2024 International Conference on Integrated Intelligence and Communication Systems (ICIICS)

Unveiling Student Success: A Multifaceted Approach with Learning Coefficients and Beyond

Nukala VijayaKumar
dept. computer science
Narasaraopeta Engineering College (of Jntuk)
Narasaraopeta, India
nvk20022001@gmail.com

Shaik Rafi
dept. computer science
Narasaraopeta Engineering College (of Jntuk)
Narasaraopeta, India
shaikrafinrt@gmail.com

Tumpala Mahith
dept. computer science
Narasaraopeta Engineering College (of Jntuk)
Narasaraopeta, India
mahitumpala65@gmail.com

D.Venkata Reddy
dept. computer science
Narasaraopeta Engineering College (of Jntuk)
Narasaraopeta, India
doddavenkatareddy@gmail.com

Kethavath Ravi naik
dept. computer science
Narasaraopeta Engineering College (of Jntuk)
Narasaraopeta, India
rn3154773@gmail.com

Kolakani Raju
dept. computer science
Narasaraopeta Engineering College (of Jntuk)
Narasaraopeta, India
rajukolakani529@gmail.com

***Abstract*—The student performance is examined in this study using a number of methods of Educational Data Mining (EDM), Clustering and classification techniques are employed to classify the course as well as the performance in the entrance examination. The results obtained show that the Random Forest and XG Boost which are machine learning models outperform traditional methods for predicting student success. Moreover, CNN and LSTM Networks, which are deep learning models, improve prediction accuracy even further. Conducted through metrics like accuracy, precision, recall and F1-score, this study shows that any form of recognition of the pattern, in this case, the early one, helps to reduce failure rates to considerable extents. The results of this study suggest that there is a potential scope for further improving prediction algorithms and management of educational resources, which are of great relevance to the institutions to further the student success.**

Keywords—Early detection, data mining, academic performance, non-academic performance student performance prediction, graph mining Introduction

I. INTRODUCTION

Education, which is a necessary societal element, is heavily influential on various aspects of life. The integration of information and communication technologies has changed many fields of study including education. COVID-19 pandemic for instance has forced several countries to adopt e-Learning environments much faster than they would have done in normal circumstances [1][2][3]. Higher education institutions consider the academic achievement of students as one of the main indicators for the quality of service they offer. However, it might be difficult to identify what are the key elements that influence a student's performance during their early years at school. In response to issues concerning academic performance, many useful instruments have been developed but these tools are often not applicable elsewhere in education contexts [3]. Despite advances in predicting students' outcomes, there are still gaps in data-based analysis and augmentation of student results using technology across all areas. Consequently, EDM has the potential to improve education institutions whole student experience as well as teaching/learning [1]. Because academic achievement is strongly correlated with desired outcomes, it is extremely important. Academic achievement among college or university students still remains a critical indicator of institutional success. Student's academic performance can be analyzed and predicted using variables such as basic courses and non-academic performance and name of university. This study uses consequences from assignments, final examinations and primary-degree route scores [3]. This research provides a singular software of t-SNE dimensionality discount to front scores, first-level path ratings, AAT, and GAT. To the best of the researcher's expertise, that is the primary attempt to utilize machine gaining knowledge are expecting early scholar overall performance the use of attributes from each admission scores and first-degree path rankings. The researchers investigate a novel method to raising the relocation threshold that entails calculating absolutely the difference among grades previous to and following a specific point. This research uses the most recent categorization models to assess how well the researcher advised technique works. The layout of this paper is as follows: The literature on techniques for predicting student overall performance in the school room is reviewed in Section II. The dataset, together with statistics categorization and correlations, is defined extensive in Section III. Section IV affords the technique used within the observe. The advised method is classed, examined, and outcomes are pronounced in Section V. The have a look at is finally concluded in Section VI.

