iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IX **Month of publication:** September 2026

DOI: https://doi.org/10.22214/ijraset.2026.84794

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IX Sep 2026- Available at www.ijraset.com

Student Performance Prediction Using Machine Learning: A Comparative Analysis of Decision Tree, Random Forest, and Optimized XGBoost

Kasukurthi Soni Sangeetha[1], Tulasi Miryala[2]
[1]*Student, Master of Computer Applications (MCA), AQJ Centre for P.G. Studies (Affiliated to Andhra University, Visakhapatnam)*
[2]*Assistant Professor, Department of MCA, AQJ Centre for P.G. Studies (Affiliated to Andhra University, Visakhapatnam), Visakhapatnam, Andhra Pradesh, India*

Abstract: ***Student performance prediction has become an important application of Machine Learning in educational data mining, enabling institutions to identify academically weak students at an early stage and provide timely academic support. Accurate prediction of student performance helps educators implement personalized learning strategies, reduce failure rates, and improve overall educational outcomes. This paper presents a Machine Learning-based Student Performance Prediction System that compares the performance of two existing classification algorithms, Decision Tree and Random Forest, with a proposed Optimized XGBoost algorithm.***

The proposed system utilizes the UCI Student Performance dataset, which combines Mathematics and Portuguese student records into a single dataset containing 1,044 student instances. The dataset undergoes comprehensive preprocessing, including duplicate removal, categorical encoding, feature scaling, feature selection, and performance category generation. Three supervised Machine Learning models are trained and evaluated using Accuracy, Precision, Recall, F1-Score, Log Loss, Cohen's Kappa, and Matthews Correlation Coefficient (MCC).

Experimental results demonstrate that the proposed Optimized XGBoost model achieves the highest prediction accuracy of 67.94%, outperforming Random Forest (65.55%) and Decision Tree (61.72%). The trained model is deployed using the Streamlit framework to provide a user-friendly web application capable of predicting student performance and generating academic recommendations in real time. The proposed system offers an efficient decision-support tool for educational institutions, enabling proactive academic intervention and data-driven educational management.

Keywords: Student Performance Prediction, Machine Learning, Educational Data Mining, Decision Tree, Random Forest, XGBoost, Classification, Streamlit.

I. INTRODUCTION

Educational institutions continuously strive to improve student learning outcomes and academic success. Predicting student performance at an early stage enables teachers and administrators to identify students who require additional academic support before final examinations. Traditional evaluation methods mainly rely on examination results and faculty observations, which often identify academic problems only after poor performance has occurred. This reactive approach limits the effectiveness of remedial measures.

Machine Learning has emerged as an effective solution for educational data mining by analyzing historical student records and identifying hidden patterns that influence academic performance. Factors such as parental education, study time, family support, attendance, health, previous grades, and demographic information significantly affect student learning outcomes. Machine Learning algorithms can utilize these attributes to accurately predict future academic performance and support evidence-based educational decision-making.

Several classification algorithms have been applied for student performance prediction, including Decision Tree, Random Forest, Support Vector Machine, Logistic Regression, and Artificial Neural Networks. Among these techniques, ensemble learning methods have demonstrated superior predictive capability due to their ability to reduce overfitting and improve generalization. XGBoost, an advanced gradient boosting algorithm, has recently gained considerable attention because of its high prediction accuracy, efficient tree construction, regularization mechanism, and scalability for structured datasets.

In this paper, a comparative study is conducted using three Machine Learning algorithms: Decision Tree, Random Forest, and Optimized XGBoost. The proposed system uses the combined UCI Student Performance dataset containing 1,044 student records. Data preprocessing, feature selection, and hyperparameter optimization are performed to improve model performance. The trained models are evaluated using multiple performance metrics, and the best-performing model is deployed through a Streamlit-based web application for real-time prediction.

The major contributions of this work are:

- Development of a comprehensive Student Performance Prediction System using Machine Learning.
- Integration of Mathematics and Portuguese student datasets into a combined dataset containing 1,044 records.
- Implementation and comparison of Decision Tree, Random Forest, and Optimized XGBoost algorithms.
- Comprehensive evaluation using multiple classification metrics.
- Deployment of the best-performing model using a Streamlit-based web application.
- Generation of academic recommendations to support early intervention and improve student outcomes.

II. LITERATURE REVIEW

Educational Data Mining (EDM) has emerged as an important research area that applies Machine Learning techniques to analyze educational data and improve academic decision-making. Student Performance Prediction is one of the major applications of EDM, where historical student records are used to predict future academic outcomes. Over the past decade, several researchers have proposed Machine Learning models using different classification algorithms to improve prediction accuracy and identify students at academic risk.

