# Contents

## Figure List

# Introduction

The Open University (OU) is one of the world's largest open and distance education universities. This university aims to provide students with equal opportunities in education. In this context, most of its courses are conducted through a distance education model. With its international access capacity enabled by distance education, the university also offers flexible working hours to its students.

The Open University has implemented a new virtual learning environment (VLE) system to enhance students' academic performance. This report evaluates whether the VLE system has a measurable effect on students' academic success and examines the potential for predicting students' success in their courses.

The study addresses the following two questions using statistical analysis and modeling; additionally, various recommendations are presented to improve the effectiveness of the VLE system:

- Does the VLE impact students' academic success?
- Can students' success be predicted?

# Data Exploration

The datasets used in this study were sourced from the Open University Learning Analytics Dataset. Seven primary datasets were utilized in the study.

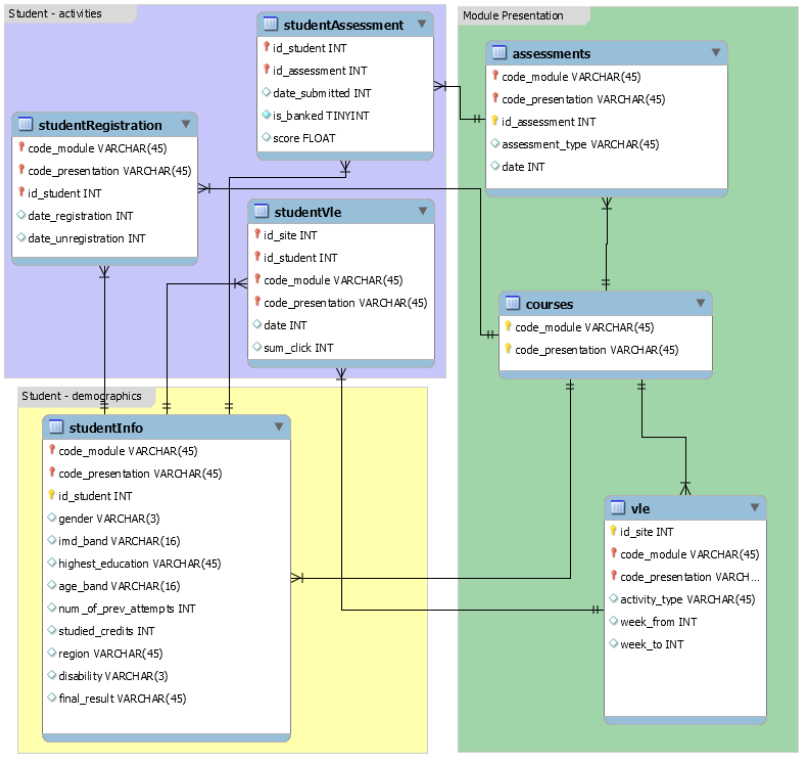



*Figure 1: Open University Learning Analytics Dataset Entity Relationship Diagram (Knowledge Media Institute, 2024).*

The data files used in the study are examined in detail below.

## courses.csv

This table contains the course names, semester information, and course length details. The data in the code_presentation column specifies the year the course is offered. If the column ends with 'B', it indicates that the course begins in February; if it ends with 'J', the course starts in October.

As shown in Figure 2, data for BBB, DDD, and FFF courses are available for each semester. Therefore, these courses have yielded more accurate results in the prediction study.

| | Module | Presentations |
|---|---|---|
| **0** | AAA | [2013J, 2014J] |
| **1** | BBB | [2013B, 2013J, 2014B, 2014J] |
| **2** | CCC | [2014B, 2014J] |
| **3** | DDD | [2013B, 2013J, 2014B, 2014J] |
| **4** | EEE | [2013J, 2014B, 2014J] |
| **5** | FFF | [2013B, 2013J, 2014B, 2014J] |
| **6** | GGG | [2013J, 2014B, 2014J] |

*Figure 2: Distribution of Presentations*

## assessments.csv

This table contains information about homework and exams in the course contents. There are three basic types of evaluations:

- Tutor Marked Assessment (TMA)
- Computer Marked Assessment (CMA)
- Final Exam (Exam)

Additionally, the last submission date for each evaluation can be accessed from the Date field.

During the data cleaning phase, the data was checked on an ID basis, and singularity was confirmed. Records with a Date field value of '?' were filled with zero due to their low impact during the modeling phase. Each type of evaluation has a specific weight effect for each course, with values ranging between 0 and 100.

The weight is 100% for exams of the "Exam" type, while the total weight of all other evaluations for a course also adds up to 100%. There are two rows of duplicate data that do not comply with this rule (Figure 3). The duplicated id_assessment values were checked against the student_assessment table. While the record with assessment ID 24290 exists in the student_assessment table, the record with assessment id 40087 (Figure 4) was not found in the student_assessment table, so this row was dropped. Similarly, the assessment ID 40088 (Figure 4) record was also dropped as it was not found in the student_assessment table.

| assessment_type | | CMA | Exam | TMA |
|---|---|---|---|---|
| code_module | code_presentation | | | |
| AAA | 2013J | 0.0 | 100.0 | 100.0 |
| | 2014J | 0.0 | 100.0 | 100.0 |
| BBB | 2013B | 5.0 | 100.0 | 95.0 |
| | 2013J | 5.0 | 100.0 | 95.0 |
| | 2014B | 5.0 | 100.0 | 95.0 |
| | 2014J | 0.0 | 100.0 | 100.0 |
| CCC | 2014B | 25.0 | 200.0 | 75.0 |
| | 2014J | 25.0 | 200.0 | 75.0 |
| DDD | 2013B | 25.0 | 100.0 | 75.0 |
| | 2013J | 0.0 | 100.0 | 100.0 |
| | 2014B | 0.0 | 100.0 | 100.0 |
| | 2014J | 0.0 | 100.0 | 100.0 |
| EEE | 2013J | 0.0 | 100.0 | 100.0 |
| | 2014B | 0.0 | 100.0 | 100.0 |
| | 2014J | 0.0 | 100.0 | 100.0 |
| FFF | 2013B | 0.0 | 100.0 | 100.0 |
| | 2013J | 0.0 | 100.0 | 100.0 |
| | 2014B | 0.0 | 100.0 | 100.0 |
| | 2014J | 0.0 | 100.0 | 100.0 |
| GGG | 2013J | 0.0 | 100.0 | 0.0 |
| | 2014B | 0.0 | 100.0 | 0.0 |
| | 2014J | 0.0 | 100.0 | 0.0 |

*Figure 3: Duplicated Data on Exam Weight*

| code_module | code_presentation | id_assessment | assessment_type | date | weight |
|---|---|---|---|---|---|
| CCC | 2014B | 24290 | Exam | ? | 100 |
| CCC | 2014B | 40087 | Exam | ? | 100 |
| CCC | 2014J | 24299 | Exam | ? | 100 |
| CCC | 2014J | 40088 | Exam | ? | 100 |

*Figure 4: Unused Duplicated Assessment*

In the final state, the total scoring for each course is ensured to be out of 100, including both the "Exam" type and other types of evaluations. (Figure 5)

| assessment_type | | CMA | Exam | TMA |
|---|---|---|---|---|
| code_module | code_presentation | | | |
| AAA | 2013J | 0.0 | 100.0 | 100.0 |
| | 2014J | 0.0 | 100.0 | 100.0 |
| BBB | 2013B | 5.0 | 100.0 | 95.0 |
| | 2013J | 5.0 | 100.0 | 95.0 |
| | 2014B | 5.0 | 100.0 | 95.0 |
| | 2014J | 0.0 | 100.0 | 100.0 |
| CCC | 2014B | 25.0 | 100.0 | 75.0 |
| | 2014J | 25.0 | 100.0 | 75.0 |
| DDD | 2013B | 25.0 | 100.0 | 75.0 |
| | 2013J | 0.0 | 100.0 | 100.0 |
| | 2014B | 0.0 | 100.0 | 100.0 |
| | 2014J | 0.0 | 100.0 | 100.0 |
| EEE | 2013J | 0.0 | 100.0 | 100.0 |
| | 2014B | 0.0 | 100.0 | 100.0 |
| | 2014J | 0.0 | 100.0 | 100.0 |
| FFF | 2013B | 0.0 | 100.0 | 100.0 |
| | 2013J | 0.0 | 100.0 | 100.0 |
| | 2014B | 0.0 | 100.0 | 100.0 |
| | 2014J | 0.0 | 100.0 | 100.0 |
| GGG | 2013J | 0.0 | 100.0 | 0.0 |
| | 2014B | 0.0 | 100.0 | 0.0 |
| | 2014J | 0.0 | 100.0 | 0.0 |

*Figure 5: Sum of Assessments' Weight*

## vle.csv

This dataset contains information about the type of documents in the VLE system and their associated courses. In 5,243 rows, the week_from and week_to fields were filled with '?'. However, no action was taken for these records as these fields are not critically important for the model.

