Al-Farabi Journal of Humanities, Special Issue of the Ninth Scientific Conference
Volume (2) December (2025)

ISSN: 2957-3874 (Print)
Journal of Al-Farabi for Humanity Sciences (JFHS)
https://iasj.rdd.edu.iq/journals/journal/view/95
مجلة الفارابي للعلوم الإنسانية تصدرها جامعة الفارابي

The Role of Machine Learning in Understanding Student Performance

Hazem Salim Abdullah a*,
Salim Abdullah Saleh b
a Directorate of Municipalities Nineveh Governorate, Mosul, IRAQ
E-Mail: hazemabdullah956@gmail.com
b Retired Assist. Prof. Dr., Tikrit University, Eng. College, Mosul, IRAQ
E-Mail: drsalimasaleh@gmail.com

Abstract
The study assesses the performance of three machine learning models logistic regression, random woodland, and Gaussian Naive Bayes to are expecting the performance of middle School students. Based on metrics which includes F1-score, accuracy, precision, and specificity, the performance turned into determined to be the highest for logistic regression, reaching accuracy and F1-rating values of 72.50% and 94.27%, respectively. In the second role, the random wooded area version scored an accuracy and F1-rating level of 69.17% and 93.64%, respectively, even as GNB scored the accuracy and F1-score degree of 69.17% and 93.19%, respectively. This final results indicates that superior, complex fashions have higher overall performance in comparison to low-complexity algorithms in the context of tutorial statistics mining. Findings support the application of state-of-the-art analytical methods in the direction of deeper understanding of the overall performance of School students, featuring that the stated fashions may be beneficial in supplying individualized learning trajectories and optimizing curricula and getting to know techniques. Appropriate predictive fashions suitable for the unique properties of the records were harassed, as is the significance that information-knowledgeable choice-making provides on the fulfillment stage of School students. By making use of these insights, gaining knowledge of institutions can doubtlessly develop holistically evolved School students, engagement, and motivation stages within the getting to know surroundings.**Keywords:** Academic performance prediction, Educational data mining, Predictive analytics, Machine learning

1. Introduction

The education area has a completely unique possibility to enhance scholar performance metrics and educational institutional consequences through using Big Data analytics. The fast expansion of scholar statistics quantity and grades is a right away end result of the virtual revolution in better schooling [1]. There are increasing expectancies for American academic establishments to enhance educational overall performance, achievement rates, and student retention rates [2]. By using greater sophisticated predictive procedures, academic institutions are starting to transport far away from their reliance on conventional metadata collection [3].The main demand is the development of robust prediction systems that take into account different factors that determine students' success, such as academic achievement, demographics, socioeconomic status, and behavioral characteristics[4].This study evaluates 5 top-performing device studying algorithms for high school student performance, comparing logistic regression, random woodland, and GNB to identify the optimal predictive modeling solution [5].This study is structured as follows: Section 2 reviews the literature on predictive analytics and educational data mining; Section 3 explains the data collection and methodology; Section 4 presents and assesses the results; and Section 5 summarizes the conclusions.

2. Literature Review

2.1 Educational Data Mining and Predictive Analytics

Educational Data Mining, or EDM for short, is a study of learning data in data mining and machine learning contexts. It is a process that attempts to facilitate statistical analysis and enhance the learning results of students. Reference indicates that educational data mining plays the most significant role in influencing how we see learning results. Essentially, EDM is intended to find valuable patterns from massive sets of data to enable educators and administrators to make proper responses. Early detection of students at risk is the first in the discipline because early recognition of these issues may lead to the changes that will enhance academic success [6] [7].

2.2 Factors Influencing Student Performance

Research has found that there are a number of related factors affecting students' performance at school. These can be found within the different areas of the student life. Where we're referring to academics, we're referring to typical standards like how students do academically in terms of grade level, how often they attend school, and what clubs or sports they participate in. How educated their parents are, as well as other factors like their age, how they self-identify gender-wise, and socioeconomic status, all contribute significantly when it comes to how students perform academically [7]. How students learn, how they interact on social media platforms, and how well they get along with teachers and peers are all analyzed through the lens of behavioral factors. Studies indicate that examining both the academic records and extracurricular activities provides a better picture of how academically well students will perform, as highlighted in [8].

