Llewy Smith
a1706836

Assessment 1B: Big Data Analysis

Table of Contents

Part 1: Data description (5 marks)

Research questions

Primary

Within the first 25% of teaching time, how accurately and transparently can we predict which enrolments will result in a failed course outcome?

Sub-questions

- **Census-date early-warning accuracy:** What ROC-AUC, F1, precision and recall does the model achieve when trained and evaluated on records available up to the census cutoff (~25% of course progress) on the OULAD dataset?
- **Post-census performance gain:** How much does predictive accuracy improve when the training window is extended to 50% progress, and is the gain large enough to justify waiting beyond Census date?
- **Explainable-AI insights for intervention:** Using SHAP and Permutation Feature Importance, which behavioural, demographic and historic-grade features drive the risk scores prior to the census date and at 50%, and how can these insights guide targeted interventions?
- **Cross-dataset consistency:** When the same pipeline is applied separately to the SED and UCI student-performance datasets, how do accuracy metrics and key explanatory features compare with those from OULAD?

Data tables

Table Name	Records	Description
courses	22	Metadata including code, schedule, and length
assessments	2,913	Details such as type, weight, and due date
vle	6,389	Virtual learning environment (VLE) activities
studentInfo	32,504	Demographic and outcome data for each student-course enrolment
studentRegistration	32,504	Dates of enrolment and unenrolment
studentAssessment	26,699	Submissions, scores, and dates
studentVLE	10.6 million	VLE interactions by student, activity, and date

Key features:

The reason behind using the OULAD datasets above is that they provide data in multiple useful categories and formats. I group features into these four categories:

(1) learning analytics (VLE engagement and assessment)
(2) demographic
(3) administrative
(4) course metadata

Learning analytics (VLE engagement and assessment)

This data is sourced from studentVLE.csv, studentAssessment.csv, vle.csv and assessments.csv. Examples are:

- **sum_click** - the number of times a student interacts with the material in that day
- **date_submitted** - the date of assessment submission, measured as the number of days since the start of the course

This data forms the time-series basis for my CNN-LSTM model. I also construct aggregates and threshold based dummy variables using these so that the CNN part can learn from these. Examples are sum_click (clicks over 25% of the course), early_weighted_avg_score (of assessments over 25% of the course), and an early_failed_assessment (flag for 1 or more failed assessments).

Demographic features

This data is sourced from studentInfo.csv. Examples are:

- **highest_education** - highest student education level on entry to course
- **imd_band** - the Index of Multiple Depravation band of student's home address

These features are static - hence having no time-series element. They can be Boolean (e.g. disability), ordinal categorical (e.g. imd_band) or nominal categorical (e.g. region).

Administrative features

This data is sourced from studentInfo.csv and studentRegistration.csv. Examples include:

- **num_of_prev_attempts** - count of student's previous attempts at this course
- **date_registration** - date of student's course enrolment relative to start date
- when they enrolled, relative to the module start

These features help identify academic overload (e.g., high credit load), prior struggle (e.g., repeat enrolments), and possible delayed start (e.g., late enrolment).

Course metadata

This data is sourced from courses.csv and assessments.csv. It gives fixed characteristics about each course. Examples include:

- **code_module** - course name (e.g., GGG, DDD)
- **code_presentation** - semester of delivery (e.g., 2013B or 2014J)
- **module_presentation_length** - total length of the course in days

These features are used to calculate the 25% cutoff point of each course, and to stratify performance at the course level. They also unlock nuance at the course level - some courses may set more assessment early on, while others may rely more on VLE clicks.

Part 2: Clustering/Pattern (6 marks)

I explored using Pandas crosstabs to detect early patterns, and later used a simple Random Forest classifier using data in the first 25% of each course to see baseline results. I assessed this using AUC and ranked the top 10 most important features.

Patterns

Learning analytics risk - assessments

Among students flagged with early_failed_assessment = 1, 67% failed the course. In contrast, those who passed all their early assessments had a 39% fail rate. This suggests that even a single failed early assessment is a strong indicator of course risk.

Learning analytics risk - VLE engagement

Students were binned into four groups based on total VLE clicks (sum_click). Pass rates increased monotonically from 38% in the lowest quartile (Q1) to 77% in the highest (Q4), confirming the predictive utility of early engagement metrics.

