Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

A Comparative and Explainable Machine Learning Approach for Student Dropout Prediction

Polepally Navaneetha
Assistant Professor
Dept. of AI & ML
Vignan Institute of Technology and Science
Deshmukhi, Pochampally, TG, India - 508284

Ijjana Prasanth Kumar
Assistant Professor
Dept. of AI & ML
Vignan Institute of Technology and Science
Deshmukhi, Pochampally, TG, India - 508284

Dr. M. Sreeramu
Assistant Professor
Dept. of MBA
Vignan Institute of Technology and Science
Deshmukhi, Pochampally, TG, India - 508284

***Abstract*—Abstract— Now a days student dropout has become major challenge for educational institutions. This is impacting on students' academic performance , their career development. It is affecting educational reputation also. Existing student dropout prediction approaches primarily focus on achieving high prediction accuracy using individual machine learning models, while providing limited model comparison and insufficient explanation of the factors influencing dropout decisions This study proposes a comparative and explainable machine learning approach for predicting student dropout. A publicly available dataset containing academic, behavioral, and socio-economic attributes was utilized for analysis. Data preprocessing techniques, including missing value imputation and categorical encoding, were applied to ensure data quality. The dataset was divided into training and testing subsets in an 80:20 ratio for model evaluation. Several machine learning algorithms, such as Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine, were implemented and compared.**

***Index Terms*—Machine Learning, Student Dropout Prediction, Random Forest, Classification**

I. INTRODUCTION

Student dropout is a significant challenge faced by educational institutions, affecting both student success and institutional performance. Early identification of students at risk of dropping out is essential to implement timely intervention strategies. However, traditional methods often fail to accurately predict dropout due to the complexity of influencing factors such as academic performance, behavioral patterns, and socio-economic conditions.

With the advancement of machine learning techniques, it has become possible to analyze large volumes of educational data and identify patterns that can help predict student dropout. Machine learning models can effectively handle complex relationships among multiple variables and provide better prediction accuracy compared to traditional approaches.

This study aims to develop a comparative and explainable machine learning approach for student dropout prediction. The objective is to evaluate multiple machine learning models and identify key factors influencing student dropout, thereby assisting educational institutions in making informed decisions and reducing dropout rates. Education analytics has emerged as an important research field due to the rapid growth of digital educational platforms and student-related data. Educational institutions continuously generate data related to attendance, academic performance, assessments, and student behavior. Proper analysis of this data can help institutions improve learning outcomes, identify at-risk students, and enhance institutional decision-making processes.

Student dropout is considered a major educational and social problem worldwide. High dropout rates negatively affect institutional reputation, financial stability, and overall educational quality. In addition, students who discontinue their education may face difficulties in career growth and employment opportunities. Therefore, accurate prediction of student dropout has become an essential task in modern educational systems.

Despite the availability of educational data, predicting student dropout remains challenging because dropout behavior is influenced by multiple interconnected factors. Academic performance alone is not sufficient for accurate prediction, as behavioral and socio-economic conditions also play a critical role. Traditional statistical methods often struggle to handle these complex relationships effectively.

Machine learning techniques provide efficient solutions for analyzing educational datasets and extracting hidden patterns from the data. Algorithms such as Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine have shown promising results in predictive analytics applications. These techniques can improve prediction accuracy and support early intervention strategies for educational institutions.

In recent years, explainable artificial intelligence has gained significant attention in educational analytics. In addition to accurate predictions, it is important to understand the factors influencing student dropout. Explainable machine learning models improve transparency and help educators identify the key attributes contributing to student dropout risk.

A. *Problem Statement*

Educational institutions face significant challenges due to increasing student dropout rates, which negatively affect academic performance, institutional reputation, and student career development. Identifying students at risk of dropping out at an early stage is difficult because multiple academic, behavioral, and socio-economic factors influence student performance. Traditional approaches are often unable to effectively analyze

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

these complex relationships. Therefore, there is a need for an efficient machine learning-based system capable of accurately predicting student dropout and identifying the key factors contributing to it.

B. Objectives

- To preprocess and analyze student-related data.
- To implement multiple machine learning models for dropout prediction.
- To compare the performance of different algorithms.
- To identify important factors influencing student dropout.

II. LITERATURE REVIEW

Several studies have explored the use of machine learning techniques for predicting student dropout. Commonly used algorithms include Logistic Regression, Decision Trees, Random Forest, and Support Vector Machines. These models have demonstrated the ability to identify students at risk based on academic and behavioral data. Some research has reported high prediction accuracy using advanced techniques such as ensemble methods and neural networks. However, many of these studies focus primarily on improving accuracy and do not provide sufficient insights into the underlying factors contributing to student dropout. This lack of interpretability limits the practical application of these models in real-world educational settings. Recent works have emphasized the importance of explainable machine learning approaches to better understand the factors influencing student dropout. Identifying key attributes such as academic performance, attendance, and socio-economic conditions is essential for effective intervention strategies. Despite these advancements, there is still a need for a comprehensive study that not only compares multiple machine learning models but also provides clear insights into the factors affecting student dropout. Therefore, this study adopts a comparative and explainable machine learning approach to address these challenges. In recent years, educational data mining and learning analytics have become important research areas due to the increasing availability of student-related data. Educational institutions generate large amounts of data related to student attendance, assessments, learning behavior, and academic performance. Analyzing this data can help institutions identify at-risk students and improve retention strategies.

Logistic Regression is widely used in student prediction systems because of its simplicity and interpretability. It performs effectively when relationships among variables are linear and provides probability-based predictions. However, its performance may decrease when handling highly complex and non-linear datasets.

Decision Tree algorithms are popular because they provide clear visualization and easy interpretation of decision-making processes. These models can identify important attributes affecting student dropout. However, Decision Trees are prone to overfitting and may produce unstable results when trained on noisy data.

Random Forest, an ensemble learning algorithm, overcomes some limitations of Decision Trees by combining multiple trees to improve prediction accuracy and reduce overfitting. Due to its robustness and capability to handle high-dimensional datasets, Random Forest has become one of the most effective techniques for educational prediction tasks.

Support Vector Machine is another widely used classification algorithm in educational analytics. SVM performs well in high-dimensional spaces and can create optimal decision boundaries for classification tasks. Nevertheless, the algorithm requires careful parameter tuning and feature scaling to achieve better performance.

Recent studies have also emphasized the importance of explainable artificial intelligence in educational prediction systems. In addition to achieving high prediction accuracy, researchers focus on understanding the factors influencing student dropout predictions. Explainable machine learning techniques improve transparency and help educators identify critical factors such as attendance, GPA, and study behavior that contribute to dropout risk.

III. 3.PROPOSED WORK:

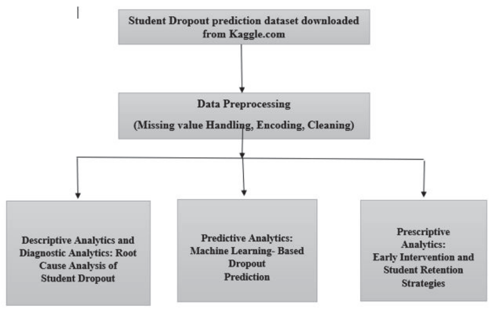


Fig. 1. Student Framework Methodology

The proposed work depicted in Figure 1 entails a comprehensive analytics-driven framework for student dropout prediction and educational risk analysis aimed at extracting meaningful insights to improve student retention, academic performance, and institutional decision-making.

3.1.1 Data Collection and Preprocessing Data collection and preprocessing are the foundational stages of the proposed student dropout analytics framework, enabling the extraction of meaningful insights from large educational datasets.

3.1.2. Descriptive and Diagnostic Analytics With the preprocessed dataset, descriptive and diagnostic analytics techniques were applied to gain detailed insights into student dropout-related factors.

3.1.3. Predictive Analytics

Building upon the insights obtained from descriptive and diagnostic analytics, robust machine learning models were developed using predictive analytics techniques to forecast student dropout probability.

3.1.4. Prescriptive Analytics Finally, leveraging the findings obtained from descriptive, diagnostic, and predictive analytics,

Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

prescriptive analytics techniques were utilized to formulate effective intervention strategies for reducing student dropout risk and improving academic performance.

