Student Dropout Analysis for School Education

Introduction:

Access to quality education is a fundamental right, and governments worldwide strive to ensure every child's enrollment and completion of their schooling. However, school dropout rates remain a persistent challenge, often influenced by various social, economic, and demographic factors. To address this issue, the Government of Gujarat has recognized the need for a comprehensive analysis of dropout patterns at the school level. By understanding the underlying causes and identifying vulnerable groups, the government aims to formulate targeted interventions that can significantly reduce dropout rates.

This project presents an in-depth analysis of student dropout trends in school education, utilising a dataset titled "Predict Students' Dropout and Academic Success - Investigating the Impact of Social and Economic Factors." **The dataset, sourced from Kaggle and contributed by thedevastator**(https://www.kaggle.com/thedevastator), encompasses many attributes that shed light on the dynamics contributing to student dropout.

Project Overview:

The primary objective of this project is to conduct a comprehensive analysis of student dropout rates in school education, with a focus on the state of Gujarat, utilising the available dataset titled "Predict Students' Dropout and Academic Success - Investigating the Impact of Social and Economic Factors." While the dataset may not include information on schools, areas, or castes, we can still extract valuable insights from the existing attributes.

The analysis aims to provide insights into the following key aspects:

Demographic Analysis: We will explore how demographic factors such as gender, age at enrollment, marital status, and nationality correlate with student dropout rates.

Economic Factors: Investigate the influence of economic factors, such as parental occupation, tuition fee payment status, and scholarship eligibility, on student dropout rates.

Academic Performance: Analyze how students' academic performance, represented by variables like curricular units and evaluations, impacts their likelihood of dropping out.

Social and Special Needs: Explore whether students with educational special needs or those facing unique challenges like displacement or debt are more susceptible to dropout.

Macroeconomic Factors: Investigate how broader economic indicators like the unemployment rate, inflation rate, and GDP growth relate to dropout rates, as these can indirectly affect education outcomes.

The expected outcome of this analysis is to provide valuable insights into the complex web of factors influencing student dropout. The government can develop targeted interventions and policies to improve student retention and foster a conducive learning environment by identifying high-risk groups and understanding the nuanced factors contributing to dropout rates.

In the subsequent sections of the project, we will delve into data preprocessing, exploratory data analysis, and the development of predictive models to aid in the dropout analysis. While we may not have school-wise, area-wise, or caste-wise information, we will use the available attributes to contribute to the government's efforts in ensuring every child's right to education and reducing dropout rates where possible.

About DataSet:

This dataset provides a comprehensive view of students enrolled in various undergraduate degrees offered at a higher education institution. It includes demographic data, socioeconomic factors and academic performance information that can be used to analyze the possible predictors of student dropout and academic success. This dataset contains multiple disjoint databases consisting of relevant information available at enrollment, such as application mode, marital status, course chosen and more. Additionally, this data can be used to estimate overall student performance at the end of each semester by assessing curricular units credited/enrolled/evaluated/approved as well as their respective grades. Finally, we have the unemployment rate, inflation rate and GDP from the region which can help us further understand how economic factors play into student dropout rates or academic success outcomes. This powerful analysis tool will provide valuable insight into what motivates students to stay in school or abandon their studies for a wide range of disciplines such as agronomy, design, education nursing journalism management social service or technologies.

Dataset Features:

Initially we had 34 features like Martial_Status, application, Course, Previous qualification, Nationality, Mothers_Qualification, Mothers_occupation, Fathers_Qualification, Fathers_Occupation, Tuition fees, Age at enrollment, Gender, Scholarship holder, Curricular units 1st sem (credited), Curricular units 1st sem (enrolled), Curricular units 1st sem (approved), Curricular units 1st sem (without Evaluation), Curricular units 1st sem (grade), Similarly for 2nd Semester, Inflation rate, Unemployment rate, GDP, etc.

