Student Dropout Predictor

Student ID: 2605549
Date: 17th July 2025

Abstract :

Student dropout is a critical challenge for higher education institutions, affecting academic outcomes, institutional reputation, and financial sustainability. This study investigates early prediction of dropout risk using the Open University Learning Analytics (OULA) dataset, framing the task as a binary classification problem in which the negative class represents dropouts and the positive class includes continuing students. Four machine learning (ML) models: Logistic Regression (LR), Support Vector Machine (SVM), Multi-Layer Perceptron Classifier (MLP), and Random Forest (RF), were trained on processed, feature-engineered data segmented into Early (first 25% of the module completion), Midpoint (50%), Late (75%), and Full phases. Hyperparameters were optimised via GridSearch and five-fold cross-validation. Statistical analysis identified lower engagement, weaker assessment performance, certain demographic characteristics, and multiple prior module attempts as significant predictors of dropout.

SVM achieved the highest macro-average F1 scores overall (65% Early, 67% Midpoint), while LR demonstrated the best early-phase dropout detection (F1 = 51%, precision = 43%, recall = 62%). RF nearly matched LR's Early-phase results but surpassed them at the Midpoint (F1 = 60%, precision = 47%, recall = 82%). MLP showed the weakest performance across phases. Feature importance analysis highlighted weighted score, fail rate, late rate, studied credits, and days active as the most influential predictors, with demographic variables contributing minimally. These findings suggest that LR is best suited for early intervention, whereas RF is more effective by the Midpoint stage, offering actionable insights for targeted dropout prevention strategies.

Index Terms: Keyword A, Keyword B, Keyword C.

1 Introduction

Student dropout continues to be a pressing concern in higher education, affecting not only students' academic progress but also the financial stability and reputation of educational institutions. Early detection of students at risk of dropping out enables timely, targeted interventions that can improve retention rates and enhance student success. When students withdraw prematurely, it disrupts their learning journey, wastes valuable resources, increases administrative workload, and poses reputational challenges for universities [1]. Effectively addressing this issue requires identifying vulnerable students early enough to provide personalised support before disengagement becomes irreversible.

In recent years, ML has emerged as a powerful tool for predicting student dropout risk across diverse academic programmes. By leveraging various data sources, such as virtual learning environment (VLE) interactions, continuous assessment scores, and demographic information, ML models can uncover patterns that signal student vulnerability. These insights empower educators to deliver tailored support and interventions [2]. However, predicting dropout remains challenging due to the complex interplay of demographic, academic, and behavioural factors influencing student outcomes. This project aims to develop a reliable, scalable ML-based dropout prediction system that assists educators and administrators in proactively managing dropout risk.

The proposed solution integrates multiple data types, including demographic characteristics (e.g., age, education level, disability status, region), academic performance indicators (such as early assignment grades), and behavioural metrics (including VLE engagement patterns) captured at specific time points within a module. The prediction task is framed as a binary classification problem, where the model forecasts whether a student will drop out or continue their studies.

To reflect the progression of student engagement over time, the analysis is divided into four phases based on the proportion of module completion: Early (first 25%), Midpoint (50%), Late (75%), and Full (100%). For instance, in a 100-day module, the Early phase corresponds to the first 25 days, the Midpoint to 50 days, the Late phase to 75 days, and the Full phase to the entire duration. Since the outcome is binary, distinguishing dropouts from continuing students (including those who pass, fail, or achieve distinction), the problem is well-suited for classification modelling. Multiple ML algorithms will be developed and compared, with hyperparameter optimisation employed to maximise predictive performance.

The project implements a comprehensive ML pipeline starting with exploratory data analysis (EDA) to identify key trends and correlations in student demographics, academic records, and VLE engagement. This is followed by data preparation steps such as dataset merging, handling missing values and time series data, outlier detection, categorical feature encoding, feature engineering, and data normalisation. The dataset will be split into training and testing subsets to ensure unbiased evaluation. Hypothesis testing on the cleaned data will also be conducted to gain deeper insights that may influence model performance.

Four supervised learning models will be trained and evaluated: LR, SVM, RF, and neural networks exemplified by the MLP classifier. Hyperparameter tuning will be performed using GridSearch with five-fold cross-validation, optimising for the macro F1-score to prioritise accurate classification of minority classes such as dropouts. Additional evaluation metrics will include accuracy, precision, recall, and class-specific F1-scores, with particular emphasis on minimising false negatives to avoid overlooking at-risk students.

Finally, model interpretability will be enhanced through SHAP (SHapley Additive exPlanations) analysis to identify which features most strongly influence predictions and explore the reasons

behind their impact. The insights gained will support academic institutions in making data-driven decisions, optimising resource allocation, and designing effective interventions to improve student retention and academic achievement.

The paper is organised into six main sections. It begins with the Introduction, followed by the Background, which includes a literature review and a discussion of related studies that have utilised the OULA dataset. Within this section, three specific case studies are examined to illustrate different approaches to the student dropout prediction problem. Next, the Project Objectives and Hypotheses are outlined in detail, providing a clear framework for the study. The subsequent subsection, Model Selection and Justification, explains the rationale for choosing LR, SVM, MLP, and RF models, with detailed reasoning for each choice.

This is followed by the Data and Methods section, which begins with a comprehensive description of the OULA dataset. The preprocessing and feature engineering steps are then discussed, including data cleaning, creation of meaningful features, handling missing values, and temporal segmentation of the dataset into the four phases: Early, Midpoint, Late, and Full. Additional processing steps, such as train/test splitting, feature scaling, and encoding are also covered to prepare the data for ML models. Once preprocessing is complete, EDA is conducted to examine each feature, identify trends, and gain insights into the dataset. Following the EDA, hypothesis testing is conducted to assess engagement, assessment performance, demographic disparities, and re-enrolment patterns, with the results presented and interpreted. The next subsection covers the Implementation of ML Models, including model and pipeline setup, hyperparameter tuning using GridSearch, and the model evaluation strategy.

The Results and Analysis section presents findings for each ML model: LR, SVM, MLP, and RF, and evaluates their performance individually. Each model's evaluation is divided into three subsections. The first subsection details the hyperparameter tuning process, showing the optimal parameters identified for each phase via 5-fold cross-validation and their corresponding macro F1 scores. The second subsection provides a thorough analysis of performance metrics, including accuracy, precision, recall, and F1 scores for each class, with comparisons drawn between models and against related work and case studies from the Background section. The third subsection examines SHAP values for each model, focusing on the default positive class (class 1 or non-dropouts).

The Conclusion is divided into four parts: a summary of the hypothesis tests, insights gained from feature analysis, an evaluation of student dropout prediction models, identification of the best-performing models and highlighting the most influential features for the Early and Midpoint phases, and a discussion of project limitations and potential directions for future work. The paper concludes with the Appendix.

2 Background

2.1 Literature Review

Student dropout in higher education has become an increasingly pressing issue across the globe, with wide-reaching implications for individuals, educational institutions, and national education systems. The expansion of access to university-level education over the past few decades has undoubtedly increased opportunities for students and provided a greater supply of skilled graduates to the labour market. However, this expansion has also led to a significant rise in the number of students who do not complete their degrees, which in turn undermines the potential benefits of such educational access [3]. According to a 2019 report by the Organisation for Economic Co-operation and Development (OECD), dropout rates in tertiary education institutions are rising by an average of approximately 30% across many countries [4]. This trend has led to growing concerns about the sustainability and effectiveness of higher education systems, especially as universities face increasing pressure to maintain academic standards amid rising enrolment figures.

Within the United Kingdom, national statistics reflect this global trend. Data released by the Student Loans Company (SLC) revealed that the number of students who took out loans but failed to complete their degrees increased sharply from 32,491 in the academic year 2018–19 to 41,630 in 2022–23, a 28% increase in just five years [5]. A variety of reasons contribute to early withdrawal from higher education, but mental health difficulties have been particularly highlighted as a major factor. Financial strain, academic pressure, social isolation, and insufficient institutional support often exacerbate these challenges [5]. This underscores the importance of developing systems that allow for early identification of at-risk students and provide tailored support to help them remain engaged and succeed in their studies.

Several studies have investigated the effectiveness of targeted intervention strategies in mitigating dropout rates. For example, Foster et al. (2019) found that proactive academic engagement, including personalised feedback emails and increased tutor involvement, led to an 11% reduction in dropout rates within certain student cohorts [6]. However, the study also acknowledged limitations, such as the influence of curriculum design and assessment structure, that were not explored in depth. Furthermore, the research pointed to the unique advantages offered by online and distance learning platforms, where detailed student interaction data (such as logins, video views, and assignment submissions) can be leveraged to track engagement and provide timely interventions.

The application of predictive modelling, particularly through ML techniques, is increasingly seen as a valuable approach to addressing the dropout problem. Predictive models can analyse vast amounts of student data, including demographic information, academic history, and engagement metrics, to identify patterns associated with increased risk of dropout. These insights can enable academic institutions to implement early-warning systems that trigger timely and personalised support measures. For instance, models can flag students with declining activity levels or poor assessment performance, prompting automated alerts or human-led outreach. This supports not only student well-being but also institutional efficiency in managing tutoring resources, optimising learning environments, and adapting curricula based on observed patterns of disengagement.

Moreover, predictive systems can aid long-term planning by helping universities identify structural issues in specific courses or programmes that exhibit persistently high dropout rates. With accurate predictions, institutions can explore preventive redesigns of such courses, develop more inclusive and adaptive learning experiences, and ensure that teaching staff are appropriately allocated to support at-risk students. In this way, data-driven approaches do not merely serve as reactive tools but can actively inform strategic

educational reforms.

To summarise, addressing student dropout through predictive modelling holds significant potential for improving student outcomes, enhancing institutional performance, and ensuring the long-term viability of higher education. The literature supports the case for integrating ML into dropout prediction pipelines, particularly in contexts like distance learning, where rich digital traces can be harnessed to deliver more effective and equitable academic support.

2.2 Related Work

This section reviews three studies that utilise the OULA dataset for student performance prediction and discusses their findings.

2.2.1 Case Study 1: The first study, conducted by N. Tomasevic, N. Gvozdenovic, and S. Vranes (2019), aimed to perform a comprehensive comparison of state-of-the-art supervised ML techniques for predicting student exam performance. Specifically, their focus was on identifying students at "high risk" of failing and forecasting outcomes such as final exam scores. While their study included both classification and regression approaches, the focus here will be on the classification component. Notably, their classification task was structured around predicting pass/fail outcomes rather than student continuation or dropout.

An important detail is that their predictions were made using all available student data prior to the final exam, including VLE engagement, assessment results, and demographic information. The models applied included a broad range of ML techniques such as k-Nearest Neighbours (k-NN), SVMs with both linear and RBF kernels, Artificial Neural Networks (ANN), Decision Trees (DT), Naive Bayes (NB), and LR. The results of their classification experiments are summarised in Figure 1.

	D	E	P	D + E	D + P	E + P	D + E + P
k-NN (no weights)	0.6173	0.9146	0.94	0.8886	0.9406	0.939	0.9423
k-NN (distance weights)	0.6136	0.9124	0.9344	0.8906	0.938	0.9423	0.9453
SVM (linear kernel)	0.7202	0.7202	0.9288	0.934	0.9382	0.9608	0.9622
SVM (RBF kernel)	0.7202	0.942	0.9377	0.932	0.946	0.9565	0.9604
ANN (2x1)	0.7127	0.9505	0.9404	0.9487	0.947	0.9662	0.9645
Decision trees	0.6436	0.9454	0.9386	0.9473	0.9388	0.9507	0.9507
Naïve Bayes	0.567	0.8458	0.9207	0.8497	0.9236	0.9172	0.9135
Regularized logistic regression	0.6739	0.8896	0.9317	0.8927	0.9335	0.9336	0.9442

**green - optimal input data for given technique; orange – optimal technique for given input data; red – overall best*

Figure 1. F1 Score for final exam result prediction – classification (D: Demographics, E: Engagement, P: Performance data). Source: [7].

Their results indicate that the highest classification performance (in terms of F1 score) was generally achieved when all three data types (D + E + P) were used together. This trend was observed across various models including k-NN (with and without distance weighting), SVMs (both linear and RBF), DT, and LR. For NB, the best performance came from using demographic and performance data (D + P), while the combination of engagement and performance data (E + P) yielded the best results for ANN and DT models. In terms of overall performance, ANNs outperformed other models with an F1 score of approximately 96–97%, closely followed by SVMs (linear and RBF) at around 96%, DT at 95%, and both k-NN and LR at 94%. NB showed the weakest performance, with an F1 score of about 91% [7].

2.2.2 Case Study 2: In 2019, H. Waheed, S.-U. Hassan, Naif Radi Aljohani, J. Hardman, Salem Alelyani, and R. Nawaz conducted a study that examines the prediction of student performance. The researchers structured the problem into four distinct binary classification tasks. To manage class imbalance, the original labels 'pass' and 'distinction' were combined into a single 'pass' category. For identifying students at risk of failure, the binary classification used 'pass' (including 'distinction') and 'fail'. Similarly, in predicting withdrawal, students who withdrew were classified against those who either passed or earned distinction, resulting in another binary task. The classification to identify students likely to achieve distinction was also formulated against the 'pass' and 'fail' labels. In total, four binary classification tasks were designed based on student outcomes, with datasets comprising over 20,000 records for each.

The early prediction strategy used in this study closely resembles the temporal segmentation method applied in the current paper. To support early intervention, each module (spanning nine months) was divided into four quartiles. Student interaction data, referred to as clickstream data, was analysed for each quartile to compute time-specific features. These features aimed to identify which periods in a module had the most impact on student performance, providing critical insights for proactive academic support [8].

A deep ANN along with SVM and LR was trained on a set of handcrafted features derived from VLE clickstream activity. As shown in Figure 2, the ANN model achieved classification accuracy ranging from 84% to 93%, outperforming the baseline LR and SVM models. The LR models recorded accuracy between 79.82% and 85.60%, while SVMs performed slightly better with results between 79.95% and 89.14% [8]. The findings align with other studies, suggesting that the inclusion of historical and assessment-related data significantly improves model accuracy. Students who regularly accessed previous learning materials were more likely to perform well.

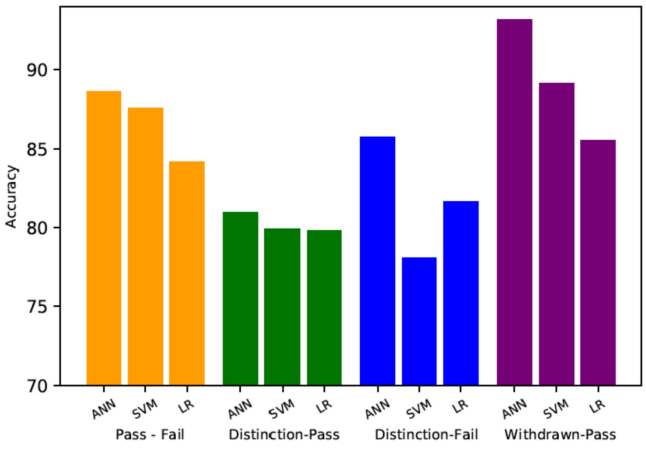

Figure 2. ML Model Accuracy Comparison By Label - Quaters 1-4. Source: [8].

This research contributes to the domain of early risk detection by identifying key indicators associated with poor academic outcomes and withdrawal. It mentions that that student activity after the start of the module had a more substantial influence on performance than engagement before the course. The ANN model

demonstrated strong results in detecting at-risk students, with a sensitivity of 69%, precision of 93%, and overall accuracy of 88%. For early withdrawal prediction, sensitivity reached 86%, precision 96%, and accuracy 93%. When identifying students likely to achieve distinction, the model achieved 74% sensitivity, 81% precision, and 85% accuracy [8]. The study highlights the effectiveness of deep learning in enabling timely institutional interventions and recommends using such predictive systems to inform student support strategies. Although no distinct pattern was found for students achieving distinction due to class imbalance, demographic and geographic factors were still shown to play a significant role in student outcomes.

2.2.3 Case Study 3: The final study, conducted by M. Adnan et al. (2021), it shares the primary objective of building a predictive model to identify challenges faced by at-risk students. The aim is to enable timely interventions from instructors, encouraging students to improve their academic engagement and performance. The study employed several ML algorithms, including RF, SVM, k-NN, Extra Trees Classifier (ETC), AdaBoost Classifier (ABC), and Gradient Boosting (GB).

Similar to other research, this study also applied temporal segmentation to the student data. To facilitate early prediction, the course duration was divided into five intervals: 20%, 40%, 60%, 80%, and 100% completion stages [9]. Predictive models were trained and evaluated at each stage using 10-fold cross-validation. Additionally, to simplify the classification task and improve model performance, the original labels were merged: 'Distinction' and 'Pass' were combined into a single 'Pass' class, while 'Withdrawn' and 'Fail' were grouped into a 'Fail' class [9]. Unlike the present work, which focuses on a continue-dropout prediction task, this study framed the problem as a binary pass-fail classification.

The study also highlighted the importance of assessment scores, VLE engagement (clickstream data), and time-sensitive features as key predictors in online learning settings [9]. Experimental results showed that RF consistently delivered the best performance. At 20% course completion, RF achieved an average precision, recall, F1-score, and accuracy of 79%. These metrics improved to 88% at 60% completion and reached their peak at 100% course completion, with precision of 92%, recall of 91%, F1-score of 91%, and accuracy of 91% [9].

Interestingly, the performance of the RF model was higher for the 'Fail' class compared to the 'Pass' class after feature engineering, likely due to class imbalance; the dataset contained 17,208 'Fail' instances versus 15,385 'Pass' instances [9]. Overall, the results demonstrate the efficacy of RF in making accurate early predictions about the performance of at-risk students.

2.2.4 Conclusion: The three reviewed studies collectively demonstrate the value of using ML methods with the OULA dataset to predict student outcomes in online education. Although each study adopts a different methodological approach, they converge on several important findings.

The first study does not incorporate temporal segmentation across the course duration. Instead, it focuses on evaluating how the inclusion or exclusion of different feature types, such as demographic, assessment, or VLE engagement data, affects model performance in predicting final exam outcomes. This offers important insight into the impact of various features on predictive accuracy. The results showed that ANN achieved the best performance, followed closely by SVM with both linear and RBF (Radial Basis Function) kernels, with F1 scores exceeding 94%. This is an impressive result.

The second study differs from the first by adopting multiple labelling schemes, such as pass-fail, distinction-pass, distinction-fail, and withdrawn-pass. It applies several ML models, with ANNs again performing best, achieving accuracy between 84-93%. SVM followed at 80-89%, and LR achieved around 80-86%. This study also implemented temporal segmentation by dividing each module into four quarters, an approach also used in this report.

The final study also adopts a pass-fail classification task but introduces a more granular segmentation of the course into six points: 20, 40, 60, 80, and 100% of module progress. This provides a broader view of performance trends and includes an especially early stage (20%), allowing early identification of at-risk students. This study found that RF performed best overall, achieving around 79% accuracy and F1 score at the 20% stage, and up to 91% when the full course data (100%) was used.

Collectively, these studies highlight three consistent insights. First, the integration of diverse data types, especially student engagement through VLE activity, assessment scores, and temporal features, is essential for improving predictive performance. All three studies show that richer datasets lead to better outcomes. Second, across the tested models, ANNs, SVMs, and RFs consistently outperform simpler methods such as LR and k-NN, although the performance advantage was relatively modest. ANNs achieved the highest F1 scores and accuracy in predicting final results and handling multiple classification tasks, while RFs showed strong and reliable performance even in early-stage prediction. Third, temporal segmentation proves to be an effective strategy for enabling earlier interventions. Both Waheed et al. and Adnan et al. used time-based partitioning successfully, allowing predictions at different points in a module's progression. This makes it possible for institutions to provide timely support before students fall too far behind.

Finally, the way student outcomes are framed, such as pass/fail versus continue/dropout, has a significant impact on model design and the insights that can be drawn. While the reviewed studies primarily focused on pass/fail predictions, this report explores the less frequently addressed task of continue-dropout classification, offering a fresh perspective in the field. In summary, the evidence strongly supports the use of ML for early identification of at-risk students, enabling timely educational interventions and ultimately improving student success in online learning contexts. The following sections of this paper will provide a detailed discussion of the project's objectives and hypotheses.

2.3 Project Objectives

The primary objective of this project is to build a model that can accurately identify students at risk of dropping out. This is intended to enable timely academic interventions by allowing institutions to provide appropriate support, tools, and resources to help students continue their studies. The task is framed as a classification problem, where the model predicts whether a student is likely to continue in the module. Moreover, Table 1 presents a concise summary of the key objectives of the student dropout project, along with brief descriptions for each.

Additionally, the OULA dataset used for the student dropout

prediction task requires thorough cleaning, preprocessing, and feature engineering to produce model-ready features. As these features directly influence model training and evaluation, careful preparation is essential. To guide this process, detailed EDA is performed to uncover patterns, assess the dataset's structure, and examine variable types.

