A WEB-BASED MACHINE LEARNING MODEL FOR PREDICTING STUDENT ACADEMIC PERFORMANCE IN TERTIARY INSTITUTIONS

E.J. Anagu[1], R. Wella[2]

[1]Faculty of Computing and Information Systems, Federal University Wukari, Nigeria.

[2]Department of Information Technology, Taraba State Polytechnic Suntai, Nigeria.

Corresponding Author's Email: [1]anague@fuwukari.edu.ng

Article History: Received 26 March 2025; Revised 30 April 2025; Accepted 5 May 2025

ABSTRACT: Educational data mining plays a crucial role in analyzing student performance to identify those at risk and enhance academic success. Traditional statistical methods often fail to capture the complex factors influencing student achievement. This research presents a machine learning-based predictive system integrated into a web application to forecast student academic performance. The study utilizes a dataset from Taraba State University, comprising students from diverse demographics. The dataset undergoes preprocessing, feature selection, and modeling using three machine learning algorithms: Random Forest, Support Vector Machine (SVM), and Decision Tree. The evaluation results demonstrate that Random Forest achieved the highest accuracy (94%), followed by SVM (93%) and Decision Tree (92%). The developed web-based system allows educators to input student data and receive real-time performance predictions, facilitating early intervention strategies. The study highlights the potential of machine learning in educational decision-making and recommends further research on ensemble learning techniques for real-time academic performance prediction.

KEYWORDS: *Web-based system; Machine learning; Student academic performance; Real-time prediction; Educational decision-making*

1.0 INTRODUCTION

The success of tertiary education heavily depends on academic results which determines institutional ranking measures and student achievement rates. Educational institutions need to determine why students leave their studies to design appropriate solutions which enhance both learning quality and operational performance. The health state of educational systems depends on student retention rates thus educators need to activate solutions for their challenges [1].

Institutions use educational data mining to examine student information which helps them discover important patterns for their decision processes [2]. The application of machine learning by researchers now brings out latent academic data connections which surpasses the predictive abilities of traditional statistical procedures. The developed technology provides substantial opportunities to customize learning approaches while enabling schools to take better institutional choices [3].

Research shows machine learning techniques have become increasingly prevalent for addressing the problem of student educational outcomes prediction throughout the past few years. The advanced analytical tools examine large educational datasets to discover subtle connections which standard analysis techniques would miss. The analysis of intricate educational data by machine learning algorithms achieves effective student outcome predictions [4]. The researchers employed SVM and Decision Trees along with other algorithms as part of their investigation while matching the direction of our research on machine learning implementations. The evaluation method utilized to detect previously undetected academic aptitude and engagement markers. The research objective directly supports the identification of unknown variables which escape traditional measurement tools during student progress analysis.

Research conducted in Nigeria on educational data mining patterns focused on student performance prediction through selected data factors ([5], [6]. The research analyzed student performance through different variables but failed to create software which tracked educational advancement and course progress. The implementation of tailored software that tracks student performance in Nigerian higher education institutions became crucial due to the lack of suitable tools since data mining packages were plentiful. The study examined one university as its research case.

Traditional methods for predicting student academic performance, which often rely on demographic factors and standardized test scores, have limitations in capturing the full range of influences on student success. There is a need for more nuanced and data-driven approaches that can better account for the complex interplay of factors within specific educational settings. Machine learning offers a promising way to address this gap by enabling the analysis of large datasets to uncover hidden patterns and relationships that might improve prediction accuracy [7].

2.0 LITERATURE REVIEW

Machine Learning (ML) and Artificial Intelligence (AI) are research fields concerned with enabling computers to learn from experience. Some researchers consider ML as a subset of AI, given that the ability to learn is a fundamental characteristic of intelligent beings. The primary objective of ML is to create computer systems capable of learning from past observations and making informed decisions. AI, on the other hand, focuses on developing intelligent agents or assistants that leverage ML techniques to provide effective solutions [8].

ML techniques are generally classified into three main categories: In Supervised Learning (SL), an algorithm maps input data (I) to corresponding outputs (O) with the goal of accurately predicting outputs for new inputs [7]. SL can be further divided into: Classification: Used for categorizing data into predefined groups, such as genotypes. Regression: Focuses on predicting continuous values. Common SL techniques include K-Nearest Neighbors (KNN), Decision Trees (DT), Neural Networks (NN), Genetic Algorithms (GA), and Support Vector Machines (SVM). Unsupervised Learning (UL), only input data (I) are available, with no corresponding output labels. The system's goal is to identify the underlying patterns or distribution within the dataset [9]. UL is categorized into: Clustering: Identifies inherent groupings in the data. Association Rule Learning: Extracts meaningful relationships between different data points.