II. RELATED WORK

Presented a deep neural network (DNN) based binary classification framework, highlighting the important thing characteristics affecting the outcome. They assessed the framework with two awesome optimizers and activation functions. According to the experimental effects, prediction accuracies of ninety-three. 43% and 94. 48%, respectively, had been attained with the Ad delta and Adara optimizers, yielding overall getting to know overall performance rankings of ninety. 65% and 99%. Used plenty of device mastering algorithms to forecast students' remaining grade averages primarily based on more than a few of factors, consisting of their first-year performance, private traits, university entrance examination effects, and gap year. The researchers also employed the CNN and LSTM fashions, which yielded 94% and 91% accuracy, respectively [4][5][6][12]. The college's student control records gadget and a ballot of graduates from 3 separate years provided the dataset that used. Their

979-8-3315-0496-0/24/$31.00 ©2024 IEEE

outcomes confirmed a dating among some of variables and the instructional achievement of college students of their 2nd year of examine. The look at also used CNN and LSTM models, which produced accuracy rates of 94% and 91%, respectively [4][5]. in order to demonstrate various contemporary methods for predicting student performance. Their study concentrated on existing techniques for forecasting student conduct in classroom settings. They came to the conclusion that because supervised learning produces accurate and consistent findings, there is a noticeable trend toward utilizing it to forecast university student success. On the other hand, because unsupervised learning predicts student behavior less accurately in the situations under study, academics have not found it as appealing [6][8]. They integrated fuzzy set rules, Lasso linear regression, and collaborative filtering—three dynamically weighted approaches. They emphasized the necessity of expanding their methodology and verifying the model's dependability with authentic scholarly datasets. Sought to estimate applicants' academic potential prior to admission in order to help better education establishments make nicely knowledgeable admissions decisions [10]. Finding the characteristics that set apart academically struggling students from high achievers was the aim. To forecast a student placement in the Information Technology industry based on their academic and non-academic performance in classes 10th, 12th, graduation, and the number of backlogs during graduation built supervised machine learning classifiers [11]. A method for predicting student grades using grades from Portuguese language courses and mathematics. They used a deep learning model, which requires additional validation on bigger and more evenly distributed datasets as it was only verified on two datasets. This machine learning method is used for ranking, classification, and regression applications. It is a member of the Boosting method family. A novel approach called the Multi-Agent System (MAS), which incorporates an Agent based totally Modelling Feature Selection (ABMFS) version. This model efficiently eliminates features that are not relevant from the prediction effects. They then built a Convolutional Neural Network (CNN) shape to predict pupil overall performance the use of deep learning strategies [3][5][6]. presented a way to use base classifiers in both homogeneous and heterogeneous, as well as selecting and ranking systems, to improve the performance of many single classification algorithms. To find the ideal algorithm parameters and configuration, model needs to be Refinement using Refinement techniques. Seven widely used group fairness criteria were evaluated in order to forecast problems with student performance [3][5]. They used two fairness-aware machine learning algorithms and four conventional machine learning models to conduct tests on five educational datasets. Nevertheless, their research was restricted to public schools and did not take into account academics at the university level. A model to forecast the final test grades of undergraduate students. Based on statistics from 5 languages college students at a Turkish public institution, the model takes into consideration branch, faculty, and midterm examination results. Artificial Neural Networks (ANN) and Random Forests (RF) outperformed preceding category fashions, correctly classifying very last exam grades with region below the curves (AUC) of 99% and 93%, respectively [7].

III. DATASET DESCRIPTION

The" Student Academic Performance" dataset in table 1 gives an intensive summary of all the variables influencing students' educational success. Academic performance signs comprise choice indices for post-secondary training establishments as well as rankings in more than a few subjects, together with well-known, FEP, and English, alongside percentile, decile, and quartile rankings. The dataset provides insights into the effect of own family records and socioeconomic role on academic outcomes and enables thorough take a look at and prediction of student overall performance.

TABLE I. DATASET DESCRIPTION

Variable	Description
COD S11	Identification code for each student
GENDER	Gender of the student
EDU FATHER	Education level of the father
EDU MOTHER	Education level of the mother
PEOPLE HOUSE	Number of people living in the household
MAT S11	Mathematics score for S11
CR S11	Critical reading score for S11
CC S11	Civic and citizenship education score for S11
BIO S11	Biology score for S11
ENG S11	English score for S11
Cod SPro	Program code for higher education
UNIVERSITY	Name of the university attended
ACADEMIC PROGRAM	Name of the academic program
QR PRO	Quantitative reasoning score for higher education
CR PRO	Critical reading score for higher education
CC PRO	Civic and citizenship education score for higher education
ENG PRO	English score for higher education
WC PRO	Writing communication score for higher education
FEP PRO	Final exam performance score for higher education
G SC	General score
PERCENTILE	Percentile rank
2ND DECILE	Second decile rank
QUARTILE	Quartile rank
SEL	Selection status
SEL IHE	Selection status in higher education institutions

