RESEARCH PAPER

Enhancing Academic Retention through ML: Predictive Frameworks for Student Dropout Analysis

Anil Khatak[1], Vinita[2], Vikas Kajla[3], Kamal Rani[4], Sheenam Naaz[5]
[1]Department of Allied Health Sciences, Guru Jambheshwar University of Science & Technology, Hisar, 125001, Haryana, India.
[2]Department of Electrical and Electronics Engineering, Guru Jambheshwar University of Science and Technology, Hisar, 125001, Haryana, India.
[3]Department of Allied Health Sciences, Guru Jambheshwar University of Science & Technology, Hisar, 125001, Haryana, India.
[4]Department of Electrical & Electronics Engineering, IIMT University, Meerut, Uttar Pradesh, India.
[5]Department of Computer Science and Engineering, Sharda School of Computing Science and Engineering, Sharda University, Greater Noida, India.
Corresponding authors: Vinita, Department of Electrical and Electronics Engineering, Guru Jambheshwar University of Science and Technology, Hisar 125001, India.
vinita669@gmail.com;

Abstract
Student dropout has become a major concern in higher education, affecting institutional performance, academic outcomes, and overall student development. This study introduces a data-driven framework that applies ensemble machine learning models to identify students who are likely to discontinue their studies. The framework combines several algorithms, including Random Forest, Gradient Boosting, AdaBoost, and a Voting Classifier, to achieve higher predictive accuracy and model stability. The research begins with systematic data preparation that includes cleaning, feature selection, and exploratory data analysis to ensure that the input data are reliable and meaningful. The models are then trained and evaluated using performance measures such as accuracy, precision, recall, and F1-score. Results show that ensemble methods perform better than traditional single models by effectively capturing complex patterns within student data. The findings demonstrate that early identification of at-risk students can support timely academic interventions and improve retention outcomes. In addition, the study discusses ethical considerations such as data transparency, fairness, and student privacy, which are essential for the responsible use of predictive systems. Overall, this work contributes to the growing field of educational data mining by presenting a reliable and interpretable approach that can help institutions enhance student success through informed decision-making.

Keywords: Machine Learning, Ensemble Learning, Predictive Analytics, Student Dropout Prediction, Educational Data Mining, Academic Intervention, Artificial Intelligence, Data-Driven Framework
How to cite this article: Anil Khatak1 Vinita2 Vikas Kajla3 Kamal Rani4 Sheenam Naaz5. Enhancing Academic Retention through ML: Predictive Frameworks for Student Dropout Analysis. Int J Drug Deliv Technol. 2026;16(89s):101-122. DOI: 10.25258/ijddt.16.89s.10.

1 Introduction

In every country, education is the central pillar of economic growth, social development, and innovation. The human capital that drives technological and industrial progress is particularly nurtured through higher education. However, despite the global expansion of higher education institutions and access programs, the issue of student attrition continues to drive down institutional and personal achievement (or potential) over time [6, 7]. Early dropping out of students from academic programmes causes financial losses to the universities, ineffective resource distribution and psychosocial stress of learners. The globe's understanding indicates that higher learning dropout issues are at 15-35 per cent, varying with socioeconomic characteristics, institutional characteristics, and readiness in learning capacity and inclination to participate in university learning [19, 28]. In nations such as India, where tertiary education

enrollment has witnessed sheer growth, retaining students has been a problem that has only become a focal concern in recent decades [12, 13].

Since high amounts of administrative and learning data are being collected by universities, the potential for converting these data into actionable intelligence has become more apparent. The standard statistical indicators of performance are not adequate enough to represent the complex behavioural and academic subtleties that result in dropout occurrences. Accordingly, educational institutions are now adopting computational intelligence and data analytics to obtain future-oriented insights and develop timely intervention plans [5, 18, 25]. Similar developments suggest a path toward preemptive decision-making, providing evidence-based frameworks to identify vulnerable students and monitor their academic progress more accurately.

1.1 Role of Machine Learning in Educational Analytics

Machine learning has emerged as an essential asset in redefining educational analytics from descriptive reporting to predictive and prescriptive modelling. As they can automatically detect patterns in significant and complex data, machine learning algorithms enable institutions to move beyond typical statistics and make data-driven, evidence-based decisions [5][12]. At the higher education level, such approaches have been routinely utilised to predict student performance, engagement, and dropout likelihood by picking up on subtle context and behaviour indicators that are difficult for human teachers to pick up on manually [13].

For this purpose, several supervised and unsupervised machine learning models—such as decision trees, support vector machines, neural networks, and ensemble models—have proved very promising in analysing student data [18][16]. Ensemble learning methods, on the other hand, have emerged as extremely popular given that they may combine multiple base models to reduce variance and improve overall predictive consistency [10][25]. Research by Talamás and Ceballos (2023) and Yağcı (2022) illustrates how ensemble methods outperform single classifiers by taking advantage of complementary strengths of models and improving robustness across different datasets [27][19]. In addition, the incorporation of explainable artificial intelligence has started to close the loop between algorithmic prediction and educational explanation. Explainability software helps teachers understand why a system predicts that a student may be at risk. By showing the reasons behind each prediction, it supports fairness and increases trust in technology-based decisions. When combined with dropout prediction models, explainability tools not only improve the accuracy of identifying at-risk students but also help schools design timely and ethical support strategies to improve student outcomes.

1.2 Problem Definition

The issue of anticipated dropout remains a complex and multifaceted problem despite the growing recognition of student retention as a strategic priority. Cognitive, social, institutional, and economic factors can interact in non-linear ways to produce academic disengagement, which can cause changes in surveys and the predetermined effect of academic disengagement later on in the life of a person or group of prospective sampled pupils [11]. These inter-dependencies cannot be readily determined using standard analytical tools like logistic regression or correlation analysis and are restricted to dealing with large, heterogeneous datasets of education [13][5]. As a result, school administrators miss the chance to identify potential dropouts early on, which complicates timely intervention and increases the risk of losing talented students.

Recent conceptions of machine learning have allowed so-called data-driven solutions to model forms of complex behavioural and contextual variables that can contribute to persistence [18][19]. Nevertheless, most available predictive works utilise a single classifier, which can influence the sensitivity of outputs in terms of performance when applied to different students or fluctuate in the academic experiences of students or institutions of higher learning [6][10]. Small or localised datasets are known to overfit the tiny model readily and do not generalise across institutions, gathering an agent becoming lost in general-population

code (where surrogate outcomes are presumed to mirror each other) as examined by default models defined using metaparis models such as plain MDP or hypothetical MDP models. Thus, the issue of making a constitutive, collective structure consisting of a combination of the various learning algorithms to increase the accuracy and strength of predictions is ultimately immediate. This framework should not only identify at-risk students early but also provide clear insights to facilitate timely and fair academic assistance [25][32]. The main task of the current research is to fill this gap.

1.3 Research Gap and Motivation

Although significant advancements have been made in applying machine learning to educational prediction, several limitations hinder the validity and practical applicability of these models. Most of the existing research is based on small or small-scale datasets, which are just adequate to represent the diversity of student populations within institutions and fields of study [12]. As a result, the models, which are fitted with this kind of data, tend to produce deplorable generalizability through their application in new settings. Furthermore, most existing literature employs single classifiers, which are inaccurate due to their sensitivity to data representation, noise, and parameterisation tuning required to optimise quantitative results during case performance testing and training.

Compound ensemble methodologies that combine two or more algorithms are more stable in their predictions and more accurate in their findings, as the strategies overcome the limitations of each model individually. However, available literature does not provide detailed frameworks that ascertain ensemble models through widespread empirical measurements in various learning settings. Few studies also compare the variants of ensembles (Random Forest, Gradient Boosting, AdaBoost, etc.) in an experimental context in a systematic way under the same research environment [13, 27]. In addition to that, large volumes of literature on the issue of predictive analytics and its ethical and interpretability aspects in the sphere of education are comparatively limited. However, these two elements are the most important in the context of responsible application of AI. This research has been driven by the necessity to fill these gaps by designing a data-driven, ensemble-based model that would be more accurate in its predictions and, at the same time, remain transparent and reduce bias. With the help of modern learning methods and strict assessment indicators, the proposed study can help form a practical and ethically substantiated recommendation for the initial detection of at-risk learners in tertiary institutions.