Reinforcement Learning (RL) refers to techniques where an agent improves its performance by interacting with the environment and optimizing its rewards over time. Unlike other ML approaches, RL does not have prior knowledge of the environment's behavior and relies on trial and error for learning [9]. This method is particularly valuable for

autonomous systems due to its adaptability to changing environments. Technological advances like the Internet of Things (IoT), artificial intelligence (AI), machine learning (ML), deep learning (DL), and big data have opened up avenues for study aimed at enhancing the educational experience for students and tackling issues the educational system faces. The ability to examine massive datasets and extract insightful information has revolutionized the prediction of academic achievement through machine learning technology, which analyzes data to identify patterns and generate predictions. Using a sizable dataset from an academic setting, [10], carried out a study that showed robust analysis produces high-quality insights for policy implementation.

In addition, [11] stressed how important it is to use contemporary data mining methods—like decision trees, naive bayes, and neural networks—instead of more conventional models to forecast student performance properly. Their research emphasizes how these models might improve instructional tactics and yield better results. The suggested algorithms that were applied. These methods are essential to educational data mining because they let heads of institutions and decision-makers maximize resources and create strategies and policies that work. Classification is the task that is most frequently used to predict student performance in predictive modeling, along with regression and categorization. For this, algorithms like artificial neural networks, naive Bayes, decision trees, and support vector machines are commonly used.

This study explores ways to enhance university admissions decisions using data mining techniques for predicting applicants' academic success. It validates the proposed methodology using a dataset of 2,039 students enrolled at a Saudi public university's Computer Science and Information College from 2016 to 2019. The findings reveal that early university performance can be forecasted before admission based on specific pre-admission criteria, including high school grades, Scholastic Achievement Admission Test scores, and General Aptitude Test scores[12]. Notably, the Scholastic Achievement Admission Test score is the most accurate predictor, suggesting it should carry more weight in admissions systems. Furthermore, the study identifies the Artificial Neural Network as the superior technique with an accuracy rate exceeding 79% compared to other considered methods like Decision Trees, Support Vector Machines, and Naïve Bayes.

Also, a novel prediction algorithm for assessing student academic performance, utilizing both classification and clustering techniques. The

algorithm is tested in realtime using student datasets from diverse academic disciplines in Kerala, India. The findings demonstrate that the hybrid algorithm, which integrates clustering and classification approaches, outperforms others in accurately predicting student academic performance [13].

|This research employed a deep artificial neural network and handcrafted features derived from virtual learning environments' clickstream data to predict students at risk and enable early intervention measures. The model achieves a classification accuracy ranging from 84 to 93%, surpassing logistic regression (79.82–85.60%) and support vector machine models (79.95–89.14%). Including legacy and assessment related data significantly impacts the model's performance, and students engaging with previous lecture content exhibit improved outcomes [14].

3.0 RESEARCH GAP

Traditional methods for predicting student academic performance in tertiary institutions, such as those relying on demographic factors (e.g., age, gender) and standardized test scores, often fail to capture the complex, non-linear interactions among variables influencing student success [4]. These methods, including basic statistical models like linear regression, are limited by their inability to handle high-dimensional datasets or uncover latent patterns that affect academic outcomes, particularly in diverse educational settings like Nigeria [7]. While prior studies in Nigeria have explored educational data mining for performance prediction ([5], [6]), they primarily focused on analyzing static variables without developing scalable, user-friendly tools for real-time application. Moreover, existing research often overlooks the integration of predictive models into accessible platforms that educators can use for timely interventions.

This study addresses these gaps by proposing a web-based machine learning system that leverages advanced algorithms Random Forest, Support Vector Machine (SVM), and Decision Tree to model complex relationships within a comprehensive dataset from Taraba State University

4.0 SYSTEM METHODOLOGY

The research adopted a quasi-experimental design to investigate the effectiveness Random Forest, Support Vector Machine (SVM) and Decision Tree models in classifying students' academic performance based on the dataset of Taraba state University. The framework consists of several stages: data acquisition and preparation, model development, training and evaluation. Cross validations strategies were employed to ensure the robustness of the model and mitigates the risks of overfitting. The aforementioned phases of the research methodology are shown in Figure 1