## studentInfo.csv

This table contains information about the students' demographic structure and results. Rows with the imd_band field filled with '?' (a total of 1,111 rows) were dropped, as this field is significant for the analysis. Additionally, the segment labeled as '10-20' was changed to '10-20%' to align with other segments.

The final_result value for 2,825 rows is labeled 'Distinction' in the dataset. To predict whether a student will 'Pass' or 'Fail,' these labels were changed to 'Pass' during the modeling process. Similarly, the final_result value for 9,920 rows is labeled as 'Withdrawn.' The impact of either dropping these rows or treating them as 'Fail' was tested.

When these rows were dropped, the model's ability to predict the 'Fail' label was significantly lower than its success in predicting the 'Pass' label due to the imbalance in the distribution of 'Fail' and 'Pass' labels. Instead of dropping, the model's performance improved when the 'Withdrawn' label was treated as 'Fail.' Given the dataset's size of 32,593 rows and the importance of these 9,920 rows, they were categorized as 'Fail' instead of dropping.

Finally, the data marked as 50<= in the age_band column was adjusted to >=50 to ensure consistency with other categories.

### studentRegistration.csv

This table contains information about students' registration and withdrawal dates from the courses. Rows with a "?" in the registration date field, representing uncertain registration dates and uncertain participation in the course (a total of 45 rows), were dropped. After this operation, 32,548 rows of data remained.

### studentAssessment.csv

This table records the evaluation results of students and indicates whether these results will have an impact on the assessment. If a student has not submitted the relevant course exam, there is no recorded result for that evaluation in the table. Evaluation results range between 0 and 100. Scores below 40 are labeled as unsuccessful, while those above 40 are labeled as successful.

Rows, where the score information was filled with a "?" (173 rows), were dropped, as they hold significant importance in the study and represent a proportionally small fraction of the data. After this operation, 173,739 rows of data remained in the table.

### studentVle.csv

This table contains information about students' interactions with the VLE system. Students' activities in the VLE system are recorded daily for each course. In this dataset, which originally had 10,655,280 rows, 787,170 rows were identified as duplicates. Only the first occurrence of each duplicate row was retained, and the others were dropped. After this operation, 9,868,110 rows of data remained.

The data from the studentVLE table is used in assessing the effect of VLE on the final results of the students. Data was grouped and aggregated by student_id, code_module, and code_presentation on the sum_click field as the total number of VLE interactions made by each student in his or her course.

| | code_module | code_presentation | id_student | total_sum_click |
|---|---|---|---|---|
| 0 | AAA | 2013J | 11391 | 922 |
| 1 | AAA | 2013J | 28400 | 1409 |
| 2 | AAA | 2013J | 30268 | 260 |
| 3 | AAA | 2013J | 31604 | 2007 |
| 4 | AAA | 2013J | 32885 | 1012 |

*Figure 6: Total Sum Click Calculation*

After cleaning, various graphs were created to gain an understanding of the data. Join of df_student_info and df_student_vle_upd was done to evaluate the students in terms of demographic structure and their interactions with the VLE system. Additionally, a join between the df_student_assessment table and the df_assessments table was created to calculate the students' weighted grades.

In the df_student_assessment table, the is_banked field indicates whether a student's exam result will be included in the final average for any reason. If is_banked=1, the exam score in that row is excluded from the final average calculation. The weighted_score column is set to 0 if is_banked=1. If is_banked!=1, the score value in the df_student_assessment table is multiplied by the percent of the weight value in the df_assessments table, and the result is stored in the weighted_score field.

| | id_assessment | id_student | date_submitted | is_banked | score | code_module | code_presentation | assessment_type | date | weight | weighted_score |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1752 | 11391 | 18 | 0 | 78 | AAA | 2013J | TMA | 19 | 10.0 | 7.8 |
| 1 | 1752 | 28400 | 22 | 0 | 70 | AAA | 2013J | TMA | 19 | 10.0 | 7.0 |
| 2 | 1752 | 31604 | 17 | 0 | 72 | AAA | 2013J | TMA | 19 | 10.0 | 7.2 |
| 3 | 1752 | 32885 | 26 | 0 | 69 | AAA | 2013J | TMA | 19 | 10.0 | 6.9 |
| 4 | 1752 | 38053 | 19 | 0 | 79 | AAA | 2013J | TMA | 19 | 10.0 | 7.9 |

*Figure 7: Calculation Weighted Score*

Then, two data frames named exam_scores and nonexam_scores are created, dividing the data into two groups based on assessment_type=Exam and assessment_type!=Exam. In both groups, the total score is calculated by grouping the data based on id_student, code_module, and code_presentation.

Afterward, both data frames are joined, and the overall_score is calculated according to the following rule:

- If both an exam score and a non-exam score exist for a course, the average of these scores is used as the overall_score value.
- If only one of these scores exists for a course, that score is directly assigned as the overall_score.

| | id_student | code_module | code_presentation | exam_weighted_score | non_exam_weighted_score | overall_grade |
|---|---|---|---|---|---|---|
| **0** | 6516 | AAA | 2014J | 0.0 | 63.50 | 63.500 |
| **1** | 8462 | DDD | 2013J | 0.0 | 34.90 | 34.900 |
| **2** | 11391 | AAA | 2013J | 0.0 | 82.40 | 82.400 |
| **3** | 23629 | BBB | 2013B | 0.0 | 16.69 | 16.690 |
| **4** | 23698 | CCC | 2014J | 80.0 | 69.97 | 74.985 |

*Figure 8: Calculation of Overall Grade*

The obtained data frame is joined with other tables containing the students' demographic and VLE interaction information to prepare the final dataset.

The data visualization steps performed with these intermediate tables are detailed below. Each graph includes a comprehensive explanation.

In all courses, the average VLE interaction for the 'Pass' group is higher than that for the 'Fail' group. (Figure 9) The FFF course has the highest average interaction for both the 'Fail' and 'Pass' labels, whereas the GGG course has the lowest average interaction in both categories. The number of VLE interactions is directly proportional to learning success. Therefore, as interaction increases, learning achievement is expected to improve.

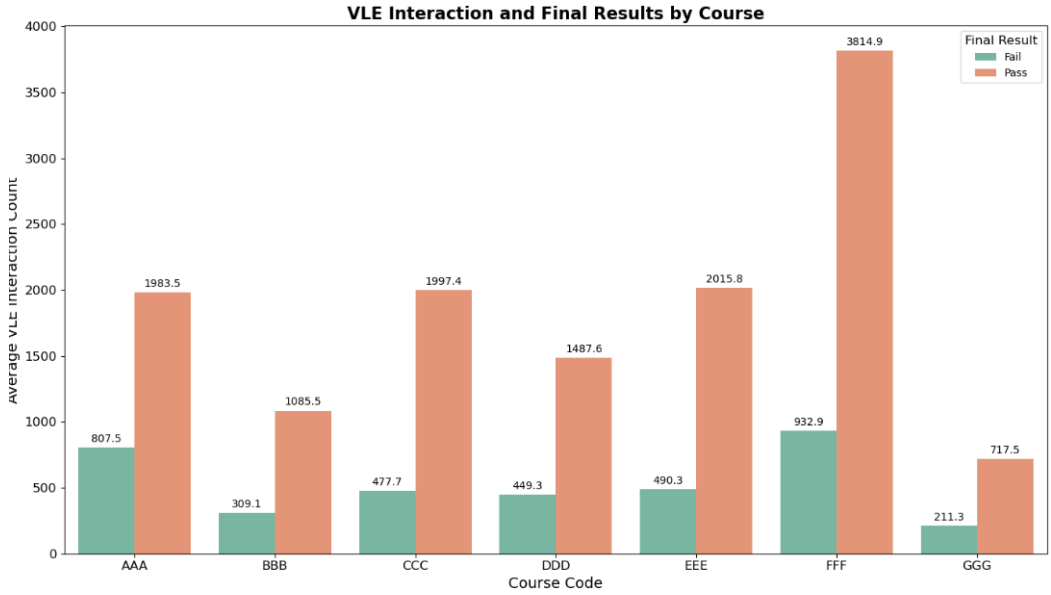



*Figure 9: VLE Interaction and Final Results by Course*

This graph shows a significant relationship between age and VLE interaction (Figure 10). Generally, VLE interaction increases with age. Based on this observation, different learning strategies should be implemented for different age groups.