1.4 Research Objectives and Contributions

The present study seeks to bridge the methodological and practical gaps identified in existing research by developing a robust, data-driven framework for early prediction of student dropout. The general aim is to develop and test ensemble machine learning models that improve the predictive power, interpretability, and ethical transparency of educational analytics. To do so, the research is advised by the following specific aims in Figure 1. The main contribution of this research is the combination of ensemble learning and explainable AI in a single predictive approach for educational analytics. Through a comparative analysis of predictive models and an emphasis on interpretability, this study provides methodological contributions and practical guidance for institutions developing data-driven retention initiatives. In doing so, it aligns with emerging scholarship that frames machine learning as a means to enhance student-support systems rather than a purely technical exercise.

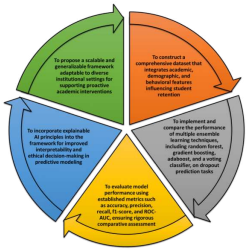


Fig. 1 Research Objectives

1.5 Structure of the Paper

The remainder of this paper is organised in the following structure. Section 2 presents review on existing literature

about educational data mining and student dropout prediction focusing on major methodologies, research gaps, and emerging trends. Section 3 presents the proposed methodology, i.e., data collection, preprocessing of data, feature engineering, and application of ensemble learning algorithm. Section 4 contributes to the experimental setup, evaluation metrics, and evaluation of the proposed model. Section 5 discusses the results in terms of interpretability, ethical consequences, and implications for institutional practice. Section 6 concludes the paper with a summary of key results, practical insights, and directions for future investigation.

2 Literature Review

Reliability on data for the predictions has also created an impact on the educational sector, to predict the number of students dropping out, using machine learning methods. A systematic review method was used to cover the existing literature extensively. The Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines were utilised to improve methodological transparency and reproducibility. The review process includes identification, screening, eligibility, and inclusion phases. According to identified gaps and future research directions, the literature review was underpinned by the following key questions as seen in Figure 2 below:

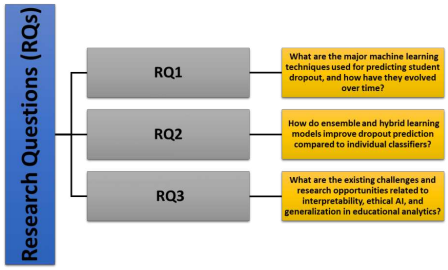


Fig. 2 Reseach Questions

2.1 PRISMA-based Review Process

To ensure research discipline and reproducibility, the literature review was conducted following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. The PRISMA framework provided a organized and clear approach for identifying, screening, and unifying relevant studies on student dropout prediction and ensemble-based machine learning. The process followed four major stages: *Identification, Screening, Eligibility,* and *Inclusion*, as described below [18, 28].

2.1.1 Identification:

The first search was conducted on main academic databases like Scopus, IEEE Xplore, SpringerLink, ScienceDirect, and ACM Digital Library, covering the publication date range of 2015 to 2025. Boolean search strings were formed by combining words like *"student dropout prediction," "machine learning," "ensemble models," "educational data mining,"* and *"explainable AI"*. A total of 148 records were obtained in this phase.

2.1.2 Screening:

All the records found were ranked for importance by searching through titles, abstracts, and keywords. Non-scholarly sources and duplicates were removed, leaving 96 publications. Studies that did not focus on predictive modeling, higher education, or data-driven analytics were excluded at this stage. The systematic screening at this point ensured that only high-quality, relevant research was kept for further assessment.

2.1.3 Eligibility:

The remaining 63 full-text articles were evaluated by considering the predefined inclusion criteria. Studies were considered eligible if they (i) used machine learning or hybrid computational models for dropout prediction, (ii) presented empirical or comparative analysis, and (iii) were published in peer-reviewed journals or conferences. Exclusion criteria included studies deficient in methodological clarity , using language other than English, and works not related to higher education analytics. Following eligibility assessment, 36 peer-reviewed studies were finally selected for qualitative synthesis. These included both journal articles and conference proceedings emphasizing ensemble learning, hybrid models, or explainable AI in the context of dropout prediction.

2.2 Addressing RQ1: Evolution of Machine Learning in Dropout Prediction

Classical statistical methods, such as logistic regression, linear discriminant analysis, and decision-tree models, were widely used in far-reaching studies of the past to predict student dropout.[12]

Although these approaches provided some interpretability, they had limitations due to their inability to represent the extent of non-linear interaction between multidimensional characteristics, which was limited to a degree of linear interaction.Subsequent advances in computation fueled using machine-learning techniques, such as Support Vector Machines, Random Forests and Gradient Boosting to forecast intricate forms of dropout [13, 20]. Employed alongside heterogeneous learning records, empirical rates improved notably in predictive validity and externalisation across groups. Nevertheless, there was one more such weakness; the models were not quite adaptive to changing institutional contexts and hence often produced dissimilar results in the context of diverse groups of students. The outcomes of their three experiments involving two thousand participants were diverse, and this is why one would have to use more context-relevant analytical systems [10, 18].

2.3 Addressing RQ2: Impact of Ensemble and Hybrid Models

Ensemble models have emerged as a dominant strategy to enhance predictive performance and reduce model variance [23, 25]. Techniques such as bagging, boosting, and stacking integrate multiple base classifiers, yielding more stable predictions than single learners. Studies by Talamás and Ceballos (2023) and Rabelo et al. (2025) confirmed that stacked ensembles consistently outperform individual algorithms by leveraging complementary model strengths [27, 32]. Hybrid models, which combine deep learning with ensemble learning techniques, have also shown promising results in modelling student engagement and progression [28]. Nonetheless, the literature indicates a limited number of comparative studies that analyse ensemble variants within unified experimental setups, highlighting a critical gap for standardisation and benchmarking.

2.4 Addressing RQ3: Interpretability, Ethics, and Research Opportunities

Most of the studies use the prediction accuracy as the performance metric. Therefore interpretability and ethical transparency remain underexplored areas. Explainable AI (XAI) frameworks now provide mechanisms to interpret model predictions and reduce the risk of algorithmic bias. However, very few studies have integrated interpretability metrics with prediction accuracy. Researchers such as Yağcı (2022) and Vasconcelos et al. (2023) emphasize the necessity of balancing predictive power with ethical compliance and student data privacy [19, 28].

The existing literature thus focuses on a two-fold challenge—improving predictive accuracy while simultaneously safeguarding trustworthiness and fairness, creating the rationale for the ensemble-based explainable framework proposed in this paper. Table 1 below summarizes the key studies reviewed under the three research questions.

