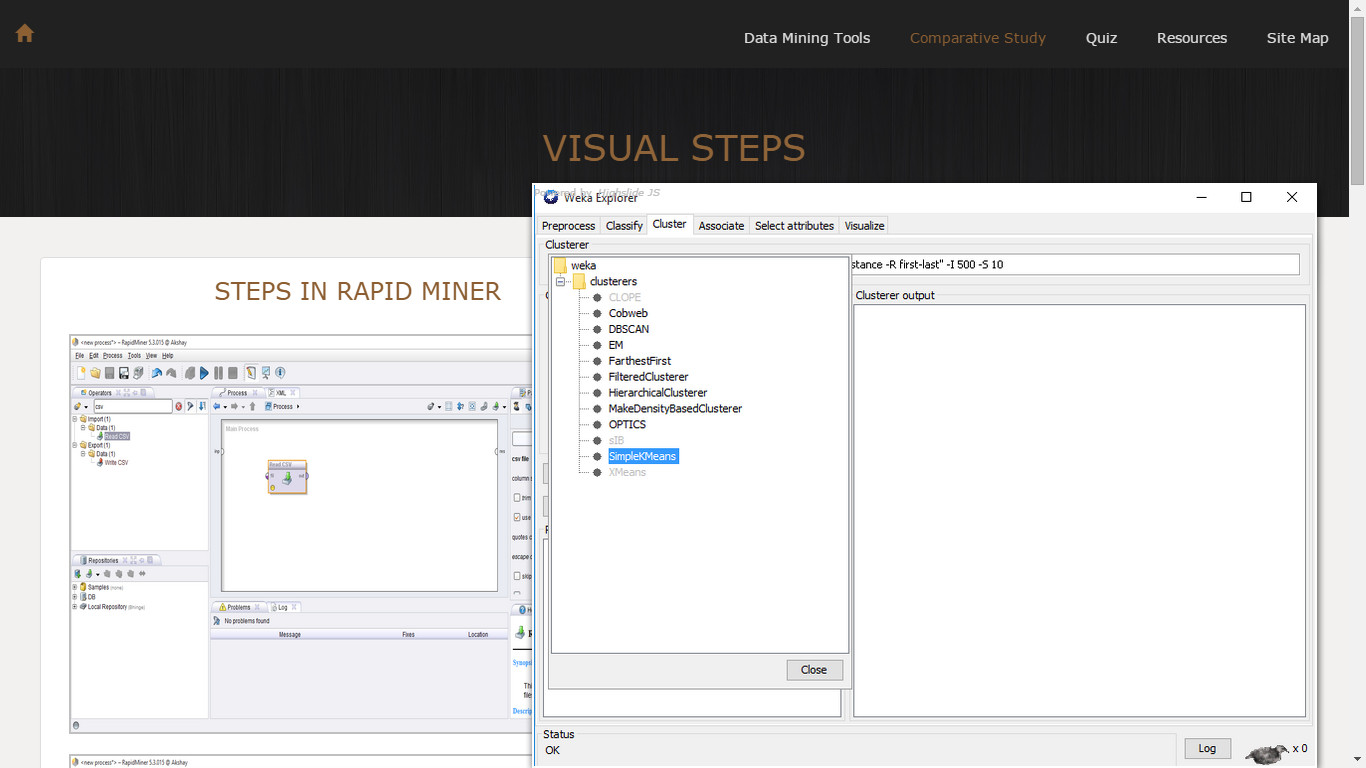
**Fig. 12 Screenshot of Classification Algorithm Implementation**

