Leveraging Random Forest for Predicting Student Academic Trajectory

Mingyuan Li 2145618

Introduction - Student academic trajectory refers to the outcome of students' educational journey, like graduation or dropout. It may be influenced by various factors, such as family background, enrollment history, and academic results. Predicting academic trajectories based on student background is meaningful as it can provide insights into optimizing educational resources. However, due to the significant individual variability among students, relying on individual-level data often causes low predictive accuracy. Thus, incorporating Big Data becomes essential.

Leveraging large datasets, combined with appropriate preprocessing and machine learning techniques, Big Data enables highly accurate predictions of student academic trajectories. In this study, a Random Forest (RF) model was trained on a dataset consisting of 3,539 samples and 36 features to perform a multi-class classification task (dropout, enrolled, or graduate). The final model achieved a competitive score of 0.81674 on the Kaggle competition.

Methodology **- Data Preprocessing.** By utilizing basic data interpretation to examine the data distribution, it revealed that there were no missing values or apparent outliers. The target feature was identified as categorical labels, which were encoded into numerical values using a label encoder. Furthermore, correlation analyses output multiple pairplots and a correlation matrix (Figure 1), indicating that most features exhibit dispersed and uneven distributions. Some features showed notable correlations, like the socioeconomic. Notably, for the target variable, while most features showed low correlations, a few still exhibited weak correlations in the range of 0.05–0.12.

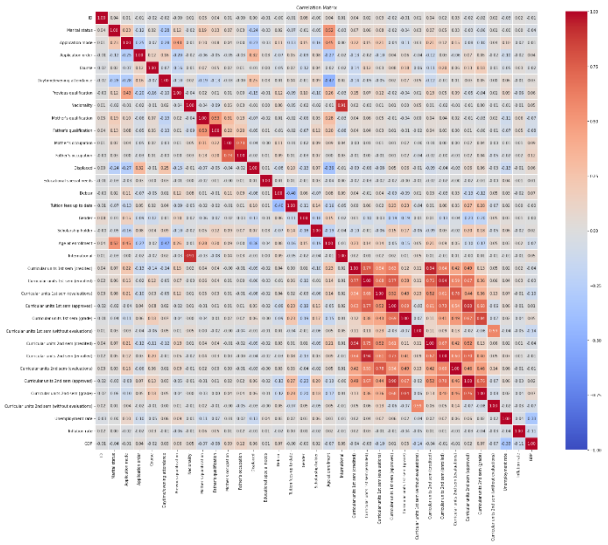

Figure 1: Correlation Matrix of the Dataset

After splitting the dataset into training and testing sets, feature selection was performed using a RF classifier to identify features with an importance score greater than 0.02. It effectively reduced dimensionality while retaining the most relevant features for prediction. Additionally, to address class imbalance in the dateset, the SOMTE was employed. It generated synthetic samples for underrepresented classes, ensuring a balanced class distribution and improving the model robustness.

Classification Algorithm. This study employed RF as the backbone (Figure 2). It operates by constructing an ensemble of decision tree, where each tree votes on the final prediction, ensuring robustness and reducing overfitting. Its ability to handle multi-class classification tasks and its adaptability to heterogeneous datasets make it highly suitable for the current dataset.

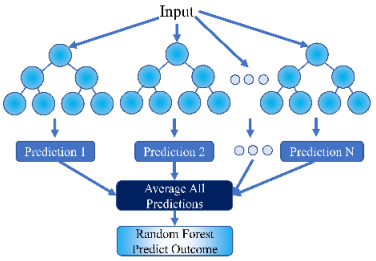

Figure 2: Random Forest Model Structure

Results - By employing grid search for hyperparameter optimization during the ablation study, all models were tested under optimal configurations. Three additional model were trained and compared with the RF. RF outperformed the others, achieving the highest Kaggle score of 0.81674, shown in Table 1.

Table 1: Model Performance Comparison

Model	Accuracy	Cross-validation Score	Kaggle Score
RF	0.79	0.81	0.81674
XGBoost	0.76	0.80	0.79638
LightGBM	0.78	0.79	0.78280
Deep Forest	0.77	0.80	0.80090

Discussion - The RF approach is lightweight, with short training times and low resource requirements. However, it suffers from overfitting, as the model performs better on the test set than on unseen data. This overfitting issue is even more severe in more complex models, indicating limited dataset diversity. A solution involving additional data collection or augmentation techniques may reduce it.

Conclusion – This study developed a RF model with suitable preprocessing techniques to perform a multi-class classification task, predicting student academic trajectory. Compared to other considered models, the proposed approach shown superior performance, achieving higher evaluation metrics and a best Kaggle score. In conclusion, leveraging Big Data is essential for achieving robust and accurate results in machine learning prediction tasks.