INTRODUCTION

I have been assigned as a data scientist to check if VLE courses provided by Open University are improving student grades and whether we can predict student's grades. I will be exploring interactions with VLE and some demographics to check if it improves scores so that I can understand which features are essential to predict grades as well as the final results. I will also be performing some hypothesis testing to check the significance of various demographics on the code modules selected by the students. Here we investigate the OULAD dataset which has information about courses, students and their interactions with Virtual Learning Environment (VLE) for seven selected courses (called modules).

Key Components of the Dataset

The dataset used for this analysis is the Open University Learning Analytics Dataset (OULAD), which provides comprehensive information on courses, student interactions, and academic outcomes. Below is an overview of the key components of the dataset.

The **Courses Information** is stored in the courses.csv file. This dataset includes fields such as code_module, which uniquely identifies each module, and code_presentation, which denotes the module's presentation (e.g., "B" for February and "J" for October). The length field specifies the duration of the module in days. Notably, February and October presentations may exhibit structural differences, necessitating separate analyses for these periods.

The **Assessments Data**, housed in assessments.csv, links to courses through code_module and code_presentation. This file contains unique assessment identifiers (id_assessment), the type of assessment (e.g., TMA, CMA, or Exam), and the date of submission relative to the course start. Additionally, the weight field indicates the percentage weight of each assessment in the overall grade. Final exams often occur at the end of the presentation, providing critical data for evaluating student performance.

The **Virtual Learning Environment Materials** data is found in vle.csv, detailing online resources available to students. Each material is identified by id_site and linked to specific courses using code_module and code_presentation. The activity_type field describes the material's role, while week_from and week_to indicate its planned usage weeks. These materials include a variety of formats, such as PDFs and HTML, crucial for analyzing the scope of learning resources.

The **Student Demographics and Results** are provided in the studentInfo.csv file. This dataset contains demographic details, including gender, region, age_band, and disability. It also includes education-specific fields such as highest_education, studied_credits, and num_of_prev_attempts. The final_result field categorizes outcomes as Pass, Fail, Withdrawn, or other statuses. Data in this file links students to their courses via code_module and code_presentation.

The **Student Registration Details**, available in studentRegistration.csv, record the registration timeline for each student. Fields include date_registration, which marks the days relative to the course start when the student registered, and date_unregistration, which denotes withdrawal

timing. This dataset correlates closely with withdrawal statuses in studentInfo.csv, helping analyze retention trends.

The **Assessment Results** in studentAssessment.csv link assessments to students through id_assessment and id_student. This file records the date_submitted, representing the days since the course start when an assessment was submitted, and the score, ranging from 0 to 100.

The **Student VLE Interactions**, captured in studentVle.csv, document engagement with online learning materials. Each entry is tied to id_site, the identifier for VLE materials, and includes the date of interaction in relation to the course start. The sum_click field quantifies the number of clicks or interactions, providing a detailed view of student engagement with the VLE.

Data Wrangling, Cleaning and Exploration:

Let's start by answering the questions asked in the report. We will keep cleaning the data as when required along the way.

1. **Top 5 modules chosen by students** – From the dataset exploration we find that the studentInfo.csv has records of the students along with the modules they are enrolled in. The dataset doesn't have any NaN values. We want the value counts of the 'code_module' column. On performing the operation, we find the top 5 modules students enrolled in as:

