Student Dropout Analysis

Overall Analysis and Dashboard Interpretation

Submitted to: Gigaversity
Data Analytics Internship Challenge

Submitted by: Vivekanand Ojha
vivekanandojha09@email.com
linkedin.com/in/vivekanand-ojha-485462289
https://github.com/vivekOJ1129

June 2025

This report provides an in-depth analysis of student dropout patterns using machine learning models and interactive Tableau visualizations to support early intervention strategies in education.

Introduction

Student dropout is a critical concern for educational systems worldwide. This analysis combines machine learning and Tableau-based visualizations to identify and explain the primary factors contributing to dropout risk, such as GPA, absences, and parental education. The objective is to enable data-informed interventions that can reduce student attrition and improve academic outcomes.

1 Objective

To analyze and predict student dropout risk by applying machine learning techniques and Tableau dashboards, enabling educational institutions to take proactive steps in reducing dropout rates through early identification of at-risk students.

2 Problem Framing

Student dropout is a multifaceted issue influenced by academic, behavioral, and socio-economic factors. Identifying students at risk before they leave the education system can significantly improve institutional performance and student outcomes.

This project approaches dropout prediction as a binary classification problem: each student is labeled as either **High** or **Low** dropout risk. The goal is to develop a data-driven model that can accurately classify students using available features such as GPA, absences, test preparation, parental education, and more.

By framing this challenge in predictive terms, institutions can shift from reactive measures to proactive interventions, ultimately increasing retention and student success rates.

3 Dataset Overview

The original dataset for this project was compiled from publicly available sources, including:

- Kaggle – Student Performance Dataset
- UCI Machine Learning Repository – Student Alcohol Consumption / Dropout Records
- World Bank and UNESCO – Dropout Trends by Region

This raw dataset contained diverse educational attributes related to student academic history, behavioral trends, socio-demographics, and institutional conditions.

The combined data was stored in the file `Student_Dropout_Dataset.csv`. It required extensive preprocessing to ensure reliability, consistency, and suitability for modeling and visualization tasks.

The final cleaned dataset used for Tableau dashboards is named: `Tableau_Ready_Dropout_Dataset.json`, which was derived from the machine learning cleaning pipeline.

4 Data Cleaning

The raw data included missing values, inconsistent labels, and mixed data types. The cleaning and preparation process was performed in Python and optimized separately for machine learning and Tableau use cases.

For Machine Learning (Logistic Regression and Random Forest)

- Numerical missing values were filled using either **mean** or **median**, based on skewness.
- Categorical variables were imputed using the **mode**.

- Label encoding was applied to convert categorical values into numeric form.
- New features such as `Average Score` were engineered for model accuracy.

For Tableau Visualizations

- The cleaned ML dataset was post-processed to retain original category labels for better human readability.
- The binary target variable `dropout_risk` was remapped from 0/1 to `Low` / `High`.
- The final version was exported as `Tableau_Ready_Dropout_Dataset.json`.

5 Descriptive and Diagnostic Analysis

Descriptive and diagnostic analysis was performed using both Python and Tableau to better understand the distribution, relationships, and potential risk indicators in the dataset.

Exploratory Data Analysis (Python)

Initial exploration using Python involved:

- Visualizing distributions of key features such as GPA, Absences, and Failures.
- Correlation heatmaps to identify relationships between variables and the target label (`dropout_risk`).
- Boxplots and bar charts to compare feature distributions across dropout categories.

These analyses confirmed that lower GPA, higher absences, and a greater number of past failures are strongly associated with high dropout risk.

Interactive Dashboards (Tableau)

Tableau was used to build dynamic dashboards offering multidimensional views of the data. Key visualizations included:

- **Dropout Risk Distribution:** A bar chart showing the proportion of students in High vs. Low dropout risk categories.
- **GPA vs Dropout Risk:** A column chart revealing that students with lower average scores are more likely to drop out.
- **Absences and Failures Boxplot:** Highlighting greater variability and higher median values in the high-risk group.
- **Heatmap by Gender and Parental Education:** Displaying how parental education and gender intersect with dropout risk levels.