Through this holistic approach, the proposed work aims to reduce Students dropouts rates with data-driven insights and actionable recommendations derived from advanced analytics techniques.

A. 4. Dataset Analysis

	Student_ID	Age	Gender	Family_Income	Internet_Access	Study_Hours_per_Day	Attendance_Rate	Assignment_Delay_Days	Travel_Time_Minutes	Part_Time_Job	Scl
0	1	22.1	Male	25000.0	Yes	3.36	86.1	2	20.4	Yes	
1	2	20.7	Male	25000.0	Yes	4.30	68.0	2	44.0	No	
2	3	22.4	Male	40183.0	Yes	4.40	70.9	0	48.9	Yes	
3	4	24.4	Male	NaN	Yes	NaN	82.2	2	38.6	No	
4	5	20.5	Female	25319.0	Yes	4.19	75.7	1	23.0	No	

Fig. 2. Dataset Overview

The effectiveness of the proposed framework is demonstrated using a student dropout dataset downloaded from Kaggle.com, illustrating how educational institutions can leverage analytics techniques for data-driven decision-making and early student intervention. The dataset comprises several important academic, behavioral, and socioeconomic attributes related to student performance and dropouts, as depicted in Figure .

B. 5. Data Preprocessing

Data preprocessing was performed to improve the quality and reliability of the dataset before model development. The preprocessing steps included missing value imputation, duplicate removal, and categorical data encoding.

Initially, missing values were identified and handled using multiple imputation techniques. Among the evaluated methods, the K-Nearest Neighbors (KNN) imputation technique produced comparatively better predictive performance, indicating its effectiveness in preserving the underlying data distribution. The comparative analysis of the imputation techniques is presented in Figure 3.

After selecting the most suitable imputation method, duplicate records were identified and removed to eliminate redundant observations. Subsequently, categorical features were transformed into numerical representations using label encoding, making the dataset suitable for machine learning algorithms. The preprocessing workflow and the resulting cleaned dataset are illustrated in Figure **??**.

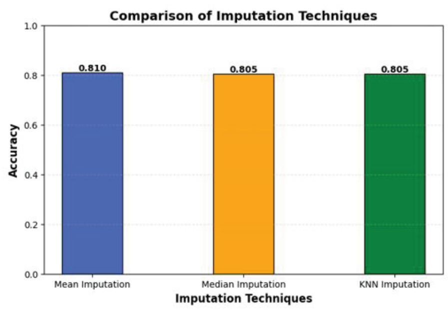


Fig. 3. Comparative Analysis of Missing Value Imputation Techniques

	Student_ID	Age	Gender	Family_Income	Internet_Access	Study_Hours_per_Day	Attendance_Rate	Assignment_Delay_Days	Travel_Time_Minutes	Part_Time_Job	Scl
0	1	22.1	Male	25000.000000	Yes	3.360000	86.1	2	20.4	Yes	
1	2	20.7	Male	25000.000000	Yes	4.300000	68.0	2	44.0	No	
2	3	22.4	Male	40183.000000	Yes	4.400000	70.9	0	48.9	Yes	
3	4	24.4	Male	38377.247474	Yes	4.014592	82.2	2	38.6	No	
4	5	20.5	Female	25319.000000	Yes	4.190000	75.7	1	23.0	No	

Fig. 4. Dataset after missing value imputation and duplicate removal.

After handling missing values and performing deduplication, label encoding is applied in order to convert the categorical into numerical ones to make the data ready to efficiently apply predictive analytics. The dataset after label encoding is depicted in Figure.

	Student_ID	Age	Gender	Family_Income	Internet_Access	Study_Hours_per_Day	Attendance_Rate	Assignment_Delay_Days	Travel_Time_Minutes
count	10000.00000	10000.00000	10000.000000	9500.000000	10000.000000	9500.000000	10000.00000	10000.000000	10000.00000
mean	5000.50000	21.02606	0.498900	38377.247474	0.876900	4.014592	81.73683	1.799700	30.17926
std	2886.89568	2.13981	0.500024	20496.232179	0.328568	1.295450	8.22093	1.344307	11.91887
min	1.00000	17.00000	0.000000	25000.000000	0.000000	0.500000	38.20000	0.000000	5.00000
25%	2500.75000	19.50000	0.000000	25000.000000	1.000000	3.160000	76.40000	1.000000	21.90000
50%	5000.50000	21.00000	0.000000	29740.500000	1.000000	4.000000	81.80000	2.000000	30.20000
75%	7500.25000	22.50000	1.000000	44520.000000	1.000000	4.870000	87.30000	3.000000	38.40000
max	10000.00000	29.60000	1.000000	316601.000000	1.000000	8.980000	100.00000	8.000000	74.90000

Fig. 5. Dataset after label encoding.

