INDIA'S LARGEST UNIVERSITY

LOVELY PROFESSIONAL UNIVERSITY

Transforming Education, Transforming India

TOPIC APPROVAL PERFORMA

School of Computer Science and Engineering

Program : P172::M.Tech. (Computer Science and Engineering) [Full Time]

COURSE CODE : CSE548 **REGULAR/BACKLOG :** Regular **GROUP NUMBER :** CSERGD0053

Supervisor Name : Nitin Umesh **UID :** 15857 **Designation :** Assistant Professor

Qualification : _____ **Research Experience :** _____

SR.NO.	NAME OF STUDENT	REGISTRATION NO	BATCH	SECTION	CONTACT NUMBER
1	Amandeep Kaur	11604922	2016	K1637	9646709044

SPECIALIZATION AREA : Database Systems **Supervisor Signature:** _____

PROPOSED TOPIC : machine learning approach to predict student academic performance and to provide recommendations on the basis of the prediction results

Qualitative Assessment of Proposed Topic by PAC		
Sr.No.	**Parameter**	**Rating (out of 10)**
1	Project Novelty: Potential of the project to create new knowledge	6.00
2	Project Feasibility: Project can be timely carried out in-house with low-cost and available resources in the University by the students.	6.67
3	Project Academic Inputs: Project topic is relevant and makes extensive use of academic inputs in UG program and serves as a culminating effort for core study area of the degree program.	6.00
4	Project Supervision: Project supervisor's is technically competent to guide students, resolve any issues, and impart necessary skills.	7.00
5	Social Applicability: Project work intends to solve a practical problem.	7.00
6	Future Scope: Project has potential to become basis of future research work, publication or patent.	6.67

PAC Committee Members		
PAC Member 1 Name: Kewal Krishan	UID: 11179	Recommended (Y/N): Yes
PAC Member 2 Name: Raj Karan Singh	UID: 14307	Recommended (Y/N): NA
PAC Member 3 Name: Sawal Tandon	UID: 14770	Recommended (Y/N): NA
PAC Member 4 Name: Dr. Pooja Gupta	UID: 19580	Recommended (Y/N): Yes
PAC Member 5 Name: Kamlesh Lakhwani	UID: 20980	Recommended (Y/N): NA
PAC Member 6 Name: Dr.Priyanka Chawla	UID: 22046	Recommended (Y/N): NO
DAA Nominee Name: Kuldeep Kumar Kushwaha	UID: 17118	Recommended (Y/N): NA

Final Topic Approved by PAC: **Machine Learning approach to predict student academic performance and to provide recommendations on the basis of the prediction results**

Overall Remarks: Approved (with minor changes)

PAC CHAIRPERSON Name: 11024::Amandeep Nagpal **Approval Date:** 04 Nov 2017

ABSTRACT

Nowadays, Data mining in the field of education is utmost important to do predict the performance of student. The reason being is that it's competition everywhere to achieve goals as well as to be smart in taking decisions. Education is the Power and by predicting performance in education by considering relevant parameters we would be able to work on the weaknesses of student at right time by using right pedagogies and approaches. In this way we would be able to build some constructive thinkers and competitors. Data Mining is a process of extracting knowledge from large amount of data. 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 would discuss how a hybrid approach based data mining model can help to improve an education system by providing more accuracy in results as compare to traditional approaches as well as how it would enable better and effective teacher-student interaction. This research would also help the other departments related to education such as research and placements because by enhancing academic performance we would be able to develop critical thinkers.

CERTIFICATE

This is to certify that AMANDEEP KAUR has completed M.Tech dissertation proposal titled "MACHINE LEARNING APPROACH TO PREDICT AND IMPROVE STUDENT ACADEMIC PERFORMANCE" under my guidance and supervision. To the best of my knowledge, the present work is the result of h i s / her original investigation and study. No part of the dissertation has ever been submitted for any other degree or diploma.

Date: **Signature of Advisor**

Name:

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. NITIN UMESH** (Assistant professor, Department of Computer science, Lovely Professional University, Phagwara). I am very much thankful to **Mr. NITIN UMESH** 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.

AMANDEEP KAUR

DECLARATION

I hereby declare that the dissertation proposal entitled, "MACHINE LEARNING APPROACH TO PREDICT AND IMPROVE STUDENT ACADEMIC PERFORMANCE" submitted for the M.Tech Degree is entirely my original work and all ideas and references have been duly acknowledged. It does not contain any work for the award of any other degree or diploma.

Date:

Investigator

Regn. No. 11604922

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 DATA MINING

Data Mining is the process to find the hidden information as well as pattern from a bulk amount of data i.e. the data should be coming from different sources such as data warehouse, data mart etc. Data is hidden taken out through techniques of data mining. It gives imperative information that is important to take appropriate decisions. Pre-processing of information contain information cleaning to decrease noise, pertinence investigation to eliminate unessential qualities, forecast accuracy, scalability and interpretability.

Information digging method is useful for a few reasons in private and additionally open areas. Numerous Enterprises utilize Information Mining strategy to remove the significant data from the extensive database to limit costs, upgrade research, and increment deals i.e. saving money, pharmaceutical, insurance, retailing and Educational Information Mining. By the increase of technology of computers the collection of data, storage of data as well as change of data have turned out to be straight forward. There is exchange off between size of information and execution time. In the event that the extent of dataset is large then execution is consequently diminished.

Data mining is process of extracting knowledge from large amount of data. The main reason for what data mining algorithms are used is that it gather relevant information which provides us better outcomes. Information mining apparatus is utilized to discover questions and relations between them. This technique incorporate measurable and additionally numerical model. Data mining process is performed on gathered data which is represented in different forms like text form, web, image processing and visuals. The very important step is to find knowledge from data by using Knowledge discovery process. It includes various steps for extracting meaningful data. Data mining is concerned with more than one areas such as database management system, statistics,

visualization etc. It merges techniques from so many fields such as image processing, ecommerce, retail, pattern mining etc. Data mining system consist of operational units for tasks such as association, classification, prediction and clustering data analysis.

Data Mining is extensively useful in Educational Data Mining. EDM is a rising field for knowledge learning finding from vast measure 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 angles may either help an understudy in exceeding expectations during scholarly period or end scholastic program of an understudy. Such disappointment is known as drop-out.

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.

