Student Dropout Predictor

Author: Ali Suhail
Student ID: 2605549
Date: 17th July 2025

August 5, 2025

Abstract :

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Praesent porttitor arcu luctus, imperdiet urna iaculis, mattis eros. Pellentesque iaculis odio vel nisl ullamcorper, nec faucibus ipsum molestie. Sed dictum nisl non aliquet porttitor. Etiam vulputate arcu dignissim, finibus sem et, viverra nisl. Aenean luctus congue massa, ut laoreet metus ornare in. Nunc fermentum nisi imperdiet lectus tincidunt vestibulum at ac elit. Nulla mattis nisl eu malesuada suscipit. Aliquam arcu turpis, ultrices sed luctus ac, vehicula id metus. Morbi eu feugiat velit, et tempus augue. Proin ac mattis tortor. Donec tincidunt, ante rhoncus luctus semper, arcu lorem lobortis justo, nec convallis ante quam quis lectus. Aenean tincidunt sodales massa, et hendrerit tellus mattis ac. Sed non pretium nibh. Donec cursus maximus luctus. Vivamus lobortis eros et massa porta porttitor.

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

1 Introduction (2)

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

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

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

2 Background (5)

2.1 Literature Review

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

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

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

2.2 Related Work

sdfdfds

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 OULAD 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 exploratory data analysis (EDA) is performed to uncover patterns, assess the dataset's structure, and examine variable types.

Next, Four ML models: Logistic Regression (LR), Support Vector Machines (SVM), Multi-Layer Perceptron Classifier (MLP) and Random Forest (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 [CITE]. 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.

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

Table 1
Summary of Project Objectives

Objective	Description
Data Preparation	Clean, preprocess, and engineer features from the OULAD dataset to ensure they are suitable for ML training and evaluation.
EDA	Conduct detailed exploratory data analysis (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.

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 covers the relationship between certain groups (by age band, region, education, IMD band, and disability) are overrepresented among dropouts. Factors like low income or IMD band region would increase 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 was used 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 [CITE] of the Data Methods. The following section 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 [6]. It performs well when the relationship between input features and target classes is approximately linear [7]. 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 [8], which are common in educational data mining. Their ability to define complex decision boundaries via kernel functions, such as the Radial Basis Function (RBF) or polynomial kernels, allows them to model non-linear relationships in the data [8]. 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 [8], 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 [9]. 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 [10]. It is particularly effective at handling non-linear relationships, missing values, and mixed data types (numerical and categorical) [11]. Random Forest 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[10], 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

We selected models from distinct algorithmic categories, linear (LR), kernel-based (SVM), neural (MLP), and ensemble tree-based (RF), to capture a wide spectrum of learning behaviours and model capacities. This diversity allows us to:

- Comparison between linear and non-linear models in predicting student dropout.
- Analysis of model performance across different temporal subsets of the data (early, midpoint, late, and full phases).
- 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 (10)

3.1 Data Description

The OULAD 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 Virtual Learning Environment (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 [12].

Figure 1 illustrates the OULAD 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`.

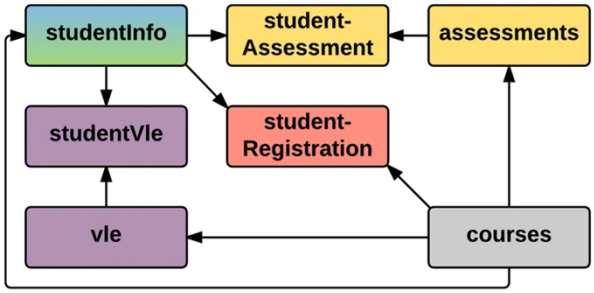

Figure 1. Student dropout dataset structure. Source: [12].

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`.

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 OULAD dataset originally consists of seven tables, which were first combined into three primary tables:

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.

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:

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

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:

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

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 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. 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 [12].

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.

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 [12]. 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

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

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

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 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 inner join into a single, comprehensive dataset, as shown in Figure 2.

