Journal of Engineering Education Transformations,
Volume No 36, January 2023, Special issue, eISSN 2394-1707

10.16920/jeet/2023/v36is2/23003

A Review of Data Mining in Education Sector

Sunita M. Dol[1], Dr. P. M. Jawandhiya[2]
[1]Computer Science and Engineering, Pankaj Laddhad Institute of Technology and Management Studies, Buldhna, Maharashtra, India
[2]Computer Science and Engineering, Pankaj Laddhad Institute of Technology and Management Studies, Buldhna, Maharashtra, India

***Abstract*— Educational Data Mining (EDM) is one of the trending areas in which various researchers are working for the betterment of the student's performance. Predicting the students' performance is considered as an important task in education sector and is of paramount importance as predicting the performance accurately may lead to great future of students by analyzing data properly. This article presents the review of 32 research articles which are from ACM, IEEE, Springer and Elsevier research database. This article analyzes these research articles based on number of research articles considered from research database, publication year, performance parameters, number of performance parameteres used by research articles, Data Mining Techniques, number of algorithms used by research articles, and dataset size. It is found that classification technique is used in EDM for analyzing students' data and in classification technique, mostly employed algorithms are Random Forest, Logistic Regression, Decision Tree, Naïve Bays, Support Vector Machine and K-nearest Neighbour. Generally the performance parameters such as accuracy, precision, recall and F-measures are used to decide the performance of the classification algorithms. This review article will be helpful to those researchers who are working in the EDM for predicting students' performance for the dataset obtained from education sector.**

Keywords—Data Mining, Educational Data Mining, Classification, Clustering, Association Rule
JEET Category—Research

I. INTRODUCTION

Among the most important developments in computer-aided education over the past few years has been data mining, which deals with extracting useful information from huge amounts of publicly available data sets relating to settings of education. Recommender Systems include techniques and methodologies borrowed from other neighboring research areas, such as Information Retrieval and Human-Computer Interaction. They also typically include an algorithm that can be categorized as Data Mining (DM). The field of data mining is one of the newest and most promising fields that have attracted the attention of both scholars and industry experts. Different types of data mining tools are available which allow specialists to predict data behavior and patterns and make dynamic decisions based on them.

Educational institutions provide a best education to their students in order to process the learning. Education environments require student performance prediction and analysis.

This paper was submitted for review on August, 13, 2022. It was accepted on November, 16, 2022.
Corresponding author: Sunita M. Dol[1], Dr. P. M. Jawandhiya[2]
[1]Computer Science and Engineering, Pankaj Laddhad Institute of Technology and Management Studies, Buldhna, Maharashtra, India
[2]Computer Science and Engineering, Pankaj Laddhad Institute of Technology and Management Studies, Buldhna, Maharashtra, India
Address: Mrs. Sunita M. Dol, Assistant Professor, Computer Science and Engineering , Walchand Institute of Technology, Solapur P.B.No.634, Walchand Hirachand Marg, Ashok Chowk, Solapur - 413006 Maharashtra, India (e-mail: sunita_aher@yhoo.com).

The data mining technique is very important for predicting student performance. In learning analytics, the aim is to collect, analyze, and extract knowledge that is related to data in order to improve learning results and the environment. Education is crucial for the progress and success of any nation. Students face two essential problems while they are learning different process of failing classes and dropping out the courses of programming. Applied data mining techniques include regression, time series analysis, classification and all associated rules of mining are employed in education data mining to analyze and evaluate the aspects over various datasets collected. A common technique in Educational Data Mining (EDM) is to develop predictive models and to identify pattern that are hidden, information that can be useful in educational settings. Student's academic success must be predicted in order to identify those students who have risk of failing early on in the semester assessment.

II. REVIEW PROCESS

This section describe the review process used to select the research articles for this review article. Four research database such as IEEE, ACM, Springer and Elsevier were considered to download the research articles. Conference paper/ proceedings from these databases are not considered for this review article. So 74 research articles which are journal articles were searched using the keyword "Educational data mining" from research databse. So such number of research articles 19, 12, 27 and 16 have downloaded from IEEE, ACM, Springer and Elsevier research databases respectively. Research articles not in English language were removed from the downloaded research articles. So such three research articles were not considered for review process. Research articles were also reviewed by going through the abstract and removed 18 articles which are not relevant to the study. So after reading the research articles, twenty research articles which are not relevant for the study, are removed. Finally, 32 research articles are considered for review. The method used to select the research articles for this study is given in Figure 1.

After selecting 32 research articles, the excel sheet is prepared which contains the analysis of these articles and contains various fields such as serial number, year of publication, Indexing in Elsevier / Elsevier Proc / Springer / Springer Proc / IEEE Transaction / IEEE Digital Explore, journal name/ conference name, year of publication, Title, Abstract, Performance Evaluation Measures, Parameter values, Dataset used, Data collection technique, Software used, Data Mining tools used, Classification algorithm employed in article, conclusion, limitations, future work, URL of article, MLA citation of article, APA citation of articles, Citation count and Cited by research article URL. Excel sheet is used to arrange the data so that filter can be

used to analyze any field considered and draw the graph for the same.