Demographic based risk – imd_band

I examined pass rates across imd_band, which categorises students' home addresses by socioeconomic disadvantage (similar to SES in Australia). Students in the most deprived category (0-10%) had a pass rate of 46%, whereas those in the least deprived category (90-100%) passed at 68%. This supports imd_band as a valuable equity-based early predictor of academic risk. Interestingly, students with an unknown imd_band passed even more frequently, at 73%. This leads me to believe that this group is special, such as international students or accelerated high school students, and that I need to look deeper.

Administrative based risk - repeat course attempts

Students taking a course for the first time passed 61% of the time. For those with two or more prior attempts, pass rates fell below 40%. These findings support the inclusion of num_of_prev_attempts and other administrative history features.

Course metadata-based risk

Fail rates varied across courses (code_module). For instance, DDD and CCC had the highest fail rates (48% and 47%), while AAA students failed only 24% of the time.

I also found variation across semesters (code_presentation). Students enrolled in 2013B and 2014B had higher fail rates (47% and 46%) compared to those in 2013J and 2014J (39% and 38%). I did not investigate, but it would be interesting to further look at the causes of this, as it could be that more difficult courses (e.g. DDD or CCC) are offered in one half of the year, causing this large differential.

Predictive modelling – classification

To validate these patterns quantitatively, I trained a Random Forest classifier using features available within the first 25% of each course. Performance was evaluated on a 20% hold-out set (n = 5,251). Metrics below are reported with respect to the fail class (i.e., predicting a course failure):

- Accuracy: 75%
- Precision: 74%
- Recall: 61%
- F1: 67%
- ROC-AUC: 0.808

Feature importance

The top 5 predictors based on Gini importance scores were:

Feature	**Importance**
early_weighted_avg_score	0.185
sum_click	0.137
days_clicked	0.136
date_registration	0.099
imd_band	0.066

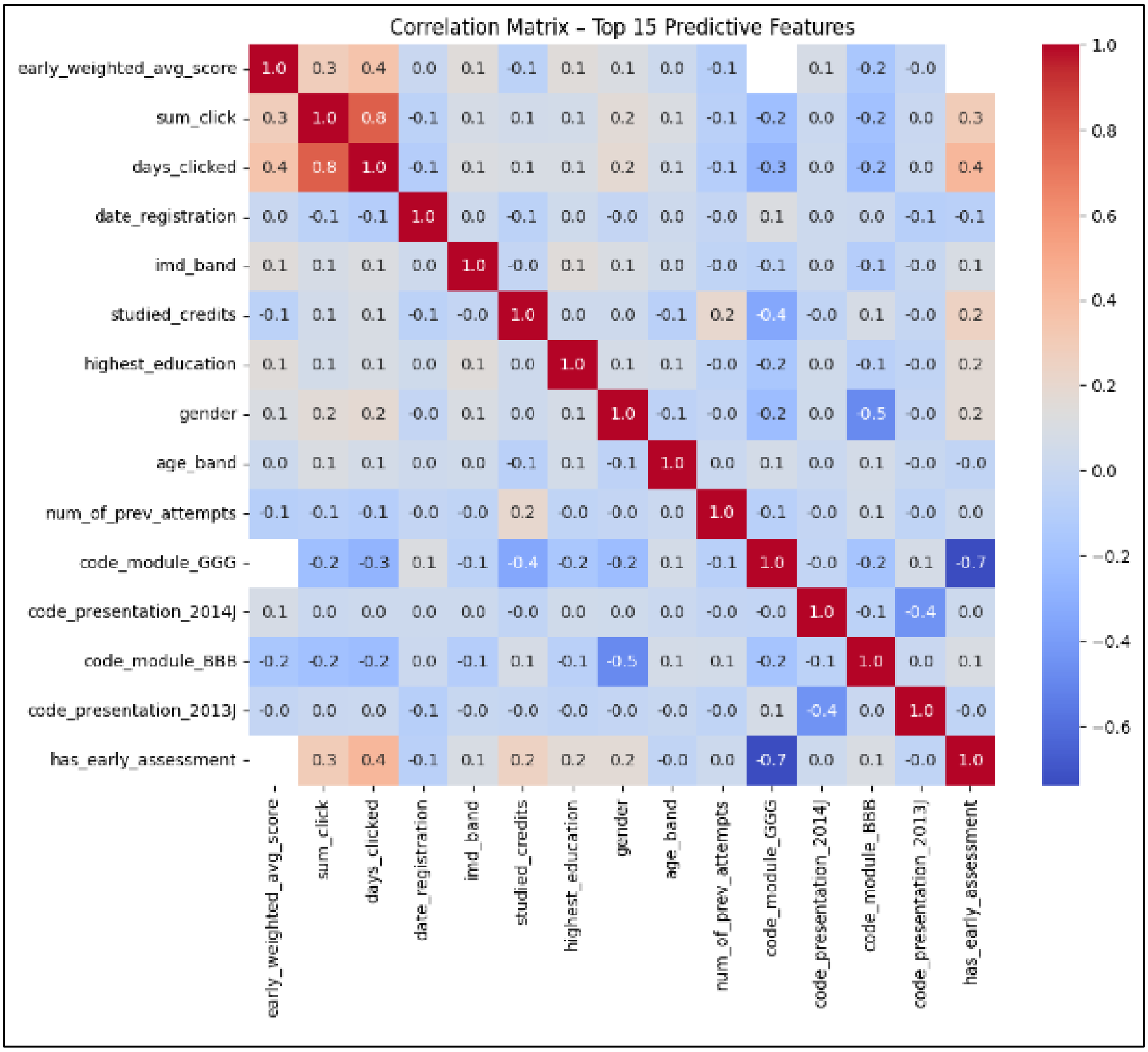
This aligns with the pattern findings - early assessment, VLE engagement, and timing of enrolment were some of the strongest early signals of risk. Socioeconomic indicators (imd_band) also contributed meaningfully but to a lesser extent.