Data Preprocessing:

a. Data Cleaning: Begin by inspecting the dataset for missing values, outliers, and inconsistencies. Handle missing data using techniques like imputation or removal, and address outliers appropriately.

b. Data Exploration: Explore the dataset to understand the distribution of features, relationships between variables, and any patterns that might emerge. Here we do not have any null values, or None type values, the ‘nationality’ column name is replaced by ‘nationality’ We have one categorical column, which is our target column, and we have mapped the unique values of that column to integer values, with the help of LabelEncoder Library.

Feature Selection:

To build an efficient classification model, several feature selection techniques were employed to identify the most relevant features while removing insignificant ones.

1. Correlation Analysis: This method evaluated the relationship between the target variable and other features. Strongly positive and negative correlated features were retained as they provided valuable information for distinguishing between classes.

2. Recursive Feature Elimination (RFE): RFE iteratively removes less important features based on their contribution to the model’s accuracy. By training and eliminating features in multiple rounds, only the most significant ones were retained, improving model efficiency.

3. Variance Threshold: Features with low variance, which contribute little to the model, were removed. Such features often display similar values across samples and don't help in-class differentiation.

4. Data Knowledge: In addition to statistical methods, domain knowledge was used to retain important features, even if they had low variance but were relevant in the context of student outcomes.

As a result, 19 important features were selected based on their high correlation, predictive power from RFE, and variance. This refined feature set enhances the model's accuracy, reduces overfitting, and improves interpretability, ensuring better performance when predicting student outcomes.

We have also calculated the Correlation heat map, i.e. correlation matrix. So finally we are neglecting the less important features, like the inflation rate", "Educational special needs", "Nationality", "Unemployment rate", "Course", "International", "Father's qualification", "Marital status", "Application order", "GDP", "Displaced", "Curricular units 1st sem (without evaluations)", "Curricular units 2nd sem (without evaluations)", "Previous qualification", "Daytime/evening attendance", & "Marital status".

We have finally Nineteen Columns as the final important feature.

Target Variable:

The classification problem at hand involves predicting student outcomes based on various features in the dataset. The target variable, labelled 'Target', has three unique categorical values: 'Dropout', 'Graduate', and 'Enrolled'. The objective is to accurately classify students into one of these categories using the dataset features.

Data Split:

To ensure robust model performance and avoid overfitting, the dataset was divided into three parts: training, validation, and test sets. The most relevant features, identified through feature selection methods, were used to create these sets. Initially, the dataset was split into training and testing sets, and then the training set was further divided into training and validation subsets. This allows for model evaluation of unseen data and helps in fine-tuning.

Model Selection:

For the task of predicting student outcomes, we experimented with two machine learning models: Random Forest Classifier and Support Vector Machine (SVM). Random Forest was selected due to its ability to handle complex datasets and reduce overfitting through ensemble learning. SVM was also used for comparison, known for its performance in classification tasks with well-separated classes.

Model Training:

The models were trained on the training data, with hyperparameters fine-tuned to optimize performance. During training, cross-validation was employed to assess generalization performance and avoid overfitting.

Model Evaluation:

Model performance was evaluated using key metrics such as accuracy and the confusion matrix. The Random Forest Classifier achieved an accuracy of 0.78 on the test set (comparing `y_test` and `y_pred`) and 0.77 on the validation set (comparing `y_val` and `y_val_pred`). The SVM model, on the other hand, achieved an accuracy of 0.77. These results indicate strong performance for both models, though further optimization could improve their predictive power.

Both models show promising potential for predicting student outcomes accurately, making them useful for real-world applications.

Summary:

The **Student Dropout Classification** project predicts student outcomes (Dropout, Graduate, Enrolled) using key features from the dataset. By applying feature selection methods like correlation analysis, Recursive Feature Elimination (RFE), and variance thresholding, we identified the top 19 features for the model. Using Random Forest and SVM classifiers, the models achieved test accuracies of 0.78 and 0.77, respectively. The model has been successfully deployed via Streamlit and is publicly accessible on GitHub for real-time student outcome predictions.