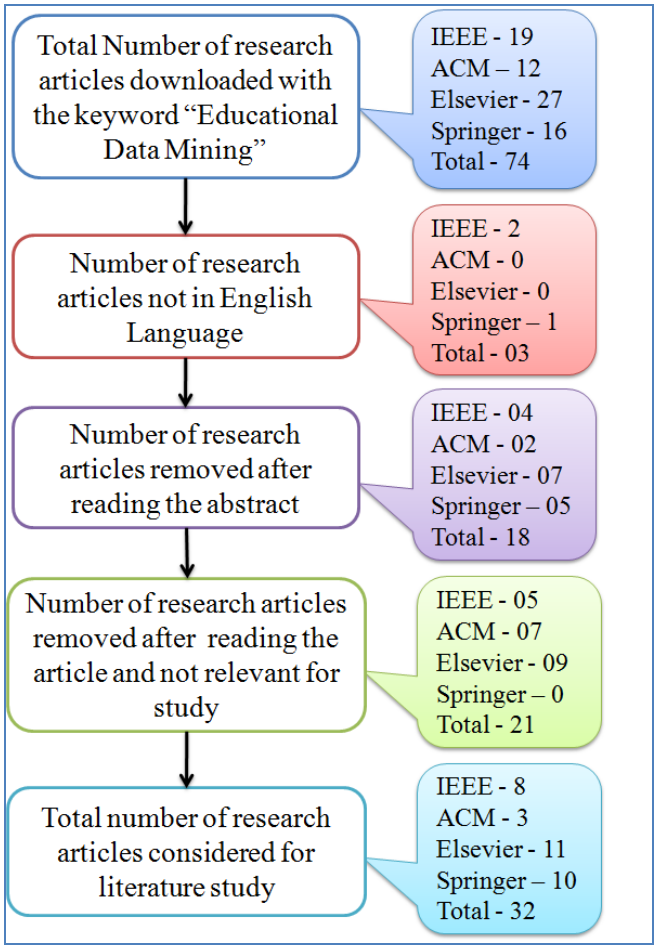

Fig. 1. Review Process for selecting the research articles for review paper.

III. Educational Data Mining

EDM is used to analyze and predict the performance of students from dataset received from education sector/ Learning management system (LMS) / Content management system (CMS). Figure 2 explains the EDM process in which data is collected from institutions/ universities/ LMS / CMS. From obtained data, dataset for appropriate sampling period is prepared. Dataset may contain the outlier/ missing values. The preprocessing involves missing value treatment, Outlier, finding out the number of features and their relationship with each other so that during the training of the model the accuracy of result should be improved. After preprocessing step, required features are extracted from datset. Data is prepared in the required format which act as input the model being developed. Model is build using Data Mining techniques such as classification, clustering, association ruls, etc. After building the model, results are analyzed based on performace parameters. Last step is the prediction of students performance by interpreting the result of developed model.

Fig. 2. EDM Process.

A. Classifcation

In Classification technique which is supervised learning, data is categorized into the classes and new category of observation is identified on the basis of training data. This technique is used various application such as document classification, classification of spam and non-spam mails, etc. Various classification techniques are Naïve Bays, Support Vector Machine, K-Nearest Neighbour, Decision Tree, Random Forest, etc.

This subsection discusses about the use of classification technique in EDM.

JARKKO LAGUS et. al. [2]: The purpose of this project was to investigate methods for predicting courses of students so that their outcome can be used as the introductory courses of programming. They have used F1 score, recall and precision technique for performance evaluation with 348 student datasets. The models are made up by observed data which is relevant to machine learning model. Algorithm used are Support Vector Machine, Random Forest, and AdaBoost (Best – Random. It is difficult to generalize these methods in educational contexts, however, due to differences between courses. In courses of introductory, methods are evaluated with the help of data collection. Transfer-learning improves predictions, especially in cases with little training data.

Concepción Burgos et. al. [5]: Data from systems of E-learning can generate tremendous large data; analysis can be the challenging work that requires computational techniques and tools. Evaluation done on was precision and recall and dataset of 104 students. They have used logistic regression. Knowledge discovery techniques could be used to analyze historical course to predict the grade data that will drop the course. Using their tool and tutoring plan, their educational institution which can decrease the dropout rate of E-learning courses. Their obtained results are only slight better than the existing models.

V.L. Miguéis et. al. [6]: Early identification of students for university was based on their strength academic performance that is useful for mitigating failure, or better result they are been encourage with achievements, and better managing resources. This paper has proposed 2 stage model, which uses the techniques of data mining to predict students' academic performance of their 1st year of an academic career. This Study proposed that based on student's segment examples over failure start at high performance of their degree programs, as well as the students' performance levels predicted by the model. The segmentation framework suggests strategies for promoting higher level performance and preventing a failure of academic based upon the framework.

Aderibigbe Israel Adekitan et. al. [12]: When predicting failure is available, it is an advantageous instructional tool that can be used to counsel students, and it may also be used in developing teachers' skills. The purpose of this study is to

determine whether there is a relationship between graduation grades and to examine the influence of ethnicity in the development of the models using Nigerian students' geopolitical zones as predictors of scores for graduation and admission criteria scores. According to their analysis, the accuracy of the pre-admission scores alone is 53.2% indicating that the score of pre-admission is not sufficient to predict a student's graduation result, but they can serve as a useful guide. Logistic Regression is considered with a dataset of 1841 instances are used which is very small. Hence, ethnicity should not be considered in admission processes, as it does not have a statistically significant effect on graduation results.

Anwar Ali Yahya et. al. [13]: Using supervised machine learning techniques, the present work presents a method for predicting students' marks and grades. It is based on the historical performance of students. In this study, the purpose is to analyze the quality of education in relation to sustainable development goals. As a result of implementing the system of more data has been collected over time that is processed appropriately for data that can be more useful for continued planning and development Secondary Education Islamabad and Federal Board of Intermediate provides the dataset that we are using in our methodology. The dataset contains 80,000 historical student data. They didn't have any sort of forecast or suggestion mechanism in place.

Aderibigbe Israel Adekitan et. al. [14]: Research has studied increasing educational data mining because of the advantages obtained through knowledge acquired through machine learning processes which enable higher education institutions to make better decisions. The main role of this study is to find out how the cumulative grade point average (CGPA) can be determined over fifth-year and Nigerian university final engineering students by using predictive analysis. Data of 2,413 students from 2002 to 2014 is taken into consideration which is very small.

Bashir Khan Yousafzai et. al. [15]: The given work is a trained machine learning-based system for predicting the grades of the students or marks. The system relies on the historic performance of students. The purpose of this study is to examine education quality in relation to long-term development goals. The system's adoption has resulted in an abundance of data that must be handled properly in order to obtain more relevant data on that is used for future growth and planning. Finally, the findings of both models are compared and contrasted. Decision tree, K-nearest neighbor data mining algorithm are used. The obtained findings demonstrate the value and usefulness of machine learning technology in predicting student performance. The Secondary Education Islamabad and Federal Board of Intermediate provided the dataset for our suggested technique. There are 80,000 historical students' data in this collection. Its limitation is that the regression is based on the predicting system so its RMSE of 5.34 is poor.