Next, four ML models: LR, SVM, MLP and RF are selected for evaluation and comparison to identify the most effective approach for detecting at-risk students, especially during the early and midpoint phases of the module. The rationale behind the selection of these models is discussed in detail in Section 2.5. The analysis focuses on maximising performance for the dropout class, with an emphasis on recall and the correct identification of true dropout cases. The primary evaluation metric is the F1 macro score, which is well-suited for imbalanced datasets and ensures fair assessment across classes. Since the dropout class accounts for approximately 20% of student records, this metric helps to prioritise minority class performance. Additional metrics such as precision, recall, and overall accuracy are also considered to provide a complete performance overview.

In addition, four hypothesis tests are carried out to better understand key trends within the dataset and uncover meaningful insights about student behaviour. These hypotheses focus on the main data categories available: online portal engagement, assessment performance, and demographic characteristics, with a fourth hypothesis added to explore patterns related to student re-enrolment. Together, they cover the primary aspects of the dataset and provide a structured basis for further analysis. Each hypothesis is described in detail in the following sections.

A comprehensive version of the dataset, which includes full assessment records and VLE activity, is used as a benchmark. This allows for a comparison between models trained on partial data and those trained on complete data to evaluate the impact of data availability on prediction accuracy. The underlying assumption is that prediction accuracy increases as more data becomes available during the module, despite the decreasing likelihood of dropout over time. The goal is to identify students at risk as early as possible using relevant behavioural and academic indicators. The final model is intended to function as an early warning system during the first half of the module. By evaluating model performance across different phases, the study aims to determine when predictions are most reliable and to identify the key factors that signal potential dropout.

Feature importance is also analysed to determine which input variables most influence model predictions. In addition to building an accurate prediction system, this project also 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.

2.4 Hypothesis Tests

This section provides a detailed explanation of the four hypotheses, outlining each case along with the underlying assumptions and expected outcomes. It also describes the methods that will be used to test each hypothesis.

Table 1
Summary of Project Objectives

Objective	Description
Data Preparation	Clean, preprocess, and engineer features from the OULA dataset to ensure they are suitable for ML training and evaluation.
EDA	Conduct detailed EDA to understand data structure, variable types, and key trends.
Dropout Prediction	Develop and tune four models (LR, SVM, MLP and RF) to accurately identify students at risk of dropping out during the phases of the module.
Model Evaluation	Compare the four ML models to identify the most effective approach, focusing on recall and F1 macro score for the dropout class.
Hypothesis Testing	Perform four hypothesis tests covering portal engagement, assessment performance, demographics, and re-enrolment patterns.
Benchmarking	Use the complete dataset as a benchmark to evaluate the effect of data availability on model accuracy.
Phase-Based Analysis	Assess model performance across different module-presentation length phases to determine when predictions are most reliable.
Feature Importance	Identify which features most strongly influence predictions to understand dropout behaviour.
Educational Insight	Provide data-driven insights to help educators implement proactive interventions to improve retention.

2.4.1 Engagement Hypothesis: The first hypothesis examines whether low engagement with the VLE during the early stages of the module is associated with an increased likelihood of dropout. Specifically, it tests whether students who withdraw within the first 25% of the module duration demonstrate significantly lower VLE activity compared to those who remain enrolled, regardless of their final outcome (pass, fail, or distinction). The underlying assumption is that reduced early engagement may reflect a lack of motivation, interest, or access.

The null hypothesis states that there is no significant difference in early VLE engagement between students who drop out and those who continue the course. To assess this, the Mann-Whitney U test will be used to compare the distributions between the two independent groups, complemented by the p-value to determine statistical significance.

2.4.2 Assessment Performance Hypothesis: This hypothesis examines the relationship between early assessment performance and student dropout risk. The underlying assumption is that low scores in early assessments, particularly within the first 25% of the module duration, may undermine a student's confidence in achieving a satisfactory result, thereby increasing the likelihood of withdrawal.

The null hypothesis asserts that there is no significant difference in early assessment scores between students who withdrew and those who remained enrolled (including those who passed, failed, or achieved distinction). As in the engagement hypothesis, the Mann–Whitney U test and the p-value will be employed to compare the distribution of scores between the two groups.

2.4.3 Demographic Disparity Hypothesis:

This hypothesis covers the demographic student data and ex-

amines the relationship between certain groups (by age band, region, education, IMD band, and disability) that are overrepresented among dropouts. Factors like low income or IMD band region would increase the chances of dropout, including low education level and disability.

This test will be conducted using the p-value and the Chi-Square Test of Independence to determine whether dropout rates vary significantly across different categories of these variables, since all involved variables are categorical.

2.4.4 Re-enrolment Hypothesis: The final test focuses on re-enrolment, examining whether students with multiple previous attempts at the module are more likely to drop out again. The assumption is that a history of dropout may indicate persistent academic difficulties.

The null hypothesis states that there is no difference in the distribution of prior attempts between students who drop out and those who continue (distinction/pass/fail). This test will also use the Mann–Whitney U test along with the p-value for significance assessment.

The choice of hypothesis testing methods for each case will be further explained and justified in detail during the evaluation of results in Section 3.5 of Data Methods. The following subsection explains the reasoning behind the choice of models used for this classification task.

2.5 Rationale for Model Selection

As this is a multi-class classification task focused on identifying students at risk of dropout, four ML models were selected: LR, SVM, MLP, and RF. These models were trained on the processed training dataset and evaluated on the held-out test set using two key metrics: accuracy and macro-averaged F1 score. The macro-F1 score was chosen in particular to account for potential class imbalance, as it gives equal weight to each class by averaging F1 scores per class, thus offering a more balanced assessment of model performance.

Logistic Regression: LR was selected as a baseline model owing to its simplicity, interpretability, and efficient training process [10]. It performs well when the relationship between input features and target classes is approximately linear [11]. In the context of student dropout prediction, LR's probabilistic outputs allow stakeholders to define intervention thresholds, and its model coefficients provide valuable insights into how different features influence predictions. However, its primary limitation lies in its linear nature, which makes it less effective in capturing complex, non-linear relationships within the data. Given the multifaceted nature of student behaviour and engagement, LR may struggle to establish accurate decision boundaries, particularly compared to non-linear models such as neural networks or tree-based methods. Evaluating its performance against these more flexible models will highlight the extent to which linear assumptions limit predictive accuracy in this setting.

Support Vector Machines: SVMs are well-suited for high-dimensional and sparse datasets [12], which are common in educational data mining. Their ability to define complex decision boundaries via kernel functions, such as the RBF or polynomial kernels, allows them to model non-linear relationships in the data [12]. something will LR will not be able to do. Furthermore, SVMs handle class imbalance effectively by incorporating class weights into the optimisation objective, ensuring fairer representation of minority classes [12], which is essential for identifying at-risk students with relatively low prevalence in the dataset.

Multi-Layer Perceptron: The MLP classifier is a type of feed-forward artificial neural network. It was selected for its capacity to model non-linear feature interactions and uncover latent patterns not captured by linear models [13]. Given its flexibility, MLP is well-positioned to learn complex relationships between engagement, performance, and demographic variables. Although neural networks require extensive tuning, such as the number of layers, neurons, activation functions, and learning rates, their adaptability to heterogeneous data types can yield strong predictive performance, especially when properly regularised.

Random Forest: RF is a robust ensemble learning method that combines multiple decision trees to improve generalisation and reduce overfitting [14]. It is particularly effective at handling non-linear relationships, missing values, and mixed data types (numerical and categorical) [15]. RF also provides feature importance metrics, offering interpretability and insight into which student characteristics most strongly influence dropout risk. Its inherent bagging mechanism makes it less sensitive to noise and outliers[14], enhancing predictive stability across diverse student profiles.

These four models were selected to provide a diverse set of learning mechanisms and interpretability levels. LR and SVM offer strong baselines and clear interpretability, while MLP and RF capture non-linear patterns and feature interactions more effectively. Together, they offer a comprehensive evaluation framework for predicting student withdrawal risk, accommodating the trade-offs between transparency, complexity, and predictive performance. This multi-model approach enables robust comparisons and supports the identification of the most effective predictive strategy for early intervention.

2.6 Rationale for Using a Diverse Set of Models

Models from distinct algorithmic categories, including linear (LR), kernel-based (SVM), neural (MLP), and ensemble tree-based (RF), were selected to capture a broad range of learning behaviours and model capacities. This diversity allows:

- Comparison between linear and non-linear models in predicting student dropout.
- Analysis of model performance across different temporal subsets of the data.
- Evaluation of trade-offs between predictive accuracy and interpretability.
- Assessment of each model's ability to generalise to new, unseen data.

The main goal is to identify students at risk of withdrawal as accurately as possible. As such, higher priority is given to recall and precision for the dropout class, since correct identification is crucial for timely support and intervention.

Hyperparameter tuning for each model type is performed using GridSearchCV with 5-fold cross-validation to ensure robust and generalisable performance estimates. Models are evaluated on four distinct dataset phases: Early, Midpoint, Late, and Full, to observe how predictive capability changes throughout the course timeline. Early and midpoint phases are of particular interest,

as these represent periods where intervention is still actionable, while the late and full datasets serve as performance benchmarks.

3 Data Methods

This section outlines the key steps undertaken to develop the dropout prediction ML models. It begins with a description of the OULA dataset, including its structure and the type of data it contains, to provide context and insight. This is followed by an overview of the main features used in the models.

The section then proceeds to outline the steps involved in preparing the dataset for modelling, explaining how the models were initialised, tuned, executed, and evaluated. It begins with the Preprocessing and Feature Engineering subsection, which covers the merging of OULA dataset tables (assessments, VLE, and demographic data) into a unified table, the temporal segmentation of the dataset into four phases, and the subsequent cleaning, preprocessing, and feature engineering procedures.

The Processing Dataset stage involves handling missing values, removing duplicates, applying data imputation, creating train/test splits for analysis, and scaling and encoding features. The following subsection presents an in-depth EDA of the features. The section concludes with the results of hypothesis testing and a detailed explanation of model implementation, including pipeline configuration, hyperparameter tuning methods, and evaluation procedures.

3.1 Data Description

The OULA dataset consists of structured, tabular student data collected during the 2013 and 2014 academic years. It includes multiple interlinked tables, each capturing different aspects of student performance and engagement, connected through shared identifiers. It provides demographic information, module registration records, assessment outcomes, and summarised daily interactions with the VLE for each student, module, and presentation combination. In total, the dataset covers 22 module presentations and includes 32,593 students. These modules are offered in two academic sessions: February and October, identified as “B” and “J” respectively. The data is distributed across several CSV files, with each file representing a relational table. Additional information is available in [16].

Figure 3 illustrates the OULA dataset’s schema. The `studentInfo` table connects to `studentAssessment`, `studentVle`, and `studentRegistration` via the `id_student` key. The `courses` table links with the `assessments`, `studentRegistration`, `vle`, and `studentInfo` tables using the `code_module` and `code_presentation` fields. Lastly, the `assessments` table connects to `studentAssessment` through `id_assessment`, while the `vle` table is linked to `studentVle` via `id_site`.

After cleaning and preprocessing, the dataset comprises 27,984 students, with 19 selected features. These features are detailed in Table 2.

The next section outlines the feature engineering process used to derive aggregated variables such as `total_clicks`, `days_active_norm`, `weighted_score`, `banked_rate`, `late_rate`, and `fail_rate`.

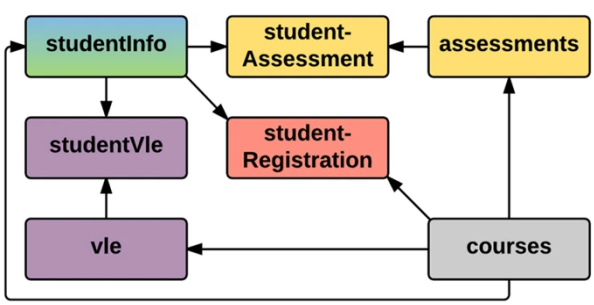

Figure 3. Student dropout dataset structure. Source: [16].

3.2 Preprocessing and Feature Engineering

The initial stage involved cleaning and preparing the dataset. This process included handling missing values, eliminating duplicate entries, and standardising data formats. An exploratory analysis was conducted to examine the distribution of key features such as gender, age group, and final result, as well as to detect any class imbalance. These observations informed subsequent steps such as scaling, encoding, and transformation, ensuring that no feature disproportionately influenced the model. Following the cleaning process, feature engineering was carried out by merging relevant datasets and combining variables into a structured format appropriate for modelling. The OULA dataset originally consisted of seven tables, which were first combined into three primary tables: VLE, assessments, and student information. These were subsequently merged into a single unified dataset for modelling purposes. In the following sections, we outline the merging process and feature engineering applied to each of the three main tables.

3.2.1 VLE Tables: The VLE data is split across two tables: `vle`, which contains activity definitions, and `studentVle`, which records individual student interactions. This includes the activity type, date of access, and the number of clicks associated with that activity.

An inner merge is suitable when combining these tables, as activities without student interaction provide no actionable insight. Furthermore, the "week_from" and "week_to" columns are dropped due to over 82% missing values and limited analytical value. To simplify the dataset, the activity_type column is also excluded; retaining it would require encoding categorical values, leading to a sparse dataset. Table 3 shows an example of the merged VLE data:

Each student has multiple VLE records distributed over time, which need to be aggregated into a single row per student to enable integration with assessment and demographic data for ML tasks. Two aggregate features are computed: the total number of clicks across the entire module and the proportion of days the student was active, expressed as a value between 0 and 1. Assuming a module duration of 10 days for AAA and 5 days for BBB, the resulting engineered features are presented in Table 4. We now proceed to the assessment tables.

3.2.2 Assessment Tables: Similar to the VLE tables, the assessment-related tables include multiple entries for various assessments and exams submitted by each student. The `assessments` table lists all assessments associated with each

Table 2
Description of dataset features used for student dropout prediction.

Feature	Description
`Code Module`	A categorical variable and an abbreviated code identifying the module.
`Code Presentation`	Also a categorical variable and an abbreviated code for the specific presentation of the module (e.g., "B" for February, "J" for October).
`Date Registration`	A numerical feature for the student's registration date relative to the module start (in days).
`Module Presentation Length`	Numerical feature for the module presentation duration in days.
`Gender`	Student's gender: Male = 1, Female = 0.
`Region`	A categorical variable for geographic region where the student resided during the module.
`Highest Education`	A categorical feature for the highest qualification held by the student at the time of enrolment.
`IMD Band`	A categorical field for the socio-economic band based on the Index of Multiple Deprivation (IMD) of the student's residence.
`Age Band`	Student's age group.
`Num. of Prev. Attempts`	Number of times the student has previously attempted the same module.
`Studied Credits`	Total credits of all modules the student is enrolled in concurrently.
`Disability`	Indicates whether the student has declared a disability.
`Final Result`	Final outcome in the module: Distinction/Pass/Fail = 1, Withdrawal = 0. (Target variable)
`Total Clicks`	Total number of interactions (clicks) with the VLE within the selected timeframe.
`Days Active Norm.`	Total number of days the student was active on the VLE, normalised for the selected timeframe.
`Weighted Score`	Student's average weighted score for assessments submitted during the timeframe.
`Banked Rate`	Proportion of assessment scores carried over from previous module presentations.
`Late Rate`	Proportion of assessments or exams submitted after the deadline.
`Fail Rate`	Proportion of assessments or exams the student failed.

module and presentation, along with details such as the assessment type, scheduled date, and assigned weight. The `studentAssessment` table records the submissions made by students, including whether the score was banked, the submission date, and the score achieved.

These features are important for calculating metrics such as fail rate, banked rate, late submission rate, and average weighted score for each student. It is important to note that missing assessment records are interpreted as assessments not submitted by the student, and are therefore treated as failures [16]. Additionally, final exam results are often missing because they are processed separately for final grading at the end of the module, which is explicitly noted in the dataset documentation [16].

Before feature engineering can be applied, the two assessment tables must first be merged. Once combined, feature extraction can proceed. To illustrate this, Table 5 presents an example showing how a student's assessment activity is represented in the merged time-series format.

Table 3
Sample merged student VLE records.

Code Module	Student ID	Activity Type	Date	Sum Clicks
AAA	0	homepage	1	4
AAA	0	forumng	7	11
BBB	1	oucontent	2	3
BBB	1	homepage	3	2
BBB	1	subpage	4	13

Table 4
Aggregated VLE features per student after feature engineering.

Code Module	Student ID	Total Clicks	Days Active
AAA	0	15	0.2
BBB	1	18	0.6

Table 5
Sample student assessment records showing assessment type, weight, score, and submission timing.

Code Module	Stu. ID	Assess. ID	Assess. Type	Weight	Score	Date Due	Date Subm.	Final Result
AAA	0	0	TMA	20	35	5	6	Pass
AAA	0	1	TMA	20	60	20	19	Pass
AAA	0	2	CMA	20	75	50	55	Pass
AAA	0	3	Exam	40	85	100	100	Pass

In this example, `final_result` is the target variable, representing whether a student passed, failed, withdrew from, or achieved distinction in the module. According to the data specifications, a score < 40 is a fail [16]. Because the data is time-dependent, we need to aggregate it to make it usable for modelling. For instance, we can summarise a student's performance as shown in Table 6:

As illustrated in Table 6, the weighted score is calculated by multiplying each assessment score by its respective weight and then dividing the sum by the total weight (which is 100 in this case). The late rate is determined by comparing the submission date with the due date; any submission made after the due date is considered late. The late rate represents the fraction of late submissions relative to the total number of submissions, ranging from 0 to 1. Lastly, the fail rate is computed by counting the number of assessments with scores below 40 (including missing scores, which are treated as failures or non-submissions) divided by the total assessments submitted. For the example in Table 5, the student failed one assessment (assessment ID 0 with a score of 35), resulting in a fail rate of 0.25 as shown in Table 6.

3.2.3 Student Information Tables: The `courses`, `studentRegistration`, and `studentInfo` tables are merged into a single dataframe using an inner join, as these tables do not contain time series data and share common identifiers. The `courses` table provides details such as the module presentation and its duration, while the `studentRegistration` table includes information on students' registration dates, the modules they enrolled in, and whether they later withdrew. The date unregistration column is excluded from the dataset because it would not be available at the point of early prediction. Including it would compromise the integrity of the modelling task, as it could

Table 6
Example of a feature-engineered student record with aggregated performance metrics.

Code Module	Student ID	Weighted Score	Late Rate	Fail Rate	Final Result
AAA	0	68	0.50	0.25	Pass

lead the model to learn direct withdrawal signals rather than underlying predictive patterns.

Following the merge of the student information dataframe, three key datasets are obtained: VLE activity data, assessment data, and student information. These are subsequently combined using an inner join into a single, comprehensive dataset, as shown in Figure 4.

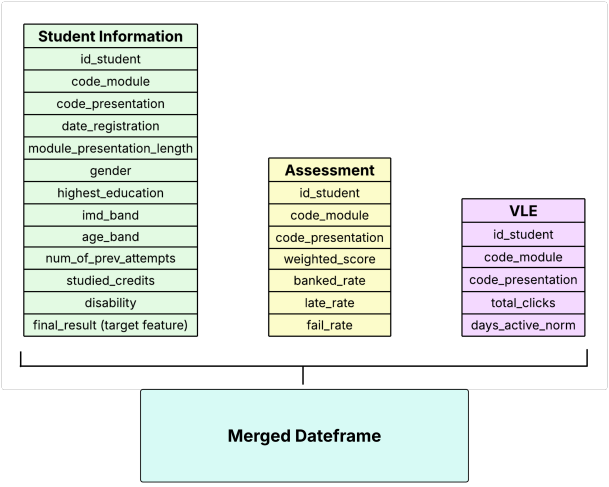

Figure 4. Final integrated dataset after merging VLE, assessment, and student information tables.

3.2.4 Time-Based Feature Limiting for Early Prediction: In real-world applications, full course histories are rarely available when attempting to predict student dropout early. Making predictions at the end of a course is typically too late for effective intervention. To address this, a timeline-based feature limitation approach is introduced. This involves restricting the available data to a specific point in time (e.g., the midpoint of a module) to emulate early-stage prediction. For instance, if the total module duration is 100 days, the dataset can be truncated to include only events occurring up to day 50. The aggregated data based on this restriction is shown in Table 7:

Table 7
Aggregated student performance metrics derived from data available up to the module midpoint.

Code Module	Student ID	Weighted Score	Late Rate	Fail Rate	Final Result
AAA	0	57	0.67	0.33	Pass

Table 7 illustrates student performance metrics based only on data available up to the midpoint of the module, in contrast to the full-course view over 100 days (Table 6). At the 50% mark, higher late and fail rates are observed, largely because the final exam has not yet occurred and is therefore not included in the data. This table represents the type of truncated data used for training ML models on thousands of student records to uncover patterns associated with dropout. Such intermediate datasets often reflect lower performance due to incomplete assessment coverage and can highlight early warning signs, such as high failure rates, frequent late submissions, or poor early engagement. These models are thus trained to recognise early behavioural indicators that correlate with eventual dropout risk.

