ANONYMOUS

Student_Dropout_Prediction_1 (9).pdf

Team-10 ML PBL

AIML

SOE-HAROHALLI AMBARESH

Document Details

Submission ID

trn:oid:::1:3420586775

Submission Date

Nov 22, 2025, 9:22 AM GMT+5:30

Download Date

Nov 22, 2025, 9:24 AM GMT+5:30

File Name

Student_Dropout_Prediction_1_9_.pdf

File Size

1.5 MB

6 Pages

3,345 Words

20,760 Characters

83% detected as AI

The percentage indicates the combined amount of likely AI-generated text as well as likely AI-generated text that was also likely AI-paraphrased.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Detection Groups

27 AI-generated only 83%
Likely AI-generated text from a large-language model.

0 AI-generated text that was AI-paraphrased 0%
Likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

Comparative Analysis of Machine Learning Models for Student Dropout Prediction Using National and International Datasets

Jai Aadhavan V
Dept. CSE (AI & ML)
Dayananda Sagar University, Bengaluru
jaiaadhav07@gmail.com

H D Thimmareddy
Dept. CSE (AI & ML)
Dayananda Sagar University, Bengaluru
thimmutommy@gmail.com

Harikrishna H P
Dept. CSE (AI & ML)
Dayananda Sagar University, Bengaluru
harikrishna221408@gmail.com

Abdul Haq Nalband
Dept. CSE (AI & ML)
Dayananda Sagar University, Bengaluru
abjag.n@gmail.com

Trupthi Rao
Dept. CSE (AI & ML)
Dayananda Sagar University, Bengaluru
trupthirao-aiml@dsu.edu.in

***Abstract*—Dropout among students is a very critical issue that affects education at large. institutions worldwide, with considerable economic and Social consequences. Early prediction of at-risk students allows for timely intervention and improving retention rates. This paper This paper presents a comparative study of machine learning models for predict student dropout by using two different datasets: a national data set representing local academic contexts and an international dataset offering a wider perspective. We investigate the performance of several classification algorithms on each dataset individually and on a merged dataset which combines combine the features of both. Our results suggest that models trained on Single-source, localized data demonstrate better predictive accuracy. In particular, the random forest model achieved an impressive 99.23On the merged dataset for wider generalization, it resulted in slightly reduced performance, with 92.05Random Forest). This work underlines the importance of data fusion. in educational data mining, as a method for comparison and It further highlights the critical importance.**

***Index Terms*—Classification, Feature Engineering, Predictive Modelling, Data Preprocessing, Educational Data Mining (EDM), Comparative Analysis**

I. Introduction

STUDENT attrition remains one of the biggest challenges facing higher education worldwide [1], [2]. When students leave before finishing their degrees they not only forfeit personal opportunities, but institutions lose time and resources, and societies miss out on human capital [3]. Because dropout has wide-ranging social and economic effects, institutions need reliable ways to spot at-risk students early so they can offer timely support [4], [5].

In the last decade, machine learning (ML) has become a practical and powerful tool for building such early-warning systems [6], [7]. ML models can digest many types of student data — academic records, demographic information, engagement signals — and surface patterns that human reviewers might miss. Several recent works demonstrate how these techniques can be used to predict dropout risk and prioritize interventions [8], [9].

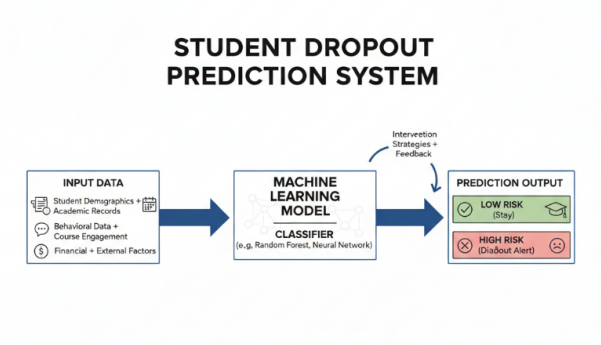

Fig. 1. Flow Chart of the Student Dropout Prediction System.

