# DESIGN AND IMPLEMENTATION OF CLASSIFICATION USING DECISION TREE IN WEB USAGE MINING

*Dissertation submitted in fulfilment of the requirements for the Degree of*

**MASTER OF TECHNOLOGY**

**in**

**COMPUTER SCIENCE AND ENGINEERING**


By

**LOVELEEN**

**(11410599)**

Supervisor

**Ms. Shilpa Sharma**




**School of Computer Science and Engineering**

Lovely Professional University

Phagwara, Punjab (India)


December 2016

December 2016

# ABSTRACT

Web Usage Mining is the implementation of data mining procedures to find interesting usage patterns from Web information, with a specific end goal to comprehend and better serve the needs of Web based applications. This work presents the summary of various techniques of web mining and Data Mining Techniques in various application domains in addition to introducing the classification based on Hybrid Clustering with Classification approach. The paper takes a step ahead in this direction and proposes a Hierarchical Clustering with Improved Bagging Algorithm classification. The hybrid approach has been simulated using WEKA. The simulation has been done in Java Net Beans. It can be concluded that by implementing J48 decision tree, performance and accuracy has been compared and improved.

# DECLARATION STATEMENT

I hereby declare that the research work reported in the dissertation entitled "**DESIGN AND IMPLEMENTATION OF CLASSIFICATION USING DECISION TREE IN WEB USAGE MINING**" in partial fulfilment of the requirement for the award of Degree for Master of Technology in Computer Science and Engineering at Lovely Professional University, Phagwara, Punjab is an authentic work carried out under supervision of my research supervisor Ms. Shilpa Sharma. I have not submitted this work elsewhere for any degree or diploma.

I understand that the work presented herewith is in direct compliance with Lovely Professional University's Policy on plagiarism, intellectual property rights, and highest standards of moral and ethical conduct. Therefore, to the best of my knowledge, the content of this dissertation represents authentic and honest research effort conducted, in its entirety, by me. I am fully responsible for the contents of my dissertation work.

*Signature of Candidate*

**Loveleen**

**Reg. No – 11410599**

# SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the M.Tech Dissertation entitled "**DESIGN AND IMPLEMENTATION OF CLASSIFICATION USING DECISION TREE IN WEB USAGE MINING**", submitted by Loveleen at Lovely Professional University, Phagwara, India is a bonafide record of her original work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

*Signature of Supervisor*

Ms. Shilpa Sharma

**Counter Signed by:** **Date:**

1) **HoD's Signature: _____**

   HoD Name: _____

   Date: _____

2) **Neutral Examiners:**

   **(i) Examiner 1**

   Signature: _____

   Name: _____

   Date: _____

   **(ii) Examiner 2**

   Signature: _____

   Name: _____

   Date: _____

# ACKNOWLEDGMENTS

This research work is made possible through the help and support from everyone, including my parents, teachers, family, friends, and in essence, all sentient beings. Especially, please allow me to dedicate my acknowledgment of gratitude toward the following significant advisors and contributors:

I am proud to express my gratitude to my supervisor Ms. Shilpa Sharma for providing me with an opportunity to work under his valuable guidance, without her inspiration, efforts and top notch advices this thesis would not have been possible. Her meticulous approach has improved the precision and clarity of my writing. I sincerely thank my family and friends for unconditional support, trust and infinite love because of their encouragement, I would have been able to complete my work with zeal. It is also a pleasure to praise my friends for their valuable help and debugging some of the problems during the studies.

# TABLE OF CONTENTS

**CONTENTS** **PAGE NO.**

# LIST OF FIGURES

**FIGURE NO. : FIGURE DESCRIPTION** **PAGE NO.**

# LIST OF TABLES

**TABLE NO. : TABLE DESCRIPTION** **PAGE NO.**

# Chapter 1

# INTRODUCTION

The recent growth of content on the Internet has made it steadily troublesome for the clients to discover and use information and content providers find it difficult to classify and catalogs documents. It gets highly tiring for users to browse with traditional web search engines as they often return hundreds or a great many results for a search. Libraries on-line, search engines, and other large repositories (e.g. customer support databases, product specification databases, press release archives, news, story archives, etc.) are growing so rapidly that it is troublesome and enormous to classify each record physically. Keeping in mind the end goal to manage these issues, analysts look toward automated methods of working with web contents. So they can all be the more adequately browsed, sorted out, and indexed with negligible human intervention.

## 1.1 Web Mining

Web information mining picks up its significance with the expanding measure of Web data that is turning out to be much bigger than any customary data sources. Web data mining includes applying different techniques of data mining to Web content. It emphasizes on Web pages link structure, their possessed content and their utilization. Web data mining is a procedure that finds the connections between Web information, generally communicated in the form of text, linkage or usage information, by means of investigating the elements of the Web and online information utilizing data mining procedures. Specially, Web usage pattern are discovered via Web usage mining and then are utilized for presenting Web users with more personalized Web contents. Web pages are Hypertext documents, which contain both text and hyperlinks to other documents. Furthermore, web data are heterogeneous and dynamic. Thus, design and the implementation of a web data mining research support system has become a challenging task for researchers in order to utilize useful information from the web.

A client interfaces with the Web where there is a wide diversity of client's anavigational inclination, which brings about requiring diverse content and presentations of data. To enhance the quality of the Internet administration and build the

user click rate on a particular site, the most vital thing for a Web developer or designer is to comprehend what the client truly needs to do, foresee which pages the client is conceivably inspired by, and present the customized Web pages to the client by learning client navigational pattern knowledge.

The web mining task can be categorized to following sub tasks:

- Resource Finding: includes the retrieval of the indented Web documents.
- Information Selection and Pre-processing: automatic pre-processing and selection of a specific information from extracted web sources.
- Generalization : This includes automatically discovering general patterns at individual web sites as across multiple sites.
- Analysis: validation and /or interpretation of the mined patterns.

The working of web mining includes three stages: preprocessing, pattern discovery and pattern analysis. The pattern discovery has been a noteworthy study for improving the efficiency of various applications based on web. These days, web based applications offer many personalized experience for their users which make it extremely important to form some kind of interaction with Web users and always be a step ahead of them when it comes to prediction of next accessed pages. For instance, knowing all about the user's browsing history on the site helps to know which one of the most frequently accessed pages will be accessed next along with some extra information like the type of users and the user's preferences. Such objectives can be achieved by extracting useful patterns and knowledge applying different techniques and tools. Each of these pattern discovery techniques has its own strengths and weaknesses.

## 1.2 Web Data

Web data can be categorized into following categories:

A. Content data are presented to the end-user in a properly structured manner. Content Data can be simple text, images, or structured data, such as information retrieved from databases.

B. Structure data is the representation of how the content is organized. They can be either data entities used within a Web page or data entities used to put a Web site together. HTML or XML tags belong to the former and the latter includes hyperlinks connecting one page to another.

C. Web site's usage is represented by Usage Data. This type of data includes a visitor's IP address, access time/date, complete path of files or directories accessed, referrers' address, and other attributes that can be included in a Web access log.

D. User profile data informs about the Web site users. A user profile usually contains demographic information for each user of a Web site, as well as information about users' interests and preferences which is acquired through registration forms or questionnaires, or can be inferred by analyzing Web usage logs.

## 1.3 Web Mining Techniques

World Wide Web has developed in a past couple of years from a little research group to the greatest and most known method for information dissemination and communication. It has lot of information and keeps on expanding enomorously in size and complexity with time. It is extremely colossal task to seek relevant information from an enormous amount of data.

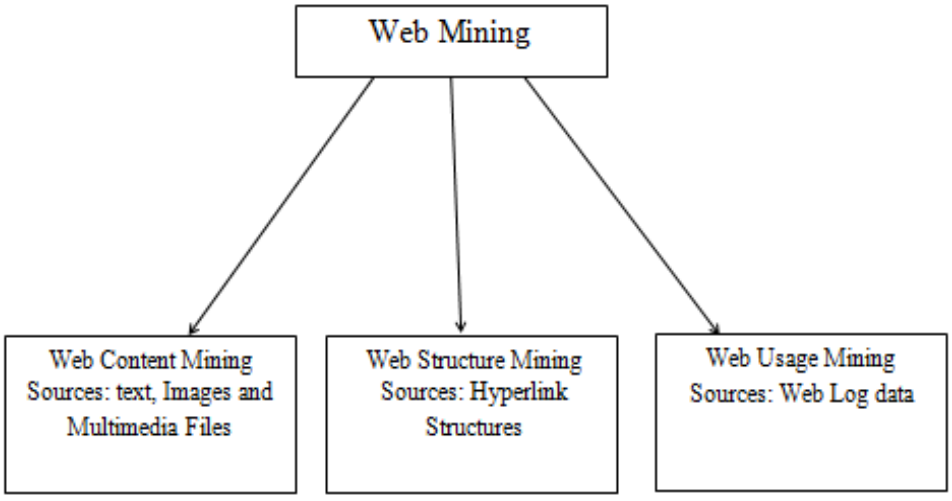



**Figure 1.1:** Web Mining Techniques

### 1.3.1 Web Content Mining:

As the name itself suggests, Web Content Mining includes extracting of useful information from the contents of Web documents. In actual, the content data is the collection of facts which are to convey to the users. It may consist of text, images, audio, video, or structured records such as lists and tables. Research activities in this field also involve using techniques from other disciplines such as Information Retrieval (IR) and natural language processing (NLP).

Web Mining involves various techniques for summarizing, classification and clustering of the web contents. It can provide useful and interesting patterns about user needs and contribution behavior.

As explained by Govind Murari Upadhyay, some of the prominent web content mining techniques are Unstructured, Structured, Semi structured and Multimedia data mining techniques. These techniques are explained in brief in this section to gain the interest of the reader in the current techniques.

#### 1.3.1.1 Unstructured data mining techniques

In unstructured category of web mining, the web pages are in the form of text. As per this technique, the data is searched and extracted. It is a bit much that the information which is recovered is significant information, it might be obscure data. Some tools or techniques needed to be used to get relevant information from that retrieved data.

#### 1.3.1.2 Structured data mining techniques

Structured data extraction is an advanced step taken ahead in information retrieval from web pages. A program called wrapper is usually used for extracting such data. Unlike the unstructured category, structured data are particularly the data records retrieved from underlying database and displayed in the web pages following some templates where the template can be a table or sometimes a form. It is useful to extract such data records as it facilitates to obtain and integrate data from multiple sources hence further able to provide value-added services.

#### 1.3.1.3 Semi structured data mining techniques

The third class semi-structured data is a state of union for the Web and database groups where the Web deals with documents while database communities deal with data. That data form is developing from rigidly organized relational tables including numbers and strings to empower the regular representation of complex real world objects like books, papers, movies and so forth, avoiding sending the application writer into further contortions. Rising representations for semi-organized information are varieties on the Object Exchange Model (OEM) where data is in the form of atomic or compound objects. Atomic objects refer to integers or strings whereas compound objects may include other objects through marked edges. HTML is a unique instance of intra-document structure. Clients not just query the data to locate a specific bit of information; however they are likewise interested in understanding the query. On account of this assortment, semi - structured DBs don't accompany a conceptual schema. To make these databases more available to clients a rich conceptual model is required. Conventional recovering systems are not straightforwardly applied on these databases.

#### 1.3.1.4 Multimedia data mining techniques

As the name itself is giving hint, Multimedia data mining can be characterized as the procedure of discovering interesting patterns from media data such as audio, video, image and text which basic queries and associated results cannot access. Multimedia data mining highly contributes to enhance decision making using discovered patterns. Hence Multimedia data mining has succeeded in attracting appropriate research endeavours in creating techniques and tools to compose, manage, search and perform domain specific tasks for data from different domains such as surveillance, meetings, broadcast news, archives, medical data, as well as personal and online media collections.

### 1.3.2 Web Structure Mining

Typical Web graph structure consist of Web pages as nodes, and hyperlinks as edges interfacing between two related pages. In addition, the content within a Web page can likewise be composed in a tree like structure format according to the various and XML tags within the page. Thus, Web Structure Mining can be viewed as the procedure of

discovering structure information from the Web. This kind of mining can be performed either at the (intra-page) document level or at the (inter-page) hyperlink level.

The biggest hurdle in the way of Web structure mining is to manage the structure of the hyperlinks within the Web itself. With the escalating interest in Web mining, the exploration of structure analysis has increased and the endeavours has brought about a recently rising exploration zone called Link Mining which is located at the intersection of the work in link analysis, hypertext and web mining, relational learning and inductive logic programming, and graph mining. In addition to Internet, there is a conceivably extensive variety of utilization scope for this new domain of research.

#### 1.3.2.1 Hyperlinks

A hyperlink is an auxiliary unit that interfaces a location in a web page to an alternate location, either inside of the same website page or on an alternate page. A hyperlink that joins with an alternate piece of the same page is called an intra-document hyperlink, and a hyperlink that interfaces two distinct pages is called an inter-document hyperlink.

#### 1.3.2.2 Document Structure

Furthermore, the content inside of a Web page can likewise be sorted out in a tree organized configuration, in light of the different HTML and XML tags inside of the page. Mining endeavours here have concentrated on consequently extracting document object model (DOM) structures out of documents.

### 1.3.3 Web Usage Mining:

Web Usage Mining is the implementation of data mining procedures to find interesting usage patterns from Web information, with a specific end goal to comprehend and better serve the needs of Web based applications. Due to the user's interactions with one or more Web sites, web usage mining involves the automatic discovery and analysis of patterns in data. It concentrates on tools and methods that are typically intended to study and comprehend the users' navigation inclinations and behaviour by discovering their Web access patterns. These strategies are highly effective means that help e-commerce businesses improve their Web sites in an efficient manner. Furthermore, usage

data captures the identity or origin of web users along with their browsing behaviour at a web site.

Taking into account the kind of usage data, web usage mining itself can be classified further into Web Server Data, Application Server Data, and Application Level Data. These are briefly explained in the section.

#### 1.3.3.1 Web Server Data

Web server collect user logs and it typically includes IP address, page reference and access time, date of accesing the data, web cache, web accesing information of the user, details,etc..

#### 1.3.3.2 Application Server Data

Commercial application servers such as Web logic, Story Server, have significant features to enable E-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs.

#### 1.3.3.3 Application Level Data

New kinds of events can be defined in an application, and logging can be turned on for them which facilitate generating histories of these events.

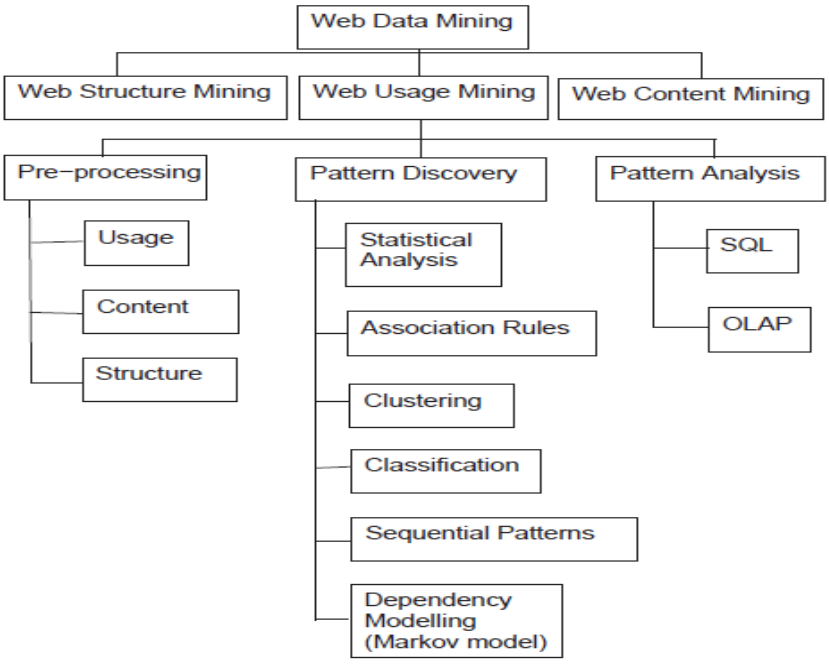



**Figure 1.2:** Web Mining Architecture defining different techniques used in web mining

### 1.3.4 Web Usage Mining Concepts

#### 1.3.4.1 Data accumulation

The first step of web usage mining is Data accumulation. The data authenticity and integrality directly affects the subsequent works smoothly carrying on and the final recommendation of characteristic service's quality.

#### 1.3.4.2 Data preprocessing

Web data could take numerous forms. The primary data sources include the server log files further including Web server access logs and application server logs. The additional data sources may include operational databases, domain knowledge, site files and meta-data. IN prior to mining technique, web data has to be cleaned, filtered, integrated and pre-processed. Preprocessing converts the data to be used by pattern discovery phase. This phenomenon is most time consuming and intensive step. The web usage mining process which explains various important pieces of information such as who accessed the web site page, what pages accessed, the pages accessed order and total time spent on each page, takes a user session file as an input. It changes Web log files into Web transaction data which are further handled by data mining tasks. The data pre-treatment

work, mainly include data cleaning, user identification, session identification and path completion.