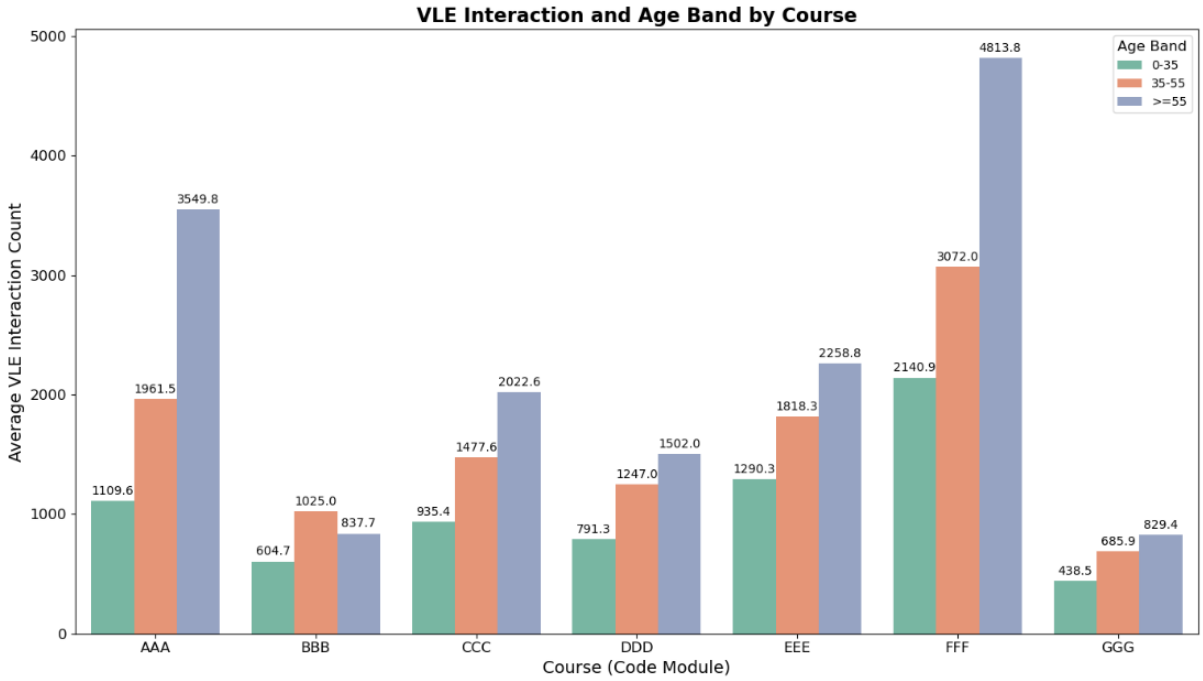



*Figure 10: VLE Interaction and Age Band by Course*

As the level of education increases, the mean VLE interaction in the category 'Pass' also rises accordingly. Figure 11 The standard deviation is higher in the 'Pass' group, indicating that the VLE interaction rates for this group are highly scattered. This graph gives an indication that the higher the level of education of a student, the more actively and efficiently students utilize online resources.

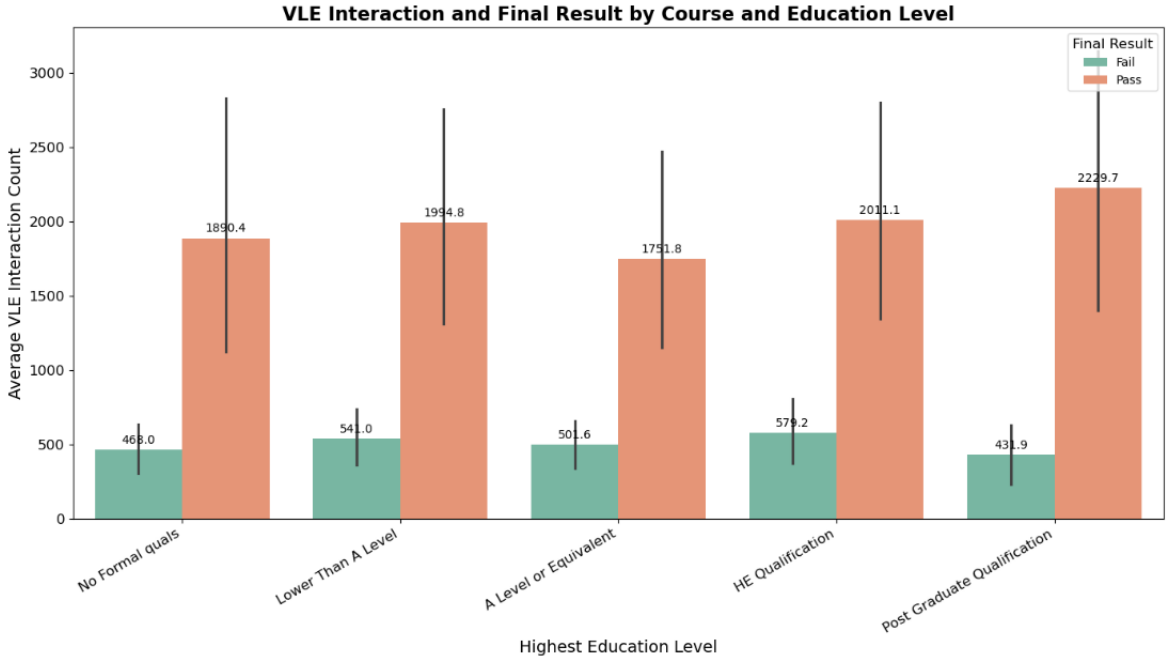



*Figure 11: VLE Interaction and Final Result by Course and Education Level*

Students have shown a greater liking for modules like FFF and BBB, while CCC and EEE modules have been less in choice. Figure 12 The reasons may be manifold: the difficulty level of the course, the instructor, or the course content itself.

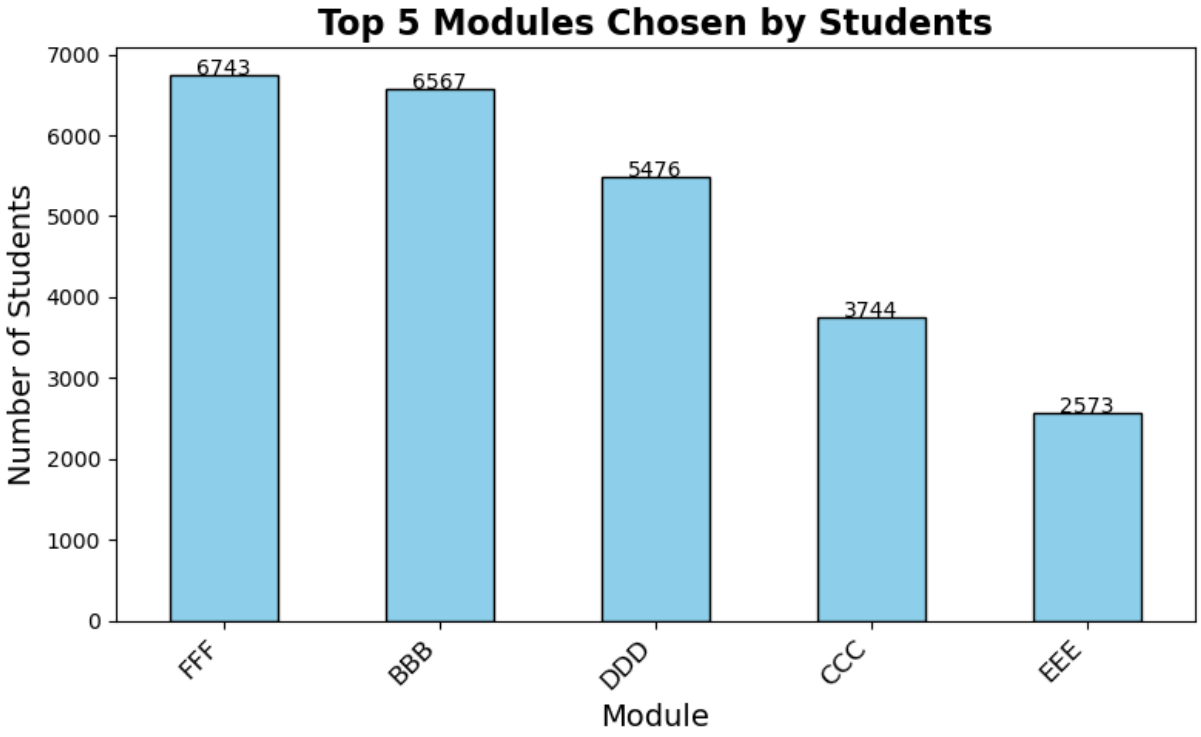



*Figure 12: Top 5 Modules Chosen by Students*

The FFF and DDD modules have the highest numbers of 'Fail' cases. (Figure 13) Therefore, various changes should be implemented to reduce the 'Fail' outcomes in these courses.

