Student Dropout Predictor

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

July 27, 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 Objectives and Hypothesis

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

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

2.2.1 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:

- Compare the effectiveness of linear versus non-linear models in predicting student dropout.
- Evaluate how model performance varies under different data conditions (e.g., early vs full datasets).
- Examine trade-offs between predictive performance and model interpretability.
- Assess each model's ability to generalise to unseen data.

The primary objective is to identify students who are likely to withdraw accurately. Accordingly, greater emphasis is placed on recall (and precision) for the dropout class, as correctly flagging at-risk students is essential for enabling timely interventions and academic support.

All models are tuned using GridSearchCV for hyperparameter optimisation, paired with 5-fold cross-validation to ensure stable and generalisable performance estimates. Each model is then evaluated across four temporal subsets of the dataset: early, midpoint, late, and full, to examine how predictive performance changes throughout a module. The early and midpoint datasets are of particular interest, as they represent periods where intervention is most effective. In contrast, the late and full datasets serve as benchmarks for assessing the maximum achievable predictive performance.

3 Methods (10)

3.1 Data Description

The dataset used in this project is the publicly available and anonymised Open University Learning Analytics Dataset (OULAD). It includes information on courses, students, and their interactions with the Virtual Learning Environment (VLE) across seven selected modules. These modules are delivered in two presentation periods: February and October, labelled as "B" and "J" respectively. The dataset is organised into multiple CSV files, each representing a table linked through unique identifiers. Further details can be found in [6]. After cleaning and preprocessing, the dataset comprises 27,984 students, with 19 selected features. These features are detailed in Table 1.

Table 1
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 enrollment.
`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.

The following 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. For example, Table 2 illustrates how the time-series dataset captures a student's assessment activity.

Table 2
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, or achieved distinction. According to the data specifications, a score < 40 is a fail [6]. 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 3:

Table 3
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

As illustrated in Table 3, 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 2, 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 3.

3.3 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 4:

Table 4
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 4 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 3). 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.

3.4 Processing Dataset

Following feature engineering and the merging of relevant tables, the 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.

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

Table 5
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.

3.4.2 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.

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.

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.

3.4.3 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.

3.4.4 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.

3.4.5 Scaling and Encoding the Dataset As detailed in Table 6, 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 logistic regression and SVM require numerical inputs and cannot handle raw categorical strings.

Table 6
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 [7]. 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 [8].

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.

3.5 Exploratory Data Analysis Findings and Hypothesis Testing

3.5.1 Exploratory Data Analysis Findings Before applying machine learning models, it is essential to gain an initial understanding of the dataset by exploring each feature individually.

3.5.2 Date registration As shown in Figure 1, 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.

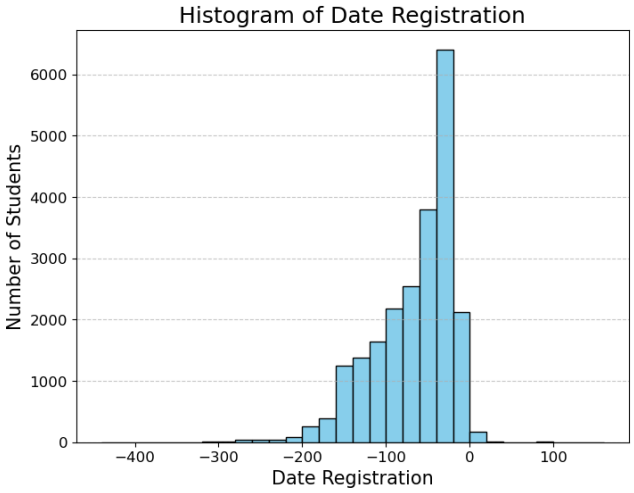

Figure 1. Histogram of Date Registration.

In Figure 2, 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.

3.5.3 Code Module and Presentation Figure 3 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 4, 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

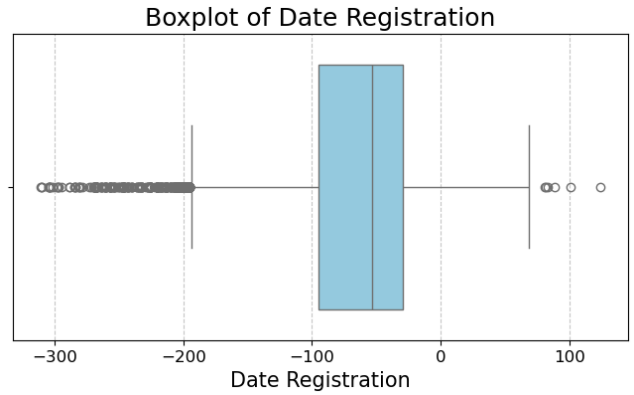

Figure 2. Boxplot of Date Registration.

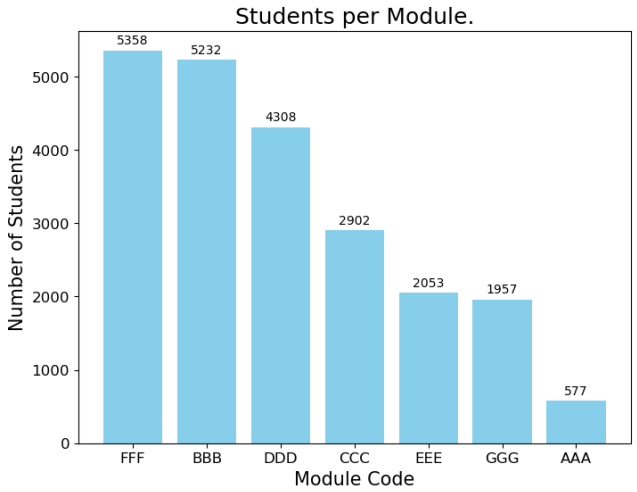

Figure 3. Students Per Module

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 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.

