Dr. Abdul Majid

Dr. Abdul- Research_Paper_CBD-7.doc

Quick Submit Quick

Submit Presidency

University

Document Details

Submission ID

trn:oid:::1:3399166907

Submission Date

Nov 5, 2025, 11:45 AM GMT+5:30

Download Date

Nov 5, 2025, 11:51 AM GMT+5:30

File Name

Research_Paper_CBD-7.doc

File Size

613.5 KB

6 Pages

3,155 Words

19,171 Characters

0% 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

- 0 AI-generated only 0%
  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

Howshould I interpretTurnitin'sAIwritingpercentageandfalsepositives?
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.

Whatdoes'qualifyingtext'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.

Towards Inclusive Education: A Web-Based Platform for School Dropout Prediction using Random Forest and Gradient Boosting Classifiers

Dr.Abdul Majid
School of Computer Science and Engineering Presidency University Bangalore
dr.majid.wahab@gmail.com

Tanushree R
School of Computer Science and Engineering Presidency University Bangalore
tanushreer222@gmail.com

Kavya J
School of Computer Science and Engineering Presidency University Bangalore
kavyaj770@gmail.com

Kavya S
School of Computer Science and Engineering Presidency University Bangalore
kavyareddy52892@gmail.com

Abstract— School dropout in India presents a rather major obstacle in India for equal and inclusive education. Though this phenomenon is complicated and multi-casual, dropout rates are especially high at the secondary level. And typically the policy making does not really solve this issue. By means of machine learning, this paper offers a proactive web based response to this problem. It entails the creation of a thorough platform with frontend website employing HTML, CSS, and JavaScript as well as strong backend supported by Random Forest and Gradient Boosting classifiers. The reason for choosing Random Forest and Gradient Boosting algorithms were they have high accuracy and capacity for feature important analysis, which helps to diagnose the underlying reasons for the predictions. This website provides a user- friendly interface where users can access and also analyze pre- trained model's results on the dashboard while exploring detailed analytics by different factors such as school, age, area and caste. This paper demonstrates how prediction analysis can be transformed into an actionable tool that helps people understand the reasons why student dropout.

Keywords - School Dropout, Machine Learning, Random Forest, Gradient Boosting, Predictive Analytics, Inclusive Education.

I. INTRODUCTION

In the Indian Education System, a monumental enterprise serving approximately 24.8 crore students have been faced by a persistent challenge in ensuring students complete their education [1]. While the government initiatives have significantly increased the Gross Enrolment Ratio at the primary level with gaps having at secondary (77.4%) and higher secondary (56.2%) levels [2]. School dropouts affect the barrier to the nation's vision of inclusive and equitable education with profound implications for individual human capital and socio-economic development [2]. Statistic has also shown that dropout rates are low i.e., 1.5% at the primary level but increases to 12.6% at the secondary level and also with a separate survey did on 2023-24 reporting an even higher rate of 14.1% [3]. This drastic increase suggests that the dropout behavior is not static. Rather they are complex, multi-casual and rooted in a combination of academics, family background and societal pressures that become more pronounced as students age [4]. This highlights the need for a consistent, data-driven analytical tool that will provide a clearer and more unified picture of thus evolving issue. Additionally traditional methods for addressing the dropout problem have been largely reactive by relying on the post-factor analysis. However the emergence of big data and machine learning offers a powerful opportunity to shift to a proactive, predictive approach [5]. By leveraging predictive models in educational institutions, it becomes easier to identify students who are at risk of dropping out, enabling targeted interventions such as tutoring, counseling or family outreach [6]. This approach transforms data from a passive report into a active, decision-support platform enabling personalized strategies at early stages of attrition [7]. This paper presents a comprehensive research and development project that involves both sociological analysis and technical implementation to create a web-based platform for school dropout analysis. Recent advances in machine learning have enabled the development of predictive models capable of analyzing diverse student attributes to predict dropout risk. While many studies say building such models relatively fewer works to help them to implement practically and produce user-friendly platforms. To address this gap, this paper introduces a web-based dropout analysis system that integrates a Random Forest Classifier and Gradient Boosting Classifier with interactive frontend. The core of this system is the machine learning model, specifically the Random Forest Classifier. These servers as a predictive backend model to analyze and forecast students' dropout risk. The system is designed the functionality of displaying the results of the backend model on the dashboard which provides a detailed analytics page. The primary contribution of this work is the creation of a functional, accessible system that translates complex research findings into an actionable tool, creating a powerful instrument for research, policy-making and direct intervention.