HANAN ABDULLAH MENGASH [17]: To choose applicants who are more to perform better in institutions academics for higher education, the system admissions are based on accurate and trustworthy admissions criteria is critical. This research focuses on how data mining techniques may be used to forecast applicants' academic achievement at university to assist institutions in making admissions decisions. Additionally, scores on the Scholastic Achievement Admission Test predict future student performances the best of all pre-admission criteria. Dataset of 2,039 students was taken into consideration but this a set is too small.

Mohammad Noor Injadat et. al. [18]: The goal of research is to support higher institutions education make better admissions decisions by forecasting applicants' academic success before they are admitted. Researchers found that prediction models are useful in university settings because making decision can employ this model to plan and optimize institutions' limited resources.

Karthikeyan et. al. [20]: Many academic institutions use data analysis to keep track of their student's records, especially their educational achievements, which are much more significant. The model is tested using the benchmark education dataset from the WEKA environment, which is available online. They have used the Hybrid Educational Data Mining model of classification models which includes Naïve Bayes and J48. The findings suggest that the proposed model outperforms previous studies in assessing student performance in EDM. They have used a dataset of 14 student attributes. There is no information about the dataset specifications. The findings were generated using an existing tool.

Mudasir Ashraf et. al. [21]: They have used the ensemble approach for boosting which is based on prediction development. They have taken dataset of 115 students. The Data mining algorithm used is j48, random tree, naïve Bayes, K-nearest neighbor, and boosting (Best - Naïve Bays). They find that the filtering approach has a substantial impact on the accuracy and prediction of classifiers. Its limitation is that they have evaluated with small size dataset.

Mohammad Noor Injadat et. al. [22]: There has been a lot of studies done in the past on forecasting student achievement in order to help them grow. They have use Gini index and p-value for the performance evaluation. The use of comparative analysis to anticipate students' performance at early phases of the instructional delivery using multiple categorization algorithms is useful. Dataset of 538 Students have been used which is in small size.

Pranav Dabhade et. al. [23]: The major goal of educational institutions is to offer students a high-quality education in order to improve their academic performance. They have used root mean squared error (RMSE) and R-Square for performance evaluation. A dataset of 85 people was gathered. They have found that the linear model can be the best fit with an accuracy of 83.44%. Other machine learning techniques, such as regression algorithms and neural networks, can be used to expand the current study. Usage of small dataset is the drawback of this paper.

Malini et. al. [24]: This research uses the EDM to describe the many aspects influencing students' performance by employing efficient algorithms to make predictions. They have used a data mining algorithm which includes Boosting algorithm, Bagging, and Artificial Neural Network (ANN). With the economic background attributes, Multi-layer Perceptron (MLP) classifiers exhibit 72 percent accuracy, Bagging classifiers show 88 percent accuracy and

MultiBoost classifiers show 86 percent accuracy. Bagging classifiers had a greater accuracy, indicating that a student's economic background influences their learning behavior in the educational system. They have not provided the dataset information.

AYA NABIL el. al. [25]: The major purpose of research is to look into the efficacy of EDM in deep learning, particularly in terms of forecasting students' academic success and identifying students who are in danger of failing. For performance, they have evaluated the accuracy, precision, recall, F1-score, classification score. They used the dataset of 4266 students. Algorithm of data mining used in these researches are, random forest, gradient boosting, decision tree, support vector classier, K-nearest neighbor and logistic regression. Other technique deep neural network (DNN) is used. Kalboard 360, a learning management system, was used to collect the data. There are 500 entries and 16 characteristics in this collection. While using the SMOTE technique as an oversampling approach, DNN surpassed the others with an accuracy of 89 percent, an F1-score of 89 percent, and a sensitivity of 89 percent but these results can be increased.

ZHONGYING ZHAO et. al. [28]: The Massive Open Online Courses (MOOC) course captions were used to create a system that included course-level requirement relationships and concept-level. They use two datasets that is 5167 amp and 23288 instances respectively and for evaluation, they have used recall, precision, and F1 measure. A random forest algorithm is used but they have not implemented the cognitive inference.

Yifan Zhu et. al. [30]: The prediction rating of proposed system is applied by Bayesian Probabilistic Tensor Factorization (BPTF) through the teaching evaluation. In this they have used the Bayesian Probabilistic Tensor Factorization algorithm and the performance evaluation was based on root mean squared error(RMSE), precision, mean absolute error(MAE), and F1 score. They have used 532 instances dataset which small in size and also the functions which they have used are not unique.

NACIM YANES et. al. [31]: According to the findings, the adaptable Machine Learning k-nearest neighbor (ML-KNN) had the least hamming cost, whereas the Support Vector Classifier (SVC) had the greatest F1 measure. They have used BR, k-nearest neighbor (KNN), random forest (RF), Gaussian Naive Bays (NB), corrected classifier (CC) and decision tree (DT) algorithm for implementation. The size of the dataset has not been mentioned and for evaluation purpose, they have used Hamming Loss, F1 Measure, Recall, and Precision.

Fernández-García et. al.[32]:In this research, they have created the model which can assist the student in the selection of best subject with more accuracy. They have used random forest algorithm, Gradient Boosting Classier, Support vector machine, Logistic regression, decision tree and multilayer perceptron with 323 instance datasets. Its limitation is that they have used very small dataset so accuracy is less and the dataset are not balance.

B. Clustering

In Clustering technique which is an unsupervised learning, dataset is divided into number of clusters such that data belonging to a cluster have same characteristic. Types of clustering are Hierarchical clustering, Partitioning methods, Density based clustering, constraint based clustering, Fuzzy clustering, and Distribution based clustering. This subsection describe the use of clustering technique to predict students' performance.

Sanyam Bharara et. al. [3]: The field of Learning Analytics (LA) uses sophisticated analytical tools to improve education and learning. They have used the concepts of Educational Data Mining and Learning Analytics, the aim of this research is to find the meaningful metrics and indicator to analyze relationship and context, evaluating the effects of different teaching methods. An important part of their project involves data mining technique that is K-means clustering for getting the clusters, which are then mapped to main features that finds a learning context and it uses the dataset of Kaggle.

