AN OPTIMIZED APPROACH TO IMPROVE THE QUALITY OF EDUCATION

Dissertation submitted in fulfilment of the requirements for the Degree of

MASTER OF TECHNOLOGY in

COMPUTER SCIENCE AND ENGINEERING (PART TIME)

By

PRIYA BABBAR

Registration number

41400002

Supervisor

MR. BARJINDER SINGH

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

June 2017

June 2017

Abstract

Data Mining is a process of extracting knowledge from large amount of data. It is used in EDM (Educational data mining). Educational data mining is a field for discovering knowledge from large amount of Educational data. The main purpose of EDM is to find the appropriate pattern of educational data so that there is improvement of qualification of education. Different aspects are evaluated like social, economic, personal, cultural, geographical, institute environment and other in education research study. Such aspects may either help a student in shining during academic period or halt academic program. Such failure is known as drop-out. Data mining algorithm helps in finding those factors; that are mostly contributing the student's performance. If we work on most contributing attribute better results can be achieved. In our research we are going to construct a hybrid model that can fit in Educational data mining. Hybrid approach is an approach which is combination of two or more techniques of data mining such as association, Clustering, Bayesian networks, neural network's machine learning technique, fuzzy logic, genetic algorithms etc. In this research, we discuss how a hybrid approach based data mining model can help to improve an education system by enabling better and effective teacher-student interaction.

DECLARATION STATEMENT

I hereby declare that the research work reported in the dissertation entitled "AN OPTIMIZED APPROACH TO IMPROVE THE QUALITY OF EDUCATION" in partial fulfillment 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 Mr. Barjinder Singh. 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

Priya Babbar

R.No. 41400002

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the M.Tech Dissertation entitled "**AN OPTIMIZED APPROACH TO IMPROVE THE QUALITY OF EDUCATION"**, submitted by **Priya Babbar** 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

Barjinder Singh

Date:

Counter Signed by:

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

   HoD Name: _____

   Date: _____

2) **Neutral Examiners:**

   (i) **Examiner 1**

   Signature: _____

   Name: _____

   Date: _____

   (ii) **Examiner 2**

   Signature: _____

   Name: _____

   Date: _____

ACKNOWLEDGEMENT

All praise in the name of almighty God, who give us in the darkness and help in difficulties. The dissertation is the result of full semester of work whereby I have been accompanied and supported by many people. It is a pleasant aspect to that I have the opportunity to express my gratitude for all of them.

I am also extremely indebted to my guide **Mr. Barjinder Singh** (Assistant professor, Department of Computer science, Lovely Professional University, Phagwara). I am very much thankful to **Mr. Barjinder Singh** for picking me as a student at the critical stage of my masters. I warmly thank him for his valuable advice, constructive criticism and his extensive discussions around my work.

I expand my thanks to my friends and family who always kept my spirits up with their extended love, affection and support at the time of my project work.

At last but not the least, I would like to pay high regards to the authors whose work I have consulted very often during my project work. And I would like to thank Lovely Professional University that provided me the road for the completion of my degree in this particular field.

Priya Babbar

TABLE OF CONTENTS

CONTENTS **PAGE NO.**

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND

Data mining technique is helpful for several reasons in private as well as public sectors. Many Industries use Data Mining technique to extract the valuable information from the large database to minimize costs, enhance research, and increase sales i.e. banking, medicine, insurance, retailing and EDM (Educational Data Mining). By the increase of technology of computers the collection of data, storage of data as well as manipulations of data have become straight forward. There is trade off between size of data and performance time. If the size of dataset is large then performance is automatically decreased. Data mining is process of extracting knowledge from large amount of data. The main reason for using data mining technique is that it collects important information which provides us better result. Data mining tool is used to find unknowns and relations between them. This method include statistical as well as mathematical model. Data mining process is performed on collected data which is represented in different forms like quantitative form, web, image processing, textual as well as multimedia forms. They contain association sequence or path analysis, classification, clustering and forecasting. The very important step is to find knowledge from data or KDD (Knowledge discovery of data). It includes various steps for extracting meaningful data. Data mining is related to more than one field which include database system, statistics, visualization, fuzzy approach. It combines techniques from computer graphics, business, picture analysis, pattern recognition. Knowledge base is used to search resulting pattern, knowledge include values of attributes, and domain knowledge include Meta data. Data mining system consist of functional units for tasks like association, classification, prediction, cluster analysis, outer analysis. Data mining involves data mining methodology, user interaction, performance, scalability. Data mining is used for financial data analysis like banking services, investment services, insurance services, industries. Data mining is used for biological data analysis, scientific applications, and intrusion detection.

Data Mining is extensively useful in EDM (Educational Data Mining). Educational data mining is an emerging field for knowledge discovering from large amount of Educational data. The purpose of EDM is to find the pattern of educational data so that qualification of education can be improved.

EDM is the educational research study of Variety of methods in which different aspects are evaluated like social, economic, personal, cultural, geographical, institute environment and other. Such aspects may either help a student in excelling during academic period or halt academic program of a student. Such failure is known as drop-out. [1]

Data mining algorithm helps in finding those factors, that mostly contributing the student's performance. If we work on most contributing attribute better results can be achieved.[31]

1.2 DATA MINING

Data is hidden taken out through techniques of data mining [20]. It provides important data which is necessary for decision making. Classification include categorical classes, prediction include valued functions. Pre-processing of data contain data cleaning to reduce noise, relevance analysis to remove unessential attributes, forecast accuracy, scalability, interpretability, computational speed.

Classification: The approach of classification includes mining processes suggest discovering rules on the basis of sub processes build. It include categorical values like discrete, unordered. It include approaches like k nearest neighbour classifier, case based learning. It is used for fraud detection and medical diagnosis. Classification is done by using k means algorithm, genetic algorithms. We can use cluster centers for data classification such that the computation load is less and the effect of noisy data is reduced.

Data compression: We can use cluster center to show the actual dataset. The numbers of clusters are less than the size of actual dataset. So goal of data compression can be achieved.

Prediction: The prediction model include continue values. It Predict numeric values in which predictor can guess value of predicted attribute for new data.

1.2.1 Types of Data mining algorithms

It is collection of various methods which can perform task. Currently lot of data mining techniques used to handle large dataset.

Association rule algorithm:- It mainly deals with search statistical relations between objects in dataset. It finds how events aggregate together.

Classification algorithm:- It can describe or classify objects related to dataset into predefined set of classes. It is supervised learning approach. It includes objects in dataset used to understand existing objects and predict behaviour of new objects.

Clustering algorithm:- It is collection of objects of similar type in one group. The cluster provides us better results.[28]

Inductive and Deductive learning:- Machine learning in mainly classify into two different types. In deductive learning, we learn something with existing knowledge and produce some new knowledge from existing knowledge. In inductive learning rules and patterns are extracted from large datasets. In clustering partition the dataset in to subsets for optimization.

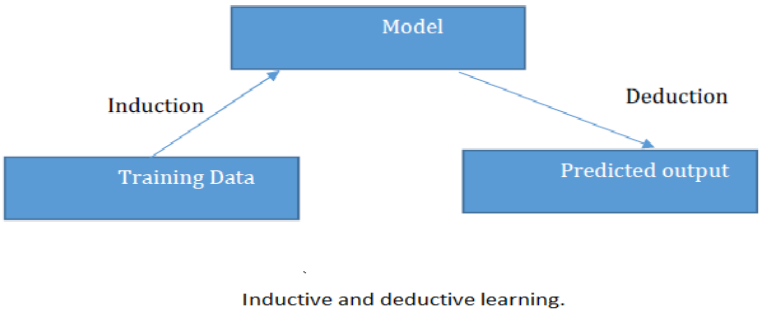


Figure 1.1 Inductive and deductive learning.

1.3 DATA MINING PROCESS:

Data mining is a process of extracting or mining knowledge from large amount of data [20]. It means knowledge extraction, knowledge mining of data, pattern analysis and data

knowledge discover from data. It is process of discover required knowledge from database. It includes various operations such as selection, processing, transformation, interpretation and evaluation. There are various steps of knowledge discover. It selects a dataset or its subset. It removes noise from data.

Data cleaning:- It is process of removing noise and inconsistent data. It can fill absent values.

Ignore the tuples.

Fill absent values manually.

Use global constant to fill in absent values.

Use attribute mean to fill in absent value.

Use attribute mean for all samples belong to same class.

Use most important nearest values to fill the absent values.

Data Integration:- It can combine multiple sources in data warehouse. It includes multiple database, data cubes and files. Redundancy is duplication of data. It is removed by correlation analysis.

Data selection:- It can retrieve data from database which is required for analysis. It can describe how to select various attributes.

Data Transformation:- In data transformation, data is transform into forms appropriate for mining. It can involve various steps:-

a) Smoothing:- It helps us to remove noise from data.

b) Aggregation:-Data aggregation is a process of gathering information and expressed in a summary form such as statistical analysis. A common aggregation purpose is to get more information about particular groups based on specific variables such as age, income.

c) Generalization of data:- In generalization there is replacing of low level data to high level concepts through use of concept hierarchies.

Pattern evaluation:-It can identify those patterns which represent knowledge based on some measures. Data mining is a process of taking out knowledge from large database. It can evaluate results in form of patterns. The large amount of knowledge is collected from different knowledge engineers.

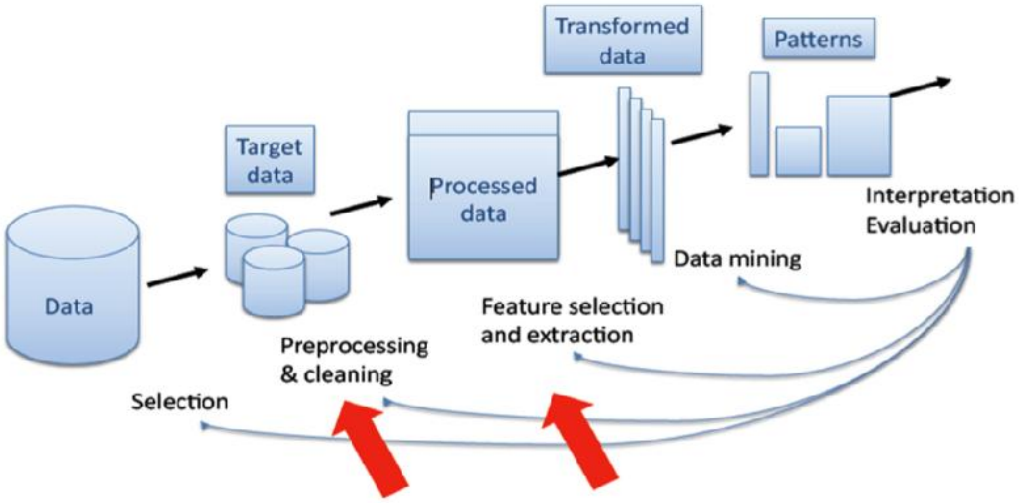


Figure 1.2. KNOWLEDGE DISCOVERY PROCESS

1.3.1 Field in which data mining used:-

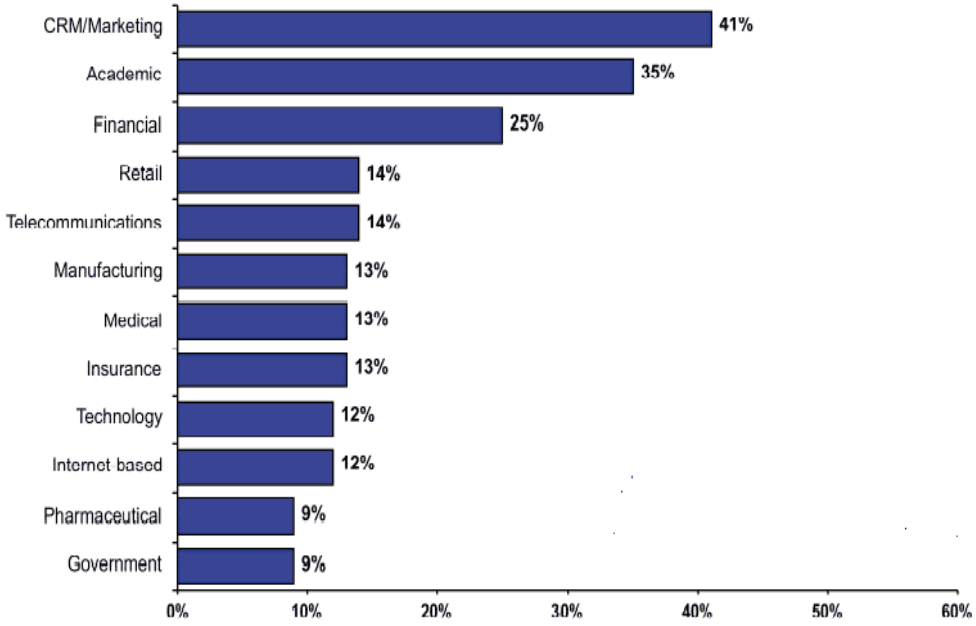


Figure 1.3 Data mining used in different fields.

1.4 DECISION TREE

It is widely used technique in data mining [20]. It is basically a representation of data which is in hierarchical shape. The top node is called root node and below the last level nodes are called leaf nodes. Between root node and leaf nodes there are internal nodes. The internal nodes in decision tree are represented by a rectangle and leaf nodes are represented by oval. Decision tree is constructed on the principle of recursion. In this process root node (main attribute) is recursively divided into sub nodes (Sub attributes). The process is repeated until some class is not reached.

Decision tree in classification: Decision trees are very useful in classification. Let's suppose a record X having no class then simply insert the record at the root then using the classification rules the class is found. Construction of decision tree is basically splitting a record into sub-records based upon some attribute. This attribute selection is done using measures of attribute selection. These measures are information gain, gain ratio and gini index. In information gain method, information gain of every attribute is calculated. Then these results are evaluated and the highest contributing independent factor is determined that effects the output of dependent variable.

Expected information needed to classify a record is calculated by the formula:

1. $\text{Info}(D) = -\sum_{i=1}^{n}(\text{pi})\log 2(\text{pi})$

The contribution of each independent attribute is measured towards the dependent variable(admission in the considered example). This is done by the formula:

2. $\text{Info}_A(D) = \sum_{j=1}^{v} \frac{|Dj|}{|D|} \times Info(Dj)$

Finally the information Gain is evaluated as:

3. $\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D)$

This factor tells us that how much it will be beneficial if we partition on A attribute.

ID3 and C4.5:- These are developed by Quinlan for inducing *Classification Models* from data that are also called decision trees. We are given a set of accounts. Each record has the same construction, consisting of a number of quality/value pairs. These attributes shows the group of the record. The problem is to decide a decision tree. This decision is done on the basis of answers to questions. These questions are about the non-category attributes predicts correctly the value of the category attribute. Usually the category attribute takes only the values {true, false}, or {success, failure}, or something equivalent. In any case, one of its values will mean failure.[29]

CHAID:- It stands for *Chi*-squared Automatic Interaction Detector. The CHAID is a kind of analysis that finds how variables are best combined to explain the effect of a given dependent variable. The model can be used in situation of market dispersion, predicting and interpreting responses or numerous of other research problems.

