ENVISIONING EDUCATIONAL SUCCESS THROUGH ADVANCED ANALYTICS AND INTELLIGENT PERFORMANCE PREDICTION

Dr T.Sureshkumar [1]
Professor,
Department of Information technology
Nehru institute of technology
Coimbatore – 641 105
suresh_technology@ yahoo.co.in

Ms. Charanya J[2]
Assistant Professor,
Department of Computer Science and Engineering,
Kongu Engineering College (Autonomous),
Erode- 638 401.
charanjagan@gmail.com

Dr.T.Kumaresan[3]
Associate Professor,
Department of Computer Science and Engineering,
Bannari Amman Institute of Technology,
Sathyamangalam - 638 401.
kumaresant@bitsathy.ac.in

Dr Rajeshkumar G[4]
Associate Professor,
Dept of CSE
Bannari Amman Institute of Technology,
Sathyamangalam.
grajesh.grk@gmail.com

Kavin kumar P[5]
UG scholar,
Department of Computer Technology,
Bannari Amman Institute of Technology,
Sathyamangalam
kavinkumar.ct21@bitsathy.ac.in

Anuj B [6]
UG scholar,
Department of Computer Technology,
Bannari Amman Institute of Technology,
Sathyamangalam
anuj.ct21@bitsathy.ac.in

Abstract-The accurate prediction of students academic success holds significance in enhancing educational outcomes. The study focuses on overcoming the challenge of assessing whether a student will successfully pass the final exam, taking into account a range of socio-economic and academic factors. Every student's journey is different, and it is influenced by various factors such as their socioeconomic background, academic abilities, and personal experiences. To evaluate their performance, we use various machine learning algorithms such as KNN, Naive Bayes, SVM, AdaBoost, and Naive Bayes— and assess their performance using key metrics such as ROC curve, confusion matrix, precision, recall, and F1 score. The F1 score offers a balanced assessment of precision and recall, offering insights into overall model accuracy. Precision and recall focus specifically on the model's ability to correctly identify positive instances and capture all actual positive instances, respectively. ROC curves demonstrate the balance between the rates of correctly identifying true positives and incorrectly identifying false positives, allowing for a nuanced assessment of classifier performance. The confusion matrix further breaks down true positives, true negatives, false positives, and false negatives, enhancing our understanding of algorithmic effectiveness. Through a comprehensive comparison of these classification methods, our goal is to identify the most accurate approach for predicting student outcomes, thus contributing valuable insights to educational institutions and policymakers aiming to implement effective strategies for student success and academic improvement.

KEYWORDS: Academic Success, Socio-economic Factors, Machine Learning Algorithms,F1 Score, Educational Outcome

I. INTRODUCTION

In the ever-changing world of education, knowing how well students will do is crucial for their future success. By using smart analytics and prediction models, teachers can anticipate challenges and offer the right support to improve student outcomes. Being able to predict performance is like having a guide that helps educators spot areas for improvement and give timely help. This not only helps individual students grow academically but also makes educational systems work better overall. Predicting student performance lets schools use their resources wisely, ensuring that each student gets the support they need for a personalized learning experience. Since education is the foundation for personal and societal growth, predicting how students will do is vital for creating a supportive learning environment.

Our research aims to go beyond existing approaches, presenting a powerful predictive model that not only considers various influencing factors but also outperforms existing benchmarks. By leveraging the strengths of various machine learning algorithms and evaluation metrics, we seek to contribute valuable insights to the ongoing discussion about educational outcomes, fostering an environment conducive to academic success for every student.

Our study embarks on a comparative analysis of machine learning algorithms, including logistic regression, KNN, SVM, AdaBoost, and Naive Bayes. We aim to assess their performance using key metrics such as F1 score, precision, recall, ROC curve, and confusion matrix. These metrics provide a comprehensive evaluation of model accuracy, precision in positive instance identification, recall in capturing all actual positive instances, and the trade-

off between both true and false positive rates. Our overarching goal is to distill insights into the most accurate approach for predicting student outcomes, contributing to the knowledge base that can empower educational institutions and policymakers in implementing effective strategies for student success and academic improvement.

II. LITERATURE SURVEY

The importance of predictive modeling in education has been highlighted by many studies.Conijn et al. [1], Bujang et al. [2]carried out a thorough analysis of the literature on approaches for predicting student grades using unbalanced categorization. The study's objectives were to evaluate and compile a collection of research on unbalanced categorization in grading prediction.