3.3 Processing Dataset

Following feature engineering and the merging of relevant tables, the final merged dataset undergoes final cleaning to prepare it for predictive modelling. This stage includes removing non-informative attributes, addressing missing values, and verifying that the dataset aligns with the intended prediction goals. For example, the date unregistration feature is excluded, as it cannot be known during early-stage prediction. If used, it would introduce data leakage, providing the model with information that would not be available at prediction time, leading to overly optimistic and misleading performance.

3.3.1 Imputation Strategy for Missing Values: Missing values in various features are handled through context-aware imputation. The following Table 8 summarises the imputation methods and the rationale behind them:

Table 8
Imputation strategies for selected features, based on student engagement and demographic relevance.

Feature	Imputation Method	Justification
`Date Registration`	Median replacement	Most missing values are for withdrawn students. The median provides a reasonable neutral estimate.
`Total Clicks`	Replace with 0	Students who fail or withdraw often have no VLE interaction. Zero engagement is a logical substitute.
`Banked Rate`	Replace with 0	Missing values typically indicate withdrawn or failing students. The feature has low coverage and impact, so 0 is a practical default.
`Weighted Score`	Replace with 0	Non-submission is represented by missing values. A 0 score reflects no submission, as per the dataset specification.
`Late Rate`	Replace with 1	No submission implies full lateness. A value of 1 reflects total disengagement with deadlines.
`Fail Rate`	Replace with 1	Missing values imply assessment failure. A value of 1 reflects complete non-completion.
`IMD Band`	Bayesian Ridge Regression	IMD is a critical socio-demographic feature. Missing values are predicted using age, education, and region for contextual accuracy.

A total of 4,609 students who withdrew either before the module commenced or within the first 19 days were excluded. This

cutoff was chosen because most modules start assessments after day 19, and these students generally exhibit no VLE activity or assessment records. Including them would introduce noise without contributing valuable information. Furthermore, early withdrawal data would not be available when predicting dropout during the early or mid-phase of the module, so retaining such records would reduce the model's realism and practical applicability.

3.3.2 Temporal Segmentation of Dataset for Dropout Prediction: An automated data processing script was developed that allows the user to specify the portion of the module timeline to include. Using this, four datasets corresponding to different time points within the module were created for training, testing, and evaluation: Early, Midpoint, Late, and Full.

As shown in Figure 5, the Early dataset includes student data up to the first 25% of the module's duration. For instance, in a 100-day module, this would cover only the first 25 days. Data beyond this point is excluded to prevent leakage and compel the models to identify early indicators of potential dropout. The Midpoint dataset covers 50% of the module duration, Late covers 75%, and Full contains the complete data for the entire module. While the Late and Full datasets will be used primarily for EDA and serve as benchmarks, the main emphasis is placed on enhancing dropout prediction performance using the Early and Midpoint datasets.

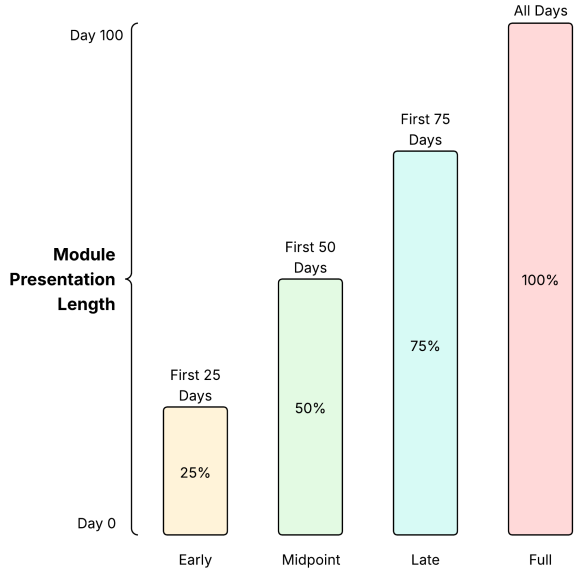

Figure 5. Diagram illustrating the temporal splits of the dataset across the module timeline.

The choice of 25%, 50%, 75%, and 100% time points reflects a balanced progression through the module, providing meaningful intervals for prediction. Selecting a very early cutoff, such as 10%, would make dropout prediction challenging due to insufficient VLE interaction and assessment data. At such an early stage, the model would have to rely heavily on demographic information alone, which is less ideal. By using 25%, 50%, 75%, and 100%, the model has more opportunity to learn from a combination of demographic data, VLE activity, and assessment performance. The 25% mark is particularly suitable for early prediction since, by this point, most modules have already involved some assessments and VLE activities, offering a solid foundation for detecting early signs of potential dropout.

3.3.3 Train/Test Split: To prevent data leakage and ensure unbiased evaluation, the dataset is split into training and testing sets before any detailed EDA. An 80/20 split is applied, with 80% of the data used for training (including all EDA) and 20% reserved as a test set for final model evaluation. The split is stratified by course module to maintain proportional representation of each module in both subsets.

3.3.4 Exploratory Data Analysis: Following the split, comprehensive EDA is conducted exclusively on the training set. This process uncovers insights about the engineered features, examines value distributions, and evaluates feature relevance. Relationships between features are explored, outliers identified, and correlation matrices generated to assess associations among features and with the target variable. The findings guide decisions on scaling continuous variables, encoding categorical features, and removing irrelevant or redundant attributes such as "id_student" that do not contribute to predictive modelling.

3.3.5 Scaling and Encoding the Dataset: As detailed in Table 9, one-hot encoding is applied to categorical variables such as gender and disability, transforming each category into its binary feature. This step is essential since models like LR and SVM require numerical inputs and cannot handle raw categorical strings.

Table 9
Feature preprocessing methods applied before model training

Feature	Scaling/Encoding Method
Code Module	One-Hot Encoding
Code Presentation	One-Hot Encoding
Date Registration	Standard Scaler
Module Presentation Length	Standard Scaler
Gender	One-Hot Encoding
Region	One-Hot Encoding
Highest Education	Standard Scaler
Age Band	Standard Scaler
Num of Prev. Attempts	Standard Scaler
Studied Credits	Standard Scaler
Days Active Norm.	Standard Scaler
Disability	One-Hot Encoding
Total Clicks	Standard Scaler
Weighted Score	Standard Scaler
Banked Rate	Standard Scaler
Late Rate	Standard Scaler
Fail Rate	Standard Scaler
Final Result (target)	Distinction/Pass/Fail=1; Withdrawn=0

Continuous numerical features are normalised using standard scaling, which adjusts values to have a mean of 0 and a standard deviation of 1 [17]. This is particularly important for models sensitive to feature magnitudes, including MLP, LR, and SVM, which rely on gradient-based optimisation. For instance, studied credits may vary between 30 and 600, whereas "num_of_prev_attempts"

ranges from 0 to 6. Without scaling, features with larger ranges could disproportionately influence the model [18].

Ordinal categorical variables such as age band, which have a meaningful order but are not inherently numeric, are first label-encoded (e.g., "0–35" → 0, "35–55" → 1, "55<=" → 2) and subsequently scaled using StandardScaler to maintain consistency with other numerical features.

The target variable "final_result" is converted into a binary outcome: students who passed, failed, or obtained distinction are labelled as 1 (indicating module completion), while those who withdrew are labelled as 0, aligning to predict dropout versus continuation.

3.4 Exploratory Data Analysis Findings

Before implementing ML models, it is essential to gain an initial understanding of the dataset by exploring each feature individually.

3.4.1 Date registration: As shown in Figure 6, the majority of students registered between 25 and 100 days before the module start date, suggesting that most students enrolled promptly. A smaller proportion registered after the module had begun, which appears to be relatively rare. To further examine unusual cases, we use a box-and-whisker plot.

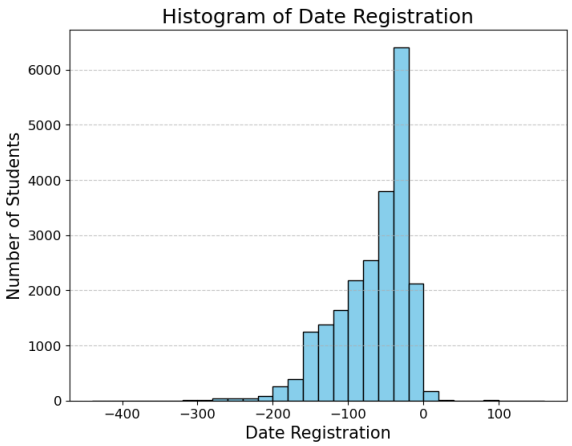

Figure 6. Histogram of Date Registration.

In Figure 7, we observe that some students registered exceptionally early, up to nearly a year in advance. Conversely, a few students enrolled very late, as much as 130 days after the course had started, by which time a considerable portion of the content would have already been completed.

3.4.2 Code Module and Presentation: Figure 8 shows the distribution of students across different modules. The highest enrolments are in modules FFF and BBB, each comprising approximately 24% of the training dataset. On the other hand, module AAA has the fewest students, representing only about 2.5% of the sample. It's also worth noting that the 2014J presentation accounts for the largest share of students (around 33%), while the smallest cohort comes from 2013B (about 15%). This indicates some imbalance in the representation across modules and presentations.

Additional insights are revealed in Figure 9, which presents the

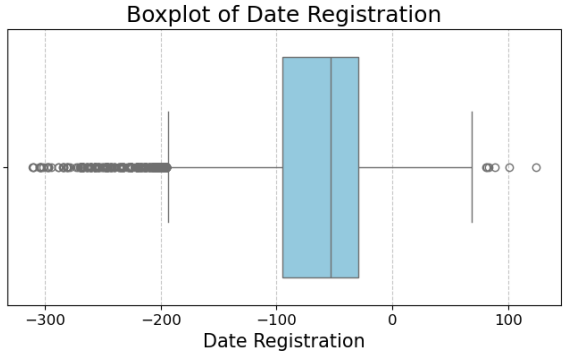

Figure 7. Boxplot of Date Registration.

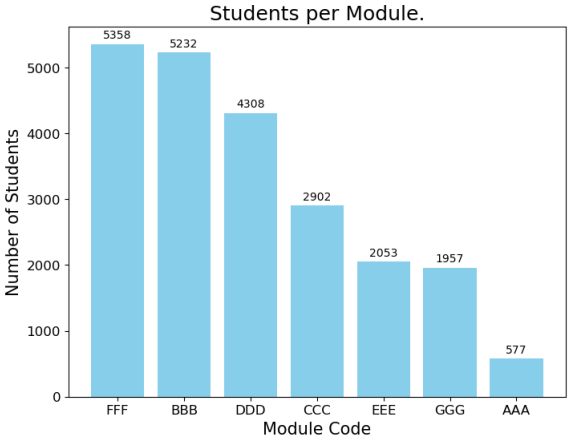

Figure 8. Students Per Module

distribution of final outcomes by module. Module GGG stands out with the highest distinction rate at 16.1%, but it also has the highest failure rate (29.1%), indicating a polarising pattern; students tend to either excel or struggle significantly. Module CCC shows concerning trends, with the highest dropout rate (32.6%) and the lowest pass rate (32.4%), suggesting serious retention challenges. In contrast, module AAA appears to be the most consistent and supportive, with the highest pass rate (66.4%) and relatively low dropout (14.4%) and failure (12.7%) rates, although it only accounts for about 2% of the training data.

Furthermore, modules CCC and DDD both have over half of their students either dropping out or failing, highlighting potential issues in their design, support mechanisms, or assessment structure. Module EEE shows a more balanced performance profile, with the second-highest distinction rate (13.3%) and a solid pass rate (51.1%), making it one of the strongest modules overall. These variations suggest that the structure and delivery of individual modules significantly affect student outcomes, and targeted interventions may be necessary for those with high failure or dropout rates.

3.4.3 Gender: The gender distribution is relatively balanced, with approximately 55% of students being female and 45% male. Moreover, it was found that female students have a slightly lower dropout rate (18.5%) compared to males (20.7%), along with a

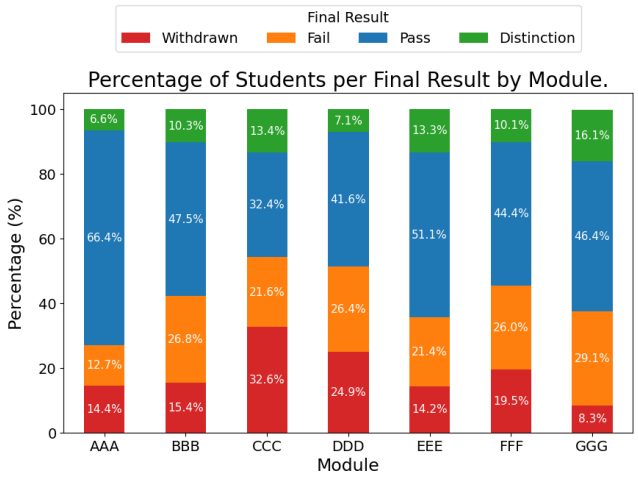

Figure 9. Percentage of Students per Final Result by Module.

slightly higher pass rate (46% vs 43%). The failure and distinction rates are also very close, with females at 24.4% and 11%, and males at 25.8% and 10.5%, respectively. These small differences indicate that gender does not have a significant effect on academic outcomes in this dataset.

3.4.4 Disability: Fewer than 10% of students in the dataset are recorded as having a disability, amounting to approximately 2,140 individuals. As shown in Figure 10, there is a noticeable gap in academic outcomes between students with and without disabilities. Those without disabilities have a lower dropout rate (18.8%) and a higher pass rate (45.2%) compared to students with disabilities, who exhibit a higher dropout rate (28.3%) and a lower pass rate (36.7%). The distinction rate is also slightly higher among students without disabilities (11%) than those with disabilities (8.4%). Failure rates are relatively similar, at 25.1% for non-disabled students and 26.6% for disabled students. These differences indicate that students with disabilities may encounter additional barriers that affect both their academic success and likelihood of course completion.

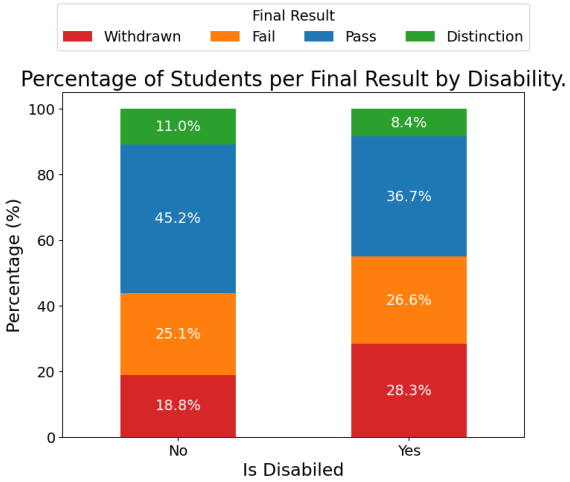

Figure 10. Percentage of Students per Final Result by Disability.

3.4.5 Region: As for the regions where the students are from in the dataset in Figure 11, Scotland has the highest proportion of students, making up around 11% of the dataset, while Ireland has the smallest share with just 901 students, representing around 4%. Other regions, like London, also have notable representation, contributing approximately 9% of the total. Next, we examine the dropout rates across different regions.

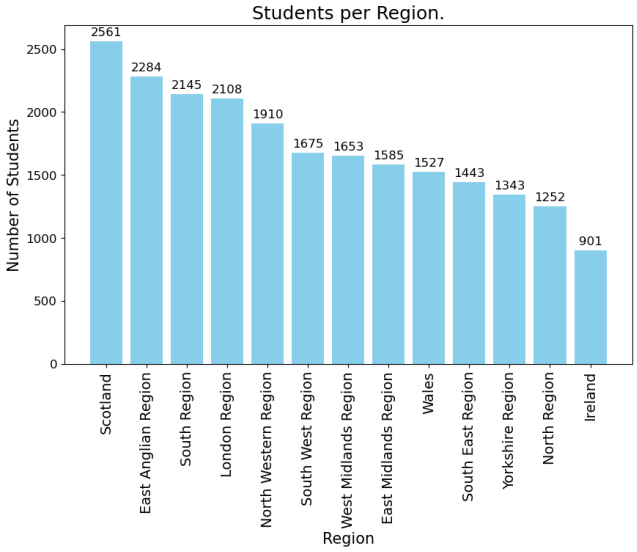

Figure 11. Students per Region.

According to Figure 12, the highest withdrawal rates are observed in the West Midlands (21.6%), East Midlands (21.4%), and North West (21.4%), suggesting that students in these regions are more likely to discontinue their studies. On the other hand, the lowest dropout rates occur in the South East (18.1%), East Anglian (17.8%), and Ireland (18.3%), indicating comparatively better student retention.

Regarding failure rates, students in Wales (32.2%), North West (29.6%), and London (28.9%) are most affected, implying they are more likely to complete the course but underperform academically. In contrast, the South East (19.9%) and South (20.5%) show the lowest failure rates, which may reflect stronger academic support or better overall student performance.

Pass rates are highest in Ireland (49.4%), South East (48.2%), and South (47.7%), highlighting stronger academic outcomes in these areas. Lower pass rates are seen in the North West (39.8%) and London (41.8%), potentially linked to the higher dropout and failure rates mentioned earlier.

Distinction rates are most prominent in the North Region (14.7%), South East (13.8%), and South (13%), suggesting higher levels of academic excellence. The lowest distinction rates are found in Ireland (8.4%), Wales (8.5%), and West Midlands (8.4%).

3.4.6 Highest Education: Most students in the dataset possess some form of educational qualification. The largest group comprises those with A Level or equivalent qualifications, making up 43% of the dataset, followed closely by students with qualifications below A Level, who account for approximately 40%. The smallest groups are postgraduates and those with no formal qualifications, each representing just 1% of the total.

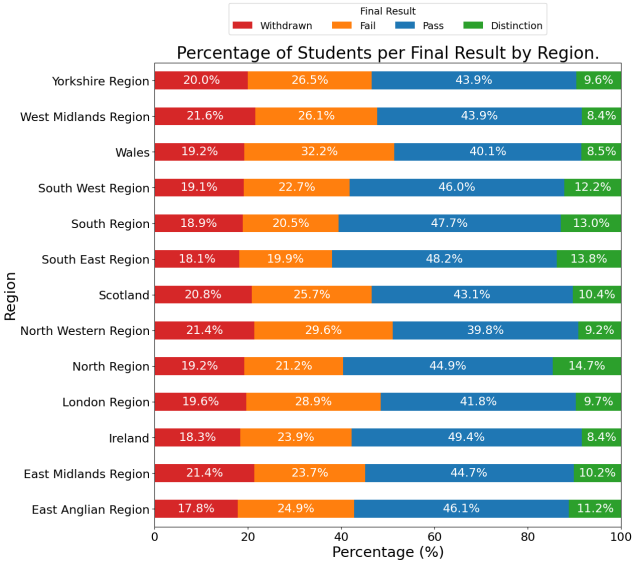

Figure 12. Percentage of Students per Final Result by Region.

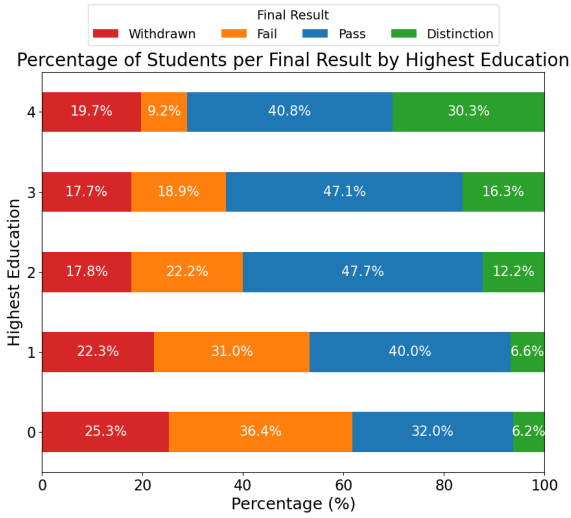

Figure 13. Percentage of Students per Final Result by Highest Education. 0 = No Formal Qualifications; 1 = Qualification Lower than A Level; 2 = A Level or Equivalent Qualification; 3 = Higher Education Qualification; 4 = Postgraduate Qualification.

As illustrated in Figure 13, students without any formal qualifications (0) show the highest withdrawal rate at 25.3% and the highest failure rate at 36.4%. Their pass rate is the lowest at 32%, and only 6.2% achieve a distinction. Students with qualifications lower than A Level (1) show some improvement, with a withdrawal rate of 22.3%, failure rate of 31%, a higher pass rate of 40%, and a slight increase in distinction rate to 6.6%.