Part 3: Visualisation (5 marks)

Able to provide at least 3-4 clear and correct visualisations of the dataset and correctly describes the visualisation.

Correlation of top features

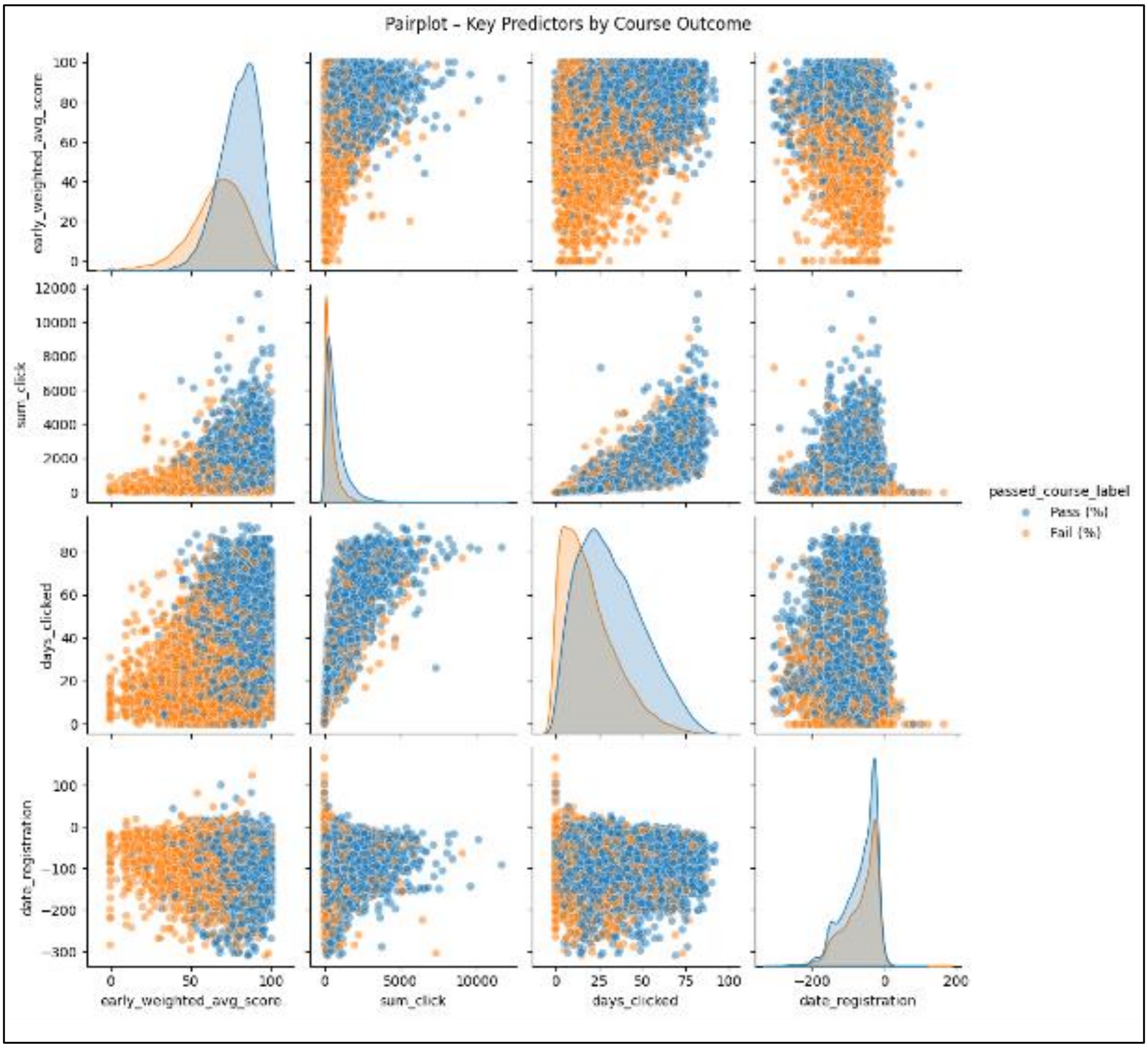
The goal of this heatmap is to detect multicollinearity among top predictors, and potentially justify dimensionality reduction (e.g., via Recursive Feature Elimination).