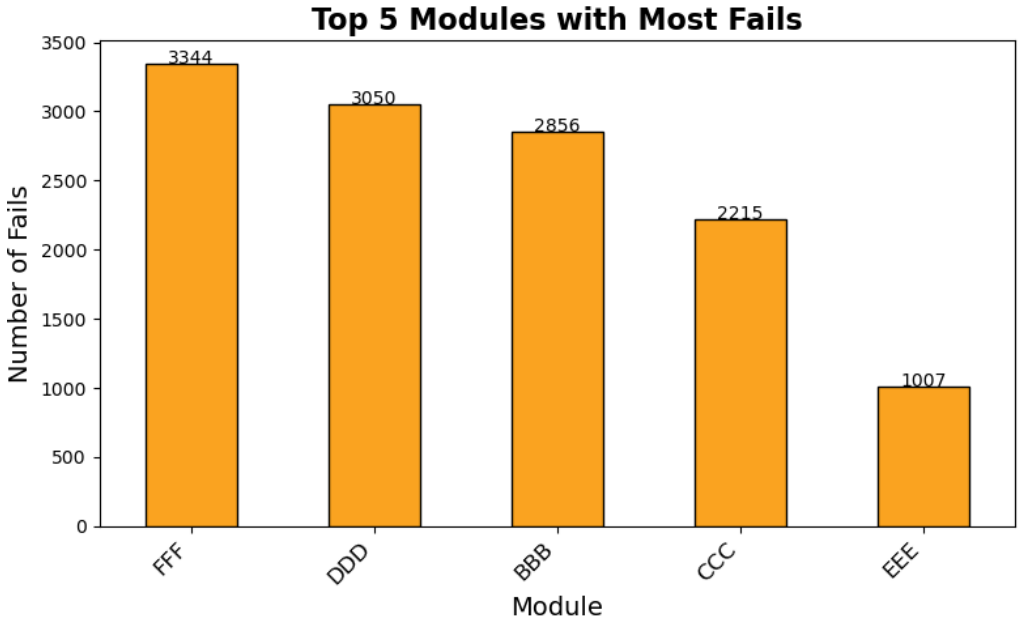



*Figure 13: Top 5 Modules with Most Fails*

As shown in the graph below, most of the population consists of young people. (Figure 14) One possible reason for this could be that online materials are designed to cater to younger users. Steps should be taken to increase the participation of older individuals. For example, making the materials more inclusive for advanced age groups will be useful.

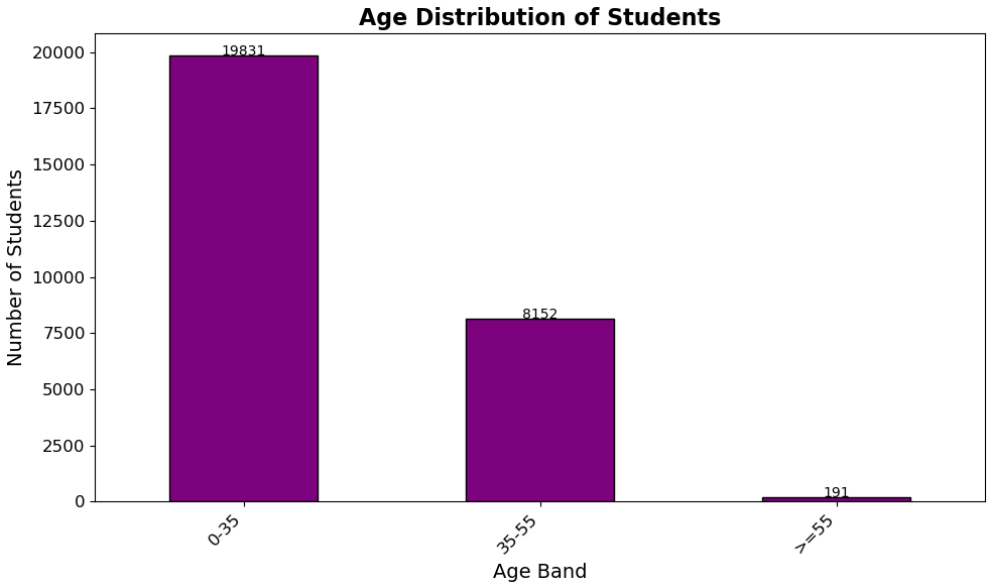



*Figure 14: Age Distribution of Students*

As the graph below (Figure 15) shows, most students do not actively use the VLE system. A very small fraction of students ever use online learning resources frequently. Several initiatives may be designed to increase usage of the VLE system in order to further facilitate student success.

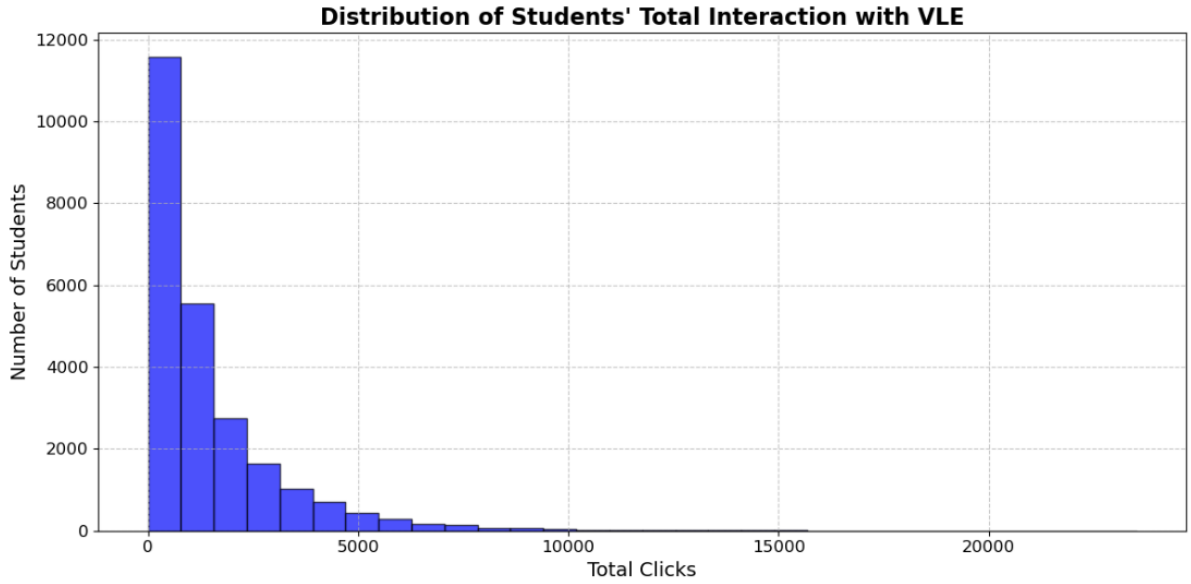



*Figure 15: Total Interaction with VLE*

There is a significant relationship between socioeconomic status and students' achievement levels. (Figure 16) For example, groups with lower socioeconomic status generally have lower success rates. This indicates that additional support should be provided to students who fall within the lower levels of the IMD band.