Mengash's [3] emphasis on data mining for informed decision-making in university admissions, a related study seeks to enhance the precision of admission processes. Kusumawardani and Alfarozi[4] aimed to predict student performance by analyzing log activities, using a Transformer encoder model. The study focused on understanding patterns in students' interactions, employing a model known for its success in sequence-related tasks, although specific dataset details and performance metrics were not disclosed. Pek et al. [5] use various algorithms like Naïve Bayes, Random Forest, and others were tested, with the best-performing ones combined into a hybrid ensemble model. Using the Chinese National College Entrance Exam as a case study, Yao et al. [6] examined the impact of high school characteristics and student skills on student development. The goal of the study was to comprehend how different elements affect the development of students. Yao et al. [7] conducted a case study that investigated how student skills and high school experiences affect academic advancement. Behavioral analysis also plays a role, with Zollanvari et al. In evaluating our model, we focus on the F1 score, a balanced measure of recalland precision. Inspired by Liang et al. [8] and Pek et al. [9] looking into self-regulatory learning behaviors for predicting students' GPAs. The temporal dimension is not overlooked, Precision and recall, focusing on correctly identifying positive instances and capturing all actual positive instances, contribute to the granularity of our evaluation. ROC curves add an extra layer of insight, illustrating the trade-off between both trueand false positive rates and also Inspired by studies such as Zollanvari et al. [9] and Alshanqiti et al. [10],with Qin et al. [14] emphasizing the importance of early warning systems for student performance. Moreno-Marcos et al. [15]analyzed factors influencing learners' performance prediction using learning analytics, providing a holistic view of the determinants of academic success. Adnan et al. [11] aimed to predict at-risk students at various course percentages for early intervention, employing machine learning algorithms like Decision Trees and Neural Networks. Czibula et al. [12]created a system called IntelliDaM, which uses machine learning to improve decision-making in educational data mining. Using machine learning, the research attempted to enhance decision-making in educational settings.

Pereira et al.'s [13] main goal was to develop a black-box predictive model to interpret the behavior of both individual and group programming students.

Feng et al. [16] addressed the challenges of academic achievement of students using data mining for education, aligning with the broader goal of leveraging data-driven insights. The study focused on predicting academic performance through various data mining algorithms.

To predict university students' GPAs, Prabowo et al. [17] suggested combining historical and tabular data in a deep-learning model. Sun et al. [18] created a model with multi-feature fusion and attention mechanisms, showing the growing complexity of prediction methods. Ahmad et al. [19] aimed to address the challenges of predicting students' cognitive skills using a multilayer approach, echoing the broader focus on effective methods for assessing cognitive abilities. To tackle the problem of forecasting student outcomes based on cognitive preferences, P. Jiang and X. Wang [20] developed a preferred cognitive diagnosis for student performance prediction.The research focuses on understanding and utilizing cognitive diagnoses and preferences for accurate predictions, contributing to personalized educational strategies. In order to forecast student performance and identify important elements, Alshanqiti and Namoun[10] used a hybrid regression and classification by multiple labels technique. Sun et al. [18] developed a university predictive model for student performance using a multi-feature fusion and Attention Mechanism (MFAPM) model that offers an innovative approach to predicting university student performance, addressing data privacy concerns. The integration of subjective and objective information was proposed by Qin et al. [14] as a precursor for student performance.

III. SYSTEM WORKFLOW

The research starts by collecting a comprehensive dataset that includes information on the academic performance of secondary education students in two educational institutions in Portugal. The dataset is a large collection of data that allows us to gain insight into various aspects of the academic performance of students in these educational institutions. The dataset contains details of different students and the academic records of these students. It is an important starting point for the study as it provides the necessary information for the analysis and drawing of insights. The next step in the research is the data processing. Here, the non-numeric information is processed using numerical_data() to ensure that the dataset is suitable for machine learning. After that, feature scaling is used to ensure that all the different information types in the dataset are in the same format, which helps in the training of the machine learning models.

Next, we smoothly move into the data processing phase.Here, we use a unique function called numerical_data() to handle non-number data. This guarantees that machine learning will

effectively use our dataset. Next, we apply feature scaling to ensure that every kind of information in our dataset has a comparable format. This aids in the more efficient training of our machine-learning models.

In the Next stage where we select the appropriate algorithm for the task. Each of our five machine learning technologies has advantages and disadvantages of its own. These are similar to many specialists that we can consult for guidance. We have AdaBoost, Naive Bayes, KNN, Logistic Regression, and SVM with various information understanding methods (Linear, Polynomial, and Gaussian).