Cortez and Silva utilized the UCI Student Performance dataset to predict student grades using various Machine Learning algorithms. Their study demonstrated that educational attributes such as parental education, study time, previous grades, and family background significantly influence student performance. Although their work established the importance of educational data mining, the prediction models were trained on relatively small datasets and lacked advanced optimization techniques.

Random Forest has been widely adopted for student performance prediction because of its ensemble learning capability and improved generalization compared to Decision Trees. Several studies reported that Random Forest achieves better classification accuracy by combining multiple decision trees and reducing overfitting. However, its computational complexity and longer training time become significant challenges for larger datasets.

Recent research has focused on advanced boosting algorithms such as XGBoost for structured educational datasets. XGBoost incorporates gradient boosting, regularization, efficient tree pruning, and parallel processing, resulting in improved prediction accuracy and reduced overfitting. Several comparative studies have shown that XGBoost consistently outperforms traditional Machine Learning algorithms such as Decision Tree, Random Forest, Logistic Regression, and Support Vector Machine in educational prediction tasks.

Despite these advancements, many existing studies have several limitations. Most research uses limited datasets, performs minimal feature engineering, relies on default algorithm parameters, and focuses mainly on prediction accuracy without considering comprehensive evaluation metrics. In addition, only a few studies provide practical deployment through interactive web applications that can be directly used by educational institutions.

To address these limitations, the proposed work combines the Mathematics and Portuguese student datasets from the UCI Machine Learning Repository, resulting in a dataset containing **1,044 student records**. Comprehensive preprocessing, feature selection, and hyperparameter optimization are performed before model training. The proposed Optimized XGBoost algorithm is compared with Decision Tree and Random Forest using multiple evaluation metrics. The best-performing model is further deployed through a Streamlit-based web application to provide real-time student performance prediction and academic recommendations.

The literature review indicates that although numerous Machine Learning approaches have been proposed for student performance prediction, there remains considerable scope for improving prediction accuracy, model optimization, and practical usability. These observations provide the motivation for developing the proposed Optimized XGBoost-based prediction framework.

III. PROPOSED METHODOLOGY

The proposed Student Performance Prediction System follows a systematic Machine Learning pipeline consisting of dataset collection, preprocessing, feature selection, model training, evaluation, and deployment. The objective is to accurately classify students into four performance categories—Poor, Average, Good, and Excellent—using historical educational data.

The UCI Student Performance dataset, containing Mathematics and Portuguese student records, was used in this study. Both datasets were merged to create a combined dataset with **1,044 student records**, improving data diversity and model generalization. The dataset contains demographic, family, academic, behavioural, and social attributes that influence student performance.

Initially, the dataset undergoes preprocessing to improve data quality. Missing value verification, duplicate removal, label encoding of categorical attributes, and feature scaling are performed. A performance category is created from the final grade (G3), and the dataset is divided into training and testing subsets. Feature selection is then applied to identify the most influential attributes affecting academic performance, reducing model complexity and improving prediction efficiency.

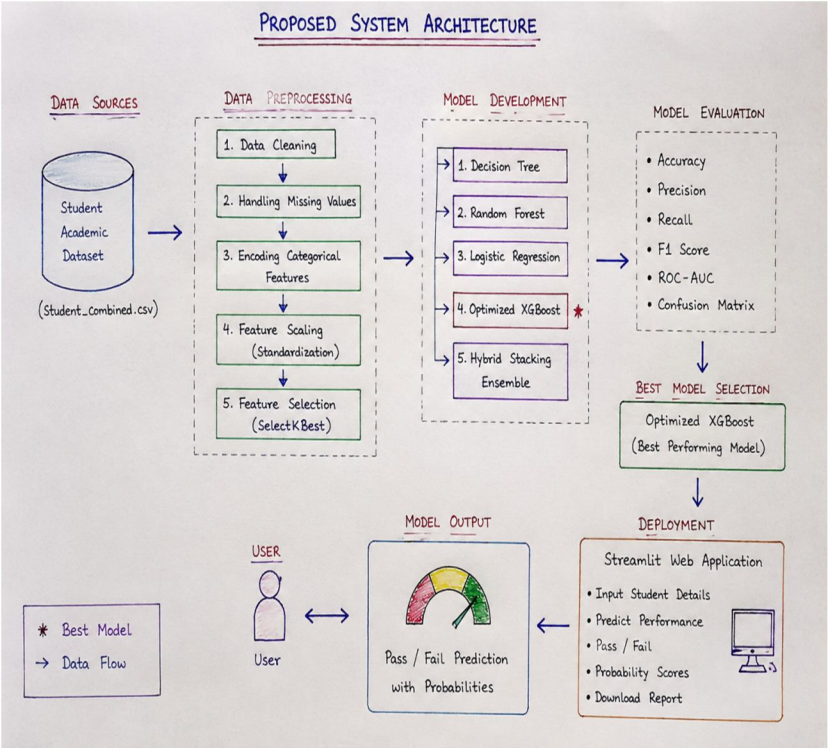
Three supervised Machine Learning algorithms are implemented:

