Journal of Universal Computer Science, vol. 28, no. 8 (2022), 777-798
submitted: 24/8/2021, accepted: 6/7/2022, appeared: 28/8/2022 CC BY-ND 4.0

A Novel Hybrid Ensemble Clustering Technique for Student Performance Prediction

Sanam Fida
(Department of Computer Science,
Capital University of Science and Technology, Islamabad, Pakistan
https://orcid.org/0000-0003-1197-8189, sonu1668@yahoo.com)

Nayyer Masood
(Department of Computer Science,
Capital University of Science and Technology, Islamabad, Pakistan
https://orcid.org/0000-0002-3450-144X, nayyer@cust.edu.pk)

Nirmal Tariq
(Department of Computer Science,
Capital University of Science and Technology, Islamabad, Pakistan
https://orcid.org/0000-0002-3713-645X, nirmal.tariq@cust.edu.pk)

Faiza Qayyum
(Jeju National University, Jeju, Republic of Korea
https://orcid.org/0000-0001-9597-2387, faizaqayyum_99@yahoo.com)

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. Among a vast number of approaches available in this field, this study mainly focuses on generating a smarter dataset through reduced feature set without compromising the number of records in it and then producing an approach which combines the strengths of classification and clustering for better prediction results. In this study it has been identified that individual features have distinct effect and that removing misclassified data can affect the overall results. Backward selection is adopted using Pearson correlation as a metric to produce smarter dataset with lesser attributes and better accuracy in prediction. After feature set selection, we have applied EMT (Ensemble Meta-Based Tree Model) classification on it to identify best performing classifiers from five families of classifiers. In hybrid approach, first the ensemble clustering is applied on smart dataset and then EMT classification is applied to reevaluate the un-clustered data, which gives a boost in performance and provides us an accuracy of 93%.

Keywords: Student academic performance, Educational data mining, Ensemble clustering, Ensemble classification
Categories: I.5.3, I.5.5
DOI: 10.3897/jucs.73427

1 Introduction

Students are deemed as main stakeholders of an institute as they contribute in economic and social growth of a country, which leads towards production of creative graduates, innovators and entrepreneurs [Ahmad, 15]. High retention rate of an institute is considered as a core element for determining academic excellence [Ma 19]. Student retention is not only prime concern of educational institutions but also positively contributes towards student's bright future. Growth ratio of students admitted to higher educational institutions has increased from 2.73% in 2003 to 9.927% in 2015 (UNESCO, 2017). At the same time, the tremendous growth has imposed a persistent pressure over higher educational institutions to provide top-notch services to the students in order to minimize student's attrition rate [Kalaivani 17].

Attrition of a student means a student leaving an institute without completing his/her degree. Around 45% of the students quit their education prior to degree completion, which becomes a major cause of emotional stress and financial loss for student community [Amrieh 16]. There could be various reasons behind increased attrition rate such as trouble while learning, distance from home to institute, poor academic record etc. These reasons are broadly categorized into three categories: demographic, pre-qualification and institutional attributes. Demographic attributes contain age, gender, financial status, balance due, permanent address, residential address, guardian, father qualification, father occupation etc. Pre-college attributes comprise of secondary school grade, higher secondary grade, SAT score, previous college, previous-board and institutional attributes enlist current academic situation of students, like, program of study, CGPA (Cumulative Grade Point Assessment), first semester courses, initial major, current major, etc.

Scientific community has employed traditional statistical measures or data mining techniques to identify reasons behind the attrition [Asif 17] [Daud 17]. These studies are conducted under the paradigm of educational data mining (EDM) wherein a set of attributes pertaining to aforementioned categories are scrutinized to reveal determinants causing increased attrition rate. Data mining techniques are used as analytical tools to extract the hidden knowledge in EDM [Marbouti 16]. 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 [Atherton 17].

Contemporary state-of-the-art on student performance prediction adopts clustering and classification based mechanisms. Some of the approaches produce good accuracy through classification while some produce good results with clustering. Researchers have combined classification and clustering to form a hybrid approach with an intention of producing increased accuracy [Al-Shehri 17] [Marbouti 15].

While conducting in-depth analysis of the existing literature, we came across two influential studies [Almasri 19] [Francis 19], wherein authors have categorized students into low, middle and high level so that student at a risk of being dropped out could be identified at the start of their academics. Individual features were analyzed for their impact on student performance prediction w.r.t different machine learning (ML)

classifiers including SVM, Nave Bayes, decision tree and neural network. They also proposed an ensemble meta based tree model (EMT)in [Amrieh 16] [Francis 19] [Ajibade 18] [Alasadi 17].

This study presents an enhancement of both the studies [Almasri 19] [Francis 19] by exploiting a set of potential features from both the approaches to examine their collective contribution in predicting student's performance. The other approaches have employed classification algorithms without considering their suitability to the problem being addressed while we have used the comprehensive set of classifiers to explore their strength in predicting student performance on the said dataset. Overall, this study focuses to identify that which of the two pre-processing approaches (threshold and category based) hold potential to produce effective feature set. Furthermore, we also investigate on how can we combine the strengths of EMT and Hybrid approaches to form a better prediction approach. We have employed "Students' Academic Performance Dataset" that enlists activities of around 500 students in a Learning Management System (LMS) environment that contains 17 attributes which have been divided into 4 categories [Francis 19]. Outcome of this study reveals that proposed hybrid method has produced better results than existing approaches.