Students with A Level or equivalent qualifications (2) perform better overall, with lower withdrawal (17.8%) and failure (22.2%) rates. Their pass rate increases to 47.7%, and the distinction rate rises to 12.2%. Those with higher education qualifications (3) continue this trend, with slightly lower withdrawal (17.7%) and failure rates (18.9%). While their pass rate is slightly lower at 47.1%, their distinction rate improves to 16.3%.

Finally, students holding postgraduate qualifications (4) perform best in certain areas. Although their withdrawal rate is slightly higher at 19.7% and their pass rate is lower at 40.8%, they have the lowest failure rate (9.2%) and the highest distinction rate at 30.3%. Overall, the analysis indicates that higher levels of prior education are generally associated with better academic performance and reduced dropout rates.

3.4.7 IMD Band: The IMD band provides insight into students' socio-economic backgrounds. As illustrated in Figure 14, the highest concentration of students falls within IMD band 3 (30–40%), representing approximately 12% of the training data. Conversely, IMD band 9 (90–100%) is the least represented, accounting for around 8%. Notably, about 10% of students originate from the most deprived areas, classified under the 0–10% IMD band.

Figure 15 further reveals a strong association between deprivation level and academic outcomes. Students from the most deprived areas (IMD 0) exhibit the one of the highest rates of withdrawal (23%) and failure (34.1%), along with the lowest rates of passing (36.9%) and achieving distinction (6%). In contrast, students from the least deprived areas (IMD 9.0) show significantly better results, with the lowest withdrawal (16.4%) and failure rates

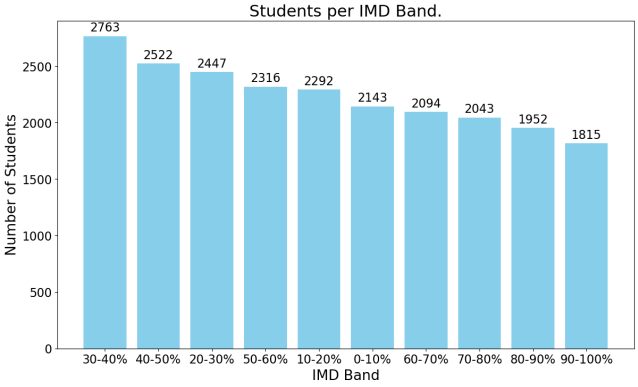

Figure 14. Students per IMD Band.

(18.4%), and the highest pass (49.4%) and distinction rates (15.8%).

In summary, there is a distinct pattern indicating that students from less deprived backgrounds tend to achieve better academic results, characterised by lower dropout and failure rates and higher levels of success.

3.4.8 Age Band: The dataset is predominantly composed of students aged between 0–35, who represent approximately 70% of the total. This is followed by those aged 35–55, making up around 30%. Students aged 55 and above form the smallest group, comprising less than 1% (roughly 150 students).

Figure 16 illustrates the relationship between age and academic performance. Students in the youngest group exhibit the highest withdrawal (20.1%) and failure rates (26.9%), alongside comparatively lower pass (45.5%) and distinction rates (9.6%). Outcomes improve for those aged 35–55, who have a slightly reduced withdrawal rate (18.9%) and failure rate (21.6%), while their pass and distinction rates rise to 46.2% and 13.3%, respectively. The best results are observed in the oldest group, with the lowest withdrawal

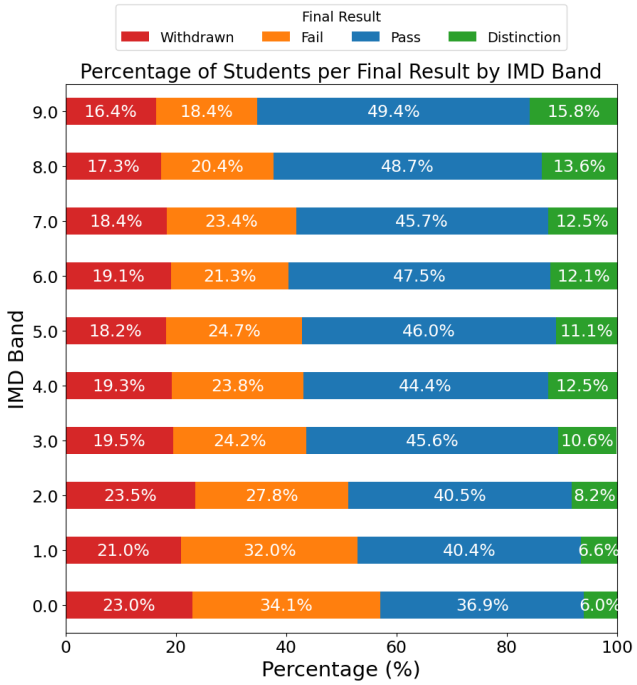

Figure 15. Percentage of Students per Final Result by IMD Band.

(18%) and failure rates (15.3%), and the highest pass (47.3%) and distinction rates (19.3%). However, this group represents a very small portion of the training data. Overall, the data suggests that academic performance tends to improve with age, with older students showing higher success rates and fewer dropouts.

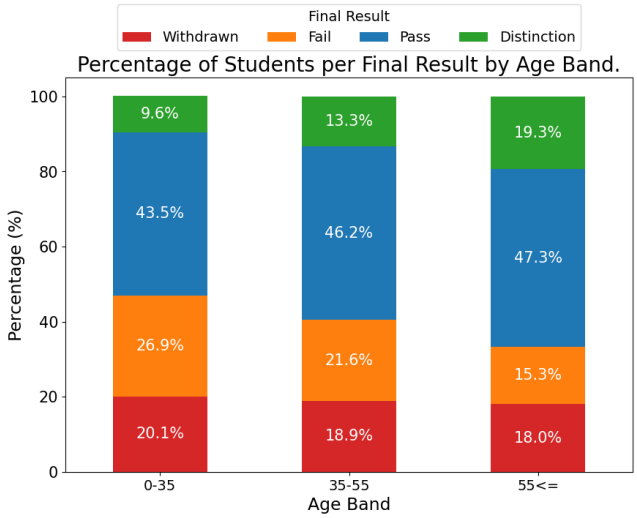

Figure 16. Percentage of Students per Final Result by Age Band.

3.4.9 Number of Previous Module Attempts: Most students, approximately 87%, are enrolled in a module for the first time. Around 10% have previously attempted the same module once, while multiple reattempts are uncommon. Only 11 students have taken a module five times, and just 3 have attempted it six times. Furthermore, students taking a module for the first time tend to perform best, with a withdrawal rate of 19.2%, a failure rate of 23.4%, a pass rate of 45.8%, and a distinction rate of 11.6%. Performance tends to decline with each additional attempt. Students retaking the module once or twice show higher withdrawal rates (23.1% and 25.5%), increased failure rates (36.6% and 38.9%), and lower pass rates (35.3% and 31.1%). Those attempting a module three or more times experience the poorest outcomes, with withdrawal rates rising to 33%, failure rates reaching 54%, and distinction rates dropping significantly or disappearing altogether.

In summary, there is a clear pattern indicating that students who retake modules multiple times are more likely to struggle, with higher dropout and failure rates and reduced academic success.

3.4.10 Studied Credits: In Figure 17, most students have around 60 credits for their module. While credit values extend up between 200 and 650, such high values are extremely rare.

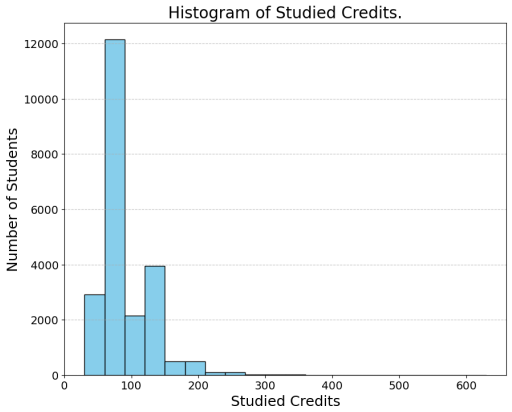

Figure 17. Histogram of Studied Credits.

When we check a box and whisker plot of the number of previous module attempts feature, the majority of students have studied between 60 and 120 credits. Credit values above 140 are uncommon and mostly considered outliers, as seen in Figure 18.

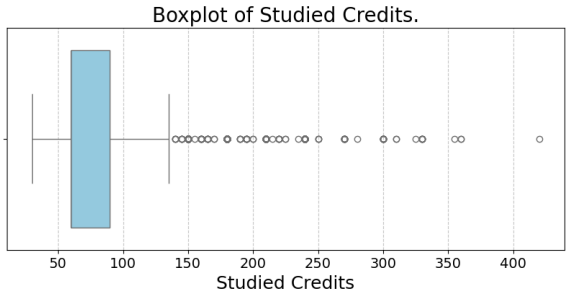

Figure 18. Boxplot of Studied Credits.

Most students with over 140 credits have either withdrawn (395) or failed (320), suggesting that having more credits increases the dropout and fail rates.

3.4.11 Total Clicks: As shown in Figure 19, the number of students with total VLE clicks in the 0 to 1,000 range gradually de-

creases over time. There are around 19,000 students in this range during the early phase, about 14,000 in the late phase, and roughly 13,000 by the end of the module. Despite this decline, most students consistently fall within the 0 to 1,000 click range at each stage of the module. On the other hand, students with over 5,000 clicks remain relatively uncommon, even in the later phases.

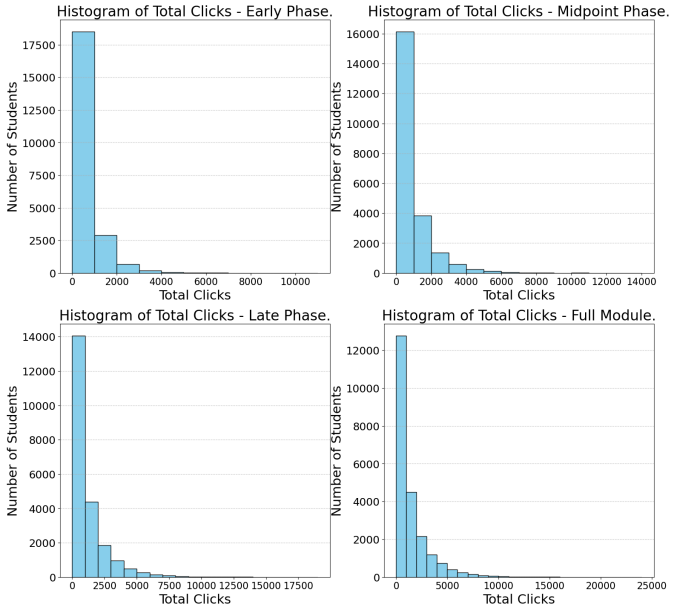

Figure 19. Histogram of Total Clicks for Each Phase.

3.4.12 Days Active: This feature is a binary value between 0 and 1 that captures how consistently a student engaged with the VLE over the course of a module phase. For instance, if the midpoint phase lasts 150 days, the value represents the proportion of those days the student was active. A value of 1 indicates daily engagement, whereas 0 means the student was entirely inactive during that phase.

Figure 20 illustrates how this metric varies across final results. Withdrawn (0) students exhibit a sharp drop in engagement over time, with the median falling from 0.25 (Early) to 0.15 (Midpoint), then to 0.10 (Late), and 0.09 (Full). Moreover, 75% of withdrawn students remain largely inactive throughout, indicating early disengagement with little recovery later on. Still, several outliers are more active than the majority of withdrawn students.

For Fail (1) students, engagement also remains low throughout. Their median drops from 0.22 (Early) to 0.15 (Midpoint), then to 0.11 (Late), and finally to 0.09 (Full). This pattern suggests a gradual loss of interest after early attempts. As with withdrawn students, some outliers demonstrate notably higher activity levels.

Pass (2) students maintain moderate engagement, though it steadily declines over time. The median falls from 0.45 (Early) to 0.36 (Midpoint), then 0.33 (Late), and 0.30 (Full). These students stay relatively active, albeit less consistently, in later phases.

Distinction (3) students are the most engaged group overall. While their activity decreases slightly over time, with medians moving from 0.57 (Early) to 0.47 (Midpoint), 0.44 (Late), and 0.40 (Full), they still demonstrate high and sustained engagement even in the final stages of the module.

Overall, student activity tends to decline over time across all outcome groups. However, the relative ranking is consistent: students who attain a distinction or pass show more persistent engagement than those who fail or withdraw. The gaps between median values also grow wider in later phases, further highlighting the strong relationship between sustained activity and academic success.

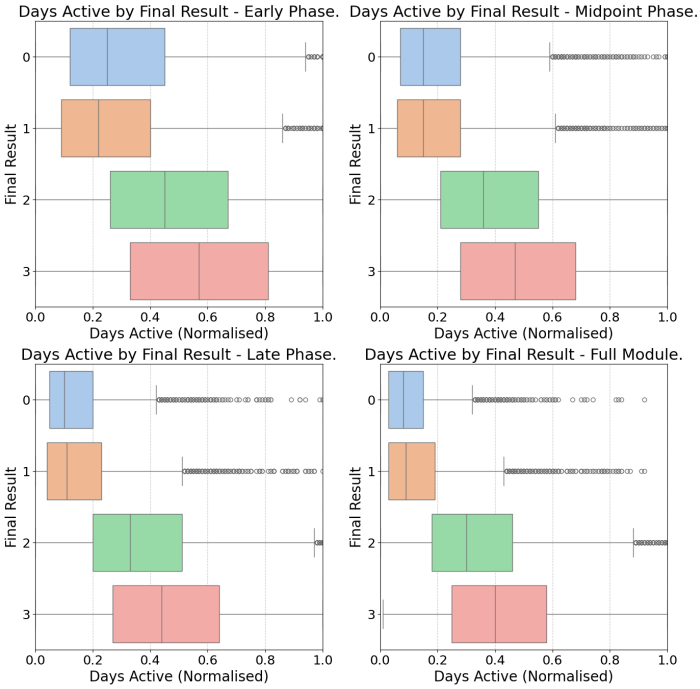

Figure 20. Days Active (Normalised) By Final Result For Each Phase.

3.4.13 Banked Rate: The majority of students do not carry over assessment results from previous presentations, with only around 250 students having a banked rate above 0.7. Moreover, students who either failed or withdrew tend to have slightly higher banked rates on average, although the difference is relatively small. Let us now turn to the `weighted_score` feature.

3.4.14 Weighted Score: Figure 21 shows that in the Early Phase, around 3,750 students fall within the 0–5 score range, suggesting minimal progress or engagement at this point. From a score of 50 onwards, the distribution begins to rise, peaking in the 70–75 band with roughly 2,300 students. Notably, a small group (about 1,000 students) have already achieved near-perfect scores (95–100).

In the Midpoint Phase, the number of students scoring 0–5 remains high at around 4,700. However, there is a noticeable increase in students scoring between 20 and 60. The mode shifts to the 80–85 range, where about 1,950 students are concentrated. This reflects improvement in performance as more assessments are completed. Fewer than 500 students achieve a perfect score at this point.

In the Late Phase, the number of students in the 0–5 range rises slightly to around 4,800, likely including those who became inactive or withdrew. The score distribution is now more concentrated

in the 70–90 range, with a peak still around 80–85. A small number of high achievers remain, with roughly 300 students scoring between 95 and 100.

By the Full Module phase, approximately 5,600 students are still in the 0–5 range, most likely those who dropped out or failed to participate fully. The score distribution has matured, showing strong peaks between 75 and 90. However, only about 200 students have weighted scores between 95 and 100.

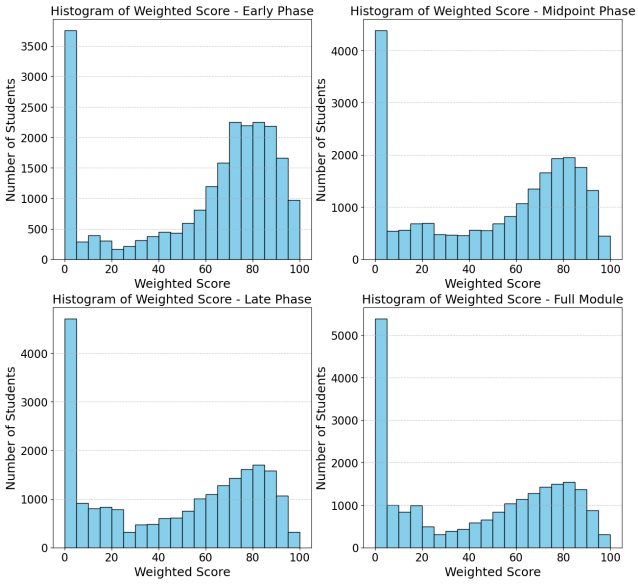

Figure 21. Histogram of Weighted Score for Each Phase.

It is also worth noting that the `weighted_score` is not a fully dependable metric in the Full Module context. This is primarily due to missing exam assessment data that typically becomes unavailable toward the end of the module. Most modules, with the exception of DDD, are affected by this issue, resulting in incomplete or skewed weighted scores that do not accurately represent student performance. Additionally, some modules such as GGG and FFF include TMAs and CMAs that carry no weight. As a result, the `weighted_score` may fail to reflect actual student performance, even if a student passes or fails those assignments, since the outcomes have little impact. For example, in module GGG, only the final exam result truly matters. Given these limitations, the `weighted_score` should be viewed with caution and not relied upon as the sole measure of student success. Let us now proceed to the late rate.

3.4.15 Late Rate: Like the weighted score, both the late rate and the related fail rate are influenced by incomplete assessment data, particularly due to missing exam results. This limits their reliability in the full dataset. Nonetheless, they still offer useful insights.

Figure 22 shows how students' submission patterns change over the phases. In the Early Phase, the majority of students (over 11,200) submitted their assignments on time, falling within the 0–10% late rate range. However, there is a smaller but distinct group of around 2,500 students who had a late rate of 50–60%, indicating frequent delays. Some intermediate ranges (such as 10–20%, 40–50%, and 80–90%) show no entries, which might be due to system-related issues or particular submission patterns. Interestingly, about 4,400 students submitted all their assessments late during this phase.

By the Midpoint Phase, although nearly 10,000 students continued to submit on time, the pattern of lateness became more spread out. Many students fell into the 20–80% late rate range, with a notable increase in the 40–60% category. This suggests a gradual rise in irregular submission habits.

In the Late Phase, the trend becomes more pronounced. While the largest group remains in the 0–10% bin, significant numbers of students now appear in the 20–60% range, especially around the 50–60% mark, which includes roughly 3,200 students. The overall pattern flattens, reflecting a broader shift toward late submissions as the module advances.

Finally, in the Full Module view, on-time submissions still form the largest group, but there is a longer tail across higher late rate intervals. The number of students with 90–100% late rates drops to under 3,700, while the 40–60% range grows more noticeable. This shift indicates increasing delays in submission, possibly due to reduced engagement or growing academic pressure as the module progresses.

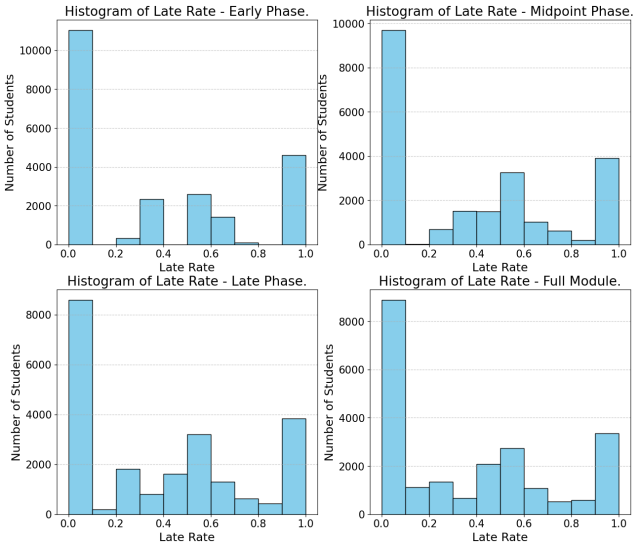

Figure 22. Histogram of Late Rate for Each Phase.

3.4.16 Fail Rate: Figure 23 illustrates how student failure rates change across the different phases of the course, showing a steady shift toward higher failure as the module progresses. In the Early Phase, most students (around 16,000) had low fail rates, staying below 10%, which suggests early academic achievement. However, a smaller group of approximately 2,100 students experienced higher failure rates in the 50–60% range. Several mid-range bins (10–20%, 40–50%, and 80–90%) recorded no students, indicating a highly polarised distribution of failure at this point. In addition, over 2,900 students had extremely high fail rates between 95–100%, reflecting significant academic difficulty from the beginning.

As the course moved into the Midpoint Phase, the number of students with low fail rates (0–10%) dropped to around 13,000. Meanwhile, failures became more evenly distributed, with over 1,400 students in each of the 20–40% bands. The group of students with near-total failure (95–100%) decreased slightly but remained

substantial.

