Early Student Dropout Prediction in Online Learning Using Explainable Machine Learning and Deep Learning Models

MSc Research Project
Data Analytics

Shakti Siva Raman Annamalai
Student ID: X23383569

School of Computing
National College of Ireland

Supervisor: Athanasios Staikopoulos

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Shakti Siva Raman Annamalai		
Student ID:	X23383569		
Programme:	MSc in Data Analytics	**Year:**	2025
Module:	Research Practicum		
Supervisor:	Athanasios Staikopoulos		
Submission Due Date:	11th December 2025		
Project Title:	Early Student Dropout Prediction in Online Learning Using Explainable Machine Learning and Deep Learning Models		
Word Count:	***7775***	**Page Count:**	**23**

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Shakti Siva Raman Annamalai

Date: 11/12/2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Early Student Dropout Prediction in Online Learning Using Explainable Machine Learning and Deep Learning Models

Shakti Siva Raman Annamalai
X23383569

Abstract

Online learning is now part and parcel of contemporary learning, but it is facing significantly greater rates of dropouts than in the real classrooms. In this study, an early prediction system that has been developed is aimed at identifying potentially at-risk students at the start of the first few weeks of a module so that a tutor can intervene before the situation gets out of control. Based on the Open University Learning Analytics Dataset (OULAD) of demographic attributes and the early engagement behavior, a number of predictive models were applied, as follows, Logistic Regression, Decision Tree, XGBoost, and an Artificial Neural Network. Out of them XGBoost was the most successful model on the whole, with good predictive stability and can be used as an early warning tool. SHAP explainability tools were utilized to achieve greater transparency and interpretability and determine the factors that contribute to dropout the most, with the early click activity and previous academic achievements seeming to be the most important factors. A probability-based risk stratification mechanism was also included in the system that allowed grouping students into valuable risk groups to facilitate targeted intervention. In general, the study represents a feasible, understandable, and scalable early-warning model that can be implemented in learning institutions to enhance student retention and the level of learner support. The work can be continued in the future to expand the feature set and investigate sequential modelling strategies to gain further insights into behavior.

Keywords: Early Dropout Prediction, Machine Learning, Deep Learning, XGBoost, Logistic Regression, Artificial Neural Network (ANN) SHAP Explainability, Student Engagement, Predictive Modelling, Risk Stratification.

1 Introduction

1.1 Background

Online learning has become a vital form of education that provides flexibility and accessibility for learners from all over the world. However, despite the benefits, online education has one nagging problem a high student dropout rate compared to traditional classroom learning. Numerous studies have tried to fix this problem through Machine Learning (ML) and Deep Learning (DL) to predict students at risk of withdrawing from their courses. These models often use demographic information, engagement behaviour and performance information gathered via virtual learning environments (VLEs). While the earlier studies reveal encouraging results in dropout prediction, most of them focus on end-of-course prediction which offer limited time to educators to take preventive actions. Additionally, the fact that many approaches lack model interpretability makes it difficult for educators to understand and respond to the drivers of student dropout.

1.2 Problem Statement

The major issue to be addressed in this project is the late identification of those students who are likely to drop out of online courses. Many predictive models identify the dropout risk after students have already done a great share of the course, taking away the chance of early intervention. Moreover, deep learning models in previous research, such as LSTM and CNN, need sequenced/ordered data (for instance, weekly or daily activity logs), which are not available in every dataset. Thus, there is a need for an early, explainable and practical dropout prediction approach that can perform well even with limited engagement data.

1.3 Motivation

Anticipating early student dropout can make a big difference for retention, completion and learning results. Educators can provide personalized support, academic guidance and motivation to students who are at risk during the first three or four weeks in order to prevent drop-out. This work is inspired by the vision of creating a predictive system that can be both accurate and interpretable, so that tutors not only know which students are at-risk, but also why they are at risk. Early, interpretable predictions could enable interventions and overall higher course completion rates.

1.4 Research Question

How accurately can early engagement and demographic data predict student dropout in online courses using both explainable machine learning and deep learning models?

1.5 Objective

- To build and evaluate Machine learning models (logistic regression, Decision Tree, and XGBoost with SHAP for interpretability) and a deep learning model (ANN/MLP) for early dropout prediction
- To compare the performance of ML and DL models I use metrics such as accuracy, precision, recall, F1-score, and AUC
- To identify the best performing models for early dropout prediction.
- To provide explainable and actionable insights that can help tutors support at risk student effectively.

1.6 Limitation

Even though LSTM or GRU deep learning architectures are frequently employed in educational prediction tasks, they have to operate with ordered or sequential interaction history (e.g. daily or weekly engagement patterns). The data in OULAD lacks time-sequenced behavioral data, and the data consists of aggregate counts of clicks, so sequential models are inappropriate in this research. This is why an MLP was chosen as the deep learning baseline as MLPs are long-standing in the process of modelling structured tabular data and can model nonlinear associations involving demographic factors and early engagement behavior. Moreover, as existing studies in the field of learning analytics have demonstrated, MLPs can compete effectively with tree-based models in cases where feature engineering is clear, thus these are an apt deep learning model to compare with the old machine learning algorithms.