CHAID analysis is mainly useful for data expressing categorized values not for continuous values. For this kind of categorized data some common statistical tools such as regression are not applicable and CHAID analysis is a perfect tool to discover the relationship between variables.

One of the advantages of CHAID analysis is that it can visualize the relationship between the target (dependent) variable and the related factors with a tree image.

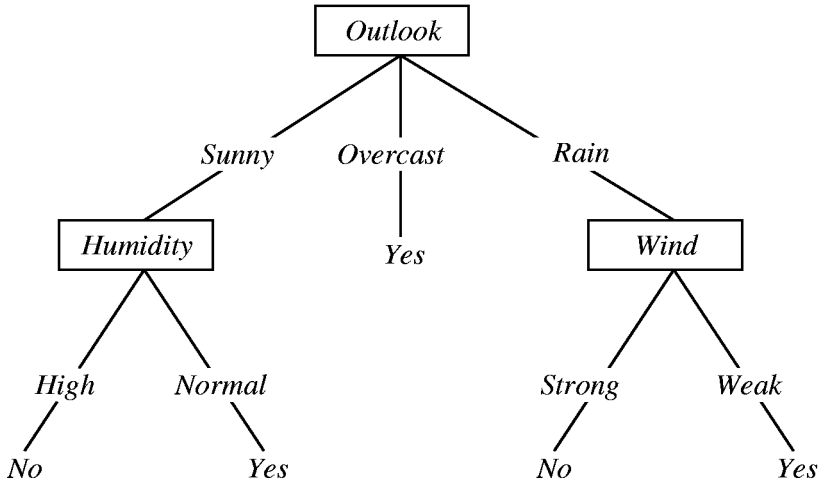


Figure 1.4: CHAID

1.5 CLUSTERING:

Clustering is a process of grouping physical objects into similar objects classes [4]. Similar objects classes are collected in one cluster and dissimilar object in other cluster. In data mining, the analysis of cluster focuses on handling large amount of data by dividing them into various groups. Various tools are used for cluster analysis. Grouping is done in such a way that objects in similar group are same and different group are different from each other. First partition is partition of data into groups. Now assign label to them according to their similarity. K means algorithm is used to calculate centroid. It shows that objects which are similar to each other and dissimilar to each other. It analyzes mean value of object for each cluster. It is subset of collection of objects of same and different size. Cluster analysis is like factor analysis, makes no distinction between dependent and independent variables. Factor analysis reduces the number of variables by grouping them into a smaller set of factors. Cluster analysis reduces the number of observations or cases by grouping them into a smaller set of clusters. It can partition the object into set of sub class. It has different qualities. These qualities depend on their size of nature. It can satisfy two conditions. First condition is that intra class similarity is high and second condition is that inter class similarity is low. The clustering problem requires calculating similarity between documents in order to assign them to a particular cluster. Cluster analysis is one of the key technologies in the field of data mining and machine learning which has been applied in many areas: data mining and knowledge discovery, pattern recognition and pattern classification, data compression and vector quantization and plays an important role in biology, geology, geography, and marketing.

1.5.1 Requirements of clustering in Data mining:- Clustering plays an important role in data mining. It can group together objects of same size. It provides us better results.

1. Scalable:- Mostly clustering algorithms handle small data set easily which contain many data objects. In large data base include billion of data objects. Clustering may deal with these large data sets to provide scalable results. There are different input parameters which provide us better result and produce scalable output.

2. To deal with different object attributes:- There are many kind of algorithms design to handle different data objects. These algorithms can handle categorical as well as ordinal data. Each object contains different kinds of attributes related to different data sets. Each attribute has different role play in cluster analysis.

3. Discover arbitrary shape clusters:- Algorithms are obtained to cluster numeric as well as categorical data. Clustering algorithm can describe clusters based on Euclidian distance. It can measures distance to find spherical clusters having same size as well as density. It is necessary to develop these kinds of algorithms that detect arbitrary shape clusters contain several data objects.

4. To determine input parameters related to domain knowledge:- Most of algorithms provides input parameters to user in cluster analysis. Results obtain from clusters become equivalent to input parameters. Data set contains high dimensional objects having difficult to determine parameters.

5. To handle noisy data:- Data base contain lot of information which may be unknown. Some kinds of algorithms handle clusters having poor quality. It is necessary to remove these outliers from data base.

6. High Dimensional:- Data ware house contain many attributes. Clustering algorithm may deal with low dimensional data easily. It can be challengeable issue to search clustering data objects in large dimensional space.

7. Constraints based clustering:- In real world applications contain many types of constraints which describe clustering. The most important issue is to find good clusters which increase efficiency of data set.

8. Usability:- Results obtained from clusters become interpretable. Clustering contain many applications which provides us more usability.

1.6 TOOLS OF DATA COLLECTION AND ANALYSIS

1.6.1 MYSQL WORKBENCH:

It is a fast and easy to use relational database management system. This system acts as an interface between any user and the database. The primary job of database is that it makes

data manipulation easy and minimizes the data inconsistency. It is becoming popular due to following reasons:

1. We do not need to pay anything to use it.
2. It can handle a large subset of the functionality of very expensive aas well as powerful database packages.
3. It uses a standard form of the well-known SQL data language.
4. It works on many OS (operating systems) with many languages like PHP, C, C++, JAVA, etc.
5. MySQL works very quickly that is it is very fast and works very well with large data sets.
6. It is very friendly to PHP.
7. It supports large databases with 50 million rows or more in a table. The default file size limit for a table in MYSQL is 4GB, but you can increase this size if needed.
8. It is customizable which means open-source GPL license allows programmers to modify the MySQL software to fit their own specific environments.[21].

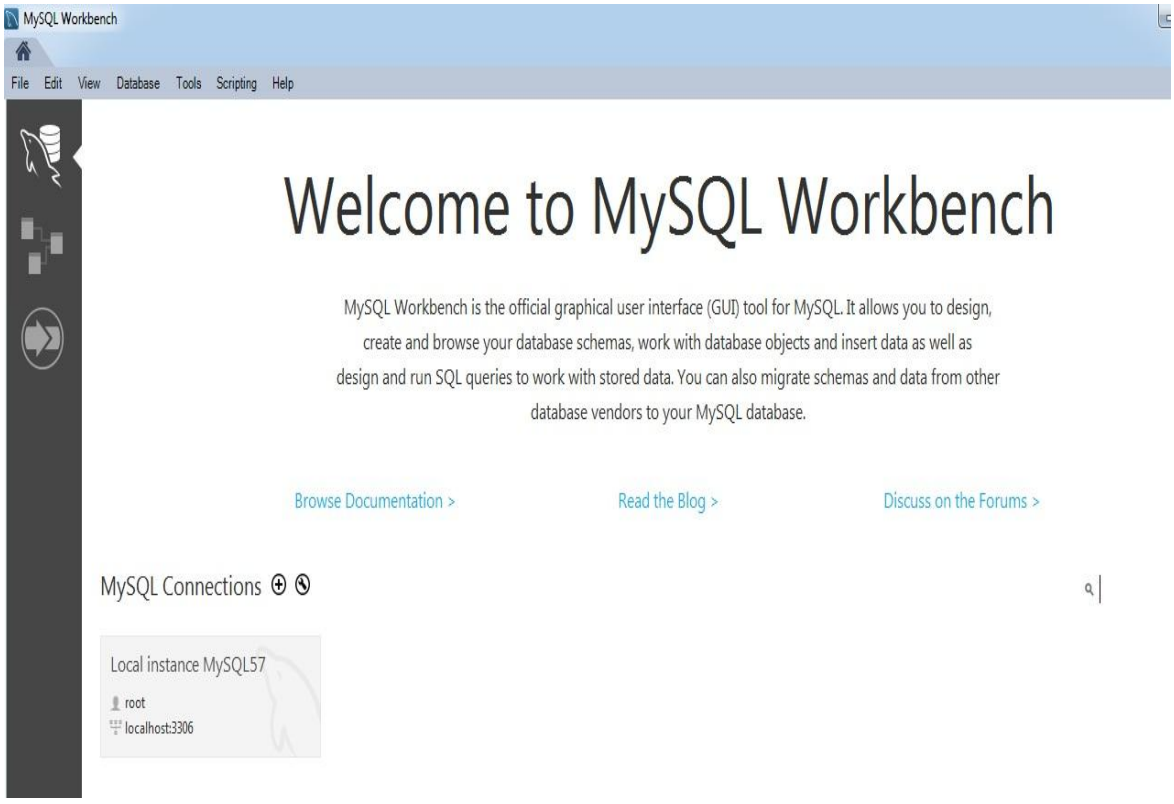


Figure 1.5: Representing outlook of MYSQL Workbench

1.6.2 NetBeans 6.0

NetBeans is an integrated development environment (IDE) [22]. Primarily it was developed to be used with Java language only. But, now days, it can also be used with other languages such as C/C++, PHP etc. In this work, Java language has been used with NetBeans 6.0 [23].

Figure 1.6: Representing Outlook of NetBeans 6.0

This tool has been used because of its interesting and easy to use features. Some of its features are:

- Best support for latest Java Technologies
- Smart and fast code editing
- Rapid User Interface Development
- Multiple languages support
- Support for Cross Platform

A connection has been established between Net beans and MYSQL using ODBC (Open database connectivity).

1.6.3 WEKA 3.6.3

WEKA is a tool which is specially designed for classification. It implements supervised learning. In supervised learning all the class labels are known in advanced. WEKA is best suitable for classification tree algorithm such as: CHAID, C4.5, ID3 (Iterative Dichotomiser 3). It can handle both discrete and continuous attributes. Due to its

popularity it is widely used in educational institutions. Other important fields in which WEKA is used are: medical, financial institutions and industries. It is supported by Windows operating system. [24] [25].

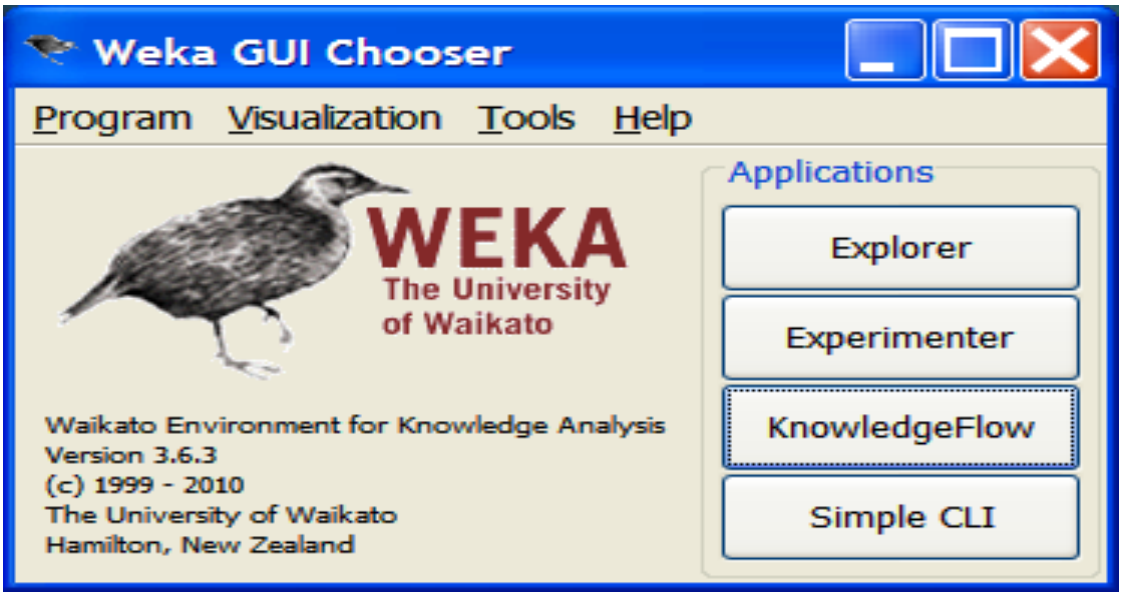


Figure 1.7: Representing outlook of WEKA 3.6.3

WEKA features: The best features supported by WEKA are as under:

- Data Access
- Feature Transformation
- Feature Selection
- Learning Algorithm
- Error Evaluation
- Classification

CHAPTER 2

REVIEW OF LITERATURE

This chapter studies the broad review of literature both at national as well as international level associated with the theme of the research work.

Hany M. Harb1[2]: In the paper feature selection technique is used to reduce the number of feature form the large attribute set. In this paper author use ASSISTments platform dataset which is a web based teaching system developed at Worcester Polytechnic institute and used with 4^{th} to 10^{th} grade math students. In this paper author used technique to remove irrelevant, redundant or noisy data. In this paper author used various classification algorithm and ranker algorithm to find top most contributed attribute and removed the less appropriate attribute. This helps to speeds up the process of data mining and improves its performance parameters such as predictive accuracy.

Carlos Marques-Vera[3]: In this research paper author used three different approaches. Cross tabulation analysis, Feature selection and Balancing imbalance data. Features selection method is used to select those attribute which are highly affected dependent variables. Classification tree is built considering all available attributes. This method finds out all possible splits that can occur for each predictor variable at each node. The search stops when the split with the largest imprudent in goodness of fit is found. It is based on the Gini measure of node impurity. Several feature selection algorithms are applied and features ranking higher in multiple algorithms are selected. In this manner 15 important features are selected from original 77 features. Misbalancing issue is resolved by using data balancing and rebalancing algorithm specifically SMOTE(Synthetic Minority Over sampling technique). Ten fold cross validation is used for establishing training and testing data from original data. This data set is prepared in three categories. First category contains data with all 77 attributes. Next category contains data with 15 important attributes. Last category contains balanced data after applying rebalancing technique in weak.

Wagstaff kiri [4]: clustering approach is used for data mining analysis. In this paper we read how k-means clustering algorithm modified using knowledge domain information. It can also apply to automatic road detection lanes from GPS system. This algorithm access set of features which describe each data object. Mostly in real world applications background knowledge must be important related to our dataset. K-means is popular algorithm which is used in different domains like segmentation, banking, and Information retrieval and solves problem related to our domain. First we develop k-mean algorithm which provide us knowledge in form of instance level constraints. Second restriction is testing with random constraints. It obtains result in form of graph where each graph describes its efficiency and accuracy. For each constraint randomly choose two instances from data set and check their labels. If they are similar we generate multi link constraints. It describes how background information utilizes in real domain, global position system. It reduces the complexity of data set in which various attributes are related to our field. Computational complexity of our constrained k-means algorithm is reducing as compare to original k-means.

