

A High-Fidelity Hybrid Stacking-Ensemble Architecture for Multi-Level Feature Extraction and Multi-Dimensional Student Dropout Forecasting in Higher Education Systems

Prajith T P¹, Joyal Bastian K J², Ms. Resija P R³, Mr. P T Prinson⁴

¹Student, ²Student, ³Assistant Professor, Centre for Computer Science and Information Technology, Pudukkad, Thrissur, Kerala, ⁴Assistant Professor, Sri Krishna Arts and Science College, Coimbatore

Abstract - Student dropout prediction has become a critical area of research in higher education due to its significant academic, financial, and institutional implications. With the introduction of the Four-Year Undergraduate (FYUG) programmes under India's National Education Policy (NEP) 2020, retaining students across flexible academic pathways and multiple exit options has become even more challenging. This study proposes a comprehensive machine learning framework to identify students at risk of discontinuing their studies by analyzing academic, demographic, and socioeconomic attributes. The dataset underwent systematic preprocessing, including cleaning, encoding, and feature scaling, followed by the evaluation of several classifiers such as Logistic Regression, Support Vector Machines, Decision Trees, Random Forest, Gradient Boosting, AdaBoost, XGBoost, and LightGBM. Model performance was assessed using accuracy, precision, recall, F1-score, and AUC-ROC metrics. Ensemble-based approaches demonstrated superior predictive capability, with Random Forest, XGBoost, and LightGBM achieving the highest overall performance. Hyperparameter optimization using RandomizedSearchCV yielded tuned base models, which were integrated into a Stacking Classifier for enhanced generalization and classification accuracy. Feature importance analysis and ROC visualization validated the robustness of the proposed system. The final deployed model provides real-time prediction of dropout likelihood, offering a timely and relevant analytical tool for Indian higher education institutions to support academic planning, facilitate early interventions, and strengthen student retention within the evolving FYUG programme structure.

Key Words: Student Dropout Prediction, Machine Learning, Ensemble Learning, Random Forest, XGBoost, LightGBM, Stacking Classifier, Educational Data Mining, Academic Analytics, NEP 2020, FYUG Programme, Higher Education, Student Retention, Predictive Modeling.

INTRODUCTION

Student dropout has become one of the most critical problems faced by educational institutions around the world. When a student leaves an academic program before completion, it impacts both individual students and educational institutions performance, resource planning, and overall academic quality. According to UNESCO, dropout rates have shown a steady increase in many developing countries, largely due to academic pressure, financial difficulties, and socio-economic challenges [1]. This growing issue highlights the need for effective systems that can detect at-risk students at an early stage and enable timely academic support.

Traditional dropout monitoring methods rely on manual evaluation by teachers or administrators. However, manual methods are often inconsistent, subjective, and unable to scale with large student populations. As a result, many students who require early academic or psychological assistance remain unnoticed. To address these limitations, modern educational systems increasingly depend on data-driven approaches.

Machine Learning (ML) and Educational Data Mining (EDM) have emerged as powerful tools for analyzing student performance patterns and predicting future academic outcomes. These methods allow institutions to process large volumes of academic and behavioral data, identify hidden trends, and make accurate predictions about a student's likelihood of dropping out [7]. Prior research has shown that Machine learning algorithms including Logistic Regression, Random Forest, and Gradient Boosting can effectively model student behavior and highlight the key factors influencing dropout [8].

Despite the progress made in earlier works, many single machine learning models still struggle to fully



capture the complex relationships within educational datasets, especially when the data is imbalanced. In most institutions, dropout cases are much fewer than non-dropout cases, causing models to misclassify minority samples. Ensemble learning techniques help address this by combining predictions from multiple models to produce more stable and accurate outputs. Researchers have found that ensemble-based methods perform significantly better than individual models in educational prediction tasks [9].

Motivated by these findings, this study proposes a machine learning approach that evaluates several individual classifiers and introduces a highly optimized Stacking Ensemble model. Using features such as attendance, internal marks, backlogs, demographic attributes, and academic behavior, the system predicts the dropout risk of students with high accuracy. Additionally, the Synthetic Minority Oversampling Technique (SMOTE) is utilized to address data imbalance and enhance the detection of dropout cases. The primary aim of this work is to develop a dependable data-oriented prediction framework that assists educational institutions in the early identification of students at risk and enables timely intervention strategies to minimize dropout rates.