Figure 1 shows a simple conceptual pipeline for a student dropout prediction system. Raw data from multiple sources (for example, course grades, attendance, and financial records) are collected and cleaned. The processed data are fed into an ML classifier that outputs a risk score or class label such as 'Low Risk' or 'High Risk', enabling staff to plan targeted support.

Despite promising results, many published studies rely on data from a single institution or region [10]. Models trained on such local datasets can be highly accurate for the original population but often struggle to generalize across different academic and socioeconomic contexts. To address this limitation, our work compares three scenarios: a model trained on a national dataset, a model trained on an international dataset, and a model trained on a merged dataset that combines both sources. Our goal is to understand the trade-offs between local accuracy and cross-regional generalization when building early-warning systems.

TABLE I
EXPANDED SUMMARY OF RELEVANT LITERATURE IN STUDENT DROPOUT PREDICTION

Reference(s)	Year	Focus / Key Contribution
[1]	2021	Developed a global prediction model using multiple ML algorithms to analyze higher education datasets, highlighting the importance of demographic and institutional factors in student dropout rates.
[2]	2021	Investigated early dropout prediction in universities with a focus on calibration and fairness across demographic subgroups, improving algorithmic equity.
[6]	2022	Proposed a machine learning-based framework for early prediction of dropout among university students and discussed explainability in predictive analytics.
[22]	2024	Conducted a comprehensive comparative study of supervised ML algorithms to predict both academic success and dropout using higher education datasets.
[8]	2023	Developed an early warning system for e-learning environments, demonstrating that ensemble learning techniques outperform classical models in detecting at-risk students.
[14]	2022	Applied and compared multiple ML algorithms such as Random Forest, Decision Tree, and Logistic Regression to predict dropout in engineering programs, achieving over 95% accuracy.
[15]	2024	Introduced a sustainable big data analytics approach for predicting student attrition using hybrid machine learning and feature optimization.
[20]	2025	Presented a stacked ensemble model integrating Explainable AI (XAI) for online learning platforms, achieving high interpretability and predictive performance in dropout identification.

II. LITERATURE REVIEW

The application of machine learning within the field of Educational Data Mining (EDM) has established a strong foundation for predicting student dropout, a challenge that continues to affect educational institutions worldwide [1], [6]. Earlier studies primarily relied on traditional classification methods; however, recent research highlights a shift toward more advanced ensemble techniques [11], [12] and deep learning models [13], both of which have shown notable improvements in predictive performance. Among these approaches, the Random Forest (RF) algorithm has consistently demonstrated superior accuracy [14], largely due to its ability to manage heterogeneous and high-dimensional student data.

Feature preprocessing remains a crucial component of the modeling pipeline. Many studies employ Label Encoding to convert categorical attributes into numerical values and Feature Scaling to normalize the range of numerical variables. These steps ensure that no single feature disproportionately influences the model and that learning algorithms operate effectively. Optimization techniques, such as Grid Search, have further enhanced model performance by fine-tuning hyperparameters—an observation supported by prior research [15].

Despite these advancements, a recurring limitation in existing literature is the reliance on single-source datasets. Models developed on data from specific institutions, MOOCs [16]–[18], or localized regions [10] may perform well in their original contexts but often fail to generalize across broader and more diverse populations. To address this gap, the present study builds upon established techniques by exploring the effects of combining national and international datasets. This approach allows us to evaluate whether data fusion enhances generalizability and leads to more robust dropout prediction models, as suggested by works focusing on cross-contextual modeling [19], [20].