The rest of the paper is organized as follows: Section II elaborates the current state-of-the art, Section III presents a comprehensive methodology to address the existing gap. Section IV presents the experiments and results obtained by applying proposed scheme. Section V compares the evaluation results with state-of-the-art and Section VI concludes the paper with some future tasks.

2 Related Work

Scientific community has presented a plethora of EDM based studies that adopt traditional statistical measures while some have enriched the approaches by incorporating evolving concepts from the paradigm of data mining and machine learning [Ahmad, 15]. This section encompasses details about contemporary state-of-the-art focusing on educational data mining (EDM) problems related to student performance prediction in order to improve retention rate in educational institutions.

In the study [Ajibade 18], authors built a statistical model to predict retention of college students. The attributes like demographic, student's understanding about particular course, enrolment time were extracted from the record of 9200 students. The data was collected from the span of four years. The employed statistical models were correlation and logistic regression model. The outcomes suggested that attributes like passing development courses, taking Internet courses and participating in the Student Support Services program were potential predictor attributes. Alkhasawneh and Hobson [Francis 19] have presented a machine learning based approach to improve retention rate in higher learning institutions. The main objective of this study was to identify the reasons that affect retention of newly admitted students. Authors suggested that retention rate could significantly be improved if certain strategies are adopted at the initial phase of student persistent in an organization. The study evaluated features like GPA, test score and total credit hours using decision tree, Bayesian classifier, neural network and Support Vector Machine (SVM) classifiers. Authors in [Bharara 18] proposed data mining-based model using freshmen student data to identify the aspects behind attrition of students. The dataset contained records of 432 students from

Department of MCA, VBS from which attributes like students' demographic information, past performance, address and contact details were extracted. The prediction model was built using ID3, C4.5, ADT algorithms. The ADT algorithm outperformed by attaining 82.8% precision rate. Among all the attributes, student's GPA and financial constraints had potential impact on student's retention. In [Marbouti 16], 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 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. ML model was constructed by using combination of best techniques. 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 leverages to improve academic outcomes [Atherton 17]. In [Francis 19] 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, Naïve 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 student's performance. This model could be extended for varieties of feature of student dataset. 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 [Sana 19] 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 Naïve Bayes and Decision Tree. It was indicated that academic assessment was a prominent factor when predicting such students. In [Nurafifah 19], 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.

Many other researches have also emphasized on the importance of feature set selection in performance prediction tasks. In [Alshabandar 20], the authors have examined the previous success rate of students to predict their performances in currently enrolled online courses. Two predictive models were introduced and examined. In the first model, called 'grade assessment model', regression analysis was applied to predict grades of students on the basis of dynamic behavioral features. In the second model classification analysis was performed to predict the final performance of students in the course using behavioral features, temporal features and demographic features as input. It was called as 'final students' performance model'. Same set of classifiers were used in both cases which include Random Forest, Multi-layer Perceptron, Neural Network with one hidden layer, Generalized Linear Model and Gradient Boosting Machine. The simulation results showed that the highest accuracy of 0.868 was achieved by Gradient Boosting Machine in classification problem while in regression problem Random forest acquire the lowest value of RSME which is 8.131. Classification model was considered more reliable on the basis of feature set selection in regression the temporal features could not be used with dynamic behavioral features.

The authors in [Aggarwal 21] have studied and compared the effect of academic and non-academic features on predicting students' performance at early stage. The dataset was gathered from a college in India. The academic features include Secondary school %age, higher secondary school %age, passing year, gap etc while non-academic features consist of demographic features such as Gender, Age, category etc. They applied same eight Machine Learning algorithms in 2 sets of experiments, one with only academic features and the other one with all features. Through experimental results they have suggested that performance prediction of students is highly by influenced demographic features so the combination of academic and non-academic features should be considered.

An analysis of related literature shows that student attrition is a critical issue for institutions, parents, students and for society at large. Different studies performed on different data sets have shown varying levels of attrition. However, the common factor highlighted through all studies is that it is a serious issue, and needs to be studied at institution level so that appropriate measures could be taken to improve student retention. To establish an approach that is applicable to a broader group of institutions, more work is required on the features set. Features like, family income, subjects understanding and others like this are not available in many institutions. There is a need to build an approach that works on features that are commonly available in educational

institutions. The work in this field should not only be limited to predicting the students attrition, the reasons for the attrition, but also the steps to reduce the attrition rate and improve the retention rate. The research community needs to work on these lines and share the experience so that others could benefit our of it.

3 Methodology

This section presents detailed analysis of methodical steps proposed to address following two research problems:

RQ1: Which of the two preprocessing approaches (threshold and category based) hold potential to produce effective feature set?
RQ2: How to combine the strengths of EMT and Hybrid approaches to form a better prediction approach?

The proposed study is very close to [Almasri 19][Francis 19] with intention of improving their results by incorporating the aspects that have previously been overlooked by state-of-the-art approaches. We have looked features selection in two different perspectives such as, individual or categorical way and then applied a better prediction approach accordingly. We have applied Pearson correlation feature selection technique to select the list of appropriate features to ensure the model's accuracy. Afterwards, classification and clustering are performed in hybrid mode and results are evaluated using standard evaluation measures. The architecture of the proposed methodology is shown in [Figure 1].