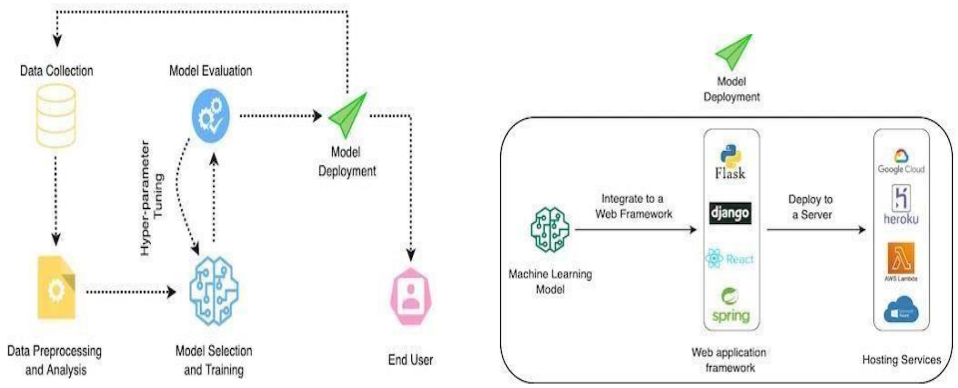

Figure *1*: Machine Learning Model Integration to a Web Application

4.1 Data Preprocessing and Normalization

Data preprocessing and normalization significantly enhance machine-learning model effectiveness. Categorical variables, such as Gender and Marital Status, were encoded using label encoding to facilitate machine learning algorithms. The researchers chose attributes to predict student academic performance using criteria of theoretical importance, true-to-life evidence and relevant application in Nigerian tertiary education. Among the important predictors discovered are Age, Gender, Marital Status and First-Year Cumulative Grade Point Average (CGPA), because they have a strong impact on student results in educational data mining studies. A person's age when they begin their studies matters a lot, as it affects their studies due to changes in maturity, motivation and what they have done before enrolling. Younger ones seem to handle academic matters well but could be less disciplined than older peers who usually focus better on what

they want to do in the future. Gender is taken into account due to its links to differences in how students participate in academic work, meet social demands and perform in various fields.

Continuous numerical features, including Age and First-Year CGPA, were standardized through feature scaling to maintain consistency. Outliers were identified using statistical techniques and were either transformed or removed to minimize their impact on the dataset. This research employs the MinMaxScaler function from Scikit Learn for normalization. This method is beneficial as it boosts algorithm performance. The dataset contains varied measuring units, causing a disparity that can negatively affect model performance, as high-value features may overshadow lower ones. The dataset was split into training (70%) and testing (30%) sets to effectively evaluate the model's performance.

To evaluate the performance of the adopted models. This study proposed some standard evaluation metrics such as accuracy, precision, recall, and F1-score Model deployment refers to the process of integrating the developed machine learning model into a system or application that allows users—such as educators or administrators—to interact with it. This research deploys a predictive system on a web-based interface, enabling users to input relevant student data and receive predictions regarding their academic performance. This would allow educational stakeholders to make informed decisions based on real-time predictions, helping identify students at risk and providing timely interventions. The deployment process ensures that the model is accessible, reliable, and compatible with various platforms, enhancing its usability in real-world educational environments. It also involves continuous monitoring and updates to ensure that the model adapts to new data and remains accurate over time. This step ensures that the model is not only theoretically sound but also practical for real-world use, providing educational institutions with a tool that can be accessed from any web browser platform.

5.0 MATERIALS AND METHODS

This research adopts a quasi-experimental design to evaluate the effectiveness of Random Forest, SVM, and Decision Tree models in

classifying student academic performance. The methodology includes data acquisition, preprocessing, model development, training, and evaluation. Cross-validation techniques ensure model robustness and mitigate overfitting risks.

The dataset utilized in this study comprises primary data extracted from Taraba State University's student database. The dataset includes both demographic and academic attributes, specifically focusing on students' Cumulative Grade Point Averages (CGPA) from their first to fourth year of study, along with their final overall CGPA. It spans 16 academic programs across the Faculty of Agriculture, Faculty of Social & Management Sciences, Faculty of Arts, and Faculty of Science, covering a five-year period (2015–2019) with a total of 3,046 students. Data preprocessing involved encoding categorical variables and normalizing numerical features to enhance model performance. The dataset was split into training (70%) and testing (30%) subsets. Feature selection focused on key academic indicators, including first-year CGPA, age, gender, and marital status. To evaluate model performance, this study employs standard metrics such as accuracy, precision, recall, and F1-score. The final predictive model is deployed as a web-based application, allowing educators to input student data and receive real-time academic performance predictions. This system provides actionable insights for early intervention and student support.

