ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep648

Comparative Analysis of Predictive Model for Student Academic Performance

Gabriel Adaugo Cynthia[1]; Prema Kirubakaran[2]; Ridwan Kolapo[3]

[1,2,3]Information Technology Department Nile University of Nigeria

Publication Date: 2026/10/17

Abstract: Early detection of students who are likely to have low performance levels is a major issue in learning institutions. This paper aims at creating a predictive model that can be depended on as well as interpreted to deal with this problem using machine learning. Another strength of this research is that it focuses on accuracy, among other factors, which is unlike many other studies that focus on accuracy only but not on model robustness and contextual relevance, which are essential to meaningful educational decision-making. To enhance the credibility of the findings, the primary and secondary data were also used, including demographic, socio-educational, and academic characteristics of students. This two-dataset methodology made it possible to evaluate and generalise better. The methodology referred to five supervised machine learning models, including Logistic Regression, Decision Tree, Random Forest, eXtreme Gradient Boosting (XGBoost), and Artificial Neural Networks (ANN). The models were assessed and compared in terms of standard performance measures. The findings indicate that the most appropriate and the most effective model in this study was logistic regression with an optimal performance on all the evaluation measures. XGBoost, ANN, the Decision Tree and the Random Forest models showed good predictive power and were competitive. The originality of the proposed research is in the use of both primary and secondary data, a focus on the interpretability of the models in addition to their accuracy, and the systematic comparison of the traditional and advanced methods of machine learning. This study refutes the widely held belief that more complex models are always better at educational prediction tasks by showing that a transparent model like logistic regression can be better than other more complex models when the educational data is well-prepared. In practice, this study offers a data-driven, easy-to-scale, and simple tool. To the students, it allows identifying the ones who are at risk at an earlier age and thus being able to intervene with them; to the teachers and institutions, it provides a powerful support system to track the learning outcomes, tailor teaching methods and increase the overall quality of education.

Keywords: *Student Academic Performance, Machine Learning, Educational Data Mining, Learning Analytics.*

How to Cite: Gabriel Adaugo Cynthia; Prema Kirubakaran; Ridwan Kolapo (2026) Comparative Analysis of Predictive Model for Student Academic Performance. *International Journal of Innovative Science and Research Technology*, 11(9), 3457-3463. https://doi.org/10.38124/ijisrt/26sep648

I. INTRODUCTION

The modern educational landscape is changing very fast with the introduction of data analytic technologies and artificial intelligence (AI) technologies into the teaching and learning systems. The introduction of Educational Data Mining (EDM) and Learning Analytics (LA) is at the stage of maturity, and it should be regarded as the main design paradigm in contemporary educational research. These paradigms use computational methods on learning data to create practical insights on how a student behaves and performs and how well the institution is functioning (Romero & Ventura, 2020). Although EDM and LA are different disciplines, they have a lot of overlap.

Educational data mining is more concerned with algorithmically discovering facts based on the data, whereas learning analytics is more involved in making sense of the data that is used in pedagogical decisions (Lampropoulos, 2023).

The forms of evaluation currently utilised by educational institutions around the world are considered to be mostly based on the schemes which monitor student progress manually, rely on description, and utilise retrospective schemes. The systems that have been in use are heavy on summative assessments that is, the final examinations only when the learning period is over. This means that teachers are frequently notified about the poor academic performance of a student once it is too late to take appropriate action (Christou et al., 2023). This brings up a fundamental inefficiency: as in the present state of the higher education industry, the volumes of data being generated on a daily basis are monumental; there is a gaping void between the data that is being produced and how it can be used strategically and proactively to predict and enhance student performance. The student outcomes prediction by machine learning is a practice which has experienced significant improvement in the recent years but also has several significant methodological shortcomings which impede application into practice. The main problem is that the majority of the current research implies using single

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep648

predictive models, i.e., simple decision trees, logistic regression or naive Bayes classifiers. These standalone models are, however, very interpretable but often fail to achieve high-accuracy plateaus and have low generalisability when used on new data (Albreiki et al., 2021).