Tzortzis.F Grigorious and likas.C Aristids[5]: Kernel k-means algorithm is extension of standard k-means clustering algorithm. It can describe non linear differentiated cluster. In this paper we proposed global kernel k-means clustering algorithm is developed to overcome cluster initialization. It consists of many executions of kernel k-means from best initial centroid point. Two modifications done to reduce computational cost and different data set help us for compare kernel k-means for random initialization. Kernel k-means is extension of standard k-means which maps data from input space to feature space through non linear transformation and minimize cluster error rate. This algorithm works in incremental manner by solving all intermediate problems. The basic concept behind proposed method is to choose near optimal path with k-1 clusters and initialization of k-means cluster. Drawback of global k-means is its high computational complexity. It requires running kernel k-means n times when solves clustering problems. To obtain solution of this problem we need to run weighted kernel k-means instead of k-mean. But important issue is change the way of select data point which reduces cluster error. It optimize clustering error rate in feature space by locate their optimal solution.

K.A Abdul Nazzar, M.P Sebastian[6]: With new technology scientific methods used for collect results in large scale accumulation of data related to different fields. Conventional data base method is used to extract useful information from data banks. Cluster analysis is important data analysis technique used in many application areas. This paper can represent proposed method for making our algorithm more effective and efficient which helps us to reduce complexity. It is practically impossible to extract useful information by using conventional database analysis techniques. In k-means clustering algorithm main idea is to classify set of data set in to k number of disjoint groups. It may consist of two separate phases. In first phase include k centroid of each cluster. Second phase describes each data point belongs to given data set which associate its nearest centroid point. It provides optimal solution which is dependent to select local initial centroid. It can take both numeric and continues attributes. Our proposed algorithm helps us to increase accuracy and efficiency of k-means clustering algorithm. But there is also limitation of proposed algorithm, the value of k no of clusters required is given as input regarding distribution of data points.

Clara Pizzuti and Domenico Talia[7]: On large dataset clustering algorithm usually takes more time than small data set. So this is very challenging issue to work on large data set with clustering in very short time. In this study they used the concept of scalable clustering. By this method overall computation time is reduced. In clustering task can be divided around multiple processors among multiple systems. Each processor must perform their own task and reduces execution time. It can reduce processing time along with scalable parallel clustering for handle large data set. There are many strategies for parallelize data set of p auto class. Independent parallel search: - It can use classification technique for parallelism. Every processing element performs different classification and share data set. During the computation classification performs in parallel to evaluate accuracy. Task parallel approach:- It can balance the computational cost of every processor. It can assign task parallel to every processor. It can gather results at end of classification. It can predicts execution time, speed up and efficiency. The research purposes a new algorithm p-auto class. In this algorithm, p stands for Parallel computing working of p-auto class. This class divides the final data set among the number of

processors of multicomputer. Each processor then asses its partition and they show the results between each cluster. P-auto class is based on theory of Bayesian classification and offers a parallel computing on different processors. To implement p-auto class MPI (Message passing interface) is used to exchange the messages among different processors. The efficiency of algorithm is increased by this.

Saadat Nazirova[8]: In this paper various methods that deals with spam mails are used. These methods are classified in two categories: Method to avoid spam distribution and Method to avoid spam receiving. The second method is again sub-divided into Theoretical approach and Filtration approach. Under theoretical approach three techniques: Traditional, Learning and Hybrid are explained. Similarly, Client and server approach is explained under Filtration approach.

Learning based method is used to avoid spam mails received from server. This is an intellectual method based on Data Mining Algorithms for e-mail filtration. This algorithm classifies the data into pre-defined classes. In this paper, researcher divides all mails into two categories: Spam mails and legitimate e-mail. There are some parameters that decide that received mail is spam or legitimate. The list of parameters is represented with symbol ζ. In this research, Image based spam filtering is used which detects those spam messages that are embedded into an image. Some traditional text-based information does not work on images. Three layer image-spam filtering method is purposed for analysis. First layer acts as Mail- header classifier. Second layer is an Image header classifier and the third layer acts as Visual feature classifier. A statistical feature extraction for classification of image-based spam is implemented by using artificial neural networks (ANN). The next method is Bag of words model. This is based on Natural Language Processing and Information Retrieval. In this method two bags of words are used by researcher. One bag is filled with words found in spam mails and other bag is filled with words that are found in legitimate mails. By considering e-mail as a bundle of words from any of these bags, they used Bayesian probability to determine to which bag this e-mail belongs. K-nearest neighbor, SVM (support vector machine), boosting classifiers are also applicable to the bag of words.

P.Moniza and P. Asha[9]: In this paper, researcher gives various tips to stop spam mails like Customer Revolt- forcing companies not to publicize their confidential information like e-mail, phone number, etc., Domain filters- Allow mails from specific servers only, Black listing, White Listing, Government action law implemented by government against spammers. All these are theoretical concepts which are not possible to implement in real time scenario. Some automated recognition methods for spam detection are also discussed in which machine learning algorithm is implemented. Main focus of research is on SAG (Structure Abstraction Generation) which generates an HTML tag sequence to represent each mail. This paper deals with email layout structure instead of detail content text.

Patricia Bellin Ribeiro, Luis Alexandre da Silva, Kelton Augusto Pontara da Costa[10]: In this research paper, researcher has compared various available technologies of Data Mining on SPAMBASE dataset. SPAMBASE Dataset contains 57 attributes and 4601 sample previously labeled mails. Out of which 906 instances has been used. From these instances, 453 instances are classified as non-spam and remaining 453 are labeled as spam mails. Twelve methods: Random forest, Rotation Forest, Nbtree, J48, Bagging, MLP, LogitBoost, AdaBoost, RBF, Naive Bayes, OneR and ZeroR are implemented on the data. A standard statistical method called cross- validation is chosen to assess the effectiveness of the compared techniques. This approach randomly partitions the data set into training and test sets, being the former composed by 75% of whole dataset, and the latter contains the remaining 25% of the dataset. This procedure has been executed over 10 running, being the mean accuracy employed for comparison purposes. A ROC curve has also been used to assess the classifiers performance. As a result, Rotation forest and Random forest are two classification techniques which gave maximum correctly classified instances. The accuracy of Random forest test is evaluated as 99.42% and it is 98.03 % in Rotation Forest test.

Ji-Dan and Qiu Jianlin [11]: With the popularity, facilities of technology and less cost of hardware and software the size of data is expanding in our lives. Due to large data it is very difficult to get the meaningful information from the huge dataset. For taking right decision and dig out meaningful information we use technique of data mining. In this

paper JI-DAN used two techniques of data mining. They used clustering and decision tree. Clustering is used to associate homogeneous type of data in one or more clusters or describing and decision tree is use to analyzed the pattern. Ji-Dan and Qiu Jianlin combine both the algorithm named CA(component analysis) which produce better result than the original algorithm. The study is based on agriculture in which Maize Seed breeding takes the dependent variable. They applied the CA algorithm and shows that their approach is better than the earlier one. At present, we have accumulated rich agriculture meaningful data for the vast land and variety of crop assets. However, we are using just small quantity of data for lack of useful tools. Furthermore, agriculture itself has some difficulties like the complication of crop resources and the influences on crop along with nourishment, water thickness and weather conditions which build the information dataset, high dimensional, dynamic, unfinished and doubtful ones which are hard for us to handle. Whereas data mining can express and forecast different datasets by different technologies and uncover out probable rules or models. So we can contract with these agriculture difficulties successfully, and the application and progress of data mining to agriculture will clearly be a new investigation point.

Yen-Liang Chen, Hsiao-Wei Hu, Kwei Tang [12]: So far, we have studied that the classification is based upon Boolean class labels. The Boolean or categorical class labels are not suitable in many real life problems. In this paper the researcher has purposed a new way of tree classification using hierarchical class labels. This newly purposed algorithm has been named as HCL (Hierarchical class Label classifier). The main focus of the researcher is on accuracy in the results. The researcher has considered a training set of 16 hypothetical customer records. The purpose of the study is to find the interest of a customer towards the purchase of a particular brand of a computer. Some considered [30] in the study are: Gender of the customer, Customer's Career, Customer's Income, Preferred product etc. Among these attributes preferred product is a dependent variable and all other are independent variables. The training data is further sub-divided based upon an attribute. The selection of that attribute is made on the basis of Gain ratio and entropy which calculates the maximum contributing factor and further that attribute serves

as a base for division of training data. The main drawback of this study is that, if there exists a gap between labels in the tree, then the accuracy level is not achieved.

Qiang Yang[13]: Data mining is very useful in many areas. This technology can also be applied on customer relationship management, which is helpful to figure out those customers who are unfavorable towards your products and those who are well-wisher or favorable towards your product. After getting this knowledge manually some post processing techniques are enable. These techniques show us about the behavior of many customers who are favorable or unfavorable. Pre-processing technique show you result in virtualized way. But they don't suggest any thing which helps us to increase profit. In his study Qiang Yang presents new algorithm that suggest some action which converts the opinion of customer from undesired to desired one. This increases the profit, which is objective function of his paper and he used Quiang Yang decision tree as data mining technique.

Souptik Datta[14]: It is the world of distributed database. We are accessing internet day by day for everything and it is located on distributed areas over whole world. On distributed environment data mining application is very challenging issue. Souptik Datta and his team implements data mining (k-means) over distributed environment using peer-to-peer network. They divide the whole concept into two algorithms. First algorithm is designed to produce local synchronization and second algorithm is used to combine result of all local system. The result of these algorithms produce better accuracy, speed and efficiency of the algorithm which run on centralized system. Dataset is divided into n number of systems with n processors. It shows that the efficiency of their algorithm is n time better than efficiency of centralized system. There are number of challenges that can be faced when data is distributed placed on nodes. These challenges can be node failure. When any node suddenly crashed, the whole information inside node is lost. So we need recovery process. Data change:- If data on any node is change, it is responsibility of that node to inform the other nodes that data has been changed. It mainly focuses on three factors. Accuracy: In his study various experiments done to show the overall accuracy like data coming from dynamically any node from distributed environment and result is

similar, same experiment done on 2000 nodes which shows again similar result. Communication: Different experiments also perform to show complexity of communication link over distributed environment from time to time. Some nodes are dynamically added or removed and simultaneously check the complexity of algorithm and their algorithm proved robustness and more efficient. Scalability: they show scalability experiment with peer to peer network from 200 to 2000 nodes

Jasna Soldic-Aleksic [15]: In this paper, two data-mining models Kohonen self-organizing model (SOM) and CHAID (Chi-square Automatic Interaction Detector) decision tree model are used. The basic purpose of this paper is to merge these two methods to develop a new technique. This technique is used in market analysis and clustering. This paper focuses on visualization of market trends and dividing the customers of the products into clusters. SOM is used for visualization purpose. SOM provides good clustering results and CHAID is a best interpreter of the SOM results. Due to this combined approach both techniques are purposed in this paper. This paper mainly focuses on the attributes 1) market segmentation 2) Customer attitude Analysis 3) Clustering the market for testing 4) Discovering opportunities for new product. This information is useful in the analysis of current trends and then evaluation of new approaches.

One of the effective clustering mechanisms for the market segmentation is a standard cluster analysis. But there are a set of the other clustering methods, which belong to the group of hierarchical or non-hierarchical procedures.

Olaiya Folorunsho [16]: Like other fields Medical field is also expanding in nature, in which different types of patients are involved i.e. their diseases, symptoms, medicines are different so its very difficult for expert to take decision about patient's treatment. In his study Olaiya take the medicine dataset to predict the patient's health condition. Olaiya compare two classification techniques: Artificial Neural Network(ANN) and decision tree for diabetes patients. Many performance measures are studied like kappa statistics, mean absolute error etc. Final conclusion was that Decision tree algorithm is better than

Artificial Neural Network. In his study 200 patient's dataset were collected & nine variables were used i.e. age, smoking status, blood pressure etc.

Nancy Lekhi and Manish Mahajan [17]: In this paper the researcher used the hybrid approach for outlier detection. They used two algorithms: K-mean and Neural Network. The proposed method use Integrating Semantic Knowledge (SOF- Semantic outlier factor) for outlier detection. This method detects the semantic outlier. Semantic outlier is a data point that behaves differently from other data points in the same class or same cluster. The main motive of this research was to reduce the number of outliers in clusters as well as data by improving the cluster formulation methods so that outlier rate reduces. It also decreases the error and improves the accuracy. The result showed that the hybrid algorithm performs better than that of genetic k-means. This proposed method deals with text and date dataset that has not been implemented before using genetic k-means.

Norlida Buniyamin, Usamah bin Mat, Pauziah Mohd Arshad [18]: In this paper they proposes a Neuro-Fuzzy classification method to enable the prediction and classification of students which is based on their past academic performance. Then they used it to predict their future academic performance that is then classified into various ranges from weak students to excellent students. Further, they describe the development of a tool that will enable faculty members to identify, forecast and classify students based on academic performance measured using Cumulative Grade point average (CGPA) grades. In their study they used the data for proposed research from an Electrical Engineering degree program for years 2005, 2006, and 2007. The data set of 391 students collected from the university database system. This data includes CGPA of students of every semester and GPA scores in their courses such as Mathematics, Signal and system, Digital System, and English, etc, also CGPA of every semester and GPA.

John Jacob, Kavya Jha, Paarth Kotak, Shubha Puthran [19]: In this paper various Educational Data Mining techniques are studied like regression, clustering, classification, decision trees etc. Regression is a numerical evaluation process. In this process the students' performance is predicted based on the already acquired data set like lab grade,

CGPA, attendance etc. These methods help the university teachers to know about changes that are need to be made, provide remedial courses to the weak students, identify weak students and to make learning a better experience for these students.

CHAPTER 3

PRESENT WORK:

3.1 SCOPE OF STUDY

This research checks the effectiveness of decision tree classification as well as clustering algorithms by applying them to a large scale data set. Example: Classification methods try to find those students who are likely to fail or need more attention.[32] Focus on these kinds of students can better the quality of education and decrease the dropout rate. Clustering methods try to make cluster of students according to their knowledge of subjects. This helps the student to find job according to their taste. Experiment result will also show the best accuracy, less time taken, higher robustness and generalization ability in one of the algorithm.

3.2 PROBLEM FORMULATION

This research will find better efficiency as well as limitations of traditional algorithms. The research also finds out the best possible solution which can handle large amount of high dimensional data. Our research is about checking the performance of decision tree and clustering algorithms by applying them on different data sets using data mining tool and evaluates the outcome.

3.3 OBJECTIVE OF STUDY

In this work, a technology will be used which is based on data mining algorithms for the induction of decision trees. It is well suited in this context for various reasons.

1. Study of present algorithms – One major problem in data mining process is classification of large datasets. Various algorithms have been developed like CHAID, C4.5, K-Means, DBSCAN for mining large-scale high dimensional datasets but all have their own shortcoming in term of quality. The main objective of this research is to classify the development of tree based upon some attributes to find the *quality* of the tree.