A 64-bit Windows Operating System, with an Intel(R) Corel (TM) i5-3630QM CPU @2.40GHZ with 4.00 GB of RAM was used. The programming environment utilized for implementing the program code was the Anaconda environment using the Python 3.11 software development kit as the programming language. The application programming interface utilized was Sklearn API with some other Python dependencies such as NumPy for vector operations, pandas for reading files, TensorFlow API (Application Programming Interface), and the Mat-plot library for data visualization operations. The deployment of the model on a web-based interface also relies on Python-based frameworks to ensure seamless interaction between users and the backend prediction system.

6.0 RESULTS AND DISCUSSIONS

There are various machine learning models used for different classification tasks. This research compares the effectiveness of several

models, including Random Forest, Support Vector Machine (SVM) and Decision Tree. The preprocessed data is split into two sets: the training dataset and the testing dataset, with a 10-fold cross-validation.
Feature selection in data mining helps models by looking out for the most important features in the dataset, reducing complexities and increasing the accuracy of the model. The target variable for prediction in this study was the Graduating CGPA, while four independent attributes were selected based on their relevance to academic performance: Age (student's age at admission), Gender (male or female), Marital Status (single, married, or other categories), and First-Year CGPA (academic performance at the end of the first academic year). These attributes were chosen due to their potential impact on student outcomes, as factors such as age, gender, and marital status may influence academic performance, while First-Year CGPA serves as an early indicator of future success. Table 1 presents the class attribute and their variable codes, and the corresponding values.

Table 1: Class attributes

variable codes	Corresponding values
ID No	Randomly generated number sequence
Prog Code	Program of Study
Gender	Gender
YoG	Year of Graduation
MS	Marital status
CGPA100	Cumulative Grade Point Average at the end of the first year
CGPA200	Cumulative Grade Point Average at the end of the second year
CGPA300	Cumulative Grade Point Average at the end of the third year
CGPA400	Cumulative Grade Point Average at the end of the fourth year

7.0 MODEL DEPLOYMENT

The deployment of the machine learning model as a web-based application enables educators and administrators to interact with the predictive system in real-time, facilitating timely interventions for at-risk students. The system, built using Python-based frameworks (e.g., Flask or Django) and the Scikit-learn API, allows users to input student data through a user-friendly interface and receive predictions of academic performance (e.g., Graduating CGPA or risk level). The web application is hosted on a server running a Python-based framework, leveraging libraries such as Flask for the backend,

HTML/CSS/JavaScript for the frontend, and Scikit-learn for model inference. The system ensures data security through encrypted connections (HTTPS) and user authentication to protect sensitive student information. Continuous monitoring is implemented to track model performance, with automated alerts for significant accuracy drops, prompting retraining with updated datasets. This deployment strategy ensures the system is practical, scalable, and accessible for educators in Nigerian tertiary institutions, enabling data-driven decision-making for student support.

The deployment process ensures accessibility across web browsers, compatibility with various devices, and continuous monitoring to maintain model accuracy as new data becomes available. This section describes the user interaction workflow and presents a flowchart to illustrate the process.

7.1 User Interface Workflow

1. Accessing the Application: Educators access the web application via a secure URL on any standard web browser (e.g., Chrome, Firefox) or compatible mobile device, requiring no specialized software installation.
2. Data Input: Users input student data through a form-based interface, including attributes such as Age, Gender, Marital Status, and First-Year CGPA. The interface validates inputs to ensure completeness and correct formatting (e.g., numerical values for Age and CGPA, categorical selections for Gender and Marital Status).
3. Prediction Processing: Upon submission, the system preprocesses the input data (e.g., encoding categorical variables, normalizing numerical features) and feeds it into the trained Random Forest model, which achieved the highest accuracy (94%) in our evaluation.
4. Output Display: The system returns a prediction, such as the estimated Graduating CGPA or a classification of the student's risk level (e.g., "At Risk," "Satisfactory," "High Performing"). Results are displayed with interpretable visualizations (e.g., a performance score or probability chart) to aid decision-making.
5. Actionable Insights: Educators receive recommendations based on the prediction, such as suggesting academic counseling or tutoring for at-risk students. The system logs interactions for monitoring and future model updates.

6. Continuous Updates: The application supports periodic retraining with new student data to maintain accuracy, with updates managed through a backend administrative interface.