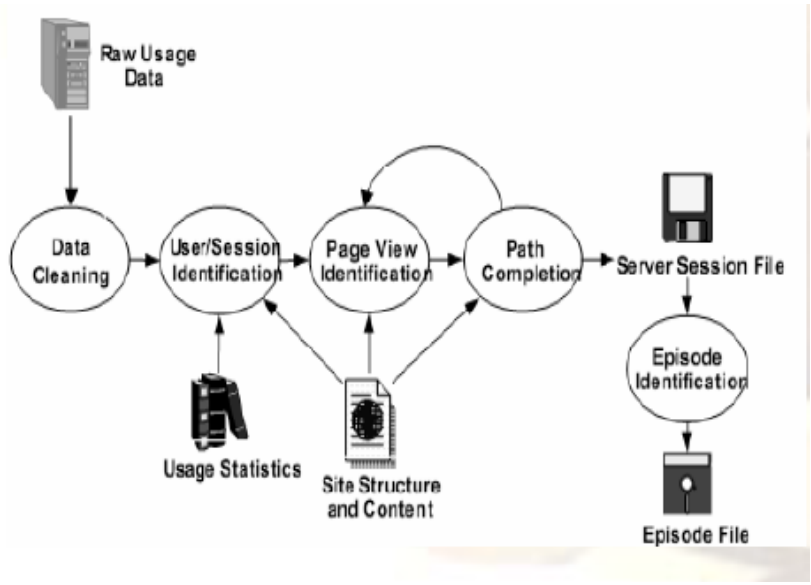



**Figure 1.3:** Preprocessing of web usage data including data cleaning, path completion, page view etc.

### 1.3.4.3 Data Cleaning

The motivation behind data cleaning is to remove the unessential irrelevant items. Such sorts of techniques are of great significance for not only data mining but also for any kind of web log analysis. As indicated by the reasons of distinctive mining applications, irrelevant records in web access log will be disposed of during data cleaning. The records have filename extensions of GIF, JPEG, CSS, and so on, which can found in the URI field of the each record**.**

### 1.3.4.4 User and Session Identification

In user and session identification task, different user sessions are discovered from the original web access log. The main definition of User's identification is to identify who access web site and which pages are accessed. The sequence of activities performed by a user from the moment he enters the website to the moment he leave the web site is referred to as a session. Session identification can be performed using time interval between consecutive log entries.

Followed by data cleaning, user identification can be performed by various heuristics:

A. A piece of knowledge can be exposed by converting IP address into domain name. For instance, a visitor's location can be figured out by just looking at the suffix of each visitor's domain name, such as .ca (Canada); .au (Australia); .cn (China), etc. That's really appreciable idea.
B. The well-known concept of cookies can also help in this. Whenever a browser connects the web server for first time, the web server randomly assigns it an Id which is called cookies. The Web browser sends the same ID back to the Web server, hence informing the Web site about the return of a particular user. Cookies contribute a lot in helping out a Web site developer to easily identify individual visitors, subsequently getting a greater understanding of the website usage. Cookies also help visitors by allowing Web sites to recognize repeat visits.

**Path Completion:** Sometimes some important accesses occurred due to local cache, agent cache, post technique or browser's back button, fail to get recorded in the access log files and moreover, the recorded count of Uniform Resource Locators (URLs)in the log may be less than the real existing ones. These circumstances lead to Path Incompletion state. Due to this, the client access paths are not completely preserved in the web access log. To find out user's travel pattern, it is must to append the missing pages in the user access path which is the main motivation behind the path completion. The better results of data pre-handling include the enhanced mined pattern's quality and reduced algorithm's running time. The final procedure of the preprocessing is formatting, which is a preparation module to properly format the sessions or transactions.

**Knowledge Discovery:** The pre-treated data is mined and analyzed using statistical method. The user or the user community's interests can be discovered to further construct the interest model. At present the usually used machine learning methods mainly have clustering, classification, the relation discovery and the order model discovery.

**Pattern analysis:** Pattern Analysis is a final stage of the whole Web usage mining. Challenges of Pattern Analysis are to filter and eliminate uninteresting information and to visualize and extract the interesting patterns. The pattern analysis methodology includes first deleting the less significant rules or models from the interested model

storehouse; then use of OLAP technology to carry on the comprehensive mining and analysis; further to let discovered data or knowledge be visible; and Finally, providing the characteristic service to the electronic commerce website.

## 1.4 Web personalization

The Web users' content personalization and recommendation of appropriate Web pages enables developers to supply users with what they require based on their previous interactions within the same Web site. If personalization of the visitor's experience is correctly implemented and executed, it makes his time on the site or in application more productive and engaging. Personalization can be highly valuable to any individual or any organization, because it drives desired business results such as increasing visitor response or promoting customer retention. A personalization mechanism is based on explicit preference declarations by the user and on an iterative process of monitoring the user navigation, collecting its requests of onto logical objects and storing them in its profile in order to deliver personalized content .

## 1.5 Data Mining

As the innovations for creating and collecting data have been progressing quickly, at the present stage, lack of data has no more being an issue rather the issue is to generate useful information from available data. Due to the explosive growth in data and database, there arises the need to develop new technologies and tools to process data into useful information and knowledge intelligently and automatically. Hence, Data mining (DM) has turned into a research territory with expanding significance. Data Mining is defined as the procedure of extracting information from extensive data sets using algorithms and techniques derived from the field of Statistics, Machine Learning and Data Base Management Systems.

Data mining, popularly referred as Knowledge Discovery in Databases (KDD), is the nontrivial extraction of implicit, already unknown and conceivably useful information from data in databases. In fact, it is actually the process of finding the hidden information/pattern of the repositories.

Data mining has been a potential tool to analyze data from distinctive points and angles and getting useful information from chunk of raw data. It can likewise help in predicting patterns or values, classification of data, categorization of data, and to find correlations, patterns from the dataset. While on the other hand, technical challenges such as data storage and data transfer are arising due to use of immeasurable measure of data. The management of data resources and flow between the storage and compute resources is becoming the primary bottleneck. The domain of data mining has been prospered and has breaded into new areas of human life with various integrations and advancements in the fields of Statistics, Databases, Machine Learning, Pattern Reorganization, Artificial Intelligence and Computation capabilities etc. Almost every corner of human life has become data-intensive making the data mining an essential component. The term data mining signifies the action of extracting new, valuable and nontrivial information from extensive volumes of data. Usually, the motive is to discoverSpatterns or construct models using specific algorithms from various scientific disciplines including artificial intelligence, machine learning, database systems and statistics. The data mining tasks can be classified into Predictive data mining and Descriptive data mining. Predictive data mining with the objective to build an executable model from data which can be efficiently used for classification, prediction or estimation, and Descriptive data mining where the goal is to discoverSinteresting patterns and relationships in data.

### 1.5.1 Data Mining Process

#### 1.5.1.1 Selection

The primary step begins with gathering the essential and important knowledge about the subject of interest and setting the objectives to be accomplished. This information is then utilized as a part of the preparation of a dataset which incorporates selecting a proper sub set of data samples and/or variables.

#### 1.5.1.2 Cleaning and Preprocessing

It includes finding incorrect or missing data. It also includes removal of noise or outliers, collecting necessary information to model or account for noise, accounting for time sequence information and known changes.

#### 1.5.1.3 Transformation

It involves transforming the data into a common format for processing. Some data may be encoded or transformed into more usable configuration. Data reduction, dimensionality reduction & data transformation method may be used to reduce the number of possible data values being considered.

#### 1.5.1.4 Data Mining

This is considered as the most elaborate step as it consists of choosing the function of data mining, choosing the right data mining algorithm and its application. Choosing the function includes deciding the purpose of the resulting data mining model, such as classification, regression, clustering and summarization. The selection of the data mining algorithm encompasses the decision which models and parameters are appropriate and matching with the criteria of the process (i.e., trade-off between the predictive power of the data mining model and its understandability).

#### 1.5.1.5 Interpretation/Evaluation

Towards the final step the discovered patterns are evaluated and their validity and relevance are assessed. Redundant and unessential patterns are removed while the remaining, relevant patterns are studied and interpreted, typically with the aid of computer-assisted visualization techniques. Application of the discovered knowledge includes resolving potential conflicts with existing knowledge, taking actions based on the obtained knowledge, such as aiding, modifying and improving existing processes and procedures, especially those involving human experts , and storing, documenting and reporting to interested parties.

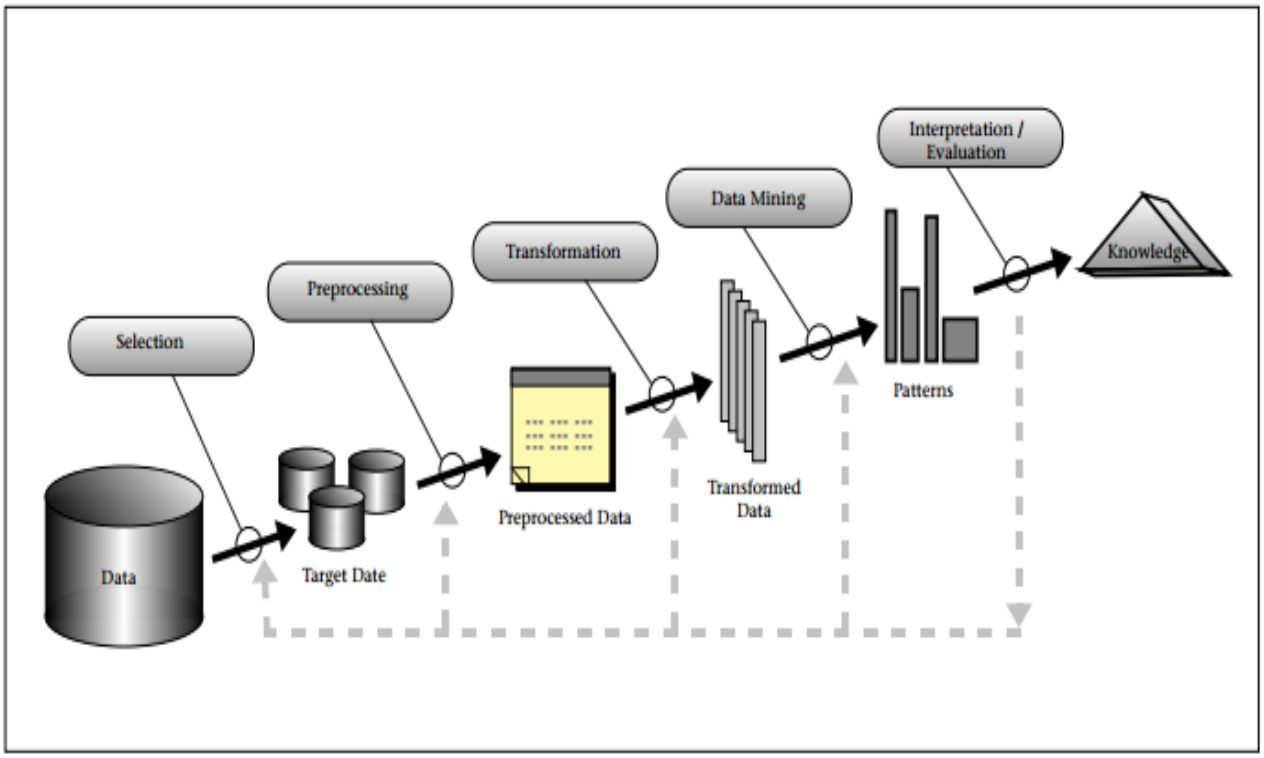



**Figure 1.4:** KDD Steps

### 1.5.2 Data Mining Techniques

The various data mining techniques are classification,clustering, prediction, association rule, and neural networks. These techniques are necessary to study and explained in this section.

#### 1.5.2.1 Classification

Classification is the most usually implemented data mining technique, which employs a set of pre-classified samples to create a model that can classify the vast population of records at a grand level. This particular sort of analysis includes fraud detection and credit risk applications. This approach in most possible ways utilizes decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. Where in Learning, the training data are analyzed by classification algorithm on the other hand in classification, test data are used to estimate the accuracy of the classification rules. If the accuracy is found to be adequate, the rules very smoothly go for the new data tuples. In case of fraud detection application, complete records of both fraudulent and valid activities are included, which are derived on a record-by-record premise. Likewise,

the classifier-training algorithm derives the set of parameters required for proper discrimination using pre-classified samples. The algorithm then encodes these parameters into a model called a classifier.

Types of classification models:

- PART
- Classification by decision tree induction
- Bayesian Classification
- Neural Networks
- Support Vector Machines (SVM)
- Classification Based on Associations

#### 1.5.2.2 Clustering

The procedure of distinguishing similar classes of objects is called clustering. Using clustering techniques, dense and sparse regions can be discovered in item space and subsequently the overall distribution pattern and relationships among data attributes can be identified. This approach can also be used for effective means of distinguishing groups or classes of object but it turns out to be expensive so clustering can be used as preprocessing approach for attribute subset selection and classification.

Types of clustering methods:

- Partitioning Methods
- Hierarchical Agglomerative (divisive) methods
- Density based methods
- Grid-based methods
- Model-based methods

#### 1.5.2.3 Prediction

For prediction, regression technique can be adapted. Regression analysis can be utilized to model the correlation between one or more independent variables and dependent variables. In data mining, already known attributes are independent variables and response

variables are what are yet to be predicted. In real world, many real-world problems are not simply prediction, for instance, sales volumes, stock prices, and product failure rates are all very tuff to predict as they may depend on complex interactions of multiple predictor variables which need more complex techniques to forecast future values. The same model types can often be used for both regression and classification. For example, the PART i.e. partial decision algorithm can be used to generate the decision list that shows the output. Neural networks too can create both classification and regression models.

Types of regression methods:

- Linear Regression
- Multivariate Linear Regression
- Nonlinear Regression
- Multivariate Nonlinear Regression

#### 1.5.2.4 Association Rule

Association and correlation is usually to find interesting relationships among data items of abundant data sets. With excessive amounts of data continuously being collected and stored in databases, many companies are interested in association mining rules from their databases to increase their profits. This sort offending highly helps businesses to make decisions, eg. catalogue design, cross marketing and customer shopping behavior analysis. Association Rule algorithms need to be able to generate rules with confidence values less than one. However quantity of conceivable Association Rules for a given dataset is for the most part substantial and a high extent of the rules are typically of little value.

Types of association rule

- Multilevel association rule
- Multidimensional association rule
- Quantitative association rule

#### 1.5.2.5 Neural Networks

Neural network is a set of connection of input/output units and every connection has a weight present with it. They can be used to demonstrate complex relationships between

inputs and outputs or to find patterns in data. Numerous data warehousing firms are harvesting information from datasets using neural networks as a tool. Neural networks have the remarkable ability to derive meaning from complicated data one can extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. These are well suited for continuous valued inputs and outputs.

# Chapter 2

# LITERATURE REVIEW

**K. Sudheer Reddy, G. ParthaSaradhi Varma, and M. Kantha Reddy, "An Effective Preprocessing Method for Web UsageMining" (2014)**

This research paper studies and presents various data preparation techniques of access stream to be performed before begin of them in process and these are utilized to enhance the performance of the data preprocessing in order to figure out the uniques essions and unique users. The methods proposed contribute to retrieve meaningful pattern and relationships from the access stream of the user and these are proved to be valid and useful by various research tests. The paper concludes by proposing the future research directions.

**Prabhjot Kaur, "Web Content Classification: A Survey" (2014)**

The paper clarifies that the increment in the measure of information on the Web has brought on the requirement for exact automated classifiers for Web pages to keep up Web indexes and to build web engine performance. The paper exhibits an outline of the web classification research regarding its elements and algorithms.

**M. Sujatha, S.Prabhakar, Dr. Lavanya Devi, ""A Survey of Classification Techniques in Data Mining (2013)**

This paper defines the different techniques of classification in web mining. These techniques are:

Genetic algorithm- association rule mining technique is used in GA, used to find undetermined solutions. It is implements on small groups of different data.

Rules sets- This classification rule is "if-then-rule".

C4.5 – The aldorithm measures numeric attributes deals with missing values and pruning noisy data. Pruning is used in C4.5 to avoid over fitting to noise in data.

CART- it is classification and regression tree based on accuracy, when data is noisy/ missing values. It takes random sample and allows handling missing values by CHAID algorithm. Data- preprocessing is not required in data mining. It automatically selects relevant attributes.

Decision Tree Induction- Decision tree generated by top-down approach and conquers approach. The objective of this algorithm is to reduce the impurity or uncertainty in data as much as possible

Bayesian network- Bayesian network is based on DAG and one to one method. This network is divided into two tasks i.e. network structure is fixed, learning the parameter in the conditional probability tables second is if the structure is unknown, a concept is scoring function that evaluates the "fitnessness" of network.