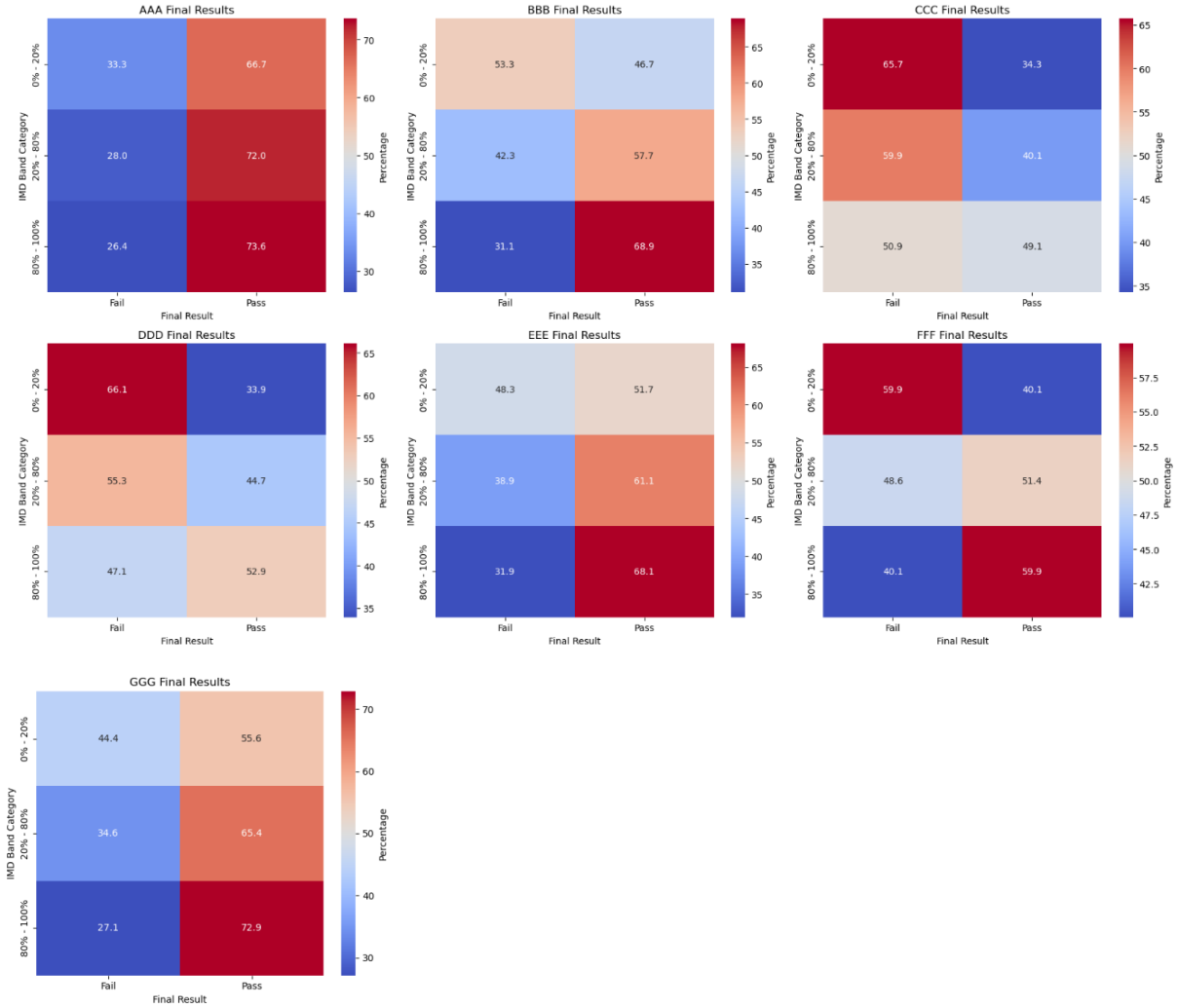



*Figure 16: Imd Band and Final Results*

The average scores of students in some courses differ significantly from those in others. (Figure 17) For instance, the EEE and AAA modules have noticeably higher average scores than other modules.

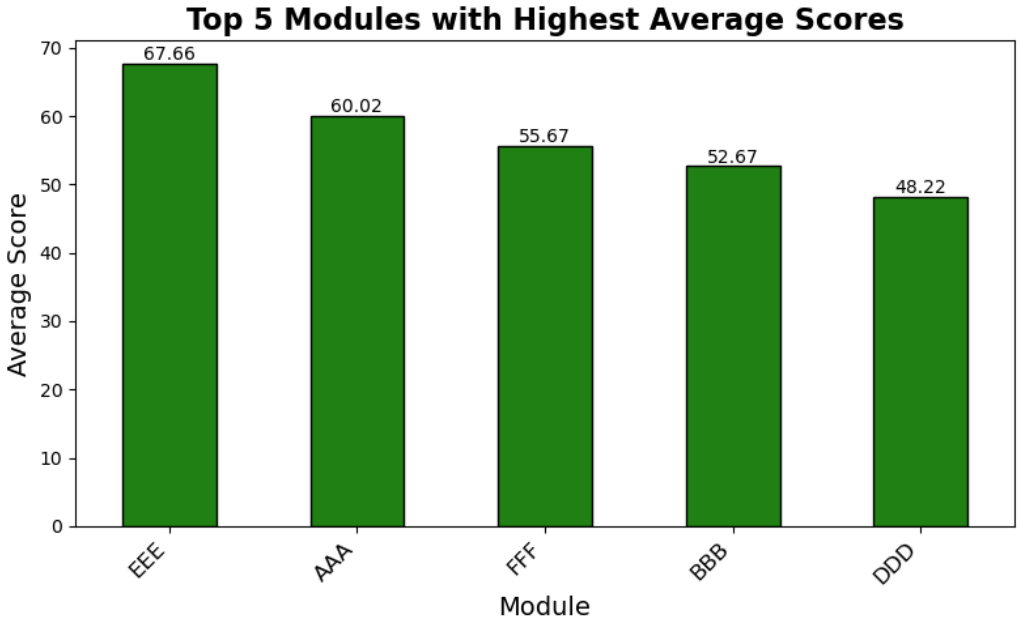



*Figure 17: Top 5 Modules with Highest Average Scores*

As shown in the graph below (Figure 18), there is an anomaly in the scoring for the GGG course. The evaluation weights for the GGG course apply only to assessment_Type='Exam', but no column is labeled 'Exam' in the student_assessment table (Figure 19). Consequently, when calculating the students' overall_grade information, the GGG course grade is zero.

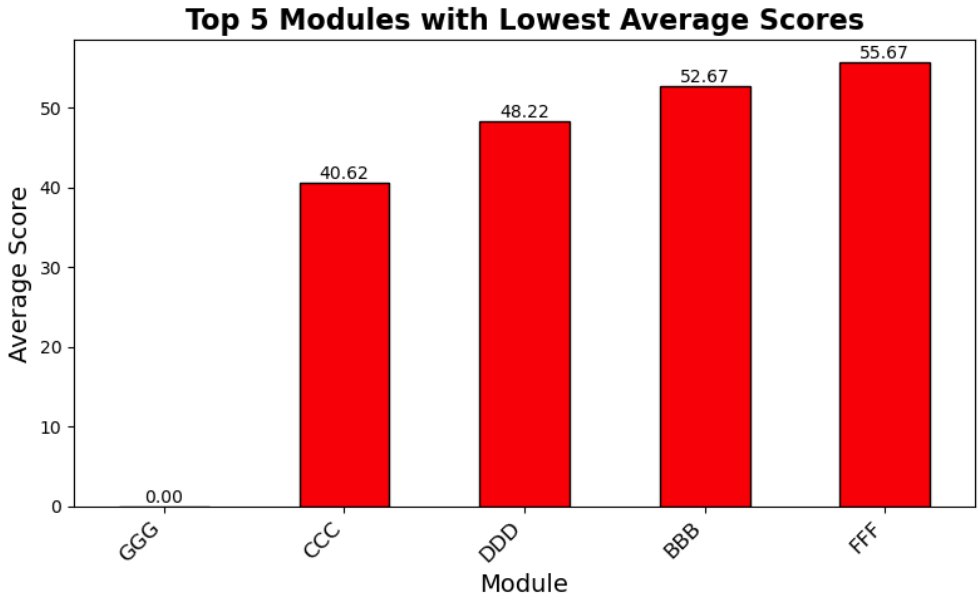



*Figure 18: Top 5 Modules with Lowest Average Scores*

| id_assessment | assessment_type | date | weight |
|---|---|---|---|
| 37418 | CMA | 229 | 0 |
| 37419 | CMA | 229 | 0 |
| 37420 | CMA | 229 | 0 |
| 37421 | CMA | 229 | 0 |
| 37422 | CMA | 229 | 0 |
| 37423 | CMA | 229 | 0 |
| 37415 | TMA | 61 | 0 |
| 37416 | TMA | 124 | 0 |
| 37417 | TMA | 173 | 0 |
| 37424 | Exam | 229 | 100 |

*Figure 19: Evaluation Criteria for GGG*

# Methods

A hypothesis test was conducted to determine whether there is a significant difference between VLE interaction and student achievement. In this hypothesis test, an independent two-sample t-test was performed to compare the total number of VLE clicks for the "Pass" (Successful) and "Fail" (Unsuccessful) groups for each module. This test determines whether there is a significant difference between the two groups.

This method is appropriate as it can also be used when the assumption of equal variance is not met.

