Dropout Defender: A Machine Learning Approach to Lower Dropout Rates

Harmi Mathukiya
Department of Information Technology
A. P. Shah Institute Of Technology
Thane, Maharashtra, India
21104044.harmi.mathukiya@gmail.com

Avantika More
Department of Information Technology
A. P. Shah Institute Of Technology
Thane, Maharashtra, India
21104033.avantika.more@gmail.com

Atharva Mohape
Department of Information Technology
A. P. Shah Institute Of Technology
Thane, Maharashtra, India
21104121mohapeatharva@gmail.com

Sahil Mohite
Department of Information Technology
A. P. Shah Institute Of Technology
Thane, Maharashtra, India
21104099.sahil.mohite@gmail.com

Ms.Apeksha Mohite
Department of Information Technology
A. P. Shah Institute Of Technology
Thane, Maharashtra, India
mam@apsit.edu.in

***Abstract*—Student dropout is a persistent issue in education, affecting both institutions and students' futures. The Dropout Defender system aims to predict and address dropout risks by analyzing various factors, including academic performance, student behavior, and engagement levels. The system utilizes machine learning techniques such as Decision Trees, Random Forest, Support Vector Machines (SVM), and Neural Networks to assess the likelihood of a student dropping out. To improve prediction accuracy, algorithms like XGBoost and Logistic Regression are employed, while K-Means clustering helps categorize students based on their risk level. A real-time, web-based dashboard allows educators, mentors, and parents to track student progress and intervene when necessary. Through data-driven insights and personalized recommendations, the system offers a more effective approach to preventing student dropouts.**

***Index Terms*—Keywords: Machine Learning, Decision Trees, Random Forest, SVM, Neural Networks, XGBoost, Logistic Regression, K-Means Clustering, Dropout Prediction, Educational Data Mining.**

I. Introduction

Student dropout continues to be a significant challenge in the education system, affecting not only the individuals involved but also society at large. Various factors contribute to students falling behind in their studies, leading to disengagement and ultimately dropping out. These factors can include struggles with academics, financial constraints, insufficient family support, or emotional and psychological difficulties. While schools and colleges attempt to provide support, many of these issues go undetected without a proactive system in place. As dropout rates continue to rise, it becomes increasingly urgent to implement a structured approach to identify at-risk students and intervene before it's too late.

To address this, machine learning offers a promising solution. By analyzing a variety of student data, including academic performance, attendance, behavioral patterns, and engagement levels, machine learning models can effectively identify students at risk of dropping out. Techniques such as Decision Trees, Random Forests, Support Vector Machines (SVM), and Neural Networks are commonly used to predict dropout risks with high accuracy. These models can process large datasets and provide actionable insights, enabling educators to implement timely interventions that support at-risk students.

The Dropout Defender system integrates these machine learning algorithms to create a comprehensive tool for educators, mentors, and parents. This system predicts dropout risks, segments students into different risk categories, and offers personalized recommendations for improvement. By providing real-time data through an interactive dashboard, the system fosters proactive decision-making, allowing institutions to reduce dropout rates effectively. Through the application of data-driven strategies, the Dropout Defender has the potential to significantly improve student retention and academic success. The system goes beyond simply predicting dropout risks by providing personalized suggestions tailored to the unique needs of each student. By examining factors like academic performance, attendance, participation in extracurricular activities, and overall engagement, Dropout Defender can offer specific recommendations for mentors or parents to act upon. This proactive approach ensures that students at risk receive the timely support necessary to help them stay in school and excel academically.

Additionally, the system features an interactive online dashboard that enables teachers, mentors, and parents to track student progress in real time. This dashboard presents in-depth reports on each student's performance, identifying the primary factors that could lead to dropout. It also allows stakeholders to monitor the effectiveness of interventions, offering valuable insights that can help refine and improve future strategies. Ultimately, Dropout Defender aims to reduce dropout rates by utilizing the power of machine learning to predict, intervene, and support students effectively. By fostering a data-driven approach to education, this system can play a crucial role in helping institutions identify at-risk students early and take proactive measures to improve retention rates. The system's potential for scalability and adaptability ensures its applicability across various educational settings, making it a promising tool in the fight against student dropout.