Instance based learning- Instance based is a type of learning algorithm. It consumes more computation time for classification

Support Vector Machine- This is a technique of linear and non-linear data. SVM uses mapping to transform the original trained data.

K-nearest Neighbour (KNN)- KNN is a simple but effective in many cases but it has many drawbacks like it is low efficient, dependency of this algorithm is based on good selection of k.

**Vivek Agarwl, Saket Thakre, Akshay Jaiswal "Survey on Classification Techniques of Data Mining" (2015)**

Classification technique processes different types of data. The various techniques of classification are: Decision Tree and K-Mean Algorithm, naïve bayes, Sequential mining optimization (SMO) support vector machine (SVM).

a. Decision tree algorithm:

A decision tree is a kind of decision support system which deploys a graph decision tree and their possible action, probability results, resources costs and value. All the possible actions, and resources are thus viewed as a tree, with the resources leading to the consequent solumns produced.

b. K-mean algorithm:

K-mean is a basic, simple partition cluster technique which works to search a user-specified k number of clusters. The centroid informs these clusters that is typically the mean of the point in the cluster.

c. Naïve Bayes:

Naïve bayes is a probability based classifier used for training a categorized set of documents. It is based on the bayes theorem. Classifier has two categories i.e. training and classification stages.

d. Sequential mining optimization (SMO):

The SMO systematically solves the optimization problem when training the support vector machines.

e. Support vector machines (SVM):

The algorithms utilizes by SVMs is based on substitution method. It can be defined as systems that use hypothesis space of linear functions and in a high dimensional feature space.

**R. Fernandes, L. J. Peo, N. Kamat, and S. Miranda (2014)**

Characterizes that the strategy of information mining to concentrate learning from the database. The diverse methods of web mining are web content mining, web structure mining and web structure mining. Content mining gathers actualities of a site page to give better results to the client. In the structure mining there are common web chart which comprises of hubs and hyperlinks to associate distinctive edges with related pages. There are different HTML and XML labels inside the page. In this procedure diverse organized data are found from the site page. This sort of mining is performed on either between page (hyperlink level) or at the intra-page (report level). Web use mining method is utilized to find fascinating use designs from the site pages to comprehend the necessities and give better results. Information base uses measurable techniques for the investigation and mining to anticipate information. We can extricate information from client's intrigued information and create distinctive model. The most regularly utilized machine learning strategies are grouping, characterization, connection and request revelation. Web personalization is a methodology to gather guest's data in view of clients went by data then utilize that information for substance conveyance

system. It can likewise be helpful for control the data you present to client. S. Vijiyarani and M. E. Suganya (2015)

The paper enlightens the major issues in web mining, which are:

- Web data contains very large data, hundreds of TB is required to store the on the database.
- It needs large number of servers to generate the data. A single server cannot do that.
- There is to properly organize the hardware software to mine terabytes of data.
- There is limited query interface to user.
- There is no automatic data cleaning.
- Overlapping of data.
- Difficult to find relevant data or information.

**K. B. Patel, A. R. Patel, and N. S. Pate (2014)**

The paper elaborated that web personalization plays an important role in Web Advertisements. Web advertisements are in the form of banners, images, graphical elements, animation videos etc. The visitors are forced to click on the image for more information. Advertiser wants to attract the visitors. The advertisements should be according to the visitor's behaviour which will increase the effectiveness of the web site. Web personalization is used to track the user browsing behaviour. Then it divides the customers into different groups. After the pages visited by users, appropriate advertisements are assigned to each active user by using fuzzy rules. Advertisers use different web sites to display the advertises and publish by paying to them. Google is one of the most popular sites to publish these advertisements. Based on some analyzed content, advertisements are displayed on Google page. There are different types of models are used for Advertising Management Process which are Agent Advertising Model, Publisher Advertisement Model, Advertiser Model of Advertising. Different types of data are required to offer customers interested information. Data can be use's IP address, browser detail i.e. send through HTTP request, navigation pattern and user profile.

**J. Srivastava, R. Cooley, M. Deshpande, and P. Tan (2015)**

The paper provided improved way to predict user's browsing behaviour. K-mean clustering and Regression Analysis algorithms are used to predict the future request. Personalization is used for filtering the information. It also customizes the data according to the users need and provides correct results to the user. By using web log files, user's navigation pattern, recommended systems recommends the web files to user in form of recommended lists. Self organizing map is used for the pre-processing of the web log files. K-mean clustering is a iterative strategy technique divide the data into different groups. Firstly data is assigned then relocation of mean is done.

**V. Dongre (2015)**

The refers web usage mining, web browsing behaviour to predict the navigation pattern to provide recommendation to the user. Different types of web mining techniques are used to extract knowledge from the data source. All the above techniques of web mining are applied for providing better results to the user. Different type of data is fetched and analyzed from user's activities stored as log files. After that it discovers the information from various applications such as web prediction. Source of logs files contains Web server, Proxy servers and client caches. To get efficient result, we need to get log files from all the above sources. The Web server can't keep track of the data, which is kept in proxy server or client cache. Knowledge is extracted from various disciplines and then browsing behaviour is discovered. It also analyzes the pages in which users stay longer, which path they follow to navigate from one page to another. In which pages there is delay in searching for some information, which pages attract more attention. There is need to analyze the server log files at some intervals. So that server can perform well. He provides solutions based on LCS algorithm to analyze and process patterns for next web page predictions. In double algorithm, session is used to track the pattern in order to increase the performance of the system. It converts the session pattern into binary form then uses some search strategy like up and down to generate double candidate frequent item sets. Self Organizing map (SOM) is referred to apply pre-processing web log.

This algorithm is based on unsupervised learning. Now in next, we have Markov model which provides scalability to the system. It apply prediction algorithm on the training data. But sometimes it can't predict accurate results. A-priori algorithm emphasizes on usage data.

Memory utilization and time usage is compared by this algorithm. But the limitation of it is that creation of candidate set is expensive. If data set is large then long patterns are required. K-mean clustering algorithm is most commonly used for clustering method and which is much efficient than other. One can improve the performance of the system by using hybrid method.

**Jin Xu Yingping Huang Gregory Madey, "A Research Support System Framework For Web Data Mining" (2012)**

Defines that the technique of data mining to extract knowledge from the database. The different techniques of web mining are web content mining, web structure mining and web structure mining. Content mining collects facts of a web page to provide better results to the user. In the structure mining there are typical web graph which consists of nodes and hyperlinks to connect different edges with related pages. There are various HTML and XML tags within the page. In this process different structured information are discovered from the web page. This type of mining is performed on either inter-page (hyperlink level) or at the intra-page (document level). Web usage mining technique is used to discover interesting usage patterns from the web pages to understand the needs and provide better results. Knowledge base uses statistical methods for the analysis and mining to predict data. We can extract data from user's interested data and develop different model. The most commonly used machine learning methods are clustering, classification, relation and order discovery. Web personalization is a strategy to collect visitor's information based on users visited information then use that knowledge for content delivery framework. It can also be useful for manipulate the information you present to user..

**Peng PengQianli Ma Chaoxiong Li, "The Research and Implementation of Data Mining Component Library System" (2013)**

To improve the efficiencyeand quality of the reusing data mining software andfreduce the period and cost of developing data mining application system, this paper proposes a new component library system of data mining. Throughecomponentization of data mining algorithm, this system implements varied core algorithms of data mining in the form of components. In this way, the quality and efficiency of developing data mining software are

improved to meet various application demands. To improve the efficiencyeand quality of the reusing data mining software andfreduce the period and cost of developing data mining application system, this paper proposes a new component library system of data mining. Throughecomponentization of data mining algorithm, this system implements varied core algorithms of data mining in the form of components. In this way, the quality and efficiency of developing data mining software are improved to meet various application demands.**R. Shukla (2013)**

Provides solution on the basis of K-Mean Clustering algorithm for analysis of web log data. It is a method of clustering analysis which partitions the data into k clusters .the observation of these clusters contains the nearest mean to that cluster. It is an iterative process in which different the iteration provides better output to the system. it assigns the elements to different groups. Calculate the mean to the centroid of that observed cluster. There are different methods of this algorithm such as Forgery method which use data as initial mean. In Random Partition method we first adding assign cluster then proceeds for further process. It will first generate random cluster within the domain. Then clusters associated by using observation with the nearest mean. This method is fast, scalable. Mainly used to process large dataset and can be used for non –numeric data. That supports to classify the observation. On the other hand we have Hierarchical cluster Analysis used to find the relative homogenous cluster of case base. On measured characteristics, it starts with separate cluster then combines the cluster sequentially. It reduces the no. of clusters at each step until only one cluster is left. It observes that how homogenously clusters are formed. It is appropriate for finding small samples. When it deal with large data, system gets slow. K mean clustering consider sample size of more than 200.

**Jigna Ashish Patel(2015)**

This paper defines the classification algorithm which is applied to clustering the data. Data mining defined the useful information from large amount of data kept in database. It contains decision support system that uses graph decision of tree like structure and its repositories. Decision tree take most suitable concept in knowledge and information discovery as well as in data mining. K- mean is a different classification and clustering algorithm that clusters the data into different groups.

**Olga Tanaseinchuk, Alira Hadj Khodabakshi, Dimitri Prtrov, Jainwei Che, Tao Jinag, Bin Zhou, Andrey Santrosyan And Yingyao Zhou (2015)**

They defined the hierarchical clustering algorithm which is widely adopted technique of unsupervised learning technique. It identify objects groups within a given dataset, where intra- groups objects seems to be more similar than inter- group objects. Hierarchal clustering and k-mean are two most popular adopted algorithms used due to their simplicity in result interpretation. Hierarchical clustering is applied on large dataset as well as challenging.

**Jayalatchumy, Dr. P.Thambidurai, "Web Mining Research Issues and Future Directions – A Survey" (2013)**

This paper includes an overview on the current techniques of web mining and the difficulties related to it. Due to the colossal and assorted data on the web, the clients by one means or another neglect to make utilization of the data viably and effectively. Information mining concentrates on non-inconsequential extraction of undeniable effectively dark and potential accommodating information from the broad measure of data. This paper furthermore reports the outline of various strategies for web mining moved closer from the going with edges like Feature Extraction, Transformation and Representation and Data Mining Techniques in different application areas. The review on information mining framework is made with respect to Clustering, Classification, Sequence Pattern Mining, Association Rule Mining and Visualization. The exploration work done by unmistakable clients portraying the points of interest and hindrances are talked about. It in like manner gives the survey of progression in investigation of web mining and some fundamental examination issues related to it.

**Pradnyesh Bhisikar1,Prof. Amit Sahu, "Overview on Web Mining and Different Technique for Web Personalisation" (2013)**

This article gives an outline and examination of current web mining structure and headways. There are three general class in web mining i) web utilization mining ii) web content mining, iii) web structure mining. It explains the errands performed by every classification, for example, undertakings performed by web utilization mining incorporate information gathering, information arrangement, route design disclosure, patter examination,

design representation and example application. Through web content mining, one can extricate valuable data from the substance of web archives. This paper clarifies the idea of web personalization as the utilization of Web mining systems.**GovindMurariUpadhyay, KanikaDhingra, "Web Content Mining: Its Techniques and Uses" (2013)**

The center of this paper is to toss light on the idea of Web Content Mining. The paper gives knowledge into its strategies, forms and its applications in the current business environment also in examination and extricating contents for educational purposes. It further clarifies how web content mining assumes a critical part by getting rich arrangement of content and further using it in the decision making in the professional workplace, instruction and examination.

**A. N. Networks, M. I. T. Press, N. Networks, P. Hall, S. Maps, I. Exploration, and U. Learning (2011)**

The research paper suggested Self Organizing Map (SOM). It is the application of neural network and a category of competitive learning network. It is mainly used for clustering data without having the information of class membership of the input data. It detects the feature inherent problem and also known as SOFM. It provides a topology preservation map. The Map units or neurons form a two dimension data. It preserves the relative distance between the points. It trains and tests the data then provides output based on it.

**K. P. Adhiya and S. R. Kolhe (2015)**

The author in his paper suggested the terminology of An Efficient And Novel Approach For web Search Personalization Using Web Usage Mining. In this approach they have used the concept of personalization. Web personalization and web search results are carried out for users according to their needs. There are different techniques of web personalization and strategies. The customized web pages presented to the users according to their interest and needs, which is one of the main facility of web personalization. There are some components of web personalization such as:

- Preprocessing of web data such as content data, user data profile, usage data and structured data.

- Extraction of statistical information and discovering of interesting usage patterns using multiple data mining techniques.
- Recommendation through personalization system.

They used web log files. These log files contains the information of users navigation behaviour such as submit of request. It is also known as web access log which is stored in a web log file. The various sources of web logs are proxy servers, client browsers and web servers. Their proposed algorithm consists of Data collection and cleaning, user Identification, Sessionization, WAP generation and user profile creation. Data cleaning includes preprocessing phase which removes irrelevant entries and data from web access logs. For example the entries with code like 200 are considered as cleaned database and entries having status of ‘error’ or ‘failure’ is unclean data. Any user on the internet is identified by their IP address. User identification is done by the unique IP address. In sessionization, activity of user on web can be segmented into different sessions. It finds different sessions for various users. WAS generation and user profile creation done after sessionization that finds the set of pages accessed together.

**A.Hannak, P. Sapie, D. Lazer, and A. Mislove (2013)**

Refer Measuring personalization of web server that describes the personalization of web search. Different users get different results from the same searched query. The increasing demand of personalization leads to the concept of Filter Bubble effects, in which many users are unable to access information that search engine algorithm discovers irrelevant. Firstly they defined the methodology of measuring personalization in web search results. Then they applied methodology to different users on Google Web Search, they found that an average 11.7% of results showed differences due to personalization but it varies widely by search query and by result ranking. In third they investigated the causes of personalization on Google Web Search. Filter Bubble effect, where users only give results that the personalization algorithm thinks they want.

**S. S. Kontamwar and M. T. Cse (2015)**

The paper defines Clustering and Preprocessing for web Log Mining in which they describes the accessed documents in web log files. Application of data mining stores the navigated data in web log files. These log files contains user identification, session identification and clustered data. They examine the problem of getting reliable data. So, the data need to be pre-treated and behaviour of the user should be constructed as transformation. They used data cleaning process to remove irrelevant data. User identification associate page references with same IP address with different users. Session identification breaks user's references into user sessions. Path completion fills the missing page references in a session. They analyse the main problem of reliability in the dataset. Web session clustering classifies the data of web visitors on the basis of user click streams and similar measures. It is useful to manage the web resources efficiently such as web personalization, schema modification and web server performance. A session could be obtained through swarm optimization and appropriate similarity and then applied to web log data. The clustering algorithm arranges the data into similar groups from a database. Different types of data are kept in different groups and same types of data are kept in same type of combinations. The existing system lacks in scalability of data.

**Saucha Diwandari, Adhistya Ena Permanasari, Indriana Hiayah(2015)**

In this paper different data preprocessing is applied on the data. Different algorithms are applied on the data to conduct the evaluation of the website according to the user's customization. Client log files, web proxy server log and log files. Client log contain remote agent files that modify the source code to improve data collection. Data cleaning, user and session identification is conducted. Implicit evaluation is carried out on use's data and browsing behaviour. Different algorithms are applied in this paper such as k-mean, PART, SMO, and Naive Bayes. They evaluate the parameters and generate the results.

**Kwang Leng Goh, Ashutosh Kumar Singh (2015)**

They provide various comparison on machine learning algorithm which detects the web spams. This paper focused on the structure of the machine learning algorithm for web spam classification. They defines various algorithms on machine learning technique, such as Boosting algorithm, bagging algorithm, dagging algorithm, rotation forest algorithm.

**Sagar S. Nikam (2015)**

Data mining is widely used in area of machine learning, network intrusion detection, spam filtration, artificial intelligence, statistics and pattern recognition for the analysis of large dataset. Various classification algorithms are implemented on different datasets such as market data, patient's data, financial data etc. According to conditions, performance and features users can select the methods. Classification techniques are used to find data each and every instance within a given dataset. It classifies the data in many different classes according to different conditions. There are various techniques which discussed in previous papers. These are C4.5, ID3, k-nearest neighbour classifier, naive bayes, SVM, ANN. They conducted survey on these classification algorithm.

**P. Mehtaa, B. Parekh, K. Modi, and P. Solanki (2012)**

Alludes web personalization utilizing web mining methods. Web mining is the branch of information mining to separate significant information from web, for example, web report, hyperlinks between records, logs of sites and so on. There are two distinct ideas characterized in web mining, for example, Process driven view that characterizes web mining as an arrangement of errands and Data driven depicts the sorts of web mining as far as web information that is being utilized as a part of the mining procedure. The web mining plays out the assignments like:

- Resource finding
- Information determination and preprocessing
- Generalization
- Analysis

The thought process of web personalization is to give clients the information they need or need, without anticipating from clients to request them unequivocally. It requires the gathering of data and utilizations that information into substance conveyance system. It deciphers and controls the data you present to the client.