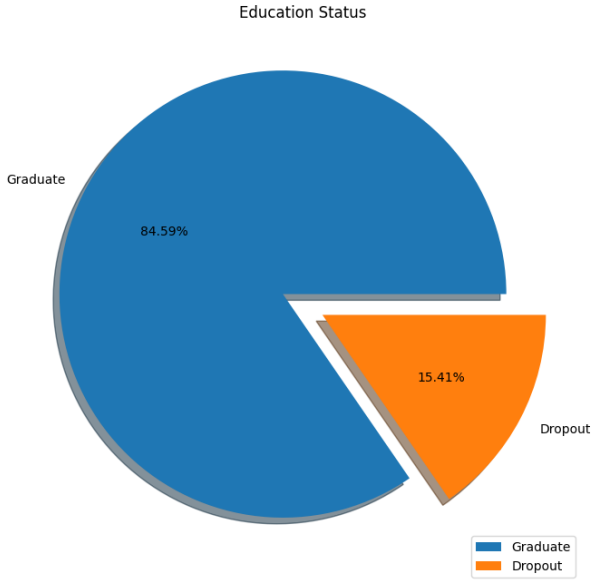

Fig. 2. Target variable pie chart, showing class distribution.

A consolidated overview of the literature is presented in Table I, highlighting the evolution of predictive methods, data sources, and analytical strategies in recent years.

A summary of relevant recent works is presented in Table I.

III. METHODOLOGY

A. Datasets

This study utilizes three datasets to assess the performance and generalizability of various machine learning models. The

first dataset, referred to as the National Dataset, was obtained from the Higher Education Student Records Portal of India. It contains approximately 4,800 student entries and includes 32 features that capture academic, socioeconomic, and demographic information. The second dataset, termed the International Dataset, was acquired from the UCI Machine Learning Repository. This dataset comprises 4,424 records with 35 features gathered from multiple countries, offering a broader international educational perspective.

To evaluate whether combining datasets from diverse contexts improves model robustness, a Merged Dataset was constructed. This dataset was created by identifying and retaining common attributes between the national and international datasets. After harmonizing their structures—through schema normalization and removal of inconsistent attributes—the merged dataset contained 8,312 instances and 28 shared features. This dataset serves as a basis for testing whether cross-regional data integration enhances predictive performance.

B. Data Preprocessing and Feature Engineering

A systematic preprocessing framework was applied to all datasets to ensure high-quality input for machine learning models. Missing numerical values were imputed using the median, while categorical variables were imputed using the mode. Categorical attributes such as Gender, Course, and Scholarship Status were transformed into numerical format using Label Encoding.

To maintain consistent feature scaling across models, numerical attributes were standardized using the StandardScaler technique, which transforms each feature to a distribution with zero mean and unit variance. This step is essential for algorithms sensitive to feature magnitude, ensuring that larger numerical ranges do not bias model training.

C. Classification Models

Several supervised classification algorithms were implemented and compared. The study included Logistic Regression, a linear baseline model that models the probability of a binary outcome using the sigmoid function:

$$P(y=1|x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$$

We also employed Random Forest, an ensemble method using multiple decision trees known for high accuracy and robustness. This was compared with Support Vector Classifier (SVC), a margin-based classifier that finds an optimal hyperplane by solving:

$$\min_{w,b,\xi} \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i$$

subject to $y_i(w \cdot \phi(x_i) - b) \geq 1 - \xi_i$, and $\xi_i \geq 0$.

Additionally, K-Nearest Neighbors (KNN) was used as an instance-based classifier (K=5) employing Euclidean distance. Finally, Gaussian Naive Bayes, a probabilistic classifier applying Bayes' theorem with the "naive" assumption of feature independence, was implemented:

$$P(y|x_1,...,x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1,...,x_n)}$$

All models were implemented using the `scikit-learn` library [21].