Classification: The approach of classification includes mining processes suggest discovering rules on the premise of sub forms fabricate. It include straight out qualities like discrete, unordered. It include approaches like k nearest neighbor classifier, case based learning. It is utilized for fraud detection and medical diagnosis. Classification is done by utilizing k means algorithm, genetic algorithms. We can utilize group habitats for data classification such that the computational load is less and the impact of noisy data is lessened.

Data compression: We can use cluster center to show the actual dataset. The numbers of clusters are not as much as the extent 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 figure estimation of predicted attribute for new data.

1.2. 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 behavior of new objects. For instance Naive Bayes, SVM, Decision Tree, KNN etc.

Clustering algorithm: It is collection of objects of similar type in one group. The cluster provides us better results. Clustering analysis has been a developing exploration issue in information mining due its assortment of uses. For instance K-means clustering, DBSCAN etc.

Machine Learning: Both data mining and machine learning used same methods. But there is difference, machine learning focused on prediction, based on known properties, whereas data mining focuses on identification of unknown properties.

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 vast datasets. In clustering partition the dataset in to subsets for optimization.

1.3 KNOWLEDGE DISCOVERY PROCESS

Data mining is a procedure of eliciting or mining knowledge from enormous amount of data. It means knowledge extraction, knowledge mining of data, pattern analysis and data knowledge discover from data. It is the process of discovering required knowledge from database. It includes various operations such as selection, processing, transformation, interpretation and evaluation. Knowledge Discovery Process is abbreviated as KDD.

There are various steps to discover knowledge. It selects a dataset or its subset. It removes noise from data.

Data cleaning: It is the process of removing noise and inconsistent data. It can fill missing values. It is a first step in which dirty data and inconsistent facts or data are eliminated or discarded.

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, information is change into forms fitting for mining. It can include different advances:-

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 purpose of aggregation 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 procedure of taking out knowledge from big data repositories or databases. It can evaluate results in form of patterns. The large amount of knowledge is collected from different knowledge engineers.

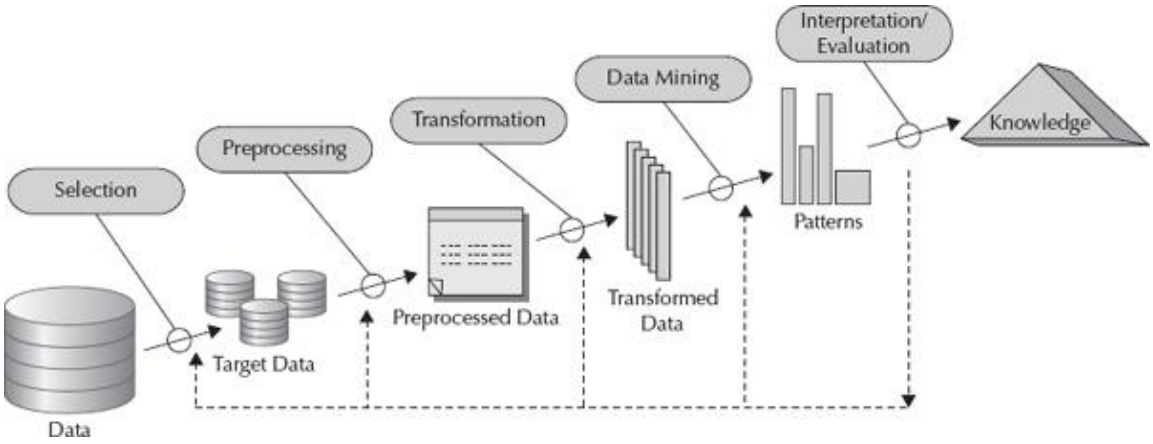

Figure 1.1: Knowledge Discovery Process

1.4 DATA MINING TECHNIQUES

1.4.1 Classification

The classification is done because of exactly guess the aimed class for all case in the data. One of the example of this model is it is help to predict the student performance. There are two stages in classification. The initial part is the learning process. In this part, the training data or facts are examined by classification algorithm and rules and design are created which are based on learned model or classifier. In the second part the model is used for classification and testing data are used for gaining the accuracy of classification design. Then, establish on the sufficient accuracy, the rules can be used for the classification of new or recently developed data or for unseen data.

1.4.2 Decision Tree

It is broadly utilized technique in data mining. 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 gain 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:

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

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

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

Finally the information Gain is evaluated as:

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

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

1.4.3 ID3 and C4.5: These are the algorithmic approaches developed by Ross Quinlan for inducing Classification Models from data that are also called decision trees. ID3 applies a top-down, greedy search approach through the space of possible branches with no backtracking. It uses entropy and Information Gain to construct a decision tree. Likewise, C4.5 also uses the concept of entropy and gain to build the decision trees.

Tree-shaped structures address sets of decisions. These decisions make rules for the request of a dataset. Particular decision tree strategies merge different approaches such as Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). It separates a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The last outcome is a tree with decision nodes and leaf nodes. Here in the following figure, a decision node for instance "Outlook" has two or more branches such as "Sunny", "Overcast" and "Rainy". Leaf node for instance "Play" represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. A decision tree can easily be transformed to a set of rules by mapping from the root node to the leaf nodes one by one.

Figure 1.2: Decision Tree Classification

R_1: IF (Outlook=Sunny) AND (Windy=FALSE) THEN Play=Yes

R_2: IF (Outlook=Sunny) AND (Windy=TRUE) THEN Play=No

R_3: IF (Outlook=Overcast) THEN Play=Yes

R_4: IF (Outlook=Rainy) AND (Humidity=High) THEN Play=No

R_5: IF (Outlook=Rain) AND (Humidity=Normal) THEN Play=Yes

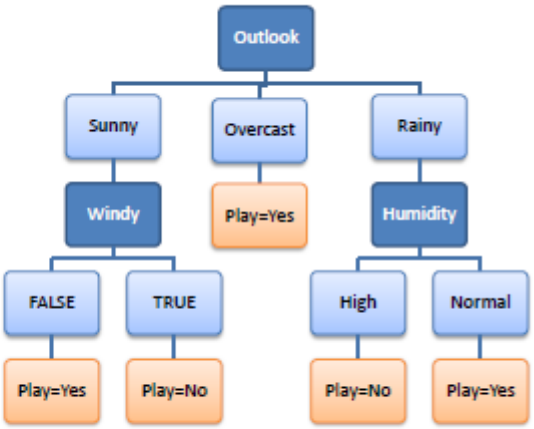

Figure 1.3: Decision Tree Rules

1.4.4 Fuzzy logic: It is a method to determine the "degree of facts" instead of the general "true or false" (1 or 0).Data mining uses different methods and assumption from a broad areas or fields for the knowledge extraction from huge amount of data. But uncertainty is a general phenomenon in data mining problems. Therefore, it is applied to manage with the uncertainty in actual world.

