A Logistics Regression-Based Student Performance Prediction System

Adebanjo, Adedoyin S.[1], Anyanwu, Chiamaka G.[2], Adeoti, Babajide E.[3], Mgbhearuike, Emmanuel[4], Oyerinde, Emmanuel I[5]

[1,3,4.] *Department of Software Engineering, Babcock University, Ogun State, Nigeria*

[2.] *Department of Computer Science, Babcock University, Ogun State, Nigeria*

[5.] *Department of Information Technology, Babcock University, Ogun State, Nigeria*

***Abstract*— Student performance prediction has become an important focus in educational data mining. Institutions are looking for ways to use data to identify at-risk learners and improve academic outcomes. Logistic Regression, a widely used statistical and machine learning model, offers a clear way to model categorical outcomes, such as pass/fail or high/low performance. This study uses Logistic Regression on student performance data to examines how demographic, behavioral, and academic features affect learning outcomes. By using its probabilistic framework, Logistic Regression predicts student success with high accuracy. It also reveals the importance of various predictors, helping educators design targeted interventions. The findings show that Logistic Regression is an effective predictive model as it balances accuracy and clarity, contributing to early warning systems and better decision-making in higher education.**

***Index Terms*— Logistics Regression, Machine Learning, Student Performance Prediction, Supervised Learning.**

I. INTRODUCTION

In recent years, the education sector in Nigeria has seen a significant decline in student performance in both internal and external exams. Poor academic results have been reported at all levels of education [1]. A recent report from the Joint Admission and Matriculation Board (JAMB) showed that about 1.4 million candidates scored below 200 in the 2024 Unified Tertiary Matriculation Examination (UTME). This highlights the seriousness of the problem [2]. While this number pertains to Nigeria, student dropout is a global issue. Dropout rates exceed 40% in some European and Latin American countries, and around 30% of first-year students in U.S. Baccalaureate Institutions do not return for their second year [3]. The United Nations has pointed out that poverty, poor infrastructure, limited funding, and low quality of life in urban slums are major factors leading to poor educational outcomes in developing countries [4]. These challenges often result in delayed graduation, low self-esteem, and, in extreme cases, withdrawal from academic programs [3].

This trend highlights the urgent need for effective solutions that can improve both student retention and performance. One promising method is using predictive analytics and data mining techniques to identify at-risk students. This allows for proactive academic support [5]. Student Performance Prediction (SPP) is becoming an important area of research in Educational Data Mining (EDM) and Machine Learning. It goes beyond grades and looks at the skills and knowledge needed for academic and societal success [6]. By using historical and behavioral data, SPP offers useful insights for everyone involved. Students can plan their learning paths, instructors can adjust their teaching methods, and institutions can improve retention with targeted support [6].

Several studies have looked into the potential of SPP using various machine learning methods. Al Husaini and Shukor [7] conducted a review of literature from 2014 to 2020. They identified a wide range of internal factors, like entry grades and family support, as well as external factors, such as socioeconomic background and e-learning engagement, that influence academic outcomes. They noted that female students generally show more persistence. Similarly, Tjandra et al. [8] analyzed over 250 studies on SPP. They found that most of these studies focused on monitoring learning activities (67.2%). Fewer studies (9.2%) addressed dropout prevention. Feng et al. [9] demonstrated the use of machine learning to predict academic

performance, along with standardized tests, teacher ratings, and classroom observations.

With more schools using digital technologies and having access to a lot of student data, Educational Data Mining (EDM) offers new chances to gather knowledge and improve learning outcomes. This study adds to this expanding area by using supervised machine learning algorithms, specifically Logistic Regression, to predict student performance in Nigerian schools. The insights gained are meant to help with curriculum design, guide academic interventions, and improve overall student success in both internal and external examinations.

The aim of the study is to develop a student performance prediction system using Logistic Regression model that predicts student performance and student Cumulative Grade Point Average (CGPA) respectively.

The specific objectives of the study are:

i. To evaluate the relationship between student performance and their grades.
ii. To identify critical features that contribute to student performance.
iii. To design and develop a Logistic Regression model to predict student performance.
iv. Evaluate the model accuracy and performance

II. LITERATURE REVIEW

A. Educational Data Mining (EDM) and Learning Analytics (LA)

Educational Data Mining (EDM) and Learning Analytics (LA) have become important methods for tackling issues like student retention, dropout rates, and overall performance in higher education. EDM uses computational and statistical techniques to find meaningful patterns in large datasets created by students. Meanwhile, LA focuses on interpreting this data to generate actionable insights [10], [11].