The hypothesis is: "There is a difference in VLE usage between successful and unsuccessful students."

| | Module | T-Statistic | P-Value | Comment |
|---|---|---|---|---|
| 0 | AAA | 8.58 | 1.45e-16 | There is a significant difference (p < 0.05). Passing students used the VLE more. |
| 1 | DDD | 30.26 | 1.62e-179 | There is a significant difference (p < 0.05). Passing students used the VLE more. |
| 2 | BBB | 28.10 | 8.42e-163 | There is a significant difference (p < 0.05). Passing students used the VLE more. |
| 3 | CCC | 27.45 | 2.03e-141 | There is a significant difference (p < 0.05). Passing students used the VLE more. |
| 4 | GGG | 21.63 | 1.50e-92 | There is a significant difference (p < 0.05). Passing students used the VLE more. |
| 5 | EEE | 32.13 | 1.13e-185 | There is a significant difference (p < 0.05). Passing students used the VLE more. |
| 6 | FFF | 56.87 | 0.00e+00 | There is a significant difference (p < 0.05). Passing students used the VLE more. |

*Figure 20: t-test result*

Since the p-value for all models is less than 0.05, it is evident that there is a significant difference between the two groups (Figure 20). As the T-statistic value is positive, it indicates that successful students utilize the VLE system more frequently. The highest T-statistic value is observed in the FFF module. Considering these findings, initiating efforts to increase VLE usage among unsuccessful groups will likely improve overall success rates.

## Results

The data used in this study were obtained from the Open University Learning Analytics Dataset. The study's primary objective is to determine whether there is an interaction between students' achievements and the VLE system and to predict students' academic performance. The following features were utilized in the model developed for this purpose:

- Demographic Information: Age band, highest education, region, and IMD band
- VLE Interactions: Total sum click
- Performance Indicators: Weighted score and overall score

Two primary algorithms were used during the modeling process:

### *Linear Regression*

A linear relationship was explored between the Total Sum Click and Overall Grade features. Thus, a linear model was constructed using these two features. However, the RMS value was measured at 27.55 (Figure 21). Due to the model's low success, it was concluded that the linear model was not suitable for the data structure. As illustrated in the graph, the data points in the dataset do not align well with the linear line. Therefore, an alternative model was required.

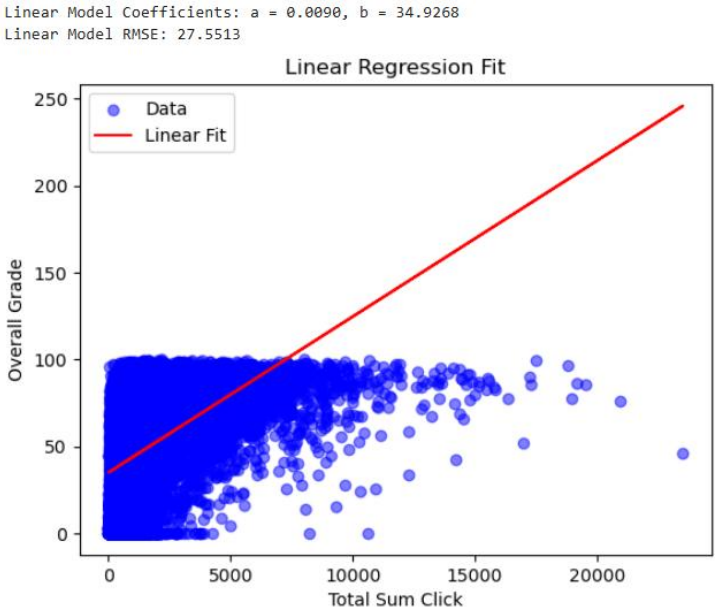



*Figure 21: Linear Regression Fit*

### *Logistic Regression*

A logistic regression model was developed to predict students' success or failure status rather than their success grade. Categorical variables in the final dataset that could have a meaningful relationship in the model were encoded. The first of the encoded columns was dropped to avoid the linear dependency risk, commonly called the dummy variable trap. The data type of the converted categorical fields was then set to an integer.

Subsequently, the student's success was modeled using the following variables: total_sum_click, code_module, code_presentation, region, imd_band, gender, age_band, highest_education, and disability. The coefficients, standard errors, z-scores, p-values, and 95% confidence intervals of the Logistic Regression model were determined based on the model results.

MNLogit Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | final_result_upd_numeric | No. Observations: | 19642 |
| Model: | MNLogit | Df Residuals: | 19602 |
| Method: | MLE | Df Model: | 39 |
| Date: | Thu, 02 Jan 2025 | Pseudo R-squ.: | 0.2815 |
| Time: | 22:14:38 | Log-Likelihood: | -9518.4 |
| converged: | True | LL-Null: | -13248. |
| Covariance Type: | nonrobust | LLR p-value: | 0.000 |

*Figure 22: Logistic Regression Results*

These results show that (Figure 22), the dataset contains 19,642 observation records. Based on the LLR p-value, the model is statistically significant.

| final_result_upd_numeric=1 | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | -1.3985 | 0.156 | -8.954 | 0.000 | -1.705 | -1.092 |
| total_sum_click | 0.0016 | 2.98e-05 | 54.081 | 0.000 | 0.002 | 0.002 |
| code_module_BBB | 0.6891 | 0.124 | 5.542 | 0.000 | 0.445 | 0.933 |
| code_module_CCC | -1.0394 | 0.128 | -8.152 | 0.000 | -1.289 | -0.790 |
| code_module_DDD | -0.4627 | 0.122 | -3.778 | 0.000 | -0.703 | -0.223 |
| code_module_EEE | -0.3530 | 0.132 | -2.671 | 0.008 | -0.612 | -0.094 |
| code_module_FFF | -2.1817 | 0.129 | -16.880 | 0.000 | -2.435 | -1.928 |
| code_module_GGG | 1.2975 | 0.133 | 9.790 | 0.000 | 1.038 | 1.557 |
| code_presentation_2013J | 0.5212 | 0.060 | 8.756 | 0.000 | 0.405 | 0.638 |
| code_presentation_2014B | 0.4131 | 0.062 | 6.663 | 0.000 | 0.292 | 0.535 |
| code_presentation_2014J | 0.5589 | 0.059 | 9.469 | 0.000 | 0.443 | 0.675 |
| region_East Midlands Region | 0.0017 | 0.085 | 0.020 | 0.984 | -0.164 | 0.168 |
| region_Ireland | 0.0190 | 0.113 | 0.168 | 0.867 | -0.203 | 0.241 |
| region_London Region | -0.0598 | 0.080 | -0.748 | 0.454 | -0.217 | 0.097 |
| region_North Region | -0.1712 | 0.113 | -1.512 | 0.131 | -0.393 | 0.051 |
| region_North Western Region | -0.1077 | 0.083 | -1.303 | 0.193 | -0.270 | 0.054 |
| region_Scotland | -0.1447 | 0.078 | -1.857 | 0.063 | -0.298 | 0.008 |
| region_South East Region | 0.2002 | 0.088 | 2.286 | 0.022 | 0.029 | 0.372 |
| region_South Region | 0.0931 | 0.079 | 1.177 | 0.239 | -0.062 | 0.248 |
| region_South West Region | 0.1506 | 0.083 | 1.810 | 0.070 | -0.013 | 0.314 |
| region_Wales | -0.1597 | 0.086 | -1.866 | 0.062 | -0.327 | 0.008 |
| region_West Midlands Region | -0.0494 | 0.084 | -0.589 | 0.556 | -0.214 | 0.115 |
| region_Yorkshire Region | -0.0560 | 0.090 | -0.622 | 0.534 | -0.233 | 0.121 |
| imd_band_10-20% | 0.0458 | 0.078 | 0.587 | 0.557 | -0.107 | 0.199 |
| imd_band_20-30% | 0.1557 | 0.077 | 2.017 | 0.044 | 0.004 | 0.307 |
| imd_band_30-40% | 0.3675 | 0.077 | 4.743 | 0.000 | 0.216 | 0.519 |
| imd_band_40-50% | 0.3317 | 0.079 | 4.194 | 0.000 | 0.177 | 0.487 |
| imd_band_50-60% | 0.3215 | 0.080 | 4.033 | 0.000 | 0.165 | 0.478 |
| imd_band_60-70% | 0.4333 | 0.082 | 5.285 | 0.000 | 0.273 | 0.594 |
| imd_band_70-80% | 0.4479 | 0.082 | 5.457 | 0.000 | 0.287 | 0.609 |
| imd_band_80-90% | 0.5358 | 0.085 | 6.336 | 0.000 | 0.370 | 0.701 |
| imd_band_90-100% | 0.6629 | 0.088 | 7.508 | 0.000 | 0.490 | 0.836 |
| gender_M | 0.2168 | 0.045 | 4.781 | 0.000 | 0.128 | 0.306 |
| age_band_35-55 | -0.3021 | 0.041 | -7.288 | 0.000 | -0.383 | -0.221 |
| age_band_>=55 | -0.7570 | 0.254 | -2.985 | 0.003 | -1.254 | -0.260 |
| highest_education_HE Qualification | 0.0150 | 0.056 | 0.266 | 0.790 | -0.096 | 0.126 |
| highest_education_Lower Than A Level | -0.6740 | 0.040 | -17.013 | 0.000 | -0.752 | -0.596 |
| highest_education_No Formal quals | -0.9231 | 0.202 | -4.566 | 0.000 | -1.319 | -0.527 |
| highest_education_Post Graduate Qualification | 0.2964 | 0.255 | 1.162 | 0.245 | -0.203 | 0.796 |
| disability_Y | -0.2993 | 0.060 | -4.948 | 0.000 | -0.418 | -0.181 |

*Figure 23: Logistic Regression Summary*

According to the output above (Figure 23), it is evident that students' use of the VLE system positively impacts their success. (This is indicated by the coef: 0.0016 value for total_sum_click in the table.)

