**School Dropout Risk Analysis**

Pusuluri Naga Anusha

Pusulurinaganusha@gmail.com,7093709617

21-06-2025

**Abstract**

Student dropout is a critical concern impacting both educational institutions and the broader society. This project explores the contributing factors behind school dropout through detailed data analysis and predictive modeling. By examining academic records, financial status, and demographic characteristics, we aim to uncover patterns that signal dropout risk. Using Python for data processing and machine learning, and Tableau for visual storytelling, this analysis provides meaningful insights and practical recommendations. The outcome supports proactive measures to enhance student retention and academic success.

**1. Introduction**

**Problem Statement**

Student dropout remains a persistent challenge in higher education, affecting not only individual academic and career outcomes but also institutional performance and policy planning. While educational institutions collect vast amounts of student data covering demographics, academic records, and financial status this data is rarely used effectively to anticipate or prevent dropout events.

This project aims to bridge that gap by applying data analytics and machine learning techniques to identify key risk factors that lead to student dropout. By transforming raw data into actionable insights, the goal is to empower educational institutions to intervene earlier and support students more effectively. This approach can help minimize dropout rates, enhance student success, and improve institutional decision-making.

**2. Objectives**

* Clean and preprocess student data for accurate analysis.
* Identify trends using exploratory and diagnostic analysis.
* Develop a predictive model to assess dropout risk.
* Build an interactive dashboard in Tableau for visual communication.
* Recommend targeted interventions to reduce dropout rates.

**3. Data Preparation**

* **Dataset Name**: Predict Students' Dropout and Academic Success
* **Source**: Kaggle
* **Tools Used**: Pandas, Matplotlib, Seaborn, Scikit-learn
* **Cleaning Steps**:
  + Renamed columns to remove special characters
  + Filled missing values:
    - Numerical: median value
    - Categorical: mode value

Check and fill missing values

numerical\_cols = df.select\_dtypes(include=['int64', 'float64']).columns categorical\_cols = df.select\_dtypes(include=['object']).columns

df[numerical\_cols] = df[numerical\_cols].fillna(df[numerical\_cols].median()) for col in categorical\_cols: df[col] = df[col].fillna(df[col].mode()[0])

**4. Exploratory Data Analysis (EDA)**

Key visual insights:

* **Age Distribution**: Majority of enrollments are between 18–22 years.
* **Unemployment Rate**: External economic factors may affect student stability.
* **Correlation Heatmap**: Highlights strong correlations among numeric features.
* **Age vs Target**: Older students show a higher dropout tendency.
* **Tuition Fee Status vs Target**: Financial dues strongly influence dropout rates.

**5. Diagnostic Analysis**

By comparing different student segments, we identified:

* Dropouts are more likely to:
  + Be older
  + Have low GPA
  + Accumulate more failed subjects
  + Miss tuition payments

**6.** **Predictive Modeling**

* These findings suggest specific student profiles that need timely attention.
* **Algorithm Used**: Random Forest Classifier
* **Target Variable**:
  + Dropout: 0, Graduate: 1, Enrolled: 2
* **Feature Set**: Numerical variables excluding the encoded target

python

CopyEdit

model = RandomForestClassifier(n\_estimators=100, random\_state=42)  
model.fit(X\_train, y\_train)  
predictions = model.predict(X\_test)

* **Model Performance**:
  + Achieved ~90% accuracy
  + Key predictors: GPA, Age at enrollment, Tuition status, Failed subjects

**7. Tableau Dashboard**

To visually communicate the insights, we developed a Tableau dashboard containing:

1. **Target Outcome Pie Chart**
2. **Gender vs Target Bar Graph**
3. **Filtered Dropout by Gender**
4. **Tuition Fee Status vs Target Outcome**
5. **Age vs Target (Boxplot/Histogram)**
6. **Application Mode vs Target**
7. **Course vs Dropout Rate (Tree Map)**
8. **Enrollment Year Trend Analysis**

Interactive filters (Gender, Target, Fee Status) allow dynamic exploration.

**8. Prescriptive Insights**

Based on the analysis:

* **Strengthen Financial Aid Programs** for students behind on tuition.
* **Develop Gender-Inclusive Policies** for retention.
* **Create Mentorship Programs** for older students.
* **Redesign High-Risk Courses** or offer additional academic support.
* **Screen Applications** and provide onboarding to high-risk applicants.

**9. Conclusion**

By combining machine learning and data visualization, this project provides a holistic approach to tackling student dropout. The model enables proactive risk detection, while the dashboard empowers institutions to monitor trends and make data-driven decisions. Together, these tools foster a more supportive learning environment and reduce the likelihood of students leaving prematurely.

**Tools Used**:

* Python (Pandas, Seaborn, Scikit-learn, Matplotlib)
* Tableau for Dashboarding
* Microsoft Excel for dataset input

**Appendix**:

* Project published on Tableau Public
* Cleaned dataset saved as cleaned\_dataset.xlsx
* Dashboard features filters, visuals, and student segmentation