By the Late Phase, the number of students in the 0–10% category fell further to about 10,700. More students began to appear in higher fail rate ranges, including 20–30%, 50–60%, and 80–90%, suggesting growing academic challenges. The distribution clearly shifted towards higher failure, reflecting cumulative difficulties experienced over time.

Finally, during the Full Module period, those with minimal failures (0–10%) declined again to roughly 10,100, although they still formed the largest single group. However, the upper end of the distribution became more populated, with more students in the 70–90% failure range. The number of students failing nearly all assessments (95–100%) reached its highest point at around 3,100, signalling the most concentrated instance of academic struggle seen across the entire module.

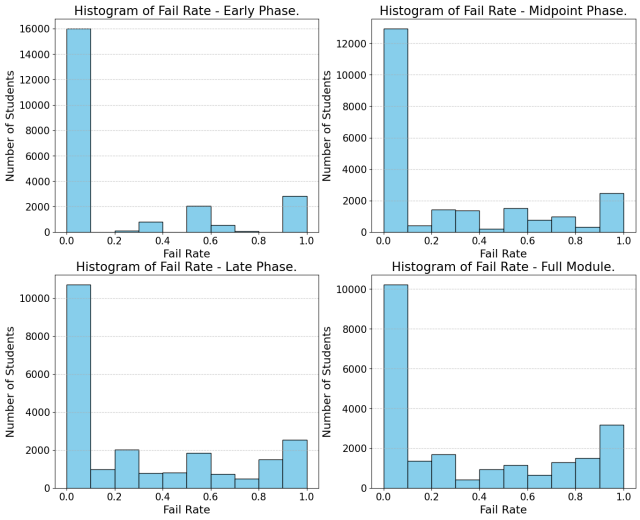

Figure 23. Histogram of Fail Rate for Each Phase.

3.4.17 Final Result: The `final_result` serves as the target variable in our analysis. As shown in Figure 24, the largest proportion of students (approximately 44%) achieved a pass. This is followed by around 25% who failed, while about 20% withdrew from the course. Students who passed with distinction make up the smallest group, accounting for roughly 11% of the training dataset.

3.4.18 Correlation Matrix (Full Phase): A correlation analysis was performed on the full dropout dataset to identify key relationships between features and final outcomes, with the correlation matrix shown in the appendix (Figure 34).

As expected, weighted score has the strongest positive correlation with final results (corr = 0.71), while fail rate shows a strong negative correlation (corr = −0.78), confirming their importance in predicting student success or failure. Engagement metrics like days active (corr = 0.54) and total VLE clicks (corr = 0.41) also positively relate to better outcomes, whereas late submission rate negatively correlates (corr = −0.36) with performance.

Background variables such as highest education level (corr = 0.14) and IMD band (corr = 0.12) have weak positive correlations, while age band and registration date show minimal or no correlation. Gender does not appear to influence final results. Moreover, other features like studied credits (corr = −0.11), previous attempts (corr = −0.10), disability, and banked rate show weak negative correlations, with banked rate offering little predictive value. Notably, higher prior education aligns with better assessment scores (corr = 0.21), and repeat module attempts correlate moderately with banked rate (corr = 0.30), reflecting retake patterns.

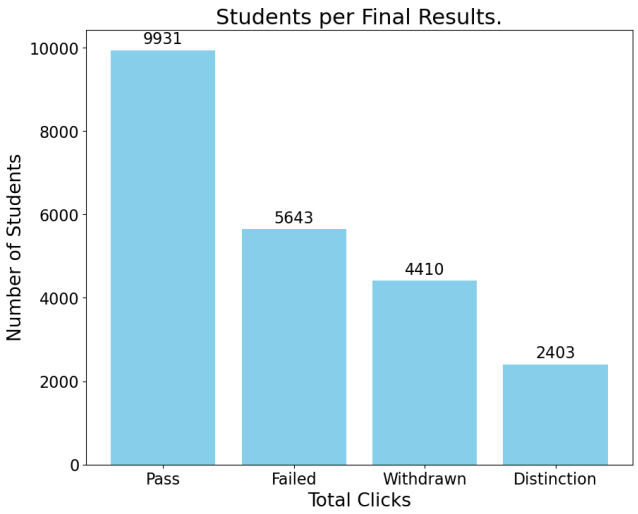

Figure 24. Students per Final Result.

Overall, assessment performance and engagement indicators are far stronger predictors of final outcomes than demographic factors, guiding feature selection for modelling. Let us now move on to hypothesis testing.

3.5 Hypothesis Testing

3.5.1 Engagement Hypothesis: This hypothesis tests whether students who drop out within the first 25% of the module show lower early VLE engagement compared to those who continue, regardless of their final result. The hypotheses for the Mann–Whitney U test are as follows:

- **Null Hypothesis (H_0):** There is no difference in early engagement (`days_active_norm`) between students who dropped out and those who remained enrolled (including those who passed or failed).
- **Alternative Hypothesis (H_1):** Students who dropped out have significantly lower early engagement than those who stayed enrolled.

Because `days_active_norm` is not normally distributed, the Mann–Whitney U test was chosen for its robustness in comparing medians of skewed data [19].

The test showed a significant difference in early engagement between the groups ($U = 29,961,144.5$, $p < 0.001$), with dropouts having notably lower activity during the first 25% of the module. This strongly supports the hypothesis that low early engagement is linked to higher dropout risk. The high U value (close to the maximum) indicates a clear difference in engagement ranks between the groups, and the very small p-value (approximately 8.82×10^{-140}) confirms this difference is statistically significant. Therefore, the null hypothesis can be confidently rejected.

3.5.2 Assessment Performance Hypothesis: This hypothesis investigates whether poor performance in early continuous assessments contributes to an increased risk of dropout. We focus on early assessment scores (e.g. `weighted_score`) collected during the first 25% of the module.

- **Null Hypothesis (H_0):** There is no difference in early assessment scores between students who dropped out and those who remained enrolled (including those who passed or failed).
- **Alternative Hypothesis (H_1):** Students who dropped out scored significantly lower in early assessments compared to those who continued the course.

To test this, a Mann–Whitney U test was applied to compare early assessment performance (`weighted_score`) between the dropout group and the enrolled group (both pass and fail). Since histogram analysis showed non-normal distributions in both groups, a non-parametric approach was appropriate.

The test revealed a statistically significant difference in assessment performance between the two groups ($U = 22,891,409.0$, $p < 0.001$), with students who dropped out scoring substantially lower in early assessments. These findings strongly support the hypothesis that weaker early assessment results are associated with a higher likelihood of dropout, possibly due to diminished confidence or motivation after initial poor academic outcomes.

3.5.3 Demographic Disparity Hypothesis: This hypothesis investigates whether certain demographic groups, defined by age band, region, highest education level, IMD band, and disability status, are disproportionately affected by student dropout. A Chi-Square Test of Independence was used to determine whether dropout rates vary significantly across different categories of these variables, since all involved variables are categorical. The results are shown in Table 10.

Table 10
Chi-Square test results for demographic variables and dropout status

Feature	Chi-Square	p-value	Findings
Age Band	4.10	0.129	No significant variation in dropout rates across age groups.
Highest Education	72.64	0.000	Strong association between prior education level and dropout likelihood.
Region	22.57	0.032	Dropout rates vary by region, possibly due to inequalities in access or support.
IMD Band	65.69	0.000	Students from more deprived areas (low IMD) are significantly more likely to drop out.
Disability	109.53	0.000	Students with disabilities face a substantially higher dropout risk.

Table 10 shows that the hypothesis is partially supported. Significant relationships were found between dropout and highest education level, region, IMD band, and disability. These findings suggest disparities in dropout risk based on these variables. However, no significant effect was found for age, indicating that age alone might not be a key factor in early dropout when other variables are considered. These results are further supported by the correlation matrix in the Appendix Figure 34.

Key findings also emerge from the correlation matrix (Figure 34):

- **Region:** Dropout rates range from 22.8% in Ireland to 35.5% in the West Midlands, indicating a clear influence of location on withdrawal.
- **Highest Education:** Students with higher prior qualifications are more likely to continue their studies.
- **Age Band:** Shows weak evidence, with only a slight correlation to dropout rates.
- **Conclusion:** The hypothesis is partially confirmed. Region, education level, and deprivation level have notable effects, while age appears to have limited impact.

3.5.4 Re-enrolment Hypothesis: The final hypothesis explores whether students who have previously attempted the same module multiple times are more likely to drop out again.

- **Null Hypothesis (H_0):** There is no difference in the distribution of the number of previous attempts (`num_of_prev_attempts`) between students who drop out and those who continue (i.e., achieve a distinction, pass, or fail).
- **Alternative Hypothesis (H_1):** The distribution of previous attempts differs between students who drop out and those who do not.

Given that `num_of_prev_attempts` is a non-normally distributed count variable and the two groups (dropouts vs. non-dropouts) are independent, the Mann–Whitney U test is an appropriate method for comparing their distributions.

The test produced a U statistic of 40,818,054.0 and a p-value of approximately 1.10×10^{-7}. This very small p-value (less than 0.05) provides strong evidence against the null hypothesis, indicating a significant difference in the number of previous attempts between dropouts and non-dropouts. In practical terms, this supports the hypothesis that students with more prior attempts are at a greater risk of dropping out again.

In summary, the analysis reveals statistically significant evidence that a student's re-enrolment history, specifically, the number of times they have previously attempted the module, is associated with an increased likelihood of dropout. The next section discusses the ML models used in this study and the rationale behind their selection for the classification task.

3.6 Implementation of Machine Learning Models

Since the task involves classification, four models were employed: LR, SVM, MLP, and RF. Each model was trained on the training dataset and evaluated on the test set using two primary metrics: accuracy and the macro-averaged F1 score, which accounts for both precision and recall across all classes. Although the models required different types of hyperparameters, their implementation followed a consistent framework, comprising model setup, pipeline construction, model initialisation, hyperparameter tuning, feature importance analysis, and performance evaluation.

Following the stages of data description, preprocessing, and feature engineering, all datasets were merged into a unified table suitable for training and evaluating ML models. The preprocessing

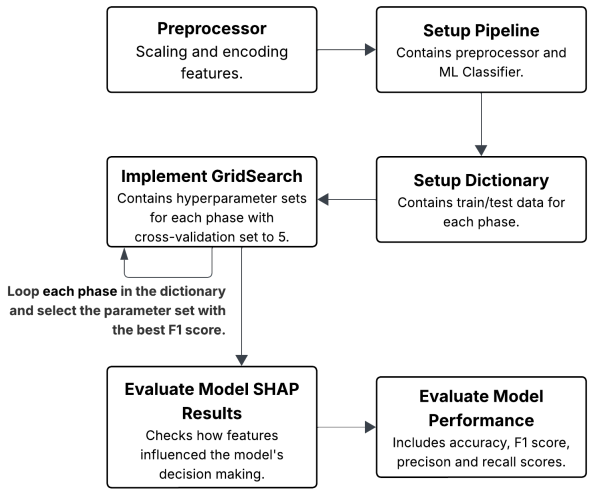

Figure 25. Model implementation and evaluation steps.

phase involved several operations, including imputation of missing values, encoding of categorical features, scaling of numerical features, and other cleaning procedures. Each step was applied with specific justification to ensure data quality and suitability for modelling. The processed dataset was then partitioned into training and testing subsets using an 80:20 split ratio. Subsequent to this, the dataset underwent a second round of exploration to further investigate feature distributions and relationships. Hypothesis testing was also conducted to inform model design and interpretation.

3.6.1 Model and Pipeline Setup: The modelling phase begins with encoding and scaling both the training and test sets using a ColumnTransformer. This transformer applied a standard scaler to numeric columns and a one-hot encoder to categorical columns. Table 9 outlines which features were encoded and scaled, with justifications provided in Section [CITE]. Figure 25 summarises the overall model implementation process.

Each model was implemented using a pipeline structure that included the preprocessor and the respective classifier, such as LR, RF, SVM, or MLP. A fixed random state was used to ensure consistent and reproducible results. To manage datasets across different stages of the module, a dictionary named `phase_sets` was created as follows:

```
phase_sets = {
    'Early': (train_class_early,
        test_class_early),
    'Midpoint': (train_class_midpoint,
        test_class_midpoint),
    'Late': (train_class_late,
        test_class_late),
    'Full': (train_class_full,
        test_class_full),
}
```

3.6.2 Hyperparameter Tuning Using GridSearch: A GridSearchCV model from sklearn was configured with 5-fold cross-validation and a predefined set of hyperparameters for each classifier. The grid search was wrapped within a loop that iterated through each phase of the dataset, allowing the model to be trained and evaluated in a single process. The selection of the best model parameters was based on the F1 score (macro). This metric was preferred over accuracy due to the imbalanced nature of the dataset, where only around 20% of students had withdrawn. As the aim was to effectively identify students at risk of dropping out, the model was optimised for both precision and recall, making the F1 score a more appropriate choice.

3.6.3 Model Evaluation: After the optimal hyperparameters were determined for each ML model across all phases, feature importance was assessed using SHAP. SHAP is a game-theoretic method that assigns each feature an importance value based on its contribution to the model's output [20]. This allows for consistent and interpretable explanations of how individual features influence predictions. SHAP values indicate whether a feature contributes positively or negatively to a prediction and quantify the magnitude of that effect [21]. One major advantage of SHAP is its model-agnostic property, meaning it can be applied to any ML model [20], including LR, SVM, MLP, and RF. Since each model type differs in structure, appropriate SHAP explainers were selected accordingly: LinearExplainer was used for LR due to its linear nature, KernelExplainer was applied to both SVM and MLP as they are non-linear and model-agnostic methods, and TreeExplainer was employed for RF given its suitability for tree-based models. In this paper, SHAP analysis was limited to the Early and Midpoint phases, as these are the primary focus for understanding early indicators of student dropout.

Finally, model performance was assessed using the F1 score (macro) as the primary metric, with additional consideration given to precision and recall. The macro F1 score was selected because it is effective for imbalanced datasets, as it assigns equal importance to all classes regardless of how frequently they appear in the data [22]. Accuracy was also reported but was not treated as a reliable indicator due to class imbalance. All models were trained using the `class_weight='balanced'` setting, which increases the weight of the minority class to improve fairness and detection capability. A classification report was generated to evaluate performance specifically on the withdrawal class (class 0), and performance metrics were visualised through plots of accuracy, precision, and recall scores.

3.7 Conclusion

This section presented the data and outlined the key features considered for model development. It detailed the preprocessing and feature engineering procedures applied to prepare the dataset for ML tasks, including the integration of relational tables, temporal segmentation for early-phase analysis, handling of missing data, and transformation of variables through encoding and scaling. The dataset was then divided into training and testing subsets to maintain data integrity and enable generalisable evaluation. EDA and hypothesis testing were conducted to guide model design and verify assumptions related to student engagement, assessment outcomes, and demographic variation. Subsequently, model development involved configuring preprocessing pipelines, performing

hyperparameter tuning through grid search, and assessing feature importance using SHAP. Model performance evaluation concluded this methodological process, establishing a sound basis for analysing the predictive capacity of the models in identifying students at risk of dropout. The next section presents the results and analysis.

4 Results and Analysis (10)

This section presents the results for each ML model used in this study (LR, SVM, MLP and RF) and evaluates their performance individually. Each model's evaluation is divided into three subsections. The first subsection details the hyperparameter tuning process, including the most optimal parameters identified for each phase using grid search with 5-fold cross-validation. The best hyperparameter combinations are displayed alongside the corresponding macro F1 scores.

Following the hyperparameter tuning, the second subsection focuses on a detailed analysis of the model's performance metrics, including accuracy, F1 score, precision, and recall for each class. As we proceed through each model, their results will be compared with one another as well as with related work and the case studies referenced in Section [CITE].

Finally, the third subsection examines the SHAP values for each model, based on the default positive class (class 1). For instance, a high weighted score shifts the model's prediction toward the positive (non-dropout) class on the right, while a low weighted score pushes it toward the negative (dropout) class on the left. After evaluating all models, we will identify the overall best-performing model for predicting student dropouts in the early and midpoint phases, and highlight the key features influencing these models.

4.1 Logistic Regression

4.1.1 Logistic Regression Hyperparameter Tuning: Grid search was applied to each dataset phase using a LR model. The following set of hyperparameter combinations was tested:

```
lr_param_grid = {
    'classifier__C': [0.01, 0.1, 1, 5, 10],
    'classifier__max_iter': [20, 30, 50,
        100, 200],
    'classifier__class_weight': ['balanced']
}
```

For each phase, the model was trained within a for loop using 5-fold cross-validation. The best-performing configuration in terms of macro F1 score was recorded and selected for that phase. Table 11 presents the best hyperparameter values and corresponding macro F1 scores for each phase of the LR model.

Table 11
Best Logistic Regression hyperparameter combinations and corresponding macro F1 scores for each learning phase.

Phase	C Value	Max. Iter.	Class Weight	Macro F1 Score
Early	0.01	20	balanced	56%
Midpoint	0.01	20	balanced	64%
Late	0.01	30	balanced	67%
Full	0.01	30	balanced	69%

All phases achieve their highest performance with a C value of 0.01, suggesting strong regularisation is effective. The early and midpoint phases converge within 20 iterations, while the late and full phases require 30. The class weight is set to "balanced" across all phases to address class imbalance in the target variable. As more student data becomes available in the later phases, the macro

F1 score improves significantly, rising from 56% in the early phase to 69% in the full phase.

4.1.2 Logistic Regression Model Evaluation: In Table 12 The LR model shows a clear improvement in overall accuracy as more student data becomes available over time, increasing from 75% in the early phase to 82% in the full phase. This upward trend aligns with the increasing availability of meaningful features such as assessment scores and engagement metrics as students progress through the module. Given the imbalanced nature of the dataset, we focus on evaluating the model using macro-averaged F1 scores, precision, and recall for both classes.

Table 12
Logistic Regression Performance by Temporal Phase.

Phase	Acc.	F1 (1)	F1 (0)	Prec. (1)	Prec. (0)	Rec. (1)	Rec. (0)
Early	75%	84%	51%	89%	43%	79%	62%
Midpoint	79%	86%	60%	93%	49%	80%	76%
Late	80%	86%	63%	95%	51%	79%	83%
Full	82%	88%	67%	97%	54%	80%	89%

The LR model performs strongly in identifying students who will continue with their studies (class 1). The F1 score rises from 84% in the early phase to 88% in the full phase, supported by consistently high precision (89–97%) and recall (79–80%). This indicates that the model is both accurate and reliable when predicting the majority class.

Performance for predicting withdrawn (class 0) students also improves across phases. The F1 score increases from 51% to 67%, driven by higher precision (from 43% to 54%) and notably better recall (from 62% to 89%). Precision for Class 0 reflects the proportion of students predicted as withdrawn who actually withdrew, while recall indicates how many of the actual withdrawn students the model successfully identified. The recall gain from 62% to 89% in the full phase is particularly encouraging, showing the model becomes substantially better at catching true dropout cases as it incorporates more data. However, precision remains relatively low, suggesting the model still incorrectly flags some continuing students as withdrawals.

Compared to findings from related work (case 2), which reported accuracies between 80% and 86% using different target definitions (e.g., four-class outcomes like pass/fail or distinction/pass), the LR model in this study does not perform particularly well, especially during the early phase. Here, the F1 score is just 51%, with a low precision of 43%, indicating that the model correctly identified only 43% of students who dropped out, and it recognised just 62% of actual withdrawals. Performance improves notably in the midpoint phase, where the F1 score rises to 60%, with a modest increase in recall to 76%, though precision remains relatively low at 49%.

The relatively weak performance of the LR model may be due to its simplicity as a linear classifier. LR often struggles to model complex, non-linear relationships within the data, which likely made it difficult to effectively separate students who withdrew from those who continued. Next, this paper will assess and interpret the SHAP values for the LR model.

4.1.3 Logistic Regression SHAP Values Evaluation: Figure 26 presents the SHAP values for the most influential features in the

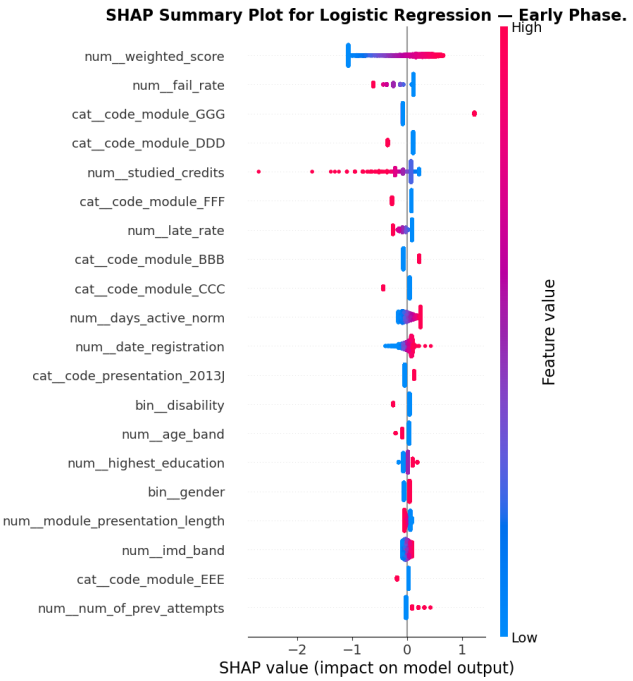

Figure 26. SHAP summary plot showing the most influential features for the Logistic Regression model in the Early phase.

