

**School of Electrical Engineering and Computer Science,
National University of Sciences and Technology (NUST)**

Gamified Online WEKA

A Thesis

By

Asma Nisar

NUST201362759MSEEC61413F

Department of Computing

Submitted in partial fulfillment of the requirements

for the degree of

Master of Science, Computer Science

July 2016

Approval

The undersigned have examined the thesis entitled '**Gamified Online WEKA**' presented by **ASMA NISAR**, a candidate for the degree of **Master of Science (Computer Science)** and hereby certify that it is worthy of acceptance.

Date

Dr. Muhammad Moazam Fraz

Date

Dr. Anis ur Rahman

Date

Dr. Muhammad Muneeb Ullah

Date

Dr. Mian Muhammad Hamayun

Certificate of Originality

I hereby declare that this submission titled “Gamified Online WEKA” is my own work. It does not contain any material previously published or written by another author except with reference was made in this thesis report. Every person whether from NUST SEECS or anywhere else who contributed in this research work, has been acknowledged explicitly in this thesis.

I also declare that the content and prototype of this research is my own work except for the help and support from others in design, development, presentation and content writing with has been duly acknowledged. I have also verified the originality of content using Turnitin.

Author Name: Asma Nisar
Signature:

Abstract

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Data generated by different industries needs to be analyzed and summarized to help in the growth of business. Data mining applications are widely used in direct marketing, health industry, ecommerce, customer relationship management (CRM), telecommunication industry and financial sector.

WEKA is one of the most commonly used open source data mining tool. Its Java API is freely available. So, it can be embedded in any java based software. It is constantly improving and new versions are being released since 2004. Its architecture is very simple and it can read data from 10 different file formats, URL of file and relational database. WEKA has more than 220 algorithms for different types of data processing (preprocess, classification, Clustering etc).

Desktop based WEKA needs installation and configuration. It uses system resources and has a maximum heap size limit. All the visualizations (trees, bar charts and scatter plots) are static in WEKA and it doesn't convey the required information. Results of the filters are displayed in a very user unfriendly way.

As a solution to this problem we have created "Gamified Online WEKA". It has improved the visualizations of WEKA. As to use it user would need to upload and store private data on server, it maintains separate and secure user accounts and stores their data for later use. With its collaborative environment, data security, data storage, interactive visualizations and all the algorithms of WEKA, "Gamified Online WEKA" is a complete data mining solution.

Acknowledgments

First of all Thanks to Allah Almighty for giving me strength to undertake and complete this thesis project. After that I would take this opportunity to thank my family for their prayers, encouragement and support to complete this work. Special thanks to my husband who was always there to encourage me throughout the process.

I would like to express my gratitude to my supervisor Dr. Muhammad Moazam Fraz whose guidance, support and patience added considerably to complete this thesis. I am highly honored to be monitored by him. I am grateful for his valuable advices and moral support throughout the thesis.

I am grateful to my thesis committee members for guiding me in many ways. Thanks to Dr. Mian M. Hamayun for his guidance and suggestions on thesis writeup on PTS and proposal defense. I would also like to thank Dr. Muhammad Muneeb Ullah and Dr. Anis ur Rahman for taking out time for my thesis and for giving valuable suggestions on refining my research and thesis topic.

I must also acknowledge Dr. Khalid Latif who helped me through research and implementation of thesis idea. Another person who deserves to be appreciated is Mr. Rizwan Mushtaq who helped me in hectic task of WEKA API exploration.

Asma Nisar

Table of Contents

Approval	ii
Certificate of Originality.....	iii
Abstract.....	iv
Acknowledgments.....	v
Table of Contents	vi
List of Tables	ix
List of Figures	x
Chapter 1 : Introduction	1
1.1 Motivation:.....	2
1.2 Challenges:.....	2
1.3 Objectives:	2
Chapter 2 : Background and Literature Review	4
2.1 Cloud Computing:.....	4
2.1.1 Cloud Computing stack:	4
2.2 Data Mining Tools:	6
2.2.1 WEKA:	6
2.2.2 Limitations of WEKA:.....	7
2.3 Comparison of WEKA with Desktop based Datamining tools:	8
2.3.1 Comparison of WEKA with R:.....	8
2.3.2 Comparison of WEKA with Orange:.....	9
2.3.3 Summary:.....	10
2.4 Comparison of WEKA with Web based Datamining tools:	11
2.4.1 Comparison of WEKA with IBM SPSS & Cognos:.....	11
2.4.2 Comparison of WEKA with Tableau:.....	11
2.5 Visualization libraries:	12
2.5.1 Data Visualization Libraries comparison:	13

2.5.2 Selected Data Visualization Libraries:.....	14
Chapter 3 : METHODOLOGY.....	16
3.1 Research Methodology Types:	16
3.1.1 Types of Research by Application:.....	17
3.1.2 Types of Research by Objectives:.....	17
3.1.3 Types of Research by Information Sought:	18
3.2 Thesis Research Workflow:.....	18
3.2.1 Define a Research Area:	19
3.2.2 Literature Review:	19
3.2.3 Identify Research Problem:.....	20
3.2.4 Prepare Research Design:	20
3.2.5 Prototype Implementation:.....	21
3.2.6 Testing:	25
Chapter 4 : RESULTS	26
4.1 Chosen dataset:	26
4.2 Account Creation:	26
4.3 Data Loading:.....	27
4.3.1 Data Loading From File:.....	28
4.3.2 Data Loading From Database:	29
4.3 Preprocessing:	30
4.3.1 Bar Charts:	31
4.3.2 Data Editor:	34
4.4 Classify:	36
4.4.1 Classify Result Panel:	37
4.4.2 Time Taken to build Classifier Model:	38
4.4.3 Classification Result Comparison:.....	39
4.4.4 Classification Tree Comparison:.....	39
4.5 Cluster:	41
4.5.1 Cluster Result Comparison:	41
4.6 Associate:	42

4.7 Select Attributes:.....	44
4.8 Visualize:	45
Chapter 5 : CONCLUSION	47
5.1 Conclusion:	47
5.2 Future directions:	48
Appendix A.....	50
Installation and Configuration of WEKA:	50
A.1 Required Softwares	50
A.2 Detailed Steps	50
Appendix B	56
Datamining with WEKA Course:	56
B.1 WEKA Book:	56
B.2 WEKA Course:	56
Bibliography	57

List of Tables

Table 2.1 Datamining tools Comparison	10
Table 2.2 Visualization Libraries Comparison	13
Table 2.3 Tree Visualization Libraries Comparison.....	14

List of Figures

Figure 1.1 Knowledge Discovery and Data Mining (KDD).....	1
Figure 2.1 Cloud Computing stack.....	5
Figure 2.2 WEKA Classification Tree.....	7
Figure 2.3 WEKA bar chart.....	8
Figure 2.4 R Datamining.....	9
Figure 2.5 Orange datamining.....	10
Figure 2.6 Tableau Datamining.....	12
Figure 2.7 Highcharts Visualization.....	15
Figure 2.8 D3 Visualization.....	15
Figure 3.1 Types of Research.....	16
Figure 3.2 Thesis Workflow.....	19
Figure 3.3 Use Case Diagram.....	21
Figure 3.4 Prototype Design and Architecture.....	24
Figure 4.1 Gamified Online WEKA Login.....	27
Figure 4.2 Gamified Online WEKA "Load Data" Tab.....	28
Figure 4.3 Gamified Online WEKA Data Loading from File.....	29
Figure 4.4 Gamified Online WEKA Data Loading from Database.....	30
Figure 4.5 WEKA "Preprocess" Tab.....	30
Figure 4.6 Gamified Online WEKA "Preprocess" Tab.....	31
Figure 4.7 WEKA Bar Charts.....	31
Figure 4.8 Gamified Online WEKA Bar Charts.....	32
Figure 4.9 Gamified Online WEKA Bar Chart Disabled class value.....	32

Figure 4.10 Gamified Online WEKA Bar Chart Distinct Intervals	33
Figure 4.11 Gamified Online WEKA Bar Chart Image Download Options	33
Figure 4.12 Gamified Online WEKA Bar Chart Tooltip.....	34
Figure 4.13 WEKA Data Editor.....	35
Figure 4.14 Gamified Online WEKA Data Editor.....	35
Figure 4.15 WEKA Classifier.....	36
Figure 4.16 Gamified Online WEKA Classifier.....	36
Figure 4.17 WEKA Classify Result Panel.....	37
Figure 4.18 Gamified Online WEKA Classify Result Panel.....	38
Figure 4.19 WEKA Classifier Model	38
Figure 4.20 Gamified Online WEKA Classifier Model	38
Figure 4.21 WEKA Classification Result.....	39
Figure 4.22 Gamified Online WEKA Classification Result.....	39
Figure 4.23 WEKA Classification Tree.....	40
Figure 4.24 Gamified Online WEKA Classification Tree.....	40
Figure 4.25 WEKA Cluster Result	41
Figure 4.26 Gamified Online WEKA Cluster Result	41
Figure 4.27 NumericToNominal filter selection.....	42
Figure 4.28 WEKA Associator.....	43
Figure 4.29 Gamified Online WEKA Associator	43
Figure 4.30 WEKA Attribute Selection.....	44
Figure 4.31 Gamified Online WEKA Attribute Selection.....	44
Figure 4.32 WEKA Visualize.....	45

Figure 4.33 Gamified Online WEKA Visualize46

Chapter 1 : Introduction

Knowledge Discovery and Data Mining (KDD) is the effective process of finding hidden useful information (knowledge) from raw data. Rapid increase in data generation by individuals and industry has increased the need of Data Mining and Data Visualization. It helps companies and other organizations in taking strategic decisions and on the basis of trends they may take corrective critical actions in future.

Data Mining and Data Visualization is widely used in internet marketing, Customer relationship management, telecommunication and finance industry. Figure 1 shows step by step process from data to information.

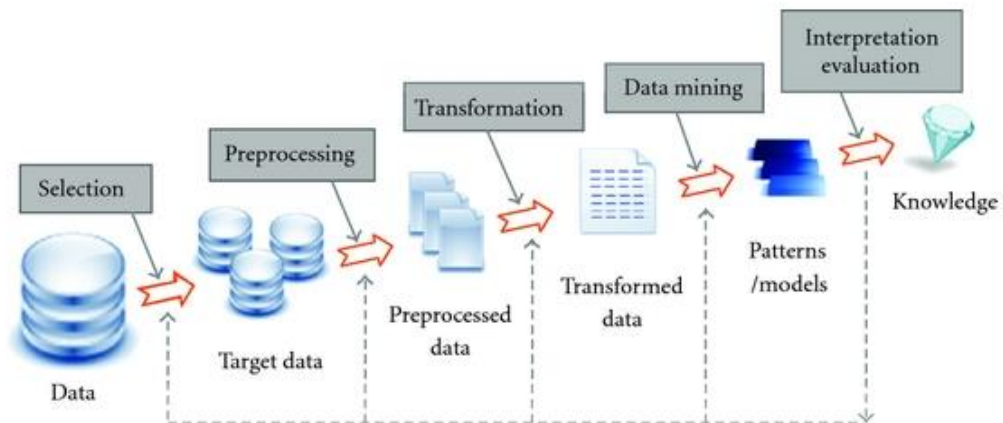


Figure 1.1 Knowledge Discovery and Data Mining (KDD)

Most of the Data Mining tools available are desktop based which need installation and configuration and have memory space issues while most of the web based tools are locally domain specific. WEKA is the 2nd most popular Data Mining and

Visualization tool. It is free under the GNU General Public License and has plenty of data analysis and predictive modeling algorithms. It is also desktop based software with memory issues and poor visualization. The goal of this thesis is to provide a web based data processing and visualization solution by integrating data mining API and visualization libraries.

1.1 Motivation:

WEKA is one of the most commonly used Data Mining tool. It can be more beneficial if it has better visualizations, easy accessibility and more processing power. This can be done by implementing it to web and using better visualization libraries.

1.2 Challenges:

We faced the following challenges during Implementation:

- Explore and use WEKA API.
- Explore interactive visualization tools which best complement our application.
- Integrate visualization libraries.
- Development of User Friendly UI.

1.3 Objectives:

These objectives have been successfully achieved

- Implementing full functionality of WEKA Explorer.
- Identifying the flaws in WEKA UI and visualizations.

- Identifying and implementing interactive visualization techniques.
- Designing user friendly interface and displaying results of filters in a more understandable way.
- Providing session maintenance and secure way to save data on server and reuse it.

Chapter 2 : Background and Literature Review

As our research revolves around cloud computing and data processing tools so in this chapter we will review these topics in detail. We want to present a datamining software as a service so we will go in details of cloud computing specifically Software as a Service (SaaS).

There are many datamining tools available both desktop based and web based. We will compare those and will choose the best to work with. We will also discuss and compare tools and technologies needed for thesis implementation purpose.

2.1 Cloud Computing:

Cloud computing can be described as a model for enabling access to the shared computing resources. End user can utilize a part of bulk resources available at the shared place. The term Cloud Computing is also used as the representation for internet. In Cloud computing applications do not run on local machine (PC, tablet etc) instead they run on a server and carry out assigned tasks through network connection. Its “as a Service” paradigm is low cost and elastic scaling.

2.1.1 Cloud Computing stack:

Three different categories in Cloud Computing are Software as a Service, Platform as a Service and Infrastructure as a Service.

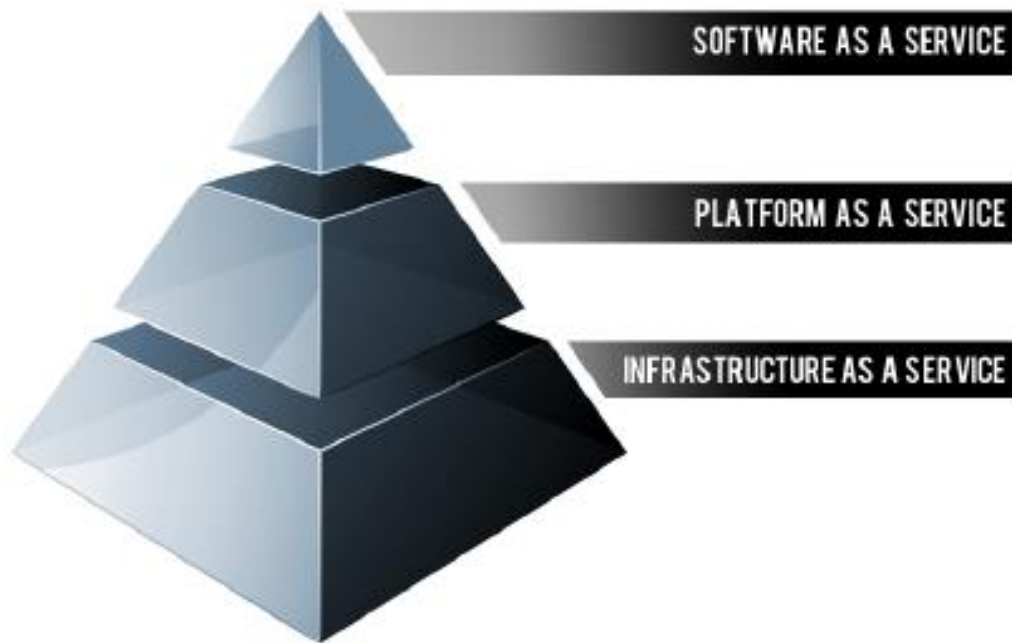


Figure 2.1 Cloud Computing stack

Top layer is Software as a Service (SaaS). SaaS applications are designed to serve a specific purpose and delivered to the user through web. Platform as a Service (PaaS) is used to deploy SaaS applications. Infrastructure as a Service (IaaS) is the software and hardware used to provide storage, network and operating system to the upper layers. We are going to present our thesis prototype as SaaS application. So, we will go in more detail of SaaS.

SaaS can be a service on demand generating revenue through users (getting paid) or through advertisements or user list sales. Recent reports predict a double digit growth of SaaS. Some of the characteristics of SaaS are

1. It is easily accessible from anywhere.
2. It follows a One to Many Model as it is centrally located and many people can use the application at once.

3. User does not require handling software installation and upgrades.

2.2 Data Mining Tools:

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining applications are widely used in direct marketing, health industry, ecommerce, customer relationship management (CRM), telecommunication industry and financial sector. Some of the most commonly used desktop based datamining and visualization apps are WEKA, R and Orange. While there are also some web based data mining applications available like IBM Cognos, IBM SPSS and for data visualization Tableau is considered top most data visualization business intelligence solution.

2.2.1 WEKA:

According to Wikato University New Zealand:

“Weka is a collection of machine learning algorithms for data mining tasks. 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.”

This software was originally developed as a mix of Tcl/Tk, C, and Makefiles in 1993 but then it was redeveloped in JAVA in 1997. The most recent stable release of WEKA is 3.8. It can load data from file, URL and database and it supports more than 10 file formats. ARFF is the default and most used data file format for WEKA. It has more than 220 data processing algorithms and many sorts of data visualizations for visualizing actual and processed data and results.

2.2.2 Limitations of WEKA:

WEKA needs installation and configuration on local machine. It uses system resources for processing and has a maximum heap size limit. It can only handle datasets with a certain size limit otherwise throw “out of memory” exception. User needs to have the required datasets on the local machine he is working with.

All the visualizations (bar charts, trees and scatter plots) are not up to the mark in WEKA and do not convey the required information and user cannot filter out the data in visual representations. Results of the filters are displayed in a very clumsy way. User has to go through the whole text to find out the desired section.