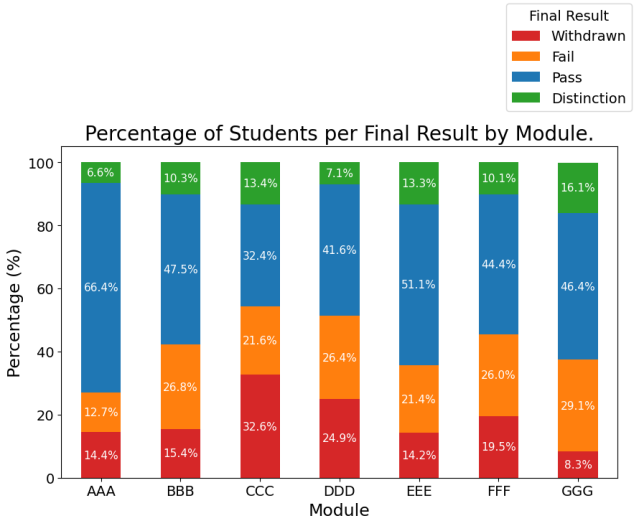

Figure 4. Percentage of Students per Final Result by Module.

3.5.4 Gender The gender distribution is relatively balanced, with approximately 55% of students being female and 45% male. As shown in Figure 5, 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.

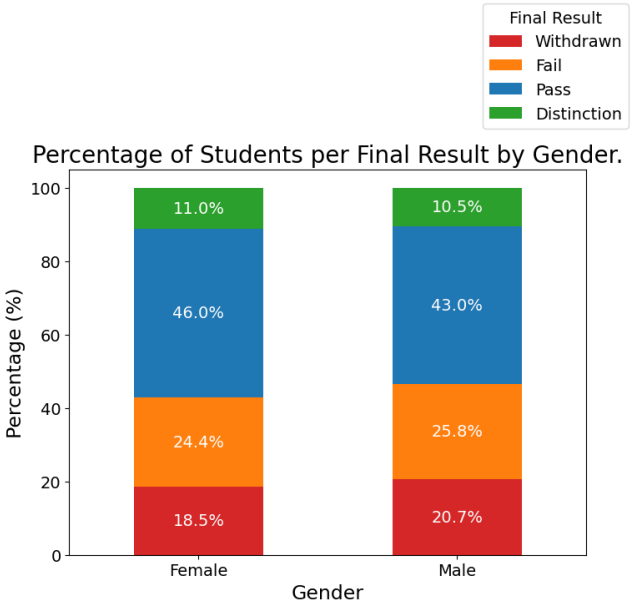

Figure 5. Percentage of Students per Final Result by Gender.

3.5.5 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 6, 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.

3.5.6 Region As for the regions where the students are from in the dataset in Figure 7, 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 dropout rate per region.

According to Figure 8, 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

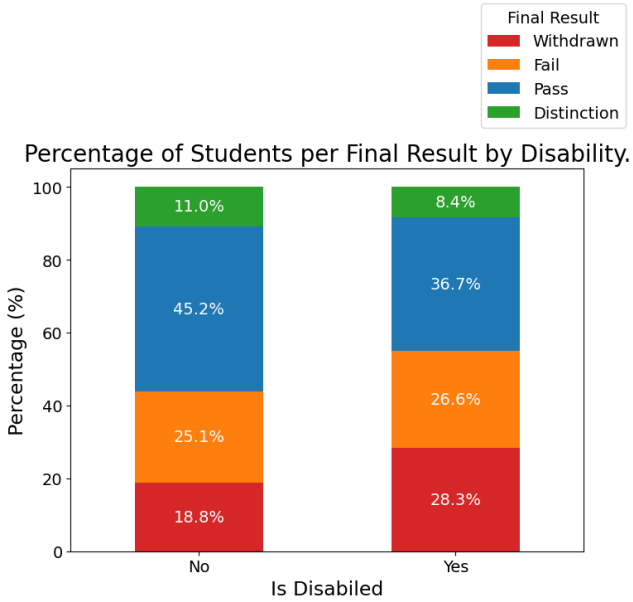

Figure 6. Percentage of Students per Final Result by Disability.

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%).

3.5.7 Highest Education As for the highest education level, the majority of students in the dataset hold some form of qualification. The largest group has A Level or equivalent qualifications, accounting for 43%, closely followed by those with qualifications lower than A Level at around 40%. Postgraduates and those with no formal qualifications represent the smallest groups, each making up just 1% of the total.

Moreover, it was found that using Figure 9 that students without formal qualifications (0) have the highest withdrawal rate at 25.3% and also the highest failure rate at 36.4%. Their pass rate is the lowest at 32%, with only 6.2% achieving a distinction. Those with qualifications lower than A level (1) see a decrease in withdrawal (22.3%) and failure (31%), with a pass rate increase to 40% and a slight improvement in distinctions at 6.6%. Students holding A level or equivalent qualifications (2) have even lower withdrawal

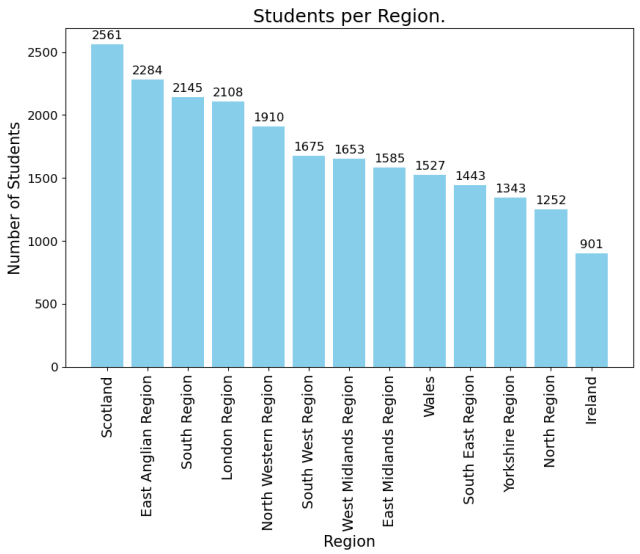

Figure 7. Students per Region.