II. LITERATURE REVIEW

A. Statistical Trends and Geographic Disparities:

The analysis of national and state-level data reveals that the school dropout problem is not monolithic. It varies by education level, gender, and geography. While the overall dropout rate for India at the primary level i.e. class 1-5 is 1.5%, this figure has notable gender differences. In 2021-22, the dropout rate for boys at this level was 1.6%, slightly higher than the 1.4% for girls [3]. However, this trend reverses at the upper primary level i.e. class 6-8, where the dropout rate for girls is 3.3% higher than for boys and its 2.7% [3]. The most significant attrition occurs at the secondary level i.e. class 9-10, with a combined rate of 12.6%, where boys have a higher dropout rate having 13% than girls having 12.3% [3]. This trend is also reflected in the 2023-24 data, which reports secondary dropout rates of 14.1% [1]. Beyond national averages, there are regional differences. States like Andhra Pradesh, Assam and Bihar report secondary dropout rates of 16.3%, 20.3%, and 20.5%, respectively, which are substantially higher than the national average. In contrast, states such as Chandigarh have zero dropout rate across all levels, and Delhi has a relatively low secondary dropout rate of 4.8%. Furthermore, a closer look at the data reveals paradoxical gender-based trends in different states. For instance, in Assam, girls in secondary school dropout at a rate of 20.7%, higher than the 19.8% for boys, while in Goa, the dropout rate for boys is 12.1% compared to 5.5% for girls [3]. These variations highlight that the determinants of dropout are highly localized and context-dependent. A single-factor solution would not address such a diverse and complex problem, which needs the model to provide detailed, granular analysis tailored to specific demographic and geographic contexts. Table 1 provides a consolidated view of national dropout rates, highlighting the trends across different educational levels and genders.

TABLE EDUCATIONAL LEVEL AND GENDER (2021-22 & 2023-24)

1: NATIONAL SCHOOL DROPOUT RATES BY

Education Level	Boys (%) (2021-22)	Girls (%) (2021-22)	Total (%) (2021-22)	Total (%) (2023-24)
Primary (1-5)	1.6	1.4	1.5	1.9
Upper Primary (6-8)	2.7	3.3	3.0	5.2
Secondary (9-10)	13.0	12.3	12.6	14.1

B. Socio-Economic and Demographic Determinants:

The reasons for school dropouts are multi-casual including socio-economic and demographic factors playing a vital role. The issue is particularly different among poor and destitute families who are unable to meet the financial demands of schooling or even provide for their basic needs. Poverty is a root cause for many things such as child labor, where children are forced to abandon their education to contribute to family income [8]. Child marriages, particularly for girls, represent another significant contributing factor to dropout, with one study identifying it as a major reason for school dropout among the female population [9]. Study has consistently shown that factors such as caste division, household wealth, and parental education have a direct influence on dropout rates [8]. Particularly, a study using a survival analysis approach found that, the monthly per capita expenditure (MPCE) quintile of a household increases, the risk of a child dropping out of school decreases. High dropout rate trends are disproportionately observed among those from low- income households and specific socio-demographic groups, including Muslim families and low-caste communities. The finding that the risk of school dropout is 50% less in private schools compared to government institutions further underscores the role of socio-economic factors and access to quality education [10]. This evidence demonstrates a crucial chain: economic poverty is not merely a barrier to paying school fees, it fundamentally alters the value proposition of education for families, turning children into economic assets and leading to outcomes like child labor or early marriage. An effective predictive model must therefore be able to quantify and weigh the influence of these complex, interlinked factors to provide an accurate risk assessment.

C. Academic and Institutional Factors:

Beyond the family and societal pressures, institutional and academic factors within the school environment are also major contributors to student dropout. A significant portion of students drop out due to a lack of interest in education or an inability to cope with their studies [10]. These reasons, while appearing to be personal failings, are often symptoms of deeper systemic issues. Study indicates that "no interest in education" and "unable to cope up with studies" are directly linked to the poor quality of education, a lack of engaging content, and inadequate infrastructure [9]. The observation that the risk of school attrition is half as likely in private schools compared to government institutions highlights the "sufficient infrastructure, efficient teacher- student ratio, and quality of teaching" in government schools as a major problem [10].