2. Enhance the efficiency of building decision tree – In this work various pruning techniques can be used that can improve the efficiency of the construction of decision trees & Clustering and evaluate the performance of new approach.

3. Scenario – This research aim to develop a new approach which discloses those attributes that mostly contribute the student's performance. This helps to find those Students who are likely to fail and also help them to secure good grade in final exam.

4. Implement tool - In this research WEKA (Waikato Environment for Knowledge) tool will be used. WEKA is considered to be a best tool for Decision tree and clustering construction. It is compatible with windows operating system.

5. Analysis between computation times - The major issue that effects the efficiency of an algorithm is the necessity of resources. Different mathematical models have been developed that aims to find the amount of resources required by different algorithms. The resources that are considered are time, memory space, number of processors etc. The main focus of this research will be to develop an algorithm that needs minimum resources.

6. Eliminate error rate – Our algorithm try to minimize the rate of errors made by a predictive model.

3.4 SOURCES OF DATA SET

The dataset is collected from two sources. These are primary sources and secondary sources. Data collected from primary sources is known as primary data. Primary sources are survey, interviews, questionnaire and data collected from secondary sources is called secondary data. Secondary sources are newspapers, journals, libraries etc.Primary data are also known as raw data. To collect the data for the purpose of this thesis, both the primary and secondary sources have been used:

Primary Sources: The primary data has been based on the response received directly from Students, Institutes.

Secondary Sources: The first step in literature study is to review the research articles that provide general understanding of topic. The next task was to refine the text or data which is related to our research, from the large pool of information available on internet.

Two different ways to collect data:

1. Collection of data by online application
2. Available data sets from web sites.

Step:1

In the research the purpose is to design another form of the CHAID (*Chi*-squared Automatic Interaction Detector) algorithm. Because, it is based on a pre pruning technique to determine the right size of the tree. These algorithms are extremely faster on large datasets. It will verify that the multithreading technique is only interesting in the growing phase for the learning algorithms such as C4.5.

Step:2

Function Split (candidate attribute) : selected attribute

Max = $-\infty$

Selected attribute = NULL

For Each Candidate Attribute

    Relevance = Goodness of split (attribute)

    If (Relevance > Max) Then

        Max = Relevance

        Selected attribute =attribute

    End if

End For

Return (selected attribute)

3.5 RESEARCH DESIGN

From the proposed methodology, secret information could be extracted from large data set of personnel. It enhances the better understanding of decision makers and visualization of required knowledge. By this knowledge they can take right decision at right time. The various points show the design with the following steps:

- Objective structuring and Problem definition
- Data collection and preparation
- Data mining model construction
- Model analysis and evaluation
- Interpretation and knowledge extraction

- Using discovered knowledge

3.6 RESEARCH METHODOLOGY

Student's performance is a great concern for academic institutions. Classification and clustering methods like decision trees, Bayesian network, k-means etc can be applied on the educational data for predicting the student's performance in examination. These classification methods will be useful to identify the weak students and help them to score better marks. Various decision tree and clustering algorithms like C4.5, ID3 (Iterative Dichotomiser 3), k-means and CART (Classification and Regression Trees) can be applied to the research.

In this study, C4.5 algorithm is applied on Students of different colleges to predict their performance in the final exam. The outcome of the clustering is to group the similar types of students and analysis with inter cluster students. The outcome of the decision tree predicts the number of students who are likely to pass, fail or promoted to next year. The results provide steps to improve the performance of the students who were predicted to fail or promoted.

In our research, the data is firstly converted to an optimal dataset by applying various Expert rules and then on this dataset feature reduction is performed. Finally from this optimal dataset decision tree has been constructed. Following steps are performed in our research:

- Primary data has been collected from 1300 students from two different schools through questionnaires and interviews.
- After that training dataset has been created from this primary data by randomly selecting 650 rows.
- Knowledge elicitation has performed from domain experts.
- From this knowledge of experts, rules have been created in Java Language.

- These rules have been applied to the training data set.
- Then feature reduction is performed to this dataset by calculating gain ratio of every attribute and the attributes having minimum gain ratio have been deducted. After deduction we obtain an optimal dataset.
- The optimal decision tree has been constructed using c4.5 algorithm.

Finally, comparison has been made between existing approach and our approach to find the outcome.

Figure 3.1: The diagram represents existing approach

The figure shown above represents the method of existing approach. In this approach, after preparing training dataset classification algorithm C4.5 is directly applied to this which results in the construction of decision trees as well as classification rules. But, in our study, various expert rules are applied to the training dataset followed by feature reduction process that makes this set an optimal dataset. This is done because by doing this accuracy has been improved and computation time of C4.5 algorithm is reduced.

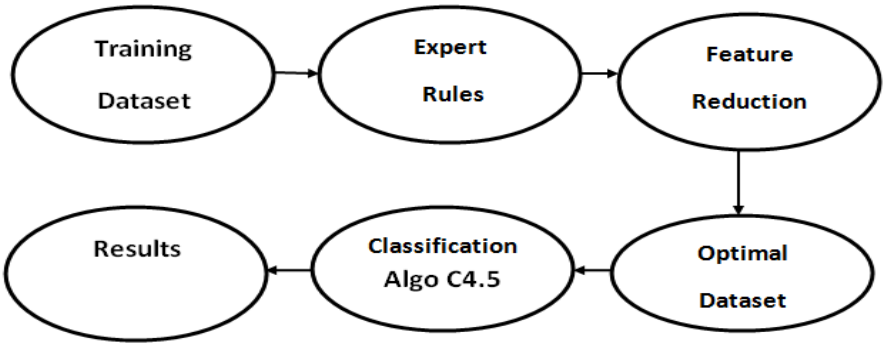


Figure 3.2: Representing our approach

The various steps which are implemented in our approach are described as below:

3.6.1 Importing packages from WEKA 3.4 to Netbeans IDE 6.0

WEKA 3.4 has been used in our research because it is considered as a good machine learning tool through which we can implement various data mining techniques. In this paper, training dataset has been converted into an optimal dataset by applying expert rules and feature reduction techniques. This is implemented on Netbeans IDE 6.0 platform by fetching packages from WEKA tool to Netbeans. 15 steps are performed to do this operation. These steps are as follows:

- Download and Install WEKA 3.4 on your machine in any drive let us suppose D:\WEKA
- Create a folder in E:\ drive which is named as backup
- Open D:\Weka folder as we can view all the extracted from weka-3-4-16jre.exe; 25.4 MB
- We will find a file called weka-src.java, after findng copy and paste it in E:\backup
- Extract all the files using WinZip in to E:\backup

- We will find here folders like lib, src, meta-inf, test and build.xml file
- Create a folder in E:\ drive name called tmp
- Open Net beans IDE and click on File\New Project, New Project dialog box appears

a. Categories: General
b. Projects: Java Application
c. Click on next button

- In the Name and location appears

    a. Project Name: weka
    b. Project Location: E:\tmp
    c. Project Folder: E:\tmp\weka
    d. Put tick mark in Set as Main project, create main class
    e. Fill the text box with "weka.gui.Main"
    f. Click on Finish button

- After creating file we can view it now in Net beans with Source package, Test package and libraries
- Go to E:\mytemp\src\main\java and copy the weka folder and then paste it in D:\tmp\weka\src.

    a. A dialog appears that shows folder name is existing, overwrite yes or cancel
    b. You can click on yes to all button.

- Now you have to switch to Net beans and click on Weka project and Source Package, here we can observe that all the earlier files such as classification, gui, clustering, filters etc are dumped into it.
- Click on Build Menu\Build Main Project, then weka project will be compiled

- Click on Run Menu \Run Main Project, then weka project will be executed

After performing these steps, inbuilt packages of WEKA 3.4 will be imported to NetBeans 6.0. Now, all the features of Data mining can be used in Netbeans IDE (Integrated Development Environment).

After importing packages the view of NetBeans is as under.

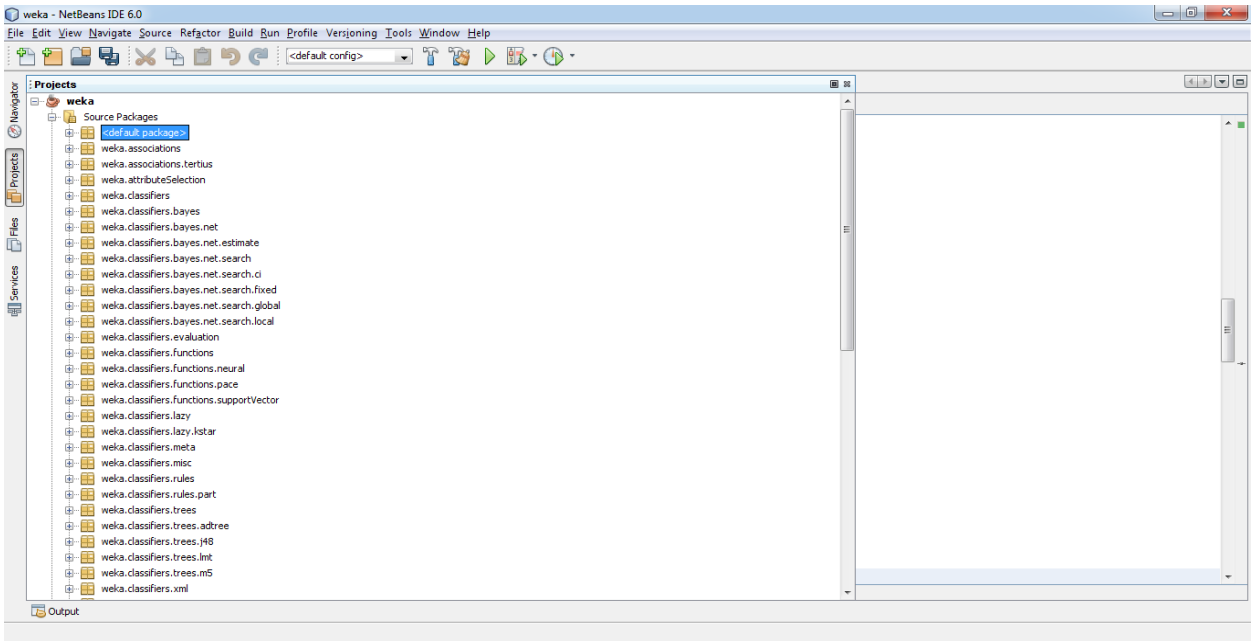


Figure 3.3: Figure represents view of Netbeans after importing packages from weka

3.6.2 Preparing Training Dataset

For training dataset 650 rows are randomly selected from the large dataset of 1300 rows, which has been collected through students by questionnaires and interviews. The collected data has been stored in only csv file format because WEKA accepts only this format files. The attributes considered for the analysis of students performance are[26]: Sex, Address, Parents status, Mother's education, Father's education, Mother's job, Father's job, Reason to choose this school, Other attribute is guardian, Travel time, Study time, Failures, School support, Family support, Paid, Activities, Nursery. This randomly selected data is the training dataset which is further processed to obtain an optimal dataset.

Attribute 'Sex' has domain values {Male, Female}

Address {Rural, Urban} which shows whether student lives in rural area or urban area, Age{15-20}, 'family size'{less than 3 or greater than 3}

Attribute Parents status{living together, apart} shows if parents are living together or not.

Mother's education {0,1,2,3,4} here 0 shows none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education , 4 – higher education

Father's education {0,1,2,3,4} here 0 shows none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education , 4 – higher education

Mother's job {nominal, health, civil services, at home, other} nominal- teacher, health care related, civil-police or administrative

Father's job {nominal, health, civil services, at home, other} nominal- teacher, health care related, civil-police or administrative

Reason to choose this school{nominal: close to "home", school "reputation", "course" preference or "other"}

Other attribute is guardian - student's guardian {nominal: "mother", "father" or "other"}

Travel time - home to school travel time {1 is <15 min., 2 is 15 to 30 min., 3 is 30 min. to 1 hour, or 4 is >1 hour}

Study time - weekly study time { 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours}

Failures - number of past class failures {n if 1<=n<3, else 4)

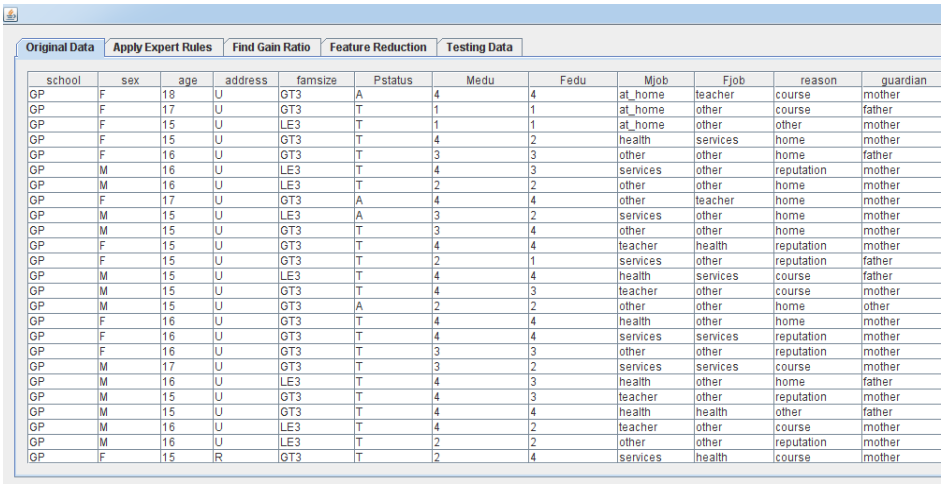
School support - extra educational support {yes or no}

Family support - family educational support {yes or no}

Paid - extra paid classes within the course subject (Math or Portuguese) {yes or no}

Activities - extra-curricular activities {yes or no}

Nursery - attended nursery school {yes or no}



Original Data | Apply Expert Rules | Find Gain Ratio | Feature Reduction | Testing Data

school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian
GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother
GP	F	17	U	GT3	T	1	1	at_home	other	course	father
GP	F	15	U	LE3	T	1	1	at_home	other	other	mother
GP	F	15	U	GT3	T	4	2	health	services	home	mother
GP	F	16	U	GT3	T	3	3	other	other	home	father
GP	M	16	U	LE3	T	4	3	services	other	reputation	mother
GP	M	16	U	LE3	T	2	2	other	other	home	mother
GP	F	17	U	GT3	A	4	4	other	teacher	home	mother
GP	M	15	U	LE3	A	3	2	services	other	home	mother
GP	M	15	U	GT3	T	3	4	other	other	home	mother
GP	F	15	U	GT3	T	4	4	teacher	health	reputation	mother
GP	F	15	U	GT3	T	2	1	services	other	reputation	father
GP	M	15	U	LE3	T	4	4	health	services	course	father
GP	M	15	U	GT3	T	4	3	teacher	other	course	mother
GP	M	15	U	GT3	A	2	2	other	other	home	other
GP	F	16	U	GT3	T	4	4	health	other	home	mother
GP	F	16	U	GT3	T	4	4	services	services	reputation	mother
GP	F	16	U	GT3	T	3	3	other	other	reputation	mother
GP	M	17	U	GT3	T	3	2	services	services	course	mother
GP	M	16	U	LE3	T	4	3	health	other	home	father
GP	M	15	U	GT3	T	4	3	teacher	other	reputation	mother
GP	M	15	U	GT3	T	4	4	health	health	other	father
GP	M	16	U	LE3	T	4	2	teacher	other	course	mother
GP	M	16	U	LE3	T	2	2	other	other	reputation	mother
GP	F	15	R	GT3	T	2	4	services	health	course	mother

Figure 3.4: Representing training dataset

3.6.3 Applying Expert Rules on training dataset

From previous studies we have found that several attempts have been made to design and develop the data mining system that is generic but still no system is found that is completely generic. For this purpose in every domain the domain expert's assistant is compulsory. The domain experts shall be guided by the system to effectively apply their knowledge for the use of data mining systems to generate required knowledge. Expert's knowledge has been gathered from various experts and from that knowledge some rules have been developed in java language and MYSQL. To implement this java code Netbeans IDE 6.0 platform is used. Data collected for our study is in csv file format which is not directly supported by MYSQL. For this reason, to apply expert rules on this dataset the dataset is firstly converted into MYSQL required format by implementing java code. These rules are implemented so that accuracy of training dataset will increase.

