EAI Endorsed Transactions
on Internet on Things

Research Article EAI.EU

Predicting Academic Success: A Comparative Study of Machine Learning and Clustering-Based Subject Recommendation Models

Kinjal[1], Sagar Mousam Parida[2], Jayesh Suthar[3] and Sagar Dhanraj Pande[4, *]

Abstract

The study of students' academic performance is a significant endeavor for higher education schools and universities since it is essential to the design and management of instructional strategies. The efficacy of the current educational system must be monitored by evaluating student achievement. For this research, we used multiple Machine Learning algorithms and Neural Networks to analyze the learning quality. This study investigates the real results of university examinations for B.Tech (Bachelor in Technology) students, a four-year undergraduate programme in Computer Science and Technology. The K-means clustering approach is used to recommend courses, highlighting those that would challenge students and those that will improve their GPA. The Linear Regression method is used to make a prediction of a student's rank among their batchmates. Academic planners might base operational choices and future planning on the findings of this study.

Keywords: K-means Clustering, Neural Networks, Academic Performance, Prediction, Recommend

Received on 18 December 2023, accepted on 04 March 2024, published on 12 March 2024

doi: 10.4108/eetiot.5378

*Corresponding author. Email: sagarpande30@gmail.com

1. Introduction

Students' academic performance is presently being evaluated at the majority of universities and colleges for a variety of reasons. A multitude of conditions influence pupils' academic success. The determinants of academic success are known to vary across different student populations, academic backgrounds, and geographical locations. Such variations may be attributed to a range of factors that are unique to each context. Despite their best efforts, students' grades may not always represent their efforts. This might result in a high failure rate, raising the training rate for these graduates. Low graduation rates and a reduction in the intake of potential users due to a shortage of places are two implications of low pass rates for students that may have a substantial financial impact on the community.

Higher education is a must for every organization in order to land high-profile positions and earn more money than is reasonable. Many studies have been done to examine the variables linked to students' academic success at different institutions, but less studies have been done for private colleges. When compared to public universities, the cost of education at private colleges is significantly higher. The mere fulfilment of admission prerequisites to a higher education institution does not guarantee a favourable outcome in the pursuit of a degree programme. However, a number of variables that might contribute to a student's success in life can also have an impact on their academic achievement. To continue to assure enhanced quality in universities, it is essential for universities and colleges to engage in tracking and assessing indicators of excellence across various domains. This approach is critical for improving the overall quality of education and ensuring that students receive an outstanding educational experience.

Specifically, educational institutions and colleges must focus on evaluating excellence in areas such as academic programmes, faculty performance, student outcomes, research activities, and community engagement. By conducting rigorous evaluations in these domains, institutions can identify areas for improvement and implement targeted interventions to enhance the quality of education and overall institutional performance.

Ultimately, this approach can help to ensure that private universities and colleges are able to offer students the utmost level of academic excellence possible, while also contributing to the broader goals of academic excellence and societal progress. In order to help students, improve their individual academic success at the institution, accessing the variables for academic achievement is also crucial. The goal of this research study is to concentrate on the interactions between various elements and the contributions they make to students' academic achievement at their college or institution.

Through a comprehensive analysis of relevant literature and empirical data, this study seeks to shed light on the determinants that contribute to students' academic success in institutions of higher learning. This study aimed to investigate the multifarious aspects that impact the performance of pupils in university examinations. The research sought to establish a connection among such variables and their influence on students' academic outcomes. The latest research analysis has produced encouraging outcomes.

The K-Means Clustering technique is used to various subject marks earned by students, distinguishing courses based on difficulty level, and may be used to recommend subjects to students as a reference for future batch students [13]. Linear Regression Model is used to build an efficient ranking system where a student can visualize their current position in comparison to their batchmates. We have used Data pre-processing and Analysis to plot various graphs. Different methodologies, such as clustering algorithms, might be used in doing analyses, such as student outcomes, to provide valuable insights into the factors that influence performance and their implications for individuals and organizations.