C. 6. Descriptive Analytics and Diagnostic Analytics: Comprehensive Insight Discovery

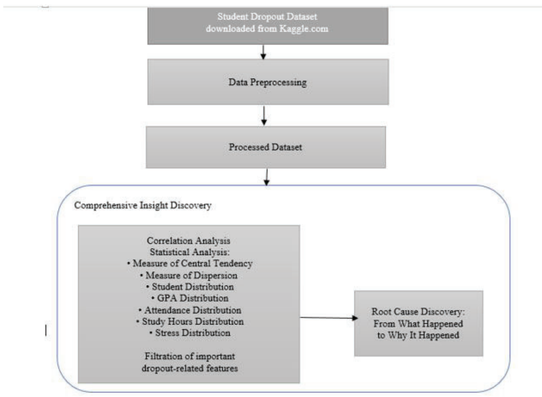


Fig. 6. Architecture of the proposed descriptive and diagnostic analytics framework.

Descriptive Analytics: Further analysis aims to identify the major academic and behavioral factors influencing student dropout and low GPA. In particular, the study investigates the impact of attendance rate, study hours per day, stress level, and academic risk on student performance and dropout probability.

D. Diagnostic Analytics:

GPA, Attendance Rate, and Stress Index are identified as the three most influential features through the correlation analysis depicted in Figure 11, and the target attribute Dropout is found to be highly correlated with GPA. The analysis indicates that students with lower GPA values exhibit a significantly higher probability of dropout compared to students with better academic performance.

Several machine learning models were implemented for comparative analysis, including Logistic Regression, Decision

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

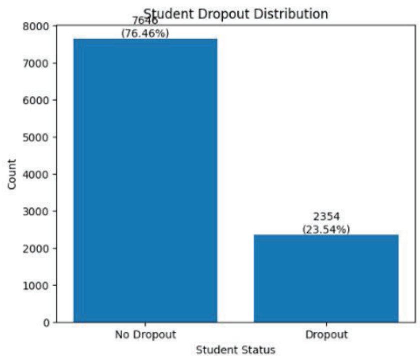


Fig. 7. Student Dropout Distribution.

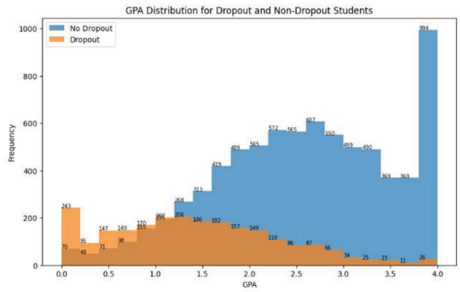


Fig. 8. Distribution of Grade Point Average (GPA) for dropout and non-dropout students.

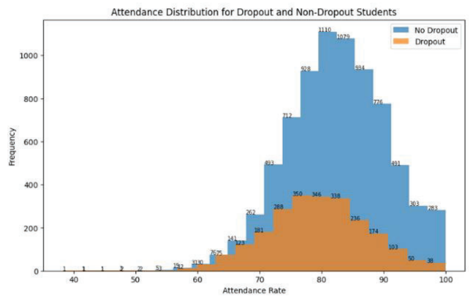


Fig. 9. Attendance rate distribution for dropout and non-dropout students.

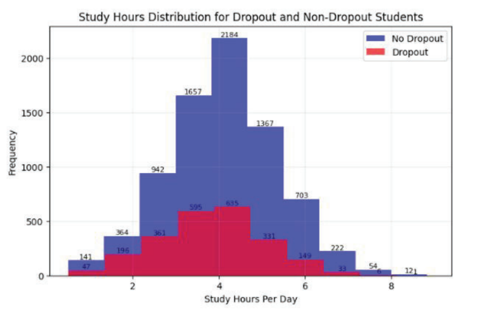


Fig. 10. Study hours per day distribution for dropout and non-dropout students.

