An Intelligent Machine Learning-Based Student Dropout Prediction System Using Interactive GUI

NUTHI BHAVESH VARDHAN

PG Scholar. Department of MCA, DNR College, Bhimavaram, Andhra Pradesh

K.Venkatesh

(Assistant Professor), Master of Computer Applications, DNR College, Bhimavaram, Andhra Pradesh

ABSTRACT

Student dropout is a critical issue affecting educational institutions worldwide, leading to financial losses, reduced institutional reputation, and negative societal impact. Early identification of students at risk of dropping out enables institutions to take preventive actions and improve student retention. This project presents an intelligent Student Dropout Prediction System that leverages machine learning algorithms and an interactive graphical user interface (GUI) to predict whether a student is likely to continue or drop out. The system utilizes structured student data containing key attributes such as marital status, application mode, course, previous qualifications, tuition fee status, scholarship information, gender, and age at enrollment. These attributes are selected as essential features influencing student retention. The dataset is preprocessed using techniques like categorical encoding and feature transformation through one-hot encoding to make it suitable for machine learning models. Three powerful classification algorithms are implemented in the system: Random Forest, Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost). These algorithms are chosen due to their robustness, accuracy, and ability to handle complex patterns in data. The system allows users to dynamically select any of these models for training and evaluation. The performance of the model is measured using accuracy, providing users with insights into prediction reliability.

A user-friendly GUI is developed using Tkinter, enabling users to upload datasets, train models, and perform predictions without requiring programming knowledge. The system dynamically generates input fields based on dataset features, allowing users to input new student details for real-time predictions. This enhances usability and accessibility for educational administrators. The prediction process involves transforming user inputs into a format compatible with trained models and ensuring consistency with training features. The output is mapped back to human-readable labels, indicating whether a student is likely to drop out or continue. This system demonstrates the potential of machine learning in educational data mining and decision support systems. By providing accurate predictions and an easy-to-use interface, the proposed system can assist institutions in identifying at-risk students early and implementing targeted interventions. Future enhancements may include integration with real-time databases, advanced deep learning models, and additional performance metrics.

KEYWORDS: Student Dropout Prediction, Machine Learning, Random Forest, SVM, XGBoost, Educational Data Mining, Classification, Tkinter GUI, Predictive Analytics

I. INTRODUCTION

Education plays a vital role in personal and societal development. However, student dropout remains a significant challenge faced by educational institutions globally. Dropout not only affects students' future opportunities but also impacts institutional performance and national development. Therefore, identifying students at risk of dropping out at an early stage is essential for implementing timely interventions. With the advancement of technology, machine learning has emerged as a powerful tool for predictive analysis in various domains, including education. Educational Data Mining (EDM) focuses on extracting meaningful insights from educational datasets to improve learning outcomes and institutional effectiveness. By analyzing historical student data, machine learning models can identify patterns and predict future outcomes such as academic success or dropout. This project introduces a Student Dropout Prediction System that combines machine learning algorithms with a graphical user interface for ease of use. The system is designed to help educational institutions predict student dropout based on key attributes such as demographic details, academic background, and financial status. These factors significantly influence a student's ability to continue education.

The system uses three classification algorithms: Random Forest, Support Vector Machine (SVM), and XGBoost. Random Forest is an ensemble learning method known for its high accuracy and robustness. SVM is effective in high-dimensional spaces and works well for classification tasks. XGBoost is a gradient boosting algorithm that provides superior performance and efficiency. By offering multiple models, the system allows users to compare and select the most suitable algorithm. A key feature of this system is its user-friendly GUI developed using Tkinter. Unlike traditional machine learning applications that require programming expertise, this system enables users to interact with the model through simple interface elements such as buttons and dropdown menus. Users can upload datasets, train models, and input new data for prediction. The system also incorporates preprocessing techniques such as handling categorical variables and feature alignment to ensure accurate predictions. The use of one-hot encoding ensures that categorical data is converted into numerical form suitable for machine learning algorithms. Overall, this project demonstrates how machine learning can be effectively applied in the education sector to address real-world challenges. By predicting student dropout early, institutions can provide support such as counseling, financial aid, and academic assistance to improve retention rates. The system serves as a decision-support tool for administrators and educators.