The foundation studies [Almasri 19] [Francis 19] have used the same dataset but with difference of features and records. In [Almasri 19] only 400 records with 13 features were considered whereas [Francis 19] has considered 480 records by dividing features into different groups and using these groups for the prediction. This study considers the original dataset rather than reduced dataset by excluding only missing and repeated records. In [Francis 19], authors have considered all the records and attained comparatively lower accuracy. In this study, we have applied classification including assembling as in [Almasri 19] but with different set of records. In [Francis 19] categorical features were used with application of hybrid approach including both classification and clustering. Also, this research includes individual features rather than categorical and applying ensemble and clustering based methods using hybrid approach. The objective of predicting student performance is to reduce the attrition rate which will ultimately be beneficial for the students, parents and educational institutes.

3.1 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 form containing redundancies like missing records, noisy data etc. In pre-processing step of [Almasri 19], dataset was reduced to 400 records because of missing and noisy data. Whereas in this paper, 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.

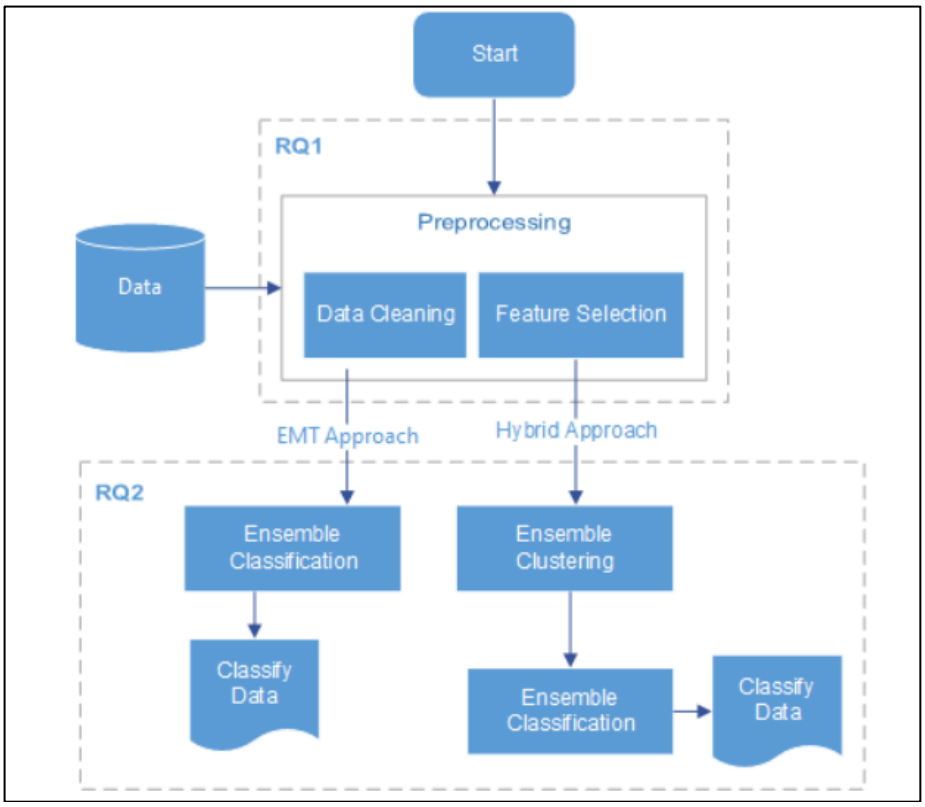

Figure 1: Flowchart of Proposed Methodology

3.1.1 Data Cleaning

In data cleaning, major task is to remove noise and irrelevant records, dealing with missing values, recognize outlier and correct inconsistent data [Alasadi 17]. The missing data can be handled through different filters. After missing data handling and outlier detection phase, final attributes have been selected to perform experiments. Such attributes can be seen in [Section 4.1]. The attributes belong to all three types of factors, demographic, pre-university, and institutional.

3.1.2 Feature Selection

As explained earlier, the approach in [Francis 19] uses groups of features to select best feature set and that in [Marbouti 16] 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. Furthermore, it does not exploit the real association of an individual attribute with the response variable. On the other hand, the threshold approach adopted in [Almasri 19] 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. Considering these issues, we have employed backward selection using Pearson Correlation (PC) as a metric for feature selection using individual attributes in prediction. The technique used for feature selection in shown in [Figure 2].