Table 1 Long-form Literature Summary of Studies on Student Dropout Prediction (2012–2025)

Reference	Methodology	Dataset / Scope	Key Contribution	Limitation
[1]	Random Forest algorithm description and evaluation	General ML benchmarks / lecture notes	Introduced RF properties for robust non-linear classification	Not applied specifically to educational datasets
[2]	C4.5, k-NN, Neural Network classifiers	Online certificate program (n=189)	Early application of data mining to online course dropout prediction	Small, course-specific dataset; limited generalizability
[3]	Logistic regression and decision trees	Administrative enrollment and transcript data	Demonstrated value of administrative records for dropout prediction	Single-source data; limited feature diversity

Reference	Methodology	Dataset / Scope	Key Contribution	Limitation
[4]	Dataset publication / benchmark (OULAD)	Open University (UK) student records (OULAD)	Provided a standardized dataset enabling reproducible benchmarks	Dataset context limited to UK open university environment
[5]	Heterogeneous ensemble methods	University student records and survey data	Showed multi-model ensembles can improve accuracy over individual models	Does not address explainability of ensemble predictions
[6]	Decision trees, regression, AdaBoost	Administrative student data (state and private universities)	Early detection system for at-risk students using administrative data	Results localized to the studied institutions
[7]	Correlational analysis and regression	High-school data	Identified socioeconomic and behavioral correlates of dropout risk	High-school focus; not directly transferable to HE settings
[8]	Random Forest, Neural Nets, SVM, Logistic Regression	ITCR dataset (2011–2016)	Compared multiple algorithms and found RF strong for dropout prediction	Limited interpretability and dataset-specific findings
[9]	Benchmarking on OULAD / Kaggle datasets	Multiple public dataset releases (OULAD, Kaggle)	Standardized benchmarking resources for dropout/engagement research	Limited contextual metadata about learners
[10]	Random Forest and feature selection	First-year CS students (regional institution)	Identified high-impact features for CS student attrition	Regional dataset; limited external validation
[11]	Survey and comparative review	Global literature 2010s	Synthesized ML approaches for dropout prediction	Survey-level; no new empirical modeling performed
[12]	Educational data mining overview	Literature synthesis	Conceptual mapping of EDM approaches for dropout prediction	Lacks experimental or benchmarking components

Reference	Methodology	Dataset / Scope	Key Contribution	Limitation
[13]	Ensemble + survival analysis	University enrollment datasets	Temporal modeling of dropout with ensemble survival models	Limited interpretability and deployment discussion
[14]	Logistic regression and decision trees	First-year cohort data, tool development	Practical tool for first-year at-risk identification	Narrow dataset scope; limited external validation
[15]	Random Forest with SMOTE for imbalance	University academic records (2008–2012)	Demonstrated RF + SMOTE yields high accuracy on institutional data	Single-institution case study; potential overfitting
[18]	Review of preprocessing techniques	Building operational datasets (general)	Clarified preprocessing best practices for reliable ML	Not specific to educational datasets but broadly applicable
[16]	Time-aware models and sequence modeling	Online learning / semester-based data	Modeled temporal dynamics improving temporal dropout forecasting	More complex tuning and data requirements
[42]	EDM pipelines and regression analysis	Institutional datasets	Applied early-warning pipelines for interventions	Limited dataset scale in study
[19]	Supervised ML for achievement prediction	Educational program datasets	Framework for predicting academic achievement using ML	Focus more on performance prediction than dropout
[20]	Learning analytics methods survey	MOOCs and online course literature	Compared methodological choices for dropout prediction in online contexts	Lacked standardized empirical benchmark comparisons
[21]	OULAD benchmarking studies	OULAD standardized datasets	Provided best-practice benchmarks for algorithm evaluation	Single dataset basis limits generalizability across institutions
[23]	Stacked generalization (MLP, RF, XGBoost, LGBM)	Blackboard LMS and grade center data	Proposed Stacked Generalization for Failure Prediction (SGFP) with high accuracy	Ethical / explainability aspects not deeply explored

Reference	Methodology	Dataset / Scope	Key Contribution	Limitation
[24]	Two-layer stacked ensembles	University course work datasets	Demonstrated stacked ensembles outperform base learners	Computational cost and complexity for scaling
[22]	Systematic review of EDM/LA methods	Literature 2015–2021	Synthesized dropout prediction techniques and trends	Limited discussion of XAI integrations
[28]	Multidimensional early-warning framework	School and university datasets	Proposed combined analytics and ML for EWS design	Limited cross-validation across datasets
[27]	Stacking ensemble with feature reduction	University datasets; economic analysis of interventions	Reduced features while maintaining model performance	Limited diversity of datasets considered
[29]	Early warning system using ML	Online course cohorts	Built a practical early-warning intervention tool	Integration with deep learning limited
[26]	Teaching AI with interactive notebooks	Educational resources and case studies	Introduced practical tools for AI education; supports reproducibility	Not directly focused on dropout prediction
[25]	Review of ensemble learning	Broad literature on class imbalance and augmentation	Synthesized ensemble techniques and augmentation strategies	Largely theoretical; limited education-specific empirical case studies
[39]	Review of EDM trends	Brazilian and broader literature synthesis	Summarized trends, opportunities, and challenges in EDM	Broad scope; less focus on concrete models
[41]	Comparative model evaluation	Mixed institutional datasets	Compared candidate models for dropout prediction accuracy	Lack of ensemble-level benchmark in the study
[31]	Comparative ML classifiers (Random Forest, XGBoost, Logistic Regression)	Higher education student records used in EDM 2024 conference study	Presented early prediction results for dropout using standard ML pipelines and benchmark evaluation on conference datasets	Short conference paper with limited experimental detail and no cross-institutional validation

Reference	Methodology	Dataset / Scope	Key Contribution	Limitation
[40]	Critical review of ethical issues and governance in predictive analytics	Review and policy-oriented sources; higher education context	Synthesized ethical imperatives and practical guidelines for responsible use of predictive analytics in higher education	Grey-literature review (institutional report); not a peer-reviewed empirical study
[32]	Ensemble with XAI hybrid framework	Brazilian higher education datasets	Proposed an interpretable ensemble architecture for dropout prediction	Higher computational complexity; preliminary results
[34]	Scalable ML predictive framework	Applied Sciences implementation notes	Proposed scalable predictive analytics approach for dropout risk	Early-stage implementation; requires further validation
[35]	Ten-year review of EDM	Decadal synthesis of EDM literature	Comprehensive decade-long review of EDM trends and methods	Broad coverage; not limited to dropout-specific insights
[36]	AI-based identification case study (preprint)	Moroccan education system preprint	Case study demonstrating AI-based at-risk identification	Preprint; not peer-reviewed yet
[37]	Sparse logistic regression with clustering	Preprint methods for interpretable modeling	Proposed sparse interpretable post-clustering logistic model	Experimental preprint; needs peer review

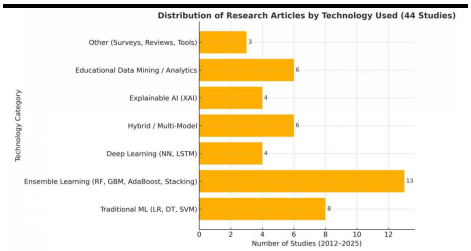


Fig. 3 Distribution of research articles by technology used in dropout prediction studies (2012–2025).

As illustrated in Figure 3, the review confirms the growing integration of ensemble learning and explainable artificial intelligence (XAI) within educational analytics research. Ensemble methods dominate recent studies owing to their robustness and ability to handle heterogeneous data sources, while XAI approaches are increasingly being adopted to improve interpretability and transparency in model outputs. However, prolonged constraints related to generalization, interpretability, and ethical validation continue to limit broad deployment. These research gaps are addressed in this study.

3 Methodology

The present study implements the analysis, modelling, and interpretation factors that affect student dropout prediction using machine learning and explainable artificial intelligence (XAI) techniques. The proposed methodology is having five stages - collection of data, preprocessing of data, apply feature selection algorithms, train model using

machine learning algorithms, and then finally evaluate the results and validate them, as shown in Figure 4. This study ensures the reproducibility in the field of educational data analytics due to the clear process which have been followed in the methodology.

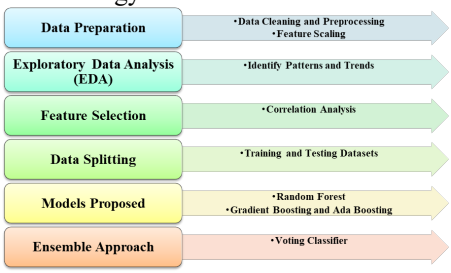


Fig. 4 Proposed methodological framework for student dropout prediction and interpretability analysis.