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:

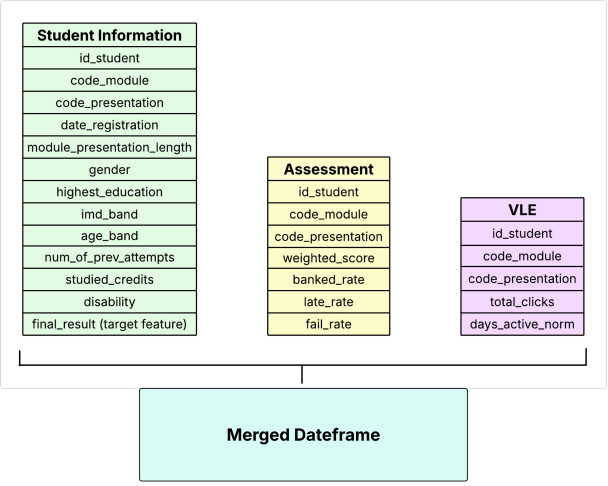

Figure 2. Final integrated dataset after merging VLE, assessment, and student information tables.

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 3 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 exploratory data analysis and serve as benchmarks, the main emphasis is placed on enhancing dropout prediction performance using the Early and Midpoint datasets.

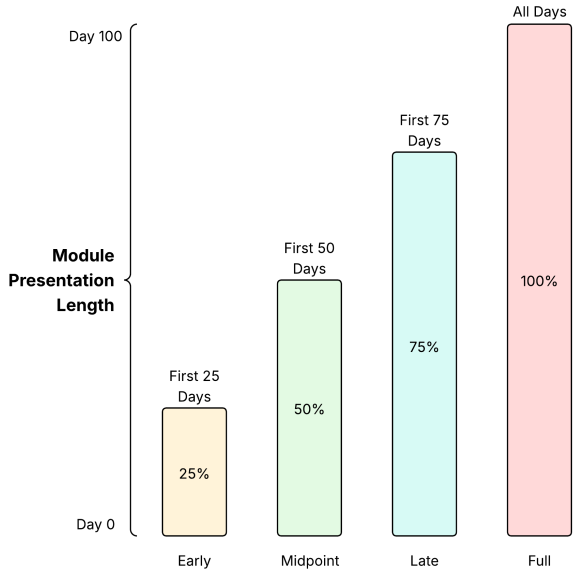

Figure 3. Diagram illustrating the temporal splits of the dataset across the module timeline.

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 exploratory data analysis (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 own 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 [13]. 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 5. Without scaling, features with larger ranges could disproportionately influence the model [14].

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 with the objective of predicting dropout versus continuation.

3.4 Exploratory Data Analysis Findings

Before applying 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 4, the majority of students registered between 25 and 100 days prior to the module start date, suggesting that most students enrolled in a timely manner. 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.

In Figure 5, 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.

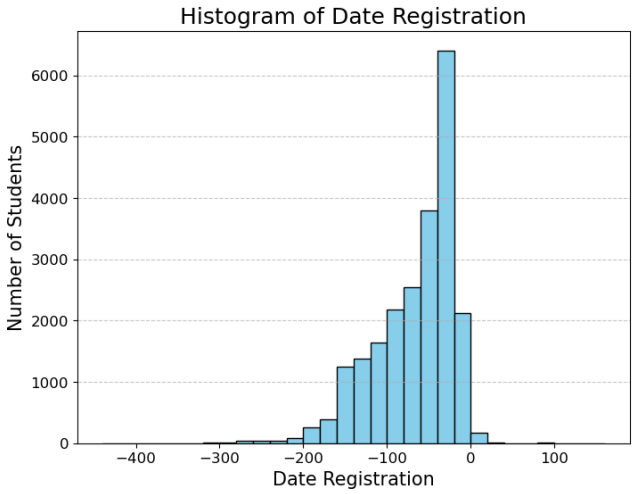

Figure 4. Histogram of Date Registration.

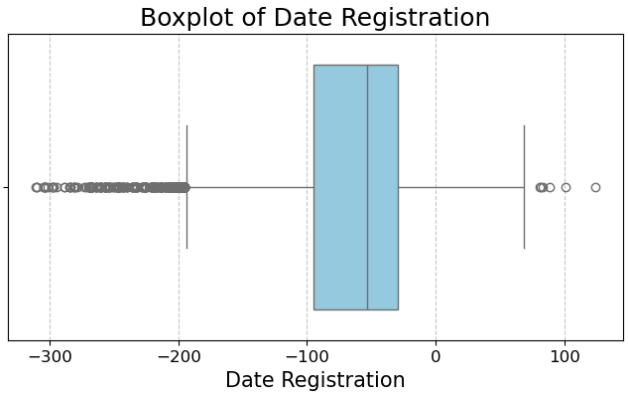

Figure 5. Boxplot of Date Registration.