AFTAB AKRAM el. al. [8]: In computer-supported learning environments, procrastination is reported that affect the performance of student. Research shows that students with higher tendencies achieve very less than those with least procrastination tendencies. The purpose of their paper is to present algorithm to predict student's performance of academic through the detection of late non-submitted homework. They have used 115 instances dataset and evaluated the performance over Kappa Statistics and Root Mean Square Error (RMSE). Different method of classification is compared with different numbers of classes in a detailed study and the best and worst methods; however, it depends on number of classes. The limitation of these paper are authors do not consider features vectors of varying duration, such as numerous classes with varying amounts of homework.

TapaniToivonen et. al. [9]: As a result of evolution, Educational Data Mining (EDM) processes now incorporate visualizations, parameter and predictive model adjustment, and open-ended data mining. EDM end-users are able to understand the dataset and the context of data collection much better with the use of adjusting and even creating decision tree classifiers, as well as creating them themselves. With accuracy, the model performance was evaluated and 64 instances of the dataset were considered. Unlike other methods similar to their model, the most important part of the model is adjusting the model, not tuning the parameters. According to the study, Augmented Intelligence method (AUI) is also capable of enabling knowledge discovery and its limitation is they have considered a very small dataset.

Bindhia K. Francis et. al. [10]: As the number of student's information grows, the need for education data mining is also growing rapidly. This can be used to infer valuable patterns to better understand the learning process of students. They have used Naïve Bayes, Support Vector Machine (SVM), Neural Network classifiers (Naïve Bayes) algorithms and Decision tree. They have study to produce a prediction algorithm for evaluation students' performances in academia that uses classification and clustering

techniques at the same time. For achieving the predicting the academic student's performance, the hybrid algorithm, which incorporates clustering and classification approaches, is clearly the more accurate algorithm. The drawback of this is that the dataset description is missing.

MIGUEL ÁNGEL PRADA et. al. [16]: This work describes a web-based software application for engineering students' tutoring help that does not require a data scientist expertise to use. Preliminary classification and fall results were acceptable, with accuracy and reliability reaching 90% in several circumstances. Dataset of 21,000 graduated and non-graduated students with 22 attributes between 2011 and 2017. This research describes a web-based software solution for student that assists tutoring professionals that do not have a background in data science. The proposed technology is intended for analyzing and forecasting student success in terms of measurable scores and degree completion.

C. Association Rules

Association rule mining has two parts – antecedent and consequent. It is rule-based algorithm used to find the relationship between the variables in the database, e.g. if a person buy the mobile then he is likely to buy the mobile cover and screenguard for the same. Different association rule algorithms are Apriori association rule, Filtered growth, relational association rule mining, etc. In this subsection, the use of this association rule algorithms to analyze students' data is discussed.

Anupam Khan et. al. [4]: Student performance is an active area of research in educational research. They have evaluated their proposed model based on confidence and support with dataset of 9072 instances. Their study also examines the impact of teaching on performance improvement in a classroom-based course. The result of the study indicates that teaching has a positive influence on student performance. More specifically, it suggests more students will reach expected or higher levels of achievement with superior teaching. The influence of instruction on failure instances may necessitate a distinct strategy.

Tao Xie et. al. [7]: The paper, represent the method that mines big learning data in a way that considers both the frequency and duration. Evaluation based on Precision, recall, and accuracy with 57717 instances of datasets. Their method evaluated the events by identifying their importance, and segmenting the using a sliding window for big uniform events (BUEs) to avoid counting bias. They have used Association rule algorithm. They also suggested that, it is very difficult to determine the weight of duration of numerical and frequency.

D. Other Techniques

This subsection illustrates the use of other techniques such as fuzzy logic, collaborative filtering, process mining, etc. in EDM.

QI LIU et. al. [1]: The goal of this study is to explore both objective and subjective scores of cognitive problems using a fuzzy cognitive diagnosis framework (FuzzyCDF). Additionally, they were able to predict examinee with respect their performance and guesses. Also, they have analyzed cognitive diagnosis which was based on the FuzzyCDF model. They have used Precision, Recall and F1 score for evaluation. Moreover, extensive experiments showed that FuzzyCDF could analyze quantitatively and interpretatively each examinee's characteristics, determining better performance and can increase the studies in the future. They have considered 8656 dataset instances. This system has a high level of computational complexity.

Yu Wang et. al. [11]: During education, we are often responsible for evaluating student learning effects based on students' final deliverables. Here, the authors presented a framework for assessing in-process student learning effect evaluation that includes an interactive component. In a case study, they examined the student online modeling behaviors collected from 24 computer science majors, which they used to demonstrate the process we developed. The limitation of this paper is that they have considered very few datasets.

Rebeca Cerezo et. al. [19]: The goal of this study was to use Educational Process Mining(EPM) approaches to measure Self-Regulated Learning(SRL) student's abilities during courses of E-learning. To accomplish so, we examined a file of log containing 21,629 occurrences from an online Spanish undergraduate course. Preprocessing was used to execute process mining. This study aimed to shine new light just on e-teaching-e-learning processes using EPM approaches and to be valuable to the essential actors in the teaching-learning process, instructors and learners, despite its limitations. The finding of this paper is that they have created a model based upon data which were supplied by graduate students of 3rd year, therefore the models may alter if first-year graduate students were included. Furthermore, the dataset is rather tiny.

LING HUANG et. al. [26]: Experimental results have already been carried out to assess the efficacy of the proposed strategy, with the findings indicating that it is capable of achieving a high annual strike rate with average was taken. They have used a cross-user collaborative filter domain algorithm. The dataset of 1166 instances were considered in this work and has been evaluated with the performance of accuracy but they have considered the dataset very small.

MOHAMMED E. IBRAHIM et. al. [27]: Collaboration-based filtering is combined with material filtration under this project. They have also considered familiar linked ideas that appear from both the graduate's as well as the school's profile when calculating their similarity. The data mining algorithm used in this paper is ontology-based hybrid-filtering with 95 instances. They have evaluated the performance based on the relevance, rank accuracy and recovery. But its limitation is that they have used very small dataset.