II. LITERATURE REVIEW

Elements of modern tendencies in the period of 2020-25 have been used to encrust machine learning strategies into the sphere of educational data mining (EDM) to predict and improve student learning performance. Other research works point to a clear change in the direction of the field by doing away with the classical means of statistical tools and using more modern ensemble and deep learning predictive models. The previous studies, including those of Abdul Bujang et al. (2021) and Razali et al. (2022), established that individual Decision Trees (DT) and Random Forest (RF) classifiers have a fixed high accuracy of over 90 per cent. However, these studies were often based on smaller-thanappropriate databases that are highly homogeneous and cannot easily be externally validated, causing serious concerns about how well the findings can be generalised. Equally, Ojajuni et al. (2021) and Zoralioglu et al. (2023) both claimed that ensemble modelling methods, including XGBoost and Random Forest, are competent enough in the delivery of highly precise prediction outcomes when they are applied to localised data.

In order to incorporate this growing body of empirical evidence, a number of new studies have been performed through systematic reviews and meta-analyses. As an example, Al-Din and Al Abdulqader (2024) and Wang et al. (2024) conducted a review of over eighty studies and found that ensemblebased methods such as Gradient Boosting and Random Forest perform better than their more recent counterparts, with averages of accuracy more than 85. However, these reviews continually indicate three primary persistent problems throughout the literature: extreme data imbalance, weak feature selection, and clear absence of explainable AI (XAI) models.

More so, those studies that incorporated both behavioural and socio-economic ones, as covered by Hoti et al. (2025) and Ibrahim et al. (2024), are actually setting the direction on a wider perspective of academic analytics, in which the non-academic factors that essentially predetermine student success make their mark.

The latest trials have also narrowed these strategies down to particular institutions and areas. As an example, the study conducted by Anagu and Wella (2025) and Edmond, Sada, and Osijirin (2025) provided the conclusive evidence of the applicability of these techniques to Nigerian universities, showing a rather high level of predictive validity of the applied techniques, such as Random Forest and Gradient Boosting models. Nevertheless, both papers observed that data heterogeneity, interpretability and scalability are a major challenge. Similarly, the trend that Vamsi and Bidikar (2025) and Bello, Ajibade, and Wonuola (2025) described involved web-based deployments based on advanced techniques of regression, and this is the future of real-time prediction platforms

III. METHODOLOGY

In this study, data sources utilised in this paper were based on the publicly available Kaggle Educational Data Mining repository that provides some demographic and academic performance related data of 1,000 students in secondary or high school. These are gender, race/ethnicity, parental education level, type of lunch, whether or not they cite a test preparation course and eventual test scores in mathematics, reading and writing.

Table 1 Description of Dataset Variables

Variable	Description
Race/Ethnicity	Ethnic group classification
Parental Level of Education	Highest educational level of parents
Lunch	Lunch type (standard or free/reduced)
Test Preparation Course	Completion status of test preparation course
Math Score	Mathematics performance score
Reading Score	Reading performance score
Writing Score	Writing performance score

Table 1 shows the description of the dataset variables (as the demographic and socio-educational variables are described, followed by the measures of academic performance). This is race/ethnicity, which categorizes the students by race. Parent education is the best educational qualification of the parent, which in most cases correlates with student achievement. The students who pay for their lunch and the students receiving free/reduced lunch are separated by lunch type; thus, it is used as a proxy of socioeconomic status, but the students who have undergone preparation courses are indicated through a test preparation course. Such academic characteristics are math score, reading score, and writing score, which represent performance in each of the subjects. This architecture is similar to traditional datasets that have been employed in education data mining as it is elaborated

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep648

Table 2 Objective vs Methodology Matching

Objectives	Proposed Methodology	Expected Results
Objective (1)	Data collection and preprocessing	Clean, reliable dataset
Objective (2)	Model design and training (Linear Regression, Decision Tree,Random Forest, XGBoost, ANN.)	Predictive model for dropout
Objective (3)	Model evaluation and comparison	Best-performing identified model

Table 2 provides a structured mapping of the research objectives to their respective methodologies and expected results. The progression from data collection and pre-processing to model development and finally to evaluation and comparison, reflects a systematic approach designed to produce a robust predictive model and identify the most effective technique.