2.3 Machine Learning Models for Student Performance Prediction

Logistic regression's baseline model is used in educational research because it is simple and straightforward. From what we can see in [9], logistic regression has good classification accuracy and model interpretability on simple datasets but fails to work well in complex educational problems and non-linear relationships. When you work with predictor variables, logical regression gives clear coefficients and meaningful probability estimates so it is useful for education data mining. Used in combination with well-designed feature selection and preprocessing logistic regression can achieve competitive prediction accuracy as complex procedures. It is also used to predict binary response in schooling. Organizations that require transparent prediction systems that comprehend the reasoning behind computational models would greatly benefit from its explain ability features [10]. Because Decision Trees combine flexible characteristics with user-friendly interfaces, they perform better than traditional statistical methods for predicting academic performance [11]. Furthermore, a similar method approach was used to assess the differences between machine learning and the conventional statistical strategy for cancer patient prediction [12]. From attendance figures to engagement variables, educational administrators and teachers may determine which aspects have a major impact on student achievement because to Decision Trees' visualization capabilities [13]. The study illustrated how Decision Trees employ binary splitting criteria to find significant patterns that can be transformed into targeted support initiatives by educational institutions. Because Random Forest ensemble learning combines many Decision Trees to lower the risk of overfitting, predictive accuracy rises [14] [15]. Random forests have shown excellent overall performance in relation to complex dataset interactions. Teachers can perceive important fulfillment elements using the model's ability to rank important traits [16]. Random Forest's disappearance mechanism, every now and then known as bootstrap aggregating (bagging), is a key component supporting its use in educational research [7]. In managing disparate instructional data units, along with identifying at-hazard students with struggling youngsters who're much less representative than regular students, researchers verified that this methodology is specific. Research suggests that Random Forest improves prediction accuracy via combining ensemble gaining knowledge of with dependable overall performance metrics, producing predictions which can be regular throughout academic environments [17]. Since SVM is now essential for predicting student overall performance in each online and combined learning environments, it plays very well when working with excessive-dimensional datasets [18]. Accessing feature choice strategies for optimization, the observe by way of [19] highlighted how SVMs established excellent capacity to predict student dropout in Massive Open Online Courses (MOOCs). Through its kernel-primarily based approach, the superior mathematical basis of SVM gives special advantages in instructional records analysis, it is claimed [20]. This paper discusses how AI can be used in the study of Monkeypox disease, and its applications can enhance the research by accelerating the process of identifying effective interventions, thus improving the overall outcomes and response to the disease by the population [21]. The scalability combined with the missing value management features of this algorithm ensure its suitability for large educational data analysis projects. The ability of these models to learn themselves makes it possible to apply them in educational contexts, which involve complex relationships between variables that are difficult for human analysts to identify [22].

2.4 Comparative Studies, Research Findings and Gaps

The discipline continuously compares and evaluates different machine learning models to ascertain how well they predict student performance [23]. According to a comprehensive evaluation by ensemble methods Random Forest in particular performed better in terms of accuracy and generalization capacities than Logistic Regression, Decision Trees, and Neural Networks [24]. For predicting university grades, SVM and Random Forest worked best together, however Logistic Regression was a useful baseline comparison for performance assessment [25]. Deep learning algorithms and ensemble approaches have been shown to outperform Random Forests, Decision Trees, and Logistic Regression in terms of accuracy and generalization. The most successful methods for university-level prediction are SVM and Random Forest with Neural Networks, whereas logistic regression offers essential performance data for comparison [26]. There are still unanswered questions in solvable models, such neural networks, especially in terms of their interpretability and applicability. These models' dependence on tiny datasets limits their use in other educational situations. Additionally, there is no attempt to forecast dynamic student performance by integrating adaptive learning systems with real-time data. A research uses university-level data to evaluate logistic regression, decision trees, random forests, support vector machines, and neural networks in order to fill these gaps.

3. Methodology

3.1 Dataset Overview

The student performance dataset is an integrated model that simulates authentic educational environments, incorporating data from 1,200 male and female students aged between 16 and 20 years in the Nimrud district of Nineveh Governorate. The dataset includes ten independent variables, including gender, age, attendance rate, homework completion, classroom participation, previous academic scores, weekly study hours, and extracurricular activities, and two interactive variables to enhance predictive model accuracy see the(Figure 1 Shows the most important features). Data generation uses sophisticated scientific methodologies, utilizing normal distributions to simulate natural student variability and introducing a 10% random noise component to replicate unobserved factors in genuine educational settings.The dependent variables, which measure educational achievement along three interrelated dimensions - cognitive, affective and psychosocial - are designed according to Bloom's comprehensive educational taxonomy. To maximize algorithmic performance, computing uses sophisticated techniques such as the development of interactive variables, numerical coding of categorical variables, and data scaling methods. Using machine learning methods, three sophisticated predictive models were created: Naive Bayesian Classification, Random Decision Forest and Multitask Logistic Regression. The dataset shows a particular ability to reproduce between-school variations in average student achievement and trait distributions, and the settings of each algorithm were adjusted to achieve optimal performance.An integrated forecasting system was developed, allowing users to enter new student data and receive personalized improvement suggestions in addition to real-time performance estimates in all three dimensions. This technique is useful for academic research and educational policy development because it makes it easier to generate random data and compare large groups of students. The findings are presented as complete reports that include ROC curves, correlation matrices, variable importance charts and performance metrics. These reports can be fully exported to an Excel spreadsheet for group analysis. For the stratified sample, a rigorous scientific process was used, and 1,147 male and female students were randomly selected from a total of 1,200. Due to the strict inclusion and exclusion criteria used in the sample design, only 4.4% of the data were excluded, demonstrating the accuracy and reliability of the data. The study sample and any biases were carefully examined in (Figure 8 shows Chart of Sample Distribution). We can see the age distribution of the sample (16-20 years), with percentages ranging from 10.4% to 25.5%. With 52.3% of the population being male and 47.7% female, the gender distribution is fairly balanced. The outstanding attendance rate of 84% indicates that the participants were highly dedicated. However, since this study was conducted in a specific area where the percentage of geographical bias was 85%, major methodological biases should be addressed in the future. Among the recommendations, the geographical distribution could be improved, and regarding age representation (75%), the target group for the preparatory study is approximately the same age group.