A correlation heatmap of the top 15 predictive features showed that sum_click and days_clicked (r = 0.80) may be redundant in combination, supporting the future use of Recursive Feature Elimination (RFE) or similar. Additionally, a strong inverse correlation between has_early_assessment and code_module_GGG (r = –0.70) suggests that this course may not have any assessment in the first 25% of the course, which is very useful to know and renders certain features useless for predictions in this course.

Distribution and patterns between key predictors

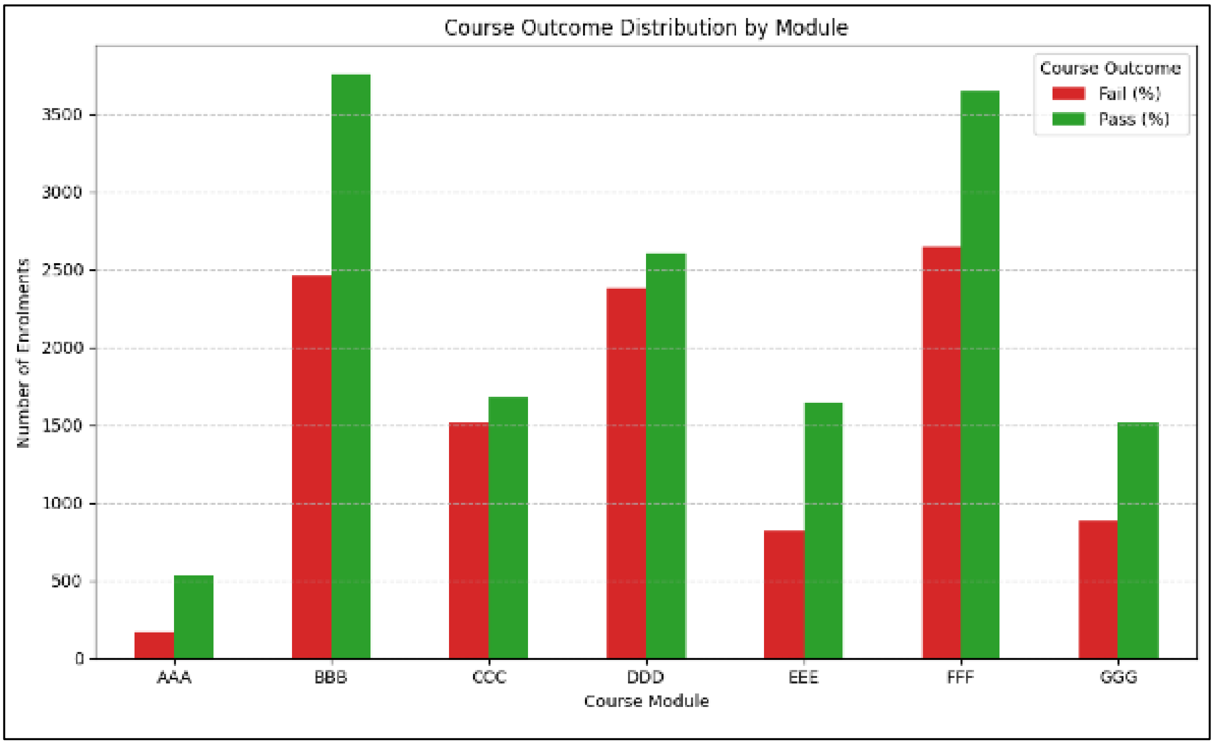
The goal of this pairplot is to explore how key early predictors of course failure relate to each other and to the target variable (pass/fail), helping to understand class separation and the potential decision boundaries of classifiers.