The ultimate objective of this research is to gain a comprehensive understanding of the method of instruction and the subject knowledge acquired by students, with the aim of identifying areas for improvement that can enhance academic achievement.

This research paper makes use of a comprehensive dataset containing the grades earned by 500 students in 10 distinct subjects. The dataset was methodically acquired and constructed, ensuring its uniqueness and dependability. This collection of grades provides a valuable resource for analysing the academic performance of the investigated student population. With its wide variety of subjects, the dataset provides a comprehensive view of the educational achievements of students, allowing researchers to investigate various aspects of their academic journey.

2. Literature Review

The prediction of academic success has been a longstanding interest in the field of education. With the advent of machine learning and clustering techniques, there have been new opportunities to delve deeper into this topic. The objective of this literature review is to present a thorough analysis of the current body of research pertaining to the prediction of academic success through theoretical and empirical research, case studies, and scholarly articles.

Table 1. Related Work with their Implementation

Papers	Implementation
Education 4.0 – Fostering Student's Performance with Machine Learning Methods	Uses Neural Networks, SVM, Decision Tress, Cluster analysis Clusters students from analysis.
Enhancing prediction of student success: Automated machine learning approach panel	Relies on Auto – ML to increase prediction accuracy Uses Data Features for prediction
Explainable Student Performance Prediction Models: A Systematic Review	Uses Forward and Backward Snowballing technique. Decision trees and deep learning models were used.
An intelligent tutoring system for supporting active learning: A case study on predictive parsing learning	In this study, a smart teaching system was designed, developed, and applied to facilitate the acquisition of predictive parsing techniques. [15]
Predicting Students' Performance Using Machine Learning Techniques	This study employs various machine learning techniques, including Artificial Neural Network, Naïve Bayes, Decision Tree, and Logistic Regression using ROC index performance index. [16]
A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs	Latent factor model-based course developed. Assemble based progressive prediction architecture developed.

