Predicting Students Performance by Using Data Mining Methods

Abbood Kirebut Jassim
Department of Computer Science,
College of Science for Women,
University of Baghdad, Baghdad, Iraq
abboodkj_comp@csw.uobaghdad.edu.iq
orcid.org/0000-0003-4211-6493

Ahmed Al-Taie
Department of Computer Science,
College of Science for Women,
University of Baghdad, Baghdad, Iraq
a.altaie@csw.uobaghdad.edu.iq

Zaid S. Naama
Department of Computer Science,
College of Science for Women,
University of Baghdad, Baghdad, Iraq
zaidsn_comp@csw.uobaghdad.edu.iq

DOI: http://dx.doi.org/10.31642/JoKMC/2018/100202

Received Jan. 28, 2023. Accepted for publication Mar. 20, 2023

***Abstract*—** *The corona pandemic disrupted the educational process, especially in universities that use traditional education. Universities were therefore obliged to move from traditional education to e-learning without adequate preparations. The aim of this research is to analyze the students' performance in the two educational environments and predict the result of any of them in the future. The k-means algorithm, an important data mining method, was used to analyze the results of the fourth-stage classes of five consecutive years of students from one Iraqi university's scientific departments. Four of these years were traditional education, while the last was e-learning to see whether the student's performance distribution is normal or abnormal. The results indicate a 100 percent of students' success rate in e-education, while the upper limit is 70 percent for the previous years. Moreover, the average class rate increased to 75 percent compared to 62 in previous years. The decision tree has been built based on a dataset created from the collected data to predict the distribution of both traditional and e-learning with a 2% error tolerance. The study shows that using the exact mechanism in e-learning will give abnormal results. Therefore, the study recommends the need for good infrastructure, the preparation of efficient staff, increasing students' skills, and appropriate software platforms for an accurate assessment of students' performance.*

Keywords*— *Data Mining, E-Education, K-means Clustering, Rand Index, Traditional Education.

I. INTRODUCTION

Data mining is a process of extracting valuable information from large datasets [1]. Data mining methods can be used to predict student performance by analyzing historical data such as past test scores, attendance records, and demographic information [2]. These methods can identify patterns and relationships in the data that can be used to make predictions about future student performance [3]. Some common data mining techniques used in educational contexts include decision trees, neural networks, and association rule mining [4]. These methods can be used to identify at-risk students, predict future success, and inform decision-making about student support and intervention [5]. K-means is a clustering algorithm that groups similar data points together [6]. In the context of predicting student performance, k-means can be used to group students with similar characteristics, such as academic background and performance on assessments [7]. Once the groups are formed, the performance of students in each group can be analyzed to identify patterns and trends [8]. This information can then be used to predict the performance of new students who are similar to those in the existing groups [9]. Decision tree is a type of algorithm that creates a model in the form of a tree structure. It uses a set of rules to make predictions about the outcome of a particular event [10]. In the context of predicting student performance, decision tree can be used to analyze data on student demographics, academic history, and performance on assessments [11]. The algorithm will identify the most important factors that contribute to student performance and create a model that can be used to predict future performance [12]. Both K-means and decision tree are powerful data mining methods that can help educators make accurate predictions about student performance [13]. These

predictions can be used to provide targeted support and interventions to help students succeed [13].

In this paper, we conducted a comparison between the performance of graduate students (in the final stages) of one of the scientific departments in an Iraqi university before and after the spread of the pandemic (that is, under both the traditional education and the e-learning environment) using data mining techniques (K-means) to find out which of the two environments gives better students' performance distribution. As well as building a decision tree to predict outcomes if education is adopted using either of the two environments. The results showed that traditional education reflects the normal distribution of students' performance while the distribution for E-learning was abnormal. Therefore, the study recommended that educational institutions study the real reasons and conduct a review of e-learning or combine the positives of the two environments.