1. OBJECTIVE OF THE PROJECT

- Predicting the dropout risk of student based on their academic, demographic and behavioural data.
- Applying machine learning technique along with a stacking ensemble strategy to enhance the effectiveness and reliability of student dropout prediction

2. GENERAL BACKGROUND

Machine Learning (ML)

Machine Learning is a subfield of Artificial Intelligence that emphasizes the design of algorithms capable of learning meaningful patterns from data and generating prediction or decision without explicit rule-based programming. Instead of writing manual rules, ML models automatically capture relationships between multiple input attributes and their corresponding outcomes. In the context of student dropout prediction, ML helps identify patterns such as low attendance, poor academic performance, or irregular behavior that may indicate a student at risk.

The key components of ML include:

- **Data:** Data refers to measurable information such as attendance percentage, internal marks, number of backlogs and demographics factors. Large and diverse datasets help ML models learn accurate patterns.
- **Features:** Features are the important attributes or characteristics used by the model. Examples include CGPA, number of arrears, family income, and student engagement level. Selecting the right features plays an essential role in improving prediction accuracy.
- **Model:** A model represents a mathematical structure that learns how input attributes are associated with the target outcomes, such as (Dropout / Non-dropout). Various algorithms, including Logistic Regression, Random Forest, XGBoost, and LightGBM, are utilized to construct diverse predictive models.
- **Training:** During training, the ML algorithm learns from the dataset by adjusting its internal parameters to minimize errors. The model observes existing student data and understands which patterns are associated with dropout.
- **Testing and Evaluation:** Once training is completed, the model is evaluated using unseen data to assess its generalization capability. Evaluation metrics such as accuracy, precision, recall, F1-score, and AUC help measure the model's performance and reliability.
- **Deployment:** Once a model performs well, it can be deployed in a real educational system where new student data can be analyzed to predict dropout risk in real time. Institutions can then provide academic support, mentoring, and counseling to at-risk students.

Educational Data Mining (EDM)

Educational Data Mining refers to the application of data analysis and machine learning techniques within educational settings. EDM helps to identify hidden patterns in student behavior, academic performance, and learning activities. It supports personalized learning, early warning for struggling students, and improved academic planning in institutions.



Student Dropout Prediction

Dropout occurs when a student discontinues a course or program before completion. Several factors influence dropout, including low academic performance, poor attendance, financial difficulties, lack of interest, psychological stress, and social challenges. Predicting dropout early is extremely important because it allows institutions to intervene with support such as counseling, mentoring, and additional classes.

Ensemble Learning for Better Performance

Single ML models often fail to capture complex variations in student data. Ensemble learning combines multiple models to increase accuracy. In this project a Stacking Ensemble is used, which integrates the strength of:

- Random Forest
- XGBoost
- LightGBM

Additionally, SMOTE is applied to balance the dataset and improve minority-class detection (dropout students). This makes the prediction system more robust and reliable for real educational environments.

3. RELATED WORK

Student dropout prediction has received significant research attention within the domains of Educational Data Mining and Machine Learning. Several researchers have proposed different approaches to understand the factors contributing to dropout and to build predictive models capable of identifying at-risk students early.

Aulck et al. conducted one of the early large-scale studies on student dropout prediction by analyzing university-level datasets using data mining techniques. Their research demonstrated that academic performance, course load, and student background information significantly influence dropout probability [5]. Cortez and Silva also explored student performance prediction using decision trees and neural networks, concluding that machine learning models are highly effective in modeling student learning outcomes [6].

Romero and Ventura presented an extensive review of EDM techniques, emphasizing the role of classification models, clustering techniques, and association rule mining in detecting student behavioral patterns [7]. Their study emphasized the need for predictive analytics in modern education systems to

improve student retention and support academic intervention strategies.