3. Methodology

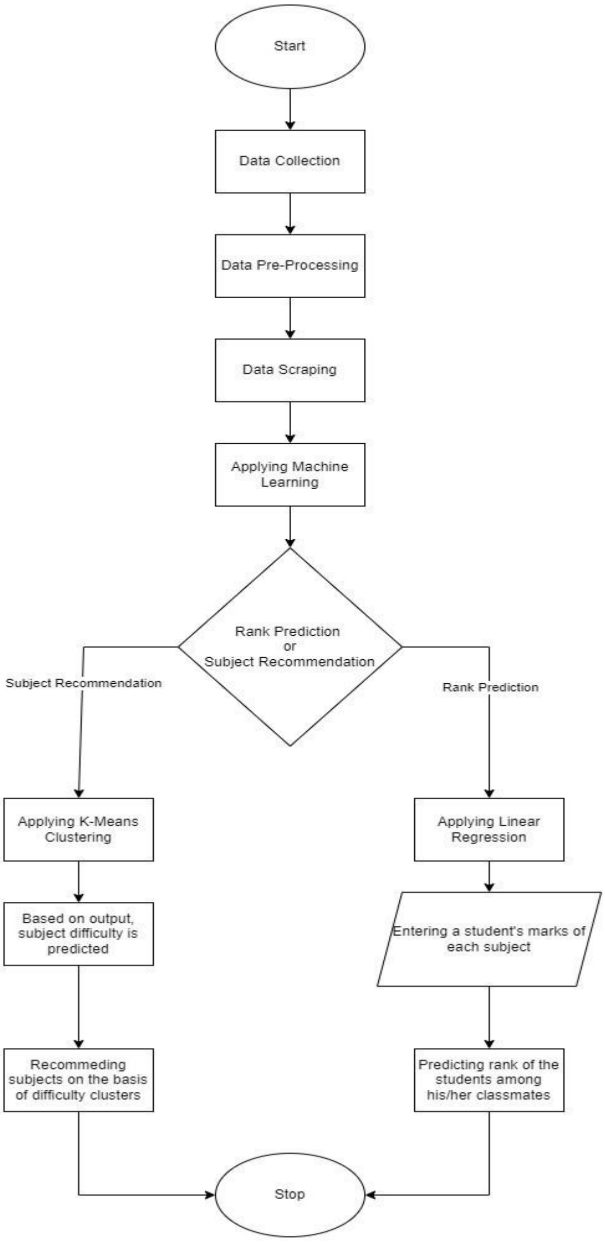

Start
Data Collection
Data Pre-Processing
Data Scraping
Applying Machine Learning
Rank Prediction or Subject Recommendation
Subject Recommendation
Rank Prediction
Applying K-Means Clustering
Based on output, subject difficulty is predicted
Recommeding subjects on the basis of difficulty clusters
Applying Linear Regression
Entering a student's marks of each subject
Predicting rank of the students among his/her classmates
Stop

Figure 1. Data Flow Diagram

3.1. Data Scraping

The dataset used in our analysis is scraped out from a larger university dataset which contains various factors like marks, grades and number of attempts of appearing the exam for each subject in every semester. We structured the dataset using Data Preprocessing methods to prepare it for further analysis. Out of which we selected factors like marks and grades of one semester for our observation.

3.2. K-Means Clustering

Clustering was employed as a statistical method in order to conduct the present research. Clustering is described as a collection of algorithms used to locate subgroups of observations inside a given data set, which in our case is of the topics in that semester in order to determine the difficulty level. When we cluster the data, we receive the results that individuals in the same cluster are of similar difficulty. The technique of clustering is a type of unsupervised machine learning.

$$K = \sum_{m=1}^{M} \sum_{n=1}^{N} a_{mn} ||i_m - l_n||^2$$

M: total number of data points
N: number of clusters
i_m: vector of measurement n
l_n: mean for cluster k
a_{mn}: an indicator variable that indicates whether to assign i_m to n
We need to determine the value of $\{a_{mn}\}$ and in that gives the least value of K.

The present study opted for the aforementioned strategy due to its capacity to unveil correlations among the gathered observations. The present study aims to investigate the factors that impact students' academic performance in university examinations. Clustering helps us to determine which subjects can be recommended, allowing students to choose subjects accordingly and enabling teachers to enhance their pedagogical skills [11]. As a result, we found the that 10 subjects were divided into 4 clusters according to their difficulty level.

3.3. Linear Regression

Predicting the outcome of a single parameter from the value of another is known as linear regression analysis, and it is a common statistical technique. The primary objective of linear regression analysis is to establish a linear relationship between two variables, where one variable is considered the dependent variable, and the other variable is considered the independent variable. In the present study, the variable of interest, namely marks, is designated as the dependent variable. The process of making predictions about future outcomes is commonly accomplished by fitting a straight line or surface to the data points. This approach involves the use of mathematical models to establish a relationship between the variables under consideration. The resulting model can then be used to make predictions about future outcomes based on the observed data. The utilization of simple linear regression calculators involves the implementation of the "least squares" methodology to obtain the optimal-fit line based on a collection of paired information. The "least squares" method is a statistical approach that seeks to minimize the sum of the squared variations between the values that were obtained and the anticipated results of the dependent variable. This technique is widely used in various fields, including economics, engineering, and social sciences, to estimate the parameters of a linear regression model.

$$s = \frac{N\sum(ab) - \sum a \sum b}{N\sum(a^2) - (\sum a)^2}$$

$$c = \frac{\sum b - s\sum a}{N}$$

$$b = sa + c$$

b = Dependent Variable
a = Independent Variable
s = Slope / Coefficient
c = Constant/Intercept

The primary objective of the least squares method is to find the line of best fit that can explain the relationship between the independent and dependent variable. This ML model aims to predict student ranks. The idea behind this analysis to predict ranks of students by their subject marks [12].