Esteban et. al. [29]: The suggested model uses several methods, including such Collaborative Filtering (CF) depending on neighborhood or Content-based Filtering (CBF), as well as text analytics, to merge input first from learner and also the program. A modified Genetic Algorithm (GA) has been applied, resulting in intelligible models wherein users can manage the importance of each criterion in the suggestions and acquire the optimum solution of all Recommendation Systems (RS) variables, including such similarity metrics or size of the network. Genetic algorithm

and collaborative and content-based algorithms have been used with 95 instances of the dataset. Limitation of this work is that they have used very small dataset.

IV. Recent Advances

Education is very crucial factor in development of a nation. If an effective system is developed for prediction of students' performance, then it can be very helpful for the students to improvise their score as well as performance in the academics. As suggested by some researchers that to carry out such educational data mining system the dataset plays a crucial role. It can be clearly seen from research work like [2],[3],[5],[8],[9],[11],[18],[19],[21],[22],[23],[27],[29],[30], [32] that there is very lack of availability of real-time datasets makes the research work in this sector more difficult. Also, the datasets are imbalanced in nature. It becomes very important to handle such kind of issues providing more precise and accurate results by using various advance sampling techniques. As the datasets are smaller in size, it is very difficult to get more accurate results. Hence a system consisting of good datasets can be designed and developed to provide more promising and accurate results. Also, few very performance metrics parameters are explored by the researchers to compare their obtained results. Due to this reason more parameters can be utilized to compare the obtained results [26]. Some research work like [7],[13], [16],[25], and [28] consist of considerably good volume of dataset instances which be further used for comparatively analysis purpose.

V. Analysis and Discussion

This section dicusses the analysis based on number of research articles considered from research database, number of research articles considered yearwise, performance parameters, number of performance parameteres used by research articles, Data Mining Techniques, number of algorithms used by research articles, dataset size, Data Mining techniques and other techniques

A. Analysis based on number of research articles considered from research database

This subsection represents analysis based on number of research journal articles considered from research databases such as IEEE, ACM, Springer, and Elsevier. From Table 1, it is noted that elevan journal articles are considered from IEEE transaction/ IEEE access while ten and nine journal articles are considered from Springer and Elsevier respectively.

Table I
Analysis on basis of number of research articles considered from research database

Research Database	Number of research articles	Reference number
IEEE Transaction/ ACCESS	8	[8], [16], [17], [25], [26], [27], [31], [32]
ACM Transaction	3	[1], [2], [18]
Springer Journal	10	[3], [4], [9], [10], [11], [14], [15], [18], [19], [20]
Elsevier Journal	11	[5], [6], [7], [12], [13], [21], [22], [23], [24], [29], [30]

B. Analysis based on Publication year

This subsection describe the analysis based on number of research articles considered yearwise. Four years research articles are considered for this review paper. From Figure 1, it is found that 13 research articles were from year 2020 while 8 research articles were considered from year 2018 and 2019.

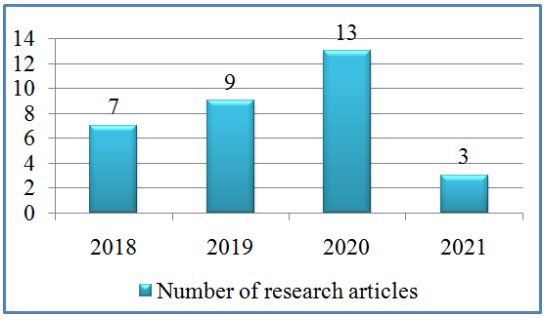

Fig. 3 - Analysis based on publication year

C. Analysis based on performance parameters

This subsection analyzed the research articles based on performance parameters of classification, clustering, association rule and any other. The performance parameters of classification considered are Accuracy, Precision,Recall, F-measures, Specificity, Kappa Statistics, Area under ROC Curve (AUC), Correctly classified instances, Incorrectly classified instances, False Positive Rate, True Positive rate, Receiving Operating Characteristics, Relative Absolute Error, RMSE, MAE, R-Square, Confusion matrix, Sensitivity, while for clustering K-value parameter is used. For association rules, Support and Confidence parameters are used while t-Test, Gini index, p-value, Pearson correlation coefficient, and Cosine similarity are other parameters. From Table, it is elucidated that performance parameters accuracy and F-measure are used by research articles [5, 9, 10, 12, 13, 14, 16, 17, 20, 22, 24, 25, 32] and [5, 6, 10, 12, 17, 18, 20, 22, 24, 25, 28, 30, 31, 32] respectively while performance parameters precision and recall are considered in research articles [5, 6, 10, 12, 17, 18, 22, 24, 25, 28, 30, 31] and [5, 6, 10, 12, 17, 18, 24, 25, 28, 30, 31].

Table II
Analysis on basis classification performance parameters

Techniques	Performance Parameters	Number of Research articles	Reference Number
Classification	Accuracy	13	[5], [9], [10], [12], [13], [14], [16], [17], [20], [22], [24], [25], [32]
	Precision	12	[5], [6], [10], [12], [17], [18], [22], [24], [25], [28], [30], [31]
	Recall	11	[5], [6], [10], [12], [17], [18], [24], [25], [28], [30], [31]
	F-measures	13	[5], [6], [10], [12], [17], [18], [20], [22], [24], [25], [28], [30], [31], [32]
	Specificity	3	[5], [20], [22]
	Kappa Statistics	1	[8]
	Area under ROC Curve	1	[12]
	Correctly	2	[14], [21]

	classified instances		
	Incorrectly classified instances	2	[14], [21]
	False Positive Rate	3	[18], [21], [24]
	True Positive rate	2	[21], [24]
	Receiving Operating Characteristics	1	[21]
	Relative Absolute Error	1	[21]
	RMSE	5	[8], [13], [23], [29], [30]
	MAE	1	[30]
	R-Square	1	][23]
	Confusion matrix	1	[24]
	Sensitivity	2	[20], [22]
Clustering	K-value	1	[3]
Association Rule	Support	2	[4], [7]
	Confidence	1	[4]
Other	t-Test	1	[15]
	Gini index	1	[22]
	p-value	1	[22]
	Pearson correlation coefficient	1	[26]
	Cosine similarity	1	[22]

D. Analysis based on number of performance parameteres used by research articles

This section illustrates the analysis on basis of number of performance parameteres used by research articles. Research articles [22] and [24] used seven performance parameters while research article [5] and [12] considered five performance parameters. Research articles [3, 7, 9, 15, 16, 26, 27], [4, 8, 13, 23, 29, 32] , [1, 2, 6, 14, 28, 31], and [10, 17, 18, 20, 25, 30] had made the use of one, two, three and four performance parameters.