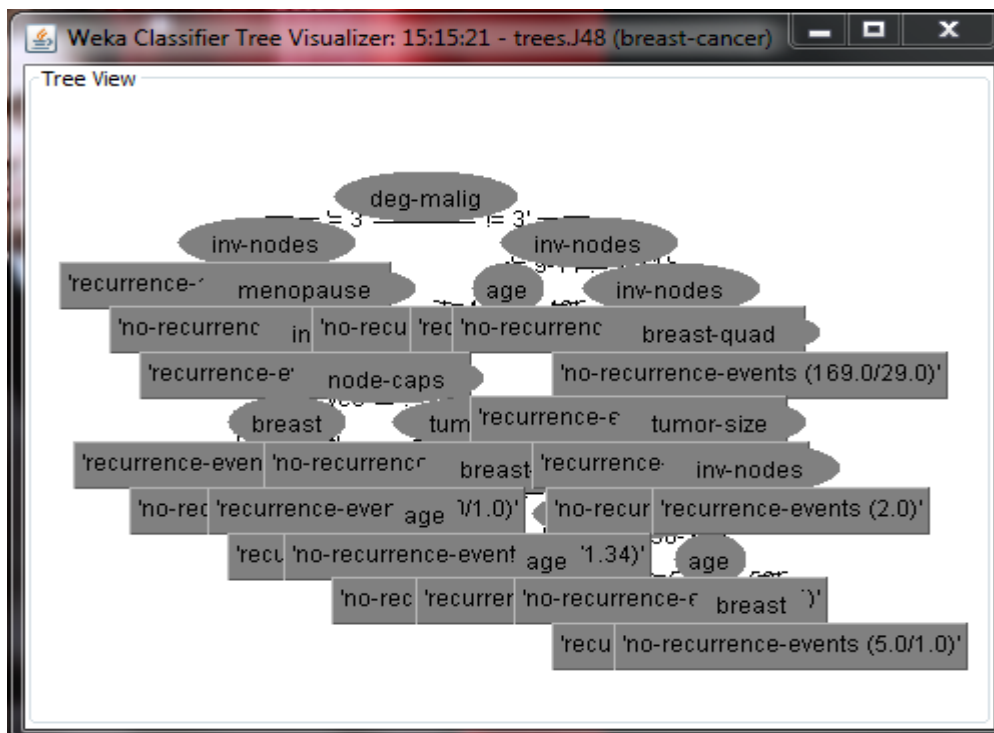


Figure 2.2 WEKA Classification Tree

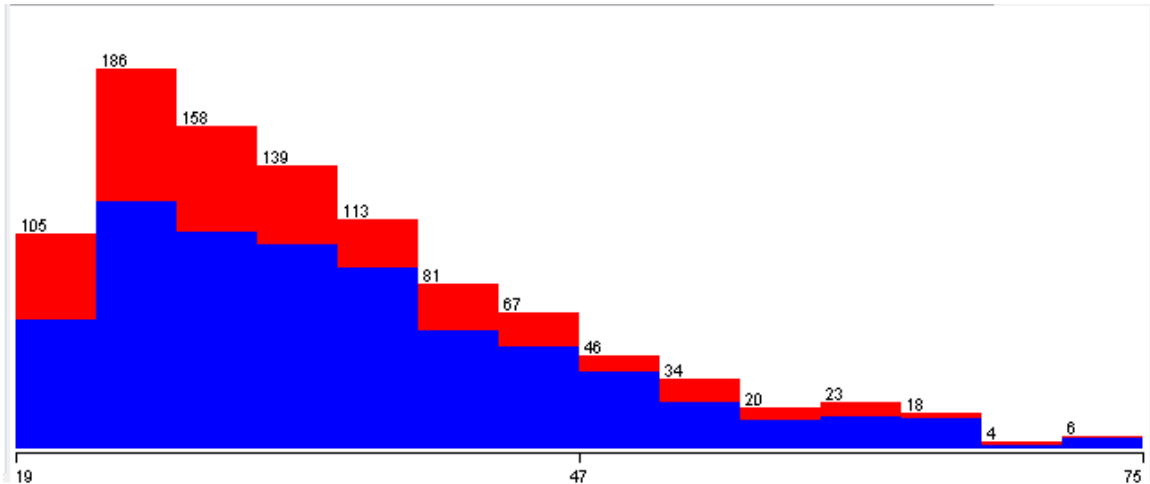


Figure 2.3 WEKA bar chart

2.3 Comparison of WEKA with Desktop based Datamining tools:

As mentioned earlier there are many desktop based and web based datamining tools available. Now we will compare those with WEKA to get more understanding of their functionality and flaws.

2.3.1 Comparison of WEKA with R:

Most of R's modules are written in R programming language but R is primarily written in FORTRAN and C. It is basically open source but R-Cloud is its paid version. Its core focus is on statistical computing and graphics. It is easy to use and extensible but as it is language specific the user needs to know R in order to extend it or to use its command line. While WEKA is written in JAVA and has an API available with documentation which can easily be embedded to any application.

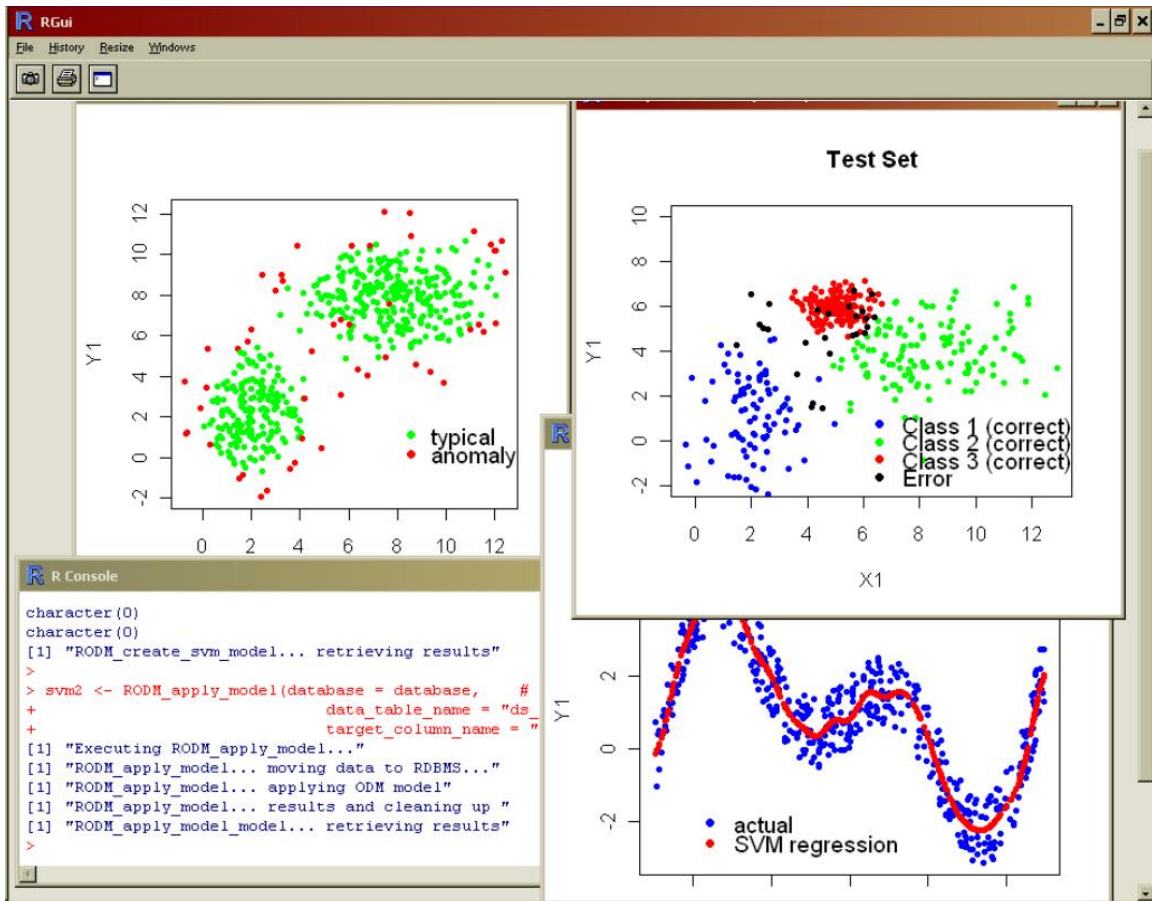


Figure 2.4 R Datamining

2.3.2 Comparison of WEKA with Orange:

Like WEKA and R Orange is also open source and platform independent. It is usually used for data analysis and visualization. It is written in python and works through python scripting. It is not very useful if you want a deep down analysis. Its visual programing interface has also been upgraded to support service oriented architecture. It is most commonly used in health sector for predictive analysis.

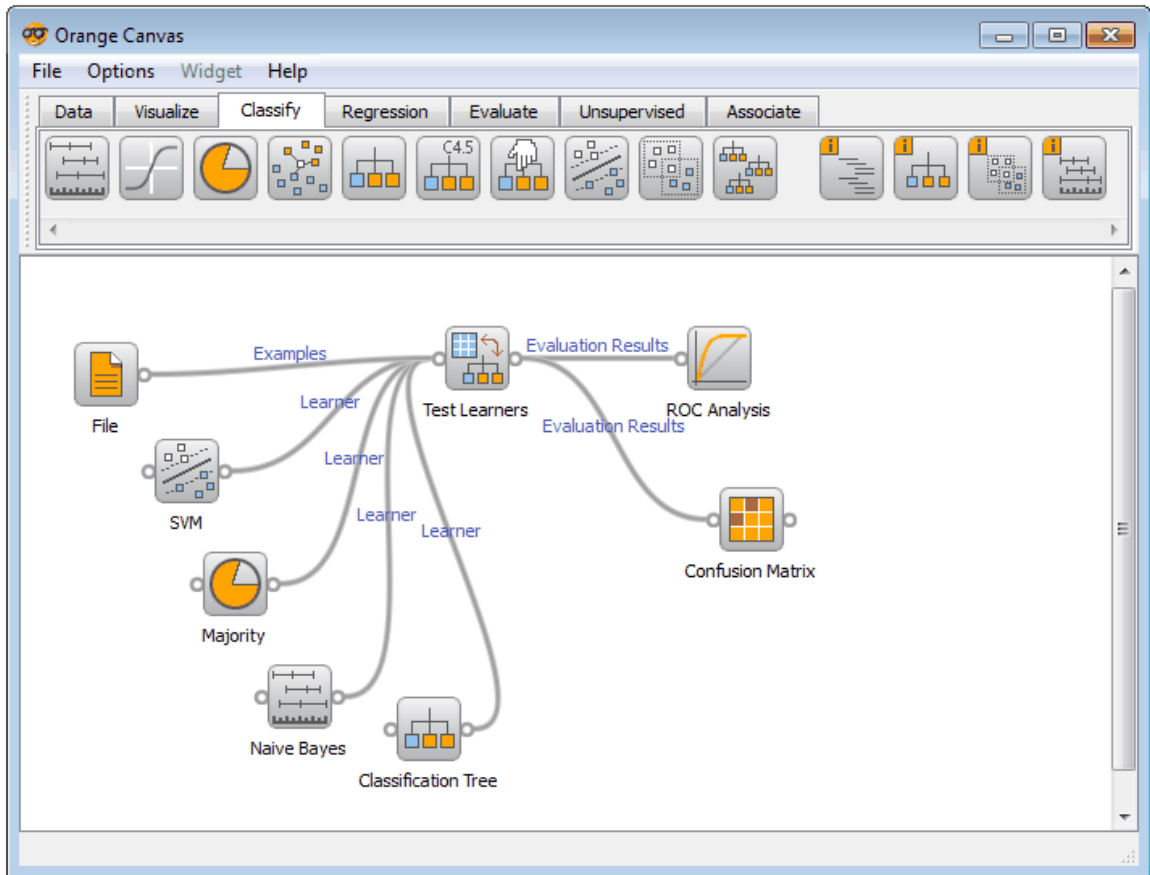


Figure 2.5 Orange datamining

2.3.3 Summary:

<u>WEKA</u>	<u>R</u>	<u>Orange</u>
Developed in java	Developed in R programming language.	Developed in Python.
Open source	Open source	Open source
Data visualization and analysis.	Graphics visualization	Data visualization and analysis
Collection of machine learning algorithms	Statistics, Text mining, Clustering	Scripting interface
Execute data files in multiple format and Database connection using jdbc		Has its own data format

Table 2.1 Datamining tools Comparison

2.4 Comparison of WEKA with Web based Datamining tools:

As there are some web based datamining tools available so it is very important to see why we need a new web based datamining tool. What are the flaws in existing web based applications which we can fix in web based WEKA.

2.4.1 Comparison of WEKA with IBM SPSS & Cognos:

IBM applications are not easy to use. They require deployment and configuration. One needs to buy them to deploy on their local network for internal use of business. IBM SPSS is a pure business analysis tool which is usually used by the companies for future trend predictions. It provides a user interface for user to interact with the predictive and datamining algorithms without any sort of programming. It has implemented algorithms for entity analytics, text analytics, decision management and optimization. IBM Cognos is a business intelligence suite which is used for business analytics, report generation, and score carding. It can generate both online and offline reports.

2.4.2 Comparison of WEKA with Tableau:

Tableau is a visualization based product with a thought that information is more understandable if it is in the form of visualizations. According to a report the world will generate 50 times more data in 2020 as compared to 2011. As the data increases it gets hard to analyze it and to present the generated information in a more useable way. It has the ability to present the data in many ways through different useful visualizations.



Figure 2.6 Tableau Datamining

The purpose to discuss it here is to realize what the world needs. In WEKA there are many sort of data analysis but they are useful only if we are able to present the results in a more understandable way. So, we need to incorporate Gamification (dynamic visualizations) to attract more users to use it.

2.5 Visualization libraries:

As we discussed earlier that we are using SaaS as a model for “Gamified Online WEKA” so we need to pick those visualization libraries which are compatible with it. A very important aspect of software as a service is load balancing between server and client. We are integrating WEKA API on server side for processing data so it will be a good option to use a client side technology to display results for load balancing purpose.

We need three different types of visualizations (bar chart, trees and scatter plot) to implement the WEKA visualization completely. There are many JavaScript based

visualization libraries which can be used for this purpose but we will choose the best suiting our requirements.

2.5.1 Data Visualization Libraries comparison:

Some of the most used JavaScript visualization libraries are highcharts, zingchart, fusioncharts, plotly, google charts and d3. All of them can be used to create different sort of visualizations. We will compare them based of their functionality and type of visualization we need.

For Stacked Bar Chart and Scattered Plot:

Attributes	<u>Highcharts</u>	<u>Zingchart</u>	<u>Fusioncharts</u>	<u>Plotly</u>	<u>Google charts</u>	<u>D3</u>
Takes json object as input	Yes	Yes	Yes	Yes	Yes	Yes
Print and download	Yes	No	No	No	No	No
Zooming	No	No	No	Yes	No	No
Is interactive (bar charts)	Yes	No	No	Yes	No	No
Is interactive (scatter plot)	Yes	Yes	Yes	No	Yes	No
Supports color class attribute	Yes	Yes	Yes	No	No	No

Table 2.2 Visualization Libraries Comparison

The above table shows the comparison of visualization libraries for the stacked bar chart and scatter plot on the basis of attributes which are more important from WEKA gamification point of view. As it is clear that **highcharts** library has most of these attributes present so we will prefer this over others.

For Tree Visualization:

For tree visualization the most important attributes are to be in the form of tree as some of the visualization tools display trees in the form of lists. Secondly it should be collapsible as if the tree is large it gets messy when fully expanded so user should have the option to expand or collapse a tree.

Attributes	<u>D3</u>	<u>Dhtml tree</u>	<u>jsTree</u>
Takes json object as input	Yes	Yes	Yes
Is in the form of tree	Yes	No	No
Is collapsible	Yes	No	No

Table 2.3 Tree Visualization Libraries Comparison

It is clear from this table that D3 is the best option for us to implement our classification trees as it provides the required attributes.

2.5.2 Selected Data Visualization Libraries:

Highcharts:

Highcharts is famous for creating many types of interactive charts and graphs. It provides the functionality of filtering out the data on the run. Moreover it provides the option to print and download the visualizations in four different formats. It uses distinct colors and shapes to make the visualizations more understandable.

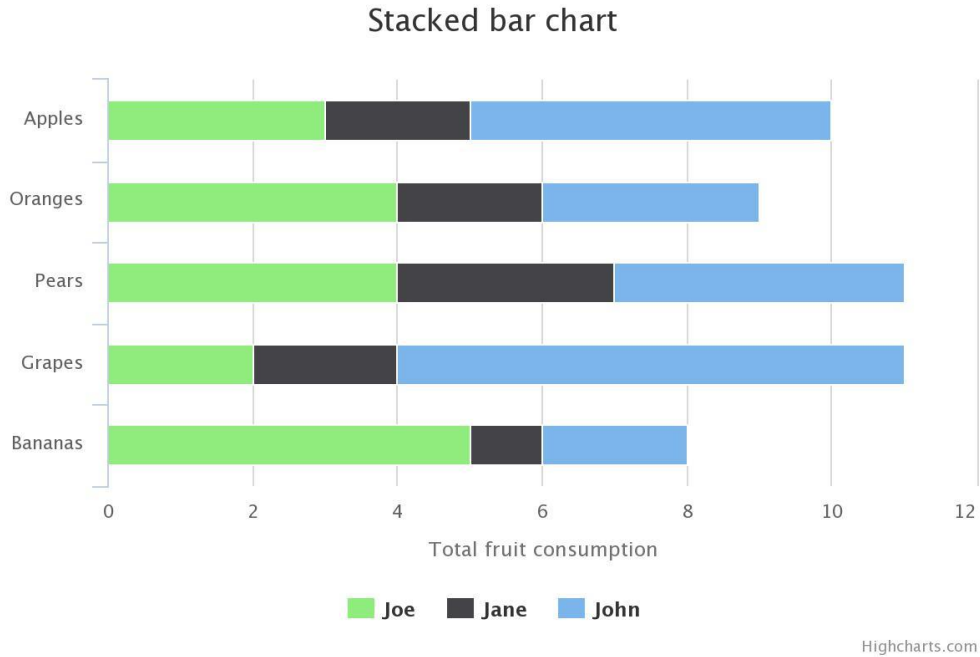


Figure 2.7 Highcharts Visualization

D3:

In WEKA the visualization of classification trees is not good. When number of nodes increase they start overlapping and tree gets messed up. D3 trees are collapsible and do not overlap. User can expand the required nodes while others remain unexpanded.