Code Module	Count
BBB	7909
FFF	7762
DDD	6272
CCC	4434
EEE	2934

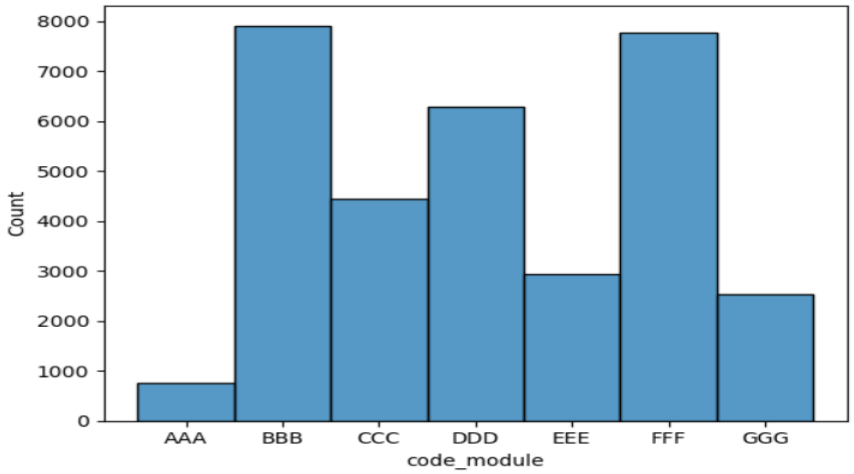
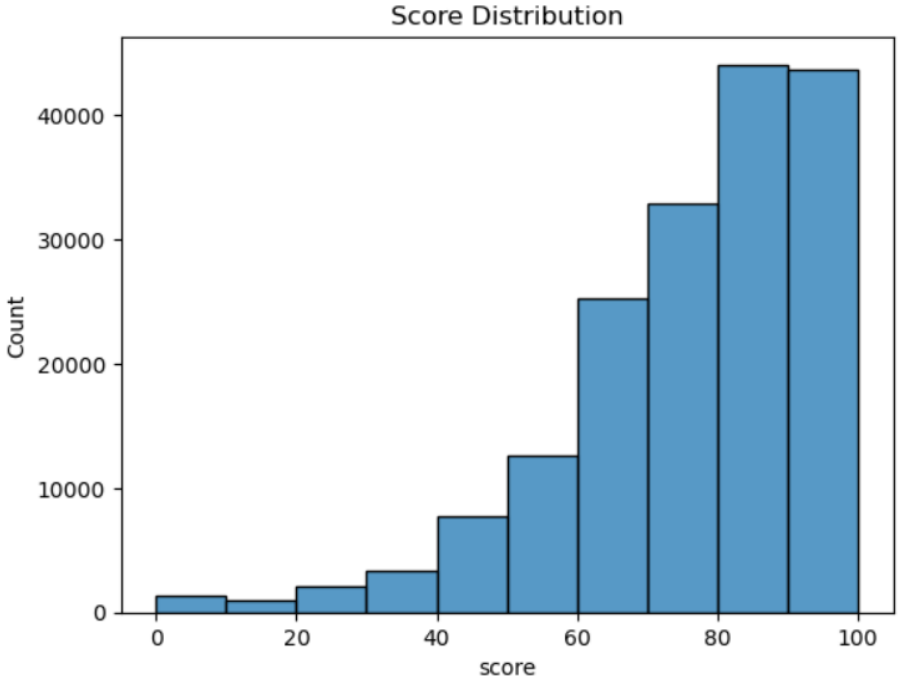


Fig 1. Top: Top 5 modules chosen by students, Bottom: Modulewise count

2. **The modules with the highest and lowest average score –** To find the highest and lowest average scores we use studentAssessment.csv and assessments.csv. In

studentAssessment.csv cleaning and Wrangling is necessary here because we can see the scores are of type object and on investigation it has '?' in it. So, we remove the rows with '?' and convert it to integer type and merge it with the assessments into a variable assessment_scores. We perform inner join to remove all missing data points. Then we group by modules and find the average of scores.



Code_ module	Average Score
AAA	69.030515
DDD	70.090800
CCC	73.261398
BBB	76.706368
FFF	77.707590
GGG	79.700493
EEE	81.180066

Fig 2. Top: Score distribution, Bottom : Module-wise average score

So modules with highest average scores are EEE, GGG, FFF, BBB and CCC and modules with lowest average scores are AAA, DDD, CCC, BBB and FFF.

3. **Top 5 modules with the most number of fails –** To find this we use studentInfo.csv again where we mark each fail as 1 if final_result is 'Fail' in a new column 'is_fail' and then group by code_module and add up the values in 'is_fail'. Drop the 'is_fail'.

Code Module	Count
BBB	1767
FFF	1711
DDD	1412
CCC	781
GGG	728

Fig 3. Modules with most number of fails.

4. **Age Distribution of Students –**

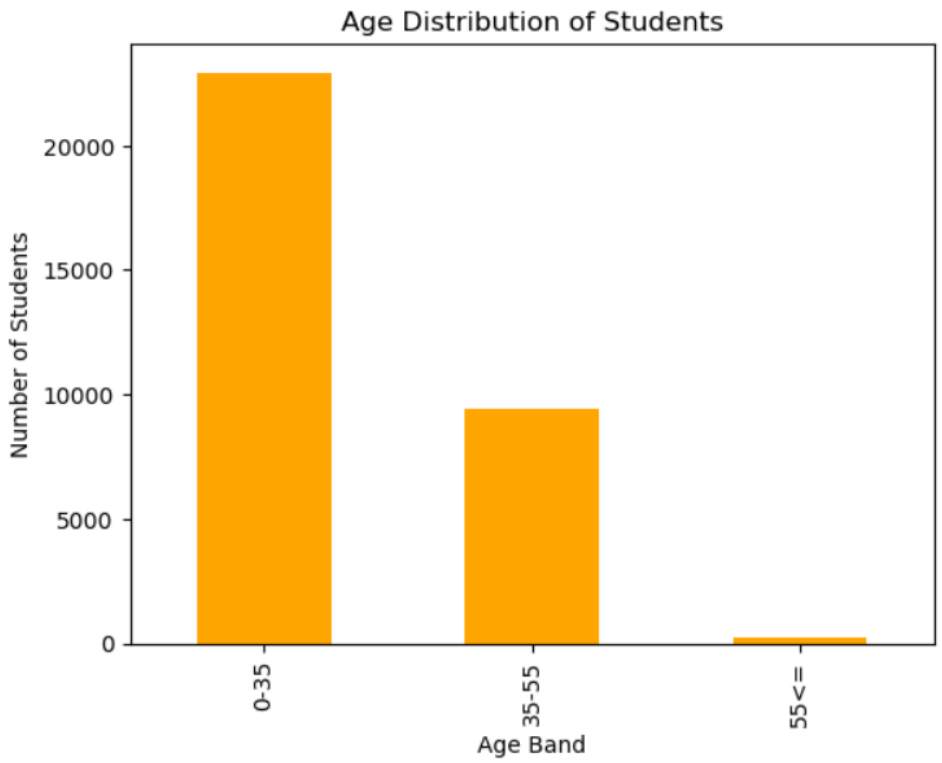


Fig 4. Age distribution

5. **Weekly activity patterns of the students' interaction with the VLE -** studentVLE.csv has data on sum_clicks and the day which is wrangled to into weeks and plotted.