The rules developed for our study are like:

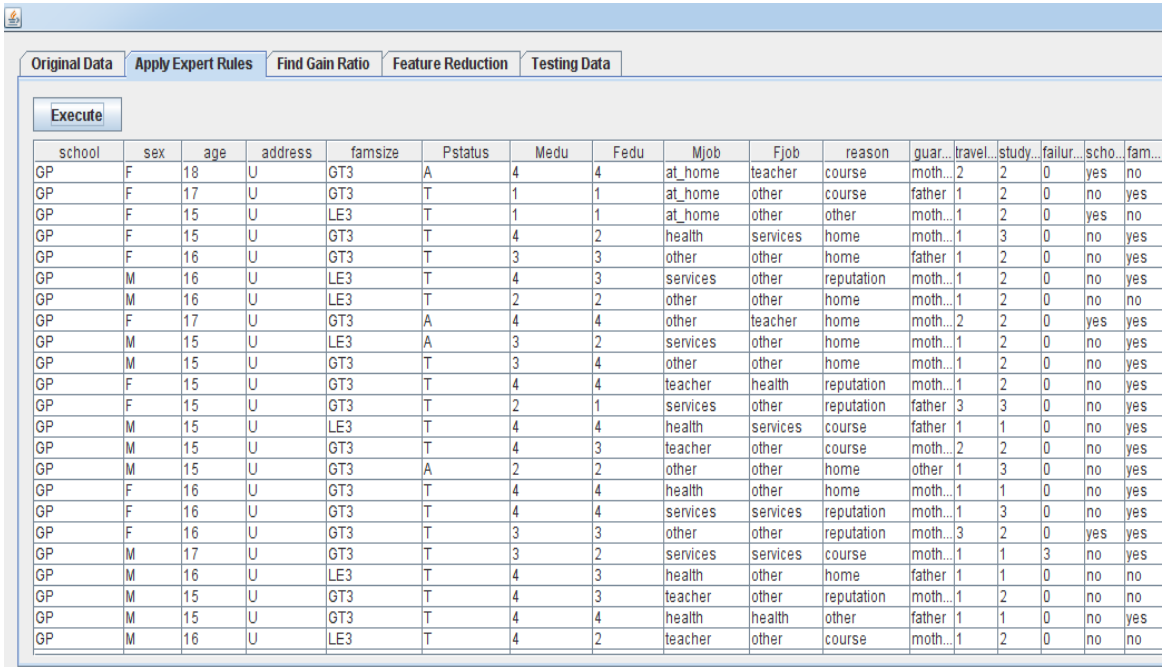
- if(G1)<9&&(traveltime)>2 && (studytime)<3&& famsup.equalsIgnoreCase ("NO") then pass='false' where G1<9 and traveltime>2 and studytime<3 and famsup='NO' '';

- if(G2)<9&&(traveltime)>2 && (studytime)<3 && famsup.equalsIgnoreCase ("NO") then pass='false' where G2 < 9 and traveltime >2 and studytime < 3 and famsup='NO' ";

These types of rules have been applied to the training dataset and then those rows have been identified from it whose attributes are more effective as well as relevant to the input parameters. After that this attributes are modified according to expert's rule.

After this step we get an optimal dataset which will increase the accuracy of the rules of classification.

The following figure represents dataset which we gain after applying expert rules on data.



school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guar...	travel...	study...	failur...	scho...	fam...
GP	F	18	U	GT3	A	4	4	at_home	teacher	course	moth...	2	2	0	yes	no
GP	F	17	U	GT3	T	1	1	at_home	other	course	father	1	2	0	no	yes
GP	F	15	U	LE3	T	1	1	at_home	other	other	moth...	1	2	0	yes	no
GP	F	15	U	GT3	T	4	2	health	services	home	moth...	1	3	0	no	yes
GP	F	16	U	GT3	T	3	3	other	other	home	father	1	2	0	no	yes
GP	M	16	U	LE3	T	4	3	services	other	reputation	moth...	1	2	0	no	yes
GP	M	16	U	LE3	T	2	2	other	other	home	moth...	1	2	0	no	no
GP	F	17	U	GT3	A	4	4	other	teacher	home	moth...	2	2	0	yes	yes
GP	M	15	U	LE3	A	3	2	services	other	home	moth...	1	2	0	no	yes
GP	M	15	U	GT3	T	3	4	other	other	home	moth...	1	2	0	no	yes
GP	F	15	U	GT3	T	4	4	teacher	health	reputation	moth...	1	2	0	no	yes
GP	F	15	U	GT3	T	2	1	services	other	reputation	father	3	3	0	no	yes
GP	M	15	U	LE3	T	4	4	health	services	course	father	1	1	0	no	yes
GP	M	15	U	GT3	T	4	3	teacher	other	course	moth...	2	2	0	no	yes
GP	M	15	U	GT3	A	2	2	other	other	home	other	1	3	0	no	yes
GP	F	16	U	GT3	T	4	4	health	other	home	moth...	1	1	0	no	yes
GP	F	16	U	GT3	T	4	4	services	services	reputation	moth...	1	3	0	no	yes
GP	F	16	U	GT3	T	3	3	other	other	reputation	moth...	3	2	0	yes	yes
GP	M	17	U	GT3	T	3	2	services	services	course	moth...	1	1	3	no	yes
GP	M	16	U	LE3	T	4	3	health	other	home	father	1	1	0	no	no
GP	M	15	U	GT3	T	4	3	teacher	other	reputation	moth...	1	2	0	no	no
GP	M	15	U	GT3	T	4	4	health	health	other	father	1	1	0	no	yes
GP	M	16	U	LE3	T	4	2	teacher	other	course	moth...	1	2	0	no	no

Figure 3.5: DataSet after implemention of Expert Eules

3.6.4 Feature Reduction

Feature reduction is a method in data mining which identify and remove those attributes that do not take part in the classification of the dataset. In our work, gain ratio technique is used to evaluate the value of any attribute with respect to its class. In our research total 6 attributes are selected by algorithm out of 33 that are more effective.

Then rank search algorithm is used in our study to arrange these attributes. The arranging of these attributes is done in descending order according to their value of gain ratio and last attributes having lowest gain ratio are removed.

After applying gain ratio and Rankers algorithm on dataset ranked attributes left are:

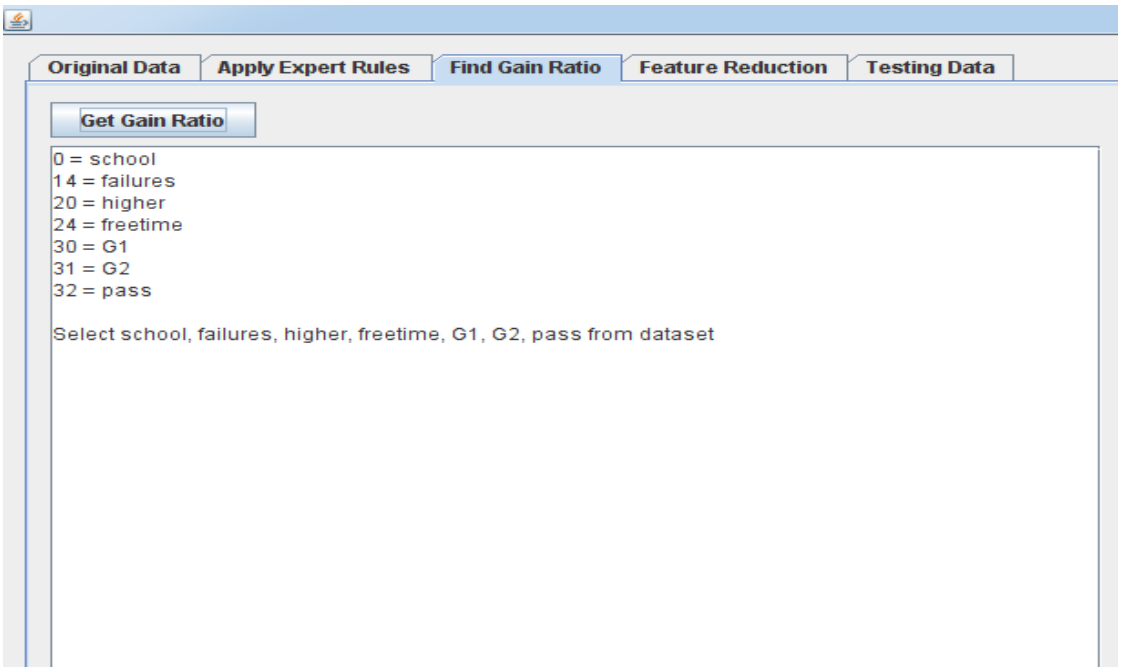
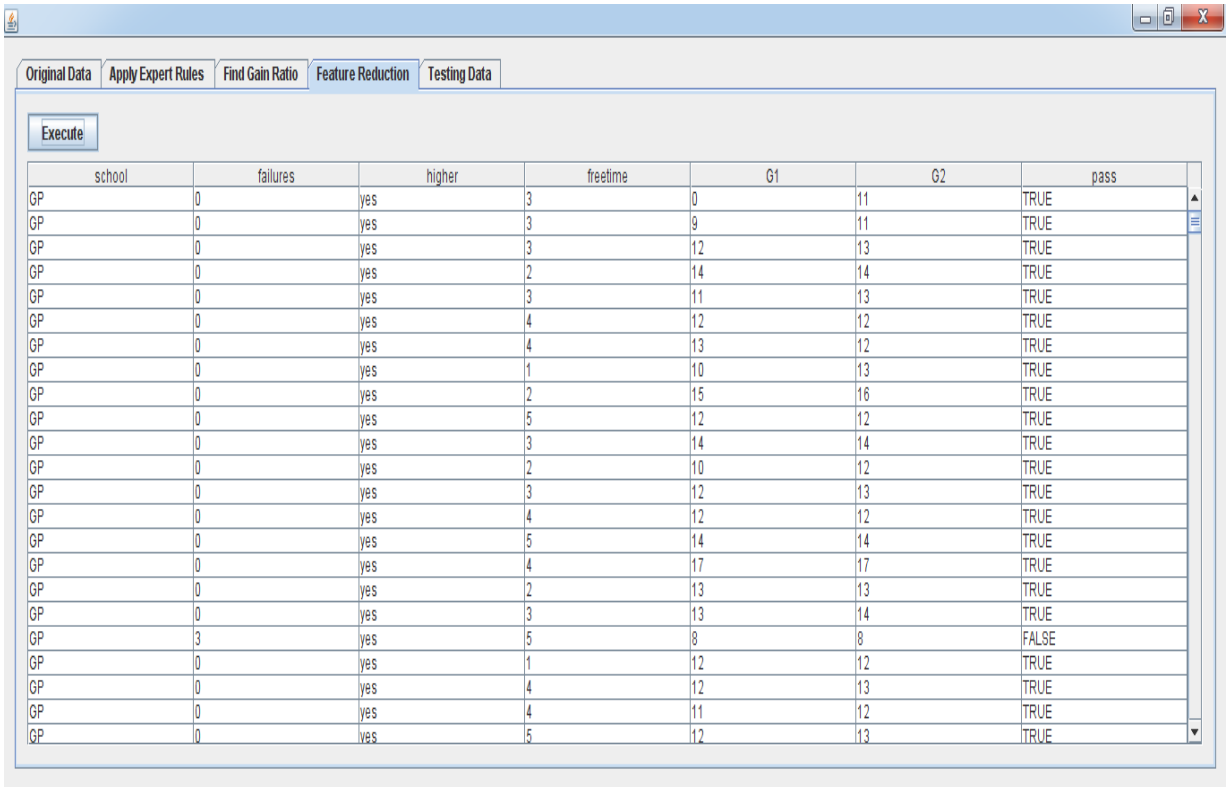


Figure 3.6: Figure represents Gain Ratio of all the attributes

From the above shown figure, the attributes school has maximum gain ratio which means it will affect the performance of students at highest mark of value. Attribute failures is at second position of series, higher, free time, G1, G2 are the total effective parameters. Other attributes with less gain ratio will be removed from the dataset. It has been assumed that the factors with minimum gain ratio do not take part in data classification which

means results we gain after implementing C4.5 algorithm will not get affected but this will result in less computational time with less memory requirement.



school	failures	higher	freetime	G1	G2	pass
GP	0	yes	3	0	11	TRUE
GP	0	yes	3	9	11	TRUE
GP	0	yes	3	12	13	TRUE
GP	0	yes	2	14	14	TRUE
GP	0	yes	3	11	13	TRUE
GP	0	yes	4	12	12	TRUE
GP	0	yes	4	13	12	TRUE
GP	0	yes	1	10	13	TRUE
GP	0	yes	2	15	16	TRUE
GP	0	yes	5	12	12	TRUE
GP	0	yes	3	14	14	TRUE
GP	0	yes	2	10	12	TRUE
GP	0	yes	3	12	13	TRUE
GP	0	yes	4	12	12	TRUE
GP	0	yes	5	14	14	TRUE
GP	0	yes	4	17	17	TRUE
GP	0	yes	2	13	13	TRUE
GP	0	yes	3	13	14	TRUE
GP	3	yes	5	8	8	FALSE
GP	0	yes	1	12	12	TRUE
GP	0	yes	4	12	13	TRUE
GP	0	yes	4	11	12	TRUE
GP	0	yes	5	12	13	TRUE

Figure 3.7: Representing Optimal dataset

CHAPTER 4
RESULT AND DISCUSSION

In our study classifiers that are provided in WEKA software are used for Automatic Evaluation of performance of students. C4.5 algorithm is applied to an optimal dataset that constructs an optimal decision tree. This tree will results in minimum error rate, low consumption of memory and less computational time as compared to the existing approach. From the student's point of view, the classification rules which are resulted will be more realistic and gives more proceedings to the students about their career. The comparison is made between existing approach and our approach by this algorithm on the basis of computation time, memory and accuracy.

The training dataset has been prepared by randomly selecting 650 instances from the actual dataset. The training dataset which is collected from students is graphically represented in WEKA 3.4 by using histograms. The view of training dataset in WEKA 3.4 is as shown below:

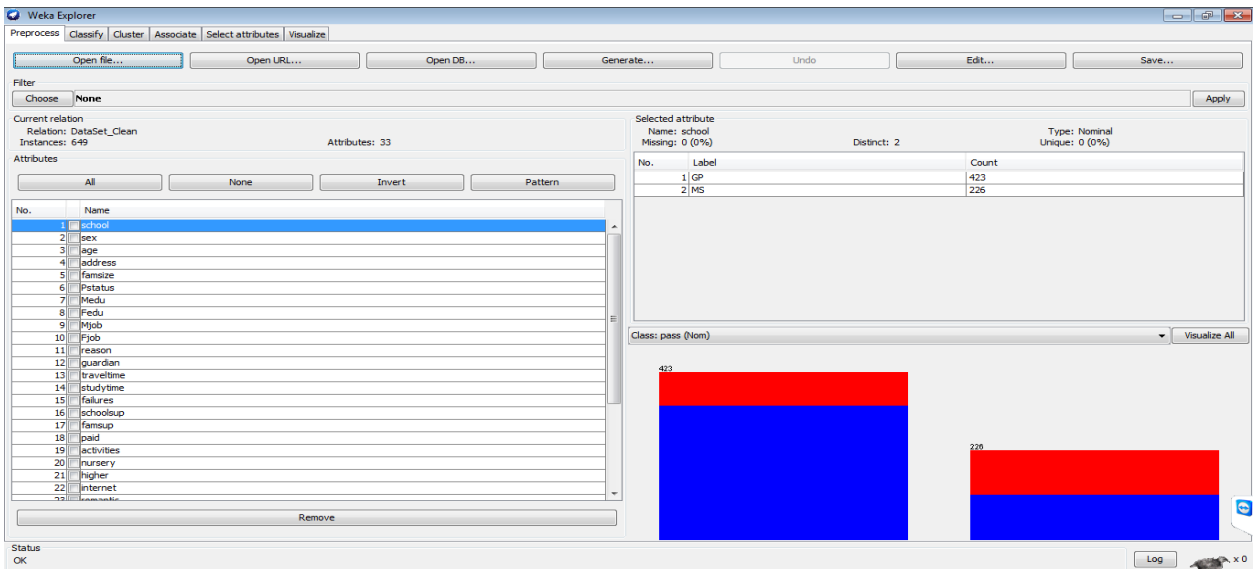


Figure 4.1: Representing training dataset in WEKA 3.4

It has been observed from analysis that most of the students prefer that subjects and coaching which they think are easy. But, this analysis may not be true because when students do not have much awareness about latest trends then they prefer to follow others.

In the next step of study, the C4.5 algorithm has been applied to this training dataset. The accuracy of this training dataset has been measured and the output of it is shown below.

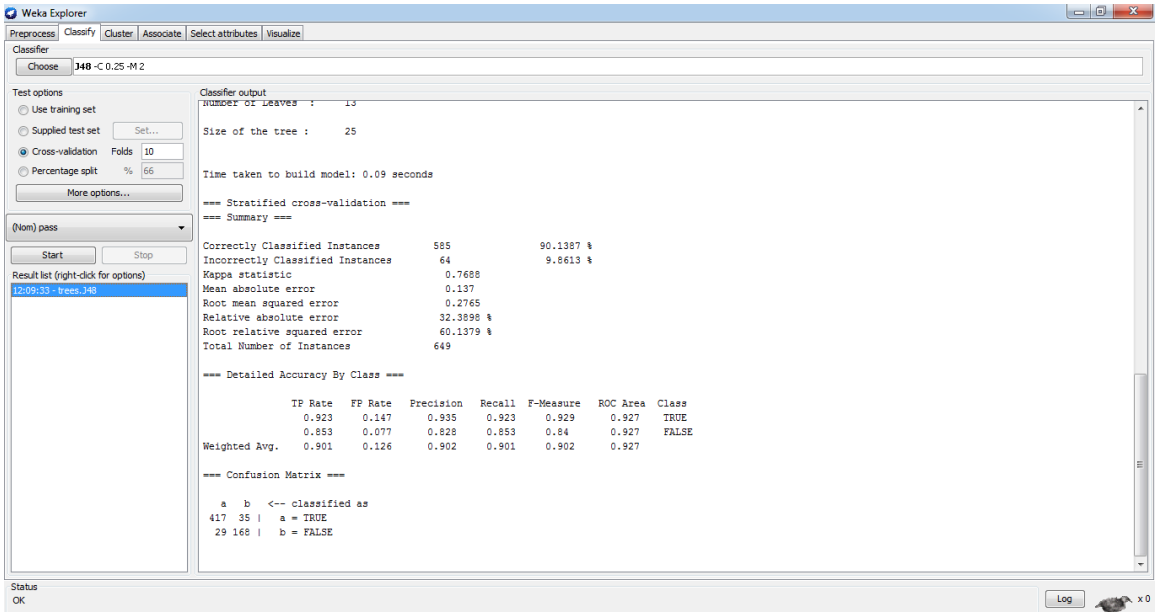


Figure 4.2: Representing accuracy of Existing Approach in WEKA tool

It has been observed that, this training dataset has 90.1387% correctly classified data and 9.8613% data is incorrectly classified.

In our work, training dataset is converted into an optimal dataset which is then fetched into WEKA 3.4 machine learning tool.

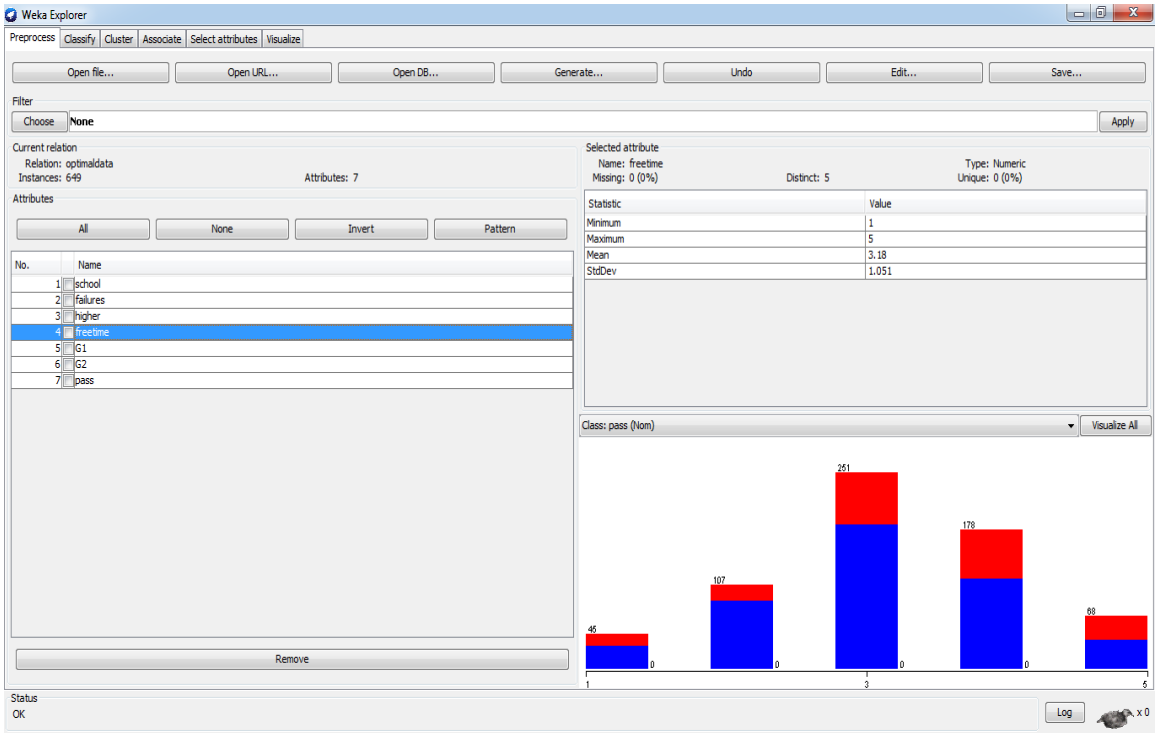


Figure 4.3: Representing view of an optimal dataset in WEKA 3.4

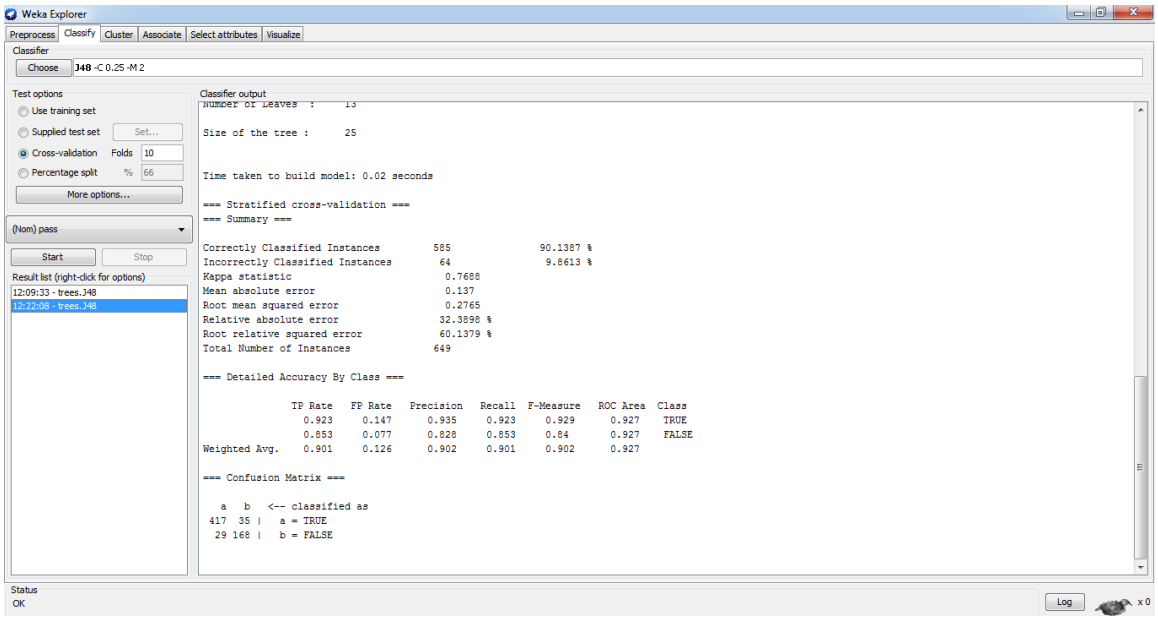


Figure 4.4: Representing classification of an optimal dataset in WEKA 3.4

This above figure shows us the view of an optimal dataset. After applying expert rules[27], less effective attributes are replaced by the more effective attributes according to expert's opinions. Further, gain ratio of all the attributes is calculated and it has been

observed that attributes having lowest gain ratio are removed from the dataset to make it an optimal dataset. This work is done on the assumption that, the attributes with lowest gain ratio does not contribute towards construction of decision trees.

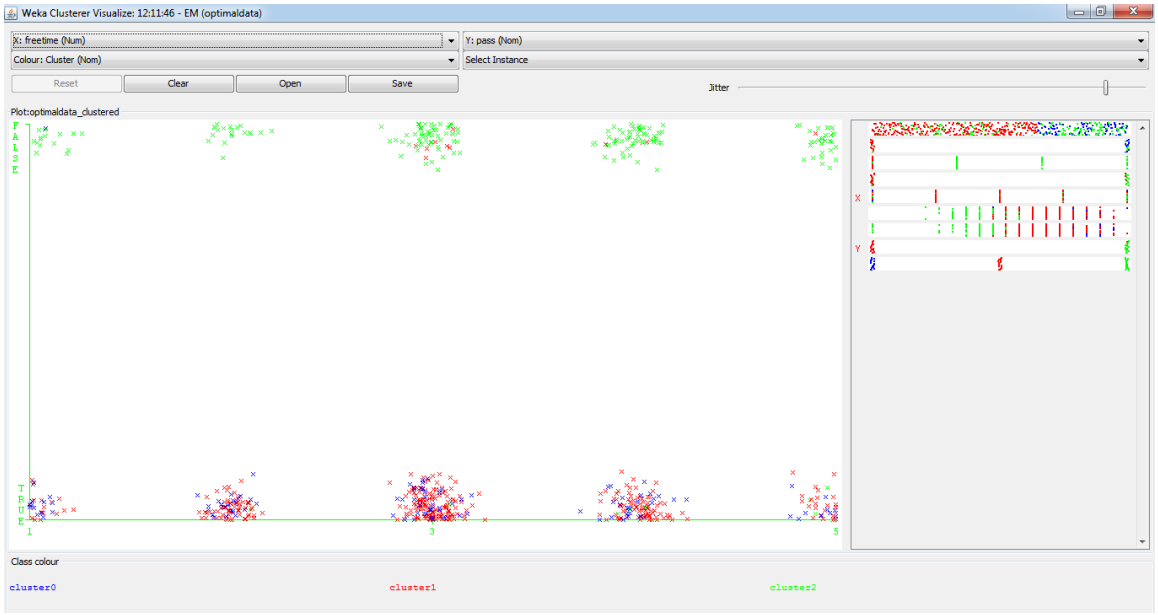


Figure 4.5: Shows the cluster size of all attributes of optimal dataset

This following discussed approach is implemented in Netbeans by using with java language.

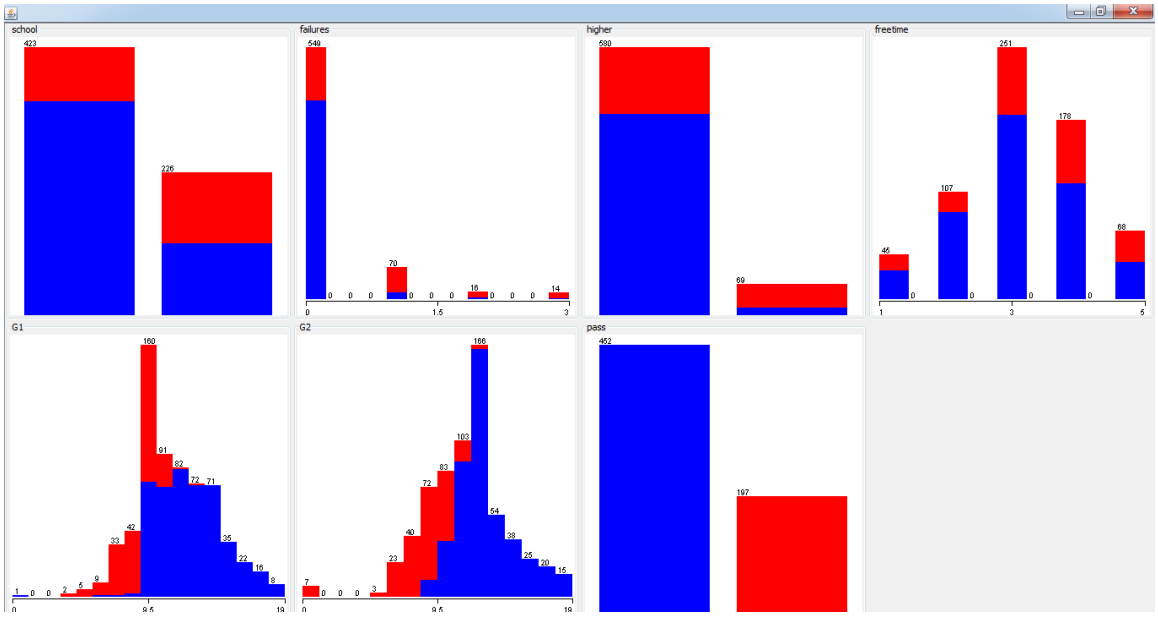


Figure 4.6: Representing all attributes showing data according to class attribute

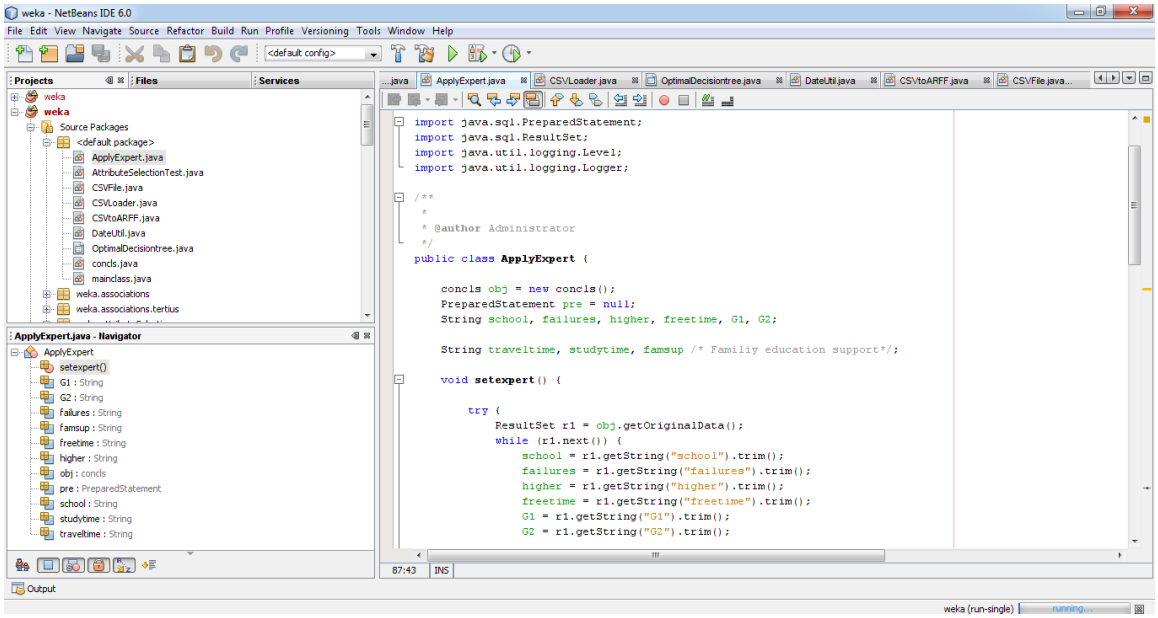


Figure 4.7: Representing view of coding implemented in java language

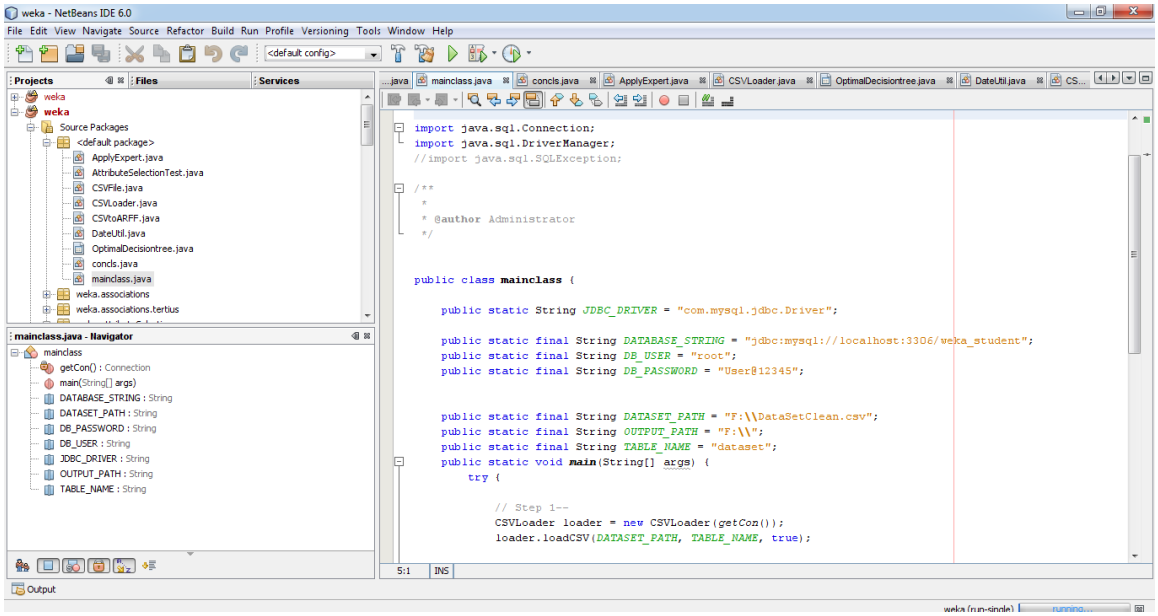


Figure 4.8: Representing view of main class coding implemented in java language

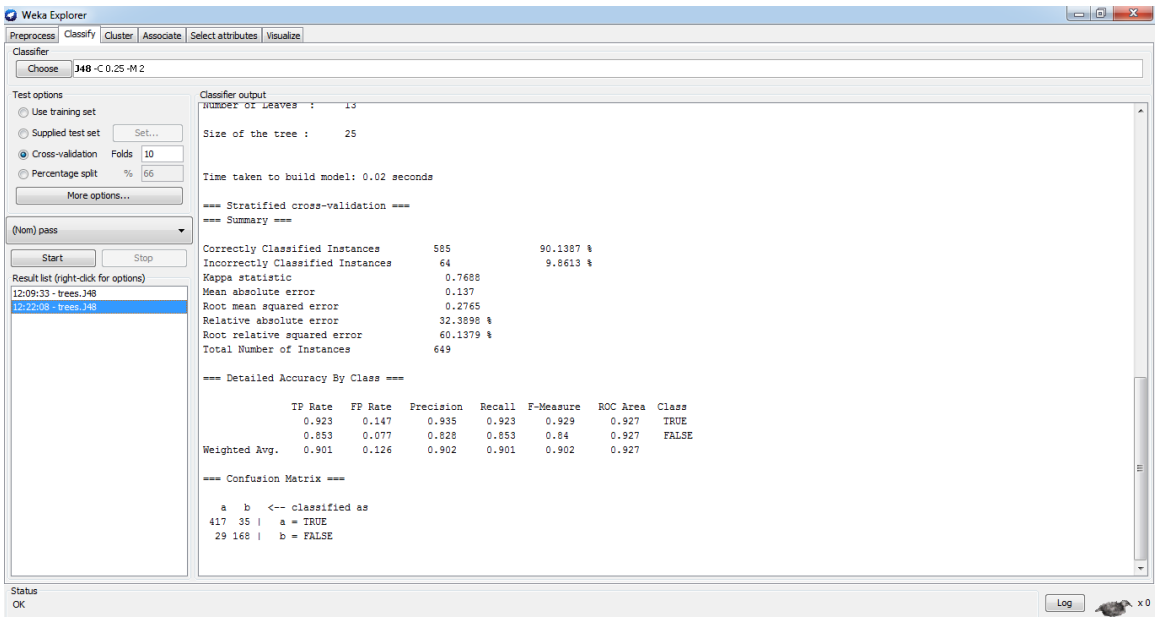


Figure 4.9: Representing accuracy of an optimal Approach on same dataset.

When C4.5 algorithm is run on an optimal dataset which is generated by applying expert rules and feature reduction technique, the percentage of accuracy has been increased as

well as time reduced. It has been observed that, correctly classified instances in this approach is 90.1387% and 9.8613% instances are in correctly classified.

Computational time has also been reduced by using this approach that is 0.09 sec to 0.02 sec. Above analysis shows the decrease in computation time by applying C4.5 algorithm on datasets. The implemented technique has been applied to the dataset of 650 rows to make it an optimal dataset and then comparison has been done between computational time of existing and our technique.

Figures given below show the execution time of an existing approach and our optimal approach. It has been observed that execution time of an existing approach is 0.09 seconds and for our approach is 0.02 seconds thus decreasing in 77% of computation time.

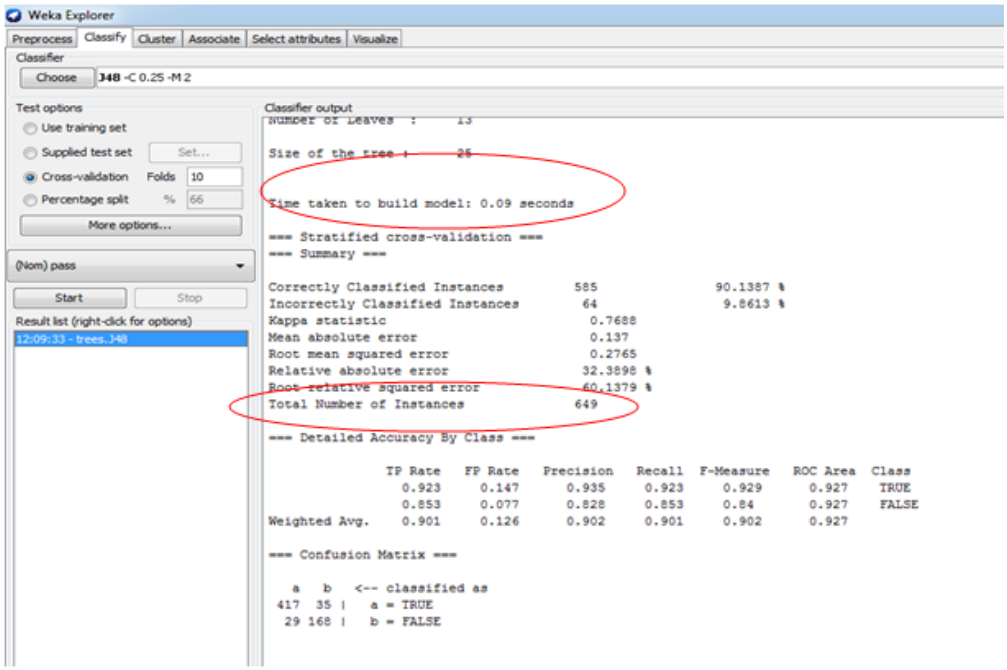


Figure 4.10: Representing Speed of an Existing Approach on Large Dataset.

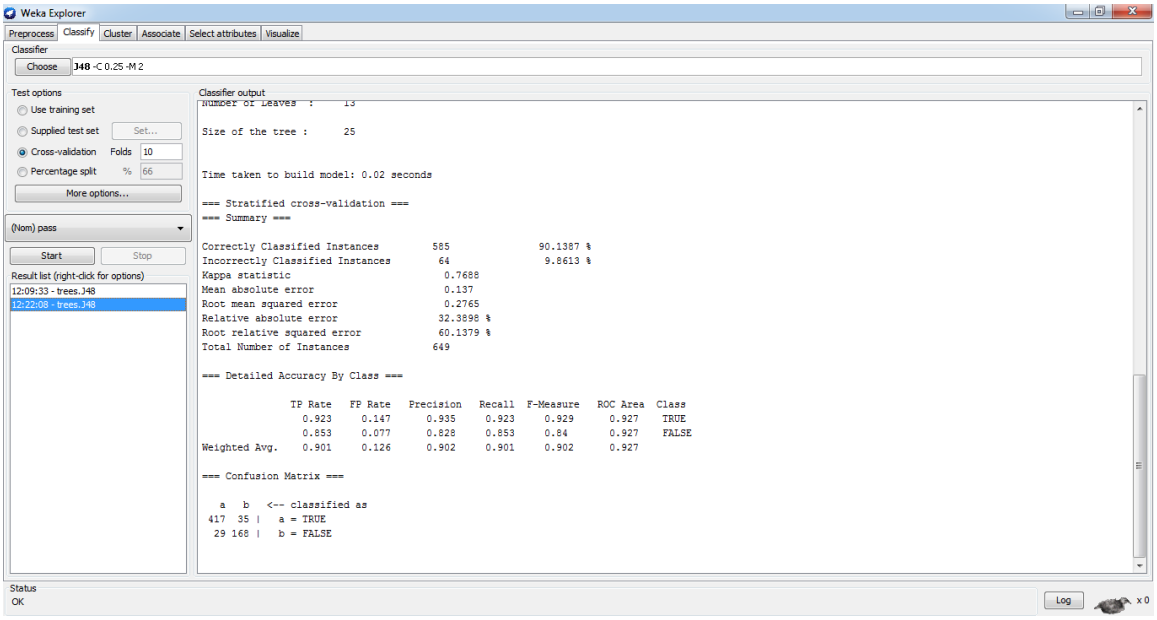


Figure 4.11: Representing Speed of an Optimal Aproach on Large Dataset.

The time comparison is also shown below by a bar graph.