****

**Fig. 13 Classification Algorithm: Visual Steps**

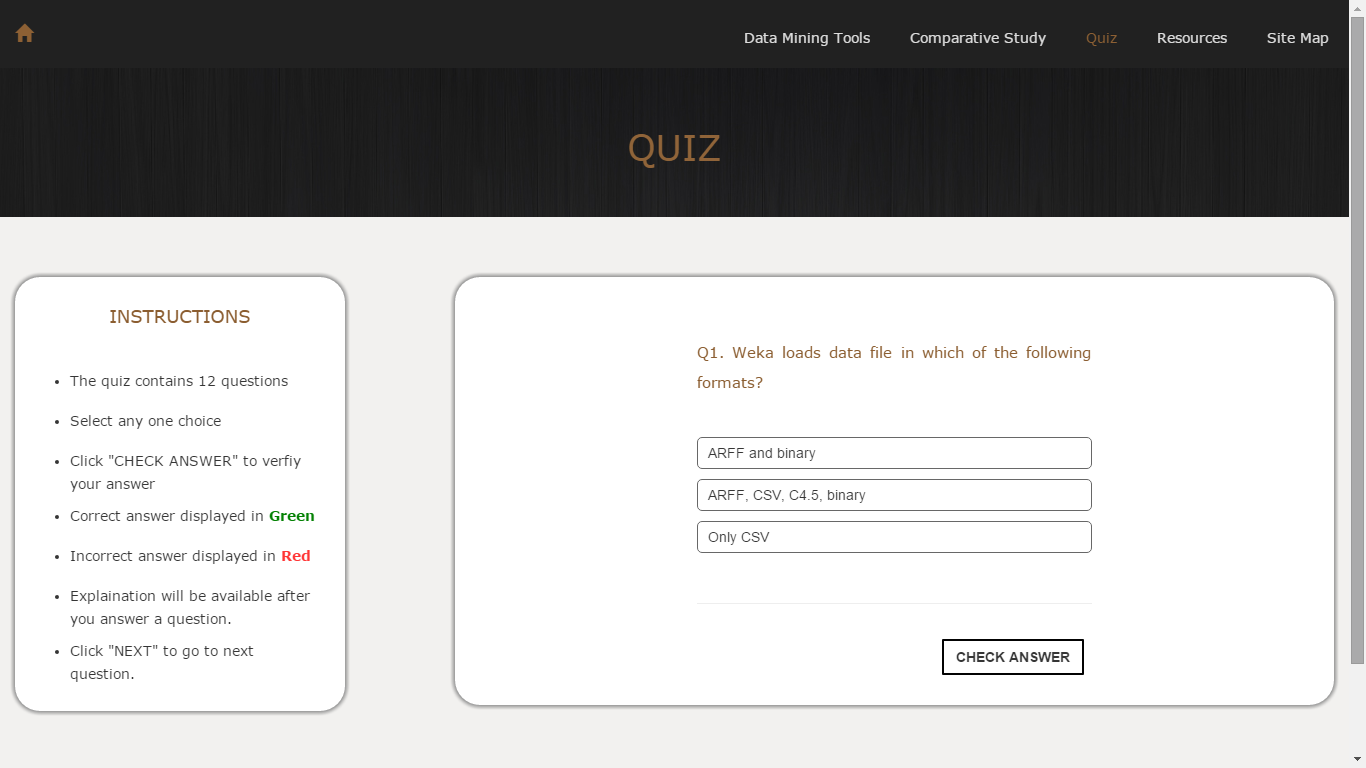


**Fig. 14 Screenshot of Clustering Algorithm Implementation**



**Fig. 15 Clustering Algorithm: Visual Steps**

Figure 16 on the next page describes the Quiz page. I have provided 12 questions for students. The quiz covers questions on Rapid Miner, Weka, Tableau and R. Students would be able to get the Explanation for each question once they answer a particular question. Based on their response, students would get response on number of questions answered correct and incorrect.

****

**Fig. 16 Screenshot of Quiz**

6. CONCLUSION

The primary focus of this comparative study is to provide a platform to learn various data mining tools without having to spend a lot of time in searching for resources and tutorials. Another important aspect is to enhance the learning experience for the CSc 177 students to help them in their Data Mining course project and their own individual research.

The comparative study also provides a quiz on the data mining tools which helps to check the student knowledge on the implementation of each tool. Additionally, a number of relevant useful resources are provided to assist in further learning. This comparative study will further encourage the students to learn a number of other data mining tools and use them for their data mining projects.

## 6.1 Learning experience

Working on this project has taught me new concepts and enhanced my web development skills like HTML5, CSS3, JavaScript, jQuery and Bootstrap. I got experience of working in systematic development process from research, design, implementation all the way up to testing the working of my project. It has also helped me to gain some insight into how professional projects should be undertaken. This project has taught me how to do comprehensive research on a number of topics and use that research to further expand my knowledge. I am more comfortable working on Rapid Miner, Weka, Tableau and R now as compared to the beginning of this project.