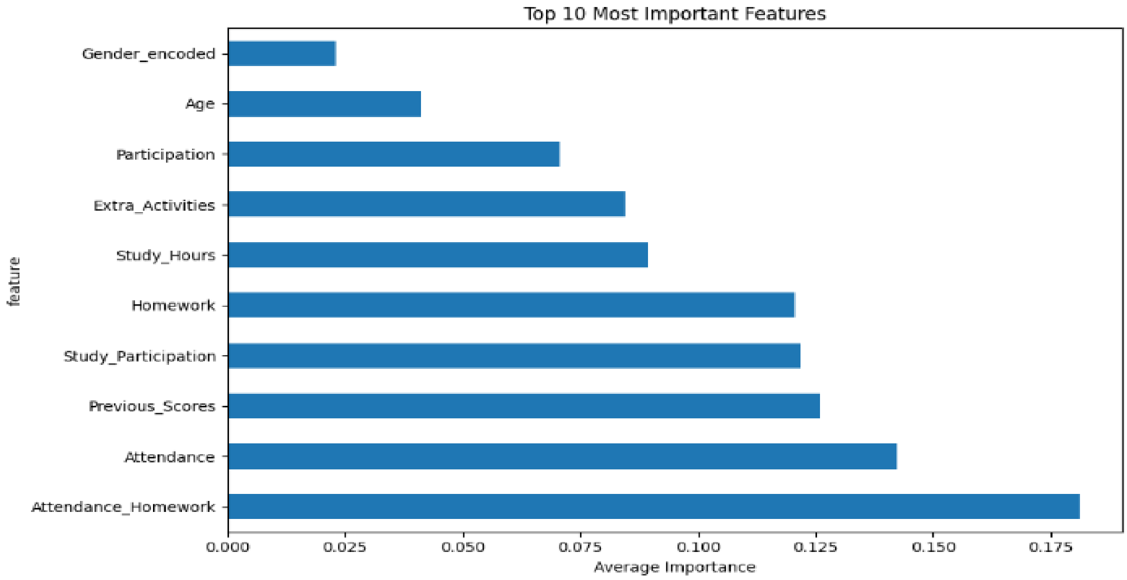

Figure 1: Top10 most Important feature

3.2 Data Preprocessing in Educational Performance Predictive Modeling

The data preprocessing step is a crucial step within the development of strong academic predictive fashions. It entails deciding on seven core academic variables, including attendance styles, homework of completion rates, magnificence participation levels, historic academic overall performance, devoted take a look at time, extracurricular involvement, and demographic age traits, that exhibit predictive validity in the academic performance literature. The gender variable undergoes label encoding to transform the textual illustration into numerically computable values even as keeping the statistical integrity of the variable. The preprocessing framework then proceeds thru the development of sophisticated interactive features that seize the complicated synergistic relationships among educational elements. The architectural design of the preprocessing pipeline ensures systematic records agency, creating duplicate records sets that hold the original statistics integrity at the same time as permitting transformative operations. This effects in a comprehensive characteristic matrix that consists of each simple educational metrics and constructed interactive variables, developing a multidimensional function space that is optimized for predicting tripartite instructional results. The dataset partitioning method makes use of the traditional 80-20 schooling-testing split, allocating 960 pupil records for version development and 240 records for validation. This cautiously designed preprocessing framework ensures dimensional consistency among the function matrix and goal variables, maintains relevance to the educational context, and permits seamless integration with computerized gadget gaining knowledge of pipelines. The best basis for building predictive models that appropriately represent the complex interplay of student academic performance across cognitive, affective, and psychomotor domains is created by this procedure, which combines statistical rigor with pedagogical awareness. The heatmap (Figure 2 Shows the correlation matrix for academic performance characteristics).

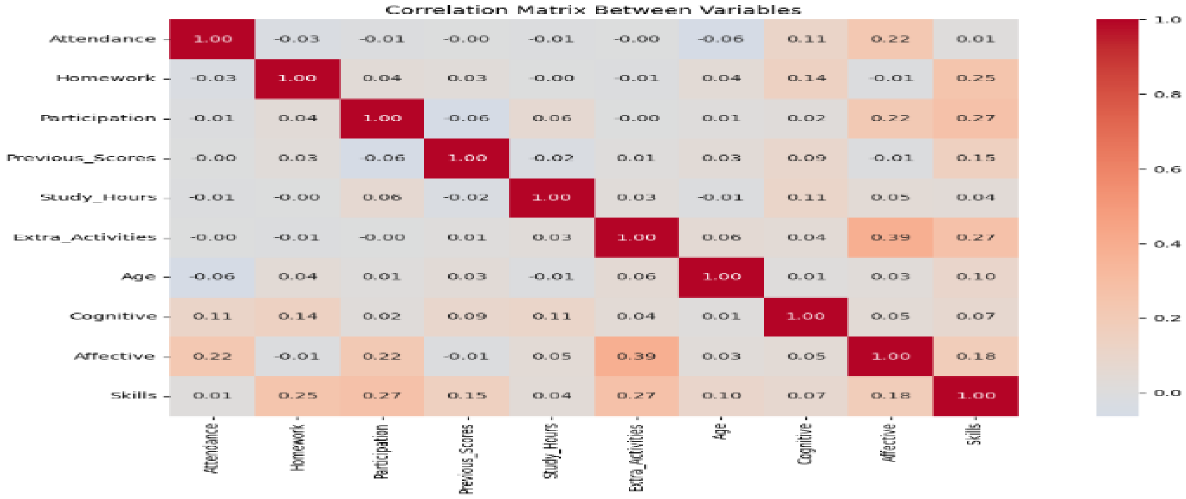

Figure 2: Correlation Matrix for Academic Performance Features

Python programming was used to develop models using popular libraries like pandas, NumPy, scikit-learn, TensorFlow, and Keras, with hyperparameters identified through cross-validation techniques. The system used for analysis was an i5 CPU, Windows 11 operating system, and 16 GB RAM.

3.3 Data Analysis Algorithms

One of the most significant shifts has been to make predictive analytics and machine learning an integral part of our daily work, displacing basic statistics and gut intuition with machine learning. were utilized in this research as predictors of student success in the available academic history. All algorithms certainly have advantage and disadvantage in terms of precision, interpretability and computational efficiency. The techniques that we employ in our study are Random Forest (combines decision trees), GNB (an interpretable, but very simple method) and logistic regression [27]. These algorithms are then selected for a variety of models, from interpretable linear models to more complex nonlinear models, in order to fully assess how well they predict student achievement. In the parts that follow:

3.3.1 Logistic Regression Algorithm

The algorithm is an optimal statistical model for predicting multi-class classification outcomes. It uses a logistic function to transform linear results into probabilistic values. The algorithm has been adapted to perform multiple tasks simultaneously, thereby mitigating risk of overfitting. The advanced L-BFGS optimizer is employed for parameter estimation and a regularization coefficient is included to govern for the complexity of the model. The proposed method yields steady parameter estimates that allow the evaluation of the impact of independent variables on the output probability, providing insight into educational performance indicators [27].

3.3.2 Random Forest Decision Algorithm

The algorithm applies ensemble learning to enhance prediction performance by constructing many independent decision trees. The algorithm generates bootstrap samples from the original data and builds independent trees using each sample and random restrictions on split features. It lowers variation and increases model stability by fostering diversity. The program handles complex interactions and non-linear correlations without changing the data. It also tracks purity improvements by split across all trees which can help researchers identify significant variables in the modeling process for educational performance predictions [28].

3.3.3 Gaussian Naive Bayes Algorithm

The algorithm is based on Bayes' theorem and assumes that the features are conditionally independent for classification purposes. It has performed well empirically, particularly with high-dimensional data. Posterior class probabilities are obtained as prior probabilities times the probabilities of the features. The Gaussian version of algorithm assumes that the continuous variables are normally distributed and estimates the parameters of this distribution from the training data. The method allows for probabilistic predictions from the model to be reported directly and is computationally efficient [29].

3.3.4 Evaluation metrics

Using metrics such as accuracy, recall, precision, the F1-score, and the False Positive Rate from the confusion matrix, the study assesses the efficacy and efficiency of a classification model. Each performance metric is calculated as follows:

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

$$\text{Recall} = TP/(TP + FN) \qquad [2]$$

$$\text{Precision} = TP / (TP + FP) \qquad [3]$$

$$F1 - \text{Score} = 2TP / (2TP + FP + FN) \qquad [4]$$

$$FPR = FP / (FP + TN) \qquad [5]$$

where, TP is the number of true positive cases, TN is the number of true negative cases, FP is the number of false positive cases, and FN is the number of false negative cases [30].

4. Results and Discussion

4.1 Logistic Regression Analysis

The logistic regression model is an extremely reliable predictive model for multidimensional educational assessment, with an overall accuracy of 72.50% and outstanding results in terms of precision (91.45%) and recall (97.32%) see the (Table 1 shows Performance Measures). The model is able to classify 233 students and made

150 errors when classifying successful students. The model was able to identify 201 successful students but it incorrectly classified 20 students see the (Figure 3 shows Confusion matrix). This suggests that affective indicators are more discernable predictors in the data structure than more sophisticated cognitive indicators. The model reports an impressive F1 score of 94.27% showing it is robust enough for educational assessment practices. Looking across particular dimensions the model is consistently impressive with cognitive dimension 82.89%, affective dimension 85.94%, and skills dimension 85.48% see the (Table 2 shows AUC by dimension). The cognitive dimension of the assessment achieves perfect recall of 100% and 97% precision. The skills dimension performed strong in predicting students with 95% recall and 85% with precision. Therefore, claims can be made for the reliability of the model for complete assessment of students and early detection of exceptional students.

Table 1. Performance Measures for Logistic Regression Analysis				
Accuracy	**Precision**	**Recall**	**F1 Score**	**Mean AUC-ROC**
72.5	91.45	97.32	94.27	84.77