IV. RECOMMENDED METHODOLOGY

A. Data Arrangement

At this stage, the key variables affecting the performance of the students by using a sample of their academic and non-academic records from a computer science institution. The goal is to anticipate students' achievement in the final examination (GPA) by presenting the data of their records and features at an early stage. It additionally offers some of overall performance signs for a number disciplines, including G-SC, CC-PRO, ENG-PRO, WC-PRO, and FEPPRO, which stand for rankings or talent levels in numerous educational domains. By analyzing these sizeable facts sets, the

researchers are hoping to discover early signs a good way to lead to kids' successful educational careers [3].

B. Used features

To forecast and enhance student performance, researchers make use of a number of features. The main characteristics include gender, Basic Tests, and results from all initial level scores. Researchers also take into account sociodemographic variables such the parents' educational backgrounds and jobs, the socioeconomic class, and the SISBEN categorization. Every attribute applied in the analysis is listed. It is found that those traits have a big effect on how well university college students do academically. Using this massive collection of statistics, researchers are hoping to beautify the precision of forecasts and create plans that inspire instructional fulfilment for university college students.

C. Preprocessing

Data cleaning: authors address missing values by way of either deleting rows or columns that have too much lacking facts or imputing them the usage of the applicable records (mean, median, and mode). To ensure the excellent of the statistics, errors or inconsistencies inside the fact's entries might be constant. Normalization and Scaling: To maintain consistency and beautify version overall performance, numerical functions like admission rankings and direction overall performance measures may be both normalized and scaled. authors use methods like Z-rating normalization and Min-Max scaling. Handling Class Imbalance: Methods like resampling (oversampling/beneath sampling) or using algorithms advanced to handle class imbalance can be used if the dataset carries an unbalanced class distribution (extra excessive-acting students than low-appearing ones, for instance). Dataset Splitting: In order to evaluate version performance, the dataset will be divided into training and trying out sets. Depending on the dataset, an 80/20 or 70/30 cut up ratio is traditional.

D. Stochastic neighbor mapper

T-distributed Stochastic Neighbor Embedding is a powerful nonlinear dimension reduction technique to visualize high dimensional data. The researchers are using the t-SNE algorithm here to find out how GAT and AAT relate with and affect student grades in terms of GPA. Researchers can use t-SNE to decrease the overall complexity that authors get in relation to visualizing the data which has many dimensions by simply representing it in 2D or 3D space.[3] fig 1 and fig 2 represents the dimensionality reduction using t-sne. It is crucial for educators and researchers to understand these associations as they seek to assess and improve learner performance [15].

E. Learning algorithms

In this section the dataset is train and test with the machine learning algorithms. This includes the most commonly used algorithms such as extreme boost, Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Random Forest (RF)[2][3][4][5][6][7][8][9][10]. Xgboost makes the image more accurate and helps stop overfitting. It does this by making an objective function better. The researchers compare these ways of grouping things to see how well authors can watch student performance. What authors found shows that supervised machine learning methods work much better than models that use old-school features.