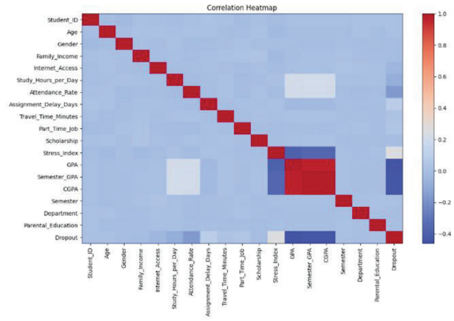


Fig. 11. Correlation heatmap of the student dropout dataset.

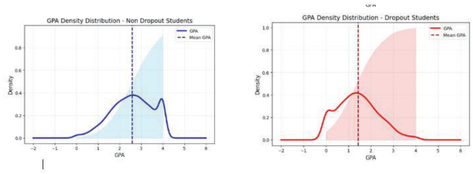


Fig. 12. GPA risk analysis for dropout and non-dropout students.

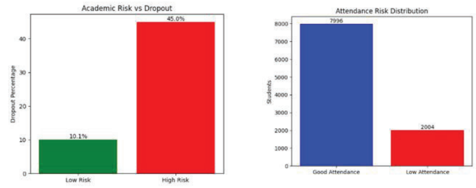


Fig. 13. Attendance risk analysis and Acedemics risk analysis for dropout and non-dropout students.

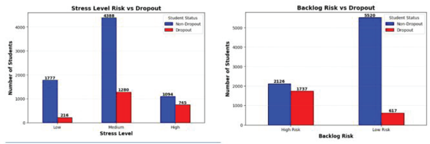


Fig. 14. Stress level and backlog risk analysis for dropout and non dropout students.

Tree, Random Forest, and Support Vector Machine. These algorithms were selected because they represent different learning approaches and are widely used in classification and predictive analytics tasks.

Logistic Regression is a supervised machine learning algorithm used for binary classification problems. It predicts the probability of a target class using the sigmoid function and performs effectively when relationships among variables are linear. Logistic Regression is widely preferred because of its simplicity, computational efficiency, and interpretability.

The Logistic Regression model is mathematically represented as:

$$P(Y=1)=\frac{1}{1+e^{-(b_0+b_1x)}} \tag{1}$$

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

where $P(Y = 1)$ represents the probability of the positive class, b_0 represents the bias term, and b_1x represents the weighted input features.

Decision Tree is a non-linear supervised learning algorithm that represents decisions in a tree-like structure. The model splits the dataset into multiple branches based on feature conditions and predicts the target class using decision rules. Decision Trees are easy to interpret and visualize but may suffer from overfitting when trained on complex datasets.

Random Forest is an ensemble learning algorithm that combines multiple Decision Trees to improve prediction accuracy and reduce overfitting. Each tree is trained using random subsets of the dataset and features. The final prediction is obtained through majority voting among the trees. Due to its robustness and ability to handle high-dimensional data, Random Forest often provides better performance compared to individual decision trees.

Support Vector Machine (SVM) is a supervised classification algorithm that identifies the optimal hyperplane for separating different classes. SVM performs effectively in high-dimensional feature spaces and is widely used for classification tasks. However, the performance of SVM depends on appropriate feature scaling and parameter tuning.

These machine learning models were trained using the prepared training dataset and evaluated using the testing dataset to compare their prediction accuracy and effectiveness for student dropout prediction.

E. Model Evaluation

The performance of the machine learning models was evaluated using accuracy as the primary evaluation metric. Accuracy measures the proportion of correctly predicted instances to the total number of instances in the dataset. The evaluation process was carried out using the testing dataset to ensure an unbiased assessment of model performance on unseen data.

Accuracy is mathematically represented as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

where TP represents True Positives, TN represents True Negatives, FP represents False Positives, and FN represents False Negatives.

In addition to accuracy, confusion matrix analysis was also performed to evaluate the classification performance of the models. The confusion matrix provides detailed information regarding correctly and incorrectly classified instances for both dropout and non-dropout classes.