Structure of the Report

Section 2: The Literature Review, provides a summary of the past studies that forecasted dropout with the help of ML and DL.
Section 3: Report on the Methodology, such as dataset, preprocessing, feature selection and model implementation.
Section 4: Discusses the findings and compares the performance of the models based on different evaluation measures.
Section 5: Finalizes the report with the conclusion and summarization of the finding, limitations and advices to the future work.

2 Related Work

One of the most significant aspects of research has become student dropout prediction because the dropout rates in online and MOOC-based learning are continuously very high. Research depicts that early engagement behaviour and demographic variables are very strong predictors of dropout risk, which necessitates predictive modelling to identify the risk and intervene in a timely manner (Jha et al., 2019; Magalhaes et al., 2021; Vaarma and Li, 2024). The recent studies have examined machine learning (ML), deep learning (DL), and the combination of both to understand the model dropout patterns of MOOCs, university LMS, and online training systems (Basnet et al., 2022; Cho et al., 2023; Rabelo and Zarate, 2025).

However, the existing literature is focused on end-course prediction, elaborate deep learning frameworks, or accuracy with little focus on the early prediction, model simplicity and interpretability. The researcher addresses these gaps in this paper by developing an early and explainable predictive dropout model using ML and ANN on the OULAD dataset.

2.1 Machine Learning Based Approaches

Machine learning has been extensively found in dropout prediction due to its applicability on structured data like demographics and VLE interactions. In Jha et al. (2019), OULAD was applied to predict dropout through regression, ensemble and ML, and demonstrated a high level of success in the case of behavioural features. Magalhaes et al. (2021) and Villar and de Andrade (2024) have shown that random forest and gradient boosting tree-based models tend to perform better than simpler models particularly when behavioural engagement data is available.

The significance of early-week LMS activity in predicting early withdrawal was emphasized by Park and Yoo (2021), whereas Zerkouk et al. (2024) ensured that behavioural features combined with demographic ones help predict it better in online training platforms. Random Forest was also found to be efficient by Dass, Gary and Cunningham (2021) in self-paced MOOCs. Despite the good performance of ML approaches, majority of the studies use full-course data. There is less research on early dropout and it is often not interpretable.

2.2 Deep Learning Based Approach

The student activity logs have been captured using deep learning models like DNN, CNN and LSTM to learn the complex behavioral patterns. Basnet et al. (2022) compared ML and DL models on MOOC data, and the results have shown that LSTM-based architecture has a strong performance. Almaazmi et al. (2023) and Jin (2023) also demonstrated that temporal models can be useful in the description of sequence-based engagement.

Nevertheless, time-series behavioral logs are typically needed by deep learning methods, and such data is not always easily available or feasible to gather in an institutional context (Torkhani and Rezgui, 2025). They also come at high calculation cost and low interpretation, which renders them inapplicable to the easy to visualize Tabular data like OULAD. ANN/MLP models provide a trade-off between the simplicity of sequential modelling and the deep learning capabilities.

2.3 Explainable and Hybrid Approaches

Explainability is an aspect that has been gaining relevance in dropout prediction studies due to the need to have educators to know why a student is likely to be at risk. Er (2023) created an explainable machine learning system to explain the dropout predictions in MOOCs, demonstrating that feature-level explanations turn out to be more trustful and beneficial to the instructors.

Emerging hybrid methods of ML, DL, and explainable methods. Alghamdi, Soh and Li (2025) have reviewed dropout-prediction systems and stated that their systems should incorporate interpretability to be adopted in a real-world. The fact that behaviour-driven prediction models should incorporate meaningful explanations to aid academic decisions was also stressed by Marcolino et al. (2025) and Rabelo and Zarate (2025).
Nevertheless, with this advancement, there is a scarcity of research that integrates features of early-stage with interpretable modelling methods into a light-weight and practical model.

2.4 Comparison between Previous Research vs My Project

Dimension	**Previous Research/Existing Work**	**My Project**
Dataset Context	Many studies use MOOC platforms, university LMS data, or OULAD (Jha et al., 2019; Magalhães et al., 2021; Marcolino et al., 2025).	We Used OULAD with a focus on early engagement + demographics for practical institutional uses.
Prediction Timing	It Often predicts full-course or late dropout (Jha et al., 2019;	We Predicts dropout within the first 3–4 weeks, for enabling early

	Dass et al., 2021; Rabelo and Zárate, 2025).	intervention.
ML methods	Mainly use of LR, DT, RF, SVM, Gradient Boosting (Villar & de Andrade, 2024; Zerkouk et al., 2024).	Many focused on Implementing LR, DT, XGBoost for fair and structured comparison.
DL methods	It Uses LSTM, CNN, DNN; requires time-series data (Basnet et al., 2022; Jin, 2023).	We Used ANN/MLP, which works with simple tabular early-week features.
Explainability	Many models lack interpretability (Alghamdi et al., 2025; Er, 2023).	Integrates SHAP for feature-level explanations of XGBoost predictions.
Feature Usage	Heavy reliance on detailed actions (forum posts, weekly logs) (Jin, 2023; Marcolino et al., 2025).	Uses demographics + total_clicks_early, simple & realistic to collect.
Practical application	Often focused on algorithmic performance (Basnet et al., 2022).	We focus on a practical, early-warning tool for tutors.