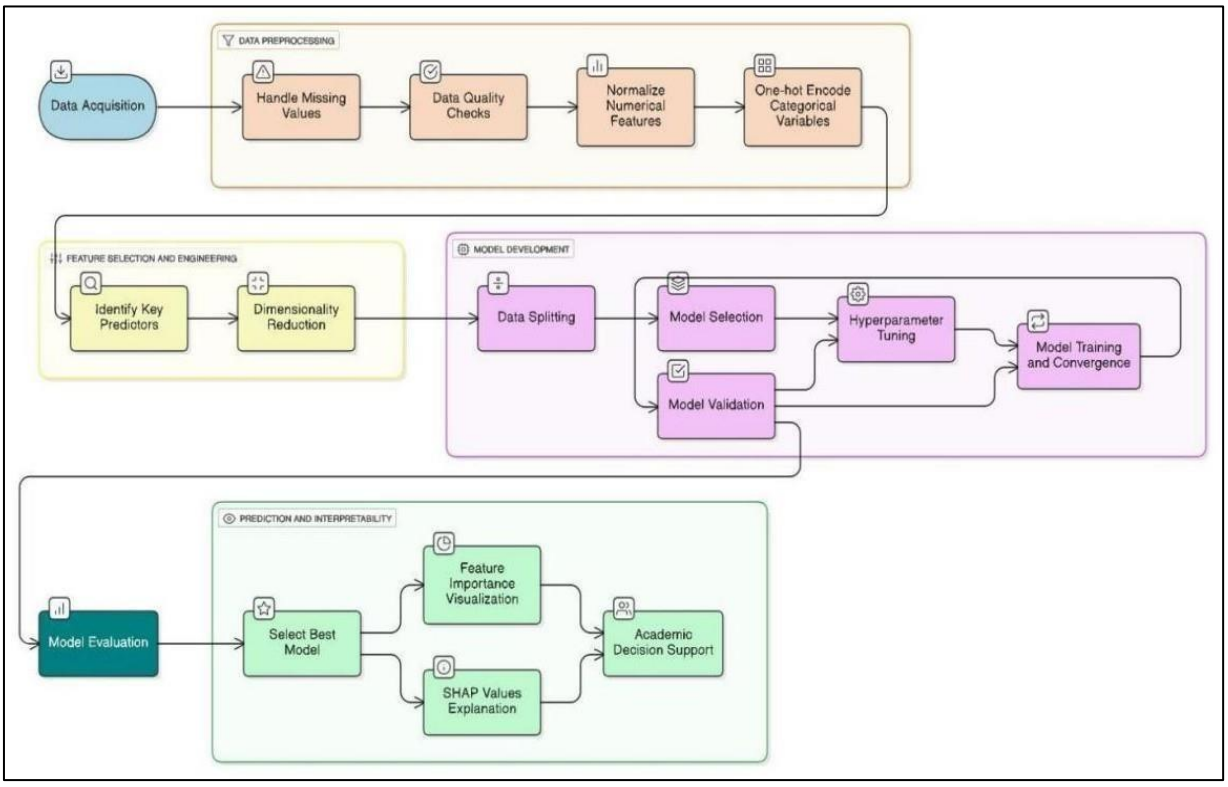


Fig 1 Academic Predictive Modelling Workflow

Figure 1 illustrates the overall research workflow adopted in this study. The process begins with dataset acquisition and pre-processing, followed by feature engineering and data transformation. The dataset is then split into training and testing sets, after which multiple machine learning algorithms, including Logistic Regression, Decision Tree, Random Forest and ANN, are applied. Finally, model evaluation is conducted to determine the best-performing predictive model. As a measure of the effectiveness of the predictive model, four conventional performance measures using a confusion matrix have been used. These indicators provide the comprehensive picture of the predictive model of student academic achievement, which makes it a fair comparison of machine learning algorithms:

- Accuracy: Measures the share of student performance outcomes that were correctly predicted.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- Precision: Evaluates the number of true positive cases among all positive predictions.

$$Precision = \frac{TP}{TP + FP}$$

- Recall (Sensitivity): Measures the share of actual positive cases that have been correctly identified.

$$Recall = \frac{TP}{TP + FN}$$

- F1-Score: The harmonic mean of precision and recall, balancing both metrics.