3.4.2 Code Module and Presentation: Figure 6 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 7, which presents the 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

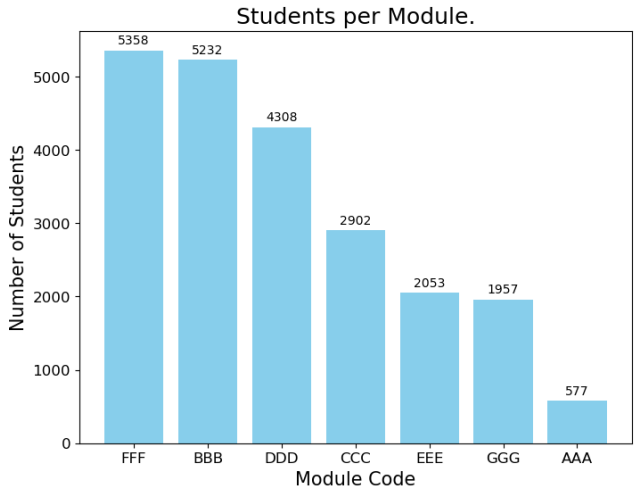

Figure 6. Students Per Module

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 stronger 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.

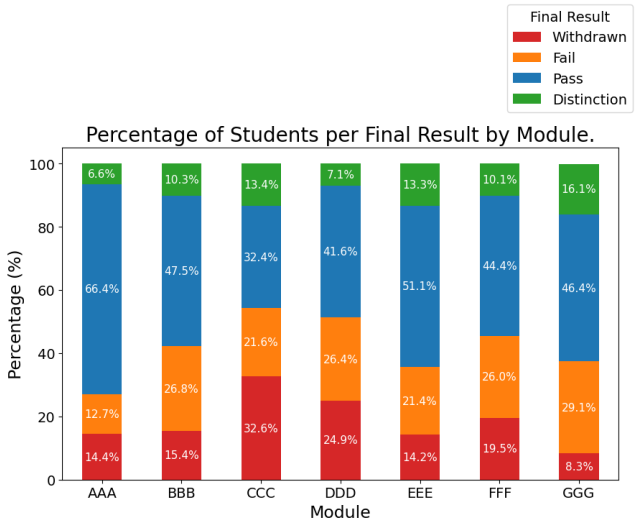

Figure 7. Percentage of Students per Final Result by Module.

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 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 8, 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.

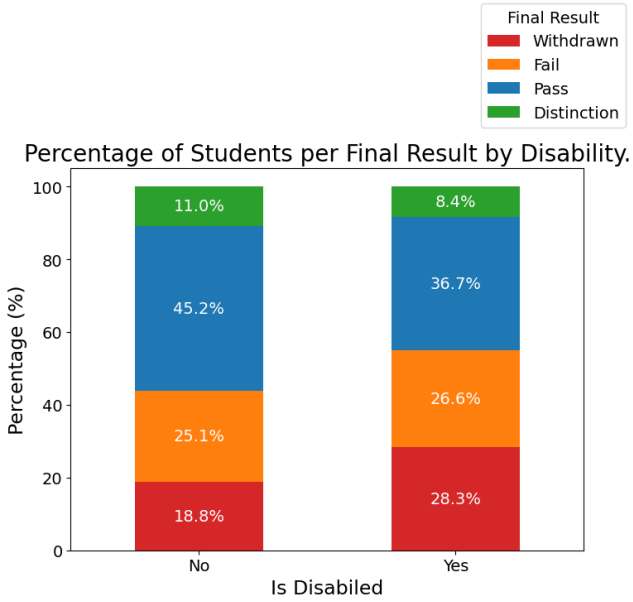

Figure 8. Percentage of Students per Final Result by Disability.

3.4.5 Region: As for the regions where the students are from in the dataset in Figure 9, Scotland has the highest proportion of students, making up around 11% of the dataset, while Ireland has the smallest share with just 928 students, representing around 4%. Other regions like London also have notable representation, contributing approximately 9% of the total. Next, let us find out the droput rate per region.

According to Figure 10, 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 stu-

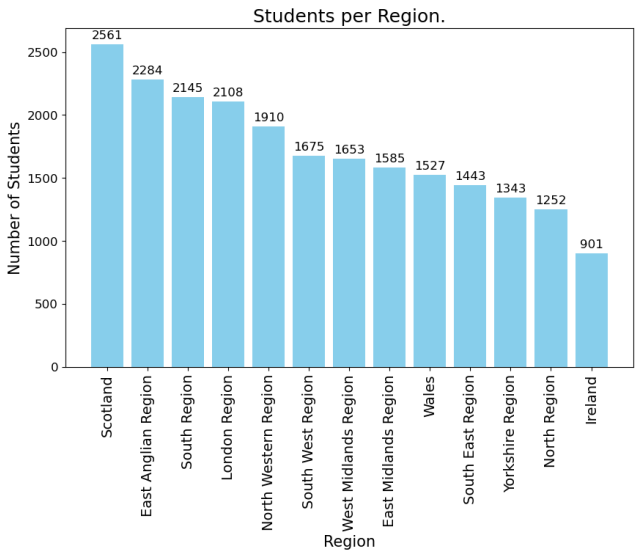

Figure 9. Students per Region.