### Module Analysis

- The BBB code module is a course that students are more likely to pass (Coef: 0.6891 for code_module_BBB.)
- The CCC code module reduces students' likelihood of passing. (Coef: -1.0394 for code_module_CCC.)
- The FFF code module is one of the riskiest courses as it significantly increases the probability of failure. (Coef: -2.1817 for code_module_FFF.)
- The GGG code module increases the probability of passing. (Coef: 1.2975 for code_module_GGG.)

### IMD Analysis

A higher IMD (indicating a higher level of well-being) significantly increases the likelihood of success. Conversely, a lower level of well-being increases the probability of failure.

### Gender Analysis

Male students are slightly more likely to succeed compared to female students.

### Age Analysis

Older students are much more likely to fail compared to younger students.

### Highest Education Analysis

Students with an education level of "Lower Than A Level" or "No Formal Qualifications" are more likely to fail.

### Disability Status Analysis

Students with any disability are at a disadvantage when it comes to achieving success.

### Z-Score Analysis

The Z-score provides insights into which variables significantly impact the dependent variables. For example, variables such as total_sum_click, code_module_BBB, imd_band_80-90%, and gender_M have significant relationships with the model as their Z-scores are below 0.05. Conversely, variables such as region_London Region and region_West Midlands Region do not have a substantial relationship with the model.

### Model Performance

When all encoded fields are used in the model, the performance metrics are as follows:

Accuracy: 0.7699

Precision: 0.7682

Recall: 0.7699

F1 Score: 0.7685

These results indicate that the model performs well overall, with 76% accuracy and balanced Precision/Recall scores. However, the model uses a large number of features. To optimize the performance of the logistic regression model, the ideal number of features was determined using the Recursive Feature Elimination (RFE) method, which identified the most impactful features.

```
Selecting feature's count: 5
Model Score: 0.7748192648406476
Selected features: ['total_sum_click', 'code_module_BBB', 'code_module_FFF', 'code_module_GGG', 'highest_education_Lower Than A Level']

Selecting feature's count: 10
Model Score: 0.7760411363404949
Selected features: ['total_sum_click', 'code_module_BBB', 'code_module_CCC', 'code_module_FFF', 'code_module_GGG', 'code_presentation_2013J', 'code_pr
esentation_2014B', 'code_presentation_2014J', 'imd_band_90-100%', 'highest_education_Lower Than A Level']

Selecting feature's count: 15
Model Score: 0.7773139191528358
Selected features: ['total_sum_click', 'code_module_BBB', 'code_module_CCC', 'code_module_DDD', 'code_module_EEE', 'code_module_FFF', 'code_module_GG
G', 'code_presentation_2013J', 'code_presentation_2014B', 'code_presentation_2014J', 'imd_band_80-90%', 'imd_band_90-100%', 'gender_M', 'age_band_35-5
5', 'highest_education_Lower Than A Level']

Selecting feature's count: 20
Model Score: 0.7798594847775175
Selected features: ['total_sum_click', 'code_module_BBB', 'code_module_CCC', 'code_module_DDD', 'code_module_EEE', 'code_module_FFF', 'code_module_GG
G', 'code_presentation_2013J', 'code_presentation_2014B', 'code_presentation_2014J', 'imd_band_30-40%', 'imd_band_40-50%', 'imd_band_50-60%', 'imd_ban
d_60-70%', 'imd_band_70-80%', 'imd_band_80-90%', 'imd_band_90-100%', 'gender_M', 'age_band_35-55', 'highest_education_Lower Than A Level']
```

*Figure 24: Selecting Features*

According to the results obtained (Figure 24), ten features selected improved the model's score significantly compared to five. However, adding more than ten features did not significantly improve the model's score. Since using fewer effective variables improves model performance, the model was built using ten features.

The model was developed with statistically significant features that contribute highly to model performance. While some features obtained through the RFE method were used for feature selection, they were not adopted directly as determined by RFE. This is because the RFE method selects features to maximize model performance. However, a disadvantage of this approach is that some selected features might not have meaningful relevance in a business context.

For this reason, the model was built using the following features:

total_sum_click, code_module_BBB, code_module_CCC, code_module_FFF, code_module_GGG, imd_band_90-100%, imd_band_70-80%, highest_education_Lower Than A Level, age_band_35-55, code_presentation_2013

MNLogit Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | final_result_upd_numeric | No. Observations: | 19642 |
| Model: | MNLogit | Df Residuals: | 19631 |
| Method: | MLE | Df Model: | 10 |
| Date: | Thu, 02 Jan 2025 | Pseudo R-squ.: | 0.2685 |
| Time: | 23:37:36 | Log-Likelihood: | -9691.2 |
| converged: | True | LL-Null: | -13248. |
| Covariance Type: | nonrobust | LLR p-value: | 0.000 |

| final_result_upd_numeric=1 | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | -1.0441 | 0.044 | -23.566 | 0.000 | -1.131 | -0.957 |
| total_sum_click | 0.0016 | 2.92e-05 | 54.852 | 0.000 | 0.002 | 0.002 |
| code_module_BBB | 0.8778 | 0.047 | 18.724 | 0.000 | 0.786 | 0.970 |
| code_module_CCC | -0.5111 | 0.059 | -8.706 | 0.000 | -0.626 | -0.396 |
| code_module_FFF | -1.7938 | 0.061 | -29.474 | 0.000 | -1.913 | -1.675 |
| code_module_GGG | 1.5812 | 0.067 | 23.570 | 0.000 | 1.450 | 1.713 |
| imd_band_90-100% | 0.4593 | 0.065 | 7.015 | 0.000 | 0.331 | 0.588 |
| imd_band_70-80% | 0.2231 | 0.060 | 3.698 | 0.000 | 0.105 | 0.341 |
| highest_education_Lower Than A Level | -0.6732 | 0.037 | -18.228 | 0.000 | -0.746 | -0.601 |
| age_band_35-55 | -0.2580 | 0.040 | -6.425 | 0.000 | -0.337 | -0.179 |
| code_presentation_2013J | 0.1577 | 0.041 | 3.877 | 0.000 | 0.078 | 0.237 |

