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

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

Dr. Abdul Majid
School of Computer Science and Engineering
Presidency University
Bangalore
dr.majid.wahab@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 like a frontend website employing Streamlit Python Framework 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

The Indian Education System has encountered an uphill task of making sure that students finish their education in a monumental enterprise providing its services to about 24.8 crore students [1]. Though the government schemes have made big improvements at the primary level in raising the Gross Enrolment Ratio with gaps at secondary (77.4%) and higher secondary (56.2%) levels [2]. The school dropouts have far-reaching consequences on both the personal human capital and the socio-economic development due to the barrier in the vision of an inclusive and equitable education in the country [2]. Statistic has also indicated that the dropout rates are low i.e 1.5% at the primary level and higher to 12.6 percent at the secondary level and also in a separate survey that was done on 2023-24 indicated a higher rate of 14.1 percent [3]. Such a drastic increase implies that the dropout behavior cannot be fixed. Instead, they are complicated, multi-casual and based on an amalgamation of both academics, family background and societal pressures, which increase with age of the students [4]. This is where the necessity of an analytical tool which will be working in a consistent and data-driven manner emerges, and will give a clearer and more coherent view of thus developing issue. As well, the traditional means of correcting the dropout problem have been very reactive and based on the post-factor analysis. But with the advent of big data and machine learning, there is a great opportunity to change to an active, predictive model [5]. Through the use of predictive models in education institutions, there is increased ease in detecting students who are likely to drop out so that interventions like tutoring and counseling of students or even the family can be done [6]. This strategy makes data not a passive report but a active and decision support platform to allow individual strategies at the initial levels of attrition [7]. The paper is a detailed research and development project, which incorporates sociological analysis as well as the technical application in developing a web based platform onto which school dropouts will be analyzed. The latest developments in machine learning have allowed the creation of predictive models that can analyze a number of various characteristics of students to predict their risk of dropping out. On the one hand, various researches claim that the creation of such models is comparatively less numerous because they need it to practice it in the field and create easy-to-use platforms. In order to fill this gap, the current paper presents a web-based dropout analysis system comprising of an interactive frontend, a Random Forest Classifier and Gradient Boosting Classifier. The machine learning model is the fundament of this system, and it is the Random Forest Classifier. These servers are a predictive backend model to examine and predict the likelihood of dropouts among students. The system is planned the functionality of showing the results of the backend model on the dashboard that offers a more detailed analytics page. The main value of this work is the establishment of a practical, easy to use system of translating the complex research results into the feasible tool creating a potent research tool, policy-making and a direct intervention.

II. Literature Review

A. Statistical Trends and Geographic Disparities:

The comparison of the national and state-level data shows that the issue of school dropout is not unitary. It differs according to the level of education, gender and geography. Although the general rate of drop out in India at the primary level i.e. class 1-5 is 1.5 percent, there is a significant difference between the genders. In 2021-22, the level of dropouts among boys was 1.6% (a bit more than 1.4% among girls) [3]. This however turns the other way around at the upper primary level i.e. class 6-8, the dropout rate is 3.3 percent more in girls than in boys and 2.7 percent [3]. At the secondary level i.e. class 9-10, the highest rate of attrition is recorded 12.6 with boys registering 13% higher than girls registering 12.3% [3]. The same tendency is observed in the 2023-24 data where the secondary dropout rates are reported to be 14.1% [1]. Outside the national means, there are regional variations. There are states such as Andhra Pradesh, Assam and Bihar in which the secondary dropout rates include 16.3, 20.3, and 20.5 respectively, which is significantly higher than the national average. Conversely, there are state like Chandigarh with zero dropout rates at all levels, and Delhi with the relatively the lowest dropout rate of 4.8 per cent at the secondary level. More so, a more detailed analysis of the data will demonstrate counter-intuitive trends in gender-based trends across states. As an example, the dropout rate of girls in secondary school is 20.7 compared to the 19.8 percent in boys in Assam, whereas, in Goa, the dropout rate of boys is 12.1 as compared to 5.5 percent in girls [3]. Such differences underscore the complexity of the determinants of dropout as being too local and specific to the context. Such a complex and multi-factored problem could not be solved by a single-factor solution, but rather requires that the model will offer detailed and granular analysis based on particular demographic and geographic situations. Table 1 gives a condensed overview of the national dropout rates, which points out the trends of various levels of education and different gender.