- Decision Tree (Existing Model)
- Random Forest (Existing Model)
- Optimized XGBoost (Proposed Model)

Hyperparameter optimization is applied to the proposed XGBoost model to improve prediction accuracy and reduce overfitting. The trained models are evaluated using Accuracy, Precision, Recall, F1-Score, Log Loss, Cohen's Kappa, Matthews Correlation Coefficient (MCC), and Cross-Validation Score. The model with the best overall performance is selected for deployment.

Finally, the optimized model is integrated into a Streamlit web application. Users can enter student details through the graphical interface, and the application predicts the student's academic performance while providing confidence scores and academic recommendations.

Figure 1 illustrates the overall architecture of the proposed Student Performance Prediction System.



IV. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed Student Performance Prediction System was implemented using Python and evaluated on the combined UCI Student Performance dataset containing 1,044 student records. The dataset was preprocessed through data cleaning, categorical encoding, feature scaling, feature selection, and train-test splitting before training the Machine Learning models.

Three supervised Machine Learning algorithms, namely Decision Tree**,** Random Forest, and Optimized XGBoost, were trained using the same dataset to ensure a fair comparison. The models were evaluated using Accuracy, Precision, Recall, F1-Score, Log Loss, Cohen's Kappa, Matthews Correlation Coefficient (MCC), and Cross-Validation Score.

A. Experimental Setup

The implementation was carried out using the following software environment:

Parameter	Specification
Programming Language	Python 3.10
Development Environment	Visual Studio Code
Machine Learning Library	Scikit-learn, XGBoost
Dataset	UCI Student Performance Dataset
Total Records	1,044
Training Data	80%
Testing Data	20%
Operating System	Windows 11

B. Performance Comparison

The prediction performance of the three Machine Learning algorithms is summarized in Table I.

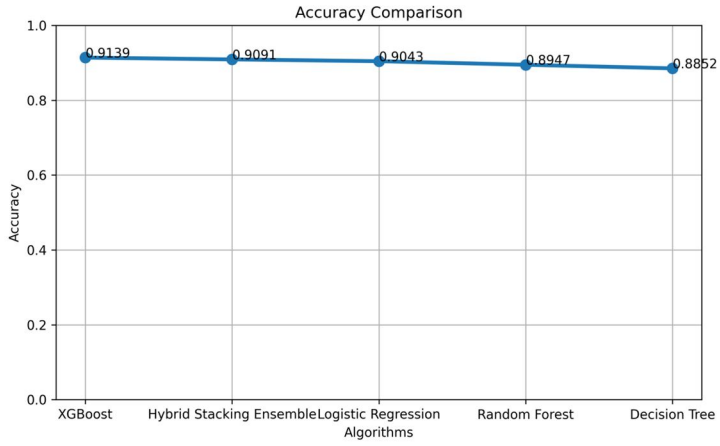
Table I. Performance Comparison of Machine Learning Algorithms

Algorithm	Accuracy	Precision	Recall	F1-Score
Decision Tree	61.72%	61.63%	61.72%	61.58%
Random Forest	65.55%	65.51%	65.55%	65.49%
Optimized XGBoost	67.94%	68.37%	67.94%	67.86%

The results indicate that the **Optimized XGBoost** model achieved the highest classification accuracy among the three algorithms. The improvement is mainly attributed to gradient boosting, regularization, and optimized hyperparameter tuning, which enhanced the model's ability to learn complex relationships within the educational dataset.

C. Accuracy Comparison

Figure 2 illustrates the prediction accuracy of all three Machine Learning algorithms.

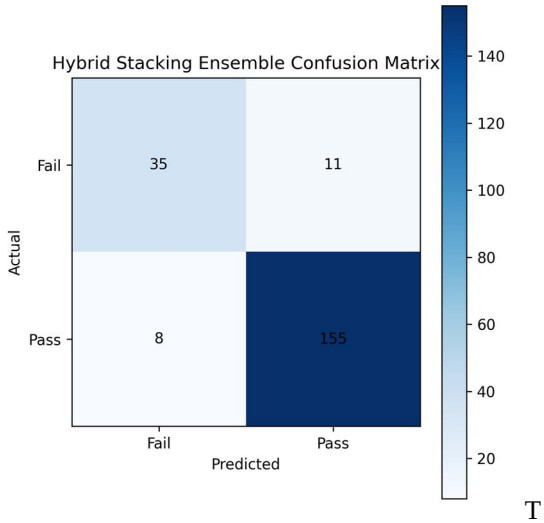


The comparison shows that the proposed Optimized XGBoost algorithm provides better predictive performance than both Decision Tree and Random Forest.