II. Literature review

[1] The research paper titled ”Predicting Student Dropout in Higher Education: A Machine Learning Approach,” authored by R. N. Goh, M. R. B. M. Isa, and M. R. Ab. Ghani, this research focuses on the use of machine learning algorithms to predict student dropout in higher education. The authors explore various classification algorithms such as decision trees, random forests, and support vector machines (SVM) to identify students at risk. The research highlights the importance of data features such as grades, attendance, and socio-economic factors in predicting dropouts. The study provides a framework for predicting student dropout using real-time data, which aligns with our work on leveraging machine learning to predict and mitigate dropout risks in Dropout Defender.

[2] The research paper titled ”Early Prediction of Student Dropout Using Data Mining Techniques,” authored by A. Gupta, S. L. Shukla, and S. Tripathi, this research investigates the use of data mining techniques to predict student dropout in a timely manner. The authors apply clustering and classification algorithms to identify early warning signs of at-risk students. By analyzing student demographic data, academic performance, and engagement levels, the paper highlights the role of data mining in early intervention. This study is directly relevant to Dropout Defender, which uses machine learning to identify students who need intervention, emphasizing predictive models to improve student retention.

[3] The research paper titled ”Student Retention and Dropout Prediction Using Machine Learning,” authored by J. P. Thomas, M. S. Ward, and K. M. Smith, In this study, the authors apply several machine learning techniques, including neural networks and random forests, to predict student retention and dropout. They analyze factors such as student behavior, academic history, and social engagement. The research demonstrates how data-driven models can help predict dropout rates and provide the foundation for interventions. The insights from this study are in line with our system's goal of using machine learning algorithms like decision trees and neural networks to predict student dropout and offer recommendations.

[4] The research paper titled ”A Review on Early Warning Systems for Student Dropout,” authored by M. K. Dube, A. S. Kumar, and D. R. Bhagat, This paper provides a comprehensive review of existing early warning systems (EWS) designed to predict student dropout. It covers various machine learning models and approaches, emphasizing the importance of personalized interventions to reduce dropout rates. The authors argue that timely data analysis can be used to provide customized support for at-risk students, a principle that is central to the design of Dropout Defender. This paper has reinforced our understanding of how different predictive models can be used for early intervention.

[5] The research paper titled ”Impact of Social Media on Student Retention and Dropout Prevention,” authored by J. F. Baker, L. B. Armstrong, and E. W. Williams, This research examines the role of social media platforms in student engagement and retention, particularly how interactions through these platforms can help prevent dropouts. It highlights how students who engage more actively in online communities and discussions tend to persist in their academic journeys. This paper complements our project by emphasizing the importance of student engagement, a critical factor considered in our machine learning models to predict dropout risks. It reinforces the need for incorporating engagement metrics into Dropout Defender's prediction system.

[6] The research paper titled ”Using Machine Learning for Predicting Student Dropout in Online Education,” authored by L. R. Park, A. M. Singh, and H. R. Gupta, The paper explores the use of machine learning to predict student dropout specifically in online education environments. By analyzing online behavior data such as course engagement, participation in discussions, and assignment completion, the study finds that machine learning models can effectively predict whether an online student will drop out. This research aligns with the use of behavioral and engagement data in Dropout Defender to predict dropout risks, showcasing the power of data-driven predictions in educational settings.

III. Methodology

A. Problem Statement

Student dropout rates continue to be a significant challenge for educational institutions, affecting academic performance, financial stability, and overall student success. Current methods for identifying at-risk students often rely on limited data, leading to missed opportunities for early intervention.

Solution - The Dropout Defender system leverages machine learning algorithms to predict student dropout risk by analyzing comprehensive data sets, including academic performance, attendance, and behavioral factors. By providing early warnings and personalized intervention strategies, the system enables educators, mentors, and parents to take proactive steps to support at-risk students and reduce dropout rates.