Through this, a study can analyze his position in the class and can use this information to identify their weak areas and improve their performance in the future semesters accordingly maintain their GPA and enhance their academic profile.

3.4. System Configuration

- Laptop Model: Lenovo IdeaPad 720s
- Processor: Intel(R) Core (TM) i5-8250U CPU @ 1.60GHz 1.80 GHz
- RAM: 8GB
- Operating System: Windows 11
- Software: Jupyter Notebook v6.5.4

4. Machine Learning Models

The assessment of students' performance is a crucial aspect of educational evaluation, as it enables educators to gauge the effectiveness of their teaching methods and identify areas for improvement. By analyzing the results of student evaluations, educators can gain valuable insights into the strengths and weaknesses of their instructional strategies and make informed decisions about how to optimize their teaching practices. The evaluation process in educational institutions can be a strenuous job due to the huge number of students enrolled in a single institution. The sheer volume of students necessitates a significant amount of manual effort to complete the evaluation process. The present study proposes an automated approach for evaluating student performance through the use of machine learning. The study explores the potential of machine learning techniques to provide a more efficient and objective means of assessing student performance. The proposed approach is expected to contribute to the development of a more accurate and reliable system for evaluating student performance. The research findings suggest that the use of ML algorithms can significantly shoot up the accuracy and efficiency of student performance evaluation. [1][16]

The expected outcomes of this research are categorical values.

Our analytical tools comprise of:

- Support Vector Machine – ‘SVM’
- Logistic Regression
- K-Nearest Neighbor - ‘KNN’
- Artificial Neural Network – ‘ANN’
- Naive Bayes classifier
- Decision Tree
- XG Boost

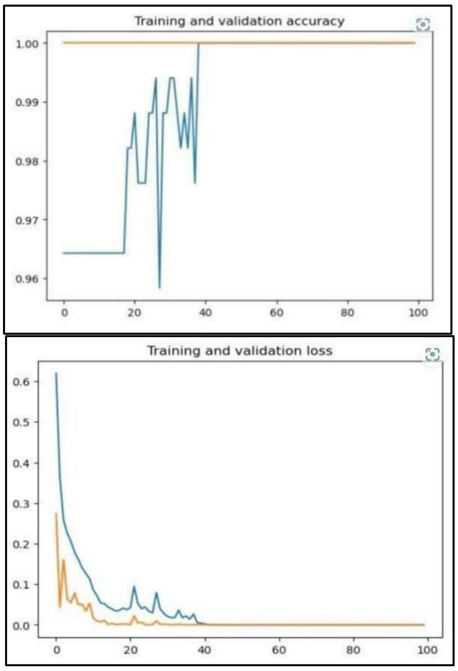


Figure 2. Performance of Artificial Neural Network Model

Table 2. Machine Learning Performance

Different Approaches	Train Accuracy (%)	Validation Accuracy (%)	Test Accuracy (%)	Precision (%)	Recall (%)	Cross Validation Accuracy (%)
ANN	95.91	99.30	97.65	---	---	---
MLP Classifier	83.92	87.50	79.16	98.25	79.16	93.33
KNN	96.42	99.80	100	99.71	88.35	97.50
Logistic Regression	95.74	97.22	97.74	98.79	94.01	94.97
Support Vector Machine	97.68	87.42	88.35	99.84	97.72	98.36
Decision Tree	99.26	83.23	85.97	86.25	85.97	92.18
Naïve Bayes	94.04	97.59	91.66	88.95	91.66	93.75
XG Boost	92.64	89.22	88.95	---	---	93.16