Poor academic performance, including failing grades and low test scores, is a strong predictor of dropout risk, as struggling students may become disengaged and lose motivation. Irregular attendance and a lack of participation in school activities are also indicators of a high time risk [6]. The paper also suggests that the phenomenon of dropout is

rarely due to a single cause but is rather the combination of multiple factors [4]. Therefore, a collective approach that considers not just socio-economic data but also academic performance and institutional context is essential for building a robust predictive model.

III. METHODOLOGY

A. Dataset Description:

More than 200 students' data with more than 10 features were collected for training the model. The features shown in the Table 2 were analyzed and collected to satisfy the problem statement and to get higher accuracy of the model.

TABLE 2: KEY DETERMINANTS OF SCHOOL DROPOUT AND CORESSPONDING MODEL FEATURES

Dropout Determinant	Corresponding Model Features
Gender Disparity	Gender
Academics	Standard, Attendence, Previous_Year_Performance
Institutional Factors	School, Distance_From_School
Family Factors	Parental_Education, Family_Income_Level, Parent_Type
Regional/Demographic Factors	Area, Caste
Target Variable	Dropout(Yes/No)

B. Data Pre-processing:

The dataset is manually cleaned by scanning all the records of the dataset and removing all the unnecessary records, inaccurate values and null values. Then this cleaned dataset is given to the model with 80:20 ratios i.e. 80% of the dataset for training and 20% of the dataset for testing. To ensure the dataset is suitable for machine learning, several preprocessing steps were applied before model training. The dataset consisted of 300 students' record with 15 attributes covering academics, institutional, family and demographic factors information. The preprocessing pipeline consisted of handling missing values, categorical encoding, feature engineering and feature scaling.

1) Handling Missing Values: Parental Education contained missing values. These were filled using mode imputation to retain categorical consistency. And for other features, they had minimal missing entries and were handled through manual removal.

2) Encoding Categorical Features: Since Random Forest requires numerical inputs, categorical variables are encoded using Label Encoding. The following features were transformed: Gender (Male=1, Female=0), Caste, School, Area, Parental Education, Family Income, Scholarship, Special Car and Parent Type.

This will ensure that each categorical value is represented with an integer while maintaining class identity.

C. Feature Engineering:

To enhance predictive power, domain-specific engineered features were introduced:

1) Age-Standard Gap: Age-(Standard+5) this captures whether a student is lagging academically relative to expected age
2) Attendance-Score Interaction: Product of attendance and Previous Score highlights the joint effect of effort and performance
3) Financial Stress Indicator: A binary feature computed as 1 if (Family Income = Low and No Scholarship), otherwise 0.

These engineered features were motivated by educational research suggesting that academic lag, poor attendance, and financial hardship strongly correlate with dropout risk.

D. FeatureScaling:

Although Random Forest is not sensitive to feature scaling, it was applied to improve training stability and visualization consistency. All numerical features were standardized using StandardScaler (mean=0, variance=1):

- Numerical features scaled: Age, Attendance, Previous Score, Distance, Age-Standard Gape, and Attendance-Score Interaction.
- Categorical features remained label-encoded and were not scaled

E. Class Balancing:

After applying rule-based dropout labeling, the dataset was found to be imbalanced (dropout case ~79%, non-dropout ~29%). To mitigate bias:

- The minority class (non-dropouts) was up sampled using random resampling until class distribution reached approximately 2:1 (dropouts: non-dropouts).
- This ensures the classifier received balanced exposure to both classes during training

F. Model Training:

The prediction of school dropout is formulated as a binary classification problem, where the target variable indicated whether a student is likely to drop out (1) or continue in school (0). After preprocessing feature engineering, the dataset is split using 80:20 ratio to ensure sufficient data for model learning and unbiased evaluation.

1) Choice of Algorithm: The Random Forest Classifier is selected as the primary predictive model due to its robustness in handling mixed data types (categorical and numerical), resistance to over fitting on small datasets, and ability to provide interpretable feature importance scores.