Table 2. AUC by dimension		
Cognitive	**Affective**	**Skills**
82.89	85.94	85.48

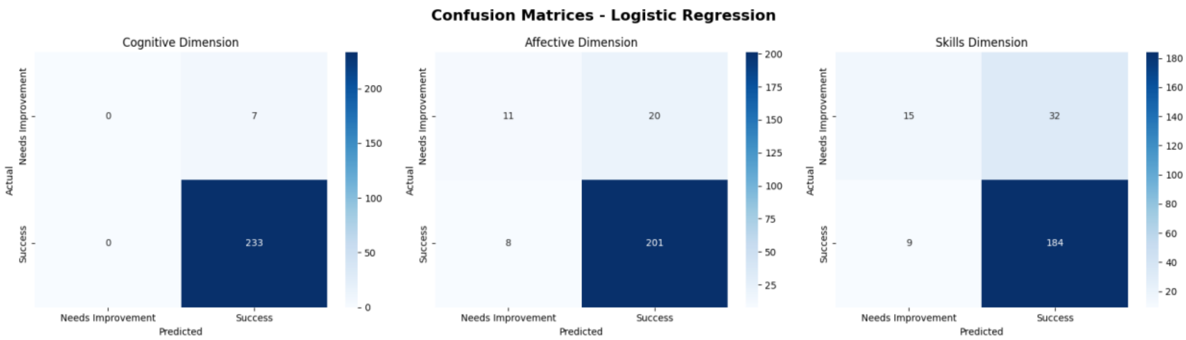

Figure 3: Confusion matrix of Logistic Regression

4.1.2 Random Forest Analysis

The Random Decision Forest model achieves a more complete balance, actively preserving its strength to identify students who may need to make improvements in the cognitive dimension (233 correct classifications), while indicating an appreciable improvement in reporting successful students, compared to the previous model. In the affective dimension, the model demonstrates 199 correct classifications of successful student outcomes versus only 23 errors, providing further evidence that behavioral indicators may be represented more transparently and effectively by tree models (see the Figure 4 shows Confusion matrix). This balanced sample hooked up for the Random Decision Forest model is specifically remarkable as it demonstrates the model to be effective in coping with the varied complexities found in instructional datasets. The version does truly display variations in performance across the academic dimensions, consultant of the complexity of the educational procedure and the special and complicated nature of every academic area. Finally, the Random Forest version gives a greater blended overall performance predicting educational overall performance normal, attaining a standard accuracy stage of 69.17%, with a high remember charge, at 97.17%. The Random Forest model can show a good stability in producing schooling performance type values, as tested by means of a balanced F1 rate of 93.64% see the (Table 3 shows Performance Measures) . The Random Forest model does show first-rate energy within the affective (86.09%) and skill (81.27%) dimensions, with lower overall performance effects shown in the cognitive size (65.97%) see the (Table 4 shows AUC by dimension). The version does receive remarkable overall performance inside the cognitive size for accuracy (97%), and recall (100%), at the same time as performance is a bit lower inside the ability measurement (83% accuracy, 96% recall). Despite the version in performance across the exclusive dimensions, the model stays a useful device for helping academic decisions, especially in figuring out behaviorally and emotionally advanced college students, with a need to improve overall performance in cognitive elements.

Table 3. Performance Measures for Random Forest Analysis				
Accuracy	**Precision**	**Recall**	**F1 Score**	**Mean AUC-ROC**

69.17	90.45	97.17	93.64	77.78

Table 4. AUC by dimension		
Cognitive	Affective	Skills
65.97	86.09	81.27

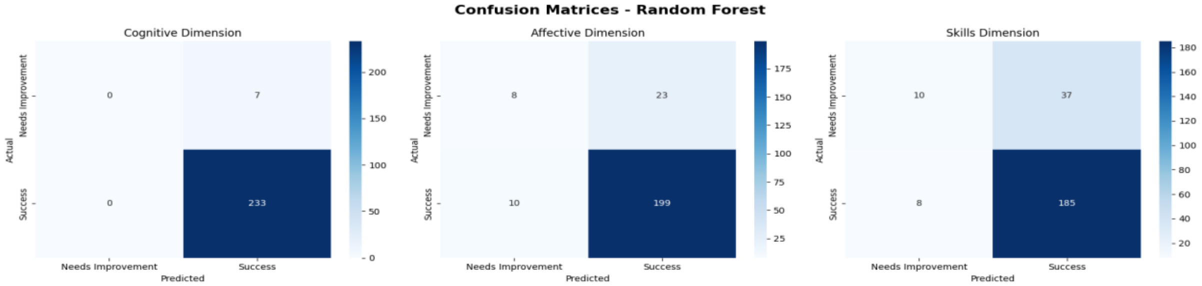

Figure 4: Performance Measures for Random Forest Analysis

4.1.3 Gaussian Naive Bayes model

The Gaussian Naive Bayes model produces different results when it analyzes the three educational dimensions. The model demonstrates a clear tendency to over-classify which reduces its ability to accurately analyze various categories. The Gaussian Naive Bayes model, however, shows a blatant systematic bias for over-classification, identifying 226 students as successful on the cognitive dimension, while missing students who needed to improve almost entirely (only 7 instances). The same trend persisted on the affective dimension, with 196 correct classifications of students who were successful compared to only 10 students who needed improvement (see the Figure 5 shows Confusion matrix). This consistency of bias, across dimensions, suggests that the assumptions of this algorithm are not appropriate for the complexity of educational data. The simple Gaussian Bayes model demonstrates balanced and stable performance with the educational assessment, while producing an overall accuracy of 69.17% and a remarkable balance in precision (91.41%) and recall (95.12%) reflected by the F1 rate of 93.19%, see the (Table 5 shows Performance Measures). Model performance in the cognitive dimension (85.47%) was complemented by consistency in the affective (85.07%) and skill (81.89%) dimensions such as (Table 6 shows AUC by dimension). The details of the report demonstrate remarkable consistency across educational dimensions and utility as an efficient way to support educational decisions and offer appropriate support to students.

Table 5. Performance Measures for GNB				
Accuracy	Precision	Recall	F1 Score	Mean AUC-ROC
69.1	91.41	95.12	93.19	84.14

Table 6. AUC by dimension		
Cognitive	Affective	Skills
85.47	85.07	81.89

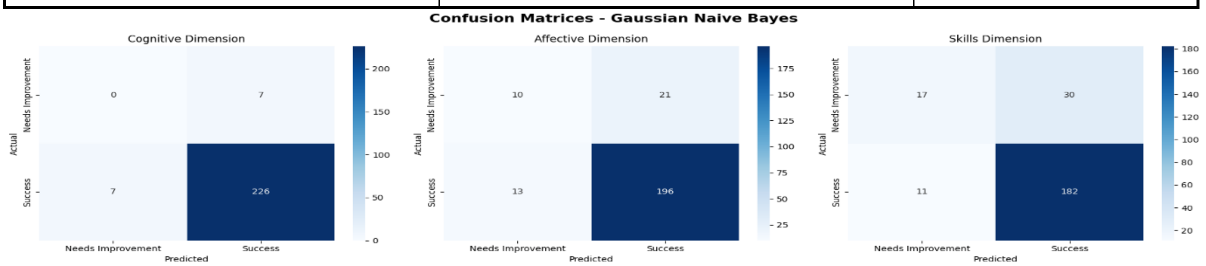

Figure 5: Performance Measures for Gaussian Naive Bayes Analysis

4.1.4 Discussion

The logistic regression model performed best, achieving a predictive accuracy of 72.50%, while the GNP model achieved 69.17%, and the random forest model had a lower accuracy of 69.17% but demonstrated robustness when dealing with emotional and behavioral traits (see Figures 6 shows Model Performance and Figures 7 shows AUC-ROC BY Dimension and model). The overall average F1 scores were 94.27%,93.19% and 93.64%, respectively, while the precisions were 91.45, 91.41, 90.45 and, respectively. Logistic regression was slightly higher than the other models, achieving excellent performance. Despite some overconfidence in cognitive classification, it is considered optimal for early detection in education. The proposed logistic regression model remains the most suitable for fair educational assessment, while all three proposed models outperform related studies on key performance indicators (see Table 7 Compared with the results of several researchers).

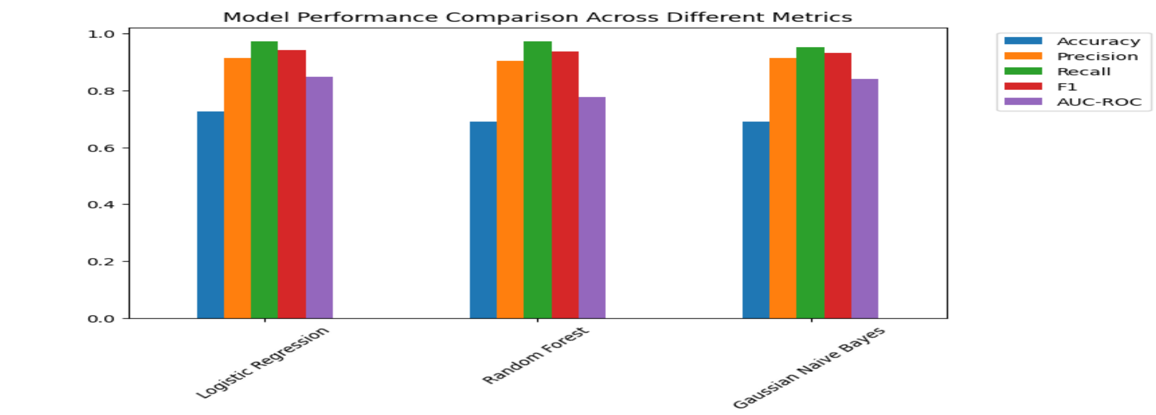

Figure 6: Model Performance Across Different Metrics

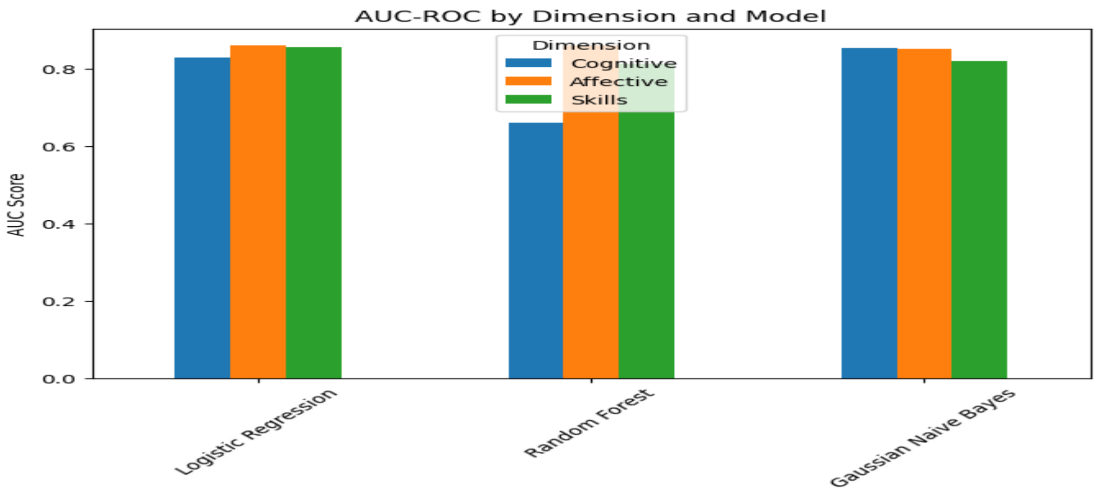

Figure 7: AUC-ROC BY Dimension and model

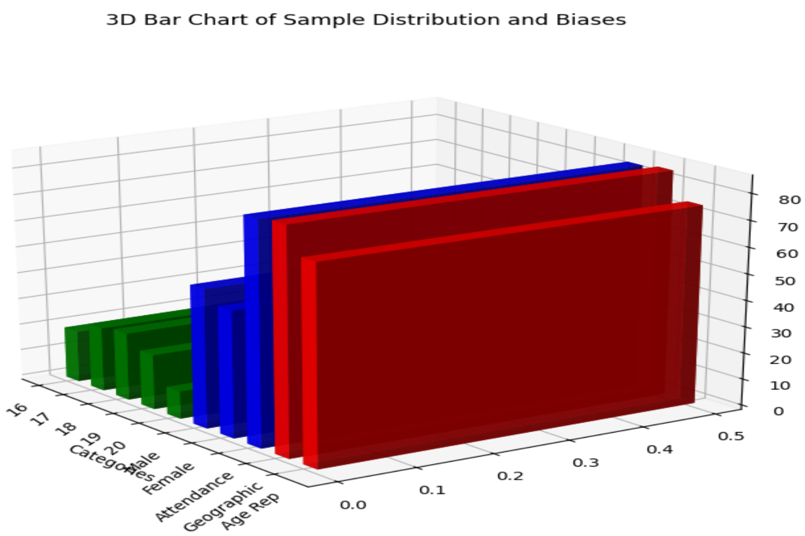

Figure 8: Chart of Sample Distribution

Table 7. A Comparison between different ML models

Ref. No.	Model	Acc. (%)	Precision	Recall	F1 Score	Model Type
Proposed	**Logistic Regression (LR)**	**72.50**	**91.45**	**97.32**	**94.27**	**Machine Learning (ML) Models**
[31]	LR	72.63	64.10	54.35	58.82	
Proposed	**GNB**	**69.17**	**91.41**	**95.12**	**93.19**	
[32]	GNB	69	70.28	63.28	60.28	
Proposed	**Random Forest (RF)**	**69.17**	**90.45**	**97.17**	**93.64**	
[31]	RF	78.13	74.36	61.70	67.44	

5. Conclusion

An extensive study of machine learning algorithms for predicting student academic achievement yielded several findings that contribute to educational data mining and analysis research. The study focuses on three different machine learning models: random forest, logistic regression, and GNB. The more advanced algorithms yield better prediction results in the educational context. The report cites the logistic regression model as the best model included in the report, achieving the best F1 value (94.27%) and average AUC-ROC (84.77%). These demonstrate a remarkable balance between excellent predictive value and sensitivity, earning the model a Predictive Excellence badge. Doubts should be avoided in this regard, as the model's prediction accuracy (72.50%) also indicates superiority, while other metrics provide absolute confidence in the model's performance. These effects dictate the primary purpose of the research. This study sets a new standard in educational predictive modeling. The analytical strategies used to evaluate it are also diverse. Furthermore, future research should focus on the challenges facing predictive models in several key areas. First, developing hybrid modeling techniques that combine the best features of multiple algorithms while minimizing their weaknesses. Second, integrating temporal dynamics and longitudinal data to capture the evolving nature of student learning processes.

References

[1] M. A. Hashim, I. Tlemsani, and R. Matthews, "Higher education strategy in digital transformation," Educ. Inf. Technol., vol. 27, no. 3, pp. 3171–3195, 2022.

[2] V. Pendakur, Ed., Closing the Opportunity Gap: Identity-Conscious Strategies for Retention and Student Success. Taylor & Francis, 2023.

[3] W. Seo and Y. Bu, "Transforming role classification in scientific teams using LLMs and advanced predictive analytics," arXiv preprint arXiv:2501.07267, 2025.

[4] M. M. Asad and A. Qureshi, "Impact of technology-based collaborative learning on students' competency-based education: Insights from the higher education institution of Pakistan," High. Educ. Skills Work-Based Learn., 2025.

[5] A. Malik et al., "Forecasting students' adaptability in online entrepreneurship education using modified ensemble machine learning model," Array, vol. 19, p. 100303, 2023.

[6] H. Almaghrabi, B. Soh, A. Li, and I. Alsolbi, "SoK: The impact of educational data mining on organisational administration," Information, vol. 15, no. 11, p. 738, 2024.

[7] V. Onker et al., "Harnessing machine learning for academic insight: A study of educational performance in Bhopal, India," Educ. Inf. Technol., pp. 1–40, 2025.

[8] A. Ahmad, S. Ray, M. T. Khan, and A. Nawaz, "Student performance prediction with decision tree ensembles and feature selection techniques," J. Inf. Knowl. Manag., p. 2550016, 2025.

[9] Y. Salunke et al., "Fraud detection: A hybrid approach with logistic regression, decision tree, and random forest," Cureus, vol. 17, no. 1, 2025.

[10] A. Leiva-Araos, C. Contreras, H. Kaushal, and Z. Prodanoff, "Predictive optimization of patient no-show management in primary healthcare using machine learning," J. Med. Syst., vol. 49, no. 1, p. 7, 2025.

[11] J. Fan, “A big data and neural networks driven approach to design students management system,” Soft Comput., vol. 28, no. 2, pp. 1255–1276, 2024.
[12] N. I. Elzayat et al., “A comparative study for survival prediction of NKI data using statistical and machine-learning approaches,” in 2024 6th Novel Intelligent and Leading Emerging Sciences Conf. (NILES), pp. 269–273, Oct. 2024.
[13] M. El Jihaoui, O. E. K. Abra, and K. Mansouri, “Factors affecting student academic performance: A combined factor analysis of mixed data and multiple linear regression analysis,” IEEE Access, 2025
[14] X. Wan, J. Zeng, and L. Zhang, “Predicting online shopping addiction: A decision tree model analysis,” Front. Psychol., vol. 15, p. 1462376, 2025.
[15] J. F. Hair Jr, L. P. Fávero, W. T. Junior, and A. Duarte, “Deterministic and stochastic machine learning classification models: A comparative study applied to companies’ capital structures,” Mathematics, vol. 13, no. 3, p. 411, 2025.
[16] X. Bai, L. Zhang, Y. Feng, H. Yan, and Q. Mi, “Multivariate temperature prediction model based on CNN-BiLSTM and RandomForest,” J. Supercomput., vol. 81, no. 1, p. 162, 2025.
[17] Z. Khoudi, N. Hafidi, M. Nachaoui, and S. Lyaqini, “New approach to enhancing student performance prediction using machine learning techniques and clickstream data in virtual learning environments,” SN Comput. Sci., vol. 6, no. 2, p. 139, 2025.
[18] S. Zhao, D. Zhou, H. Wang, D. Chen, and L. Yu, “Enhancing student academic success prediction through ensemble learning and image-based behavioral data transformation,” Appl. Sci., vol. 15, no. 3, p. 1231, 2025.
[19] S. Alghamdi, B. Soh, and A. Li, “A comprehensive review of dropout prediction methods based on multivariate analysed features of MOOC platforms,” Multimodal Technol. Interact., vol. 9, no. 1, p. 3, 2025.
[20] K. L. Du, B. Jiang, J. Lu, J. Hua, and M. N. S. Swamy, “Exploring kernel machines and support vector machines: Principles, techniques, and future directions,” Mathematics, vol. 12, no. 24, 2024.
[21] Y. L. Khaleel, M. A. Habeeb, and R. Benotsmane , Trans., “Emerging Trends in Applying Artificial Intelligence to Monkeypox Disease: A Bibliometric Analysis ”, Applied Data Science and Analysis, vol. 2024, pp. 148–164, Sep. 2024, doi: 10.58496/ADSA/2024/012.
[22] G. Alhussein, M. Alkhodari, A. H. Khandoker, and L. J. Hadjileontiadis, “Novel speech-based emotion climate recognition in peers’ conversations incorporating affect dynamics and temporal convolutional neural networks,” IEEE Access, 2025.
[23] H. Pallathadka et al., “Classification and prediction of student performance data using various machine learning algorithms,” Mater. Today: Proc., vol. 80, pp. 3782–3785, 2023.
[24] N. Rane, S. P. Choudhary, and J. Rane, “Ensemble deep learning and machine learning: Applications, opportunities, challenges, and future directions,” Stud. Med. Health Sci., vol. 1, no. 2, pp. 18–41, 2024.
[25] G. Latif, S. E. Abdelhamid, K. S. Fawagreh, G. B. Brahim, and R. Alghazo, “Machine learning in higher education: Students’ performance assessment considering online activity logs,” IEEE Access, vol. 11, pp. 69586–69600, 2023.
[26] A. Kye, “Comparative analysis of classification performance for US college enrollment predictive modeling using four machine learning algorithms (logistic regression, decision tree, support vector machine, artificial neural network),” Ph.D. dissertation, Loyola Univ. Chicago, 2023.
[27] I. A. Mohammed, “Machine learning in predictive analytics: A game changer for business decision making,” 2020.
[28] Z. Sun, G. Wang, P. Li, H. Wang, M. Zhang, and X. Liang, “An improved random forest based on the classification accuracy and correlation measurement of decision trees,” Expert Systems with Applications, vol. 237, p. 121549, 2023. doi: 10.1016/j.eswa.2023.121549.
[29] K. Gohari, A. Kazemnejad, M. Mohammadi et al., “A Bayesian latent class extension of naive Bayesian classifier and its application to the classification of gastric cancer patients,” BMC Medical Research Methodology, vol. 23, no. 190, 2023.
[30] H. S. Abdullah and S. A. Saleh, “Securing Smart Cities – The Impact of Machine Learning Technology,” j.komputer, j.informasi, j.teknologi, vol. 5, no. 1, p. 19, Jun. 2025.
[31] A. Bashir, "Understanding Student Performance in Foundation Year: Insights from Logistic Regression, Naïve Bayes, and Random Forest Models," Int. J. Inf. Educ. Technol., vol. 14, pp. 1716–1723, 2024, doi: 10.18178/ijiet.2024.14.12.2202.

[32] A. Harif and M. A. Kassimi, "Predictive Modeling of Student Performance Using RFECV-RF for Feature Selection and Machine Learning Techniques," Int. J. Adv. Comput. Sci. Appl. (IJACSA), vol. 15, no. 7, 2024.