iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.80495

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 IV Apr 2026- Available at www.ijraset.com

Student Academic Performance and Dropout Risk Prediction Using Machine Learning

Dr. R. Satya Prasad[1], Tamiri Divya[2], Relli Gnaneswari[3], Vemula Ghana Prabhas[4], Rentapalli Hemanth[5], Syed Jilani[6]
[1]*Professor & Dean, R&D, Department of CSE,* [2,3,4,5,6]*Students, Department of Computer Science and Engineering Dhanekula Institute of Engineering and Technology Vijayawada, India*

Abstract: ***Student dropout and inconsistent academic performance continue to pose serious challenges in higher education, affecting both institutional outcomes and student success. Early identification of students at risk enables institutions to implement timely interventions and improve retention rates. This study presents a machine learning-based framework for predicting student academic outcomes and dropout risk using historical educational data.***
The proposed model is developed using a dataset from the UCI Machine Learning Repository that includes academic, demographic, and socioeconomic attributes. Multiple machine learning algorithms, including Logistic Regression, Random Forest, and XGBoost, are implemented and evaluated for classifying student outcomes. Experimental results indicate that XGBoost achieves the best performance, with a test accuracy of approximately 87.45%, outperforming the other models.
To enhance interpretability, Shapley Additive explanations (SHAP) are utilised to examine feature contributions. The findings reveal that academic performance indicators, particularly second-semester results and the completion of curricular units, play a significant role in predicting student dropout.
Overall, the proposed system provides an accurate and interpretable solution for early dropout detection, supporting data-driven decision-making in educational institutions.
Keywords: Machine Learning, Student Dropout Prediction, Academic Performance, XGBoost, SHAP, Explainable AI.

I. INTRODUCTION

Student dropout and poor academic performance are major concerns in higher education, as they impact both institutional reputation and student career progression. Many students encounter academic, financial, and personal challenges that may lead to discontinuation of their studies. Traditional approaches for identifying at-risk students are often reactive and rely on manual monitoring, which limits the possibility of timely intervention. With the increasing availability of educational data, machine learning techniques provide an effective way to analyse student information and predict outcomes in advance. By leveraging historical academic records, demographic details, and socio-economic factors, predictive models can identify patterns associated with student success or failure. This enables institutions to take proactive measures such as mentoring, counselling, and academic support. In this study, a machine learning-based framework is proposed to predict student academic outcomes and dropout risk. The system utilizes a publicly available dataset containing comprehensive student-related attributes. Multiple classification algorithms, including Logistic Regression, Random Forest, and XGBoost, are implemented and compared to evaluate their performance. An ensemble approach is also incorporated to improve prediction accuracy and robustness. Furthermore, the proposed system emphasizes interpretability by integrating Shapley Additive explanations (SHAP), which helps in understanding how individual features influence model predictions. The system is also supported by an interactive dashboard that presents predictions and analytical insights in a clear and user-friendly manner. This approach aims to assist educational institutions in making informed decisions and implementing early intervention strategies.

II. LITERATURE REVIEW

Several research studies have examined the use of machine learning methods to predict academic performance and risk of dropout in students. Educational data mining has proved to be an important field of research, with previous student data being used to determine patterns and further decision-making in educational establishments [6], [7].
Conventional algorithms like Logistic Regression and Decision Trees have been popular since they are easy to learn and understand. In particular, Logistic Regression can be used to address binary classification problems, whereas Decision Tree methods offer an intuitive decision-making framework [4], [7]. Nevertheless, such models can be hard to extend to large and heterogeneous datasets to capture complex relationships. In order to overcome this weakness, ensemble methods like Random Forest have been proposed, and such methods help to enhance the accuracy of predictions as they use a set of decision trees [3].

Recent research indicates that more sophisticated algorithms (Gradient Boosting, XGBoost) have become popular in student performance prediction because they are highly accurate and have the potential to operate on non-linear relationships in data [2], [9]. Such models have proven to be better in performance than the conventional methods. Also, other studies are concerned with enhancing the efficiency of models by selecting features and hyperparameter optimization, but such approaches can lead to higher computing costs and low interpretability [9], [10].

Moreover, some of the existing systems concentrate on one activity, e.g., either dropout prediction or academic performance analysis.

