School Dropout Risk Analysis

1. Introduction

School dropout remains one of the most pressing challenges in the global education system. The consequences of students leaving school prematurely are extensive, impacting not only individual futures but also broader societal development. This project, titled *School Dropout Risk Analysis*, aims to proactively identify students who are at risk of dropping out by leveraging data analytics and machine learning techniques. Using a student dataset, we aim to build a predictive model to help educational institutions implement early interventions.

2. Dataset Overview

The dataset was obtained in CSV format (semicolon-delimited) and comprises 4424 records of student attributes. These include demographic data, family background, academic performance, and other school-related features such as:

- Demographics: sex, age, address
- Education and Family: Medu (mother's education), Fedu (father's education), failures, schoolsup, famsup
- Academic Performance: G1, G2, G3 (grades)
- Personal and Social: activities, internet, studytime

The primary outcome variable, G3 (final grade), was used to derive a binary target variable:
- Dropout = 1 if G3 < 10 (at-risk)
- Dropout = 0 otherwise

3. Data Cleaning and Preprocessing

After confirming there were no missing values, we ensured all applicable columns were converted to numeric types. Label encoding was not used since the numeric features were sufficient for effective model training. A function was applied to attempt conversion of all columns to numeric, ensuring compatibility with machine learning algorithms. We applied StandardScaler to normalize feature values for better convergence during model training.

School Dropout Risk Analysis

4. Exploratory Data Analysis (EDA)

A correlation heatmap was used to identify important features influencing dropout. High correlation was found between early academic performance (G1, G2) and the final grade (G3). Other features such as previous failures and lack of support also indicated higher dropout risks.

Key findings:

- Strong positive correlation among grades (G1, G2, G3)
- Students with multiple past failures were more likely to drop out
- Family education levels and support influenced retention

5. Model Building and Evaluation

We trained a Logistic Regression model on the processed dataset. The dataset was split into training and test sets (80:20). Feature scaling was applied using StandardScaler. The logistic regression model achieved an accuracy of approximately 89%. A confusion matrix showed strong classification performance with relatively low false negatives, which is essential in dropout prediction.

6. Model Deployment

The trained model was saved using the joblib library. This allows the model to be reused for prediction or integration with external systems like school management dashboards.

7. Recommendations

Based on the data-driven insights:

- Provide additional support to students with low G1 and G2 scores.
- Monitor students who previously failed subjects.
- Focus on improving study habits and time management through structured programs.
- Encourage parental involvement, especially in families with lower educational backgrounds.

8. Conclusion

School Dropout Risk Analysis

This project demonstrates the effectiveness of machine learning in predicting school dropout risk using available student data. With an accuracy of approximately 89%, the logistic regression model provides a valuable tool for identifying at-risk students early. Educational institutions can use such models to implement targeted interventions and reduce overall dropout rates.