TABLE 1: NATIONAL SCHOOL DROPOUT RATES BY EDUCATIONAL LEVEL AND GENDER (2021-22 & 2023-24)

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:

School dropouts are multi-casual with socio-economic and demographic factors being critical factors. This is especially different with the poor and destitute families which cannot afford the financial cost of schooling or even sustaining their basic needs. Many things including child labor are a result of poverty where children are compelled to leave their education in order to help the family earn money [8]. Girl child marriage is also another important factor that contributes to dropout with one study showing child marriage to be one of the greatest causes of school dropout amongst the female population [9]. Research has always indicated that caste division, household wealth and parental education among other factors have a direct effect on dropout rates [8]. Especially, when a survival analysis method is used, the risk of a child dropping out of school declines as the monthly per capita expenditure (MPCE) quintile of a household rises. The trends of high dropout rates are disproportionately found in the families of low income level and certain socio-demographic groups such as Muslim families and low-caste communities. The fact that a school dropout rate is half in private school than in government institutions also contributes to the importance of socio-economic factors and access to high-quality education [10]. This fact shows that there is a critical chain: economic poverty does not simply prevent the ability to pay school fees; it essentially changes the value proposition of education of the family, making children their economic resources, which result in the appearance of such phenomena as child labor or early marriage. A good predictive model should then be in a position to measure and balance the effect of these multifaceted interconnected factors to give a correct risk assessment.

C. Academic and Institutional Factors:

Other than the family and societal pressure, institutional and academic pressures in school setting also contribute significantly to school dropout. A substantial number of learners dropout because they are not interested in education or fail to manage with their studies [10]. These causes, though seeming to be individual weaknesses, are mostly manifestations of some underlying systemic problems. Research has shown that no interest in studying and unable to keep up with the studies are directly connected with the level of poor quality of education, lack of exciting content and the lack of proper infrastructure [9]. This observation that the threat of school dropout is one half in the case of the private schools and twice in the case of the government institutions points to the adequate infrastructure, the teacher-student relationship and the quality of teaching in the government schools as a key issue [10]. Low academic achievement, such as poor grades and low test scores, is a high predictor of dropout potential as low-achieving students could become bored and unmotivated. Abnormal attendance and the absence of school activities are also signs of a high time risk [6]. The paper also hypothesizes that the phenomenon of dropout is hardly caused by a single concept but is instead the convergence of numerous components [4]. Thus, the collective strategy, which will take into account not only the socio-economic information but also academic achievement and institutional background, is the key to the development of the robust predictive model.

III. METHODOLOGY

A. Dataset Description:

The data of over 200 students having over 10 characteristics were gathered to train the model. The features presented in the Table 2 were processed and gathered in order to meet the problem statement as well as achieve greater 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 cleaned manually through scanning all the records of the dataset and deleting all the unwanted records involved, incorrect values, and null values. This purified dataset is then fed to a model with 80:20 ratios i.e. 80 percent of dataset being fed in the training and 20 percent of the dataset in the testing. In order to make the dataset fit in the machine learning, various preprocessing measures were undertaken prior to the model training. The dataset comprised of 300 records of students having 15 attributes comprising of the information of academics, institutional, family and demographic factors. The preprocessing pipeline was composed of dealing with missing data, categorical encoding, feature engineering and feature scaling.

1) Preprocessing Missing Values: Parental_Education field had missing values. Mode imputation was used to fill these in order to maintain categorical consistency. And on other features, they had few entries to remove and were done manually.
2) Encoding Features: Random forest feeds on numerical data so, categorical values are encoded with Label Encoding. These characteristics were changed as follows: Gender (Male=1, Female=0), Caste, School, Area, Parent Education, Family Income, Scholarship, Special car and parent Type.

This will make sure that the categorical values are represented by integer but not by changing classes.