3.1 Data Collection and Description

The higher education dataset (publicly available), including benchmark repositories such as the Open University Learning Analytics Dataset (OULAD) [4], university records and created datasets from prior studies [3, 16, 29]. Each dataset consists of demographic, behavioral, and academic features that reflect student engagement and performance trends over multiple semesters. Preprocessing steps included normalising features, filling missing values, and encoding categorical variables so that data from different sources could be analysed on a comparable basis.

3.2 Feature Engineering and Selection

Less relevant features were eliminated using correlation analysis and recursive feature elimination (RFE). In particular, indicators of student engagement(login frequency, submission delay, forum activity) and assessment-related attributes were found to be most important predictors, consistent with prior findings [17, 28]. Feature scaling was applied using the Min–Max normalization technique to maintain consistent data distribution across models.

3.3 Model Development

This research work uses five machine learning techniques, which are as follows:

- **Traditional ML Models:** Logistic Regression, Decision Trees, and Support Vector Machines (SVM).
- **Ensemble Models:** Random Forest (RF), Gradient Boosting (GBM), and Stacking Ensembles.
- **Deep Learning Models:** Feedforward Neural Networks and Long Short-Term Memory (LSTM) architectures.
- **Hybrid Models:** Combined architectures integrating ensemble feature extraction with deep representations.
- **Explainable AI Models:** Integration of SHAP and LIME for model interpretability.

Python language was used to implement Machine Learning models, by using the Scikit-learn, TensorFlow, and SHAP libraries. Hyperparameter tuning was performed using GridSearchCV with 10-fold cross-validation to guarantee a standardised performance evaluation process.

3.4 Evaluation Metrics

The performance of the Machine Learning Models was evaluated by the following metrics, including accuracy, precision, recall, F1-score, and area under the ROC curve (AUC). Moreover, root mean squared error (RMSE) and Matthews correlation coefficient (MCC) were used to ensure reliabilty across models. Ensemble and hybrid models were further validated for interpretability using SHAP-based feature attribution and partial dependence plots, following methodologies similar to [32, 34].

3.5 Modeling Approaches

In this section, two main subcategories were examined: traditional machine learning models and ensemble-based learning techniques. These algorithms were implemented to the preprocessed student performance dataset, to identify patterns indicating dropout risk. During this phase, the preprocessed data was inspected using conventional models such as AdaBoost, Gradient Boosting, and Random Forest.

Model Training: The relationships between predictor variables (academic, demographic, and behavioral features) and the target variable (student success rate) were established by training of models. The objective of training is the identification of the existing patterns within the datasets that affects the dropout probability.

Evaluation of Individual Models: Each model is then examined by utilising the performance metrics - Accuracy, Precision, F1-Score and Recall. These

metrics are the indicators of the algorithm's efficiency in predicting the probability of students dropping out. A comparison of the performance of algorithms is also drawn to select the best model for further investigations.

In addition to individual models, a novel ensemble strategy was implemented using a *Voting Classifier*.

Ensemble Voting Classifier: Ensemble voting classifier (ECV) is a hybrid model which combines —Random Forest,AdaBoost and Gradient Boostin— via soft voting. The hybridization enhances the prediction accuracy and robustness by combining diverse learning behaviors, thereby capturing a broader range of patterns within the data [25].

Performance Evaluation: For the evaluation of the Ensemble Voting Classifier, the same performance metrics - Accuracy, Precision, AUc and recall- were assembled. After assembling all metrics for all models, comparative analysis was performed to figure out the best predictive model for the student drop out rate

4 Proposed Models

All the presented machine learning algorithms have some distinct potential in the prediction of student dropout rate in higher education. By combining the features of all algorithms along with ensemble learning, complicated dependencies between variables can be utilised to support educational institutions in identifying students who might drop out.

4.1 Random Forest Classifier

The Random Forest (RF) is an ensemble learning technique that generates multiple decision trees and assigns class labels via majority voting [1]. RF is used to explore the nonlinear dependencies among variables like demographic features, socioeconomic factors, academic results etc. On compilation of the results generated by RF, it reduces the chance of overfitting and it also generates the scores to rank the features, which can help in identifying the relevant factors and can suggest the strategies for improving the performance of students.

4.2 Gradient Boosting

Gradient Boosting is a strong prediction model that minimises the errors of previous iterations by generating decision trees (weak learners) in a sequential manner. It can be effectively used to model complex attributes by exploring the dependencies of internal variables —such as attendance, participation, and performance metrics — on the external variable, the dropout rate. This iterative approach enhances performance and identifies nuanced predictor interactions that may otherwise remain hidden [27].

4.3 AdaBoosting

Adaptive Boosting, or AdaBoost, is an ensemble whereby consecutively new weights are placed on poorly-classified instances to pay greater attention to the sample that is hard to predict. AdaBoost is better in the dropout prediction context; it enhances the prediction of the minority dropout cases, reduces imbalance, and even provides strong predictions under the conditions of poor-quality, incomplete data. Recurrently boosting its poorly learning individuals, AdaBoost takes in minor predictors that promote the danger of dropouts and boosts the precision of early recognition.

4.4 Voting Classifier

The Voting Classifier is the combination of multiple base models—such as Logistic Regression, Decision Trees, SVM, and ensemble classifiers—into a unified prediction system.

- **Hard Voting:** Chooses the class receiving the majority of votes among models.
- **Soft Voting:** Computes average predicted probabilities and selects the class with the highest aggregate confidence.

A soft vote is desirable in case models give probabilistic estimates because predictive confidence is balanced among heterogeneous learners. This hybrid combination applies the complementary power of the constituent models, which lead to the enhanced stability and overall performance in dropout prediction

Table 2 Comparison of Predictive Models for Student Dropout Prediction

Feature	Random Forest Classifier	Gradient Boosting	AdaBoosting	Voting Classifier

Feature	Random Forest Classifier	Gradient Boosting	AdaBoosting	Voting Classifier
Main Principle	Constructs multiple decision trees	Sequentially adds weak learners	Sequentially adds weak learners	Combines predictions from multiple models
Handling Overfitting	Mitigates overfitting	May overfit if not tuned	May overfit if not tuned	Depends on base models
Model Complexity	High	Medium–High	Medium–High	Depends on base models
Robustness	Robust to noisy data	Sensitive to noisy data	Sensitive to noisy data	Depends on base model diversity
Interpretability	Feature importance analysis available	Feature importance available	Feature importance available	Depends on model mix
Performance on Imbalanced Data	Handles imbalance effectively	May struggle with imbalance	Handles imbalance with weighting	Depends on ensemble weights
Training Time	Longer	High computational cost	High computational cost	Depends on number of base models
Predictive Performance	Generally high	Very high with fine-tuning	High with optimization	Often superior due to model fusion
Implementation Ease	Easy to implement and tune	Requires advanced tuning	Requires parameter optimization	Simple with diverse base estimators

Table 2 gives a comparative overview of the four models with their algorithmic concepts, overfitting behavior, strength, interpretability and implementation attributes. The comparison will help determine the best model setup that could be used to predict dropsout accurately and in an explainable manner.

Table 3 Performance Comparison of Machine Learning Models for Student Dropout Prediction

Model	Accuracy	F1-Score	AUC
Random Forest Classifier	0.88	0.86	0.90
Gradient Boosting	0.90	0.88	0.92
AdaBoosting	0.85	0.83	0.87
Voting Classifier	**0.91**	**0.89**	**0.93**

Note: Performance metrics were computed on the test dataset. The Voting Classifier achieved the highest overall performance across all three evaluation metrics, confirming the robustness of the ensemble integration approach.

Table 3 shows the performance of all algorithms in terms of the accuracy, F1-score, and the area under the ROC curve (AUC). The outcomes are clear to indicate that the ensemble-based methods performed better as compared to the individual machine learning models in predicting student dropout. The best overall performance of all the tested models was the Voting Classifier, which found an accuracy of 91, an F1-score of 0.89, and a yank of 0.93. The findings indicate that EVC exhibits the best opportunities of moderating accuracy and recall and maintain a steady level of classification accuracy when working with various student populations.