**S. S. Kontamwar and M. T. Cse (2015)**

Characterizes Clustering and Preprocessing for web Log Mining in which they depicts the got to reports in web log records. Use of information mining stores the explored information in web log documents. These log records contains client recognizable proof, session ID and bunched information. They look at the issue of getting solid information. Along these lines, the information should be pre-treated and conduct of the client ought to be developed as change. They utilized information cleaning procedure to expel immaterial information. Client recognizable proof partner page references with same IP address with various clients. Session distinguishing proof breaks client's references into client sessions. Way culmination fills the missing page references in a session. They investigate the primary issue of unwavering quality in the dataset. Web session grouping characterizes the information of web guests on the premise of client snap streams and comparable measures. It is valuable to deal with the web assets productively, for example, web personalization, outline adjustment and web server execution. A session could be acquired through swarm advancement and suitable closeness and after that connected to web log information. The bunching calculation orchestrates the information into comparative gatherings from a database. Distinctive sorts of information are kept in various gatherings and same sorts of information are kept in same kind of mixes. The current framework needs in adaptability of information.

**Neelamadhab Padhy1, Dr. Pragnyaban Mishra 2, and Rasmita Panigrahi3, "The Survey of Data Mining Applications and Feature Scope" (2012)**

The paper focused on an assortment of techniques, systems and differing scopes of the investigation which are valuable and set apart as the basic field of information mining Technologies. The paper immediately overviewed the distinctive information mining

applications which is found valuable to researchers to focus on the diverse issues of information mining. The particular systems for information mining are used to remove the examples and along these lines the learning from a blended sack databases. Decision of data and schedules for information mining is a basic task in this technique and necessities the learning of the zone. A couple attempts have been made to diagram and develop the bland information mining structure yet no system found absolutely non specific. Along these lines, an area master's associate is obligatory for each domain.

**M. P. Jarkad and P. Mansi (2015)**

Provides solution on classification, clustering and backtracking algorithm. The clustering algorithm groups the heterogeneous users and homogenous groups. In Backtracking algorithm can be used to reduce the prediction time. In the PUCC system different steps will be performed:

- Data cleaning or pre-processing which removes unwanted data from the log files.
- Users are categorized into potential users and non potential users. Only potential users are considered for processing step.
- For clustering the data, Graph Partitioning Algorithm is applied.
- Then user's request is predicted using LCS Algorithm.

For the prediction of user's browsing behaviour FCM (fuzzy clustering method) and is applied. The steps are as follows:

i. Read the web log files.
ii. Pre-process the web log files.
iii. Divide the clusters using Fuzzy clustering mean and Kernelized FCM

The Backtracking algorithm searches the data from the database that is no longer accessed by the user. If user wants to access that data, then it would be difficult for him/her and will be very time consuming. This algorithm divides the data sequentially at every stage and repeats the steps until get the final result.

# Chapter 3

# PRESENT WORK

## 3.1 Problem Formulation

Typically, standard substance mining and information recuperation rely on upon word organizing and the stem of the word to speak to record content. These procedures don't investigate the likeness of words and the structure of the archive inside the body of the content. Counting semantics, a content can be assembled on a higher semantic level than single words. The greater part of the literary data on the web is unstructured, and in addition, the world is attempting to sort out, dissect and hunt the perpetually aggregating mass of archives. Frameworks that naturally arrange content reports into predefined classes and along these lines contextualize data, offer a promising approach to manage taking care of this multifaceted nature. The prior information examination handle used to include much manual work henceforth, the clarification of information was moderate, exorbitant, and very inborn. The information mining apparatuses getting the information are utilized for taking conclusions in view of the information together. Grouping technique is regulated and allots objects into sets of predefined classes. There are differing sorts of arrangement methodologies being used in information mining, for example, rules trees and capacity. The essential goal of grouping is to precisely compute the estimation of every class variable. This order strategy includes two phases i.e. testing and preparing. The initial step i.e. to manufacture the model from the preparation set, the calmly tests are precisely browsed the information set. In the second step the information qualities are dispensed to the model and approve the model's precision

In the proposed approach PART calculation has been utilized for the execution of multi class grouping. PART is a Decision Tree calculation. It is utilized to prune the tree. Credulous Bayes is an arrangement calculation that which freely accept between expectations. It is simple and best helpful in huge datasets. SMO is additionally a sort of classifier that takes care of quadratic programming issues (QP).It effectively enhances the preparation set of SVM. Packing calculation produces many preparing sets of homogeneous size. It enhances the precision and dependability of the classifier. The machine learning

strategy is for enhancing the execution of learning calculation. Creating a very precise forecast manage by mix of different powerless classifiers and non-fitting guidelines. Packing and boosting calculation with great results in numerous zones, however in content characterization. Inquire about around there hasauncovered connections to different calculations, for example, Support Vector Machines.

### 3.2 Objectives

1. To apply preprocessing phase on the web related raw data and convert into formatted dataset (content, structure, usage etc.) using Replace Missing Value filter(RMVF)
2. To apply existing Hierarchical K-Means algorithm for clustering the formatted data.
3. To apply Improved Bagging algorithm on clustered data for classifying the clusters of Web data.
4. To compare and analyze the results of proposed technique with K means clustering on the basis of parameters viz. Accuracy Rate(AR), False Positive Rate(FPR), True Positive Rate(TPR), Precision, Clustering time.

# Chapter 4

# RESEARCH METHODOLOGY

## 4.1 Phases of proposed algorithm

### 4.1.1 Data Filters

The data filters are of four types. These are ReplaceMissingValues unsupervised Filter, NumericToNominal Filter and Discretization filter.

#### 4.1.1.1 ReplaceMissingValues unsupervised Filter

ReplaceMissingValues unsupervised Filter has been utilized to supplant every single missing worth utilizing means and modes. Missing information or missing qualities happen when no information esteem is put away for an occasion in the late record. Missing information may be happen in light of the fact that esteem is not significant to specific case, couldn't be recorded when information was gathered or overlooked by clients as a result of security concern. Mosteinformation framework as a rule makes them miss values because of inaccessibility of information. Some of the time information is not introduced or get adulterated because of irregularity of information documents. Missing information is nonattendance of information things that conceal some data that might be vital.

#### 4.1.1.2 Numeric To Nominal Filter

It is utilized for transforming numeric properties into ostensible ones. Not at all like discretization, it just takes every numeric esteem and adds them to the rundown of ostensible estimations of that characteristics.

#### 4.1.1.3 Discretization Filter

This filter used transform a continuous attribute into a categorical attribute. This involves dividing the range of possible values into sub-ranges called buckets.

### 4.1.2 Hierarchical clustering algorithm

The Hierarchical bunching additionally rang the base approach begins with every protest shaping a different gathering. It progressively joins the articles or gatherings that are near each other, until each gatherings are converged into one chain of command. The

divisive approach is additionally called the top down approach, it begins with the entire protest in the sameecluster. For every cycle, a bunch is part up into littler groups, untilaeventually every protest is in one group, or until an end all the condition that it holds. Various leveled bunching does not oblige us to pre indicate the no. of bunches and the greater part of the various leveled calculations are deterministic.

` Key points of hierarchal bunching are:

i. Comes at the cost of lower proficiency.

ii. It has a sensible structure, is anything but difficult to peruse and translate.Improved bagging technique

Classifying the clustered data by using **improved bagging technique** decreases the variance of the prediction using dataset using combinations with repetitions to produce multi-sets of same size of the dataset. For each multi set the Decision tree algorithm is applied to classify the instances and a model is created and a vote related to that model is generated. The average of all the predicted votes is considered to be the result of the classifier. This will classify the data.

### 4.1.3 Comparative Performance Analysis

Analyse the performance parameters like FP rate, TP rate, Recall, Precision of existing algorithm and new proposed algorithm then Compare the results of both.

## 4.2 Flow chat of proposed work

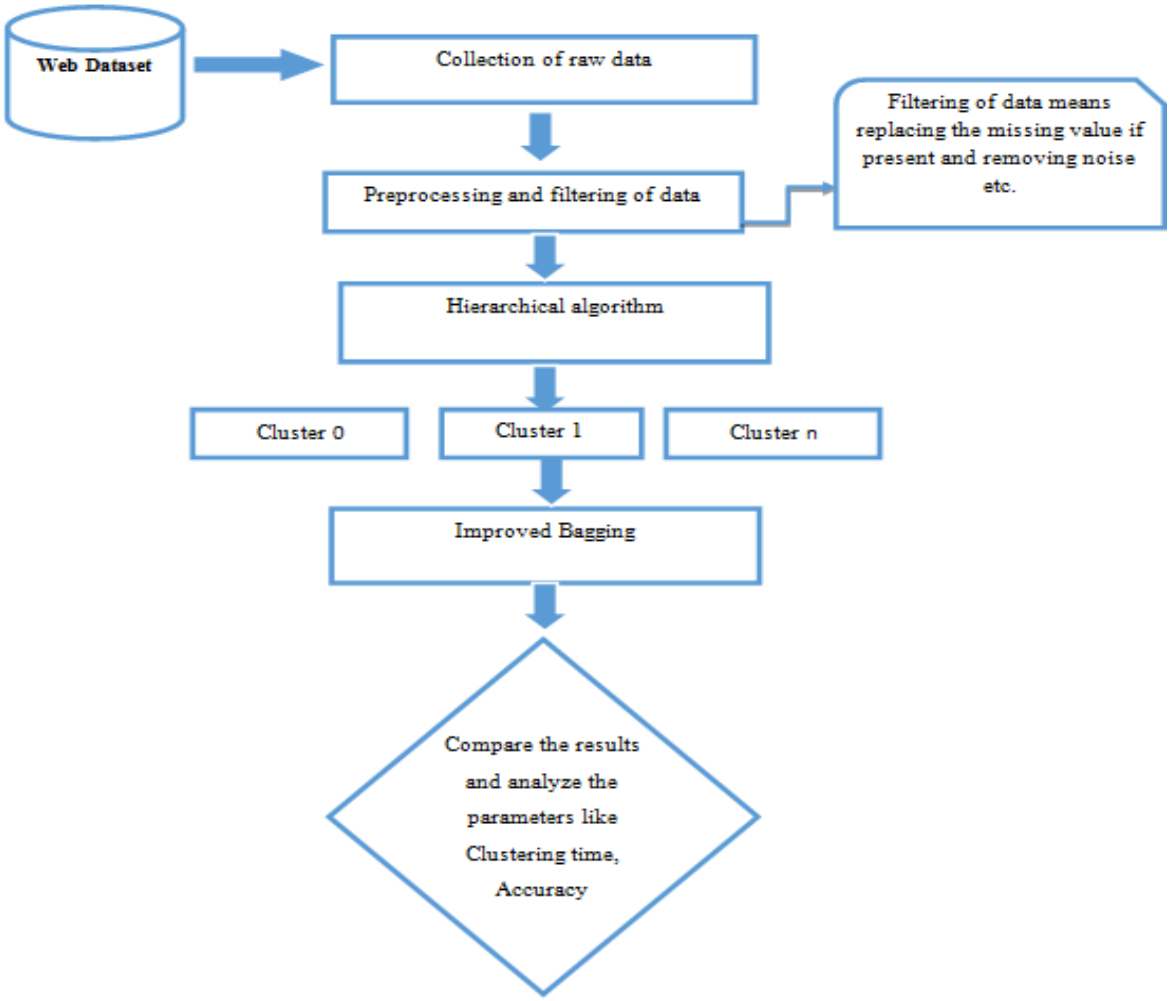



**Figure 4.1 : Flow chart of proposed algorithm**

## 4.3 Proposed algorithm

Data mining deals with different techniques that are applied on data and it provide various methods to cluster, classify, predict, analyze, etc on the data. Hierarchical clustering technique and improved bagging algorithm is applied in this approach.

1) Collection of raw data and then apply filtering techniques to make that raw data into structured format: Filtering techniques like Replace Missing Value filter

2) Applying and comparing the result with different algorithms on this proposed algorithm.

The goal of filtering is to quickly filter out the likely irrelevant data and to reduce large unmatched data. The retrieved cleaned data can then be undergone into more sophisticated data processing. As a result, Web mining system could be more efficient to deliver the users with more relevant results. The classical information filtering used the term-based user profile, based on which, it gets highly difficult to define the threshold of filtering. High percentage of useful information is filter out, thus greatly compromise the system effectiveness. Therefore, one of the key issues in developing an effective filtering system is to construct accurately and comprehensive user profiles that can describes the user needs and information searching intentions.

In real experience, a data set may contain noisy or redundant data items and large number of features, many of them may not be relevant for the objective function at hand. In this manner the noisy data may degrade the accuracy and performance of the classification models. Thus, dealing with missing values in data pre-processing is an important step in building an effective and efficient classifiers. It is a process by which the missing values are replaced by suitable values according an objective function or the noisy data may be filtered. It leads to better performance of the classification models contains predictive or descriptive accuracy, diminishing of computing time needed to build models as they learn faster, and better understanding of the models. Data cleaning particularly improves the predictive capability of the classifiers. This algorithm consist of is a hierarchy of nested clusters, every cluster consist of union of two smaller sub-clusters. It has two further clustering methods such as Bottom-up and top-down. Bottom-up starts with a single-object cluster and combine them into larger clusters. Top-down clustering starts with a cluster and contain all data and then divide it into smaller clusters. Bagging algorithm is machine learning technique that ensambles meta learning algorithm. It improves the availability and accuracy of machine learning algorithms. It avoid over fitting and reduce variances.

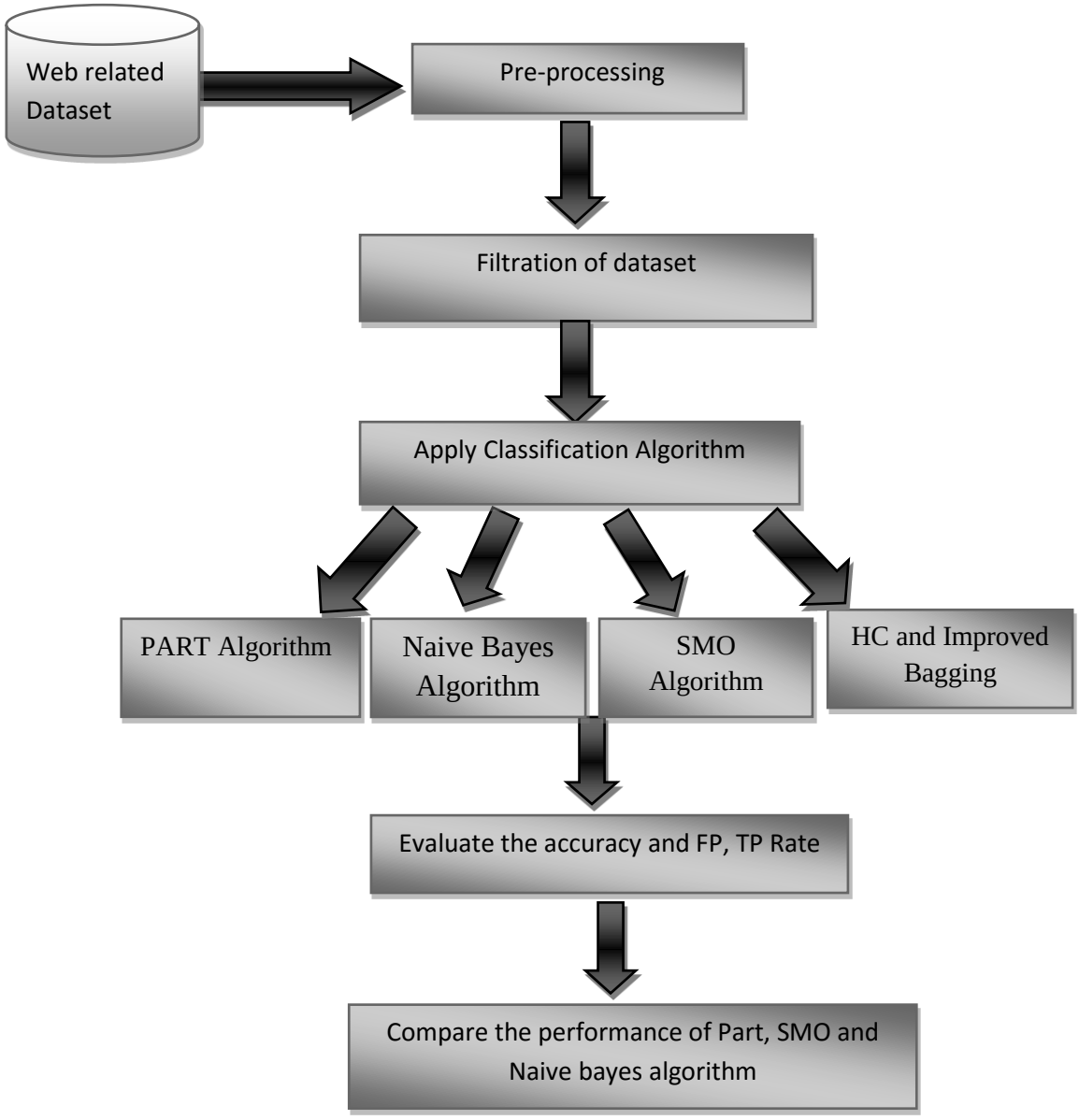



**Figure 4.2 : Flow of Proposed algorithm**

# Chapter 5

# RESULTS AND DISCUSSION

## 5.1 Tool Used For Implementation

### 5.1.1 WEKA

WEKA (Waikato Environment for Knowledge Analysis) is used to evaluate the performance. It is a collection of machine learning algorithms for data mining. WEKA is created by researchers by the University of Waikato in New Zealand. It is written in Java and runs on any platform. It is used for ML algorithms pre-processing, classifiers, clustering, association rule and visualization.