5. Result

Table 3. Abbreviations

Abbreviation	Subject Name
JAVA	Programming using Java
DMS	Discrete Mathematical Structures
CG	Computer Graphics
FDA	Fundamentals of Data Analysis
TOC	Theory of Computation
COA	Computer Organization & Architecture
CN	Computer Networks
DAA	Design & Analysis of Algorithms
OS	Operating Systems
MAD	Mobile Application Development

5.1. Clustering

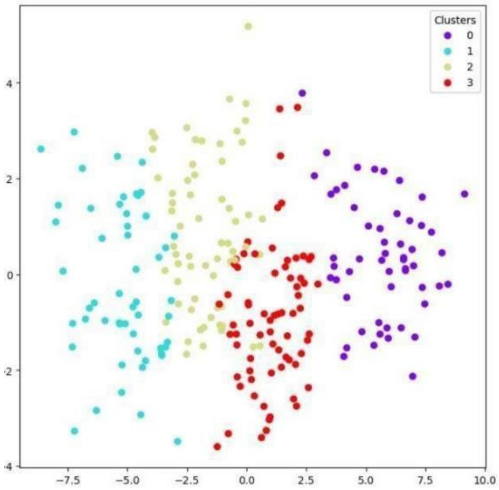


Figure 3. Subject Clusters

From Fig.2, it is clear that four clusters are formed from the ten subjects which are 'JAVA', 'DMS', 'CG', 'FDA', 'TOC', 'COA', 'CN', 'DAA', 'OS' and 'MAD'.

{'COA M': 9, 'DMS M': 8, 'MAD M': 10, 'DAA M': 10, 'TOC M': 9, 'CG M': 8, 'CN M': 10, 'FDA M': 7, 'OS M': 9, 'JAVA M': 8}

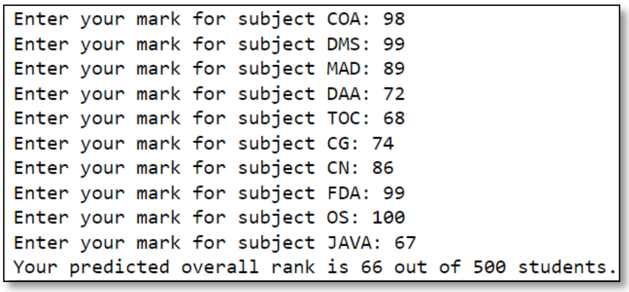
Table 4. Subject Difficulties

Difficulty Cluster	Subject
10	DAA, CN, MAD
9	COA, TOC, OS
8	JAVA, CG, DMS
7	FDA

‘MAD’ and ‘CN’ have a difficulty ranking of 10, which means they were the most difficult subjects based on the clustering analysis. On the other hand, ‘FDA’ has a difficulty ranking of 7, which means it was the least difficult subject.

Overall, the analysis suggests that ‘DAA’ and ‘CN’ were the most difficult subjects, while ‘FDA’ was the least difficult subject.

5.2. Linear Regression



```
Enter your mark for subject COA: 98
Enter your mark for subject DMS: 99
Enter your mark for subject MAD: 89
Enter your mark for subject DAA: 72
Enter your mark for subject TOC: 68
Enter your mark for subject CG: 74
Enter your mark for subject CN: 86
Enter your mark for subject FDA: 99
Enter your mark for subject OS: 100
Enter your mark for subject JAVA: 67
Your predicted overall rank is 66 out of 500 students.
```

Figure 4. Rank Prediction

Here, students enter their marks of each subject. Out of this, they can get the ranking position among their entire batch.