C. Feature Engineering:

To raise the predictive power domain specific engineered features were added:

1) Age-Standard Gap: Age- (Standard +5) this is used to measure how a student is performing relative to his age.
2) Attendance-Score Interaction: The product of attendance and previous score points identifies the combined influence of the performance of endeavor and output.
3) Financial Stress Indicator: Binomial trait, based on (Family Income = Low and No) Scholarship), 1 otherwise.

The basis of these artificial capabilities was the learner studies which showed that academic lagging, poor attendance and economic disasters are directly linked to the threat of dropping-out.

D. Feature Scaling:

Random Forest is not sensitive to the scaling properties, but it was applied in order to increase the stability of training and visualization stability. Before any numerical characteristics were used, they were normalized StandardScaler (mean=0, variance=1):

- Numerical features scaled: Age, Attendance, Previous score, Distance, Age-Standard Gape and Interaction of age and Standard Score on attendance.
- Categorical features were retained and did not undergo transformation.

E. *Class Balancing:*

The dataset was subsequently also passed through rule-based dropout labeling which was followed by the determination of The dataset to be unbalanced (dropout case slightly more than-~79%, non- dropout slightly less than-~29%). To mitigate bias:

- The A minority class (non-drop outs) was up sampled up to the point of class distribution random resampling being approximately 2:1 (dropouts: non dropouts).
- It will ensure that both classes were presented equally with the classifier being fed during training.

F. *Model Training:*

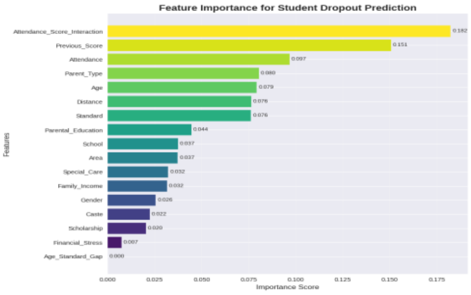
The School dropout is a binary prediction formulation problem, whereby the target variable was indicated by the high likelihood of dropping out (1) or not school (0). After feature engineering (preprocessing feature engineering) the 80: 20 ratio are employed to sub-classify the dataset and create 2 equal parts in order to have sufficient data learning models and objective evaluation.

1) Selection of Algorithm: The main predictive model chosen is the Random Forest Classifier because it is strong in taking mixed type of data (both categorical and numerical), it will not overfit well on small datasets, and it will offer interpretable scores of feature importance. To compare, a Gradient Boosting Classifier was also trained to compare the difference in performance.
2) Random Forest Settings: To improve the Random Forest model, a cross-validation close to 5-fold hyperparameter tuning process using a GridSearchCV procedure was used. The hyperparameters that were tuned are as follows: number of estimators: 100,200, maximum tree depth: 10,15, None, minimum samples per split: 2,5, minimum samples per leaf: 1,2.
3) Gradient Boosting Settings: The default parameters of the Gradient Boosting model were used in the form of learning rate as 0.1 and shallow decision trees as base learners. Whereas Gradient Boosting can be effective with structured data, its learning capabilities were smaller with the current data in terms of sample size and potential overfitting.
4) Training Procedure: The two models were trained using the training set which was preprocessed. Cross-validation is done in order to reach generalizability. To assess the capability of prediction in the real world, the trained models were tested on the held out test set.