The machine learning models were compared based on their prediction accuracy to identify the most effective algorithm for student dropout prediction. Experimental results demonstrated that Random Forest achieved the highest prediction accuracy among all implemented models, followed closely by Logistic Regression.

TABLE I
MODEL COMPARISON

Model	Accuracy
Random Forest	81.2%
Logistic Regression	81.0%
Support Vector Machine	77.25%
Decision Tree	74.1%

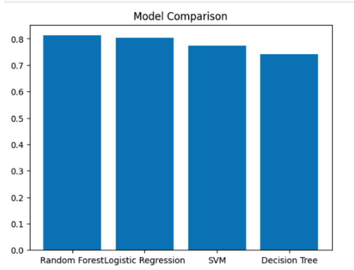
IV. RESULTS AND DISCUSSION

A. Model Performance Analysis

The performance of different machine learning models was evaluated using accuracy as the metric. Among the models, Random Forest achieved the highest accuracy of 81.2

The comparable performance of Logistic Regression and Random Forest suggests that both linear and non-linear relationships exist in the dataset. Random Forest performs slightly better due to its ability to capture complex feature interactions and reduce overfitting. The lower performance of Decision Tree may be due to its tendency to overfit the training data. Support Vector Machine shows moderate performance, indicating sensitivity to feature distribution and scaling.

B. Graphical Analysis



Comparison of Machine Learning Models

Fig.15 illustrates the performance comparison of different machine learning models used for student dropout prediction. Among all the models, Random Forest achieved the highest accuracy of 81.2

C. Feature Importance Analysis

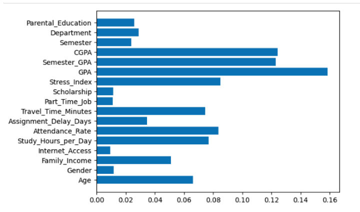


Fig. 15. Feature Importance for Student Dropout Prediction

Fig. 16 shows the importance of different features in predicting student dropout. Academic factors such as GPA and attendance have higher importance, indicating their strong influence on student retention. Behavioral attributes such as study hours also contribute significantly. These findings provide useful insights for early intervention strategies.

D. *Confusion matrices Analysis*

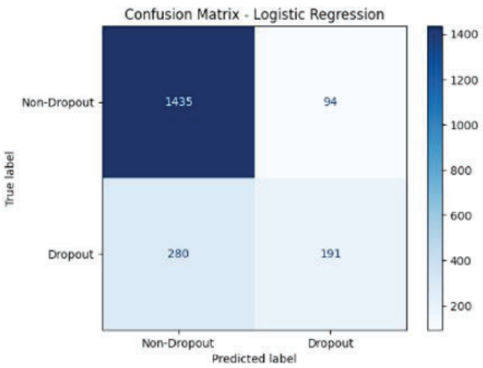


Fig. 16. Confusion matrices of the evaluated machine learning models for student dropout prediction.

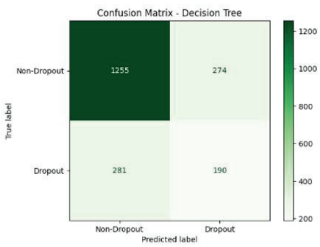


Fig. 17. Confusion matrices of the evaluated machine learning models for student dropout prediction

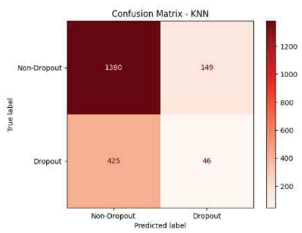


Fig. 18. Confusion matrices of the evaluated machine learning models for student dropout prediction

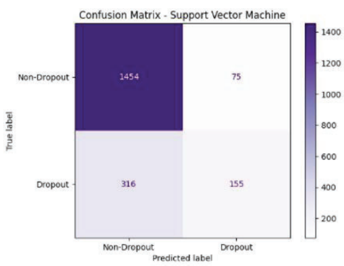


Fig. 19. Confusion matrices of the evaluated machine learning models for student dropout prediction

The confusion matrix was used to analyze the classification performance of the Random Forest model in detail. The matrix provides information about correctly classified and misclassified instances for both dropout and non-dropout classes.

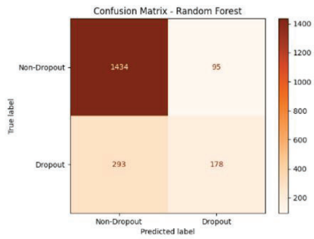


Fig. 20. Confusion matrices of the evaluated machine learning models for student dropout prediction

The results indicate that the Random Forest model correctly classified the majority of students in both categories, demonstrating strong predictive capability. The number of false predictions was comparatively low, which indicates that the model can effectively identify students who are at risk of dropping out.

The confusion matrix analysis also shows that the model achieved balanced classification performance without significant bias toward a particular class. This is important in educational prediction systems because both dropout and non-dropout students must be identified accurately for effective intervention strategies.

E. *Comparative Discussion*

The comparative analysis of machine learning models demonstrates that ensemble learning methods provide better performance for student dropout prediction. Random Forest achieved the highest accuracy because it combines multiple decision trees and reduces the risk of overfitting. The algorithm can effectively capture complex relationships among academic, behavioral, and socio-economic factors.

Logistic Regression also achieved competitive performance, indicating that the dataset contains significant linear relationships among the features. Due to its simplicity and interpretability, Logistic Regression remains a useful model for educational prediction tasks.

Support Vector Machine achieved moderate performance and required proper feature scaling for effective classification. Although SVM performs well in high-dimensional spaces, its performance depends heavily on parameter tuning and feature distribution.

Decision Tree produced the lowest accuracy among the evaluated models. This may be due to overfitting and sensitivity to variations in training data. Individual decision trees often fail to generalize effectively on unseen data compared to ensemble methods such as Random Forest.

F. *ROC Curve:*

ROC Curve Analysis: Figure X illustrates the ROC curves of the top-performing machine learning algorithms. Logistic Regression achieved the highest Area Under the Curve (AUC) of 0.819, demonstrating the strongest capability to distinguish

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

between dropout and non-dropout students. AdaBoost and CatBoost followed closely with an AUC of 0.812, while Random Forest achieved 0.804. SVM obtained the lowest AUC (0.773) among the selected models. Since all AUC values are significantly greater than 0.5, the evaluated models exhibit effective predictive performance, with Logistic Regression providing the best balance between sensitivity and specificity for the student dropout prediction task.

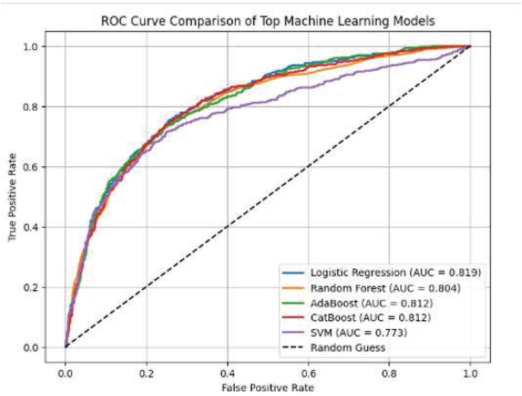


Fig. 21. ROC curve comparison of the top-performing machine learning models for student dropout prediction.

G. *Practical Implications*

The results of this study can help educational institutions identify students who are at risk of dropping out at an early stage. By analyzing important factors such as GPA, attendance, and study behavior, institutions can develop targeted intervention strategies to improve student retention.

Machine learning-based prediction systems can support educators and administrators in making data-driven decisions. Early identification of at-risk students enables institutions to provide academic counseling, mentoring, and additional support services to improve student success rates.

The explainable nature of the proposed approach also improves transparency and trust in educational analytics systems. Understanding the factors influencing dropout prediction can help educators implement more effective policies and student support programs.

V. LIMITATIONS AND FUTURE ENHANCEMENTS

Although the proposed machine learning approach achieved promising results for student dropout prediction, the study has certain limitations. The prediction models were trained and evaluated using a single dataset, which may limit the generalization capability of the models across different educational environments.

The dataset mainly focused on academic, behavioral, and socio-economic factors. Additional attributes such as psychological factors, emotional well-being, family support, and extracurricular activities were not included in the analysis. Incorporating such features may further improve prediction performance and provide deeper insights into student dropout behavior.

