Enhancing Educational Assessment: Predicting and Visualizing Student Performance using EDA and Machine Learning Techniques

R. Parkavi[1], P. Karthikeyan[2], S.Sujitha[3] and A. Sheik Abdullah[4]
[1,2,3]Department of Information Technology, Thiagarajar College of Engineering, Madurai
[4]School of Computer Science and Engineering (SCOPE), Vellore Institute of Technology, Chennai
[1]rpit@tce.edu [2]karthikit@tce.edu [3]ssiit@tce.edu [4]sheikabdullah.a@vit.ac.in

***Abstract*— Educational institutions must manually evaluate student's performance, which requires a significant amount of faculty time and effort. Teachers can easily monitor and document student's learning behavior if they utilize a learning management system (LMS). It can be utilized by coaching sessions or educational institutions to quickly analyze student's performance. Due to the enormous number of pupils, huge data must be analyzed; teachers frequently encounter challenges. The qualities of learners who are students are described analytically in this analysis. In this work, student's cognitive and psychomotor skills are predicted and visualized using Exploratory Data Analysis (EDA) and Machine Learning techniques like KNN and Multiple Regressions. The maximum accuracy of 99% has been obtained in Multiple Regression.**

***Keywords*—Analytics; Education; Learning Management System; Performance; Students;**

JEET Category - Practice

I. INTRODUCTION

Education is a process that deals with knowledge transmission, and the development of skills and character. For scholars to discriminate between education and indoctrination, critical thinking plays an important role. Education helps to improve the student's cognitive and psychomotor skills. The efficiency of education can be analyzed by student's performance. Student performance is affected by factors such as the learning ability of students, interest, gained skills, parental background, peer influence, teachers' quality, and learning infrastructure. To overcome these challenges, we need a system that tracks student's performance. Several analytical techniques were used to enhance the performance. Multiple machine learning algorithms are used to evaluate the most accurate model. This work aims to provide a system to predict the performance of the student by using visualization. This system would help the institutions to improve the student's skills. Also, visualization helps in finding the variation of the class. The proposed system offers student performance prediction. It also gives various insights into the dataset. The dataset used in this work holds the marks secured by students from our college. The dataset is collected from around 400 students. The primary objective of the work is to visualize students' ability and predict their performance. It is helpful to analyze the impact of student's affective, cognitive, and psychomotor skills due to the switch from traditional classrooms to full-time online classes in educational institutions. The work focuses on the Sustainable Development Goal 4 - Education goal. It aims to "ensure inclusive and equitable quality education and promote lifelong learning opportunities for all."

This paper was submitted for review on August 31, 2023 . It was accepted on November, 15, 2023.
Corresponding author: R. Parkavi, Department of Information Technology, Thiagarajar College of Engineering, Tamil Nadu, India.
e-mail: rpit@tce.edu

II. RELATED WORKS

Analysis was conducted by Pandey M et al (2013) regarding the forecasting of students' academic performance in higher education. It appears to be a crucial prerequisite for raising educational standards. The study's findings will aid in identifying the weak students so that the institution's management may take the necessary measures. Angeline D. Magdalene Delight (2013) worked on the Apriori technique which is used to classify the students by extracting the set of rules unique to each class and analyzing the supplied data. Based on things like their participation in assignments, performance on internal assessments, attendance, etc., students are divided into groups. Based on the pattern gleaned from the database, it aids in predicting the student's performance.

Numerous analysis methods that can evaluate student performance, student growth, and student potential in various ways have been developed by FanYang et al. (2018). To examine student performance, identify factors that can affect it, and determine if pupils can do better, more analysis tools than those currently available are needed. The previous works are defeated by this work. Evaluation outcomes are displayed to demonstrate that the suggested instruments can produce accurate results.

McDonald, M. et al (2004) conducted a study that focused on the authors' experiences instructing upper-division computer science students. As a result of technology making the delivery of such courses possible, there are now a lot more courses available online. It presents the findings of a study contrasting online learners' achievement with that of conventional learners.

Edgcomb Alex Daniel et al (2015) examined the possibility of raising student grades by integrating an interactive textbook into introductory engineering classes. Four introductory computer programming courses were examined; they were offered at three research universities and were in C++, C/C++, and MATLAB, respectively. The study compared the final course grades and performance on certain course assignments, such as tests and projects, between the two programs. As a result, interactive textbooks considerably raised students' course marks, including those for exams and projects, especially for the least prepared students. Edgerton H. A (1940) aims to explain how R. A. Fisher's contributions are applied in the design of educational research in a logical and non-mathematical manner. Basic ideas in sampling theory, using the X2 distribution to test hypotheses, and small sample error theory are the topics covered in each chapter.