This restricts their relevance in practical educational settings, which are highly interrelated in aspects of both aspects. Furthermore, there are numerous studies with inefficient visualization and interfaces that complicate the interpretation and effective use of the results [6].

To address these drawbacks, the current work suggests an integrated model that fuses several machine learning models to forecast academic success and dropout probability. Explainability is also implemented in the system with the help of Shapley Additive Explanations (SHAP) to investigate the contribution of features and increase transparency of model predictions [5]. Moreover, an interactive dashboard is created to display insights clearly and in an easy way to assist in making decisions based on data in educational institutions.

III. METHODOLOGY

This section describes the dataset, preprocessing techniques, machine learning algorithms, and evaluation methods used to predict student dropout risk and academic performance.

A. Dataset Description

The dataset used in this study is obtained from the UCI Machine Learning Repository, which provides detailed information on students' academic, demographic, and socio-economic characteristics **[1]**. The dataset includes features such as curricular unit performance, grades, admission scores, financial status, and personal background.

The dataset comprises 36 input features and is divided into training and testing sets. The training set contains 3539 samples, while the test set includes 885 samples. The dataset is moderately imbalanced, with approximately 67.87% non-dropout instances and 32.13% dropout instances.

The data is structured with 36 input features and is divided into training and testing sets. The training set consists of 3539 samples, while the test set contains 885 samples. The dataset is moderately imbalanced, with approximately 67.87% non-dropout instances and 32.13% dropout instances.

B. Data Preprocessing

Data preprocessing is performed to ensure data quality and improve model performance. The following steps are

- Handling Missing Values: Missing or inconsistent data entries are identified and handled appropriately.
- Encoding Categorical Variables: Categorical features are converted into numerical values using encoding techniques.
- Feature Scaling: Numerical features are normalized to maintain consistency across different ranges.
- Train-Test Split: The dataset is divided into training and testing subsets for model evaluation. These preprocessing steps ensure that the dataset is suitable for training machine learning models.

C. Model Development

Several machine learning models are implemented to predict student dropout risk:

- Logistic Regression: A basic model used for binary classification due to its simplicity and interpretability [4].
- RandomForest: An ensemble learning technique that builds multiple decision trees and impr oves prediction accuracy by reducing overfitting [3].
- XGBoost: A gradient boosting algorithm that effectively captures complex patterns and delivers high predictive performance **[2]**.

Cross-validation is conducted to assess model stability. The results show that all models perform well, with XGBoost achieving the highest average accuracy.

D. Model Evaluation

The models are assessed using standard classification metrics, including:

- Accuracy
- Precision
- Recall
- F1-score
- Confusion Matrix

The confusion matrix provides detailed insight into model performance by showing the number of correctly and incorrectly classified instances. The results indicate that XGBoost achieves the highest test accuracy of approximately 87%, outperforming Logistic Regression and Random Forest.

E. Explainability using SHAP

To enhance model interpretability, Shapley Additive Explanations (SHAP) are applied to analyze the impact of each feature [5]. SHAP provides both global and local explanations:

- Global Explanation: Identifies the most influential features across the entire dataset.
- Local Explanation: Explains individual predictions by showing how each feature contributes to the final output.

The SHAP analysis reveals that features like second-semester performance, number of approved curricular units, and tuition fee status play a significant role in predicting student dropout.

F. System Workflow

The overall system follows a structured pipeline from data collection to final prediction and decision support. Initially, student data is collected and preprocessed. The processed data is then used to train machine learning models. Once trained, the models generate predictions for academic performance and dropout risk.

The results are evaluated using performance metrics and further interpreted using SHAP-based explainability techniques. Finally, the predictions and insights are presented through a user-friendly dashboard, helping institutions to identify at-risk students and implement early intervention strategies.

IV. RESULTS AND DISCUSSION

This section presents the experimental results of the implemented machine learning models and provides a detailed analysis of their performance in predicting student dropout risk.

A. Model Performance

The performance of the models was evaluated using standard classification metrics, including accuracy, precision, recall, F1-score, and ROC-AUC. The experimental results indicate that all models perform effectively, with noticeable differences in predictive capability.