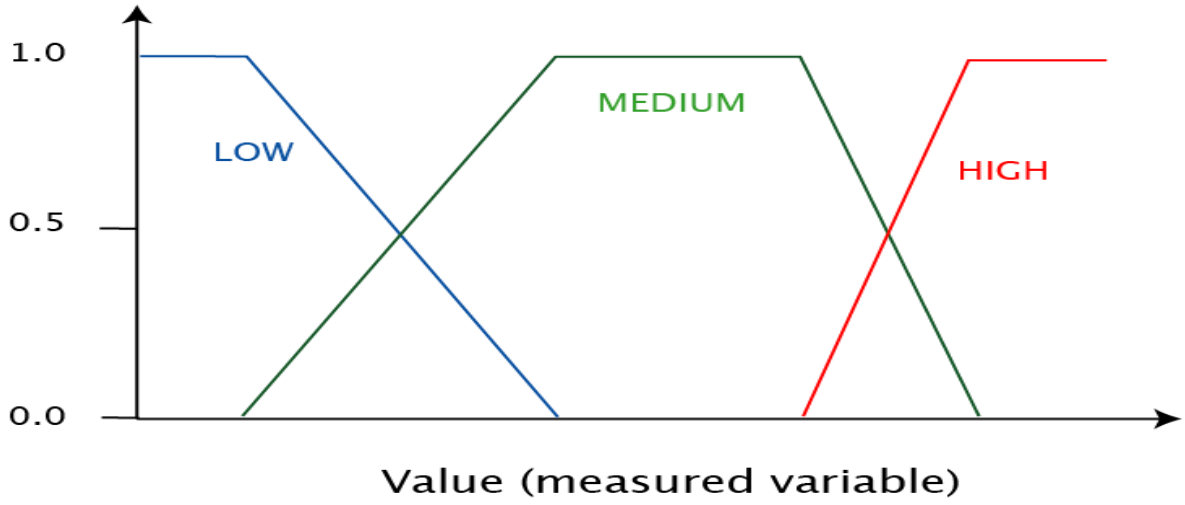

Figure 1.4: Graphical Representation of Fuzzy Logic

1.4.5 Clustering: Clustering is a procedure of dividing a gathering of information into an arrangement of significant family, called cluster groups. Clustering can be utilized as stand–alone tool to get inside into information distribution or it can be utilized as pre-processing step for different calculations.

The process of organizing objects into groups whose individuals are comparable somehow. While doing cluster analysis, we initially segment the arrangement of information into bunches based on data similarity and then assign the labels to the groups. It is versatile to changes and helps single out valuable features that recognize distinctive gatherings. It is used in various applications in the real world. Such as data mining, voice mining, image processing, etc. It is important in real world in certain fields. It is use for study the internal structure of a complex data set.

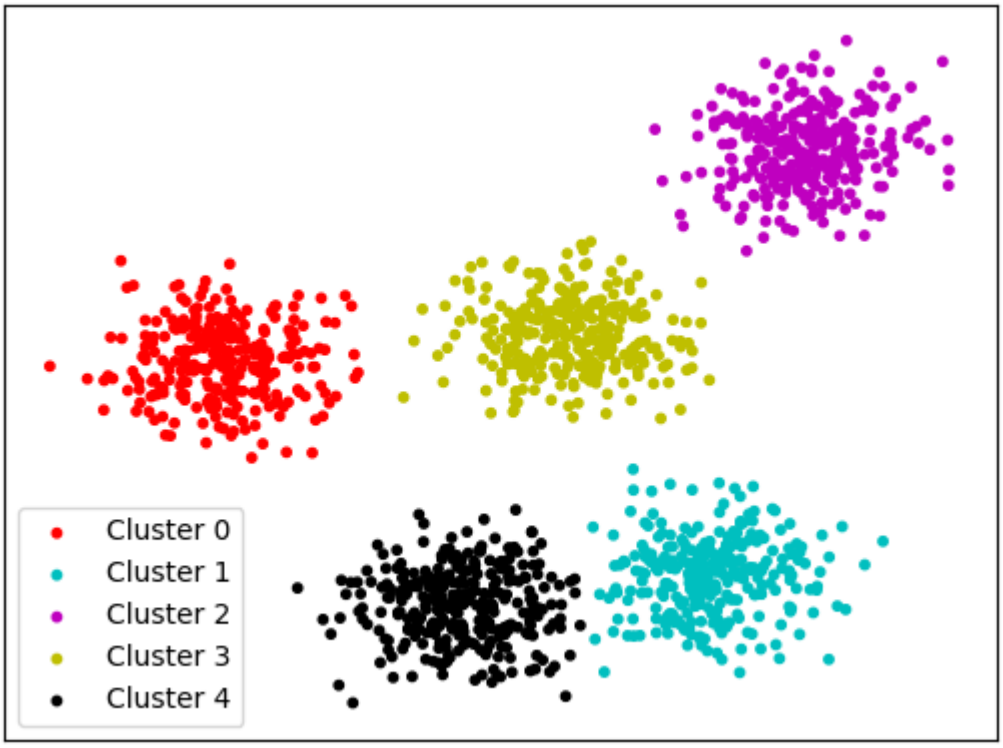

Figure 1.5: Process of Clustering

1.4.6 Correlation and Regression Analysis

Before going into complex model building, by seeing at data relation is a sensible advance step to see how your different variables cooperate together. Correlation take a gander at patterns shared between two factors, and regression take a gander at causal connection between an indicator (independent) and a reaction(dependent) variable.

Correlation: As specified above relationship take a gander at worldwide development shared between two factors, for instance when one variable increments and alternate increments also, at that point these two factors are said to be positively correlated. The other way round when a variable increment and the other reduced then these two factors are negative correlated. On account of no correlation no pattern will be seen between the two variable. The following Figure depicts four hypothetical scenarios in which two different variables are plotted along the different axis.