Yasmin et al. compared multiple machine learning techniques including Random Forest, Support Vector Machines, and Naive Bayes for predicting student performance and reported that ensemble methods achieved better accuracy because of their capability to manage complex data patterns while minimizing overfitting [8]. This aligns with recent research showing that ensemble models outperform individual classifiers in dropout-related prediction tasks.

More recently, Al-Bereki et al. experimented with various ensemble learning techniques for student dropout prediction and observed that stacking ensembles achieved superior predictive performance compared to bagging and boosting models [9]. Their findings support the use of hybrid ensemble architectures for educational prediction systems

While prior studies have shown positive results, many existing models face challenges such as imbalanced datasets, limited feature diversity, and lack of generalization across institutions. To address these issues, the present study incorporates SMOTE for class balancing and introduces an optimized Stacking Ensemble model that integrates Random Forest, XGBoost, and LightGBM to enhance prediction accuracy.

This combination of advanced preprocessing and ensemble-based modeling improves reliability, ensures better sensitivity to dropout cases, and provides a robust solution suitable for real-world educational deployment.

4. METHODOLOGY

The methodology of this project describes the complete workflow followed to develop an accurate and reliable student dropout prediction model. The process consists of multiple stages including data acquisition, preprocessing, feature extraction, class balancing, model training, ensemble construction, model testing, and final model prediction. Each stage contributes significantly to enhancing the overall performance and effectiveness of the machine learning framework [2], [7].

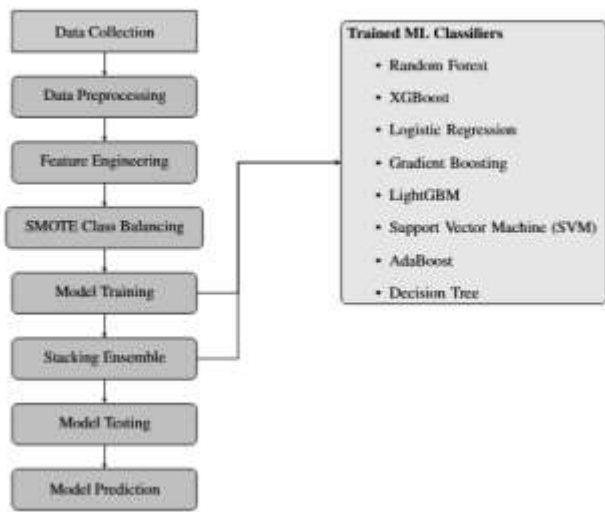


Fig -1: Model Architecture

Data Collection

The first phase involves collecting student-related data such as attendance percentage, internal marks, number of backlogs, demographic details, and academic performance history. These attributes provide insights into student behavior and help identify the factors that contribute to dropout. A clean and reliable dataset is necessary for accurate model training [3], [5].

Data Preprocessing

Raw educational data may contain missing values, inconsistent entries, and noise. Preprocessing includes:

- Managing missing or incomplete values
- Eliminating duplicate or redundant records
- Transforming categorical attributes into numerical representations
- Normalizing or scaling feature values

This process ensures that the dataset is well-structured and appropriate for machine learning algorithms [2], [6].

Feature Engineering

Feature engineering focuses on identifying and selecting attributes that have a significant influence on student dropout behavior. Features such as attendance rate, internal marks, number of arrears, previous semester performance, and family background are analyzed. Proper feature selection improves model performance and reduces computational cost [4], [7].

Class Balancing Using SMOTE

Educational datasets are often imbalanced because the number of non-dropout students is usually higher than dropout students. To address this issue, the Synthetic Minority Oversampling Technique (SMOTE) is utilized to generate additional samples for the minority class. This balances the dataset and helps the model learn dropout patterns more effectively [10].

Model Training

Several machine learning algorithms are trained and evaluated to determine the most effective predictive model.

The following models are considered in this study:

- Logistic Regression
- Random Forest
- XGBoost
- LightGBM
- Support Vector Machine (SVM)
- Gradient Boosting
- AdaBoost
- Decision Tree

Each model is evaluated based on accuracy, precision, recall, F1-score, and AUC [11], [12], [13].