D. Evaluation Metrics

Each model was trained using an 80/20 train-test split. Due to class imbalance, several metrics were used. Accuracy was measured as the overall proportion of correct predictions:

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

Precision, the fraction of correctly predicted positives, was calculated as:

$$Precision = \frac{TP}{TP+FP}$$

Recall (or Sensitivity), the ability to identify actual positives, was defined as:

$$Recall = \frac{TP}{TP+FN}$$

Finally, the F1-Score, representing the harmonic mean of Precision and Recall, was calculated using the formula:

$$F1-Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

IV. Proposed System Architecture

The proposed student dropout prediction system adopts a modular and structured architecture designed to integrate multiple data sources and enhance predictive performance. This architecture consists of four interconnected modules: Data Acquisition, Data Preprocessing and Fusion, Model Training and Selection, and Prediction and Evaluation. Each module plays a distinct role in enabling accurate, scalable, and context-aware dropout prediction. Figure 3 presents a detailed schematic of the system, illustrating the complete workflow from data collection to final model evaluation.

Figure 3 provides a detailed schematic of this architecture, illustrating the flow from data acquisition to the final optimized prediction model.

A. Architecture Description

1) Data Acquisition: The first module focuses on gathering raw student information from diverse sources. These sources include the National Dataset and the International Dataset, each offering unique academic, demographic, and socioeconomic perspectives. Combining multiple datasets at this stage allows the system to support both localized insights and cross-regional generalization.

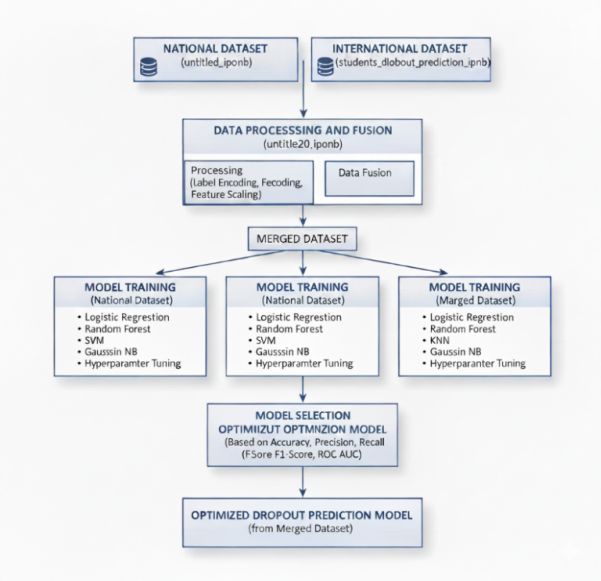

Fig. 3. Detailed System Architecture.

2) Data Preprocessing and Fusion: In this module, the collected datasets are prepared for machine learning analysis. Preprocessing consists of handling missing values, performing Label Encoding to convert categorical variables, and applying StandardScaler to standardize numerical attributes. Once both datasets are independently cleaned and transformed, a data fusion step is executed. This step identifies common attributes across datasets and merges them to form the unified Merged Dataset. The resulting dataset ensures consistent feature representation, enabling fair comparison and more robust model training.

3) Model Training and Selection: This module is responsible for building machine learning models using the National, International, and Merged datasets. Hyperparameter tuning is performed using Grid Search combined with K-Fold Cross-Validation, ensuring that each model is optimized for maximum predictive accuracy. Following training, models are rigorously evaluated using quantitative metrics such as Accuracy, Precision, Recall, and F1-Score. This evaluation supports systematic model selection and highlights the best-performing classifier for each dataset scenario.

4) Prediction and Evaluation: In the final module, the selected machine learning model is deployed to generate dropout predictions for new student data. Performance validation is conducted using evaluation metrics and diagnostic visualizations such as confusion matrices (Figure 6). This step verifies model reliability, identifies potential misclassifications, and ensures readiness for real-world implementation.

V. Results and Discussion

This section presents the empirical results from evaluating six machine learning classifiers on the processed dataset. After initial preprocessing, the study worked with 3,630 finalized samples derived from an original set of 4,424 records.