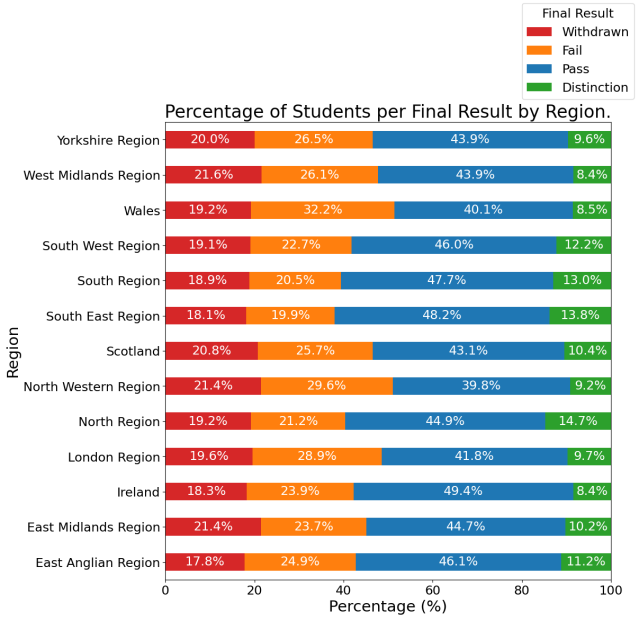

Figure 8. Percentage of Students per Final Result by Region.

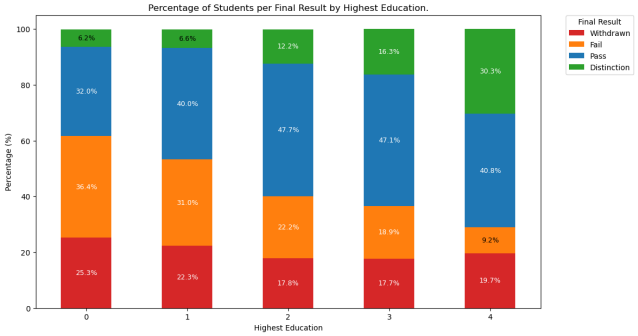

Figure 9. Percentage of Students per Final Result by Highest Education.

(17.8%) and failure rates (22.2%), while their pass rate rises to 47.7% and distinctions to 12.2

Those with a higher education qualification (3) experience a small reduction in withdrawals (17.7%) and failures (18.9%), with a slighly smaller pass rate around 47.1% and distinctions increasing to 16.3%. Finally, postgraduate qualification holders (4) perform slighly worse than level 3 in terms of withdrawal rate (19.7%) and pass rate (40.8%), but have the lowest failure rates (9.2%), and the highest distinction rate at 30.3%. Overall, higher prior educational attainment is associated with better academic outcomes and lower dropout rates.

3.5.8 IMD Band The next feature is the IMD band and as we can see in Figure 10 the largest proportion of students fall into IMD band 3 (30–40%), making up around 12% of the training set, while band 9 (90–100%) is the least common, accounting for about 8%. Additionally, around 10% of students come from the most deprived areas, classified in the 0–10% IMD band.

Moreover, as shown in Figure 11 Students in the most deprived areas (IMD 0.0) have the highest withdrawal rate at 23% and the highest failure rate at 34.1%, with the lowest pass (36.9%) and distinction rates (6%). As deprivation decreases (higher IMD bands), withdrawal and failure rates gradually decline, while pass and distinction rates increase. For example, students in the least deprived areas (IMD 9.0) have the lowest withdrawal rate at 16.4% and failure rate at 18.4%, alongside the highest pass rate (49.4%) and distinction rate (15.8%).

Overall, there is a clear trend showing that students from less deprived areas tend to perform better academically, with fewer dropouts and failures and more passes and distinctions.

3.5.9 Age Band The majority of students are aged between 0–35, making up about 70% of the dataset. This is followed by the 35–55 age group, which accounts for approximately 30%. The 55 and above age group is the smallest, with fewer than 1% of students, or 150 students.

As shown in Figure 12, we see that younger students aged 0-35 (Age Band = 0) have the highest withdrawal rate at 20.1% and fail rate at 26.9%, with lower pass (45.5%) and distinction rates (9.6%). Moreover, Students aged 35-55 (Age Band = 1) show better outcomes with a lower withdrawal rate of 18.9% and fail rate of 21.6%, while pass and distinction rates improve to 46.2% and 13.3%, respectively. The oldest group, aged 55 and above (Age Band = 2), performs best overall, having the lowest withdrawal rate (18%) and fail rate (15.3%), slightly higher pass rate than the 35-55 age band at 47.3% and distinction rates (19.3%). But keep in mind that this age group only represent less than 1% of the train set. This indicates that older students tend to achieve better results and have lower dropout rates compared to younger students.

3.5.10 Number of Previous Module Attempts The vast majority of students (around 87%) are taking their module for the first time. About 10% have attempted it once before, while multiple reattempts are rare. Only 11 students have taken the module five times, and 3 students have attempted a module six times.

In Figure 13, Students with no previous attempts have the best outcomes, with a withdrawal rate of 19.2%, fail rate of 23.4%, pass

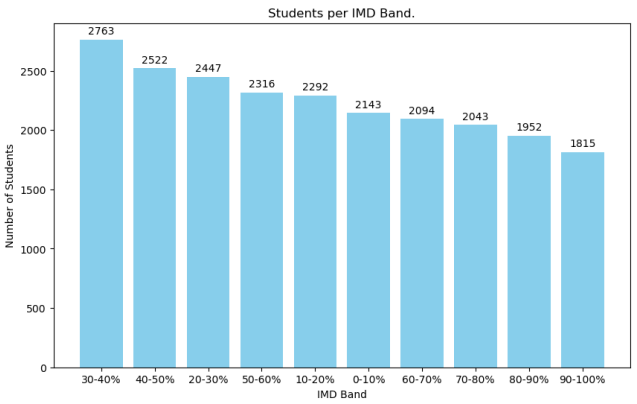

Figure 10. Students per IMD Band.

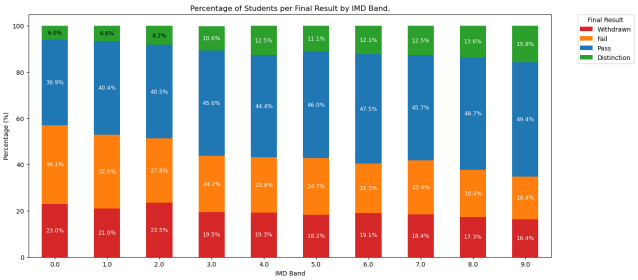

Figure 11. Percentage of Students per Final Result by IMD Band.