II. RELATED WORK

The COVID-19 pandemic has fundamentally altered how people learn. The learning has moved from offline to online throughout this pandemic. Educational data mining has become an active tool in discovering hidden relationships in educational data and forecasting students' academic performance. Therefore, many researchers try to analyze and predict student performance, admissions decision-making for universities and analyzing e-education. In this section a brief exploring for these works.

Verma et al. proposed some machine learning approaches to predict student learning performance. Educational data were analyzed at the Open University (OU) based on performance, engagement, and demographic criteria. Throughout the experimental analysis All the compared methods on the OU dataset fared best in the experimental analysis, with the k-NN approach performing best in certain circumstances and ANN performing best in others. [14]

Mohamed et al. collected data on online learning behavior in different countries during the COVID-19 pandemic. The final results of three critical regression approaches, such as linear regression, multilayer regression, and SMO regression, were applied. The results show that the method for SMO regression generates fewer errors with improved accuracy compared to others. Furthermore, different scholars might enhance this study by using various approaches. [15]

De Oca et al. sought to identify professorial concerns after the shift to distance education in the first 15 months of confinement. The results evidenced that the professors created a kind of social network, sharing tips and digital media as educational resources, which led to a natural learning curve for developing online teaching competencies. Other relevant findings included the need to provide the professors with continuous training in communication and learning management platforms to engage in ongoing discussions on topics such as whether turning on the cameras should be compulsory during online lectures. The results of this work have value for higher education institutions and professors seeking a better understanding of their requirements and decision-making to improve education delivery under current and future constraints. [16]

Alsharhan et al., worked out to forecast overall performance in this study and looked into a practical method for analyzing a dataset of 480 Middle Eastern students using three supervised machine learning techniques: artificial neural networks, decision trees, and Naive Bayes. Efficiency increased when using SPSS. According to the results, the artificial neural networks algorithm had the lowest variance, with a standard deviation of 2.37, while the naive Bayes approach had the best accuracy, at 89.85%. In addition, there are other useful insights beyond that precision. Additional machine learning methods, such visual representation and normalized relevance of independent variables, can be offered in the SPSS environment. Additionally, the data set was assessed to see if e-learning factors alone could predict student success in the case of missing student data. [17]

The researchers focused on feelings and satisfaction with the direction of distance learning or students' performance in e-learning. The proposed paper dealt with studying the differences in performance between traditional and e-learning.

III. THE PROPOSED SYSTEM METHODOLOGY

The Proposed System has six distinct steps; data collection, data Cleaning, distributing student mark's vectors in three levels, combining data for each level from all years into one dataset, and clustering each level into two clusters then build decision tree to predict of results as shown in Fig. 1.

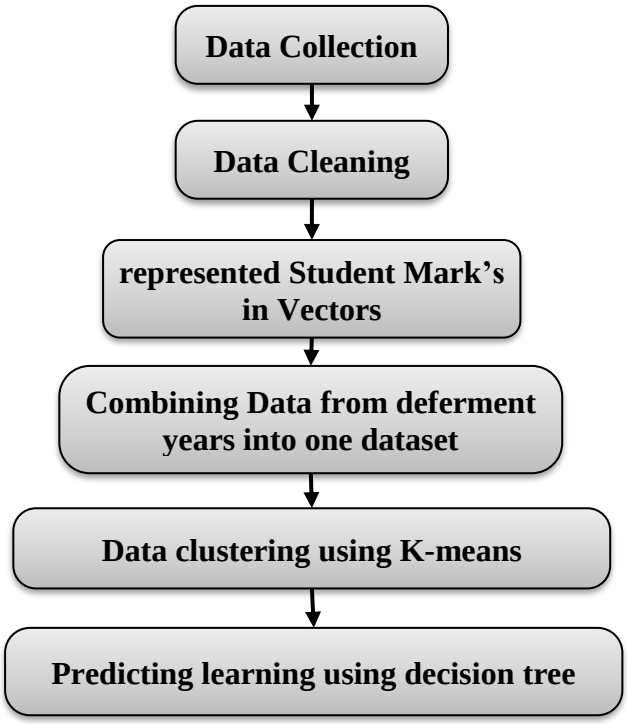

Fig. 1. Block Diagram of Proposed System

A. data collection

For the first step, data of students' marks are collected from one of the science departments in Iraqi universities, as shown in Table (1). The marks represent the latest five years. Corona pandemic occurred in the last year of them. Therefore, the Traditional learning are suspended, but distance education applies with electronic systems, such as Google Classroom and Google Meet.