K-Means clustering was applied by Ratnapala I.P et al (2014) to assess and categorize the access behavior of a group of 412 students using an e-learning system. More than 40% of the student group is a passive online learner, according to the results, in both graded and ungraded learning contexts. Devasia, T et al (2016) proposed a system that uses the Naive Bayesian mining. Around 700 students with 19 qualities participated in the study from Amrita Vishwa Vidyapeetham in Mysuru. Using Naive Bayesian, the system tries to improve students' success graphs while also keeping track of their admissions information, course information, subject information, marks information, attendance information, etc. It uses the academic background of the student as input and forecasts the student's upcoming performances by semester.

Kapur, B. et al. (2019) worked to examine and compare effective algorithms. Student's potential is based on a variety of factors, including their backgrounds and past results. All these are determined to conclude which approach is more effective in accomplishing the objective. So, regardless of whether they are public or private, educational institutions can create their curriculum and methods of teaching depending on what is most effective. Zhiwu Liu et al (2010) developed a classification rule and an analysis-forecasting model for student grades. It explains how the outcome might be utilized to determine the factors that might have an impact on students' grades. Additionally, the administration of the classroom can be helped, and the teacher's teaching impact can be evaluated.

Data were compiled from the eighth and ninth-grade records of 214 students who were in the ninth grade during one of the years between 1931–1932 and 1935–1936 for Clifton, L. L. (1940) analysis. The same data was gathered for a group of 57 students who studied algebra the following year after the five-year period had ended. This was done to evaluate the effectiveness of a table of "chances of success" and a regression equation. Performance prediction was the subject of analysis by Kabra, R.R et al (2011). To create the model, decision tree algorithms are applied to the historical performance data of engineering students. The teacher will be able to give the right inputs by being able to predict which kids are most likely to fail.

Anuradha, C. et al (2015) conducted a comparison study. Data mining has a wide range of uses in the subject of education. The findings of the students' final-year performances are examined in this study using data mining. The major objective of this study is to forecast students' performance on semester-ending university exams using categorization algorithms. A group of characteristics for students majoring in computer science in several undergraduate colleges in Kolkata was introduced by Acharya, A. et al (2014). This data collection is subjected to five kinds of machine learning algorithms (MLAs). The decision tree class of algorithms produced the best results.

Parkavi R et al (2017) discuss the role of mobile phones in classrooms, With this another study is conducted to measure the performance of students with online courses. Sheik Abdullah A et al. (2018) used a statistical approach to analyze the learning outcome of students with real-time datasets and concluded that incorporating active learning strategies improves student learning ability.

Sahlaoui H et al. (2021) proposes a streamlined and comprehensible student performance prediction approach, enhancing accuracy and providing insights into score determinants. Employing machine learning models, this approach surpasses prior models on the same dataset, achieving an impressive 98% accuracy, a 20.3% improvement. Utilizing Synthetic Minority Over-sampling TEchnique (SMOTE) for data upsampling and ensemble methods for balance, optimize parameters through a scikit grid search algorithm and validate with ten-fold cross-validation. Introducing a novel visualization method using SHapley Additive exPlanations (SHAP) values enhances interpretability, aiding educators in identifying at-risk students early for timely intervention. The combination of predictive accuracy and interpretability makes our model reliable for educational decision-makers.

Srivastava N.K et al. (2023) leveraging a machine learning platform and Python, EDA delves into a dataset encompassing students' demographic features, including social and academic attributes. The EDA employs various analyses to unveil insights, paving the way for predicting students' performance. By scrutinizing the dataset, aim to discern the impact of different demographic features on student performance, shedding light on attributes that influence outcomes positively and negatively. This comprehensive analysis contributes to a nuanced understanding of the factors shaping students' academic

success.

Yagci M (2022) introduces an innovative educational data mining approach, employing machine learning algorithms to predict undergraduate students' final exam grades based on midterm exam results. The study evaluates the performance of various algorithms, including random forests, nearest neighbor, support vector machines, logistic regression, naive bayes, and k-nearest neighbor, using a dataset of 1854 students from a Turkish university. The proposed model, utilizing only midterm exam grades, department, and faculty data, achieves a notable classification accuracy of 70–75%. This data-driven analysis is crucial for developing a learning analysis framework in higher education and enhancing decision-making processes. The study also contributes to early identification of high-risk students and identifies the most effective machine learning methods for prediction.