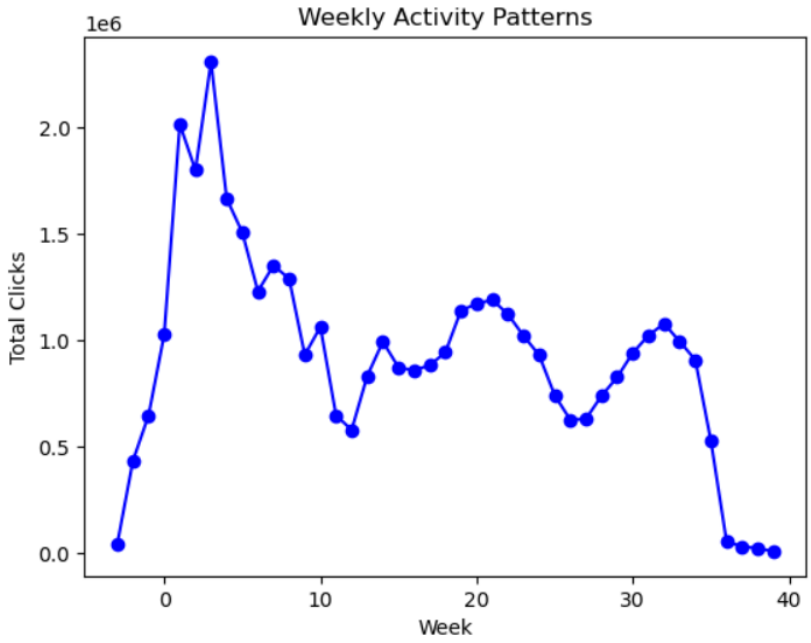


Fig 5. Weekly click activity pattern

Data wrangling and cleaning for statistical analysis –

We perform a series of joins to prepare the dataset before conducting hypothesis testing, statistical modelling, and data cleaning. During the analysis, we discovered that some assessments had not been taken by any students, so these assessments need to be removed. This is necessary before calculating the aggregate score for each student in each code module and for individual code presentations, as well as determining the total weights per module and per code presentation.

Next, we calculate the total weights and compute the weighted scores in the assessment_scores dataset by multiplying the weight of each assessment by the student's score. We then sum the scores for each student within each code module and code presentation, dividing by the total weight associated with those assessments, which is stored in the aggregated_weighted_scores dataset.

We then merge this data with student_info and student_vle datasets, thus consolidating all the required information for each student. Finally, we examine the distribution of the final_result.