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

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep648

IV. RESULTS AND DISCUSSIONS

Accuracy, precision, recall, and F1-score are some of the indicators shown in Table 3 to summarise the performance of the five machine learning models under evaluation for predicting student academic performance. In terms of accuracy, the logistic regression model scored perfectly, with a value of 1.0000. The model's precision, recall, and F1 scores were also all equal to 1.0000. In contrast, the other model, Decision Tree, with an accuracy of 0.9700, precision of 0.9703, recall of 0.9700, and F1-score of 0.9701, came in second just after Logistic Regression. According to Random Forest records, accuracy 0.9650, precision 0.9666, recall 0.9650, F1-score 0.9651. The XGBoost model gives impressive results with an accuracy of 0.9950, precision of 0.9951, recall of 0.9950, and an F1-score of 0.9950. ANN model's performance in terms of precision, recall, and F1-score is 0.9902, 0.9900, 0.9900, respectively with accuracybeing 0.9900. Almost perfect results from the logistic regression model and XGBoost suggest that the models fit the data well; however, perfection in logistic regression might be indicative of some component of overfitting or high separability in the pre-processed features consistent with observations in highly structured educational datasets.

Table 3 Summary of Model Performance

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	1.0000	1.0000	1.0000	1.0000
Decision Tree	0.9700	0.9703	0.9700	0.9701
Random Forest	0.9650	0.9666	0.9650	0.9651
XGBoost	0.9950	0.9951	0.9950	0.9950
ANN	0.9900	0.9902	0.9900	0.9900

The confusion matrix of the three performance classes, which are High, Low, and Medium, was obtained with the logistic regression model and is displayed in Figure 2. The matrix obtains that 86 high cases were well classified. Misclassification of low or medium cases as high cases was 0, but 3 cases of medium ones were falsely forecasted as high cases. In the case of the low class, it had 22 true positives and zero misclassifications. The medium class also had 89 true positives where there were not any false positives of other classes.