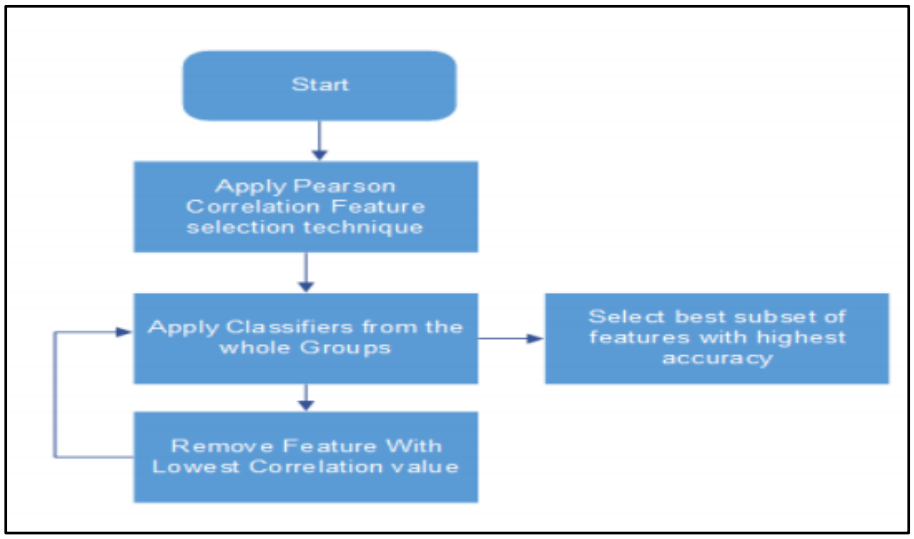

Figure 2: Technique used for Feature Selection

3.2 Proposed Approach

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 applied EMT based classification proposed in [Almasri 19] 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 [Purba 18], 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.

3.2.1 Proposed EMT based Classification

Ensemble Meta-Based Tree Model is an ensemble technique which combines multiple sets of weak learning classifiers into final prediction model either by using weighting or voting techniques [Almasri 19]. 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. The five families which contain different set of algorithms on the basis of like properties which are Bayes, Functions, Lazy, Rules and Trees. In proposed approach all algorithms from mentioned families are applied on our smart dataset from which best performing algorithm is selected manually. In the next

step bagging and boosting is applied on best performed five selected algorithms and highest result either from bagging or boosting is selected. Working of the proposed approach is shown in the [Figure 3].

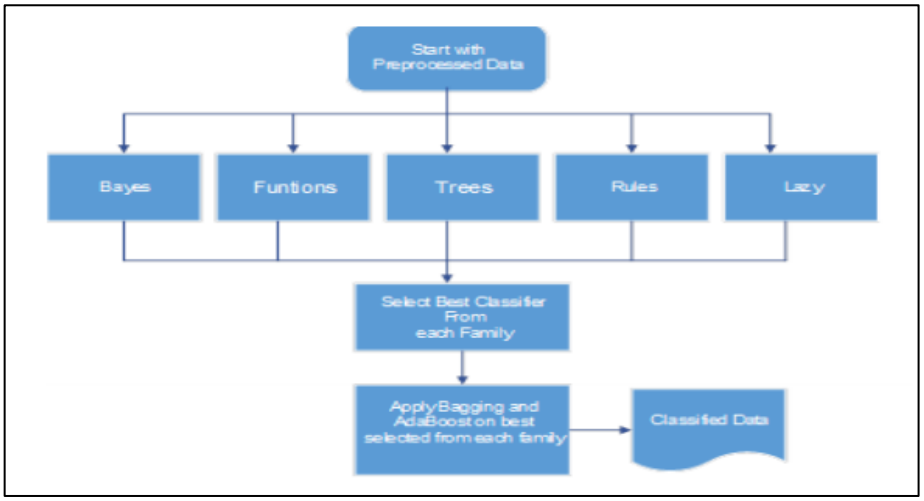

Figure 3: Proposed EMT approach

3.2.2 Proposed Hybrid Approach

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 algorithms are applied such as PAM (Partition Around Medoids), EM (Expectation Maximization) and K. Means. The result of these algorithms is aggregated on the basis of majority voting. As a result, clustered or un-clustered data are attained. In aggregation on which neither of three algorithms shows agreement is considered as un-clustered data. In the next step we provide this data to classification unit to classify more data. Here we have used EMT based classification based on our research. So that more data can be grouped. Here clustered data is considered as a training set and un-clustered data as a testing set. The proposed Hybrid approach is shown in the [Figure 4].