II. LITERATURE SURVEY (WITH EXISTING METHODS)

Student dropout prediction has been widely studied in the field of Educational Data Mining (EDM). Various researchers have proposed machine learning and statistical approaches to identify at-risk students and improve retention rates. Early studies focused on statistical methods such as logistic regression and decision trees. Logistic regression was commonly used due to its simplicity and interpretability. It models the probability of dropout based on independent variables such as academic performance and socio-economic factors. However, its limitation lies in handling non-linear relationships in data. Decision trees gained popularity as they provide easy-to-understand rules for classification. They split data based on feature importance and generate interpretable models. However, decision trees often suffer from over fitting, reducing their generalization ability. To overcome these limitations, ensemble methods such as Random Forest were introduced. Random Forest combines multiple decision trees to improve accuracy and reduce over fitting. Studies have shown that Random Forest performs well in predicting student performance and dropout due to its ability to handle large datasets and complex relationships.

Support Vector Machines (SVM) has also been widely used in dropout prediction. SVM works by finding the optimal hyper plane that separates data into classes. It is effective in high-dimensional spaces and provides good classification performance. However, it requires careful parameter tuning and may not scale well with very large datasets. In recent years, boosting algorithms such as XGBoost have gained significant attention. XGBoost is an optimized gradient boosting technique that provides high accuracy, speed, and scalability. Research studies indicate that XGBoost often outperforms traditional algorithms in classification tasks, including student dropout prediction. Deep learning approaches have also been explored, including Artificial Neural Networks (ANNs) and Recurrent Neural Networks (RNNs). These models can capture complex patterns in data but require large datasets and computational resources. They are less interpretable compared to traditional machine learning models. Several studies emphasize the importance of feature selection in improving prediction accuracy. Features such as attendance, grades, financial status, and demographic factors are commonly used. Proper preprocessing techniques, including handling missing values and encoding categorical data, are crucial for model performance.

In addition to prediction models, recent research highlights the importance of user-friendly systems for practical implementation. Many existing systems lack intuitive interfaces, limiting their usability for non-technical users. The proposed system builds upon these research findings by integrating multiple machine learning algorithms and providing a GUI-based interface. This combination enhances both prediction accuracy and usability, making the system suitable for real-world applications in educational institutions.

III. EXISTING SYSTEM

Existing systems for student dropout prediction primarily rely on traditional statistical methods or standalone machine learning models without user-friendly interfaces. Many institutions use basic data analysis techniques to identify at-risk students, such as manual evaluation of grades and attendance records. These methods are time-consuming, error-prone, and lack predictive capabilities. Some systems use logistic regression and decision trees to predict student dropout. While these models provide moderate accuracy, they are limited in handling complex relationships between features. Additionally, these systems often require technical expertise to operate, making them inaccessible to non-technical users such as academic administrators. Another limitation of existing systems is the lack of flexibility in model selection. Most systems rely on a single algorithm, which may not perform well across different datasets. This reduces the overall effectiveness of prediction. Furthermore, many existing solutions do not include proper data preprocessing techniques such as encoding categorical variables or aligning feature columns. This leads to inconsistent predictions and reduced accuracy. The absence of real-time prediction capabilities also limits their practical usability.

User interface is another major drawback. Most systems are either command-line based or integrated into complex software environments, making them difficult to use. Users cannot easily input new student data or visualize results. In contrast, the proposed system addresses these limitations by integrating multiple machine learning algorithms, implementing proper preprocessing techniques, and providing an interactive GUI. This improves prediction accuracy, usability, and accessibility, making it a more effective solution for educational institutions.

IV. PROPOSED METHOD

The proposed system is an intelligent Student Dropout Prediction System that utilizes machine learning techniques combined with a user-friendly graphical interface to identify students at risk of dropping out. Unlike traditional systems, this solution integrates multiple classification algorithms and real-time prediction capabilities, making it both accurate and accessible. The system takes input data containing essential student attributes such as demographic details, academic background, and financial status. These features are carefully selected based on their strong influence on student retention, as identified in recent studies. Research shows that factors like academic performance, demographic characteristics, and engagement data significantly impact dropout prediction accuracy.

The proposed system implements three powerful machine learning algorithms: Random Forest, Support Vector Machine (SVM), and XGBoost. These algorithms are capable of handling complex and non-linear relationships in educational data. Studies indicate that ensemble methods and boosting algorithms often achieve higher prediction accuracy compared to traditional methods. A key enhancement in this system is the integration of a Tkinter-based GUI, which allows users to upload datasets, select models, train them, and make predictions without requiring technical expertise. The system dynamically

generates input fields based on dataset features, ensuring flexibility and adaptability. Additionally, the system includes preprocessing steps such as categorical encoding and feature alignment to maintain consistency between training and prediction phases. This ensures reliable outputs and minimizes errors. The proposed system aims to provide early detection of at-risk students, enabling educational institutions to implement timely interventions such as counseling, financial support, and academic assistance. Overall, it enhances decision-making, improves retention rates, and contributes to the advancement of educational data mining.

V. IMPLEMENTATION

The implementation of the Student Dropout Prediction System is carried out using Python, leveraging libraries such as Tkinter for GUI development, Pandas for data manipulation, and Scikit-learn and XGBoost for machine learning. The system begins with the dataset loading phase. Users can upload a CSV file through the GUI using a file dialog. The dataset is read using Pandas, which allows efficient handling of structured data. Once the dataset is loaded, the system extracts essential features required for training. Data preprocessing is a crucial step in implementation. The target variable, labeled as "Target," is converted into categorical format and encoded into numerical values using label encoding. Feature variables are transformed using one-hot encoding to convert categorical data into numerical form suitable for machine learning algorithms. This ensures compatibility with classification models. After preprocessing, the dataset is split into training and testing sets using the train_test_split function. Typically, 80% of the data is used for training and 20% for testing. This allows the system to evaluate model performance on unseen data.

The system supports three machine learning models: Random Forest, Support Vector Machine (SVM), and XGBoost. Users can select the desired model from a dropdown menu in the GUI. Based on the selection, the corresponding model is initialized and trained using the training dataset. Once the model is trained, predictions are made on the test dataset. The performance of the model is evaluated using accuracy score, which is displayed on the GUI. Accuracy is a widely used metric in classification tasks and provides a quick assessment of model performance. For prediction, the system generates input fields dynamically based on essential features. Users can select values from dropdown menus corresponding to each feature. The input data is converted into a Data Frame and processed using the same encoding techniques applied during training. To ensure consistency, the system aligns the input data with the training feature columns by adding missing columns with default values. This step is critical to avoid mismatches between training and prediction data. The trained model is then used to predict the output class, which is mapped back to its original label using label mapping. The result is displayed on the GUI as either "Dropout" or "Continue." The entire system is designed to be interactive and user-friendly, making it suitable for non-technical users. By combining machine learning with an intuitive interface, the implementation bridges the gap between complex analytics and practical usability.

VI. ALGORITHMS

The proposed system utilizes three supervised machine learning algorithms: Random Forest, Support Vector Machine (SVM), and XGBoost. These algorithms are widely used in classification problems and have demonstrated strong performance in student dropout prediction tasks. Random Forest is an ensemble learning algorithm that combines multiple decision trees to improve prediction accuracy and reduce overfitting. Each tree is trained on a random subset of the data, and the final prediction is determined by majority voting. This approach enhances robustness and handles large datasets effectively. Research indicates that Random Forest achieves high accuracy in dropout prediction due to its ability to capture complex relationships.

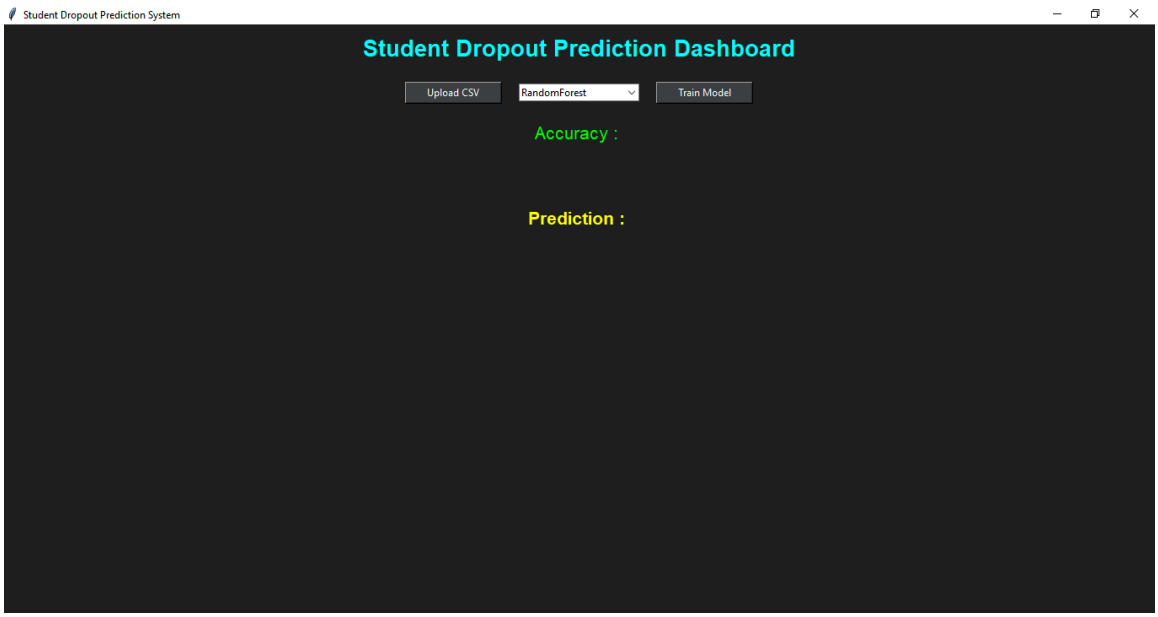
Support Vector Machine (SVM) is a powerful classification algorithm that finds the optimal hyper plane to separate data into different classes. It is particularly effective in high-dimensional spaces and works well when the number of features is large. SVM can handle both linear and non-linear classification using kernel functions. It has been widely applied in educational data mining for predicting student outcomes. XGBoost (Extreme Gradient Boosting) is an advanced boosting algorithm that builds models sequentially by minimizing prediction errors. It is known for its high efficiency, scalability, and accuracy. XGBoost incorporates regularization techniques to prevent over fitting and supports parallel processing. Recent studies show that boosting algorithms often outperform traditional models in classification tasks, including dropout prediction. These algorithms are selected to provide flexibility and allow users to compare performance across different models. By integrating multiple algorithms, the system ensures better prediction accuracy and adaptability to different datasets.