Examining the SVM tool in more detail, we investigate its many information-processing models or kernels. We have Gaussian, which is even more complex, Polynomial, which is a little more complicated, and Linear, which is straightforward. We evaluate each of these using critical metrics such as the ROC curve, confusion matrix, and F1 score.

Once we've figured out the best way for SVM to understand our data, we broaden our view. Now, we compare how well this best SVM method performs against other popular methods like AdaBoost, Naive Bayes, KNN, and Logistic Regression. This detailed comparison, using various measurements, gives us a clear picture of what each method is good at and where it might struggle. In the end, we make our decision on which method is the winner. This choice is based on looking at all the different measurements like F1 score, accuracy, ROC curve, and confusion matrix. The whole process is like solving a puzzle, and our step-by-step approach ensures we make smart and well-informed predictions about student success. This way, we contribute useful information to the field of educational analytics.

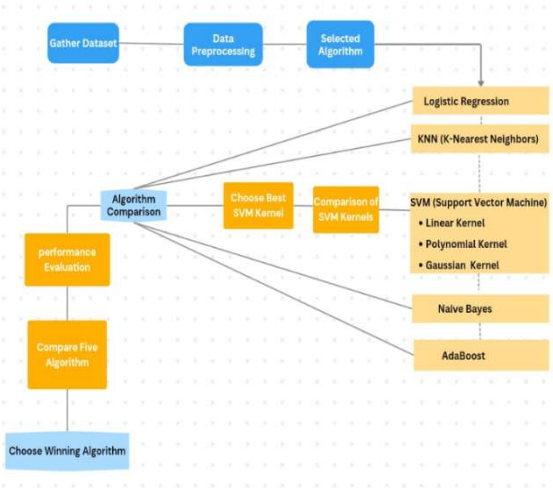

Figure :1 Workflow Overview

IV. OVERVIEW OF THE DATASET

In this research, we examine a dataset centered on secondary education student performance in two Portuguese schools, featuring 395 records and 31 columns. The dataset is comprehensive, containing no missing values, thus obviating the necessity for data imputation. This rich dataset provides insights into demographic, social, and school-related attributes, offering a robust foundation for further analysis and modeling in the context of students' academic achievements.

A. Data Processing

The first crucial step in data processing involves handling non-numeric values. We accomplished this by implementing the numerical_data() function, which maps each non-numeric string to a corresponding integer. Since a lot of machine learning models need numerical inputs, this translation is essential. Following the conversion of non-numeric values, we addressed another critical aspect of data preprocessing – feature scaling. Feature scaling normalizes the range of independent variables, aiding in the convergence of learning algorithms. The feature_scaling(df) function was employed for this purpose, iterating through each column and replacing the values based on a scaling equation. This scaling not only ensures numerical stability but also facilitates faster convergence of machine learning models during training.

As a result of these preprocessing steps, the dataset is now in a form suitable for training models to predict student performance. These procedures lay the foundation for robust machine learning algorithms by transforming and standardizing the data, ultimately enhancing the models' ability to generalize patterns and make accurate predictions.

V. METHODOLOGY

A. Dataset Selection

The first step in our methodology involves the careful curation of a dataset that encapsulates a diverse range of socio-economic and academic attributes. Recognizing the uniqueness of each student's journey, we gather relevant information that may influence academic success. This collection of data is then separated into distinct training and testing sets, ensuring a robust evaluation of the machine-learning models.

B. Data visualization

Data visualization plays a pivotal role in extracting meaningful insights from the dataset, shedding light on key factors influencing student academic success. The following visualizations provide a comprehensive overview of significant observations and patterns:

The bar chart Figure 1 illustrates that individuals who passed the exam tended to spend fewer hours going out with friends. This implies a potential correlation between reduced socializing and improved academic outcomes, suggesting the need to limit social activities for enhanced performance.

The bar chart in Figure 2 showcases the association between the number of failures and exam success. Notably, individuals who passed the exam had minimal failures, suggesting a positive correlation between academic diligence and success.