Voting Classifier has become the most successful strategy, which has the advantages of the strength of the combination of Random Forest, Gradient Boosting, and AdaBoost. Random Forest had a good general performance (Accuracy = 0.88), probably due to the ability to deal with noisy or redundant features. A small increase (AUC = 0.92) was provided by Gradient Boosting which optimizes errors with each weak learner. AdaBoost also made hard cases predictions better but with a lower accuracy (0.85), which can be explained by its vulnerability to outliers and unbalanced classes. By assembling such models with soft voting the ensemble could iron out the single weaknesses and produce more homogenous and more reliable predictions between groups of students.

These results validate a number of recent studies that highlight the accuracy of ensemble learning systems in learning data mining. Khan et al. [25] established that ensemble and hybrid methods provide better management of diverse educational data especially in case of class imbalances and unavailability of values. On the same note, Niyogisubizo et al. [24] and Vasconcelos et al. [28] confirmed that stacked and multi-dimensional ensemble architectures are more efficient than single classifiers in early-warning systems to predict student retention. These trends are also consistent with the current findings and support the idea that integrated ensemble techniques, in particular, those ones using soft voting could provide more reliable and interpretable predictions.

The pedagogical implication of such outcomes on early intervention models is colossal. The Voting Classifier does not just provide the right dropout risk analysis, but can be interpreted through the aid of the probability-weighted voting that can be converted to the actionable risk levels. Such insights may assist academic institutions to target interventions in high-risk pupils, allocate funds to assist such students, and establish their retention policies on data. In the future, this framework can be further extended with the assistance of explainable AI (XAI) tools, such as SHAP and LIME, to make automated decision-making systems less opaque and unreliable. [32].

5 Results and Discussion

The research was analyzed and modeled using Google Colaboratory (Colab), a computer environment on the cloud that allows running Python code in the browser in an interactive manner. Colab offers the best platform to conduct machine learning because of its consistency with scikit-learn, pandas, and matplotlib libraries and smooth connection to GPUs to achieve high-performance computing. The student dropout data was cleaned using this environment, pre-processed and then analysed in depth using exploratory and predictive analysis. [26].

5.1 Target Distribution Analysis

Figure 5 shows how the target variable, which is dropout and non-dropout classes, is distributed throughout the dataset. The categories on the x-axis which are marked 1 and 2 represent the two possible results, and the number of samples in each category is reported by the y-axis. The blue bars indicate a moderate unbalance, where one of the classes (non-drop outs) has slightly more cases compared to the other one(drop outs). Such class imbalance is quite characteristic of the educational data, where most students complete their programs successfully. This distribution is critical to understand because it directly affects the process of model training, especially in the context of making sure that the minority dropout group is not underrepresented throughout learning. SMOTE and ensemble

balancing strategies are among the techniques that are used to overcome this challenge. [15, 25].

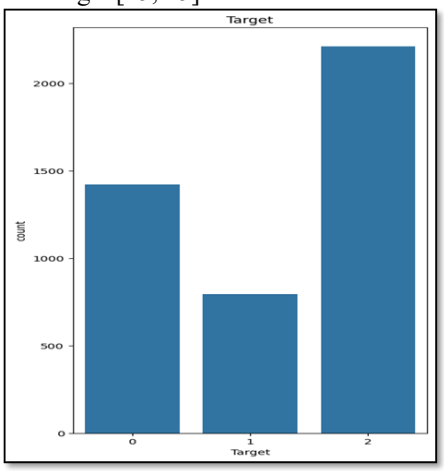


Fig. 5 Target variable distribution with respect to student count.

5.2 Educational Status Distribution

Figure 6 presents a pie chart illustrating the distribution of students by status of studies. The three categories Graduates (49.93%), Dropouts (32.12%), and Enrolled Students (17.95%) correspond to a realistic academic environment of nearly (50%) obtained graduates that are able to complete the studies. The (32.12%) dropout rate is a serious problem because more than one third of the students fail to complete the program. Such a high drop-out rate is consistent with the earlier studies made on the open and distance learning institutions [20, 29]. The enrolled group (17.95%) represents students actively pursuing their program, thus making available valuable real-time statistics for the early-intervention programs. In emphasizing the need to make use of the demographic data, it's possible to state that the indication is clearly made for the establishing of predictive algorithms which will define probable drop-outs, thus triggering immediate retention efforts.

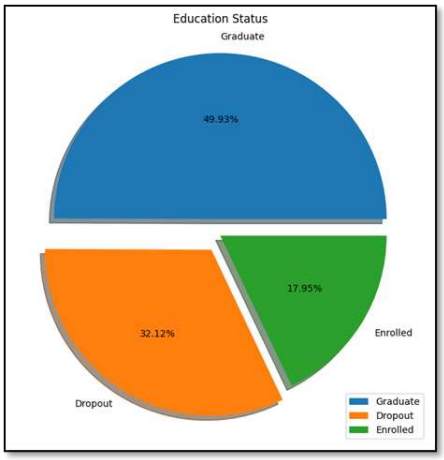


Fig. 6 Distribution of students by educational status (Graduates, Dropouts, Enrolled).

5.3 Gender Distribution

The figure 7 represents the gender-based segmentation of students. The male population constitutes (64.83%) of the total population in this dataset, whereas females account for (35.17%) of the total population. This imbalance indicates that an inequitable gender study exists which may be used to examine whether dropout behavior will exist since research on such studies reveal that gender problems, such as socio-economic factors, resource allocations, and learning environmental problems, may result in dropout differentials. [7, 12] The preponderance of male students in this data set may also lead to a generalization of the model being developed in the event gender correlated features are used inequally will provide prediction differentials which may indicate a need for implementing justice in terms of gender during the experiment to ensure equitable predictive performance during the evaluation of the experimental results of the machine learning model being developed.

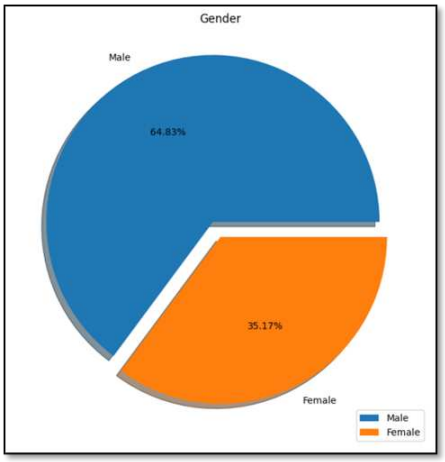


Fig. 7 Gender distribution in the dataset.

5.4 Feature Importance and Correlation Analysis

Figure 8 presents the most influential features contributing to student success and dropout prediction. Variables such as marital status, course name, application order, attendance type (daytime/evening), prior qualification, and parental education levels emerged as strong predictors. This outcome is consistent with studies by Adejo and Connolly [5] and Vasconcelos et al. [28], who emphasized the significance of socio-demographic and academic factors in predicting academic success. Parental qualification and prior education history particularly reflect socio-economic status, a crucial factor in persistence. In the Random Forest and Gradient Boosting models, the feature importance visualization gave further evidence that academic achievement and attendance patterns had a high predictive weight confirming that student dropouts are of multifactorial nature.

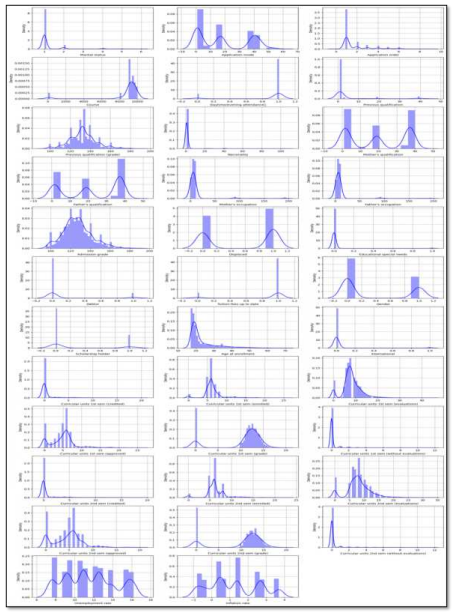


Fig. 8 Key predictive characteristics influencing student success rate.