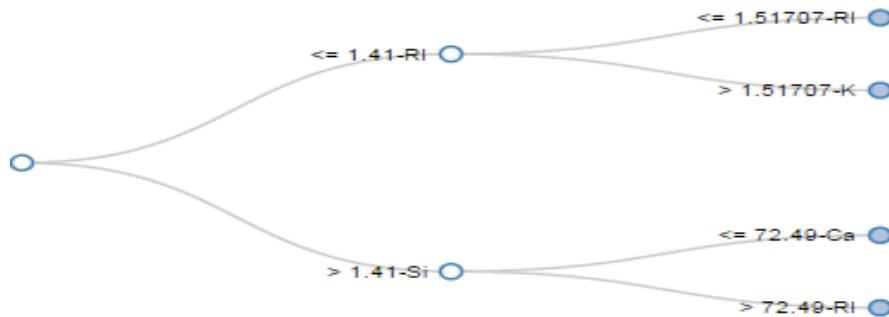


Figure 2.8 D3 Visualization

Chapter 3 : METHODOLOGY

This chapter presents methodology used to carry out this research work.

Research methodology can be categorized into different types. In this chapter we will discuss different research methodologies and which methodology we used and why.

3.1 Research Methodology Types:

Types of research based on different categories are listed below.

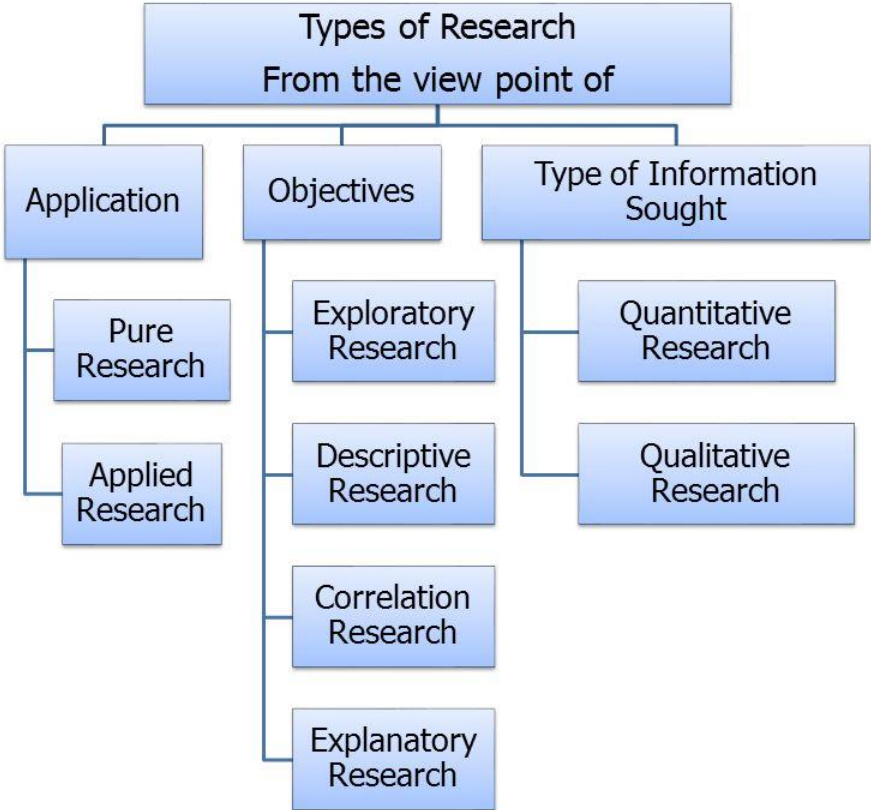


Figure 3.1 Types of Research

3.1.1 Types of Research by Application:

Pure research also known as fundamental research is carried out just to explore a research area and to create new information. Its focus is to gather fundamental knowledge, explore it in detail and generate new information in that specific area.

On the other hand **Applied research** focuses on developing a solution for a particular problem. It is most commonly used to find solution for industrial problems. Applied research has been used for this thesis problem because we are developing a system to overcome the flaws in an existing datamining tool (WEKA).

3.1.2 Types of Research by Objectives:

Exploratory research is the informal research used to gain background information about the research problem. It clarifies and defines the nature of the problem. It is used to set a ground for further research.

Descriptive research is used to describe the characteristics of a phenomenon. It is usually undertaken after having background knowledge of the problem. It can be used to determine the perception of product characteristics. We have partially used this in this thesis to clarify the problems users were having with existing WEKA. We used personal interaction and internet survey for this purpose.

Correlation research is used is there to determine the relationships between different variables. It is carried out to test the reliability and predictive validity of a product. Explanatory research is carried out to determine the accuracy of a theory, exploring the related knowledge and building a new theory based on it.

3.1.3 Types of Research by Information Sought:

Quantitative research can be applied to every phenomenon which can be expressed in terms of quantity while **Qualitative research** deals with the quality like actions and behaviors of a product. We have used this research method to determine the quality of existing software and predicted the possible solution.

3.2 Thesis Research Workflow:

A hybrid research methodology has been used to carry out this research work as different research techniques have been combined to accomplish this. These are the steps used to formalize our research methodology:

- i. Explore existing WEKA and identify the issues and challenges.
- ii. Narrow down the study to pinpoint the areas which need to be improved.
- iii. Explore existing solutions to specified problem and how we can use them.
- iv. Derive a hypothesis based on the existing literature and propose a solution.
- v. Validate the proposed hypothesis by implementing Gamified Online WEKA.

We conducted the systematic analysis on the existing software focused on the space, accessibility and visualization issues of WEKA. Our research comprises of different steps. Here is a workflow of our research work.

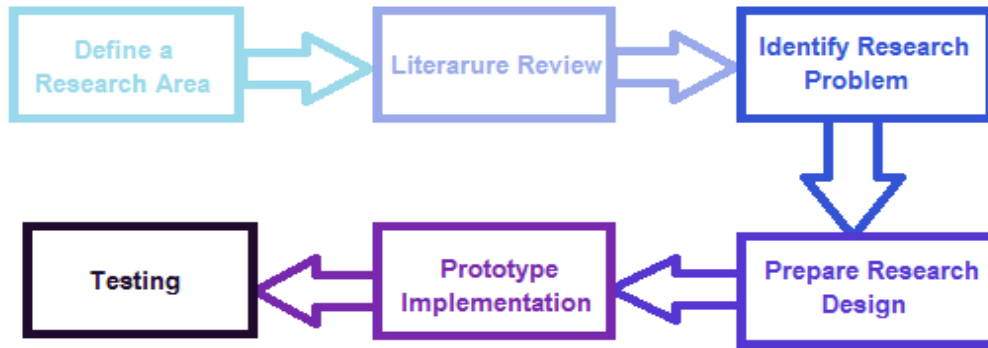


Figure 3.2 Thesis Workflow

3.2.1 Define a Research Area:

This research started with defining the research area. By taking the current trends into consideration data science is the hot topic of 21st century as the data generated and retained by the companies has been massively increased over a decade. Data science is about data analysis and data modeling.

Cloud computing is another most talked about and most used technology these days. Cloud based applications also called Software as Service (SaaS) run on server, use server’s computing power and display results to users through their web browsers. Based on the discussion above we defined our research area to be “Combining Data Science and Cloud computing to generate a Processing tool which is easy to access and has more computational power”.

3.2.2 Literature Review:

To identify a particular research problem we had to go through all the available data mining and visualization tools. There are some web based tools but they are not as good at computing data as some of the desktop based tools. WEKA is one of the best data

mining tools which supports many data formats and have a strong set of algorithms for all sort of data processing. But being desktop based it has some installation issues and data processing limitations. Another thing we noticed it does not have really good visualization.

In this age when visualizing data and processing results is vital for any data processing software even some products like tableau are purely visualization based, It is important for WEKA to not only process data well but also to visualize it in a more user friendly way.

3.2.3 Identify Research Problem:

After extensive research on datamining tools and web based technologies we defined our Research problem as “To provide a web based data processing and visualization solution by integrating data mining API and visualization libraries”.

3.2.4 Prepare Research Design:

Once we identified our research problem the next thing was to solve our problem using best tools and techniques. So, we identified the most suitable technologies to create our solution. As we have Software as a Service solution and privacy and data security are most important aspects of a web based Software. So, we needed to secure each user’s sessions. Load balancing is another important aspect so we used JavaScript as client Side script to balance load b/w server and client. All the computation is performed on server side then the result is handed over to JavaScript for display purpose.

WEKA has a JAVA based API which can be used in any java based application. So we choose the server side technology to be JAVA server pages (JSP) and for

visualization purpose we explored all the available visualization libraries with best visualization features and can be combined with JAVA based application. We choose highcharts and D3 visualization libraries. Both are JavaScript based.

3.2.5 Prototype Implementation:

As we discussed the methodology and all the tools and technologies to implement our prototype in the last section, in this section we will talk about the steps involved in Prototype Implementation.

Use Case Diagram:

Below is a Use Case Diagram to show what tasks a user can perform in Gamified Online WEKA.

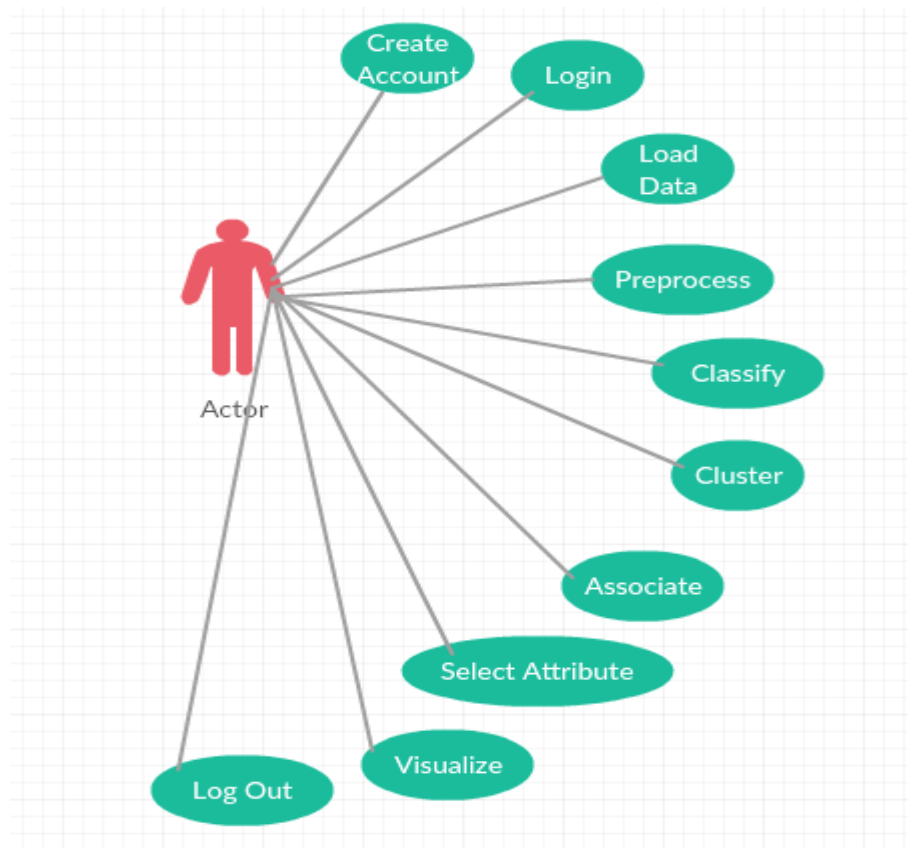


Figure 3.3 Use Case Diagram

User can perform all the data processing, mining and visualization after creating an account and logging in. After performing the desired tasks he/she can logout to end up session.

Factors Considered For Designing Gamified Online WEKA:

i. Highly Available

A web-based Implementation of data mining tool is done keeping in mind the availability factor. i.e. Tomcat application server is used in order to provide high availability. Users will be able to access this application from anywhere anytime.

ii. Complete

This application provides complete functionality of all the data mining tasks supported by the WEKA Explorer including preprocessing, classification, clustering, association, attribute selection and visualization.

iii. Interactive

An interactive user interface has been designed so that the users using the application get engaged to it and perform mining tasks efficiently. Visualizations are highly interactive; users can filter out the class values to make the visualization more understandable.

iv. Intuitive

The interface is similar to WEKA desktop application interface with just a few changes. People who already work with WEKA will face no problem in switching to web-based implementation rather they will find it to more simple.

v. Secure

The security of this web application has been ensured using authentication i.e. encryption is done for both email id and password so nobody can access any other user account even if he gets access to the database.

vi. Memory Issues Resolved

As we discussed that WEKA cannot load a file up to 12 MB without increasing java heap size. It gives memory leakage issue on loading a file of size 14.4 MB and having 501200 instances. We loaded same file in “Gamified Online WEKA”. As a result it uploaded the file and took 5 seconds to completely process it.

Instances were classified and result was displayed in 2 seconds and it took 2:30 seconds to generate clusters and display results.

Design and Architecture:

We have used restful service architecture to implement Software as service architecture. WEKA request controller interacts with session manager, database and dataset manager. Session manager keeps track of the session, expires it after thirty minutes of interactive session and redirects user to login page. Database holds user credentials while dataset manager loads and maintains data in the application. WEKA API is used in the base layer to process data.

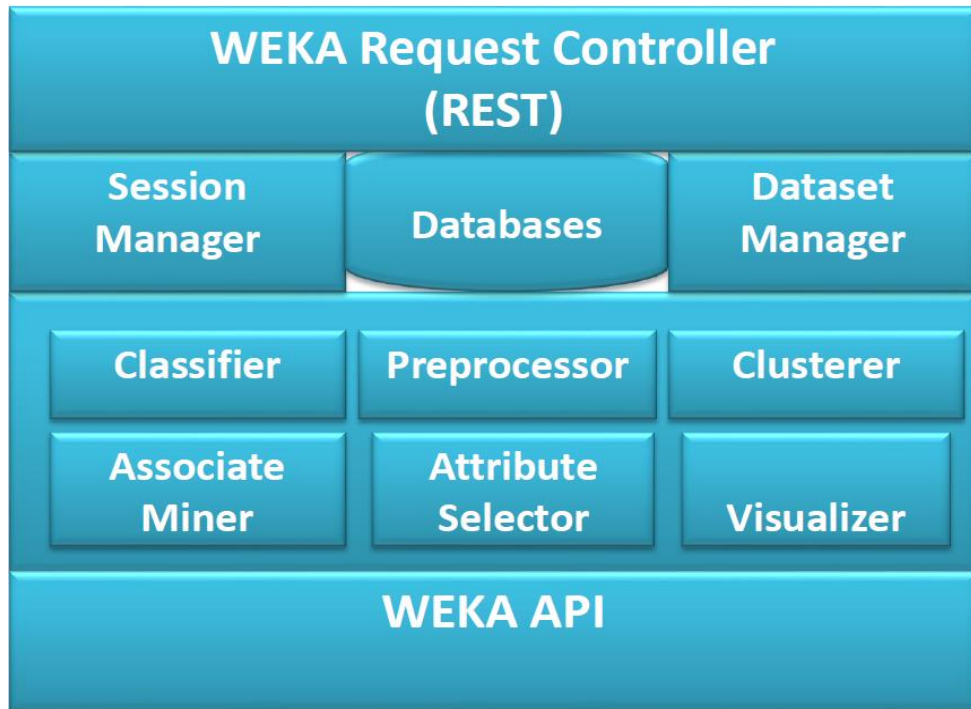


Figure 3.4 Prototype Design and Architecture

Tools and Technologies Used:

Java Server Pages (JSP) is a Java technology that helps to serve dynamically generated web pages. JSP is used as server side technology in our application.

MySQL is used as backend database to store and maintain users information.

jQuery is a JavaScript library which has been created with the motive of “Write Less, Do More”. It is a client side technology. We have used this on user interface to make it understandable and user friendly.

Highcharts is a charting library written in JavaScript offering interactive and intuitive charts. We have used this to generate stacked bar charts and scatter plot.

D3 is another JavaScript library for visualization. We are using it to generate classification trees.

Tomcat has been used as Web Application server to deploy “Gamified Online WEKA”.

IDE used to create this application is NetBeans and the application is deployed on Tomcat Server.

OPENSIFT is Red Hat's application hosting platform we have currently deployed our application on it for testing purpose.

3.2.6 Testing:

Last step of our research work was testing our newly created prototype. We tested Gamified Online WEKA for its security, data storage and data integrity. Three types of comparisons were done on WEKA AND Gamified Online WEKA.

- i. Comparison of the results. (Results of all the processing algorithms were same).
- ii. Comparison of the visualizations (Gamified Online WEKA has better visualizations).
- iii. Comparison of processing power (Gamified Online WEKA can process large data sets and processing is more efficient).

Chapter 4 : RESULTS

After developing the prototype of Gamified Online WEKA, we have deployed it on <http://gamified-onlineweka.rhcloud.com>. Anyone can create a free account there and can perform all the available tasks. For comparison purposes it is important to know that current release of Gamified Online WEKA uses WEKA 3.6-stable API. Any comparison with any other version of desktop based WEKA can give different results.

We will go step by step exploring Gamified online WEKA. First we have a login screen. To compare Gamified Online WEKA with WEKA we used WEKA 3.6-stable desktop version.

4.1 Chosen dataset:

Before performing any sort of processing on data and comparing it with WEKA we have chosen a dataset “german_credit” for testing purpose. This dataset has been taken from UCI machine learning repository. This dataset classifies people described by a set of attributes as good or bad credit risks.

4.2 Account Creation:

User can create a free account by his unique email Id and at least 5 characters long password. After account creation user will Sign in to get access to WEKA Explorer.