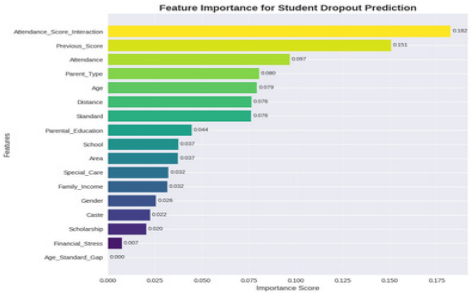
For comparison, a Gradient Boosting Classifier was also trained to evaluate performance differences.

2) Random Forest Configuration: To optimize the Random Forest model, a GridSearchCV-based hyperparameter tuning procedure was employed with 5-fold cross-validation. The following hyperparameters were tuned- number of estimators: [100,200], maximum tree depth: [10,15,None], minimum samples per split: [2,5], minimum samples per leaf: [1,2] Gradient Boosting
3) Configuration: The Gradient Boosting model was implemented with default parameters, using a learning rate of 0.1 and shallow decision trees as base learners. While Gradient Boosting is powerful for structured data, its performance was comparatively weaker on the current dataset due to limited sample size and risk of overfitting. Training Procedure: Both models were trained on the preprocessed training set. Cross-
4) validation is performed to access generalizability. The trained models were then evaluated on the held-out test set to measure real-world prediction capability.

IV. DISCUSSION

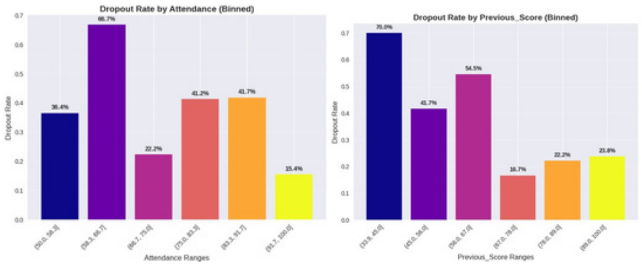
A. FeatureImportanceAnalysis:

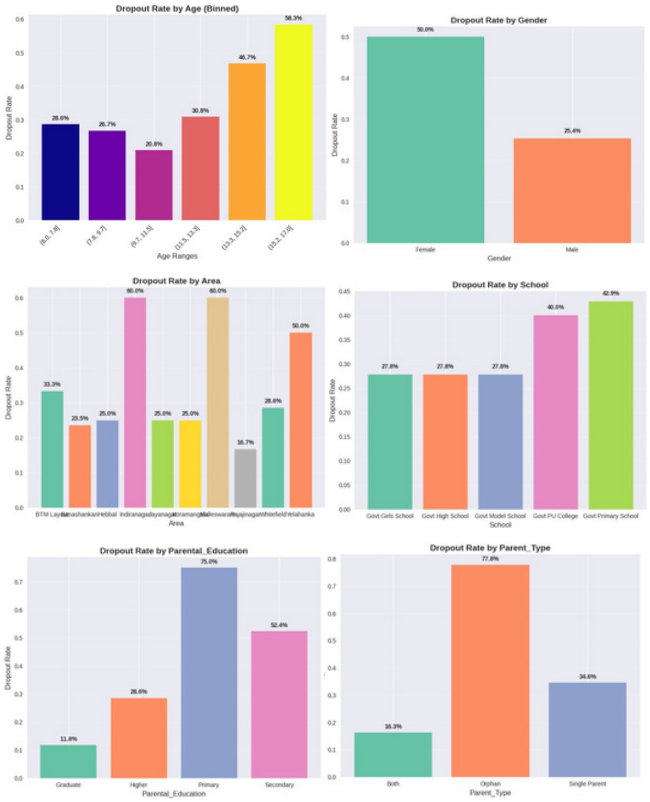
The Random Forest model revealed that attendance, previous academic score, parental education, and family income were the strongest predictors of school dropout. Attendance and Previous Score contributed significantly to model decision making confirming that consistent classroom participation and performance are the keys determinant of retention. Similarly, students with low prior scores and lower parental education levels were at a higher risk of leaving school, emphasizing the joint influence of academic and scioeconomic factors.