LR model during the early phase. The plot shows that weighted score is the most influential variable. A high weighted score (red) is associated with positive SHAP values, meaning it strongly pushes predictions toward the positive class (pass). This aligns with intuition, since students with higher scores are performing well academically and are more likely to pass. Conversely, a low weighted score (blue) shifts predictions toward the negative class (dropout), indicating a higher likelihood of withdrawal.

The fail rate is the next most influential feature, where higher failure rates push predictions toward dropout and reduce the probability of the positive class. Module type (from the one-hot encoded `code_module` feature) also appears prominently, showing that certain modules influence the likelihood of passing or dropping out.

An interesting pattern is seen in studied credits. The SHAP values indicate a threshold effect in which taking too many credits can increase the risk of dropout, while maintaining a moderate workload improves the chances of passing. Similarly, late rate behaves like fail rate, with higher lateness pushing predictions toward dropout. However, its overall influence is weaker compared to weighted score, meaning that even extreme values only slightly affect the final decision.

Other features, such as days active and date registration, have smaller overall impacts but still contribute to the model's decision-making. Demographic variables, including disability, age band, and highest education, also influence predictions to a lesser extent. While these features are not as impactful as weighted score, fail rate, studied credits, or late rate, they still provide additional context for the model's assessment.

In the midpoint phase plot shown in Figure 27, weighted score remains by far the most influential feature. Its SHAP value range is

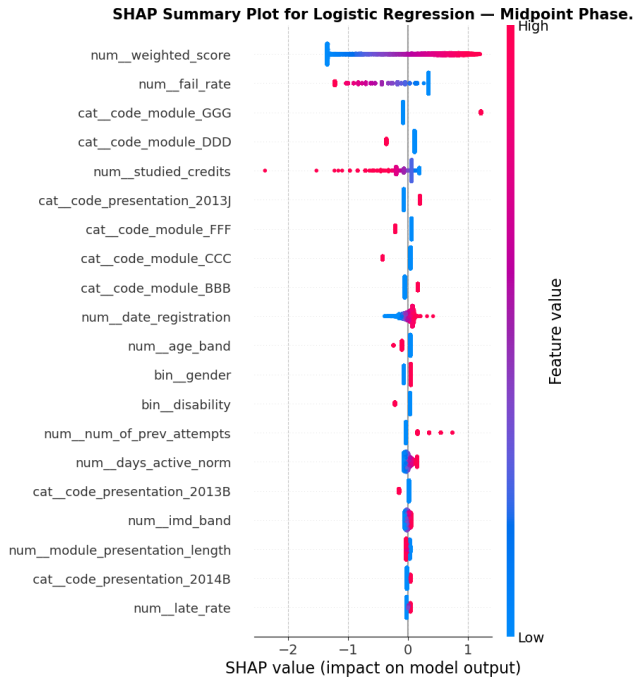

Figure 27. SHAP summary plot showing the most influential features for the Logistic Regression model in the Early phase.

slightly wider than in the early phase, extending to approximately ±1.2 in extreme cases. This indicates that as the course progresses, weighted score becomes even more predictive of final outcomes. This is expected, since more assessments have been completed by this stage, and students who consistently perform well (red) are more likely to succeed academically, while those with lower scores (blue) are more likely to fail.

Fail rate is still the second most influential feature, with a broader spread (around −1 to 0.2) compared to the early phase, suggesting it has gained predictive power and generally shifts the model's output toward the dropout class.

Studied credits shows little change in influence, which is expected since this feature does not vary across phases. A notable shift occurs with late rate, which drops from being among the top seven features in the early phase to around the twentieth position at midpoint. This decline may be due to the fact that students still submitting work by midpoint are generally engaged and have a higher likelihood of success. Many students who dropped out early due to late submissions are already absent from the dataset by this stage. Consequently, weighted score becomes a stronger determinant, as the key question shifts from timeliness to actual academic performance.

Days active experiences a similar drop in influence, likely for the same reason as late rate. Demographic features such as disability status, age band, and highest education remain unchanged in influence because they are static variables. Next, we examine the performance of the SVM model and conduct an evaluation using SHAP values.

4.2 Support Vector Machine

4.2.1 SVM Hyperparameter Tuning: As with the LR model, the SVM was tuned using a grid search applied separately to each dataset phase. Only the RBF kernel was used in the final tuning process. While experiments were conducted with the polynomial kernel, these led to excessive computation times, frequent crashes due to hardware limitations, and overall mediocre performance, so they were not pursued further.

```
svm_param_grid = {
    'classifier__kernel': ['rbf'],
    'classifier__C': [0.01, 0.1, 0.5, 1, 5],
    'classifier__gamma': ['scale', 0.001,
        0.01, 1],
    'classifier__class\_weight': ['balanced'
        ]
}
```

The optimal configuration for each phase was determined based on the highest macro F1 score. Table 13 summarises the best-performing hyperparameters along with the corresponding macro F1 scores.

Table 13
Best SVM hyperparameter combinations and corresponding macro F1 scores for each learning phase.

Phase	C Value	Gamma Value	Class Weight	Macro F1 Score
Early	0.1	0.001	balanced	65%
Midpoint	0.1	0.001	balanced	67%
Late	0.1	0.001	balanced	69%
Full	0.1	0.001	balanced	71%

From Table 13, it is clear that all phases achieved their highest macro F1 score when $C = 0.1$ and $\gamma = 0.001$. Notably, the SVM model consistently outperformed LR across all phases, achieving a substantial 9% improvement in the Early phase and a smaller 3% gain at the Midpoint phase. These results suggest that the SVM model is generally more effective at distinguishing between dropout and non-dropout cases. A more detailed analysis of accuracy, macro F1, precision, and recall per class follows in the subsequent subsection.

Table 14
SVM (RBF) Performance by Temporal Phase.

Phase	Acc.	F1 (1)	F1 (0)	Prec. (1)	Prec. (0)	Rec. (1)	Rec. (0)
Early	78%	86%	48%	87%	46%	84%	51%
Midpoint	78%	86%	57%	91%	48%	81%	70%
Late	79%	86%	62%	95%	49%	78%	84%
Full	81%	86%	66%	97%	51%	78%	92%

4.2.2 SVM Model Evaluation: Table 14 provides a detailed comparison of model performance. In terms of accuracy, SVM achieves slightly higher accuracy than LR during the Early phase, reaching 78% compared to LR's 75%. Accuracy improves across phases for both models as more student data becomes available, with LR gaining a slight advantage by the Full phase (82% vs. 81%).

For the continue class (non-dropout, class 1), both models exhibit similar F1 scores in the range of 84–88%. However, the more noticeable differences appear in the dropout class (class 0), where SVM's F1 scores are lower than LR's in the Early phase (48% vs. 51%) but gradually improve to nearly match LR by the Full phase (66% vs. 67%).

SVM shows slightly lower precision for the positive class in almost all phases, indicating it is somewhat less effective at correctly identifying students who will continue without producing many false positives, particularly in the early and midpoint phases. Conversely, LR demonstrates higher recall for the dropout class in the Early phase (62% vs. 51%), suggesting it better captures actual dropout cases early on. Although recall improves for both models in later phases, LR maintains a modest lead.

Overall, the SVM model performs marginally better in most metrics during the early phase, showing a small advantage in accuracy, F1 score for class 1, precision for class 0, and recall for class 1, but this advantage diminishes as LR catches up in subsequent phases. As a result, SVM's overall performance is comparable to, or slightly behind, LR in general.

Comparing to related work, this SVM model shows somewhat lower F1 scores (65–67%) in the early and midpoint phases compared to another study (referred to as Case 1 in Section [CITE]), where an SVM with RBF kernel achieved approximately 94% F1 score for full data predicting pass/fail rather than dropout/continue. Another study (Case 2) reported SVM accuracies between 80–90% across four segmented phases, though their task did not distinguish dropout versus non-dropout. The current study's SVM results, with F1 scores between 65–71% and accuracy of 78–81%, are somewhat lower than Case 1 but closer to Case 2's performance. These comparisons highlight how different label definitions and prediction targets impact model performance.

The next section will examine the SHAP values to better understand the SVM model's decision-making process.

4.2.3 SVM SHAP Values Evaluation: It is important to note that the SHAP analysis for the SVM model was conducted using a reduced training subset of size 50. This subset was passed to the KernelExplainer, which utilised the model's predicted probabilities. This approach was necessary due to the extensive processing time and computational demands. While the results may not be perfectly precise, they provide a reliable approximation of each feature's relative influence.

Figure 28 shows that the fail rate is the most significant feature, though its SHAP values are confined to a narrow range between -0.02 and 0.10, which is notably smaller than the range observed in the LR model. This could be attributed to the nature of the SVM's decision boundary, which relies more on support vectors and margins rather than on explicit feature weights. Another possible explanation is that the use of a smaller sample for the SHAP approximation, due to high computational demands, may have reduced the variability in the SHAP values. The weighted score follows as the next most influential feature, displaying a similar limited range of impact. Other relevant features include studied credits, late rate, and the module BBB, while variables such as date of registration and days active have comparatively less influence.

Demographic features like gender, IMD band, age band, and disability play an even smaller role in the SVM model's decision-making process compared to LR. The SHAP values for the mid-

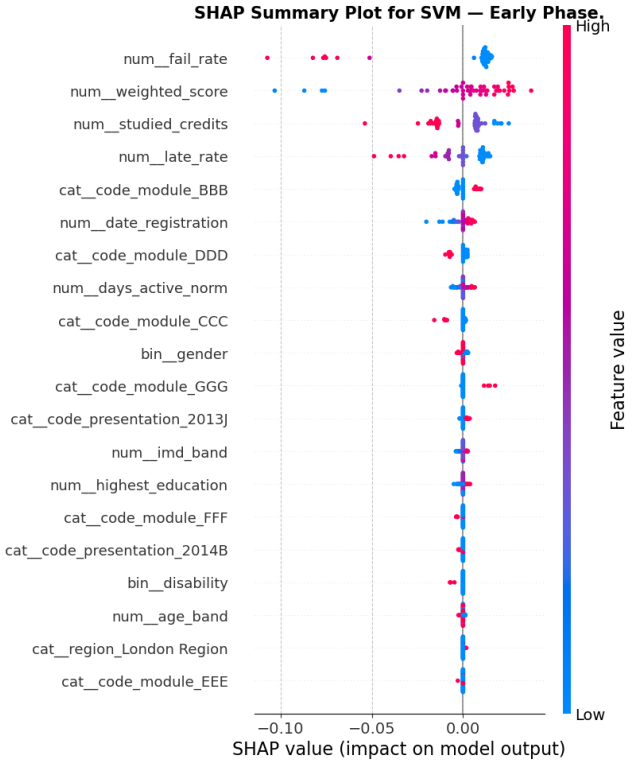

Figure 28. SHAP summary plot showing the most influential features for the SVM (RBF) model in the Early phase.

point phase will be examined next.

Figure 29

Again similar to SVM Early phase, fail rate is the most influential feature followed by weighted score and then studied credits. Beyond this things change again, like the late rate falling from the 4th place to the 12th place being taken over by date registration a static feature. this is a similar behavious obseverved from LR midpoint phase SHAP values. Again the reasoning may be similar to why this has happened in the LR model as well may be due to the fact that students still submitting work by midpoint are generally engaged and have a higher likelihood of success. Many students who dropped out early due to late submissions are already absent from the dataset by this stage. Moreover the SHAP values here still remain low between -0.1 to 0.2. apart from the fail rate, weighted score, studied credits and date registration, other features and demographic features have small influence in the SVM models decision making process. Next, let us evaluate the MLP classifier results.

Similar to the SVM early phase, the fail rate remains the most influential feature, followed by the weighted score and studied credits. Beyond these, the rankings shift, with the late rate dropping from fourth place to twelfth, replaced by date registration, which is a static feature. This mirrors the pattern observed in the LR midpoint phase SHAP values. A possible explanation is that students still submitting work by the midpoint are generally engaged and more likely to succeed, while many who dropped out earlier due to late submissions are no longer present in the dataset. The SHAP values remain relatively low, ranging from -0.1 to 0.2. Aside

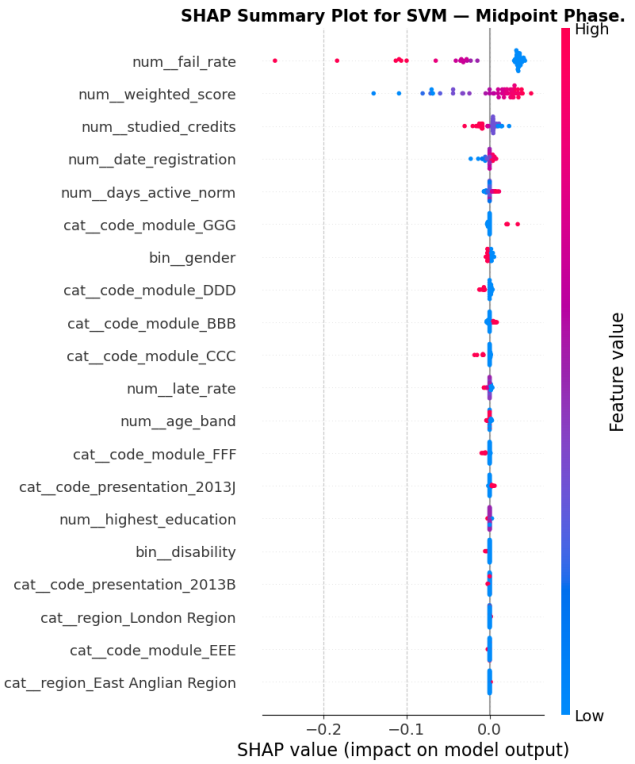

Figure 29. SHAP summary plot showing the most influential features for the SVM (RBF) model in the Midpoint phase.

from fail rate, weighted score, studied credits, and date registration, other features, including demographic variables, have minimal influence on the SVM model's decision-making process. Next, the results for the MLP classifier will be examined.

4.3 Multi-Layer Perceptron Classifier

4.3.1 MLP Hypermeter Tuning: Grid search with five-fold cross-validation was applied to identify the optimal hyperparameter configurations for the MLP model across each learning phase. The parameter grid used is shown below:

```
mlp_param_grid = {
    'classifier__hidden_layer_sizes': [(100,
        50), (100, 50, 25), (100, 100, 50)
       ],
    'classifier__alpha': [1e-4, 1e-3, 1e-2],
    'classifier__learning_rate_init':
       [0.001, 0.01],
    'classifier__max_iter': [350]
}
```

After running the MLP model using gridsearch model the highest macro F1 score for each phase was obtained and displayed in 15 along with the best combinations of hyperparameters. In general all phases converged at $\alpha = 0.01$, learning rate at 0.01 and maximum iteration at 350. The only differencerecorded was in the hidden layer sizes where the early phase had 3 layers with 50 addtional neurons on the third instead of 2 for the remaining phases with 100 each. As for the macro F1 scores, the perforamnce is

Table 15
Best MLP Classifier hyperparameter combinations and corresponding macro F1 scores for each learning phase.

Phase	Alpha Value	Hidden Layer Size	Learning Rate	Max. Iter.	Macro F1 Score
Early	0.01	(100, 100, 50)	0.01	350	43%
Midpoint	0.01	(100, 50)	0.01	350	47%
Late	0.01	(100, 50)	0.01	350	49%
Full	0.01	(100, 50)	0.01	350	57%

pretty weak especially compared with both LR and SVM, the performance remains below 50% from early to late phases with the full phase being the only one with the macro F1 score more than 50%. Let us see MLP models perforamnce in more detail.

The results in Table 15 show that all phases converged on the same α value of 0.01, a learning rate of 0.01, and a maximum iteration count of 350. The main variation occurred in the hidden layer sizes: the early phase used three layers with an additional 50 neurons in the third layer, whereas the remaining phases used two layers of 100 neurons each.

Macro F1 scores indicate relatively weak performance compared to LR and SVM. From the early to late phases, the scores remain below 50%, with the full phase being the only stage to surpass the 50% threshold (57%). A closer look at the MLP model's performance is presented in the next section.

Table 16
MLP Classifier Performance by Temporal Phase.

Phase	Acc.	F1 (1)	F1 (0)	Prec. (1)	Prec. (0)	Rec. (1)	Rec. (0)
Early	77%	86%	38%	84%	43%	89%	34%
Midpoint	80%	88%	47%	86%	51%	89%	43%
Late	82%	89%	56%	88%	56%	89%	55%
Full	83%	89%	58%	89%	57%	89%	59%

4.3.2 MLP Model Evaluation: A shown in Table 16, The MLP classifier consistently underperforms relative to both SVM and LR across all temporal phases, particularly in terms of F1 score for the dropout class, which remains noticeably lower. Although MLP maintains strong F1 scores for the positive class (above 86% in every phase, comparable to SVM and LR), its F1 (0) begins at only 38% in the early phase and rises to just 58% in the full phase. By comparison, LR achieves 51–67% for F1 (0), and SVM reaches 48–66% over the same period.

For dropout class precision, MLP starts at 41% in the early phase and improves into the low-to-high 50% range in later stages, ultimately surpassing both SVM and LR in the full phase (57% for MLP versus 54% for LR and 51% for SVM). However, its recall for the dropout class is particularly weak in the early and midpoint phases (34–43%), reflecting low sensitivity to this class. SVM records higher recall in these stages (51–70%), while LR achieves 62–76%.

Regarding accuracy, MLP shows a slight advantage over SVM and LR in the late and full phases, with a peak full-phase accuracy of 83%, compared to 82% for LR and 81% for SVM. Nonetheless, these small accuracy gains obscure the more significant imbalances in class-level F1, precision, and recall, where MLP struggles

to deliver consistent performance across both classes. Additionally, the ANN model from case 2 records accuracy between 84–93% across time segments, which is not far ahead of MLP's 77–83% range. Nevertheless, this study places greater emphasis on macro F1, precision, and recall, particularly for class 0 (the dropout class).

Overall, while MLP shows competitive accuracy and high F1 for the positive class, its inability to effectively capture the negative class, evident from consistently lower F1 (0), precision (0), and recall (0) values, results in weaker macro performance compared to SVM and LR. This imbalance indicates a bias towards the positive class, reducing its robustness in scenarios with class asymmetry. We now turn to the evaluation of the MLP SHAP values.

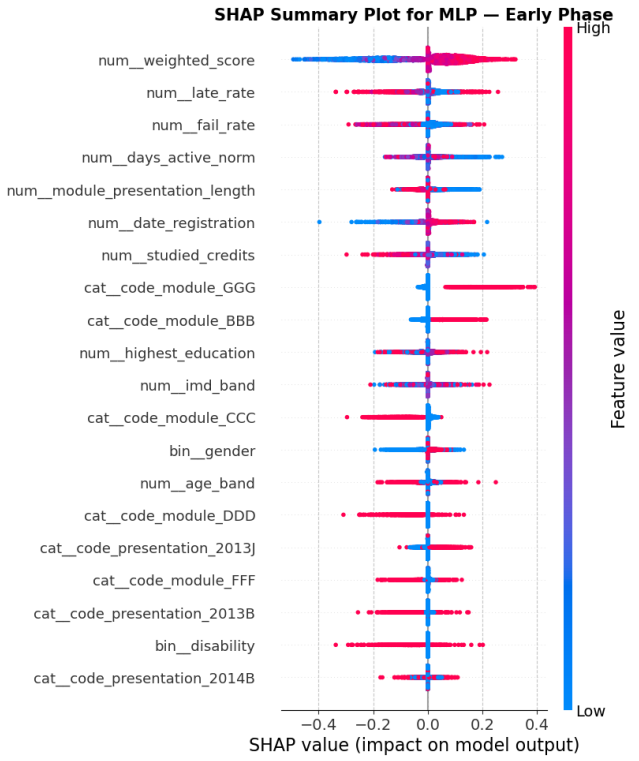

Figure 30. SHAP summary plot showing the most influential features for the MLP Classifier model in the Early phase.

4.3.3 MLP SHAP Values Evaluation: From the SHAP summary plot in Figure 30, the weighted score emerges as the most influential feature in the early phase, followed by late rate, fail rate, days active, module presentation length, and date registration. The weighted score shows the widest spread, roughly 0.4 to +0.4 on the x-axis, with red points concentrated on the right and blue points on the left. This indicates that higher weighted scores push predictions toward the positive class, while lower scores increase the likelihood of a dropout prediction.