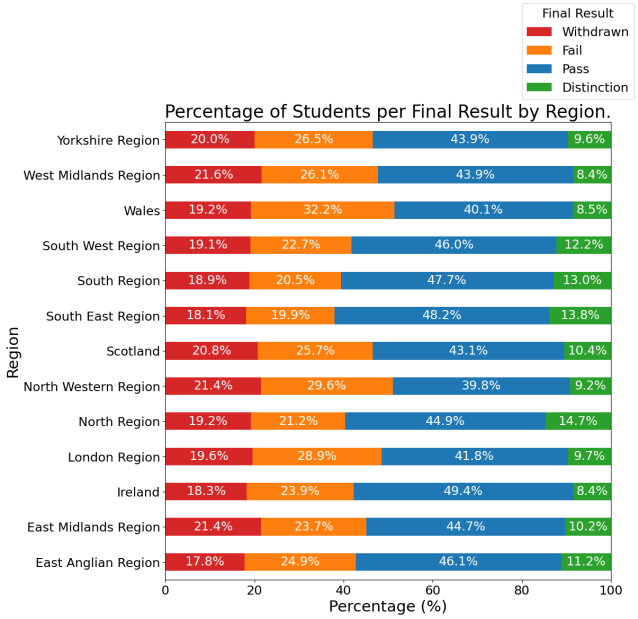

Figure 10. Percentage of Students per Final Result by Region.

dent 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.

As illustrated in Figure 11, 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 12, 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 13 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 (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).

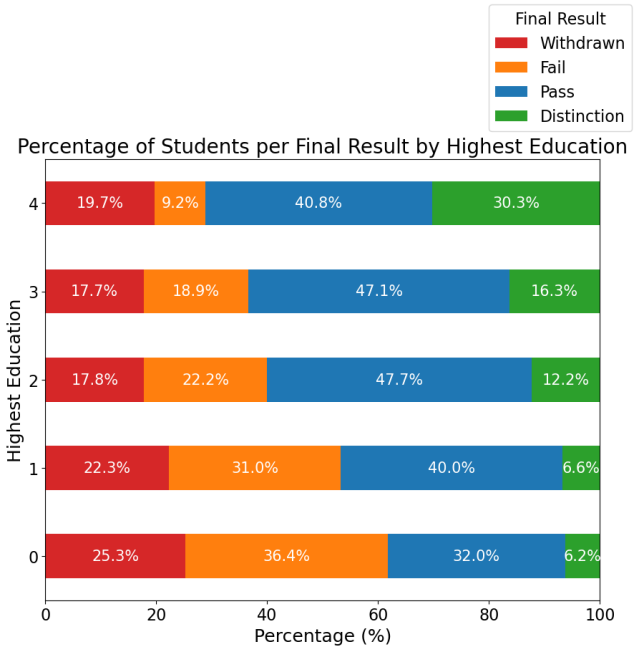

Figure 11. Percentage of Students per Final Result by Highest Education.

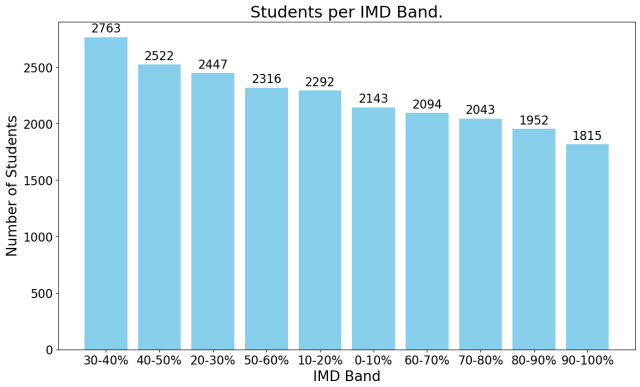

Figure 12. Students per IMD Band.

Figure 14 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 (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.

