A Machine Learning Approach to Identifying and Predicting Student Dropout in Higher Education

Ali Suhail

Supervised by: Felipe Campelo

Abstract:

Student dropout is a persistent and growing challenge in higher education, with wide-reaching implications for students' prospects and the strategic planning and funding of academic institutions. Withdrawal from studies not only disrupts academic progression but also results in wasted resources, increased administrative burden, and reputational risks for universities. Addressing this issue requires the timely identification of at-risk students so that effective, personalised interventions can be delivered before disengagement becomes irreversible.

This project proposes a data-driven solution by developing a machine learning (ML) based dropout prediction system. The system will analyse multiple data sources, including student demographic information (such as age, education level, and region), academic performance indicators (such as early assignment grades), and behavioural metrics (such as virtual learning environment or VLE engagement patterns) collected at specific points within a module. The task is framed as a binary classification problem where the model predicts whether a student will drop out or continue.

A complete ML pipeline will be implemented, beginning with exploratory data analysis to uncover trends and correlations, followed by data preprocessing such as handling missing values, detecting outliers, encoding categorical features, and normalising data. The dataset will be divided into training and testing subsets to ensure reliable performance evaluation. Several supervised learning models will be developed, including logistic regression (LR), support vector machines (SVM), random forests (RF), and neural networks like the Multilayer Perceptron (MLP). Each model will undergo hyperparameter tuning and be evaluated using metrics such as accuracy, precision, recall and F1-score. Special attention will be given to reducing false negatives to ensure that students at risk are not overlooked.

The ideal outcome is a reliable, interpretable, and scalable model that can identify students likely to drop out by the midpoint of the module or earlier, enabling timely and effective support. In addition to accurate prediction, the project also seeks to identify the most influential factors contributing to dropout and to understand how their impact varies over time. These findings will help academic institutions make data-informed decisions, allocate resources more efficiently, and design targeted interventions that enhance student retention and academic success.

Ethics statement:

This project fits within the scope of ethics pre-approval process, as reviewed by my supervisor Felipe Campelo and approved by the faculty ethics committee as application 15208.

Project plan:

Introduction

Student dropout remains a significant issue in higher education, impacting not only students' academic success but also the financial health and reputation of universities. Early identification of students at risk of dropping out of their studies enables timely interventions, which can enhance retention and improve student outcomes. In recent years, ML has become a valuable approach for predicting at-risk students across various courses. These models utilise diverse data sources such as VLE interactions, continuous assessment results, and demographic details [1]. By analysing this information, predictive models can reveal which students are vulnerable and the reasons behind their struggles, enabling educators to offer targeted, personalised support [1]. Nonetheless, dropout prediction is challenging due to the complex interplay of demographic, academic, and behavioural factors. This project is motivated by the goal of creating a dependable and scalable ML-based dropout prediction system to assist educators and administrators in better understanding and mitigating dropout risks.

This project focuses on predicting student dropout by analysing data available up to a specified point within the duration of a module. Since the outcome is binary, indicating whether a student drops out or not, the task is framed as a classification problem. A variety of ML models will be developed and compared, with hyperparameter tuning applied to improve their performance. The model that demonstrates the highest effectiveness will be selected for final use.

A comprehensive ML pipeline will be developed specifically for predicting student dropout. The process will begin with exploratory data analysis to identify patterns and relationships within the data, including student demographics, academic history, and engagement with the VLE. Next, the data will be pre-processed by handling missing values, managing outliers, and encoding categorical variables. After preparation, the dataset will be split into training and testing sets to ensure an unbiased evaluation of model performance. Several machine learning models will then be implemented, covering traditional approaches such as LR, SVM, and RF, along with neural network models like the MLP Classifier. Each model will be optimised using hyperparameter tuning and evaluated using metrics such as accuracy, precision, recall, and F1-score. Based on this evaluation, the model with the best performance will be chosen to predict student dropout.