The late rate is the second most important feature but shows a narrower range of influence compared to the weighted score. Notably, red points appear on both sides of the x-axis, indicating a non-linear or interacting effect between lateness and the predicted outcome. In some cases, high lateness is associated with success, while in others it correlates with failure—an inconsistent pattern that suggests the MLP model struggles to distinguish dropouts from non-dropouts based solely on late rate. Similar behaviour is observed for fail rate, age band, disability, and other features. Due to the MLP's non-linear capabilities, it can capture such complex conditional relationships that linear models cannot, which also helps explain the model's low precision and recall for the dropout class.

Days active, module presentation length, and date registration show moderate influence, with SHAP values clustered closer to zero, indicating usefulness but not decisiveness. Demographic features, including gender, age band, and IMD band, have minimal spread, reflecting minor direct impact in the early phase, consistent with patterns observed in LR and SVM models.

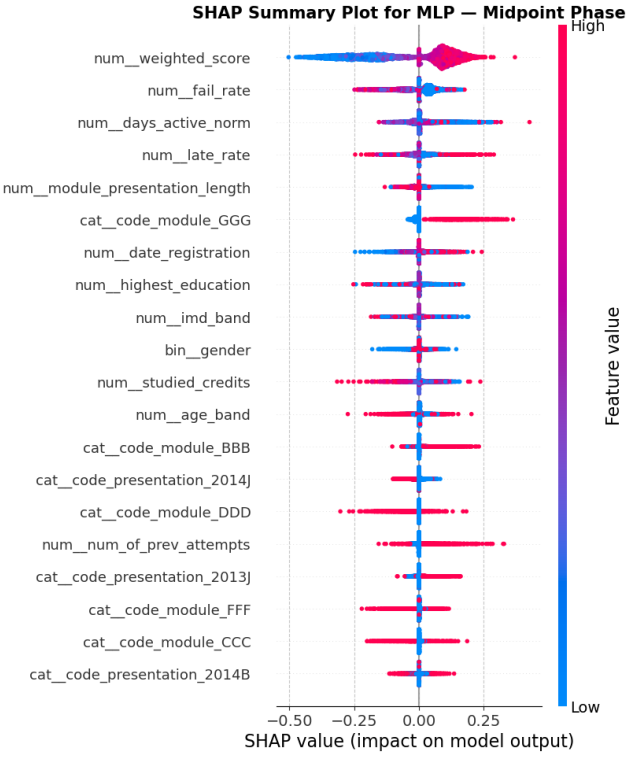

Figure 31. SHAP summary plot showing the most influential features for the MLP Classifier model in the Midpoint phase.

Comparing the MLP SHAP values at the midpoint phase (Figure 31) with those from the early phase, it is evident that the weighted score remains the most influential factor initially, with late rate, fail rate, and days active having smaller effects. During the early phase, the model's predictions rely heavily on early performance indicators, though these metrics have limited variability due to fewer completed assessments. By the midpoint phase, performance-related features, especially weighted score and fail rate, become significantly more prominent. The wider range of SHAP values for weighted score at this stage indicates increased confidence in its impact. Meanwhile, the influence of certain behavioural and demographic features diminishes or remains unchanged.

Moerover, Fail rate becomes the second most important feature, showing a slight negative association. By this stage, enough failed attempts have occurred for fail rate to serve as a reliable predictor. High fail rates (red points) tend to cluster on the dropout side,

though a few low fail rate cases (blue points) also appear in the dropout group, suggesting either model difficulty in separating classes or dropouts occurring for other reasons.

Days active remains among the top four features, while late rate drops from second to fourth place and shows a slightly narrower spread compared to the early phase. This decline in predictive power likely occurs because late submissions lose significance once direct performance data is available. Demographic variables continue to play a minimal role in model predictions, mirroring the trend seen in LR and SVM results. Let us now move on to evaluating the RF model.

4.4 Random Forest

4.4.1 RF Hyperparamter Tuning: The final model to be evaluated is the RF classifier. Similar to the previous models, RF hyperparameters were optimised using grid search combined with 5-fold cross-validation. The sets of hyperparameter combinations tested are presented below:

```
rf_param_grid = {
    'classifier__n_estimators': [100, 200,
        300],
    'classifier__max_depth': [None, 5, 10,
        20],
    'classifier__min_samples_split': [2, 3,
        5],
    'classifier__min_samples_leaf': [1, 2],
    'classifier__class_weight': ['balanced']
}
```

Table 17
Best Random Forest hyperparameter combinations and corresponding macro F1 scores for each learning phase.

Phase	No. of Estim.	Max. Depth	Min. Samp. Split	Min. Samp. Leaf	Macro F1 Score
Early	100	5	2	2	48%
Midpoint	300	5	5	2	59%
Late	100	5	2	2	68%
Full	100	5	2	1	67%

Table 17 presents the optimal hyperparameter settings that yielded the best macro F1 scores. Most phases converged on using 100 estimators, except for the midpoint phase which required 300. The maximum tree depth was consistently set to 5 across all phases. The minimum samples required to split a node was 2 for all phases except the midpoint, where it was set higher. Similarly, the minimum samples per leaf settled at 2 for all phases except the full phase, where it was reduced to 1.

Regarding macro F1 performance, RF displays the lowest scores among all models except for MLP across all phases. Starting at 48% in the Early phase, its performance rises to 59% at Midpoint, then to 68% in the Late phase, before slightly declining to 67% in the Full phase. Although RF's scores improve over time, they remain below those of LR and SVM. However, in the later phases, RF narrows the gap and approaches the performance levels of LR and SVM, which range between 67% and 71% for the Late and Full phases, while RF scores 67% to 68%.

Moreover, LR performs significantly better, with macro F1 scores starting at 56% in the Early phase and steadily increasing to 69% by the Full phase. This reflects a more balanced classification capability, particularly in earlier phases. SVM consistently outperforms all other models, achieving the highest macro F1 scores across every phase, beginning at 65% in the Early phase and reaching 71% in the Full phase. This demonstrates SVM's effectiveness in balancing precision and recall for both classes throughout the timeline. Meanwhile, MLP shows the weakest results, starting at 43% in the Early phase and improving only to 57% in the Full phase, indicating greater difficulty in distinguishing classes, especially during the early stages.

In conclusion, SVM leads in macro F1 performance throughout all phases, followed by LR and RF. MLP trails behind. This comparison suggests that, for this dataset and prediction task, SVM and LR offer more reliable and balanced classification results than the other models. Next, the RF model's performance will be examined in greater detail, focusing on accuracy, F1 score, precision, and recall for each class.

Table 18
RF Classifier Performance by Temporal Phase.

Phase	Acc.	F1 (1)	F1 (0)	Prec. (1)	Prec. (0)	Rec. (1)	Rec. (0)
Early	76%	84%	50%	89%	43%	80%	60%
Midpoint	77%	84%	60%	94%	47%	76%	82%
Late	79%	85%	63%	97%	49%	76%	89%
Full	81%	86%	67%	98%	52%	77%	94%

4.4.2 RF Model Evaluation: Based on the results in **??**, RF achieves 76% accuracy in the Early phase, increasing steadily to 81% in the Full phase. LR starts slightly lower at 75% but overtakes RF from the Midpoint onwards, ending at 82% in the Full phase. SVM opens stronger than RF at 78%, maintains a consistent lead, and finishes alongside LR at 81% in the Full phase. MLP surpasses all other models from the Midpoint phase, reaching the highest accuracy of 83% in the Full phase. Overall, while RF improves over time, it remains behind MLP and SVM, and only manages to match LR in the Late phase before slipping slightly in the Full phase.

Looking at F1 scores for Class 1 (continue class), RF performs consistently well, moving from 84% to 86%. LR slightly outperforms RF in later stages, peaking at 88% in the Full phase. SVM stays at 86% across all phases, matching RF in the later phases but starting stronger in Early. MLP remains ahead of RF throughout, starting at 86% and climbing to 89% from the Late phase onward. In short, RF holds its ground for Class 1 but trails LR and MLP towards the end. For Class 0 (dropout class), RF starts at 50% in Early and improves to 67% in Full. LR only outperforms RF in the Early phase, then matches it from the Midpoint to the Full phase. SVM generally performs slightly worse than RF, ending at 66%. MLP lags far behind in Early (38%) and, despite improving to 58% in Full, still remains below RF and the others. Here, RF consistently outperforms MLP and catches up with LR by the final phase.

For precision, RF starts strong for Class 1 at 89% and climbs to 98% in the Full phase, the highest among all models. LR is close behind, ending at 97%, while SVM also peaks at 97% but begins lower. MLP trails with a maximum of 89%. In the case of Class 0, RF is weaker, starting at 43% and only reaching 52% in Full. LR maintains a slight lead in all phases. SVM begins ahead in Early

and Midpoint, but RF narrowly surpasses it in the Full phase (52% vs 51%). MLP generally performs best for Class 0 precision (except in Early, where SVM leads), finishing ahead of RF in the Full phase (57% vs 52%). This suggests that RF struggles with precision for dropouts, often misclassifying them as non-dropouts.

In terms of recall for Class 1, RF begins at 80% but drops to the mid/high 70s after the Early phase. LR and SVM converge with RF in the later stages (around 78–80%), with SVM also declining from 84% to 78%. MLP clearly dominates here, holding a steady 89% across all phases. Thus, RF is reliable but clearly outperformed by MLP in identifying actual non-dropouts. For Class 0 recall, RF starts at 60% and climbs to 94% in the Full phase, the best among all models. LR follows closely, peaking at 89% and slightly outperforming RF in the Early phase (62% vs 60%). SVM improves from 51% to 92%, making it the third-best at peak performance. MLP records the lowest recall for Class 0, starting at 34% and reaching only 59% by Full. This shows RF's clear strength in correctly identifying non-dropouts, particularly from Midpoint onwards.

In related work, Case 3 identified RF as the top-performing model, achieving 79% accuracy and F1 score at the early stage (20% of course completion) and improving to 91% for both metrics when using the full dataset. It's important to note that their task focused on pass–fail classification. In our study, the RF model records a 76% accuracy in the early phase, close to Case 3's early performance and peaks at 81% in the full phase. We now turn to examining the RF SHAP values.

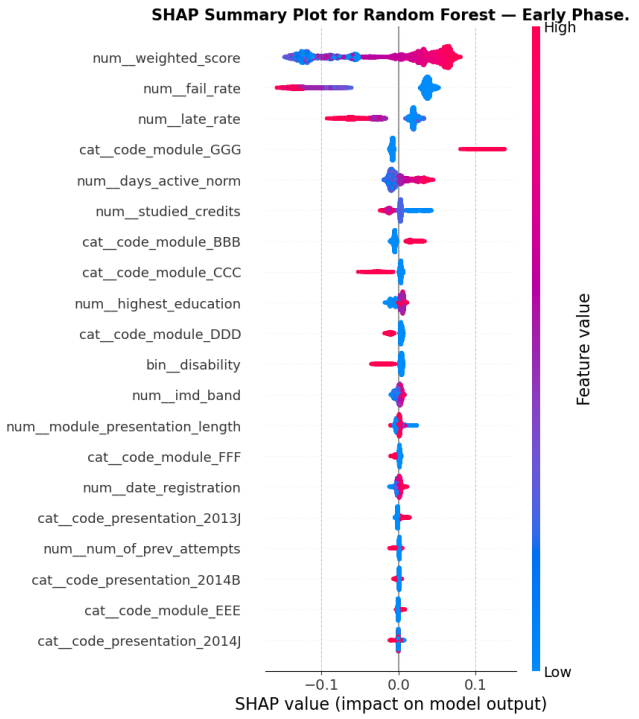

Figure 32. SHAP summary plot showing the most influential features for the RF Classifier model in the Early phase.

4.4.3 RF SHAP Values Evaluation: Figure 32 shows that, in the early phase, the SHAP values for the RF model have a noticeably smaller absolute range compared to the LR and MLP plots discussed earlier (around −0.15 to +0.15), with only the SVM showing slightly smaller values. This indicates that individual feature effects are generally weaker for RF at this stage, which is expected since RF distributes predictive influence across many trees and features. Nevertheless, a few variables still emerge as clearly important, similar to the other ML models evaluated in this paper: weighted score, fail rate, and late rate.

The weighted score remains the top predictor early on. Higher scores (red points) push predictions strongly towards the positive class, while lower scores (blue points) push them negative. Its SHAP range spans roughly from −0.15 to +0.08, which is smaller than in the MLP and LR but still the largest within the RF model at this stage. The fail rate is the second most influential feature: higher fail rates (red) drive sharply negative predictions, while lower fail rates (blue) have a positive effect. This shows that even early in the course, failing a large proportion of assessments is already a strong warning signal. The late rate ranks third, with high lateness pushing predictions negative and low lateness pushing them slightly positive, highlighting that timeliness is meaningful even at this early stage.

The categorical variable code module GGG surprisingly holds a relatively large effect size compared to other categorical features. Students in module GGG appear to influence predictions strongly in a single direction (likely negative, judging by the clustering of one color on the left), suggesting that this module may historically have lower pass rates. The days active variable is the fifth most important and follows a familiar pattern seen in LR, SVM, and MLP results: more active days generally correlate with positive predictions, while low activity predicts poorer outcomes. This underlines the role of engagement as a proxy for success before many graded assessments are completed. Finally, studied credits shows a moderate influence in the early phase, possibly reflecting differences in workload, with higher credit loads potentially correlating differently with performance. Additionally, Demographic features once again have only minor influence, with SHAP ranges within ±0.05. or less. This confirms that, for RF in the early phase, predictions are driven far more by performance and engagement metrics than by demographic factors, a pattern also seen in the other models tested.

By the midpoint of the course (Figure 33), the model's feature importance shifts noticeably. Fail rate overtakes weighted score as the most influential predictor. In the early phase, weighted score dominated, with higher scores driving positive predictions and lower scores driving negative ones. At the midpoint, fail rate takes the lead, showing a clear split: high fail rates (red) strongly push predictions negative, while low fail rates (blue) push them positive. This shift is logical, by this stage, students have completed more graded work, making fail rate a direct indicator of performance trends. Weighted score drops to second place but remains highly influential, with a similar SHAP range of about ±0.15, meaning its overall impact has changed little, though it is now slightly overshadowed by fail rate.

Late rate stays in third place but its pattern becomes more pronounced: high lateness correlates with negative predictions, low lateness with positive ones. This suggests punctuality remains consistently important and becomes more diagnostic as deadlines accumulate. Days active moves up to fourth place, still positively linked to better outcomes, though it now has less predictive power than direct performance measures, a typical trend as engagement

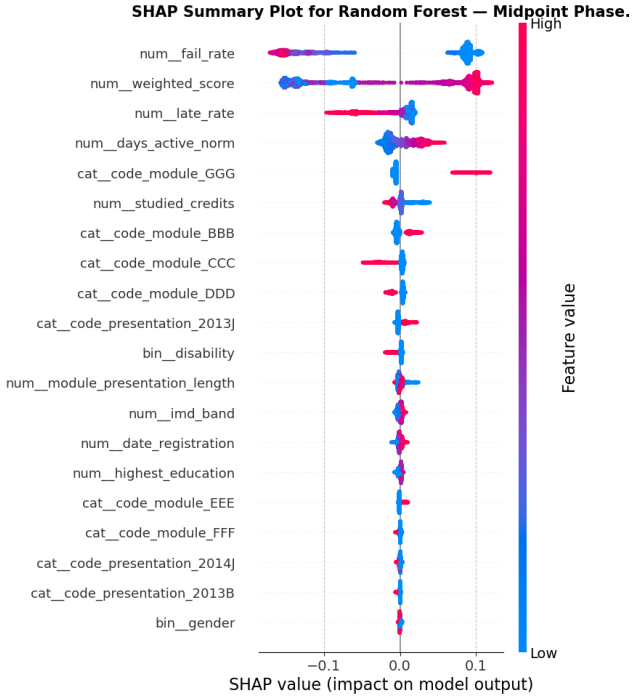

Figure 33. SHAP summary plot showing the most influential features for the RF Classifier model in the Midpoint phase.

indicators give way to actual results.

Categorical module effects (GGG, BBB, CCC, DDD) remain stable, with GGG still showing the strongest influence, suggesting persistent module-level differences. The other module codes cluster around the middle in importance, with narrow SHAP ranges. Demographic variables (disability, IMD band, highest education, gender) continue to have minimal impact, with narrow ranges centred near zero, contributing far less to predictions than performance metrics. We will now proceed to the final conclusion of the Results and Analysis section.

4.5 Best Student Dropout Model

After reviewing the optimal hyperparameters and macro F1 scores obtained via GridSearch with 5-fold cross-validation, SVM achieved the highest overall macro F1 scores (65% Early, 67% Midpoint), followed by LR (56–64%), RF (48–59%), and MLP (43–47%).

Focusing specifically on dropout detection (class 0) during the Early and Midpoint phases, LR emerged as the most effective model, followed by RF, SVM, and MLP. LR achieved an Early-phase F1 score of 51%, precision of 43%, and recall of 62%, improving at Midpoint to an F1 of 60%, precision 49%, and recall 76%. These results demonstrate LR's strong ability to identify dropouts early, with a good balance between precision and recall.

RF showed competitive performance, with an Early-phase F1 of 50% and recall of 60%, and Midpoint F1 of 60% with an impressive recall of 82%. While RF captures most dropout cases by Midpoint, its precision is slightly lower than LR, reflecting a higher number of false positives. SVM ranked third, with lower recall (Early 51%, Midpoint 70%), meaning it misses more dropout cases compared to LR and RF.

MLP performed the weakest in identifying dropouts, with an Early-phase recall of 34% and Midpoint recall of 43%, detecting relatively few actual dropouts despite moderate precision.

In summary, LR is most effective for early dropout detection, whereas RF outperforms LR by the Midpoint phase, making it better suited for interventions later in the module.

5 Conclusion

The project began with a comprehensive exploration of the OULA dataset, examining each of its seven tables through initial EDA to gain insights into the data types, structure, and features involved. Following this, the objectives, project goals, and hypothesis tests were clearly defined and subsequently achieved. The primary aim was to develop a model capable of accurately identifying students at risk of dropping out, enabling timely academic interventions so that institutions can provide the necessary support, tools, and resources to encourage students to continue their studies. The problem was framed as a binary classification task, predicting whether a student is likely to persist in or withdraw from a module.

Key project tasks included:

- Data preparation and definition of the classification problem (dropout vs. continuation).
- Conducting initial EDA, merging tables, and performing data cleaning (removal of null values, duplicates, and irrelevant records).
- Feature engineering, which produced variables such as weighted score, late rate, fail rate, total clicks, and days active.
- Temporal segmentation of the dataset into four phases (Early, Midpoint, Late, and Full) to evaluate model performance across different stages and facilitate early dropout prediction.
- Hypothesis testing to better understand the dataset and provide actionable insights for proactive retention strategies.
- Development of four ML models (LR, SVM, MLP, and RF) for dropout prediction, incorporating pipeline construction and hyperparameter optimisation via GridSearch with five-fold cross-validation for each phase.
- Model evaluation using metrics such as accuracy, macro F1 score, precision, and recall per class and phase, as well as comparisons with findings from other studies utilising the OULA dataset.
- Feature importance analysis using SHAP values for the Early and Late phases to identify the most influential predictors for each model.

The subsequent sections present demographic findings, summarising the hypothesis testing results and general data insights, followed by an evaluation of model performance with rankings based on dropout detection effectiveness. Influential features are then discussed in the context of SHAP analysis, concluding with the identified limitations of the study.

5.1 Dataset Insights

5.1.1 Hypothesis Tests Summary: The engagement hypothesis tested whether students who drop out within the first 25% of the module show lower early VLE engagement than those who continue. Using the Mann–Whitney U test, a significant difference was found ($U = 29,961,144.5$, $p < 0.001$), with dropouts exhibiting substantially less activity. The very small p-value (8.82×10^{-140})

leads to rejecting the null hypothesis, confirming that low early engagement is strongly linked to higher dropout risk.

The assessment performance hypothesis examined if poor early continuous assessment scores increase dropout risk. The Mann–Whitney U test showed a significant difference ($U = 22,891,409.0$, $p < 0.001$), with dropouts scoring notably lower, supporting the link between weak early performance and increased dropout likelihood, possibly due to reduced motivation or confidence.

For the demographic disparity hypothesis, Chi-Square tests (Table 10) assessed dropout differences across age, region, education, IMD band, and disability. Significant associations were found for education, region, IMD band, and disability, but not for age. This suggests that factors like prior education, location, deprivation, and disability affect dropout rates, while age alone is not a strong predictor.