**Figure 5.1:** WEKA Platform

### 5.1.2 NET BEANS

This section presents the simulation results of the work done and the proposed approach. The simulation has been done in Java Net Beans. Net beans is an open-source project dedicated to providing rock solid software development products (the Net beans IDE and the Net beans Platform) that address the needs of developers, users and the businesses who rely on Net beans as a basis for their products; particularly, to enable them to develop these products quickly, efficiently and easily by leveraging the strengths of the Java platform and other relevant industry standards.

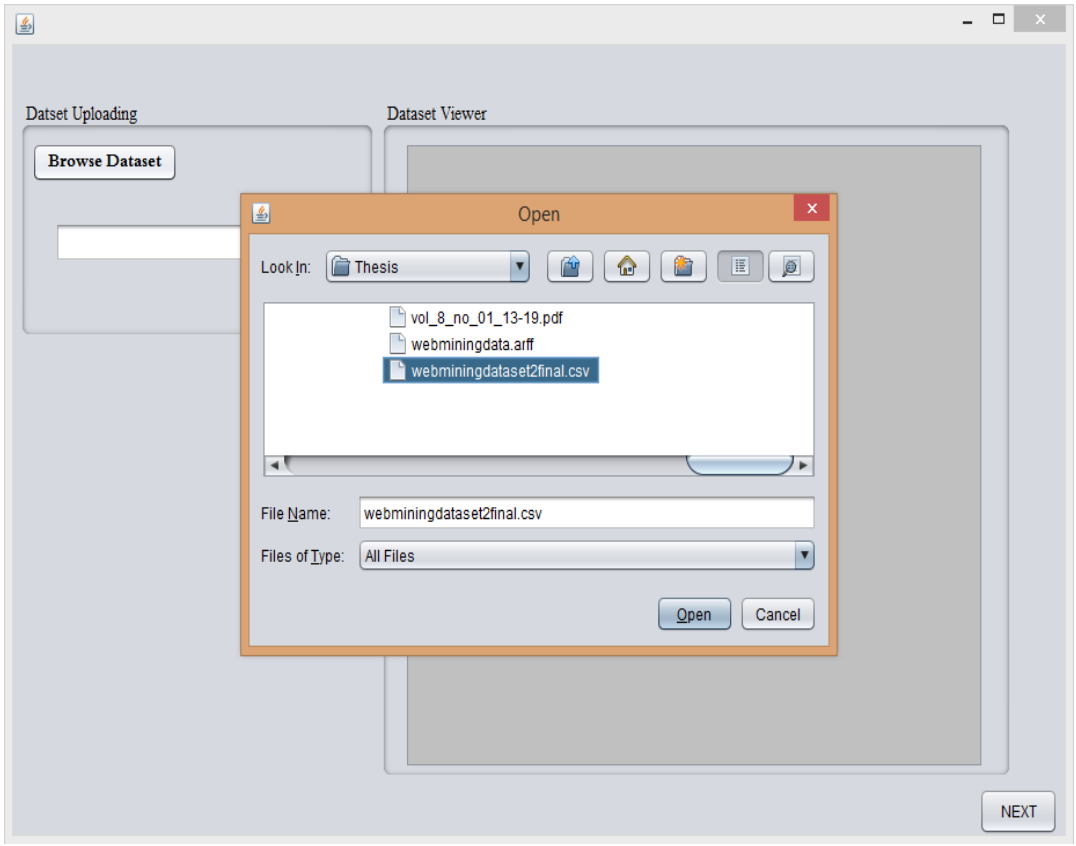



**Figure 5.2:** Net Beans Platform

Net bean is platform that enables the users to conduct the simulation with platform like WEKA.

You can select the database by the link given by browse database.

## 5.2 Experimental Results

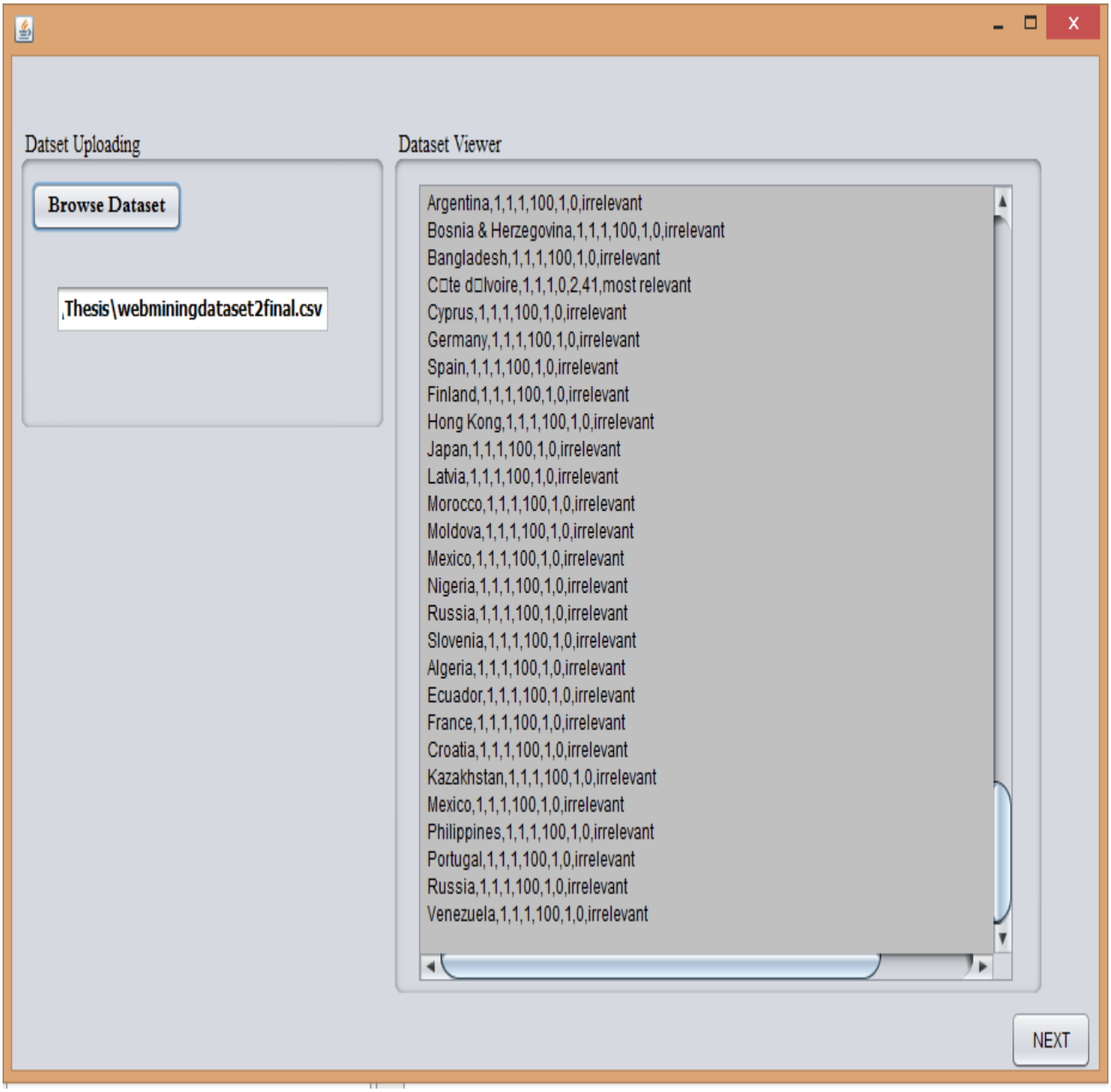



**Figure 5.3:** Uploaded dataset to database

After uploading the data to the database it user will be able to see the dataset view. It shows the different forms of data are being used in the dataset i.e. relevant, irrelevant, most relevant.

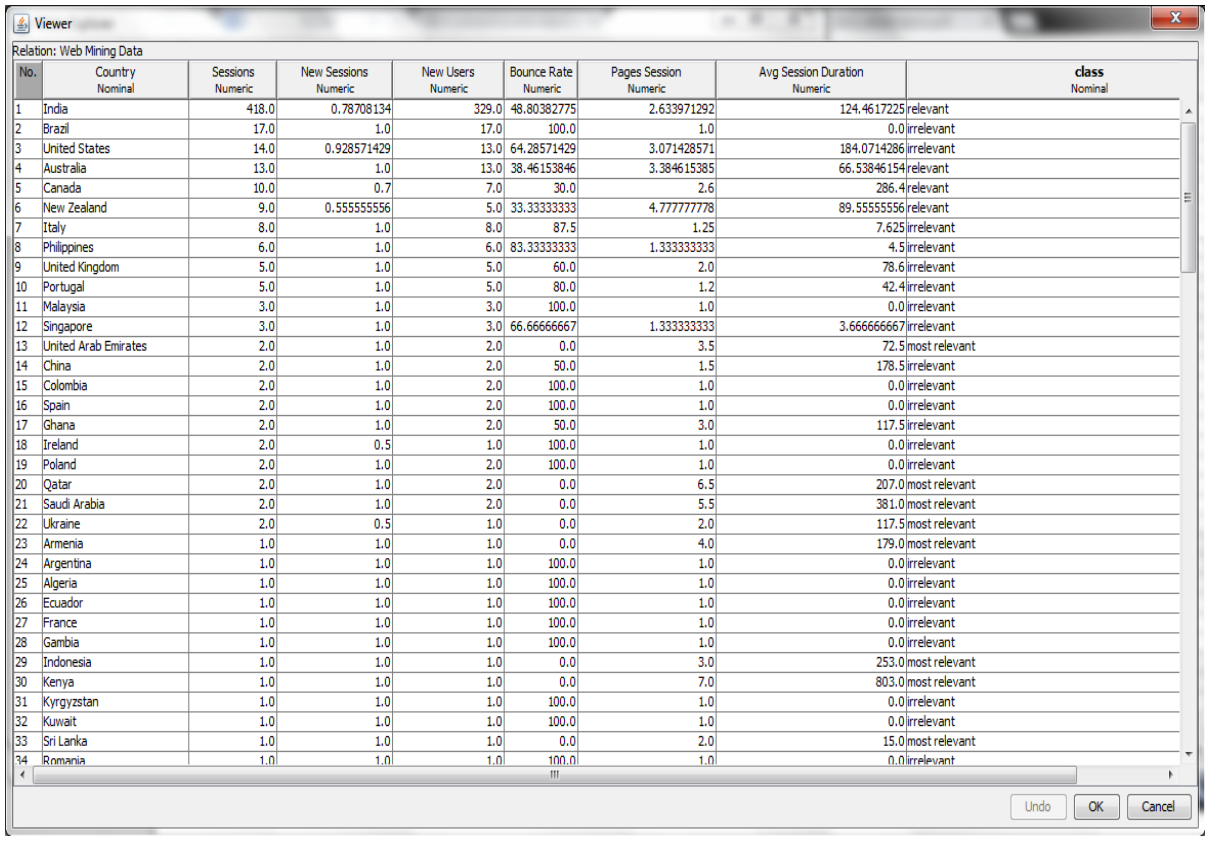
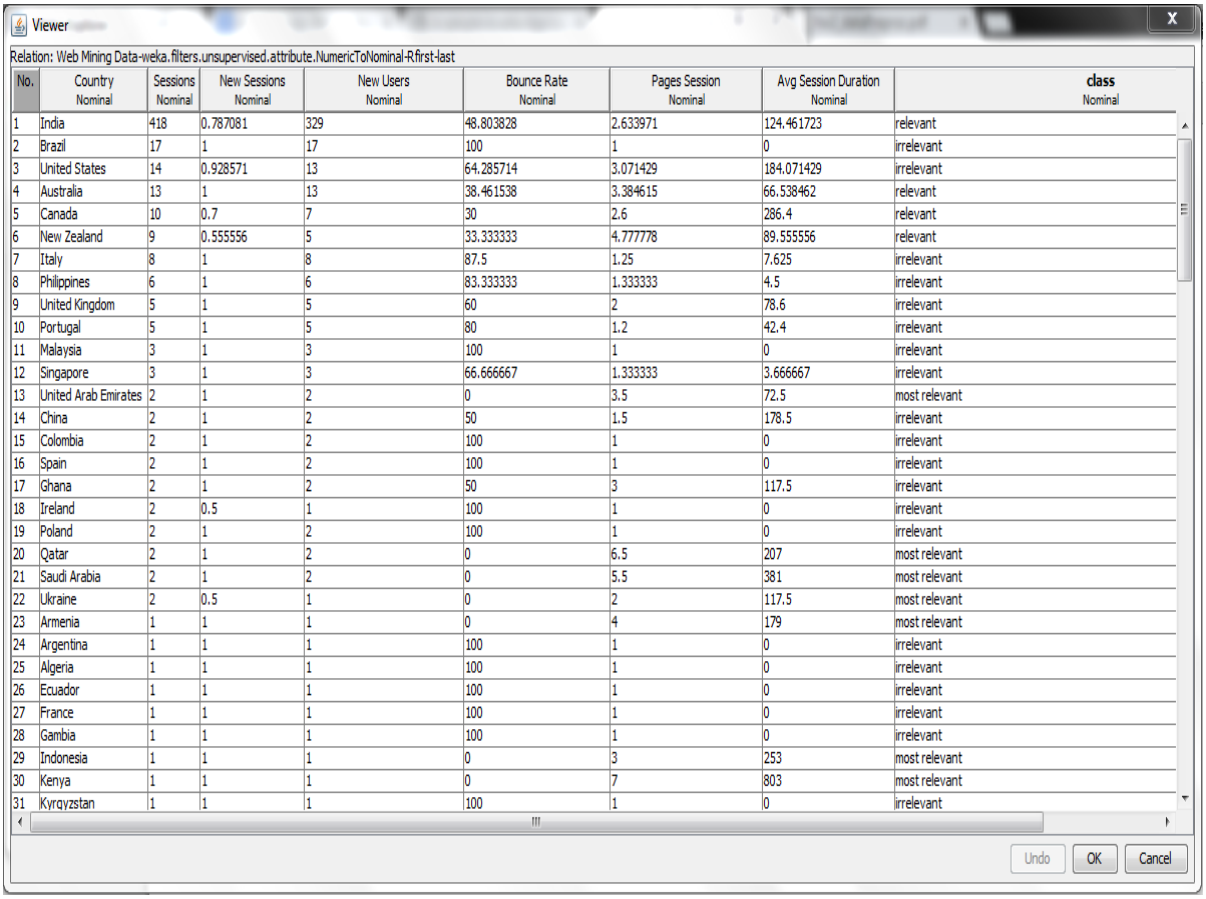



**Figure 5.4:** Dataset before applying Numeric to Nominal Filter

Here in this the attributes like sessions, bounce rate, page sessions, average session duration are all of numeric type; means they can be any combination of numbers between 0-9. Therefore after applying numeric to nominal filter the dataset will be converted to nominal attribute. In nominal attribute format, each attribute is defined a category of some values; the instances of that attributes should be form that category or from any other like in numeric.




| No. | Country Nominal | Sessions Nominal | New Sessions Nominal | New Users Nominal | Bounce Rate Nominal | Pages Session Nominal | Avg Session Duration Nominal | class Nominal |
|---|---|---|---|---|---|---|---|---|
| 1 | India | 418 | 0.787081 | 329 | 48.803828 | 2.633971 | 124.461723 | relevant |
| 2 | Brazil | 17 | 1 | 17 | 100 | 1 | 0 | irrelevant |
| 3 | United States | 14 | 0.928571 | 13 | 64.285714 | 3.071429 | 184.071429 | irrelevant |
| 4 | Australia | 13 | 1 | 13 | 38.461538 | 3.384615 | 66.538462 | relevant |
| 5 | Canada | 10 | 0.7 | 7 | 30 | 2.6 | 286.4 | relevant |
| 6 | New Zealand | 9 | 0.555556 | 5 | 33.333333 | 4.777778 | 89.555556 | relevant |
| 7 | Italy | 8 | 1 | 8 | 87.5 | 1.25 | 7.625 | irrelevant |
| 8 | Philippines | 6 | 1 | 6 | 83.333333 | 1.333333 | 4.5 | irrelevant |
| 9 | United Kingdom | 5 | 1 | 5 | 60 | 2 | 78.6 | irrelevant |
| 10 | Portugal | 5 | 1 | 5 | 80 | 1.2 | 42.4 | irrelevant |
| 11 | Malaysia | 3 | 1 | 3 | 100 | 1 | 0 | irrelevant |
| 12 | Singapore | 3 | 1 | 3 | 66.666667 | 1.333333 | 3.666667 | irrelevant |
| 13 | United Arab Emirates | 2 | 1 | 2 | 0 | 3.5 | 72.5 | most relevant |
| 14 | China | 2 | 1 | 2 | 50 | 1.5 | 178.5 | irrelevant |
| 15 | Colombia | 2 | 1 | 2 | 100 | 1 | 0 | irrelevant |
| 16 | Spain | 2 | 1 | 2 | 100 | 1 | 0 | irrelevant |
| 17 | Ghana | 2 | 1 | 2 | 50 | 3 | 117.5 | irrelevant |
| 18 | Ireland | 2 | 0.5 | 1 | 100 | 1 | 0 | irrelevant |
| 19 | Poland | 2 | 1 | 2 | 100 | 1 | 0 | irrelevant |
| 20 | Qatar | 2 | 1 | 2 | 0 | 6.5 | 207 | most relevant |
| 21 | Saudi Arabia | 2 | 1 | 2 | 0 | 5.5 | 381 | most relevant |
| 22 | Ukraine | 2 | 0.5 | 1 | 0 | 2 | 117.5 | most relevant |
| 23 | Armenia | 1 | 1 | 1 | 0 | 4 | 179 | most relevant |
| 24 | Argentina | 1 | 1 | 1 | 100 | 1 | 0 | irrelevant |
| 25 | Algeria | 1 | 1 | 1 | 100 | 1 | 0 | irrelevant |
| 26 | Ecuador | 1 | 1 | 1 | 100 | 1 | 0 | irrelevant |
| 27 | France | 1 | 1 | 1 | 100 | 1 | 0 | irrelevant |
| 28 | Gambia | 1 | 1 | 1 | 100 | 1 | 0 | irrelevant |
| 29 | Indonesia | 1 | 1 | 1 | 0 | 3 | 253 | most relevant |
| 30 | Kenya | 1 | 1 | 1 | 0 | 7 | 803 | most relevant |
| 31 | Kyrgyzstan | 1 | 1 | 1 | 100 | 1 | 0 | irrelevant |

**Figure 5.5:** Dataset after applying Numeric to Nominal Filter

After applying numeric data to nominal filter you will notice that the data type of the dataset will be changed, but there will be no change in the values.