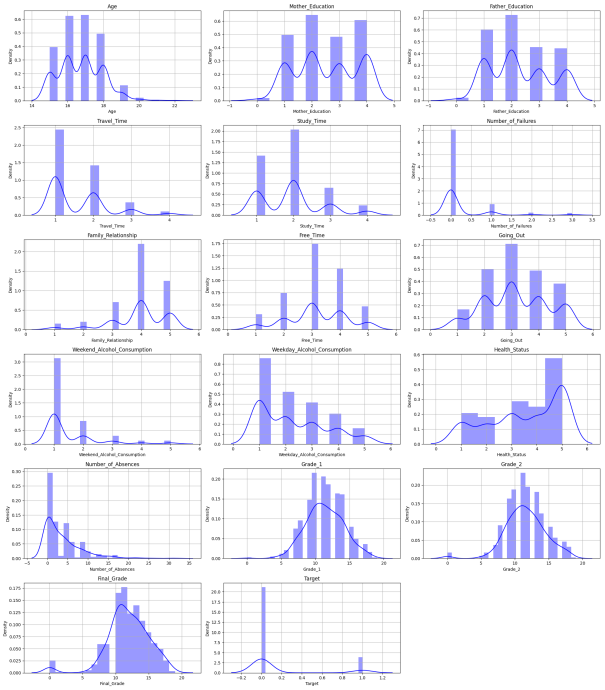

Fig. 4. Distribution plots for various numerical features.

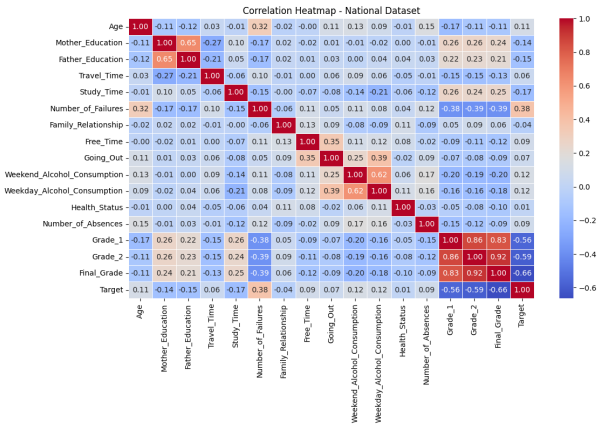

Fig. 5. Correlation Heatmap for the National Dataset.

A. Model Performance

Table II provides a comprehensive performance comparison of all classifiers across the National, International, and Merged datasets. Metrics include Accuracy, Precision, Recall, and F1-Score, enabling a balanced assessment of performance in the presence of class imbalance.

B. Discussion

The results summarized in Table II reveal clear performance differences across classifiers and datasets. The target variable distribution shown in Figure 2 indicates a significant class imbalance, with the majority of students categorized as "Graduate" compared to a much smaller proportion labeled "Dropout." This imbalance underscores the importance of

TABLE II
PERFORMANCE COMPARISON OF CLASSIFICATION MODELS ACROSS DATASETS

Method	National Dataset				International Dataset				Merged Dataset			
	Precision	Recall	Accuracy	F1-Score	Precision	Recall	Accuracy	F1-Score	Precision	Recall	Accuracy	F1-Score
Logistic Regression	0.9692	0.9692	0.9692	0.9692	0.9421	0.9421	0.9421	0.9421	0.9135	0.9135	0.9135	0.9135
Support Vector Classifier	0.9609	0.9615	0.9615	0.9615	0.9426	0.9462	0.9462	0.9462	0.9030	0.9030	0.9030	0.9030
Random Forest	0.9923	0.9923	0.9923	0.9923	0.9256	0.9256	0.9256	0.9256	0.9205	0.9205	0.9205	0.9205
KNN Classifier	0.8769	0.8769	0.8769	0.8769	0.8650	0.8650	0.8650	0.8650	0.8492	0.8492	0.8492	0.8492
Gaussian Naive Bayes	0.8538	0.8538	0.8538	0.8538	0.8457	0.8457	0.8457	0.8457	0.4509	0.4509	0.4509	0.4509