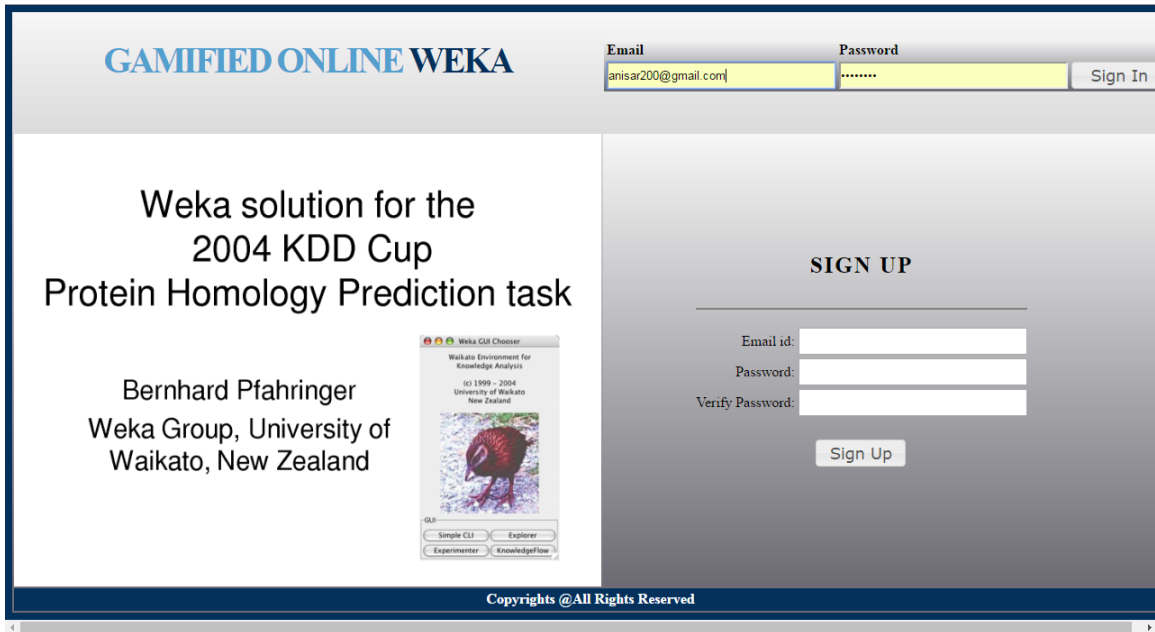


Figure 4.1 Gamified Online WEKA Login

It is evident from above figure that nobody can access this application without having a personalized account. This makes sure that all your loaded data is secure. After you sign in to your account WEKA Explorer interface is loaded and Load Data is selected as default tab.

4.3 Data Loading:

WEKA performs data loading tasks in preprocess tab. But we have separated data loading from data processing. Data loading from URL and from database is similar to WEKA. In WEKA only local database is accessible but in Gamified Online WEKA you can load any publically accessible database.

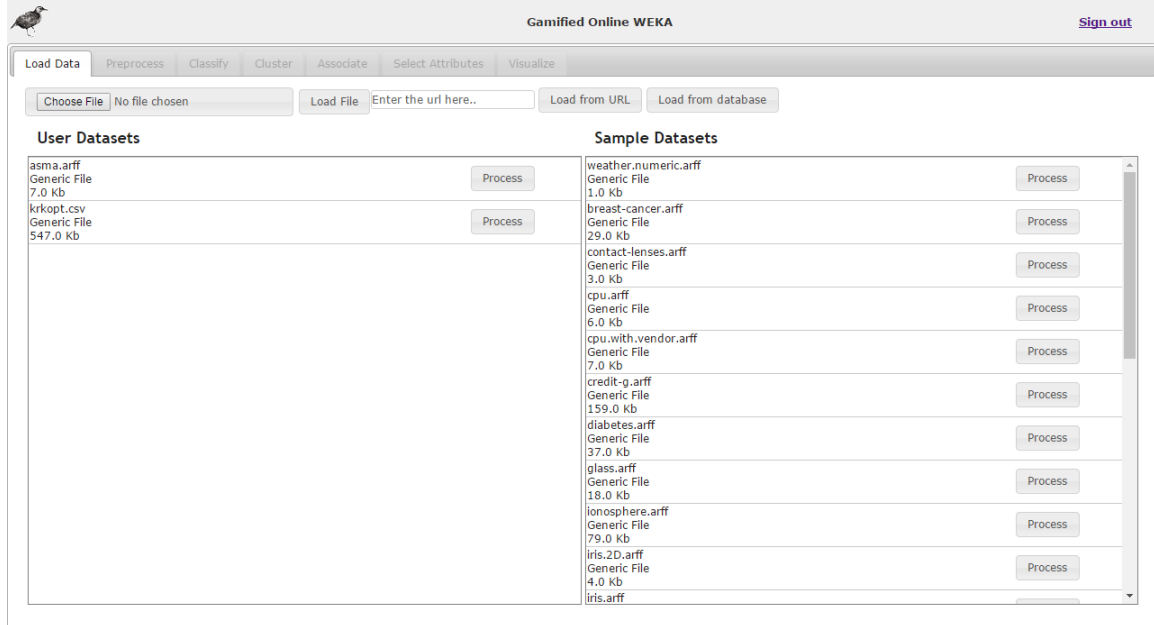


Figure 4.2 Gamified Online WEKA "Load Data" Tab

WEKA has some sample datasets which come with WEKA in the installation folder. We have displayed all those datasets in “Sample Datasets” panel. These data sets are open for everybody. On the other hand “User Datasets” are personalized datasets which a user can add by loading a dataset file. Once loaded a user can access these data sets from anywhere by logging in to his account. But no one can see datasets loaded by other users.

4.3.1 Data Loading From File:

In desktop based WEKA when a file is uploaded it directly gets loaded in WEKA explorer but in case of Gamified Online WEKA the data file is first uploaded to server. This file is displayed in User Datasets portion of Load Data tab. User can process it by clicking the process button right next to it. By doing so Preprocessor tab will be loaded and all the other tabs will be enabled.

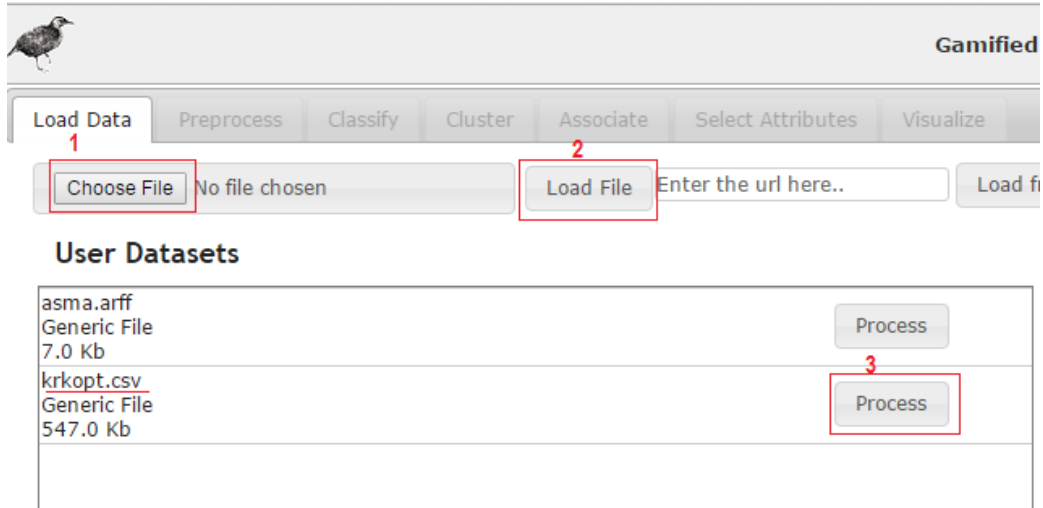


Figure 4.3 Gamified Online WEKA Data Loading from File

4.3.2 Data Loading From Database:

In case of file loading file is first uploaded to server then you need to process it but data from database is loaded directly. WEKA can only load data from local database but Gamified Online WEKA can load data from an online database if credentials are valid. Here I have connected it to a database placed on phpmyadmin. The URL value is “jdbc:mysql://SERVER_NAME/SERVER_IP:SERVER_PORT/DATABASE_NAME”. User credentials can be specified by clicking on “User...” button. Click the connect button to establish connection. Once the connection is established you can write and execute the query.

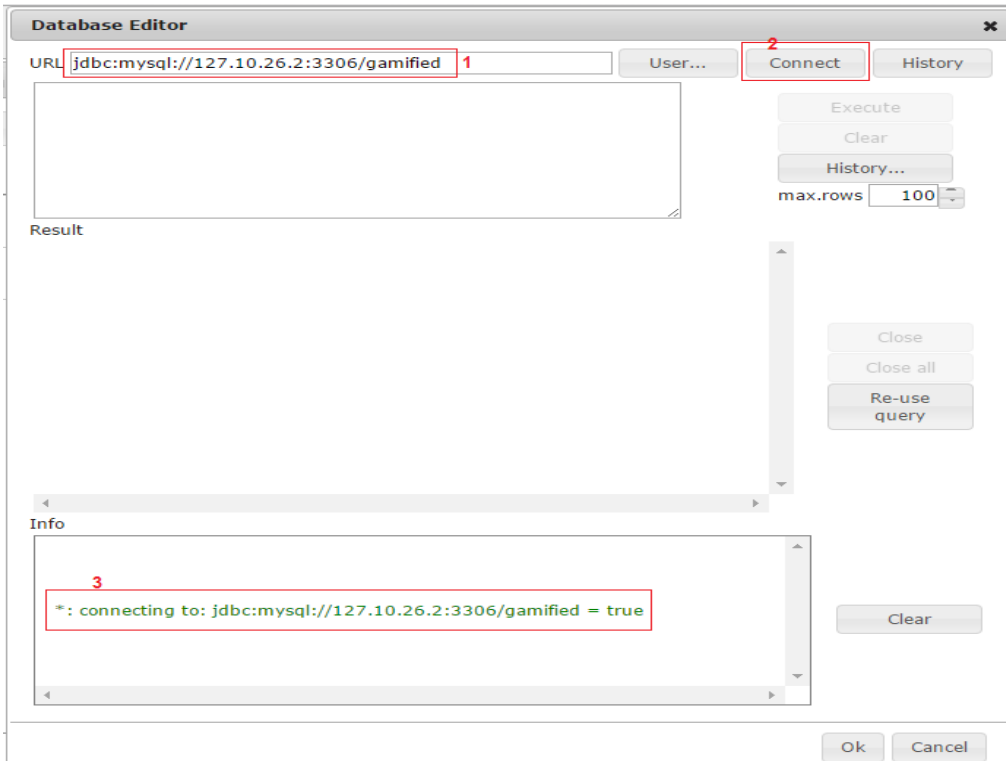


Figure 4.4 Gamified Online WEKA Data Loading from Database

4.3 Preprocessing:

Preprocessing tab has almost same interface as in WEKA except that data loading options have been moved to Load Data tab.

In WEKA

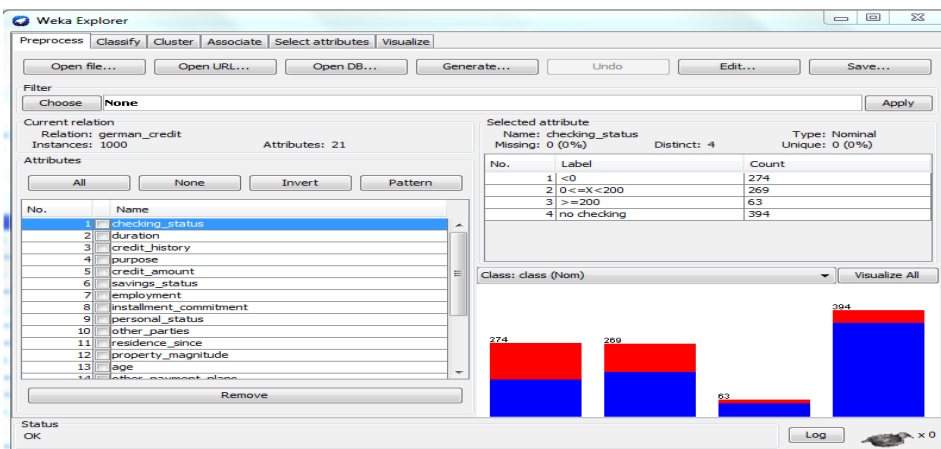


Figure 4.5 WEKA "Preprocess" Tab

In Gamified Online WEKA

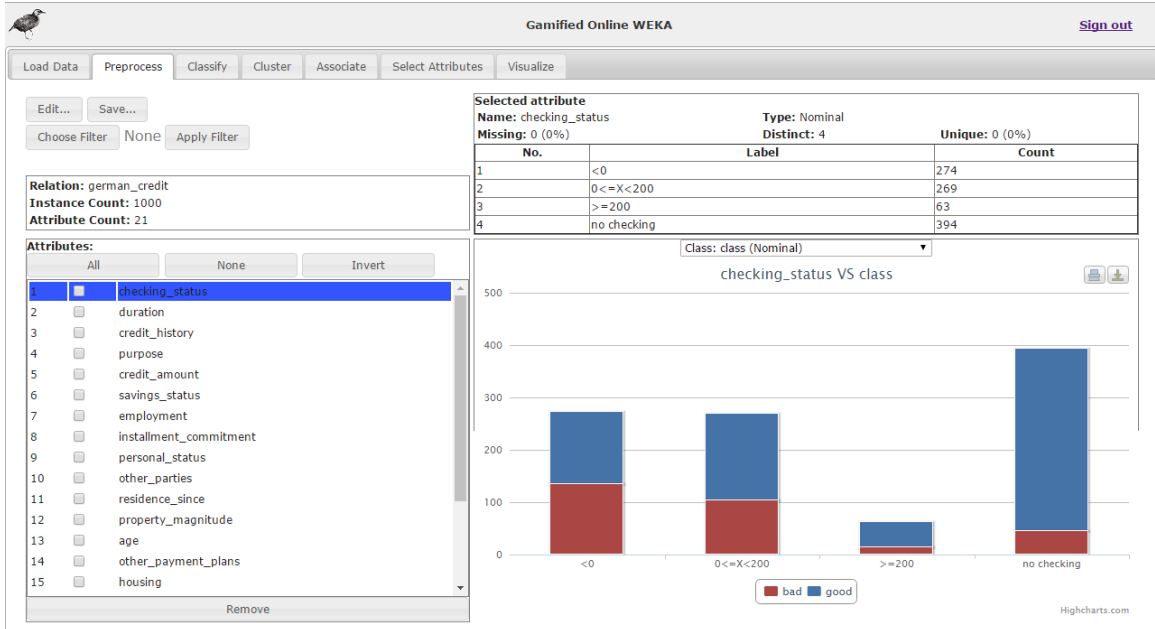


Figure 4.6 Gamified Online WEKA "Preprocess" Tab

Filter choosing and applying mechanism is same, Attribute selection is same but bar charts are different.

4.3.1 Bar Charts:

In WEKA we have got static bar charts

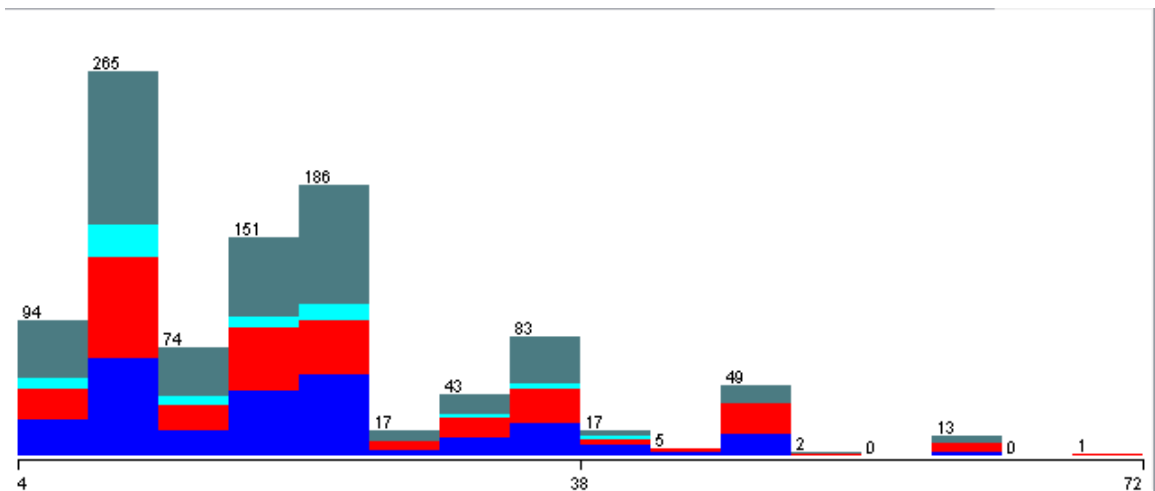


Figure 4.7 WEKA Bar Charts

In Gamified Online WEKA bar charts are dynamic

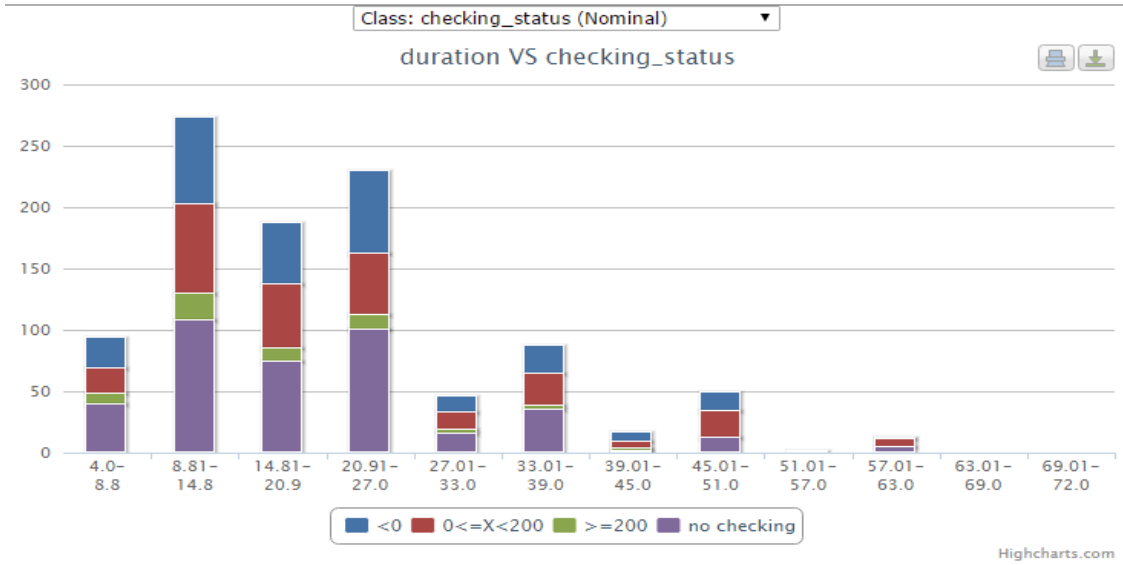


Figure 4.8 Gamified Online WEKA Bar Charts

Some of the properties of Gamified Online WEKA bar charts are listed below.