TABLE I. FINAL YEAR STUDENTS' MARKS SUMMARIZATION

Final Year Students' Marks Summarization				
Year	***Class Students***	***First Attempt Class Success Rate***	***Second Attempt Class Success Rate***	***Class Marks Average***
2015-2016	40	%50.94	%77	62.30
2016-2017	31	%31.25	%60.41	60.09
2017-2018	48	%50.87	%82	59.59
2018-2019	38	%70.00	%95	61.0
2019-2020	20	%100	-	75.85

From Table (I), an apparent problem arises in the discrepancy between the e-learning year and traditional learning years related to success rates and the class average. This problem has been analyzed in depth through data mining techniques to develop appropriate solutions for it in the future through competent academic authorities.

The data set consists of attribute vectors. Each vector consists of final year subject's marks which are (operating systems, cloud computing, computers security, mobile computing, computer networks, website design, data security, and communications). Because of the differences between students in abilities, giving, initiative, response to changes, and adaptation, students are divided into three levels.

B. Data Cleaning

Data cleaning is the process of preparing raw data for analysis by removing bad data, organizing the raw data, and filling in the null values. In this step, the students' records were processed by deleting the data of the dismissed and absent students, as well as processing and correcting errors in the data and placing them in a revised file from all defects.

C. Distributing Student Mark's Vectors

A vector was created for each student which containing the student's grades. where each cell represents the student's grade in a specific subject, so that the students' grade in each subject correspond to the same location in the cell for all vectors in the dataset. The use of vectors for the purpose of forming that necessary database when dealing with data mining algorithms. Due to the differences between students in abilities, giving and taking initiative, response to changes, and adaptation, students are divided into three levels (High performance students, middle performance students and low performance students). To ensure neutrality, K-means is used to automatically divide each year's data into three levels. Also, this algorithm is used later in the main objective of this research as a data mining lustering-task. Distribution was only three levels due to the limitation of volume of data available. The High performance students data for all years was collected in one dataset, along with middle performance students and low performance students. Therefore, there are three datasets, one dataset for each level. there are three datasets.

D. Combining Data

Data combining, also known as data integration, is the process of bringing data together from multiple sources into a single dataset. Data from each level of student performance in different years was combined into one dataset (one dataset for each level). in order to conduct mining operations on each dataset individually.

E. Data Clustering

The K-means mining algorithm groups each level into two groups. Determines if there is an overlap between student records from both environments (traditional or e-learning) or separate them. If there is an overlap, that is, the presence of students from both environments in each cluster, this means that the evaluation is fair. But if the students of each environment or the majority are present in an independent cluster, this indicates that the evaluation is biased.

The less the overlap between the grades of the students in the two clusters, the more this leads to the different performance of the students in the two environments, and vice versa.

Results of clusters were evaluated using RAND index methods, which is a measure of similarity between two sets of data.

F. Performance Predicting

The concentrated data was extracted from the results of the Kmeans algorithm to create a dataset that was used to build a prediction model It relies on the decision tree to determine the expected paths. for both traditional education and e-learning.

IV. RESULTS AND DISCUSSION

K-means algorithm distributes student records (attributes vector) of each into three levels and The dataset of each level was collected from all years that means they represent the dataset of High performance students, middle performance students and the dataset of low performance students as shown in Table (II).

TABLE II. THREE LEVELS OF STUDENTS FOR EACH YEAR

Three Levels of Students for Each Year				
Year	***High performance students***	***Middle performance students***	***Low performance students***	***Total***
2015-2016	5	24	11	40
2016-2017	8	17	6	31
2017-2018	4	20	24	48
2018-2019	7	13	18	38
2019-2020	4	8	8	20
All years	28	82	67	177

K-means is used for its majority task (mining task), which is to divide each dataset into two clusters. The result of clustering of all datasets are as follows in Table (III).