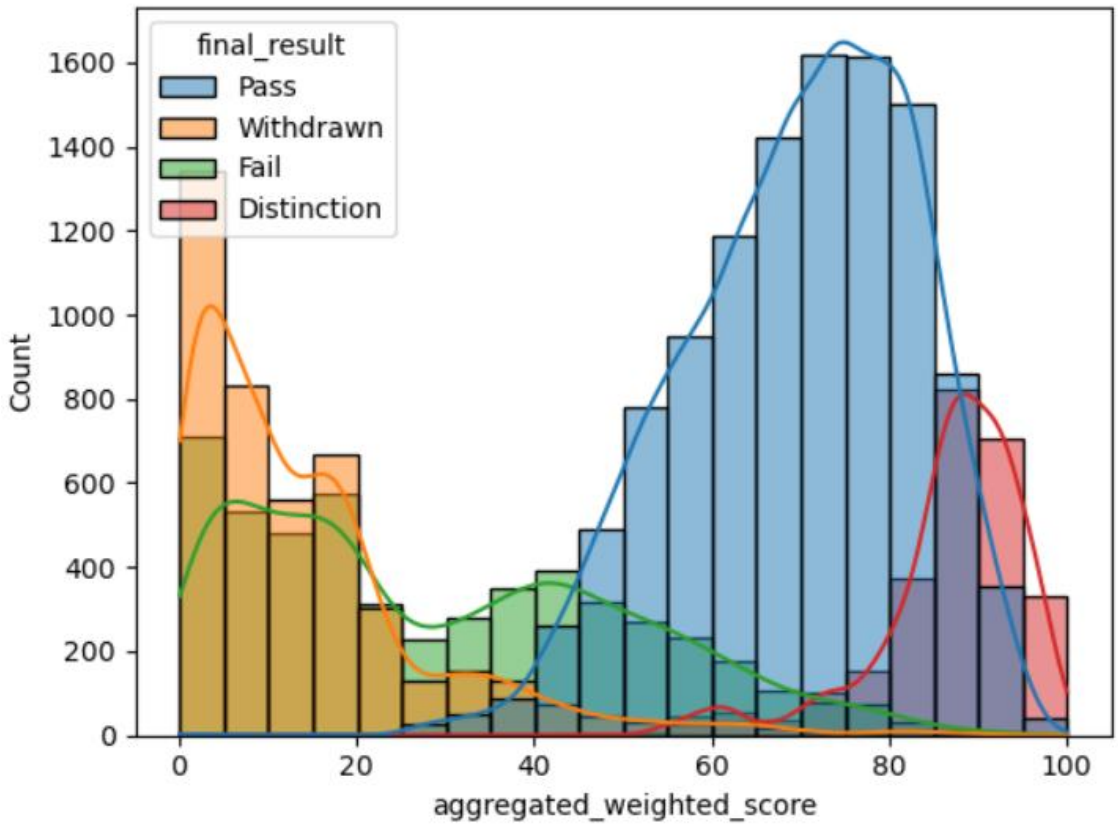


Fig 6. Histogram and KDE for the aggregate score with final result demarcation

Mathematical Modelling

Hypothesis Testing for checking if VLE interaction is improving the scores:

We use the pearsonr test to check there is correlation between sum_click and the aggregated_weighted_score. I want to check whether a greater number of clicks can result in higher scores.

Hypothesis Testing:

- Null Hypothesis (H_0): Interaction with VLE has no effect on students' scores.
- Alternative Hypothesis (H_1): Higher interaction with VLE is associated with higher scores and there is a linear relationship.

One – hot encoding some of the categorical and ordinal data:

We one-hot encode some of the columns like gender, disability and age_band and check for correlation. In this analysis, features such as imd_band, highest_education, and region were not one-hot encoded to avoid overcrowding the dataset with additional variables, as their correlation with the target variable was found to be relatively low. However, it is important to note that these

features can be modelled similarly to evaluate their significance. While they were excluded from the current focus due to their lower correlation, they remain valuable for further exploration and could be incorporated in future analyses to assess their potential impact.

Correlation Check for module AAA:

We want to check for multicollinearity for code module AAA. We do a spearman correlation test as well as pearson test for this.

Statistical Modelling:

OLS and Logistic Regression:

Now we start statistically modelling our data based on the above information. We iterate over the features to find a statistically suitable parameters for setting up a linear regression model. We find that sum_clicks is the only statistically significant parameter here. We also investigate the Recursive Feature Elimination to select features to fit our statistical modelling. We begin by fitting an OLS model and look into the residuals too followed by a classification by logistic regression.

Decision Tree Classifier Evaluation

Model Overview

We trained a decision tree classifier with a maximum depth of 4 using a training dataset created via a train-test split. The following results summarize the model's performance on the test dataset, broken down by class, with overall accuracy, precision, recall, and F1-score metrics.

Results

Interpretation of Hypothesis Testing:

A p-value of 0.0 (or effectively very close to zero) means the result is statistically significant at any reasonable significance level (e.g., $\alpha=0.05$ or $\alpha=0.01$) and Pearson Correlation: 0.4837320001517453 signifies positive correlation.

We **reject the null hypothesis** and conclude that there is a statistically significant linear relationship between the two variables hence suggest that interaction with the VLE is improving students grades.

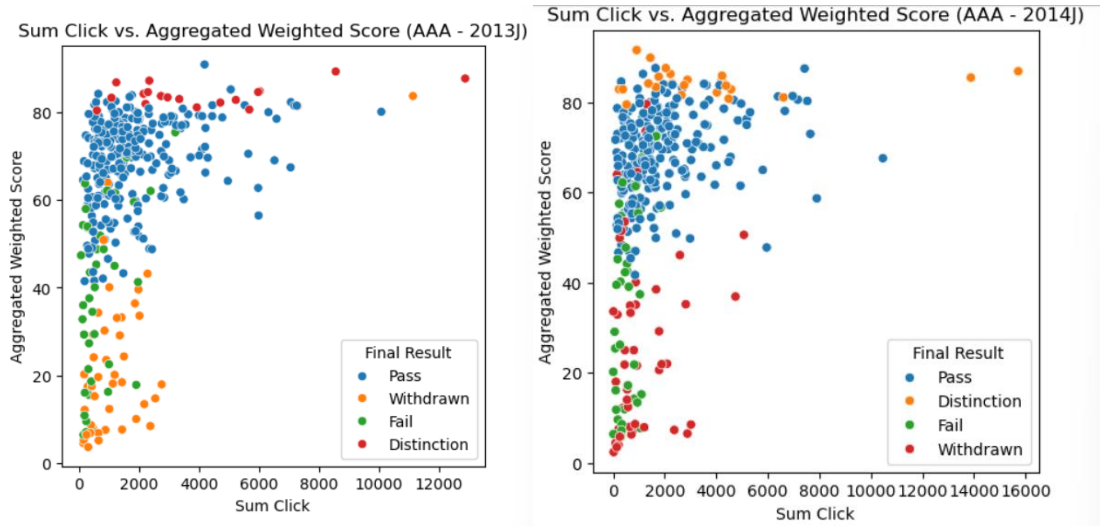


Fig .7 Exploratory analysis between Sum_click and aggregated weighted score for module AAA