Project Background and Motivation

The rising dropout rates in higher education have become an increasing concern globally, with serious implications for students, educational institutions, and policymakers. Although greater access to university education has created a larger pool of graduates for the labour market, it has also resulted in a notable increase in the number of students leaving before completing their degrees [2]. According to the OECD (2019), dropout rates are increasing by an average of around 30% across many countries [3]. This highlights the need for effective strategies to identify and support students who are at risk of disengaging, while still maintaining academic standards despite growing enrolment figures.

In the United Kingdom, data from the Student Loans Company (SLC) highlights the issue, showing a 28% rise in university dropouts over five years. The number of students who took out loans but failed to complete their courses increased from 32,491 in 2018–19 to 41,630 in 2022–23 [4]. Mental health challenges have been identified as a major cause of early withdrawal [4]. These statistics emphasise the need for early intervention and predictive tools to help academic staff identify students who may require additional support. Previous research has shown that targeted academic measures, such as personalised emails and proactive tutor involvement, can reduce dropout rates by 11% in affected classes [5]. While the study acknowledged that factors like course design might also affect outcomes, it did not explore these in depth. Additionally, it pointed out that distance learning provides valuable opportunities to monitor and respond to student engagement.

Predicting student dropout is also important for managing academic resources and improving learning outcomes. Accurate predictions allow institutions to provide timely support, such as tutoring or customised learning pathways. They also enable better planning, such as adjusting teaching staff levels or identifying courses that may need revision. By implementing machine learning models that forecast dropout risk, universities can take data-driven actions to reduce attrition, improve retention, and enhance the overall quality of education.

Objectives and Hypothesis

This project assumes that prediction accuracy improves as more data becomes available during the module, though dropout likelihood generally decreases over time. The focus is on early identification of at-risk students by uncovering key dropout-related features. The model will be designed for use around the module midpoint or earlier as an early warning system. By analysing dropout indicators at different stages, the project aims to find out when predictions are most accurate and understand performance variations. Beyond accuracy, it will identify major dropout factors to inform academic support. The chosen model will support institutions in improving retention and success. A table of additional assumptions and hypotheses will accompany the analysis.

Topic	Hypothesis	Rationale	Test
Engagement Hypothesis	Students with low VLE engagement in the early stages of the module are more likely to drop out.	Low interaction with online content may indicate a lack of motivation/resources.	Analyse dropout rates based on engagement levels within the first 2–3 weeks.
Assessment Performance Hypothesis	Poor performance on early continuous assessments increases dropout risk.	Early low grades may discourage students and lower their confidence in passing.	Correlate early assessment scores with dropout outcomes.
Demographic Disparity Hypothesis	Certain demographic groups (based on age, region, and education level) are disproportionately represented among dropouts.	External factors like work, childcare, or lack of prior academic support may contribute.	Analyse dropout distribution across demographic variables.
Re-enrolment Hypothesis	Students who previously dropped out of another module or presentation have a higher chance of dropping out again.	Past dropout may be a predictor of future academic risk or instability.	Track student IDs across modules and measure repeat dropout patterns.

In addition to building an accurate prediction system, this project aims to uncover the underlying factors driving student dropout at various points in the module. These insights will help educators implement targeted interventions earlier, where they are likely to be most effective. The goal is to support proactive, data-informed decision-making that improves retention and enhances student success.

References:

[1] Markson Rebelo Marcolino *et al.*, "Student dropout prediction through machine learning optimization: insights from moodle log data," *Scientific Reports*, vol. 15, no. 1, Mar. 2025, doi: https://doi.org/10.1038/s41598-025-93918-1.

[2] "Factors influencing academic performance and dropout rates in higher education," *Oxford Review of Education*, 2025, doi: https://doi.org/10.1080//03054985.2024.2316616.

[3] "Executive summary," *OECD*, 2025. https://www.oecd.org/en/publications/education-at-a-glance-2019_f8d7880d-en/full-report/component-5.html#execsumm-d1e1370 (accessed Jun. 13, 2025).

[4] J. Bryson, "University dropout rates reach new high, figures suggest," *BBC News*, Sep. 28, 2023. https://www.bbc.co.uk/news/education-66940041 (accessed Jun. 13, 2025).

[5] "A systematic review on the deployment and effectiveness of data analytics in higher education to improve student outcomes," *Assessment & Evaluation in Higher Education*, 2020, doi: https://doi.org/10.1080//02602938.2019.1696945.