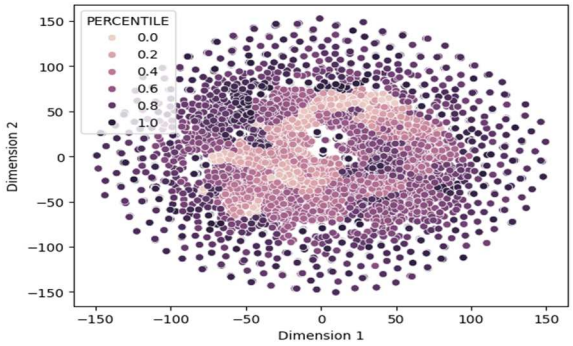

Fig. 1. Dimensionality reduction using T-sne

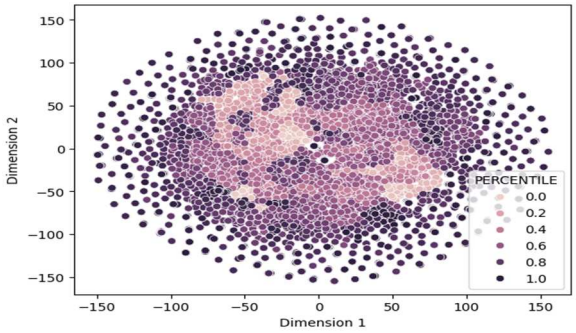

Fig.2.Dimensionality reduction using T-sne

F. Deep learning algorithms

In addition to these machine learning algorithms, The researchers additionally utilize deep mastering algorithms along with Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNN). LSTM networks are exceptionally powerful in managing time-collection facts and sequences because of their ability to maintain lengthy-time period dependencies, making them perfect for tasks involving temporal styles. CNNs, on the other hand, are particularly effective in spotting patterns and structures in information, particularly while managing grid-like information together with images or, on this context, based academic statistics [2][5].

- The dataset contains information on academic performance and various socio-demographic factors of 12,411 individuals.
- Strong correlations between PERCENTILE, 2NDDECILE, and QUARTILE are seen in the heatmap, demonstrating their resemblance. G-SC and CR-PRO likewise have a strong positive correlation. Comparatively speaking, QR-PRO exhibits lesser correlations with other variables, indicating that it assesses a different facet of performance than the others.
- The above table represents the features of the dataset which was used for research

G. Evaluation Framework

The researchers conduct experiments using varying sets of features, including admission scores alone, basic tests combined with gender, and mid-term exams along with all basic-level scores. The purpose of these analysis is to demonstrate the importance of the introduced features in achieving higher accuracy, thereby proving that the

improvements are not coincidental. This underscores the important function these capabilities play in reaching good accuracy in getting good results. Authors increase the model to expect pupil performance. The version is initially educated the usage of suitable samples extracted from the dataset. Specifically, researchers accumulate scholar data, focusing on admission scores, gender, and all first-degree required path rankings. These functions, as previously mentioned, are used to put together the training samples. The equal system is carried out while predicting the overall performance of recent, unknown samples. This consistency ensures that the model appropriately reflects the have an effect on of these features on student performance. Fig 3 represents the flowchart of student performance prediction.

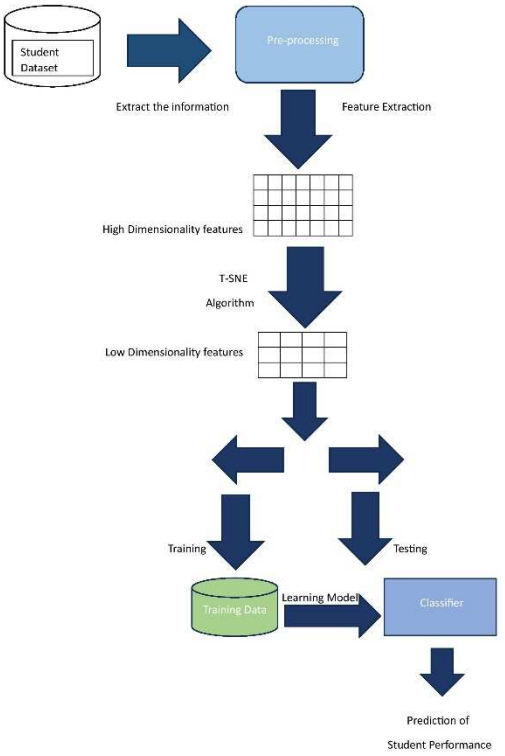

Fig.3. work flow of student performance prediction

H. Admission Trails

SCORES FEATURES: Additionally, the researchers incorporated deep mastering fashions into the experiment: Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. The inclusion of those models aimed to explore their effectiveness in contrast to traditional classifiers. the accuracy of the models the use of best admission rating capabilities outperformed those using a mixture of admission rankings and gender features. This result underscores the efficacy of the admission rating capabilities in reaching better prediction accuracy, demonstrating that the inclusion of deep gaining knowledge of models in addition complements performance.