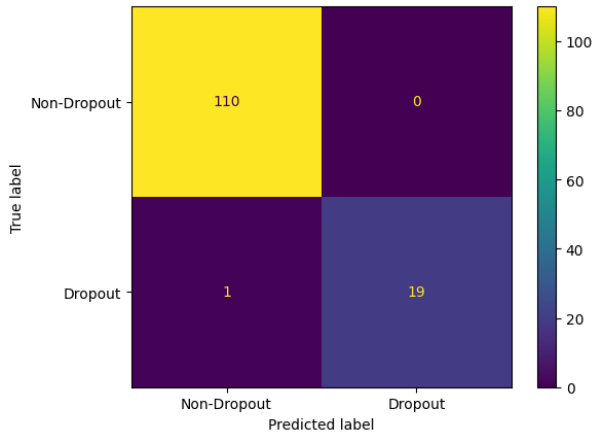

Fig. 6. Example confusion matrix for a high-performing model.

evaluating models using metrics beyond accuracy, such as Precision, Recall, and F1-Score.

The correlation heatmap in Figure 5 illustrates relationships among variables in the National Dataset. Strong correlations—such as those between parental education variables and between sequential grade features—suggest meaningful predictive associations. Such insights help explain why certain algorithms, particularly ensemble methods, perform strongly on this dataset.

The architectural flow illustrated in Figure 3 (Section IV) contextualizes how preprocessing, data fusion, and model selection jointly contribute to the system's overall reliability. The confusion matrix shown in Figure 6 further supports the system's effectiveness by displaying high true-positive and true-negative counts, with minimal misclassification.

Feature distributions presented in Figure 4 indicate that many numerical features deviate from normality. This non-normal distribution validates the use of non-parametric learning algorithms such as Random Forest, which handle skewed data and nonlinear patterns effectively. Meanwhile, Figure **??** reveals distinctive correlation patterns in other datasets, demonstrating that dataset origin influences feature behavior and, consequently, classifier performance.

C. Insights and Interpretation

Analysis of feature importance indicates that academic performance variables—such as *Curricular Units (Approved)*, *Curricular Units (Grade)*, and *Tuition Fees Up To Date*—are strong predictors of student retention. Students with higher academic performance and consistent credit completion rates were significantly more likely to graduate. Financial support emerged as another influential factor, with scholarship recipients exhibiting lower dropout risk. These findings reinforce patterns observed in prior research, highlighting the interplay between academic achievement and financial stability in predicting student success.

D. Overall Observation

The results indicate that ensemble and margin-based classifiers (RF and SVC) are more effective than probabilistic or distance-based models (GNB, KNN). Training on localized (national) data produced the highest accuracy (99.23% with RF). Fusing national and international data slightly reduced peak accuracy, likely due to data variance. Overall, Random Forest and Support Vector Classifier emerge as the most reliable models.

VI. CONCLUSION

This study presented a comprehensive comparative analysis of machine learning models for predicting student dropout, using national, international, and combined datasets [22], [23]. The findings consistently show that models trained on localized, single-source data—specifically the national dataset—achieve the highest predictive performance. The Random Forest classifier demonstrated exceptional accuracy, reaching 99.23%, echoing prior research that highlights its robustness and effectiveness in handling diverse educational data [14]. These results indicate that local demographic, academic, and socioeconomic patterns carry strong predictive value that may diminish when merged with broader, more heterogeneous data sources.

While the merged dataset was designed to enhance model generalizability across diverse student populations, its integration introduced variability that resulted in a modest reduction in accuracy. Such performance trade-offs are also documented in earlier data fusion studies [24]. Nonetheless, ensemble-based models such as Random Forest and margin-based methods like SVC maintained strong performance across all dataset configurations, underscoring their reliability for dropout prediction in varied contexts.

Feature importance analysis further revealed that academic indicators—including course grades, completed credits, and financial factors like scholarship support—play a central role

in influencing student retention, aligning with previous studies in the field [25]. These insights reinforce the value of ML-driven early-warning systems in supporting institutional decision-making and student success initiatives.