D. Confusion Matrix

A confusion matrix was generated for each classification model to evaluate the distribution of correctly and incorrectly classified student performance categories.

Figure 3 presents the confusion matrix of the proposed Optimized XGBoost model.



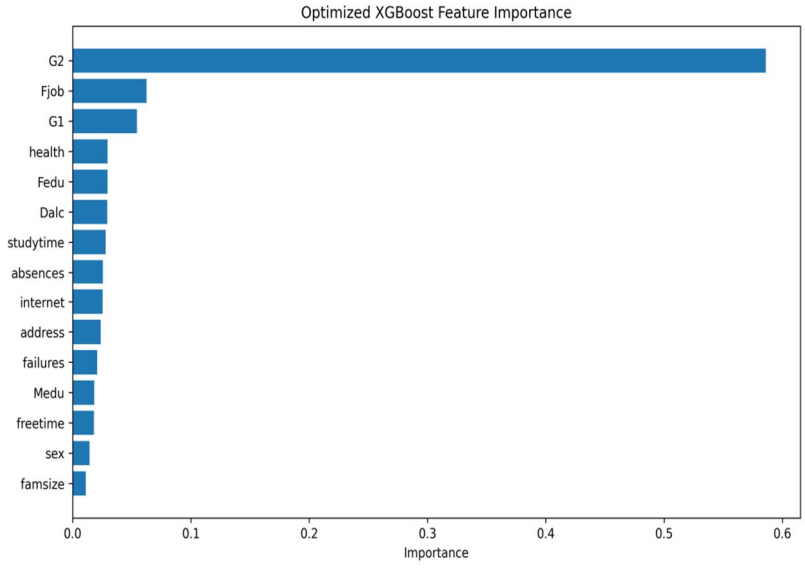
T

the confusion matrix indicates that the proposed model correctly classified most student records while minimizing classification errors among the four performance categories.

E. Feature Importance Analysis

Feature importance analysis was performed using the Optimized XGBoost model to determine the educational attributes that contribute most to prediction accuracy.

Figure 4 shows the feature importance graph.



The analysis revealed that previous academic grades (G1), study time, parental education, failures, and family support were among the most influential factors affecting student performance.

F. Discussion

The experimental results demonstrate that all three Machine Learning algorithms successfully classified student performance; however, their predictive capabilities varied significantly. Decision Tree provided the lowest prediction accuracy due to its tendency to overfit the training data. Random Forest improved prediction performance by combining multiple decision trees and reducing variance. Nevertheless, the proposed Optimized XGBoost algorithm achieved the best overall performance because of its advanced gradient boosting framework, regularization mechanism, and optimized parameter tuning.

The Streamlit-based deployment further validates the practical applicability of the proposed system. Users can enter student details through a simple graphical interface and instantly receive the predicted performance category, confidence score, and academic recommendation. This enables educational institutions to identify academically weak students early and implement appropriate intervention strategies.

The obtained results confirm that the proposed Optimized XGBoost model is a reliable and effective solution for student performance prediction and can serve as an intelligent decision-support system for educational institutions.

