CAPITAL UNIVERSITY OF SCIENCE AND
TECHNOLOGY, ISLAMABAD

Student Performance Prediction by Using Cluster Analysis

by

Sanam Fida

A thesis submitted in partial fulfillment for the
degree of Master of Science

in the

Faculty of Computing
Department of Computer Science

2020

My dissertation work is devoted to My Family, My Teachers and My Friends. I have a special feeling of gratitude for My beloved parents, brothers. Special thanks to my supervisor whose uncountable confidence enabled me to reach this milestone.

CERTIFICATE OF APPROVAL

Student Performance Prediction by Using Cluster Analysis

by

Sanam Fida

(MCS163011)

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr. Mussarat Yasmin	CUI, Wah Campus
(b)	Internal Examiner	Dr. Abdul Basit Siddiqui	CUST Islamabad
(c)	Supervisor	Dr. Nayyer Masood	CUST, Islamabad

Dr. Nayyer Masood
Thesis Supervisor
30 November, 2020

Dr. Nayyer Masood
Head
Dept. of Computer Science
30 November, 2020

Dr. Muhammad Abdul Qadir
Dean
Faculty of Computing
30 November, 2020

Author's Declaration

I, **Sanam Fida** hereby state that my MS thesis titled "**Student Performance Prediction By Using Cluster Analysis**" is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my MS Degree.

(Sanam Fida)

Registration No: MCS163011

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled "**Student Performance Prediction By Using Cluster Analysis**" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been dully acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS Degree, the University reserves the right to withdraw/revoke my MS degree and that HEC and the University have the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized work.

(Sanam Fida)
Registration No: MCS163011

Acknowledgements

"And your God is one God. There is no deity [worthy of worship] except Him, the Entirely Merciful, the Especially Merciful" [2:163]. First and foremost, I wish to say thanks to Allah (S.W.T) for giving me blessings, power and knowledge to finish this research. Secondly, I wish to express my gratitude to my supervisor ***Dr. Nayyar Masood*** for his help, precious time and supervision. I pay my thanks to him sincerely for his assistance, motivation and advice in this field of research. He helped me from the understanding of this subject till the write up of final thesis.I am deeply indebted to my family and my parents for their support and encouragement till the end of my MS thesis.Their prayers and guidance have lead me here. A special thanks to my son and husband for his support and encouragement in the completion of my research work. I pray to Allah that may he bestows me with true suucess in all fields in both worlds and shower his blessed knowledge upon me for betterment of all muslims and whole mankind.

(Sanam Fida)

Registration No: MCS163011

Abstract

Educational Data Mining (EDM) is a branch of data mining that focuses on extraction of useful knowledge from data generated through academic activities at school, college or at university level. The extracted knowledge can help to perform the academic activities in a better way, so it is useful for students, parents and institutions themselves. One common activity in EDM is students grade prediction with an aim to identify weak or at-risk students. An early identification of such students helps to take supportive measures that may help students to improve. There are many prediction approaches proposed in literature, each with its own strengths and weaknesses. In this study, the characteristics of three recent prediction approaches have been studied, their respective weaknesses and strengths have been identified. The weaknesses have been removed and strengths have been combined to propose two prediction approaches that generate better results. While studying weaknesses, a shortcoming in the pre-processing phase has been identified that the preprocessing can remove some valid data rows as well this on one hand reduces the total number of objects in the dataset and on the other hand accuracy of prediction; the resultant feature set contains less number of attributes and produces better accuracy of prediction. Another weakness was found in the selection of classifiers in an Ensemble approach which has been modified by adopting a comprehensive approach for selection of classifiers in Ensemble. These two proposed improvements have been used to produce two different grade prediction approaches. In the first prediction approach, the reduced feature set identified in pre-processing phase has been used in an Ensemble approach achieved accuracy upto 80%. The second prediction approach is a hybrid one that combines clustering and classification for better prediction results. The outcomes of both these approaches have been compared with the base techniques and improvement in results up to 93% has been observed.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the current era, data has become primary ingredient in multifarious disciplines of life. For instance, various organizations like educational institutions, medical centers, engineering, marketing, sports and warehouses contain rich amount of data about different objects or individuals [1]. The amount of data is getting increased rapidly and this bulk is explored by the concerned departments to know different facts about the data or extract hidden patterns. Analysis of data is performed in different forms such as customers purchasing history, track user activities on website, keep track of patients in medical [2] etc. For any industry or organization, data is considered as a valuable asset. To identify requisite information, industries focus on identifying valuable insights from the data. Those insights can help markets to get competitive advantages. For instance, in medial field, it provides early detection of certain diseases, future growth of companies by analyzing finances data, students performance prediction using historical data and so on [3].

Nowadays, the field of data science has become an emerging discipline that provides a wide range of methods to process and analyze data [3]. One of the major fields of data science is data mining that specifically focuses on extracting the interesting patterns or knowledge from data. Data mining is a concept that performs extensive exploration on large data sets to discover patterns and establishes relationships among them to reveal fact about data through data analysis[1].

Educational data Mining (EDM) is an emerging discipline that uses data mining

techniques to extract and analyze hidden knowledge from educational data [4][29]. It is interesting field for research wherein effective knowledge is explored from large scale educational database to predict students performance. The educational data is collected from different sources such as education repositories, web-based education and surveys [4]. This data can be helpful in identifying previous and current performance pattern of students, forecast students graduation time and make a prediction for future use [5][16]. It is essential to effectively convert the tremendous collection of data into knowledge which will be helpful for teachers, administrators and policy makers for analysis [6][29][15].

The immense amount of students data can be mined and analyzed through data mining techniques [7]. This process is referred as Educational Data Mining (EDM) wherein interesting data pattern can be extracted and utilized to improve education system. Data mining techniques are used as analytical tools to extract the hidden knowledge in EDM [7]. These techniques have been used not only to predict students performance but they can also play a significant role in determining that how the prediction algorithm can be used to identify the most influential attributes of a student. The main purpose of EDM is to improve the quality of educational institutes. The improvements can be achieved by using some predictive models to predict the performance of a student, especially those who are the risk of being dropped out [1].

Students are the main stakeholders of any institute as they contribute in economic and social growth of a country, which leads towards production of creative graduates, innovators and entrepreneurs [8]. The educational institutions contain bulk of data in a digital format. That data contains personal information about students, details about their academic progress etc.

State-of-the-art in EDM indicates that most of the students leave institute within first 2 years of their academic degree while only 45% of the them complete their degree [9]. This can become an alarming situation for educational institutes and student as both invest a lot of their resources to acquire positive output. The educational data contains rich range of parameters that can be harnessed to beneficiate the educational institutions and students [10]. Forecasting student performance

is essential to acquire prior feedback and take instant actions to enhance their performance[5]. Such extracted knowledge helps in improving teaching methods and learning process[4]. Moreover, the result of such analysis can help instructors in improving teaching methods to help students in their education and overall improvement in educational output[9][15][17].

Research community has been focusing on proposing methods to improve student retention rate in higher learning institutions. The primary focus of these approaches is to help educational institutes and students as drop out of a student leaves adverse impact on both students and institutes. The first step in contemporary approaches is to identify those students who are at the high risk of being dropped out [9]. The reason behind degrading performance of the a student involves various aspects such as learning difficulties or behavior[3], financial issues[8], low level of confidence[4], health issues, social activities, employment commitments [11] etc. There is a need to identify such problems in advance so that necessary actions can be taken well before time.

In literature, researchers have presented various studies that employ EDM techniques to identify different patterns and facts from the data. There are different EDM methods such as Prediction Models, Structure Discovery, Relationship Mining and Discovery with Models. There are four common mining approaches provided by data mining, (1) association rule mining, (2) sequential pattern mining, (3) correlation mining, and (4) causal data mining[9][29]. The existing state-of-the-art that focuses on identifying reasons behind student attrition has revealed various facts. There exist many features that affect student performance in educational system. There is a difference in aptitude of students that depend upon gender [4][29]. A strong correlation exists between parents education and student performance, therefore, family background-based features play an important role in students academic performance [1][8][4][11][10]. Similarly, students attendance [7] has a major influence on student achievements. Behavioral features [10] also play a significant role in the learner engagement with education [12]. Influential features are identified by using different approaches namely Pearson correlation, gain ratio, information gain etc. [8][4][9][7][13]. Clustering is also used to find

important features to assess students performance[14]. Ensemble method is an approach that combines multiple models to solve the said problem [7][4][10][12]. Ensemble method in contrast to other traditional learning approaches, trains data with set of models, and then combines them by taking vote on their result. Mostly, predictions made by ensemble methods produce more accurate results as compared to predictions done by a single model [4]. Varying datasets are being employed in the studies focusing on student performance prediction[2]. Research can be conducted on face to face learning or can be on learning management system (LMS). LMS is an e-learning system that captures students, track progress and gives targeted outcomes. LMS is an innovative idea as compared to books, PDFs which helps in developing online learning methods for student and institutes [15].

However, all of prediction algorithms fall under two major methods, clustering and classification. Some of the approaches produce good accuracy through classification while some produce good results with clustering. Some of the researchers have combined classification and clustering to form a hybrid approach with the intention of producing more accurate results [13][21][20].

In [7] students were categorized into low, middle and high level so that student at a risk of being dropped out could be identified initially and provide guidance. The individual features were analyzed for their impact on student performance prediction and studied the performance of comprehensive classifiers. They also proposed an ensemble meta based tree model (EMT)in [4][13][18][24]. The features were analyzed as a category without considering impact of each feature by using SVM, Nave Bayes, decision tree and neural network. These best features were taken as an input into a hybrid model and k-means clustering was applied using majority voting. This thesis presents an enhancement of both the studies [7][13] by merging best and effective features of both the approaches. Accuracy of a model built to predict students performance depends on the feature set and the particular method employed for classification. Before adopting a particular approach for performance prediction, it is important to validate the accuracy with different perspective. This study analyzes each feature individually taken from the same data set as of [7][13] and predicts the student performance using ensemble

methods and clustering.Contribution in this research work is:-

- Detailed literature reviewed
- Deficiencies in the previous work have been highlighted
- Strength of different approaches have been combined to develop an improved prediction approach
- Cluster analysis is used as part of prediction problem to enhance the performance

1.1 Problem Statement

The two base papers approaches report different results on the same dataset [9][13]. This significant difference in reported results needs to be investigated and validated using different approaches. These differences need to be probed so that we could have a dependable approach that can be used for this critical task of students performance prediction.

(i) **RQ1** : Which of the two preprocessing approaches either threshold based or category based produces better feature set?

(ii) **RQ2** : How can we combine the strengths of EMT and Hybrid approaches to form a better prediction approach?

1.2 Purpose

The main purpose of this study is to identity low performing students at the early stage of their academic degree so that maximum amount of student retains in the institutions and a student excels in his academics. Early detection of such students will help in forming a strategy to provide special attention to them for improving their academic performance. The proposed early warning system will assist

the students, teachers and institutes to timely focus on the alarming situations. The features employed by the proposed system include Demographic, Academic background Features, social, behavioral features etc. This study will help in improving quality of education, teaching methods, learning process. It can further contribute in decision making regarding teacher hiring and improving performance of a students which leads to overall improvement in educational output.

1.3 Scope

The system proposed in this study will be applicable for all the educational institutions as they will be able to early predict the performance of enrolled students. Thereafter, they will be able to form a strategy to improve academic performance for improving the retention rate. In general, this prediction model is generalized and can be used by schools, colleges and universities or any other educational platform.