IV. DISCUSSION

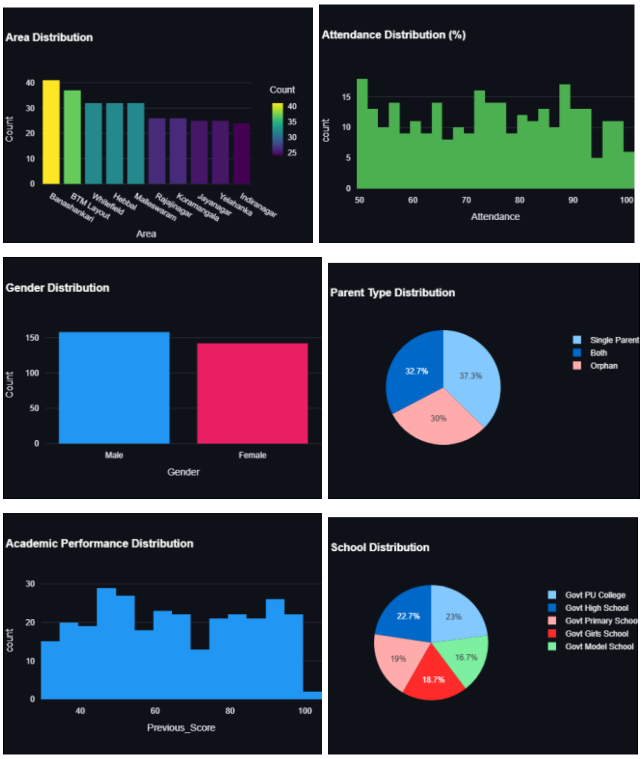
A. *Feature Importance Analysis:*

The model of Random Forests showed attendance, previous academic score, education level of parents, and family income as the most predictive variables of school dropout. The model decision-making included an important part of attendance and Previous Score that proves the fact that classroom attendance and previous performance rank as the most significant determinant of retention. In the same way, the students, who scored lowly and had low education level were also more likely to drop out, which indicates that both socioeconomic and academic causes may have an interactive effect.



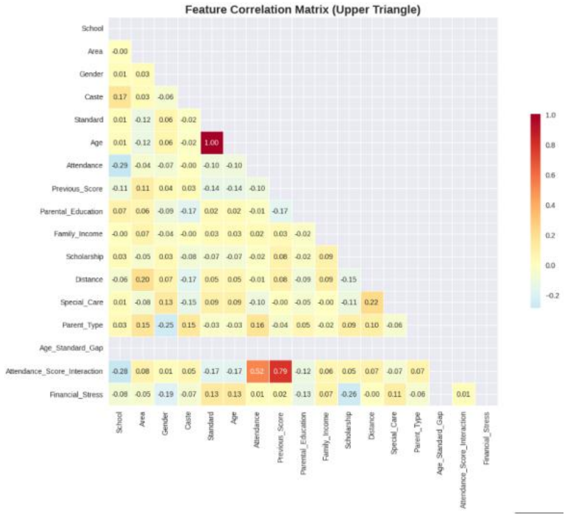
B. *Dropout Rate by Demographic and Academic Features:*

The below figure indicates dropout rates of various attributes such as Parent Type, attendance, previous score, school, area and gender.



C. *Correlation Analysis:*

The intuitive relationships were proven by the correlation coefficient of the dropout, and the input features. One was the negative correlation between the attendance and dropout as well as the positive correlation between the age standard gap and distance to school. The other social economic variable that was highly illustrative and connected to the dropout status was the family income and parental education.

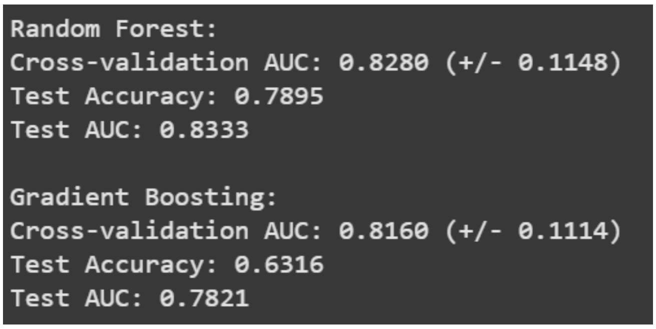


D. *Model Performance Evaluation:*

The cross-validation and the test set were considered to be the performance measures of the trained models. The best parameters of the Random Forest classifier that were obtained in the process of the GridSearchCV were

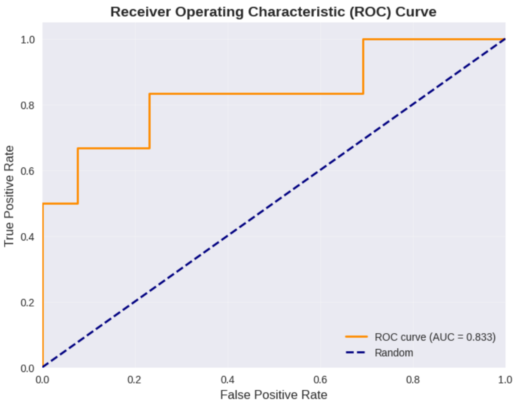
Best Random Forest parameters: {'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 200}