2.5 Identified Research Gap and Justification

Despite many studies using machine learning and deep learning to predict student dropout, much of the current literature is based on full-course behavioural logs, multifaceted sequential models, or only achieves high accuracy but not early intervention (Jha et al., 2019; Basnet et al., 2022; Rabelo and Zarate, 2025). Few studies investigate the risk of dropouts during the first few weeks, and even less of them take into account straightforward, institutionally available tabular data, including demographic data and initial click activity (Park and Yoo, 2021; Magalhaes et al., 2021). Moreover, a lot of deep learning models used in previous studies require time-series or sequential interaction histories, unavailable in such datasets as OULAD and, thus, such models are challenging to implement in practise. Such models may also generally not be interpretable, so educators are unable to interpret what they predict or rely on it (Jin, 2023; Er, 2023).

This paper attempts to fill these gaps by specifically examining the issue of early-stage dropout prediction based on the first three to four weeks of engagement data, which allows initiating intervention at the appropriate moment. It also adds to the domain by comparatively assessing machine learning models to a deep learning MLP on structured tabular data- a field where empirical investigation has not been considerably investigated in previous research. Also, the SHAP explainability can be used to have a clear, easy-to-understand, insights into the factors that cause dropout, in a practical will-to-use way of guidance to effective academic decision making. This early prediction, model comparison, and explainability is a significant contribution to the existing state-of-the-art research.

3 Research Methodology

3.1 Introduction to Methodology

This section describes the research procedure that was applied to create a model of early dropout prediction based on machine learning (ML) and deep learning (DL) methods. The paper is based on the CRISP-DM model (Cross-Industry Standard Process of Data Mining) that offers a structured and scientific cycle that involves data understanding, data preparation, data modeling, data evaluation, and data interpretation. The primary goal will be to forecast the student dropout on an early basis-before the first four weeks of online course- by examining student activity and demographic trends. The methodology will make all the steps transparent, repeatable and in line with the overall aim of the research being to identify at-risk students in advance to avoid them dropping out

3.2 Architecture Diagram for Early Dropout Prediction

This study was to be carried out through four consecutive stages of implementation. Phase 1 involved the identification and preparation of the data sources which comprised the three fundamental OULAD tables: studentInfo, studentVLE and studentRegistration. Phase 2 entailed initial data preprocessing, where the interaction between VLE was limited to first 28 days of activity and later aggregated, merged and transformed to form a single modelling dataset. Phase 3 involved the stage of model development, where a few machine learning models, including Logistic Regression, Decision Tree, and XGBoost, and a deep learning model (ANN), were trained and tested on the basis of the standard performance measurements. Lastly, Phase 4 generated the interpretability and practical outputs of the system, such as SHAP-based explainability visualizations and high-risk students, i.e., those with a high probability of dropout identified.