TABLE III
ANALYSIS ON BASIS OF NUMBER OF PERFORMANCE PARAMETERS USED BY RESEARCH DATABASE

Number of parameters	Number of research articles	Reference Number
1	7	[3], [7], [9], [15], [16], [26], [27]
2	6	[4], [8], [13], [23], [29], [32]
3	6	[1], [2], [6], [14], [28], [31]
4	6	[10], [17], [18], [20], [25], [30]
5	2	[5], [12]
6	1	[26]
7	2	[22], [24]

E. Analysis based on Data Mining Techniques and Algorithms

In this subsection, Data Mining techniques and algorithms referred by research articles are analyzed. Classification techniques referred are Random Forest, Random Tree, AdaBoost, Logistic Regression, ZeroR, OneR, ID3, J48, Decision Stump, Jrip, PART, NBTree, Prism, Neural Network, Decision Tree, Naïve Bayes,Support Vector Machine, Probabilistic Neural Network, Particle Swarm Classification, K-nearest neighbor, Artificial Neural Network, Multi-Layer Perceptron, Multiple linear regression, Tree Ensemble, Boosting algorithm, Bagging, Bayesian Probabilistic Tensor Factorization, and Fuzzy cognitive diagnosis framework while K-means algorithgm is used for clustering. Table 4 represents aalysis on basis of Data Mining techniques and algorithms. From Table 4, it is noted that Support Vector Machine, Random Forests and Naïve Bays are mostly used classification techniques while Random Tree, AdaBoost, ZeroR, OneR, ID3, Decision Stump, Jrip, PART, NBTree, Prism, Probabilistic Neural Network, Particle Swarm Classification, Multiple linear regression, Tree Ensemble, Bagging, Bayesian Probabilistic Tensor Factorization, and Fuzzy cognitive diagnosis framework are least used classification techniques.

TABLE IV
ANALYSIS ON BASIS OF DATA MINING TECHNIQUES AND ALGORITHMS

Technique	Algorithm	No. of Research Article	Reference Number
Classification	Random Forest	9	[2], [6], [8], [12], [14], [18], [22], [25], [31]
	Random Tree	1	[21]
	AdaBoost	1	[2]
	Logistic Regression	6	[12], [18], [22], [25], [31], [32]
	ZeroR	1	[8]
	OneR	1	[8]
	ID3	1	[8]
	J48	3	[8], [20], [21]
	Decision Stump	1	[8]
	Jrip	1	[8]
	PART	1	[8]
	NBTree	1	[8]
	Prism	1	[8]
	Neural Network	4	[9], [10], [14], [18]
	Decision Tree	7	[10], [12], [14], [15], [17], [25], [31]
	Naïve Bayes	9	[10], [12], [14], [17], [18], [20], [21], [22], [31]
	Support Vector Machine	10	[2], [10], [16], [17], [18], [22], [23], [25], [31], [32]
	Probabilistic Neural Network	1	[12]
	Particle Swarm Classification	1	[13]
	K-nearest neighbor	7	[15], [18], [21], [22], [25], [31], [32]
	Artificial Neural Network	2	[17], [24]
	Multi-Layer Perceptron	2	[22], [32]
	Multiple linear regression	1	[23]
	Tree Ensemble	1	[12]
	Boosting algorithm	3	[21], [25], [32]
	Bagging	1	[24]
	Bayesian Probabilistic Tensor Factorization	1	[30]
	Fuzzy cognitive diagnosis framework	1	[1]
Clustering	K-means	5	[3], [8], [9], [10], [16]
Association Rule	Apriori Association Rules / Association Rule	2	[4], [7]
Other	Process Mining	2	[11], [19]
	Collaborative filtering	2	[26], [29]
	Content-based Filtering	1	[29]
	Ontology-based hybrid-filtering system	1	[27]

F. Analysis based on number of algorithms used by research articles

Genrally research articles uses various classification techniques and based on various performance parameters, best classification technique is used for given dataset for the application. In this approach, various classification techniques are compared to find the best algorithm. Table 5 discusses the analysis of research articles based on number of algorithms used. Research article [8] used 11 classification algorithms ZeroR, OneR, ID3, J48, Random Forest, decision stump, JRip, PART, NBTree and Prism and found Random Forest to be the best algorithm. Five research articles [12, 18, 22, 25, 31] compared six classification algorithm to find the best. Research article [14] considered the classification algorithms Tree, Naive Bayes, Random forest, and Neural network.

TABLE V
ANALYSIS ON BASIS OF NUMBER OF ALGORITHMS USED

Number of algorithm used	Number of Research articles	Reference Number
1	14	[1], [3], [4], [5], [6], [7], [11], [13], [16], [19], [26], [27], [28], 30]
2	5	[9], [15], [20], [23], [29]
3	2	[2], [24]
4	1	[14]
5	4	[10], [17], [21], 32]
6	5	[12], [18], [22], [25], [31]
11	1	[8]

G. Analysis based dataset size

Dataset size plays an important role in deciding the performance of model being developed. Table 6 discusses analysis based on dataset size. The dataset size range considered are 1-100, 101-200, 301-400, 401-500, 501-600, 601-700, 1001-2000, 2001-2500, 4001-4500, 8501-9500, 21000-30000,55001-60000, and 80000. There are five research articles in which dataset size is not mentioned. Only research article [13] considered dataset size 80,000 for analysis of students' performance. Four research articles [9, 11, 23, 29] used the dataset in the range 1-100.