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,

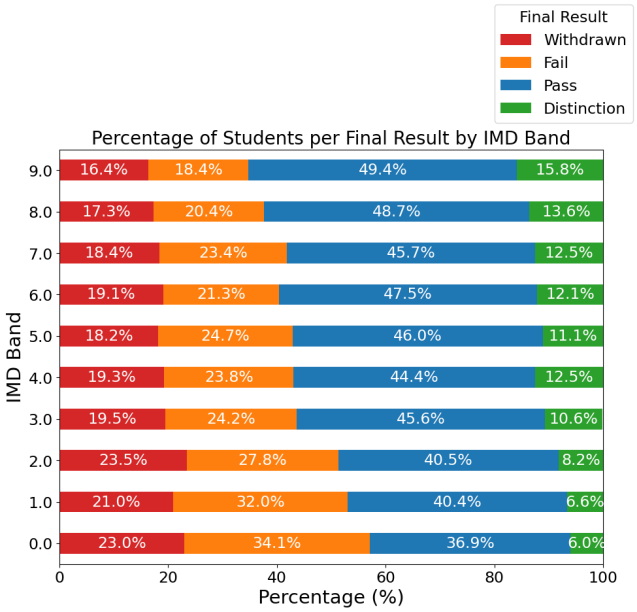

Figure 13. Percentage of Students per Final Result by IMD Band.

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 15, most students have around 60 credits for their module. While credit values extend up between 200 and 650, such high values are extremely rare.

When we check a box and whisker plot of 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 16.

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 17, the number of students with total VLE clicks in the 0 to 1,000 range gradually decreases 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

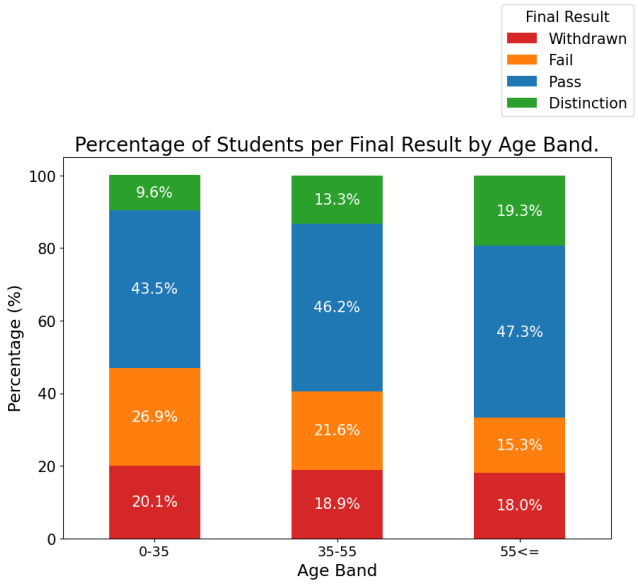

Figure 14. Percentage of Students per Final Result by Age Band.

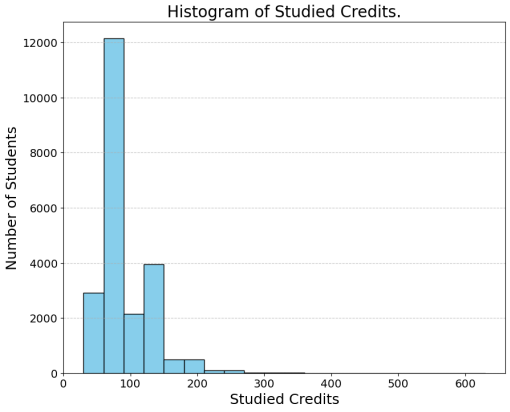

Figure 15. Histogram of Studied Credits.

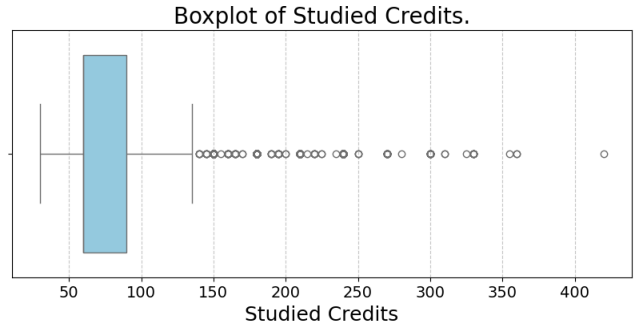

Figure 16. Boxplot of Studied Credits.