B. DropoutRatebyDemographicandAcademicFeatures:

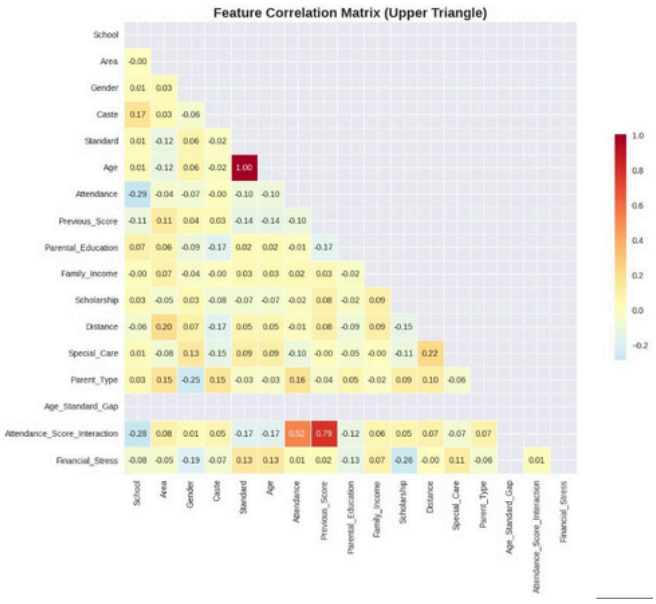
The visualization below shows dropout rates across different features like Age, Parent Type, Attendance, Previous Score, Parent Education, School, Area and Gender.





C. CorrelationAnalysis:

The correlation matrix between input features and dropout confirmed several intuitive relationships. Attendance was negatively correlated with dropout, while age–standard gap and distance from school were positively correlated. Socioeconomic indicators such as family income and parental education also exhibited strong associations with dropout status.



D. ModelPerformanceEvaluation:

The trained models were evaluated using cross-validation and test set performance metrics. The optimal parameters for the Random Forest classifier obtained through GridSearchCV were

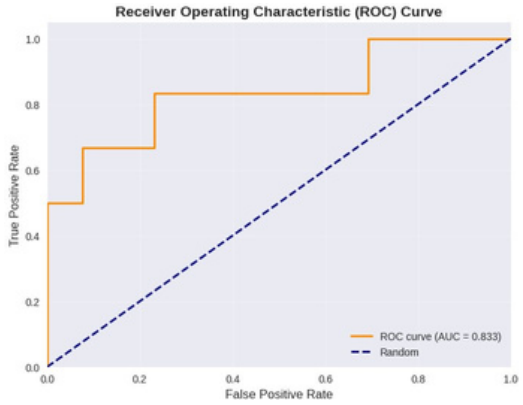
```
Best Random Forest parameters: {'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 200}
```

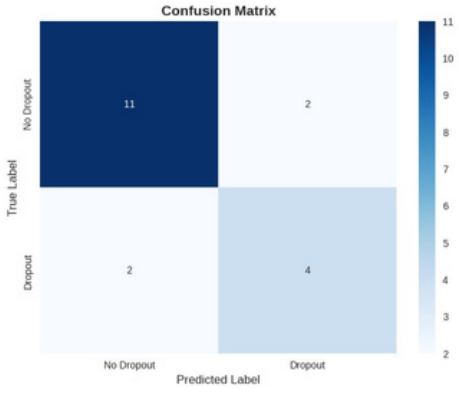
The performance comparison of Random Forest Gradient and Boosting is summarized below:

```
Random Forest:
Cross-validation AUC: 0.8280 (+/- 0.1148)
Test Accuracy: 0.7895
Test AUC: 0.8333

Gradient Boosting:
Cross-validation AUC: 0.8160 (+/- 0.1114)
Test Accuracy: 0.6316
Test AUC: 0.7821
```

The Random Forest model consistently outperformed Gradient Boosting across all metrics. The ROC curve showed an AUC of 0.83 for Random Forest, indicating a strong discriminatory ability between dropout and non- dropout classes. The confusion matrix further confirmed that the model was effective in correctly classifying both categories, though slight misclassifications persisted due to overlapping feature distributions.