Romero and Ventura [12] examined how EDM can predict academic performance. They emphasized the use of classification, clustering, and association rules. These methods help educational institutions identify risks of failure, suggest personalized learning paths, and support adaptable curricula. Similarly, Siemens and Long [11] explained that LA allows institutions to monitor students in real time, which enables early interventions to enhance outcomes.

Both EDM and LA are now critical in developing predictive models for student outcomes, enhancing resource allocation, and designing interventions for retention and success [10]-[14]. Frameworks in EDM typically emphasize predictive modeling using algorithms such as Support Vector Machines (SVM), Random Forests (RF), and Gradient Boosting (GB), while LA frameworks focus on systematic data collection and reporting to inform pedagogical decisions [10], [13]-[15]. Hybrid approaches that integrate EDM and LA are increasingly popular, providing comprehensive pipelines that include preprocessing, feature selection, model validation, and deployment for predictive and recommendation systems [15]. The relationship between EDM and LA lays the groundwork for predictive analytics in education. This combination equips stakeholders with data-driven strategies aimed at improving learning outcomes and the effectiveness of institutions.

B. Predictive Modeling in Education

Predictive modeling techniques are increasingly used to estimate student performance indicators like grade point average (GPA), exam scores, and dropout risk. Earlier models mostly relied on regression analysis, but recent studies show a growing preference for machine learning methods, including random forests, gradient boosting, and support vector machines [16]-[21]. These algorithms effectively capture nonlinear relationships, making them a good fit for complex educational data.

However, "black-box" models have a drawback: they are hard to interpret. Because of this, regression-based approaches still hold value, especially in academic settings where transparency and explanation matter as much as accuracy [13], [15], [22]-[25]. Logistic regression, in particular, strikes a balance between predictive performance and interpretability. It helps stakeholders pinpoint the relative importance of factors like academic preparation, socioeconomic background, and learning engagement in determining student success.

C. Logistic Regression in Student Performance Prediction

Logistic regression (LR) is commonly used to predict binary outcomes like pass/fail, success/dropout, or retention/non-retention. The logistic function that forms the basis of LR is expressed as:

$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$ (i)

where θ represents the parameter vector, and x is the input feature vector [24], [26]. The model parameters are improved by reducing the cost function:

$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log\left(h_\theta\left(x^{(i)}\right)\right) + \left(1 - y^{(i)}\right)\log\left(1 - h_\theta\left(x^{(i)}\right)\right)\right] + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2$ (ii)

Here, the inclusion of the regularization term $\frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2$ helps prevent overfitting, which is common in high-dimensional educational datasets [25], [27].

Logistic regression is a common method in educational data mining for binary classification tasks, like predicting whether a student will succeed or fail. It calculates the probability of an input belonging to a specific class, usually by applying a logistic function to a linear combination of features. The output can be seen as a probability, and by setting a threshold, often at 0.5, instances are sorted into one of two categories. The model is simple and easy to understand. It can also include regularization techniques like Lasso (L1) or Ridge (L2), making it effective for dealing with noisy or small datasets often found in educational research. These benefits keep logistic regression a popular option for tasks such as predicting student performance and identifying at-risk students early [24], [27].

Logistic regression is popular in educational research because it can handle categorical variables and provide clear model parameters. This helps educators make decisions based on data. The model is also affordable to run, offers probabilities, and is flexible for both binary and multiclass situations. This makes it useful for predicting student engagement and timely submissions. However, it has some drawbacks. It can overfit, especially with small datasets, and it struggles to capture complex, nonlinear relationships found in educational data. While logistic regression is transparent, this can sometimes lower predictive accuracy compared to more complex models like deep learning or random forests. Additionally, problems with data quality and the model's limited assumptions may lessen its effectiveness in various educational settings [13], [25], [28].

III. METHODOLOGY

This study followed a structured approach to create and use a predictive model for student performance through Logistic Regression. The process included the following stages: data collection, data pre-processing, model development, model evaluation, and deployment. Each stage is explained in detail below.