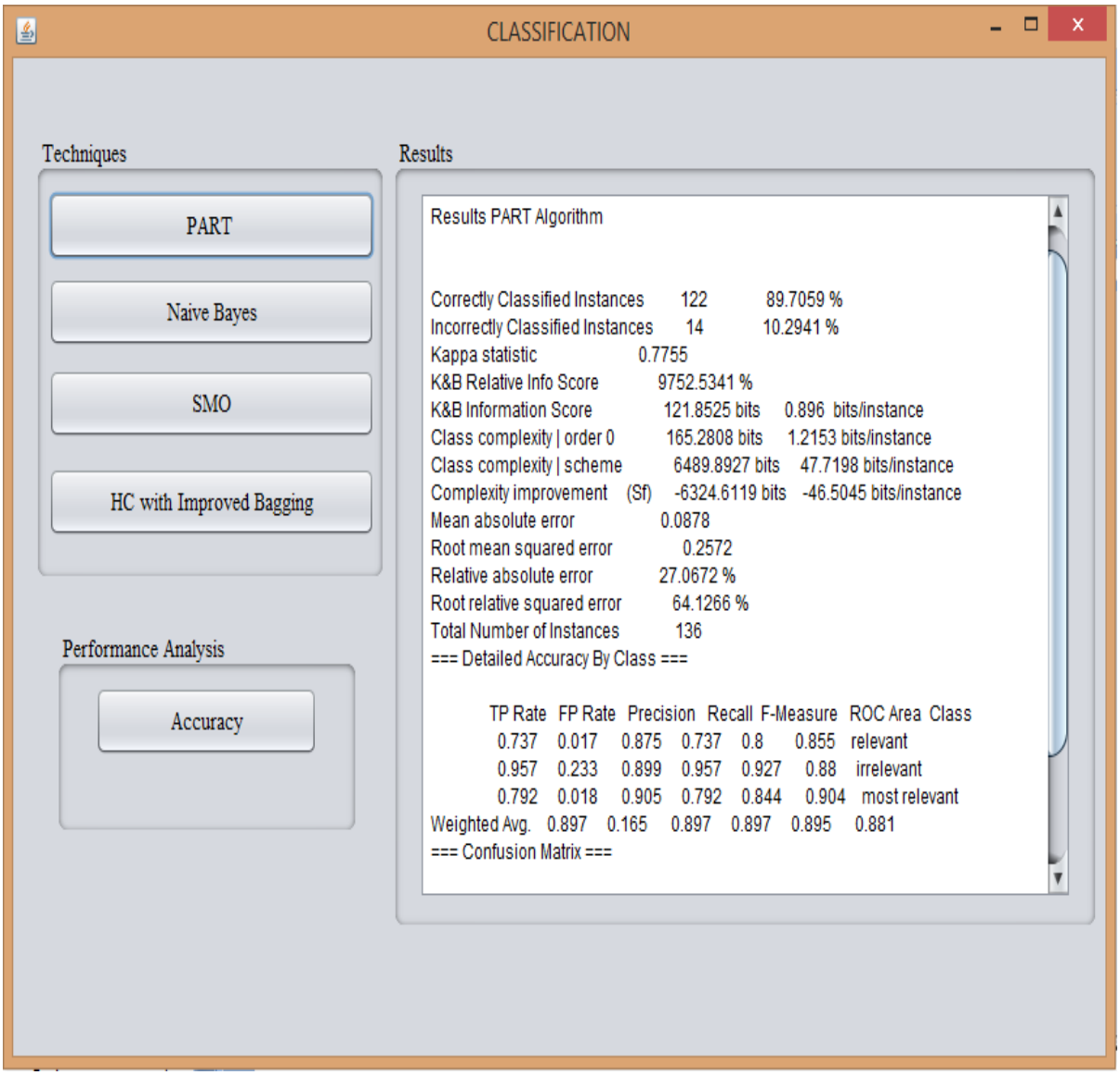



**Figure 5.6:** Performing classification using PART Results

The PART algorithm contains the latest version of C4.5 algorithm. It is also known as patition decision tree algorithm. In the implementation of this algorithm, it shows the correct instances of about 89.70%.

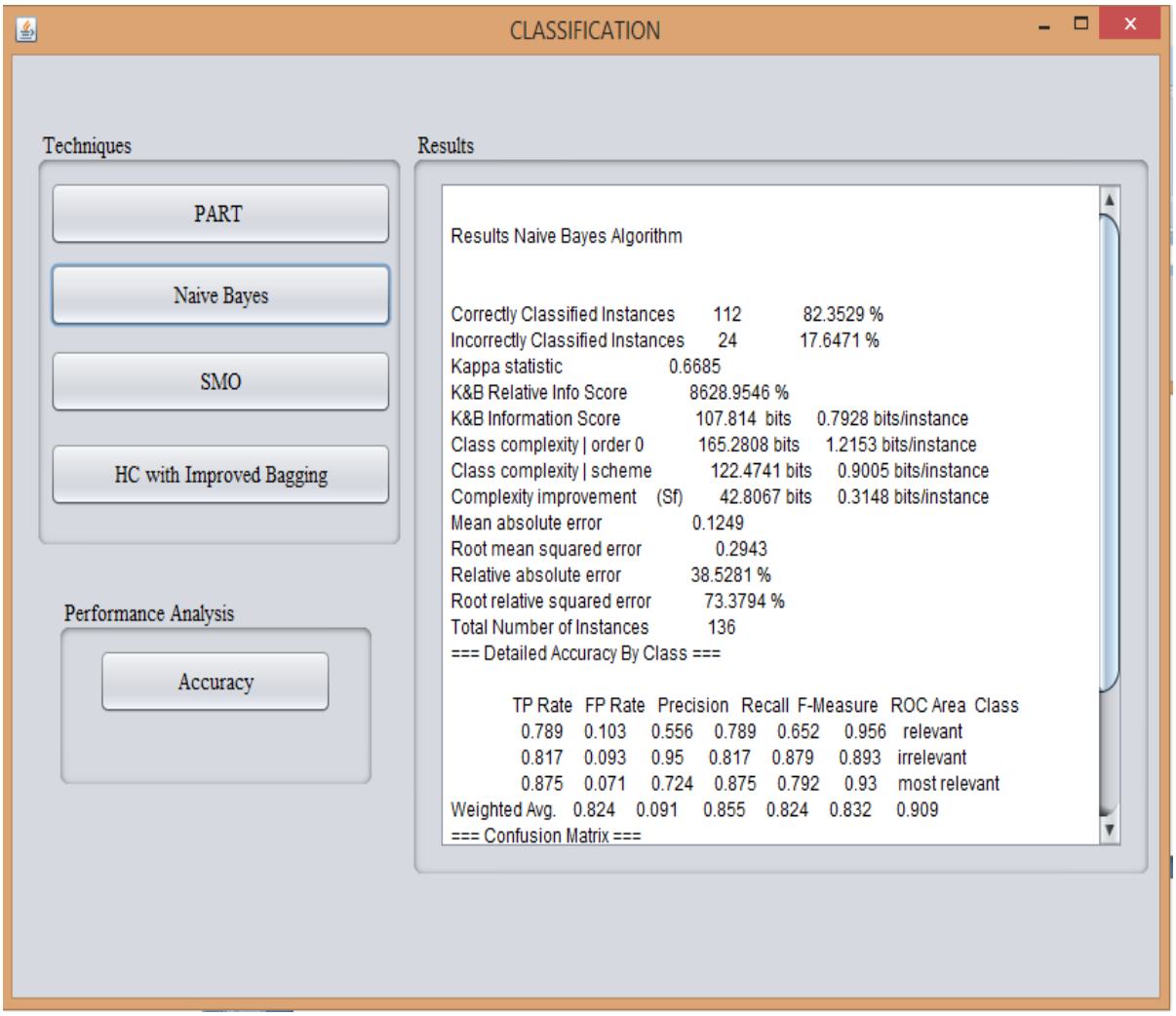



**Figure 5.7:** Performing classification using Naive Bayes Results

It contains quadratic problem in the algorithm**.** it is a classification algorithm that defines the result of about 82.35%.

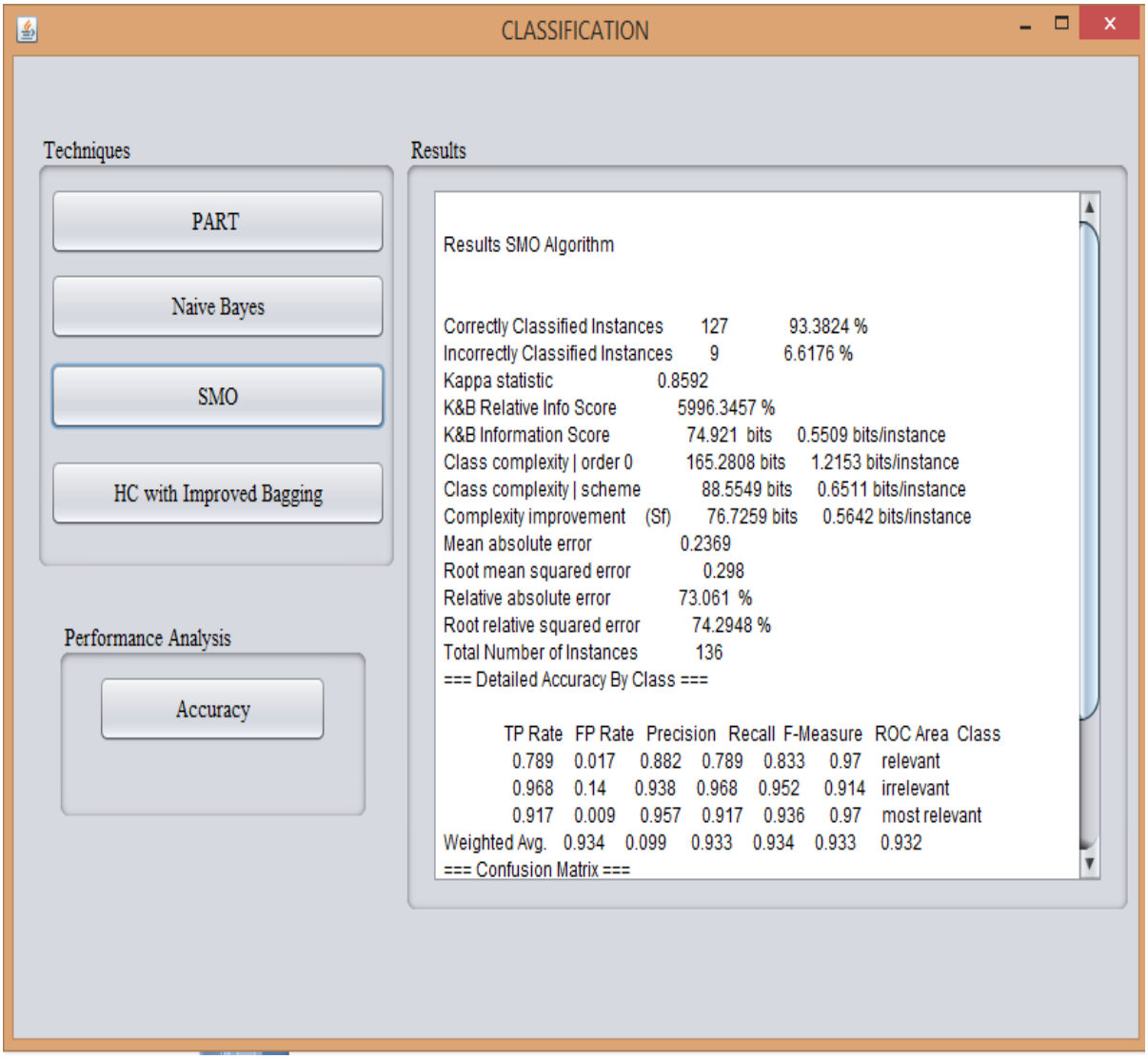



**Figure 5.8:** Performing classification using SMO Results

SMO algorithm is a classification algorithm. This algorithm is used for the decision tree contruction. The accuracy of the dataset is about 93.38%.

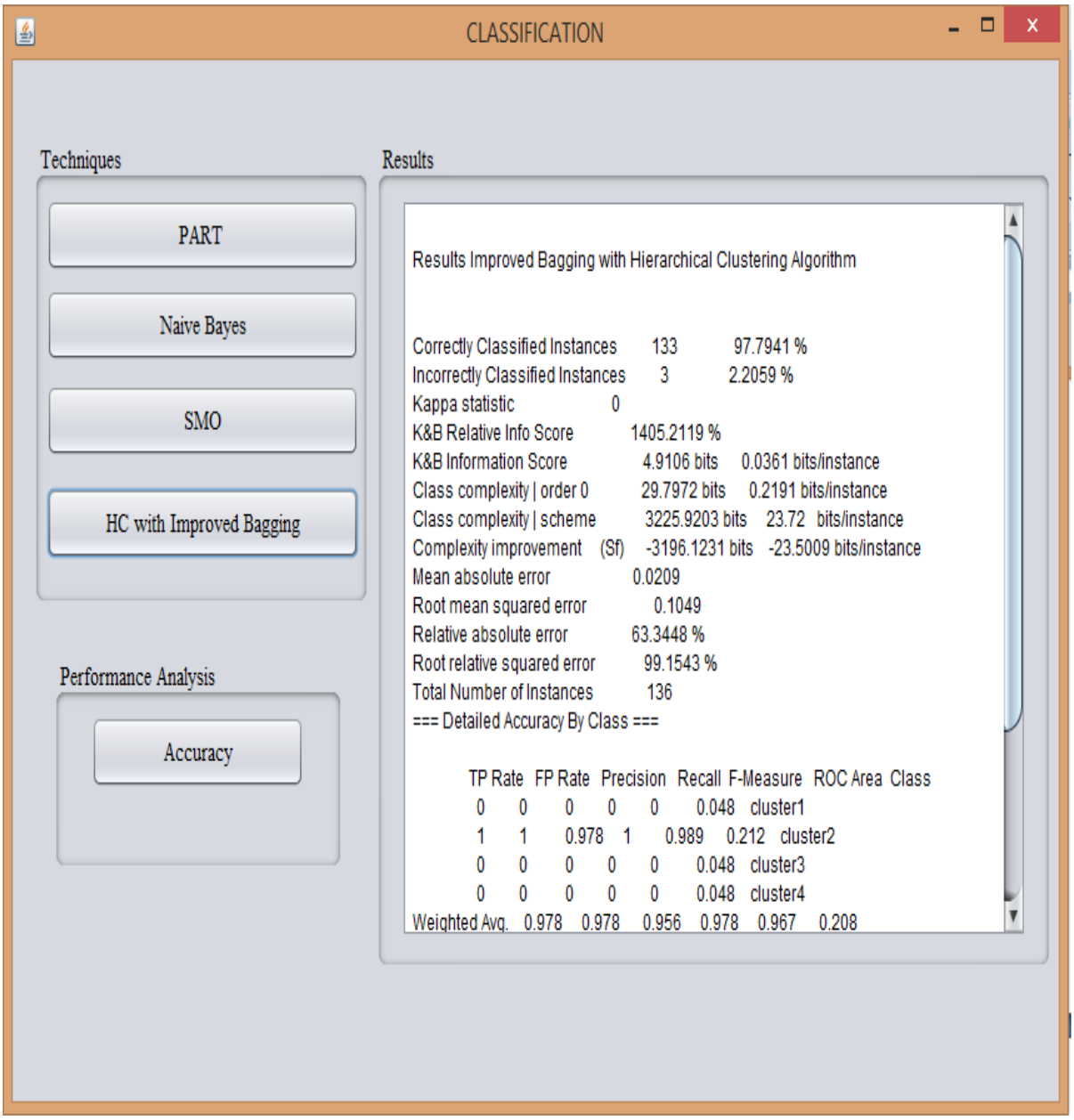



**Figure 5.9:** Performing classification using Improved Bagging with Hierarchical Clustering Results

The hierarchal clustering and improved bagging algorithm is the advanced technique that is used to analyse and predict the data. It shows the result of data is about 97.80%.