E. Final Model Summary:

1) Academic engagement (attendance and previous scores) and socioeconomic conditions are the most influential dropout factors. Dropout risk is not uniform—it varies
2) significantly across schools, regions, and social groups. The Random Forest model achieved reliable performance
3) (Accuracy = 78.95%, AUC = 0.83) and is well suited for deployment in the proposed website.
4) Visual analytics provide actionable insights, enabling educators and policymakers to design targeted interventions for at-risk groups.

```
================================================
FINAL MODEL SUMMARY
================================================
Final Model: Random Forest
Test Accuracy: 0.7895
Test AUC: 0.8333

Classification Report:
              precision    recall  f1-score   support

  No Dropout       0.85      0.85      0.85        13
     Dropout       0.67      0.67      0.67         6

    accuracy                           0.79        19
   macro avg       0.76      0.76      0.76        19
weighted avg       0.79      0.79      0.79        19


Final Dropout Distribution:
Dropout
0    0.666667
1    0.333333
Name: proportion, dtype: float64

Top 5 Most Important Features:
Age: 0.0793
Parent_Type: 0.0803
Attendance: 0.0967
Previous_Score: 0.1507
Attendance_Score_Interaction: 0.1823
```

V. CONCLUSION

This paper presented a web-based system for school dropout prediction and analysis, integrating a Random Forest machine learning model with an interactive front-end. The methodology combined rule based labelling, feature engineering, and resampling techniques to prepare the dataset, followed by training and evaluation of Random Forest and Gradient Boosting classifiers. Experimental results demonstrated that the Random Forest model achieved superior overall performance, validating its selection as the core predictive engine. The model attained a Test Accuracy of 78.95% and a Cross-validation AUC of 0.8333, confirming its reliability and strong ability to generalize across the student population. Crucially, the functional utility of the platform was verified by the Feature Importance analysis, which identified the academic and behavioural factors led by Attendance Score Interaction 0.182 and Previous Score 0.151 as the primary drivers of dropout risk. The system's architecture, including its three-tier structure and detailed data flow, was designed to translate complex analytical findings into a user-friendly, actionable tool. The ultimate finding is that a data-driven, predictive approach can transform the problem of school dropouts from a static statistic into an active, manageable challenge, enabling targeted and effective interventions.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to their faculty mentors and academic advisors for their valuable guidance throughout the course of this project. We also acknowledge the support provided by our institution in terms of resources and infrastructure that made this work possible. Finally, we appreciate the availability of open- source libraries and tools such as Scikit-learn, Pandas, NumPy, and Matplotlib, which were instrumental in the implementation of the machine learning models and data visualization components of this work.

REFERENCES

[1] Chaudhary, Mr Hukam Chand. "Major Challenges of Education System in India and Efforts for Solutions." (2023).

[2] Sajjad, Haroon, et al. "Socio-economic determinants of primary school dropout: Evidence from south east Delhi, India." European Journal of Social Sciences 30.3 (2012): 391-399."

[3] Mehta, A. "Dropout rates in schools in India: an analysis of UDISE+ 2021-22 data." Education for All, available at: https://educationforallinindia. com/dropout-rates-in-schools-in-india/(accessed 4 August 2024) (2022).

[4] Song, Zihan, et al. "All-year dropout prediction modeling and analysis for university students." Applied Sciences 13.2 (2023): 1143.

[5] SULAK, Süleyman Alpaslan, and Nigmet KOKLU. "Predicting student dropout using machine learning algorithms." Intelligent Methods In Engineering Sciences 3.3 (2024): 91-98.

[6] Alameri, Fatma. Predicting Student Dropout Risk using Machine Learning. Rochester Institute of Technology, 2025.

[7] Behr, Andreas, et al. "Early prediction of university dropouts–a random forest approach." Jahrbücher für Nationalökonomie und Statistik 240.6 (2020): 743-789.

[8] Sajjad, Haroon, et al. "Socio-economic determinants of primary school dropout: Evidence from south east Delhi, India." European Journal of Social Sciences 30.3 (2012): 391-399.

[9] Kurian, Aju, et al. "School Dropouts: Reasons and Prospective Solutions-Teachers' Perspective." International Journal of Creative Research Thoughts (IJCRT) 11.3 (2023): 2-3.

[10] Garg, Mausam Kumar, Poulomi Chowdhury, and Illias Sheikh. "Determinants of school dropouts in India: a study through survival analysis approach." Journal of Social and Economic Development 26.1 (2024): 26-48.

[11] Behr, Andreas, et al. "Early prediction of university dropouts–a random forest approach." Jahrbücher für Nationalökonomie und Statistik 240.6 (2020): 743-789.

[12] Andrade-Girón, Daniel, et al. "Predicting Student Dropout based on Machine Learning and Deep Learning: A Systematic Review." EAI Endorsed Transactions on Scalable Information Systems 10.5 (2023).