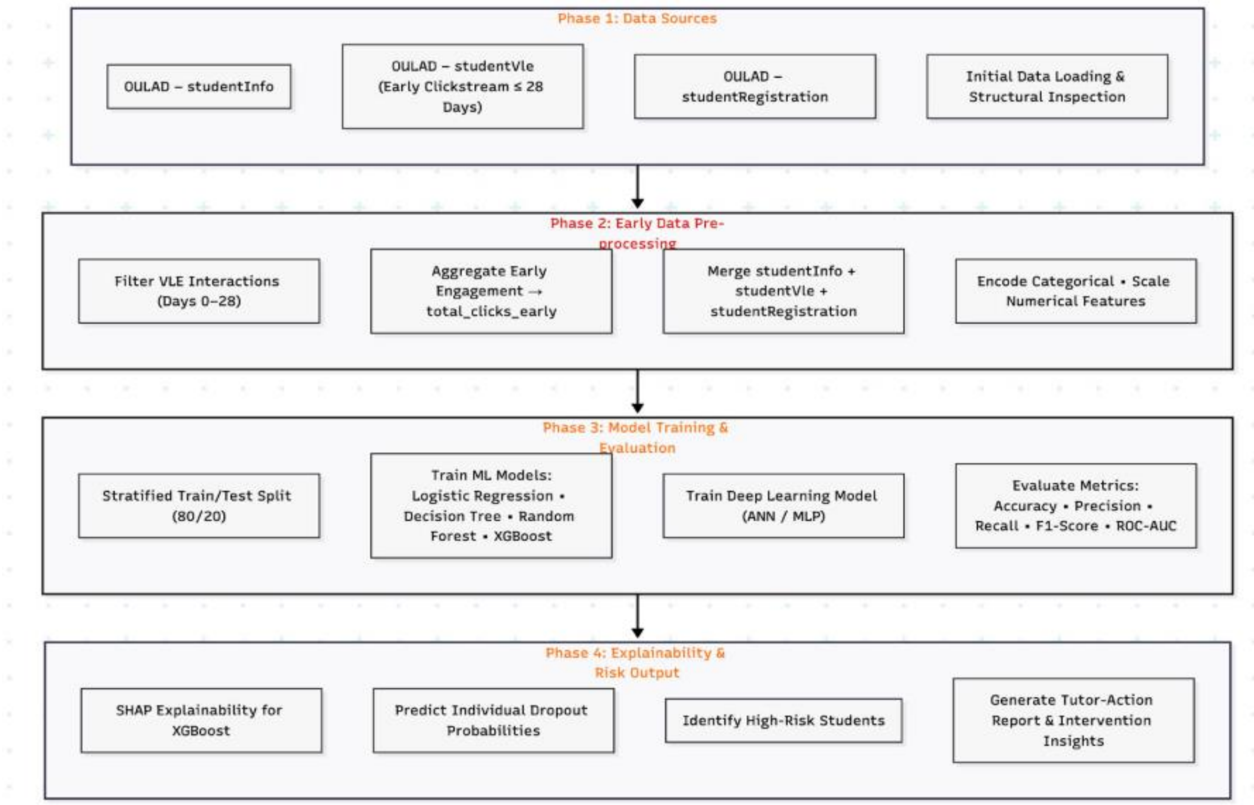

Figure 1: Architecture Diagram

3.3 Data Description and Preparation

This study uses the Open University Learning Analytics Dataset (OULAD), a publicly available resource, as the dataset in the research. It has the demographic, behavioural and registration data of 32593 students taking online courses. The raw data is represented as a number of interconnected tables, of which three will be mostly utilised in the present project. StudentInfo.csv file gives demographic information of gender, age band, highest education, disability status and final result. StudentVle.csv file has comprehensive Virtual Learning Environment (VLE) interaction logs, which capture the frequency of visit and date of visit of students to course materials. The studentRegistration.csv file contains the date of registration and unregistration of every student and is utilised in determining the drop out of any student.

A number of preprocessing steps were undertaken in order to prepare a clean and structured dataset to be used in the modelling. To begin with, the initial VLE interactions were pre-filtered to comprise the first 28 days of activity because this is the stage that represents the initial engagement phase during which dropout tendencies have the greatest likelihood of developing. Second, the student engagement was summarized as the sum of the overall number of clicks in the first 28 days, resulting in the creation of the feature total_clicks_early that reflects the early behavioural trends. Third, the three main data sets were connected to one another (with the help of common identifiers) studentVle, studentInfo, and studentRegistration to create one common dataset. Fourth, a binary target variable dropped_out was designed with the value of 1 representing the case of dropout by a student (against date_unregistration) and 0 representing the case of continuity. Fifth, any gaps in the engagement values were assigned 0 as a value to signify no activity, and any gaps in the demographic variables (gender or highest education) were coded as Unknown to maintain the completeness of the data.

Lastly, gender, age band, highest education, disability, and total_clicks_early were chosen as pertinent features to be modelled whereas other features like final_result and date_unregistration which might lead to data leakage were avoided at the expense. The cleaned dataset was saved as prepared_dropout_dataset_early.csv, which has one record per student with all the features needed and the target value to be predicted.

3.4 Research Design and Process:

The study workflow was also designed and replicable, and it was scientifically right and allowed the entire process to be repeated. The first stage was data preprocessing in which the raw OULAD tables were cleansed, merged, and narrowed down to some similar modelling data. This was then followed by feature encoding and data division where the categorical variables were coded into numbers and the data sample was further divided into a training (80) and test (20) sample using stratified sampling in order to maintain the initial class distribution. The second phase was devoted to the development of models in which several machine learning and deep learning models have been trained on the same data to allow the fair performance comparison to be made. Once the model was trained, model evaluation was performed with the help of quantitative measures to determine the predictive accuracy and the quality of the whole performance of the model. Lastly, the workflow also included an explainability analysis based on SHAP, which showed the relevant factors that affect the prediction of dropouts and better interpretability of the system.

3.5 Techniques and Algorithms Applied:

The study adopted three machine learning algorithms and a single deep learning model to identify student dropout. Linear baseline model was used on the logistic regression (LR), which is highly interpretable, with its coefficients, which determine the impact of each feature on likelihood of dropout. Decision Tree (DT) was a rule-based classifier that can divide students into small groups based on their attributes so that the process of decision making is easily represented in the form of a tree. The XGBoost, which is an advanced ensemble algorithm, is used to combine several weak learners successfully to achieve greater predictive accuracy and robustness, which can also work with imbalanced datasets and allow interpretation with the help of SHAP values. The deep learning model is an Artificial Neural Network (ANN/MLP), which is made up of an input layer, a hidden layer that contains 16 neurons with ReLU-activation, and a sigmoid output layer, which allows it to learn nonlinear relationships that other models might fail to recognise. All the models were trained on the same feature set and similar parameter settings so that performance and interpretability were fairly and equally assessed.

3.6 Evaluation Methodology:

Various standard measures were used to compare the model performance they are Accuracy, Precision, Recall, F1-score and AUC (Area Under the ROC Curve). The highest priority was given to Recall since the primary aim of the early dropout prediction is to find as many at-risk students as possible with some false positives. A five-fold cross-validation strategy was adopted in order to maintain healthy and impartial performance and the means of all the folds were noted.

Additionally, a Confusion Matrix was created in each model, to plot the visualization of true positives (predicted dropouts correctly), true negatives, false positives and false negatives. This assisted in interpreting model effectiveness in practical values. SHAP (SHapley Additive exPlainations) was implemented in order to increase interpretability, especially on the XGBoost model, to learn about the aspects that impacted the most on every prediction. As an example, reduced early engagement (total clicks) and lower educational level were found to be important predictors of the risk of dropout. This interpretability gives useful insights to the tutors to take preventive measures.

3.7 Relation to previous studies:

This study is consistent with the past studies, which have utilised both OULAD dataset and LMS behavioural data in predicting student dropout (Jha et al., 2019; Magalhaes et al., 2021). But this research does not look at a duration of 12 weeks or more like most other previous studies, but rather the initial four weeks which makes it possible to make true early predictions. It also compares the model of ML and ANN directly and includes SHAP-based interpretability. This makes the model more practical and helps in early intervention of the real online learning situations.

4 Design Specification

4.1 System Architecture

The system is organised in the form of an analytical pipeline that processes raw OULAD data into meaningful features to predict early dropout. The architecture comprises of three light weight stages. Data Processing Layer, at this stage, unites three basic OULAD tables which include studentInfo, studentVLE, and studentRegistration and makes them ready to be modelled. This involves the filtering of VLE interactions to the initial 28 days to capture early contact, aggregation of click behaviour into the total_clicks_early feature, categorical encoding, missing values, and creating a binary dropped out label through registration status. Such measures are taken to avoid any data leakage to avoid making predictions based on data obtained at the very beginning of the course. The second step is Feature Engineering and Encoding, which is categorical encoding and numerical scaling where necessary, which yields a single clean feature set that will be used uniformly across all models. Lastly, the Unified Modelling Framework cross-trains every model with 80/20 stratified split and a common preprocessing pipeline, the fairness and reliability of comparing the performance of every machine learning and deep learning model.

4.2 Modelling Techniques

The system has three machine-learning algorithms and a deep-learning model, and no new algorithm is suggested; the functional behaviour is determined according to the system requirements. Logistic Regression (LR) is a linear probabilistic classifier that approximates the connexion between features and the likelihood of dropout which can be used as a good basis of interpretability using its coefficients. Decision Tree (DT) is a rule-based classification method, which groups students into decision paths, including determining low engagement as a risk of increased dropout, to provide the decision logic in an easily understandable manner. XGBoost or gradient-boosting ensemble is a machine learning algorithm that creates a series of weak learners in order to be highly predictive and robust enough to address non-linear interactions and missing data. It also works with SHAP to allow explainable insights to the tutors. The Artificial Neural Network (ANN/MLP) deep-learning component, which is also dense hidden layers with ReLU activation and a sigmoid output, binary classification can be used, therefore, to capture complex non-linear engagement patterns and weighting of classes to deal with disproportionate data. A combination of these models creates a systemic model of assessing early dropout predictability.

4.3 Implementation Requirements and Design Justification

The implementation and design of the predictive system are determined by three key objectives that guide the project. The first goal, early intervention is concerned with the utilization of first-month engagement statistics and demographic characteristics in order to find students who are at risk of dropping out but still have enough time to allow tutors to intervene. The second, which is interpretability, provides that the predictive procedure can be viewed as transparent; including trees-based models and SHAP analysis, the educator can see exactly why a particular student was classified as at risk. Scalability and reproducibility are the last goal that is met by using a modular preprocessing and modelling pipeline to enable other models or data can be added without the need to make significant changes to the system design. Collectively, such goals make the system viable, explicable, and long-term educational application.

5. Implementation

5.1 Integrated Workflow, Model Outputs, and System Explainability

The system has a full end-to-end analysis workflow starting with the conversion of raw OULAD tables into a structured modelling dataset. The first stage of preprocessing the data included the combination of the studentInfo, studentVle and the studentRegistration databases and transforming it into a single dataset (prepared_dropout_dataset_early.csv), where each row is an individual student. This data will consist of demographic data, early engagement behaviour in the form of the total number of clicks within the first 28 days and a dropout label that is binary. All missing data were standardized, redundant variables were eliminated and the features engineered in uniform, model ready form. Based on this cleaned data, four predictive models (Logistic Regression, Decision tree, XGBoost and an Artificial Neural Network (MLP) were fitted and assessed with the same input features to provide fairness. The outputs of these models were dropout predictions, classification labels and performance measures like Accuracy, Precision, Recall, F1 score and AUC. Additional visual deliverables like ROC curves, confusion matrices, and comparative performance charts were also generated in order to have a clear picture of the behaviour and performance of each model.

Actionable insight and explainability of the system were a key aspect of the pipeline. The XGBoost model with the best and most consistent performance was also further discussed through SHAP explainability to determine the most influential factors in the prediction of dropouts. The variables of early clicking and demographic qualities were identified in SHAP bar charts and beeswarm plots as influential in model choices, and serve as a way to give transparency to the educator. Moreover, the system established a risk stratification system that identified students assuming high, medium and low dropout likelihoods as High, Medium and Low-risk, which allowed the system to intervene in practice at an early stage. One more table with Top-50 High-Risk Students was created to help tutors make decisions. All the code is written in Python in the Jupyter Notebook, with help of Pandas and NumPy to manipulate the data, Scikit-learn, XGBoost, and TensorFlow/Keras to build the models, SHAP to obtain interpretability, and Matplotlib/Seaborn to visualize the data. The workflow is modular and reproducible which guarantees that the system is scalable and can be further extended with more models or data sources in future.

5.2 Final Deliverables

The implementation generated a full early-dropout prediction system comprising of some main deliverables. To create a fully processed and model-ready dataset, first, extensive preprocessing and feature engineering was created. This data set was used to build several predictive models both in conventional machine-learning and ANN model, which were all trained and tested at the same time. Moreover, the outputs of SHAP-based interpretability were obtained to give transparent answers on factors that affect dropout predictions. The system also provided stratified risk-adjusted outputs, which assigned students to High, Medium and Low-risk groups and allowed early intervention to be practical. All these elements constitute the ultimate integrated solution that is made in this research.

6 Evaluation

The chapter compares four models- Logistic Regression model, Decision Tree, ANN (MLP), and XGBoost that are aimed at predicting early student dropout during the initial three-four weeks of course participation. Accuracy, Precision, Recall, F1-score and ROC-AUC were used to compare their performance with the support of ROC curves and confusion matrices. The SHAP analysis was used to explain the impact of behavioural and demographic factors. This is aimed at finding the most viable model that can be used practically in early-intervention.

6.1 Experiment – 1 Logistic Regression (LR) Evaluation

The statistical model that was used as a baseline and that was implemented to assess the capacity of simple linear relationships in explaining early dropout behaviour was Logistic Regression.

Metric	Score
Accuracy	0.6013
Precision	0.4161
Recall	0.7673
F1-score	0.5396
AUC	0.6993

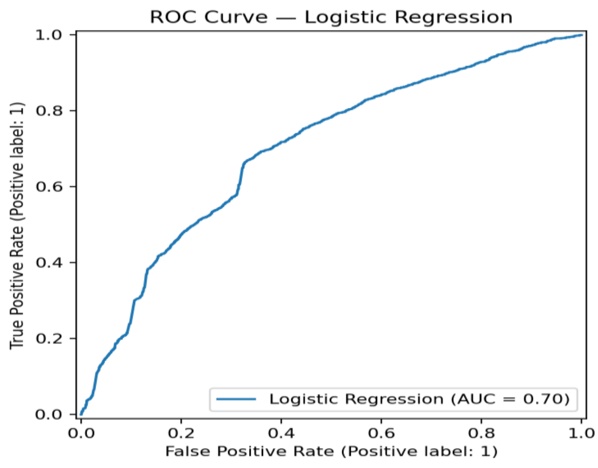

Figure – 2 ROC Curve for Logistic Regression

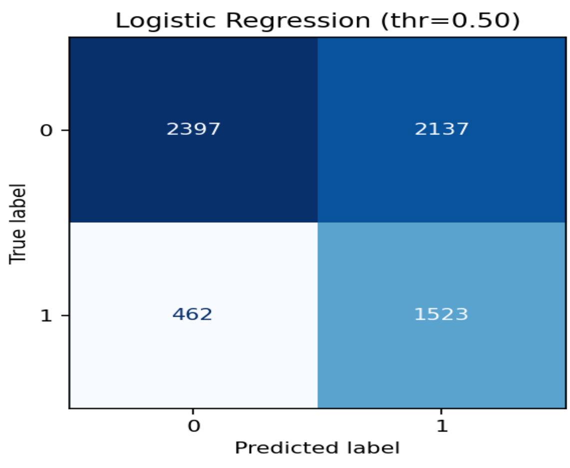

Figure – 3 Confusion Matrix for Logistic Regression

Logistic Regression exhibited a high recall (0.7673) i.e. it was able to single out the students who are at the risk of dropping out- helpful in the case of early-warning systems. Its low accuracy (0.6013) however suggests a high number of false positive and moderate F1 (0.5396) and AUC (0.6993) which is poor reliability in predicting dropout and non-dropout cases. All in all, though applicable to initial detection, Logistic Regression does not have the reliability required to undertake sure early intervention.

6.2 Experiment -2 Decision Tree (DT) Evaluation

The Decision Tree model has been used to achieve rule-based patterns on early engagement and demographic character and this allows the model to construct interpretable decision paths to classify the dropout risk.

Metrics	**Score**
Accuracy	0.7082
Precision	0.5195
Recall	0.5567
F1-Score	0.5375
AUC	0.7213

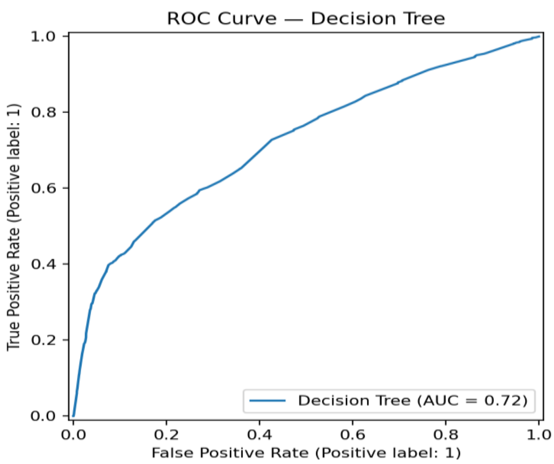

Figure – 4 ROC Curve for Decision Tree

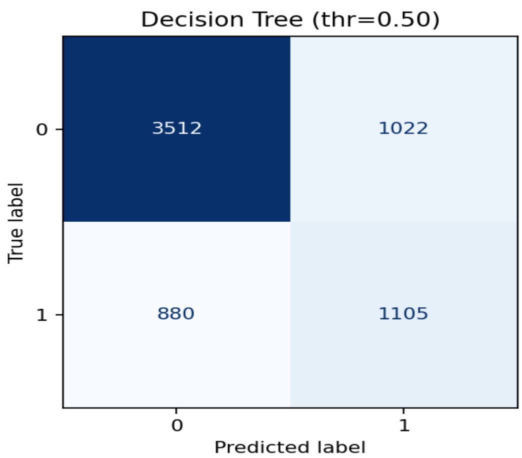

Figure 5 – Confusion Matrix for Decision Tree

The Decision Tree was relatively good with the highest accuracy (0.7082) and AUC (0.7213) compared to the Logistic Regression thus it is more valid to indicate the presence of at-risk students with reduced false positives. This however, has a low recall (0.5567), which indicates that it has missed some real drop outs. Easy to interpret, its predictive stability and generalization are moderate, and it is useful, but not the best as a whole.

6.3 Experiment 3 – Artificial Neural Network (ANN/MLP) Evaluation

Artificial Neural Network (ANN/MLP) was used to model non-linear relationships in the early engagement data and is a more flexible structure to model than conventional machine learning methods.

Performance Table

Metric	**Score**
Accuracy	0.7052
Precision	0.5142
Recall	0.5743

F1-score	0.5426
AUC	0.7265

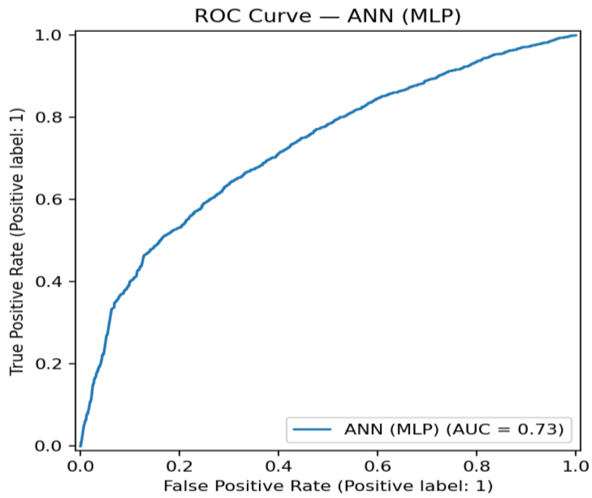

Figure -6 ROC Curve for ANN (MLP)

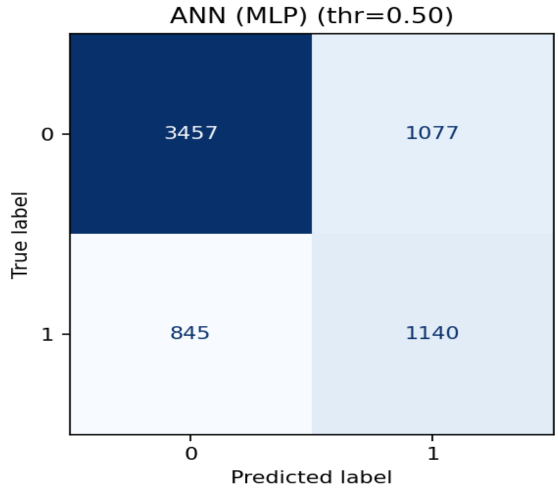

Figure – 7 Confusion Matrix for ANN (MLP)

The ANN performed in balance, with decent accuracy and AUC, which means that it is only able to distinguish between high-risk and low-risk students to a reasonable degree. Nevertheless, it had lower recall (0.5743) compared to Logistic Regression (0.7673), which is why it was not the best model to use when the early identification of the students at risk is the focus. ANN based on its consistent performance did not prove to be the most efficient model on early dropout detection.

6.4 Experiment 4 – XGBoost Model Evaluation

The XGBoost model was applied in order to utilise the complex and non-linear relationships in early behavioural and demographic variables and maximise the predictive accuracy on unequal educational data sets using gradient-boosted decision trees.

Metric	**Score**
Accuracy	0.7191
Precision	0.5381
Recall	0.5481
F1-score	0.5430
AUC	0.7303

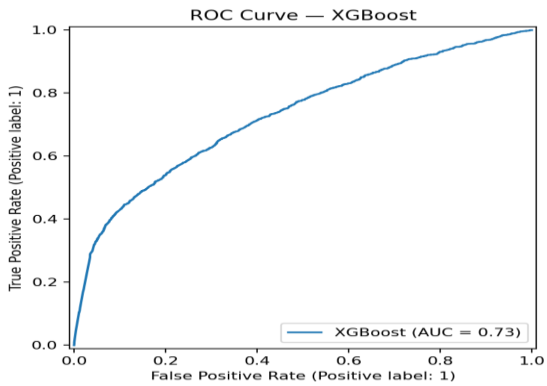

Figure -8 ROC Curve for XGBoost

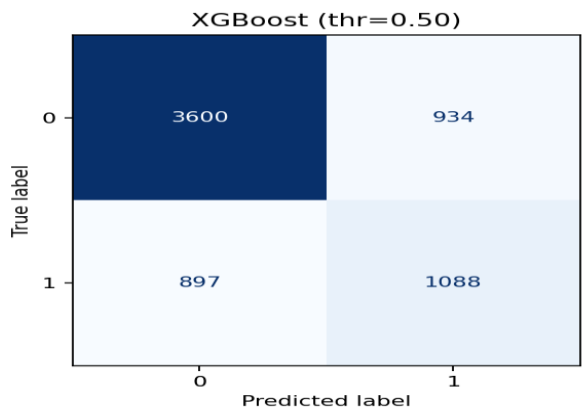

Figure – 9 Confusion Matrix for XGBoost

XGBoost produced the best and most consistent results with the highest accuracy (0.7191) and AUC (0.7303) and demonstrated better capability to classify dropout and non-dropout. It had a trade-off balance between precision and recall so that it identified a large number of students under threat without false alarms. Its ability to model non-linear and complex interactions is what makes it suitable when dealing with early engagement data. In general, XGBoost is the best model to be used to predict early dropouts.