Among the evaluated models, XGBoost achieved the highest test accuracy of approximately 87.45%, outperforming Logistic Regression and Random Forest. Logistic Regression achieved an accuracy of approximately 85.99%, while Random Forest achieved around 87.01%. Cross-validation results further confirm the robustness of the models, with XGBoost achieving the highest average accuracy.

B. Model Comparison

A comparative evaluation of the applied machine learning models reveals clear differences in predictive performance. Logistic Regression is used as a baseline model due to its simplicity and interpretability. However, its ability to capture complex relationships in the data is limited. Random Forest improves prediction performance by leveraging multiple decision trees and effectively modeling non-linear patterns present in student data. Among all the models, XGBoost achieves the best overall performance. Its gradient boosting approach iteratively minimizes prediction errors, resulting in improved accuracy and robustness. This makes it particularly suitable for handling structured educational datasets with diverse features.

The results indicate that advanced ensemble and boosting techniques are more effective than traditional models for predicting student dropout risk. These methods provide better generalization and are capable of capturing intricate relationships between academic, demographic, and socio-economic factors.

Table 1. Performance Comparison of Machine Learning Models

Model	Accuracy	Precision (Dropout)	Recall (Dropout)	F1-Score (Dropout)	ROC-AUC
Logistic Regression	0.8599	0.78	0.79	0.78	0.91
Random Forest	0.8701	0.89	0.65	0.75	0.92
XGBoost	0.8746	0.87	0.72	0.79	0.91

As shown in Table 1, XGBoost achieves the highest accuracy among all models. Random Forest demonstrates higher precision for dropout prediction, while Logistic Regression provides balanced performance with better interpretability.

C. Confusion Matrix Analysis

The confusion matrix provides a detailed evaluation of the classification performance of the XGBoost model. A significant number of non-dropout instances are correctly classified, indicating strong predictive capability in identifying students likely to continue their education.

However, some dropout cases are misclassified, which can be attributed to class imbalance in the dataset. Despite this, the model maintains a good balance between precision and recall, ensuring reliable performance across both classes.