1. They can be filtered out on the basis of class attribute value.

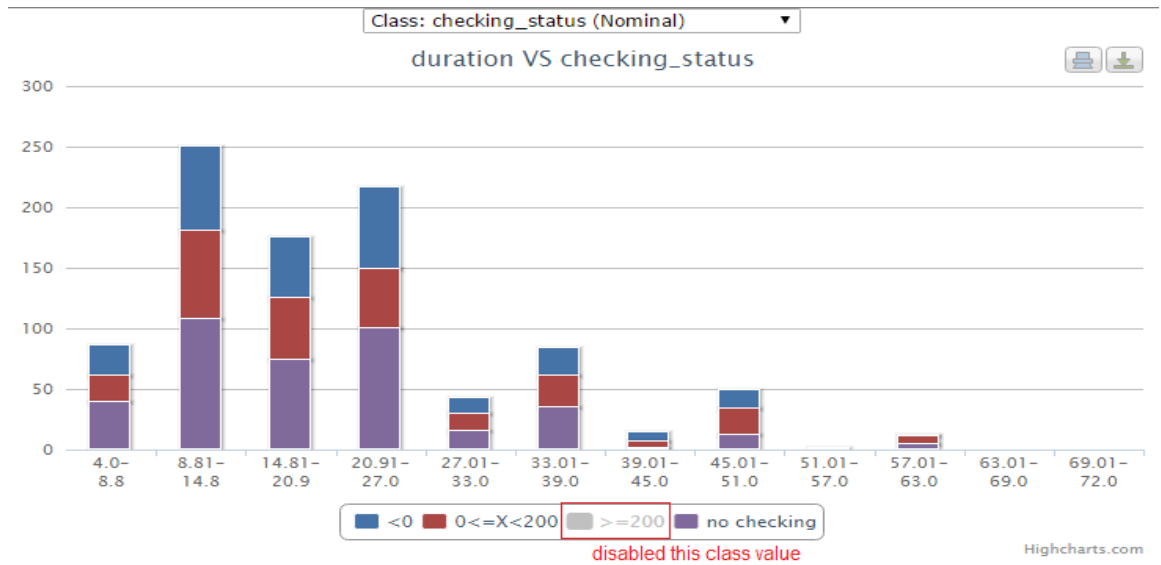


Figure 4.9 Gamified Online WEKA Bar Chart Disabled class value

- They have specified intervals even for continuous numeric attribute values.

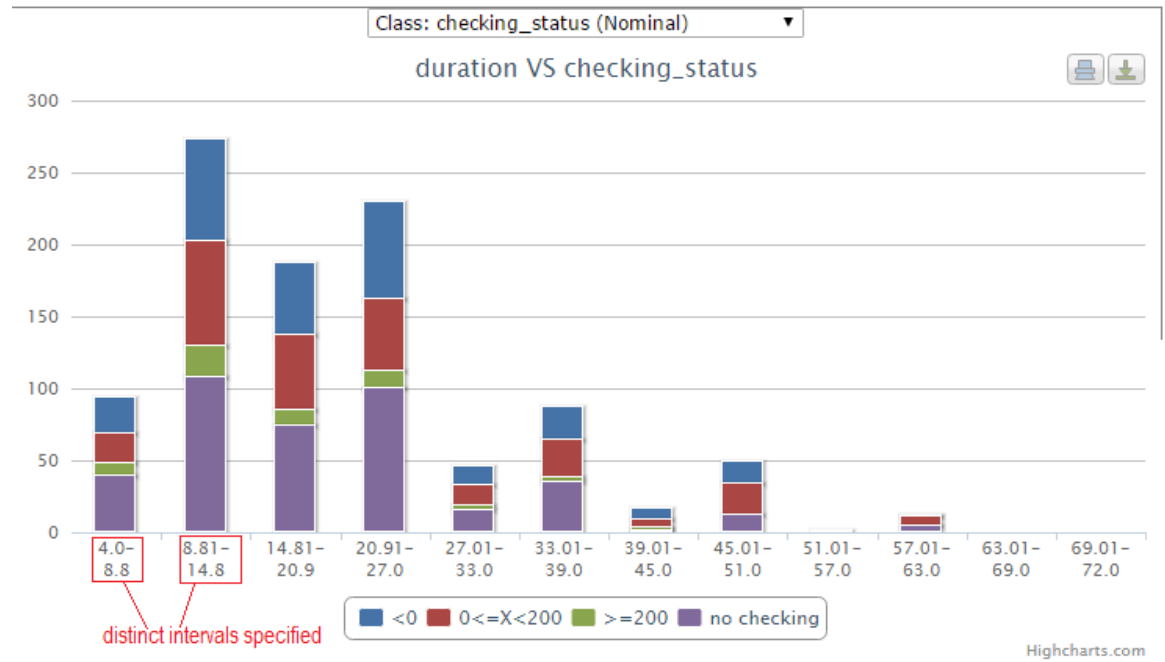


Figure 4.10 Gamified Online WEKA Bar Chart Distinct Intervals

- The image can be printed and downloaded in four different formats.

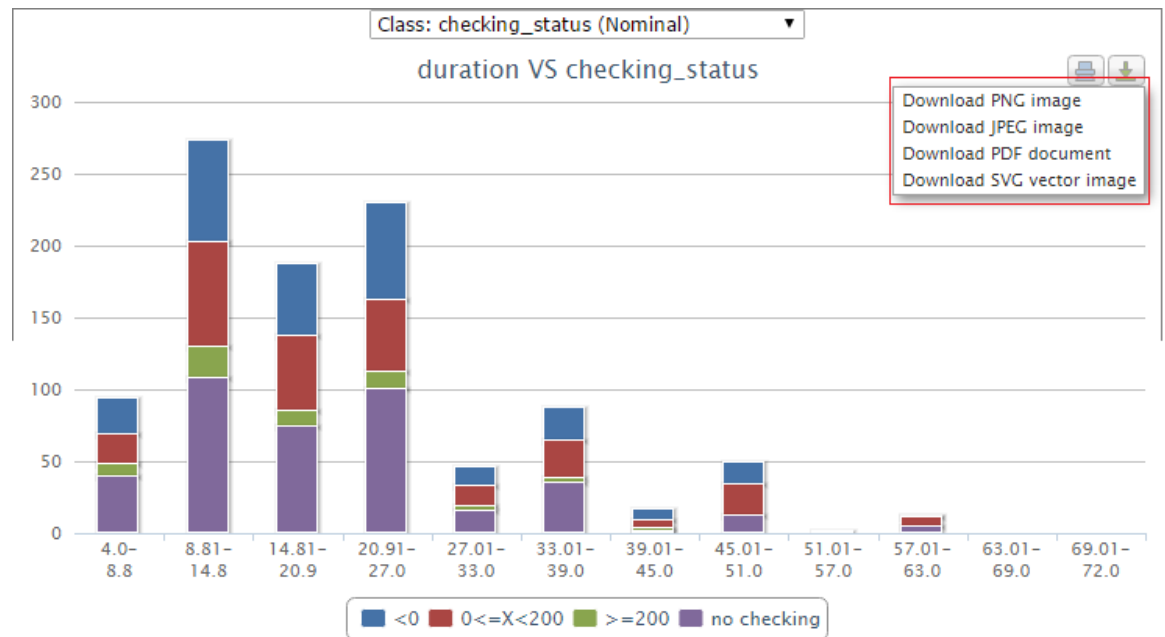


Figure 4.11 Gamified Online WEKA Bar Chart Image Download Options

- They display class name and count in a tooltip on hover.

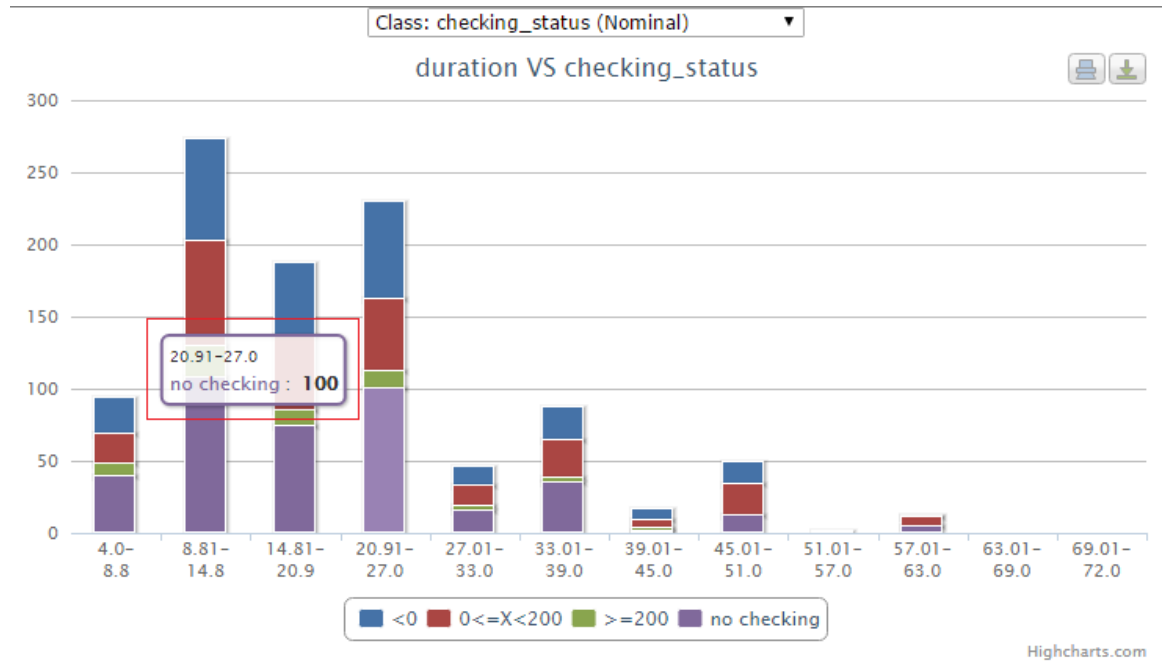


Figure 4.12 Gamified Online WEKA Bar Chart Tooltip

4.3.2 Data Editor:

The editor in WEKA displays all the data at once in a small popup. It gets hard for user to find the desired instance. Numeric value can be changed through in place editing while to change nominal value a dropdown is provided in which all the possible class values are listed. This mechanism is same for both WEKA and Gamified Online WEKA.

Viewer

Relation: german_credit

No.	checking_status Nominal	duration Numeric	credit_history Nominal	purpose Nominal	credit_amount Numeric	savings_ Nom
1	{0	6.0	critical/other...	radio/tv	1169.0	no knowr
2	0(=X(200	48.0	existing paid	radio/tv	5951.0	{100
3	no checking	12.0	critical/other...	education	2096.0	{100
4	{0	42.0	existing paid	furnitu...	7882.0	{100
5	{0	24.0	delayed pre...	new car	4870.0	{100
6	no checking	36.0	existing paid	education	9055.0	no knowr
7	no checking	24.0	existing paid	furnitu...	2835.0	500(=X(
8	0(=X(200	36.0	existing paid	used car	6948.0	{100
9	no checking	12.0	existing paid	radio/tv	3059.0)=1000
10	0(=X(200	30.0	critical/other...	new car	5234.0	{100
11	0(=X(200	12.0	existing paid	new car	1295.0	{100
12	{0	48.0	existing paid	business	4308.0	{100
13	0(=X(200	12.0	existing paid	radio/tv	1567.0	{100
14	{0	24.0	critical/other...	new car	1199.0	{100
15	{0	15.0	existing paid	new car	1403.0	{100
16	{0	24.0	existing paid	radio/tv	1282.0	100(=X(
17	no checking	24.0	critical/other...	radio/tv	2424.0	no knowr
18	{0	30.0	no credits/all...	business	8072.0	no knowr
19	0(=X(200	24.0	existing paid	used car	12579.0	{100
20	no checking	24.0	existing paid	radio/tv	3430.0	500(=X(
21	no checking	9.0	critical/other...	new car	2134.0	{100
22	{0	6.0	existing paid	radio/tv	2647.0	500(=X(
23	{0	10.0	critical/other...	new car	2241.0	{100

Undo OK Cancel

Figure 4.13 WEKA Data Editor

While in Gamified online WEKA we have implemented pagination.

Weka File Editor

Show 25 entries

checking_status (Nominal)	duration (Numeric)	credit_history (Nominal)	purpose (Nominal)	credit_amount (Numeric)	savings_status (Nominal)	employment (Nominal)	instalment_commitment (Numeric)	personal_status (Nominal)	other_parties (Nominal)	residence_since (Numeric)	property_magnitude (Nominal)	age (Numeric)	other_p (I
no checking	4.0	critical/otl	radio/	1544.0	<100	4<=X<7	2.0	male single	none	1.0	real estate	42.0	none
no checking	4.0	critical/otl	new c	3380.0	<100	4<=X<7	1.0	female div/c	none	1.0	real estate	37.0	none
no checking	4.0	existing p	furnit	601.0	<100	<1	1.0	female div/c	none	3.0	real estate	23.0	none
>=200	4.0	existing p	new c	1494.0	no known :	<1	1.0	male single	none	2.0	real estate	29.0	none
no checking	4.0	critical/otl	radio/	1503.0	<100	4<=X<7	2.0	male single	none	1.0	real estate	42.0	none
no checking	4.0	critical/otl	new c	1455.0	<100	4<=X<7	2.0	male single	none	1.0	real estate	42.0	none
no checking	5.0	existing p	busin	3448.0	<100	4<=X<7	1.0	male single	none	4.0	real estate	74.0	none
0<=X<200	6.0	existing p	radio/	590.0	<100	<1	3.0	male mar/w	none	3.0	real estate	26.0	none
0<=X<200	6.0	delayed p	new c	1209.0	<100	unemplo	4.0	male single	none	4.0	life insurance	47.0	none
0<=X<200	6.0	existing p	radio/	368.0	no known :	>=7	4.0	male single	none	4.0	life insurance	38.0	none
0<=X<200	6.0	all paid	new c	931.0	100<=X<!	<1	1.0	female div/c	none	1.0	life insurance	32.0	stores
<0	6.0	existing p	used	1352.0	500<=X<!	unemplo	1.0	female div/c	none	2.0	life insurance	23.0	none
<0	6.0	critical/otl	radio/	1169.0	no known :	>=7	4.0	male single	none	4.0	real estate	67.0	none
0<=X<200	6.0	all paid	educa	433.0	>=1000	<1	4.0	female div/c	none	2.0	life insurance	24.0	bank
0<=X<200	6.0	existing p	radio/	484.0	<100	4<=X<7	3.0	male mar/w	guarantol	3.0	real estate	28.0	bank
>=200	6.0	delayed p	radio/	683.0	<100	<1	2.0	female div/c	none	1.0	life insurance	29.0	bank
>=200	6.0	critical/otl	educa	1047.0	<100	1<=X<4	2.0	female div/c	none	4.0	life insurance	50.0	none
0<=X<200	6.0	delayed p	furnit	1050.0	<100	unemplo	4.0	male single	none	1.0	life insurance	35.0	stores
>=200	6.0	critical/otl	new c	1323.0	100<=X<!	>=7	2.0	male div/se	none	4.0	car	28.0	stores
0<=X<200	6.0	existing p	repair	454.0	<100	<1	3.0	male mar/w	none	1.0	life insurance	22.0	none
0<=X<200	6.0	existing p	radio/	753.0	<100	1<=X<4	3.0	female div/c	guarantol	3.0	real estate	64.0	none

OK Cancel

Figure 4.14 Gamified Online WEKA Data Editor

4.4 Classify:

Both the user interface and results are same for Classify tab in both WEKA and Gamified Online WEKA.