TABLE III. TWO CLUSTERS RESULTS OF PERFORMANCE STUDENTS

Cluster id	**High performance students**		
	Cluster1	***Cluster2***	***Total***
2015-2016	4	1	5
2016-2017	0	8	8
2017-2018	0	4	4
2018-2019	0	7	7
2019-2020	4	0	4
Cluster id	**Middle performance students**		
	Cluster1	***Cluster2***	***Total***
2015-2016	4	20	24
2016-2017	0	17	17
2017-2018	0	20	20
2018-2019	0	13	13
2019-2020	8	0	8
Cluster id	**Low performance students**		
	Cluster1	***Cluster2***	***Total***
2015-2016	0	11	11
2016-2017	0	6	6
2017-2018	0	24	24
2018-2019	0	18	18
2019-2020	8	0	8

For the high-performance students of the first cluster, all data on e-learning was collected in the first cluster by (%100) and (%16.6) of the data on traditional education was collected in the first cluster too, while (%83.4) of traditional education was grouped in the second cluster.

The following chart shows the results in Fig. 2.

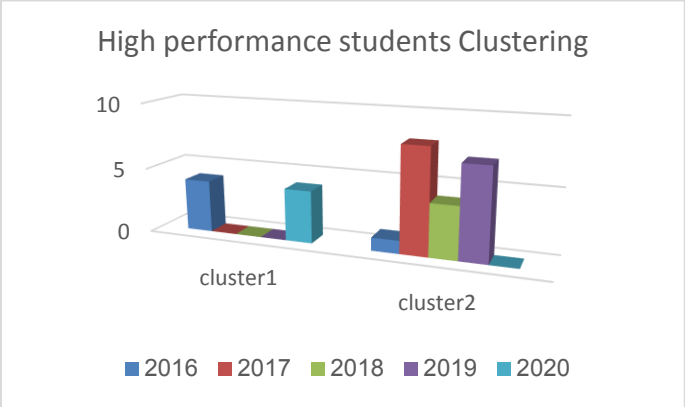

Fig. 2: Chart of Level 1 Clustering

In the middle performance students, all data of the e-learning were grouped in the first cluster by (%100) and (%0.057) of the data of the traditional education were grouped in the first cluster too, while (%99.943) of the traditional education was collected in the second cluster. The following chart shows the results in Fig. (3).

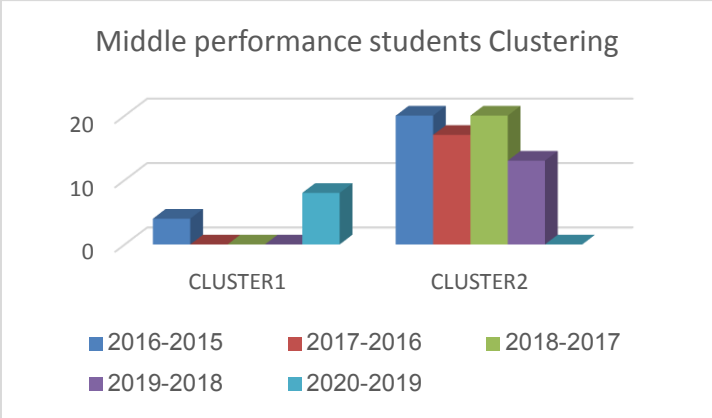

Fig. (3): Chart of Level 2 Clustering

For the low performance students, all data on e-learning was collected in the first cluster at a rate of (%100), while (%100) of the data on traditional education was collected in the second cluster. The following chart shows the results in Fig. (4).