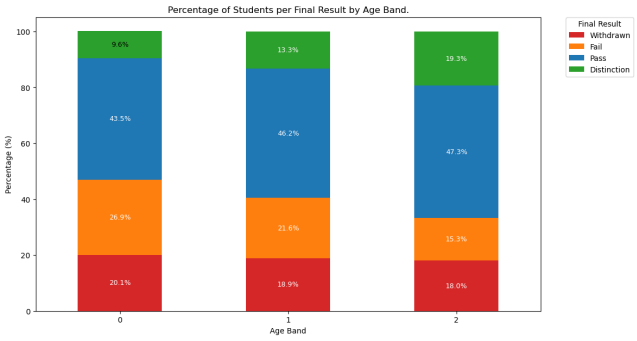

Figure 12. Percentage of Students per Final Result by Age Band.

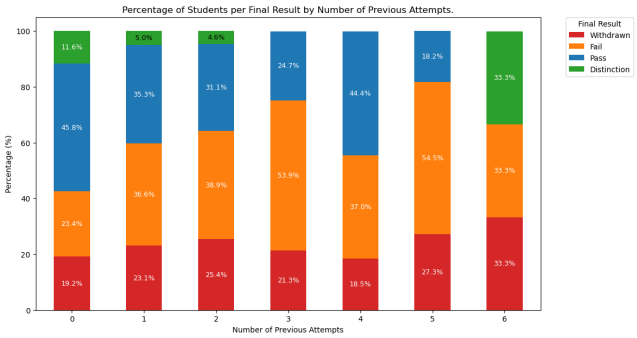

Figure 13. Percentage of Students per Final Result by the Number of Previous Attempts.

rate of 45.8%, and distinction rate of 11.6%. As the number of previous attempts increases, withdrawal and fail rates generally rise, while pass and distinction rates decline. For example, students with one or two previous attempts show higher withdrawal (23.1% and 25.5%) and fail rates (36.6% and 38.9%), and lower pass rates (35.3% and 31.1%). Students with three or more previous attempts face the worst outcomes, with withdrawal rates reaching up to 33%, fail rates up to 54%, and very low or zero distinction rates. Overall, repeated attempts correlate with poorer final results, indicating challenges for students retaking modules multiple times.

3.5.11 Studied Credits In Figure 14, 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 15.

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

3.5.12 Total Clicks Moving on to total clicks, in Figure 16, As the module progresses, the number of students with total VLE clicks in the 0–1,000 range steadily declines, from around 19,000 at the early phase, to around 14,000 in the late phase, and approximately 13,000 by the end of the module. Despite this decline, the majority of students consistently fall within the 0–1,000 click range throughout the module, while those with over 5,000 clicks remain relatively rare, even during the later stages.

3.5.13 Days Active This feature is a binary value ranging from 0 to 1 that represents how consistently a student engaged with the VLE relative to the module's timeline. For example, if the midpoint phase spans 150 days, the value reflects the proportion of those days the student was active. A value of 1 indicates daily engagement throughout the period, while 0 means the student did not engage at all.

In Figure 17, * Withdrawn (0) students witnessed a sharp decline in activity over time. as the median drops from 0.25 (Early) → 0.15 (Midpoint) → 0.1 (Late) → 0.09 (Full). Additionally, 75% of withdrawn students remain very inactive, indicating early disengagement and no recovery later. However, there are still many outliers where students are more active than the top 75% of students who have withdrawn. For students who failed (1), they have consistently low engagement across phases. this is reflected on the median which falls from 0.22 (Early) → 0.15 (Midpoint) → 0.11 (Late) → 0.09 (Full). Activity fades over time, possibly indicating a slow drop-off after initial attempts. Similar to the withdrawed students, there are many outlier students who have more engagement than the top 75% of students who have failed.

As for passed students (2), they display moderate engagement, but it gradually declines over phases as the median drops from 0.45 (Early) → 0.36 (Midpoint) → 0.33 (Late) → 0.30 (Full). These students remain fairly active, but less so as the module progresses. Finally for distinction (3) student, they are the most active and consistent group. However their median also decreases from 0.57 (Early) → 0.47 (Midpoint) → 0.44 (Late) → 0.40 (Full). But they still show strong activity even in late phases, though with a slight downward trend.

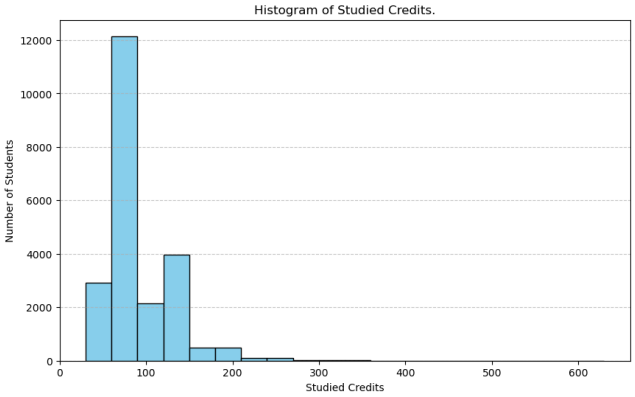

Figure 14. Histogram of Studied Credits.

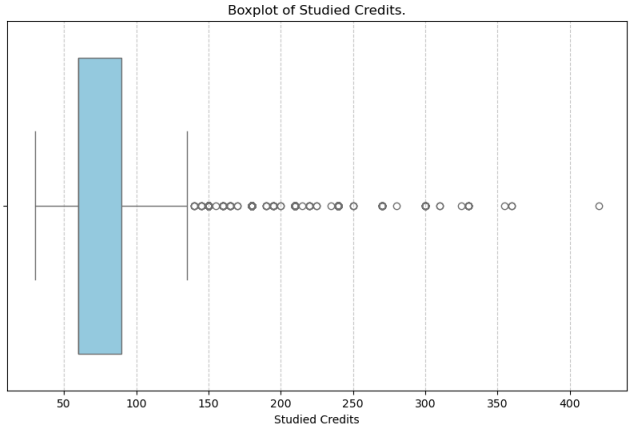

Figure 15. Boxplot of Studied Credits.