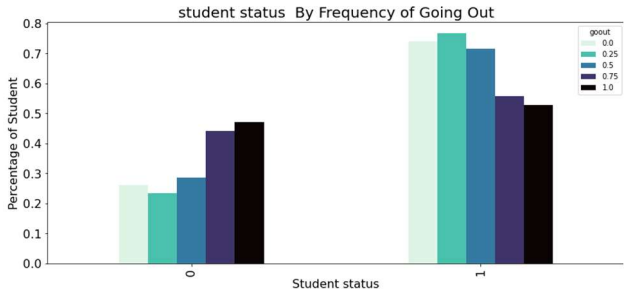

Figure 1: Hours of Going Out and Exam Performance

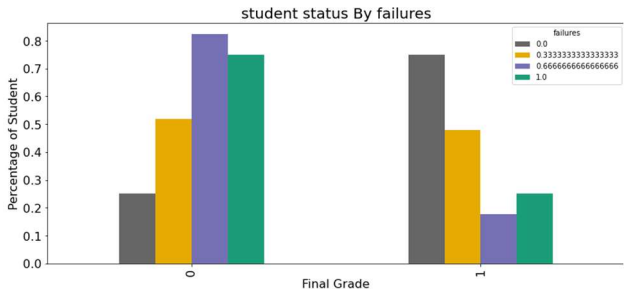

Figure 2: Failures and Exam Performance

Contrary to expectations, Figure 3 demonstrates that the area of residence does not significantly impact student performance. Even individuals with excellent results reside in rural areas, challenging preconceived notions about the correlation between urban living and academic success.

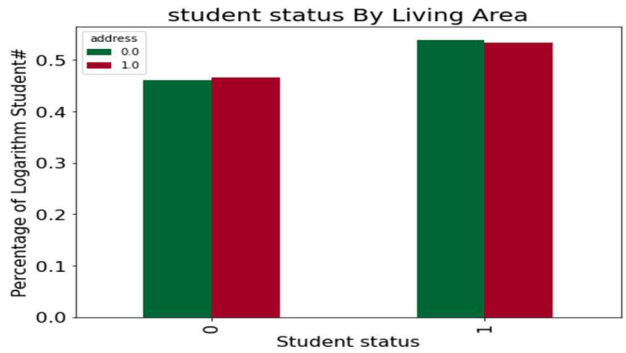

Figure 3: Area of Residence and Academic Performance

The bar chart in Figure 4 underscores the importance of internet accessibility. A higher percentage of individuals who passed the exam had access to the Internet, emphasizing the need for equitable access to educational resources.

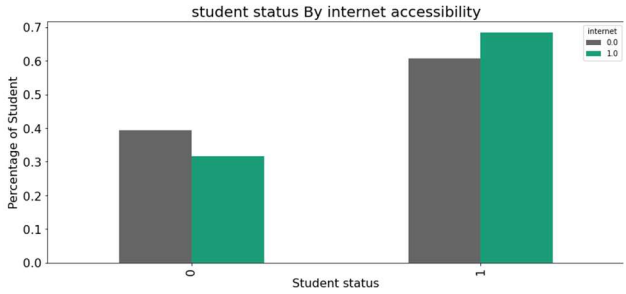

Figure4: Internet Accessibility and Exam Success

Examining weekly study hours in Figure 5 reveals that most successful individuals study between 5 to 10 hours per week. This suggests an optimal range of study hours that correlates with better academic performance.

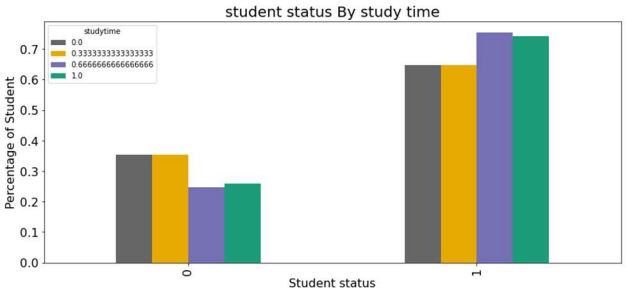

Figure5: Weekly Study Hours and Academic Achievement

These visualizations provide valuableperspectives for academic institutions and decision-makers, aiding in the development of targeted strategies to enhance student success. By addressing factors such as social activities, relationships, parental influence, and resource accessibility, institutions can tailor interventions to improve academic outcomes and promote a conducive learning environment.

C. Machine learning algorithms

The study employs a range of machine learning algorithms to predict student academic success, considering socio-economic and academic factors. The algorithms used include logistic regression, KNN, SVM, AdaBoost, and Naive Bayes. The performance of these algorithms is evaluated using key metrics such as F1 score, precision, recall, ROC curve, and confusion matrix.By identifying the best accurate method for anticipating student outcomes, the thorough comparison aims to provide valuable information to educational institutions and policymakers.