This pairplot of four top-ranked predictors - being early_weighted_avg_score, sum_click, days_clicked, and date_registration - shows visual separation between pass and fail classes. One example is the peak and skew of the Fail class in the 'early_weighted_avg_score' plot being further left than that of the Pass class.

Class imbalance

This visualisation aims to show the class imbalance of the target variable (course outcome), both overall and across different courses, to explore the need to use SMOTE (Synthetic Minority Over-sampling Technique) to balance the training data during model development as is done by Adefemi & Mutanga (2025) and other researchers. I split it by course, due to the possibility that my future model runs on course specific time-series data.



The plot reveals greater class imbalance in some modules than others, for example course ‘DDD’ likely does not require SMOTE, however ‘AAA’ and ‘EEE’ would likely benefit. Without balancing, my models could underfit the fail class, particularly in the more imbalanced courses.

Part 4: Problem refinement (4 marks)

I am satisfied with my progress toward answering my primary question: “Within the first 25% of teaching time, how accurately and transparently can we predict which enrolments will result in a failed course outcome?” However, I need to make more progress towards the sub-questions.

Census-date early-warning performance (Sub question 1): While my metrics (e.g., ROC-AUC = 0.81) are promising, there is always room for optimisation - particularly by refining feature selection (e.g., using RFE to address the multicollinearity I have between top variables) and addressing class imbalance with SMOTE, which affects precision and recall for fail cases.

Post-census performance gain (Sub question 2): To answer this question, I plan to expand to include data up to the 50% course progress point. Comparing model performance across these two windows will help evaluate whether delaying intervention yields significantly better predictions, and will come in Part C or D.

Explainable-AI (Sub question 3): My initial feature importance analysis highlighted early assessments, VLE clicks, and enrolment timing as top predictors. However, I need to undertake further SHAP and Permutation Feature Importance analyses to unpack individual-level explanations and guide targeted support strategies more transparently, in order for this whole project to have impact.

Cross-dataset consistency (Sub question 4): This remains a work-in-progress. I have not yet applied the pipeline to the SED or UCI datasets, but I aim to do so in Part C and D. This step is essential to test generalisability across contexts, and again, in order for this whole project to have impact.

Part 5: References

References are progressive over Part A and B of this report.