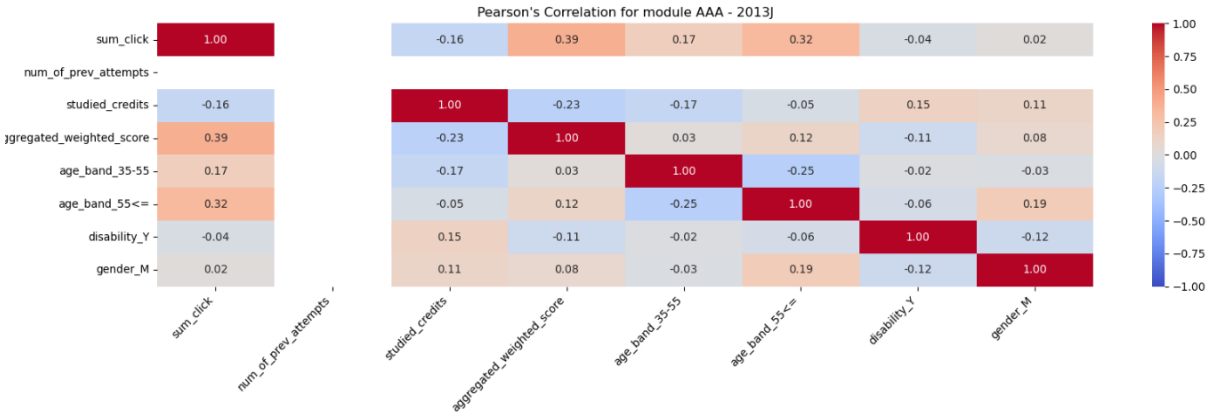


Fig .8 Pearson's Correlation Matrix for course AAA-2013J

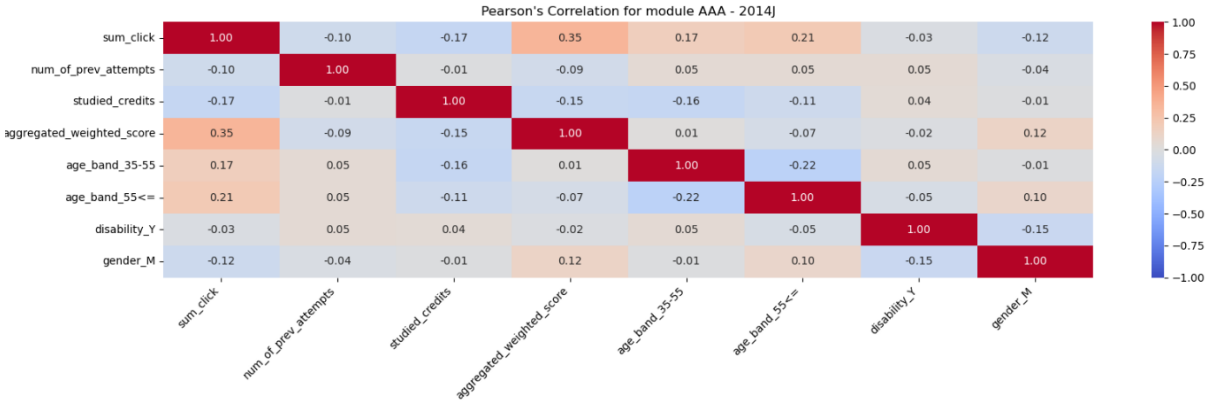


Fig .9 Pearson's Correlation Matrix for course AAA-2014J

We find that 2013J has only one value in the num_of_prev_attempts column. The collinearity is high in case of sum_clicks and score and between sum_clicks and age_band_55<=. On fitting the OLS

we find that the result was giving a very bad R^2 value, so we remove the error term. Now the resultant OLS regression result is better.

Interpretation of OLS:

OLS Regression Results

Dep. Variable:	aggregated_weighted_score	**R-squared (uncentered):**	0.554
Model:	OLS	**Adj. R-squared (uncentered):**	0.554
Method:	Least Squares	**F-statistic:**	874.9
Date:	Sat, 04 Jan 2025	**Prob (F-statistic):**	1.52e-125
Time:	16:19:36	**Log-Likelihood:**	-3645.9
No. Observations:	704	**AIC:**	7294.
Df Residuals:	703	**BIC:**	7298.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
sum_click	0.0190	0.001	29.580	0.000	0.018	0.020

Omnibus:	338.192	**Durbin-Watson:**	1.077
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	2358.453
Skew:	-2.046	**Prob(JB):**	0.00
Kurtosis:	10.978	**Cond. No.**	1.00

The Ordinary Least Squares (OLS) regression revealed that 55.4% of the variance in the dependent variable (aggregated_weighted_score) is explained by the independent variable (sum_click) when no intercept is included. This R-squared value indicates a moderate fit to the data. The adjusted R-squared remains the same, as there is only one predictor in the model.