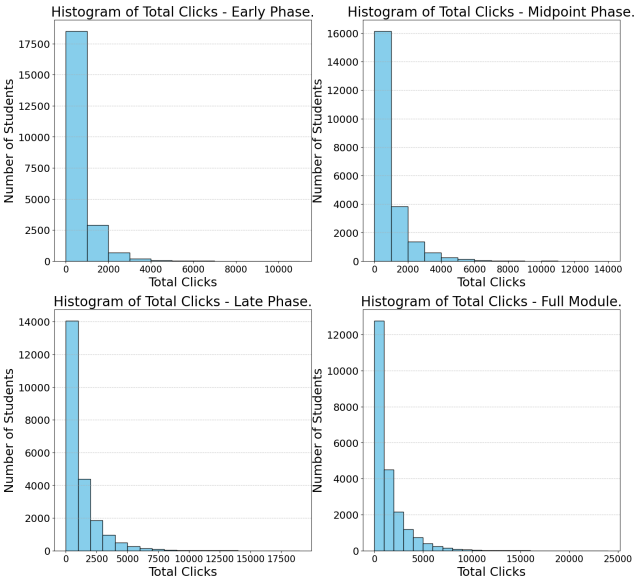

Figure 17. Histogram of Total Clicks for Each Phase.

clicks remain relatively uncommon, even in the later phases.

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 18 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 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.

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,

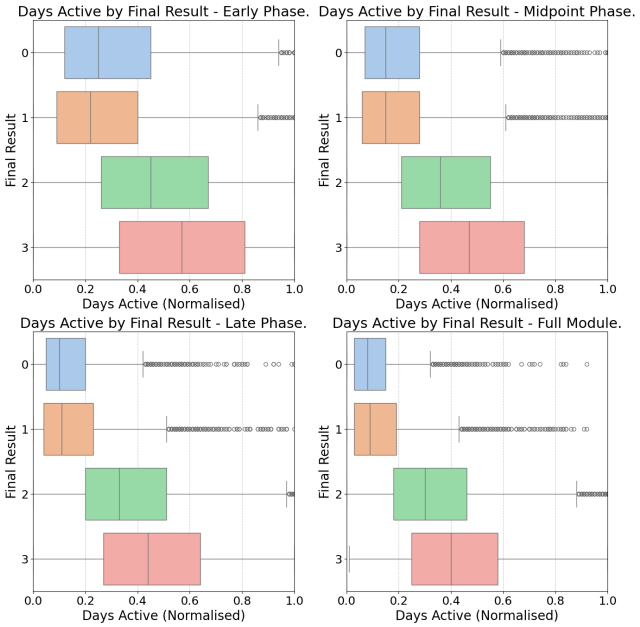

Figure 18. Days Active (Normalised) By Final Result For Each Phase.

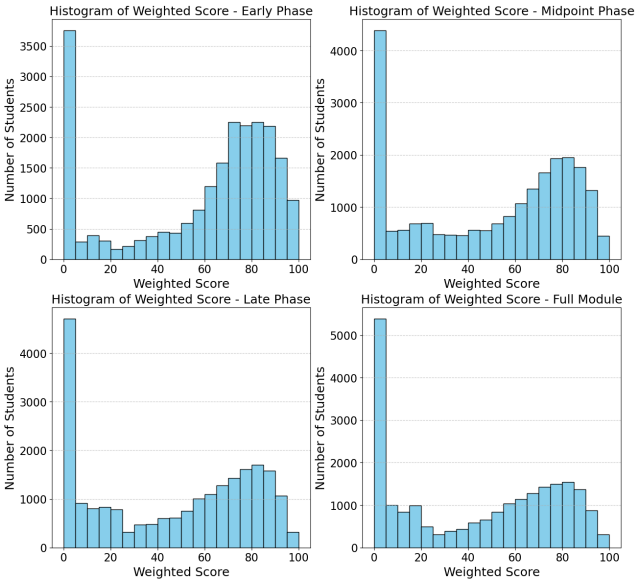

Figure 19. Histogram of Weighted Score for Each Phase.

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 19 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.

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 20 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.

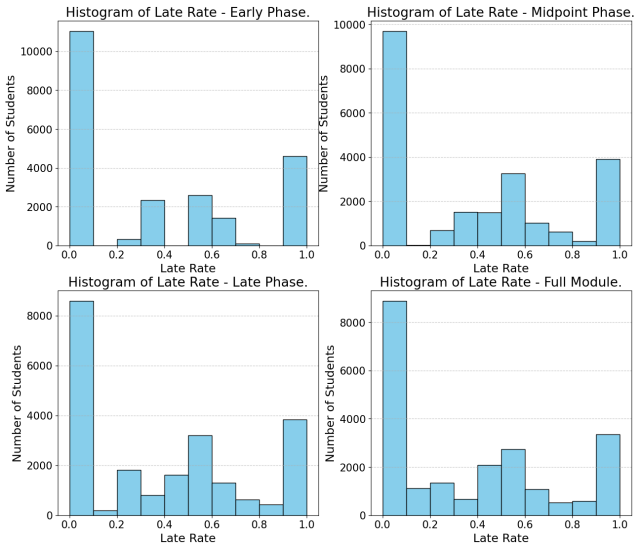

Figure 20. Histogram of Late Rate for Each Phase.