The re-enrolment hypothesis tested if students with more prior attempts at the same module are more likely to drop out. Using a Mann–Whitney U test due to non-normal data, a significant result was found ($U = 40,818,054.0$, $p = 1.10 \times 10^{-7}$), indicating that multiple previous attempts increase dropout risk, supporting the hypothesis.

5.1.2 Feature Insight Summary: The dataset is imbalanced, with 80% non-dropouts (11% distinction, 44% pass, 25% fail) and 20% dropouts. After cleaning, 27,984 students remain, excluding those who dropped out before or early in the module with minimal data.

Most students hold qualifications: 43% A Levels, 40% below A Level, 1% postgraduates, and 1% no formal qualifications. Those without qualifications have the highest withdrawal (25.3%) and failure (36.4%) rates, and lowest pass (32%) and distinction (6.2%) rates. Performance improves with higher education levels, with postgraduates showing the lowest failure (9.2%) and highest distinction (30.3%) despite slightly higher withdrawal (19.7%).

Module enrollment is uneven, with 24% in FFF and BBB, and 2.5% in AAA. Modules CCC and DDD have over half of students dropping out or failing, while EEE performs better with 13.3% distinction and 51.1% pass rates, suggesting module design impacts outcomes. Addtionally, most students (87%) take a module first time; repeat attempts are rare and linked to worse outcomes, withdrawal and failure rates rise, pass and distinction rates fall with each retake.

Under 10% of students have disabilities, who face higher dropout (28.3% vs. 18.8%) and lower pass (36.7% vs. 45.2%) and distinction rates (8.4% vs. 11%) than non-disabled students, with similar failure rates.

Regionally, the highest withdrawal rates are in West Midlands, East Midlands, and North West ($\sim$ 21.4–21.6%), lowest in South East, East Anglia, and Ireland ($\sim$ 17.8–18.3%). Failure rates peak in Wales, North West, and London, while pass and distinction rates are highest in Ireland, South East, and South. Moreover, Deprivation (IMD band) shows clear links: most deprived areas have higher withdrawal (23%) and failure (34.1%), and lower pass (36.9%) and distinction (6%) rates; least deprived (IMD band 9) areas perform best (withdrawal: 16.4% , fail: 18.4% rates, pass: 49.4% and distinction: 15.8%).

Gender is fairly balanced, 55% female, 45% male. Females show a slightly lower dropout rate (18.5% vs. 20.7%) and higher pass rate (46% vs. 43%), while failure and distinction rates are similar. Overall, gender has minimal impact on academic outcomes. Age groups show improving outcomes with age: youngest (0–35) have highest withdrawal (20.1%) and failure (26.9%), oldest (55+) the lowest withdrawal (18%) and failure (15.3%), and highest pass (47.3%) and distinction (19.3%).

Engagement metrics reveal that dropouts have low total clicks and days active, with the median days active for withdrawn students falling from 0.25 in the early phase to 0.09 by the full phase. In contrast, pass and distinction students maintain higher activity levels, ranging from 0.45 to 0.30 and 0.57 to 0.40 respectively, remaining relatively high throughout. This pattern shows that sustained engagement is closely linked to better academic outcomes, with differences between groups becoming more pronounced over time.

Weighted score results show that in the Early Phase, around 3,750 students scored between 0–5, with a peak at 70–75 (2,300 students) and about 1,000 near-perfect scores (95–100). By Midpoint, low scores (0–5) increased to 4,700, more students scored between 20 and 60, and the mode shifted to 80–85 (1,950 students), while fewer than 500 achieved perfect scores. In the Late Phase, the 0–5 group grew slightly to 4,800, with scores concentrated between 70–90 and peaking at 80–85; roughly 300 students scored 95–100. At the Full Module stage, 5,600 students remained in the 0–5 range, likely dropouts, with strong peaks between 75–90 and about 200 near-perfect scores.

Regarding late submission rates, over 11,200 students submitted on time (0–10% late), about 2,500 had high late rates (50–60%), and 4,400 submitted all assessments late. By Midpoint, on-time submissions dropped to around 10,000, with more students submitting late in the 20–80% range, especially 40–60%, indicating rising lateness. In the Late Phase, while the 0–10% late group remained largest, many students fell within the 20–60% late range, including 3,200 in the 50–60% bracket. This shows a growing trend toward late submissions. At the Full Module stage, on-time submissions were still the largest group, but the 40–60% late range expanded, and those submitting 90–100% late dropped below 3,700, reflecting increasing delays possibly linked to reduced engagement or mounting pressure.

Finally, failure rates rise as the module advances. In the Early Phase, about 16,000 students had low failure rates (below 10%), while approximately 2,100 experienced high failure rates (50–60%) and 2,900 had near-complete failure (95–100%). Several mid-range failure categories had no students, indicating a polarised distribution. By Midpoint, the number of students with low failure rates decreased to 13,000, with failures more evenly spread across the 20–40% range. Near-total failures slightly declined but remained notable. In the Late Phase, low failure rates dropped further to 10,700, while higher failure rates increased, reflecting growing academic challenges. At the Full Module stage, the group with minimal failures shrank to 10,100, while more students fell into the 70–90% failure range, and near-total failures peaked at 3,100.

With these insights, we now move to evaluating the student dropout model.

5.2 Student Dropout Model Evaluation

After reviewing the optimal hyperparameter settings for each ML model and the best macro F1 scores achieved through Gridsearch with 5-fold cross-validation. Among the models, SVM attained the

highest macro F1 scores, 65% in the Early phase and 67% in the Midpoint phase, followed by LR with 56% and 64%, RF with 48% to 59%, and finally MLP scoring the lowest with 43% to 47%.

Focusing specifically on the models' ability to accurately identify dropout students (class 0) during the Early and Midpoint phases, we assessed performance based on F1 score, precision, and recall. Overall, LR emerged as the strongest model for dropout detection, followed by RF, then SVM, with MLP performing the weakest.

LR demonstrated robust results across both phases. In the Early phase, LR achieved an F1 score of 51%, precision of 43%, and recall of 62%. By the Midpoint phase, these figures improved to an F1 score of 60%, precision of 49%, and recall of 76%. The notably high recall indicates LR's effectiveness at correctly identifying most actual dropouts, which is critical for timely intervention. The balance of precision and recall suggests LR maintains fewer false positives and false negatives compared to the other models.

RF ranked second, showing competitive recall performance, essential for dropout identification. At the Early phase, RF's F1 score was 50%, precision 43%, and recall 60%, closely mirroring LR's Early phase metrics. In the Midpoint phase, RF improved with an F1 score of 60%, precision of 47%, and an impressive recall of 82%. This high recall shows RF captures the majority of dropout cases by midpoint, although with slightly lower precision than LR. RF's strength lies in identifying more true dropouts but at the cost of more false positives.

SVM came third in dropout detection during these phases. In the Early phase, it recorded an F1 score of 48%, precision of 46%, and recall of 51%, with recall notably lower than LR and RF. By the Midpoint phase, its F1 score rose to 57%, precision to 48%, and recall to 70%. While these results are reasonable, SVM's lower recall indicates it misses more dropout cases, which is less ideal for early detection.

MLP performed the poorest for dropout detection in the Early and Midpoint phases. Its Early phase F1 was 38%, precision 43%, and recall only 34%, showing limited ability to detect dropouts. At Midpoint, it improved somewhat with an F1 of 47%, precision of 51%, and recall of 43%, but still lagged far behind the other models. Despite relatively better precision, the low recall means many actual dropouts go undetected, reducing its utility for early intervention.

In summary, LR is more effective at identifying dropouts in the Early phase with a recall of 62% compared to RF's 60%. However, by the Midpoint phase, RF surpasses LR with a recall of 82% versus 76%. Thus, LR performs better for early dropout detection, while RF takes the lead at the Midpoint phase.

5.3 Most Influential Features in ML Models

Regarding influential features in model decisions, LR identified weighted score, fail rate, the GGG code module, and studied credits as the top factors in both the Early and Midpoint phases. SVM's key features included fail rate, weighted score, studied credits, late rate, and days active. MLP's most important features were weighted score, late and fail rate, and days active. RF's decisions were influenced most by weighted score, fail and late rate, GGG code module, and days active. Across models, weighted score, late and fail rates, studied credits, and days active stood out as the most significant predictors, with demographic features playing only a minor role.

5.4 Limitations

While the project achieved its goals, it also encountered notable limitations and challenges. One major limitation lies in the OULA dataset. Although it offers valuable insights for predicting student dropout, it originates from a single institution within a single online learning environment, which may limit the broader applicability of the results. In addition, the dataset lacks several key variables typically included in university records, such as attendance rates, seminar participation, and financial details like tuition payment history, scholarships, or bursary status, along with records of suspensions due to behavioural or mental health issues. Given that mental health is a leading cause of dropouts in the UK [5], the absence of these factors could substantially impact the accuracy of student retention predictions.

Additionally, the dataset is imbalanced, with an 80–20 split between non-dropouts and dropouts. Although this was addressed by using macro F1 to give greater weight to the dropout class, most models still exhibited low precision and recall for dropout predictions. Similar imbalances were also present across features such as module code, presentation, region, age band, and highest education level.

Another important limitation is the absence of exam results in the dataset. For modules where final exams are the primary basis for passing or failing, this missing information restricts the accuracy of the full-phase analysis and the overall reliability of the dataset.

From a modelling perspective, SVM and MLP were challenging to evaluate due to the high computational cost of extensive hyperparameter tuning. This led to a reduced search space and longer processing times. For SHAP analysis of the SVM model, only a small subset of 50 training instances was used with KernelExplainer to manage computational demands, providing an approximate—though not exact—representation of feature importance.

The temporal segmentation approach, while helpful for early prediction, may have struggled to capture complex patterns of dropout risk developing throughout the course. Finally, model performance itself was a constraint: in the early phase, the best model (LR) achieved a dropout recall of 62% but a precision of only 43%. At the midpoint phase, recall improved substantially (82% for RF) but precision still remained under 50%.

5.5 Future Work

Future research could address the current limitations in several ways. Expanding the dataset to include multiple institutions and a variety of online learning platforms would enhance the generalisability of results. To better handle class imbalance, techniques such as synthetic oversampling (e.g., SMOTE), which generates new minority class samples through linear interpolation between a minority class instance and one of its k nearest neighbours [23], or class-weight adjustments could be applied. These methods would also help mitigate imbalance across under-represented demographic and module-related categories. Furthermore, integrating final examination scores would provide a more comprehensive picture of student performance, particularly for modules where passing depends primarily on the final exam.

From a modelling perspective, future studies could investigate computationally efficient algorithms or leverage distributed computing frameworks, enabling broader hyperparameter exploration

for resource-intensive models such as SVM and MLP. Additionally, more granular temporal segmentation could capture complex dropout patterns over time, improving accuracy across different course stages. Finally, hybrid prediction approaches that combine early-phase indicators with continuously updated learning data may yield models capable of both timely and precise dropout risk detection.

6 Acknowledgements

This research received support during the MSC Data Science course, instructed by Professor Felipe Campelo, PhD at the School of Engineering Mathematics and Technology, University of Bristol.

References

[1] G. Crosling, M. Heagney, and L. Thomas, "Improving student retention in higher education: Improving teaching and learning," *Australian Universities' Review*, vol. 51, no. 2, pp. 9–18, 2009. [Online]. Available: https://files.eric.ed.gov/fulltext/EJ864028.pdf.

[2] M. R. Marcolino, T. R. Porto, T. T. Primo, *et al.*, "Student dropout prediction through machine learning optimization: Insights from moodle log data," *Scientific Reports*, vol. 15, p. 9840, 2025. DOI: 10.1038/s41598-025-93918-1. [Online]. Available: https://doi.org/10.1038/s41598-025-93918-1.

[3] Á. Kocsis and G. Molnár, "Factors influencing academic performance and dropout rates in higher education," *Oxford Review of Education*, vol. 51, no. 3, pp. 414–432, 2024. DOI: 10.1080/03054985.2024.2316616. [Online]. Available: https://doi.org/10.1080/03054985.2024.2316616.

[4] OECD, *Education at a Glance 2019: OECD Indicators*. Paris: OECD Publishing, 2019. DOI: 10.1787/f8d7880d-en. [Online]. Available: https://doi.org/10.1787/f8d7880d-en.

[5] J. Bryson. "University dropout rates reach new high, figures suggest." BBC News. (Sep. 28, 2023), [Online]. Available: https://www.bbc.co.uk/news/education-66940041.

[6] C. Foster and P. Francis, "A systematic review on the deployment and effectiveness of data analytics in higher education to improve student outcomes," *Assessment & Evaluation in Higher Education*, vol. 45, no. 6, pp. 822–841, 2019. DOI: 10.1080/02602938.2019.1696945. [Online]. Available: https://doi.org/10.1080/02602938.2019.1696945.

[7] 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. 103 676, 2020, ISSN: 0360-1315. DOI: 10.1016/j.compedu.2019.103676. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360131519302295.

[8] H. Waheed, S.-U. Hassan, N. R. Aljohani, J. Hardman, S. Alelyani, and R. Nawaz, "Predicting academic performance of students from vle big data using deep learning models," *Computers in Human Behavior*, vol. 104, p. 106 189, 2020, ISSN: 0747-5632. DOI: 10.1016/j.chb.2019.106189. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0747563219304017.

[9] M. Adnan, A. Habib, J. Ashraf, *et al.*, "Predicting at-risk students at different percentages of course length for early intervention using machine learning models," *IEEE Access*, vol. 9, pp. 7519–7539, 2021. DOI: 10.1109/ACCESS.2021.3049446.

[10] H. P. Singh and H. N. Alhulail, "Predicting student-teachers dropout risk and early identification: A four-step logistic regression approach," *IEEE Access*, vol. 10, pp. 6470–6482, Jan. 2022. DOI: 10.1109/ACCESS.2022.3141992.

[11] F. Lee. "Logistic regression." Accessed: 2025-08-04, IBM. (May 14, 2025), [Online]. Available: https://www.ibm.com/think/topics/logistic-regression.

[12] "1.4. support vector machines." Accessed: 2025-08-04, scikit-learn. (2025), [Online]. Available: https://scikit-learn.org/stable/modules/svm.html.

[13] GeeksforGeeks. "Multilayer feedforward neural network in data mining." Accessed: 2025-08-04, GeeksforGeeks. (Sep. 7, 2022), [Online]. Available: https://www.geeksforgeeks.org/machine-learning/multilayer-feed-forward-neural-network-in-data-mining/.

[14] L. Barreñada, P. Dhiman, D. Timmerman, A.-L. Boulesteix, and B. V. Calster, "Understanding overfitting in random forest for probability estimation: A visualization and simulation study," *Diagnostic and Prognostic Research*, vol. 8, no. 1, Sep. 2024. DOI: 10.1186/s41512-024-00177-1. [Online]. Available: https://doi.org/10.1186/s41512-024-00177-1.

[15] F. Tang and H. Ishwaran, "Random forest missing data algorithms," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 10, no. 6, pp. 363–377, Jun. 2017. DOI: 10.1002/sam.11348. [Online]. Available: https://doi.org/10.1002/sam.11348.

[16] J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open university learning analytics dataset," *Scientific Data*, vol. 4, p. 170 171, 2017. DOI: 10.1038/sdata.2017.171. [Online]. Available: https://doi.org/10.1038/sdata.2017.171.

[17] scikit-learn developers. "Standardscaler." scikit-learn. (2025), [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html (visited on 07/26/2025).

[18] A. Géron and P. E. Central, *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems*, eng, Third edition. Sebastopol: O'Reilly Media, Incorporated, 2022–2023, Referenced: Chapter 2, p. 75 – End-to-End Machine Learning Project, ISBN: 9781098122461.

[19] E. McClenaghan. "Mann-whitney u test: Assumptions and example." Accessed: 2025-07-31. (Jul. 6, 2022), [Online]. Available: https://www.technologynetworks.com/informatics/articles/mann-whitney-u-test-assumptions-and-example-363425.

[20] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17, Long Beach, California, USA: Curran Associates Inc., 2017, pp. 4768–4777, ISBN: 9781510860964.

[21] A. Cooper. "Explaining machine learning models: A non-technical guide to interpreting shap analyses." Accessed: Aug. 07, 2025. (Nov. 2021), [Online]. Available: https://www.aidancooper.co.uk/a-non-technical-guide-to-interpreting-shap-analyses/.

[22] K. De Angeli, S. Gao, I. Danciu, *et al.*, "Class imbalance in out-of-distribution datasets: Improving the robustness of the textcnn for the classification of rare cancer types," *Journal of Biomedical Informatics*, vol. 125, p. 103 957, Jan. 2022, Epub 2021 Nov 22. DOI: 10.1016/j.jbi.2021.103957.

[23] D. Elreedy, A. F. Atiya, and F. Kamalov, "A theoretical distribution analysis of synthetic minority oversampling technique (smote) for imbalanced learning," *Machine Learning*, vol. 113, pp. 4903–4923, 2024. DOI: 10.1007/s10994-022-06296-4. [Online]. Available: https://doi.org/10.1007/s10994-022-06296-4.

7 Appendix

7.1 Student Dropout Features Correlation Matrix

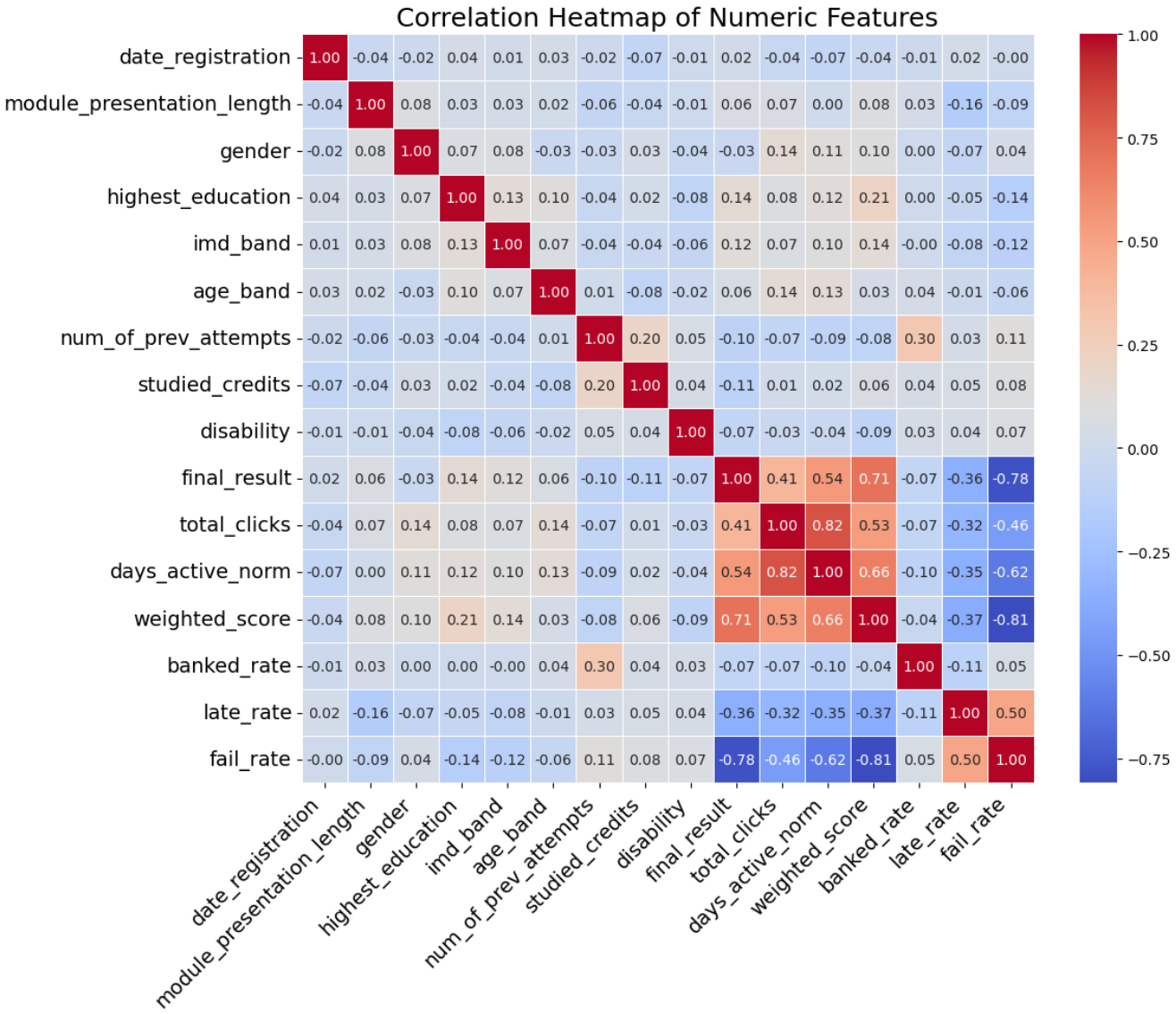

Figure 34. Correlation Matrix of Dropout Features.