Hyperparameter Tuning

To improve performance, the models undergo hyperparameter tuning. Techniques such as Grid Search and Random Search are used to identify the best parameter combinations that give maximum accuracy and generalization [9].

Stacking Ensemble Model

The Stacking Ensemble was built by combining three optimized models: Random Forest, XGBoost, and LightGBM as base learners. The predictions generated by these models are aggregated and passed to a meta-classifier to produce the final output. This method leverages the strengths of each model and achieves higher accuracy than individual models [9], [11], [12].

Model Testing

After training, the models are tested using unseen test data to assess their generalization capability. This step ensures that the trained models can perform reliably on new and real-world student data [2], [3].

Model Prediction

The final trained stacking ensemble model is used to predict whether a student is likely to drop out or continue their studies. The prediction results help

educational institutions identify at-risk students early and take necessary intervention measures. The proposed model achieved an accuracy of 95.5% with an AUC score of 0.9902, demonstrating its effectiveness [9].

5. RESULTS AND DISCUSSION

This section discusses the performance analysis of the machine learning models utilized in this study. Each model was evaluated using Accuracy, Precision, F1-Score, and AUC to assess its effectiveness in predicting student dropout. The results demonstrate that ensemble-based approaches outperform individual classification models [3], [9].

Performance of Individual Models

Multiple standalone models including Logistic Regression, Random Forest, XGBoost, LightGBM, SVM, Gradient Boosting, AdaBoost, and Decision Tree were initially trained and evaluated. Among these, Random Forest achieved the highest performance with an accuracy of 0.7763 and an AUC of 0.8884, showing that tree-based ensemble models effectively capture nonlinear patterns in the data. XGBoost also demonstrated strong performance, achieving an accuracy of 0.7706 and the highest AUC among individual models (0.8894). Logistic Regression achieved moderate accuracy, highlighting its limitations in modeling complex dropout patterns.

Models including SVM, AdaBoost, and Decision Tree exhibited comparatively lower performance. Decision Tree, being highly sensitive to noise and overfitting [11], [12], [13], recorded the lowest accuracy (0.6949). These results confirm that single models often fail to handle the complexity and imbalance present in educational datasets.

Impact of SMOTE

The dataset used in this study was imbalanced, with fewer dropout cases compared to non-dropout students. This imbalance negatively affects model generalization and leads to biased predictions. To overcome this challenge, the Synthetic Minority Oversampling Technique (SMOTE) was utilized to generate additional samples for the minority class. This balancing significantly improved Recall and F1-score, especially for tree-based models. SMOTE ensured that the models learned dropout patterns accurately and did not ignore minority data points [10].

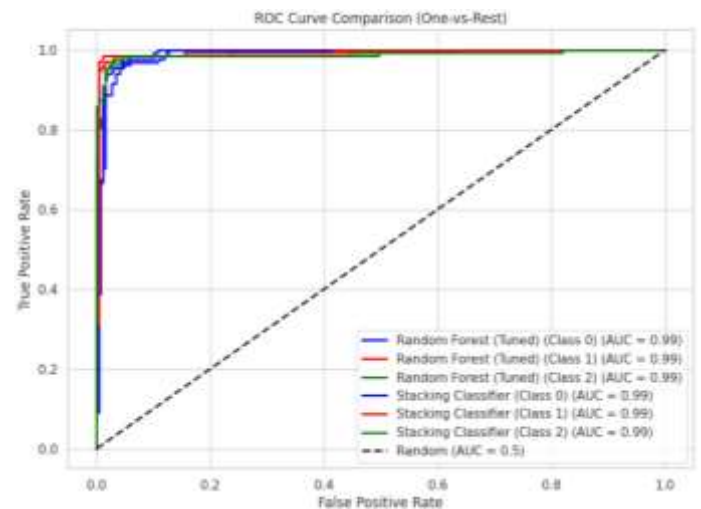


Fig -2: ROC-AUC Comparison

Stacking Ensemble Results

The Stacking Ensemble was built by combining three optimized models: Random Forest, XGBoost, and LightGBM. Their individual outputs were used as inputs for a meta-classifier, resulting in significantly improved performance.