TABLE VI
ANALYSIS ON BASIS OF DATASET SIZE

Dataset range	Number of research articles	Reference number
1-100	4	[9], [11], [23], [29]
101-200	5	[5], [8], [19], [21], [27]
301-400	2	[2], [32]
401-500	1	[3]
501-600	2	[22], [30]
601-700	1	[18]
1001-2000	2	[14], [26]
2001-2500	3	[6], [12], [17]
4001-4500	1	[25]
8501-9500	2	[1], [4]
21000-30000	2	[16], [28]
55001-60000	1	[7]
80000	1	[13]
Not mentioned	5	[10], [15], [20], [24], [31]

H. Analysis based on Data Mining and other techniques

This subsection illustrates analysis of research articles based on Data Mining techniques and other techniques such as Process Mining, Collaborative/ Content-based Filtering, Fuzzy cognitive diagnosis framework, and Ontology-based hybrid-filtering system.

Classification technique is mostly used technique by research articles [2, 5, 6, 12, 13, 14, 15, 17, 18, 20, 21, 22, 23, 24, 25, 28, 30, 31, 32] for predicting students' performance from Table 7.

TABLE VII
ANALYSIS ON BASIS OF DATA MINING AND OTHER TECHNIQUES

Technique	Number of Research Articles	Reference Number
Classification	19	[2], [5], [6], [12], [13], [14], [15], [17], [18], [20], [21], [22], [23], [24], [25], [28], [30], [31], [32]
Classification and Clustering	4	[8], [9], [10], [16]
Clustering	1	[3]
Association Rule Mining	2	[4], [7]
Process Mining	2	[11], [19]
Collaborative/ Content-based Filtering	2	[26], [29]
Fuzzy cognitive diagnosis framework	1	[1]
Ontology-based hybrid-filtering system	1	[27]

VI. RESEARCH GAP AND FUTURE DIRECTION

After reviewing the research articles, findings of these articles are-

- 19 research articles [1], [3], [4], [5], [6], [7], [9], [11], [13], [15], [16], [19], [20], [23], [26], [27], [28], [29], and [30] applied one or two algorithms to build the model in EDM
- 5 research articles [5], [6], [9], [13], and [16] considered only one classification algorithms to build the model
- 9 research articles [2], [10], [14], [15], [17], [20], [23], [24], and [30] used 2, 3, and 4 number of classification algorithms for comparison to select the best one.
- 22 research articles [1], [2], [3], [4], [5], [6], [7], [9], [10], [13], [14], [15], [16], [17], [18], [20], [25], [26], [27], [28], [30], [31], and [32] examined less number of performance parameters.
- Use of Small dataset size in 20 research articles [2], [3], [5], [6], [8], [9], [11], [12], [14], [17], [18], [19], [21], [22], [23], [26], [27], [29], [30], and [32]
- Also, very few performance parameters in 23 research articles [1], [2], [3], [4], [5], [6], [7], [9], [10], [13], [14], [15], [16], [17], [18], [20], [26], [25], [27], [28], [30], [31], and [32], are explored by the researchers to compare their obtained results. Due to this reason, more parameters can be utilized to compare the obtained results [26].
- Some research work [1], [4], [7], [10], [13], [15], [16], [20], [25], [28], and [31] consist of considerably good volume of dataset instances which can be further used for comparatively analysis purpose using more number of performance parameters.

Several research papers from educational data mining background were reviewed and analyzed thoroughly. After exploring these papers below mentioned research gap was observed.

- Lack of availability of real-time dataset in the educational data mining domain.
- Very few machine learning and deep learning algorithms were explored for prediction of student's performance.
- Majority of related work has considered few parameters for prediction the performance of the students.
- Lack of efficient recommendation system to carry out further research in this domain.
- As per the study none of the work is carried out with explainable AI approach.

From several state-of-art based on educational mining, it has been observed that

- The work is performed only on limited size of dataset which itself is not justifiable.
- Also, very few data mining techniques are implemented in the majority of the related work.

Inspite of the research gap, future direction for working in EDM are mentioned below-

- Mostly used Data Mining Technique is Classification as referred in research articles [2], [5], [6], [8], [9], [10], [12], [13], [14], [15], [16], [17], [18], [20], [21], [22], [23], [24], [25], [30], [31], and [32]
- Mostly used classification algorithms are
  - Random Forests - [2], [6], [8], [12], [14], [18], [21], [22], [25], and [31]
  - Decision Tree - [10], [12], [14], [15], [17], [25], and [31]
  - Naïve Bays - [10], [12], [14], [17], [18], [20], [21], [22], and [31]
  - Support Vector Machine - [2], [10], [16], [17], [18], [22], [23], [25], [31], and [32]
  - K-nearest neighbour - [15], [18], [21], [22], [25], [31], and [32]
  - Logistic Regression - [12], [18], [22], [25], [31], and [32]
- Mostly used classification performance parameters
  - Accuracy - [5], [9], [10], [12], [13], [14], [16], [17], [20], [22], [24], [25], and [32]
  - Precision - [1], [2], [5], [6], [10], [12], [17], [18], [22], [24], [25], [28], [30], and [31]
  - Recall - [1], [2], [5], [6], [10], [12], [17], [18], [24], [25], [28], [30], and [31]
  - F-measures - [1], [2], [5], [6], [10], [12], [17], [18], [20], [22], [24], [25], [28], [30], [31], and [32]
- Mostly used clustering algorithm is K-means clustering algorithm [3, 8, 9, 10, 16].
- Generally used Data Mining Tool – Weka [10, 13, 17, 20, 24]
- Generally used software to build the model in EDM – R-Programming [4, 18, 22], and Python [16, 23, 25].

VII. Conclusion

Several state-of-art based on educational mining has been reviewed and analyzed in this work. It is observed that the classification technique is mostly used to analyze the performance of students . It is also noted that performance parameters accuracy, precision, recall, and F-measure are used by most of the research articles. Mostly used classification techniques are Support Vector Machine, Random Forests and Naïve Bays. It has been observed that the work is performed only on limited size of dataset which itself is not justifiable. So this review article will helpful to the researcher working in EDM for further research in the same domain for analyzing and predicting the performance of students.

References

[1] Liu, Q., Wu, R., Chen, E., Xu, G., Su, Y., Chen, Z., & Hu, G. (2018). Fuzzy cognitive diagnosis for modelling examinee performance. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *9*(4), 1-26.

[2] Lagus, J., Longi, K., Klami, A., & Hellas, A. (2018). Transfer-learning methods in programming course outcome prediction. *ACM Transactions on Computing Education (TOCE)*, *18*(4), 1-18.

[3] 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.