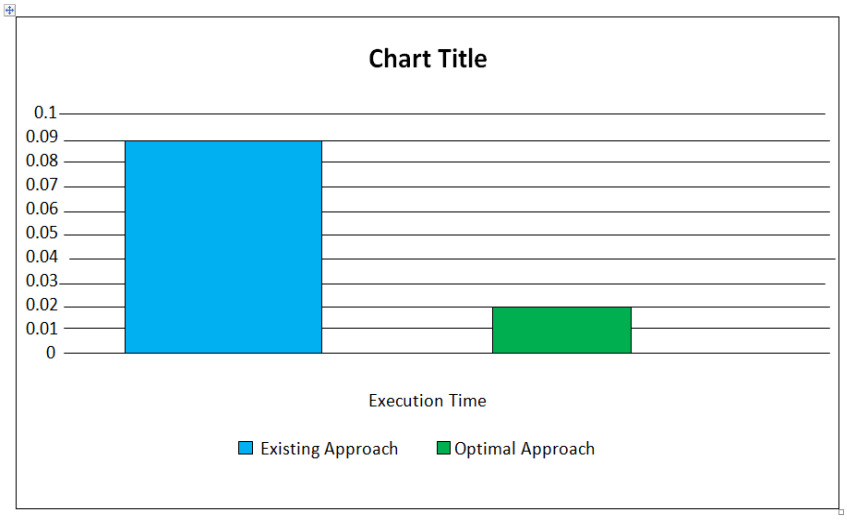


Figure 4.12: Representing time comparison between Existing Approach and Optimal Approach graphically

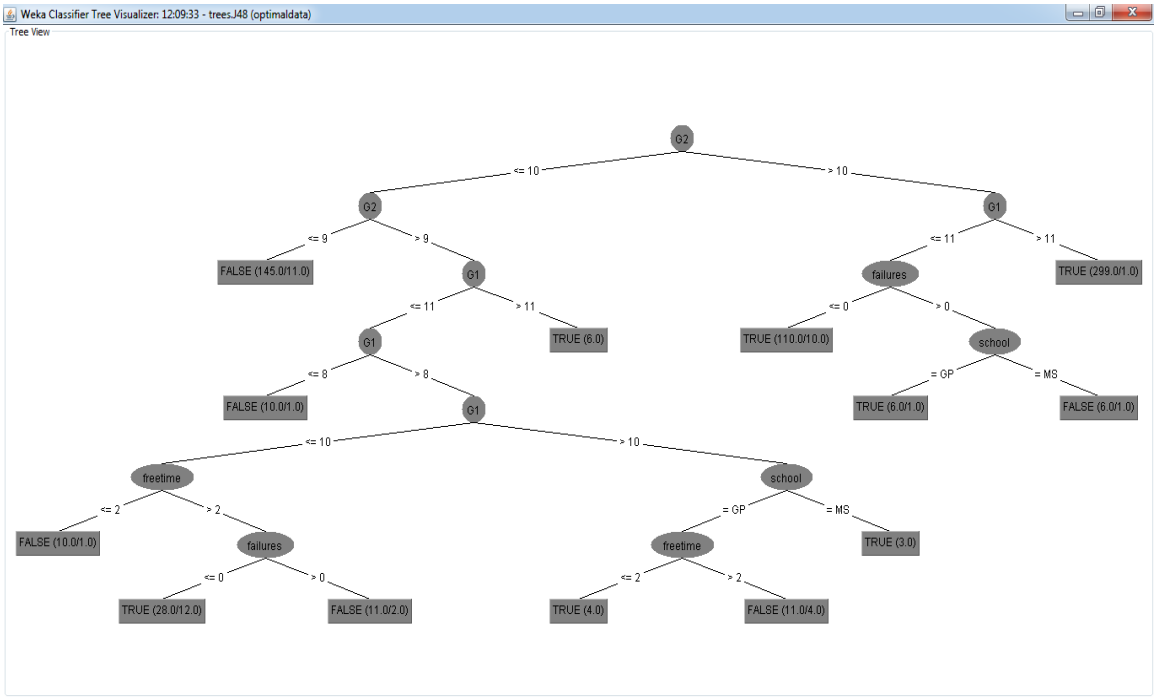


Figure 4.13 : Representing Decision tree constructed from optimal dataset

After conversion of training dataset to an optimal dataset, C4.5 algorithm has been run on this optimal dataset and finally construction of decision tree has been done. Classification rules have been generated from this tree. These classification rules will help to increase the performance of students based upon some input parameters.

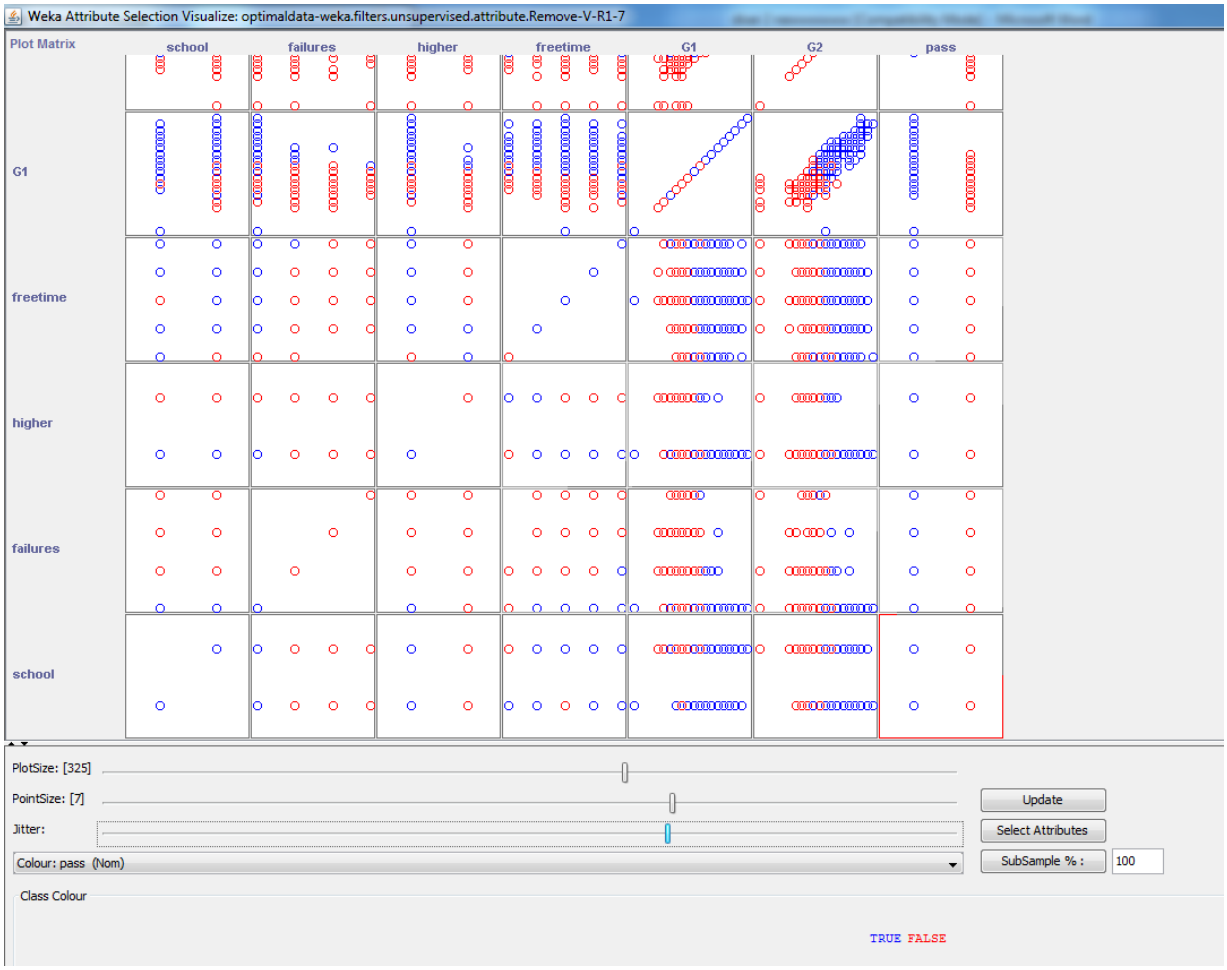


Figure 4.14 : Representing visualization of 6 attributes after removal

The main objective of our research is to finding the most effective attributes and ignoring that attributes that have least effect on performance of students.

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

The optimal algorithm is based on the C4.5 methods. This method introduces new characteristics such as implementation of Expert rules and technique of feature reduction on large dataset. Several attempts have been made to design and develop the specific data mining system but no system is found completely specific or generic. Thus the domain expert's support is compulsory for every domain. The domain experts shall be guided by the system to effectively apply their knowledge for the use of data mining systems to create required knowledge. The domain experts are required because they determine the category of data that should be composed of in the specific problem domain, selection of particular type of data for data mining, cleaning of data, transformation of data, extracting various patterns for generation of knowledge and finally interpretation of the patterns and knowledge generation. By applying this optimal approach on student's data, we can find out the best features or attributes that will influence student's performance.

Along with that, supervised learning has been implemented. A (GUI) Graphical user interface has been provided to the user to input various types of parameters. Based upon these input parameters an output is showing best attributes for the students. This new approach is better than an existing approach. The accuracy of performance has been improved by 77%. Computational time as well as memory requirements has also been reduced by this.

5.2 Future Scope:

This approach is able to handle variable data which makes it acceptable for many other applications. This research is not bounded to specific type of area. In this work, the optimal technique is applied to improve performance of weak students to their best level. But it can be also used for other applications like Human talent management, Analysis of education patterns, risk evaluation etc.

CHAPTER 6

REFERENCES

[1] Jiawei Han, Micheline Kamber, Jian Pei, "DATA MINING- Concepts and Techniques", Morgan Kaufmann

[2] Hany M. Harb,Malaka A. Moustafa, "Selecting Optimal Subset of Features of Student Performance Model", IJCSI International Journal of Computer Science Issue, Vol. 9, Issue 5, No, September 2012.

[3] Carlos Márquez-Vera, Cristóbal Romero Morales, and Sebastián Ventura Soto "Predicting School Failure and Dropout by Using Data Mining Techniques" IEEE Journal Of Latin-American Learning Technologies, Vol. 8, No. 1, February, 201, pp. 7-14

[4]Kiri Wagstaff, Claire Cardie "Constrained K-means Clustering with Background Knowledge" Proceedings of eighteenth international conference on machine learning (2001) pp. 577-584.

[5] Grigorios F. Tzortzis and Aristidis C. Likas, *Senior Member, IEEE* "The Global Kernel K-Means Algorithm for Clustering in Feature Space" IEEE transactions on neural networks, VOL. 20, NO. 7, JULY 2009, pp.1181-1194.

[6] K.A Abdul Nazeer, M.P Singh "Improving the accuracy and efficiency of k means, kohenon self organizing map and hierarchical agglomerative clustering*". Proceedings of world congress on engineering*. Volume 1, London u.k, (2002)

[7]Pizzuti Clara and Talia Domenico "P-Auto class: Scalable parallel clustering for mining large data set". IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 15, NO. 3, MAY/JUNE 2003, Pp. 629-641.

[8]Saadat Naziova “Survey on Spam Filtering Techniques”, Communication and Network, 2011, 3, 153-160, doi: 10.4236/cn.2011.33020 Published online August 2011

[9] P. Moniza and P. Asha “An Assortment of Spam Detection System”, International Conference on Computing, Electronics and Electrical Technologies [ICCEET] 2012

[10] Patricia Bellin Ribeiro, Luis Alexandre da Silva, Kelton Augusto Pontara da Costa “Spam Intrusion Detection in Computer Networks Using Intelligent Techniques”, IFIP IEEE IM Workshop: 1st International Workshop on security for Emerging Distributed Network Technologies (DISSECT) 2015

[11] Ji Dan, Qiu Jianlin, 2010 10th IEEE International Conference on Computer and Information Technology (CIT 2010).

[12] IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO. 11, NOVEMBER 2009

[13] Qiang Yang, Senior Member, IEEE, Jie Yin, Charles Ling, and Rong Pan, “Extracting Actionable Knowledge from Decision Trees” IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 20, NO. 1, JANUARY 2007.

[14] Souptik Datta, Chris R. Giannella, and Hillol Kargupta, Senior Member, IEEE IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO. 10, Approximate Distributed K-Means Clustering over a Peer-to-Peer Network, OCTOBER 2009

[15] Jasna Soldic-Aleksic , Journal of Economics and Engineering, ISSN.: 2078-0346, Vol. 3. No.1, April 2012.

[16] Comparative Study of Different Data Mining Techniques Performance in knowledge Discovery from Medical Database Volume 3, Issue 3, ISSN: 2277 128X, March 2013

[17] I.J. Modern Education and Computer Science, 2015, 5, Published Online May 2015 in MECS, pp- 43-49

[18] Educational Data Mining for Prediction and Classification of Engineering Students Achievement IEEE 7th International Conference on Engineering Education (ICEED) 2015

[19] Educational Data Mining Techniques and their Applications IEEE International Conference on Green Computing and Internet of Things (ICGCloT) 2015

[20] Data Mining: Concepts and Techniques Second Edition Jiawei Han *University of Illinois at Urbana-Champaign* Micheline Kamber

[21] https://www.tutorialspoint.com/mysql/mysql-introduction.htm

[22] S. Saravana Kumar in his article "An Analysis of Investor Preference Towards Equity and Derivatives" published in The Indian journal of commerce, July-September 2010

[23] "NetBeans IDE 6.0 - New Core Features in Depth." Online available: https://netbeans.org/community/magazine/html/03/nb06/. [Accessed: 06-Apr-2016].

[24] I. Russell, "An Introduction to the WEKA Data Mining System."

[25] "Weka 3 - Data Mining with Open Source Machine Learning Software in Java." Online available- http://www.cs.waikato.ac.nz/ml/weka/documentation.html. [Accessed: 26-Mar-2016].

[26] John Jacob, Kavya Jha, Paarth Kotak, Shubha Puthran 'Educational Data Mining Techniques and their Applications' International Conference on Green Computing and Internet of Things (ICGCloT), 2015, pp. 1344-1348

[27] "Data Mining Concepts." [Online]. Available: https://technet.microsoft.com/en-us/library/ms174949.aspx. [Accessed: 21-Jan-2016].

[28] Ashish Dutt, Saeed Aghabozrgi, Maizatul Akmal Binti Ismail, and Hamidreza Mahroeian 'Clustering Algorithms Applied in Educational Data Mining' International Journal of Information and Electronics Engineering, 2015, pp. 112-116

[29] Md. Fahim Sikder, Md. Jamal Uddin and Sajal Halder 'Predicting student's yearly performance using Neural Network' 5th International Conference on Informatics, Electronics and Vision (ICIEV), 2016 pp. 524-529

[30] Parneet Kaura,Manpreet Singhb,Gurpreet Singh Josanc(ICRTC) 'Classification and prediction based data mining algorithms to predict slow learners in education sector' 3rd International Conference on Recent Trends in Computing, 2015, pp. 500-508

[31] Hind Almayan, Waheeda Al Mayyan, 'Improving Accuracy of Students' Final Grade Prediction Model Using PSO' 6th International Conference on Information Communication and Management, 2016 pp. 35-39

[32] Norlida Buniyamin, Usamah bin Mat, Pauziah Mohd Arshad 'Educational Data Mining for Prediction and Classification of Engineering Students Achievement' IEEE 7th International Conference on Engineering Education (ICEED), 2015, pp. 49-53