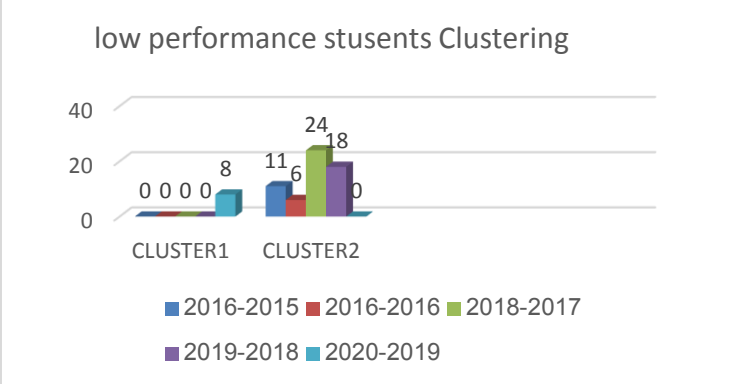

Fig. (4): Chart of Level 3 Clustering

Through the previous information, the results have been summarized for each level to compare attendance education and e-learning. The summarization is as follows in Table (IV).

TABLE IV. TWO CLUSTERS RESULT OF E-LEARNING AND ATTENDANCE LEARNING

Cluster id	Level-1		
	Cluster1	*Cluster2*	*Total*
Elearning year	4	0	4
Attendance years learning	4	20	24
Cluster id	**Level-2**		
	Cluster1	*Cluster2*	*Total*
Elearning year	8	0	8
Attendance years learning	4	70	74
Cluster id	**Level-3**		
	Cluster1	*Cluster2*	*Total*
Elearning year	8	0	8
Attendance years learning	0	59	59

From the result of data clustering task, it turns out that overlapping at higher performance is relatively greater because outstanding students have more ability than others to adapt to changing education methods.but In the middle performance, the overlap decreased. While in the lowperformance, the overlap has completely disappeared between the results of the performance of students in electronic and traditional education. Accordingly, the results clearly indicate that a fair measure of student performance is not achieved in the traditional and electronic learning environments, despite the fact that they are of the same educational level and deal with the same materials.

TABLE V. FINAL RESULTS SUMMARIZATION

Final Results Summarization				
Cluster id	*Elearning year*	*Percentage rate*	*Attendance years*	*Percentage rate*
Cluster 1	20	%100	12	%0.764
Cluster 2	0	%0	145	%92.236

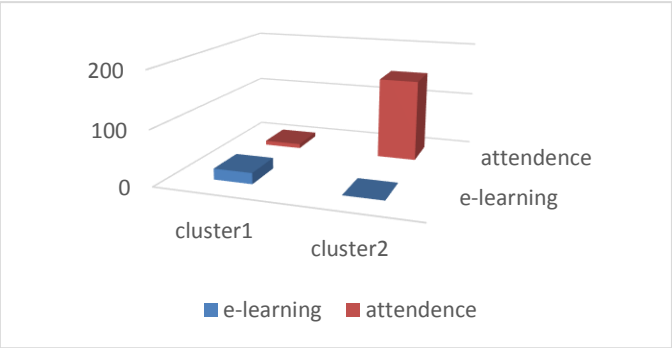

Fig. 5. : The Chart of Final Results Summarization

The final result evaluated the accuracy by rand index using Eq. (1) as follow

$$RI=TP+TN/TP+FP+TN+FN \quad (1)$$

$$RI=20+145/20+0+145+12$$

$$RI=0.932$$

The Rand Index (RI) should be interpreted as follows: -

RI >= 0.90 indicates excellent recovery; 0.80 = RI 0.90 indicates good recovery; 0.65 = RI 0.80 indicates moderate recovery; and RI 0.65 indicates poor recovery. Therefore, the results of the clustering evaluation are the highest ranking. After reviewing the above results, a questionnaire was conducted for the department's lecturers about the reasons for this discrepancy in evaluation, in addition to the high success rates and high class average. The reasons were as follows in table (VI).