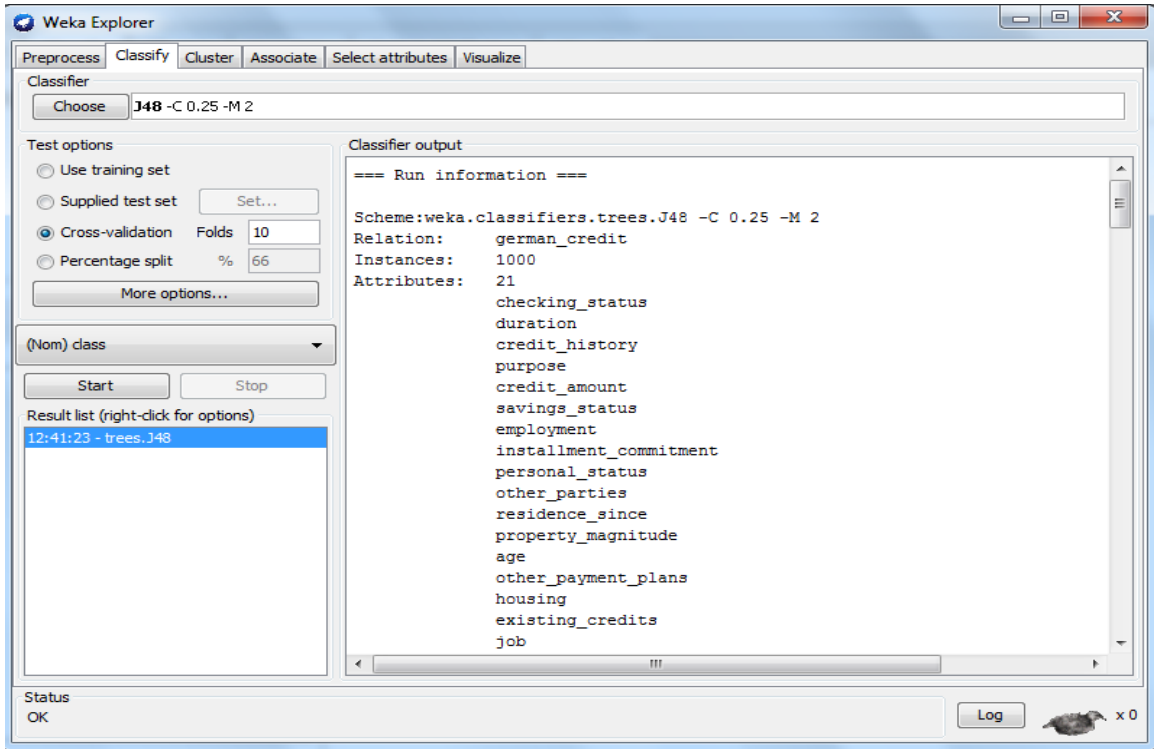


Figure 4.15 WEKA Classifier

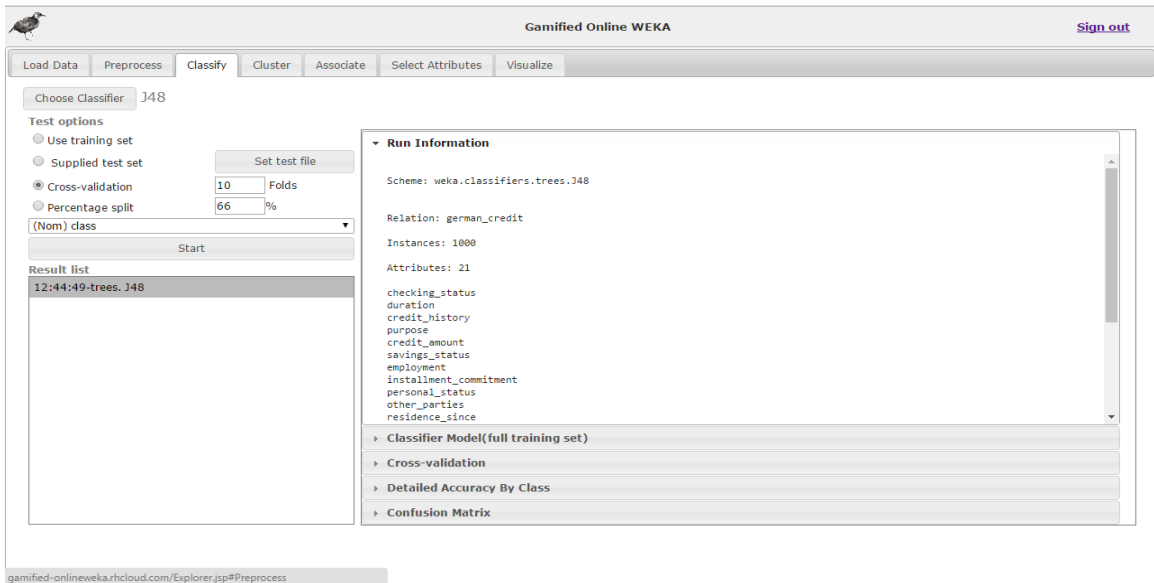
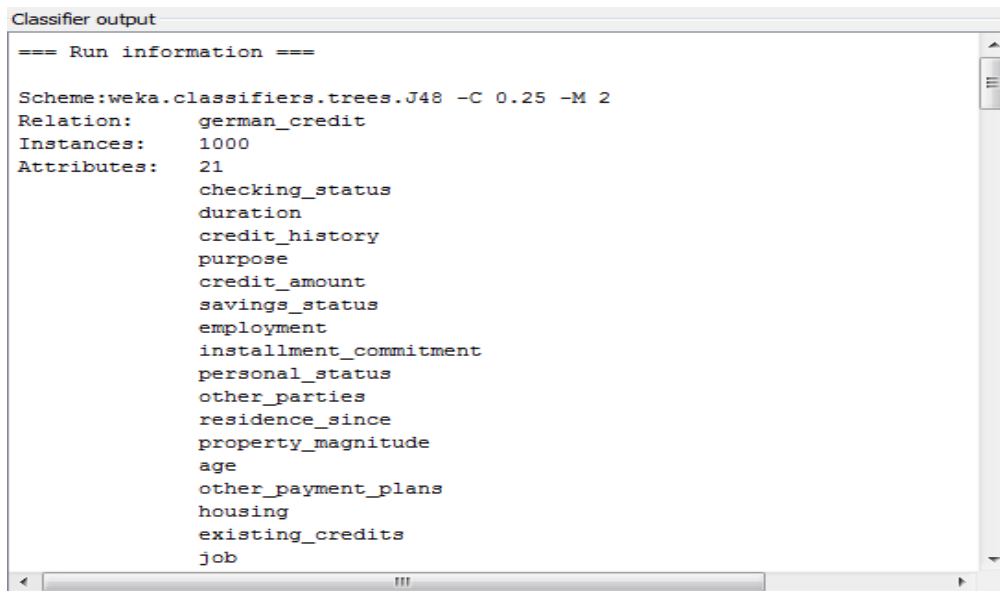


Figure 4.16 Gamified Online WEKA Classifier

It is clear from the comparisons of these two interfaces that everything is same except the result panel.

4.4.1 Classify Result Panel:

WEKA displays the result set in a panel. All sections are included in it. User has to search for required information by scrolling the result panel.



```
Classifier output
=== Run information ===
Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:  german_credit
Instances:  1000
Attributes:  21
            checking_status
            duration
            credit_history
            purpose
            credit_amount
            savings_status
            employment
            installment_commitment
            personal_status
            other_parties
            residence_since
            property_magnitude
            age
            other_payment_plans
            housing
            existing_credits
            job
```

Figure 4.17 WEKA Classify Result Panel

In case of Gamified Online WEKA results are distributed in proper sections. User can click a section to expand it.

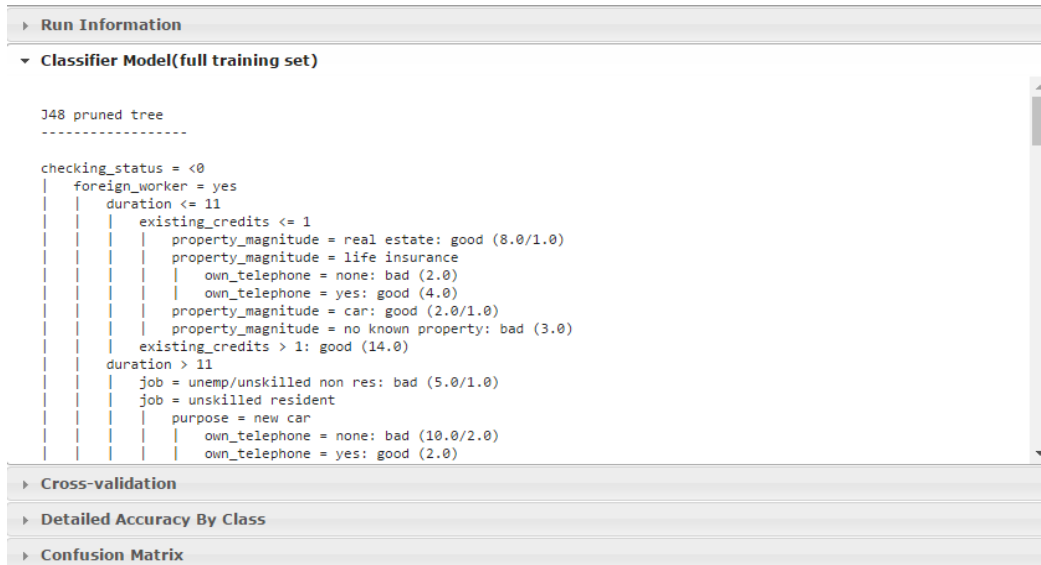


Figure 4.18 Gamified Online WEKA Classify Result Panel

4.4.2 Time Taken to build Classifier Model:

We applied J48 classification tree on both WEKA and Gamified Online WEKA. WEKA apparently took less time but there are two factors. First WEKA is running on local machine internet speed does not effect it. Secondly WEKA counts the time just to calculate model, other result processing time is not included. But Gamified Online WEKA counts the time from start of building model to display of result.

```
Number of Leaves : 103
```

```
Size of the tree : 140
```

```
Time taken to build model: 0.05 seconds
```

Figure 4.19 WEKA Classifier Model

```
Number of Leaves : 103
```

```
Size of the tree : 140
```

```
Time taken to build model: 0.17 seconds
```

Figure 4.20 Gamified Online WEKA Classifier Model

4.4.3 Classification Result Comparison:

Classification results are same for both as we have just used the WEKA API. We did not change any algorithm in it. Below is the cross validation section comparison.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      705           70.5    %
Incorrectly Classified Instances    295           29.5    %
Kappa statistic                     0.2467
Mean absolute error                 0.3467
Root mean squared error            0.4796
Relative absolute error             82.5233 %
Root relative squared error        104.6565 %
Total Number of Instances          1000
```

Figure 4.21 WEKA Classification Result