These visual insights supported the findings from Python and helped communicate results in an intuitive, stakeholder-friendly format.

6 Machine Learning Models

6.1 Logistic Regression

Logistic Regression was implemented as a baseline model to classify students into `Low` or `High` dropout risk categories. It is a well-established, interpretable model suitable for binary classification problems.

The model was trained on features such as:

- GPA (average academic score)
- Number of failures
- Absences
- Parental education
- Test preparation course status

Model Performance:

- **Accuracy:** 0.865
- **Cross-validation Accuracy:** 0.883

Logistic Regression provided a solid starting point for binary classification and was useful for benchmarking against more complex models such as Random Forest.

6.2 Random Forest Classifier

Random Forest is an ensemble-based machine learning algorithm that combines multiple decision trees to improve prediction accuracy and reduce overfitting. It is particularly effective in capturing non-linear patterns and complex interactions between features.

The Random Forest model was trained using the same cleaned feature set used in the Logistic Regression model, including GPA, failures, absences, parental education, and test preparation status.

Model Performance:

- **Accuracy:** 0.985
- **Cross-validation Accuracy:** 1.000[1]

Classification Report (Summary):

- **Class 0 (Low Risk):** Precision = 1.00, Recall = 0.98, F1-Score = 0.99
- **Class 1 (High Risk):** Precision = 0.96, Recall = 1.00, F1-Score = 0.98
- **Macro F1 Score:** 0.98
- **Weighted F1 Score:** 0.99

Top Features by Importance:

- GPA
- Absences
- Failures
- Test Preparation Course

Model Commentary:
Despite high accuracy and consistent performance, the cross-validation score of 1.000 may indicate limited variability in the dataset or potential overfitting. To improve generalizability, further enhancement of data diversity and quality from original sources (Kaggle, UCI, UNESCO) is planned.

[1] While cross-validation accuracy appears perfect, this may be due to redundancy or bias in the dataset. Further refinement and testing on diverse data is planned.

7 Prescriptive Insights

Based on the outcomes of the descriptive analysis and machine learning models, several actionable insights were identified that can help educational institutions mitigate dropout risks.

- **Monitor GPA Trends:** Students with lower GPA consistently showed higher dropout risk. Early academic support and targeted tutoring programs should be deployed for underperforming students.
- **Track Absenteeism:** Absences were strongly correlated with dropout. Institutions should implement automated attendance monitoring systems with follow-up interventions for students with frequent or prolonged absences.
- **Address Repeated Failures:** A history of multiple subject failures was a key predictor of dropout. Remedial coursework, peer mentoring, and personalized learning plans can help address learning gaps.
- **Promote Test Preparation Programs:** Students who completed test preparation were less likely to drop out. Expanding access to prep resources and encouraging participation may boost student confidence and performance.
- **Engage Parents with Lower Education Backgrounds:** Parental education level also influenced dropout risk. Schools should prioritize outreach and engagement with families where parents have limited formal education, providing them with tools to support their child's academic journey.
- **Use Visual Dashboards for Monitoring:** Tableau dashboards developed in this project offer an intuitive way to monitor trends and identify at-risk students dynamically. These can serve as an early-warning system for administrators and counselors.

By operationalizing these insights, schools can take a proactive, data-informed approach to improving student retention and educational outcomes.

8 Conclusion

This project successfully combined machine learning techniques and interactive visualizations to analyze student dropout risk. Through data cleaning, exploratory analysis, classification modeling, and dashboard interpretation, critical insights were uncovered regarding the academic and behavioral factors influencing dropout rates.

Both Logistic Regression and Random Forest models were implemented, with Random Forest achieving superior accuracy and feature interpretability. Key predictors such as GPA, absences, and failure history were consistently found to correlate strongly with student risk levels.

Tableau dashboards further enhanced the analysis by making complex relationships accessible to stakeholders through intuitive, real-time visuals. These dashboards can be deployed in institutional settings as early-warning systems to monitor and support at-risk students.

Going forward, enhancing dataset diversity and integrating real-time data pipelines could further strengthen model robustness. This data-driven framework demonstrates how education systems can leverage predictive analytics to make informed, impactful decisions that reduce dropout and promote student success.