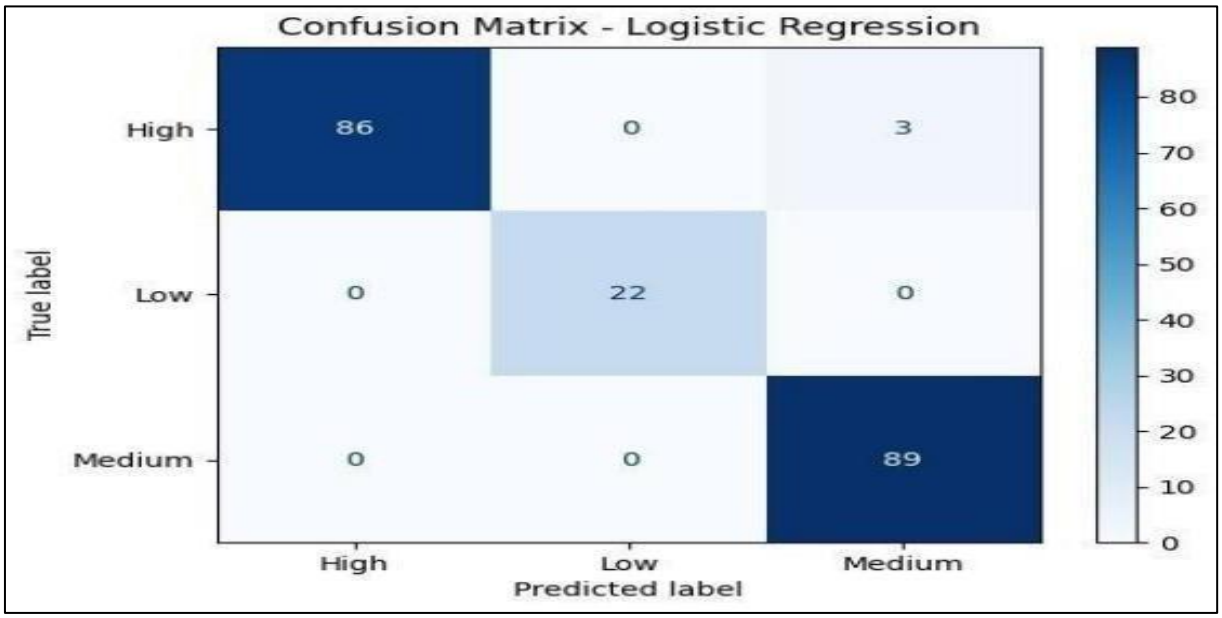


Fig 2 Logistic Regression - Confusion Matrix

Figure 2 describes the confusion matrix for the decision tree model. The High category has it right for 86 positive cases when it counts true positives. Besides that 3 times Medium category has been wrongly identified as High while there are 0 Low instances that have been wrongly identified as High. Regarding the low class, it has 21 true positives, 1 medium has beenmisclassified as low and 0 from high. Medium has 87 true positives and 2 low samples have been misclassified as medium. These small mistakes are mere indications of slight overfitting or splitting problems that are very usual with a single-tree model