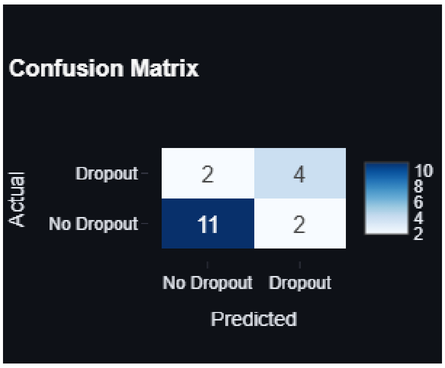
The results of the Random Forest and the Gradient Boosting can be summarized as the following way:



```
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
```

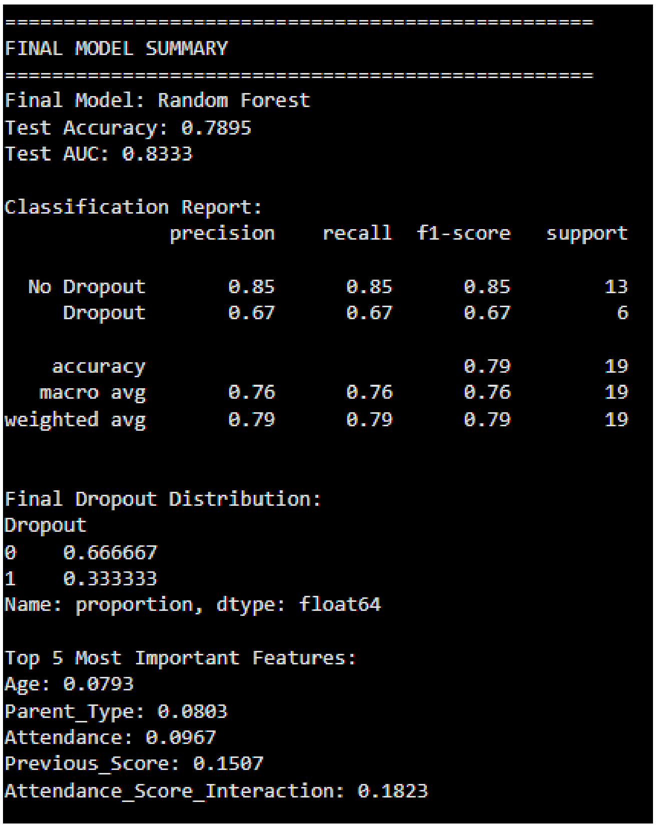
Random Forest model performed better in Gradient Boosting in both measures. The ROC graph AUC of Random Forest was 0.83, which is a strong predictive variable of the dropout and non-dropout classes. The confusion matrix also served to prove that the model was effective in the sufficient classification of the two classes though there was also a small misclassification since the perceived feature distributions were somewhat overlapping.





E. *Final Model Summary:*

1) The most important factors that may affect dropouts are academic participation (attendance and past grades) and socioeconomic status.
2) The drop out threat is not homogeneous: It differs considerably across schools, regions and social classes.
3) As it has been noted, the Random Forest was very good (Accuracy = 78.95% AUC = 0.83), and would be a good model to use in the proposed site.
4) Visual analytics unleash the proactive information which can be utilized by educators and policy makers to receive the specific measures that must be undertaken in order to be utilized against the risky groups.



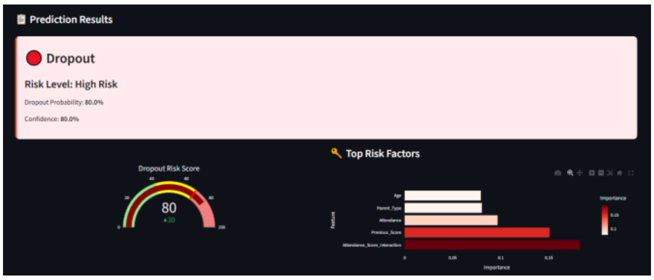
```
==================================================
FINAL MODEL SUMMARY
==================================================
Final Model: Random Forest
Test Accuracy: 0.7895
Test AUC: 0.8333

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
```