B. System Architecture

The system architecture is designed to predict student dropouts by utilizing multiple layers of data processing and analysis. Initially, the Student Grades and Student Layers represent the raw input data, which include academic performance and other student-related factors like attendance or socio-economic status. The Data Collection Layer is responsible

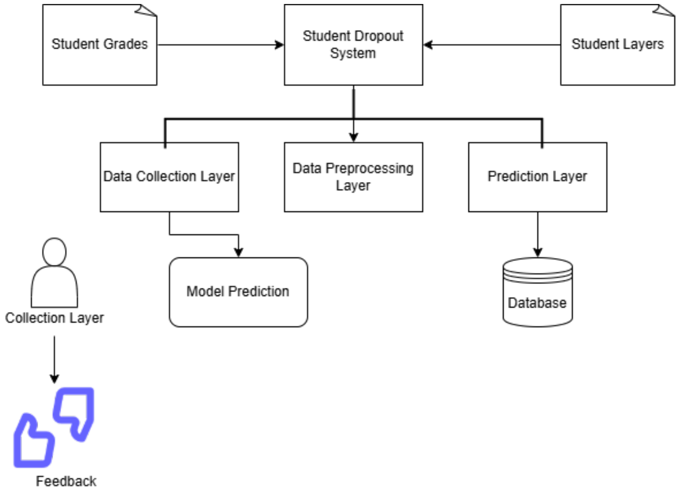


Fig. 1. System Architecture

for gathering this data from various sources, serving as the bridge between the raw inputs and the system's processing components. In the Data Preprocessing Layer, the collected data undergoes cleaning and transformation to ensure it is in a usable format for analysis, addressing issues like missing data or normalization. The Prediction Layer applies machine learning models or statistical techniques to the processed data, making predictions about whether a student is at risk of dropping out. The results of these predictions are then passed to the Model Prediction component, which outputs the likelihood of dropout for each student. Finally, the Database stores all the data, processed information, and predictions, enabling the system to make ongoing updates and analyses. This architecture ensures that student dropout risks can be effectively identified and addressed by educational institutions.

C. *Workflow Diagram*

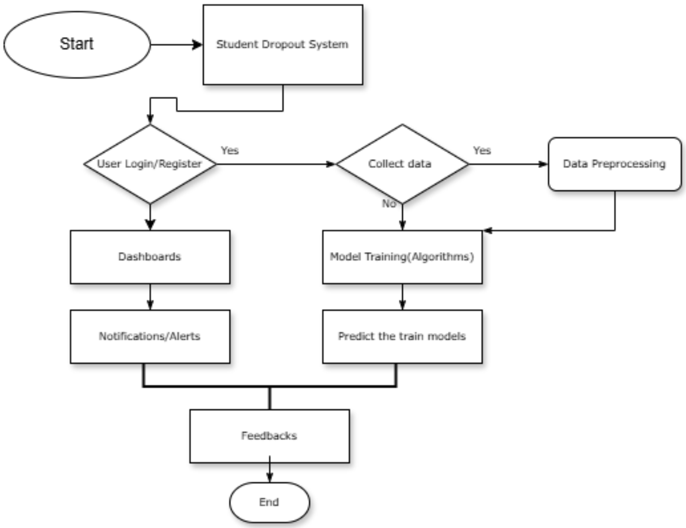


Fig. 2. Workflow Diagram

Figure describes the ”Dropout Defender: A Machine Learning Approach to Lower Dropout Rates” workflow involves several key steps: First, students, parents, and mentors register and log in to the system. The system then collects relevant data, such as academic records, attendance, and behavior patterns. This data is preprocessed for machine learning by cleaning and normalizing it. A predictive model is trained using historical data to estimate the likelihood of a student dropping out. The system then displays progress, predictions, and recommendations on dashboards for students, mentors, and parents. Notifications are sent when dropout risk is high, along with suggested interventions. Feedback from users helps refine predictions, and continuous monitoring ensures that predictions and interventions remain up-to-date.

D. *Algorithms*

1. Random Forest

Random Forest is an ensemble learning method that combines multiple decision trees to improve predictive performance and reduce overfitting. It builds a collection of decision trees by using bootstrapped samples from the training data and random feature selection for each tree.