## 5.3 Accuracy comparison of different algorithms

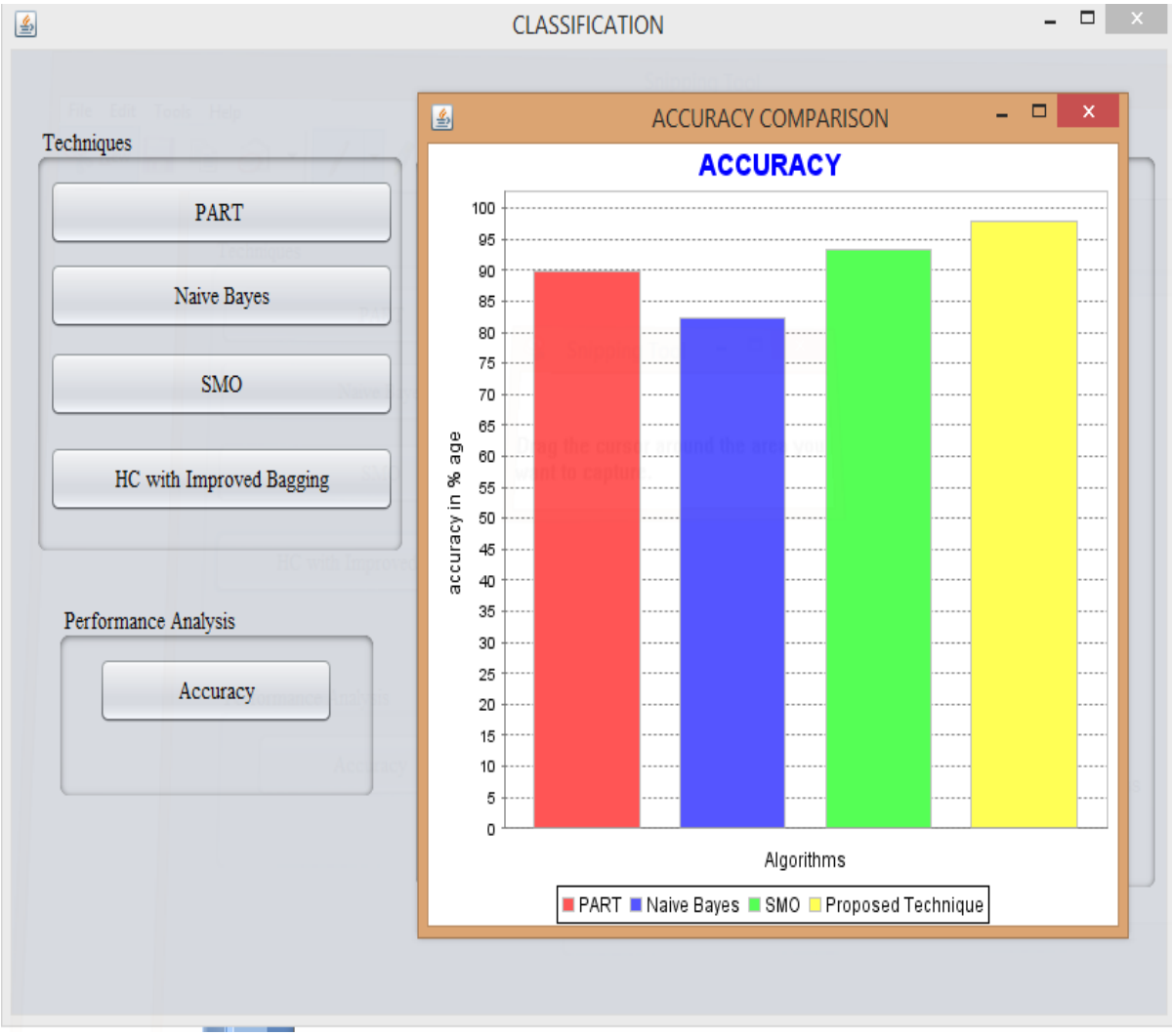



**Figure 5.10:** Accuracy comparison of different algorithms

The figures show the results of different mining algorithms like PART, Naïve Bayes, SMO and hierarchical clustering with improved bagging algorithm. PART is showing the accuracy 89.706 %, Naïve Bayes is having accuracy 82.353%, SMO is having accuracy 93.382 % and Hierarchical clustering with improved bagging algorithm is having 97.794%. Therefore it is clearly shown from the figures that the HC with Improves Bagging is better among the other three mining algorithms.

## 5.4 Database Visualization

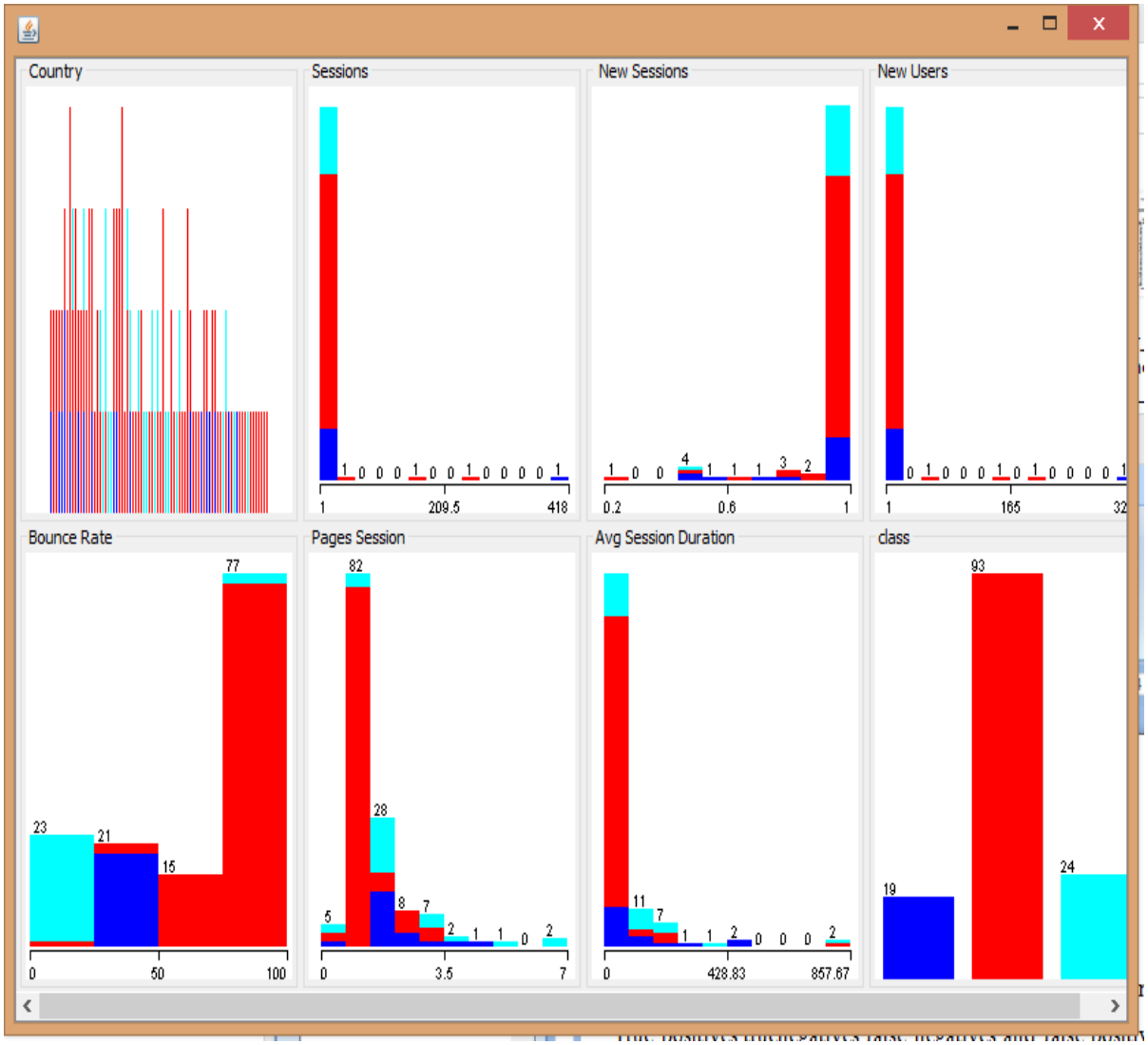



**Figure 5.11:** Database Visualization in Weka tool

The view of the dataset in the netbeans platform is different that in weka tool. The view of dataset about the bounce rate, page session, average session duration, class is given in the above figure.

## 5.5 Evaluation parameters

The parameters for the evaluation of sentiment analysis include various terms. The terms are True positives,truenegatives,false negatives and false positives.These are the terms that are used to compare the class labels assigned to documents with the classes the items actually belong to by a classifier.True positive terms are truly classified as positive terms.False

positive are not labelled by the classifier as positive class but should have been.True negative terms are correctly labelled as in negative class by the classifier.False negative terms are those terms that are not labelled by the classifier as belonging to negative class but should have be classified.Confusion Matrix contains these terms that are used for evaluation.

**Table 5.1:** Contingency Table

| | | Correct Labels | |
|---|---|---|---|
| | | Positive | Negative |
| Classified Labels | Positive | True positive | False positive |
| | Negative | False negative | True negative |

Following are the parameters for evaluation of performance

### i. Precision and recall

Precision and recall are the two metrics that are widely for evaluating performance in text mining, and in text analysis field like information retrieval. These parameters are used for measuring exactness and completeness respectively.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive+False Positive}} \qquad \text{Eq. (1)}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive+False Negative}} \qquad \text{Eq.(2)}$$

### ii. F-measure

F-Measure is the harmonic mean of precision and recall. The value calculated using F-measure is a balance between precision and recall.

$$\text{F measure} = \frac{2*\text{recall}*\text{precision}}{\text{precision+recall}} \qquad \text{Eq. (3)}$$

### iii. Accuracy

Accuracy is the common measure for classification performance. Accuracy can be measured as correctly classified instances to the total number of instances, while error rate uses incorrectly classified instances instead of correctly classified instances.

$$\text{Accuracy} = \frac{\text{True Positive+True Negative}}{\text{True Positive+False Positive+True Negative+False Negative}} \qquad \text{Eq. (4)}$$

**Table 5.2:** Performance of mining algorithm

| | PART | Naive Bayes | SMO | HC with Improved Bagging |
|---|---|---|---|---|
| Correctly classified | 122 | 112 | 127 | 133 |
| Incorrectly classified | 14 | 24 | 9 | 3 |

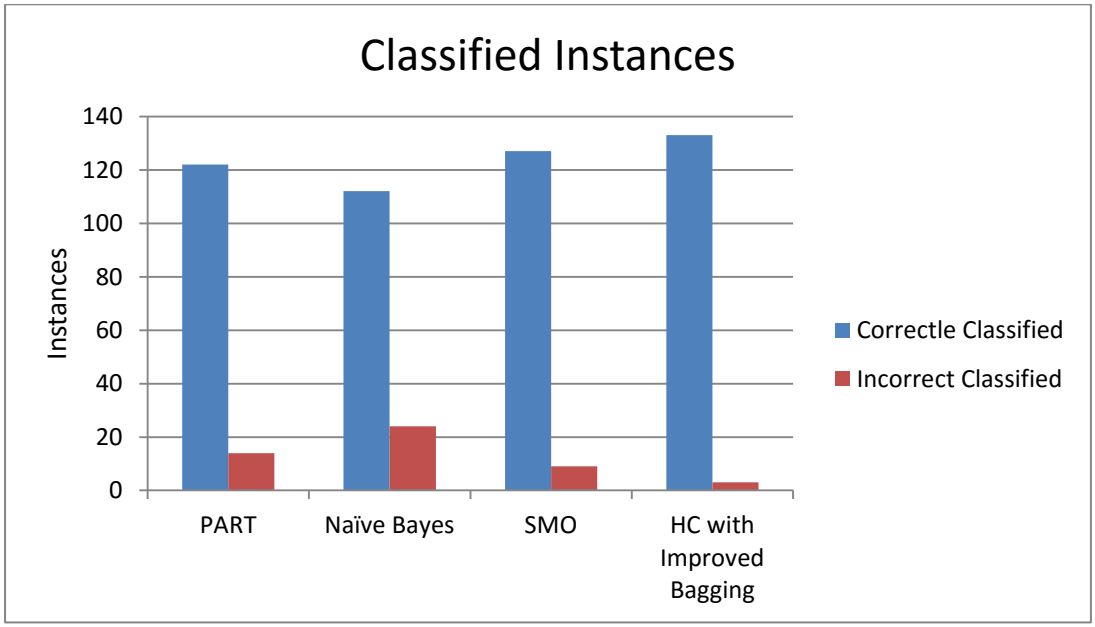



**Figure 5.12:** classification instances

This figure shows correctly and incorrectly classified instances for each algorithm. In PART, 122 instances are correctly classified and 25 are incorrectly classified, In, Naive

Bayes 112 instances are correctly and 24 are incorrectly classified. Similarly for SMO and HC with Improved Bagging 127 are correct and 9 are incorrectly classified and 133 instances are correctly classified and 3 are incorrectly classified instances.

**Table 5.3:** Detailed comparison of accuracy by the class attribute

| | PART | Naïve Bayes | SMO | HC with improved Bagging |
|---|---|---|---|---|
| Kappa Statistics | 0.7755 | 0.6685 | 0.8592 | 0 |
| Mean absolute error | 0.0878 | 0.1249 | 0.2369 | 0.0209 |
| Root mean square error | 0.2572 | 0.2943 | 0.298 | 0.1049 |
| Root absolute error | 27.0672 | 38.5281 | 73.061 | 63.3448 |
| Root relative squared error | 64.1266 | 73.3794 | 74.2948 | 99.1543 |

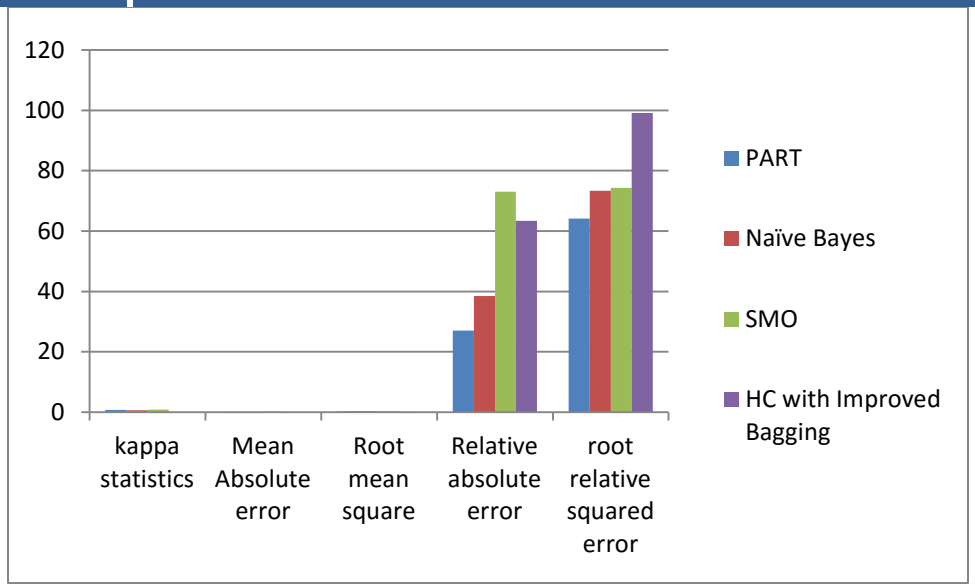



**Figure 5.13:** Accuracy Evaluation of Parameter

For Mining Algorithms detailed accuracy were calculated which includes parameter as TP rate, FP rate, Precision, Recall, F-measure and ROC area.