D. Performance Metrics

To gauge the effectiveness of each algorithm, we employ a set of key performance metrics. The F1 rating takes into account both recall and precision, providing a balanced evaluation of the model's accuracy.Recall evaluates the model's capacity to catch all real positiveinstances, whereas precision measures the accuracy of positive predictions.

1. Precision and Recall Analysis

Precision and recall metrics offer a more granular understanding of the model's behavior. In order to address the question of how many anticipated positive instances are actually positive, precision explores the accuracy of positive predictions. Conversely, recall concentrates on the model's capacity to recognize every true positive example, illuminating its comprehensiveness.

2. ROC Curve Evaluation

The difference between the rate of true

positives and false negatives is represented graphically by Receiver Operating Characteristic (ROC) curves. This curve allows us to plot the model's performance over a variety of decision thresholds. A useful metric for assessing classifier performance is the area under the ROC curve, or AUC-ROC.

3. *Confusion Matrix Analysis*

The confusion matrix is a tabular representation of a model's predictions, breaking down results into four categories: true positives, true negatives, false positives, and false negatives. This analysis offers detailed insights into the strengths and weaknesses of the algorithm, aiding in a comprehensive understanding of its effectiveness.

E. Comparison of the Five Algorithms

In this rigorous comparative study, our team evaluated the performance of five diverse classifiers, namely logistic regression, KNN, SVM, AdaBoost, and Naive Bayes. A comprehensive analysis was conducted using a set of key metrics, including F1 score, accuracy score, confusion matrix, precision, recall, ROC curve, and ROC score. The outcomes are visually depicted in Figure 6, presenting a nuanced understanding of each algorithm's strengths and weaknesses across these critical metrics.

The F1 value provides a comprehensive assessment of model correctness by offering a balanced measure of recall and precision. Precision measures how well positive predictions come true, whereas recall evaluates how well the model captures all of the real positive cases.

A detailed evaluation of each classifier's performance is made possible by the ROC curve, which offers a graphical depiction of the trade-off between true positive and false positive rates. Classifier performance is measured by the area under the ROC curve (AUC-ROC).

The confusion matrix delves into granular details, breaking down predictions into categories such asfalse positives, false negatives, true positives, and true negatives. This breakdown provides valuable insights into the algorithm's efficacy.Our thorough comparison goes beyond a simplistic determination of accuracy, offering a comprehensive understanding of each algorithm's handling of various aspects in predicting student outcomes. The visual representation in Figure: 6 encapsulates these findings, aiding in identifying the most effective algorithm.

VI. CONCLUSION

In conclusion, our study comprehensively assessed five machine learning algorithms for predicting student academic success.Through a comprehensive evaluation using key metrics, including F1 score, accuracy, ROC curve, and confusion matrix, we identified SVM as the most accurate classifier.

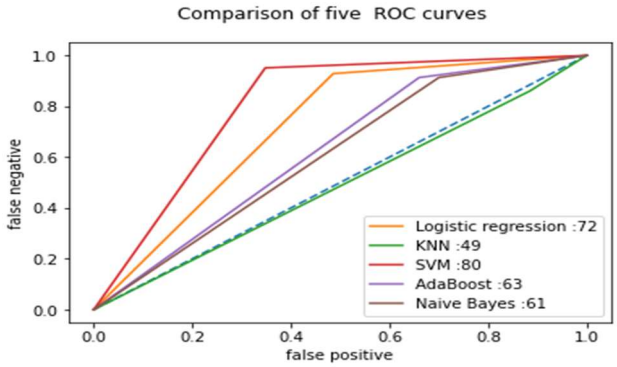

Figure6: Comparison of Five ROC Curves

The findings, illustrated in Figure 6, provide valuable insights for educators, school management, and teachers. SVM's robust performance makes it an optimal choice for early identification of students at risk, enabling targeted interventions and contributing to improved academic outcomes. The research emphasizes the significance of adopting data-driven approaches in education, fostering a proactive and tailored support system. As we navigate the complexities of student success prediction, this work serves as a practical guide for implementing effective strategies in educational institutions.

VII. **REFERENCES**

[1] R. Conijn, C. Snijders, A. Kleingeld and U. Matzat, "Predicting Student Performance from LMS Data: A Comparison of 17 Blended," *IEEETransactions On Learning Technologies,* pp. 17-29, January-March 2017.

[2] S. D. A. Bujang, A. Selama, O. Krejcar, F. Mohamed, L. K. Cheng, P. C. Chiu, and H. Fujita, "Imbalanced Classification Methods for Student Grade Prediction: A Systematic Literature Review," *IEEE Access,* Vols. 11, 2023.