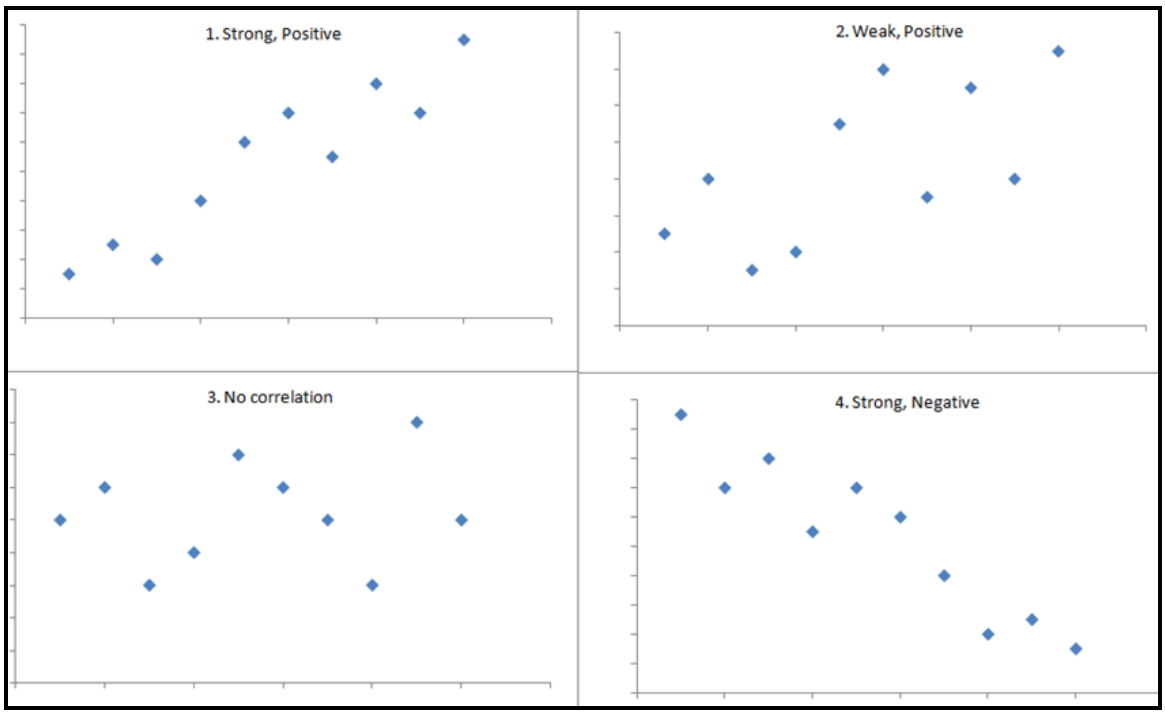

Figure 1.6: Different Scenarios of Clustering

In the above Figure, First scenario shows a strong positive correlation between two variables. Moreover, the Second scenario describes a weaker association between two variables. Furthermore, Third scenario might depict the lack of association, where r is approximately 0. The last Scenario shows the strong negative association observed between two variables, where r is supposed to be negative.

The correlation coefficient of two variables in a data set equals to their covariance divided by the product of their individual standard deviations. It is a standardized estimation of how the two are linearly related.

Formally, the correlation coefficient is determined by the following formula, where s_x and s_y are the sample standard deviations, and s_{xy} is the sample covariance.

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Regression: Regression analysis is a very widely used statistical tool to establish a relationship model between two variables. One of these variable is called predictor variable or independent variables whose value is gathered through experiments. The other variable is called response variable or dependent variable whose value is derived from the predictor variable. For example Profit or Loss predictions from the last few years available sales records of any business. The general mathematical formulation for a linear regression is: Y=aX+b

Where y is the response variable (dependent), x is the predictor variable (independent) and “a” and “b” are constants which are called the coefficients.

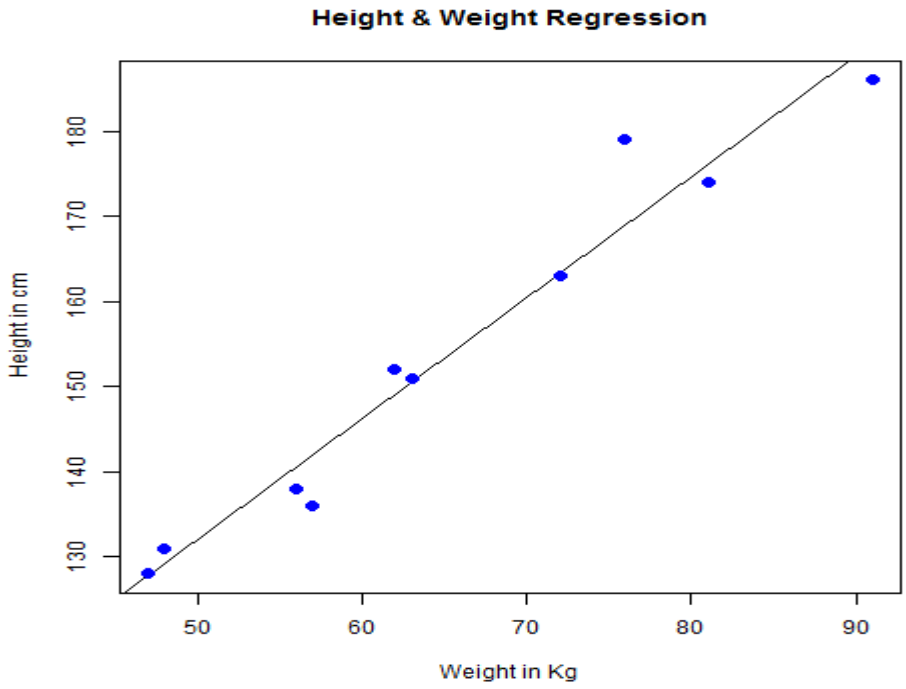

Figure 1.7: Regression Process

1.5 DATA MINING APPLICATIONS

Data mining is extremely use because of its many benefits. In this there are numerous low cost techniques to collect and manage the data or facts, but they are some approaches for extracting helpful knowledge from this data. Data mining has different applications in multiple fields.

- **Marketing and Retailing:** Marketers can make scheme to fulfil the each and every requirements and understand their buying behaviours with the help of data mining.
- **Banking:** Financial organizations can acquire the help of data mining in credit and loan details. A credit card issuer can detect fraud credit card transaction.
- **Research and Development:** Using data mining approaches researchers extract the knowledge by analyzing the data and precede their research work.
- **Education:** Data mining is very helpful in educational organizations or institutes because there is a large number of unused collected data and this data can be used in a proper way using data mining.