```
▼ Cross-validation

Correctly Classified Instances      705           70.5    %
Incorrectly Classified Instances    295           29.5    %
Kappa statistic                     0.2467
Mean absolute error                 0.3467
Root mean squared error            0.4796
Relative absolute error             82.5233 %
Root relative squared error        104.6565 %
Total Number of Instances          1000
```

Figure 4.22 Gamified Online WEKA Classification Result

4.4.4 Classification Tree Comparison:

It is one of the most important visualization. In WEKA the large trees get messed up. We generated the classification tree for J48 algorithm on the selected dataset.

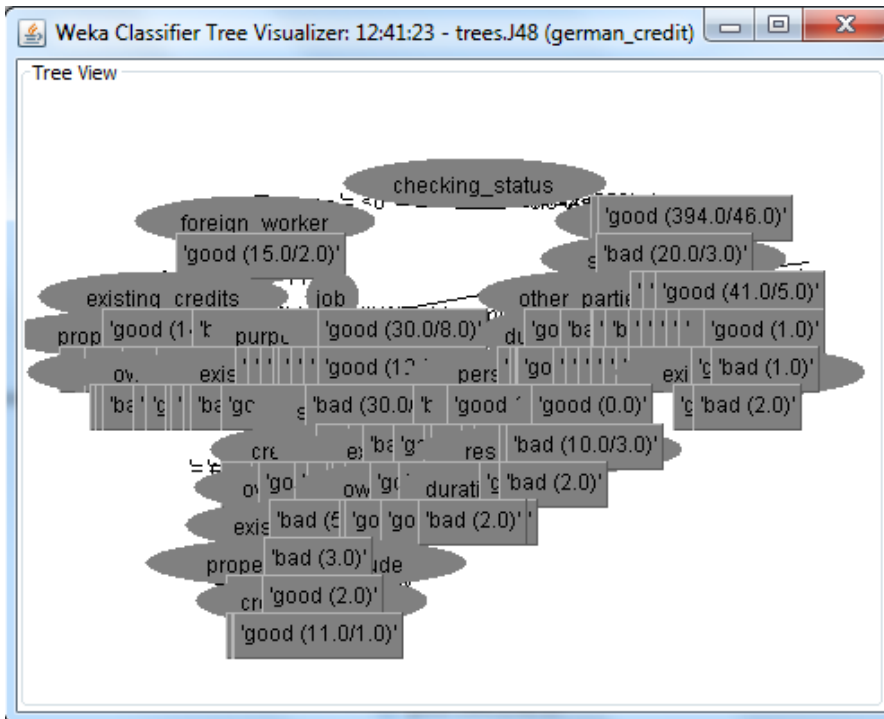


Figure 4.23 WEKA Classification Tree

We generated the same tree for Gamified Online WEKA which is shown below.

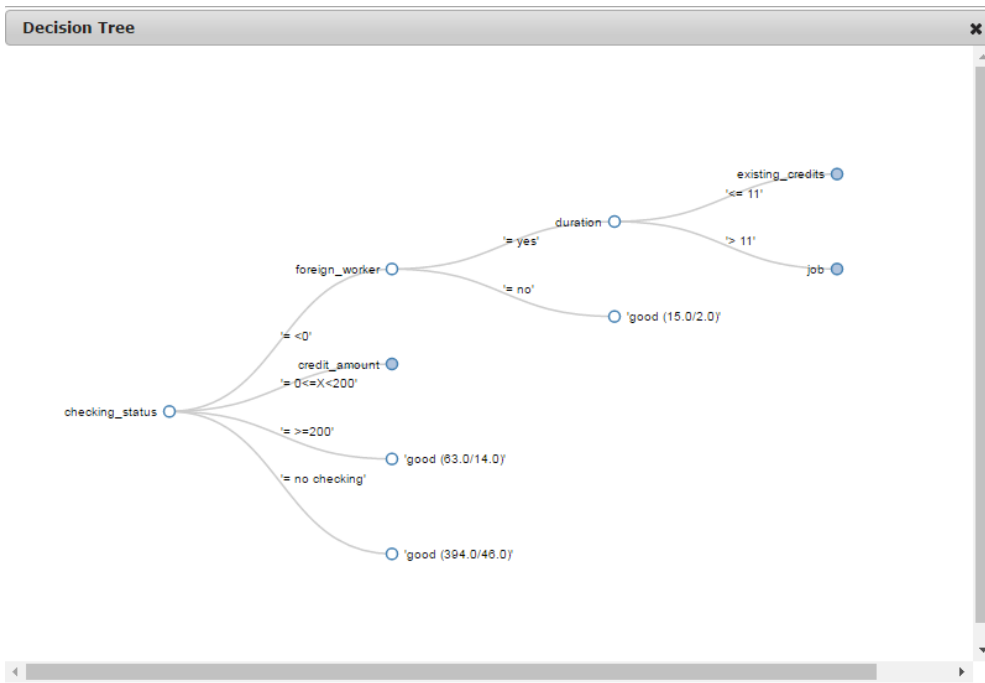


Figure 4.24 Gamified Online WEKA Classification Tree

In this tree hollow nodes are expanded nodes while nodes filled with blue color are collapsed nodes. We can expand or collapse any node by clicking on it.

4.5 Cluster:

Just like Classify tab interface is same for Cluster except the result panel with distinct sections.

4.5.1 Cluster Result Comparison:

Instead of comparing all the sections we will just compare the resulting clusters and their log likelihood. For this purpose we applied the EM Clusterer on both with percentage split as Cluster mode. Results were same for both.

The screenshot shows the WEKA Clusterer interface. The 'Clusterer' dropdown is set to 'EM -I 100 -N -1 -M 1.0E-6 -S 100'. Under 'Cluster mode', 'Percentage split' is selected with a value of 66%. The 'Store clusters for visualization' checkbox is checked. The 'Clusterer output' panel displays the following data:

Clustered Instances		
0	160	(47%)
1	48	(14%)
2	96	(28%)
3	36	(11%)

Log likelihood: -33.69341

The 'Result list' shows two entries: '13:41:17 - EM' and '17:56:24 - EM', with the latter selected.

Figure 4.25 WEKA Cluster Result

The screenshot shows the Gamified Online WEKA Clusterer interface. The 'Choose Clusterer' dropdown is set to 'EM'. Under 'Test options', 'Percentage split' is selected with a value of 66%. The 'Start' button is visible. The 'Run Information' panel displays the following data:

Clustered Instances		
0	160	(47%)
1	48	(14%)
2	96	(28%)
3	36	(11%)

Log likelihood: -33.69341

The 'Result list' shows one entry: '17:56:41-EM'.

Figure 4.26 Gamified Online WEKA Cluster Result

It is clear that both have generated 4 clusters with same amount of instances and same log likelihood.

4.6 Associate:

All the association rules in WEKA can only be applied to nominal attributes. To test the results of this tab we applied “NumericToNominal” filter under the category of Unsupervised --> Attribute from Preprocess tab. It changed all the numeric attributes to nominal.

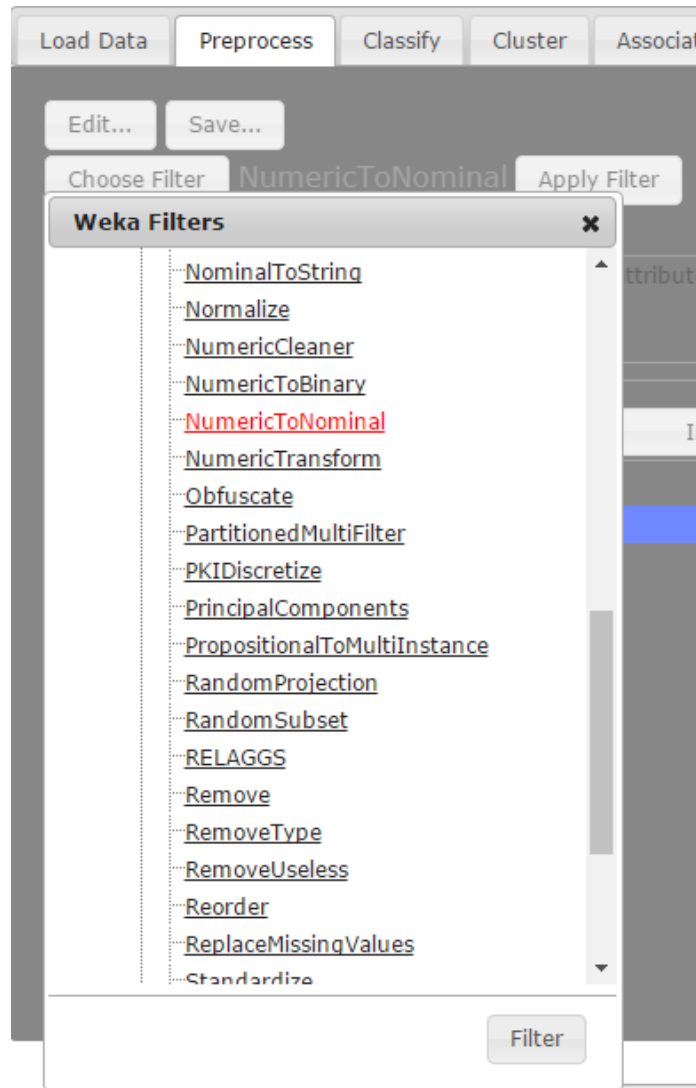


Figure 4.27 NumericToNominal filter selection

After filter application we can use any of the Association rule on our data.

By applying FilteredAssociator we got the same results.

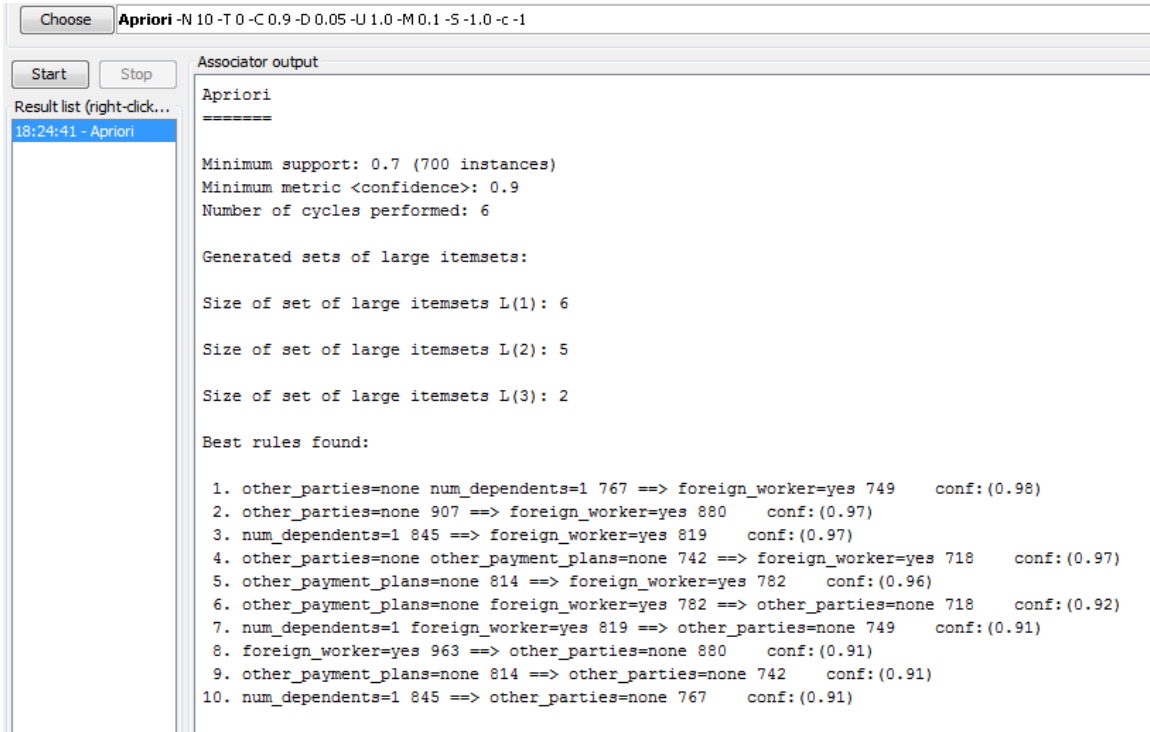


Figure 4.28 WEKA Associator

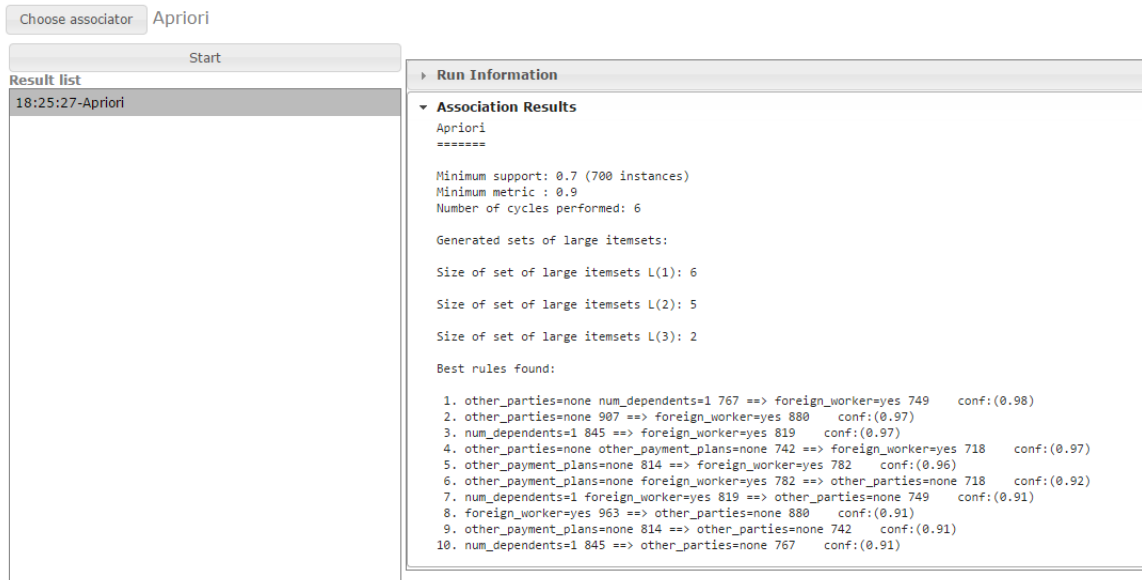


Figure 4.29 Gamified Online WEKA Associator

Above figures prove that the result set is totally same for both applications.

4.7 Select Attributes:

We evaluated our dataset using CfsSubsetEval as attribute evaluator and BestFirst as search method. The results were same for both applications.

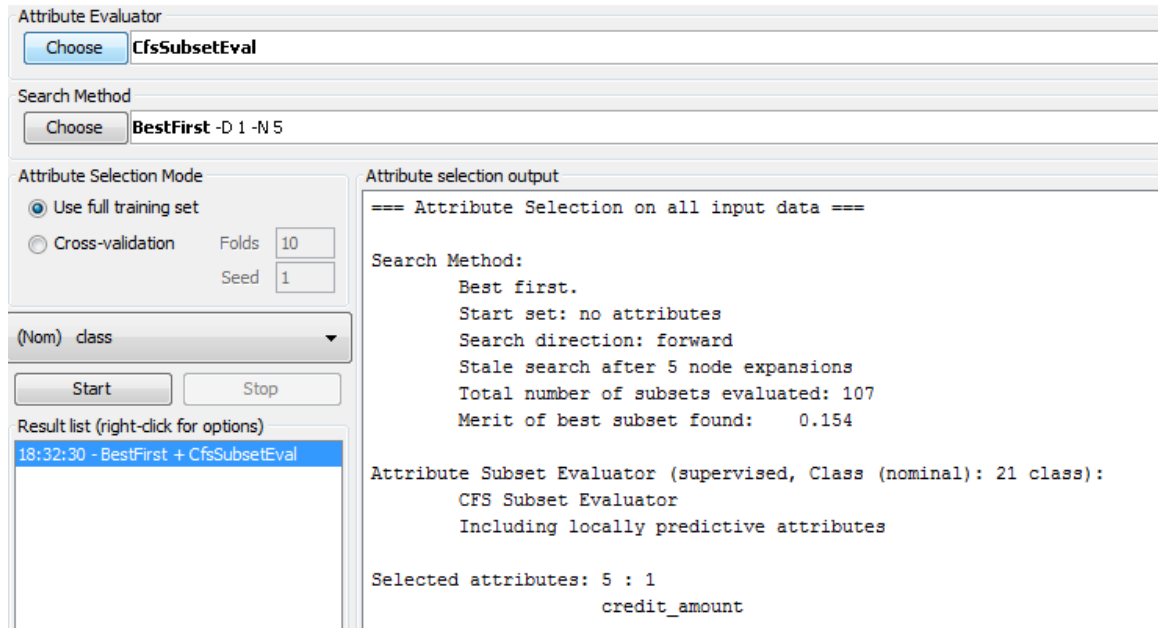


Figure 4.30 WEKA Attribute Selection

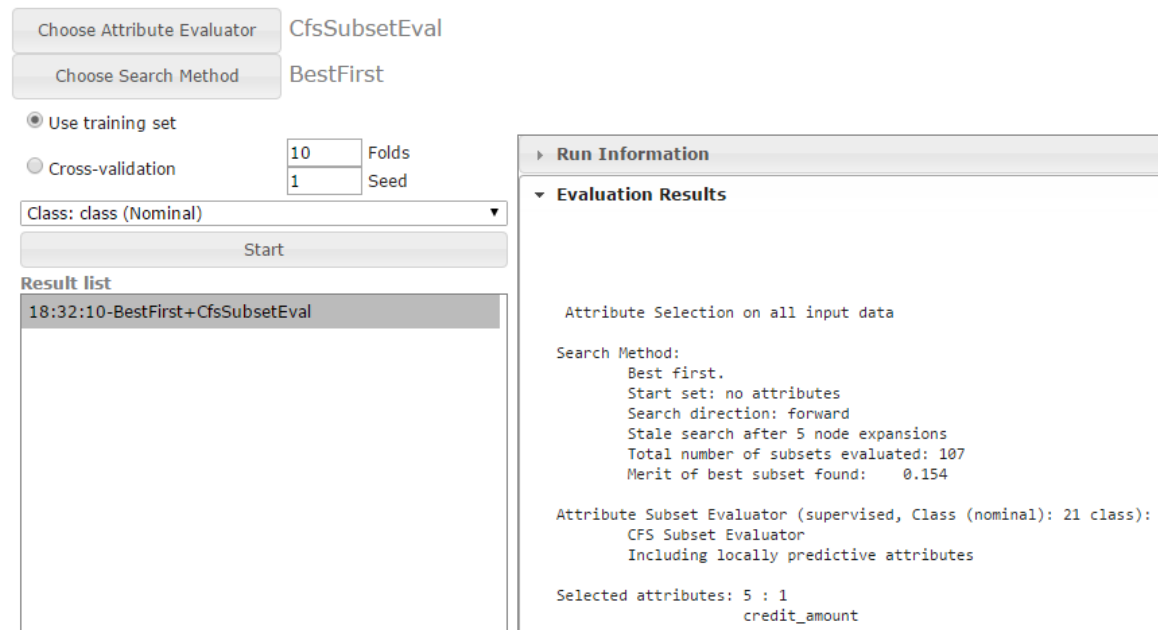


Figure 4.31 Gamified Online WEKA Attribute Selection

Selected attributes are same for both the Applications.

4.8 Visualize:

In visualize tab relationship b/w attributes is visualized using scatterplot. In WEKA distinct class intervals are not specified for numeric class value because of which it gets hard to get instances for a specific interval. Just like bar charts there is no class value filtering in WEKA.

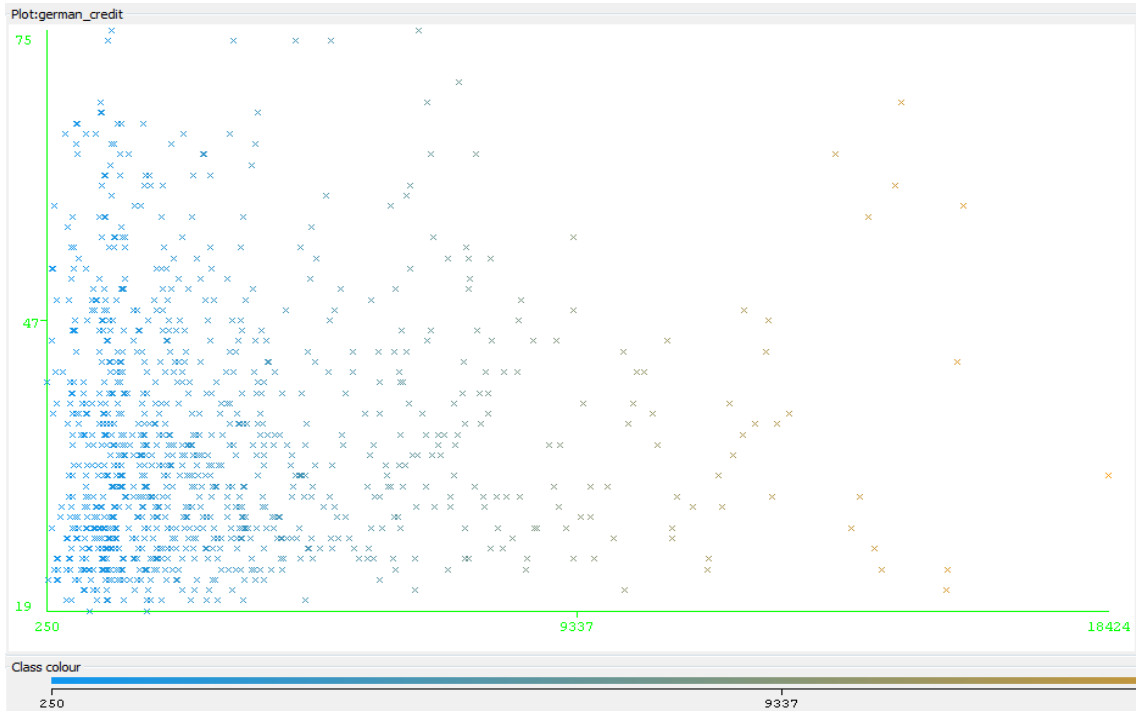


Figure 4.32 WEKA Visualize

It is clear from above figure that class intervals are not defined in WEKA and there is no way to filter out instances for a specific interval.

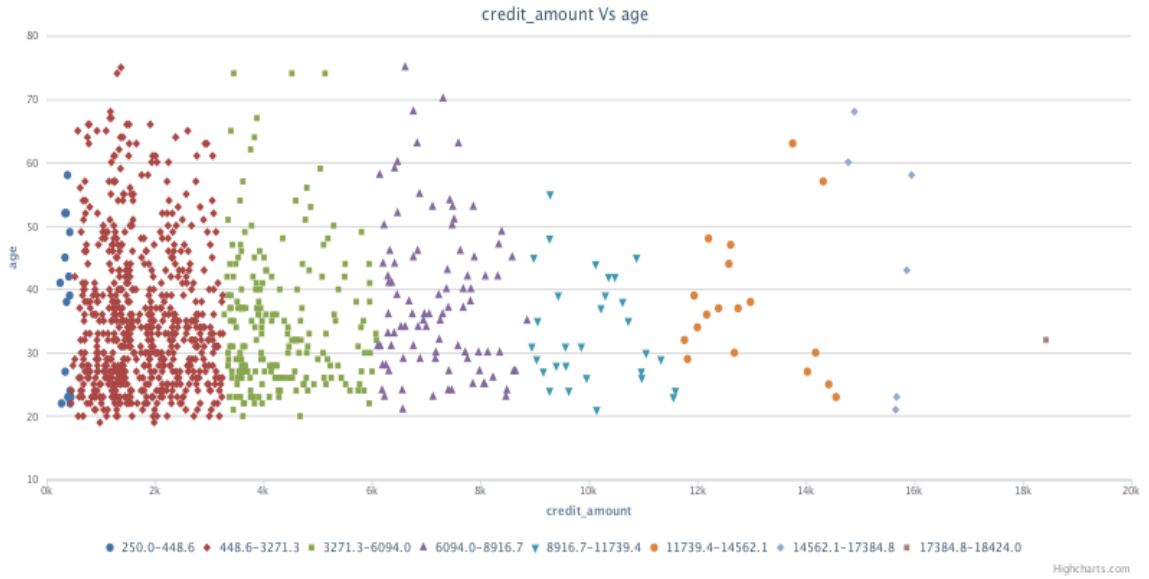


Figure 4.33 Gamified Online WEKA Visualize

In Gamified Online WEKA class intervals are not only defined but also represented by using different colors and shapes. It also provides a way to print and download charts.

Chapter 5 : CONCLUSION

This chapter concludes our thesis. We will summarize our work and contributions in Datamining domain. We are also going to discuss future research prospects of this thesis idea.

5.1 Conclusion:

In today's world, we are generating data at a large scale. Every time we swipe our credit card, make a call, send a text or pass by a security camera we generate a little bit of data. Datamining is about taking this raw data and transforming it into information which can be used in a useful way in real world. By combining datamining with Cloud computing we combine the two most powerful technologies of today's world to generate something more useful.

WEKA (Waikato Environment for Knowledge Analysis) is the second most commonly used datamining tool in the world. It is a machine learning project going on for 20 years in University of Waikato. New and improved datamining algorithms are added constantly to it. A summary of our contributions towards WEKA and datamining is as follows.

1. We identified the need of a powerful online datamining solution which most of the people from this domain were looking for. Through our research we came to know that many people are looking for an online version of WEKA. So, we provided an online solution to the datamining community.
2. We proposed, designed and implemented the solution with the following qualities.
 - It helps users to get rid of installation and configuration issues of WEKA.

- It increases processing power of WEKA and to makes it more accessible.
- It provide user with secure environment to save and process data.
- User interface has been improved to make it more user-friendly yet very close to WEKA interface which makes it easy for user to switch to Gamified online WEKA.
- There is no change in underlying datamining algorithms so the results are same for both applications. (If you use the same API).
- By gamifying our application we are able to provide better dynamic visualization instead of clumsy static visualizations.

5.2 Future directions:

The presented solution is just a prototype. We can improve and extend it in many ways. Here are some future directions to improve Gamified Online WEKA and to keep it updated.

1. WEKA is evolving software. A new snapshot is released every day and a new stable version comes in two years with new algorithms and improvements. We have used WEKA API 3.6-stable for this purpose. So we need to keep track of every new stable version and embed it in our Application.
2. We can take suggestions from users after releasing it commercially and can improve it accordingly.
3. Currently our structure is user based we can make it organization based. So that data stored and processed by one user gets accessible to all the registered members of that organization.

4. Currently we have implemented WEKA explorer as it is the main module but later on we can bring Experimenter and Knowledge Flow to our online version.
5. Efficiency and processing power can further be increased by implementing more powerful processing algorithms.

Appendix A

Installation and Configuration of WEKA:

Latest stable version of WEKA is WEKA-3.8.stable. It has been launched recently but it does not have any related documentation and manual yet. The API used for “Gamified Online WEKA” is WEKA 3.6 stable which can be downloaded from <http://www.cs.waikato.ac.nz/ml/weka/downloading.html>

A.1 Required Softwares

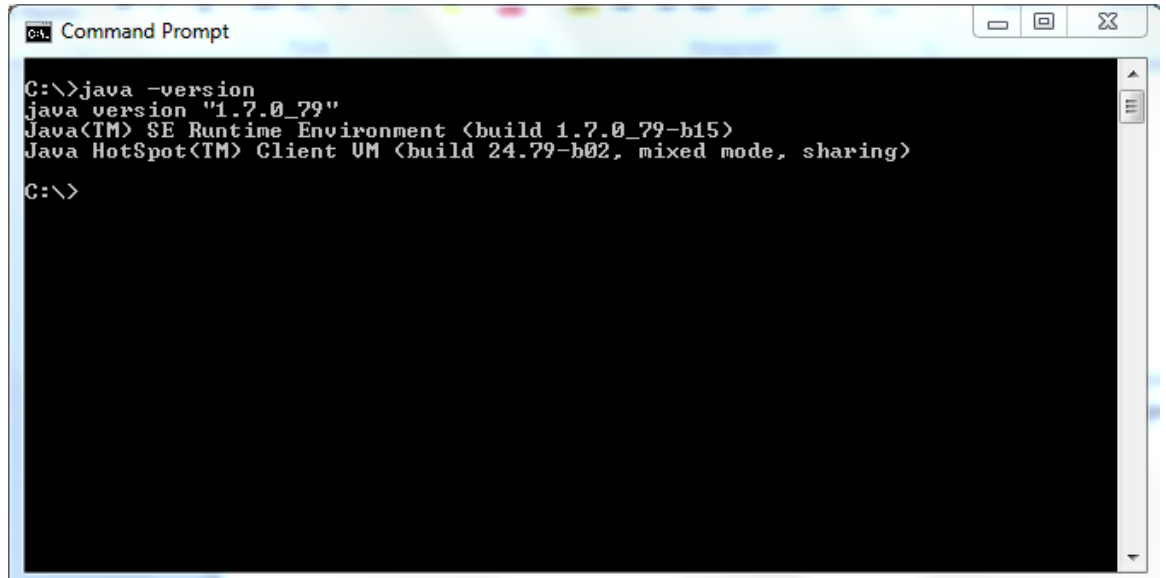
- Java 1.7
- WEKA 3.6-stable

A.2 Detailed Steps

Step 1: Java Installation:

1. First you need to verify if java is installed on system or not. If yes then is it the version we require for WEKA installation. To verify open the command prompt. Go to C (default installation directory) and type

```
java -version
```
2. If the result is “Java is not recognized as internal or external command, operable program or batch File”.
3. Otherwise it will result in the java version as shown in figure.