Ensemble Prediction: The final prediction of the Random Forest is the average (for regression) or the majority vote (for classification) from all individual decision trees.

For classification, the prediction is:

$$\hat{y} = \text{Majority Vote}(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)$$

where $\hat{y}_i$ is the prediction from the i-th decision tree.

Decision Tree Prediction: Each decision tree makes a prediction by recursively splitting the dataset based on features until a terminal node is reached. The decision rule is:

$$\text{Split feature} = \arg\max_{j} \sum_{i=1}^{n} \nVdash(x_i^j \in S)$$

where $\nVdash(x_i^j \in S)$ is the indicator function that returns 1 if x_i^j belongs to set S, and 0 otherwise.

2. Logistic Regression

Logistic Regression is a statistical model used for binary classification. It models the probability that a given input $\mathbf{X}$ belongs to a particular class.

Logistic Function (Sigmoid Function): The logistic regression model applies a sigmoid function to the linear combination of input features $\mathbf{X}$ and their weights $\boldsymbol{\beta}$:

$$P(y = 1 \mid \mathbf{X}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}}$$

where β_0 is the intercept, $\beta_1, \beta_2, \ldots, \beta_n$ are the weights, and $x_1, x_2, \ldots, x_n$ are the input features.

Log-Likelihood Function: The model is trained by maximizing the log-likelihood function, which for binary classification is given by:

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{m} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]$$

where $p_i = P(y_i = 1 \mid \mathbf{X}_i)$ is the predicted probability of the positive class.

3. Naive Bayes

Naive Bayes is a probabilistic classifier based on Bayes' theorem, assuming that the features are conditionally independent given the class.

Bayes' Theorem: The probability of a class C_k given the features $\mathbf{X} = (x_1, x_2, \ldots, x_n)$ is calculated using Bayes' Theorem:

$$P(C_k \mid \mathbf{X}) = \frac{P(C_k) \prod_{i=1}^{n} P(x_i \mid C_k)}{P(\mathbf{X})}$$

where:

- $P(C_k)$ is the prior probability of class C_k,
- $P(x_i \mid C_k)$ is the likelihood of feature x_i given class C_k,
- $P(\mathbf{X})$ is the marginal likelihood of the data.

Class Prediction: The model predicts the class $\hat{C}$ that maximizes the posterior probability:

$$\hat{C} = \arg\max_k P(C_k) \prod_{i=1}^{n} P(x_i \mid C_k)$$

where $\hat{C}$ is the predicted class.

4. KNN

K-Nearest Neighbors (KNN) is a straightforward machine learning algorithm for classification and regression tasks. It determines the class or value of a data point based on the majority vote or average of its nearest neighbors in feature space. The choice of the parameter k, representing the number of neighbors to consider, is crucial for its performance. While KNN is easy to understand and implement, it can be computationally expensive, especially for large datasets, and sensitive to irrelevant features.[8]

The training dataset is represented as:

$$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

where x_i represents the feature vector of the i-th data point and y_i represents its corresponding class label. The query point is denoted as x_q for classification or regression.

The steps of the KNN algorithm can be represented as follows:

1. Calculate the distance between the query point x_q and all data points x_i in the training dataset. Typically, Euclidean distance is used, but other distance metrics such as Manhattan or Minkowski distance can also be used:

$$d(x_q, x_i) = \sqrt{\sum_{j=1}^{d} (x_{q,j} - x_{i,j})^2} \tag{3}$$

where d is the dimensionality of the feature vectors.

2. Select the k nearest neighbors of x_q based on the calculated distances.

3. For classification: Assign the class label y_q to x_q by majority voting among the class labels of its k nearest neighbors:

$$y_q = \text{argmax}_y \sum_{i=1}^{k} I(y_i = y) \tag{4}$$

where $I(\cdot)$ is the indicator function.

4. For regression: Assign the value y_q to x_q by taking the average of the target values of its k nearest neighbors:

$$y_q = \frac{1}{k} \sum_{i=1}^{k} y_i \tag{5}$$

These steps outline the mathematical representation of the K-Nearest Neighbors algorithm for both classification and regression tasks.

5. Decision Tree

A Decision Tree is a machine learning algorithm that recursively partitions data based on input features, aiming to maximize homogeneity within resulting subsets. It selects features for splitting based on criteria like impurity reduction or variance. The process continues until a stopping criterion is met, producing a tree-like structure of decision rules. Decision Trees are easy to interpret but prone to overfitting. They are commonly used for classification and regression tasks due to their simplicity and ability to capture complex relationships in data.[1]

The decision-making process in a Decision Tree can be represented by the following formula [12]:

$$\text{Decision}(x) = \text{SplitCriterion}(x, \theta) \tag{6}$$

In this formula:

- x represents the input features.
- θ represents the split criterion, such as a feature threshold or impurity measure.
- SplitCriterion is a function that determines how to split the data based on the input features and the threshold.

IV. Result and Discussion

In our project Random Forest Algorithm has more accuracy then others which is a supervised machine learning algorithm that is used for Classification and Regression problems in Machine Learning. We know that a forest comprises numerous trees, and the more trees more it will be robust. Similarly, the greater the number of trees in a Random Forest Algorithm, the higher its accuracy and problem-solving ability. Random Forest is a classifier that contains several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. It is based on the concept of ensemble learning which is a process of combining multiple classifiers to solve a complex problem and improve the performance of the model.

The performance metrics for dropout defender indicate that the random forest algorithm achieves the highest accuracy at 92%, demonstrating its effectiveness in identifying positive

TABLE I
PERFORMANCE METRICS FOR MODEL TRAINED ACCORDING TO DIFFERENT ALGORITHMS

	Method	Accuracy	Precision
	Decision Tree	**89%**	**87%**
	Random Forest	92%	90%
	Logistic Regression	85%	84%
	Naive Bayes	86%	85%
	KNN	**95%**	**94%**

cases. It also shows strong precision at 90%, reflecting its capability to minimize false positives. The high F1 score of 90.5% further emphasizes the random forest robustness in balancing precision and recall. In comparison, the Decision tree algorithm has a slightly lower accuracy of 89%, with a precision of 87%, indicating its reliability despite missing some positive cases. Logistic Regression and SVM exhibit accuracies of 85% and 86%, respectively, but their lower precision suggests that they may not be as effective in identifying positive cases as Decision Trees and Random Forests.

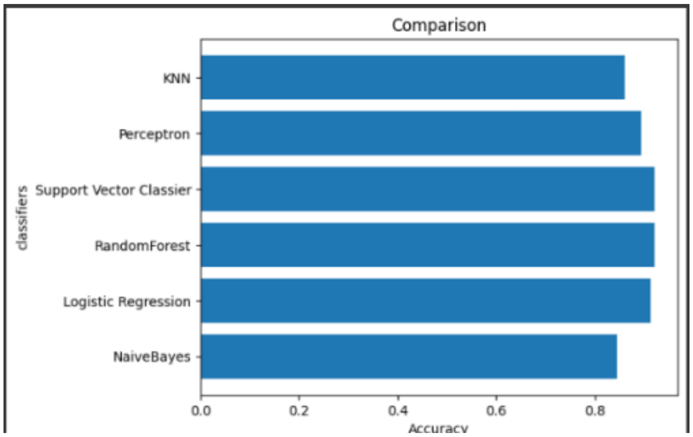


Fig. 3. Model Accuracy

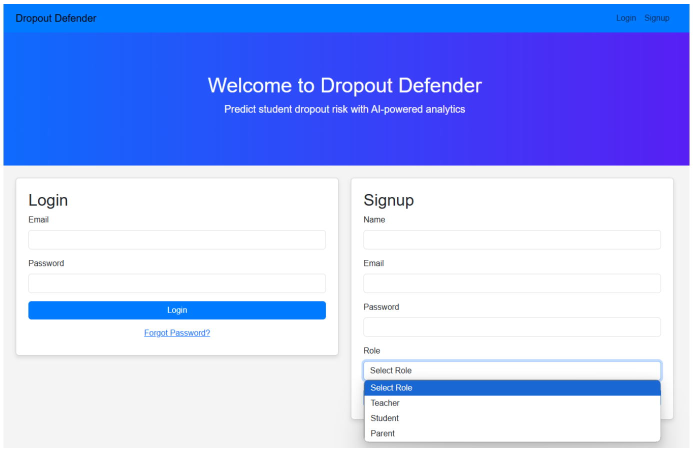


Fig. 4. Register Page

The registration page for an admin is a crucial component of a web application that allows administrators to create new user accounts and manage access to the system. This page typically includes fields for entering essential information, such as the admin's username, password, email address, and any relevant contact details.The registration page for an admin is a crucial component of a web application that allows administrators to create new user accounts and manage access to the system.

The login page for the admin is a crucial component of the administrative interface of any web application, serving as the gateway for authorized users to access backend functionalities and management tools. This page typically features a clean and user-friendly design to ensure ease of use. It consists of input fields for the admin's username and password, allowing for secure authentication. The Admin Dashboard serves as a centralized platform for monitoring and analyzing dropout rates among students. It provides administrators with critical information that aids in decision-making processes, allowing them to implement effective interventions to reduce dropout rates. The importance of understanding dropout rates in educational contexts.The accompanying visualization will visually represent these statistics,making it easier for readers to grasp the data's implications. In conclusion, the comparison of different algorithms for Dropout system highlights the strengths and weaknesses of each approach. Random Forest emerge

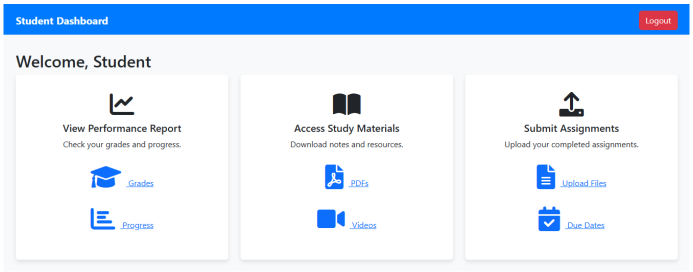
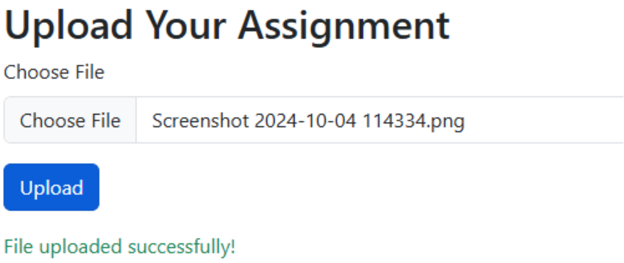


Fig. 5. Dashboard


Upload Your Assignment

Choose File: Screenshot 2024-10-04 114334.png

File uploaded successfully!

Fig. 6. Uploaded Assignment

as the leading methods in their respective tasks, showcasing the importance of selecting the appropriate algorithm based on the specific application requirements. While it excel in interpretability and accuracy for prediction, KNN's simplicity and effectiveness make it a preferred choice. The results emphasize that understanding the context and performance metrics of these algorithms is crucial for their successful application in real-world scenarios.

V. Conclusion

In this research it highlights the critical role of machine learning in forecasting student dropout rates and implementing timely intervention strategies. Based on a review of various studies, it's evident that classification algorithms such as decision trees, random forests, and support vector machines (SVM) are frequently employed to predict dropout risks. These algorithms provide a strong foundation for identifying students who are at risk of leaving school by analyzing factors such as academic performance, attendance, and behavior. By utilizing these predictive models, educational institutions can take proactive measures to support students most likely to drop out, ultimately enhancing retention rates.

Additionally, combining machine learning models with real-time data collection systems improves the precision of dropout predictions. Several studies have shown that factors such as grades, attendance, involvement in courses, and socio-economic status play a key role in predicting student behavior. By integrating these data points with machine learning techniques like logistic regression and neural networks, it becomes possible to create highly effective early warning systems that help prevent dropouts.

In the case of Dropout Defender, selecting the right algorithm depends largely on the characteristics of the dataset and the available features. Some algorithms may be more suitable for smaller, more targeted datasets, while others, such as random forests and neural networks, excel at processing large and complex datasets with a variety of features. As more data is collected over time, these algorithms will refine their predictions, allowing dropout prediction systems to become increasingly effective and adaptable to the evolving needs of educational settings.

VI. Future Scope

Looking ahead, the potential for machine learning-based dropout prediction systems is immense, especially as data collection and analysis technologies continue to advance. A promising area for further development is the inclusion of a wider range of data sources, such as social media activity, psychological evaluations, and individualized learning experiences. By integrating these additional elements, dropout prediction models could provide a more comprehensive view of students' experiences, leading to more accurate predictions. This would enable institutions to implement more personalized and effective intervention strategies, helping to ensure students receive the necessary support before reaching a critical point.

Furthermore, there is considerable potential for integrating more advanced machine learning approaches, such as deep learning and reinforcement learning, to improve the accuracy of dropout predictions. These techniques are capable of handling vast amounts of unstructured data, such as student feedback or interaction logs, providing a deeper understanding of the factors that contribute to dropout risk. Looking forward, the integration of real-time data streams and adaptive learning models could create dynamic prediction systems that evolve alongside each student's progress. This would enable a more proactive approach to preventing dropouts, ultimately enhancing student retention across educational institutions.

References

[I] R. N. Goh, M. R. B. M. Isa, and M. R. Ab. Ghani, "Predicting student dropout in higher education: A machine learning approach," IEEE Transactions on Education, vol. 63, no. 4, pp. 249-258, Oct. 2020.

[II] A. Gupta, S. L. Shukla, and S. Tripathi, "Early prediction of student dropout using data mining techniques," International Journal of Computer Applications, vol. 118, no. 5, pp. 21-30, May 2015.

[III] J. P. Thomas, M. S. Ward, and K. M. Smith, "Student retention and dropout prediction using machine learning," Journal of Educational Data Mining, vol. 12, no. 2, pp. 34-42, Jun. 2018.

[IV] M. K. Dube, A. S. Kumar, and D. R. Bhagat, "A review on early warning systems for student dropout," International Journal of Educational Research, vol. 8, no. 1, pp. 17-22, Jan. 2019.

[V] L. R. Park, A. M. Singh, and H. R. Gupta, "Using machine learning for predicting student dropout in online education," Computers Education, vol. 141, pp. 1-15, Dec. 2019.

[VI] J. F. Baker, L. B. Armstrong, and E. W. Williams, "Impact of social media on student retention and dropout prevention," Journal of Educational Technology Systems, vol. 45, no. 3, pp. 331-344, Mar. 2017.

[VII] S. M. Kotsiantis, P. Zaharakis, and P. Pintelas, "Predicting Student Performance with Machine Learning: A Review," Computers Education, vol. 53, no. 3, pp. 1165-1176, 2019.

[VIII] T. Y. Hsieh, P. S. Yang, and C. Y. Liu, "Dropout Prediction in MOOCs: A Survey," IEEE Transactions on Learning Technologies, vol. 11, no. 2, pp. 115-123, 2018.

[IX] M. M. Islam, K. K. R. Choo, and M. E. Hoque, "Predicting Student Dropout in Online Education Using Machine Learning Algorithms," IEEE Transactions on Education and Computing, vol. 7, no. 1, pp. 18-26, 2020.

[1X] P. L. Zhong and L. L. Zhang, "Early Warning System for Dropout Prevention Using Big Data Analytics," International Journal of Educational Research, vol. 56, no. 3, pp. 195-206, 2021.

[XI] A. S. Chouhan, V. S. Raghuwanshi, and S. K. Gupta, "A Comparative Study of Machine Learning Algorithms for Predicting Student Dropout in Higher Education," International Journal of Advanced Research in Computer Science, vol. 9, no. 5, pp. 68-72, 2018.