8.0 CONCLUSION

The research demonstrates a machine learning strategy for tertiary institution student academic performance prediction. Educational data and machine learning models allow the study to deliver enhanced predictive power and immediate analysis of student performance results. The developed web-based application serves educational establishments by helping them detect potentially at-risk students after connecting them to appropriate interventions. Additional future research needs to examine more sophisticated ensemble learning methods and deep learning algorithms for improving prediction accuracy. Implementing a standardized prediction framework in Nigerian higher education institutions requires expanding the database to include participating institutions.

9.0 RECOMMENDATIONS FOR FURTHER STUDY

To enhance the predictive accuracy and practical utility of the web-based machine learning system for student academic performance, the following recommendations for further study and future research are proposed:

1. Future work could use methods such as Gradient Boosting and stacking to mix the features of Random Forest, SVM and Decision Tree which should enhance how stable and accurate the models are.

2. Additional studies are needed to assess RNNs or LSTM models further may support the analysis of repeated sequences in the history of CGPA to spot ongoing trends.

3. Future research could use student logs (library apps or online learning software) to observe how other influences can affect performance. In future work, focusing on feature importance will help determine what variables are most important.

4. More institutions should be added to the dataset for the research to be more easily applied to different parts of Nigeria. So far, models have

yet to be built for flow behavior using data-sharing protocols and further research is needed.

In the future, research should enable the web version to upload data automatically, use interactive dashboards and be usable on mobile devices. More investigations are needed to gather educator opinions to make the platform more usable and secure.

ACKNOWLEDGMENTS

The authors sincerely appreciate Taraba State University for providing the dataset used in this study. Their support and commitment to educational research greatly contributed to the successful development of this web-based machine learning model for predicting student academic performance.

REFERENCES

[1] A. A. Saa, M. Al-Emran, and K. Shaalan, "Mining student information system records to predict students' academic performance," in Proc. Int. Conf. Adv. Mach. Learn., 2020.

[2] M. V. Martins, L. Baptista, J. Machado, and V. Realinho, "Multi-class phased prediction of academic performance and dropout in higher education," Appl. Sci., vol. 13, p. 4702, 2023.

[3] H. A. Mengash, "Using data mining techniques to predict student performance to support decision making in university admission systems," IEEE Access, vol. 8, pp. 55462-55470, 2020.

[4] A. M. Adeyemi and S. B. Adeyemi, "Institutional factors as predictors of students' academic achievement in colleges of education in South Western Nigeria," Int. J., 2014.

[5] V. E. Adeyemo, A. Abdullah, N. Z. JhanJhi, M. Supramaniam, and A. O. Balogun, "Ensemble and deep-learning methods for two-class and multi-attack anomaly intrusion detection: An empirical study," Int. J. Adv. Comput. Sci. Appl., vol. 10, 2019.

[6] S. A. Oyebade and C. Dike, "Restructuring Nigerian tertiary (university) education for better performance," presented at the 11th Annu. Meeting Bulg. Comp. Educ. Soc., Plovdiv, Bulgaria, 2013.

[7] K. Taherkhani, C. Eischer, and E. Toyserkani, "An unsupervised machine learning algorithm for in-situ defect-detection in laser powder-bed fusion," J. Manuf. Process., vol. 81, pp. 476-489, 2022.

[8] D. Krotov and J. J. Hopfield, "Unsupervised learning by competing hidden units," Proc. Natl. Acad. Sci., vol. 116, pp. 7723–7731, 2019.

[9] D. Silver et al., "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play," Science, vol. 362, pp. 1140–1144, 2018.

[10] O. J. Abiodun and A. I. Wreford, "Students' performance evaluation using ensemble machine learning algorithms," Eng. Technol. J., vol. 9, 2024, doi: 10.47191/etj/v9i08.23.

[11] K. Alalawi, R. Athauda, and R. Chiong, "Contextualizing the current state of research on the

use of machine learning for student performance prediction: A systematic literature review," Eng. Rep., vol. 5, p. e12699, 2023.

[12] W. Zhang, Y. Wang, and S. Wang, "Predicting academic performance using tree-based machine learning models: A case study of bachelor students in an engineering department in China," Educ. Inf. Technol., vol. 27, pp. 13051-13066, 2022.

[13] B. K. Francis and S. S. Babu, "Predicting academic performance of students using a hybrid data mining approach," J. Med. Syst., vol. 43, p. 162, 2019.

[14] H. Waheed et al., "Predicting academic performance of students from VLE big data using deep learning models," Comput. Hum. Behav., vol. 104, p. 106189, 2020.