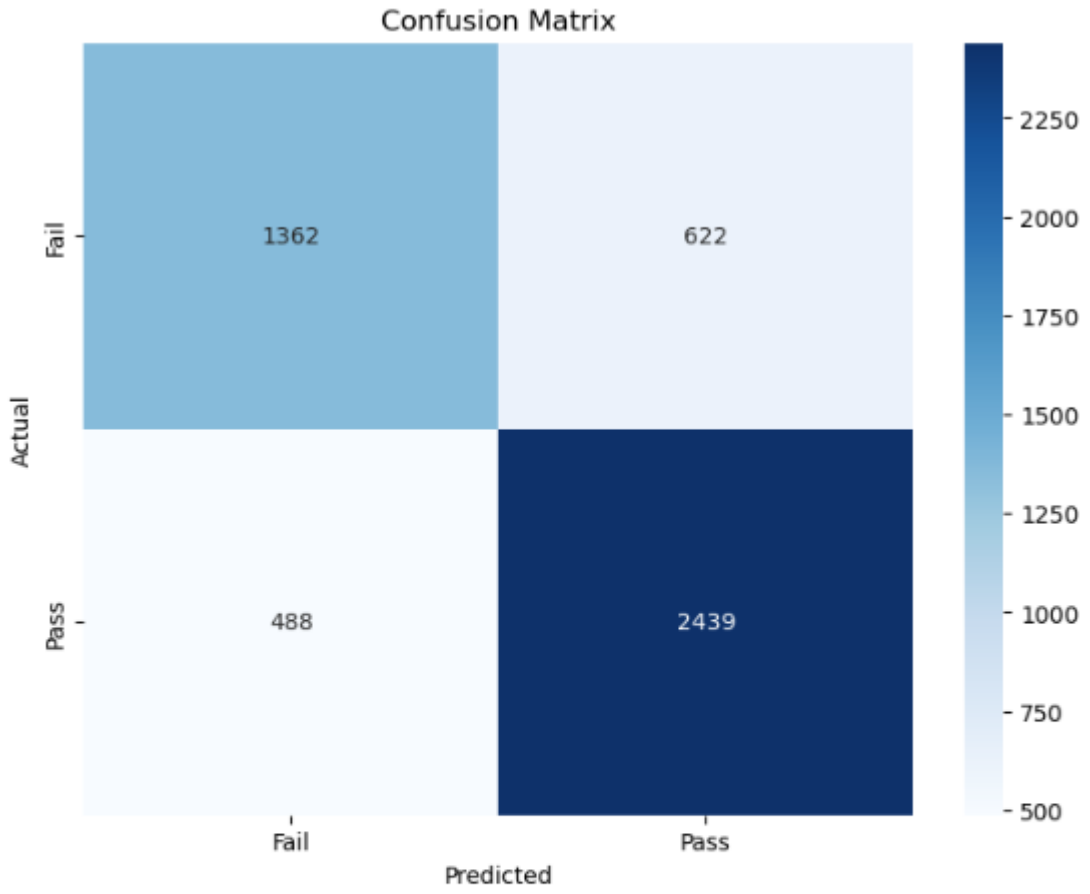
*Figure 25: Logistic Regression Results*

Features such as total_sum_click, code_module_BBB, and imd_band_90-100% have strong positive effects on the probability of success, while features like highest_education_Lower than A Level and age_band_35-55 are factors that increase the likelihood of failure among students (Figure 25).

The other performance metrics of the model are as follows:

- Accuracy: 0.7739767868051314
- Precision: 0.7723237595398864
- Recall: 0.7739767868051314
- F1 Score: 0.7725559648934477

These results give an indication that the model's performance is highly consistent. Also, it was well-balanced in distinguishing classes 'pass' and 'fail'.




Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Fail | 0.74 | 0.69 | 0.71 | 1984 |
| Pass | 0.80 | 0.83 | 0.81 | 2927 |
| accuracy | | | 0.77 | 4911 |
| macro avg | 0.77 | 0.76 | 0.76 | 4911 |
| weighted avg | 0.77 | 0.77 | 0.77 | 4911 |

*Figure 26: Logistic Regression Confusion Matrix*

The overall accuracy rate of the model (77%) indicates good model performance (Figure 26). While the model demonstrates higher performance in the 'Pass' class, it makes more errors in the 'Fail' class. Nevertheless, the accuracy rate remains at a satisfactory level.

# Discussion

According to the model results:

- Features such as Total_sum_click, code_module_BBB, and imd_band_90-100% are strong predictors that increase the likelihood of student success.
- Conversely, factors like Highest_education_Lower Than A Level and age_band_35-55 are key contributors to reducing success rates.

The success observed in the 'Pass' class highlights the positive impact of VLE usage on academic achievement.

## Limitations

The model has certain limitations, the first being the distribution of 'Pass' and 'Fail' students in the dataset. Since students with the 'Fail' label are in the minority, the recall value for the 'Fail' class is low. This indicates that the 'Fail' class is not predicted as accurately as desired.

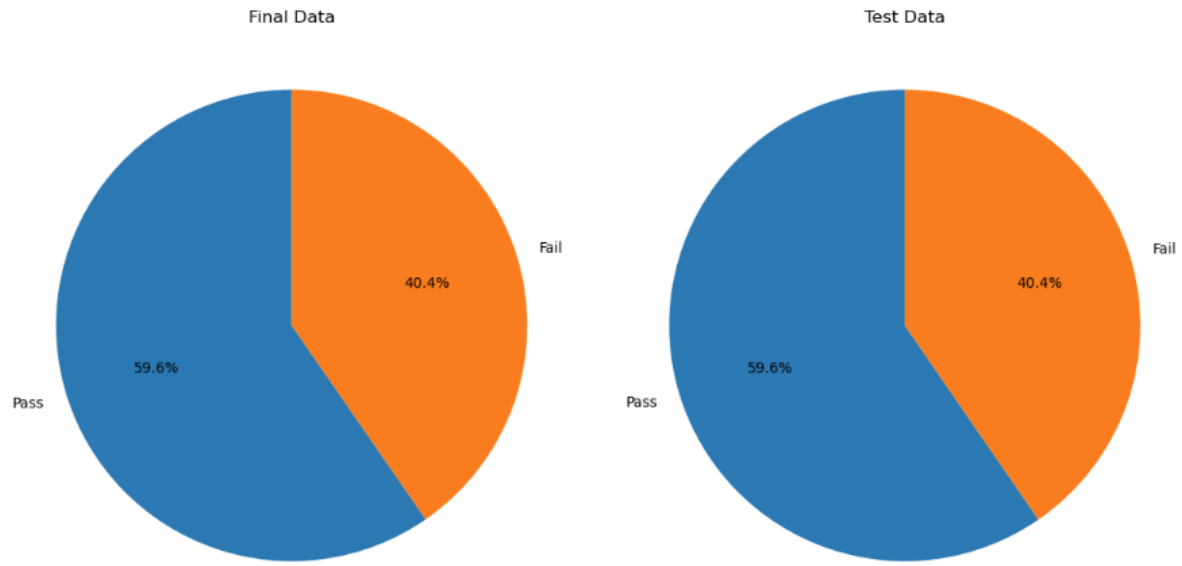



*Figure 27: Distribution of Fail&Pass*

At the same time, considering that this study was conducted exclusively with the Open University dataset, it is clear that it would not be appropriate to draw a general conclusion regarding the impact of VLE activity. A larger and more diverse dataset is required to expand the scope of the study.

In addition, although the dummy variable trap was avoided for the independent variables used in the model, some variables exhibited high correlation rates. Certain imd_band categories showing similar effects could lead to biases in the model's coefficient estimates, which has limited the model's interpretability.

Furthermore, the GGG course predictions fail to produce realistic outcomes in the model because the overall_score values for the GGG course are consistently recorded as 0 across the dataset. This stems from the fact that, as mentioned earlier in the report, the GGG course lacks any assessment types contributing to the final average other than Exam. Additionally, no Exam-type assessment scores for the GGG course are present in the dataset for any student.

Finally, in order to clearly determine how much VLEs are being used, the quality and nature of the clicks of the interactions should be analyzed as well. This is because some clicks from students that do not bear necessary or efficient material would not count toward a student's success. Thus, the nature of students' interaction details must be factored into the data set.

# Conclusion

The research was performed to understand and predict the success status of Open University students. Logistic Regression model analyses show that online interactions, module choices, and demographic factors are very influential on students' success. The model yielded an acceptable performance with an accuracy rate of 77%.

## Key Findings:

### *Influence of Virtual Interactions:*

The total clicks in the VLE site positively relate to the academic performances of the students. This implies that effective utilization of online interactions promotes success.

### *Course Modules:*

Some modules, such as BBB and GGG, improve students' chances of success, whereas others, like FFF and CCC, increase the risk of failure.

### *Demographic Data:*

As indicated by imd_band_90-100%, high well-being levels enhance the likelihood of success. Conversely, students with lower education levels (Lower Than A Level, No Formal qualifications) and those in older age groups face higher risks of failure.

### *Students with Disabilities:*

The success rates of students with disabilities are notably lower compared to other students.

## Recommendations:

### *Improving Online Interactions:*

Guidance can be given through user guides on how to use the VLE platform or services more effectively for them. Building comprehensive but user-friendly content may bring huge success to the students.

### *Module-Specific Support:*

Additional resources and teaching support services can be introduced for challenging modules like FFF and CCC. This would help students grasp module content more effectively.

### *Support Based on Educational Background:*

Programs that are preparation for enhancing weak points in their basic knowledge could support students who come from a weaker educational background. Such initiatives would greatly improve performance.

### *Age-Appropriate Learning:*

Technology training or personalized counseling could lighten the burden of adjustment to digital platforms, making the learning process much more effective for mature students.

### *Assistance for Students with Disabilities:*

Tailored solutions to meet the unique needs of students with disabilities can increase their success rates. Examples include accessible learning materials and individualized counseling services.

# References

Knowledge Media Institute (KMI), 2024.*Open University Learning Analytics Dataset Entity Relationship Diagram*. Available at: https://analyse.kmi.open.ac.uk/open_dataset [Accessed 29 December 2024].