In the competitive landscape of today, accurately forecasting student performance is vital for institutions aiming to identify talents, guide students strategically, and enhance academic outcomes. Early advice based on performance patterns helps mitigate failure rates, focusing students on specific areas for improvement. This study by Pallathadka H et al. (2023) utilizes data mining techniques to predict course performance by analyzing students' prior achievements in similar courses. The experimental study employs the UCI machinery student performance dataset, evaluating algorithms based on parameters such as accuracy and error rate. This approach aids institutions in informed decision-making and personalized guidance for students.

We discovered gaps between ongoing investigations when examining the aforementioned studies. A review of the literature reveals that different researchers have addressed various ways of improving the student's learning ability and predicting the performance of students but the gender based study on the performance evaluation has been ignored. The proposed methodology aims at evaluating the performance of students based on gender and race and predicting the performance using KNN and multiple regression algorithms.

III. PROPOSED METHODOLOGY

The dataset used in this work consists of attributes such as Gender, color, parental education, lunch, test preparation, math, reading, and writing scores are a few of the columns of the dataset. The attributes of math score and writing score are related to the cognitive domain whereas reading score is related to the affective domain. On the other hand, the psychomotor domain deals with the demonstration or description of any physical examination. It is shown in Figure 1. The mark details of each student have been collected directly from the faculty. The dataset uploaded is visualized using various graphs and the same has been studied and analyzed. The gender-based learning and preparation-based classification are also visualized. Thus, this analysis helps to find the impact of education on affective, cognitive psychomotor skills. The marks of the students are predicted using a multiple regression model since it provides maximum accuracy than the KNN algorithm.

Analysis of the data has been done using Google Colab. The system is developed using various packages like Pickle, Joblib, Seaborn, Matplotlib, Pandas, and NumPy. A web-based front end has been developed using Streamlit to provide a user-friendly interface for all kinds of users to carry out their respective actions. The prediction module and the visualization module have been integrated into it for ease of use. The overall methodology is given in Figure 2.

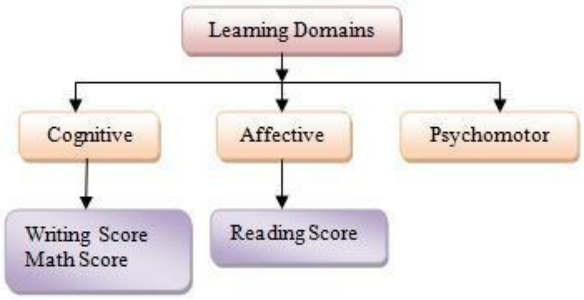

Fig.1. Different types of learning domains

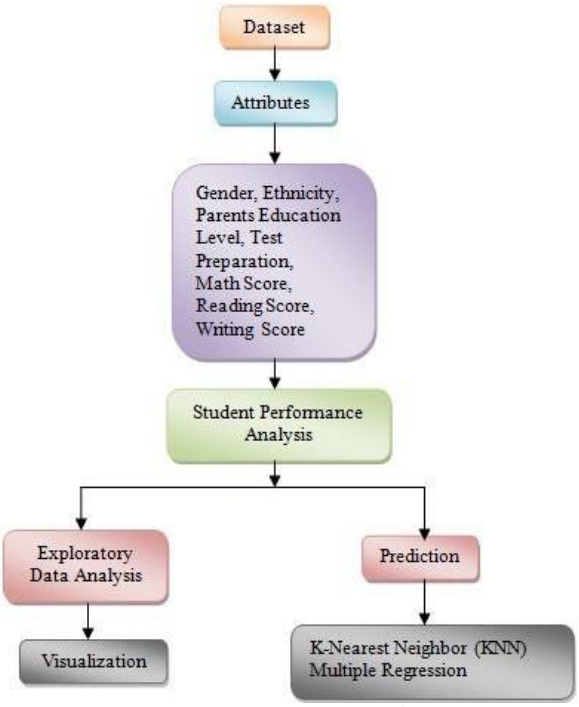

Fig 2. Proposed Methodology Overflow

IV. EXPERIMENT RESULTS AND DISCUSSIONS

The data has been gathered from the 400 engineering students of our college for analysis. In addition to the results of three examinations, this data collection also contains a range of interacting personal, social, and economic variables. All the rows of records are taken into consideration. There are

no missing or duplicate records, as determined by the completion of the initial processing stage. Bar plots, histograms, and scatter plots, among other types of charts, have all been used in the visualization. Both model generation and prediction involve machine learning algorithms, such as Multiple Regression and KNN.

