Team Progress Report: Developing a Machine Learning Model to Predict Student Dropout Rates or Academic Success

Introduction:

Our team has been working on developing a machine learning model that can predict student dropout rates or academic success based on a variety of factors, such as attendance, grades, and demographic data. The goal of the project is to identify students who are at risk of dropping out or falling behind and provide targeted interventions and support.

Team Members:

Due to unforeseen circumstances, only one team member has been able to actively work on the project for the time being.

Work Completed:

- **Dataset Retrieval:** I retrieved the dataset from Kaggle and cited the paper it was scraped from.
- **Exploratory Data Analysis (EDA):** I performed EDA on the dataset, creating histograms and plots to visualize the data. The dataset has a large number of features.
- **Model Development:** I created a baseline model using a Random Forest Classifier (SVC), which achieved around 70% accuracy. Due to the high dimensionality of the data, the team will probably be using tree-based and SVC-based algorithms in the future. A Python script was also written to run the preprocessed data on these models and return results for robust testing in the future.

The repository with the corresponding scripts:
https://github.com/sulphatet/academic-predictor/tree/main

Next Steps:

- **Feature Engineering:** One of the next steps in our project is to engineer highly correlated features in the dataset. This will help us improve the accuracy of the model and identify important factors that contribute to student dropout rates or academic success.
- **Algorithm Tuning:** We will also be trying out different algorithms to improve the accuracy of the model. As mentioned before, we will be using tree-based and SVC-based algorithms in the future. However, tree based models appear to overfit on the training data.
- **Dealing with the 'Enrolled' class:** One of the classes in our dataset is 'Enrolled,' which is not very useful for our problem set. We are contemplating what to do with this class, whether to remove it or merge it with another class.

Conclusion:

Despite some unforeseen circumstances that have limited the availability of team members, progress has been made in developing a machine learning model to predict student dropout rates or academic success. I have completed EDA, created a baseline model, and written a Python script to run the preprocessed data on different models for future testing. In the future, we will be working on Feature Engineering, Algorithm Tuning, and addressing the issue of one of the classes being not very useful for our problem set. We are looking forward to having our entire team back and continuing our work on the problem set to improve the accuracy of the model.