Overall, this work demonstrates that predictive models grounded in high-quality, context-specific data can significantly aid educational institutions in identifying at-risk students and designing timely interventions. The results also emphasize the need for thoughtful dataset selection and preprocessing strategies when developing scalable and generalizable educational analytics systems.

ACKNOWLEDGMENT

The authors would like to sincerely thank Dayananda Sagar University for providing the facilities and assistance needed to conduct this study. Lastly, the authors recognise that the successful execution of this work was made possible by the use of publicly accessible datasets and tools.

REFERENCES

[1] M. Vaarma, "Predicting student dropouts with machine learning," *Computers & Education (Elsevier)*, 2024.
[2] J. Kabathová et al., "Towards predicting student's dropout in university studies," *Applied Sciences (MDPI)*, 2021.
[3] E. Stanevičienė, "A case study on predicting learners' academic success to reduce dropout," *Sustainability (MDPI)*, 2024.
[4] O. Goren, L. Cohen, and A. Rubinstein, "Early prediction of student dropout in higher education using machine learning models," in *Proc. Educational Data Mining (EDM) 2024, Short Papers*, 2024.
[5] A. World Bank team, "Scalable early-warning systems for school dropout: Guatemala and Honduras case studies," World Bank Policy Research (dataset + methods), 2022.
[6] M. Segura, "Machine learning prediction of university student dropout," *Mathematics (MDPI)*, vol. 10, no. 18, 2022.
[7] M. Yağcı, "Educational data mining: predict student performance with machine learning," *Smart Learning Environments*, 2022.
[8] D. Bãneres et al., "Early warning system for online dropout learners (LIS enhancements)," *International Journal of Educational Technology in Higher Education*, 2023.
[9] B. Carballo-Mendívil, "Predicting student dropout from day one: XGBoost-based early warning," *Applied Sciences (MDPI)*, 2025.
[10] (Study) "Predicting student dropout rates using supervised ML: insights from Somaliland National Education Survey," 2024.
[11] S. Malik et al., "Advancing educational data mining for enhanced student retention: ensemble and deep models," *Scientific Reports*, 2025.
[12] G. Kumar et al., "Ensemble deep learning network model for dropout prediction," *International Journal of Engineering and Computer Education Studies (IJECES)*, 2023.
[13] X. Wen, "Early prediction of MOOC dropout in self-paced courses: clustering and deep learning approaches," *Interactive Learning Environments*, 2024.
[14] K. Devi, "Predicting student dropouts using Random Forest: a longitudinal study," *International Journal of Educational Development*, 2022.
[15] S. A. Sulak and N. Koklu, "Predicting student dropout using machine learning algorithms," *IMIENS / journal*, 2024.
[16] Z. Chi, "Analysis and prediction of MOOC learners' dropout using machine learning," *Applied Sciences (MDPI)*, 2023.
[17] J. Swacha, "Predicting dropout in programming MOOCs through interaction features," *Electronics (MDPI)*, 2023.
[18] Z. Chi et al., "MOOC dropout prediction — statistical analysis + ML," *Applied Sciences (MDPI)*, 2023.
[19] A. M. Rabelo et al., "A model for predicting dropout of higher education students," *Elsevier / Data in Brief / Comput. Educ.*, 2025.
[20] V. R. Prakash and S. R. Mehta, "Predicting dropout risk in online learning using stacked ensembles and XAI," *IJCA / conference paper*, 2025.
[21] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[22] B. Bouihi, "Prediction of higher education student dropout based on hybrid models," *ETASR*, 2024.
[23] R. (author), "A real-life ML experience for predicting university dropout at different stages," *IEEE Access / IJISAE summary*, 2022.
[24] "Predictive modeling of dropout in MOOCs using machine learning (OULAD),", *IJISAE*, 2024.
[25] M. Rebelo Marcolino et al., "Student dropout prediction through machine learning: a Moodle activity-log CatBoost study," *Scientific Reports (Nature)*, 2025.