Figure 3 explains the count of males and females in each group. The blue color represents female and orange represents male. From the figure, it is clear that Group C has the maximum number of male and female students.

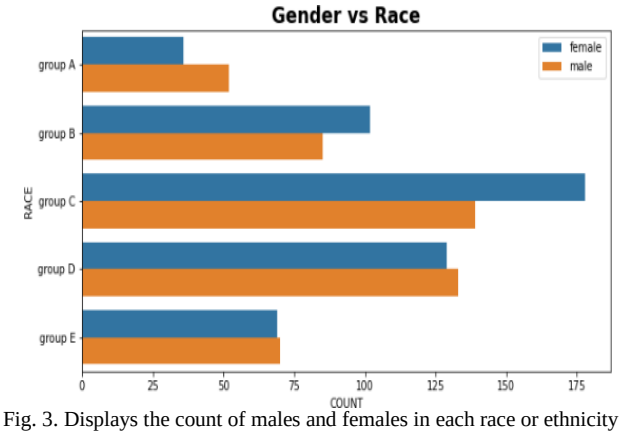

Fig. 3. Displays the count of males and females in each race or ethnicity from the dataset explored

Figure 4 explains the distribution of total scores across male and female students. The blue represents female and the orange represents male. The figure shows the maximum male has obtained less than 200 marks and the maximum female has obtained a score above 200. From the figure it is clear that the girls are the one with good marks and boys are the one with low marks also the number of girl students are more in scoring high marks compared to boys.

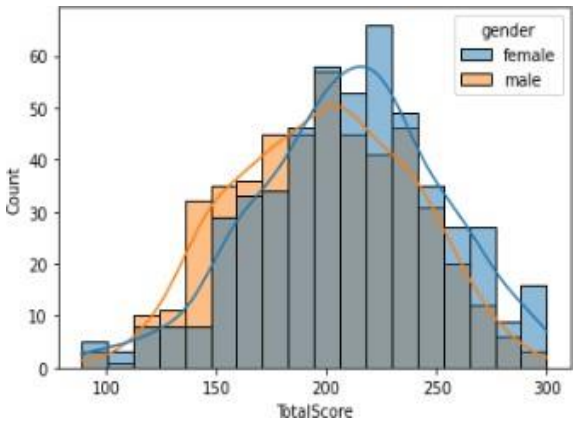

Fig. 4. Displays the total count of students who scored maximum marks. Also differentiates gender based on color.

Figure 5 shows the scatter plot of marks obtained by males and females in Math. The blue represents females and the red represents males. From the figure, it is clear that most boys are brilliant in mathematics. This proves males are good at cognitive skills. Figure 6 shows the scatter plot of marks obtained by males and females in Reading. The blue represents females and the red represents males. From the figure, it is clear that most girls are brilliant in Reading. This proves those females are good in psychological skills.

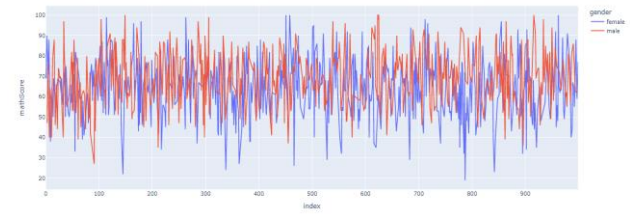

Fig. 5. Displays the variation in math scores of males and females. Proves that most boys are good in maths when compared to girls.

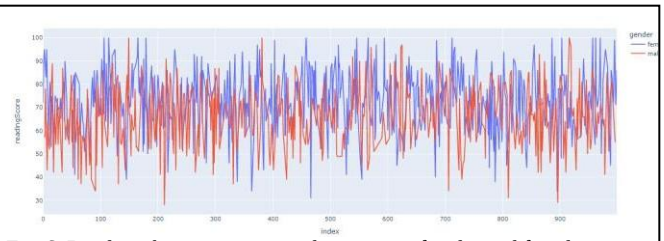

Fig. 6. Displays the variation in reading scores of males and females. Proves that most girls are good in reading when compared to boys.

Figure 7 shows the scatter plot of total marks obtained by males and females in Reading, Writing, and Mathematics. The blue represents females and the red represents males. From the figure, it is clear that most girls are brilliant in all skills. This proves that female is good at cognitive psychological skills.