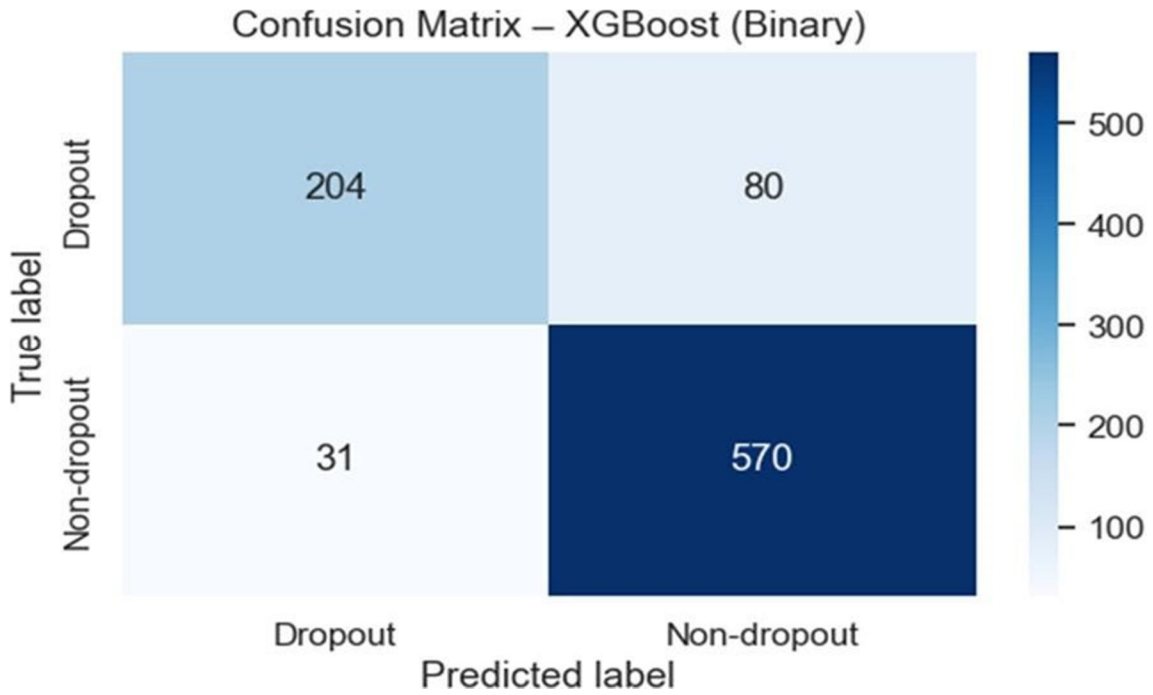


Fig. 2. Confusion matrix of the XGBoost model

As shown in Fig. 2, the model correctly classifies a large number of non-dropout students while maintaining reasonable performance in identifying dropout cases.

D. SHAP-Based Interpretation

To enhance model transparency, Shapley Additive Explanations (SHAP) are employed to interpret the predictions of the machine learning model. SHAP provides insights into how individual features contribute to the model output.

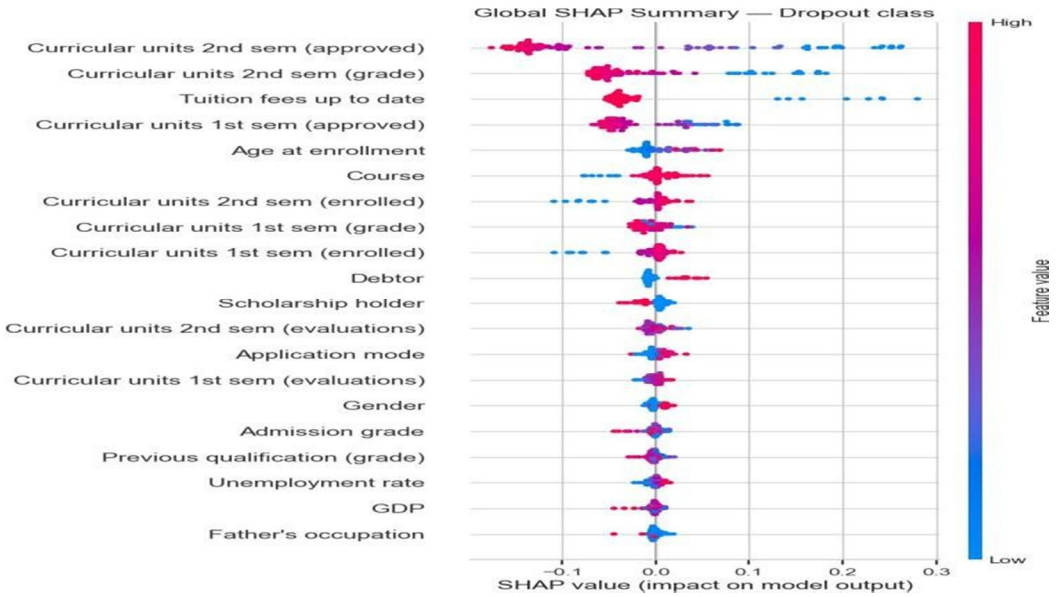


Fig. 3. SHAP summary plot for feature impact analysis

The SHAP summary plot illustrates the contribution of each feature to the prediction outcome. Features such as second-semester grades, number of approved curricular units, and tuition fee status exhibit a strong influence on student dropout prediction. The distribution of SHAP values indicates both positive and negative impacts of features on the model output.