3.4.16 Fail Rate: Figure 21 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.

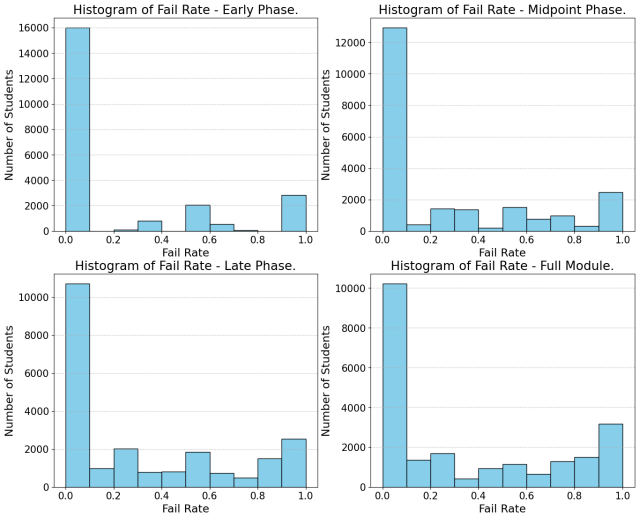

Figure 21. Histogram of Fail Rate for Each Phase.

3.4.17 Final Result: The `final_result` serves as the target variable in our analysis. As shown in Figure 22, 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.

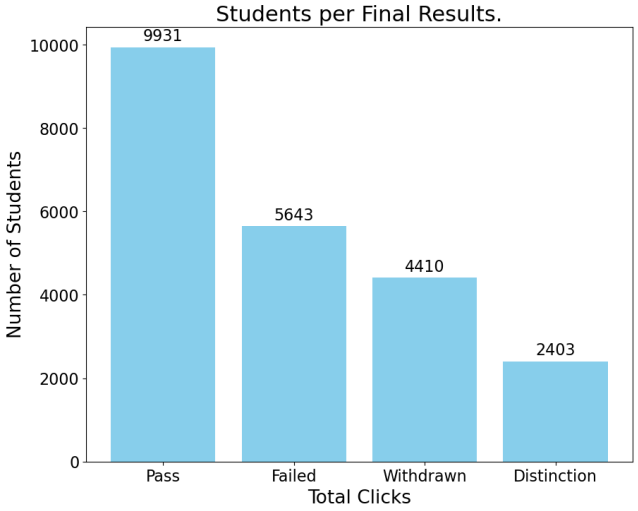

Figure 22. Students per Final Result.

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 24).

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.

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 [15].

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 24.

Key findings also emerge from the correlation matrix (Figure 24):

- **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 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 23 summarises the overall model implementation process.

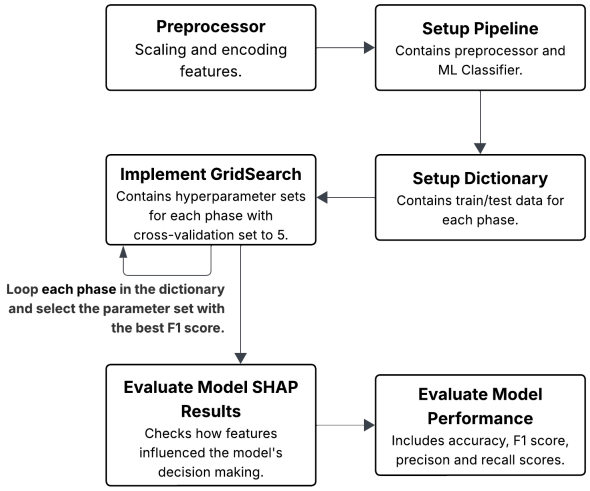

Figure 23. Model implementation and evaluation steps.