**Table 5.4:** Detailed Accuracy of PART Algorithm

| Class | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area |
|---|---|---|---|---|---|---|
| Relevant | 0.737 | 0.017 | 0.875 | 0.737 | 0.8 | 0.855 |
| Irrelevant | 0.957 | 0.233 | 0.899 | 0.957 | 0.927 | 0.88 |
| Most Relevant | 0.792 | 0.018 | 0.737 | 0.792 | 0.844 | 0.904 |

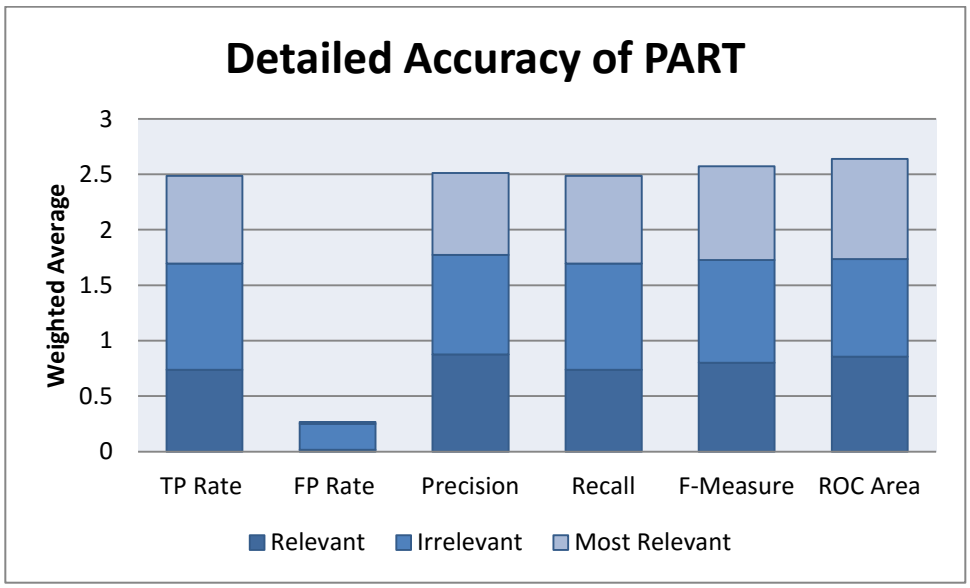



**Figure 5.14:** Detailed Accuracy of PART Algorithm

This figure contains detailed class accuracy of PART algorithm including TP rate, FP rate, Precision, Recall, F-Measure and ROC Area

**Table 5.5:** Detailed Accuracy of Naive Bayes Algorithm

| Class | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area |
|---|---|---|---|---|---|---|
| Relevant | 0.789 | 0.556 | 0.103 | 0.789 | 0.652 | 0.956 |
| Irrelevant | 0.817 | 0.093 | 0.95 | 0.817 | 0.879 | 0.893 |
| Most Relevant | 0.875 | 0.071 | 0.724 | 0.875 | 0.792 | 0.93 |

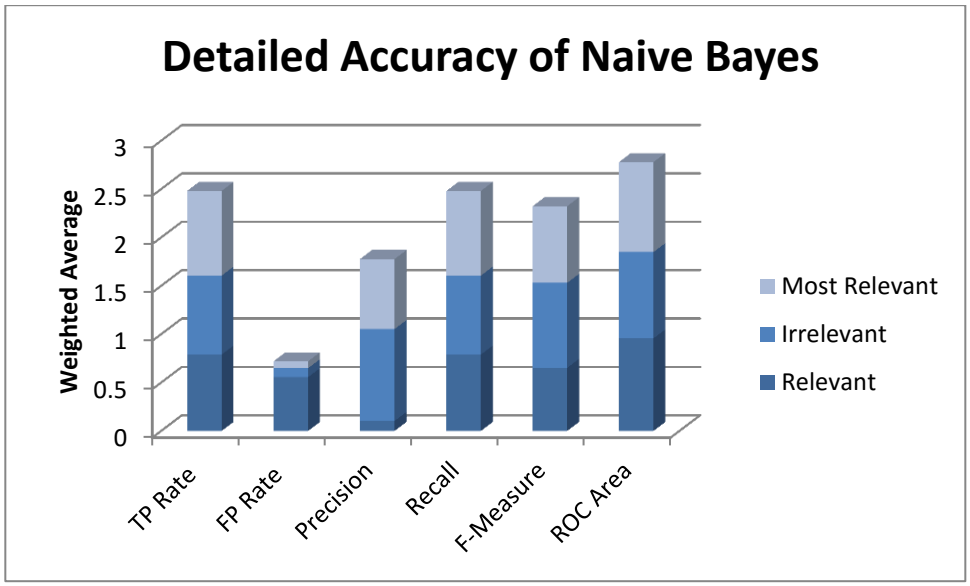



**Figure 5.15:** Detailed Accuracy of Naive Bayes Algorithm

This figure contains detailed class accuracy of Naive Bayes algorithm including TP rate, FP rate, Precision, Recall, F-Measure and ROC Area.

**Table 5.6:** Detailed Accuracy of SMO Algorithm

| Class | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area |
|---|---|---|---|---|---|---|
| Relevant | 0.789 | 0.017 | 0.882 | 0.789 | 0.833 | 0.97 |
| Irrelevant | 0.968 | 0.14 | 0.938 | 0.968 | 0.952 | 0.914 |
| Most Relevant | 0.917 | 0.009 | 0.957 | 0.917 | 0.936 | 0.97 |

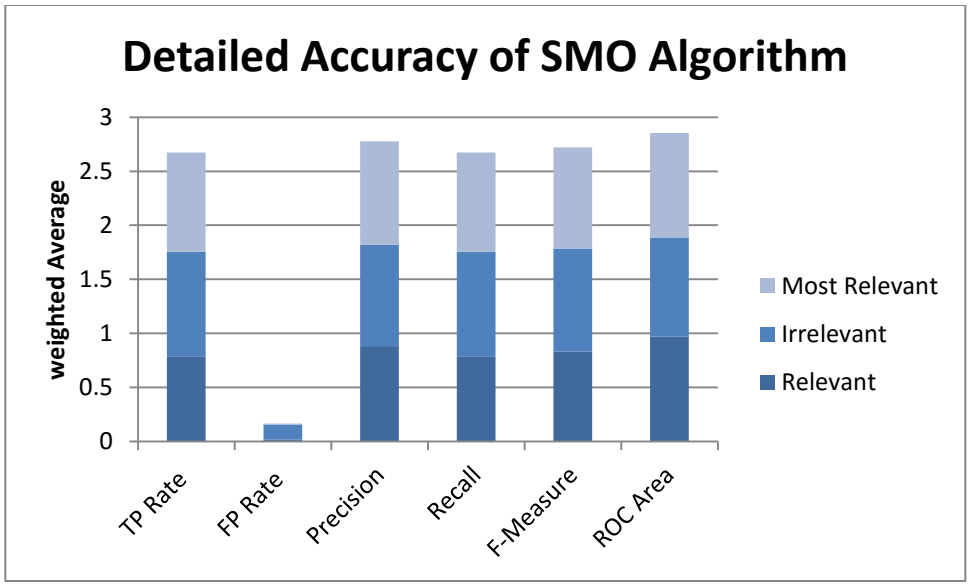



**Figure 5.16:** Detailed Accuracy of SMO Algorithm

This figure contains detailed class accuracy of Naive Bayes algorithm including TP rate, FP rate, Precision, Recall, F-Measure and ROC Area

**Table 5.7:** Detailed Accuracy for Decision Tree

| Class | TP rate | FP rate | Precision | Recall | F-measure | ROC Area |
|---|---|---|---|---|---|---|
| Relevant | 0 | 0 | 0 | 0 | 0 | 0.048 |
| Irrelevant | 1 | 1 | 0.978 | 1 | 0 | 0.212 |
| Most Relevant | 0 | 0 | 0 | 0 | 0 | 0.048 |
| Weighted Average | 0 | 0 | 0 | 0 | 0 | 0.048 |

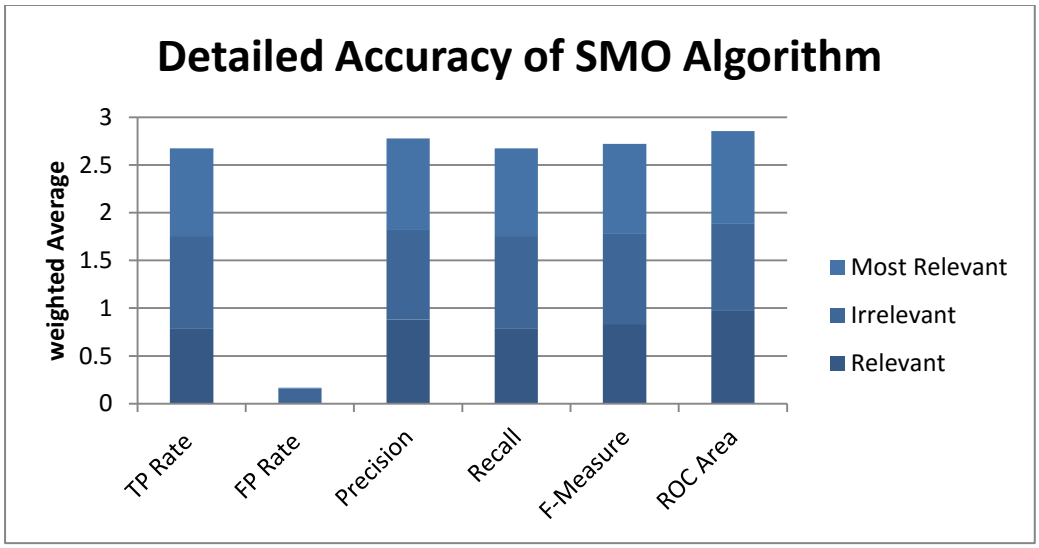



**Figure 5.17:** Detailed class Accuracy of HC with Improved Bagging

This figure contains detailed class accuracy of Decision Tree algorithm including TP rate, FP rate, Precision, Recall, F-Measure and ROC Area.

# Chapter 5

# CONCLUSION

Web mining is the application of data mining techniques to extract knowledge from Web. Web mining has been explored to a vast degree and different techniques have been proposed for a variety of applications that includes Web Search, Classification and Personalization etc. In this work, three different modes of web mining, namely web content mining, web structure mining and web usage mining have been explained. Needless to say, these three approaches cannot be independent, and any efficient mining of the web would require a judicious combination of information from all the three sources. This work also presents the summary of various techniques of web mining and Data Mining Techniques in various application domains. The survey on data mining technique is made with respect to Clustering, Classification, Sequence Pattern Mining, Association Rule Mining and Visualization. Research in text mining is at the moment very general in nature hence to deal with data an algorithm named PART has been used for text classification. The main motive is to analyze the relevant information of different websites and find patterns and make predictions. In addition to this, in this work, an enhanced version of has also been proposed and implemented. The simulation results using PART, Naive Bayes, SMO and Hierarchical clustering with Improved Bagging Algorithm the results have been presented. It has been observed that Accuracy for Enhanced Hierarchical clustering with Improved Bagging is around 97.79% and for PART is around89.70%, Naive Bayes is around 82.35%, SMO is around 93.38% which clearly shows that the proposed approach is better in performance.

## 5.1 Future Work

This work can be further extended by including a larger Dataset and considering more instances which may help in more accurate prediction analysis.

# REFERENCES


[1] K. Sudheer Reddy, G. ParthaSaradhi Varma, and M. Kantha Reddy, "An Effective Preprocessing Method for Web UsageMining", International Journal of Computer Theory and Engineering, Vol. 6, No. 5, October 2014

[2] Prabhjot Kaur, "Web Content Classification: A Survey", International Journal of Computer Trends and Technology (IJCTT) – volume 10 number 2 – Apr 2014

[3] M. Sujatha, "A Survey of Classification Techniques in Data Mining," vol. 2, no. 4, pp. 86–92.

[4] Vivek Agarwl, Saket Thakre, Akshay Jaiswal "Survey on Classification Techniques of Data Mining" 2015

[5] R. Fernandes, L. J. Peo, N. Kamat, and S. Miranda, "New Approaches to Web Personalization Using Web Mining Techniques .," vol. 5, no. 2, pp. 2195–2201, 2014.

[6] S. Vijiyarani and M. E. Suganya, "Research issues in web mining," vol. 2, no. 3, pp. 55– 64, 201

[7] K. B. Patel, A. R. Patel, and N. S. Patel, "Web Advertising Personalization using Web Content Mining and Web Usage Mining Combination," vol. 3, no. 1, pp. 8–12, 2014.

[8] J. Srivastava, R. Cooley, M. Deshpande, and P. Tan, "Web Usage Mining : Discovery and Applications of Usage Patterns from Web Data," vol. 1, no. 2, pp. 12–23, 2000.

[9] V. Dongre, "An Improved User Browsing Behavior Prediction using Regression Analysis on Web Logs," vol. 120, no. 19, pp. 19–23, 2015

[10] Jin Xu Yingping Huang Gregory Madey, "A Research Support System Framework for Web Data Mining"

[11] Peng PengQianli Ma Chaoxiong Li, "The Research and Implementation of Data Mining Component Library System"

[12] R. Shukla, "Web Personalization Systems and Web Usage Mining : A Review," vol. 72, no. 21, pp. 6–13, 2013.

[13] J. A. Patel, "Classification Algorithms and Comparison in Data Mining," vol. 4, no. May, pp. 206–210, 2015.

[14] A. Journal, O. Tanaseichuk, and A. H. Khodabakshi, “An Efficient Hierarchical Clustering Algorithm for Large Datasets,” vol. 2, no. Figure 1, pp. 1–6, 2015.

[15] D. Jayalatchumy, Dr. P.Thambidurai, “Web Mining Research Issues and Future Directions – A Survey”, IOSR Journal of Computer Engineering (IOSR-JCE)e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 14, Issue 3 (Sep. - Oct. 2015), PP 20-27

[16] Pradnyesh Bhisikar1,Prof. Amit Sahu, “Overview on Web Mining and Different Technique for Web Personalisation”, International Journal of Engineering Research andApplications (IJERA) ISSN: 2248-9622 Vol. 3, Issue 2, March -April 2013, pp.543-545

[17] GovindMurariUpadhyay, KanikaDhingra, “Web Content Mining: Its Techniques and Uses”, Volume 3, Issue 11, November 2013 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering

[18] A. N. Networks, M. I. T. Press, N. Networks, P. Hall, S. Maps, I. Exploration, and U. Learning, “Self-Organizing Maps ( SOMs ),” pp. 187–202, 1997.

[19] K. P. Adhiya and S. R. Kolhe, “AN EFFICIENT AND NOVEL APPROACH FOR WEB SEARCH PERSONALIZATION USING WEB USAGE MINING,” vol. 73, no. 2, pp. 321–335, 2015

[20] A. Hannak, P. Sapie, D. Lazer, and A. Mislove, “Measuring Personalization of Web Search.”y

[21] S. S. Kontamwar and M. T. Cse, “A Review – Clustering and Preprocessing For Web Log Mining,” pp. 2471–2473.

[22] S. Diwandari, A. E. Permanasari, and I. Hidayah, “Performance Analysis of Naïve Bayes , PART and SMO for Classification of Page Interest in Web Usage Mining,” pp. 39–44, 2015.

[23] K. L. Goh and A. K. Singh, “Comprehen nsive Literature Re view on Machine Learning structure s fo or Web Spam Classsification n,” vol. 70, pp. 434–441, 2015.

[23] S. S. Nikam, “ORIENTAL JOURNAL OF A Comparative Study of Classification Techniques in Data Mining Algorithms,” 2015.

[25] P. Mehtaa, B. Parekh, K. Modi, and P. Solanki, “Web Personalization Using Web Mining : Concept and Research Issue,” vol. 2, no. 5, pp. 510–512, 2012.

[26] S. S. Kontamwar and M. T. Cse, “A Review – Clustering and Preprocessing For Web Log Mining,” pp. 2471–2473.

[27] R. Kumar, “Classification Algorithms for Data Mining : A Survey,” vol. 1, no. 2, pp. 7–14.

[28] Neelamadhab Padhy1, Dr. Pragnyaban Mishra 2, and RasmitaPanigrahi, “The Survey of Data Mining ApplicationsAnd Feature Scope”, International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012

[29] A. K. Santra1, S. Jayasudha, “Classification of Web Log Data to Identify Interested UsersUsing Naïve Bayesian Classification”, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 2, January 2012ISSN (Online): 1694-0814

[30] M. P. Jarkad and P. Mansi, “Improved Web Prediction Algorithm Using Web Log Data,” pp. 4902–4907, 2015