TABLE VI. LECTURERS A QUESTIONNAIRE

Lack of lecturer training in the e-learning environment	%78.5
The lack of suitable infrastructure	%67.8
There is no strict monitoring of the student's performance during the exam.	%92.8
The failure of traditional education	%0
Excellence in e-learning	%0

According to the results of the above questionnaire of the opinions of the lecturers, there is no effect of the nature of education, whether traditional or electronic, on the clear difference in students' performance. But the effect was due to the reasons for the lack of lecturer training in the e-learning environment, the lack of suitable infrastructure, and the lack of strict monitoring of the student's performance during the exam.

To predict the student's performance if that was done under the same educational conditions. Based on the clustering results and the summary in Table (1) were used to build the decision tree for their first attempt with a tolerance (±2) as shown in Table (VII).

TABLE VII. A SUMMARIZATION THAT CAN BE USED TO BUILD THE DECISION TREE

A Summarization that can be used to build the decision tree			
Learning	*Class pass Rate*	*Class marks Average*	*Target*
Traditional education	64.57±2	6.75±2	Yes
Traditional education	99±1	75.85±2	No
E-learning	64.57±2	6.75±2	No
E-leaning	99±1	75.85±2	Yes

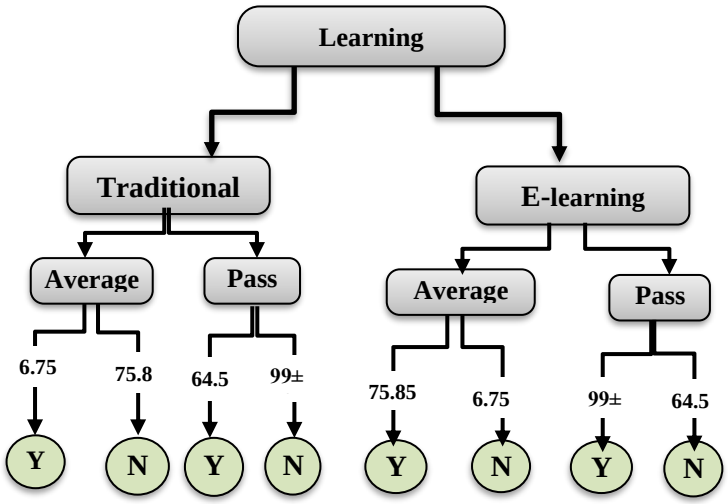

Fig. 6. :The significant contrast between Traditional education and E-education

The results in Figure (6) showed significant contrast between Traditional education and E-education, which shows the importance of reassessing the experience of e-learning by

education officials. This model proves the contrast in the results of the performance between e-learning and traditional education in this department, which is an example of what happened in all universities that applied electronic education suddenly and had not studied it due to the Corona pandemic.

V. CONCLUSION

The evaluation of educational institutions is critical in measuring the sobriety of universities, the quality of education, and the preservation of their academic reputation. After the emergence of the coronavirus pandemic, most universities switched from traditional education to e-e-leaning. This transformation was sudden and unplanned. This skewed in results in pass rates and grade point averages between years of traditional education and e-learning. Research using data mining contributes to data analysis and prediction of outcomes when either of the two educational environments is adopted. Accordingly, the academic decision maker can address the imbalance in the use of e-learning by training faculty and creating a sound infrastructure that maintains the continuity of natural outcomes in student assessment. The deviation in the evaluation of performance in the success rate of e-learning compared to traditional education was %30, while the deviation from the average of the class was %13. The limitation of the research was the lack of available data due to the fact that it belonged to one department.

REFERENCES

[1] J Han, m kamber, & j pei, .Data mining practical machine learning tools and techniques 3rd ed, Morgan Kaufmann Puplishers,Burrlington,2011.