1.6 TOOLS OF DATA COLLECTION AND ANALYSIS

1.6.1 R Studio

R is a framework which provides the statistical tools and different packages for the analysis of the data. It is a statistical computing and graphics language which is very similar to S language developed in Bell laboratory. It was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand. It provides some strong features like

- Huge data handling capacity and storage.
- Different library packages for analysis as well as graphical visualization.
- Inbuilt different classification algorithm for supervised learning
- Provide a good environment for the data mining analysis.

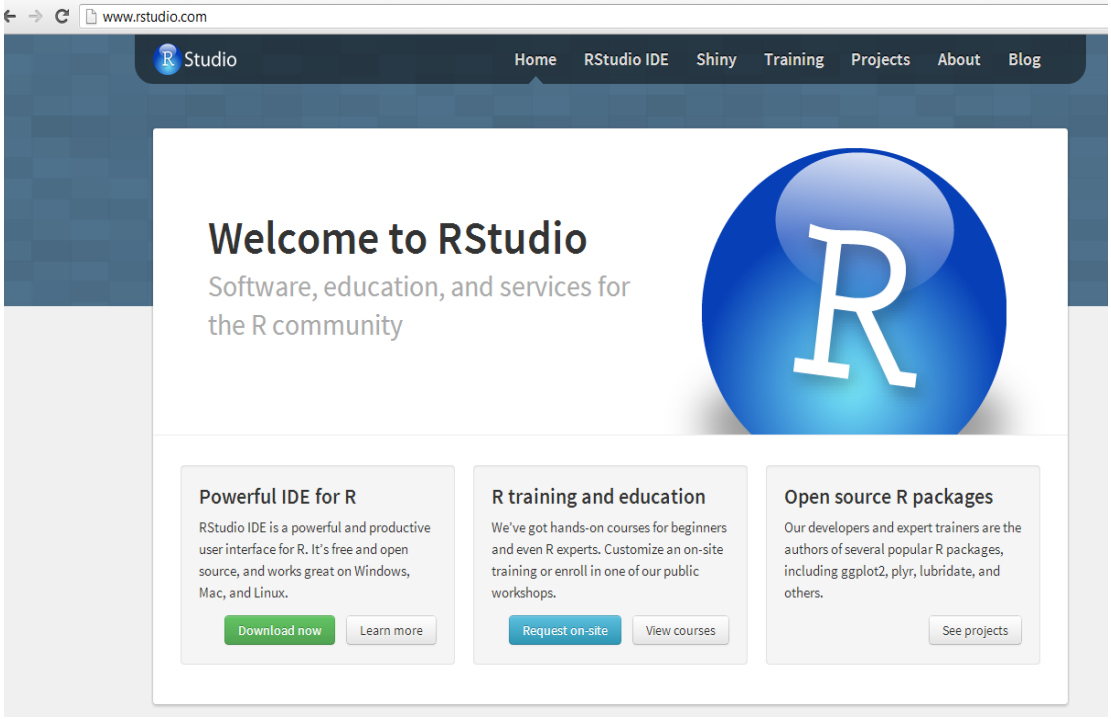

Figure 1.8: Representing Outlook of R Studio

1.6.2 MATLAB

MATLAB is the tool which is used to perform mathematical complex computations. In this MATLAB simplified C is used as the programming language. The MATLAB has various inbuilt toolboxes and these toolboxes are mathematical toolbox, drag, and drop based GUI, Image processing, Neural networks etc. The MATLAB is generally used to implement algorithms, plotting graphs, and design user interfaces. The MATLAB has high graphics due to which it is used to simulate networks. The MATLAB has various versions by current MATLAB version is 2015. The MATLAB process elements in the form of MATRIXs and various other languages like JAVA, PYTHON, and FORTRAN are used in MATLAB.

Figure 1.9: Representing Outlook of MATLAB

1.6.3 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.

Figure 1.10: 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:

- It provides support for latest Java Technologies
- It provides support for fast code editing
- It provides support for Rapid User Interface Development
- Multiple languages support
- It provides support for Cross Platform

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

1.6.4 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.

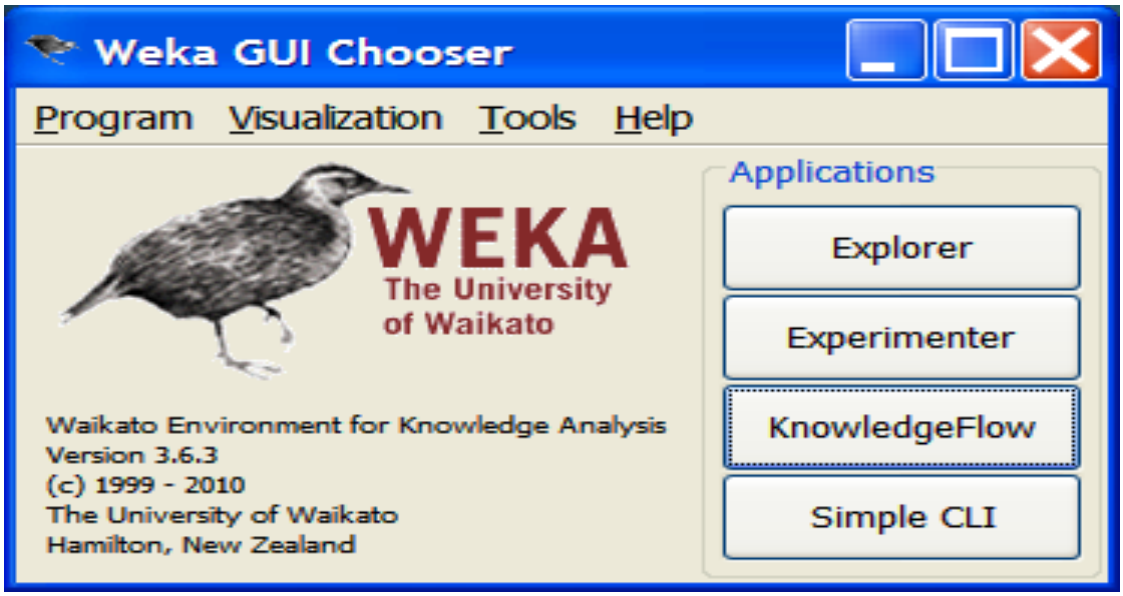

Figure 1.11: Representing outlook of WEKA 3.6.3

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

- It provides support for data preprocessing.
- It provide easy data Access to the files such as ARFF, CSV etc.
- It provides tools for Feature Selection.
- It provides algorithmic support for Clustering and Classification.
- It provides support for Association Rules.
- It provides Data Visualization.

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. Harb [1]: In this study, 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 [2]: 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 indicator variable at each node. The search stops when the split with the largest imprudent in goodness of fit is found. A few element choice calculations are connected and includes positioning higher in numerous calculations are chosen. In this way 15 vital parameter are chosen from unique 77 attributes. 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 [3]: 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 [4]: 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. The essential idea driving proposed technique is to pick close ideal way with k-1 groups and introduction of k-implies bunch. Downside of worldwide k-implies is its high computational multifaceted nature. It requires running part k-implies n times when tackles grouping issues. To get arrangement of this issue we have to run weighted part k-implies rather than k-mean. Be that as it may, essential issue is change the method for select information point which decreases bunch mistake. It enhance bunching blunder rate in highlight space by find their ideal arrangement.

K.A Abdul Nazzar, M.P Sebastian [5]: With new innovation logical strategies utilized for gather brings about expansive scale amassing of information identified with various fields. Ordinary information base strategy is utilized to remove valuable data from information banks. Bunch examination is vital information investigation method utilized as a part of numerous application ranges. This paper can speak to proposed technique for making our calculation more compelling and productive which encourages us to lessen intricacy. It is basically difficult to separate valuable data by utilizing regular database investigation strategies. In k-implies bunching calculation primary thought is to characterize set of informational index in to k number of disjoint gatherings. 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.

Saadat Nazirova [6]: 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.

P.Moniza and P. Asha [7]: 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 [8]: 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. 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.

Yen-Liang Chen, Hsiao-Wei Hu, Kwei Tang [9]: 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 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 [10]: 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.

Jasna Soldic-Aleksic [11]: 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.

Olaiya Folorunsho [12]: 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 [13]: 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. This technique identifies the semantic anomaly. Semantic exception is an information point that acts uniquely in contrast to other information focuses in a similar class or same bunch or 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 strategy manages content and date dataset that has not been executed before using genetic k-means.

John Jacob, Kavya Jha, Paarth Kotak, Shubha Puthran [14]: 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.

Chinmayee C, Manohar M, Bhavana S, Sayeeqa Anjum [15]: A student's academic performance is influenced by several factors. Studies have been done in the educational data mining area to search out what are all the factors that have an effect on a student's academic performance. There are many factors which may have an effect on a student's scholastic achievement but our study aims to search out the major factors that may have an effect on a student's academic performance. In our study we have taken a normal student, who can be a primary school student or an undergraduate student. Predicting student's performance becomes tougher attributable to the big volume of information in academic databases. More number of students, large amount of data to be stored and more the responsibility of the institutions to shape the student's career creatively.
Teacher's responsibility increases, they must be aware of the student's activities and behaviour. To make the jobs of a teacher easy, we have identified few factors/attributes affecting student's academic performance the most. Our study might bring edges and impacts to students, educators/lecturers and tutorial establishments. The factors which we have researched in our study are – student's background, financial status of student's family, college/school surrounding, class environment, faculty support, parent's support, family stress, friends circle.

Subaira.A.S [16]: This paper describes the various approaches such as Neural network, K- Nearest Neighbour, Bayesian Classifier, Fuzzy Logic and decision tree classification Algorithms for implementation of intrusion Detection system. With the help of this paper, it is clear that the data mining methods are used to perform the intrusion detection system But this paper don't describe which technique is best for all of these.

CHAPTER 3
SCOPE OF THE STUDY

3.1 SCOPE

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. 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 study will predict students' future performance on the basis of their past and current academic records, which is extremely important for effectively carrying out necessary pedagogical approaches as well as to emphasize on student's weaker zones. This would also help in curriculum design, education policy design as well as in placements strategies. To sum up, this research is about predicting the performance of student by applying decision tree classifier and clustering algorithms on collected data sets using data mining tools and assesses the result.

CHAPTER 4
OBJECTIVES OF THE STUDY

4.1 OBJECTIVES

1) To gather datasets and pre-process the datasets for the experiment.

2) To generate the rules using fuzzy and ANN kind of methodologies for the machine learning models.

3) To build the Clustering model, Regression model as well as SVM classifier model for classification and for making predictions.

4) To predict students academic performance based on their past and current academic records and as well as the other important factors.

5) To predict the parameters that can affect student's academic performance and leads to dropout in all the programs.

6) To predict performance in elective courses and on the basis of prediction results to recommend courses to students.

7) To achieve the accuracy in prediction results and compare it with other algorithmic approaches.

CHAPTER 5
RESEARCH METHODOLOGY

5.1 SOURCES OF DATA SET

The dataset is gathered from two different data sources. These are primary sources and secondary sources. Data collected from primary sources is termed as primary data. Primary sources are survey, interviews, and questionnaire. On the other hand, data collected from secondary sources is termed as secondary data. Secondary sources are like newspapers, journals, libraries etc. Primary data collected using different primary data sources is also termed as raw facts and figures. Collection of the data is preliminary and very important task for the purpose of this research work, following primary and secondary data sources have been used:

Primary Data Sources: The primary data is the main source here and it is supposed to be based on the response got straightforwardly from Student, Research departments and Institutes.

Secondary Data Sources: The initial phase in gathering information from secondary sources is to survey the research articles, conferences and journals that give general comprehension of point. The subsequent stage here is to refine the facts and figures points concerned with our research from the big archives of data accessible on web.

Different ways to collect data are as follows:

- Collection of information and facts by online application such as University management system
- Available data sets from web sites
- Data collection through surveys

5.2 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, Hybrid approach would be used on collected Students dataset of different colleges to predict their overall performance in academics. 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 sort of classifiers predicts the number of students who are probably going to pass, fail or promoted to next year. The outcomes give ventures to enhance the performance of the students who were anticipated to fail or promoted.

In our research, Firstly, the data is collected through different approaches and then preprocessed. After preprocessing a cleaned dataset would be able to collect. Moreover, attributes are selected on the basis of attribute selection method like chi square etc. It will results in reduced and refined datasets. Furthermore, Correlation Regression models, Clustering and Classification algorithms are used for prediction and detailed analysis of results. In addition to that, Fuzzy logic and formulation is used to define rules.

Finally, on the basis of this analyzed data and defined rules, we would be able to get accuracy in results and can compare it with basic algorithmic approaches already used in earlier research as well as would be able to provide recommendations on the basis of that.

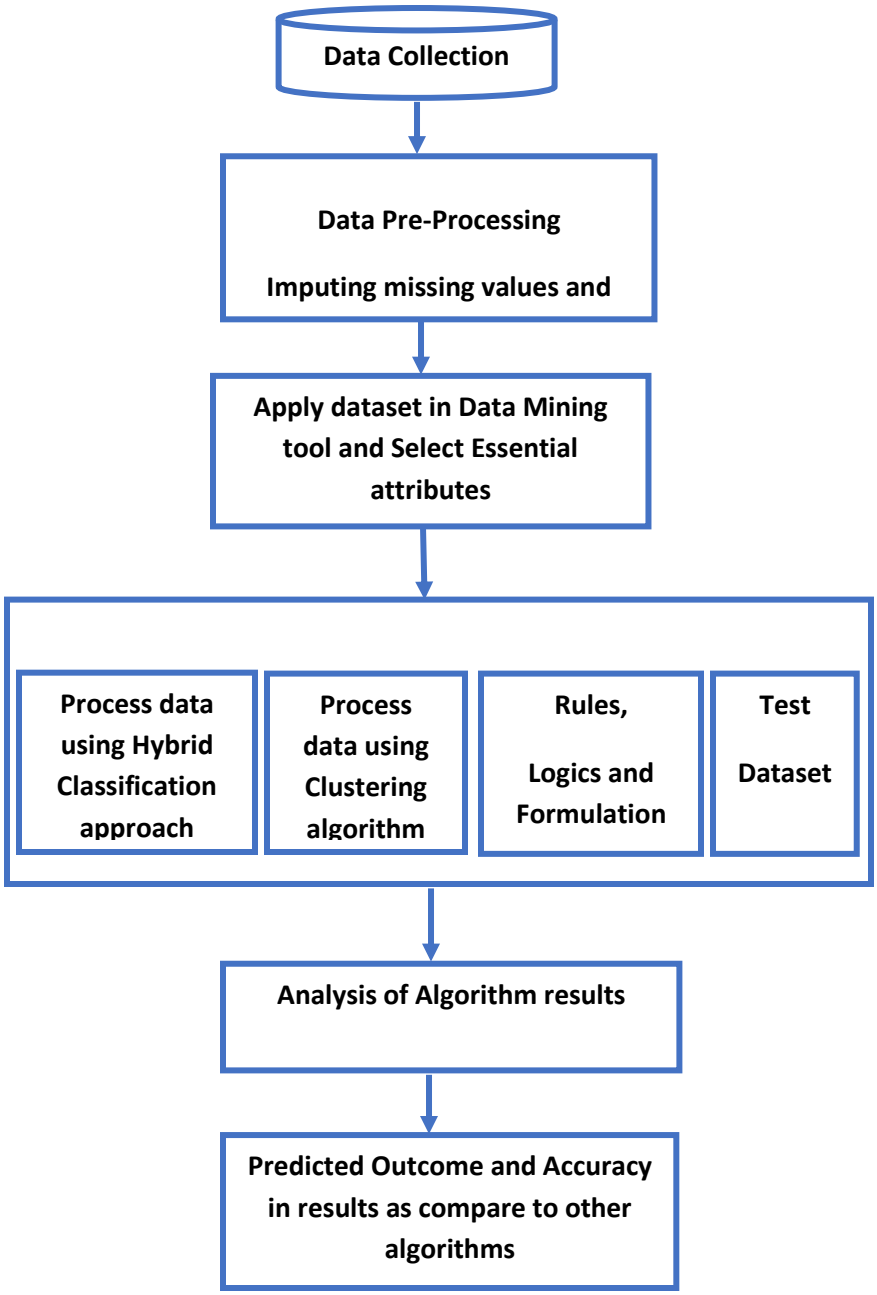

Figure 5.1: Research Methodology

The proposed method would use a Fuzzy system for the prediction of the Students Overall performance by considering different parameter and Dropout rate. Figure 5.2 demonstrates the fundamental structure of a Fuzzy Framework. In this structure a Fuzzy framework comprises of four segments: Fuzzifier, Rule Base, Inference engine, Defuzzifier.

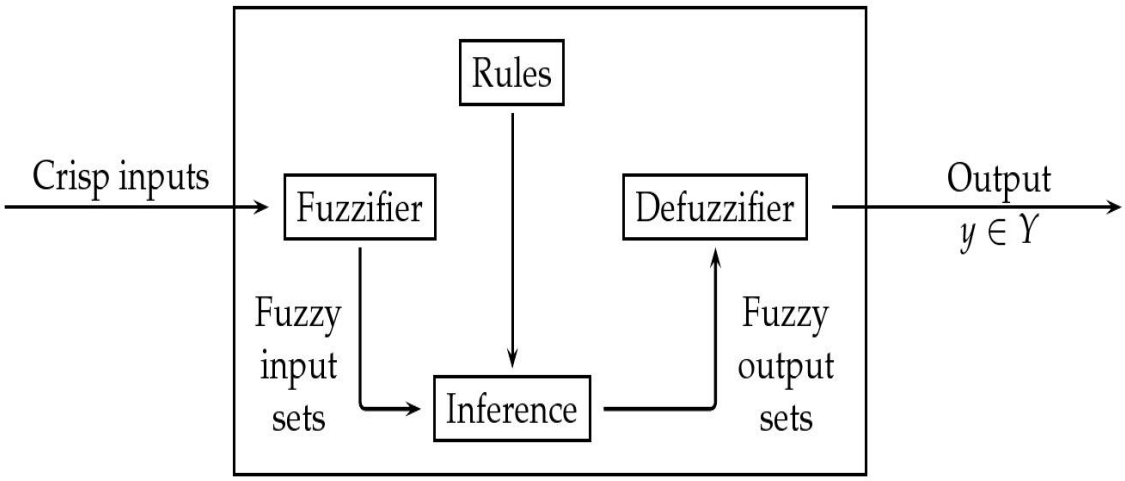

Figure 5.2: Fuzzy Process

A fuzzy system is an information-based rule system. The core of this system is a database which is configured with if-then rules. A fuzzy inference system (FIS) tries to conclude answers from a knowledgebase by utilizing a fuzzy inference engine. The inference engine which is analysed to be the mind of the master frameworks gives the systems to thinking around the information in the knowledgebase and clarifies the outcomes. Fuzzy Inference Systems are extremely essential.

A FIS comprises of an input or information stage, a preparing stage, and output stage. The initial information stage maps the inputs to the appropriate membership functions and truth values. The preparing stage invokes each appropriate rule and generates a result

for each. In the next step it joins the outcomes of the rules. Finally, the output stage converts the combined result back into a specific output value.

5.3 Overview of the Proposed Algorithm

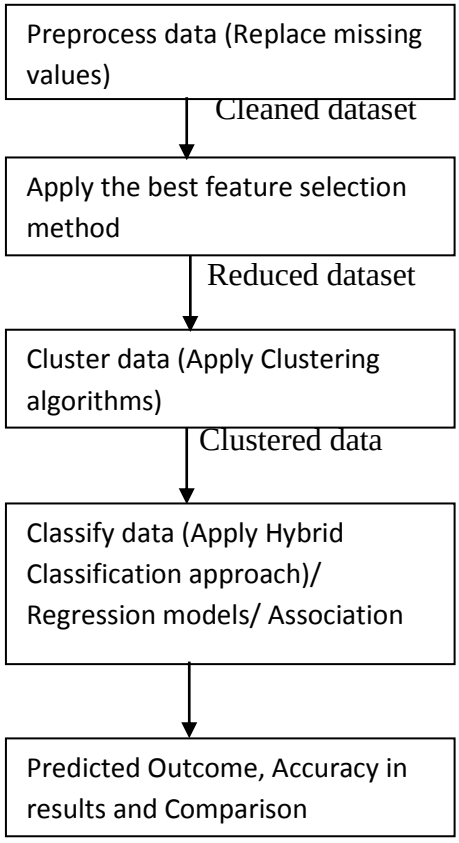

Figure5.3: Process Diagram of Hybrid Approach

Table 5.1 Performance Parameters

Variable	**Description**	**Possible Values**
CA	Continue assessment	{First Second Third Fail}
MTE	Mid-term marks	{First Second Third Fail}
ETE	End term marks	{First Second Third Fail}
ATT	Attendance	{Poor , Average, Good}
HW	Assignment/Home Work	{Yes, No}
LW	Lab work	{Yes, No}
FHS	Feel home sick	{Yes, No}
CS	Communication skills	{Poor , Average, Good}
CG	communication gap	{Yes, No}
LC	low confidence	{Yes, No}
MDP	more dependence on ppts	{Yes, No}
LTR	lack of text book reading	{Yes, No}
PLS	poor listening skills	{Yes, No}
PRS	poor reading skills	{Yes, No}
PWS	Poor writing skills	{Yes, No}
G	Gender	M or F
FE	father's education	{Poor , Average, Good}
ME	mother's education	{Poor , Average, Good}
MJ	mother's job	{Yes, No}
FJ	father's job	{Yes, No}
ACC	Accompany(Friend circle)	{Low, Medium, High}
LM	lack of maturity	{Yes, No}
LP	lack of patience	{Yes, No}
UC	uncertainties	{Yes, No}
LC	lack of concentration (focus)	{Yes, No}
BHA	bad habits	{Yes, No}
BHE	bad health	{Yes, No}
LOC	lack of consciousness, alert	{Yes, No}
BEH	bad eating habits	{Yes, No}
ID	Indiscipline	{Yes, No}
LCR	lack of creativity (innovation)	{Yes, No}
CISYS	complexities in system (ums etc)	{Yes, No}
LCG	lack of counselling	{Yes, No}
IF	infrastructural facilities	{Yes, No}
QE	quality of education	{Poor , Average, Good}
EO	employment opportunities	{Low, Medium, High}
MISC	participation in science fairs, quiz's, competitive exams, MOOC's	{Yes, No}

CHAPTER 6
SUMMARY AND CONCLUSION

6.1 Conclusion

Education System data mining is very relevant to do analyze the performance of students in academics by considering different performance factors. It plays a major role in constructive development of student. The study would help the concerned systems to do improve the student performance and hopefully, would leads to decline in dropouts and rise in placements. Moreover, Education Mining will help in the analysis of various data related to education in terms of how various factors affecting overall performance of student. In addition of that, students can choose the right courses of their interest to enhance learning outcome and to obtain maximum benefit out of this proposed system. Furthermore, this purposed hybrid classification approach would be able to enhance the results in terms of accuracy as compare to other algorithmic approaches. Overall, these outcomes can be used by various research scholars who want to involve themselves in the area of education to achieve their respective objectives.

6.2 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.

LIST OF REFERENCES

[1] Hany M. Harb and 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, pp. 253-262.

[2] Carlos Marquez, Cristobal Romero Morales and Sebastian Ventura Soto "Predicting School Failure and Dropout by Using Data Mining Techniques" IEEE Journal Of Latin-American Learning Technologies, Vol. 8, No. 1, February, 2013, pp. 7-14

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

[4] 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.

[5] K.A Abdul Nazeer and 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),

[6] Saadat Naziova "Survey on Spam Filtering Techniques", Communication and Network, August 2011, pp. 153-160

[7] P. Moniza and P. Asha "An Assortment of Spam Detection System", International Conference on Computing, Electronics and Electrical Technologies [ICCEET] 2012, pp.77-83

[8] Patricia Bellin Ribeiro, Luis Alexandre da Silva and 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, pp. 304-311

[9] Yen-Liang Chen, Hsiao-Wei Hu and Kwei Tang, "A Novel Decision-Tree Method for Structured Continuous-Label Classification" IEEE Transactions on Cybernetics, 2013, pp. 1734 - 1746

[10] 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

[11] Jasna Soldic-Aleksic , Journal of Economics and Engineering, ISSN.: 2078-0346, Vol. 3. No.1, April 2012, pp. 241-248

[12] Olaiya Folorunsho, "Comparative study of different data mining techniques performance in knowledge discovery from medical database", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 3, March 2013, pp. 11-15

[13] Nancy Lekhi and Manish Mahajan, " Outlier Reduction using Hybrid Approach in Data Mining" I.J. Modern Education and Computer Science, 2015, pp. 43-49

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

[15] Chinmayee C, Manohar M, Bhavana S, Sayeeqa Anjum, "Analysis on factors affecting student academic performance using data mining techniques", International

Journal Of Advance Research And Innovative Ideas In Education, Volume 1 Issue 5, 2016, pp. 149-155

[16] Subaira.A.S and Anitha.P, "A Survey: Network Intrusion Detection System based on Data Mining Techniques", International Journal of Computer Science and Mobile Computing, Vol. 2, Issue. 10, October 2013, pp. 145 – 153

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

[18] https://www.tutorialspoint.com/matlab/

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

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

[21] "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].