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep648

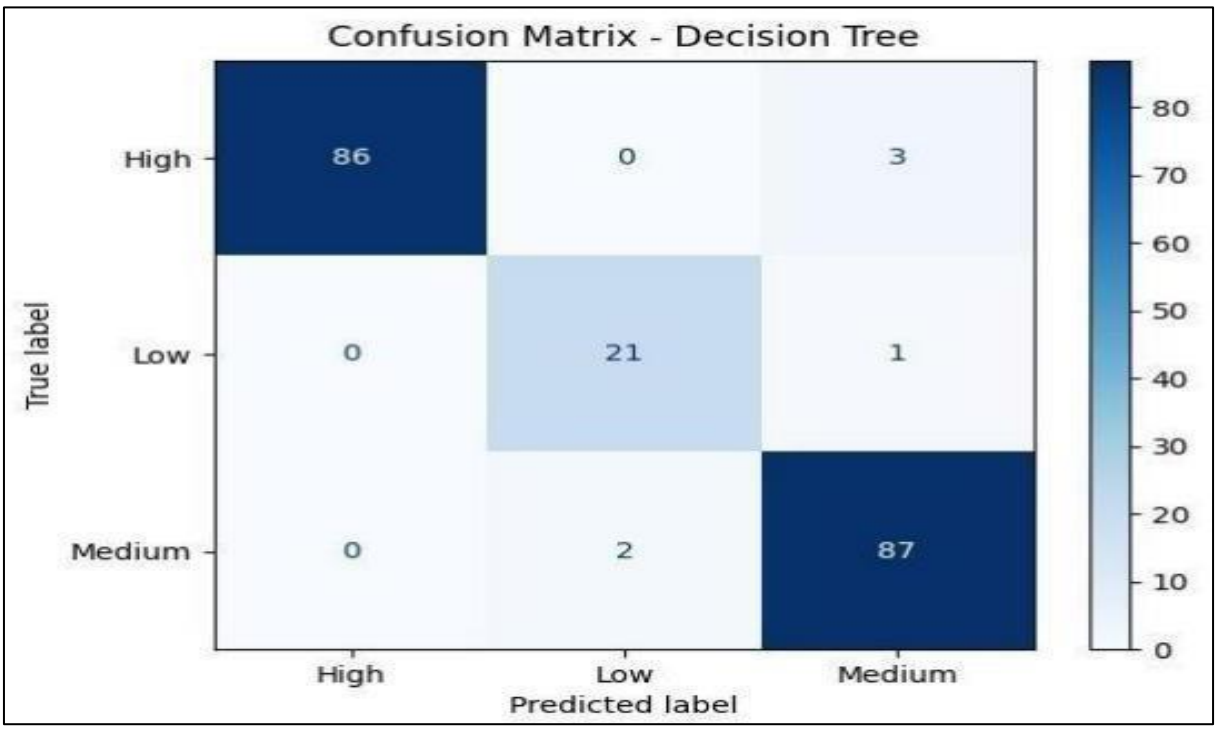


Fig 3 Decision Tree - Confusion Matrix

Figure 3 shows the confusion matrix for the Random Forest model. High has 84 truly positive cases, and 5 medium cases are incorrectly classified as high. Low: 21 truly positive cases and 1 medium wrongly classified as low. Medium: 88 truly positive cases and 1 low wrongly classified as medium. Increased misclassifications compared to boosting methods show the obvious variance reduction benefits of ensembles, while remaining errors reflect the much-documented difficulty in minority class boundaries.

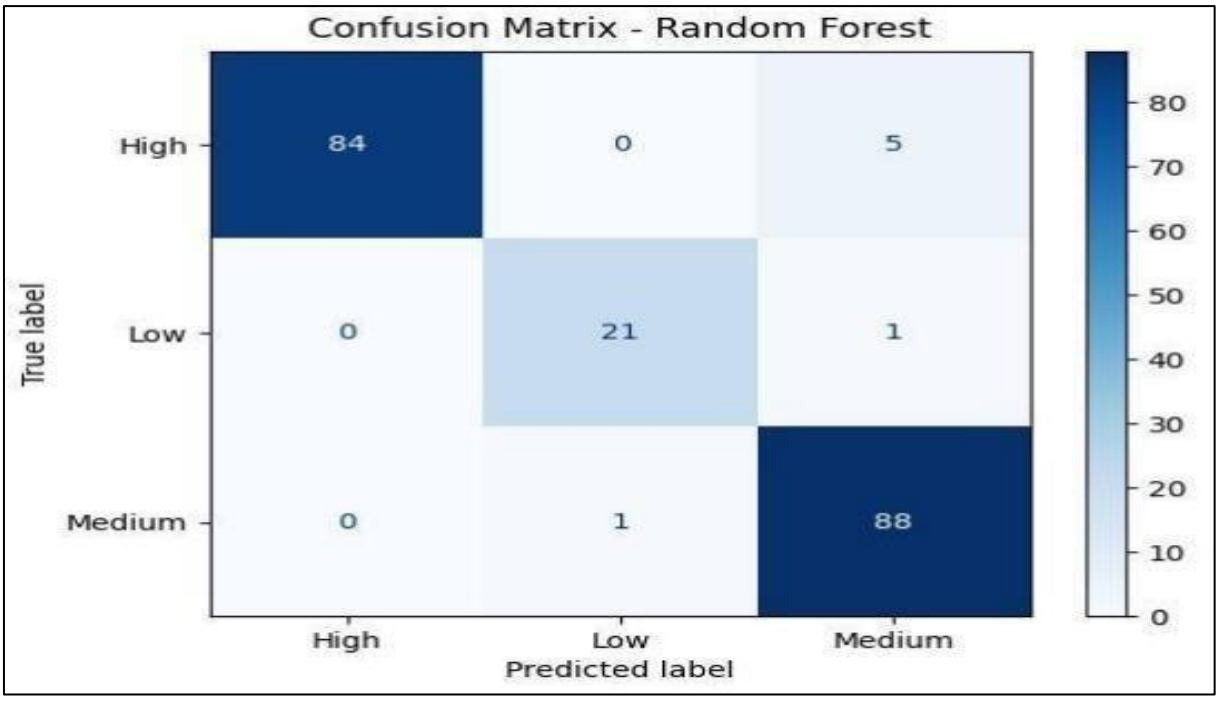


Fig 4 Random Forest - Confusion Matrix

Figure 4 gives confusion result matrices for an XGBoost model, out of which 88 true positives belong to High with only 1 Medium misclassified as High; this figure also presents 22 true positives for Low with no errors of misclassification, and, lastly, 89 true positives for Medium with no errors at all. Such matrices have the fewest off-diagonal entries when compared to tree- based ones, showcasing the superiority of this boosting technique in terms of minimising classification errors.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep648