[3] H. A. Mengash, "Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems," *IEEE Access,* Vols. 8, 2020.

[4] S. S. Kusumawardani and S. A. I. Alfarozi, "Transformer Encoder Model for Sequential Prediction of Student Performance Based on Their Log Activities," *IEEE Access,* Vols. 11, 2023.

[5] R. Z. Pek, S. T. Özyer, T. Elhage, T. Özyer, and R. Alhajj, "The Role of Machine Learning in Identifying Students At-Risk and Minimizing Failure," *IEEE Access,* Vols. 11, 2023.

[6] Y. Yao, Z. Zhang, H. Cui, T. Ren, and J. Xiao, "The Influence of Student Abilities and High School on Student Growth: A Case Study of

Chinese National College Entrance Exam," *IEEE Access,* Vols. 7, 2019.

[7] Z. Chen, G. Cen, Y. Wei, and Z. Li, "Student Performance Prediction Approach Based on Educational Data Mining," *IEEE Access,* Vols. 11, 2023.

[8] J. Liang, R. Hare, T. Chang, F. Xu, Y. Tang, F.-Y. Wang, S. Peng and M. Lei, "Student Modeling and Analysis in Adaptive Instructional Systems," *IEEE Access,* Vols. 10, 2022.

[9] A. Zollanvari, R. C. Kizilirmak, Y. H. Kho and D. Hernández-Torrano, "Predicting Students' GPA and Developing Intervention Strategies Based on Self-Regulatory Learning Behaviors," *IEEE Access,* Vols. 5, 2017.

[10] A. Alshanqiti and A. Namoun, "Predicting Student Performance and Its Influential Factors Using Hybrid Regression and Multi-Label Classification," *IEEE Access,* Vols. 8, 2020.

[11] M. Adnan, A. Habib, J. Ashraf, S. Mussadiq, A. A. Raza, M. Abid, M. Bashir, and S. U. Khan, "Predicting at-Risk Students at Different Percentages of Course Length for Early Intervention Using Machine Learning Models," *IEEE Access,* Vols. 9, 2021.

[12] G. Czibula, G. Ciubotariu, M.-I. Maier and H. Lisei, "IntelliDaM: A Machine Learning-Based Framework for Enhancing the Performance of Decision-Making Processes. A Case Study for Educational Data Mining," *IEEE Access,* Vols. 10, 2022.

[13] F. D. Pereira, S. C. Fonseca, E. H. T. Oliveira, A. I. Cristea, H. Bellhäuser, L. Rodrigues, D. B. F. Oliveira, S. Isotani and L. S. G. Carvalho, "Explaining Individual and Collective Programming Students' Behavior by Interpreting a Black-Box Predictive Model," *IEEE Access,* Vols. 9, 2021.

[14] K. Qin, X. Xie, Q. He, and G. Deng, "Early Warning of Student Performance With Integration of Subjective and Objective Elements," *IEEE Access,* Vols. 11, 2023.

[15] P. M. Moreno-Marcos, T.-C. Pong, P. J. Muñoz-Merino and C. D. Kloos, "Analysis of the Factors Influencing Learners' Performance Prediction With Learning Analytics," *IEEE Access,* Vols. 8, 2020.

[16] G. Feng, M. Fan, and Y. Chen, "Analysis and Prediction of Students' Academic Performance Based on Educational Data Mining," *IEEE Access,* Vols. 10, 2022.

[17] H. Prabowo, A. A. Hidayat, T. W. Cenggoro, R. Rahutomo, K. Purwandari, and B. Pardamean, "Aggregating Time Series and Tabular Data in Deep Learning Model for University Students' GPA Prediction," *IEEE Access,* Vols. 9, 2021.

[18] D. Sun, R. Luo, Q. Guo, J. Xie, H. Liu, S. Lyu, X. Xue, Z. Li, and S. Song, "A University Student Performance Prediction Model and Experiment Based on Multi-Feature Fusion and Attention Mechanism," *IEEE Access,* Vols. 11, 2023.

[19] S. Ahmad, K. Li, A. Amin, M. S. Anwar and W. Khan, "A Multilayer Prediction Approach for the Student Cognitive Skills Measurement," *IEEE Access,* Vols. 6, 2018.

[20] P. JIANG and X. WANG, "Preference Cognitive Diagnosis for Student Performance Prediction," *IEEE Access,* Vols. 8, 2020.