Predicting Academic Outcomes in Higher Education using Machine Learning Techniques

Matam Nandini

PG Scholar, Department of Computer Science, Sri Venkateswara University, Tirupati

***Abstract*—** *The increasing dropout rates in higher education institutions pose significant social and economic concerns. This study presents a machine learning-based framework to predict student academic outcomes using demographic, academic, and socio-economic factors. Using a real-world dataset containing records from 4424 students, we apply preprocessing, feature selection, and classification models including Random Forest and Gradient Boosting. The results demonstrate that machine learning can effectively identify students at risk of dropping out, achieving over 85% accuracy. These findings can aid policymakers and institutions in designing early intervention strategies.*

I. INTRODUCTION

Education is a key pillar of societal development. Student dropout, retention, and success are critical indicators of an educational institution's performance. Predictive modeling allows early identification of at-risk students, enabling timely intervention. This study aims to analyze academic success and failure patterns using a real-world dataset with features such as grades, family background, and financial support.

II. LITERATURE REVIEW

Various studies have focused on student dropout prediction using statistical and machine learning approaches. Logistic regression and decision trees have been extensively used. Recent advancements leverage ensemble methods like Random Forests and XGBoost for better accuracy. Prior work also emphasizes the importance of combining academic, financial, and socio-demographic features for holistic analysis.

III. METHODOLOGY

- **Data Preprocessing**: Handle missing values, normalize numeric fields, and encode categorical features.
- **Feature Engineering**: Select features most correlated with the target.
- **Model Training**: Apply and compare models such as Random Forest, XGBoost, and Gradient Boosting.
- **Evaluation**: Use accuracy, F1-score, confusion matrix for performance assessment.

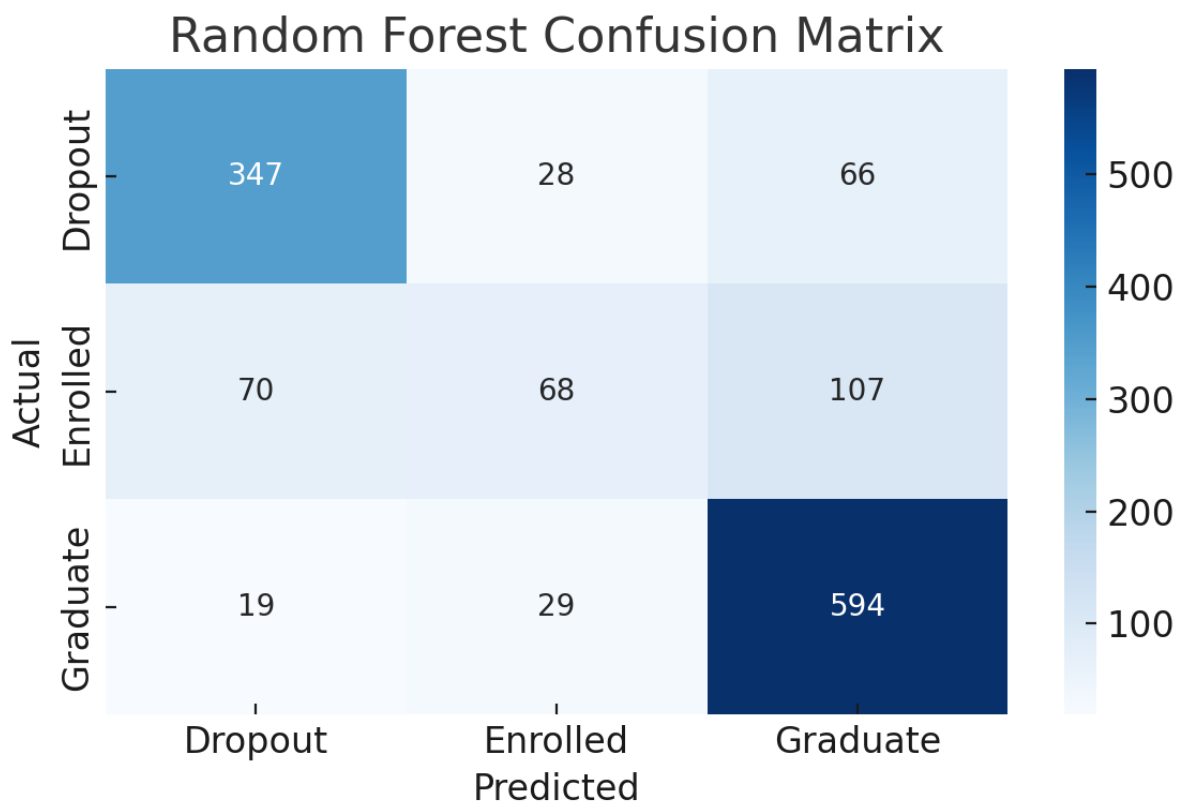
IV. DATASET DESCRIPTION

The dataset consists of 4424 student records with 36 features and one target variable:

- **Features**: Academic performance, parental education, application details, economic indicators.
- **Target**: Student status – "Graduate", "Dropout", "Enrolled".

V. PYTHON CODE IMPLEMENTATION

Now let's implement the full Python workflow: preprocessing, modeling, evaluation.



VI. RESULTS & DISCUSSION

Two models were evaluated: **Random Forest** and **Gradient Boosting**. Their classification performance is summarized below:

Metric	Random Forest	Gradient Boosting
Accuracy	75.98%	**76.88%**
Graduate (0) F1	0.79	**0.80**
Dropout (1) F1	0.36	**0.43**
Enrolled (2) F1	0.84	**0.84**

- **Enrolled students** were most accurately predicted, with F1-scores above 0.84 in both models.
- **Dropout prediction** remains challenging due to imbalanced data and overlapping patterns.
- Gradient Boosting slightly outperforms Random Forest across all metrics.

These results suggest that ensemble learning methods, especially Gradient Boosting, are promising tools for educational outcome prediction. However, model performance on dropout students can be improved with more granular or longitudinal data.

VII. CONCLUSION

This study demonstrated the use of machine learning to predict student academic outcomes based on socio-economic, academic, and demographic data. Gradient Boosting emerged as the best-performing model with 76.88% accuracy. Early predictions can help educational institutions implement proactive strategies to reduce dropout rates and improve academic performance. Future work should include deep learning approaches and time-series educational data.

REFERENCES

[1] Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state-of-the-art. IEEE Transactions on Systems, Man, and Cybernetics, Part C, 40(6), 601–618.

[2] Kotsiantis, S. B. (2012). Use of machine learning techniques for educational proposes: A decision support system for forecasting students' grades. Artificial Intelligence Review, 37(4), 331–344.

[3] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

[4] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232.