Datasets

Open University Learning Analytics Dataset (OULAD)

Kuzilek, J, Hlosta, M & Zdrahal, Z 2017, 'Open University Learning Analytics dataset', *Scientific Data*, vol. 4, article 170171, doi:10.1038/sdata.2017.171, viewed 22 June 2025, https://analyse.kmi.open.ac.uk/open-dataset\

Universiti Malaya Student Engagement Dataset (SED)

Kassim, MSS, Azizul, ZH & Fuaad, AAH 2025, 'Student Engagement Dataset (SED): An online learning activity dataset', *IEEE Access*, early access, viewed 22 June 2025, https://ieeexplore.ieee.org/document/10844083\

University of California-Irvine Dataset (UCI)

Cortez, P & Silva, A 2008, 'Using data mining to predict secondary school student performance', *Proceedings of the European Simulation and Modelling Conference (EUROSIS)*, EUROSIS, Ghent, pp. 117–122, viewed 22 June 2025, https://www.researchgate.net/publication/228780408_Using_data_mining_to_predict_secondary_school_student_performance/

Other References

Huang, Q & Chen, J 2024, ‘Enhancing academic performance prediction with temporal graph networks for massive open online courses’, *Journal of Big Data*, vol. 11, no. 52, doi:10.1186/s40537-024-00918-5.

Alnasyan, B, Basheri, M, Alassafi, M *et al*. 2025, ‘Kanformer: an attention-enhanced deep learning model for predicting student performance in virtual learning environments’, *Social Network Analysis and Mining*, vol. 15, article 25, doi:10.1007/s13278-025-01446-7.

Khoudi, Z, Hafidi, N, Nachaoui, M *et al*. 2025, ‘New approach to enhancing student performance prediction using machine learning techniques and clickstream data in virtual learning environments’, *SN Computer Science*, vol. 6, article 139, doi:10.1007/s42979-024-03622-6.

Adnan, M, Ahmad, NH, Aziz, ME *et al*. 2021, ‘Predicting at-risk students at different percentages of course length for early intervention using machine learning models’, *IEEE Access*, vol. 9, pp. 7519–7539, doi:10.1109/ACCESS.2021.3050103.

Jihaoui, ME, El Kheir Abra, O & Mansouri, K 2025, ‘A deep learning regression model for predicting students’ performance with Shapley additive explanations-based interpretability’, in *Proceedings of the 5th International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, Fez, Morocco, 1–8, viewed 22 June 2025, https://ieeexplore.ieee.org/document/11008069/.

Mohammad, AS, Al-Kaltakchi, MTS, Alshehabi Al-Ani, J & Chambers, JA 2023, ‘Comprehensive evaluations of student performance estimation via machine learning’, *Mathematics*, vol. 11, no. 14, article 3153, doi:10.3390/math11143153.

Adefemi, KO & Mutanga, MB 2025, ‘A robust hybrid CNN–LSTM model for predicting student academic performance’, *Digital*, vol. 5, no. 2, article 16, doi:10.3390/digital5020016.

Australian Government Department of Education 2023, *Higher Education Provider Guidelines 2023*, Federal Register of Legislation, Canberra, viewed 22 June 2025, https://www.legislation.gov.au/F2023L00400. legislation.gov.au

University of Adelaide 2023, *Support for Students Policy*, University of Adelaide, Adelaide, viewed 22 June 2025, https://www.adelaide.edu.au/policies/5003/

Department of Education (Australian Government) 2024, Success rate – Glossary, Tertiary Collection of Student Information (TCSI), viewed 22 June 2025, https://www.tcsisupport.gov.au/glossary/glossaryterm/Success%20rate.

Department of Education (Australian Government) 2024, Retention rate – Glossary, Tertiary Collection of Student Information (TCSI), viewed 22 June 2025, https://www.tcsisupport.gov.au/glossary/glossaryterm/retention-rate.