The coefficient for sum_click is 0.019, indicating that for every unit increase in sum_click, the predicted aggregated_weighted_score increases by 0.019. This relationship is statistically significant, with a p-value effectively equal to 0 and a t-statistic of 29.580, demonstrating the strength of the predictor. The 95% confidence interval for the coefficient ranges from 0.018 to 0.020, suggesting high precision in the estimate.

Model diagnostics show an F-statistic of 874.9, confirming the overall significance of the model (p-value = 1.52e-125). However, residual diagnostics indicate non-normality, with left-skewed residuals (Skew = -2.046) and heavy tails (Kurtosis = 10.978), as shown by the significant Jarque-Bera test (p-value = 0.00).

Summary of Feature Selection Results using Recursive Feature Elimination

Linear Regression (Prediction Task)

The best-performing linear regression model achieved an R-squared of 0.1826 with is very low suggesting linear regression is not a good model for the problem. Key features included:

- sum_click
- num_of_prev_attempts
- studied_credits
- age_band_35-55
- age_band_55<=
- disability_Y
- gender_M

The predictive performance was relatively low, with sum_click, studied_credits, and num_of_prev_attempts emerging as the most influential predictors of aggregated_weighted_score. Demographic features contributed minimally but provided slight improvements when combined with behavioural features.

Logistic Regression (Classification Task)

The best-performing logistic regression model achieved a score of 0.7691. The most influential features included:

- sum_click
- num_of_prev_attempts
- studied_credits
- age_band_35-55
- age_band_55<=
- disability_Y
- gender_M

Adding sum_click significantly improved the model's performance from 0.678 to 0.683. While demographic features like age bands, disability status, and gender provided incremental improvements, their interaction with behavioural features like click data and credits had the most substantial impact on performance.

Summary of Logistic Regression Results

The logistic regression model analyzed the likelihood of students passing their course based on demographic and behavioural features. The dependent variable was final_result_Pass (Binary: Pass vs. Not Pass), with 704 observations included in the analysis. The pseudo R-squared was 0.02496, indicating that a small percentage of the variability in the dependent variable is explained by the predictors.

Feature Coefficients and Interpretation

Feature	**Coefficient**	**P-value**	**Significance**	**Interpretation**
Intercept	1.0681	0.000	Significant	Baseline log-odds of passing when all predictors are 0.
age_band_35_55 (True)	-0.1818	0.312	Not significant	Age group 35-55 has no significant effect on the likelihood of passing.
age_band_55__ (True)	-0.8612	0.023	Significant	Students aged 55+ are less likely to pass.
disability_Y (True)	-0.3206	0.375	Not significant	Disability status does not significantly affect passing likelihood.
gender_M (True)	0.2409	0.165	Not significant	Gender (Male) has no significant effect on passing likelihood.
sum_click	0.0001	0.018	Significant	Higher engagement (clicks) slightly increases the likelihood of passing.
num_of_prev_attempts	-0.2636	0.463	Not significant	Number of previous attempts does not significantly affect passing likelihood.

Feature	Coefficient	P-value	Significance	Interpretation
studied_credits	-0.0056	0.003	Significant	Higher credits studied slightly decrease the likelihood of passing.

Fig. 10 - Table showing features along with coefficients and p-values

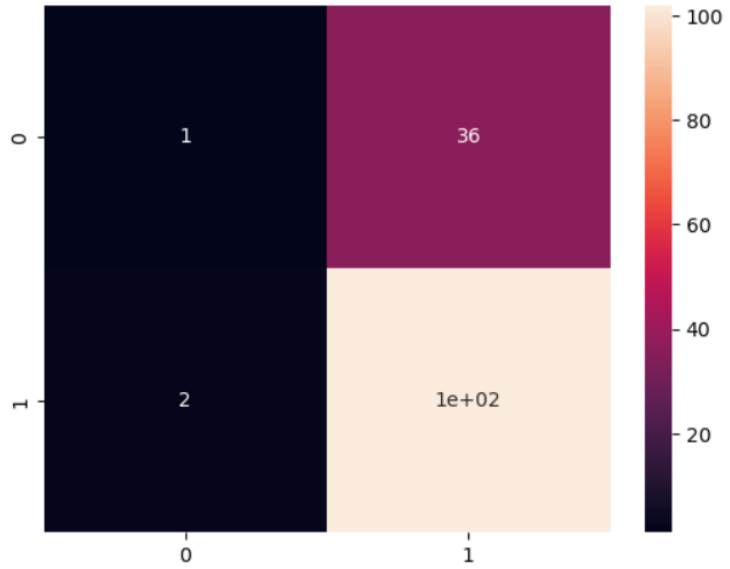


Fig. 11 Confusion Matrix for Students with module AAA 1 denotes Pass and 0 denotes not pass

Classification Performance Metrics

The classification model's performance metrics for predicting pass outcomes are as follows:

- **Accuracy**: 0.7305 (73.05%)
- **Precision**: 0.7391 (73.91%)
- **Recall**: 0.9808 (98.08%)
- **F1-Score**: 0.8430 (84.30%)

These results highlight that the model effectively identifies most true positive cases (high recall) while maintaining a reasonable balance between precision and recall (F1-score). The accuracy indicates that approximately 73% of predictions are correct, suggesting the model's practical utility in identifying students likely to pass.

