Predicting Student Academic Performance Using Machine Learning

Njinju Zilefac Fogap

November 24, 2025

1 Introduction

Academic performance plays an essential role in a student's educational path, showing how successfully they meet their learning objectives. It includes factors such as grades, exam results, and participation in academic activities. In today's fast-moving world, gaining insight into academic performance links closely to areas like productivity, efficient study techniques, and personal growth [Focnd]. Early identification of students at risk of academic failure or dropout is essential for universities aiming to enhance academic support, increase retention rates, and improve student wellbeing. With the growth of learning management systems and student information systems, machine learning (ML) provides powerful tools for modelling patterns in demographic, academic, and socio-economic data [rwaafta23].

This project aims to develop and evaluate machine learning models to predict student dropout or academic success using the publicly available dataset Predict Student' Dropout and Academic Success from the UCI Machine Learning Repository (Dataset ID: 697). While waiting for institutional data from the University of Wolverhampton, this data set serves as a strong foundation for exploring modeling techniques, evaluation methods, and deployment strategies.

2 Problem Statement

Many Universities worldwide often face challenges in identifying students who are most likely to struggle academically [RANATa23]. Traditional methods rely heavily on end-of-semester performance, making interventions reactive rather than proactive. The goal is to build a predictive system by deploying the best machine learning predictive model on a streamlit application capable of identifying at-risk students early, enabling timely academic support and institutional planning.

3 Research Aim

To develop, evaluate and deploy machine learning models that accurately predict student academic outcomes (dropout, enrolled, graduate) based on socio-economic, demographic, and educational factors.

4 Research Objectives

- To import and perform a comprehensive exploratory data analysis (EDA) on the data set.
- Data wrangling, by handling missing values, encoding categorical variables, feature engineering, and normalising numerical data.
- Train, evaluate, and compare multiple machine learning models (Support Vector Machines, Logistic Regression, Random Forest, XGBoost or Gradient Boosting).
- Address class imbalance using methods such as SMOTE, class weights, or re-sampling.
- Select the best-performing model using cross-validation, hyperparameter tuning, and performance metrics (accuracy, F1-score, ROC-AUC, confusion matrix).

- Deploy the final model using Streamlit as an interactive web application.
- Document findings, implications for higher education institutions, and recommendations for real-world adoption.

5 Research Questions

1. Which machine learning algorithms provide the highest predictive performance on this dataset?
2. Which student characteristics most strongly influence academic outcomes?
3. How can the predictive model support early intervention strategies?
4. What are the challenges and ethical considerations in modelling sensitive student data?

6 Dataset Description

- **Name**: Predict Students' Dropout and Academic Success.
- **Source**: UCI Machine Learning Repository.
- **Instances**: 4,424.
- **Features**: 36 (Real, Categorical, Integer).
- **Target Classes**: Dropout, Enrolled, Graduate.
- **Year**: 2021.
- **Key Attribute Types**: Demographics (age, gender, nationality), Social-economic factors, Enrollment information, Academic path.

7 Methodology

7.1 Data Preprocessing

- Load and inspect dataset.
- Check for inconsistencies and anomalies.
- Encode categorical variables (LabelEncoder, One-hot).
- Feature scaling (standardisation/normalisation).
- Handle imbalance (SMOTE, class weights).

7.2 Exploratory Data Analysis

- Visualise distributions of features.
- Correlation heatmap.
- Identify most influential predictors.
- Detect outliers and unusual patterns.

7.3 Model Development

- Models to train:Support Vector Machines, Logistic Regression, Random Forest, XGBoost or Gradient Boosting.

7.4 Hyperparameter Tuning

- GridSearchCV with 5-fold cross-validation.
- Performance evaluation metrics: Accuracy, Precision, Recall, F1-score, ROC-AUC (one-vs-rest), Confusion Matrix.

7.5 Model Interpretability

- SHAP value analysis.
- Feature importance plots.

7.6 Model Deployment

- Build a Streamlit web application. Inputs: student demographics + academic data , Output: prediction + risk probability and Include model explainability visuals (SHAP, feature importance).

7.7 Ethical Considerations

- Avoid algorithmic bias.
- Protect sensitive information.
- Ensure transparency and interpretability.
- Clearly state limitations.

7.8 Expected Outcomes

- A validated predictive machine learning model for early detection of at-risk students.
- Insights into key demographic and socio-economic factors impacting academic outcomes.
- A Streamlit-based decision support tool for educational institutions.
- Recommendations for integrating ML systems in academic settings responsibly.

7.9 Tools and Technologies

- Python.
- Github for version control.
- Matplotlib, Seaborn, Plotly.
- Streamlit.
- Jupyter Notebook for development.

7.10 Project Timeline

Phase	Activities	Duration
Week 1	Literature review, dataset understanding, research questions	1 week
Week 2	Data cleaning, preprocessing, handling imbalance	1 week
Week 3	Exploratory Data Analysis	1 week
Week 4	Baseline model training & evaluation	1 week
Week 5–6	Advanced models + hyperparameter tuning	2 weeks
Week 7	Model comparison & selection, interpretability analysis	1 week
Week 8	Streamlit app development (MVP)	1 week
Week 9	Final testing, validation, and documentation	1 week
Week 10	Write-up of report & final submission	1 week

Table 1: Project timeline and duration of activities

7.11 Risk Assessment

Risk	Mitigation Strategy
Dataset imbalance	Use SMOTE, weighted models, and careful metric selection
Overfitting	Apply cross-validation and regularisation techniques
Ethical concerns	Enforce strict anonymisation and transparent model reporting
Time constraints	Follow weekly milestones and use Git version control
Lack of institutional data	Use UCI dataset until Wolverhampton data becomes available

Table 2: Risk assessment and mitigation strategies

7.12 Conclusion

This project will contribute to understanding how machine learning can enhance early identification of at-risk students in higher education. By combining rigorous modelling, interpretability techniques, and user-friendly deployment, the research aims to provide actionable insights for academic support teams and contribute to improving student retention and success.

References

[Focnd] Focus Keeper. What is academic performance? https://focuskeeper.co/glossary/what-is-academic-performance/, n.d. Accessed: February 2025.

[RANATa23] Rizwan Gitay c Abdel-Salam G. Abdel-Salam d Khalifa Al Hazaa e Ahmed BenSaid f Michael H. Romanowsk Rusol Adil Naji Al-Tameemi a, Chithira Johnson b. Determinants of poor academic performance among undergraduate students. *Discover Education*, X(X):XX–XX, 2023. Accessed: 2025-02-24.

[rwaafta23] Please replace with actual authors (from the article). Predicting student academic performance using machine learning techniques: A systematic review. *Procedia Computer Science*, XXX:XXX–XXX, 2023. Accessed: 2025-02-24.