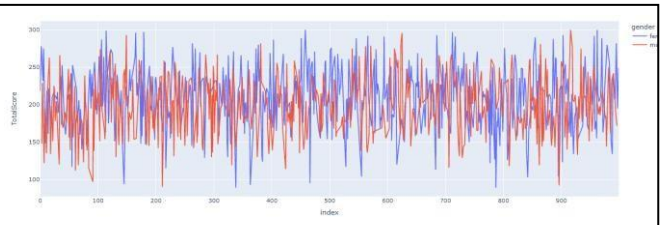

Fig. 7. Displays the variation in Overall score of male and female. Proves that most girls are good in learning and scoring. Overall performance is good for girls than boys

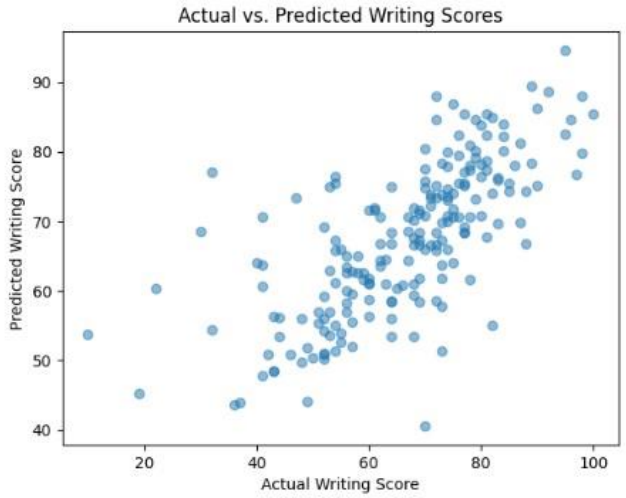

Fig. 8. Displays the variation between actual and predicted values based on the KNN algorithm.

Figure 8 shows the plot that predicts the variation in actual and predicted values. From the figure, it is evident that the analysis has obtained 48% accuracy with KNN. Figure 9 shows the plot that predicts the variation in actual and predicted values. From the figure, it is evident that the analysis has obtained 99% accuracy.

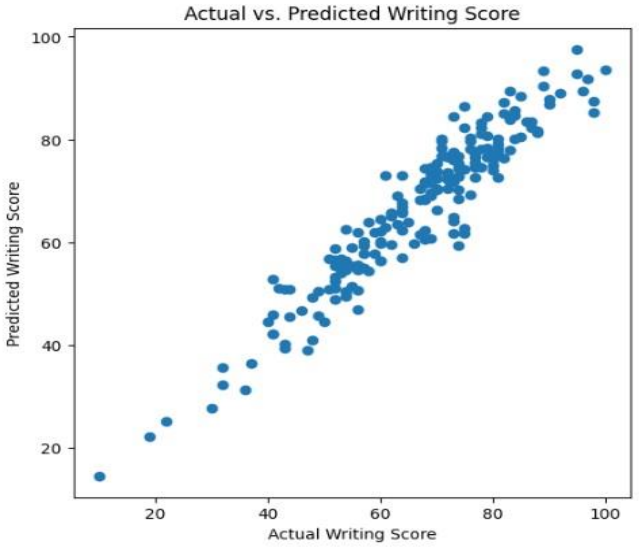

Fig. 9. Displays the variation between actual and predicted values based on the Multiple regression algorithm.

V. CONCLUSION

This work deals with the analysis of student records and visualizing the details. The reading, writing, and math scores are considered for the analysis of cognitive psychomotor skills. Multiple regression is the best suited method for this dataset as multiple attributes values are compared with a single value to estimate a result between the predicted and actual values of the single value based on the multiple values. KNN might not be capturing the underlying patterns as effectively in this case, so it is not as effective. The total score of the student is predicted using multiple regressions. Our future work is to expand our system based on Deep learning concepts. This could improve the accuracy of the results. Also, we have thought of analyzing our real-time dataset with an existing dataset for comparison studies.