Appendix: Project Timeline

TASK	PRIORITY	PROGRESS	START	END
Project Initiation				
Read Project Description & Literature	Must Have	100%	02/06/2025	06/06/2025
Gantt Chart & Identify Risks (MoSCoW)	Must Have	100%	06/06/2025	08/06/2025
Setup Trello Kanban & GitHub Repository	Must Have	100%	06/06/2025	09/06/2025
Ethics Review & Test	Must Have	100%	06/06/2025	09/06/2025
Project Plan	Must Have	100%	10/06/2025	22/06/2025
Data Exploration and Preprocessing				
Exploratory Data Analysis (EDA)	Must Have	20%	14/06/2025	17/06/2025
Data Cleaning & Preprocessing	Must Have	0%	18/06/2025	21/06/2025
Establish Machine Learning (ML) Pipelines	Must Have	0%	22/06/2025	24/06/2025
Train/Test Split	Must Have	0%	25/06/2025	26/06/2025
Data Scaling & Transformation	Should Have	0%	27/06/2025	29/06/2025
Model Development & Evaluation				
Run Baseline ML Models	Must Have	0%	30/06/2025	05/07/2025
Model Tuning & Optimisation	Should Have	0%	06/07/2025	12/07/2025
Evaluate Model Results	Must Have	0%	13/07/2025	17/07/2025
Analyse & Interpret Results	Must Have	0%	18/07/2025	24/07/2025
Real-time Dropout Prediction Dashboard	Could Have	0%	24/07/2025	28/07/2025
Analysis and Reporting				
Write Results Section	Must Have	0%	25/07/2025	31/07/2025
Write Methodology Section	Must Have	0%	01/08/2025	07/08/2025
Write Introduction & Literature Review	Must Have	0%	08/08/2025	14/08/2025
Write Conclusion & Abstract	Must Have	0%	15/08/2025	21/08/2025
Write Future Work (Optional)	Could Have	0%	17/08/2025	21/08/2025
Final Edits & Proofreading	Must Have	0%	22/08/2025	27/08/2025

Appendix: Risk Assessment

Risk	Likelihood	Impact	Mitigation
Project Timeline Delays	Low	Unexpected delays in data preprocessing or model development may impact final delivery.	Maintain a detailed project plan with buffer periods. Regularly monitor progress and adjust priorities as needed. Communicate early about delays.
Travel Plans to Kuwait	Low	Travel to Kuwait (18^{th} July to 25^{th} July). Possible delays in project tasks like data preprocessing, model development and report writing	Follow the project plan thoroughly and complete some tasks early if possible
Data Quality Issues	Medium	Poor data quality, such as missing values, inconsistencies, or incorrect records, could lead to unreliable models or biased results.	Conduct thorough exploratory data analysis (EDA) to identify and address missing or inconsistent data early. Use data imputation, validation, and cleaning techniques. Document assumptions clearly.
Computational Resource Constraints	Low	Training complex models, especially neural networks, could require more computational power or time than is available.	Plan for efficient resource use by starting with simpler models. Use cloud services if needed. Optimise code and leverage batch processing.
Insufficient or Imbalanced Data	Medium	A lack of sufficient examples of dropout cases or imbalanced classes may reduce model accuracy and generalizability.	Employ data augmentation or resampling methods (e.g., SMOTE) and carefully tune models to handle class imbalance. Use cross-validation to evaluate robustness.
Incomplete Feature Set	Low	Features critical for prediction may be missing or poorly defined, limiting model performance.	Collaborate with domain experts and stakeholders to identify key features. Iteratively refine feature engineering based on model feedback and data availability.
Limited Time for Iterative Refinement	Low	Delays could affect the quality or completeness of the final model.	Adopt an agile approach. Prioritise building a baseline model early and iteratively improve it. Schedule weekly progress reviews to stay on track.
Software or Tooling Issues	Low	Unexpected bugs or compatibility issues with libraries or platforms may disrupt development.	Use widely supported, stable tools. Regularly back up work and document dependencies. Allocate buffer time for troubleshooting.