I. Performance modelling after rounding results

In this research, an in-depth performance evaluation was conducted using key metrics such as precision, recall, F1 score, and accuracy. These metrics were calculated across several tests to ensure a comprehensive evaluation. Precision measures the proportion of correct positive predictions among all positive outcomes predicted by the model. For example, in the CR-S11 course, high precision indicates that most positive predictions are accurate, which is critical for courses requiring critical reasoning where false positives can be misleading. Recall, or sensitivity, measures the proportion of actual positives correctly identified by the model. This metric is particularly important in the BIO-S11 course, where identifying all relevant instances is crucial due to the nature of biological research. The F1 score, calculated as the harmonic mean of precision and recall, provides a balance between accuracy and recall, making it especially useful when class distribution is unbalanced. The F1 score ensures that both false positives and false negatives are considered, as in QRPRO, where quantitative reasoning is essential. Accuracy, the most intuitive performance measure, represents the percentage of correct predictions (both true positives and true negatives) among the total number of cases evaluated. The results of this evaluation are shown in Tables II, III, and IV.

TABLE II. PRECISION, RECALL, F1SCORE AND SUPPORT FOR SVM MODEL

S. NO	Precision	Recall	F1 Score	Support
1.	0.94	0.92	0.93	340
2.	0.91	0.88	0.89	664
3.	0.96	0.98	0.97	1479

TABLE III. PRECISION, RECALL, F1SCORE AND SUPPORT FOR KNN MODEL

S.NO	Precision	Recall	F1 Score	Support
1.	0.89	0.91	0.90	340
2.	0.84	0.80	0.82	664
3.	0.94	0.95	0.95	1479

TABLE IV. PRECISION, RECALL, F1SCORE AND SUPPORT FOR RANDOM FOREST MODEL

S.NO	Precision	Recall	F1 Score	Support
1.	0.97	0.99	0.98	340
2.	1.00	0.98	0.99	664
3.	1.00	1.00	1.00	1479

J. Success Indicators

The insight breakdown of machine learning and deep learning models is important in understanding and enhancing their performance.[9] In this analysis, researchers adopt widely used evaluation metrics to evaluate the models. Researchers use these metrics: prediction accuracy and F1-score that give insight into model behavior when inputs vary. Generally, classification accuracy is the key criterion used to get the potency of machine learning and deep learning models. This

offers a simple way to determine how often the model's predictions match up with real outcomes. Additionally, confusion matrices are used to compare in more detail prediction accuracy and errors including correct classifications as well as misclassifications. Such a comprehensive approach helps ourselves make a strong evaluation of the classifiers hence identifying areas where researchers need improvement on hence refining them accordingly.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

$$\text{F1-Score} = \frac{2\times Precision\times Recall}{Precision+Recall} \quad (4)$$

Taken together these metrics represent a complete view of model quality by taking into account both overall accuracy and handling positive cases

K. Comparitive Analysis

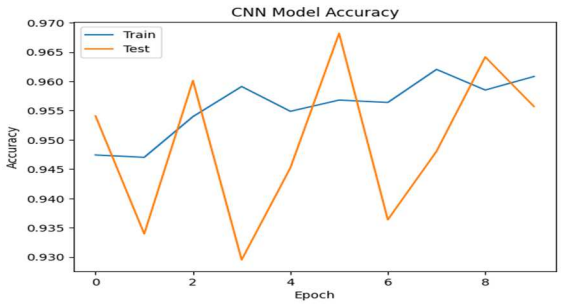

Fig. 4. Accuracy comparison between the train, test for CNN

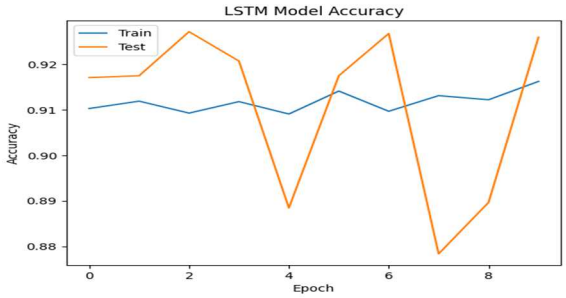

Fig. 5. Accuracy comparison between the train, test for LSTM

Methods like t-Distributed Stochastic Neighbor Embedding (t-SNE) help reduce dimensions allowing ourselves to see and assess high-dimensional data to spot patterns. The researcher test machine learning models such as Random Forest, Xgboost, SVM, and KNN as well as deep learning models like CNN and LSTM, to check how well they predict [1][2][3][4][5][6][7]. Researchers use measures like precision, recall, accuracy, and F1-score to do this. The results show that when Researchers mix advanced feature engineering with deep learning methods, these models get better at making predictions. This could play a key role in spotting college students at risk on and taking targeted steps to boost their academic results. The number of training samples used are 9928 and the number of test samples used are 2483.The below fig 4 and fig 5 shows the comparative analysis of CNN and LSTM models [13][14].