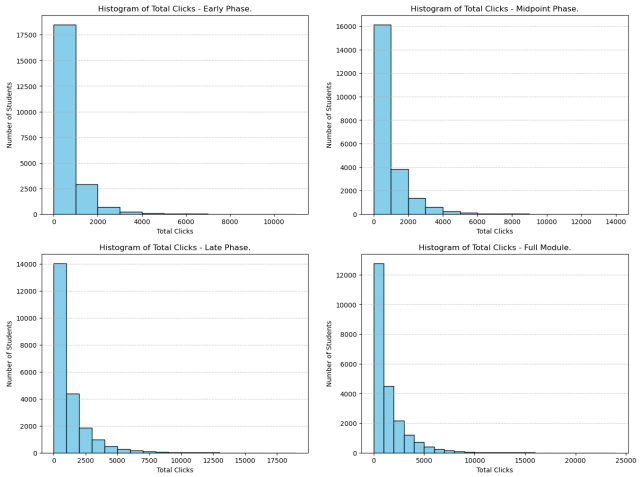

Figure 16. Histogram of Total Clicks for Each Phase.

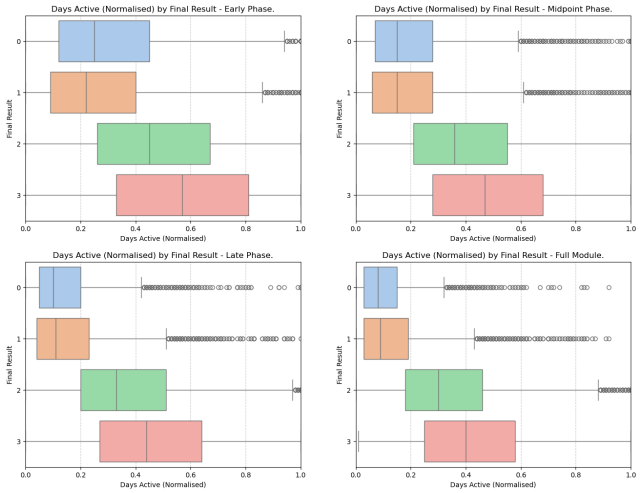

Figure 17. Days Active (Normalised) By Final Result For Each Phase.

Across all outcomes, students show a general decline in activity over time, but the relative order remains clear, those who achieve distinction or pass consistently stay more engaged than those who fail or withdraw. The gaps in median values between the groups widen in later phases, reinforcing how sustained activity relates to success.

3.5.14 Banked Rate For banked rate, the vast majority of students do not have assessment results carried over from a previous presentation, with only about 250 students with a banked rate greater than 0.7. In Figure 18, students who fail or withdraw appear to have a marginally higher banked rate, although the difference is minimal. Let us move on to 'final_results'.

3.5.15 Weighted Score In Figure 19, the plot tells us that in the Early Phase about 3,250 students are in the 0–5 score range, indicating very low engagement or progress at this point. Moreover, score distribution begins to grow noticeably from 60 onwards, peaking at 70–75 (Around 2,300 students). By this stage, some students already have high scores (e.g. about 1000 students with perfect 95-100). In the Midpoint Phase, there are still about 4,700 students in the 0–5 range, though slightly more students are scoring in the 20–60 band. The mode (most common score band) shifts to 80–85 (approximately 1,950 students), showing improvement in score distribution as assessments accumulate. Finally, very few students achieve full marks yet in the midpoint phase (less than 500 students with 95-100 score).

For the Late Phase, the number of students in the 0–5 band increases slightly to about 4,800, which might include withdrawn or inactive students. A stronger concentration appears in the 70–90 range. Peak still lies around 80–85 and a small number of high performers remain (roughly 300 students with 95-100). and for the Full Module, we see that about 5,600 students still remain in the 0–5 range, most likely those who dropped out or failed to participate fully. The score distribution has now matured, with strong peaks around 75–90, showing the final result of all assessments. Overall, only about 200 students have weighted scores beteen 95-100.

Another thing to keep in mind, that the weighted score is not a very reliable metric in the Full version case. These discrepancies stem from missing exam assessment data that occur during the end of the module, most modules (except for CCC and DDD) have missing exam data, leading to incomplete or skewed weighted scores that do not fully reflect students' true performance. As a

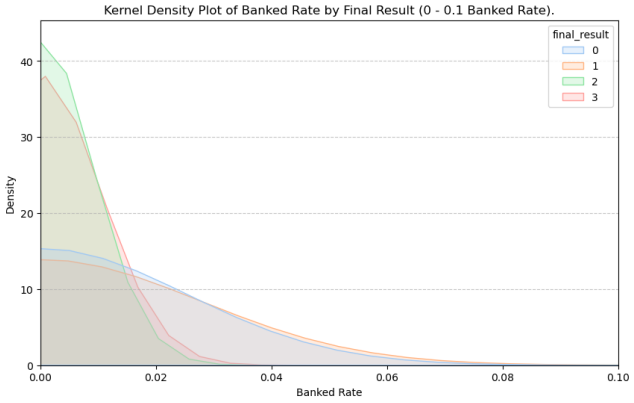

Figure 18. Kernel Density Plot of Banked Rate by Final Result (0—0.1 Banked Rate).

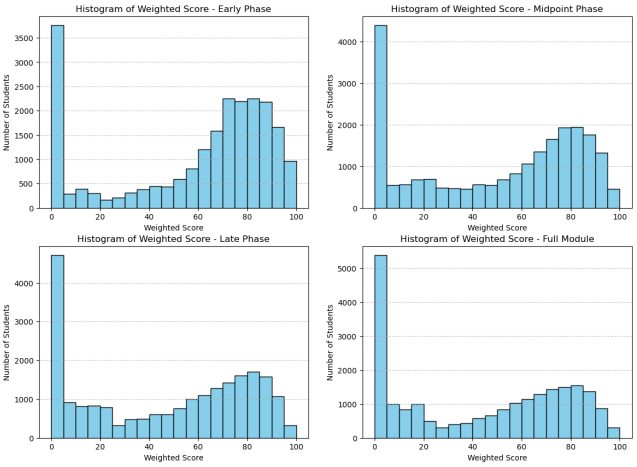

Figure 19. Histogram of Weighted Score for Each Phase.

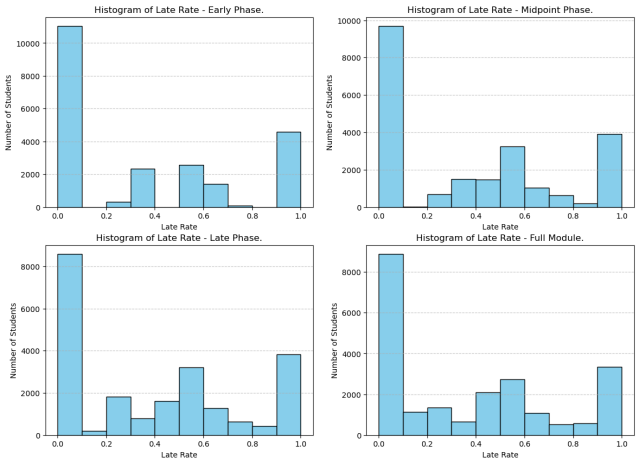

Figure 20. Histogram of Late Rate for Each Phase.

result, 'weighted_score' in the full version should be treated with caution and not used as a standalone indicator of student success. Let us move on to the late rate.

3.5.16 Late Rate Similar to the weighted score, both the late rate and, by extension, the fail rate depend on incomplete assessment data, particularly missing exam results, making them less than fully reliable in the full version of the dataset. However, they can still provide some interesting insights.

Across all phases (Figure 20), student behaviour regarding assignment lateness reveals several evolving patterns. In the Early Phase, the vast majority of students (over 11,200) had a low late submission rate (0–10%), indicating timely completion of assessments. However, a secondary peak at the 50–60% late rate marks a notable group (around 2,500 students) who were consistently late. Some intermediate bins (e.g., 10–20%, 40–50%, 80–90%) show no recorded activity, possibly reflecting either system artefacts or clustered submission habits. Interestingly, roughly 4,400 students submitted all of their assignments late during this period. By the Midpoint Phase, while most students (just under 10,000) still submitted on time, the late submission pattern became more varied. Substantial numbers of students appeared across the 20–80% late rate bins, with particularly noticeable increases in the 40–60% range. This indicates a gradual shift towards more irregular submission behaviour.

In the Late Phase, the spread of late submissions becomes even more apparent. Although the highest number of students still fall within the 0–10% bin, larger groups now emerge across the 20–60% range, especially around the 50–60% mark (approximately 3,200 students). The overall distribution flattens, suggesting that submitting assignments late becomes increasingly common as the module progresses. Finally, during the Full Module, on-time submission remains the dominant pattern, but there is a clear extended tail across higher late rate bands. The number of students with 90–100% late rates declines to below 3,700, while the 40–60% bands become more prominent, indicating accumulated delays over the course. This evolution in submission behaviour highlights growing disengagement or increasing workload struggles as time progresses.

3.5.17 Fail Rate In Figure 21, student failure rates across the phases show a gradual yet clear shift toward higher failure as the course progresses. In the Early Phase, the majority of students (around 16,000) maintained low fail rates (below 10%), indicating early academic success. However, a smaller cluster of students—approximately 2,100, experienced higher failure rates in the 50–60% range. Interestingly, several intermediate bins (10–20%, 40–50%, and 80–90%) recorded no students, suggesting that failure patterns at this stage were sharply polarised. Additionally, over 2,900 students had very high fail rates (between 95–100%), highlighting a group that struggled significantly from the outset. By the Midpoint Phase, the number of students with a low fail rate (0–10%) declined to around 13,000. Meanwhile, failure began to spread more evenly across the distribution, with significant student counts appearing in the 20–40% range (each with over 1,400 students). The number of students failing nearly all assessments (95–100%) showed a slight decrease but remained a noteworthy group.

In the Late Phase, the downward trend in low fail rate students continued, dropping to around 10,700 in the 0–10% bin. More students appeared in higher fail rate brackets such as 20–30%, 50–60%, and 80–90%, illustrating a shift toward greater academic difficulty.

The overall distribution showed a clearer lean toward higher failure rates, indicating cumulative challenges over time. During the Full Module, the number of students with minimal failures (0–10%) declined further to about 10,100, though this group still formed the largest single bin. However, the distribution's tail thickened considerably, with more students occupying high fail rate bins, especially between 70–90%. Notably, the count of students in the 95–100% fail range reached its peak, rising to around 3,100, marking the highest observed concentration of severe academic struggle by the end of the course.

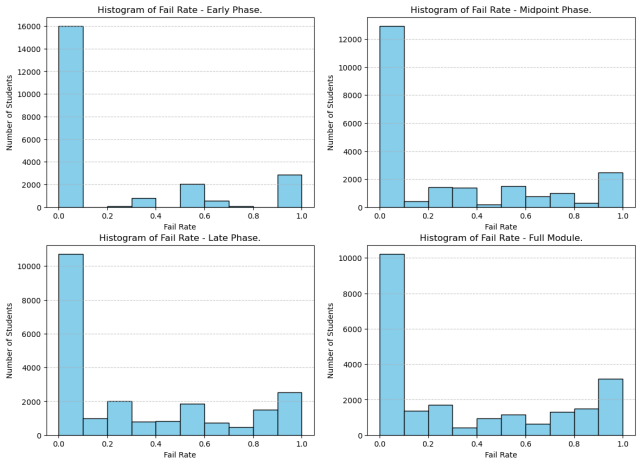

Figure 21. Histogram of Fail Rate for Each Phase.

3.5.18 Final Result The 'final_result' feature is our target variable. In Figure 22, the majority of students have passed, making up about 44%, followed by failed students at 25%. Withdrawn students represent about 20%, while those who passed with distinction have the smallest share around 11% of the train set.

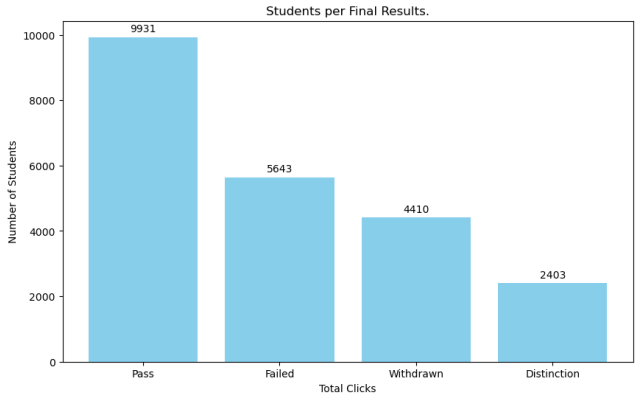

Figure 22. Students per Final Result.

3.5.19 Correlation Matrix (Full Dataset) We will be using a correlation matrix of the full version of the dropout dataset, as it gives us a whole picture on which features correlate with others the most, and what can be expected by the models.

Using a correlation matrix in Figure 23, a correlation analysis was conducted to explore how various student-related features relate to the final course outcome (final_result). As expected, weighted score shows the strongest positive correlation (= 0.71), indicating that students with higher assessment scores are much more likely to pass or perform well. Similarly, fail rate has a strong negative correlation (= −0.78), confirming that consistent assessment failure is highly predictive of poor final outcomes.

Student engagement metrics also show meaningful associations. Days active, a normalised measure of consistent platform use, has a moderately strong positive correlation (= 0.54), suggesting that students who are more engaged over time tend to achieve better results. Total VLE clicks also correlate moderately (= 0.41) with final performance, reflecting that higher levels of interaction with learning resources can support academic success. On the other hand, late submission rate shows a moderate negative correlation (= −0.36), indicating that students who frequently submit assignments late are less likely to succeed.

Sociodemographic and background variables show weaker, though still interesting, relationships. Highest education level (= 0.14) and IMD band (= 0.12) both exhibit weak positive correlations, implying that students from more educated or less deprived backgrounds perform slightly better. Age band displays a very weak positive correlation (= 0.06), suggesting a minimal advantage for older students. Date of registration is effectively uncorrelated (= 0.02), indicating that when a student registers has little bearing on their final outcome. Additionally, gender shows no meaningful correlation with final results, suggesting gender-neutral academic performance across the cohort.

Some behavioural and structural features show slight negative correlations. Studied credits (= −0.11) and number of previous attempts (= −0.10) are weakly associated with lower performance, potentially due to overcommitment or previous disengagement. Disability and banked rate both have very weak negative correlations (= −0.07), with the latter showing little predictive value — supporting the decision to exclude it from modelling.

Beyond individual correlations with the final result, a few other patterns are noteworthy. Students with higher prior education levels tend to score better in assessments (highest education vs. weighted score: = 0.21), reinforcing the value of academic preparedness. Additionally, students who have retaken modules multiple times tend to reuse past results more often (previous attempts vs. banked rate: = 0.30), reflecting the behaviour of repeat enrolments.

Overall, this analysis highlights that performance metrics (such as assessment scores and fail rate) and engagement behaviour are far more predictive of student outcomes than static demographic variables. This insight helps guide feature selection and model interpretation in subsequent predictive modelling stages.

3.6 Hypothesis Testing

3.6.1 Engagement Hypothesis Students who drop out within the first 25% of the module have significantly lower early VLE engagement compared to those who continue the course (regardless of whether they ultimately pass or fail).

Null and Alternative Hypotheses for Mann–Whitney U test: * Null Hypothesis (H): There is no difference in early engagement (days_active_norm) between students who dropped out and those who remained enrolled (distinction/passed/failed).

* Alternative Hypothesis (H): Students who dropped out exhibit significantly lower early engagement than those who remained

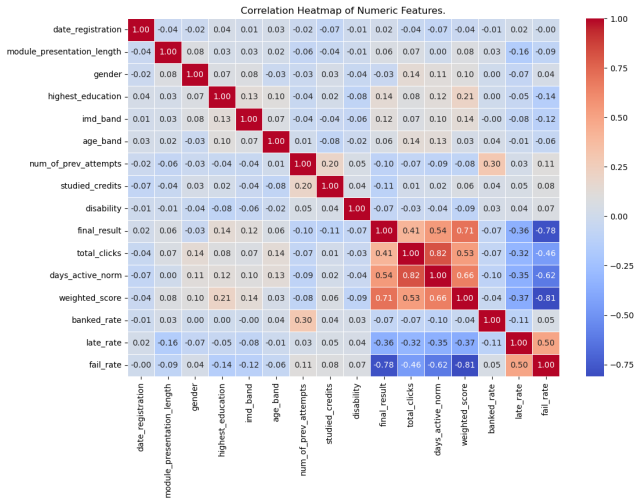

Figure 23. Correlation Matrix of Dropout Features

enrolled.

Since days_active_norm is not normally distributed, we shall proceed with the Mann-Whitney test, as it does not assume normality and is more robust when comparing medians from skewed or non-Gaussian distributions.

A Mann–Whitney U test was conducted to compare early engagement (‘days_active_norm‘) between students who dropped out and those who remained enrolled (including both pass and fail outcomes).

The results revealed a statistically significant difference in engagement levels between the two groups (U = 29,961,144.5, p < 0.001), with dropouts exhibiting substantially lower activity during the first 25% of the module timeline. This provides strong evidence supporting the hypothesis that lower early VLE engagement is associated with an increased likelihood of dropout.

The high U statistic (approaching the maximum possible value) indicates a clear separation in the ranked engagement scores between groups, while the extremely low p-value (p 8.8×10^{1}) confirms that this difference is statistically significant. Thus, the null hypothesis, that dropouts and non-dropouts have similar early engagement distributions, can be confidently rejected.

3.6.2 Assessment Performance Hypothesis The next hypothesis is on on poor performance on early continuous assessments increases dropout risk.