[2] [2] B Albreiki,, N Zaki,& H Alashwal, A systematic literature review of student'performance prediction using machine learning techniques ,Education Sciences ,VOL 11, NO 9,PP 552, MDPI,2021.
https://doi.org/10.3390/educsci11090552

[3] N A Yassein, R G Helali,S B,& others, Predicting student academic performance in KSA using data mining techniques, Journal of Information Technology \& Software Engineering, VOL 7, NO 5,PP 1-5,2017

[4] W F Yaacob, WF Wan,N M Sobri, S M Nasir,N H Norshahidi,& WZ Wan,Predicting student drop-out in higher institution using data mining techniques, Journal of Physics: Conference Series},VOL 1496,NO 1,PP 012005,2020.
https://DOI 10.1088/1742-6596/1496/1/012005

[5] H VanDer , M B Amanda ,& K Matthew,Improving decision making in school psychology: Making a difference in the lives of students, not just a prediction about their lives, School Psychology Review, VOL 47,NO 4, PP 385-395, National Association of School Psychologists 4340 East West Highway, Suite,2018.

[6] Kansal, Tushar and Bahuguna, Suraj and Singh, Vishal and Choudhury,& Tanupriya ,Customer segmentation using K-means clustering,{2018 international conference on computational techniques, electronics and mechanical systems (CTEMS),PP 135-139,IEE,2018.
http//DOI: 10.1109/CTEMS.2018.8769171

[7] A Iatrellis, I K Savvas, P Fitsilis,& V C Gerogiannis, A two-phase machine learning approach for predicting student outcomes, Education and Information Technologies,VOL 26,PP 69-88,Springer,2021

[8] J Lee,Racial and ethnic achievement gap trends: Reversing the progress toward equity?,Educational researcher, VOL 31 ,NO 1, 3-12,PP 3-12, Sage Publications Sage CA: Thousand Oaks, CA,2002.

[9] E Osmanbegovic, ,& M Suljic, Data mining approach for predicting student performance, Economic Review: Journal of Economics and Business, VOL 10, NO 1, PP 3-12, Tuzla: University of Tuzla, Faculty of Economics,2012

[10] P Odeyar, B Apel, B Derek, R Hall, B Zon, & K Skrzypkowski, A Review of Reliability and Fault Analysis Methods for Heavy Equipment and Their Components Used in Mining, Energies, VOL 15, NO 17, MDPI,2022.
https://doi.org/10.3390/en15176263

[11] Y Baashar, G Alkawsi, Gamal,N Ali, H Alhussian,& H T Bahbouh, Predicting student's performance using machine learning methods: A systematic literature review}, 2021 International Conference on Computer \& Information Sciences (ICCOINS),PP 357-362, IEEE,2021.
http//DOI: 10.1109/ICCOINS49721.2021.9497185

[12] Y Baashar, G Alkawsi, A Mustafa, A A Alkahtani, Y A Alsariera, A Q Ali, Abdulrazzaq W Hashim, T Wahidah ,& S K Tiong, Toward predicting student's academic performance using artificial neural networks (ANNs), Applied Sciences, VOL 12, NO 13, PP 1289, MDPI,2022. https://doi.org/10.3390/app12031289

[13] B K Francis, & S S Babu, Predicting academic performance of students using a hybrid data mining approach , VOL 43, PP 1-15, Springer,2019

[14] B. K. Verma ,N. and Srivastava &, H. S.K Singh,Prediction of Students' Performance in e-Learning Environment using Data Mining/Machine Learning Techniques, vol 23,pp 586--593,2021.

[15] D. Y. Mohammed, "The web-based behavior of online learning: An evaluation of different countries during the COVID-19 pandemic," Advances in Mobile Learning Educational Research, vol. 2, no. 1, pp. 263–267, 2022. DOI 10.25082/AMLER.2022.01.010

[16] S. M. De Oca, M. Villada-Balbuena, and C. Camacho-Zuniga, "Professors' concerns after the shift from face-to-face to online teaching amid covid-19 contingency: An educational data mining analysis," 2021 Machine Learning-Driven Digital Technologies for Educational Innovation Workshop, 2021. DOI: 10.1109/IEEECONF53024.2021.9733778

[17] A. Alsharhan, S. A. Salloum, & A. Aburayya, Using e-learning factors to predict student performance in the practice of precision education, Journal of Legal, Ethical and Regulatory Issues, vol 24, pp1--14, Jordan Whitney Enterprises, Inc,2021.