A. Data Collection

The dataset was gathered from anonymous undergraduate students across various departments at Babcock University. We used an online questionnaire shared on student platforms to collect data. This questionnaire recorded academic details such as grades, cumulative performance, and attendance, as well as lifestyle-related factors like age and study habits. Including lifestyle factors was deliberate, as they can greatly affect students' academic results.

B. Data Pre-processing

Data pre-processing was done to ensure the quality and readiness of the data for model training. This stage included:

i. Data cleaning: Addressing missing values, fixing inconsistencies, and removing irrelevant attributes.
ii. Feature selection: Finding which variables are most relevant for predicting performance.
iii. Transformation: Encoding categorical variables and normalizing data when needed to improve model stability.

C. Data Splitting

To ensure a fair evaluation of the predictive model, the dataset was divided using the 75/25 ratio into two parts: a training set for fitting the model and a testing set for checking its performance. This split is crucial to how well a model generalizes to new, unseen data and avoid overfitting, which occurs when a model memorizes the training data instead of learning general patterns.

Table 1: Description of part of the dataset

#	COLUMN	TYPE	DESCRIPTION
0	Age	int	Student's Age
1	Gender	object	Student's Gender
2	Family income	object	Family's Income
3	Parent.educationl evel	object	Parent's Education Level
4	School	object	Student's Faculty
5	Level	int	Student's Level
6	100_level_cgpa	int	Student's CGPA at 100 level
7	Student's CGPA at 100 level	int	Student's CGPA at current level

D. Model Development

The Logistic Regression model was created using the Scikit-learn Python library. The training aimed to reduce prediction error by adjusting model parameters through repeated optimization. The logistic function and cost function, with a regularization term, guided this process to balance accuracy and generalization.

E. Model Evaluation

After training, the model was tested on the test set, with accuracy as the main performance measure. Additional evaluations looked at true positives, false positives, true negatives, and false negatives to give a full performance assessment. Functions from Scikit-learn were used to calculate these metrics and ensure reliable interpretation of the results.

F. Data Visualization

Exploratory data visualization was conducted using Seaborn in Python. Pair plots were created to show the relationships between academic and lifestyle features. These visualizations helped to identify correlations, trends, and potential predictor variables that could impact student performance.

G. Model Deployment

The final phase involved deploying the trained model with Django, a Python web framework. The Logistic Regression model was serialized using Pickle to allow it to be stored as a binary file and reused within the web interface. This integration lets users enter student attributes and receive real-time predictions about academic risks, which supports timely interventions.

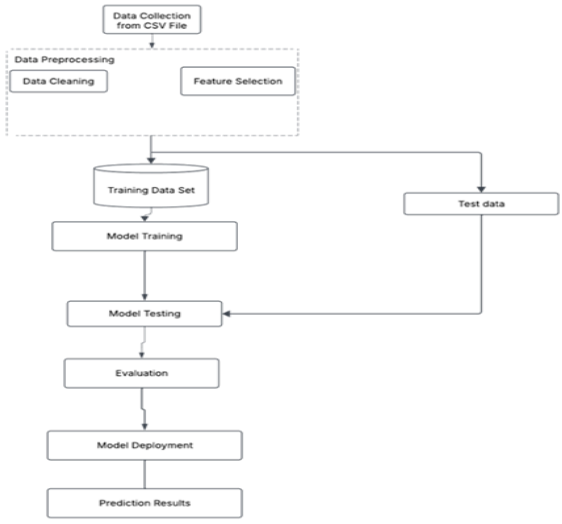

Figure 1: System Architecture

IV.RESULTS

Figures 2 and 3 presents the results of the model. Fig 2 shows the confusion matrix for the Logistics Regression algorithm, which describes the performance of the algorithm by summarizing the correct and incorrect predictions for each category. Fig. 3 demonstrates the feature importance tool which was added to provide insights on features which have significant impact on the prediction model. The model was evaluated to determine how well the model will perform on new, unseen data. The evaluation metrics used in this project is accuracy, precision, recall and F1-score. These metrics help evaluate the model by analyzing the number of true positives, false positives and false negatives. With the use of functions from Scikit-learn Python package, the metrics results are as follows:

i. Accuracy: 0.95
ii. F1 Score: 0.91
iii. Precison:0.96
iv. Recall: 0.89

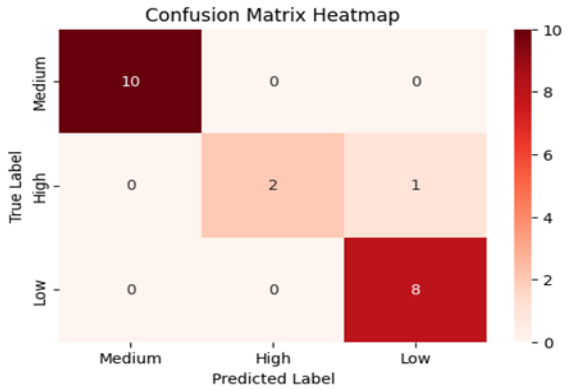

Figure 2: Confusion Matrix for Logistics Regression

Figure 3: Feature Importance for Logistics Regression

When compared to other classification algorithms such as Random Forest, Support Vector Machine and Gradient Boosting Regresso, Logistic Regression showed a better degree of accuracy of 95% making it the best choice model for predicting student performance. This makes the system a veritable tool for educators to not just recognize students in danger of failing but also understand patterns and variables that affect their performance.

V. CONCLUSION

This study successfully developed a Student Performance Prediction System that applies machine learning algorithms and techniques to predict student's grade based on previous academic performance, behavioural attributes, family relationship and demographic information. Logistic Regression was used to predict overall student performance based on selected features and the overall performance rating of students. It has a high accuracy of 95% making it a tool for educators in identifying students who are at danger of failing and providing them with required support to enhance academic support. The results demonstrated that machine learning algorithms can effectively classify student into different performance categories, helping educators and administrators make decisions.

REFERENCES

[1] S. Tete and O. M. Wizoma, "Education in Nigeria: Challenges and Way Forward," vol. 8, no. 1, pp. 42–48, 2020.

[2] D. Tolu-Kolawole, "1.4 million UTME candidates scored below 200 – JAMB," Punch, Apr. 29, 2024. Accessed: Sep. 6, 2025. [Online]. Available: https://punchng.com/1-4-million-utme-candidates-scored-below-200-jamb/.

[3] F. Ofori, E. Maina, and R. Gitonga, "Using machine learning algorithms to predict students' performance and improve learning outcome: A literature-based review," Journal of Information and Technology, vol. 4, no. 1, pp. 33–55, 2020.

[4] P. Sokkhey and T. Okazaki, "Developing web-based support systems for predicting poor-performing students using educational data mining techniques," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 7, 2020, doi: 10.14569/IJACSA.2020.0110704.

[5] A. F. Núñez-Naranjo, "Analysis of the determinant factors in university dropout: a case study of Ecuador," Frontiers in Education, vol. 9, Art. no. 1444534, 2024, doi: 10.3389/feduc.2024.1444534.

[6] R. Cariaga, "What is student performance?" J. Uniq. Crazy Ideas, vol. 1, no. 1, pp. 42–46, Aug. 2024, doi: 10.5281/zenodo.13410584.

[7] Y. Al Husaini and N. S. Ahmad Shukor, "Factors affecting students' academic performance: A review," Res Militaris: European Journal of Military Studies, vol. 12, pp. 284–294, 2023

[8] E. Tjandra, S. S. Kusumawardani, and R. Ferdiana, "Student performance prediction in higher education: A comprehensive review," AIP Conference Proceedings, vol. 2468, Art. no. 050005, 2022, doi: 10.1063/5.0080187.

[9] G. Feng, M. Fan, and Y. Chen, "Analysis and prediction of students' academic performance based on educational data mining," IEEE Access, vol. 10, pp. 19558–19571, 2022, doi: 10.1109/ACCESS.2022.3151652.

[10] R. H. Ali, "Educational data mining for predicting academic student performance using active classification," Iraqi Journal of Science, vol. 63, no. 9, pp. 3954–3965, 2022.

[11] G. Siemens and P. Long, "Penetrating the fog: Analytics in learning and education," EDUCAUSE Review, vol. 46, no. 5, pp. 30–40, 2011.

[12] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 40, no. 6, pp. 601–618, 2010.

[13] F. E. El Habti, M. Hiri, M. Chrayah, A. Bouzidi, and N. Aknin, "Enhancing student performance prediction in e-learning ecosystems using machine learning techniques," International Journal of Information and Education Technology, vol. 15, no. 2, pp. 301–311, 2025.