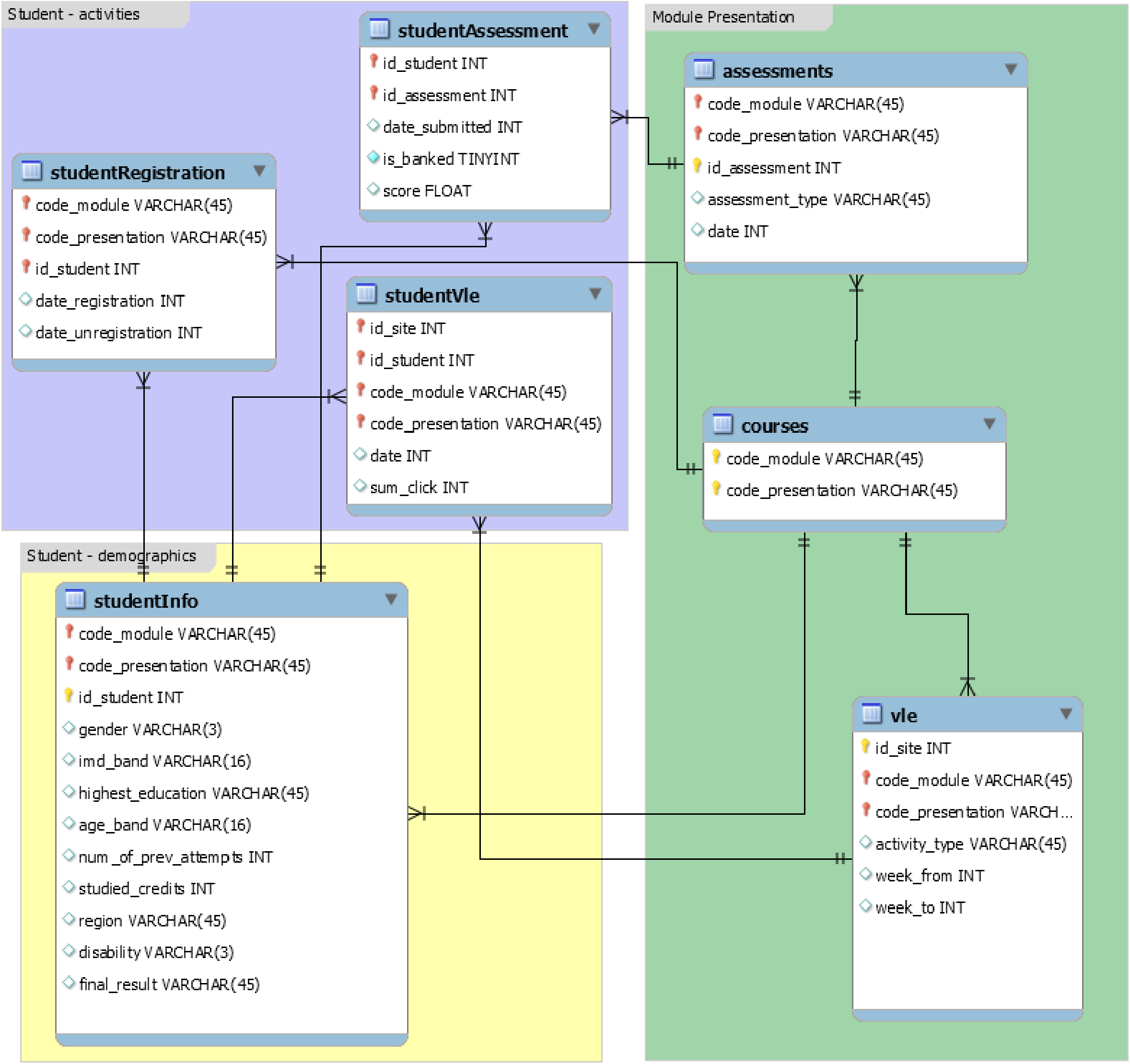
Appendix 1: Abbreviations

- **EDM** - Educational data mining
- **OULAD** - Open University Learning Analytics Dataset
- **UCI** - University of California-Irvine
- **SED** - Student Engagement Dataset (SED) - Universiti Malaya
- **VLE** - Virtual Learning Environment
- **LMS** - Learning Management System
- **CNN** - Convolutional Neural Network
- **LSTM** - Long Short-Term Memory
- **PFI** - Permutation Feature Importance
- **SHAP** - Shapley Additive Explanations
- **XAI** - Explainable Artificial Intelligence
- **APP-TGN** - Attention-based Personalised Propagation - Temporal Graph Network
- **GRU** - Gated Recurrent Unit

Appendix 2: Definitions

Success Rate – The Success rate for year(x) is the proportion of actual student load (EFTSL) for units of study that are passed divided by all units of study attempted (passed + failed + withdrawn) (Department of Education 2024).
Retention rate - Retention rate is a measure of the proportion of students who continue their studies after their first year. A more detailed definition is given here (Department of Education 2024).

Appendix 3: Dataset Schema



courses.csv

File contains the list of all available modules and their presentations. The columns are:

- code_module - code name of the module, which serves as the identifier.
- code_presentation - code name of the presentation. It consists of the year and "B" for the presentation starting in February and "J" for the presentation starting in October.
- module_presentation_length - length of the module-presentation in days.

The structure of B and J presentations may differ and therefore it is good practice to analyse the B and J presentations separately. Nevertheless, for some presentations the corresponding previous B/J presentation do not exist and therefore the J presentation must be used to inform the B presentation or vice versa. In the dataset this is the case of CCC, EEE and GGG modules.

assessments.csv

This file contains information about assessments in module-presentations. Usually, every presentation has a number of assessments followed by the final exam. CSV contains columns:

- code_module - identification code of the module, to which the assessment belongs.
- code_presentation - identification code of the presentation, to which the assessment belongs.
- id_assessment - identification number of the assessment.
- assessment_type - type of assessment. Three types of assessments exist: Tutor Marked Assessment (TMA), Computer Marked Assessment (CMA) and Final Exam (Exam).
- date - information about the final submission date of the assessment calculated as the number of days since the start of the module-presentation. The starting date of the presentation has number 0 (zero).
- weight - weight of the assessment in %. Typically, Exams are treated separately and have the weight 100%; the sum of all other assessments is 100%.

If the information about the final exam date is missing, it is at the end of the last presentation week.

vle.csv

The csv file contains information about the available materials in the VLE. Typically, these are html pages, pdf files, etc. Students have access to these materials online and their interactions with the materials are recorded.
The vle.csv file contains the following columns:

- id_site - an identification number of the material.
- code_module - an identification code for module.
- code_presentation - the identification code of presentation.
- activity_type - the role associated with the module material.
- week_from - the week from which the material is planned to be used.
- week_to - week until which the material is planned to be used.

studentInfo.csv

This file contains demographic information about the students together with their results. File contains the following columns:

- code_module - an identification code for a module on which the student is registered.
- code_presentation - the identification code of the presentation during which the student is registered on the module.
- id_student - a unique identification number for the student.
- gender - the student's gender.

- region - identifies the geographic region, where the student lived while taking the module-presentation.
- highest_education - highest student education level on entry to the module presentation.
- imd_band - specifies the Index of Multiple Depravation band of the place where the student lived during the module-presentation.
- age_band - band of the student's age.
- num_of_prev_attempts - the number times the student has attempted this module.
- studied_credits - the total number of credits for the modules the student is currently studying.
- disability - indicates whether the student has declared a disability.
- final_result - student's final result in the module-presentation.

studentRegistration.csv

This file contains information about the time when the student registered for the module presentation. For students who unregistered the unregistered date is also recorded. File contains five columns:

- code_module - an identification code for a module.
- code_presentation - the identification code of the presentation.
- id_student - a unique identification number for the student.
- date_registration - the date of student's registration on the module presentation, this is the number of days measured relative to the start of the module-presentation (e.g. the negative value -30 means that the student registered to module presentation 30 days before it started).
- date_unregistration - the student's unregistered date from the module presentation, this is the number of days measured relative to the start of the module-presentation. Students, who completed the course have this field empty. Students who unregistered have Withdrawal as the value of the final_result column in the studentInfo.csv file.

studentAssessment.csv

This file contains the results of students' assessments. If the student does not submit the assessment, no result is recorded. The final exam submissions is missing, if the result of the assessments is not stored in the system.
This file contains the following columns:

- id_assessment - the identification number of the assessment.
- id_student - a unique identification number for the student.
- date_submitted - the date of student submission, measured as the number of days since the start of the module presentation.
- is_banked - a status flag indicating that the assessment result has been transferred from a previous presentation.

- score - the student's score in this assessment. The range is from 0 to 100. The score lower than 40 is interpreted as Fail. The marks are in the range from 0 to 100.

studentVle.csv

The studentVle.csv file contains information about each student's interactions with the materials in the VLE.
This file contains the following columns:

- code_module - an identification code for a module.
- code_presentation - the identification code of the module presentation.
- id_student - a unique identification number for the student.
- id_site - an identification number for the VLE material.
- date - the date of student's interaction with the material measured as the number of days since the start of the module-presentation.
- sum_click - the number of times a student interacts with the material in that day.