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 im-

portance was assessed using SHAP (SHapley Additive exPlanations). SHAP is a game-theoretic method that assigns each feature an importance value based on its contribution to the model's output [CITE]. 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 [CITE]. One major advantage of SHAP is its model-agnostic property, meaning it can be applied to any ML model, 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. 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 [CITE]. 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 machine learning 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. Exploratory data analysis 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.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

[8] "1.4. support vector machines." Accessed: 2025-08-04, scikit-learn. (2025), [Online]. Available: https://scikit-learn.org/stable/modules/svm.html.

[9] 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/.

[10] 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.

[11] 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.

[12] 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.

[13] 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).

[14] 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.

[15] 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.

4 Appendix

4.1 Student Dropout Features Correlation Matrix

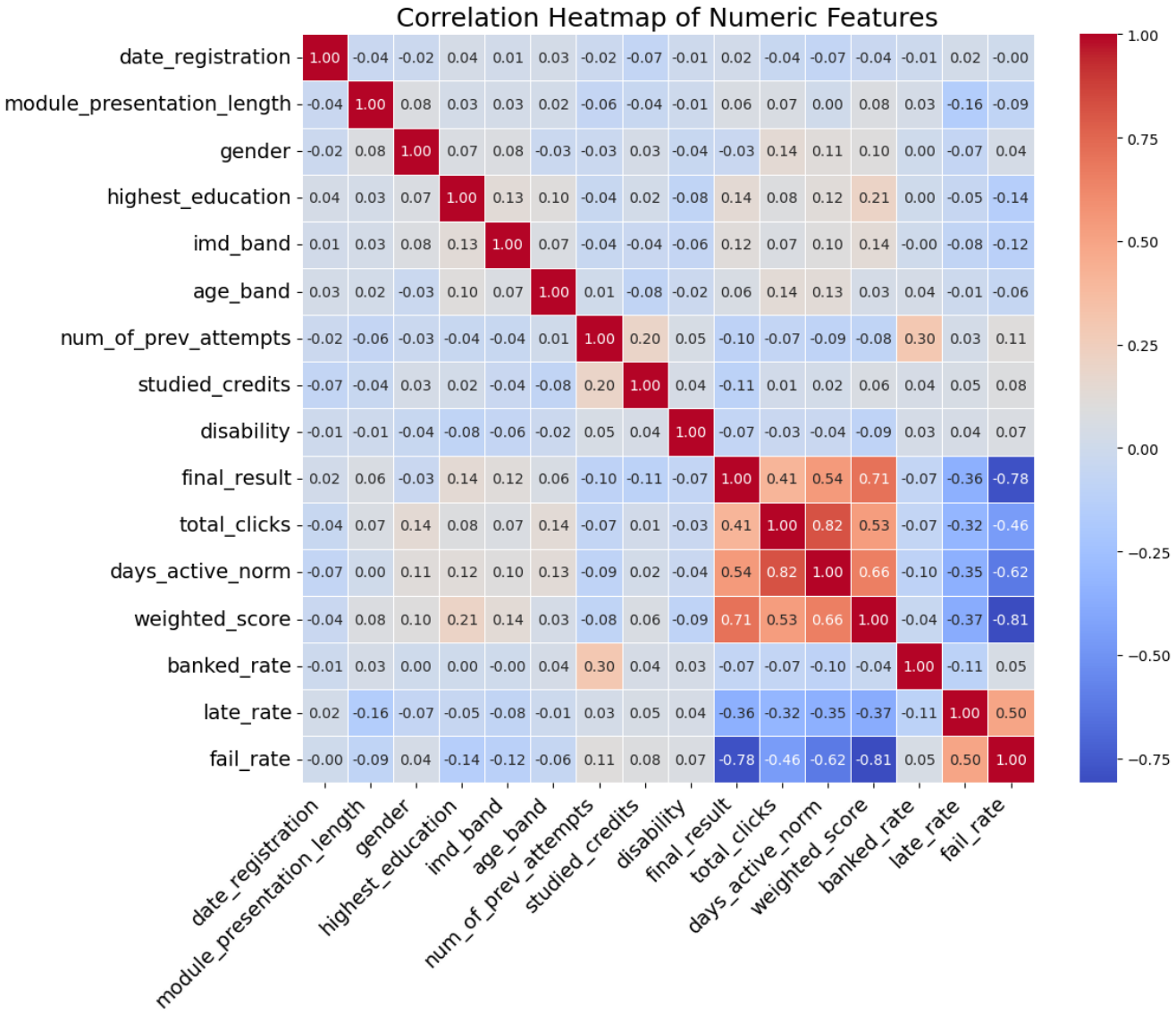

Figure 24. Correlation Matrix of Dropout Features.