**6.2 Future work**

The comparative study is restricted to the data mining tools such as Rapid Miner, Weka, Tableau and R. In the future, more data mining tools could be studied and those results could be incorporated. Also, another innovative feature like discussion forum could be added to the comparative study which would provide a platform for the students of CSc 177 to ask their queries on data mining tools and collaborate with each other online to get solutions or help from their peers. This might also help professors to check class participation of the students based on the online interaction.

# BIBLIOGRAPHY

APPENDIX A

QUIZ.JS Source Code

<script type="text/javascript">

$(document).ready(function()

{

$("#button1").hide();

$("#button2").hide();

var correct = 0;

var wrong = 0;

i = 0;

var selected\_answer;

var no\_ans\_seleted=true;

var q =

[

{

"question":"Q1. Weka loads data file in which of the following formats?",

"choice": [" ARFF and binary", " ARFF, CSV, C4.5, binary", "Only CSV"],

"answer": " ARFF, CSV, C4.5, binary",

"index": "2",

"Explanation": "Apart from CSV file format, Weka supports datasets stored in ARFF(Attribute-Relation File Format), C4.5 and binary format. Check official documentation of Weka(provided in reference) for further information"

},

{

"question":"Q2. The knowledge discovery process is described in Rapid Miner using which format?",

"choice": ["XML format", " XSLT format", "None of these"],

"answer": "XML format",

"index": "1",

"Explanation": "The process structure is described internally by XML in Rapid Miner. This process structure defined is implemented using a graphical user interface."

},

{

"question":"Q3. Rapid Miner Studio uses which perspective for creating, editing and managing all analysis processes?",

"choice": ["Design perspective", "Result perspective", "Wizard perspective"],

"answer": "Design perspective",

"index": "1",

"Explanation": "Rapid Miner supports 4 different perspectives. Design perspective is used for executing all the analysis processes. Result perspective displays the results in the form of data models, statistics or charts. The Wizard perspective allows users to apply sample data mining algorithms to your data set."

},

{

"question":"Q4. Operator View, Repository View, Process View and Parameters View is provided in which of the following data mining tools?",

"choice": ["Tableau", "Weka", "Rapid Miner"],

"answer": "Rapid Miner",

"index": "3",

"Explanation": "Rapid Miner uses Operator View for , Repository View to display all the data repositories, Process View to do actual process implementations. The Parameters View is used to display the various parameters for a particular operator selected inside the Process View."

},

{

"question":"Q5. Consider a variable <b>'read\_file'</b>. Consider a sample command on the interface, <br><b>read\_file = read.csv(\"Desktop/Sample.csv\").</b><br>It is a valid command to read CSV file in which of the following?",

"choice": ["R-programming", "Weka", "Rapid Miner"],

"answer": "R-programming",

"index": "1",

"Explanation": "read.csv is a command for data import in R and it takes a csv file and stores the contents in the variable. The data can be displayed after calling the variable. Weka and Rapid Miner involve option to select data input file using a FILE menu."

},

{

"question":"Q6. Rserve is used to do which of the following?",

"choice": ["Allows other programs to use functions and capabilities of R from other languages without initializing R.", "Perform association rule mining", "None"],

"answer": "Allows other programs to use functions and capabilities of R from other languages without initializing R.",

"index": "1",

"Explanation": "Rserve is a TCP/IP server which allows other programs to use functions and capabilities of R from other languages without the need to initialize R. Rserve cannot perform implementation of various data mining algorithms. That can be done using Rstudio which is an IDE for R."

},

{

"question":"Q7. Tableau is a business intelligence and analytics tool used for ?",

"choice": ["Graphically analyze structured data", "Produce charts, graphs and interactive dashboards for a large data set", "Both"],

"answer": "Both",

"index": "3",

"Explanation": "Tableau is a very useful tool for graphically analyzing structured data. It uses interactive dashboards to produce charts and graphs."

},

{

"question":"Q8. To implement K-Means algorithm in Weka, what are the available distance functions? ",

"choice": ["ManhattanDistance", "EucledianDistance", "Both"],

"answer": "Both",

"index": "3",

"Explanation": "Weka uses ManhattanDistance and EucledianDistance as the distance function to find out the similar objects to store them in clusters."