F. *Webpage Implementation and Deployment using Streamlit Framework:*

The finished system is published as one integrated web application with the use of the Streamlit Python framework instead of the use of classical three-layers system with separate HTML/CSS/JS components. The mode of implementation will be the application layer combined with the presentation layer with such an outcome that the ML core will be full of the interactive interface so that implementation will be simple and development time will be reduced. The architecture is grounded on the following fundamental components of operation:

1) Model Persistence and Loading: The trained model of the Random Forest and the pre-processing assets (Scaler and Label Encoders) are persisted by using joblib library. The class that is set-up with a secure loading of all the elements that are efficient is the backend class DropoutPredictor.

2) Real-Time Prediction Service: The prediction algorithm of the DropoutPredictor class will be the one that will be capturing the essence of the prediction logic. The specified approach performs the preprocessing of the prediction made by the random forest, and, most to the point, the calculation of the Top 5 Most Important Features in the specified input of a certain student serving as a diagnosis result along with the prediction (Dropout/No Dropout) and the corresponding risk probability.



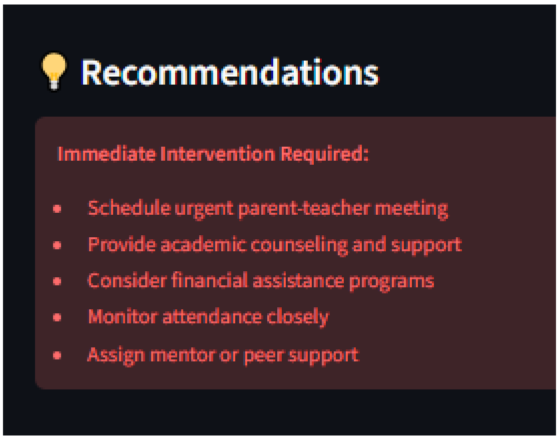
Prediction Results

Dropout

Risk Level: High Risk

Dropout Probability: 80.0%

Confidence: 80.0%



Recommendations

Immediate Intervention Required:

- Schedule urgent parent-teacher meeting
- Provide academic counseling and support
- Consider financial assistance programs
- Monitor attendance closely
- Assign mentor or peer support

3) User Interface and Analytics: The workload of Streamlit application (app.py) is connected with multi-page navigation (Dashboard, Predict Risk, Analytics) and the application of Plotly to develop all the interactive representations of data. The general model performance metrics (Accuracy, AUC) and demographic distributions would be revealed in the Dashboard, whereas the raw feature input and their present prediction and diagnostic outcomes would be shown simultaneously in the Predict Risk page based on the Streamlit widgets.

V. CONCLUSION

The paper has described a Web-based School dropout prediction using Streamlit Python Framework with analysis system, which combines a Random Forest machine learning algorithm with an interactive front-end Web site. The analysis flow comprised of rule based labelling, feature engineering, and resampling methods to prepare the input, training, and test of the Random Forest and Gradient Boosting models. The experimental findings proved that the overall performance of the Random Forest model was better, which allowed selecting the model as the main predictive engine. The model was able to achieve a Test Accuracy of 78.95% and Cross-validation AUC of 0.8333, which supported the reliability and high generalizability of the model on a student population. To confirm the functional utility of the platform, most importantly, the Feature Importance analysis identified the academic and behavioural factors with attending score interaction 0.182 and previous score 0.151 as the main causes of dropout risk. The system also had a structure with a three tier framework and elaborate data flow, which was meant to ensure that the complex findings of the analytics were converted into a user friendly and an actionable mechanism. The final conclusion is that a predictive, data-driven strategy can help to change the issue of school dropouts not a fixed number but a dynamic process, allowing administering specific and effective interventions.

ACKNOWLEDGEMENT

The authors would also like to convey their deepest appreciation of their faculty mentors and academic advisors who have been of great help to them during the period of this project. We also recognize the assistance that our institution gave with regard to resources and infrastructure that enabled us to do this. Lastly, we are grateful that there exist open-source libraries and tools like Scikit-learn, Pandas, NumPy, and Matplotlib that played a significant role in the implementation of the machine learning model and data visualization elements 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).