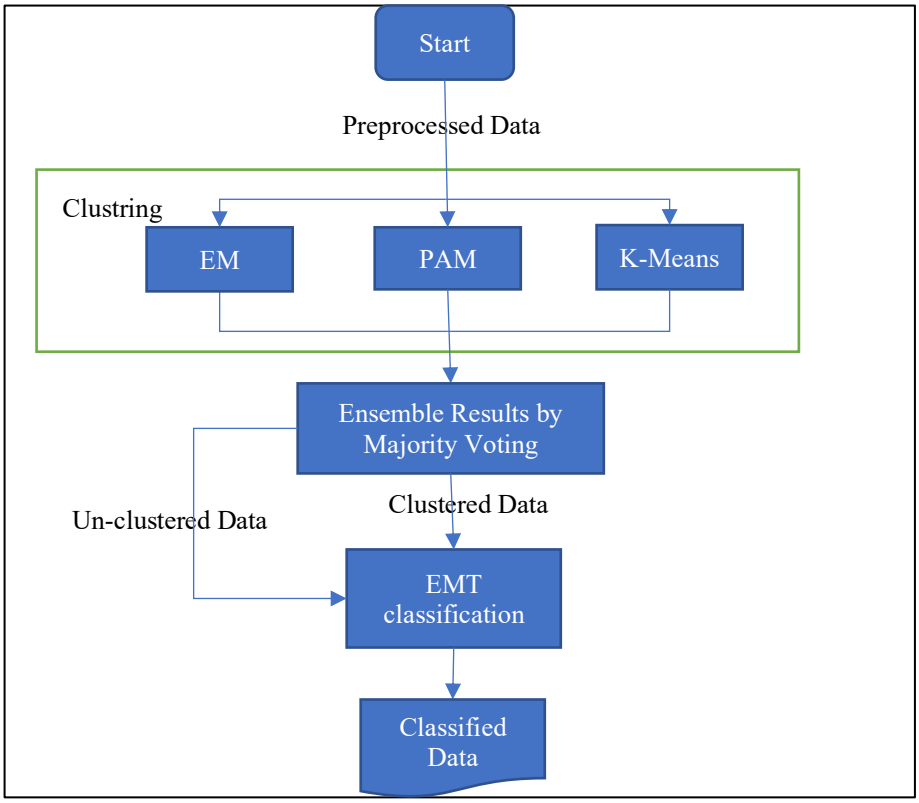

Figure 4: Proposed Hybrid Approach

4 Experimental Results and Analysis

The proposed methodology contains Feature Selection, EMT based Classification and Hybrid Prediction. Experimental details are mentioned in this section.

4.1 Data Set

Data set employed in this study is "Students Academic Performance Dataset" (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 [Francis 19]. 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 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 [Almasri 19],9,14] [Mishra 17]. [Table 1] illustrates description of features employed in this study.

Sr.No	Feature	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	Stage_ID	Educational stage of a student (high, medium, low)	Academic features
6	Semester	Student's 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-2……GL-12	
10	Student_Absence_Days	Students' educational stage (high, medium, low)	
11	Raised_Hands	These are all features concerned with student's behavior {while interacting with kalboard 360 E-learning websites}	Behavioral features
12	Visited_Resources		
13	Announcements		
14	Discussion		
15	Parent_Answering_Survey		Extra features
16	Parent_School_Satisfaction		
17	Class	A class Label	

Table 1: Description of Features Used in Student Performance Prediction

4.2 Feature selection

In order to select best features subset as mention above we have used Pearson Correlation. [Table 2] shows the PC values of all attributes with the response variable. It is observed that the attribute Visited Resource is highly correlated with 0.37 value and at the same time Section ID is least correlated to response variable with value of 0.037. We have applied multiple classifiers to find the best set of features through prediction values against multiple attributes. First, we applied the technique on complete set of attributes and then we removed the lowest correlated features and applied the classifiers. We have applied this process iteratively to obtain the minimum best possible set of attributes. The [Table 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 is selected for the further processing. The 9 attributes that have been selected through

this process are Visitedresource, StudentAbsentDays, RaisedHand, AnnouncementView, Relation, ParentAnsweringSurvey, ParentSchool-Satisfaction, Discussion, Gender and a response variable (Class).

Sr. 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

Table 2: Results of Pearson Correlation values of attributes with Response Variable

In our research we have proposed a smarter dataset with reduced number of features. The proposed feature set is better as features are selected on individual correlation rather than group wise selection. It is found that most of the influential features are removed by adopting group wise features approach which can contribute well in the prediction.

Sr. No.	Total attributes	Accuracy before	minimum PC	Accuracy After
1	17	-	-	78.0
2	16	78.0	Section ID	77.8
3	15	77.8	Grade ID	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

Table 3: Results of different subsets of Features

4.3 Results of EMT based Classification

Here we are proceeding with the solution 1 of the proposed work. When all classification algorithms from a group are applied it has shown variety of results. 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 well on ensemble techniques. It is also observed that most of the cases Tree groups performance is outstanding. [Table 4] shows the results.

Algorithm family	**Algorithm Name**	**Accuracy**
Bayes	BayesNet	72.1
	NaiveBayes	70.7
	Nave multinominalTest	44.1
Functions	Logistic	74.4
	Nave Bayes updateable	70.7
	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

Table 4: Accuracies of multiple classifiers on proposed dataset

We then applied ensemble approach and used bagging and boosting, in most of the cases result is improved. Here we have applied AdBoostM1 and Bagging on the best selected algorithm 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 5] it can be seen that result is increased from 72.1 to 74.6,73.4 to 76.7 in BayesNet and Rules. Whereas function, lazy and tree doesn't show any improvement in the result. Random forest does not show any improvement due to bagging but still its result is best from all of others. EMT based classification when applied on our proposed dataset has shown improvement.

Best Algorithm from Each Family	Algorithm Name	Proposed Approach	Boosting Method	Bagging Method
Bayes	BayesNet	72.1	72.1	74.6
Functions	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

Table 5: Results of Bagging and Boosting on best selected classifiers

4.4 Results of Hybrid Approach

In this approach ensemble clustring with EMT classification as hybrid approach for the prediction. K-means, PAM and EM algorithms are used in ensemble clustring. K-means clustering works by assigning the data points among k clusters. Where PAM characterizes clusters by their medoids (centres). EM assigns a probability distribution to each instance which indicates the probability of it belonging to each of the clusters. 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 [Bharara 18][Veeramuthu 14][Park 09]. 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 6] shows that k-means is performing well with 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 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 un-clustered 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 [Figure 5].