Decision tree classifier Class-wise Performance and Overall Metrics

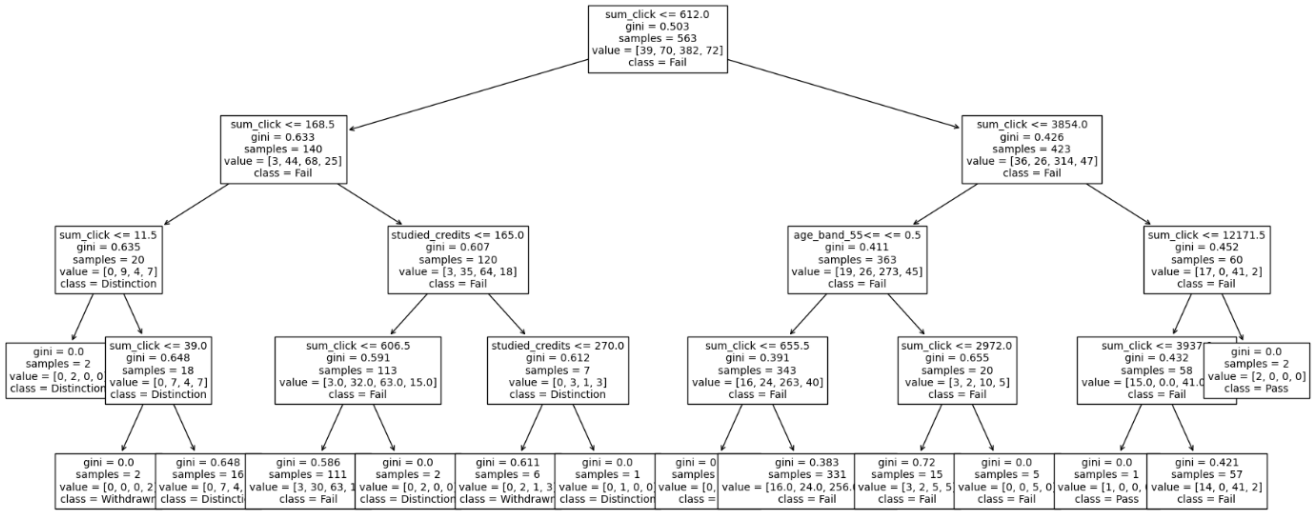


Fig. 12 Decision tree

The decision tree model's performance was evaluated using precision, recall, F1-score, and support for each class: "Distinction," "Fail," "Pass," and "Withdrawn." These metrics offer insights into how well the model performs for each category and overall.

Class-wise Performance using decision trees:

	precision	recall	f1-score	support
Distinction	1.00	0.20	0.33	5
Fail	0.36	0.36	0.36	11
Pass	0.78	0.94	0.86	104
Withdrawn	0.50	0.10	0.16	21
accuracy			0.74	141
macro avg	0.66	0.40	0.43	141
weighted avg	0.72	0.74	0.70	141

- **Distinction:** The model achieved perfect precision (1.00), meaning that all instances predicted as "Distinction" were correct. However, the recall was only 0.20, indicating that only 20% of actual "Distinction" cases were identified. The F1-score, which balances precision and recall, was 0.33, highlighting under-detection.
- **Fail:** For the "Fail" class, both precision and recall were moderate at 0.36. This balance resulted in an F1-score of 0.36, suggesting consistent but limited performance for this class.
- **Pass:** As the majority class, "Pass" exhibited the best performance, with a precision of 0.78, a recall of 0.94, and an F1-score of 0.86. The model effectively identified most "Pass" cases while maintaining high precision.

- **Withdrawn:** The model struggled with the "Withdrawn" class, achieving a precision of 0.50 and a very low recall of 0.10. This led to an F1-score of only 0.16, reflecting significant under-detection.