REFERENCES

Acharya, A., & Sinha, D. (2014). Early Prediction of Student's Performance Using Machine Learning Techniques. International Journal of Computer Applications, 107(1), 37– 43. https://doi.org/10.5120/18717-9939

Angeline, D. M. D. (2013). Association Rule Generation for Student Performance Analysis using Apriori Algorithm. The SIJ Transactions on Computer Science Engineering & Its Applications (CSEA), 01(01), 16–20. https://doi.org/10.9756/sijcsea/v1i1/01010252

Anuradha, C., & Velmurugan, T. (2015). A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Students Performance. Indian Journal of Science and Technology,8(15). https://doi.org/10.17485/ijst/2015/v8i15/74555

Clifton, L. L. (1940). Prediction of High School Marks in Elementary Algebra. The Journal of Experimental Education, 8(4),410–413. https://doi.org/10.1080/00220973.1940.11010179

Devasia, T., Vinushree TP, & Hegde, V. (2016). Prediction of student's performance using Educational Data Mining. 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE). https://doi.org/10.1109/sapience.2016.7684167

Edgcomb, A., Vahid, F., Lysecky, R., Knoesen, A., Amirtharajah, R., & Dorf, M. L. (n.d.). Student Performance Improvement using Interactive Textbooks: A Three-University Cross-Semester Analysis. 2015 ASEE Annual Conference and Exposition Proceedings. https://doi.org/10.18260/p.24760

Edgerton, H. A., & Lindquist, E. F. (1940). Statistical Analysis in Educational Research. Journal of the American Statistical Association, 35(212), 716. https://doi.org/10.2307/2279266

Kapur, B., Ahluwalia, N., & Sathyaraj, R. (2017). Comparative Study on Marks Prediction using Data Mining and Classification Algorithms. International Journal of Advanced Research in Computer Science, 8, 632-636.

Liu, Zhiwu & Zhang, Xiuzhi. (2010). Prediction and Analysis for Students' Marks Based on Decision Tree Algorithm. 10.1109/ICINIS.2010.59.

McDonald, M., Dorn, B., & McDonald, G. (2004). A statistical analysis of student performance in online computer science courses. Proceedings of the 35th SIGCSE Technical Symposium on Computer Science Education - SIGCSE ‟04. https://doi.org/10.1145/971300.971327

Pandey, M., & Kumar Sharma, V. (2013). A Decision Tree Algorithm Pertaining to the Student Performance Analysis and Prediction. International Journal of Computer Applications, 61(13), 1–5. https://doi.org/10.5120/9985-4822

Pallathadka, H., Wenda, A., Ramirez-Asís, E., Asis-Lopez, M., Flores-Albornoz, J., & Phasinam, K. (2023). Classification and prediction of student performance data using various machine learning algorithms. Materials Today: Proceedings, 80, 3782–3785. https://doi.org/10.1016/j.matpr.2021.07.382

Parkavi, R., Sheik Abdullah, A., Sujitha, S., Karthikeyan, P., (2017), „Mobile Devices in Class Room‟, Handbook of Research on Mobile Devices and Smart Gadgets in K-12 Education, IGI Global Book Publishers.

Parkavi, R, & Karthikeyan, P., & Sheik Abdullah, A. (2018) Impact of Massive Open Online Courses and Best

Practices: A Case Study on Social Network Analysis Course, Journal of Engineering Education Transformations, Vol.31, No.3, DOI: 10.16920/jeet/2018/v31i3/120782.

Ratnapala, I. P., Ragel, R. G., & Deegalla, S. (2014). Students behavioral analysis in an online learning environment using data mining. 7th International Conference on Information and Automation for Sustainability. https://doi.org/10.1109/iciafs.2014.7069609

Sahlaoui, H., Alaoui, E. A. A., Nayyar, A., Agoujil, S., & Jaber, M. M. (2021). Predicting and Interpreting Student Performance Using Ensemble Models and Shapley Additive Explanations. IEEE Access, 9, 152688–152703. https://doi.org/10.1109/access.2021.3124270

Sheik Abdullah, A., Gayathri, N, Abirami, AM Parkavi, R, (2018) A Statistical Approach for setting SLO Targets over Outcome Based Education-A Case Study, Journal of Engineering Journal of Engineering Education Transformations, Volume 34, January 2021, Special issue, eISSN 2394-1707 85 Education Transformations, DOI: 10.16920/jeet/2018/v0i0/120898

Srivastava, N. K., Pandey, P., & Mishra, V. (2023). Exploratory Data Analysis (EDA) based on demographical features for students' performance prediction. Artificial Intelligence, Blockchain, Computing and Security Volume 1, 126–133. https://doi.org/10.1201/9781003393580-19

Yagci, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. Smart Learning Environments, 9(1). https://doi.org/10.1186/s40561-022-00192-z

Yang, F., & Li, F. W. B. (2018). Study on student performance estimation, student progress analysis, and student potential prediction based on data mining. Computers & Education, 123, 97–108. https://doi.org/10.1016/j.compedu.2018.04.006