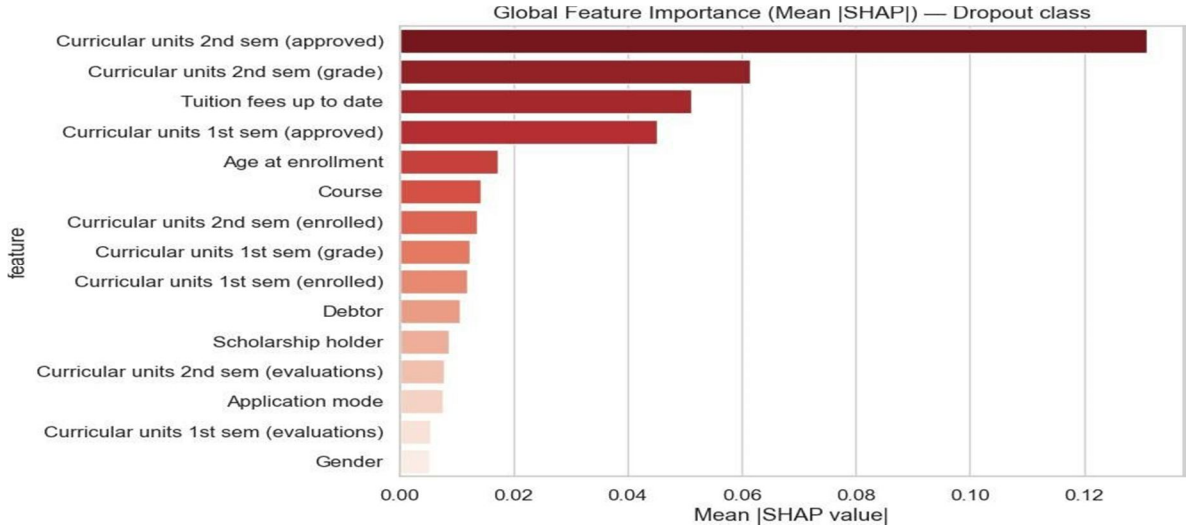


Fig. 4. Global feature importance based on SHAP values

As shown in Fig. 4, the global feature importance analysis highlights that second semester approved units, second semester grades, and tuition fee status are the most influential factors in predicting student dropout. This confirms that academic performance plays a critical role in determining student outcomes.

E. Discussion

The experimental results demonstrate that machine learning techniques can effectively model and predict student dropout risk using academic and socio-economic data. Among the evaluated models, XGBoost achieved the highest overall accuracy, indicating its strong capability in capturing complex patterns within the dataset. The ensemble-based nature of Random Forest also contributed to improved performance compared to traditional methods, while Logistic Regression provided a stable baseline with interpretable results.

The confusion matrix analysis highlights that the model performs particularly well in identifying non-dropout students, with a high number of correctly classified instances. However, some dropout cases are misclassified, which can be attributed to class imbalance in the dataset. This imbalance affects recall for the dropout class, suggesting that future improvements could focus on handling imbalanced data more effectively.

The integration of SHAP-based explainability significantly enhances the interpretability of the proposed system. The results reveal that academic-related features, especially second-semester performance and tuition fee status, play a dominant role in determining student outcomes. These insights are valuable for educational institutions, as they enable data-driven decision-making and early identification of at-risk students.

Overall, the findings confirm that combining predictive modeling with explainable AI techniques provides both high performance and transparency. This combination is essential for real-world applications, where understanding the reasoning behind predictions is as important as achieving high accuracy. The proposed approach can serve as a practical tool for supporting academic interventions and improving student retention rates.