```
C:\>java -version
java version "1.7.0_79"
Java(TM) SE Runtime Environment (build 1.7.0_79-b15)
Java HotSpot(TM) Client VM (build 24.79-b02, mixed mode, sharing)
C:\>
```

Figure A. 1 Test Java Installation

4. If Java is not installed then download it from [here](#).
5. Install it from the downloaded setup.
6. After installation open file explorer and click on system properties.

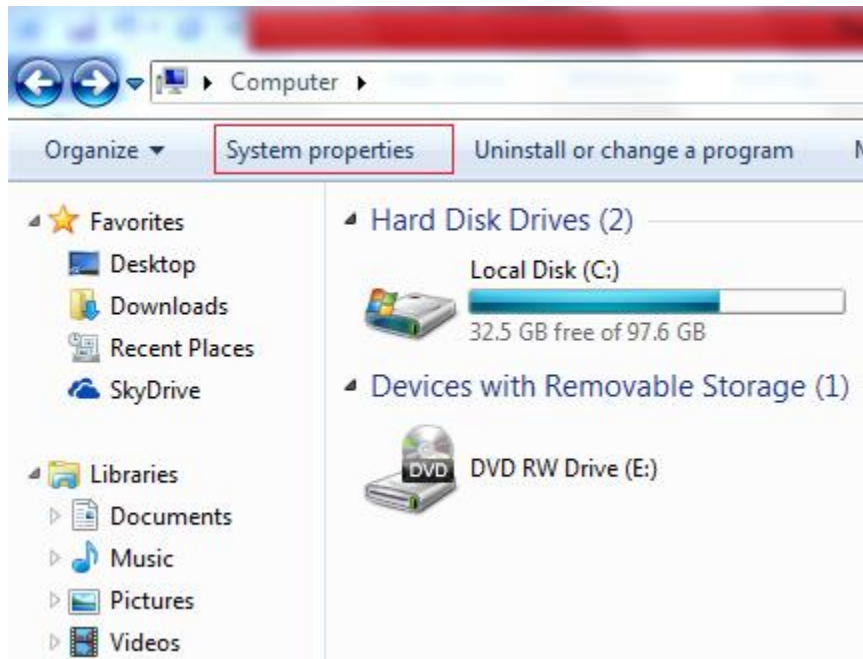


Figure A. 2 System Properties

7. Click on “Advanced System Setting” and then on “Environment variables...” in it.

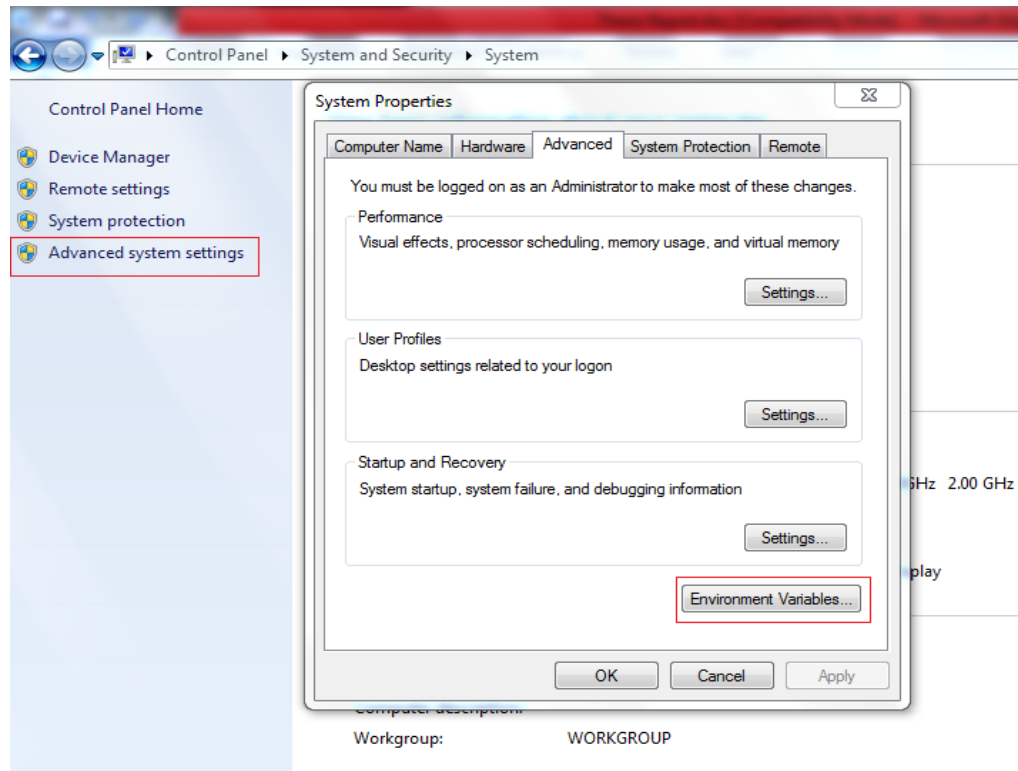


Figure A .3 Windows Environment Variables

8. Click “New...” for System Variables.

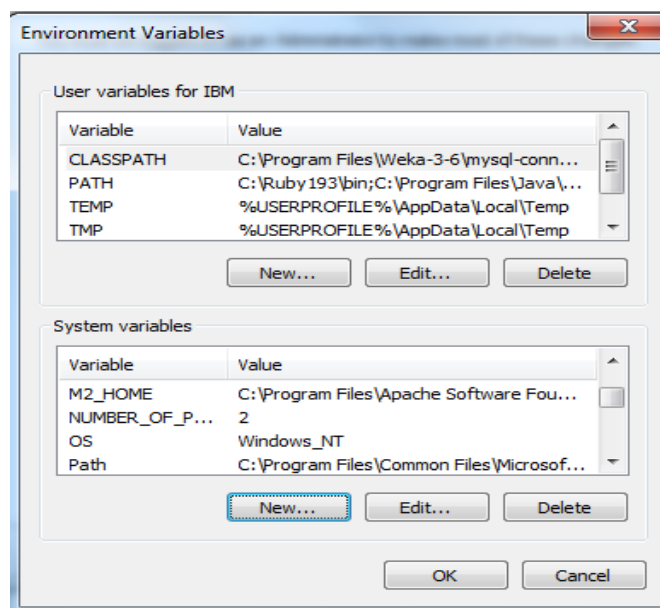


Figure A .4 New System Variable

9. Add a new variable “JAVA_HOME” and set its variable value to JDK root directory.

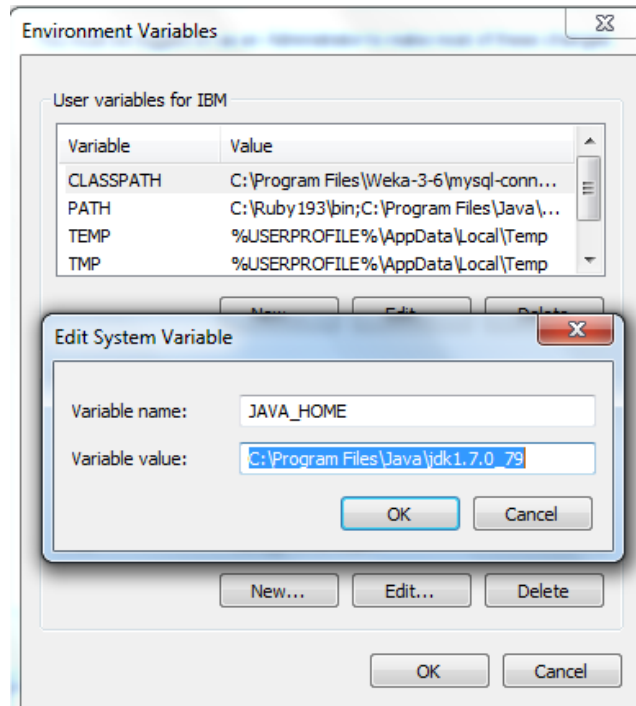


Figure A. 5 Set JAVA_HOME variable

10. Now edit Path variable from System Variables and append “JAVA_HOME%bin;” in it.
11. Now to confirm the installation open a Command Prompt and run the command “java -version” as described in step 1.

Step 2: WEKA Installation:

1. Download WEKA from [here](#).
2. Run the executable file.



Figure A. 6 WEKA Installation Setup Wizard

3. Click on next and proceed.

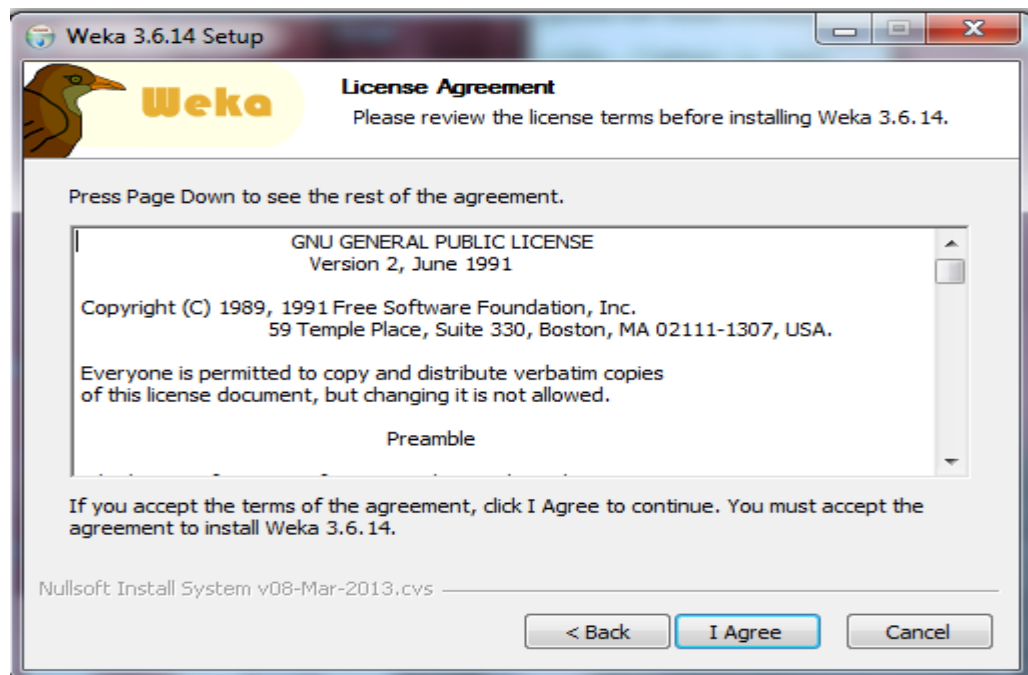


Figure A. 7 WEKA Installation

4. Open weka.jar to confirm installation.

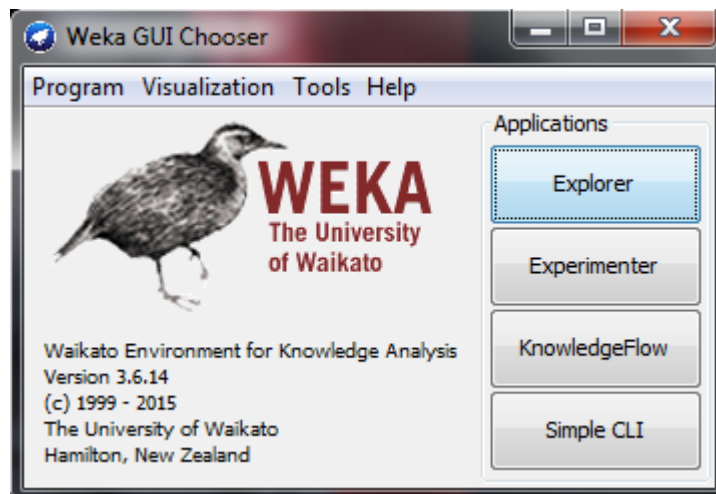


Figure A. 8 WEKA GUI Chooser

5. Weka is ready to use. Default datasets are present in “data” folder of WEKA installation directory.

NOTE: All the Installation has been done on Windows 7. Steps may vary for other operating systems.

Appendix B

Datamining with WEKA Course:

B.1 WEKA Book:

Waikato University has not only developed WEKA toolkit but they are constantly working on improving it. They have also published a book named “**Data Mining** Practical Machine Learning Tools and Techniques” written by Ian H. Witten, Eibe Frank and Mark A. Hall. This book’s latest edition uses WEKA3.6-stable as toolkit which we have used as backend API for Gamified Online WEKA. You can order this book from [here](#).

B.2 WEKA Course:

Waikato University also offers an online course named “Data Mining with Weka” which provides video lessons, Practical exercises with WEKA, reading from datamining book, online assessment system and Course Completion Certificate. The details of the course are [here](#). Anybody can take this course to improve his datamining skills and to learn WEKA datamining in detail. Professor Ian H. Witten from Department of Computer Science Waikato University is instructor for this course.

Bibliography

- [1] Fernando Perez. (n.d.). Retrieved from Plotly: <https://plot.ly/>
- [2] Frank, T. C. (2016). Introducing Machine Learning Concepts with WEKA. In E. D. Mathé, *Statistical Genomics: Methods and Protocols* (pp. 353-378). New York.
- [3] *Google Charts*. (n.d.). Retrieved from <https://developers.google.com/chart/interactive/docs/gallery?hl=en>
- [4] Hall, R. R. (2010). WEKA-Experiences with a Java Open-Source Project. *Journal of Machine Learning Research*, 2533-2541.
- [5] Highsoft. (n.d.). *Stacked Bar Chart*. Retrieved from HighCharts: <http://www.highcharts.com/demo/bar-stacked>
- [6] Hunt, L. a. (2011). Clustering mixed data. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 352-361.
- [7] Ian H. Witten, E. F. (2011). *Data Mining Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers .
- [8] *Js tree JQuery Tree Plugin*. (n.d.). Retrieved from <https://www.jstree.com/>
- [9] Kepes, B. (n.d.). *Understanding the cloud computing stack saas,paas and iaas*. Retrieved from RACKSPACE SUPPORT NETWORK: <https://support.rackspace.com/white-paper/understanding-the-cloud-computing-stack-saas-paas-iaas/>
- [10] Kosorus, H. (2011). Using R, WEKA and RapidMiner in Time Series Analysis of Sensor Data for Structural Health Monitoring. *Database and Expert Systems Applications (DEXA), 22nd International Workshop*.
- [11] Kruse, M. (n.d.). *Tree Structure*. Retrieved from JavaScript ToolBox: <http://www.javascripttoolbox.com/lib/mktree/>
- [12] Mark Hall, E. F. (2009, June 1). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*.
- [13] *OpenShift Web Hosting*. (n.d.). Retrieved 2016, from OPenShift: <https://www.openshift.com/>
- [14] *Research Methodology Types*. (n.d.). Retrieved from <http://www.alzheimer-europe.org/Research/Understanding-dementia-research/Types-of-research/The-four-main-approaches>

- [15] Schmuecker, R. (n.d.). *D3.js Drag and Drop, Zoomable, Panning, Collapsible Tree with auto-sizing*. Retrieved from D3 Data-Driven Documents:
<http://bl.ocks.org/robschmuecker/7880033>
- [16] Science, F. o. (2013). Orange: Data Mining Toolbox in Python. *Journal of Machine Learning Research* 14.
- [17] f. (n.d.). *Chart Fiddles*. Retrieved from JavaScript charts for web & mobile:
<http://www.fusioncharts.com/javascript-chart-fiddles/>
- [18] WAIKATO, U. o. (n.d.). *Weka 3: Data Mining Software in Java*. Retrieved from WEKA: <http://www.cs.waikato.ac.nz/ml/weka/>
- [19] WIKATO, T. U. (n.d.). *Use WEKA in your Java code*. Retrieved from WEKA wiki:
<http://weka.wikispaces.com/>
- [20] Zhao, Y. (2015, May 8). *R and Data Mining*. Retrieved from Introduction to Data Mining with R: <http://www.rdatamining.com/docs/introduction-to-data-mining-with-r>
- [21] ZingChart. (n.d.). *ZingChart demo*. Retrieved from ZingChart:
<http://www.zingchart.com/gallery/chart/#!stacked-bar-hooked-labels>