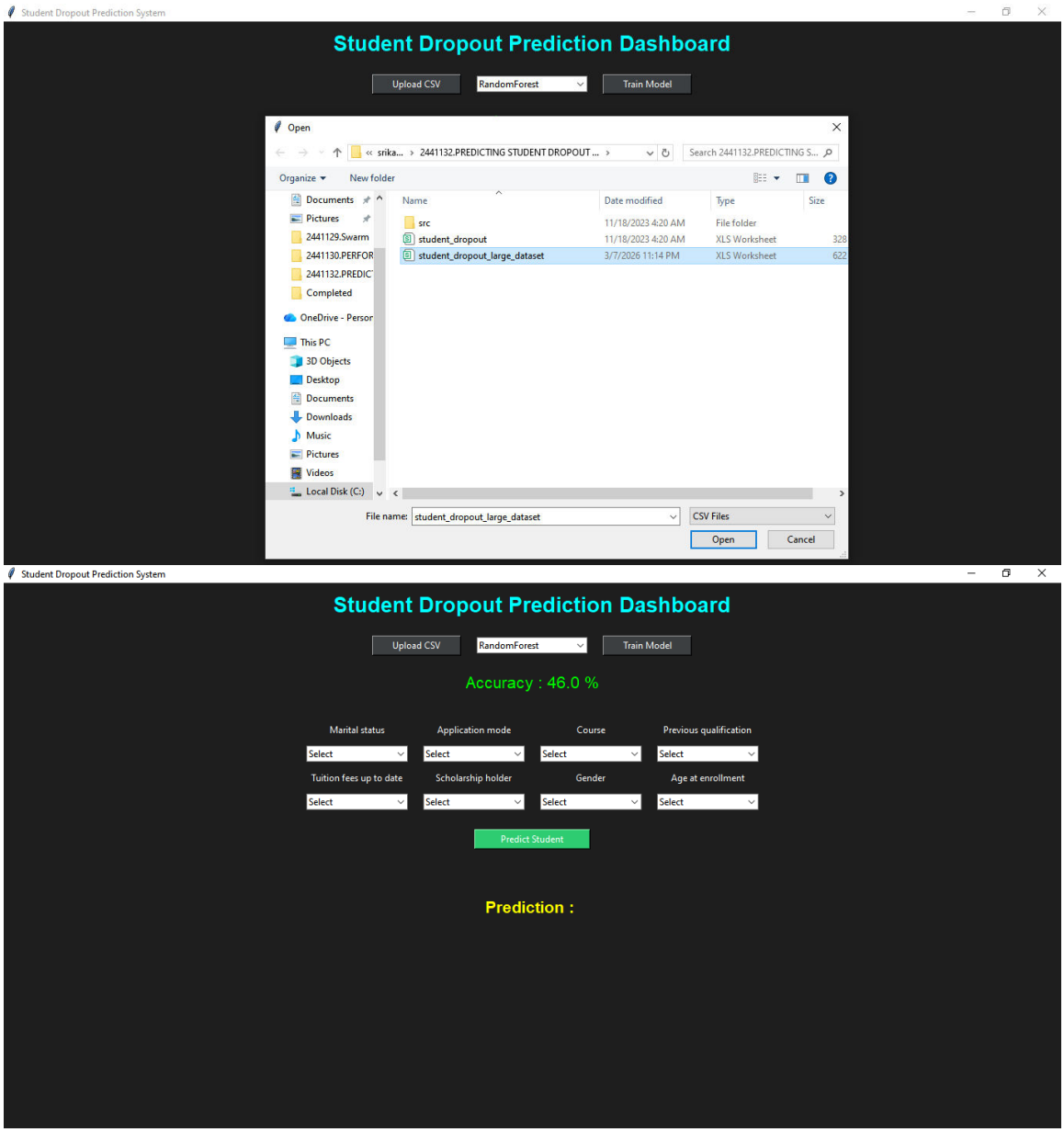
VII. SYSTEM DESIGN

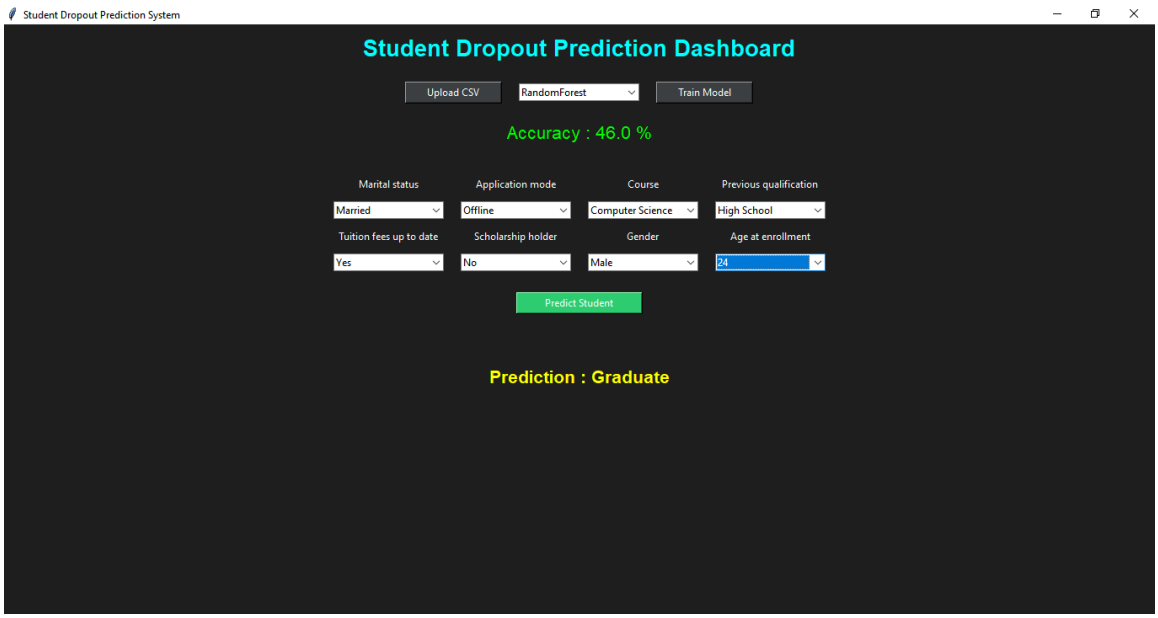
The system design of the Student Dropout Prediction System follows a modular architecture that integrates data processing, machine learning, and user interface components. The design ensures scalability, usability, and efficiency. The system consists of three main modules: Data Management Module, Machine Learning Module, and User Interface Module. The Data Management Module is responsible for handling dataset input and preprocessing. It allows users to upload CSV files containing student data. Once the data is loaded, the module extracts essential features and performs preprocessing steps such as handling missing values, encoding categorical variables, and feature selection. One-hot encoding is applied to convert categorical data into numerical form. The Machine Learning Module is responsible for model training, evaluation, and prediction. It takes preprocessed data as input and splits it into training and testing sets. The module supports multiple algorithms, including Random Forest, SVM, and XGBoost. Users can select the desired model through the interface. The module trains the selected model and evaluates its performance using accuracy metrics. During prediction, the module processes user input data and ensures it matches the format of the training data. Feature alignment is performed by adding missing columns and maintaining the same order as the training dataset.

This ensures consistent and accurate predictions. The User Interface Module is developed using Tkinter and serves as the front-end of the system. It provides an interactive dashboard where users can upload datasets, select models, train models, and input new data for prediction. Dropdown menus are dynamically generated based on dataset features, improving usability. The system follows a sequential workflow: dataset upload → preprocessing → model training → evaluation → prediction. Each step is interconnected, ensuring smooth data flow and functionality. From a design perspective, the system emphasizes modularity, allowing each component to function independently. This makes it easier to maintain and extend the system. For example, additional algorithms or features can be integrated without affecting other modules. The architecture also supports scalability. As more data becomes available, the system can handle larger datasets and improve prediction accuracy. Furthermore, the GUI design ensures accessibility for non-technical users. Overall, the system design provides a robust framework for student dropout prediction, combining efficiency, flexibility, and ease of use.

SYSTEM DESIGN IMAGES




Student Dropout Prediction System
Student Dropout Prediction Dashboard
Upload CSV
RandomForest
Train Model
Open
Organize
New folder
Name
Date modified
Type
Size
src
11/18/2023 4:20 AM
File folder
student_dropout
11/18/2023 4:20 AM
XLS Worksheet
student_dropout_large_dataset
3/7/2026 11:14 PM
XLS Worksheet
File name: student_dropout_large_dataset
CSV Files
Open
Cancel
Student Dropout Prediction System
Student Dropout Prediction Dashboard
Upload CSV
RandomForest
Train Model
Accuracy : 46.0 %
Marital status
Application mode
Course
Previous qualification
Tuition fees up to date
Scholarship holder
Gender
Age at enrollment
Select
Predict Student
Prediction :



VIII. CONCLUSION

The Student Dropout Prediction System demonstrates the effective application of machine learning techniques in addressing a critical issue in the education sector. By leveraging algorithms such as Random Forest, Support Vector Machine, and XGBoost, the system provides accurate predictions of student dropout risk based on key attributes. One of the major strengths of the system is its integration of a user-friendly graphical interface. This allows users, including educators and administrators, to interact with the system without requiring technical expertise. The ability to upload datasets, train models, and generate predictions in real-time enhances the practicality of the solution. The system also incorporates essential preprocessing techniques such as feature encoding and alignment, ensuring consistency between training and prediction phases. This improves the reliability and accuracy of the predictions. Research shows that machine learning models can achieve high accuracy in predicting student dropout when relevant features are used. The proposed system aligns with these findings by selecting important features and applying robust algorithms.

Despite its advantages, the system has certain limitations. It relies on the quality and completeness of the input dataset. Inaccurate or missing data may affect prediction performance. Additionally, the system currently uses basic evaluation metrics, and future improvements can include advanced metrics such as precision, recall, and F1-score. Future enhancements may include integration with real-time institutional databases, implementation of deep learning models, and development of web-based interfaces for wider accessibility. Incorporating explainable AI techniques can also improve transparency and trust in predictions. In conclusion, the proposed system provides a practical and efficient solution for early detection of at-risk students. By enabling timely intervention, it contributes to improving student retention rates and overall educational outcomes.

REFERENCES

1. Marcolino et al., "Student dropout prediction through machine learning optimization," *Scientific Reports*, 2025.
2. Cho et al., "Dropout Prediction for University Students Using ML," *Applied Sciences*, 2023.
3. Hassan et al., "Predicting Student Dropout Rates Using Supervised ML," *Applied Sciences*, 2024.
4. Villar et al., "Comparative Study of ML Algorithms," *Discover AI*, 2024.
5. Fitriana et al., "Prediction of Student Dropout and Academic Achievement," *IJCS*, 2024.
6. Sulak & Koklu, "Predicting Student Dropout Using ML," 2024.
7. Osemwegie & Amadin, "Student Dropout Prediction Using ML," 2023.
8. Jiménez-Gutiérrez et al., "ML Techniques for School Dropout Prediction," *Scientific Reports*, 2024.
9. Sandivar-Rosas et al., "Systematic Review on Dropout Prediction," 2023.
10. Finnish Study on ML-based Dropout Prediction, 2024.
11. Dasi & Kanakala, "ML Techniques for Dropout Prediction," 2022.
12. Krüger et al., "Explainable ML for Dropout Prediction," 2023.
13. Ndunagu et al., "Deep Learning for Attrition Prediction," 2024.
14. Jimenez Martinez et al., "Early Detection of At-Risk Students," 2024.
15. Kim et al., "Dropout Prediction with Factor Analysis," 2023.