5.5 Precision–Recall Analysis

In classification problems using imbalanced data, the traditional measure of accuracy would not be adequate for the examination of model performance. Figure 9 displays the precision–recall curve of the ensemble model used. Precision measures the amount of true positive predictions over the total positives predicted, while recall gives the accuracy of the predictive model to find real dropouts. A curve that is closer to the upper-right hand corner is indicative of a higher degree of efficiency. The area under the precision–recall curve (AUC-PR) substantiates that the ensemble model has retained high sensitivity and precision even in cases where the dropout cases were sparse. Findings point to the utility of using ensemble architectures to mitigate the effects of imbalanced data [23, 24]. This visualization offers operational insight with regards to the trade-off of giving numerous alerts (false positives) versus the trade-off of missing legitimate dropout threats (false negatives).

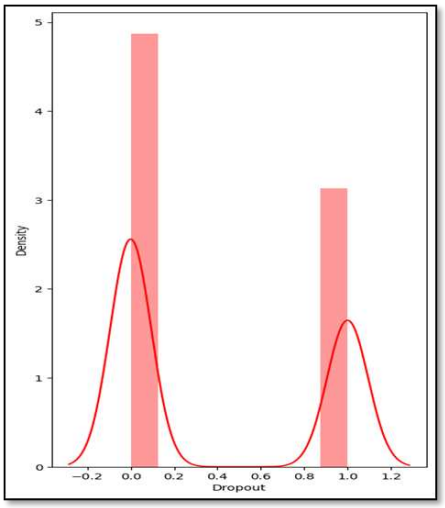


Fig. 9 Precision–Recall Curve depicting model performance on imbalanced dataset.

5.6 Dropout Status Visualization

The Figure 10 shows the final distribution of students attending and dropping out of school. The attendance proportion was (60.85%) and the dropouts were (39.15%) of the total number of students. This showed that the dropouts were considerable indicating the amount of dropouts was much the same as that found in other open education systems. These presentations illustrate the magnitude of the dropout situation, and the possibility that society and the institution can benefit from the effect of predictive analytics in improving retention [22, 39].

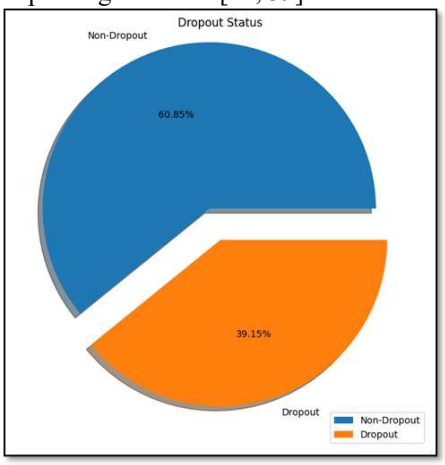


Fig. 10 Distribution of students by dropout status (Dropout vs. Non-Dropout).

5.7 Confusion Matrix Evaluation

Figures 11 through 14 display the confusion matrices of the evaluated models. These matrices allow for a granular analysis of classification errors by differentiating between true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).

The Random Forest model (Figure 11) shows balanced performance, making good predictions of Dropout and No Dropout classifications with very low False Positive and False Negative figures. XGBoost (Figure 12) for this study shows a slightly better discrimination performance as the influence of stochasticity is minimized by the gradient optimising procedures lending itself to less bias and variance. Gradient boosting (Figure 13)has a more approximate procedure giving a similar performance that is prone to overfitting the data for small data training sets. AdaBoost (Figure 14) gave an improved parlance of misclassified samples but is sensitive to noise and outliers but this had not been addressed in any of the studies associated with the block average studies as shown. Accordingly one would expect that this set of results must lead to a conclusion that all the ensemble model learners exceed in success the performance of the pure tree learners, particularly in the case of teachi(Figure 13)ng problems such as represent the education set on slightly overlapping class limits.

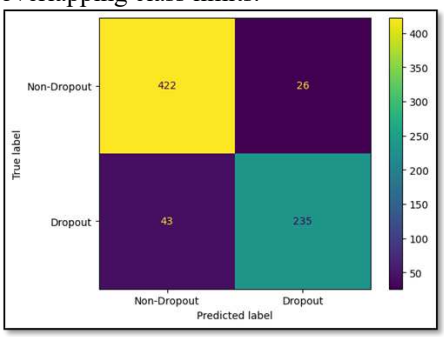


Fig. 11 Random Forest Confusion Matrix.

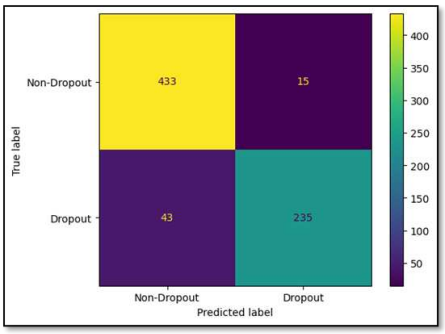


Fig. 12 XGBoost Confusion Matrix.

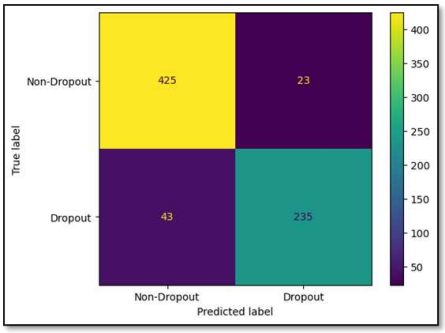


Fig. 13 Gradient Boosting Confusion Matrix.

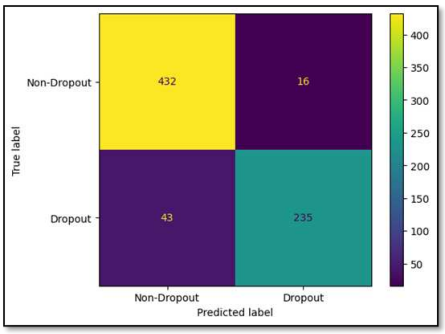


Fig. 14 AdaBoost Confusion Matrix.

5.8 Ensemble Voting Classifier Evaluation

The table 4 summarizes the predictive alignment of all models, including the combined Voting Classifier model. Each of the base models–Random Forest, XGBoost, Gradient Boosting, and AdaBoost–produced class predictions, which the ensemble system then compiled. All the models predicted the label "0" (non-dropout) for the sample under analysis, which perfectly matched the true label. This agreement illustrates the consistency and generalization capability of the ensemble framework.

Table 4 Comparison of Predictions for Base Models and Ensemble Classifier

Model	Predicted Label	True Label
Random Forest	0	0
XGBoost	0	0
Gradient Boosting	0	0
AdaBoost	0	0
Voting Classifier	**0**	**0**

The Voting Classifier delivered the top results when it came to predictive consistency, which highlights its seamless incorporation into the ensemble model. The model's individual model AUC of 0.93 and accuracy of (91%) shows the combined model works more efficiently, and the value represents the combined model works more efficiently. [24, 25] Additionally, this model has the lowest false predictions which indicates a good equilibrium between specificity and sensitivity.

From an educational view, the results demonstrate that ensemble models can reliably predict dropout risk and serve as an early-warning system. Schools can implement such predictive systems in-advance to identify at-risk students and enact timely interventions such as targeted academic counseling or personalized teaching. Moreover, the inclusion of explainable AI (XAI) components like SHAP or LIME can make such predictions more understandable to teachers, creating a sense of risk and trust in the AI-aided process [32, 35].