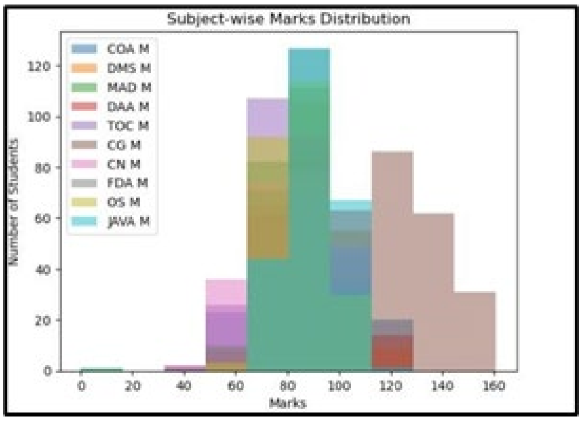


Figure 5(a). Marks Correlation

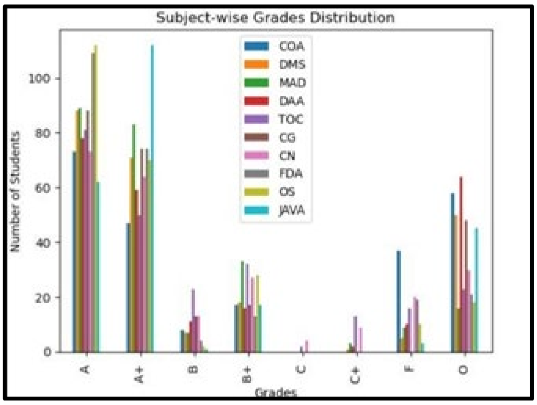


Figure 5(b). Grades Correlation

5.3. Marks Distribution Analysis

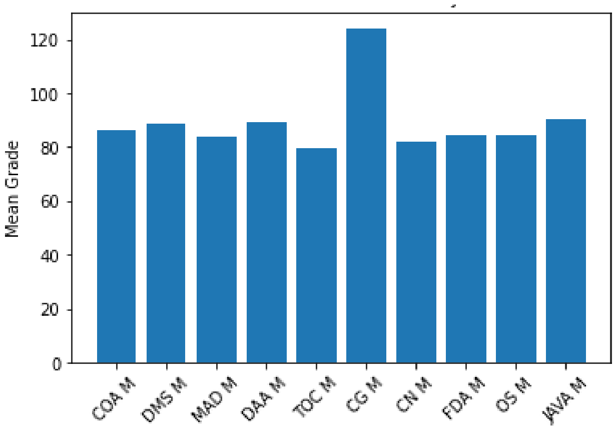


Figure 6(a). Mean Grades for each Subject

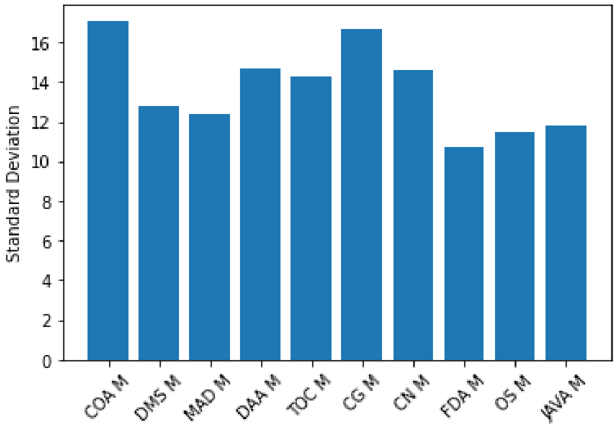


Figure 6(b). Standard Deviation for each Subject

6. Conclusion

Based on the analysis of student grades using K-Means clustering and Linear Regression, it is possible to infer that there is a substantial connection between student performance in different areas. K-Means clustering was used to arrange subjects based on student grades. The results were used to propose subjects in which a student is likely to excel. The topic recommendation model based on K-Means clustering was discovered to be an excellent way for providing students with suggestions on which courses to take based on their prior performance.

Linear regression was also used to predict students' achievement in a certain topic based on their performance in other courses. The results of this study suggest that students' performance in one topic can be used to predict their success in another subject with fair accuracy.

Finally, the study illustrates the feasibility of applying data analysis techniques such as Linear Regression and K-Means clustering to analyze student performance and provide subject suggestions. The findings of the study can help educators and policymakers build effective academic programmes and make informed judgements regarding students' academic paths.