},

{

"question":"Q9. Among the following data mining tool, which one requires less memory to install and thus works faster? ",

"choice": ["Weka", "Rapid Miner", "Both"],

"answer": "Weka",

"index": "1",

"Explanation": "Weka uses less packages as compared to Rapid Miner. Also, Rapid Miner supports more algorithms than Weka. Thus, Rapid Miner requires much more memory to operate."

},

{

"question":"Q10. What are some of the dashboard actions in Tableau ",

"choice": ["URL,Highlight,Filter", "Set,Reset", "None"],

"answer": "URL,Highlight,Filter",

"index": "1",

"Explanation": "Tableau uses \"Highlight\",\"URL\" and \"Filter\" as some of the dashboard actions."

},

{

"question":"Q11. Is there any package to integrate Rapid Miner or Weka with Tableau like R Integration",

"choice": ["Yes", "No", "Not possible"],

"answer": "No",

"index": "2",

"Explanation": "Currently there is no package to integrate Tableau with Rapid Miner or Weka."

},

{

"question":"Q12. R uses a command-driven interface to perform operations?",

"choice": ["True", "False", "Not possible. R uses graphical objects"],

"answer": "True",

"index": "1",

"Explanation": "Unlike Rapid Miner or Weka, R is based on extensive coding to perform simplest tasks of data analysis based on C Language."

}

];

console.log(q.length);

document.getElementById("quest").innerHTML=q[0].question;

document.getElementById("choice1").innerHTML=q[0].choice[0];

document.getElementById("choice2").innerHTML=q[0].choice[1];

document.getElementById("choice3").innerHTML=q[0].choice[2];

$("#click").click( function()

{

var value =$(this).text();

if(value =="NEXT" || value =="SUBMIT")

{

next();

}

else

check();

});

function check()

{

if(no\_ans\_seleted)

{

var test = "#choice" + q[i].index;

console.log(test);

$(test).css("background-color","green");

}

else if( selected\_answer.text() == q[i].answer)

{

selected\_answer.css("background-color","green");

++correct;

}

else

{

selected\_answer.css("background-color","#FF3333");

var test = "#choice" + q[i].index;

console.log(test);

$(test).css("background-color","green");

wrong++;

}

$("#click").html('NEXT');

$("#button1").show();

$("#button2").show();

document.getElementById("expl1").innerHTML=q[i].question;

document.getElementById("expl2").innerHTML=q[i].Explanation;

if(i+1 == q.length)

{

$("#click").html('SUBMIT');

}

};

function next()

{

if(i+1 == q.length)

{

$("#result").html("<h4 style='color:#8A5C2E'><span class='glyphicon glyphicon-education'></span> Quiz on Data Mining Tools</h4><hr><u>Here is Your Score!</u><br><h5 style='color:green'>Correct: " + correct +"</h5><h5 style='color:#FF3333'>Wrong: "+ wrong + "</h5>");

$("#result").css("h4","#8A5C2E");

$("#total").css("margin-top", "20px");

$("#result").css("padding-left", "20px").css("padding-right", "20px");

$("#quest").css("display","none");

$("#choice1").css("display","none");

$("#choice2").css("display","none");

$("#choice3").css("display","none");

$("#button1").css("display","none");

$("#click").css("display","none");

$("#break1").css("display","none");

}

else

{

$("#click").html('CHECK ANSWER');

console.log($("#click").text());

no\_ans\_seleted=true;

if(i+1 < q.length)

{

$("#choice1").css("background-color","#FFF");

$("#choice2").css("background-color","#FFF");

$("#choice3").css("background-color","#FFF");

$("#button1").hide();

$("#button2").hide();

console.log("Next button pressed!");

document.getElementById("quest").innerHTML=q[i+1].question;

document.getElementById("choice1").innerHTML=q[i+1].choice[0];

document.getElementById("choice2").innerHTML=q[i+1].choice[1];

document.getElementById("choice3").innerHTML=q[i+1].choice[2];

i++;

}

}

};

$("#choice1, #choice2, #choice3").click(function()

{

no\_ans\_seleted=false;

console.log("Clicked");

$("#choice1").css("background-color","#FFF");

$("#choice2").css("background-color","#FFF");

$("#choice3").css("background-color","#FFF");

selected\_answer = $(this);

$(this).css("background-color","#F1F1F1");

});

}); // End of document.ready function

</script>

REFERENCES

[1] *CSc 177 Data Warehousing and Data Mining Course Website*, Available:

<http://athena.ecs.csus.edu/~mei/177/csc177.html> in February 2015.