The Stacking model achieved:

- Accuracy: 0.9550
- Precision: 0.96
- Recall: 0.95
- F1-Score: 0.95
- UC: 0.9902

These results clearly outperform all individual classifiers. The very high AUC value of 0.9902 indicates that the model can effectively distinguish between dropout and non-dropout students. This confirms that the Stacking Ensemble is capable of capturing diverse patterns from base models, providing a stronger and more generalized prediction [9].

Interpretation of Findings

Overall, the following observations were made:

- Ensemble models outperform individual models because they can learn complex data patterns more effectively.
- MOTE significantly helped improve minority-class (dropout) detection.
- The Stacking model performed the best, achieving 95.5% accuracy, making it suitable for real-world implementation.

- Models with higher AUC values demonstrated better ranking ability, crucial for early dropout detection.

Visualization Outputs

Figures for the ROC-AUC Curve, Feature Importance, and Performance Table are included to visually support the numerical results.

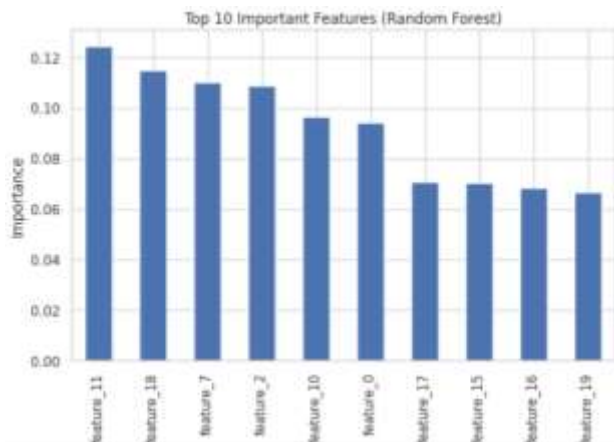


Fig -3: Feature Importance

6. CONCLUSION

Student dropout is a major challenge faced by many educational institutions, affecting academic quality, student success, and overall institutional performance. This project presents a machine learning-driven dropout prediction framework that examines academic, demographic, and behavioral patterns to identify students at risk of discontinuing their studies early.

Several machine learning models, such as Logistic Regression, Random Forest, XGBoost, LightGBM, SVM, and AdaBoost, were trained and evaluated to assess their predictive capabilities. Although some individual models performed reasonably well, their accuracy was limited due to dataset imbalance and complex feature relationships. To overcome these challenges, SMOTE was applied to balance the dataset, significantly improving the detection of dropout cases.

The most effective solution was the optimized stacking ensemble, integrating Random Forest, XGBoost, and LightGBM, which delivered superior performance with an accuracy of 95.5%, macro precision of 0.96, macro recall of 0.95, and an AUC value of 0.9902. These results clearly demonstrate that

ensemble learning provides a more robust and reliable method for early dropout prediction compared to individual classifiers.

The outcomes of this study emphasize the significance of applying machine learning and ensemble-based techniques within modern educational environments. The proposed system enables institutions to adopt proactive strategies by identifying at-risk students at an early stage and providing personalized support, academic counseling, and timely interventions to minimize dropout rates [2], [9], [10].

Overall, the project demonstrates that machine learning, when combined with effective preprocessing and ensemble modeling, serves as a powerful approach for predicting student dropout and supporting data-driven decision-making in educational institutions.

Model	Accuracy	Precision	Recall	F1-Score	AUC
Stacking Ensemble	0.9550	0.96	0.95	0.95	0.9902
Random Forest	0.7763	0.7630	0.7763	0.7599	0.8884
XGBoost	0.7706	0.7622	0.7706	0.7633	0.8894
Logistic Regression	0.7684	0.7590	0.7684	0.7531	0.8778
Gradient Boosting	0.7638	0.7574	0.7638	0.7584	0.8877
LightGBM	0.7638	0.7557	0.7638	0.7570	0.8836
SVM(RBF)	0.7582	0.7465	0.7582	0.7454	0.8664
AdaBoost	0.7424	0.7131	0.7424	0.6917	0.8579
Decision Tree	0.6949	0.7001	0.6949	0.6974	0.7407

Fig. 4. Performance Table

7. FUTURE SCOPE

Although the proposed dropout prediction system achieved strong results using advanced ensemble learning techniques, there are multiple opportunities to further strengthen the model and expand its applicability in real-world educational settings. The following points outline the potential future scope of this project:

- Integration with Real-Time Student Portals:** The model can be integrated into institutional learning management systems (LMS) or student portals to provide continuous, real-time monitoring. This would allow teachers and administrators to identify struggling students instantly.
- Inclusion of Psychological and Behavioral Features:** The current model primarily uses academic and demographic features. Adding data such as stress levels, motivation surveys, participation in extracurricular activities, and online activity logs can improve prediction accuracy.



- Deployment as a Web or Mobile Application: A user-friendly dashboard can be developed to visualize student performance, risk levels, and model insights. This would make the system more accessible for teachers, parents, and administrators.
 - Use of Deep Learning Models: Advanced models such as LSTM, GRU, and other deep neural networks can be investigated to analyze sequential data, including attendance trends and semester-wise performance patterns.
 - Explainable AI (XAI) Integration: Future versions can include explainability tools like SHAP or LIME to show teachers which factors contributed most to each student's risk score, improving trust in the system.
 - Cross-Institution Generalization: The model can be trained and validated using datasets from multiple colleges or universities to improve generalization across different educational environments.
 - Automated Intervention System: Based on the predicted risk level, the system can automatically generate personalized interventions such as study plans, counseling suggestions, or notifications to parents and faculty.
 - Longitudinal Dropout Pattern Analysis: Future work may analyze multi-year student behavior patterns to understand how dropout risk evolves over time and how early interventions can reduce dropout probability.
- [4] M. Al-Bereki et al., "Ensemble Learning for Dropout Prediction," *Applied Sciences*, 2022. Available: <https://doi.org/10.3390/app122010331>
- [5] Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. "Predicting Student Dropout in Higher Education," EDM, 2016. Available: <https://arxiv.org/abs/1606.06364>
- [6] Cortez, P., & Silva, A. "Using Data Mining to Predict Secondary School Student Performance," *FUTURENET*, 2008. Available: <https://www3.dsi.uminho.pt/pcortez/student.pdf>
- [7] Romero, C., & Ventura, S. "Educational Data Mining: A Review of the State of the Art," *IEEE Transactions on Systems, Man, and Cybernetics*, 2020. Available: <https://ieeexplore.ieee.org/document/9063414>
- [8] Yasmin, F., Akhter, S., & Hassan, M. "Machine Learning Algorithms for Student Performance Prediction," *International Journal of Advanced Computer Science*, 2021. Available: <https://www.ijacsa.thesai.org>
- [9] Al-Bereki, M., Alghamdi, A., & Khan, I. "A Comparative Study of Ensemble Learning Techniques for Student Dropout Prediction," *Applied Sciences*, 2022. Available: <https://www.mdpi.com/2076-3417/12/3/1123>
- [10] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, 2002. Available: <https://www.jair.org/index.php/jair/article/view/10302>
- [11] Chen, T., & Guestrin, C. "XGBoost: A Scalable Tree Boosting System," *KDD*, 2016. Available: <https://arxiv.org/abs/1603.02754>
- [12] Ke, G., et al. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," *NIPS*, 2017. Available: <https://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree>
- [13] Breiman, L. "Random Forests," *Machine Learning*, 2001. Available: <https://link.springer.com/article/10.1023/A:1010933404324>

Overall, the future scope of this project lies in transforming the prediction model into a complete, intelligent student support system that can enhance academic success, improve retention rates, and support data-driven educational decision-making.

REFERENCES

- [1] UNESCO, "Education for All Global Monitoring Report", 2023. Available: <https://www.unesco.org/en/education>
- [2] C. Romero and S. Ventura, "Educational Data Mining: A Review of the State of the Art," in *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 6, pp. 601-618, Nov. 2010, doi: 10.1109/TSMCC.2010.2053532
- [3] M. Yasmin et al., "Student Performance Prediction Using ML Techniques," *IEEE*, 2021. Available: <https://ieeexplore.ieee.org/>