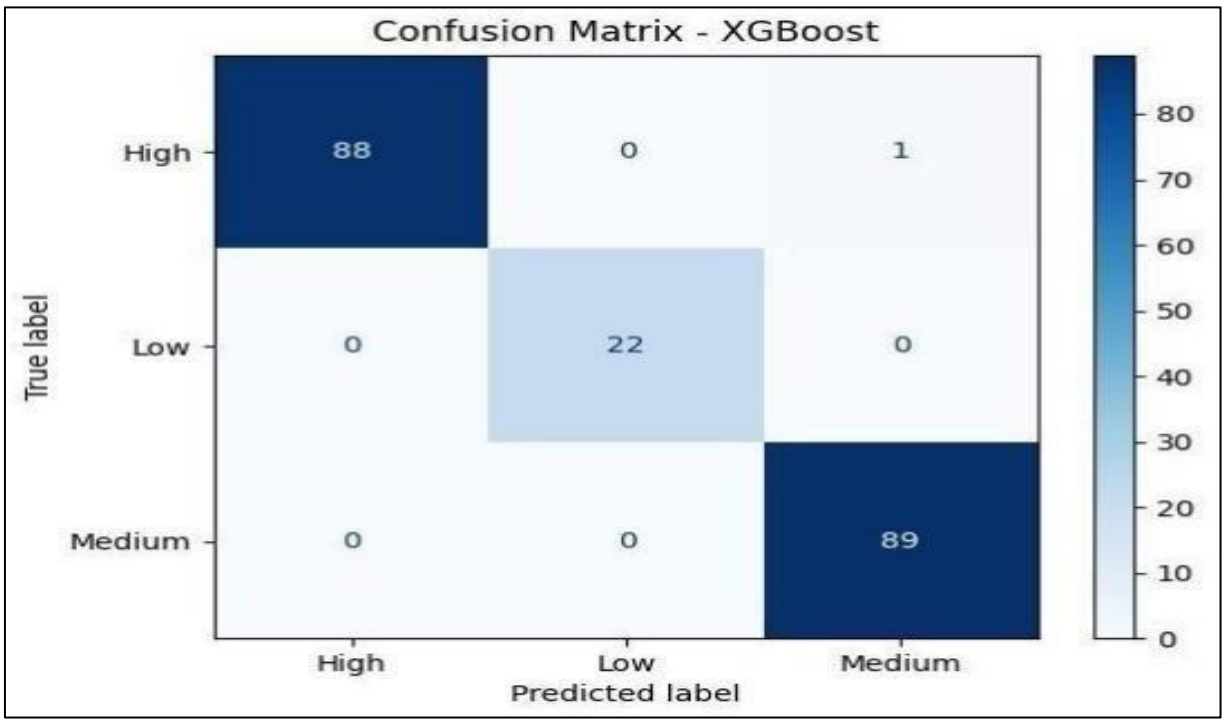


Fig 5 XGBoost - Confusion Matrix

Figure 5 presents the confusion matrix for the Artificial Neural Network (ANN) model; 86 true positives are for High, and 3 Medium are misclassified as High; 22 true positives for Low have no errors from others, and 88 true positives for Medium have one Low misclassified as Medium. This pattern is very much like that of logistic regression, which shows the model's great potential for non-linear learning, but also hints at some space for further fine-tuning