[2] *Weka - About*, Available: <https://weka.waikato.ac.nz/dataminingwithweka/preview> in March 2015.

[3] *Rapid Miner*, Available: <https://rapidminer.com/> in February 2015.

[4] *R*, Available: <http://www.r-project.org/> in June 2015.

[5] *Tableau,* Available: <http://www.tableau.com/> in May 2015.

[6] Rangra et al., “Comparative Study of Data Mining Tools”, in *International Journal of Advanced Research in Computer Science and Software Engineering*, pages 216-223, June 2014.

[7] Luis C. Borges et al., “Comparison of Data Mining Techniques and tools for data classification”, in *C3S2E ’13 Proceedings of the International C\* Conference on Computer Science and Software Engineering*, pages 113-116, 2013.

[8] Mesut Ozkan, “The Comparison of Data Mining Tools”, Department of Computer Engineering Istanbul Kultur University, November 2013.

[9] Abdullah H. Wahbeh et al., “A Comparison Study between Data Mining Tools over some Classification Methods”, in *International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence*, pages 18-26, 2011.

[10] *Bootstrap*, Available: <http://getbootstrap.com/> in June 2015.

[11] *Data Mining*, Available: <https://en.wikipedia.org/wiki/Data_mining> in February 2015.

[12] *Data Mining definition,* Available: <http://techathon.mytechlabs.com/what-is-data-mining/> in August 2015.

[13] *Rapid Miner*, Available: <https://en.wikipedia.org/wiki/RapidMiner> in June 2015.

[14] *Rapid Miner Manual*, Available:

<http://docs.rapidminer.com/downloads/RapidMiner-v6-user-manual.pdf> in August 2015.

[15] Trilok Chand Sharma et al., “WEKA Approach for Comparative Study of Classification Algorithm”, in *International Journal of Advanced Research in Computer and Computer Engineering Vol. 2, Issue 4*, pages 1925-1931, April 2013.

[16] Harshwardhan Solanki, “Comparative Study of Data Mining Tools and Analysis with Unified Data Mining Theory”, in *International Journal of Computer Applications* (0975-8887), Volume 75, No.16, pages 23-28, August 2013.

[17] *Weka*, Available: <https://en.wikipedia.org/wiki/Weka_(machine_learning)> in May 2015.

[18] *Tableau- About*, Available: <https://en.wikipedia.org/wiki/Tableau_Software> in August 2015

[19] *R*, Available: <http://www.r-statistics.com/about/> in August 2015.

[20] Sharon Christa et al., “A Comparative Analysis of Data Mining Tools in Agent Based Systems”, [arXiv: 1210.1040v1](http://arxiv.org/abs/1210.1040v1), Cornell University Library, October 2012.

[21] *Highslide JS*, Available: <http://highslide.com/download.php> in July 2015.

1. Classification and Clustering algorithms: <http://iasir.net/IJEBEApapers/IJEBEA13-140.pdf> [↑](#footnote-ref-1)
2. See Appendix [↑](#footnote-ref-2)
3. Case Study: <http://athena.ecs.csus.edu/~srilatha/> [↑](#footnote-ref-3)
4. Comparative Study Website: <http://athena.ecs.csus.edu/~akbhinge/> [↑](#footnote-ref-4)
5. List of data mining tools: <http://www.predictiveanalyticstoday.com/top-free-data-mining-software/> [↑](#footnote-ref-5)
6. Classification and Clustering algorithms: <http://iasir.net/IJEBEApapers/IJEBEA13-140.pdf> [↑](#footnote-ref-6)
7. See Appendix [↑](#footnote-ref-7)
8. Data Mining algorithms: <https://msdn.microsoft.com/en-us/library/ms175595.aspx> [↑](#footnote-ref-8)
9. See Appendix [↑](#footnote-ref-9)
10. See Appendix [↑](#footnote-ref-10)