7. Future Scope

The future scope of research in student performance analysis using machine learning is vast and exciting, with numerous potential areas for further exploration and innovation. We can focus various other factors such as assignments, class presentations, topic debates, project work and practical for an enhanced, detailed and efficient model for student academics' prediction and recommendation.

The utilisation of machine learning algorithms has been proposed as a viable approach to develop customised learning plans for students on an individual basis. The potential application of machine learning in identifying the learning style, preferences, strengths, and weaknesses of students is a promising area of research. By leveraging this technology, customised learning plans can be created to optimise academic performance. Further exploration of this approach could yield valuable insights into the effectiveness of personalised education.

References

[1] Ciolacu M, Tehrani AF, Beer R, Popp H. Education 4.0—Fostering student's performance with machine learning methods. In2017 IEEE 23rd international symposium for design and technology in electronic packaging (SIITME) 2017 Oct 26 (pp. 438-443). IEEE.

[2] Xu J, Moon KH, Van Der Schaar M. A machine learning approach for tracking and predicting student performance in degree programs. IEEE Journal of Selected Topics in Signal Processing. 2017 Apr 7;11(5):742-53.

[3] Zeineddine H, Braendle U, Farah A. Enhancing prediction of student success: Automated machine learning approach. Computers & Electrical Engineering. 2021 Jan 1;89:106903.

[4] Wang X, Zhao Y, Li C, Ren P. ProbSAP: A comprehensive and high-performance system for student academic performance prediction. Pattern Recognition. 2023 May 1;137:109309.

[5] Delavari N, Beikzadeh MR, Phon-Amnuaisuk S. Application of enhanced analysis model for data mining processes in higher educational system. In2005 6th international conference on information technology based higher education and training 2005 Jul 9 (pp. F4B-1). IEEE.

[6] Thakar P, Mehta A. Performance analysis and prediction in educational data mining: A research travelogue. arXiv preprint arXiv:1509.05176. 2015 Sep 17.

[7] Agaoglu M. Predicting instructor performance using data mining techniques in higher education. IEEE Access. 2016 May 13;4:2379-87.

[8] Ashfaq U, Booma PM, Mafas R. Managing student performance: A predictive analytics using imbalanced data. International Journal of Recent Technology and Engineering. 2020 Mar 18;8(6):6.

[9] Asif R, Merceron A, Pathan MK. Investigating performance of students: a longitudinal study. InProceedings of the fifth international conference on learning analytics and knowledge 2015 Mar 16 (pp. 108-112).

[10] Asif R, Merceron A, Ali SA, Haider NG. Analyzing undergraduate students' performance using educational data mining. Computers & education. 2017 Oct 1;113:177-94.

[11] Feng G, Fan M, Chen Y. Analysis and prediction of students' academic performance based on educational data mining. IEEE Access. 2022 Feb 15;10:19558-71.

[12] Sravani B, Bala MM. Prediction of student performance using linear regression. In2020 International Conference for Emerging Technology (INCET) 2020 Jun 5 (pp. 1-5). IEEE.

[13] Aggarwal D, Sharma D. Application of clustering for student result analysis. Int J Recent Technol Eng. 2019 Apr;7(6):50-3.

[14] Alamri R, Alharbi B. Explainable student performance prediction models: a systematic review. IEEE Access. 2021 Feb 23;9:33132-43.

[15] Castro-Schez JJ, Glez-Morcillo C, Albusac J, Vallejo D. An intelligent tutoring system for supporting active learning: A case study on predictive parsing learning. Information Sciences. 2021 Jan 12;544:446-68.

[16] Altabrawee H, Ali OA, Ajmi SQ. Predicting students' performance using machine learning techniques. JOURNAL OF UNIVERSITY OF BABYLON for pure and applied sciences. 2019 Apr 1;27(1):194-205.