Another limitation is that the study primarily used traditional machine learning algorithms. More advanced deep learning techniques and ensemble boosting methods such as XGBoost and LightGBM may provide improved prediction accuracy for larger and more complex datasets.

Future work can focus on implementing explainable artificial intelligence techniques in greater depth to improve transparency and interpretability of prediction systems. Real-time educational analytics systems can also be developed to continuously monitor student performance and provide early intervention recommendations for educational institutions.

VI. CONCLUSION

This study presented a comparative and explainable machine learning approach for student dropout prediction using educational data. Multiple machine learning algorithms, including Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine, were implemented and evaluated to identify the most effective model for predicting student dropout.

The experimental results demonstrated that Random Forest achieved the highest prediction accuracy of 81.2%, followed closely by Logistic Regression with 81%. Support Vector Machine and Decision Tree showed comparatively lower performance. The findings indicate that ensemble learning techniques are more effective in capturing complex relationships among academic, behavioral, and socio-economic factors.

Feature importance analysis revealed that attributes such as GPA, attendance, and study behavior significantly influence student dropout prediction. These insights can help educational institutions identify at-risk students at an early stage and implement effective intervention strategies to improve student retention and academic success.

The proposed approach also emphasizes explainability in machine learning-based educational analytics. In addition to achieving strong prediction performance, the study provides meaningful insights into the factors contributing to student dropout. This improves transparency and supports data-driven decision-making in educational institutions.

Overall, the results confirm that machine learning techniques can play a significant role in educational prediction systems and assist institutions in reducing dropout rates through early identification and targeted support mechanisms.

VII. FUTURE WORK

Future work may focus on improving predictive performance by incorporating additional student-related attributes, including socioeconomic factors, extracurricular activities, learning management system interactions, and psychological indicators. Advanced deep learning architectures, hybrid ensemble models, and explainable artificial intelligence (XAI) techniques such as SHAP and LIME can also be explored to further enhance both prediction accuracy and interpretability. Additionally, deploying the proposed framework as a real-time

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

early warning system for educational institutions represents a promising direction for future research.

REFERENCES

[1] L. Breiman, ”Random Forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[2] C. Cortes and V. Vapnik, ”Support-Vector Networks,” Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.
[3] F. Pedregosa et al., ”Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
[4] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed., Morgan Kaufmann, 2012.
[5] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed., Springer, 2009.
[6] I. H. Witten, E. Frank, M. Hall, and C. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed., Morgan Kaufmann, 2017.
[7] T. Mitchell, *Machine Learning*, McGraw-Hill, 1997.
[8] S. B. Kotsiantis, ”Machine Learning: A Review of Classification and Combining Techniques,” Artificial Intelligence Review, vol. 26, no. 3, pp. 159–190, 2006.
[9] S. Kotsiantis, K. Pierrakeas, and P. Pintelas, ”Predicting Students’ Performance in Distance Learning Using Machine Learning Techniques,” Applied Artificial Intelligence, vol. 18, no. 5, pp. 411–426, 2004.
[10] S. B. Kotsiantis, ”Use of Machine Learning Techniques for Educational Purposes: A Decision Support System for Forecasting Students’ Grades,” Artificial Intelligence Review, vol. 37, no. 4, pp. 331–344, 2012.
[11] F. Del Bonifro, M. Gabbrielli, L. Lisanti, and S. Zingaro, ”Student Dropout Prediction,” Proceedings of the International Conference on Artificial Intelligence in Education, Springer, 2020.
[12] J. Kabathova and M. Drlik, ”Towards Predicting Student Dropout in University Courses Using Different Machine Learning Techniques,” Applied Sciences, vol. 11, no. 7, 2021.
[13] J. Berens et al., ”Early Detection of Students at Risk: Predicting Student Dropouts Using Administrative Student Data and Machine Learning Methods,” International Journal of Intelligent Systems and Applications in Engineering, vol. 10, no. 1, 2022.
[14] H. Dasi and S. Kanakala, ”Student Dropout Prediction Using Machine Learning Techniques,” International Journal of Intelligent Systems and Applications in Engineering, vol. 10, no. 4, pp. 408–414, 2022.