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	H	70	50
	Cluster 1	113	L	98	69
	Cluster 2	180	M	106	94
				274	213

Table 6: Results of Clustering algorithms

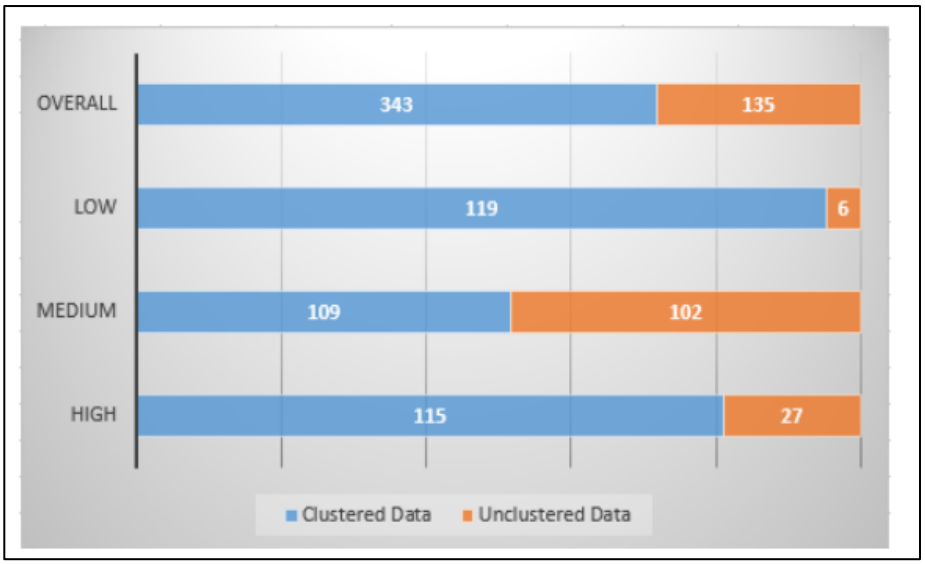

Figure 5: Results of Ensemble Clustering through majority voting

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. Here clustered data is used as a training data and classification algorithm will be trained on it. 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 items as training set and135 un-clustered items as testing set. After selection of best classifier from each

family. We have used five best classifiers with Booting and bagging methods as shown in [Table 7].

Selected Best Algorithm from each family	**Algorithm Name**	**Precision**	**AdaBoost (Precision)**	**Bagging (Precision)**
Bayes	BayesNet	0.252	0.617	0.201
Functions	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

Table 7: Result of EMT based Classification to classify Un-clustered Data

After the application of classification algorithms more instances are assigned to one of the groups. AdaBoost and bagging techniques are also applied on each algorithm. The result with best precision is selected and aggregated with the previous clustering result. We used precision as performance metric since clustered data (training data) had 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 102 more students were classified. We have combined the result of clustered data and EMT classification in order to obtained aggregate result as shown [Table 8].

	Clustered Data	Un-clustered 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

Table 8: Aggregate Result of Hybrid Approach

5 Comparison with State-of-the-art

In this section we have shown comparison with the two base papers [Almasri 19] and [Francis 19] in respect of Features used and data mining techniques that are applied in it.

5.1 Feature Analysis

The original dataset file consists of 480 records with 17 features including class label. After analysing and removing duplicate entries manually, we have used this dataset with 478 records for our research. As compared to features selected by proposed work

to the base papers [Francis 19], 12 features were used with the elimination of demographic group. whereas in [Almasri 19] it is seen individual features are selected on the bases of correlation with response variable. It is observed that demographic features are included with high correlation such as gender and relation contribute in high accuracy. In this research we have also adopted techniques used in [Almasri 19] 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. So, it is not wise to eliminate whole group containing useful features. Considering individual feature according to their effectiveness is more reliable approach. [Table 9] gives a comparison of features used in base papers with respect to our approach.

Sr.No	Name of Features	Features used		
		12 Features[Almasri 19]	12 Features [Francis 19]	9 Features (our approach)
1	Gender	✓		✓
2	Nationality			
3	Birthplace			
4	Relation	✓		✓
5	Stage_ID	✓	✓	
6	Semester	✓	✓	
7	Topic	✓	✓	
8	SectionID		✓	
9	Grade ID		✓	
10	Student_Absence_Days	✓	✓	✓
11	Raised_Hands	✓	✓	✓
12	Visited_Resources	✓	✓	✓
13	Announcements	✓	✓	✓
14	Discussion	✓	✓	✓
15	Parent_Answering_Survey	✓	✓	✓
16	Parent_School_Satisfaction	✓	✓	✓
17	Class	✓	✓	✓

Table 9: Comparisons of Features Used in base papers and proposed approach

In our research, we have adopted smarter dataset with lo number of features with achievement of high accuracy. We have evaluated the results by applying a whole family of classifiers consisting of 23 algorithms on our dataset as well as dataset used in [Almasri 19] and [Francis 19]. The results shown in [Table 10] depicts that our dataset in most of the cases out-performs the base approaches.

Algorithm family	Algorithm Name	No of Features w.r.t Base paper		
		12 Features[Almasri 19]	12 Features [Francis 19]	9 Features (our approach)
Bayes	BayesNet	71.5	69	72.1
	NaiveBayes	69.4	66.9	70.7
	Naïve Bayes multinominalTest	44.1	44.1	44.1
	Naïve 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		68.1304348	65.41304348	68.68695

Table 10: Comparison of Selected Feature Set with others

5.2 Evaluation of EMT Classification

In the proposed EMT based classification, we have used reduced feature set here with original dataset. As compared to the previous work [Almasri 19] and [Francis 19], our proposed approach has obtained better accuracy with a smarter dataset. To prove our point, we have applied ensemble techniques Boosting and Bagging on base papers feature set as well and comparison is shown in [Table 11]. We have used best classifier from each family as discussed earlier. It is observed that in different classifiers the behaviour is different but if we evaluate the average accuracy rate then there is visible increase as proposed approach gives highest accuracy of 75.6 in Bagging method. 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 grouping. It is also shown that ensemble based classification technique is

more productive than simple classification as it grasps the effective attributes of various approaches which enhances the overall result.

Sr. No	Best Algorithm from each family	Algorithm Name	No of Features w.r.t Base paper								
			Base Paper 1 [Almasri 19]	Boosting Method	Bagging Method	Base Paper 2 [Francis 19]	Boosting Method	Bagging Method	Proposed Approach	Boosting Method	Bagging Method
1	Bayes	Bayes Net	71.5	73	72.8	69	69.8	70.9	72.1	72.1	74.6
2	Functions	Logistic							74.4	74.4	73.8
		SMO	76.9	76.9	75.5	73.4	73.6	74.6			
3	Lazy	IBk	69.6	69.6	71.9				74.4	74.4	74.4
		KStar				69.8	66.7	69.4			
4	Rules	JRip	73.2	73.2	72.8						
		PART				67.9	73.0	73.2	73.4	74.2	76.7
5	Trees	Random Forest	78.2	77.8	78	74.4	74.6	74.2	79.7	77.6	78.6
Overall Average			73.8	74.1	74.2	70.9	71.54	72.46	74.8	74.54	75.6

Table 11: Comparison of proposed EMT on smarter dataset with base papers

5.3 Evaluation of Hybrid Approach

In this section we have evaluated the results of proposed hybrid approach in light of previous approaches. We have seen earlier in [Table 10] that Random Forest has given highest accuracy as compared to rest of the classifiers with proposed feature set. Also, EMT has its own benefits rather than simple classification. As described earlier ensemble clustring through majority voting played a vital role in prediction process. When Hybrid approach is used and EMT classification is used over Ensemble clustering the results increase drastically to accuracy of 93%. The results of the discussion can be seen in the [Figure 6].

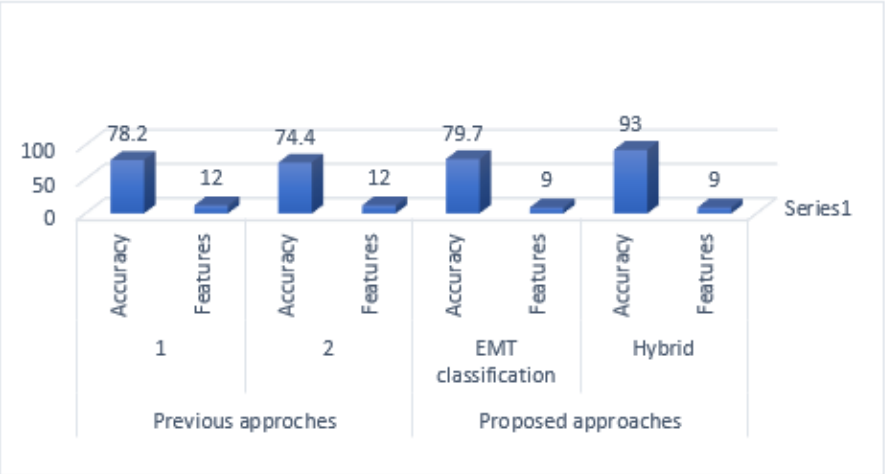

Figure 6: Overall Comparison with other approaches

6 Conclusion and Future Work

The student 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 out. 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.

In this research, the contribution towards predicting students' performance is made using a combination of tools from data mining fields. The research deals with 2 base research question: one includes the selection of features based on appropriate measure and other one mainly focused on combining two different techniques to form a better prediction approach. For feature selection, we have used pearson correlation where we have calculated the correlation of each feature with response variable. We have produced a smarter and better dataset with original 478 records and maintained a high accuracy rate which previous approaches could not get. To get better prediction results unlike the previous approaches we have used a comprehensive set of classifiers and explored their strength. We then selected the best performing classifiers among the family of classifiers. To enhance the accuracy, we have used ensemble methods during classification and clustering and it has improved the results. In hybrid approach we have explored the strengths of clustering and then boosts the performance by applying EMT on the clustered data. Through this approach we have attained an accuracy of 93% without compromising on the original dataset.

In future we would like to extend the research by testing it rigorously on more datasets and applying deep learning for the prediction. Also the strengths of hybrid architecture can be explored by using different feature selection techniques.

References

[Aggarwal 21] Aggarwal, D., Mittal, S., & Bali, V. (2021). Significance of Non-Academic Parameters for Predicting Student Performance Using Ensemble Learning Techniques. International Journal of System Dynamics Applications (IJSDA), 10(3), 38-49. http://doi.org/10.4018/IJSDA.2021070103

[Ahmad 15] Ahmad, F., Ismail, N. H., & Aziz, A. A. (2015). "The prediction of students' academic performance using classification data mining techniques". Applied Mathematical Sciences, 9(129), 6415-6426.

[Ajibade 18] Ajibade, S. S. M., Ahmad, N. B., & Shamsuddin, S. M. (2018, December). "A data mining approach to predict academic performance of students using ensemble techniques". In International Conference on Intelligent Systems Design and Applications (pp. 749-760). Springer, Cham.

[Al-Shehri 17] Al-Shehri, H., Al-Qarni, A., Al-Saati, L., Batoaq, A., Badukhen, H., Alrashed, S., Alhiyafi, J., & Olatunji, S. O. (2017, April). "Student performance prediction using support vector machine and k-nearest neighbour". In 2017 IEEE 30th Canadian conference on electrical and computer engineering (CCECE) (pp. 1-4). IEEE.

[Alasadi 17] Alasadi, S. A., & Bhaya, W. S. (2017). "Review of data preprocessing techniques in data mining". Journal of Engineering and Applied Sciences, 12(16), 4102-4107.

[Almasri 19] Almasri, A., Celebi, E., & Alkhawaldeh, R. S. (2019). "EMT: Ensemble meta-based tree model for predicting student performance". Scientific Programming, 2019.

[Alshabandar 20] Alshabandar, R., Hussain, A., Keight, R., & Khan, W. (2020, July). Students performance prediction in online courses using machine learning algorithms. In 2020 International Joint Conference on Neural Networks (IJCNN) (pp. 1-7). IEEE.

[Amrieh 16] Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016). "Mining educational data to predict student's academic performance using ensemble methods". International Journal of Database Theory and Application, 9(8), 119-136.

[Asif 17] Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). "Analyzing undergraduate students' performance using educational data mining". Computers & Education, 113, 177-194.

[Atherton 17] Atherton, M., Shah, M., Vazquez, J., Griffiths, Z., Jackson, B., & Burgess, C. (2017). "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, 32(2), 119-136.

[Bharara 18] Bharara, S., Sabitha, S., & Bansal, A. (2018). "Application of learning analytics using clustering data Mining for Students' disposition analysis". Education and Information Technologies, 23(2), 957-984.

[Daud 17] Daud, A., Aljohani, N. R., Abbasi, R. A., Lytras, M. D., Abbas, F., & Alowibdi, J. S. (2017, April). "Predicting student performance using advanced learning analytics". In Proceedings of the 26th international conference on world wide web companion (pp. 415-421).

[Francis 19] Francis, B. K., & Babu, S. S. (2019). "Predicting academic performance of students using a hybrid data mining approach". Journal of medical systems, 43(6), 1-15.

[Kalaivani 17] Kalaivani, S., & Nalini, S. (2017). "Analyzing student's academic performance based on data mining approach". International Journal of Innovative Research in Computer Science & Technology (IJIRCST) ISSN, 2347-5552.

[Ma 19] Ma, Y., Cui, C., Nie, X., Yang, G., Shaheed, K., & Yin, Y. (2019). "Pre-course student performance prediction with multi-instance multi-label learning". Science China Information Sciences, 62(2), 1-3.

[Marbouti 15] Marbouti, F., Diefes-Dux, H. A., & Strobel, J. (2015, June). "Building course-specific regression-based models to identify at-risk students". In 2015 ASEE Annual Conference & Exposition (pp. 26-304).

[Marbouti 16] Marbouti, F., Diefes-Dux, H. A., & Madhavan, K. (2016). "Models for early prediction of at-risk students in a course using standards-based grading". Computers & Education, 103, 1-15.

[Mishra 17] Mishra, A., Bansal, R., & Singh, S. N. (2017, January). "Educational data mining and learning analysis". In 2017 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence (pp. 491-494). IEEE.

[Nurafifah 19] Nurafifah, M. S., Abdul-Rahman, S., Mutalib, S., Hamid, N. H. A., & Ab Malik, A. M. (2019). "Review on predicting students' graduation time using machine learning algorithms". International Journal of Modern Education and Computer Science, 11(7), 1.

[Park 09] Park, H. S., & Jun, C. H. (2009). "A simple and fast algorithm for K-medoids clustering'. Expert systems with applications, 36(2), 3336-3341.

[Purba 18] Purba, W., Tamba, S., & Saragih, J. (2018, April). "The effect of mining data k-means clustering toward students profile model drop out potential". In Journal of Physics: Conference Series (Vol. 1007, No. 1, p. 012049). IOP Publishing.

[Sana 19] Sana, B., Siddiqui, I. F., & Arain, Q. A. (2019). "Analyzing students' academic performance through educational data mining".

[Veeramuthu 14] Veeramuthu, P., Periyasamy, D. R., & Sugasini, V. (2014). "Analysis of student result using clustering techniques". International Journal of Computer Science and Information Technologies, 5(4), 5092-5094.