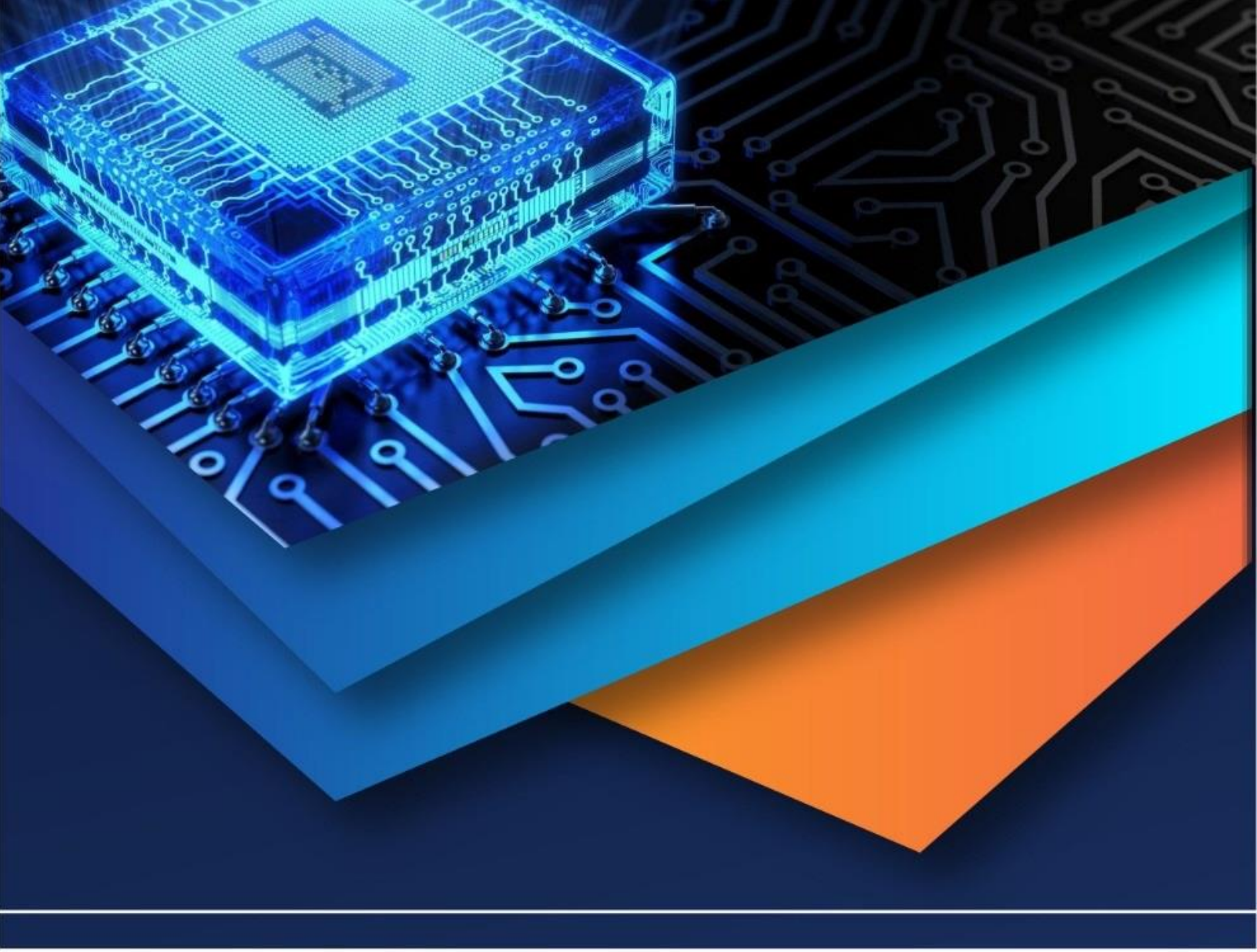
V. CONCLUSION

This study presents a machine learning-based framework for predicting student dropout risk using academic, demographic, and socio-economic data. Multiple models, including Logistic Regression, Random Forest, and XGBoost, were implemented and evaluated, with XGBoost demonstrating the best overall performance. The results indicate that ensemble and boosting techniques outperform traditional approaches in terms of accuracy and robustness. The integration of Shapley Additive Explanations (SHAP) enhances model interpretability by identifying the key features influencing predictions.

The proposed system provides a reliable and interpretable solution for early identification of at-risk students. It can assist educational institutions in implementing timely interventions, ultimately improving student retention and academic success.

REFERENCES

[1] M. S. A. N. Araújo, P. J. G. Lisboa, and A. M. P. de Carvalho, “Predict Students’ Dropout and Academic Success,” UCI Machine Learning Repository, 2020. [Online]. Available: https://archive.ics.uci.edu/
[2] 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 (KDD), 2016, pp. 785– 794.
[3] L. Breiman, “Random Forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[4] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, “Applied Logistic Regression,” 3rd ed., Wiley, 2013
[5] S. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 4765–4774.
[6] C. Romero and S. Ventura, “Educational Data Mining: A Review of the State of the Art,” IEEE Transactions on Systems, Man, and Cybernetics, vol. 40, no. 6, pp. 601–618, 2010.
[7] J. Han, M. Kamber, and J. Pei, “Data Mining: Concepts and Techniques,” 3rd ed., Morgan Kaufmann, 2012.
[8] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
[9] A. Géron, “Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow,” 2nd ed., O’Reilly Media, 2019.
[10] H. Abdi and L. J. Williams, “Principal Component Analysis,” Wiley Interdisciplinary Reviews: Computational Statistics, vol. 2, no. 4, pp. 433–459, 2010.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY