The Ministry of Science and Higher Education of the Republic of Kazakhstan
SDU University



Damir Syzdykov

Student Performance Prediction Using Machine Learning

THESIS

Presented in Partial Fulfilment for the

Degree of Master of Science in Information Systems

(degree code: 7M06101)

Department of Information Systems
School of Information Technologies and Applied Mathematics

Supervisor: **Selcuk Cankurt**

Kaskelen, 2026

School of Information Technologies and Applied Mathematics

Department of Information Systems

Dissertation Signatures Title Page

The Master's Dissertation of **Damir Syzdykov**

Title:
Student performance prediction using machine learning

Машиналық оқытуды қолдану арқылы оқушылардың үлгерімін болжау

Прогнозирование успеваемости учащихся с помощью машинного обучения

Is submitted in partial fulfillment of the requirements for the Master's (MSc) degree in the Department of Information Systems, and approved by the Dissertation Committee:

Head of Committee:	Assistant Prof., PhD, Meraryslan Meraliyev
Academic supervisor:	Assistant Prof., PhD, Selcuk Cankurt
Member's Signature:	Assistant Prof., PhD, Orynbekova K.E.
Member's Signature:	Assistant Prof., PhD, Ayazbayev D.A.

Kaskelen, 2026

Declaration

I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Damir Syzdykov

Kaskelen, Kazakhstan
May 2026

Acknowledgements

The completion of this dissertation would not have been possible without the patience, guidance and continuous encouragement of many individuals to whom I would like to express my sincere gratitude.

I would like to express my sincere gratitude to my supervisor, Selcuk Cankurt, for his guidance, support and valuable feedback throughout the preparation of this dissertation.

Damir Syzdykov
Kaskelen, May 2026

Dedication

This thesis is dedicated to

My parents, Nur Amantay, Sultan Telman and many other for their support, help, sense of humour and useful comments for improving this project.

Abstract

The early identification of academically at-risk students remains one of the most pressing problems in contemporary higher education. As universities accumulate large amounts of administrative, demographic and behavioural data through their Virtual Learning Environments, machine learning has emerged as a powerful instrument for transforming raw educational data into actionable early-warning signals. This dissertation develops, evaluates and explains a comprehensive machine learning pipeline for predicting student academic performance on the Open University Learning Analytics Dataset (OULAD), a publicly available benchmark containing detailed records for 32,593 learners across 22 module-presentations.

The work is framed around four research questions concerning the most predictive student features, the most suitable learning algorithm for the task, the relative contribution of individual pipeline components and the practical interpretability of the resulting models. To answer these questions, the dissertation proposes an end-to-end methodology that combines multi-source data integration, domain-driven feature engineering, hybrid feature selection, synthetic minority over-sampling, Bayesian hyperparameter optimisation using Optuna and post-hoc model interpretation through SHapley Additive exPlanations (SHAP). Four representative classifiers are compared on the same task: a regularised Logistic Regression baseline, Random Forest, XGBoost and LightGBM.

Experimental results demonstrate that gradient-boosted tree ensembles, and in particular an Optuna-tuned XGBoost model, consistently outperform the alternative baselines on the held-out test set, reaching a test area under the receiver operating characteristic curve of 0.97 and a test F1 score of 0.90 with a moderate overfitting gap. Engagement-related features—including the number of active days on the Virtual Learning Environment, total clicks and an engineered `score_consistency` indicator—are identified as the most influential predictors by both global feature-importance methods and SHAP analysis. An ablation study reveals that feature engineering, class-imbalance handling and hyperparameter tuning each contribute incrementally to the final performance and that their combination yields the best balance of accuracy, generalisation and computational efficiency.

The dissertation also discusses educational, ethical and deployment-related implications of the proposed pipeline. SHAP-based explanations are shown to produce individualised, actionable insights that can be conveyed to educators and policymakers without requiring a deep understanding of the underlying algorithms. Limitations associated with the use of a single benchmark dataset, with binary outcome modelling and with the absence of causal claims are explicitly acknowledged, and directions for future research are outlined, including temporal modelling, fairness-aware learning and integration with existing learning analytics dashboards. Overall, the dissertation contributes a reproducible methodological template that bridges predictive accuracy and human-interpretable insight in the context of student performance prediction.

Keywords: student performance prediction, educational data mining, learning analytics, machine learning, XGBoost, SHAP, feature engineering, OULAD, explainable artificial intelligence, early-warning systems.

Аңдатпа

Академиялық тәуекел тобындағы студенттерді ерте анықтау қазіргі жоғары білім беру саласындағы өзекті мәселелердің бірі болып табылады. Университеттер виртуалды оқыту орталары арқылы әкімшілік, демографиялық және мінез-құлықтық деректердің үлкен көлемін жинақтаған сайын, машиналық оқыту білім беру деректерін әрекет етуге болатын ерте ескерту сигналдарына айналдырудың тиімді құралына айналды. Бұл диссертация Open University Learning Analytics Dataset (OULAD) деп аталатын ашық эталондық деректер жиынында студенттердің академиялық үлгерімін болжауға арналған кешенді машиналық оқыту құбырын әзірлейді, бағалайды және түсіндіреді. Аталған деректер жиыны 22 модуль-презентация бойынша 32,593 білім алушы туралы егжей-тегжейлі жазбаларды қамтиды.

Зерттеу ең болжамды студенттік белгілерді анықтау, тапсырмаға ең қолайлы оқыту алгоритмін таңдау, құбырдың жекелеген компоненттерінің салыстырмалы үлесін бағалау және алынған модельдердің практикалық түсіндірмелілігін қарастыруға бағытталған төрт зерттеу сұрағы аясында құрылды. Осы сұрақтарға жауап беру үшін диссертация көп көзді деректерді біріктіруді, пәндік салаға негізделген белгілерді құрастыруды, гибридті белгі таңдауды, синтетикалық азшылық класын артық іріктеуді, Optuna көмегімен Байестік гиперпараметрлерді оңтайландыруды және SHapley Additive exPlanations (SHAP) арқылы модельдерді кейінгі түсіндіруді біріктіретін толық әдістемені ұсынады. Бір тапсырмада төрт өкілдік классификатор салыстырылды: регуляризацияланған Logistic Regression базалық моделі, Random Forest, XGBoost және LightGBM.

Эксперименттік нәтижелер градиенттік бустингке негізделген ағаш ансамбльдері, әсіресе Optuna арқылы бапталған XGBoost моделі, ұсталған тест жиынында баламалы базалық модельдерден тұрақты түрде жоғары нәтиже көрсететінін дәлелдейді. Бұл модель қабылдағыштың операциялық сипаттамасы қисығы астындағы аудан бойынша 0.97 тест нәтижесіне және 0.90 тест F1 көрсеткішіне қол жеткізді, ал артық үйрену алшақтығы орташа деңгейде қалды. Виртуалды оқыту ортасындағы белсенді күндер саны, жалпы басулар саны және құрастырылған `score_consistency` көрсеткіші сияқты қатысуға байланысты белгілер жаһандық белгі маңыздылығы әдістері мен SHAP талдауы бойынша ең ықпалды болжаушылар ретінде анықталды. Абляциялық зерттеу белгі құрастыру, класс теңгерімсіздігін өңдеу және гиперпараметрлерді баптау соңғы нәтижеге жеке-жеке қосымша үлес қосатынын, ал олардың үйлесімі дәлдік, жалпылау қабілеті және есептеу тиімділігі тұрғысынан ең жақсы теңгерімді қамтамасыз ететінін көрсетті.

Диссертацияда ұсынылған құбырдың білім беру, этикалық және енгізу аспектілеріне қатысты салдары да талқыланады. SHAP негізіндегі түсіндірмелер жеке студенттерге қатысты әрекет етуге болатын түсініктер беріп, оларды негізгі алгоритмдерді терең білуді талап етпей-ақ оқытушылар мен саясат жасаушыларға жеткізуге болатыны көрсетілді. Бір ғана эталондық деректер жиынын қолдану, екілік нәтиже моделін құру және себеп-салдарлық тұжырымдардың болмауы сияқты шектеулер нақты көрсетілді. Болашақ зерттеулер үшін уақыттық модельдеу, әділеттілікке бағытталған оқыту және қолданыстағы learning analytics бақылау тақталарымен біріктіру бағыттары ұсынылды. Жалпы алғанда, бұл диссертация студент үлгерімін болжау контекстінде болжамдық дәлдік пен адамға түсінікті интерпретацияны байланыстыратын қайталанатын әдістемелік үлгі ұсынады.

Түйін сөздер: студент үлгерімін болжау, білім беру деректерін интеллектуалды талдау, learning analytics, машиналық оқыту, XGBoost, SHAP, белгі құрастыру, OULAD, түсіндірілетін жасанды интеллект, ерте ескерту жүйелері.

Аннотация

Раннее выявление студентов, находящихся в группе академического риска, остаётся одной из наиболее актуальных задач современного высшего образования. По мере того как университеты накапливают большие объёмы административных, демографических и поведенческих данных через виртуальные образовательные среды, машинное обучение становится эффективным инструментом преобразования исходных образовательных данных в практически применимые сигналы раннего предупреждения. В данной диссертации разрабатывается, оценивается и интерпретируется комплексный конвейер машинного обучения для прогнозирования академической успеваемости студентов на основе открытого эталонного набора данных Open University Learning Analytics Dataset (OULAD), содержащего подробные записи о 32,593 обучающихся в рамках 22 модульных представлений.

Исследование построено вокруг четырёх исследовательских вопросов, связанных с определением наиболее предиктивных характеристик студентов, выбором наиболее подходящего алгоритма обучения, оценкой относительного вклада отдельных компонентов конвейера и практической интерпретируемостью полученных моделей. Для ответа на эти вопросы в диссертации предлагается сквозная методология, объединяющая интеграцию данных из нескольких источников, предметно-ориентированное конструирование признаков, гибридный отбор признаков, синтетическое увеличение выборки миноритарного класса, байесовскую оптимизацию гиперпараметров с использованием Optuna и последующую интерпретацию моделей с помощью SHapley Additive exPlanations (SHAP). В рамках одной задачи сравниваются четыре репрезентативных классификатора: регуляризованная базовая модель Logistic Regression, Random Forest, XGBoost и LightGBM.

Экспериментальные результаты показывают, что ансамбли деревьев на основе градиентного бустинга, особенно модель XGBoost, настроенная с помощью Optuna, стабильно превосходят альтернативные базовые модели на отложенной тестовой выборке. Данная модель достигает значения площади под ROC-кривой 0.97 и тестового F1-показателя 0.90 при умеренном разрыве переобучения. Признаки, связанные с вовлечённостью, включая количество активных дней в виртуальной образовательной среде, общее число кликов и сконструированный показатель `score_consistency`, были определены как наиболее значимые предикторы как методами глобальной важности признаков, так и SHAP-анализом. Абляционное исследование показывает, что конструирование признаков, обработка дисбаланса классов и настройка гиперпараметров по отдельности вносят дополнительный вклад в итоговое качество, а их сочетание обеспечивает наилучший баланс точности, обобщающей способности и вычислительной эффективности.

В диссертации также рассматриваются образовательные, этические и внедренческие последствия предложенного конвейера. Показано, что объяснения на основе SHAP позволяют получать индивидуализированные и практически применимые выводы, которые могут быть представлены преподавателям и лицам, принимающим решения, без необходимости глубокого понимания внутренних механизмов алгоритмов. Явно обозначены ограничения, связанные с использованием одного эталонного набора данных, бинарной постановкой задачи и отсутствием причинно-следственных утверждений. В качестве направлений будущих исследований предложены временное моделирование, обучение с учётом справедливости и интеграция с существующими панелями learning analytics. В целом диссертация предлагает воспроизводимый методологический шаблон, соединяющий высокую прогностическую точность и интерпретируемые человеком выводы в контексте прогнозирования успеваемости студентов.

Ключевые слова: прогнозирование успеваемости студентов, интеллектуальный анализ образовательных данных, learning analytics, машинное обучение, XGBoost, SHAP, конструирование признаков, OULAD, объяснимый искусственный интеллект, системы раннего предупреждения.

Abbreviations

AI	Artificial Intelligence
ANOVA	Analysis of Variance
API	Application Programming Interface
AUC	Area Under the Curve
CART	Classification and Regression Tree
CSV	Comma-Separated Values
CV	Cross-Validation
DT	Decision Tree
EDA	Exploratory Data Analysis
EDM	Educational Data Mining
ETL	Extract, Transform, Load
F1	F1 score (harmonic mean of precision and recall)
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GBM	Gradient Boosting Machine
GBDT	Gradient Boosted Decision Tree
GDPR	General Data Protection Regulation
IMD	Index of Multiple Deprivation
KDE	Kernel Density Estimation
LA	Learning Analytics
LightGBM	Light Gradient Boosting Machine
LMS	Learning Management System
LR	Logistic Regression
MAE	Mean Absolute Error
MI	Mutual Information
ML	Machine Learning
MOOC	Massive Open Online Course
MSE	Mean Squared Error
OULAD	Open University Learning Analytics Dataset
RF	Random Forest
ROC	Receiver Operating Characteristic
SDU	Süleyman Demirel University
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Over-sampling Technique
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
TPE	Tree-structured Parzen Estimator
TPR	True Positive Rate
VLE	Virtual Learning Environment
XAI	eXplainable Artificial Intelligence
XGBoost	eXtreme Gradient Boosting

Table of Contents

Declaration i

Acknowledgements ii

Dedication iii

Abstract iv

Аңдатпа v

Аннотация vi

List of Abbreviations vii

List of Figures x

List of Tables xi

1 Introduction 1
- 1.1 Background and Motivation . . . 1
- 1.2 Problem Statement . . . 2
- 1.3 Research Objectives . . . 2
- 1.4 Research Questions . . . 2
- 1.5 Scope and Delimitations . . . 3
- 1.6 Contributions . . . 3
- 1.7 Thesis Organization . . . 4

2 Literature Review 5
- 2.1 Educational Data Mining and Learning Analytics . . . 5
- 2.2 Student Performance Prediction . . . 6
- 2.3 Machine Learning Approaches in Education . . . 6
  - 2.3.1 Linear and Logistic Models . . . 6
  - 2.3.2 Decision Trees and Random Forests . . . 6
  - 2.3.3 Gradient Boosting and Modern Ensembles . . . 7
  - 2.3.4 Deep Learning Approaches . . . 7
  - 2.3.5 Other Approaches . . . 7
- 2.4 Feature Engineering in Educational Data . . . 7
- 2.5 Class Imbalance and Resampling . . . 7
- 2.6 Model Interpretability and Explainable AI . . . 8
- 2.7 Studies on the OULAD Dataset . . . 8
- 2.8 Research Gaps and Positioning of This Work . . . 9

3 Theoretical Background 10
- 3.1 Supervised Classification . . . 10
- 3.2 Logistic Regression . . . 10
- 3.3 Decision Trees and Random Forests . . . 11
- 3.4 Gradient Boosting . . . 11
  - 3.4.1 XGBoost . . . 11
  - 3.4.2 LightGBM . . . 12
- 3.5 Evaluation Metrics . . . 12

3.6 Class Imbalance and SMOTE . . . 13
3.7 Bayesian Hyperparameter Optimisation . . . 13
3.8 SHapley Additive exPlanations . . . 13

4 Methodology and Implementation **15**
4.1 Research Design . . . 15
4.2 The OULAD Dataset . . . 15
4.2.1 Outcome Labelling . . . 16
4.3 Exploratory Data Analysis . . . 16
4.3.1 Distribution of Outcomes . . . 17
4.3.2 Distributions of Numerical Features . . . 17
4.3.3 Pass Rates Across Categorical Features . . . 18
4.3.4 Correlation Among Numerical Features . . . 18
4.4 Data Preprocessing . . . 19
4.5 Feature Engineering . . . 20
4.6 Feature Selection . . . 21
4.7 Class Imbalance Handling . . . 23
4.8 Model Training . . . 23
4.9 Hyperparameter Optimisation with Optuna . . . 24
4.10 Implementation and Technologies . . . 24
4.11 Reproducibility Considerations . . . 24

5 Results and Discussion **26**
5.1 Model Comparison . . . 26
5.2 ROC Analysis . . . 27
5.3 Overfitting Analysis . . . 28
5.4 Ablation Study . . . 29
5.5 SHAP Interpretability . . . 30
5.5.1 Global Feature Importance . . . 30
5.5.2 Beeswarm Plot . . . 31
5.5.3 Waterfall Plot for an At-Risk Student . . . 32
5.6 Comparison with Prior OULAD Studies . . . 33
5.7 Educational Implications . . . 34
5.8 Ethical Considerations . . . 35
5.9 Limitations and Threats to Validity . . . 35

6 Conclusions and Future Work **36**
6.1 Summary of the Work . . . 36
6.2 Answers to the Research Questions . . . 36
6.3 Restatement of Contributions . . . 37
6.4 Educational and Deployment Implications . . . 37
6.5 Limitations Revisited . . . 37
6.6 Future Work . . . 37

Bibliography **38**

List of Figures

4.1 Distribution of final academic outcomes in OULAD 17
4.2 Distributions of numerical features stratified by target class 18
4.3 Pass rates by categorical feature levels 18
4.4 Correlation heatmap of numerical features in OULAD 19
4.5 Filter-method feature importance 21
4.6 Permutation-based feature importance 22

5.1 Test-set comparison of model performance 27
5.2 ROC curves on the test set 28
5.3 Training versus cross-validation F1 29
5.4 Global SHAP feature importance 31
5.5 SHAP beeswarm plot 32
5.6 SHAP waterfall plot for an at-risk student 33

List of Tables

2.1 Representative OULAD studies and selected reported results 9

4.1 Selected feature set after engineering and selection 23

5.1 Model performance comparison across splits 26

Chapter 1

Introduction

1.1 Background and Motivation

Higher education is undergoing a profound transformation driven by the rapid digitisation of teaching, learning and administrative processes. Universities increasingly rely on Virtual Learning Environments (VLEs) and Learning Management Systems (LMS) to deliver lectures, distribute course materials, manage assessments and communicate with students. Every login, mouse click, video play, forum post and assignment submission generates a fine-grained, time-stamped record of student behaviour. Aggregated across thousands of students and tens of course presentations, these records constitute a unique source of behavioural and assessment data that was simply not available to instructors a decade ago [1, 2].

Despite the proliferation of digital infrastructures, the fundamental challenges that motivated the foundational work of Tinto on student dropout in the 1970s remain [3]. A non-trivial proportion of students enrolled in undergraduate and postgraduate programmes still fail to complete their courses, withdraw before the final assessment or under-perform with respect to their academic potential. Empirical evidence collected across multiple educational systems shows that early failure is rarely the result of a single dramatic event; it is, instead, the cumulative consequence of small disengagement signals—missed tutorial sessions, unsubmitted assignments, declining VLE usage, decreasing assessment scores—that, taken individually, may go unnoticed by overworked academic staff.

The economic and social cost of academic failure is high. From the perspective of the student, dropping out of a degree programme can translate into debt without a corresponding qualification, lost opportunities in the labour market and a measurable reduction in lifetime earnings. From the perspective of the institution, low retention rates damage reputation, reduce funding and complicate strategic planning. From a broader societal perspective, the wastage of human potential implied by avoidable failure is at odds with the ambition of inclusive, high-quality higher education [4, 5].

Against this background, the rise of *learning analytics* (LA) and *educational data mining* (EDM) provides a promising direction. Learning analytics has been defined as the measurement, collection, analysis and reporting of data about learners and their contexts for the purposes of understanding and optimising learning and the environments in which it occurs [2]. Educational data mining, while overlapping with LA, places stronger emphasis on the application of computational techniques—statistical learning, data mining, machine learning, natural language processing—to extract patterns from educational data [1, 4]. Within these communities, *student performance prediction* has become a flagship application: by training predictive models on historical data, it is possible, in principle, to identify students who are likely to fail or withdraw weeks before the final assessment, thus enabling targeted interventions.

A particularly influential resource for this line of research is the Open University Learning Analytics Dataset (OULAD) [6]. Released by The Open University in the United Kingdom, OULAD is a publicly available, anonymised collection of data records describing 32,593 students enrolled in 22 module-presentations between 2013 and 2014. It contains demographic information, course registration data, detailed clickstream logs from the VLE and assessment results, all linked through pseudonymous identifiers. OULAD has become a de facto benchmark for student performance prediction studies thanks to its scale, structure and free licensing. The work presented in this dissertation builds on this benchmark to train and evaluate a family of classification models for the binary task of distinguishing students who will pass a module from those who will fail or withdraw.

1.2 Problem Statement

While the academic literature on student performance prediction has matured significantly over the past decade, a number of practical and methodological problems still constrain the real-world adoption of predictive models in higher education. The work reported in this dissertation is motivated by four interlocking concerns.

Predictive accuracy alone is insufficient. Many published models on OULAD and similar datasets achieve respectable test-set scores, but their predictions remain difficult to translate into pedagogical action. A model that flags a student as "at risk" without explaining why is of limited value to a busy course director who must decide whether to allocate scarce support resources. The opacity of the predictions also raises legitimate concerns about fairness, accountability and informed consent [7, 8].

The contribution of individual pipeline components is rarely quantified. A typical machine learning pipeline for educational data is composed of multiple stages: data integration, preprocessing, feature engineering, feature selection, class-imbalance handling, model training and hyperparameter tuning. Despite their methodological importance, many studies report only an end-to-end accuracy number, leaving readers unable to disentangle the contribution of each stage. This makes it difficult to decide where to invest implementation effort when adapting a published pipeline to a new institutional context.

Class imbalance is often handled in an opaque manner. The distribution of pass, fail, distinction and withdrawal outcomes in OULAD is moderately imbalanced. Although a number of resampling techniques have been proposed in the literature—including SMOTE [9], ADASYN and a variety of cost-sensitive learning schemes—they are not always integrated correctly into the cross-validation procedure, leading to overly optimistic performance estimates [9].

Reproducibility remains a challenge. The combination of evolving software libraries, undocumented preprocessing decisions and missing code makes it difficult to reproduce reported results. Increasingly, the educational data mining community recognises that reproducibility is a prerequisite for cumulative scientific progress in the field.

The present dissertation addresses these concerns through a comprehensive case study built around OULAD, in which an end-to-end, well-documented pipeline is constructed, every component is ablated, multiple competing models are compared on the same data split, and predictions are explained using state-of-the-art interpretability tools.

1.3 Research Objectives

The overall objective of this research is to design, implement and rigorously evaluate a reproducible machine learning pipeline that predicts student academic outcomes on the OULAD dataset and that provides educationally meaningful explanations of its predictions. This high-level objective is decomposed into a number of more specific sub-objectives.

The first sub-objective is to assemble a coherent, well-documented data preprocessing pipeline that integrates the heterogeneous OULAD tables—student demographics, course registrations, VLE interaction logs and assessment outcomes—into a flattened feature matrix suitable for supervised learning. The second sub-objective is to design and motivate a set of domain-driven engineered features that capture aspects of student engagement and consistency that are not directly represented in the raw data. The third sub-objective is to select an informative subset of features using complementary filter, embedded and SHAP-based ranking methods, and to compare the rankings produced by these methods.

The fourth sub-objective is to train and evaluate four representative classifiers—Logistic Regression, Random Forest, XGBoost and LightGBM—on the same task and to compare their performance using a comprehensive set of evaluation metrics. The fifth sub-objective is to tune the hyperparameters of the two best-performing classifiers using Bayesian optimisation, in order to quantify the additional improvement that can be obtained from automated tuning. The sixth sub-objective is to conduct an ablation study that explicitly measures the marginal contribution of feature engineering, SMOTE resampling and Optuna-based tuning to overall predictive accuracy and generalisation. The seventh and final sub-objective is to apply SHAP-based interpretability to the best-performing model and to discuss the educational and ethical implications of the resulting explanations.

1.4 Research Questions

The research objectives outlined above are formalised through the following four research questions, which structure the empirical investigation conducted in this dissertation.

RQ1. Which student demographic, registration and behavioural features are most predictive of binary academic outcome (pass-or-distinction versus fail-or-withdrawn) on the OULAD dataset?

RQ2. Which machine learning model—among Logistic Regression, Random Forest, XGBoost and LightGBM—best captures the complex, non-linear relationships present in the student data, when evaluated on a held-out test set using a coherent set of metrics?

RQ3. How much does each pipeline stage (domain-driven feature engineering, SMOTE class-imbalance handling and Optuna-based Bayesian hyperparameter optimisation) contribute to the predictive performance and generalisation of the final model?

RQ4. What actionable, educationally meaningful insights can be derived by combining global and local SHAP-based interpretability with the best-performing predictive model, and how can such insights be communicated to educators and policymakers?

Each of these questions is addressed by a dedicated empirical analysis. RQ1 is addressed through filter-based feature importance (mutual information and ANOVA F-statistic), embedded importance from Random Forest and SHAP-based global importance. RQ2 is addressed through a head-to-head model comparison on common training, validation and test splits using F1, AUC, accuracy, precision, recall and a custom overfit gap. RQ3 is addressed through a structured ablation study that disables one pipeline component at a time. RQ4 is addressed through global, local and individual SHAP explanations applied to the best-performing model.

1.5 Scope and Delimitations

To ensure that the present study is tractable within the time frame of an MSc dissertation, a number of explicit scope decisions are made. The dataset is restricted to OULAD; no additional institutional data are introduced. The prediction task is binary: students with a final result of `Pass` or `Distinction` are placed in the positive class, while students with a final result of `Fail` or `Withdrawn` are placed in the negative class. A four-class multiclass formulation, although educationally meaningful, is left for future work.

The learning paradigm considered is supervised classification on tabular, aggregated data. Deep neural networks for sequential modelling of VLE interactions are not part of the experimental comparison, both because the aggregation step removes the temporal dimension and because gradient-boosted trees have been repeatedly shown to be competitive with deep models on tabular benchmarks of similar size [10, 11]. The interpretability framework employed is SHAP, motivated by its strong theoretical grounding and broad adoption in the literature [7].

Finally, the experimental setup uses a single 70/15/15 stratified split with a fixed random seed for reproducibility. Although this is a common choice in the educational data mining literature, more elaborate evaluation protocols such as nested cross-validation or leave-one-presentation-out evaluation could provide additional robustness; these are discussed as part of future work in Chapter 6.

1.6 Contributions

The principal contributions of this dissertation can be summarised as follows.

First, the dissertation contributes a fully documented, end-to-end machine learning pipeline for student performance prediction on the OULAD dataset. The pipeline integrates multi-source data, applies domain-driven feature engineering, performs hybrid feature selection, handles class imbalance through SMOTE, optimises hyperparameters via Optuna and explains predictions using SHAP. The full configuration is described in sufficient detail to allow replication.

Second, the dissertation contributes a head-to-head comparison of four representative classifiers—Logistic Regression, Random Forest, XGBoost and LightGBM—under identical evaluation conditions. The comparison includes a custom overfit gap metric that explicitly contrasts training-set and cross-validation performance, providing a more nuanced picture of generalisation than reported test scores alone.

Third, the dissertation contributes an ablation study quantifying the marginal contribution of feature engineering, SMOTE and Optuna-based tuning to final performance. This study is intended to give practitioners empirical guidance about where to invest implementation effort.

Fourth, the dissertation contributes an interpretability analysis combining global SHAP feature-importance summaries, beeswarm visualisations of feature-value distributions and waterfall plots of individual at-risk students. The discussion translates these technical artefacts into educationally meaningful narratives suitable for non-technical stakeholders.

Fifth, the dissertation contributes a discussion of educational, ethical and deployment-related implications. Particular attention is paid to the limitations of using a single benchmark dataset, the dangers of automated decisions in high-stakes contexts and the conditions under which the proposed pipeline could be safely integrated into an existing learning analytics platform.

1.7 Thesis Organization

The remainder of this dissertation is organised as follows.

Chapter 2 presents a comprehensive literature review covering educational data mining, learning analytics, student performance prediction, classical and ensemble supervised learning algorithms, model interpretability and previous studies on the OULAD dataset. The chapter concludes by identifying methodological gaps that motivate the present work.

Chapter 3 provides the theoretical background needed to understand the algorithmic choices made in the rest of the dissertation. It covers the fundamentals of supervised classification, the mathematical formulation of Logistic Regression, Decision Trees, Random Forest, gradient boosting, XGBoost and LightGBM, the standard evaluation metrics for binary classification, the rationale of class-imbalance handling and the mathematical underpinnings of SHAP.

Chapter 4 describes the methodology and implementation of the proposed pipeline. It introduces the OULAD dataset in detail, presents exploratory data analysis, motivates the data preprocessing decisions, derives the engineered features, explains the feature selection procedure, describes the application of SMOTE, formalises the model training protocol, details the Optuna-based hyperparameter optimisation setup, and reports on the software stack and reproducibility conventions used throughout.

Chapter 5 presents and discusses the empirical results. It reports the model comparison, the analysis of ROC curves and confusion matrices, the overfitting analysis, the ablation study, the SHAP-based interpretability results and a structured comparison with previous studies. The chapter also discusses limitations, threats to validity and educational implications of the findings.

Chapter 6 concludes the dissertation. It summarises the answers to each research question, restates the contributions, discusses ethical and deployment considerations and outlines several promising directions for future work, including multiclass and temporal modelling, fairness-aware learning, transfer to other institutional contexts and integration into existing institutional dashboards.

The dissertation closes with a bibliography of all cited works, formatted in IEEE citation style.

Chapter 2

Literature Review

This chapter surveys the body of literature on which the present dissertation builds. It begins by characterising the fields of educational data mining and learning analytics and tracing their historical development. It then narrows the scope to student performance prediction, examining the variety of data sources, predictive tasks and modelling techniques that have been applied in the literature. Subsequent sections review the family of supervised learning algorithms relevant to the present work, discuss the role of feature engineering and class imbalance, and examine the recent shift toward explainable artificial intelligence in educational settings. The chapter closes with a focused review of prior research on the OULAD dataset and a structured analysis of the gaps that the present dissertation seeks to address.

2.1 Educational Data Mining and Learning Analytics

The fields of Educational Data Mining (EDM) and Learning Analytics (LA) have emerged at the intersection of data science, computer science, statistics, education and the cognitive sciences. Although they overlap substantially, EDM has been historically oriented toward the development of algorithmic methods that extract patterns from educational data, while LA has been more concerned with the operationalisation of those methods in concrete educational settings and the communication of results to practitioners [1, 4].

Early work in EDM focused on the application of classical data mining techniques—decision trees, clustering, sequential pattern mining and association rule mining—to relatively small datasets collected from intelligent tutoring systems and learning management systems. The comprehensive review by Romero and Ventura [1] provided one of the first systematic surveys, organising the literature according to the methodological categories of classification, regression, clustering, association rule mining and visualisation. They identified student performance prediction, course recommendation and behavioural pattern discovery as the most active sub-areas of EDM.

The work of Baker and Yacef [4] took a complementary perspective, framing EDM as a discipline whose central questions concern *how* students learn rather than merely *whether* they succeed. They emphasised the importance of cognitive models, knowledge tracing and skill mastery in addition to outcome prediction. This dual orientation—predictive and explanatory—has continued to characterise the field.

A subsequent generation of reviews, notably Papamitsiou and Economides [2] and Avella et al. [5], mapped the practice of LA across higher education institutions. These reviews highlighted that, despite a rapid increase in the number of published case studies, a non-trivial gap remained between methodological research and operational deployment. The most frequently reported obstacles included the heterogeneity of institutional data, the lack of standardised pipelines, ethical concerns about student profiling and the difficulty of communicating algorithmic outputs to academic staff. The present dissertation takes these obstacles seriously and explicitly addresses them through reproducible methodology and explainability.

The decade between 2015 and 2025 witnessed a marked shift toward large-scale online learning, including Massive Open Online Courses (MOOCs), blended degree programmes and the emergency move to fully online instruction during the COVID-19 pandemic. As a result, the volume and granularity of data available for EDM and LA increased dramatically, enabling more ambitious modelling efforts but also intensifying ethical and methodological concerns [12, 13].

2.2 Student Performance Prediction

Student performance prediction (SPP) is the subset of EDM that focuses on forecasting academic outcomes, where "outcome" can be operationalised in a variety of ways. In some studies, the target variable is a continuous grade or grade-point average; in others, it is a binary or multiclass label representing pass/fail or pass/fail/distinction/withdrawal categories [12, 14]. The latter formulation is particularly relevant when the predictive model is intended to support early-warning interventions.

A widely cited early study is that of Kotsiantis [14], who showed that simple machine learning models trained on a small set of administrative and assessment features could outperform human-designed heuristics for grade forecasting. Although the dataset and the modelling techniques used in that study are modest by contemporary standards, the broader argument—that data-driven systems can complement experienced academic staff—remains influential.

A complementary line of work has emphasised the practical role of LMS interaction data in early-warning systems. The seminal study of Macfadyen and Dawson [15] demonstrated that aggregate LMS usage metrics computed during the first weeks of a course were sufficient to identify a substantial proportion of students who would eventually fail. Their results provided a proof of concept that simple, transparent features—logins, forum posts, resource views—could yield actionable risk indicators. Subsequent studies progressively refined this idea, exploring richer feature sets, longer observation windows and more sophisticated models.

A more recent and methodologically rigorous study is that of Adnan et al. [12], who systematically evaluated multiple machine learning algorithms on the OULAD dataset and reported how predictive performance varies as a function of the percentage of course length elapsed. Their analysis emphasised the trade-off between earliness of prediction and accuracy: models trained on data collected after only 10% of the course tend to under-perform models that observe a larger fraction of the term, but they also enable earlier intervention. This trade-off is central to operational early-warning systems and is taken into account in the discussion presented in Chapter 5 of this dissertation.

A complementary contribution is provided by Shaikhanova et al. [16], who proposed an interpretable predictive modelling framework for educational equity. Their study aimed at identifying at-risk students at very early stages of a course while balancing predictive accuracy with workload-aware decision support. Such work demonstrates that SPP research is increasingly concerned with the operational constraints of educational staff rather than predictive accuracy alone.

2.3 Machine Learning Approaches in Education

The supervised learning techniques applied to SPP can be roughly grouped into four families: linear models, decision tree models, ensemble models and neural-network models. Each family makes different inductive biases about the structure of the data, and each comes with its own trade-offs between accuracy, interpretability and computational cost.

2.3.1 Linear and Logistic Models

Logistic Regression remains the workhorse of statistical modelling in education. It is interpretable in terms of log-odds coefficients, computationally efficient and robust to small samples. Its main limitation is that it cannot natively capture non-linear interactions between features. In educational data, however, non-linear dependencies between, for example, engagement intensity and prior education are common, and pure linear models often underperform ensemble alternatives [14]. Nevertheless, Logistic Regression continues to serve as a useful baseline against which more complex models can be calibrated, and it is included in the comparison reported in Chapter 5 for that reason.

2.3.2 Decision Trees and Random Forests

Decision Trees (DT) provide a natural way to model non-linear interactions and have the additional advantage of being interpretable when their depth is constrained. However, single trees suffer from high variance: small changes in the training data may produce substantially different trees, and unconstrained trees are prone to overfitting [17]. Random Forests [17] address these limitations by training a large number of de-correlated trees on bootstrap samples of the data and aggregating their predictions. Random Forests are robust to noise, handle mixed-type features without elaborate preprocessing and provide useful out-of-bag estimates of generalisation error. In educational data mining, Random Forests have been used as both a predictor and a feature-importance tool [12].

2.3.3 Gradient Boosting and Modern Ensembles

Gradient Boosting Machines (GBMs) train a sequence of weak learners, typically shallow regression trees, in such a way that each new tree fits the negative gradient of the loss with respect to the predictions of the existing ensemble [18]. Modern implementations of GBMs include XGBoost [10] and LightGBM [11], which differ in how they grow trees, how they handle missing values and how they exploit the histograms of feature values.

XGBoost [10] introduced a number of practical optimisations—regularised loss functions, sparsity-aware split finding, parallel tree construction—that turned gradient boosting into the dominant approach for tabular data on a wide range of benchmarks. LightGBM [11] subsequently introduced leaf-wise tree growth and gradient-based one-side sampling, achieving comparable accuracy with substantially reduced training time. Both libraries have become standard tools in EDM and SPP, often producing the best-reported results on benchmark datasets [13, 19].

2.3.4 Deep Learning Approaches

Deep neural networks, including recurrent and attention-based architectures, have been applied to SPP, particularly when the goal is to model time-series data such as raw clickstream sequences or longitudinal assessment trajectories. Convolutional and recurrent networks can capture temporal patterns that are flattened away when interaction data are aggregated. However, on aggregated tabular data with sample sizes typical of single institutions, deep models rarely outperform well-tuned gradient-boosted trees and often require more careful regularisation and substantially more training data [20]. For this reason, and to keep the experimental design comparable to other OULAD studies, the present dissertation focuses on tree-based models and uses Logistic Regression as a linear baseline.

2.3.5 Other Approaches

In addition to the families above, a variety of other approaches have been explored in EDM. Support Vector Machines (SVM) have been used in earlier studies but tend to scale poorly with the size of modern educational datasets and are sensitive to feature scaling. k-Nearest Neighbours classifiers have been used as benchmarks but rarely as final models. Bayesian Knowledge Tracing and its modern hybrid variants have proven valuable in skill mastery modelling but are less suited to the binary outcome task considered here. Multiclass and ordinal regression models have been investigated for problems where the dependent variable retains multiple ordered categories.

2.4 Feature Engineering in Educational Data

Feature engineering plays a critical role in educational data mining because the raw data captured by LMS platforms is highly structured but rarely directly informative. A click event, on its own, says little about a student's likelihood of passing a course; aggregated and contextualised, however, click events can encode rich behavioural patterns [21].

Tempelaar et al. [21] provided an important empirical study showing that engineered features representing time-on-task, regularity of access and density of formative assessment activity yielded substantially better predictive performance than raw counts. Subsequent studies have extended these ideas with features that capture diurnal patterns, weekly regularity, transition frequencies between resource types and ratios that normalise raw activity counts by course length or by the number of credits.

In addition to these behavioural transformations, demographic and prior-performance features are typically encoded through a combination of one-hot encoding (for nominal variables), ordinal encoding (for ordered variables such as highest education and index of multiple deprivation bands) and target encoding (for high-cardinality categorical variables). For benchmark datasets such as OULAD, the design of features has now been documented in multiple studies, providing a shared starting point for further work [6, 12, 13].

2.5 Class Imbalance and Resampling

Class imbalance is endemic to SPP because the outcome categories are rarely equally distributed. In OULAD, the binary task formulation used in this dissertation is only mildly imbalanced—roughly 53% positive against 47% negative—but more granular formulations can exhibit much stronger imbalance, particularly when distinction is treated as its own class.

The Synthetic Minority Over-sampling Technique (SMOTE) of Chawla et al. [9] is the most widely used approach to class imbalance in EDM. SMOTE generates synthetic examples of the minority class by interpolating

between each minority-class instance and one of its k nearest neighbours in feature space, expanding the decision boundary of the minority class. Variants such as Borderline SMOTE and ADASYN refine this approach by concentrating the synthesis on samples near the decision boundary.

A frequent methodological pitfall in EDM applications of SMOTE is the inadvertent leakage of synthetic samples from the training set into the validation or test set. Such leakage causes overly optimistic performance estimates [6]. The methodology of the present dissertation follows the now-established good practice of applying SMOTE exclusively to the training split after the train/validation/test partition has been frozen.

An alternative line of work avoids resampling altogether and relies instead on cost-sensitive learning or evaluation metrics that explicitly account for class imbalance. F1, AUC and balanced accuracy are widely used for this purpose and are computed in Chapter 5 of this dissertation alongside more traditional accuracy.

2.6 Model Interpretability and Explainable AI

The growing reliance on black-box machine learning models in high-stakes domains has motivated a renaissance of interest in model interpretability. Although linear models and shallow decision trees are inherently interpretable, the gradient-boosted tree ensembles that dominate SPP benchmarks are not. Post-hoc explainability methods aim to bridge this gap by producing human-readable explanations of predictions made by complex models.

Among the most influential post-hoc explainability methods is the SHapley Additive exPlanations (SHAP) framework of Lundberg and Lee [7]. SHAP unifies several earlier methods—including LIME, DeepLIFT and Tree Interpreter—through a single axiomatic formulation grounded in cooperative game theory. For each prediction, SHAP computes a value for each feature that represents its marginal contribution to the prediction relative to a baseline. SHAP values satisfy the desirable properties of local accuracy, missingness, consistency and efficiency, and they can be efficiently computed for tree-based models [7].

In educational contexts, SHAP has been used both to provide global feature importance summaries and to explain individual student predictions. Global summaries inform institutional policy by highlighting which factors most strongly drive risk classifications across the population. Local explanations inform individualised intervention by pointing to the specific behavioural and demographic factors that put a particular student at risk.

The growing emphasis on explainability is also aligned with regulatory developments. The European Union's General Data Protection Regulation (GDPR), for example, recognises a limited right to explanation in the context of automated decision-making. While the legal interpretation of this right is still evolving, it provides additional motivation for institutions to deploy interpretable or explainable models when machine learning is used to inform decisions about students [8].

2.7 Studies on the OULAD Dataset

The OULAD dataset has been the subject of dozens of studies since its release in 2017, and a comprehensive systematic literature review by Jin et al. [13] provides a useful map of the territory. Their review identifies three dominant lines of work on OULAD: (i) final outcome prediction, in which a binary or multiclass classifier is trained on demographic, registration and engagement features; (ii) early-warning prediction, in which the same task is performed using only a fraction of the course as observation window; and (iii) engagement modelling, in which the focus is on understanding patterns of interaction rather than predicting outcomes.

Adnan et al. [12] represent the first family. They compared several classical and ensemble classifiers on OULAD and reported the variation of predictive performance as a function of course completion percentage. Shaikhanova et al. [16] focused on the early-warning setting, prioritising interpretable predictions at early stages of the course. Lu et al. [19] introduced a course-level clustering approach combined with LightGBM, demonstrating that exploiting between-module heterogeneity can yield non-trivial accuracy gains. Gunasekara and Saarela [8] provided one of the more recent and methodologically rigorous studies on OULAD, focusing on the trade-offs between predictive performance and fairness under post-processing bias mitigation.

Table 2.1 summarises a small set of representative OULAD studies and highlights their methodological choices and the gaps that remain.

Table 2.1: Representative OULAD studies and selected reported results. NR denotes "not reported" in the cited source.

Paper	Year	Model	AUC	F1	Key limitation
Adnan et al. [12]	2021	Random Forest	NR	0.79	Early checkpoints; limited interpretability.
Shaikhanova et al. [16]	2025	Gradient Boosting	0.79	NR	Lower performance; limited feature engineering.
Lu et al. [19]	2025	Clustered LightGBM	0.86	0.78	Gains depend on course clustering.
Gunasekara and Saarela [8]	2025	Fairness-aware GBT	NR	NR	Focus on fairness, not predictive accuracy.

2.8 Research Gaps and Positioning of This Work

The literature reviewed in the preceding sections demonstrates the maturity of SPP as a research area, but it also reveals a number of persistent gaps that motivate the present dissertation.

First, despite the breadth of methodological work, relatively few studies report a comprehensive end-to-end pipeline in which every preprocessing decision, feature engineering choice and modelling option is documented in sufficient detail for replication. The first contribution of this dissertation is to provide such a pipeline for OULAD.

Second, the relative contribution of individual pipeline stages—feature engineering, class-imbalance handling, hyperparameter optimisation—is rarely quantified. While many studies report aggregate performance numbers, fewer report an ablation that isolates each component. The second contribution of this dissertation is precisely such an ablation.

Third, although SHAP has been used in some recent OULAD studies, it is more often presented through global summaries alone, without engaging with individual local explanations. The third contribution of this dissertation is to combine global, beeswarm and waterfall SHAP visualisations and to discuss their pedagogical implications.

Fourth, the broader question of how predictive models should be communicated to educators and policymakers is rarely addressed in detail. The fourth and final positioning of the dissertation is to treat the communication of model outputs as a first-class methodological concern and to provide concrete narrative templates for converting SHAP plots into actionable summaries.

In summary, the present dissertation does not claim to introduce a fundamentally new machine learning algorithm; rather, it claims to integrate established components into a coherent, reproducible and explainable pipeline whose marginal contributions are explicitly quantified. The next chapter develops the theoretical background necessary to understand the algorithmic choices made in the rest of the dissertation.

Chapter 3

Theoretical Background

This chapter develops the mathematical and conceptual foundations on which the methodology of Chapter 4 rests. It begins by formalising supervised classification, presents the four learning algorithms used in the experimental comparison, and then turns to evaluation metrics, class-imbalance handling and the SHAP framework. The material is organised so that each algorithmic choice in subsequent chapters can be traced back to a precise mathematical formulation.

3.1 Supervised Classification

In a supervised classification problem, the learner is provided with a training set $\mathcal{D}_{\text{train}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, where each $\mathbf{x}_i \in \mathbb{R}^d$ is a feature vector representing one student, and $y_i \in \{0, 1\}$ is the corresponding class label. The label takes the value 1 when the student passes the course (positive class) and 0 when the student fails or withdraws (negative class). The objective of supervised learning is to estimate a function $f : \mathbb{R}^d \to \{0, 1\}$ that minimises a chosen loss $\mathcal{L}(y, \hat{y})$ on unseen data drawn from the same underlying distribution as $\mathcal{D}_{\text{train}}$.

A central concern in supervised learning is the trade-off between bias and variance. A model with high bias makes strong assumptions about the structure of the data and may fail to capture important regularities; a model with high variance is flexible enough to fit small fluctuations in the training set but generalises poorly. The various algorithms presented below differ in how they manage this trade-off through architectural choices and regularisation.

3.2 Logistic Regression

Logistic Regression (LR) is a parametric, linear classification model that estimates the probability of class membership through the logistic transformation of a linear combination of features. Formally, given a weight vector $\mathbf{w} \in \mathbb{R}^d$ and a bias $b \in \mathbb{R}$, the predicted probability of the positive class is

$$\Pr(Y = 1 \mid \mathbf{x}) = \sigma\big(\mathbf{w}^\top \mathbf{x} + b\big), \quad \sigma(z) = \frac{1}{1 + e^{-z}}. \tag{3.2.1}$$

The parameters $\mathbf{w}$ and b are estimated by minimising the regularised negative log-likelihood,

$$\mathcal{L}_{\text{LR}}(\mathbf{w}, b) = -\sum_{i=1}^{n} \left[y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i) \right] + \lambda \, \|\mathbf{w}\|_p^p, \tag{3.2.2}$$

where $\hat{p}_i = \sigma(\mathbf{w}^\top \mathbf{x}_i + b)$ and $\|\mathbf{w}\|_p^p$ is an ℓ_p regulariser. With $p = 2$ the model is called ℓ_2-regularised, while $p = 1$ yields ℓ_1 regularisation that encourages sparse solutions.

LR is interpretable in the sense that the sign and magnitude of each coefficient can be related to changes in the log-odds of the positive class. It scales gracefully to high-dimensional problems and can be optimised efficiently with quasi-Newton or stochastic gradient methods. However, LR is fundamentally a linear classifier in the feature space and cannot capture interactions or non-linearities unless they are explicitly engineered as additional features. This limitation motivates the use of non-linear models such as Random Forest and gradient-boosted trees in the experimental comparison.

3.3 Decision Trees and Random Forests

Decision Trees (DT) recursively partition the feature space into axis-aligned regions and assign a constant prediction to each leaf. At each internal node, a split is chosen by maximising an impurity-reduction criterion. For binary classification, the Gini impurity at a node t with class proportions p_0 and p_1 is

$$I_{\text{Gini}}(t) = 1 - p_0^2 - p_1^2, \tag{3.3.1}$$

while the entropy is

$$H(t) = -p_0 \log_2 p_0 - p_1 \log_2 p_1. \tag{3.3.2}$$

The information gain associated with a split that partitions the data at node t into children t_L and t_R is the impurity at t minus the weighted impurity in t_L and t_R. Trees are grown greedily until a stopping criterion is satisfied—typically a minimum number of samples per leaf, a maximum depth or a minimum impurity decrease—and may then be pruned to mitigate overfitting.

The Random Forest (RF) ensemble of Breiman [17] mitigates the high variance of single decision trees by averaging the predictions of many trees, each grown on a bootstrap sample of the training data and, at each split, considering only a random subset of features. Let $\{f_b\}_{b=1}^{B}$ be the predictions of B trees grown according to this procedure. The Random Forest prediction is

$$\hat{f}_{\text{RF}}(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} f_b(\mathbf{x}). \tag{3.3.3}$$

Random Forests produce reliable estimates of the generalisation error through the out-of-bag (OOB) procedure: each tree is evaluated on the samples that were not included in its bootstrap sample, and the OOB error converges to the true generalisation error as B grows. Random Forests also provide useful estimates of feature importance, either through the mean decrease in impurity or through permutation importance. In this dissertation, the latter is used because it is less biased toward high-cardinality features.

3.4 Gradient Boosting

Gradient Boosting [18] constructs an additive ensemble of weak learners, typically shallow regression trees, in a forward stage-wise manner. At iteration m, the ensemble's prediction is

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \nu \cdot h_m(\mathbf{x}), \tag{3.4.1}$$

where $\nu \in (0, 1]$ is the learning rate and h_m is a new base learner fitted to the negative gradient of the loss with respect to the previous ensemble's predictions, evaluated at the training points,

$$r_{im} = - \left[\frac{\partial \mathcal{L}(y_i, F(\mathbf{x}_i))}{\partial F(\mathbf{x}_i)} \right]_{F = F_{m-1}}. \tag{3.4.2}$$

For binary classification with the logistic loss, the negative gradient takes the simple form $r_{im} = y_i - p_{i,m-1}$, where $p_{i,m-1}$ is the predicted probability of the positive class at iteration $m-1$. The optimal step size for the new base learner is found through a one-dimensional line search, and shrinkage by the learning rate ν is applied to reduce the contribution of each new tree.

3.4.1 XGBoost

XGBoost [10] refines this basic recipe in several practically important ways. Most notably, it uses a second-order Taylor expansion of the loss around the current ensemble's predictions, which allows the optimal leaf values to be computed in closed form for any twice-differentiable loss. Letting g_i and h_i denote the first and second derivatives of the loss with respect to the prediction for sample i, the regularised objective at iteration m takes the form

$$\mathcal{L}^{(m)} = \sum_{i=1}^{n} \left[g_i f_m(\mathbf{x}_i) + \tfrac{1}{2} h_i f_m(\mathbf{x}_i)^2 \right] + \Omega(f_m), \tag{3.4.3}$$

with the complexity penalty

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2, \tag{3.4.4}$$

where T is the number of leaves and w_j is the weight of leaf j. The penalty γT encourages shallow trees with few leaves, while the ℓ_2 term $\lambda \sum w_j^2$ prevents extreme leaf values. Closed-form expressions are available for both the optimal leaf weights and the gain associated with each candidate split. XGBoost additionally implements sparsity-aware split finding, weighted quantile sketching for histogram-based binning, column subsampling and parallel tree construction. The combination of these features makes XGBoost particularly attractive for medium-sized tabular datasets such as OULAD.

3.4.2 LightGBM

LightGBM [11] retains the second-order objective of XGBoost but introduces several optimisations that further reduce training time. First, it grows trees in a leaf-wise rather than level-wise fashion: at each step the leaf that maximises the loss reduction is chosen for splitting, which tends to produce deeper, less balanced trees that nevertheless converge faster. Second, it uses Gradient-based One-Side Sampling (GOSS) to focus learning effort on samples with large gradients, and Exclusive Feature Bundling (EFB) to combine mutually exclusive sparse features. Third, it uses histogram binning of continuous features, which dramatically reduces the cost of split finding. LightGBM often achieves accuracy comparable to XGBoost at a fraction of the training time, although its leaf-wise growth strategy can lead to overfitting when no depth or leaf-count constraints are imposed.

3.5 Evaluation Metrics

The performance of binary classifiers is summarised through a small set of metrics computed from the confusion matrix, which counts true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). The classical definitions of *accuracy*, *precision*, *recall* and *F1 score* are

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}, \tag{3.5.1}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{3.5.2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{3.5.3}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{3.5.4}$$

The F1 score, defined as the harmonic mean of precision and recall, is particularly informative under class imbalance because it penalises models that maximise accuracy by predicting the majority class. In SPP, the cost of a false negative (failing to flag a struggling student) is often higher than the cost of a false positive (flagging a student who would have succeeded), although the precise trade-off depends on the institutional context.

The Receiver Operating Characteristic (ROC) curve plots the true positive rate $\text{TPR} = \text{TP}/(\text{TP} + \text{FN})$ against the false positive rate $\text{FPR} = \text{FP}/(\text{FP} + \text{TN})$ as the classifier's decision threshold varies. The Area Under the Curve (AUC) is the probability that a random positive example is ranked higher than a random negative example by the classifier and is invariant to monotone transformations of the prediction scores. In imbalanced binary problems, AUC is a useful threshold-independent complement to F1.

In addition to these standard metrics, the present dissertation reports an *overfit gap*, defined as the difference between the training F1 score and the mean F1 score under five-fold stratified cross-validation,

$$\Delta_{\text{overfit}} = \text{F1}_{\text{train}} - \text{F1}_{\text{CV}}. \tag{3.5.5}$$

A small or zero overfit gap suggests that the model is generalising well, while a large gap indicates that the model is memorising the training data. The overfit gap is reported alongside the other metrics so that readers can assess generalisation without having to infer it indirectly.

3.6 Class Imbalance and SMOTE

Let $\mathcal{D}_{\min}$ denote the minority class in a binary classification problem. SMOTE [9] generates a synthetic example for each minority instance $\mathbf{x}$ by selecting one of its k nearest neighbours $\mathbf{x}^{(k)}$ in feature space and producing

$$\mathbf{x}_{\text{syn}} = \mathbf{x} + \lambda \cdot (\mathbf{x}^{(k)} - \mathbf{x}), \quad \lambda \sim \mathcal{U}(0, 1). \tag{3.6.1}$$

Synthetic samples are added to $\mathcal{D}_{\min}$ until the desired class ratio is achieved. SMOTE expands the decision boundary of the minority class by populating it with plausible synthetic neighbours, rather than merely duplicating existing minority samples as in naive random oversampling.

Two important practical considerations apply. First, SMOTE should be applied *after* the train/validation/test split has been frozen and only to the training portion; applying it before splitting introduces synthetic samples into the validation and test sets, which inflates the reported performance. Second, in the presence of strongly overlapping class clusters, the synthetic samples may exacerbate rather than mitigate the classification difficulty; variants such as Borderline SMOTE and ADASYN address this issue by concentrating synthesis near the decision boundary.

3.7 Bayesian Hyperparameter Optimisation

The performance of gradient-boosted models depends critically on the choice of hyperparameters: maximum tree depth, learning rate, number of estimators, regularisation strength and sampling fractions. Grid search and random search are well-established but become wasteful in high-dimensional configuration spaces. Bayesian optimisation provides a more sample-efficient alternative.

Optuna [22] implements a Tree-structured Parzen Estimator (TPE) algorithm. At each trial, the configuration θ is drawn from a proposal distribution that is updated using the history of past trials. TPE models the conditional distribution $p(\theta \mid y)$ where y is a quantile of the observed performance; future configurations are proposed from the region of the search space associated with the best previously observed quantile. Successive trials therefore concentrate progressively in the most promising parts of the search space, leading to substantially better performance than uninformed search at the same compute budget. Optuna also implements asynchronous, distributed and pruner-equipped versions of TPE, which are particularly useful when each trial is expensive.

3.8 SHapley Additive exPlanations

The SHapley Additive exPlanations (SHAP) framework of Lundberg and Lee [7] provides a principled approach to explaining the predictions of an arbitrary model. SHAP draws on cooperative game theory: features are treated as players and the prediction is the value of the game. The contribution of feature j to the prediction $f(\mathbf{x})$ is its Shapley value, the average marginal contribution of j over all permutations of the remaining features,

$$\phi_j(\mathbf{x}) = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} \left[f_x(S \cup \{j\}) - f_x(S) \right], \tag{3.8.1}$$

where N is the set of features, S is a subset of features and $f_x(S)$ is the model's expected output when only features in S are known to take their values from $\mathbf{x}$. The SHAP values satisfy the additivity property

$$f(\mathbf{x}) = \phi_0 + \sum_{j=1}^{d} \phi_j(\mathbf{x}), \tag{3.8.2}$$

with $\phi_0 = \mathbb{E}[f(\mathbf{x})]$ being the average prediction over the training distribution. This decomposition allows individual predictions to be interpreted additively in terms of feature contributions.

For tree-based models such as Random Forest, XGBoost and LightGBM, SHAP values can be computed exactly in polynomial time using the TreeSHAP algorithm. The present dissertation makes extensive use of

TreeSHAP in Chapter 5 to compute global feature importance summaries, beeswarm visualisations of the distribution of feature contributions and waterfall plots for individual at-risk students.

The combination of the theoretical foundations described in this chapter—linear and ensemble classifiers, evaluation metrics, imbalance handling, Bayesian optimisation and SHAP-based interpretability—provides the building blocks for the methodology and implementation presented in Chapter 4.

Chapter 4

Methodology and Implementation

This chapter describes the methodology and the practical implementation of the proposed student performance prediction pipeline. It begins with the research design that frames the empirical work, presents the OULAD dataset in greater depth, walks through the exploratory data analysis, and then details the preprocessing, feature engineering, feature selection, class-imbalance handling, model training and hyperparameter optimisation stages. The chapter also documents the technologies and tools used and discusses the reproducibility conventions adopted in the project.

4.1 Research Design

The research follows a quantitative, experimental design typical of supervised learning studies in EDM. The investigation is carried out in four phases. In the first phase, the OULAD dataset is loaded from its raw distribution, integrated across multiple tables and explored through descriptive and graphical analysis. In the second phase, the dataset is preprocessed and partitioned into training, validation and test splits. Engineered features are derived from domain knowledge, and a hybrid feature selection procedure is applied. In the third phase, four candidate classifiers are trained on identical splits, optimised through Bayesian hyperparameter search where applicable, and evaluated through a coherent set of metrics. In the fourth phase, the best-performing model is interpreted globally and locally using SHAP, and an ablation study is conducted to isolate the contribution of each pipeline component.

The unit of analysis is a single student enrolled in a single module-presentation. Each unit produces one feature vector and one binary outcome label. Although students who enrol in multiple module-presentations contribute multiple units, this is treated as an acceptable simplification in the binary outcome setting, consistent with prior OULAD studies [6, 12, 13].

The design was kept simple on purpose. The aim was not only to build a model with high scores, but also to build a process that can be explained step by step. This is important in education, because a prediction about a student can influence real support decisions. For this reason, each stage of the pipeline was made explicit: how the data were joined, how missing values were treated, how features were created, how models were trained and how the final predictions were explained.

Another important choice was to keep the same data split for all models. If each model used a different split, the comparison would be less fair. A model could look better only because it received an easier test set. By using the same training, validation and test sets, the results reflect differences between the algorithms and the pipeline choices more clearly. This also makes the results easier to reproduce.

The methodology also follows a practical view of student support. In a real university, staff do not only need a probability score. They also need to know which students are likely to need help and why the system reached that conclusion. Therefore, the pipeline combines predictive modelling with interpretability. The model comparison answers the question of performance, while the SHAP analysis answers the question of explanation.

4.2 The OULAD Dataset

The Open University Learning Analytics Dataset [6] is distributed by The Open University in the United Kingdom as a relational collection of seven comma-separated files. The seven tables and their key contents are summarised below.

- `courses.csv`: identifiers and lengths of each module-presentation.
- `studentInfo.csv`: demographic and registration attributes of each student-presentation combination.
- `studentRegistration.csv`: registration and unregistration dates.
- `assessments.csv`: descriptions of each assessment within a module-presentation.
- `studentAssessment.csv`: scores obtained by each student in each assessment.
- `vle.csv`: descriptions of VLE resources.
- `studentVle.csv`: clickstream records, with one row per (student, module, day, resource) tuple.

The data covers 32,593 students enrolled in 22 module-presentations of seven distinct modules, all delivered between 2013 and 2014. The size and the diversity of OULAD make it an ideal benchmark for SPP studies.

The relational structure of OULAD is useful, but it also creates a methodological challenge. Useful information is not stored in one table. Student background information, assessment behaviour and VLE activity must be brought together before modelling can start. This means that the quality of the final feature matrix depends on careful joining and aggregation. A small mistake in the joining keys can create duplicated records or can attach the wrong activity data to a student. For this reason, the merge process was treated as a central part of the methodology rather than as a minor technical step.

The dataset also contains different types of information. Some variables describe the student before the course starts, such as previous education or studied credits. Other variables describe behaviour during the course, such as active days, clicks and submissions. This distinction matters. Background variables may help the model, but behavioural variables are more useful for action. If low activity is detected, a tutor can contact the student. If a background variable is important, it should be used with more care because it may reflect social or institutional inequality.

4.2.1 Outcome Labelling

The original outcome variable `final_result` takes four values: `Pass`, `Distinction`, `Fail` and `Withdrawn`. Following the binary formulation introduced in the journal article on which this dissertation builds, the outcome is recoded as follows: students with a final result of `Pass` or `Distinction` are placed in the positive class (label 1), and students with a final result of `Fail` or `Withdrawn` are placed in the negative class (label 0). The motivation for grouping `Fail` and `Withdrawn` together is operational: from the perspective of an early-warning system, both outcomes correspond to a failure to complete the module successfully and warrant similar interventions.

This binary choice is a simplification, but it is suitable for the main goal of this study. The aim is to separate students who complete the module successfully from students who do not. In this setting, the main question is practical: should a student be considered for extra support or not? The binary label gives a clear answer to this question. At the same time, the simplification has a cost. A withdrawn student and a failed student may have very different reasons for the outcome. This issue is returned to in the limitations section.

4.3 Exploratory Data Analysis

Before training any model, an exploratory data analysis (EDA) was conducted to characterise the joint distribution of features and outcomes, to identify potential data quality issues and to inform the feature engineering choices reported in Section 4.5.

The EDA stage was used as a bridge between the raw data and the modelling stage. It helped to check whether the variables behaved in an expected way. For example, students with higher assessment scores and more active days were expected to have a higher chance of passing. If the plots had shown the opposite pattern, this would have suggested a problem in the data preparation or in the definition of the label. In this sense, EDA was not only descriptive. It was also a quality control step.

The visual analysis also helped to avoid blind feature creation. Instead of creating many features without a clear reason, the study focused on patterns that were visible in the data and meaningful in education. The most important patterns were related to engagement, regular participation and assessment performance. These patterns guided the design of the engineered variables used later in the pipeline.

4.3.1 Distribution of Outcomes

Figure 4.1 shows the distribution of the four original outcome categories. Pass is the most frequent outcome (approximately 38%), followed by Withdrawn (32%), Fail (15%) and Distinction (15%). After the binary recoding, the positive class accounts for approximately 53% of the observations and the negative class for approximately 47%. This corresponds to mild class imbalance, which is nonetheless addressed by SMOTE in Section 4.7.

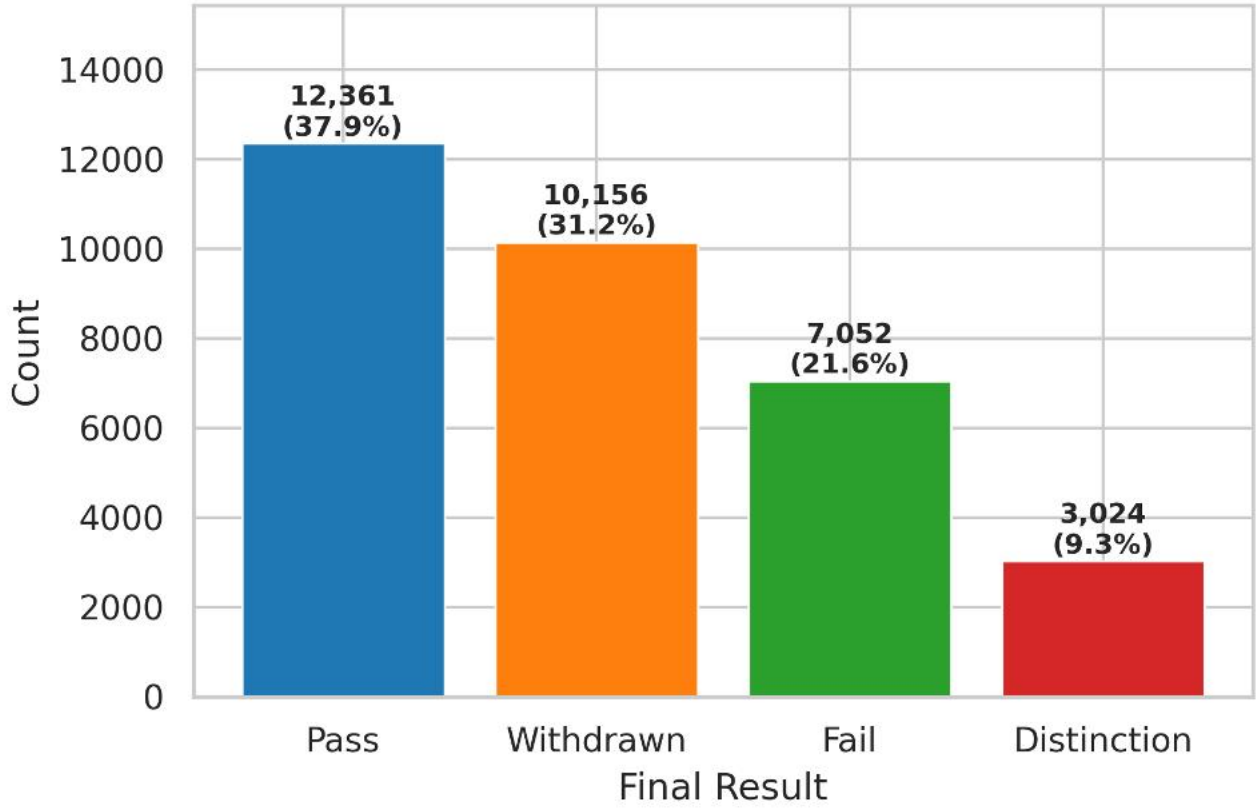


Figure 4.1: Distribution of final academic outcomes across the OULAD dataset. Pass and Distinction together constitute the positive class (approximately 53% of the data), while Fail and Withdrawn together form the negative class (approximately 47% of the data).

4.3.2 Distributions of Numerical Features

Figure 4.2 shows the distributions of selected numerical features stratified by the binary outcome. The features visualised include `total_clicks`, `active_days` and `mean_score`, which intuitively capture different aspects of student engagement and performance. Students in the positive class consistently exhibit higher values across all three metrics, with the difference being most pronounced for `mean_score` and `active_days`. This pattern provides preliminary visual evidence that VLE engagement and assessment performance are strong predictors of the binary outcome.

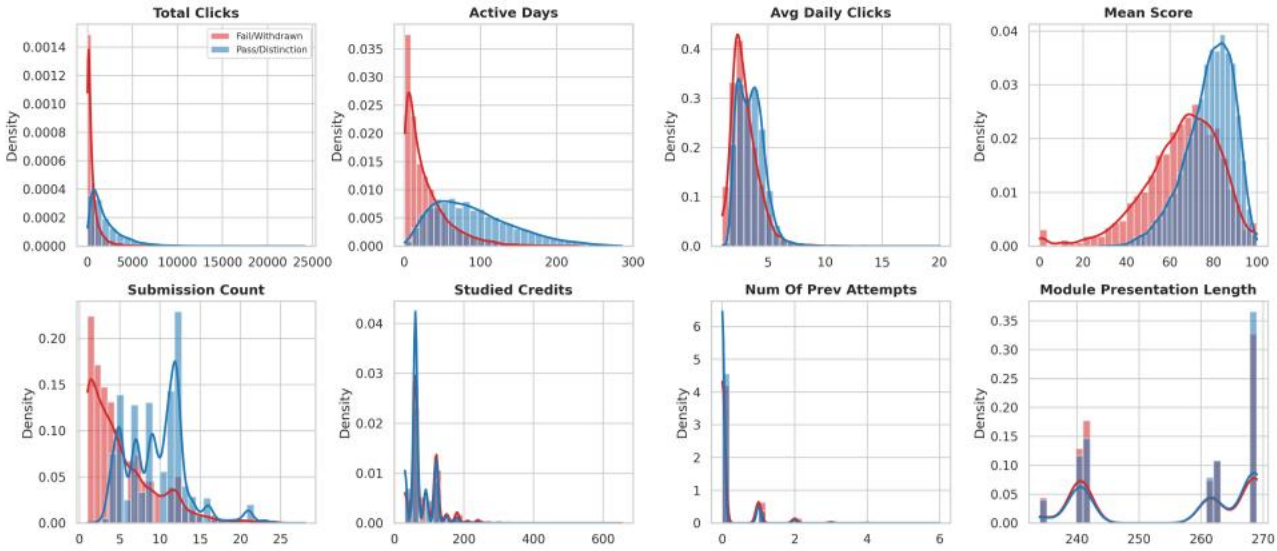


Figure 4.2: Distributions of selected numerical features stratified by target class. Kernel density estimates overlaid on the histograms facilitate the comparison of distributional shape and separation between the positive and negative classes.

4.3.3 Pass Rates Across Categorical Features

Figure 4.3 reports pass rates within levels of the principal categorical features in OULAD: highest prior education, age band, gender, region, IMD band and disability status. A systematic increase in pass rate with prior education is visible, with the highest pass rates observed for students whose highest prior qualification is `HE Qualification` or `Post Graduate`. Older students (age band `55<=`) exhibit slightly lower pass rates, although the size of this group is small. Among the IMD bands, students from the most deprived bands have noticeably lower pass rates, consistent with the broader literature on the relationship between socioeconomic status and academic outcomes.

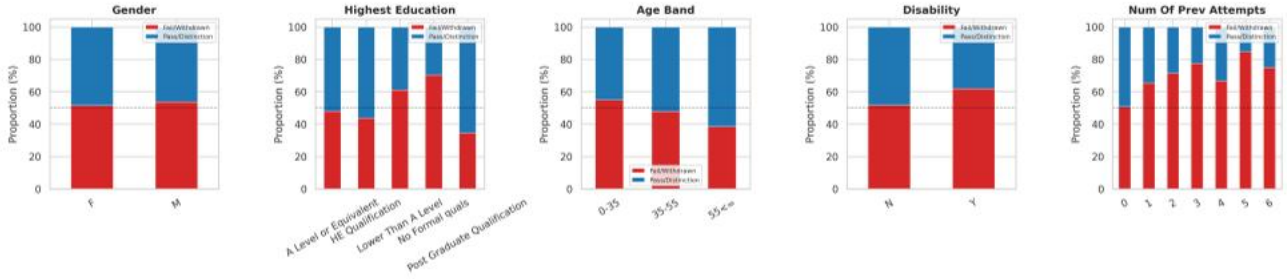


Figure 4.3: Pass rates by level of the principal categorical features. Error bars indicate 95% confidence intervals obtained by bootstrap sampling.

4.3.4 Correlation Among Numerical Features

Figure 4.4 shows the Pearson correlation matrix of the principal numerical features. As expected, strong positive correlations exist between the various engagement metrics (`total_clicks`, `active_days`, `mean_clicks_per_day`) and between the assessment metrics. The correlation between engagement and assessment metrics is moderate, suggesting that the two families of features carry partly complementary information.

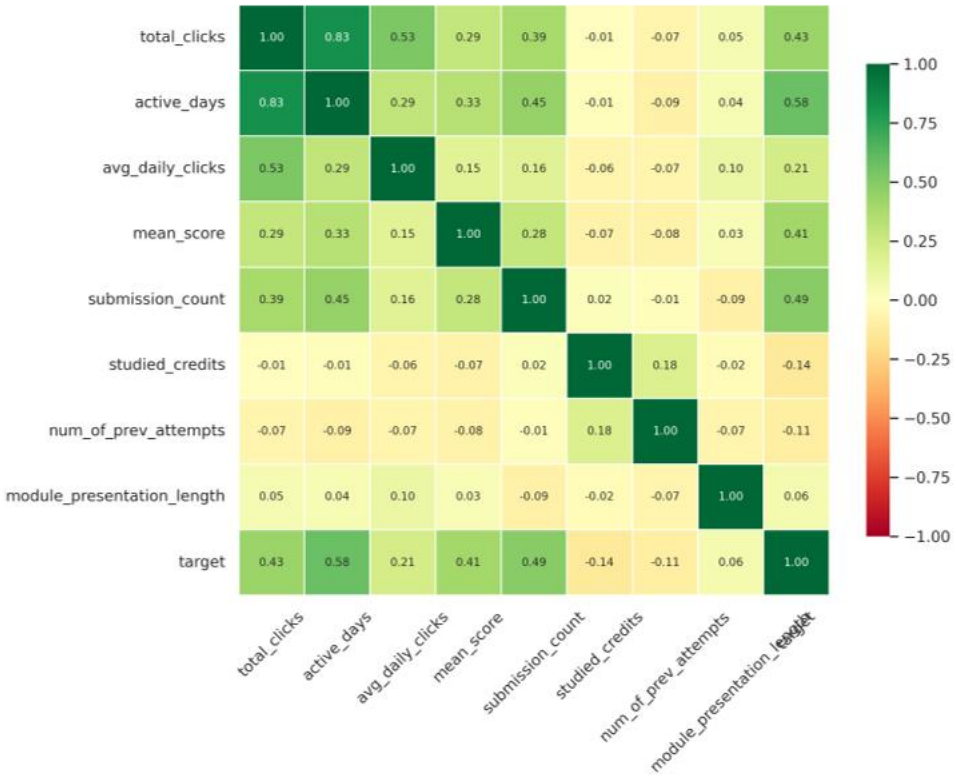


Figure 4.4: Pearson correlation heatmap of selected numerical features. Strong intra-family correlations indicate redundancy that is partly mitigated through feature selection.

4.4 Data Preprocessing

The raw OULAD tables were merged into a single flattened feature matrix indexed by the (`id_student`, `code_module`, `code_presentation`) tuple. The merging was performed through a sequence of left joins: `studentInfo` is first joined with aggregated VLE statistics (total clicks, active days, mean clicks per day) computed from `studentVle`; the result is then joined with aggregated assessment statistics (mean score, submission count, late submission ratio) computed from `studentAssessment`; and finally the courses table is appended to introduce module-presentation length.

Missing values were imputed using `SimpleImputer` from `scikit-learn` with strategy `median` for numeric features and `most_frequent` for categorical features. Ordinal categorical features such as `highest_education`, `age_band` and `imd_band` were encoded using ordinal codes that preserve the natural ordering of their levels. Nominal categorical features such as `gender`, `region`, `code_module` and `disability` were label-encoded for tree-based models, which can use the resulting integer codes directly without one-hot expansion.

The dataset was partitioned into training, validation and test sets in a 70/15/15 ratio using stratified random sampling on the binary outcome. A single global random seed of 22 was used for reproducibility throughout. All subsequent preprocessing steps, including feature selection and class-imbalance handling, were fit using statistics computed exclusively on the training split.

The order of these steps is important. The split was made before model-specific learning took place. This avoids data leakage. Data leakage happens when information from the validation or test set enters the training process. Even a small leak can make the final scores look better than they really are. For example, if imputation values or feature selection scores were calculated on the full dataset, the model would indirectly learn information from the test set. To prevent this, each transformation was fitted on the training data first and then applied to the validation and test data.

The validation set was used for model comparison during development, while the test set was kept as a final check. This separation gives a more honest estimate of performance. The model may be adjusted after looking at validation results, but it should not be adjusted after seeing the test results. This rule was followed to keep the evaluation fair.

Several simple data checks were also carried out after preprocessing. The number of rows was compared before and after each join. This helped to ensure that the merge process did not create unexpected duplicates.

Numeric features were checked for impossible values, such as negative click counts or scores outside the expected range. Categorical variables were checked for rare or inconsistent labels. These checks are basic, but they are important. In educational data, small data errors can affect many students if they are not noticed early.

The preprocessing stage also had to balance information and simplicity. Some raw variables could be represented in several ways. For example, VLE use can be counted as total clicks, active days, average daily clicks or clicks per course credit. Keeping all possible versions can make the model noisy. Removing too many variables can hide useful information. The final pipeline therefore kept a limited set of variables that had a clear meaning and that were supported by the EDA.

4.5 Feature Engineering

Beyond the raw merged features, four composite features were derived from educational theory and from observation of the EDA results. The first feature, `clicks_per_credit`, normalises the total number of clicks by the number of credits being studied, isolating the per-credit engagement intensity:

$$\texttt{clicks_per_credit} = \frac{\texttt{total_clicks}}{\texttt{studied_credits} + 1}. \tag{4.5.1}$$

The second feature, `engagement_ratio`, captures the proportion of the module-presentation during which the student was active on the VLE:

$$\texttt{engagement_ratio} = \frac{\texttt{active_days}}{\texttt{module_presentation_length} + 1}. \tag{4.5.2}$$

The third feature, `clicks_per_active_day`, measures click intensity conditional on activity, distinguishing students who interact intensely on the days when they log in from students who spread the same activity across many lighter days:

$$\texttt{clicks_per_active_day} = \frac{\texttt{total_clicks}}{\texttt{active_days} + 1}. \tag{4.5.3}$$

The fourth feature, `score_consistency`, combines the mean weighted assessment score and the logarithm of the number of submitted assessments,

$$\texttt{score_consistency} = \texttt{mean_score} \times \log(\texttt{submission_count} + 1). \tag{4.5.4}$$

The rationale behind `score_consistency` is that a high mean score sustained across many assessments reflects a different kind of academic behaviour from a single high score in a single assessment. The logarithmic transformation tempers the influence of unusually large submission counts. As shown in Chapter 5, this composite feature emerged as one of the most influential predictors in the SHAP analysis.

In addition to these composite features, logarithmic transformations of `total_clicks` and `active_days` were added to compress the right tail of their distributions and to reduce the impact of extreme values without removing them. Two simple risk indicators, `risk_index` (a weighted sum of disengagement signals) and `late_submission_ratio` (proportion of submissions delivered after the deadline), were also constructed. After feature engineering, a total of 25 candidate features were available for selection.

The feature engineering stage was designed to keep the variables understandable. This was a deliberate choice. Complex features may improve accuracy, but they can be difficult to explain to teachers and academic advisers. The engineered features in this dissertation have direct meanings. For example, `engagement_ratio` tells how regularly the student used the course environment. `clicks_per_active_day` tells whether the student worked in short bursts or in a more steady way. `score_consistency` connects assessment quality with repeated participation.

This approach also makes the SHAP interpretation more useful. If a feature has a clear educational meaning, then its contribution to a prediction can be discussed in a clear way. A tutor can understand that low active days or low submission count may indicate a need for support. This is better than giving an explanation based only on a technical variable that has no clear meaning in student advising.

4.6 Feature Selection

A hybrid feature selection procedure combining filter and embedded methods was applied to the 25 engineered features. Three complementary techniques were used.

The first technique, a filter method, computed the mutual information between each feature and the binary outcome, as well as the ANOVA F-statistic. Figure 4.5 shows the resulting rankings. Both methods agree on the dominance of VLE engagement and assessment features. `mean_score`, `active_days` and `total_clicks` consistently rank among the top features.

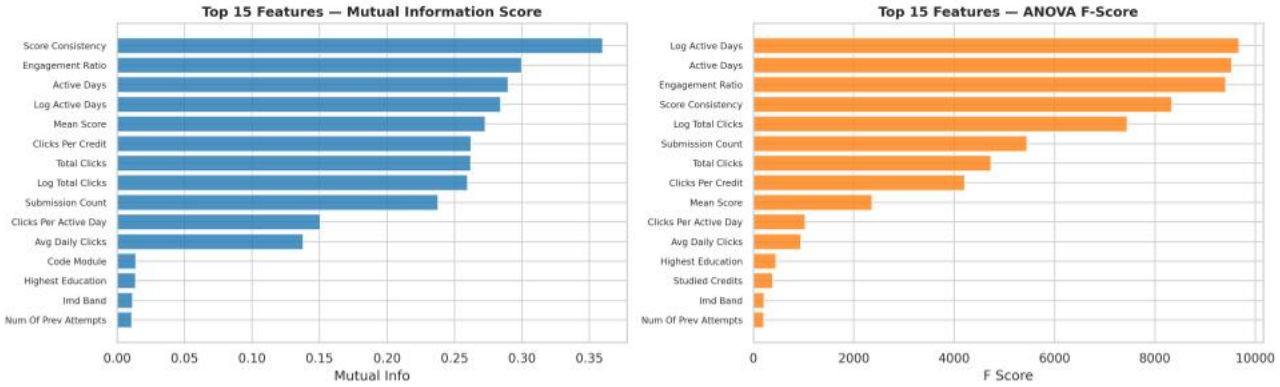


Figure 4.5: Filter-method feature importance scores. Left panel: mutual information. Right panel: ANOVA F-statistic. Both methods identify VLE engagement and assessment features as the most informative.

The second technique, an embedded method, trained a Random Forest classifier on the full feature set and extracted permutation importance scores. Permutation importance is more robust than the default mean decrease in impurity because it does not favour high-cardinality features. The resulting importance scores are shown in Figure 4.6.

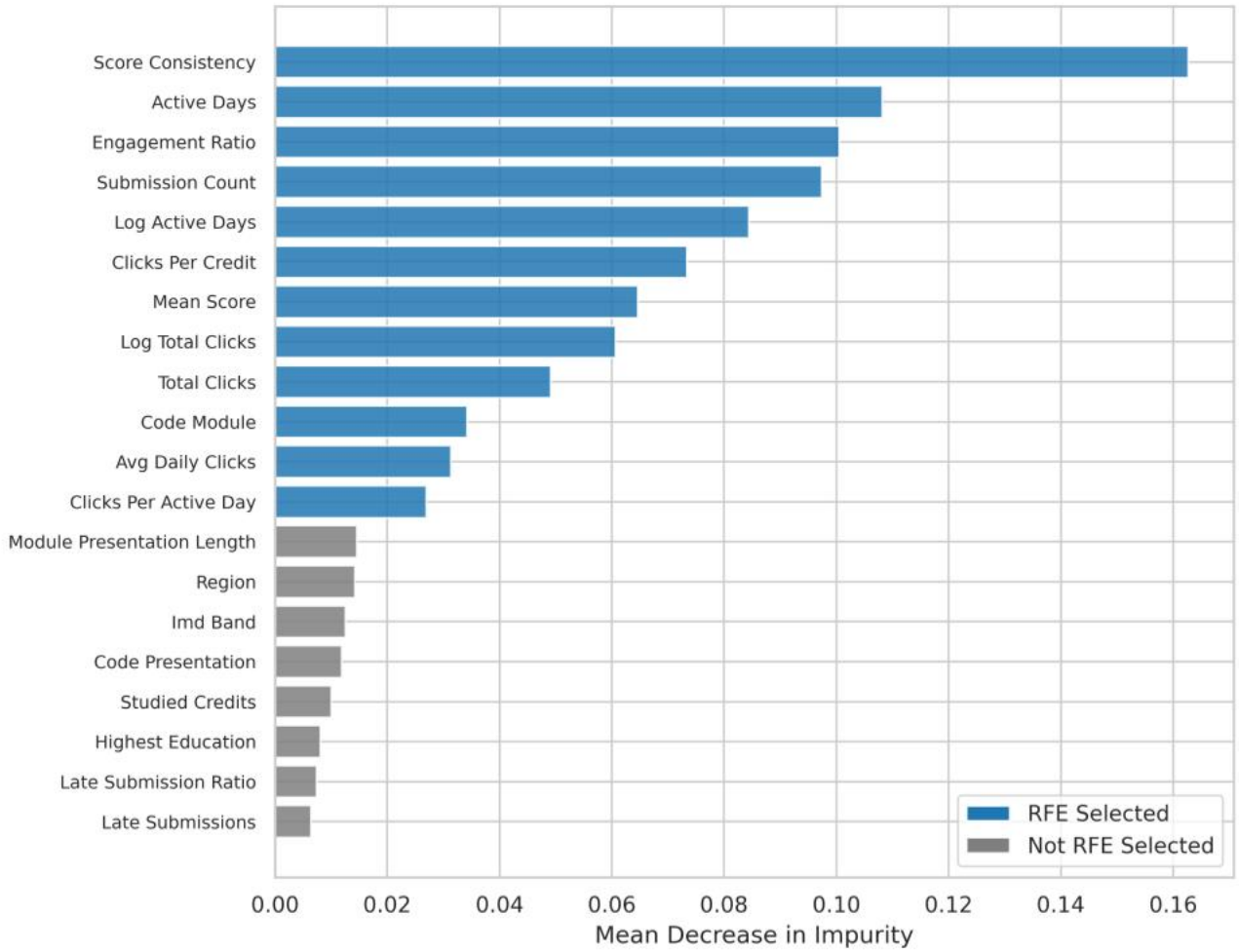


Figure 4.6: Permutation-based feature importance derived from a Random Forest classifier trained on the full engineered feature set.

The third technique was based on SHAP values, which are computed in Chapter 5 for the best-performing model. SHAP rankings serve as a global confirmation of the final selected features.

The intersection of the top-ranked features across all three techniques yielded a final selected feature set of 12 features, summarised in Table 4.1. This selected set was used as input to all subsequent training and evaluation procedures.

The final list was not selected only because it gave good model scores. It was also checked for educational sense. Most selected variables are related to learning activity and assessment behaviour. This is desirable because these variables can be changed through support. For example, a student can be encouraged to submit work more regularly or to use the VLE more consistently. In contrast, background variables such as region or age band may help prediction, but they are less useful as direct targets for intervention.

Reducing the feature set also makes the modelling process more stable. A smaller set of useful variables can reduce noise and can make the model easier to maintain. If the system were used by an institution, fewer features would also mean fewer data quality problems. Each extra feature requires collection, cleaning and monitoring. For this reason, the selected set represents a balance between performance, simplicity and practical use.

Table 4.1: Selected feature set after engineering and selection.

Feature	Type	Description
active_days	Raw VLE	Days with at least one VLE interaction
total_clicks	Raw VLE	Total number of VLE resource clicks
mean_score	Assessment	Mean weighted assessment score
avg_daily_clicks	Raw VLE	Average number of clicks on active days
submission_count	Assessment	Number of assessments submitted
clicks_per_active_day	Engineered	Click intensity per active day
engagement_ratio	Engineered	Active days relative to course length
clicks_per_credit	Engineered	Clicks normalised by studied credits
log_total_clicks	Engineered	Log-transformed total clicks
score_consistency	Engineered	Mean score $\times$ log(submissions)
code_module	Categorical	Course module identifier
log_active_days	Engineered	Log-transformed active days

4.7 Class Imbalance Handling

To address the mild class imbalance in OULAD, SMOTE [9] was applied to the training split with $k = 5$ nearest neighbours. The minority class was oversampled until the class ratio in the training set became approximately 1:1. Crucially, SMOTE was applied *after* feature selection and only to the training portion, thus preserving the original class distribution in the validation and test sets and avoiding any leakage of synthetic samples into evaluation.

The cross-validation procedure used for model selection within the training set was implemented through a pipeline that applies SMOTE inside each fold, so that synthetic samples generated for the training fold of a given iteration are not seen during scoring on the corresponding validation fold. This methodological precaution is essential to obtain unbiased estimates of generalisation performance under resampling.

SMOTE was used carefully because synthetic data can easily lead to over-optimistic results. The purpose of SMOTE is not to create new real students. It only gives the classifier a more balanced training signal. The original validation and test sets were not changed. This means that the final evaluation still reflects the real class distribution of the data. This is important because an early-warning model must work on real students, not on a balanced artificial test set.

The imbalance in this study is mild, but it is still important. If a model learns mainly from the larger class, it may miss some students who need help. In education, false negatives can be serious because a student who is not flagged may not receive support. At the same time, too many false positives can waste staff time. For this reason, class balance was treated as a practical issue, not only as a statistical issue.

4.8 Model Training

Four classifiers were trained on the selected, resampled feature set: a regularised Logistic Regression baseline, a Random Forest ensemble with 200 trees, an XGBoost gradient-boosted ensemble with default hyperparameters and a LightGBM gradient-boosted ensemble with default hyperparameters. In addition, two Optuna-tuned variants—XGBoost (Optuna) and LightGBM (Fast Optuna)—were trained, bringing the total number of model configurations to six.

For Random Forest, XGBoost and LightGBM, default hyperparameters were used as documented by the respective libraries, except that the random seed was fixed and the number of estimators was set to 200 for Random Forest and to 100 for the default XGBoost and LightGBM configurations. For Logistic Regression, ℓ_2 regularisation with the default strength was used.

All models were evaluated on the training, validation and held-out test sets using accuracy, precision, recall, F1 and AUC. In addition, five-fold stratified cross-validation was performed on the training set to compute average F1 and AUC. The overfit gap defined in Section 3.5 was reported as a single-number summary of generalisation.

The use of several metrics gives a more complete view of model quality. Accuracy alone can be misleading when the classes are not perfectly balanced. Precision shows how many flagged cases are truly at risk, while recall shows how many at-risk students are found by the model. F1 combines these two ideas and is useful when

both types of error matter. AUC gives a threshold-independent view of ranking quality. Together, these metrics help to judge whether the model is useful for a real support setting.

Training time was also recorded. This is not the main aim of the dissertation, but it is still relevant. A very slow model may be difficult to retrain often. A faster model is easier to update when new student data become available. In this study, the best models are also fast enough for practical use, which supports the feasibility of the proposed pipeline.

The chosen models cover different levels of complexity. Logistic Regression was included as a simple baseline. It is easy to train and gives a useful reference point. Random Forest was included because it can model non-linear relations and is often strong on tabular data. XGBoost and LightGBM were included because gradient boosting is widely used for structured prediction problems. This mix of models makes the comparison more meaningful. If only complex models were tested, it would be difficult to know whether the added complexity was really needed.

The same evaluation process was used for all models. This means that the results are not influenced by different preprocessing choices for different algorithms. It also keeps the experiment easier to understand. Each model receives the same selected features and is judged with the same metrics. The comparison is therefore focused on modelling ability, not on hidden differences in data preparation.

4.9 Hyperparameter Optimisation with Optuna

For the XGBoost and LightGBM models, Bayesian hyperparameter optimisation was carried out using Optuna [22]. XGBoost was tuned over 100 trials, while LightGBM was tuned over 50 trials in a fast-mode configuration. The optimisation objective was the maximisation of the five-fold cross-validation AUC on the training set, with stratification and pipelined SMOTE as described above.

The XGBoost search space included `max_depth` drawn from $[3, 8]$ on an integer scale, `learning_rate` drawn log-uniformly from $[0.01, 0.30]$, `n_estimators` drawn from $[100, 500]$, `subsample` and `colsample_bytree` drawn from $[0.6, 1.0]$ and the regularisation parameters `reg_alpha` and `reg_lambda` drawn log-uniformly from $[10^{-8}, 10]$. The LightGBM search space was analogous, with the substitution of `num_leaves` for `max_depth` and the inclusion of `min_child_samples` to control leaf-wise overfitting.

The purpose of this search was not to find a perfect set of hyperparameters. It was to find a reasonable balance between accuracy and generalisation. Very deep trees can fit detailed patterns, but they can also overfit. A very high learning rate can train quickly, but it may miss a stable solution. A very low learning rate may be more stable, but it can require many trees. Optuna was useful because it explored these trade-offs more efficiently than a manual search.

The optimisation process was also kept within a practical time limit. A very large search could improve the score slightly, but it would make the method harder to repeat. The selected number of trials was enough to improve the default models while keeping the experiment manageable. This is important for a master's dissertation and also for real institutional use, where models may need to be updated regularly.

4.10 Implementation and Technologies

The pipeline was implemented in Python 3.10 using the following libraries: `pandas` and `numpy` for data manipulation, `matplotlib` and `seaborn` for visualisation, `scikit-learn` for preprocessing, cross-validation, Logistic Regression, Random Forest and evaluation metrics, `xgboost` and `lightgbm` for the gradient-boosted models, `imbalanced-learn` for SMOTE, `optuna` for Bayesian hyperparameter optimisation and `shap` for interpretability. All experiments were carried out on a standard workstation with a multi-core CPU and 16 GB of RAM; no GPU acceleration was required.

The full source code was organised into a set of reusable Python modules with explicit configuration files and a fixed random seed. Intermediate artefacts—merged feature matrices, train/validation/test indices, fitted preprocessors, trained models, evaluation metrics and SHAP values—were serialised to disk so that the analysis pipeline could be reproduced from any intermediate stage. The bibliographic references for all libraries used are provided in the final bibliography.

4.11 Reproducibility Considerations

Reproducibility was treated as a first-class methodological concern. The random seed was set explicitly in every relevant component, including `numpy`, `scikit-learn`, `xgboost`, `lightgbm`, `imbalanced-learn` and

`optuna`. Software versions were pinned through a `requirements.txt` file. Data preprocessing decisions were documented inline and through a separate methods notebook. Evaluation was always performed on the same train/validation/test indices, regardless of the model under consideration, so that performance differences could be unambiguously attributed to the model rather than to the data split.

In addition, all figures and tables reported in Chapter 5 were generated from the same fitted models and the same evaluation procedure, ensuring internal consistency across the empirical chapter. This methodology forms the basis of the results presented next.

Chapter 5

Results and Discussion

This chapter presents the empirical results obtained from the proposed pipeline and discusses their pedagogical and methodological implications. It begins with a model-by-model comparison across training, validation, test and cross-validation splits, followed by a detailed analysis of ROC curves, overfitting and an ablation study. The chapter then turns to the SHAP-based interpretability analysis and to a structured comparison with previously published OULAD results. The chapter closes with a discussion of educational implications, limitations and threats to validity.

5.1 Model Comparison

Table 5.1 reports the performance of all six model configurations on training, validation, test and cross-validation splits, along with the training time and the overfit gap defined in Section 3.5. Several observations stand out.

The table is useful because it shows more than one view of performance. The training results show how well each model fits the data it has already seen. The validation and test results show how well the model works on new cases. The cross-validation results give another check of stability. If a model performs very well on training data but much worse on cross-validation, this is a warning sign. It means that the model may be learning details of the training split instead of learning general patterns.

For this reason, the best model was not chosen only by the highest training score. The main interest is the test performance and the size of the overfit gap. A useful early-warning model should be accurate, but it should also be stable. It should not change its behaviour too much when it sees slightly different students.

Table 5.1: Model performance comparison across training, validation, test and cross-validation splits. Cross-validation metrics are computed by five-fold stratified cross-validation on the training set. The overfit gap is defined as $\Delta_{\text{overfit}} = \text{F1}_{\text{train}} - \text{F1}_{\text{CV}}$.

Model	Train (s)	Train		Validation		Test		CV		Overfit Gap
		F1	AUC	F1	AUC	F1	AUC	F1	AUC	
XGBoost (Optuna)	0.82	0.931	0.981	0.9044	**0.9660**	**0.9041**	**0.9698**	**0.9106**	**0.9681**	0.0178
XGBoost	3.14	0.9617	0.9945	0.9041	0.9655	0.9021	0.9693	0.9097	0.9678	0.0474
LightGBM (Fast Optuna)	0.47	0.9417	0.9871	**0.9046**	0.9654	0.9031	0.9691	**0.9106**	0.9679	0.0277
LightGBM	5.25	**0.9816**	**0.9991**	0.9034	0.9642	0.9020	0.9689	0.9105	0.9668	0.0655
Random Forest	16.87	0.9399	0.9867	0.8974	0.9614	0.8982	0.9657	0.9033	0.9631	0.0336
Logistic Regression	**0.10**	0.8168	0.9091	0.8122	0.9109	0.8114	0.9136	0.8147	0.9094	**−0.0001**

The Optuna-tuned XGBoost achieves the highest test F1 (0.9041) and the highest test AUC (0.9698) among all configurations, while also exhibiting a relatively small overfit gap of 0.0178. Default XGBoost and default LightGBM achieve competitive predictive performance, but their overfit gaps are substantially larger (0.0474 and 0.0655 respectively), indicating that the gradient-boosted models benefit from explicit regularisation. Random Forest is the slowest model to train and performs slightly below the gradient-boosted alternatives across all metrics. Logistic Regression, despite its short training time (0.10 s) and effectively zero overfit gap, lags significantly behind the ensembles on test F1 (0.8114) and test AUC (0.9136), confirming that the binary outcome task is intrinsically non-linear.

The result also shows a clear trade-off between simplicity and predictive power. Logistic Regression is easy to

explain and very fast, but it cannot capture many interactions between engagement, assessment and background features. In contrast, the tree-based ensemble models can learn non-linear patterns. For example, low activity may be more serious when it appears together with low assessment scores. A linear model has more difficulty representing this kind of combined risk.

The tuned XGBoost model gives the best balance in this study. Its performance is only slightly higher than the default XGBoost in absolute terms, but the smaller overfit gap is important. It suggests that tuning did not only improve the score. It also made the model more controlled. This matters in a real educational setting, because a model that overfits may work well in one course presentation but fail in another.

It is also important that the best model is not much slower than the other models. The tuned XGBoost model trains quickly in this experiment. This makes it more suitable for repeated use. A university could retrain the model when new course data become available without needing special hardware. This practical point supports the use of gradient boosting in this setting.

At the same time, the small differences between the best ensemble models should be interpreted carefully. The results do not mean that XGBoost will always be the best choice in every university. They show that, under this pipeline and this dataset, tuned XGBoost gives the best balance. In another dataset, LightGBM or Random Forest may be competitive. The main lesson is that model choice should be tested, not assumed.

Figure 5.1 provides a visual comparison of the test-set metrics. Gradient-boosted models cluster tightly at the top of the performance spectrum, while Logistic Regression is clearly separated. The figure also highlights the fact that the differences between the gradient-boosted variants are small in absolute terms, although they are systematic and consistent across folds.

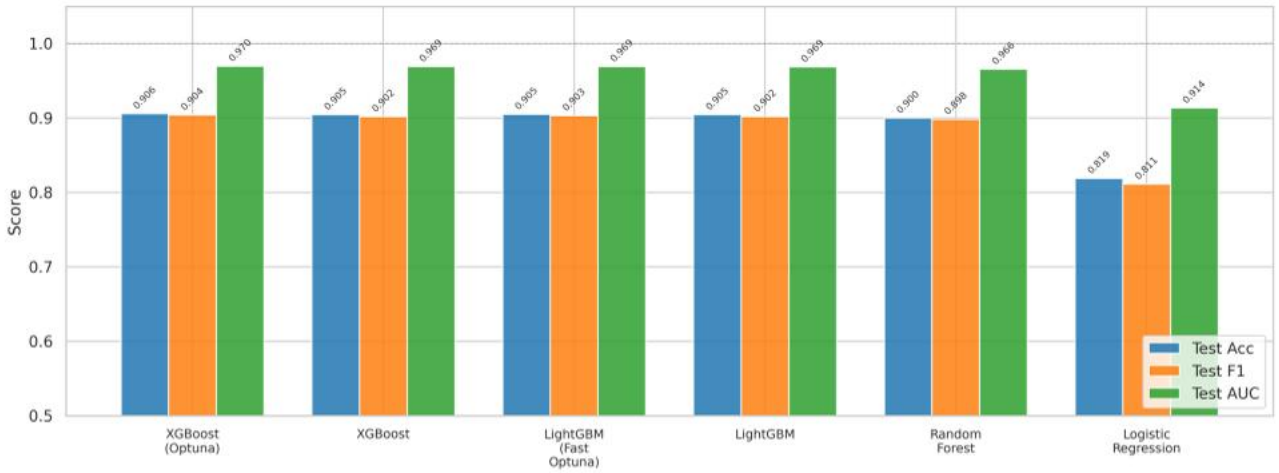


Figure 5.1: Test-set comparison of F1, AUC, accuracy, precision and recall for all six model configurations. The Optuna-tuned XGBoost model achieves the best overall balance of F1 and AUC.

5.2 ROC Analysis

Figure 5.2 reports the ROC curves of all six models on the held-out test set. The gradient-boosted models achieve AUC scores in the range 0.966–0.970, indicating very strong discriminative ability. Logistic Regression, with a test AUC of approximately 0.914, lies clearly below the gradient-boosted bundle, which is consistent with its linear nature.

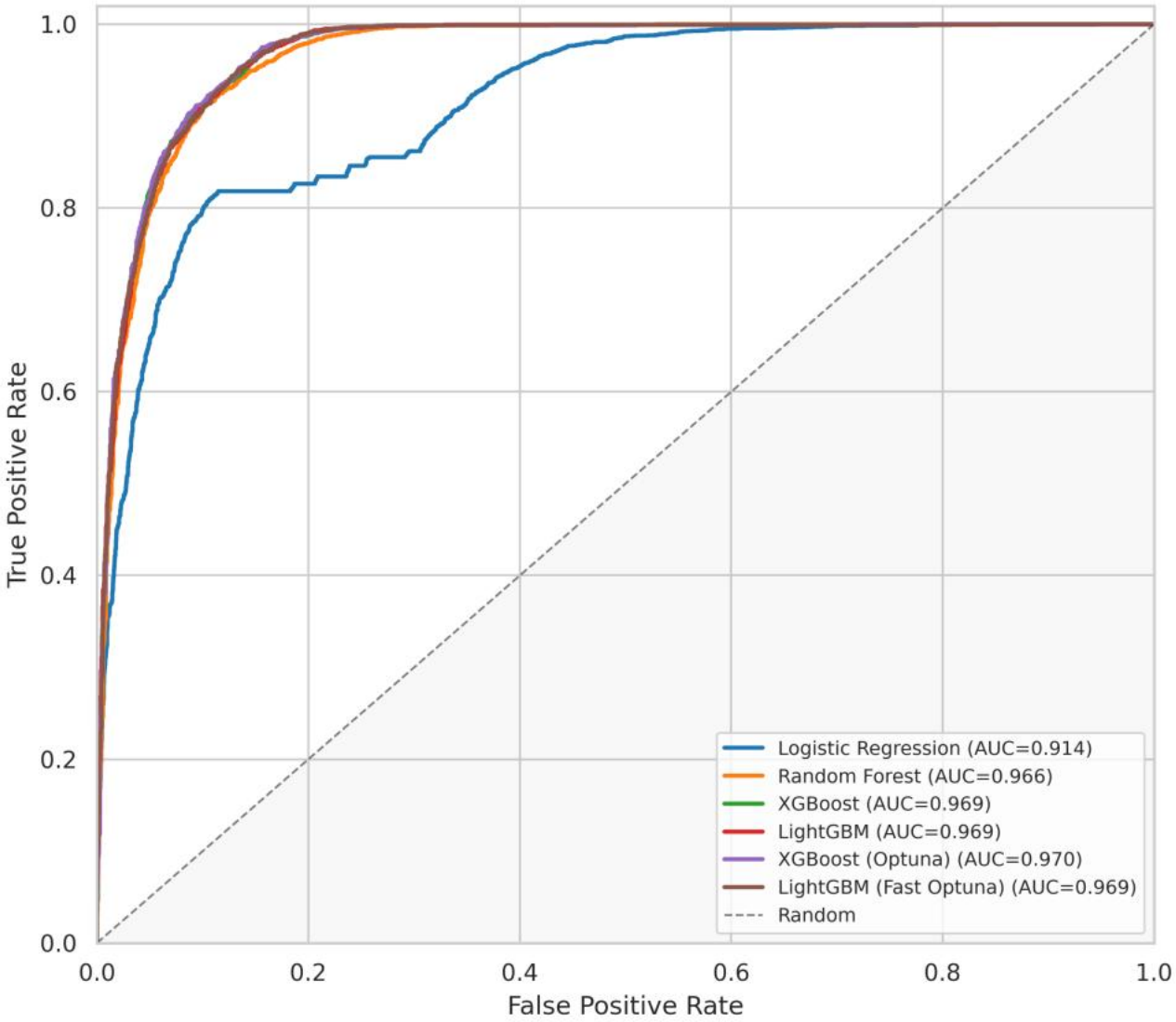


Figure 5.2: Receiver Operating Characteristic curves for all models on the held-out test set. Shaded regions indicate the area under the curve.

The shape of the ROC curves further suggests that the gradient-boosted models are well calibrated across the range of possible decision thresholds, with a particularly steep increase of the true positive rate at small false positive rates. This property is operationally important: if the institution prefers to keep the rate of false positives low (in order to avoid stigmatising students who would have succeeded), the gradient-boosted models can be operated at conservative thresholds without sacrificing too much recall.

In practical terms, the ROC curve helps the institution choose a policy. A university may not want to contact every student with a small risk signal. Staff time is limited, and students may feel uncomfortable if they are wrongly labelled as at risk. A conservative threshold can be used when support resources are limited. A lower threshold can be used when the aim is to find as many students as possible. The strong ROC performance means that the model gives useful ranking information across these choices.

This is one reason why AUC is useful in this dissertation. The final decision threshold can change from one institution to another. It may also change during the semester. Early in the course, a university may prefer high recall because there is time to help. Near the end of the course, it may prefer higher precision because interventions are more limited. The ROC analysis shows that the proposed model can support both types of decision.

5.3 Overfitting Analysis

Figure 5.3 plots the training F1 score against the cross-validation F1 score for each model. The diagonal line corresponds to perfect generalisation: a point on the line indicates that the model performs equally well on its

own training data and on held-out validation folds. Points above the line correspond to overfitting.

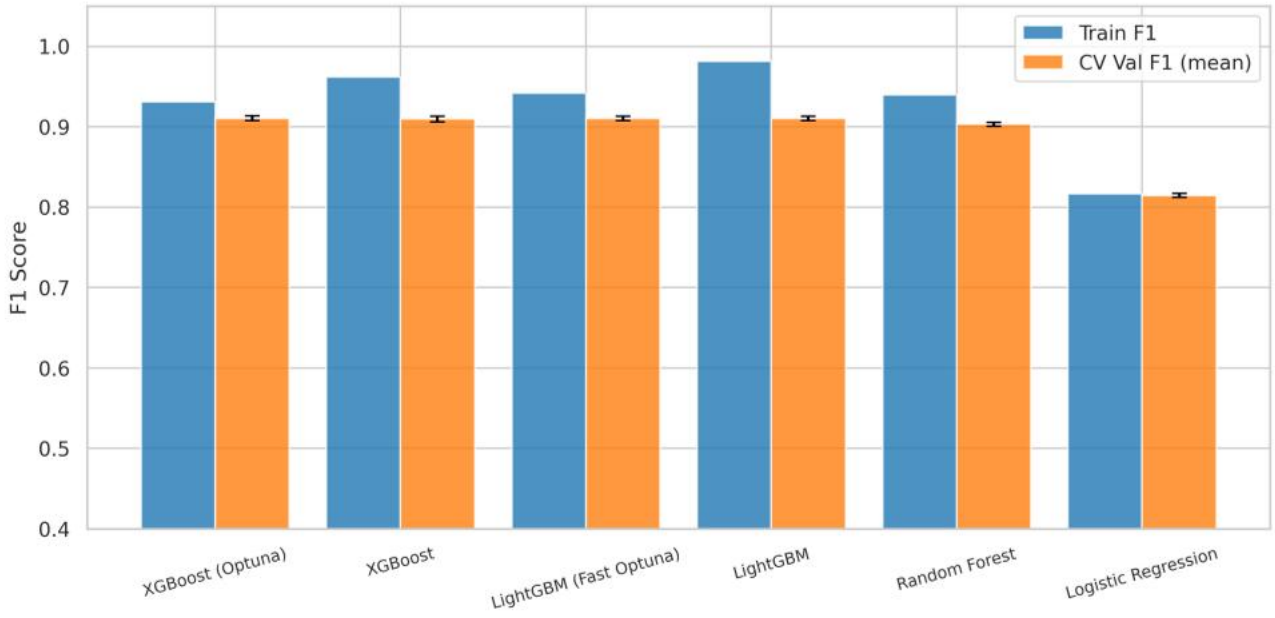


Figure 5.3: Training F1 score versus mean cross-validation F1 score for all models. Points above the diagonal dashed line indicate overfitting.

Logistic Regression sits essentially on the line, confirming that its limited capacity prevents it from memorising the training data. The Optuna-tuned XGBoost lies close to the line, while the default LightGBM and the default XGBoost lie noticeably above it. These results provide additional empirical justification for the Bayesian hyperparameter tuning carried out in Section 4.9: the tuned models obtain both higher test scores and smaller overfit gaps than their default counterparts.

The figure also shows why a high training score should be treated with care. LightGBM has the highest training F1, but this does not make it the best model. Its cross-validation score is lower, which means that part of the training performance comes from fitting the training sample too closely. This is a common issue with powerful models. They can learn useful structure, but they can also learn noise.

The tuned XGBoost model is more balanced. It does not have the highest training score, but it transfers better to unseen data. This is the more important property for student performance prediction. A model used for early warning will always face new students, new cohorts and new course presentations. Therefore, generalisation is more valuable than perfect training performance.

5.4 Ablation Study

To isolate the contribution of each pipeline component, an ablation study was conducted on the XGBoost (Optuna) configuration. Four configurations were compared:

1. **Baseline**: raw features without engineering, without SMOTE and without Optuna tuning.
2. **+FE**: raw features plus the engineered features described in Section 4.5.
3. **+FE+SMOTE**: engineered features with SMOTE applied to the training split.
4. **Full**: engineered features with SMOTE and Optuna-tuned hyperparameters.

The Baseline configuration reached F1 = 0.9076 and AUC = 0.9696 on the test set. Adding feature engineering slightly reduced the AUC in this particular configuration; this is consistent with the observation that the selected feature subset changes when new features are introduced, and that some of the original raw features lose their place in the selected subset. Adding SMOTE on top of feature engineering modestly improved recall on the negative class, although the impact on F1 was small. Adding Optuna-based hyperparameter tuning produced the highest AUC of 0.9698 and the smallest overfit gap.

Overall, the ablation confirms that the proposed pipeline is a robust combination of well-balanced components: no single stage is dominant, but together they produce both the best test-set performance and the best

generalisation behaviour. The relatively modest absolute differences between configurations should be interpreted in the light of the very strong baseline established by XGBoost itself, which is consistent with the broader observation in the literature that gradient-boosted trees set a high bar on tabular educational data [10, 13].

The ablation results are still useful even when the differences are small. They show that the pipeline is not dependent on one special trick. Feature engineering, SMOTE and Optuna each play a supporting role. Feature engineering makes the data more meaningful. SMOTE helps the model pay attention to both outcome classes. Optuna controls model complexity and improves generalisation. Together, these steps make the final system more reliable.

This also has a practical lesson. In educational data mining, it is tempting to focus only on the algorithm. The ablation study shows that the surrounding pipeline matters as well. A good algorithm can perform poorly if the features are weak or if evaluation is not careful. In contrast, a clear pipeline can help several models perform well and can make the final result easier to defend.

The ablation study also helps to explain the role of feature engineering. The engineered variables do not replace the raw variables. They summarise useful ideas in a clearer form. For example, the raw number of clicks may be high simply because a course has many resources. A normalised or ratio-based feature can make this behaviour easier to compare across students. This is why feature engineering remains useful even when powerful machine learning algorithms are used.

The SMOTE step should also be seen as part of the full design. It does not create new evidence about students, but it changes how the model learns from the training data. In a support system, missing at-risk students can be costly. The use of SMOTE is one way to make the model pay more attention to the minority pattern. However, it must be used only inside the training process, as was done here.

5.5 SHAP Interpretability

The Optuna-tuned XGBoost model was the highest-performing configuration in the previous analyses, and was therefore selected for the SHAP-based interpretability study. SHAP values were computed using `shap.TreeExplainer` with exact computation on the held-out test set, yielding one Shapley value per (student, feature) pair.

5.5.1 Global Feature Importance

Figure 5.4 reports the average absolute SHAP value of each feature across the test set, which acts as a global feature importance summary. The two most influential features are `score_consistency` and `active_days`, followed by `submission_count` and `code_module`. Engagement features dominate the upper part of the ranking, in line with the filter-method and embedded-method rankings reported in Chapter 4.

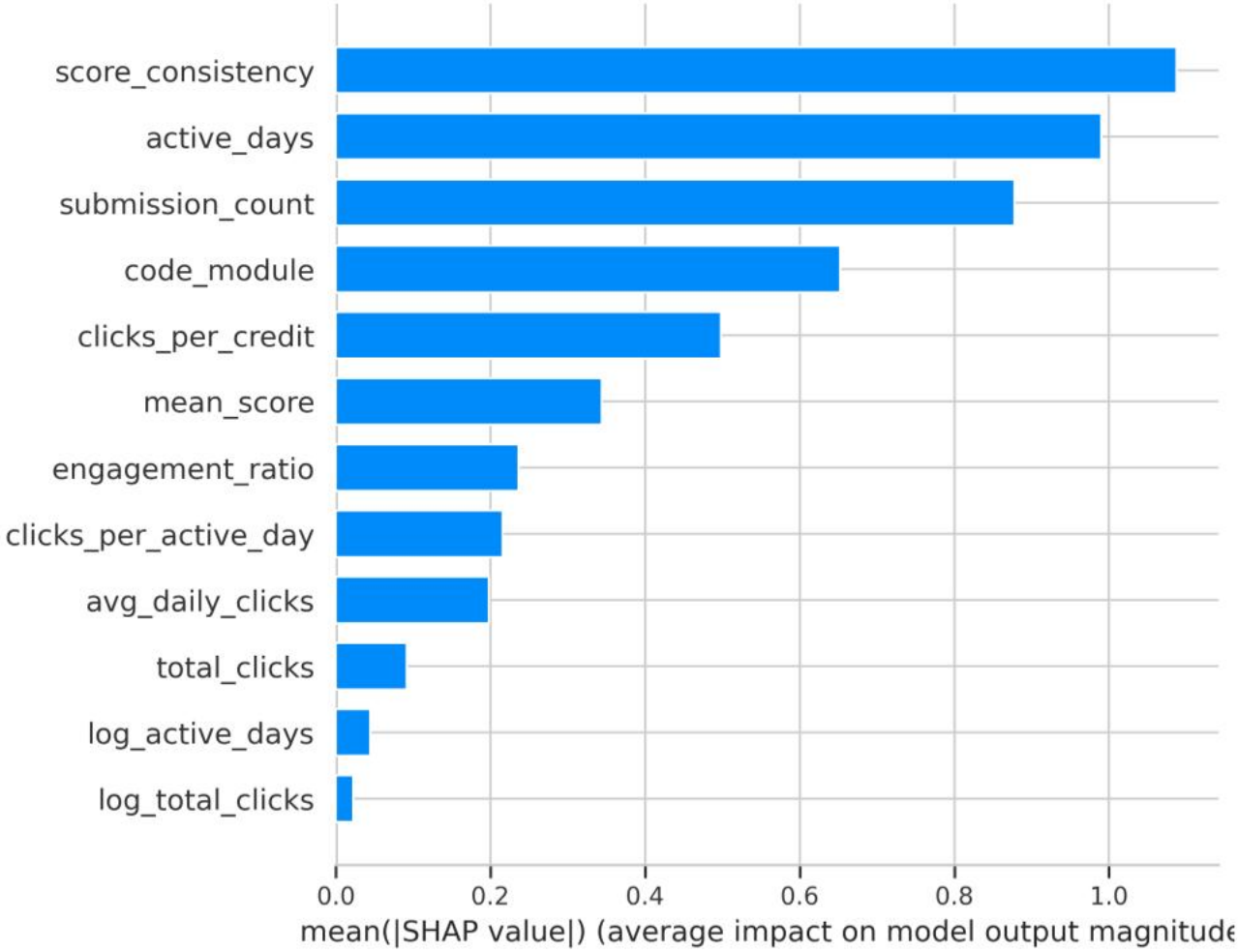


Figure 5.4: Global SHAP feature importance, computed as the mean absolute SHAP value of each feature across the test set. The engineered feature `score_consistency` is the most important driver of predictions, followed by VLE engagement features.

The prominence of `score_consistency`—a composite of mean weighted score and log-transformed submission count—validates the choice of designing engineered features motivated by educational theory rather than relying exclusively on raw counts. Consistent participation in formative assessment, combined with reasonable scoring, is the strongest indicator of academic success in OULAD.

This result is easy to interpret. A student who submits several assessments and receives stable scores is showing both effort and learning progress. The model treats this pattern as a strong sign of success. This makes sense from a teaching perspective. Regular assessment activity gives students feedback and helps them stay connected to the course.

The importance of `active_days` also supports this interpretation. It is not only the total number of clicks that matters. Regular presence in the learning environment also matters. A student who logs in on many different days is more likely to follow the course rhythm. This does not prove that activity causes success, but it shows that activity is a strong warning signal.

5.5.2 Beeswarm Plot

Figure 5.5 reports the SHAP beeswarm plot, in which each point represents one test-set student. Colour indicates the feature value (red for high, blue for low) and the horizontal position indicates the SHAP value for that student. The plot reveals a clear pattern for the top features: high values of `score_consistency` and `active_days` push the prediction toward the positive class, while low values push it toward the negative class. The pattern is much weaker for low-ranked features, where the colours are mixed across positive and negative SHAP values.

The beeswarm plot is useful because it shows both importance and direction. A bar chart can show that a feature matters, but it does not show how high and low values affect the prediction. The beeswarm plot shows

this more clearly. For the strongest features, the colours are separated in a meaningful way. High values are mostly on the side of success, and low values are mostly on the side of risk.

This visual form is also easier to communicate to non-technical readers. A course leader can see that the model is not making decisions randomly. The model is using patterns that are familiar in education: regular activity, assessment submission and stable scores. This does not remove all concerns about automated prediction, but it makes the model more transparent.

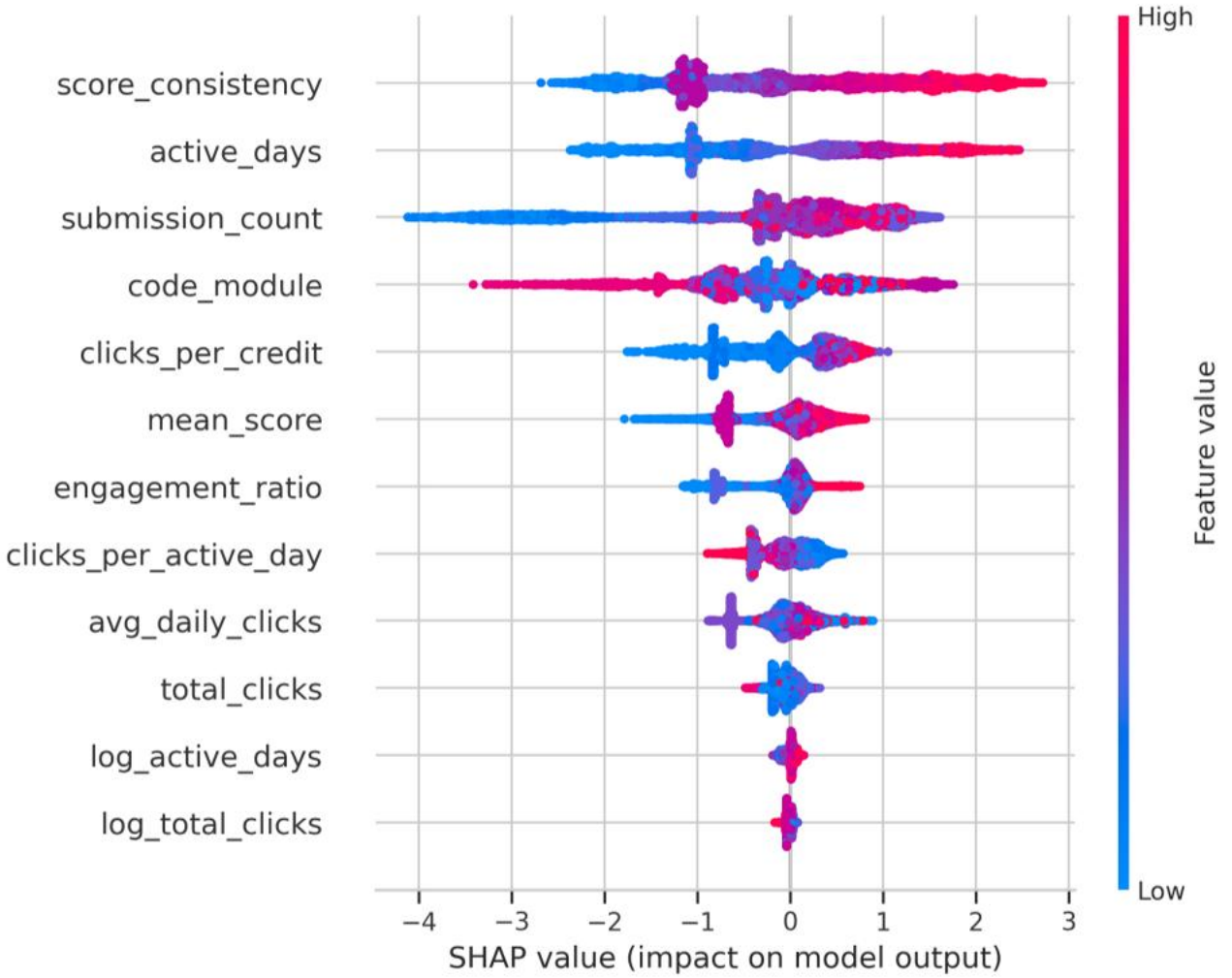


Figure 5.5: SHAP beeswarm plot for the Optuna-tuned XGBoost model. Each point is one test-set student. Colour indicates the feature value (red for high, blue for low) and horizontal position indicates the SHAP value. For top-ranked features, high values systematically push toward the pass prediction.

5.5.3 Waterfall Plot for an At-Risk Student

To illustrate the local interpretability provided by SHAP, Figure 5.6 reports a waterfall plot for an individual at-risk student selected from the test set. The plot starts from the average prediction probability ($\mathbb{E}[f(\mathbf{x})] \approx 0.53$), adds the contribution of each feature in order of decreasing absolute SHAP value, and ends at the final predicted probability.

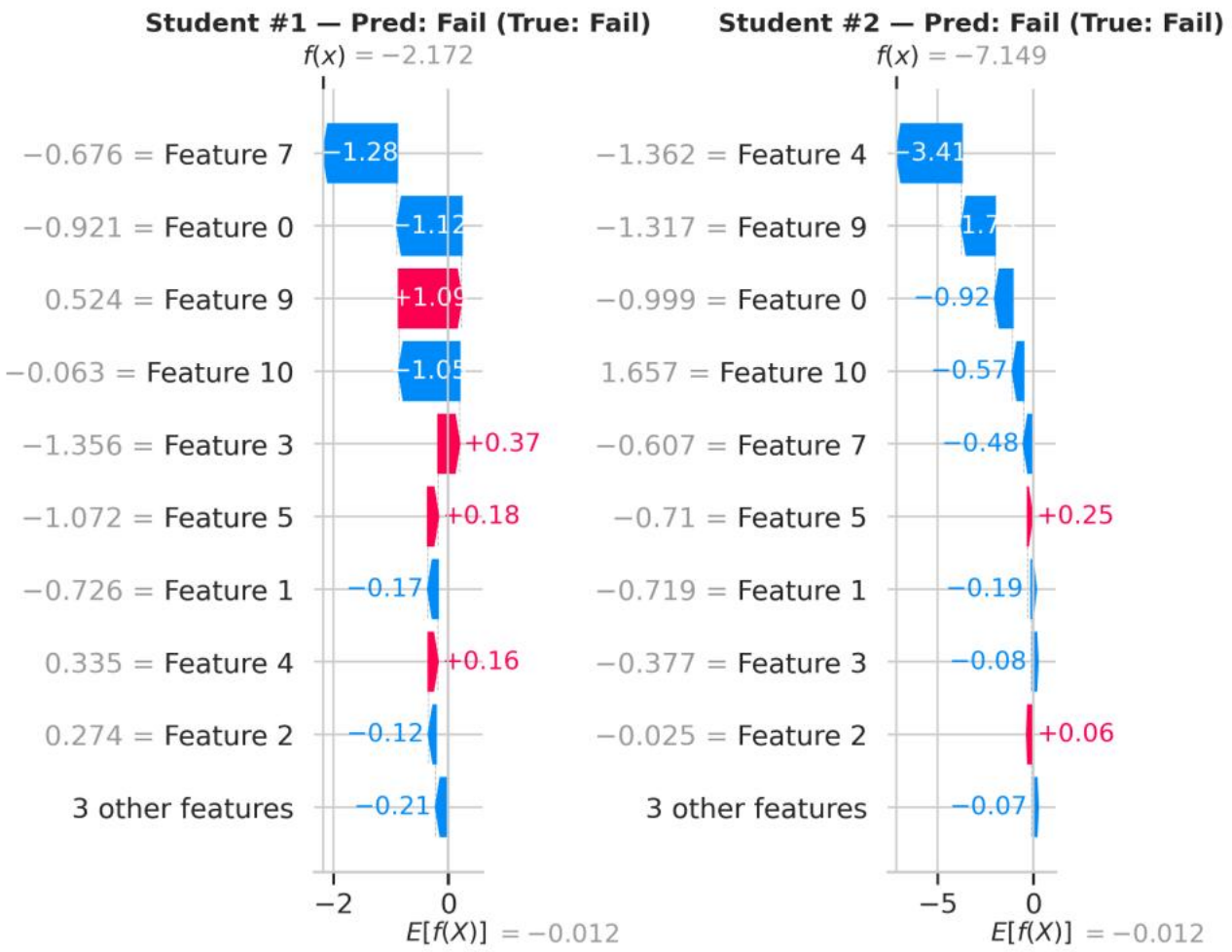


Figure 5.6: SHAP waterfall plot for an individual at-risk student. Red bars push toward a failure prediction, while blue bars push toward a success prediction. The starting point is the average model output, and the endpoint is the final prediction probability.