1.4 Definitions, Acronyms, and Abbreviations

- BP : Base Paper
- CGPA: Cumulative Grade Point Average
- DT: Decision Tree
- EDM: Educational Data Mining
- EMT: Ensemble Meta-Based Tree Model
- IBK: Instance Based learning k-nearest neighbour
- JRip: Java Repeated Incremental Pruning
- LA: Learning Analytics
- LMT: Logistic Model Tree

- LWL: Locally Weighted Learning
- MOOCs: Massive Open Online Courses
- NB: Nave Bayes
- OneR: One Rule
- PAM: Partitioning Around Medoids
- PART: Partial C4.5 decision Tree
- PA: Proposed Approach
- RB: Rule Based
- SEM: Semester
- SMO: Sequential Minimal Optimization
- SPP: Student Performance Prediction
- SVM: Support vector machine
- WEKA : Waikato Environment for Knowledge Analysis

Chapter 2

Literature Review

Researchers have conducted various studies to build Student Academic Performance prediction model for particular courses or subjects. These studies employ different types of students data with a variety of parameters to identify and classify their students. In Malaysia, researchers have conducted study on first year bachelor students of computer science from UniSZA. The study performed comparative analysis among three selected classification algorithms; Decision Tree (DT), Nave Bayes (NB), and Rule Based (RB). The data set selected for experiments comprises of a duration of 8 years that contains 497 record from July 2006/2007 to July 2013/2014. The data includes various aspects of students record including family background, previous academic record and other demographic features. RB showed the highest accuracy value of 71.3%. The model will allow the lecturers to take early actions to help and assist the poor and average category students to improve their results. Authors claim that poor and average result can be identified earlier by using this model to improve their future performance. The dataset was quite small due to incomplete and missing values. The study can be expanded by adding more data to increase the accuracy[1].

In the study [2], researchers have proposed Multi-Instance Multi-label (MIML) algorithm by using k-nearest neighbor techniques. The research was conducted on academic warning system for detection of students who find difficulty in their prior courses. The paper further demonstrated that college courses were correlated

therefore, it is better to predict them simultaneously. Result was not only compared with traditional supervised learning method but also with citation KNN. Mostly researchers heavily relied upon online learning activities but here focus is on traditional face to face learning. Student performance was predicted prior to the start of each course. Corse correlation was fully utilized. There are many other features which can affect student performance such as family, health, philological status. Moreover, the quantity of employed dataset used was quite small. The experiment could be conducted on large dataset.

In another paper [3], the final semester marks were predicted from the internal marks of students. Dataset with 1938 instance were used in the experiments. In order to increase the accuracy, this system has introduced reweight enhanced boosting algorithm. The outcomes were compared to existing algorithms like Adaboost (Decision Stump). The Adaboost(J48) produced much better accuracy. The classification techniques were applied to the students data. This model has shown improvement in student performance and class imbalance problem was addressed.

In [4], prediction model was proposed on dataset collected from LMS (learning management system). Behavioral features, demographic features, academic background features were considered. Filter method using information gain based on selection algorithm was used. The outcome was critically analyzed with and without behavioral features. The result showed strong impact on academic achievements of students. Ensemble method was used to improve accuracy and recall after applying traditional classification methods namely: DT, NB and ANN. Boosting method came up with good result. More features could be analyzed by using this model. This model performed very well with 480 records and showed 80% of accuracy but it should be verified with large dataset. This model has not handled face to face learning used in the classrooms.

In this study [7], researchers have proposed a prediction model for evaluating student performance by using dataset of 400 records with 13 features. This study analyzed correlation and relationship of features to their corresponding labels (student performance). Several Machine Learning(ML) techniques were examined on

predicting the student performance that indicate how diversity was using these techniques and to what extent they help to improve the performance. Most effective techniques were PART, A2DE, multilayer perceptron, LocalKnn, and J48 algorithms have Accuracy values of 91.8%, 89.5%, 91%, 92.8%, and 94.3%, respectively. The most effective classifiers in predicting the student performance were tree-based classifiers as compared to the other families of classifiers with high value in accuracy and F-measure. The experiments were conducted to improve the result of best classifier by using ensemble method and voting the results with the tree family technique. The results have shown a significant improvement using the proposed EMT model algorithm with 98.5 % accuracy. Student could be examined with more features such as how student could be affected by social media regarding academic performance. The concrete set of ML techniques could be used here to improve the performance. More data mining techniques could be applied such as clustering etc. with same dataset for comparison. It is specific model that cannot handle diversity of different courses.

The researchers have analyzed the performance of students in 4 years bachelor degree. The study took only marks as an input without considering any other feature. Naive Bayes performed outstanding with accuracy of 83.65% followed by 1-NN and Random forest. NB, 1-NN and Random forest were not human understandable so decision tree was used to derive the courses which are effective indicator. Typical progression of student performance was analyzed by X. mean clustering and Euclidean distance. The employed data set was based upon a sample of 210 undergraduate students. The result showed that proposed pragmatic policy was reliable which showed early sign of struggle and opportunity, graduation performance of other degree program could be analyzed. The courses identified as indicator for high or low could be investigated for student performance. This prediction system was proposed for annual system. Further research could be conducted for semester system on the same parameters, which will be giving the university another leverage to improve academic outcomes[6].

In this study [8], researchers have evaluated the impact of each feature for the prediction of performance of scholarship holding students. The dataset consists

of 23 selected features and 776 student record. New features other than academic, mostly related to family expenditure and student personal information were considered. The features like students natural gas expenditure, electricity expenditure, self-employed and location characteristics were most influential for prediction of students academic prediction. The classifiers like BN, NB, SVM, C4.5, and CART were used to build the learning model. The SVM was proved as the best classifier with F1 score of 0.867 in comparison with BN, NB, C4.5, and CART. The concrete feature can be used to attain maximum accuracy. More data mining techniques can be applied to increase accuracy. Social media content that are basis for personal expenditure can be used in prediction.

In [9] Marbouti et al. have identified student at risk by using in course performance during the semester. The authors have employed logistic regression, support vector machines (SVMs), decision trees (DTs), ANNs and a Nave Bayes classifier (NBC) for experiments. The study aimed to keep false negative error with decreasing false positive error. This study used input features, such as grades, project milestones, team participation, attendance, quizzes, weekly homework, mathematical modeling activity tasks, and exams. The best models for predicting student who passed are KNN with 99.7%, MLP with 96.7 and DT with 96.1 accuracy. These models have low accuracy with respect to identifying failed students. as research are dealing with student at risk, best models for identifying such are NBC with 86.2%,SVM with 72.4 and logistic regression with 58.6. Finally, researchers have used ensemble method including two models with the lowest false negative errors, the NBC and the SVM, and the KNN model with lowest false positive errors. The NBC and the Ensemble model were the best models with the highest F1.5 scores 0.61. The NBC performed better because data set contains only 10% of the students. The NBC is a simple model with high bias and low variance. However, NBC may perform poorly for different dataset. Pedagogical decisions made by the course designers can vary from course to course. Another problem is finding optimal time for prediction.

In [10], researcher have analyzed students diversity and the way in which course are delivered to meet the standards and outcomes. This study concluded that

student engagement with technology improved the academic outcomes. The online course learning material played a significant role in the result of students. Accessing online learning material was an indicator of good performance of students. The student s enrolled in full-time consuming one semester were compared to part-time student lasting two semesters in duration. The result showed that the student accessing online courses regularly had better assessment and marks. It was also observed that there was significant relationship between proficiency and participation level. Moreover, the male students were dominant in full-time course and female in half-time course. The authors have shown the comparisons results through graph, however, overall outcomes are not fully elaborated. There was a minority group which showed high level of access but their progress result in exam was worst. Further research could be conducted to what extent the time is spending on online material to improve academic research. Female and male students behavior could be studied more for impact on academic performance.

In [11], researchers have predicted the final scores of Mathematics so that right student can be selected for the certain tasks by using two methods, k-nearest neighbor model and SVM prediction model. Feature selection process was carried out by finding correlation between the grade and target value. The result of K-nearest neighbor model and the SVM prediction model was compared. The result showed that both classifiers performed very well for this kind of scenario but SVM slightly outperformed KNN with the correlation coefficient of 0.96 and 0.95 respectively. This work predicted actual value which is regression problem that is more complicated than classification. Performance of proposed techniques should be evaluated considering classification aspects. The SVM has not performed well with large database as it requires long time for training. Here outstanding performance of SVM is due to concise dataset but in the case of large data it may perform poorly.

In [12], the study was conducted on use of early warning system based on course to increase student success. Three models were built for week 2, 4 and 9 to predict students at different time. Models ware optimized to identify student at risk. Models achieved accuracy up to 79% for week2, 90% for week 4 and 98% for week

9 by Analyzing effective time for performane prediction of student.

In [13] new predictions approach was adopted based on both classification and clustering techniques. This study carried out the experiments on Learning Management System (LMS) 16 features using SVM, Nave Bayes, Decision tree and Neural Network classifiers. After applying the K-means clustering plus majority voting the four classifiers were compared. The best accuracy of 75% was found when applied to Academic, Behavior and Extra features. The result showed that the hybrid approach yielded good results in term of accuracy in prediction of students performance. This model could be extended for varieties of feature of student dataset.

Another study [14] uses disposition analysis to understand student association with a course. This research explores meaningful features affecting most the students performance. Learning Management System (LMS) of 489 records was used as dataset with 16 features. K-Means clustering was used in this model to assess the effect of students interactional features and students parental involvement features on student academic performance. The result showed stronger effect of these features on academic performance. Clustering performs well for heterogeneous type of data. More features should be studied along with other clustering techniques.

Another study [16] reviewed on prediction of student graduation time. It was seen student were unable to manage to complete their study on time. This paper focused on various factors and method used to predict graduation time. The result confirmed that of Neural Network and Support Vector Machine performed well as compared to Nave Bayes and Decision Tree. It was indicated that academic assessment was a prominent factor when predicting such students.

In [17], researchers have kept the track of academic record to make decision whether a student needs the educational intervention or not. The dataset contains data of 2015-19 batch students of Computer Science by considering academic features. They have used regression model instead of classification model. The proposed system has predicted the result in the numeric way by using KNN, Decision tree, SVM, Random forest, Linear Regression and multi linear Regression by analyzing the result. It is seen that multi linear regression is an optimal solution.

The researchers [15] conducted the study on on-learning management system and analyzed the features as punctuality of student and participation of parents regarding students learning activities. This study used gain ration as a feature selection technique which showed high impact of these features. The result revealed that these features contributed in 10% to 15% increase in accuracy.

Another study [18] analyzed the effect of behavioral effect on student performance prediction. Most influential features were selected by filter-based method by using information gain. This experiment was conducted by using five classifiers: Nave Bayes (NB), Decision Tree (DT), K-Nearest Neighbor (KNN), Discriminant Analysis (Disc), Pairwise Coupling (PWC). To enhance the performance, different ensemble methods such as AdaBoost, Bag and RUSBoost were used. The result showed that behavioral features played important role in student performance prediction and DT performed outclass by achieving 94% accuracy with assembling.

P. Veeramuthu et al. [19] proposed method for enhancing the learning capabilities of the learners in educational institute. This study provided a guideline to the higher education system in improving decision making.The study aims to analyses about the different factors that are affecting a learners learning behavior and their performance using k-means clustering algorithm.

Agrawal et al. [20] proposed a framework which couple ensemble clustering with ensemble classification for the identification of core group applied on world breast cancer dataset. The results showed adaptation of this approach have improved the result. Here result of ensemble clustering was gone through another stage called ensemble classification by providing cluster data as training set and unclustered data as testing set. In this way, improved result with high accuracy is achieved.

In [21] researchers proposed a hybrid machine learning for network intrusion detection. The ConsistencySebsetEvel and Genetic search algorithms have been applied to select specific features. hybrid approach by combination of K means and SMO is applied by achieving 97.3% and reducing the false alarm rate (1.2%).

In [27], researchers have conducted the study by using different datasets. It is also shown that students performance can be predicted and enhanced by preprocessing. Here social environment and behavior features were analyzed by using five different

TABLE 2.1: Literature Review of Student Performance Prediction

Ref No.	Features Category					Classifier /Model							C ~	Ensemble			Data set
	A	B	C	D	E	A	B	C	D	E	F	G		A	B	C	A
1				✓	✓				✓	✓	✓	✓	✓	✓	✓	✓	86%
2	✓		✓	✓	✓			✓			✓		✓	✓	✓	✓	78%
3	✓		✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓		✓	70
4	✓	✓	✓	✓	✓	✓		✓	✓		✓		✓	✓	✓	✓	87%
5	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓
7	✓			✓				✓	✓	✓	✓		✓				
8	✓		✓	✓	✓	✓		✓			✓		✓	✓	✓		95%
9	✓					✓	✓	✓		✓			✓			✓	98%
10			✓	✓	✓			✓		✓	✓		✓	✓	✓		84%
11	✓					✓	✓	✓			✓	✓	✓	✓	✓	✓	-
12	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	98%
13	✓							✓	✓		✓			✓	✓	✓	75%
14	✓					✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	-
15	✓					✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	-
16	✓		✓	✓	✓		✓	✓			✓		✓	✓	✓	✓	94%
17	✓				✓			✓	✓	✓	✓		✓	✓	✓	✓	78%
19	✓			✓				✓			✓		✓				94.1

classifiers such as Backpropagation(BP), Support Vector Regression(SVR), Long-Short Term Memory and Boosting Classifier. BP was superior to all by achieving 87.8 accuracy.

TABLE 2.2: Decription of Table 2.1

Features Category					Classifier /Model							C	Ensemble			Data set
A	B	C	D	E	A	B	C	D	E	F	G		A	B	C	A
A:- Pre Academic Record B:-Academic C:-Demographic D:-Family E:-Behavior					A:-DecisionTree B:NaveBayes C:RuleBase D:-K-Nearest Neighbor E:-SupportVector Machine F:-J48G:-Others							Clustering	A:-Bagging B:-Boosting C:-others			A:-Accuracy

2.1 Literature Review Summary

Students are the main stakeholders of any institutions and they contribute in economic and social growth of a country which leads to produce creative graduates, innovators and Entrepreneurs. It is obvious that the use of learning management system has been increased and as a result institutes contain extensive dataset storing various aspects related to students academic performance.Instead of relying on experience, this performance data can be helpful to enhance student success by taking instant steps based on prior feedback. This analysis can help instructors, students or educational institutes to predict. Students success or failure and overall improvement in educational output.

The educational systems are facing the problem of low performances of students. There are many factors causing academic degrading performance of a student. It has been observed that some students do not complete their studies and leave during the session. Student can be seen spending a lot of time in completing degree due to the poor performance, as they have to repeat the courses as per institute requirements[16][20]. Therefore, it is essential to identify the student at risk at early stage so that necessary steps could be taken to improve their performance in future.

We have performed in-depth analysis of literature on student performance prediction using a set of varying features. The existing prediction model are of generic nature that are unable to handle course to course diversity, as each course is differently designed by the instructor. Time of conduction of student predictions is also quite influential in performance as it can be taken before the session starts, during the semester, end of study. The existing studies have not focused on time of conduction[6][12]. The table 2.1 show the literature review of student performance prediction and Table 2.2 provides the description of previous Table.

This thesis focuses on performance/applicability of different prediction models of students performance with an objective of identifying students that are on risk so that appropriate measures could be considered to improve forthcoming academic performance. The proposed study is based on two papers [7] and [13], which are

considered as a baseline. The reasons of these papers being the baseline are as follows:

- These papers are relatively recent
- They are published at reasonably reputed journals
- They propose different approaches but are using the same data set
- The data set used is publicly available
- The results claimed in the papers for the same data set are quite different

The study in [7] performs student performance prediction by using ensemble methods attaining accuracy of 98%. Here dataset is reduced to 13 features and 400 records due to inconsistent and missing values. In [13] when another approach is applied on the original dataset, the reported accuracy was quite low. So record removed during preprocessing without any reason causes result very high which is quite unrealistic. In [13] Feature selection is applied on categories rather than doing on individual attribute and original dataset is used. Conversely, the proposed study will focus on individual attribute and use hybrid approach combining both clustering (EMT) and classification approaches. The results will be compared to analyze best techniques in term of student performance prediction.
This thesis focuses on the prediction models by evaluating student performance using data set containing 480 records with 17 features. The proposed study relies on two papers [7]and [13] considered as a baseline for this study as the dataset employed by these studies is openly available and published in a reputed journal. The study [7] predicts students performance by using ensemble methods attaining accuracy of 98%. Here dataset is reduced to 13 features and 400 records due to inconsistent and missing values. As original dataset is available, it is visualized that missing values are very low in numbers and removal of records can cause imbalance problem. In [13] when another approach is applied on the original dataset, the model yielded low accuracy. In [13] Feature selection is applied on categories

rather than individual attribute. The prime focus of our proposed study is identifying smarter feature set and achieving benefits of classification and clustering forming a hybrid approach. The results are compared to analyze best techniques in term of student performance prediction.

Chapter 3

Research Methodology

The main purpose of Learning Management System (LMS) is to manage learners, keep track of their progress and their overall participation. This is quite helpful in the scenarios where in student at risk are required to be identified earlier so that necessary steps could be taken well before time.These identified students could be given extra guidance to improve their performance in lacking aspects, thus maintaining their retention rate.

As explained earlier in chapter 1, this thesis focuses on identifying influential features effecting student performance and applying prediction approaches to predict their academic performance. The proposed approach will be able to identify low performing students to help them prior in their respective lack. This thesis holds a strong relevance to the studies [7][13] with intention of improving their results by incorporating the aspects that have previously been overlooked. This chapter presents the methodology adopted to implement the proposed study. We will look features selection in two different perspectives such individual or categorical way and then apply a better prediction approach.

The base papers [7][13] have used the same dataset but with difference of features and records. In [7] only 400 records with 13 features were considered whereas [13] has considered 480 records by dividing features into different groups and using these groups for the prediction. This thesis considers the original dataset rather reduced dataset by excluding only missing and repeated records. The accuracy

attained by [7] is quite high, which is attained by eliminating most of the misclassified records, which leads to biased results. On the other hand, in [13], authors have considered all the records and attained comparatively lower accuracy. In this thesis, we have applied classification including assembling as in [7] but with different set of records.

This study adopts a hybrid approach by considering both clustering and classification. In [13]categorical features were used with application of hybrid approach including both classification and clustering. In another paper [20] clustering and classification ensemble is used that contains strength of both approaches. Our research includes individual features rather than categorical and applying ensemble and clustering based methods. The objective of predicting student performance is to reduce the attrition rate which will ultimately be beneficial for the students, parents and educational institutes. The major focus of the proposed methodology is that we consider an original dataset rather than reduced dataset. We apply Pearson correlation feature selection technique on dataset. After data collection and its pre-processing, different data mining techniques including clustering and classification are employed to evaluate the proposed study as described flow in figure 3.1.

3.1 Dataset

Increased internet usage in various disciplines of life has formed a new scheme of study as web-based learning or LMS. The LMS allocates different learning resources as registration and other learning activities. It can be used to monitor student engagement and keep track of students progress. The dataset that we have used in this thesis is Students Academic Performance Dataset and it has been taken from Kaggle.com (Kaggle 2016). The dataset models activities of around 500 students in a Learning Management System (LMS) environment that contains 17 attributes which have been divided into 4 categories [13]. These categories are 1) Demographic, 2) Academic background, and 3) Behavioral and other extra features. The feature set covers all the features that can cover the satisfaction level of

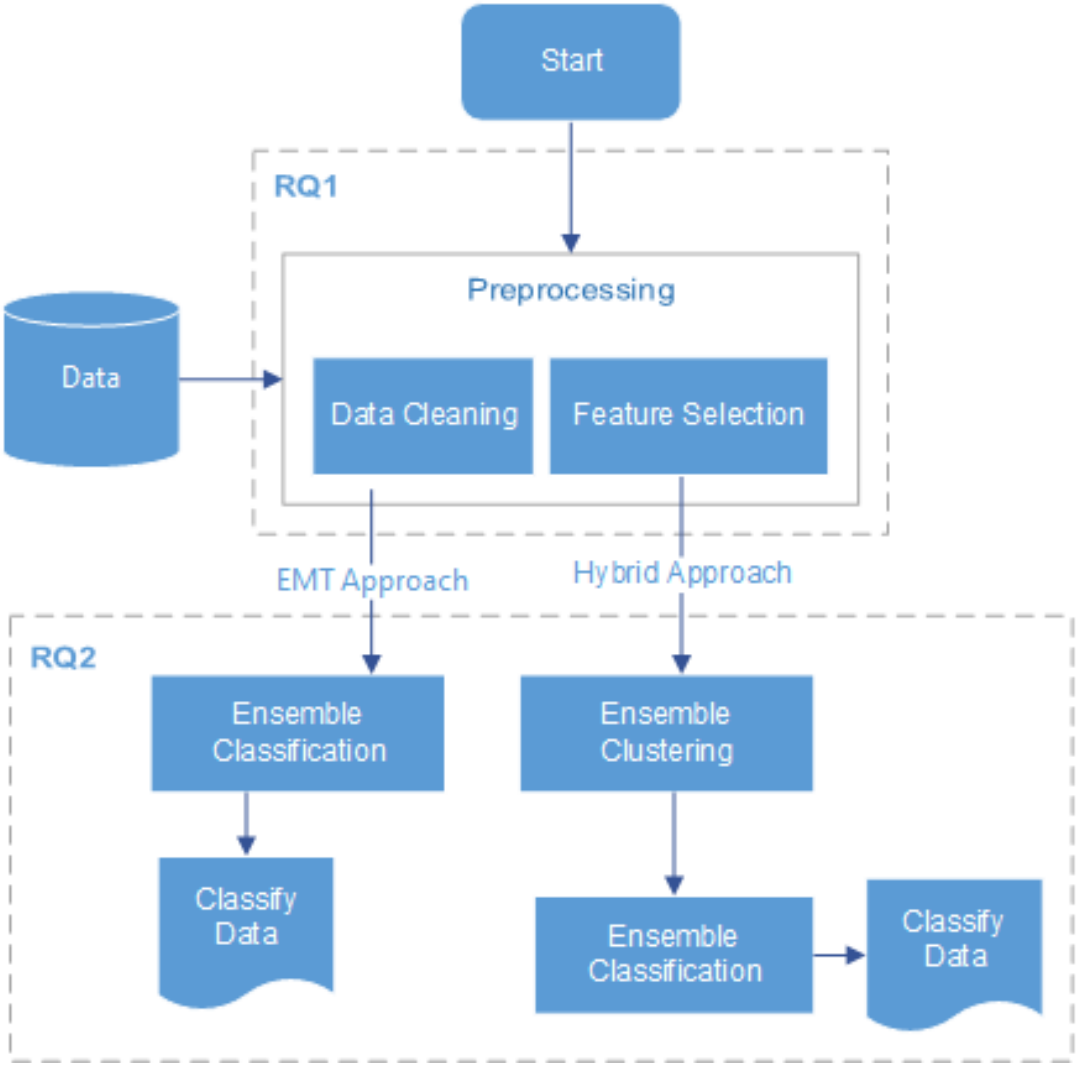

FIGURE 3.1: Abstract flow chart of proposed methodology

both students and parents. The attribute Class is the response variable that takes the values as Low, Middle and High. This dataset has been employed by many EDM approaches [4][7][14][10]. The features and their description are illustrated in the Table 3.1.

Demographic features allow to understand certain background characteristics of a student like age, gender, country, race etc [1]. Specific features related to their studies can be very helpful to predict their performance in near future. Different academic features are grade, attendance, quiz marks, assignment etc [2][3][6][10]. The emotional and interactive characteristics are behavioral features[4]. The expenditure and personal information are family features[8].

TABLE 3.1: Description of Features Used in Student Performance Prediction

S.No	Features	Description of Features	Category of Features
1	Gender	Gender of a student (male, Female)	Demographic Features
2	Country	students belonging country	
3	Birthplace	Students born place	
4	Relation	Parent responsible for the student (father or mother)	
5	StageID	Educational stage of a student (high, medium, low)	Academic features
6	Semester	Students semester (1st or 2nd)	
7	CTopic	Offered courses(IT, Math, English, Arabic,Science, Quran)	
8	SectionID	Class section i.e., A, B, C of a student	
9	Grade ID	Grade level of a student (GL-1,G-2GL-12	
10	Student Absence Days	Students educational stage (high, medium, low)	
11	Raised hands	These are all features concerned with students behavior {ile interacting with kalboard 360 E-learning websites}	Behavioral features
12	Visited_ources		
13	Announcements		
14	Discussion		
15	Parent Answering Survey		Extra features
16	Parent School Satisfaction		
17	Class	A class Label	

Since the data set employed for this study has been taken from LMS, therefore, features like raised hand, announcement discussion, by which students certain behavior could be analyzed are not part of the employed data set [7][13].

The dataset includes attendance feature like Student Absence Days. The dataset also includes parent participation features such as ParentAnsweringSurvey and ParentSchoolSatisfaction. Learning Management system (LMS) has been designed

to facilitate learning through the use of Leading-edge technology. The data is collected using a learner activity tracker tool called experience API (xAPI). The xAPI is enables to monitor learning activities like watching a training video and reading an article.

3.2 Pre-processing

Any constructed model for performance prediction depends upon historical data which is given as a training set. Most of the time, the historical records are arranged in unstructured forma containing redundancies like missing records, noisy data etc. Preprocessing is a technique used to convert raw data into process able form to be accepted by Machine Learning (ML) Algorithms [15]. In order to simplify the job of Machine Learning algorithms, it is important to convert the data set into a proper form.There are different data preprocessing technique to normalize data and remove outliers such as data cleaning, integration, transformation and reduction [24]. .After collection of data set, preprocessing is an important step which includes data cleaning, feature selection, data transformation and reduction. In pre-processing step of [7], dataset was reduced to 400 records because of missing and noisy data. Whereas in this thesis, all the instances of original data set (i.e., 480) are included. We have discovered a very low percentage of noisy data. Mostly removed records were misclassified which could cause biased results. Such an uncertainty can cause misleading prediction results.

3.3 Data Cleaning

Data cleaning is a crucial step of data pre-processing. In data cleaning, major task is to remove noise and irrelevant records, dealing with missing values, recognize outlier and correct inconsistent data [24] . The problem of missing data arises due to absence of data for any attribute. While data collection, irrelevant data is called noise. The missing data can be handled through different filters in WEKA.

Its quite important to handle the missing data and outlier before feature selection. After missing data handling and outlier detection phase, final attributes have been selected to perform experiments. Such attributes have been narrated in Table 3.2. The attributes belong to all three types of factors, demographic, pre-university, and institutional. After completion of data pre-processing, classification algorithms have been applied.

3.4 Feature selection

RQ1: Which of the two preprocessing approaches either threshold based or category based produces better feature set?

As explained earlier, the approach in [13] uses groups of features to select best feature set and that in [7] uses threshold value for this purpose. Based on critical analysis, we have identified some drawbacks in both of these approaches and have adopted an approach to address these drawbacks. The grouping of features evaluates the combined effect of a set of features in the prediction process so it ignores the effect of individual attributes. As every feature in the dataset contains individual effect. This does not exploit the real association of an individual attribute with the response variable. On the other hand, the threshold approach adopted

in [7] calculates the correlation of all attributes with the response variable once and applies a threshold to remove certain attributes without a proper justification of the threshold value.

In order to address the shortcomings of these two approaches, firstly, we have rejected the grouping of attributes to exploit the impact of individual attributes in the prediction. Secondly, we have adopted the backward selection using Pearson Correlation (PC) as a metric.

In order to select best features subset as mention above we have used Pearson Correlation. This feature selection calculated the correlation of each features with response variable. Table 3.2 above shows the PC values of all attributes with the

TABLE 3.2: Result of Feature Selection by Pearson Coorelation

S.No	Ranked Attributes	Correlation Value with Response Variable
1	Visited Resource	0.3788
2	Student Absence Days	0.3565
3	Raised Hand	0.3251
4	AnnouncemnentView	0.2863
5	Relation	0.2357
6	Parent Anwering Survey	0.2328
7	Parent School Satisfaction	0.1804
8	Discussion	0.1465
9	Gender	0.126
10	BirthPlace	0.0915
11	Nationality	0.1815
12	Semester	0.0652
13	Stage ID	0.0631
14	Topic	0.0505
15	Grade ID	0.044
16	SectionID	0.0374

response variable. It is observed that the attribute visited resource is highly correlated with 0.37 value.at the same section ID is less correlated to response variable with value of 0. 037.we have applied classifiers from all the group or family and observed the accuracy. Next we removed the lowest correlated features and again applied the whole classifiers. We have repeated these step until a stable and high value of accuracy is achieved. Figure 3.2 has shown these steps.

We have applied multiple classifiers available in Weka to find the best prediction value. In the first iteration, the Random Forest classifier gave the best result, i.e., (78.0). Then we removed the attribute with the lowest PC value that is SectionID, and applied the whole set of classifiers to find the best result. Once again, the random forest produced best result with accuracy of 77.8. The second result shows decrease in the value. The removal of Grade ID did not have any impact on the results. The third result also indicates the improvement in prediction and effectiveness of the feature selection process. We have applied this process iteratively to obtain the minimum best possible set of attributes.

The table 3.3 shows the attributes which are removed one by one and the accuracy of the classifier at each stage and the final set of features selected for the

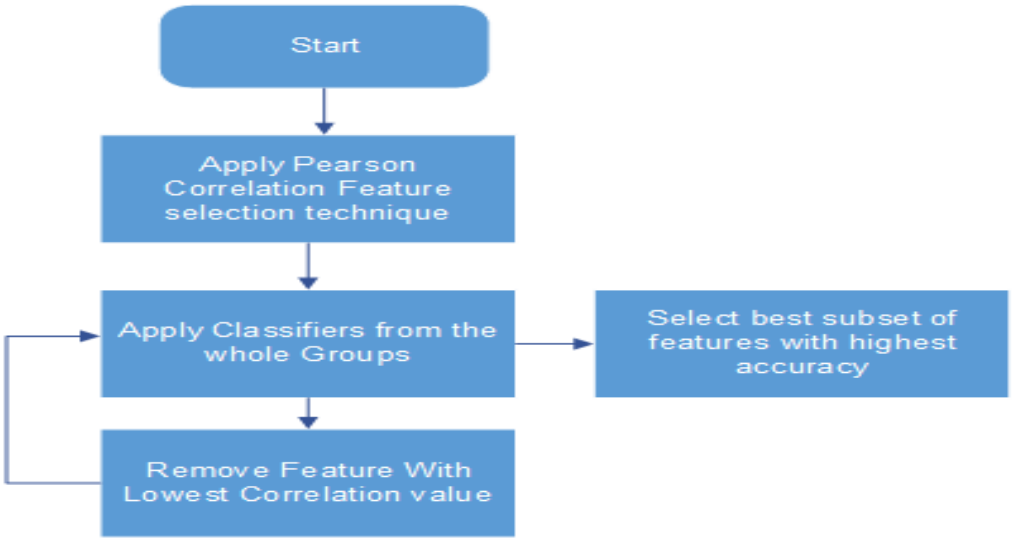

FIGURE 3.2: Feature Selection Technique

TABLE 3.3: Result of Different Subset of Features

S.No~	Total Attributes	Accuracy Before	Minimum PC	Accuracy After
1	17	-	-	78.0
2	16	78.0	SectionID	77.8
3	15	77.8	GradeID	77.8
4	14	77.8	Topic	78.0
5	13	78.0	StageID	79.4
6	12	79.4	Semester	78.9
7	11	78.9	Nationality	77.6
8	10	77.6	PlaceOfBirth	**79.7**
9	9	79.7	Gender	76.1
10	8	76.1	Discussion	74.6
11	7	74.6	ParentSchoolSatisfaction	74.4
12	6	74.4	ParentAnswerinfSurvey	73.6
13	5	73.6	Relation	68.8

further processing. One interesting phenomenon that can be observed that the accuracy of the feature sets decreased initially. However, the accuracy increased gradually but then fluctuated at a certain level when accuracy started to decline. Surprisingly, it was observed during experiments that continuation of the same process tends to improve the results further. The final 9 attributes that have been selected through this process are visitedResource, studentAbsentDays, raisedHand, announcementView, relation,parentAnsweringSurvey, parentSchoolSatisfaction, discussion, gender and a response variable(class). The comparison

and evaluation of selected feature set is presented in chapter 4.

In this work, we used the dataset of LMS used in previous research. We have used Pearson correlation approach on original dataset rather than less records. In the dataset it is uncover that mostly the records not included are misclassify. Which is the reason of extra high accuracy. Two duplicate records are found which are removed in the preprocessing step. In our research we have proposed reduced feature set compromise of smarter dataset. The proposed feature set is better as features are selected on individual correlation rather than group wise selection.it is found by adopting group wise features approach most of the influential features are removed when can be contribute well in the prediction.

Our finding regarding RQ1 is that individual-feature based feature selection [7] produces better feature set as compared to group-based one [13] However, we have modified the former approach by using Pearson Correlation based selection rather than using a threshold value. In this work fluctuation of accuracy is seen when lower correlational features are removed. Best features are selected on the bases of stability regarding accuracy. This produced a smarter feature set compromised of better result.

3.5 Data Mining Techniques

Data miming techniques are used to analyse the data and make better prediction. To evaluate the proposed study, different data mining techniques like classification, clustering, ensemble and hybrid are used to mine data for different aspects. For this purpose, we have used WEKA, as machine learning tool, which is an open source tool. The data file containing feature set was converted into CSV format to import in the WEKA.

RQ2: How can we combine the strengths of EMT and Hybrid approaches to form a better prediction approach?

Classification and clustering are two major categories of Machine Learning algorithms. The former is also called supervised learning while the latter is called non-supervised learning. We present a brief introduction of these two categories and then present our proposed approach.

3.5.1 Classification

A classifier produces a model based on training data, which carries objects described by the values they have on a set of attributes; one attribute is distinguished as a class label. The generated model should fit well with the training data and suitably predict the class label of unknown data. Classification is vastly used in data mining techniques for performance prediction. A major part of our proposed work includes classification model on the training set and new data (testing set) is classified bases upon that model. We have used 10-fold cross validation for training and testing. The contemporary state- of-the-art has applied different classification algorithms such as Naive Bayes, decision trees, random forest trees, k-nearest neighbors and rule induction[6].

Ensemble Meta-Based Tree Model (EMT) is an ensemble technique which combines multiple sets of weak learning classifiers into final prediction model either by using weighting or voting techniques [7]. It combines the best-selected techniques as strong predictive model. The ensemble also balances the under fitting and over fitting with the aim of improving overall accuracy. It is seen as a greatest gain in predictive performance when combining diverse predictions [21]. There are many EMT techniques such as AdaBoost, Bagging, stacking, Random forest etc. Ensemble methods combine different machine learning techniques into a single predictive model to decrease variance, bias and improving predictions[3][7][18][20][**?**][27].

3.5.2 Clustering

Clustering is a data mining technique that works on finding hidden pattern by exploring data in the data analysis process. It is the process of grouping objects

into classes of similar objects. Clustering can play important role in finding important traits among a group of objects based on their common characteristics. This grouping of objects is based on the concept of proximity measure which can be similarity or dissimilarity/distance among objects[22][26]. There are many distances metrics that are used for clustering like, Euclidean Distance etc. and also different types of clustering algorithms such as k-mean,Partitioning Around Medoids (PAM), expectationmaximization (EM) etc. The K-means clustering is largely used algorithm for pattern recognition for the data set of heterogeneous nature[14][19][20]. PAM characterizes clusters by their mediods (centers). EM assigns a probability distribution to each instance which indicates the probability of it belonging to each of the clusters[22]/cite23. In this work we have used three clustering algorithm PAM, k-means and EM on the smarter dataset. Each clustering algorithm is applied individually and evaluated. Afterwards we ensembled the result of three clustering algorithm.

3.5.3 Hybrid Approach

As has been mentioned in at the start of this section that clustering and classification algorithms belong to two different categories in Machine Learning. Their target and purpose are entirely different. Classification is mainly used for the prediction based on the training data set where class label is known. Whereas the clustering is used in the applications where there is no concept of class label and data is grouped based on the common properties of objects. Clustering is used independently as a function and also as a pre-processing step in other class of Machine Learning algorithms. However, there is a recent trend in Machine Learning where classification and clustering algorithms are combined for the prediction purpose. This hybrid class of approaches are in which clustering and classification are both applied in order to get better result. This is quite productive approach in which benefits of two machine learning are achieved. Hybrid approach can be used in different ways as in [13]] features are selected on the bases of accuracy by using classification and then applying clustering on that best features. Where in

other way, clustering can be applied and result is divided into training and testing phase. Afterward data is classified on the bases of that training and testing sets. In [21][20] EMT hybrid approach is used where ensemble clustering is used and on that training and testing data ensemble classification is applied. This is how best result can be obtained by achieving benefits of classification, Clustering and ensemble.

3.5.4 Proposed Approach

Before presenting the proposed approach we have above conquer different aspects of data mining techniques like classification, clustering and hybrid approaches. In this section, we are presenting the proposed approach adopted to address the RQ2. We have adopted two different approaches to address this question. In the first approach, we have used the feature set that we have selected through our pre-processing approach, and have applied EMT approach proposed in [7]. Our contribution in this approach is application of EMT based classification on a feature set that is smartly produced and gives overall better results. In the second approach, we have adopted a hybrid technique for prediction that is presented in [20], however, we have improved the previous work by including EMT based classification in the approach which produces better results. In this way, both of our proposed approaches introduce modifications in the previous work and improve the performance in both cases.

3.5.4.1 Proposed EMT Based Classification

In the study [7][18], authors have claimed that EMT plays essential role in prediction student performance with the achievement of high accuracy. The accuracy is calculated on the basis of applying classification algorithms from all families of classifiers available in WEKA.

In WEKA there are five families or group containing different set of algorithm on the basis of like properties. Bayes, Functions, Lazy, Rules and Trees are such

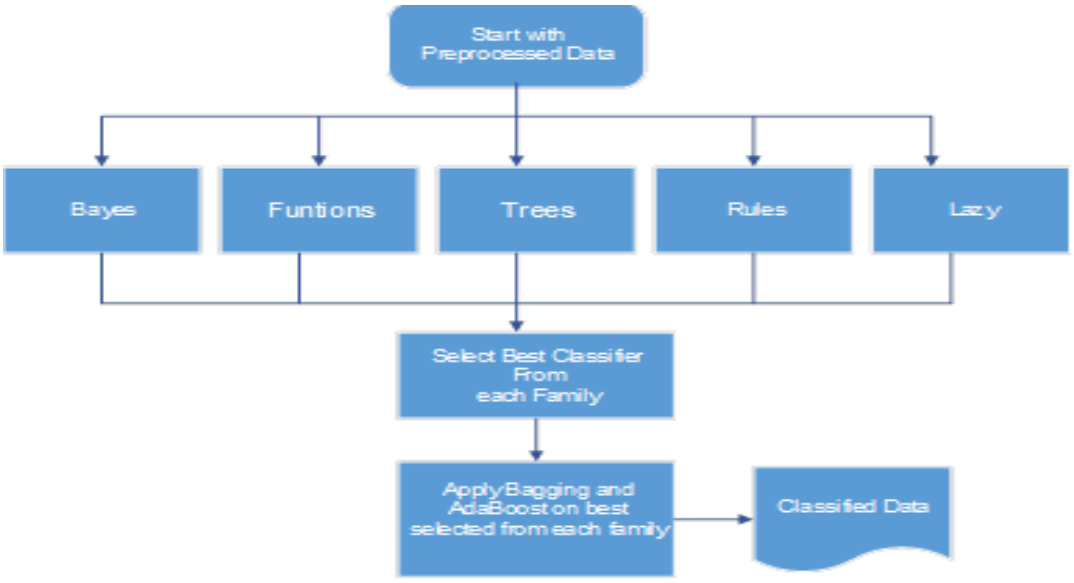

FIGURE 3.3: Proposed EMT Approach

families. In the proposed approach all the algorithm from each family are applied. Then from each family or category the best performing algorithm is selected manually. These five selected algorithm used in further step. In the next bagging and boosting methods are applied on the best five algorithms. Highest result either from the bagging or boosting method is selected from that five classifiers. Working of the proposed approach is shown below in the Figure 3.3.

In the first experiment as mention above, we applying all classification algorithm from each family on reduced features set. As there are various classification algorithms in machine learning which have different assigning label capabilities.it is quit beneficial to use different data mining algorithm instead of selected techniques. Here in this approach most affective data mining algorithm could be selected.

Here we are proceeding with the solution 1 of the proposed work. When all classification from a group are applied its shows varieties of result. BayesNet gives accuracy from group of 72.1. Logistic (used multinomial logistic regression model) and IBK performs same with the accuracy of 74.4 from Functions and Lazy respectfully.in Rules group, PART perform with the accuracy of 73.4. Trees group is on the top of all groups as Random forest perform best with 79.7 accuracy. Random forest works on ensemble techniques. It is also observed that most of the cases Tree groups performance is outstanding. Table 3.4 shows the result below.

TABLE 3.4: Result of All Groups of Classifiers

Algorithm family	**Algorithm Name**	**Accuracy**
Bayes	BayesNet	72.1
	NaiveBayes	70.7
	Nave multinominalTest	44.1
	Logistic	74.4
	Nave Bayes updateable	70.7
Funtions	Multilayer Perceptron	73.2
	Simple Logistic	72.8
	SMO	74.2
Lazy	IBk	74.4
	KStar	72.1
	LWL	70.2
Rules	Decision Table	69.0
	JRip	73.2
	OneR	60.4
	PART	73.4
	ZeroR	44.1
Trees	Decision Stump	52.2
	Hoeffding Tree	70.9
	J48	72.3
	LMT	72.5
	RandomForest	79.7
	˜Random Tree	69.4
	REPTree	73.8

We then applied ensemble approach and used bagging and boosting, in most of the cases result is improved. Then bagging and boosting methods are applied on the best classifier selected from each family. Here we have applied AdBoostM1 and Bagging on the best selected values from each group of classifiers.

Result shows that in the case of boosting method there is no improvement of accuracy.But bagging methods contributes in improvement of result. Here in Table 3.5 below result is increased from 72.1 to 74.6,73.4 to 76.7 in BayesNet and Rules. Whereas function, lazy and tree doesnt show any improvement in the result. Random forest doesnt show any improvement due to bagging but still its result is best from all of others. Result is shown in Table 3.5.

TABLE 3.5: Boosting and Bagging Classifier Results

Best Algorithm from Each Family	Algorithm Name	Proposed Approach	Boosting Method	Bagging Method
Bayes	BayesNet	72.1	72.1	74.6
Funtions	Logistic	74.4	74.4	73.8
Lazy	IBk	74.4	74.4	74.4
Rules	PART	73.4	74.2	76.7
Trees	Random Forest	79.7	77.6	78.6
Average		74.8	74.54	75.62

In the above proposed EMT based classification, we have used reduced feature set here with original dataset. As compare to the previous work, our proposed approach has obtained better accuracy with a smarter dataset. It is discovered that every feature has individual effect so it is not wise to look the features group wise. We have achieved high accuracy by adopting individual feature effects rather than group.it is also shown that ensemble base classification is more productive than simple classification.as EMT grasps the effective attributes of various approaches which enhances the overall result.

3.5.4.2 Proposed Hybrid Prediction

The second proposed part of the research is hybrid approach. Here we are using reduced feature set of the dataset. Hybrid approach is proposed on the original dataset including ensemble clustering and classification. Clustering is considered an outstanding machine learning technique in discovering common characteristics which leads to a group [19][22][23]. This work proposed a hybrid machine learning technique which is combination of classification, clustering and ensemble approaches. The potential of two machine learning approaches could be exploited to enhance the performance of prediction. Hybrid approach is able to decrease the false negative rate, false positive rate and improve the detection rate. In the proposed hybrid approach, three clustering algorithm are applied such as PAM, EM and K. Means. The result of these algorithms is aggregated on the basis of majority voting. As a result, clustered or unclustered data are attained. In

aggregation on which neither of three algorithms shows agreement is considered as unclustered data. In the next step we provide this data to classification unit to classify more data. So that more data can be grouped. Here clustered data is considered as a training set and clustered data as a testing set. The proposed Hybrid approach is shown in the Figure 3.4.

In this work PAM, EM, K-means are applied to discover the best performing method. The outcome is categorized into three clusters. A cluster representing the majority is assigned that label. k-means is applied on best subset of features selected before in the processes of classification[14][19][22]. As it is applied on 478 records of dataset which is mention before.357 students are correctly clustered into 125 as high, 112 as medium and l120 as low. In the same way when PAM clustering Algorithm is applied.it has given result into three clusters representing 112 as high,99 as low and 127 as medium. After K-means and PAM now EM has run on the dataset. EM has clustered the data into high as 70, medium as 106 and low as 98. Table 3.6 shows the result below.

TABLE 3.6: Performance of All Clustering Algorithm

Clustering Algorithms		**Total Objects**	**Majority Class Label**	**Majority Class Total**	**Minority Class Total**
K MEANS	Cluster 0	193	H	125	68
	Cluster 1	159	L	120	39
	Cluster 2	126	M	112	14
				357	121
PAM	Cluster 0	200	H	112	73
	Cluster 1	167	L	99	14
	Cluster 2	111	M	127	53
				338	140
EM	Cluster 0	185	M	106	94
	Cluster 1	113	L	98	69
	Cluster 2	180	H	70	50
				274	213

When we critically analyzed the result and make comparison of class to the PAM, EM and K-means algorithm.it is observed that k-means is performing well with

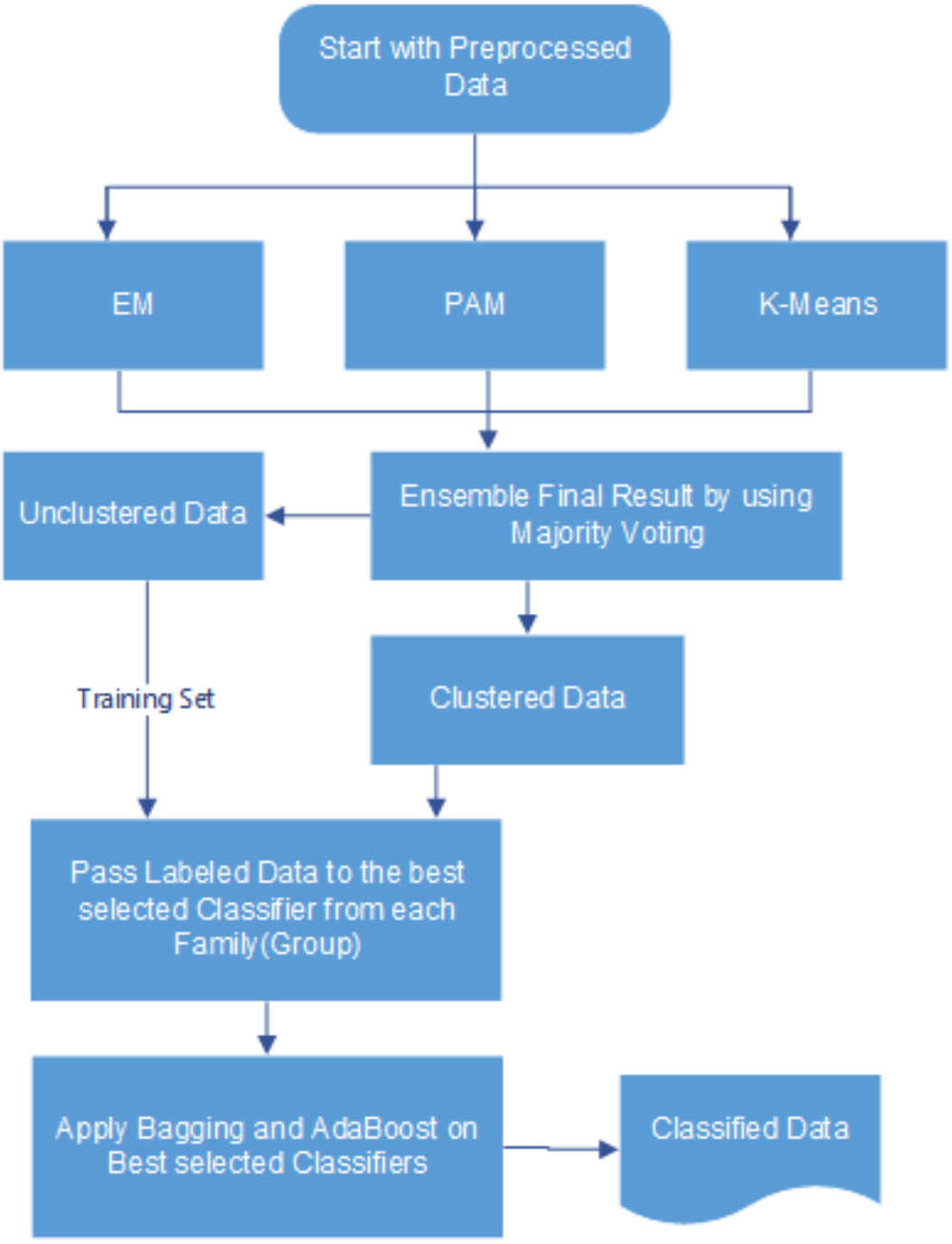

FIGURE 3.4: Proposed hybrid Approach

identification of high and low students but not with medium level students. K-means has clustered 125 high level ,120 low level and 112 medium level student out of 142,125 and 211 respectively. EM is showing 213 miss clustered data which is high rate.as our target is to identify low and medium student more than high level. PAM in this case performs out class from others as its identifies 112 highs, low and 127 mediums which comparatively good rate. Here k-means shows low rate of miss clustered data as compare to others but PAM has high rate if identify low level students.

So it is seen that algorithms perform well in different perspective. Here we have adopted ensemble clustering for the identification of groups so we can get benefits of different algorithm into one. Ensemble clustering methods merge results of multiple clustering algorithms to form core groups. Students are distributed among the groups by using ensemble clustering. The group obtained is by merging the result of three clustering algorithms such as PAM and K-means and EM (Expectation maximization). The Students are selected as a maximum agreement of all clustering algorithm. The students on which on which neither of clustering algorithm agreed would be selected as unclustered data.

TABLE 3.7: Result of Ensemble Clustering

Ensemble Clustered Data			**Unclustered**
H	**M**	**L**	
115	109	119	135

The above Table 3.7 shown that 343(115 H,109 M,119 L) out of 478 are form clustered and remaining 135 are considered unclustered data. Next step is to apply classification on the result of ensemble clustering. Classification (EMT) will be applied by considering clustered data as training set and unclustered data as testing set. The results are critically analyzed and compared with the base papers. Next step is to apply EMT classification on the result of ensemble clustering. The ensemble classification is introduced for targeting the unclustered data and refining the clustered result. This is used to select novel model for improvement of result and to reuse knowledge. Here clustered data is used as a training data

and classification algorithm will be trained on it. As mention before 343 instances are used in training phase and 135 the unclustered data. The unclustered data will be given a test data to assign them to one of the previously identified groups. It is observed that most performing classification algorithm from each family are BayesNet, Logistic, IBK, PART and Random forest. We have built the model by considering clustered 343 as training data and135 unclustered data as testing dataset. After selection of best classifier from each family. We have used five best classifiers with Booting and bagging methods as shown in table 3.8.

TABLE 3.8: Result of EMT Classification to Classify Unclustered Data

Best Algorithm from each Family	**Algorithm Name**	**Precision**	**AdaBoost (Precision)**	**Bagging (Precision)**
Bayes	BayesNet	0.252	0.617	0.201
Funtions	Logistic	0.351	0.351	0.495
Lazy	IBk	0.347	0.347	0.406
Rules	PART	0.393	0.228	0.08
Trees	Random Forest	0.106	0.108	0.124

After apply classification algorithm more instances are assigned to one of the groups. AdaBoost and bagging techniques have also applied on each algorithm. The result with best precision is selected and aggregated with the previous clustering result.We are using precision as performance metric as clustered data (training data) has 100% precision. In the ensemble clustering 115 students as high,109 as medium and 119 as low are correctly identified. Now more 17 students are in high,84 in medium and 1 in low are classified. So due to BayesNet+AdaBoost we are able to classify more 102 students. We have combine the result of clustered data and EMT classification in order to obtained aggregate result as shown Table 3.9.

TABLE 3.9: Aggregate Result of Hybrid Approach

	Clustered Data	**Unclustered Data**	**EMT Classification (BayesNet)**	**Hybrid Approach**
H	115	27	17	132
M	109	102	84	193
L	119	6	1	120
~	343	135	102	445
Accuracy				93%
Average precision				0.88

In this chapter, we have calculated the result of two approaches including EMT classification and hybrid approach.it is mention above that we have considered own proposed feature set and all classification algorithm from each family.it is beneficial to adopted the feature selection technique which evaluate features individually rather than group. Because group feature analysis can cause the influential features removed. EMT is being used widely to enhance the result. We have used two EMT methods as bagging and boosting. Random forest performs well here as it is ensemble approach by achieving highest accuracy of 79.7%. In the proposed hybrid approach, we have applied 3 clustering K-Means, PAM and EM. then we have ensemble the result of all used clustering algorithm dividing data in to clustered and unclustered. EMT Classification is applied on the result of ensemble clustering.so we have exploited the useful aspects of classification, clustering and EMT with achievement of 93% accuracy.

Chapter 4

Results and Evaluation

In the field of Educational Data Mining (EDM), researchers have made remarkable contributions that are quite beneficial for mining useful hidden patterns from the large institutes databases. Educational data mining is applied on different categories of dataset as face to face learning or online learning management system with varieties of features set. In this chapter, we evaluate our proposed work in order to address the research questions. The focus of this research is to accurately predict the student performance of first two semesters. This study is beneficial for the parents, teachers and educational institute by maintaining retention rate, improving student performance, reducing attrition rate, etc. It is easy to achieve such objectives when student performance can be predicted well before time.

Data collection and preprocessing are the initial and basic steps toward the research analyses. For this purpose, demographic, academic, behavioral and other features of LMS dataset are collected from Kaggle.com. The original dataset file is downloaded in the excel format containing 480 records with 17 features including class label. In the context of preprocessing, lack of missing values, outliers, lack of useful attributes can lead the research towards inaccurate result that might affect the overall performance. When this file is critically analyzed, there were multiple records seen. We have removed these records manually which resulted into 478 records. The focus of RQ1 is to establish a preprocessing approach. After dataset

collection, it has become necessary to convert the file into CSV in order to import into WEKA to apply filters for the preprocessing. After data collection and preprocessing; this section discusses the evaluation of feature selection and result. Best performing features are considered which are based on different classification algorithms. All experiments are conducted in the WEKA.

4.1 Evaluation and Comparison of Results

To evaluate the effectiveness of proposed model, we have employed four standards evaluation measures including precision, recall, F-measures and accuracy to evaluate result of proposed study.

1. **Recall**: This measure determines the ability of a classification model to identify all relevant instances
   Recall = true positive/true positive + false negative

2. **Precision**: This measure reports ability of a classification model to return only relevant instances.
   Precision = true positive/true positive + false positive

3. **F1 score**:It is a single metric that combines recall and precision using the harmonic mean.
   F1 = 2* Precision * Recall/ Precision + Recall

4. **Accuracy**: The accuracy can be defined as the percentage of correctly classified instances
   Accuracy = true positive + true negative/total

4.2 Feature Selection

The feature selection is performed in pre-processing phase of data mining and it is the focus of our RQ1 which states that Which of the two preprocessing approaches

produces better feature set? The methodology adopted to address this RQ and the results obtained have been mentioned in previous chapter. In this section, we are presenting the validation of our results regarding RQ1. First thing that we addressed is the difference in number of rows in [7] and [13] where the same dataset is shown to have 400 and 480 respectively. On manual observation of the dataset we found out that the existence of dirty data in [7] is not valid. Moreover, we found out that the rows removed in [7] are actually the ones that are misclassified when we apply the approach of [7]. The Table 4.1 below highlights some of the rows that are not included in the dataset of [7] and it can be seen that the data in these rows is just normal or matching with the other values in the dataset.

TABLE 4.1: Dataset Representing Misclassification

Instance No	**Gender**	**Semester**	**Topic**	**Relation**	**Predicted Class**	**Class**
105	M	F	IT	Father	L	M
89	F	F	IT	Father	M	L
84	M	F	IT	Father	L	M
109	M	F	IT	Father	L	M
214	M	S	Spanish	Father	M	L
81	M	F	Math	Father	M	L
99	F	F	IT	Father	L	M
269	M	F	English	Father	L	M
256	M	S	History	Father	L	H
367	M	F	Arabic	Father	L	M
379	M	F	Arabic	Father	M	L
333	F	F	French	Mum	H	M
129	M	F	IT	Father	M	L
244	M	S	Science	Father	L	M

However, their predicted and actual class labels are different and to us there is no valid reason to remove these rows from the dataset and also the same reason caused a high accuracy of approach of [7]. We also found out 2 duplicated rows in

the dataset as shown in Table 4.2

TABLE 4.2: Duplication of Records

Instance No	Gender	Semester	Topic	Relation	Predicted	Class
323	F	F	Spanish	Mother	M	L
324	M	S	French	Father	L	L
325	M	F	French	Father	M	M
326	M	S	French	Father	M	M
327	M	F	French	Father	L	L
328	M	S	French	Father	L	L

The comparison of finally selected rows and columns in [7], [13] and from our pre-processing approach is given below: Table 4.3 shows rows comparison of selected rows and column with our selection.it is clear that we have adopted smarter dataset with low of attribute.

TABLE 4.3: Total Number of Rows and Column

Dataset	Actual	Base paper 1 [9]	Base paper 2 [13]	our Approach
No of records	500	400	500	500
No of Features	17	12	12	9

This section evaluates the impact of each feature for students performance perdiction. 16 features and a class label are selected for experiments after application feature selection technique. This work used all classifiers comprehensively to find the best predictors for the desired students performance. We adopted Pearson Correlation feature selection techniques with ranker method. Afterwards backward feature selection method is used by analysing each feature effect. It is also seen best combination of feature subset contributes in increasing accuracy. This research focuses on high accuracy with best features selection and best data mining technique.

TABLE 4.4: Comparison of Features Used in All Approaches

S.No	Name of Features	Features Used		
		12 Features	12 Features	9(Proposed) Features
1	Gender	✓		✓
2	Nationality			
3	BirthPlace			
4	Relation	✓		✓
5	StageID	✓		
6	Semester	✓		
7	Topic	✓		
8	SectionID		✓	
9	GradeID		✓	
10	Studen tAbsence Days	✓	✓	✓
11	Raised Hand	✓	✓	✓
12	Visited Resource	✓	✓	✓
13	Announcemnent View	✓	✓	✓
14	Discussion	✓	✓	✓
15	ParentAnwering Survey	✓	✓	✓
16	ParentSchool Satisfaction	✓	✓	✓
17	Class	✓	✓	✓

4.2.1 Feature Analysis

RQ1: Which of the two preprocessing approaches produces better feature set?

This research also evaluated the features effect with 478 records, Table 4.5 shows the result of conduction of experiment. The highest influential feature is visited resource 0.37 and lowest is section ID 0.037 as shown in Table 3.2. These features selected as result of Pearson Correlation are used by all classifiers of each family. The result of each classifier is evaluated on the bases of accuracy. After that the features with the lowest rank is removed and again all classifiers of each family were applied and so on. This process is repeated until the stable and high accuracy is achieved.

As compared to features selected by proposed work to the base papers [13], 12 features were used with the elimination of demographic group. whereas in [7] it

is seen individual features are selected on the bases of correlation with response variable. It is seen that demographic features are included with high correlation such as gender and relation contributes in high accuracy. In this research we have also adopted techniques used in [7] but with different set of records. Nationality and birthplace are not used in all approaches as they are contribution in the achievement of high accuracy.S o it is not wise to eliminated whole group containing useful features. Considering individual feature according to their effectiveness is more reliable approach. In Selection of features refer back to chapter 3, give a table comprising all attributes and every attribute contains a tick selected by each approach shown in Table 4.4

. In this research, we have adopted smarter dataset with the low number of features with achievement of high accuracy. We have used proposed approach by applying whole family of classification including 23 algorithms. It is done to make comparison of features adopted by proposed approach or their approaches in [7] 12 features with a response variable are used as shown in table 4.4. When these features are used in our approach random forest achieved highest accuracy with 78.2%. When features used in [13] are applied in our approach, again Random Forest shows 74.4%, the best result from all of other classification algorithm. So as compared to our features selection techniques, only nine features with a response variable are selected showing high result of 79.7%. Our proposed features are smarter than as compared to other approchec. It also found that Random Forest is performing best in al comparison. When we looked in a perspective of average, again our approach shows highest average with 68.6% as shown in Table 4.5. However our approch has given average a little more as compared to other approch[7] but with smarter dataset .

In this study, our proposed features have performed better. As compared to other approach proposed reduced features set have highest accuracy on the original dataset. We have used comprehensive classifier to select the feature on the basis of accuracy after applying Pearson correlation approach. We have not depend upon on specific data mining classifier but used all classifiers from each family. When whole set of classifier from each family is applied, Random forest has achieved

Table 4.5: Comparison of Proposed Approach With Others

Algorithm Family	**Algorithm Name**	**No of Features w.r.t Base Paper**		
		12 Features	**12 Features**	**9 Features (Our Approach)**
Bayes	BayesNet	71.5	69	72.1
	NaiveBayes	69.4	66.9	70.7
	Nave Bayes multinominal Test	44.1	44.1	44.1
	Nave Bayes updateable	69.4	66.9	70.7
Funtions˜	Logistic	76.7	72.8	74.4
	Multilayer Perceptron	76.1	69.6	73.2
	Simple Logistic	75.3	72.5	72.8
	SMO	76.9	73.4	74.2
Lazy	IBk	69.6	61.7	74.4
	KStar	72.1	69.8	72.1
	LWL	68.2	64.6	70.2
Rules˜	Decision Table	69	64	69
	JRip	73.2	66.9	73.2
	OneR	60.4	60.4	60.4
	PART	70	67.9	73.4
	ZeroR	44.1	44.1	44.1
Trees˜	Decision Stump	52	52	52.2
	Hoeffding Tree	69.4	67.3	70.9
	J48	73	69	72.3
	LMT	75.3	72.8	72.5
	Random Forest	**78.2**	**74.4**	**79.7**
	Random Tree	67.9	70	69.4
	REPTree	65.2	64.4	73.8
AVERAGE		68.13	65.41	**68.68**

74.4 and 78. 2 and 79.9%. Result has shown highest of all with 79.7%. We have calculated overall average, of all approaches achieving 65.4, 68.1 and 68.6. We have achieved highest average with smarter dataset.

4.3 Data Mining Techniques

As it is discussed earlier the use of different data mining techniques for different aspects.in this section we are going to evaluated the result of the proposed work.in weka we have calculated the results which is an open source tool. **RQ2:** How can we combine the strengths of EMT and Hybrid approaches to form a better prediction approach?

4.3.1 Evaluation of Proposed EMT Approaches

EMT plays essential role in prediction student performance with the achievement of high accuracy. The accuracy is calculated on the basis of applying classification algorithms from all families of classifiers available in WEKA. In the first experiment, we applying all classification algorithm from each family on reduced features set. As there are various classification algorithms in machine learning which have different assigning label capabilities. It is mention above that we have presented our proposed approach that is the solution to the RQ2. In the first we have used our proposed features and then EMT is applied on it. As the dataset has become smarter due to the proposed features. After application of EMT it is seen that this smart dataset has produces better overall result.

It is observed that the proposed features have performed well in EMT classification technique. In [7] 12 features were used with 400 records and achieved highest accuracy of 78.2. when it is compared to our approach it has achieved accuracy of 79.7. this is not a big difference but the contribution is smarter feature set compromised of 9 features and a response variable. Bagging and boosting methods have produced more result. Result is increased from 71.5 to 73 in Bayes in boosting whereas in bagging it is increased from 71.5 to 72.8,69.6 to 71.9 in BayesNet and IBK respectively. We have gained more accuracy with smarter dataset. Whereas our result is gained from 72.1 to 74.6,73.4 to 76.7 in bagging. Where boosting doesnt show any improve in our case. In [13] there was different feature set in groups containing 12 features with 74.4 accuracy. When we applied EMT on it,

TABLE 4.6: Comparison of Proposed EMT Approach

Algo Family	**Algo**	**Comparison of Result w.r.t EMT Approach**								
		BP	**Boosting**	**Bagging**	**BP**	**Boosting**	**Bagging**	**PA**	**Boosting**	**Bagging**
Bayes	Bayes Net	71.5	73	72.8	69	69.8	70.9	72.1	72.1	74.6
Funtions	Logistic							74.4	74.4	73.8
	SMO	76.9	76.9	75.5	73.4	73.6	74.6			
Lazy	IBk	69.6	69.6	71.9				74.4	74.4	74.4
	KStar				69.8	66.7	69.4			
Rules	JRip	73.2	73.2	72.8						
	PART				67.9	73.0	73.2	73.4	74.2	76.7
Trees	Random Forest	78.2	77.8	78	74.4	74.6	74.2	**79.7**	77.6	78.6
Average		73.8	74.1	74.2	70.9	71.5	72.4	74.8	74.5	**75.6**

boosting method has improved result from 69 to 69.8,73.4 to 73.6 and 67.9 to 73.0 in bayesNet, function and rule respectively. Result is increased from 69.8 to 70.9 in bayesNet, 73.4 to 74.6 in SMO, 67.9 to 73.2 in PART by Bagging method.it is observed that EMT has not affected the random forest result. Random forest has an outstanding performance with achievement of highest result. When we have taken the average of all the result of 3 approaches, again proposed approach is performing with 75.6 highest in Bagging method.Result is shown in Table 4.6.

4.4 Evaluation of Hybrid Approach

In this section we will evaluate the proposed hybrid approach. we have adopted ensemble clustering with EMT classification for the identification of students. The k means, EM and Partitioning Around Medoids (PAM) algorithms are used in the ensemble clustering. K-means clustering works by assigning the data points among k clusters. Where PAM characterizes clusters by their mediods (centers).EM assigns a probability distribution to each instance which indicates the probability of it belonging to each of the clusters. In figure 3.6, it is observed that K-means has

perform well with k=3. And it has less number of unclustered data Means when compared to the original class label, low level students have clustered less. Low and medium student are the target of this work. PAM performs well with high and low but poor for medium level students. EM has high ratio of unclustered data comparatively with other clustering algorithm. Ensemble clustering methods merge results of multiple clustering algorithms to form core groups.

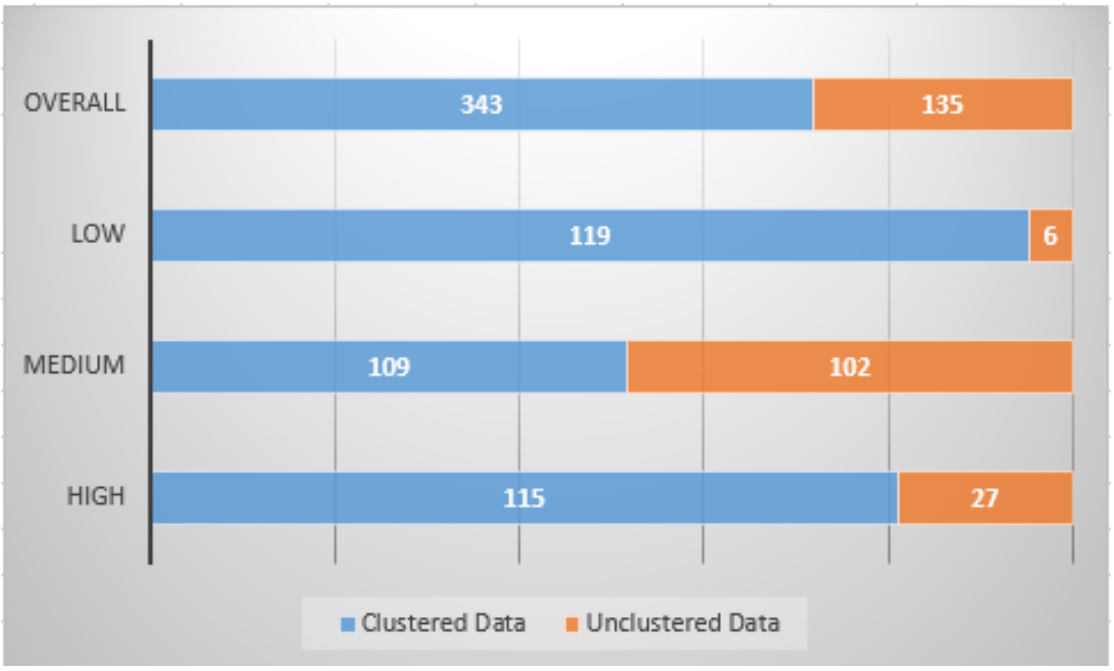

FIGURE 4.1: Result of Ensemble clustering

The group obtained is obtained by merging the result of three clustering algorithms such as PAM and K-means and EMT. The Students are selected as a maximum agreement of all clustering algorithm. The students on which on which neither of clustering algorithm agreed would be selected as unclustered data. Here we have applied ensemble clustering which gives 343 clustered data and 135 unclustered data.More than half of medium level student are unclustered.82% of the high student are identitfied.it is observed mostly medium level students are left to be identified. Overall 71% data has been clustered and remaining uncluttered as shown in 4.1. The ensemble classification is introduced for targeting the unclustered data and refining the clustered result. Now according to the adopted approach remaining 144 have used as testing dataset and 334 instances as training dataset in next phase. This is used to select novel model for improvement of

result and to reuse knowledge. Here clustered data is used as a training data and classification algorithm will be trained on it as shown in 3.9. After application of classifiers we also have applied AdaBoost and Bagging methods. bayesNet performs outstanding in all of others by classifying more 102 students. So over all 445 students are correctly identified out of 478 students.

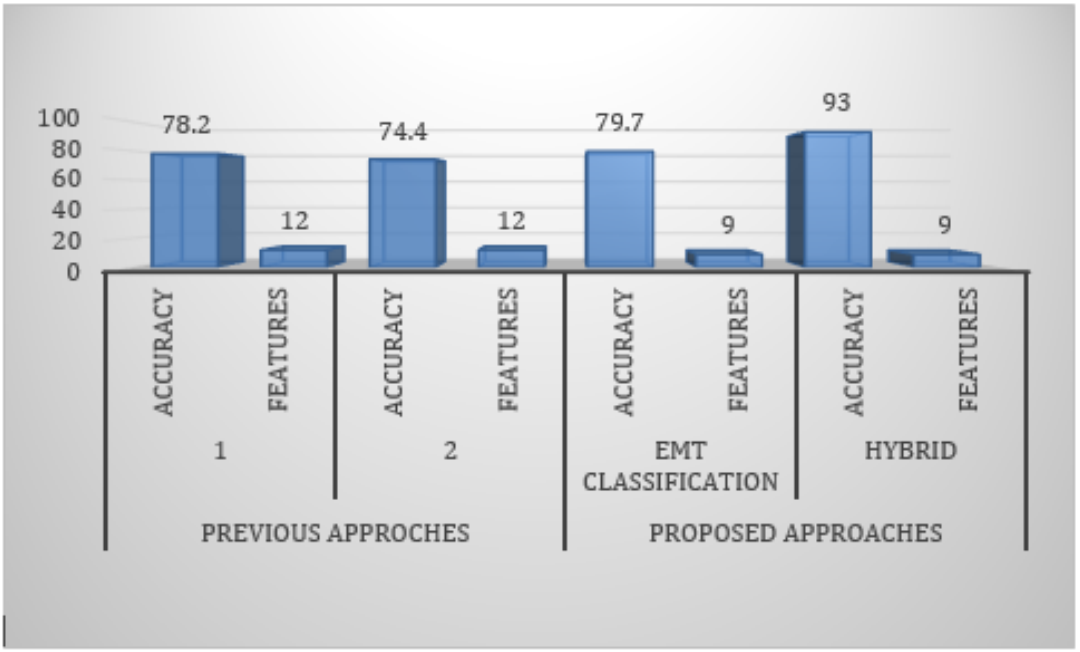

FIGURE 4.2: Comparison of All Approaches

In the below figure 4.2 all approaches are compared with respect to accuracy and features. In the 1st approach with 12 features set, its accuracy is 78.2%. where the 2nd approach abstained 74.4% with 12 different feature set. When this result is compared with our proposed approaches. MT show better result from these obtaining 79.7%. Hybrid approach has done excellent job with the achievement of accuracy of 93%.

In this chapter, we have evaluated the proposed work with others approaches. We have proposed smarter features set and make a comparison with the features. The proposed features set has shown high result as compared to other approaches. After feature selection, we have applied our proposed approaches as EMT classification and hybrid. We have evaluated the performance of approaches on smarter dataset. When proposed approaches result is compared, it has proved that they have produced better result up to 79.7%. Result is also evaluated by using hybrid

approach. Our both proposed approaches EMT classification and hybrid have perform better than other previous approaches.in this work hybrid has given outstand performance.it has enhance result from 79.7% to 93%.

Chapter 5

Conclusion and Future Work

In this chapter, we are finally concluding the work that we have carried out and presented in detail in the previous four chapters. This chapter also elaborates the directions in which we would recommend to extend this work in future.

5.1 Conclusion

The students performance prediction under the umbrella of Educational Data Mining (EDM) is an effective process as it helps to identify in advance the students who are at risk due to their unsatisfactory performance. This early identification can help to take appropriate measures and ultimately save students from being dropped or failed. Some of the appropriate measures that can be taken for such students include counselling, taking tutorials, extra time and even recommending appropriate courses. An evidence of the significance of performance prediction is the huge amount of work that is being carried out in this field. These approaches mainly perform steps included in the data mining process, use different classification algorithms and present their results by applying their approach on different data sets. These approaches can be a great help for an educational institution and consequently also for the students and parents. However, it is equally critical that the approach that is used for performance prediction is dependable and it has

been analyzed. An approach that has flaws in its steps may produce misleading results leading to wrong decisions.

In this study, we have analyzed the proposed approaches and claimed results of three different performance prediction approaches. One approach claimed high accuracy but missed some valuable data during pre-processing. If we include those missed rows in the original dataset then the accuracy does not remain that high.we have used the original dataset by including the missed valuable data and achieved. We have also analyzed that the features selection techniques and used the selected the better and smarter dataset.The other approaches made selection of classifiers without testing the rest or without proper justification of their selection.In our case we have used the comprehensive set of classifiers and explore their strength. Here the result of the classifiers is compared and best performing classifiers are selected.We have also analyzed different ensemble methods in both case either classification or clustering.As ensemble methods combines the strength of different algorithm in to one.Both classification and clustering have different end result.We have combined the strength of Classification ,Clustering and ensemble method by proposing hybrid approach.this latest hybrid approach seems to be more effective than others by producing better result results than the previous ones.

In the future we would like to extend this work by testing it rigorously on more datasets and applying deep learning for the prediction.further more feature selection techniques could be used and analyzed the result when used in hybrid approach.

Bibliography

[1] F. Ahmad, N. H. Ismail, and A. A. Aziz, "The prediction of students academic performance using classification data mining techniques," *Appl. Math. Sci.*, vol. 9, no. 129, p. 64156426, 2015.

[2] Y. Ma, C. Cui, X. Nie, G. Yang, K. Shaheed, and Y. Yin, "Pre-course student performance prediction with multi-instance multi-label learning," *Science China Information Sciences*, vol. 62, no. 2, pp. 29–101, 2019.

[3] S. Kalaivani, B. Priyadharshini, and B. S. Nalini, "Analyzing students academic performance based on data mining approach," *Int. J. Innov. Res. Comput. Sci. Technol.*, vol. 5, no. 1, pp. 194–197, 2017.

[4] E. A. Amrieh, T. Hamtini, and I. Aljarah, "Mining educational data to predict students academic performance using ensemble methods," *International Journal of Database Theory and Application*, vol. 9, no. 8, pp. 119–136, 2016.

[5] A. Mishra, R. Bansal, and S. N. Singh, "Educational data mining and learning analysis," *Proc. 7th Int. Conf. Conflu. 2017 Cloud Comput. Data Sci. Eng.*, vol. 22, no. 6, pp. 491–494, 2017.

[6] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, "Analyzing undergraduate students' performance using educational data mining," *Computers & Education*, vol. 113, no. 8, pp. 177–194, 2017.

[7] A. Almasri, E. Celebi, and R. S. Alkhawaldeh, "Emt: ensemble meta-based tree model for predicting student performance," *Scientific Programming*, vol. 20, no. 6, pp. 10–15, 2019.

[8] A. Daud, N. R. Aljohani, R. A. Abbasi, M. D. Lytras, F. Abbas, and J. S. Alowibdi, "Predicting student performance using advanced learning analytics," *In Proceedings of the 26th international conference on world wide web companion*, vol. 7, no. 4, pp. 415–421, 2017.

[9] F. Marbouti, H. A. Diefes-Dux, and K. Madhavan, "Models for early prediction of at-risk students in a course using standards-based grading," *Computers & Education*, vol. 103, no. 6, pp. 1–15, 2016.

[10] M. Atherton, M. Shah, J. Vazquez, Z. Griffiths, B. Jackson, and C. Burgess, "Using learning analytics to assess student engagement and academic outcomes in open access enabling programmes," *Open Learning: The Journal of Open, Distance and e-Learning*, vol. 32, no. 2, pp. 119–136, 2017.

[11] H. Al-Shehri, A. Al-Qarni, L. Al-Saati, A. Batoaq, H. Badukhen, S. Alrashed, J. Alhiyafi, and S. O. Olatunji, "Student performance prediction using support vector machine and k-nearest neighbor," *2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE)*, vol. 11, no. 5, pp. 1–4, 2017.

[12] M. F. Marbouti and H. A. Diefes-Dux, "Building course-specific regression-based models to identify at-risk students," *age*, vol. 26, no. 1, pp. 6–14, 2015.

[13] B. K. Francis and S. S. Babu, "Predicting academic performance of students using a hybrid data mining approach," *Journal of medical systems*, vol. 43, no. 6, p. 162, 2019.

[14] S. Bharara, S. Sabitha, and A. Bansal, "Application of learning analytics using clustering data mining for students disposition analysis," *Education and Information Technologies*, vol. 23, no. 2, pp. 957–984, 2018.

[15] B. Sana, I. F. Siddiqui, and Q. A. Arain, "Analyzing students academic performance through educational data mining," *3C Tecnol. innovacin Apl. a la pyme*, vol. 9, no. 4, pp. 402–421, 2019.

[16] M. S. Nurafifah, S. Abdul-Rahman, S. Mutalib, N. H. A. Hamid, and A. M. Ab Malik, "Review on predicting students graduation time using machine learning algorithms," *International Journal of Modern Education and Computer Science*, vol. 11, no. 7, p. 1, 2019.

[17] P. P. R. M. Anusha, K. Karthik and V. Srikanth, "Prediction of student performance using machine learning," *Int. J. Eng. Adv. Technol.*, vol. 8, no. 6, pp. 247–255, 2019.

[18] S.-S. M. Ajibade, N. B. Ahmad, and S. M. Shamsuddin, "A data mining approach to predict academic performance of students using ensemble techniques," *International Conference on Intelligent Systems Design and Applications*, vol. 4, no. 7, pp. 749–760, 2018.

[19] P. Veeramuthu, R. Periyasamy, and V. Sugasini, "Analysis of student result using clustering techniques," *International Journal of Computer Science and Information Technologies*, vol. 5, no. 4, pp. 5092–5094, 2014.

[20] Purba, W. and Tamba, S. and Saragih, Jepronel, "The effect of mining data k-means clustering toward students profile model drop out potential," *Journal of Physics: Conference Series*, vol. 1007, no. 1, pp. 1–012049, 2018.

[21] S. M. A. M. Gadal and R. A. Mokhtar, "Anomaly detection approach using hybrid algorithm of data mining technique," *2017 International Conference on Communication, Control, Computing and Electronics Engineering (ICCCCEE)*, vol. 8, no. 13, pp. 1–6, 2017.

[22] H.-S. Park and C.-H. Jun, "A simple and fast algorithm for k-medoids clustering," *Expert systems with applications*, vol. 36, no. 2, pp. 3336–3341, 2009.

[23] U. Agrawal, D. Soria, C. Wagner, J. Garibaldi, I. O. Ellis, J. M. Bartlett, D. Cameron, E. A. Rakha, and A. R. Green, "Combining clustering and classification ensembles: A novel pipeline to identify breast cancer profiles," *Artificial intelligence in medicine*, vol. 97, no. 4, pp. 27–37, 2019.

[24] Alasadi, S. A and Bhaya, S. Wesam, "Review of data preprocessing techniques in data mining," *Journal of Engineering and Applied Sciences*, vol. 12, no. 16, pp. 4102–4107, 2017.

[25] Sekeroglu, B. and Dimililer, K. and Tuncal, "Student performance prediction and classification using machine learning algorithms," *Proceedings of the 2019 8th International Conference on Educational and Information Technology*, pp. 7–11, 2019.

[26] Muhamad, Sari, I. and Maseleno, A. and Satria, F. and Muslihudin, "Application model of k-means clustering: insights into promotion strategy of vocational high school," *International Journal of Engineering & Technology*, vol. 7, no. 2.27, pp. 182–187, 2018.

[27] Han, M. and Tong, M. and Chen, M. and Liu, J. and Liu, Chunmiao, "Application of Ensemble Algorithm in Students' Performance Prediction," *2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, vol. 7, no. 2.27, pp. 735–740, 2017.

[28] d. Baker, R. S. Joazeiro and Inventado, P. Salvador, "Chapter X: Educational Data Mining and Learning Analytics," *Computer Science*, vol. 7, no. 2.27, pp. 1–16, 2014.

[29] Tomasevic, N. and Gvozdenovic,N. and Vranes,Sanja, "An overview and comparison of supervised data mining techniques for student exam performance prediction," *Computers & Education*, vol. 143, no. 2.27, pp. 103676, 2020.