[4] Khan, A., & Ghosh, S. K. (2018). Data mining based analysis to explore the effect of teaching on student performance. *Education and Information Technologies*, *23*(4), 1677-1697.

[5] Burgos, C., Campanario, M. L., de la Peña, D., Lara, J. A., Lizcano, D., & Martínez, M. A. (2018). Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. *Computers & Electrical Engineering*, *66*, 541-556.

[6] Miguéis, V. L., Freitas, A., Garcia, P. J., & Silva, A. (2018). Early segmentation of students according to their academic performance: A predictive modelling approach. *Decision Support Systems*, *115*, 36-51.

[7] Xie, T., Zheng, Q., & Zhang, W. (2018). Mining temporal characteristics of behaviors from interval events in e-learning. *Information Sciences*, *447*, 169-185.

[8] Akram, A., Fu, C., Li, Y., Javed, M. Y., Lin, R., Jiang, Y., & Tang, Y. (2019). Predicting students' academic procrastination in blended learning course using homework submission data. *IEEE Access*, *7*, 102487-102498.

[9] Toivonen, T., Jormanainen, I., & Tukiainen, M. (2019). Augmented intelligence in educational data mining. *Smart Learning Environments*, *6*(1), 1-25..

[10] 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.

[11] Wang, Y., Li, T., Geng, C., & Wang, Y. (2019). Recognizing patterns of student's modeling

behaviour patterns via process mining. *Smart Learning Environments*, *6*(1), 1-16.

[12] Adekitan, A. I., & Salau, O. (2020). Toward an improved learning process: the relevance of ethnicity to data mining prediction of students' performance. *SN Applied Sciences*, *2*(1), 1-15.

[13] Yousafzai, B. K., Hayat, M., & Afzal, S. (2020). Application of machine learning and data mining in predicting the performance of intermediate and secondary education level student. *Education and Information Technologies*, *25*(6), 4677-4697.

[14] Adekitan, A. I., & Salau, O. (2019). The impact of engineering students' performance in the first three years on their graduation result using educational data mining. *Heliyon*, *5*(2), e01250.

[15] Yahya, A. A. (2019). Swarm intelligence-based approach for educational data classification. *Journal of King Saud University-Computer and Information Sciences*, *31*(1), 35-51.

[16] Prada, M. A., Domínguez, M., Vicario, J. L., Alves, P. A. V., Barbu, M., Podpora, M., ... & Vilanova, R. (2020). Educational data mining for tutoring support in higher education: A web-based tool case study in engineering degrees. *IEEE Access*, *8*, 212818-212836.

[17] Mengash, H. A. (2020). Using data mining techniques to predict student performance to support decision making in university admission systems. *IEEE Access*, *8*, 55462-55470.

[18] Injadat, M., Moubayed, A., Nassif, A. B., & Shami, A. (2020). Multi-split optimized bagging ensemble model selection for multi-class educational data mining. *Applied Intelligence*, *50*(12), 4506-4528.

[19] Cerezo, R., Bogarín, A., Esteban, M., & Romero, C. (2020). Process mining for self-regulated learning assessment in e-learning. *Journal of Computing in Higher Education*, *32*(1), 74-88..

[20] Karthikeyan, V. G., Thangaraj, P., & Karthik, S. (2020). Towards developing hybrid educational data mining model (HEDM) for efficient and accurate student performance evaluation. *Soft Computing*, *24*(24), 18477-18487.

[21] Ashraf, M., Zaman, M., & Ahmed, M. (2020). An intelligent prediction system for educational data mining based on ensemble and filtering approaches. *Procedia Computer Science*, *167*, 1471-1483.

[22] Injadat, M., Moubayed, A., Nassif, A. B., & Shami, A. (2020). Systematic ensemble model selection approach for educational data mining. *Knowledge-Based Systems*, *200*, 105992..

[23] Dabhade, P., Agarwal, R., Alameen, K. P., Fathima, A. T., Sridharan, R., & Gopakumar, G. (2021). Educational data mining for predicting students' academic performance using machine learning algorithms. *Materials Today: Proceedings*.

[24] Malini, J., & Kalpana, Y. (2021). Investigation of factors affecting student performance evaluation using education materials data mining technique. *Materials Today: Proceedings*, *47*, 6105-6110.

[25] Nabil, A., Seyam, M., & Abou-Elfetouh, A. (2021). Prediction of Students' Academic Performance Based on Courses' Grades Using Deep Neural Networks. *IEEE Access*, *9*, 140731-140746.

[26] Huang, L., Wang, C. D., Chao, H. Y., Lai, J. H., & Philip, S. Y. (2019). A score prediction approach for optional course recommendation via cross-user-domain collaborative filtering. *IEEE Access*, *7*, 19550-19563.

[27] Ibrahim, M. E., Yang, Y., Ndzi, D. L., Yang, G., & Al-Maliki, M. (2018). Ontology-based personalized course recommendation framework. *IEEE Access*, *7*, 5180-5199.

[28] Zhao, Z., Yang, Y., Li, C., & Nie, L. (2020). GuessUNeed: Recommending Courses via Neural Attention Network and Course Prerequisite Relation Embeddings. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *16*(4), 1-17.

[29] Esteban, A., Zafra, A., & Romero, C. (2020). Helping university students to choose elective courses by using a hybrid multi-criteria recommendation system with genetic optimization. *Knowledge-Based Systems*, *194*, 105385.

[30] Yifan Zhu, Hao Lu, Ping Qiu, Kaize Shi, James Chambua, Zhendong Niu. (2020), Heterogeneous teaching evaluation network based offline course recommendation with graph learning and tensor factorization, *Neurocomputing*,415,84-95.

[31] Yanes, N., Mostafa, A. M., Ezz, M., & Almuayqil, S. N. (2020). A Machine Learning-Based Recommender System for Improving Students Learning Experiences. *IEEE Access*, *8*, 201218-201235.

[32] Fernández-García, A. J., Rodríguez-Echeverría, R., Preciado, J. C., Manzano, J. M. C., & Sánchez-Figueroa, F. (2020). Creating a Recommender System to Support Higher Education Students in the Subject Enrollment Decision. *IEEE Access*, *8*, 189069-189088.