Let us assume we are using early assessment scores (e.g. weighted_score) from the first 25% of the module.

- Null Hypothesis (H):There is no difference in early assessment scores between students who dropped out and those who remained enrolled (distinction/passed/failed).

- Alternative Hypothesis (H): Students who dropped out had significantly lower early assessment scores than those who remained enrolled.

A Mann–Whitney U test was carried out to compare early assessment performance (measured using weighted_score) between students who dropped out and those who remained enrolled (including both pass and fail outcomes). Visual inspection of the distributions via histograms indicated non-normality in both groups, justifying the use of a non-parametric test. The results revealed a statistically significant difference in early assessment performance (U = 22,891,409.0, p < 0.001), with dropouts achieving markedly lower weighted scores in the first 25% of the module.

This provides strong support for the hypothesis that early poor assessment performance increases the risk of dropout, potentially due to loss of confidence or reduced motivation following weak early results.

3.6.3 Demographic Disparity Hypothesis The next hypothesis is about the demographic disparity hypothesis where certain demographic groups (based on age, region, education level, IMD band and disability) are disproportionately represented among dropouts. A Chi-Square Test of Independence test was conducted where both variables are categorical and it checks whether distributions of dropout vary by category [CITE]. Thre results are summarised in the following Table 7

Feature	Chi Square Result	p-test	Findings
Age Band	4.10	0.129	Little evidence that dropout rate var
Highest Education	72.64	0.000	Strong evidence that prior education
Region	22.57	0.032	Region is significantly related to dro
IMB Band	65.69	0.000	Students from more deprived areas (
Disability	109.53	0.000	Disability status is very strongly ass

Table 7
Demographic Disparity Hypothesis: Chi square and p-test results

Table 7 tells us that the hypothesis is partially supported by the results. Significant associations were found between dropout status and highest education level, region, IMD band, and disability status, indicating disparities in who is more likely to drop out. However, age band did not show a significant effect, suggesting age alone may not be a major factor in early dropout when other variables are considered. Moreover, the correlation matrix Section [CITE] also supports the hypothesis tests conducted. Here are the findings:

- ‘Region‘ shows strong evidence of disparity, with dropout rates ranging from 22.8% (Ireland) to 35.5% (West Midlands), suggesting location significantly influences withdrawal likelihood. - ‘Education level‘ offers moderate support, as students with higher prior qualifications tend to perform better. - ‘Age band‘ provides weak evidence, with only a slight positive correlation to performance. - Overall, The hypothesis is partially supported. The region and education level have some influence, but age has minimal impact.

Re-enrollment Hypothesis

The final hypothesis is on the re-enrollment where students who have previously attempted the module more times are more likely to drop out again.

- Null hypothesis (H): There is no difference in the distribution of the number of previous attempts $(\text{num}_o f_p rev_a ttempts) between students who dropout and those who continue (distinc$

- Alternative hypothesis (H): There is a difference in the distribution of previous attempts between dropouts and non-dropouts.

Since $\text{num}_o f_p rev_a ttempts is a non - normally distributed count data, and the two groups (dropouts vs. non - dropouts) are independent, the Mann ˇ Whitney U test is appropriate to compare the di$

Mann–Whitney U test yielded with U statistic as 40,818,054.0 and p-value = 1.10 × 10 (approximately).

The very small p-value ($p < 0.05$) indicates strong evidence to reject the null hypothesis. This suggests that the distribution of previous attempts significantly differs between students who drop out and those who do not. In practical terms, this supports the hypothesis that students with more prior attempts are more likely to drop out again.

In conclusion, there is statistically significant evidence that re-enrolment history, as measured by the number of previous attempts, is associated with an increased risk of dropout. Now, let us move on to ther overview of machine learning models and why some ML models were justified for use ifor this classification task

3.7 Overview of Machine Learning Models

As this is a classification task, we employ Logistic Regression (LR), Support Vector Machine (SVM), Multilayer Perceptron Classifier (MLP), and Random Forest (RF) models. These models are trained on the training dataset, then evaluated on the test set using two key metrics: accuracy and macro-averaged F1 score, which balances precision and recall across all classes.

3.7.1 Model Selection Justification The models selected for this task: Logistic Regression (LR), Support Vector Machines (SVM), Multi-Layer Perceptron (MLP), and Random Forest (RF), represent a diverse set of classification strategies. Each model brings distinct advantages suitable for the challenges of student dropout prediction:

Logistic Regression (LR): LR serves as a strong baseline due to its simplicity, interpretability, and computational efficiency [CITE]. It establishes a benchmark for evaluating more complex models and provides insight into the linear separability of the data. Moreover, its probabilistic outputs are useful for setting risk thresholds when identifying students at risk of dropping out [CITE].

Support Vector Machines (SVM): SVMs are effective for high-dimensional data and can perform well even in the presence of class imbalance, particularly when using techniques such as class weighting [CITE]. By employing kernel functions like RBF or polynomial kernels, SVMs can model non-linear decision boundaries, making them capable of capturing more intricate patterns in student behaviour [CITE].

Multi-Layer Perceptron (MLP): The MLP, a feedforward neural network, is chosen for its ability to learn non-linear relationships and complex feature interactions that linear models might overlook [CITE]. It is especially valuable when underlying patterns in the data are not easily separable or explicitly encoded. Although MLPs require careful hyperparameter tuning and are more computationally intensive, they can uncover hidden structures in student engagement or performance indicators [CITE].

Random Forest (RF): RF is a robust ensemble learning method that combines multiple decision trees to improve generalisation and reduce overfitting [CITE]. It is particularly effective at handling non-linear relationships, missing values, and mixed data types (numerical and categorical) [CITE]. 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, enhancing predictive stability across diverse student profiles.

3.8 Implementation of Machine Learning Models

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] J. Kuzilek, M. Hlosta, and Z. Zdrahal, “Open university learning analytics dataset,” Scientific Data, vol. 4, p. 170171, 2017. DOI: 10.1038/sdata.2017.171. [Online]. Available: https://doi.org/10.1038/sdata.2017.171.

[7] 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).

[8] 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.