[14] M. Kumar, V. Bhardwaj, D. Thalkral, A. Rashid, and M. T. B. Othman, "Ensemble learning-based model for student performance prediction," Ingénierie des Systèmes

information, vol. 29, no. 5, pp. 1925–1935, 2024, doi: 10.18280/isi.290524.

[15] R. Bertolini, S. J. Finch, and R. H. Nehm, "Enhancing data pipelines for forecasting student performance: Integrating feature selection with cross-validation," International Journal of Educational Technology in Higher Education, vol. 18, no. 1, Art. no. 45, 2021, doi: 10.1186/s41239-021-00279-6.

[16] P. Chakrapani and D. Chitradevi, "Academic performance prediction using machine learning: A comprehensive & systematic review," in Proc. 2022 Int. Conf. Electron. Syst. Intell. Comput. (ICESIC), 2022, pp. 335–340, doi: 10.1109/ICESIC53714.2022.9783512.

[17] H. Goss, "Student learning outcomes assessment in higher education and in academic libraries: A review of the literature," J. Acad. Librariansh., vol. 48, no. 2, Art. no. 102485, 2022, doi: 10.1016/j.acalib.2021.102485.

[18] S. F. A. Hossain, Z. Xi, M. Nurunnabi, and B. Anwar, "Sustainable academic performance in higher education: A mixed method approach," Interact. Learn. Environ., vol. 30, no. 4, pp. 707–720, 2019, doi: 10.1080/10494820.2019.1680392.

[19] B. Sekeroglu, R. Abiyev, A. Ilhan, M. Arslan, and J. B. Idoko, "Systematic literature review on machine learning and student performance prediction: Critical gaps and possible remedies," Appl. Sci., vol. 11, no. 22, Art. no. 10907, 2021, doi: 10.3390/app112110907.

[20] H. Sahlaoui, E. A. A. Alaoui, A. Nayyar, S. Agoujil, and M. M. Jaber, "Predicting and interpreting student performance using ensemble models and Shapley additive explanations," IEEE Access, vol. 9, pp. 152688–152703, 2021, doi: 10.1109/ACCESS.2021.3124270.

[21] Y. Qu, F. Li, L. Li, X. Dou, and H. Wang, "Can we predict student performance based on tabular and textual data?" IEEE Access, vol. 10, pp. 86008–86019, 2022, doi: 10.1109/ACCESS.2022.3198682.

[22] F. Afrin, M. Hamilton, and C. Trevathan, "On the explanation of AI-based student success prediction," in Proc. Int. Conf. Comput. Sci., 2022, pp. 252–258, Doi: 10.1007/978-3-031-08754-7_34.

[23] P. Balaji, S. Alelyani, A. Qahmash, and M. Mohana, "Contributions of machine learning models towards student academic performance prediction: A systematic review," Appl. Sci., vol. 11, no. 21, Art. no. 10007, 2021, doi: 10.3390/app112110007.

[24] A. Kord, A. Aboelfetouh, and S. Shohieb, "Academic course planning recommendation and student performance prediction multi-modal based on educational data mining techniques," J. Comput. Higher Educ., 2025, doi: 10.1007/s12528-024-09426-0.

[25] H. Farhood, I. Joudah, A. Beheshti, and S. Muller, "Evaluating and enhancing artificial intelligence models for predicting student learning outcomes," Informatics, vol. 11, no. 3, Art. no. 46, pp. 1–17, 2024, doi: 10.3390/informatics11030046.

[26] M. Yağcı, "Educational data mining: Prediction of students' academic performance using machine learning algorithms," Smart Learn. Environ., vol. 9, no. 11, Art. no. 192, 2022, doi: 10.1186/s40561-022-00192-z.

[27] N. Alruwais and M. Zakariah, "Evaluating student knowledge assessment using machine learning techniques," Sustainability, vol. 15, no. 7, Art. no. 6229, pp. 1–25, 2023, doi: 10.3390/su15076229.

[28] A. Merchant, N. Shenoy, A. Bharali, and M. A. Kumar, "Predicting students' academic performance in virtual learning environment using machine learning," 2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T), 2022, pp. 1-6, doi: 10.1109/ICPC2T53885.2022.9777008.