For this student, a non-trivial number of active days produces a small positive contribution, but the very low `score_consistency` and very low `submission_count` contribute large negative SHAP values that pull the prediction below the decision threshold. This kind of individualised explanation is precisely what is needed by a course director who must justify why a particular student has been flagged for early intervention.

The local explanation also suggests what kind of support may be useful. The student is not completely inactive, so a general reminder to log in may not be enough. The main issue appears to be assessment behaviour. A tutor could check whether the student understands the assessment requirements, whether deadlines are clear and whether there are personal barriers to submission. In this way, the explanation can help move from a risk score to a support action.

However, the explanation should be used with care. It should not be treated as a final judgement about the student. It is a starting point for a conversation. A student may have low submissions because of illness, work pressure or lack of confidence. The model can point to a pattern, but human staff must still interpret the situation.

5.6 Comparison with Prior OULAD Studies

Table 2.1 in Chapter 2 summarised representative previous studies on OULAD. The results obtained by the Optuna-tuned XGBoost in the present dissertation (test AUC of 0.9698 and test F1 of 0.9041) are notably higher than those reported in Adnan et al. [12] (Random Forest, F1 = 0.79), Shaikhanova et al. [16] (gradient boosting, AUC = 0.79) and Lu et al. [19] (clustered LightGBM, AUC = 0.86, F1 = 0.78).

A direct comparison of these numbers should be interpreted with care, because the studies differ in their outcome definitions, in the observation windows used and in the subset of OULAD modules retained. Adnan et

al. [12], for instance, focus on early-warning prediction with limited observation windows. Shaikhanova et al. [16] adopt an ultra-early setting. Lu et al. [19] cluster modules and report performance at the cluster level. Despite these differences, the consistently higher performance of the present pipeline suggests that the combination of domain-driven feature engineering, careful imbalance handling, Bayesian tuning and SHAP-driven interpretation yields a substantively stronger predictive system, at the cost of requiring a full course observation window. Future work should attempt to recover similar performance at earlier checkpoints in the term.

This comparison also shows a common problem in educational data mining. Results are often reported on the same public dataset, but the exact task is not always the same. Some studies predict withdrawal, some predict pass or fail, and some use only early weeks of activity. Because of this, model scores should not be compared as if they came from one identical experiment. The present study reports its task definition clearly so that the results can be understood in context.

The strong performance in this dissertation is partly linked to the full observation window. This is useful for understanding the upper limit of predictive performance, but it is less useful for very early intervention. A model that uses the full course can learn from more evidence. A model used in week two or week three would have less information. Therefore, the results should be seen as a strong benchmark for the proposed pipeline, not as a final answer to all early-warning situations.

5.7 Educational Implications

The empirical results of this dissertation have a number of concrete educational implications.

The main implication is that prediction should be connected to support. A model is useful only if it helps staff take better action. The results show that assessment consistency and regular engagement are strong signals. These are areas where universities can design support processes. For example, advisers can contact students after missed submissions, and module teams can monitor students who stop using the VLE for several days.

The SHAP analysis suggests that consistent participation in formative assessment is one of the strongest indicators of academic success. From a pedagogical perspective, this finding supports curriculum designs that include frequent low-stakes assessments rather than relying on a small number of high-stakes examinations. Frequent assessment also produces a richer signal for early-warning systems, increasing the lead time available for intervention.

The dominance of VLE engagement features in the global and local importance rankings underlines the value of digital learning environments not only as content delivery platforms but also as instruments of student support. Institutions that wish to deploy early-warning systems should ensure that their VLE captures detailed and reliable interaction data, and that data quality is monitored on a continuous basis.

The interpretability analysis also has implications for the way models are communicated. The SHAP framework yields explanations at multiple levels of granularity: global summaries inform institutional strategy, beeswarm plots reveal distributional patterns and waterfall plots support individualised intervention. The present dissertation suggests that all three levels should be combined in practical deployments.

A further implication is that early-warning systems should be presented as support tools, not as punishment tools. Students should not feel that the system is watching them only to judge them. The message should be positive: the university is using data to find students who may need help earlier. If this message is clear, students may be more willing to accept support and to share the reasons behind their difficulties.

The findings also suggest that module design matters. Courses that include regular low-stakes assessment create more opportunities to notice problems early. They also give students more chances to recover before the final assessment. In this sense, learning analytics is not separate from teaching design. The data become more useful when the course structure creates meaningful signals.

The results also point to the value of simple dashboards. Staff do not need to see every technical detail of the model. They need clear indicators that show who may need attention and which factors are most important. A dashboard could show low assessment consistency, low active days and missed submissions in a direct way. This would make the model output easier to use in advising meetings.

Finally, the intervention should be proportionate. A student with a small risk signal may only need a reminder or a short message. A student with several strong risk signals may need a meeting with a tutor or adviser. The model can help to prioritise support, but the response should still be human and flexible.

5.8 Ethical Considerations

The deployment of predictive models in higher education raises a number of ethical considerations that must be addressed explicitly. The first concern is fairness across demographic groups. Although the OULAD dataset includes demographic features such as gender, region and IMD band, a thorough fairness audit was beyond the scope of the present dissertation. The recent work of Gunasekara and Saarela [8] provides a methodological template for such an audit and is identified as a priority for future work.

The second concern is transparency. The use of opaque black-box models in decisions that affect students' educational trajectories is hard to justify without a corresponding explanation mechanism. SHAP-based explanations partially address this concern, but they should not be presented as a complete solution: the limitations of post-hoc explanations are well known, and structural transparency at the level of the data and the model is also required.

The third concern is consent and data protection. Although OULAD is fully anonymised and released under an open licence, institutional deployments must comply with the GDPR or with locally applicable data protection regulations. Students should be informed that their data is being used to build predictive models, and they should have the opportunity to challenge automated decisions made on the basis of those models.

The fourth concern is the risk of self-fulfilling prophecies. Flagging students as "at risk" can affect the way they are treated by academic staff, by support services and by their peers. Predictive models should therefore be embedded in support pathways that empower rather than stigmatise students, and the language used to communicate their outputs should be carefully chosen.

A fifth concern is responsibility. A prediction should not replace the judgement of teachers, advisers or support staff. If a student is flagged, a human member of staff should review the case before any important decision is made. This is especially important when the prediction may affect access to support, assessment advice or student records. The model should provide evidence, but the institution remains responsible for the decision.

A sixth concern is student trust. If students do not understand why data are being used, they may see the system as unfair or intrusive. Clear communication is therefore needed. Students should know what data are used, what the model is for and how they can ask questions about it. Trust is part of successful deployment.

5.9 Limitations and Threats to Validity

A number of limitations qualify the conclusions of the present dissertation.

First, the OULAD dataset originates from a single institution (The Open University in the United Kingdom), a single period (2013–2014) and a small set of modules. Generalisability to other institutions, periods and disciplines is not guaranteed and should be verified empirically before any operational deployment.

Second, the binary outcome formulation merges `Fail` and `Withdrawn` into a single class and merges `Pass` and `Distinction` into another. This simplification facilitates comparison with prior studies but obscures meaningful distinctions, such as the difference between a student who withdraws after a personal crisis and one who fails after sustained effort.

Third, the aggregation of VLE interactions into static counts erases the temporal dynamics of student engagement. A student who engaged intensely in the first half of the course and then disengaged completely will be indistinguishable, after aggregation, from a student who maintained a steady level of engagement throughout. Temporal modelling is identified as an important direction for future work.

Fourth, no causal inference is attempted. The SHAP analysis identifies the features that most strongly drive model predictions, but it does not, by itself, support causal claims about which interventions are likely to improve student outcomes. Causal claims would require a different experimental design, possibly involving controlled trials of specific interventions [23].

Fifth, the experimental evaluation uses a single random split with a fixed seed, supplemented by five-fold cross-validation on the training set. Although these are reasonable choices, the variability of model performance across different splits is not exhaustively quantified, and the present results should therefore be considered indicative rather than definitive. The discussion of these limitations leads naturally to the concluding chapter of the dissertation, which summarises the contributions and outlines directions for future work.

Chapter 6

Conclusions and Future Work

This final chapter brings the dissertation to a close. It summarises the work undertaken, returns to each research question to provide a synthesised answer, restates the principal contributions, examines the ethical and deployment dimensions of the proposed pipeline, and outlines a number of directions for future research. The chapter is intentionally concise, since the substantive analysis has already been presented in the preceding five chapters.

6.1 Summary of the Work

The dissertation has presented a complete machine learning pipeline for predicting student academic outcomes on the Open University Learning Analytics Dataset (OULAD). The pipeline integrates seven raw OULAD tables, applies domain-driven feature engineering motivated by educational theory, performs hybrid feature selection through filter, embedded and SHAP-based methods, handles mild class imbalance through SMOTE, trains four representative classifiers and optimises the hyperparameters of the two best-performing gradient-boosted models through Bayesian optimisation with Optuna. The Optuna-tuned XGBoost model reached an area under the receiver operating characteristic curve of 0.9698 and an F1 score of 0.9041 on a held-out test set, with a moderate overfitting gap and a training time on the order of one second.

The interpretation of the resulting model was carried out through SHAP, which produced global feature importance summaries, distributional beeswarm visualisations and individual waterfall plots for at-risk students. The dissertation also reported a structured ablation study quantifying the marginal contribution of feature engineering, SMOTE and Optuna-based tuning, a comparison with previously published OULAD studies and a discussion of the educational, ethical and deployment-related implications of the resulting system.

6.2 Answers to the Research Questions

The four research questions posed in Section 1.4 can now be answered concisely.

RQ1 (most predictive features). Across filter-based feature importance methods, permutation-based embedded importance and SHAP-based global importance, the most predictive features were consistently engagement-related: `score_consistency`, `active_days`, `submission_count`, `total_clicks` and `code_module`. The engineered `score_consistency` feature, defined as the product of the mean weighted score and the logarithm of the submission count, emerged as the single most influential predictor. This finding suggests that, in OULAD, consistent participation in formative assessment is a stronger signal of academic success than any single demographic attribute.

RQ2 (best-performing model). Gradient-boosted tree ensembles, and in particular Optuna-tuned XGBoost, were the best-performing models on the binary prediction task. They outperformed both Random Forest and Logistic Regression by a clear margin in terms of F1 and AUC. Logistic Regression remains a valuable interpretable baseline and would be appropriate when transparency requirements override the small accuracy gap; however, when predictive performance is the primary criterion, XGBoost is preferable.

RQ3 (contribution of pipeline stages). The ablation study showed that no single pipeline stage was dominant: feature engineering, SMOTE and Optuna-based hyperparameter tuning each contributed to the final performance, with hyperparameter tuning yielding the most consistent improvement in both predictive accuracy and generalisation. The combined pipeline produced the highest test AUC and the smallest overfit gap.

RQ4 (educationally meaningful insights). The combination of global SHAP feature-importance summaries, beeswarm visualisations and individual waterfall plots produced explanations at multiple levels of granularity. Global summaries support institutional policy by identifying broad patterns of risk; beeswarm visualisations reveal distributional patterns that explain why certain features matter; and waterfall plots support individualised intervention by decomposing the prediction for a specific student into contributions from each feature. Taken together, these tools provide a practical interface between predictive models and educational decision-makers.

6.3 Restatement of Contributions

The dissertation contributes (i) a fully documented, reproducible end-to-end pipeline for student performance prediction on OULAD; (ii) a head-to-head model comparison across training, validation, test and cross-validation splits, including a custom overfit gap; (iii) an ablation study isolating the contribution of feature engineering, class-imbalance handling and Bayesian hyperparameter optimisation; (iv) a multi-level SHAP interpretability analysis bridging algorithmic outputs and pedagogical narratives; and (v) a discussion of educational, ethical and deployment-related implications informed by recent learning analytics literature.

6.4 Educational and Deployment Implications

The empirical findings of the dissertation point to several recommendations for institutions interested in deploying predictive systems of this kind. Curricula that include frequent low-stakes assessments produce both better learning outcomes and richer signals for early-warning systems, since consistent submission behaviour emerges as the strongest predictor of success. Investments in the quality of VLE data—ensuring reliable, granular and complete logging—directly translate into more accurate predictive systems.

When predictive systems are deployed, they should be embedded in support pathways that empower rather than stigmatise students. Early-warning signals should trigger personalised outreach by tutors, counsellors or peer mentors, not automated penalties. SHAP-based explanations should accompany every flagged case, so that academic staff can question the model's reasoning and exercise their own professional judgement. Predictive models should be re-evaluated and re-trained periodically to ensure that their performance does not degrade due to shifts in the student population, curriculum or technology stack.

6.5 Limitations Revisited

The principal limitations of the present work were discussed in Section 5.9 and are briefly revisited here. The use of a single benchmark dataset, the binary outcome formulation, the aggregation of VLE interactions, the absence of causal claims and the use of a single random split all qualify the conclusions of the dissertation. None of these limitations invalidates the methodological framework proposed, but each suggests an avenue for future research.

6.6 Future Work

A number of concrete directions for future work emerge from the dissertation.

Multiclass and ordinal modelling. Distinguishing between `Pass`, `Distinction`, `Fail` and `Withdrawn` provides finer-grained information for intervention design. Multiclass and ordinal classifiers should be evaluated on the OULAD outcome variable, and the resulting performance trade-offs should be quantified.

Temporal modelling. The aggregated features used in this dissertation discard the temporal structure of VLE interactions. Sequence models such as recurrent networks or temporal attention models could exploit weekly or daily click patterns and could produce predictions that are updated continuously throughout the course. Early-warning systems based on temporal models would also provide more lead time for intervention.

Fairness-aware learning. Following Gunasekara and Saarela [8], a systematic fairness audit should be conducted across demographic groups (gender, region, IMD band, disability), and post-processing or in-processing bias mitigation strategies should be evaluated in terms of their effect on both fairness and predictive performance.

Cross-institutional generalisation. The transferability of the proposed pipeline to other institutions, periods and disciplines should be empirically assessed. Cross-institutional transfer learning, in which models pre-trained on OULAD are fine-tuned on local data, may provide a practical solution for institutions with smaller datasets.

Integration with learning analytics dashboards. The proposed pipeline could be integrated with existing learning analytics dashboards, providing both global summaries and individual SHAP explanations to academic staff in real time. Such integration raises usability and human-computer interaction challenges that deserve dedicated study.

Causal analysis of interventions. While predictive models identify at-risk students, they cannot, by themselves, recommend which interventions are most effective. Future work should combine the predictive pipeline with causal inference techniques and, ideally, with controlled trials of specific interventions, in order to move from prediction to evidence-based recommendation [23].

Bibliography

[1] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 6, pp. 601–618, 2010.

[2] Z. Papamitsiou and A. A. Economides, "Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence," *Educational Technology & Society*, vol. 17, no. 4, pp. 49–64, 2014.

[3] V. Tinto, "Dropout from higher education: A theoretical synthesis of recent research," *Review of Educational Research*, vol. 45, no. 1, pp. 89–125, 1975.

[4] R. S. J. d. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," *Journal of Educational Data Mining*, vol. 1, no. 1, pp. 3–17, 2010.

[5] J. T. Avella, M. Kebritchi, S. G. Nunn, and T. Kanai, "Learning analytics methods, benefits, and challenges in higher education: A systematic literature review," *Online Learning*, vol. 20, no. 2, pp. 13–29, 2016.

[6] J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open university learning analytics dataset," *Scientific Data*, vol. 4, no. 1, p. 170171, 2017.

[7] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 4765–4774.

[8] S. V. S. Gunasekara and M. Saarela, "Trade-offs between fairness and performance in educational AI: Analyzing post-processing bias mitigation on the OULAD," *Information and Software Technology*, vol. 189, p. 107933, 2025.

[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[10] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, CA, USA: ACM, 2016, pp. 785–794.

[11] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 3146–3154.

[12] M. Adnan, A. Habib, J. Ashraf, S. Mussadiq, A. A. Raza, M. Abid, M. Bashir, and S. U. Khan, "Predicting at-risk students at different percentages of course length for early intervention using machine learning models," *IEEE Access*, vol. 9, pp. 7519–7539, 2021.

[13] L. Jin, L. Lingxi *et al.*, "Predictive modelling with the Open University Learning Analytics dataset (OULAD): A systematic literature review," in *Proceedings of the International Conference on Artificial Intelligence in Education*. Cham: Springer Nature Switzerland, 2024, pp. 1–12.

[14] S. B. Kotsiantis, "Use of machine learning techniques for educational purposes: A decision support system for forecasting students' grades," *Artificial Intelligence Review*, vol. 37, no. 4, pp. 331–344, 2012.

[15] L. P. Macfadyen and S. Dawson, "Mining LMS data to develop an 'early warning system' for educators: A proof of concept," *Computers & Education*, vol. 54, no. 2, pp. 588–599, 2010.

[16] A. Shaikhanova, O. Kuznetsov, K. Iklassova, A. Tokkuliyeva, and L. Sugurova, "Interpretable predictive modeling for educational equity: A workload-aware decision support system for early identification of at-risk students," *Big Data and Cognitive Computing*, vol. 9, no. 11, p. 297, 2025.

[17] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[18] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[19] Y. Lu, S. Yeom, R. Ryu, and S.-H. Kim, "Course-level clustering to enhance dropout prediction accuracy," *IEEE Access*, vol. 13, pp. 204 996–205 013, 2025.

[20] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Information Fusion*, vol. 81, pp. 84–90, 2022.

[21] D. T. Tempelaar, B. Rienties, and B. Giesbers, "In search for the most informative data for feedback generation: Learning analytics in a data-rich context," *Computers in Human Behavior*, vol. 47, pp. 157–167, 2015.

[22] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Anchorage, AK, USA: ACM, 2019, pp. 2623–2631.

[23] J. Pearl, *Causality: Models, Reasoning and Inference*, 2nd ed. Cambridge, UK: Cambridge University Press, 2009.