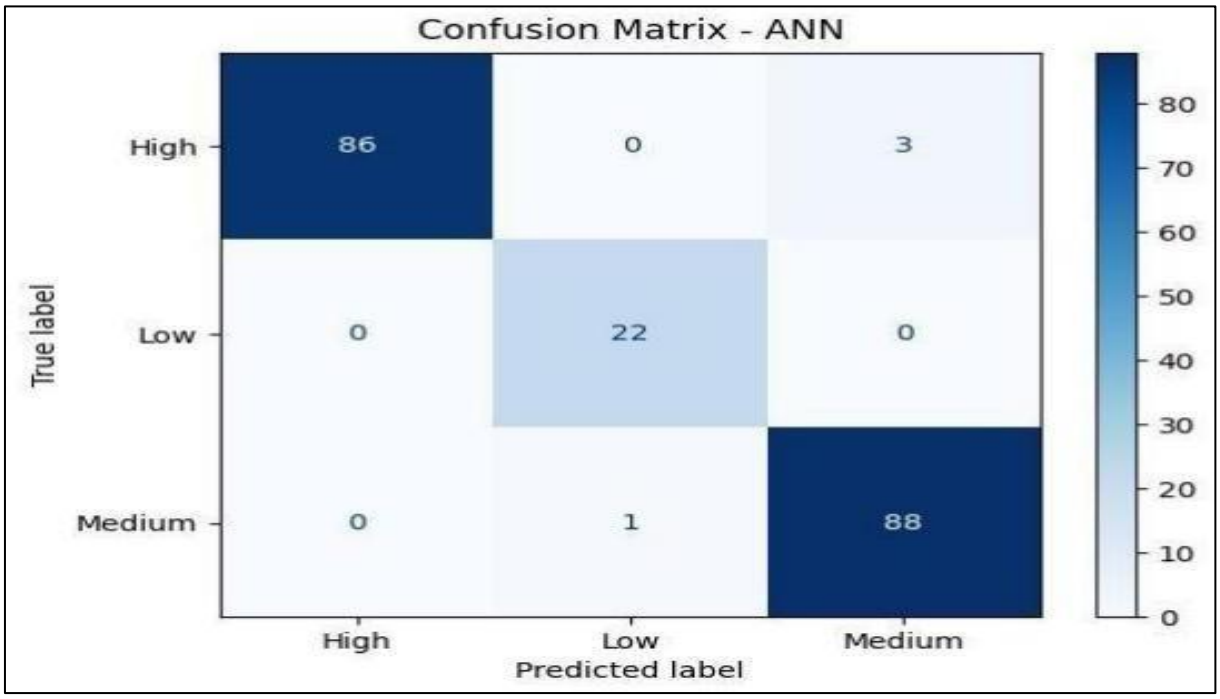


Fig 6 ANN - Confusion Matrix

Implementation of multiple machine learning models (Logistic Regression, Decision Tree, Random Forest, XGBoost, ANN) convincingly demonstrates the possibility of a model that is able to predict student performance levels (High, Medium, and Low) with acceptable accuracyLogistic Regression and XGBoost were the undisputed top performers in terms of performance with results that were almost perfect, as can be seen through their high-precision, recall, and F1scores.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep648

V. CONCLUSIONS

The primary objective of this research was to investigate the use of a machine learning algorithm in forecasting students' academic performance through a publicly available dataset of student performance. The research carried out pattern analysis on students' academic records, develop predictive models using demographic and educational variables, and evaluate the effectiveness of different machinelearning algorithms in predicting the students' performance levels. The approach embraced exploratory data analysis, data cleansing, feature selection, training the model, and evaluating the model's performance based on several metrics such as accuracy, precision, recall, F1-score, confusion matrix, and ROC curves. Five machine learning models were implemented, including logistic regression, decision tree, random forest, XGBoost, and artificial neural network. The outcomes demonstrated that all models had high levels of predictive accuracy.

Besides, logistic regression and XGBoost showed almost perfect performance, thus validating that machine learning can also be used to predict academic performance and identify at-risk students.Further proposed studies could possibly investigate advanced models with deep learning architectures and hybrid types that are much more advanced in the capturing of complex non-linear learning behaviours.

REFERENCES

[1]. Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*(3), e1355. https://doi.org/10.1002/widm.1355

[2]. Lampropoulos, G. (2023). Educational data mining and learning analytics in the 21st century. In *Encyclopedia of data science and machine learning* (pp. 1642–1651). IGI Global.

[3]. Christou, V., Tsoulos, I., Loupas, V., Tzallas, A. T., Gogos, C., Karvelis, P. S., et al. (2023). Performance and early drop prediction for higher education students using machine learning. *Expert Systems with Applications, 225*, 120079. https://doi.org/10.1016/j.eswa.2023.120079

[4]. Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student performance prediction using machine learning techniques. *Education Sciences, 11*(9), 552. https://doi.org/10.3390/educsci11090552