L. Conclusion

The document gives a complete analysis of predicting student academic performance using different data mining and machine learning methods. The study puts an emphasis on the need for detecting early any potential academic problems, which helps address such in a timely manner to improve results. Also, it helps to identify patterns and clusters that are closely related to students' performance trends. In addition of this study evaluates the effectiveness of various machine learning models and some deep learning modes including recently used XGBoost algorithm, K-Nearest Neighbors (KNN), Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression and deep learning models like CNN and LSTM in predicting student achievement. Moreover, to enhance prediction accuracy deep learning techniques like Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks have been employed. The findings indicate that advanced feature engineering when combined with deep learning approaches enhances the models' predictiveness over more traditional methods.

V. FUTURE SCOPE

The future scope of this work is to consider advanced feature engineering based on real-time student engagement data, emotion analysis, and socio-demographics for improved predictions. There is also a potential for building hybrid models that incorporate traditional machine learning and reinforcement or transformer techniques to improve performance. At-risk students can be guaranteed improvement in outcomes through personalized learning interventions based on real-time data. However, it will be necessary to verify the scalability of the models across multiple cases in educational practice and address the interpretability and ethical issues related to AI based predictions.

REFERENCES

[1] Roy, K., & Farid, D. M. (2024). An Adaptive Feature Selection Algorithm for Student Performance Prediction. IEEE Access.

[2] Qin, K., Xie, X., He, Q., & Deng, G. (2023). Early Warning of Student Performance With Integration of Subjective and Objective Elements. IEEE Access.

[3] Alhazmi, E., & Sheneamer, A. (2023). Early predicting of students performance in higher education. IEEE Access, 11, 27579-27589.

[4] Liu, D., Zhang, Y., Zhang, J. U. N., Li, Q., Zhang, C., & Yin, Y. U. (2020). Multiple features fusion attention mechanism enhanced deep knowledge tracing for student performance prediction. IEEE Access, 8, 194894-194903.

[5] Butt, N. A., Mahmood, Z., Shakeel, K., Alfarhood, S., Safran, M., & Ashraf, I. (2023). Performance Prediction of Students in Higher Education Using Multi-Model Ensemble Approach. IEEE Access, 11, 136091-136108.

[6] Bujang, S. D. A., Selamat, A., Ibrahim, R., Krejcar, O., Herrera-Viedma, E., Fujita, H., & Ghani, N. A. M. (2021). Multiclass prediction model for student grade prediction using machine learning. IEEE Access, 9, 95608-95621.

[7] Alshanqiti, A., & Namoun, A. (2020). Predicting student performance and its influential factors using hybrid regression and multi-label classification. IEEE Access, 8, 203827-203844.

[8] 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

[9] Pelima, L. R., Sukmana, Y., & Rosmansyah, Y. (2024). Predicting university student graduation using academic performance and machine learning: a systematic literature review. IEEE Access.

[10] Vives, L., Cabezas, I., Vives, J. C., Reyes, N. G., Aquino, J., Condor, J. ´ B., & Altamirano, S. F. S. (2024). Prediction of Students' Academic Performance in the Programming Fundamentals Course Using Long Short-Term Memory Neural Networks. IEEE Access.

[11] Alamri, R., & Alharbi, B. (2021). Explainable student performance prediction models: a systematic review. IEEE Access, 9, 33132-33143.

[12] Sahlaoui, H., 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.

[13] Yaakub, T. N. T., Ahmad, W. R. W., Husaini, Y., & Burham, N. (2018, November). Influence factors in academic performance among electronics engineering student: Geographic background, mathematics grade and psycographic characteristics. In 2018 IEEE 10th International Conference on Engineering Education (ICEED) (pp. 30-33). IEEE.

[14] Sagala, T. N., Permai, S. D., Gunawan, A. A. S., Barus, R. O., & Meriko, C. (2022, December). Predicting Computer Science Student's Performance using Logistic Regression. In 2022 5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI) (pp. 817-821). IEEE.

[15] Sa, C. L., Hossain, E. D., & bin Hossin, M. (2014, November). Student performance analysis system (SPAS). In The 5th International Conference on Information and Communication Technology for The Muslim World (ICT4M) (pp. 1-6). IEEE.