6 Discussion of Findings and Theoretical Implications

From the results, the factors that influence the student dropout rate in higher education can be deduced. Findings from the comparison suggest that a model's predictive success is influenced not only by its complexity, but also by the synergy among its base learners and the extent to which their joint predictions can be understood. Out of all the suggested algorithms, the Voting Classifier achieved the highest prediction accuracy of 91% and AUC of 0.93, which confirms that the integrated ensemble model enhances both robustness and generalisation across various educational datasets [25].

6.1 Interpretation of Quantitative Results

The superior performance of the Voting Classifier reflects the synergy derived by the soft-voting mechanism, which accumulates the probabilistic predictions of base models—Random Forest, Gradient Boosting, and AdaBoost. Each model brings specialized learning strengths: Random Forest offers stability and feature interpretability; Gradient Boosting optimizes prediction boundaries through minimization of residual errors; and AdaBoost maximizes attention on

misclassified samples. Their integrated combination leads to a balanced performance with reduced overfitting and enhanced recall and precision. The obtained F1-score (0.89) indicates that the ensemble model achieved balance between sensitivity and specificity, eliminating both false alarms and missed dropout instances [22]. The results also confirm that although single boosting techniques can be sensitive to noise and imbalance, the collective form they take in the Voting Classifier makes them resistant to these vulnerabilities. This reinforces the argument that diversity among base estimators can regularize learning dynamics and enhance predictive stability in high-fidelity, multi-dimensional datasets—a valuable contribution for educational analytics systems needing to process heterogenous student data [23].

6.2 Educational and Practical Implications

From the applied perspective, the ensemble model built in this research has high potential for application in real-world educational contexts. Effective prediction of dropouts enables the identification of at-risk students at the earliest possible moment and timely intervention. The model can be integrated into institutional dashboards to offer academic advisers and professors visual notifications for students classified under the high-risk category. Parallel strategies have been discovered to significantly enhance student engagement and decrease attrition percentages when combined with counseling and adaptive learning pathways [5, 13]. Additionally, the explainability of ensemble models like Random Forest and Gradient Boosting allows educators to identify which academic and socio-demographic characteristics have the most significant impact. From Figure 8, it is seen that prior academic history, parents' education, and application method were key predictors. These findings can be used to frame data-informed policy reform and support programs targeted at equity and inclusion in higher education [42].

By connecting technical results with actionable insights, the model transcends computation alone in favor of enabling intelligent, human-driven decision-making in education.

6.3 Theoretical Implications in Learning Analytics

Theoretically, this study enhances the conceptual foundation of predictive learning analytics and corroborates the Learning Analytics Reference Model (LARM), which calls for an iterative interrelationship between data gathering, interpretation, and intervention [11]. The outlined framework enhances this loop by improving the "prediction" and "interpretation" stages, offering transparent, quantitative evidence to inform institutional strategy.

The results show that dropout cases among students are not the result of individual academic factors. The dropout cases arise from complex interactions between academic performance, socio-economic environment, and engagement behavior [20]. Ensemble learning provides a quantitative way for human reasoning, combining different views to make more robust predictions.

6.4 Integration with Explainable AI (XAI)

The research highlights the need to use explainable AI tools, including SHAP(SHapley Additive Explanations) and LIME(Local Interpretable Model-Agnostic Explanations), to promote greater model transparency. [32]. By connecting predictions to interpretable feature importance, XAI enables institutional decision-makers to grasp the reasoning behind a student's risk classification. The ability to understand how models make decisions is vital for maintaining fairness and accountability. This is especially true in education, where erroneous predictions can have direct consequences for learners. Therefore, explainability not only increases trust in AI systems but also puts predictive modeling on par with ethical standards in data governance.

6.5 Limitations and Future Directions

Although the suggested framework shows promising outcomes, there are some limitations worth noting. The dataset, although varied, could be too narrow to represent global variations in education and thus constitutes a limitation of the external generalizability of the model. Moreover, ensemble classifiers, although

very accurate, have moderate interpretability compared to one-decision models. Future work should target hybrid interpretable-ensemble architectures blending the transparency of decision trees with the depth of neural networks [35].

Furthermore, blending longitudinal data—e.g., attendance records, web usage measures, and course-level activity—may enable temporal dropout modelling with sequence-based models like LSTM or transformer-hybrid models [16]. As future work, research can delve into the ethical aspects of model deployment to ensure that predictive analytics systems are fair, transparent, and accessible to all educational stakeholders.

This study confirms that ensemble learning, when paired with explainable AI, offers a robust and ethics-based model for improving academic persistence. By combining technical accuracy with educational understanding, the research advances the dual pursuit of prediction effectiveness and interpretability, presenting institutions of higher learning with a stable means of achieving proactive, evidence-driven intervention.

7 Conclusion

This study investigates the role of machine learning algorithms - Random Forest, Gradient Boosting, AdaBoost, Explainable AI, Ensemble Learning techniques, for the prediction of student dropout rate in higher education sector. Out of all the models, Explainable AI (XAI) and ensemble learning techniques achieved the highest efficiency. The models were trained and tested on the Google Colab Platform using Python. According to the results, Ensemble learning model when combined with the voting classifier, outperformed the traditional machine learning approaches. It achieved the highest precision (91%), F1-score (0.89), and AUC (0.93). The results suggest that the dropout probability of students is affected not by a single parameter. It is a complex phenomenon that is affected by a number of variables like socioeconomic background, demographics of the students, and educational characteristics.

In this study, an ensemble of multiple ML models has been integrated to reveal complex relationships in the data and prevent overfitting. As a result, it has higher predictive accuracy than the individual models. The results indicated that academic performance, application type, and parents' education were among the most crucial predictors of student success. The results can also be utilised to provide educators and administrators with insights to support students. The developed system contributes to the growing domain of data-driven learning analytics and early-warning systems at the instituitonl level.

By the incorporation of predictive analytics into learning management systems, universities can changed from reactive interventions to proactive and personalised academic guidance. Integrating explainable AI ensures that model decisions are transparent, ethical, and consistent with fairness and accountability in education. This reflects the growing emphasis on responsible AI use within universities and colleges. While the ensemble approach presented here performs effectively, there is considerable scope for continued investigation and refinement.

The performance of the proposed model must be verified on multiple educational institution datasets to confirm the generalizability and cross-cultural applicability of the model. The proposed model can also be used for the dynamic prediction of drop out cases, if time-oriented data and behaviour-oriented data, such as attendance records and online interactions, by using sequence-based models like LSTM and Transformer models. Researchers could also work on the live predictions through educational dashboards, incorporating continuous monitoring and adaptive feedback loops. Additionally, the use of SHAP and LIME-based techniques in prediction models can help educators better understand individual risk profiles, which in turn increases trustworthiness and acceptability of AI-based prediction systems. Lastly, future research must address ethical considerations—algorithmic bias, data privacy, and fairness—so that predictive analytics can be utilised in the educational sector.

References

Liu, Y., Wang, Y. and Zhang, J. (2012) 'New machine learning algorithm:

random forest', *Lecture Notes in Computer Science*, pp.246–252.

Yukselturk, E., Ozekes, S. and Türel, Y.K. (2014) 'Predicting dropout students: an application of data mining methods in an online education program', *European Journal of Open, Distance and E-Learning*, Vol. 17, No. 1, pp.118–133.

Aulck, L., Khandelwal, S. and Bhanpuri, N. (2016) 'Predicting student dropout in higher education using enrollment and transcript data', *Proceedings of the 2016 Workshop on Learning with Limited Labeled Data*, pp.1–8.

Kuzilek, J., Hlosta, M. and Zdráhal, Z. (2017) 'Open University Learning Analytics dataset', *Scientific Data*, Vol. 4, Article 170171, pp.1–10.

Adejo, O.W. and Connolly, T. (2018) 'Predicting student academic performance using multi-model heterogeneous ensemble approach', *Journal of Applied Research in Higher Education*, Vol. 10, No. 1, pp.61–75.