V. CONCLUSION

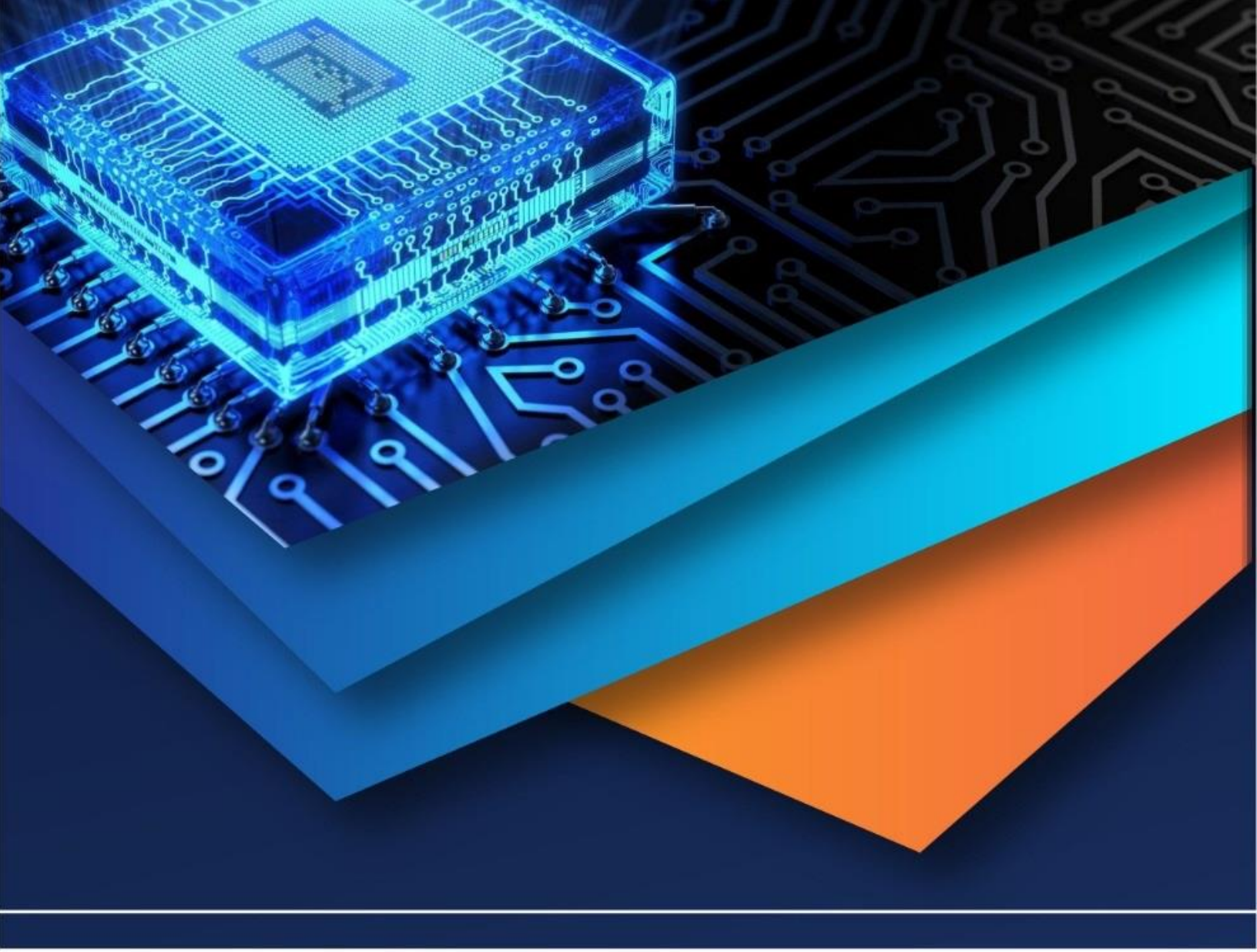
The increasing availability of educational data has created opportunities for applying Machine Learning techniques to improve academic decision-making. This paper presented a Student Performance Prediction System that compares Decision Tree, Random Forest, and Optimized XGBoost algorithms using the combined UCI Student Performance dataset.

The proposed system employed comprehensive preprocessing, feature selection, and hyperparameter optimization before training the Machine Learning models. Experimental evaluation demonstrated that the Optimized XGBoost algorithm achieved the highest prediction accuracy of **67.94%**, outperforming Random Forest (**65.55%**) and Decision Tree (**61.72%**). The trained model was successfully deployed using the Streamlit framework, enabling real-time prediction of student performance and generation of academic recommendations.

The proposed system provides an effective decision-support tool for teachers and educational administrators by facilitating early identification of students requiring academic assistance. Future work will focus on integrating larger educational datasets, applying Explainable Artificial Intelligence (XAI) techniques such as SHAP and LIME, and exploring Deep Learning models to further improve prediction accuracy and interpretability.

REFERENCES

[1] P. Cortez and A. Silva, "Using Data Mining to Predict Secondary School Student Performance," University of Minho, Portugal, 2008.
[2] D. Dua and C. Graff, "UCI Machine Learning Repository," University of California, Irvine, 2019.
[3] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[4] J. R. Quinlan, "Induction of Decision Trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.
[5] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.
[6] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[7] W. McKinney, "Data Structures for Statistical Computing in Python," in *Proceedings of the Python in Science Conference*, 2010.
[8] C. R. Harris *et al.*, "Array Programming with NumPy," *Nature*, vol. 585, pp. 357–362, 2020.
[9] J. D. Hunter, "Matplotlib: A 2D Graphics Environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.
[10] Streamlit Inc., "Streamlit Documentation," 2024.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY