Predicting Student Dropout in a Virtual Learning Environments: An Analysis Incorporating Engagement and Socio-Economic Factors

MSc student: Carlos Duran Calle

Supervisor: Professor Felipe Campelo

Abstract

Online learning offers great access to education; however, it still struggles with students leaving programs early and having different levels of success in their studies, which causes problems for both students and the schools [1]. Therefore, it is very important to understand and accurately predict how students will perform, meaning if they will 'Pass', 'Fail', or 'Withdraw' from a course. This allows for quick help and better learning support systems [2]. This project aims to solve this problem by creating a model to predict student dropout, using data from the Open University Learning Analytics Dataset (OULAD) [3]. A key part of this research involves adding a 'Student Engagement' variable, which will be a simple yes/no (binary) measure, as suggested by Mushtaq et al. [4]. This engagement variable will be calculated from detailed information about student activity in the online learning system (VLE), high scores on assignments, and final academic results. The purpose of this is to understand how much students participate, and then identify strong connections between engagement and dropout, which will make the dropout prediction model more accurate. Furthermore, this project will look into how social and economic factors, especially the deprivation index, influence student outcomes. The deprivation index is a measure that shows how disadvantaged an area is, taking into account things like income, employment, and education. We will use statistical tools, like Analysis of Variance (ANOVA), to understand the subtle effects these factors have. ANOVA is a statistical test that helps us compare the average values of different groups to see if there are meaningful differences between them. We will also study how learning changes over time by looking at student performance at different points, such as after important assignment deadlines. This also includes examining how specific features of each course impact the chances of students achieving different outcomes.

The main goal of this project is to build a prediction system. This system will not only identify students who are likely to drop out or fail early in their studies, but it will also help explain why these outcomes might happen. Having this predictor would give educators and administrators the power to quickly adjust the support they offer to students. Additionally, it would allow them to improve course designs. Ultimately, this approach is expected to lead to a more engaging and successful experience for students learning online, which should then result in more students completing their courses and achieving better academic results.

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

1. Introduction

Web-based Online learning is now a common part of education, appearing in various forms like massive open online courses (MOOCs) and virtual learning environments (VLEs), also known as learning management systems (LMSs). For instance, MOOCs give students the freedom to study whenever and wherever they choose [5]. These platforms introduce new ways to train students, change traditional study methods, attract students from all over the world, and have greatly helped higher education [6]. Despite these benefits, a major challenge within online learning systems, like VLEs, is the high number of students who do not complete their courses. About 78% of students do not finish their courses because there is a lack of face-to-face interaction, and dropout is a main problem that researchers have consistently tried to solve [7]. Therefore, a strong prediction system, which looks at what students do within an online course, could identify students who are not very engaged. This would then give instructors an extra tool to motivate and support learners [8].

2. Machine Learning for Prediction

Machine Learning (ML), a part of artificial intelligence, allows computer programs to automatically find complex patterns in features taken from existing data. This helps in making smart decisions about new data [9]. In this project, we will use five common ML models: K-Nearest Neighbours (K-NN), Support Vector Machines (SVM), Naive Bayes, Logistic Regression, and Decision Tree. These models are well-known because they are useful for many different prediction tasks. The data for this project will come from the Open University Learning Analytics Dataset (OULAD), which includes information about what students do and their personal details related to the courses.

3. Problem Statement

The main problem this project addresses is predicting when a student is likely to withdraw from a course. This prediction will consider how engaged a student is and will also work to find out which demographic features are strongly connected to students dropping out. Student engagement will be measured as a simple yes/no (binary) variable, following the method described by Mushtaq et al. [4]. This variable will be calculated using information like scores on course assignments, whether the student passed or failed the course, and how active the student was in the online learning environment (VLE).

Earlier studies that used OULAD data, such as those by Tomasevic et al. [10] and Hussain et al. [11], have looked at student engagement and predicted student dropout. However, these studies did not fully examine which demographic variables showed stronger connections with dropout. The main reason for this project is to provide useful insights into the importance of demographic variables by improving the student dropout model with the addition of student engagement data. Additionally, this project will include the timing of assessments to help make predictions at an early stage of the course. Adding this analysis aims to offer another way to improve models that predict student dropout. For this project, the prediction results will be grouped into three categories: 'Withdrawal' (0), 'Fail' (1), or 'Pass' (2).

4. Project Objectives

This project has two objectives:

1. To determine if there are corresponding relationships between student engagement and student dropout, and to identify any demographic variables linked to student dropout.
2. To establish a predictor capable of identifying students likely to withdraw at early stages of their course.

5. Methodology

The methodology will commence with an understanding of the OULAD dataset, which comprises seven interconnected CSV files, as illustrated in Figure 1. To calculate student engagement, the final results and scores will be extracted from the Assessment table, and learning behaviour from the Student VLE table. The calculation will follow the method presented in the work of Hussain et al. [11], defined as:

$$\text{Student Engagement} = \text{Excellent\_Score} \vee (\text{Final\_Result} \wedge \text{Active\_in\_VLE})$$

Where:

- Student Engagement indicates 0 (Low Engagement) or 1 (High Engagement).
- Excellent_Score signifies a score ≥ 90.
- Final_Result is 0 (Fail) or 1 (Pass).
- Active_in_VLE indicates VLE activity ≥ the average number of clicks on VLE activities.

Once the binary student engagement outcome is calculated and stored in a DataFrame, it will serve as a reference to enhance the Student Dropout Predictor (SDP), which will have values of 'Withdrawal' (0), 'Fail' (1), or 'Pass' (2). This will also facilitate new insights linked to additional features. An investigation will be conducted to determine if students who withdraw or fail were categorised as having low student engagement. The SDP values will be derived from the final_result column of the Student Info table. Initial correlations will be established with the demographic background of students from the same Student Info table, aiming to uncover potential causes for the final result. One potential insight to be explored is whether students who fail or drop out exhibit low engagement within the VLE platform and share similar demographic patterns. It is important to note that student engagement will be calculated considering the dataset collected over the full duration of the course, meaning after the course has concluded and the student's final status (pass, fail, or withdraw) is known.

An additional analysis will involve predicting dropout at early stages, for instance, after the completion of the first assessment. For this part, the selected ML classifier models—K-NN, SVM, Naive Bayes, Logistic Regression, and Decision Tree—will be employed. These models are currently proposed, but this selection may be revised or expanded in future discussions with the supervisor to achieve a more comprehensive understanding and comparison. Potential metrics for model comparison will include recall and accuracy of the predictor model.

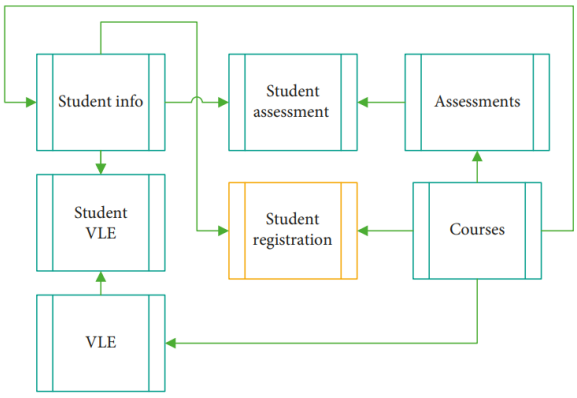

Figure 1: Relationship of the tables in the OULAD dataset, taken from Hussain et al. [11]

6. Bibliography/References:

[1] B. Means, M. Bakia, and R. Murphy, *Learning Online: What Research Tells Us About Whether, When and How*, Routledge, 2014, doi: https://doi.org/10.4324/9780203095959.

[2] E. R. Kahu, "Framing student engagement in higher education," *Studies in Higher Education*, vol. 38, no. 5, pp. 758–773, Jun. 2013, doi: https://doi.org/10.1080/03075079.2011.598505.

[3] J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open University Learning Analytics dataset", *Scientific Data*, vol. 4, p. 170171, Nov. 2017, doi: https://doi.org/10.1038/sdata.2017.171.

[4] M. Hussain, W. Zhu, W. Zhang, and S. M. R. Abidi, "Student Engagement Predictions in an e-Learning System and Their Impact on Student Course Assessment Scores," *Computational Intelligence and Neuroscience*, vol. 2018, pp. 1–21, Oct. 2018, doi: https://doi.org/10.1155/2018/6347186.

[5] S. Shishehchi, S. Y. Banihashem, N. A. M. Zin, and S. A. M. Noah, "Review of personalized recommendation techniques for learners in e-learning systems," in *2011 International Conference on Semantic Technology and Information Retrieval*, 2011, doi: https://doi.org/10.1109/stair.2011.5995802.

[6] I. Maiz Olazabalaga, "Research on MOOCs: Trends and Methodologies," *Porta Linguarum Revista Interuniversitaria de Didáctica de las Lenguas Extranjeras*, Sep. 2016, doi: https://doi.org/10.30827/digibug.54092.

[7] O. Simpson, "'22% -can we do better?'-The CWP Retention Literature Review'," Unpublished, 2010, doi: http://dx.doi.org/10.13140/RG.2.2.15450.16329

[8] O. Corrigan, A. F. Smeaton, M. Glynn, and S. Smyth, "Using educational analytics to improve test performance," in *Design for Teaching and Learning in a Networked World*, G. Conole, T. Klobučar, C. Rensing, J. Konert, and E. Lavoué, Eds., vol. 9307, *Lecture Notes in Computer Science*. Cham: Springer, 2015, doi: https://doi.org/10.1007/978-3-319-24258-3_4.

[9] J. H. Holland, *Adaptation in Natural and Artificial Systems*. The MIT Press, 1992. doi: https://doi.org/10.7551/mitpress/1090.001.0001.

[10] N. Tomasevic, N. Gvozdenovic, and S. Vranes, "An overview and comparison of supervised data mining techniques for student exam performance prediction," *Computers & Education*, vol. 143, p. 103676, Jan. 2020, doi: https://doi.org/10.1016/j.compedu.2019.103676.

[11] M. Hussain, W. Zhu, W. Zhang, and S. M. R. Abidi, "Student Engagement Predictions in an e-Learning System and Their Impact on Student Course Assessment Scores," *Computational Intelligence and Neuroscience*, vol. 2018, pp. 1–21, Oct. 2018, doi: https://doi.org/10.1155/2018/6347186.

Appendix:

1. Project Activities Timeline

Task	June				July				August			
	1	2	3	4	1	2	3	4	1	2	3	4
First meeting with JISC partner and Supervisor												
Familiarisation with OULAD dataset												
Review relevant literature												
First ideas to address the dropout problem												
Writing my project plan, aims and methodology												
Exploratory Data Analysis (EDA)												
Data pre-processing												
Development of code for objective 1 (demographic correlation analysis)												
Implementation of ML models for the objective 2 (Early Dropout Prediction)												
Initial ML Training												
Evaluation of the ML models												
Model refinement based on performance analysis and feedback from the Supervisor												
Final improvements												
Consolidation of findings and report structuring												
Report writing and submission preparation												
Preparation of final project presentation												

2. Project Risk Assessment

This section is about risks that could stop the project from succeeding or being accurate.

Incomplete or Inconsistent OULAD Dataset

The dataset might have missing information or errors, which could make the analysis and predictions unreliable. To mitigate this, thoroughly check and clean the data, and set up automated processes to handle missing values and inconsistencies.

Contradiction in "Early Stage Prediction" and "Full Course Engagement" Calculation

The project aims to predict early dropout, but the current plan defines "engagement" using data from the *entire* course, which isn't available early on. To mitigate this, it is key to redefine "engagement" to be measurable at early stages or use only early available features for the early prediction model.

Suboptimal Machine Learning Model Performance

The chosen machine learning models might not be accurate enough, especially in identifying students who will withdraw. To mitigate this, optimise the models or explore more advanced techniques if needed.

Bias and Misinterpretation of Predictive Outputs

The model might inadvertently show biases from the data, leading to unfair predictions or misinterpretations, especially with sensitive demographic factors. To mitigate this, actively check for and address biases, use tools to understand how the model makes decisions, and clarify that correlations in the data do not necessarily imply causation.

Insufficient Computational Resources:

Not having enough computing power (CPU, RAM, GPU) to process the large dataset or train complex machine learning models efficiently. To mitigate this, assess needs early, optimise code, and explore cloud computing.

Lack of Reproducibility:

Difficulty in replicating the project's results due to undocumented steps, inconsistent settings, or lack of version control for code and data. To mitigate this, use Git to document all data processing and model settings, and set fixed random seeds for experiments.

External Collaboration Dependencies (JISC Partner)

Delays or issues arising from the involvement of the JISC partner, such as communication breakdowns. To mitigate this, define the collaboration scope and maintain open communication.