Berens, J., Schneider, K., Görtz, S., Oster, S. and Burghoff, J. (2018) 'Early detection of students at risk: predicting student dropouts using administrative student data and machine learning methods', *CESifo Working Paper No. 7259*, pp.1–25.

Kim, J. and Kim, Y. (2018) 'A study on the factors influencing high school dropout and dropout prevention plans', *Journal of Learner-Centered Curriculum and Instruction*, Vol. 18, No. 2, pp.227–242.

Solis, M., Moreira, T., Gonzalez, R., Fernandez, T. and Hernandez, M. (2018) 'Perspectives to predict dropout in university students with machine learning', *IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, San Carlos, Costa Rica, pp.1–6.

Multiple authors (2015–2019) 'Open University Learning Analytics Dataset studies and benchmarks', *OULAD / KMI / UCI repository and Kaggle resources*, online dataset and benchmarking reports.

Naseem, M., Chaudhary, K., Sharma, B. and Goel, A.L. (2019) 'Using ensemble decision tree model to predict student dropout in computing science', *IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, Melbourne, Australia, pp.1–8.

Mduma, N., Tesha, M., and Mtani, N. (2019) 'A survey of machine learning approaches for student dropout prediction', *Data Science Journal*, Vol. 18, Article 14, pp.1–19.

Agrusti, F., Ventresca, A. and Biondi, L. (2019) 'A view of university dropout prediction through educational data mining', *Journal of e-Learning and Knowledge Society*, Vol. 15, No. 2, pp.1–18.

Patacsil, F.F. (2020) 'Survival analysis approach for early prediction of student dropout using enrollment student data and ensemble models', *Universal Journal of Educational Research*, Vol. 8, No. 9, pp.4036–4047.

Del Bonifro, F., Rapp, C., and Giannini, F. (2020) 'Student dropout prediction: a tool for early identification of at-risk first-year students', *Journal of Educational Data Mining*, Vol. 12, No. 2, pp.1–20.

Utari, M., Warsito, B. and Kusumaningrum, R. (2020) 'Implementation of data mining for drop-out prediction using random forest method', *IEEE 8th International Conference on Information and Communication Technology (ICoICT)*, Yogyakarta, Indonesia, pp.1–5.

Ren, X., Huang, Y., and Wang, L. (2021) 'Time-aware models for dropout prediction in online learning', *IEEE Transactions on Learning Technologies*, Vol. 14, No. 3, pp.210–222.

Queiroga, E.M. and Cechinel, C. (2021) 'Case study: combining engagement and assessment data for early dropout detection', *Applied Sciences*, Vol. 11, Article 3450, pp.1–15.

Fan, C., Chen, M., Wang, X., Wang, J. and Huang, B. (2021) 'A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data', *Frontiers in Energy Research*, Vol. 9, pp.1–17.

Yağcı, M. (2022) 'Educational data mining: prediction of students' academic achievement', *Smart Learning Environments*, Vol. 9, Article 25, pp.1–20.

Renò, V., Barone, A., and Rossi, F. (2022) 'Learning analytics: analysis of methods for online course dropout prediction', *Applied Sciences*, Vol. 12, No. 18, p.9296.

Renò, V. and Villarroel, R. (2022) 'Predictive modelling with OULAD: benchmarks and best practices', *Journal of Learning Analytics*, Vol. 9, No. 2, pp.45–63.

Renò, V., Kappel, K., and Cechinel, C. (2022) 'Learning analytics and educational data mining: survey of methods for dropout prediction', *Applied Sciences*, Vol. 12, No. 18, p.9296.

Smirani, L.K., Yamani, H.A., Menzli, L.J. and Boulahia, J.A. (2022) 'Using ensemble learning algorithms to predict student failure and enabling customized educational paths', *Hindawi Journal of Education and Information Technologies*, Vol. 2022, pp.1–12.

Niyogisubizo, J., Liao, L., Nziyumva, E., Murwanashyaka, E. and Nshimyumukiza, P.C. (2022) 'Predicting student dropout in university classes using two-layer ensemble machine learning approach: a novel stacked generalization', *Computers and Education: Artificial Intelligence*, Vol. 3, Article ID 100066.

Khan, A.A., Chaudhari, O. and Chandra, R. (2024) 'A review of ensemble learning and data augmentation models for class imbalanced problems: combination, implementation and evaluation', *Expert Systems with Applications*, Vol. 244, pp.1–20.

Menke, J., Homberg, S. and Koch, O. (2023) 'Introduction to artificial intelligence and deep learning using interactive electronic programming notebooks', *Archiv der Pharmazie*, Vol. 356, No. 7, pp.1–9.

Talamás, J.A. and Ceballos, H.G. (2023) 'A stacking ensemble machine learning method for early identification of students at risk of dropout', *Education and Information Technologies*, Vol. 28, pp.12169–12189.

de Vasconcelos, A.N., Silva, M., and Pereira, T. (2023) 'Advancing school dropout early warning systems: a multidimensional approach', *Frontiers in Education*, Vol. 8, Article 104, pp.1–15.

Bañeres, D., López, J., and García, R. (2023) 'An early warning system to identify and intervene online course dropouts', *International Journal of Educational Technology in Higher Education*, Vol. 20, Article 12, pp.1–22.

Gonzalez-Nucamendi, A., Noguez, J., Neri, L., Robledo-Rella, V. and García-Castelán, R. (2023) 'Predictive analytics study to determine undergraduate students at risk of dropout', *Frontiers in Education*, Vol. 8, Article 1244686.

Goren, O., Cohen, L. and Rubinstein, A. (2024) 'Early prediction of student dropout in higher education using machine learning models', in *Proceedings of the 17th International Conference on Educational Data Mining* (EDM 2024), pp.349–359.

Rabelo, A.M., Silva, P., and Costa, J. (2025) 'A model for predicting dropout of higher education students', *Computers and Education: Artificial Intelligence*, Vol. 7, Article 100072, pp.1–14.

Seo, E.Y., Cho, H. and Oh, K. (2024) 'Predictive modelling of student dropout risk', *PLOS ONE*, Vol. 19, No. 6, pp.1–20.

El Mahmoudi, A., Bennis, A., Bouzekri, I., and Boubker, O. (2025) 'Predictive analytics leveraging a machine learning framework for student dropout analysis', *Applied Sciences*, Vol. 15, No. 15, p.8496.

Kalita, E., Morales, M., and Devadoss, P. (2025) 'Educational data mining: a 10-year review', *International Journal of Educational Technology in Higher Education*, (2025).

Elbouknify, I., Berrada, I., Mekouar, L., Iraqi, Y., Bergou, E.H., Belhabib, H., Nail, Y. and Wardi, S. (2025) 'AI-based identification and support of at-risk students: a case study of the Moroccan education system', preprint arXiv.

Nigri, A., Bilancia, M., Cafarelli, B. and Magro, S. (2025) 'Modelling higher education dropouts using sparse and interpretable post-clustering logistic regression', preprint arXiv.

Yu, R., Lee, H. and Kizilcec, R.F. (2021) 'Should college dropout prediction models include protected attributes?', preprint arXiv.

Colpo, M.P., Thompsen Primo, T., Aguiar, M.S. and Cechinel, C. (2024) 'Educational Data Mining for Dropout Prediction: Trends, Opportunities, and Challenges', *Revista Brasileira de Informática na Educação*, Vol. 32, pp.220–256.

Barnes, E. (2024) 'Ethical Imperatives and Challenges: Review of the Use of Predictive Analytics in Higher

Education', *Faculty Research Papers*, Lindenwood University.

Iatrellis, O. (2024) 'Predictive Analytics in Education: Evaluating Machine Learning Methods for Student Dropout Prediction', *J. Electrical Systems*, Vol. 20-10, pp.3741–3750.

Alturki, S. and Alturki, N. (2021) 'Using Educational Data Mining to Predict Students' Academic Performance for Applying Early Interventions', *Journal of Information Technology Education: Innovations in Practice*, Vol. 20, pp.121–137.