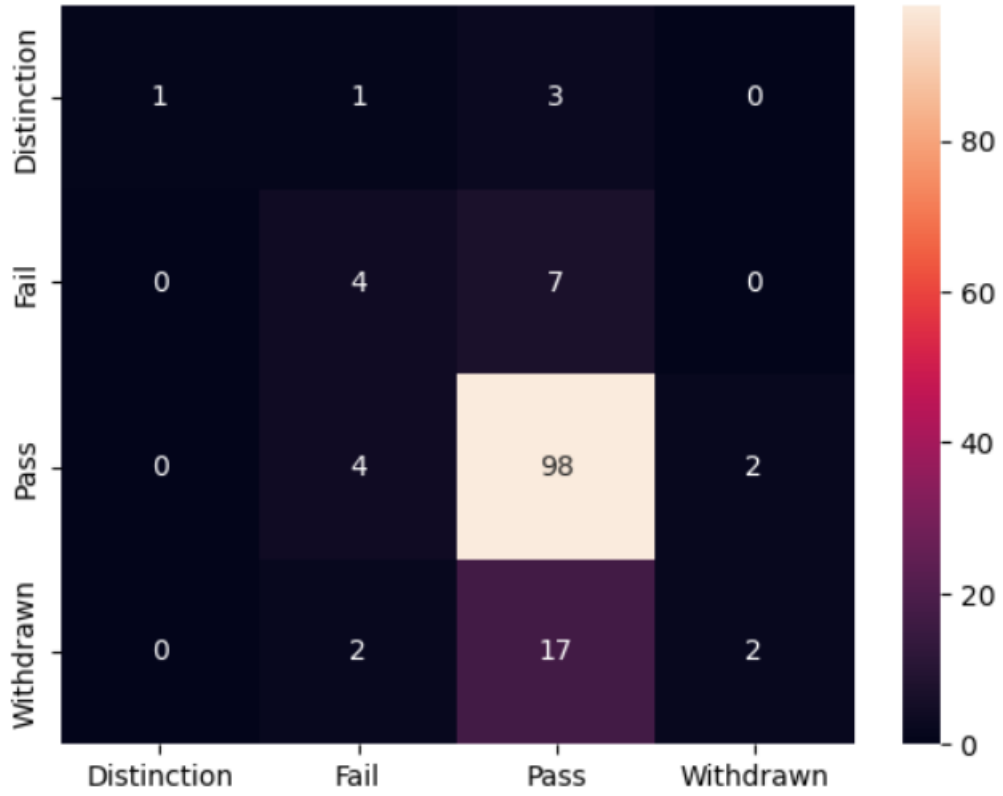


Fig. 13 – Confusion matrix for decision tree

Overall Performance

The overall accuracy of the decision tree model was 74%, indicating that it correctly classified 74% of the instances in the dataset. However, performance varied significantly across classes, with the majority class ("Pass") driving the accuracy.

Discussions and Observations

1. Hypothesis Testing :

There is a moderate positive linear relationship between the variables.This relationship is statistically significant, meaning it is unlikely to have occurred by random chance.

2. Correlation check observations:

Significant positive correlation between sum_click and aggregate_weighted_score. On fitting an OLS the model had good R^2 score as well as pearsonr test giving a low p-value signifying positive correlation among

3. Results from Recursive Feature Elimination:

Behavioural features such as sum_click and academic effort (e.g., num_of_prev_attempts, studied_credits) are critical predictors for both tasks. Demographic features, including age bands, disability status, and gender, had limited standalone impact but provided minor improvements when combined with behavioral metrics. The consistency of feature selection across tasks indicates the importance of these features in predicting both categorical outcomes (pass/fail) and continuous scores. Logistic regression demonstrated superior performance compared to linear regression, suggesting that predicting categorical outcomes is more effective with this dataset.

4. Observations from applying logistic regression on the AAA module data:

Significant predictors identified by the logistic regression model include sum_click, age_band_55__, and studied_credits. Sum_click exhibited a small but positive effect on passing likelihood, where each additional click slightly increased the odds of passing. Conversely, students aged 55 and older were significantly less likely to pass, highlighting potential challenges for this demographic. An increase in studied_credits was associated with a slight decrease in passing likelihood, potentially indicating a workload issue.

The model's precision for predicting non-pass outcomes is relatively poor due to class imbalance, as the dataset contains significantly more pass cases. To address this, alternative modules with less pronounced class imbalances should be analysed. These adjustments may help improve the model's precision for predicting non-pass outcomes while maintaining overall performance.

5. Observations from Decision tree classification:

The evaluation highlights key observations regarding the model's strengths and weaknesses. A significant class imbalance, with the majority class ("Pass") dominating the dataset, led to skewed metrics. While the model performed well for the "Pass" class, it struggled with minority classes such as "Fail," "Pass," and "Withdrawn," as evidenced by their low recall and F1-scores. These results emphasize the need for strategies to address class imbalance, such as oversampling minority classes or using ensemble methods to improve performance across under-represented categories or using SMOTE.

Conclusion

This analysis explored the relationships between student engagement, demographic characteristics, and academic outcomes using the Open University Learning Analytics Dataset. Logistic regression revealed that behavioural metrics such as sum_click and aggregated_weighted_score are strong predictors of student success, while demographic features like age and gender have limited standalone impact. The decision tree classifier, while accurate overall, struggled with class imbalance, particularly in predicting minority classes such as "Fail" and "Withdrawn."

Key findings include the significance of VLE interaction metrics and aggregated scores in predicting outcomes, as well as the challenges faced by older students and those with higher credit loads.

Addressing class imbalance and exploring alternative models with better handling of minority classes are recommended to enhance prediction accuracy and fairness.

Additionally, it was noted that the weights in the GGG module are all zero, making it impossible to compute weighted scores. To proceed, the mean of all scores by each student within this module will be used as a substitute for the aggregated score. This adjustment ensures consistency and allows the analysis to include the GGG module in meaningful ways.