6.5 Experiment 5 - Overall Model Comparison

All four models were tested on the same test set with a fixed probability threshold of 0.50 to compare the four models. The following table is a summary of the important metrics; Accuracy, Precision, Recall, F1-score and AUC of the Logistic Regression, Decision Tree, ANN (MLP) and XGBoost. Bar chart was also created to give a visual comparison of these measures among models.

Model	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	0.6013	0.4161	**0.7673**	0.5396	0.6993
Decision Tree	0.7082	0.5195	0.5567	0.5375	0.7213
ANN (MLP)	0.7052	0.5142	0.5743	0.5426	0.7265
XGBoost	**0.7191**	**0.5381**	0.5481	**0.5430**	**0.7303**

Figure – 10 Comparison Table

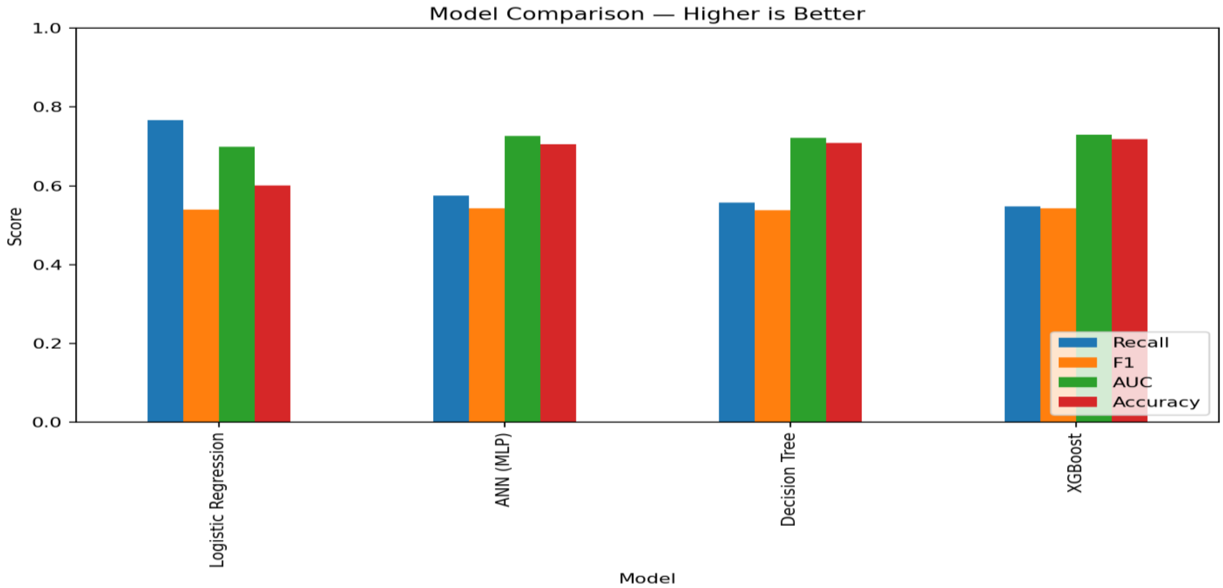

Figure 11 – Model Comparison

It is evident that there is trade-off between Logistic Regression and the better models. The most successful model in terms of recall (0.7673) is the Logistic Regression, which detected the largest percentage of real dropouts, however, at the expense of low precision (0.4161) and the lowest accuracy and AUC of all the models. In theory, this would produce multiple false alarms, sending tutors several of the students who are not actually at high risk, which would be hard to deal with in an actual institutional context. However, XGBoost offered the most balanced performance: it had the highest accuracy (0.7191), highest precision (0.5381), highest F1-score (0.5430) and highest AUC (0.7303), and yet a decent recall. This shows that the XGBoost is more effective in ranking students in terms of risk and distinction between probable and unlikely dropouts with fewer false alarms. ANN models and Decision Tree did a similar job but always slightly worse than XGBoost on the primary metrics.

XGBoost was chosen as the main AI model of early dropout risk prediction, and this choice is based on this comparison. It provides a better trade-off between falsely identifying at-risk students and false positives, and its predictions can be further interpreted as SHAP analysis in the section below.

6.6 Experiment 6 – SHAP Explainability Evaluation for – XGBoost

To guarantee the interpretability of the predictive model, SHAP (SHapley Additive exPlanations) was implemented on the XGBoost classifier, which showed the highest overall performance in terms of evaluation measures. In contrast to other conventional feature importance techniques, SHAP offers both global and local explanation, which measures the contribution of each feature to the prediction of the model to each individual student.

Global Feature Importance (SHAP Bar Plot)

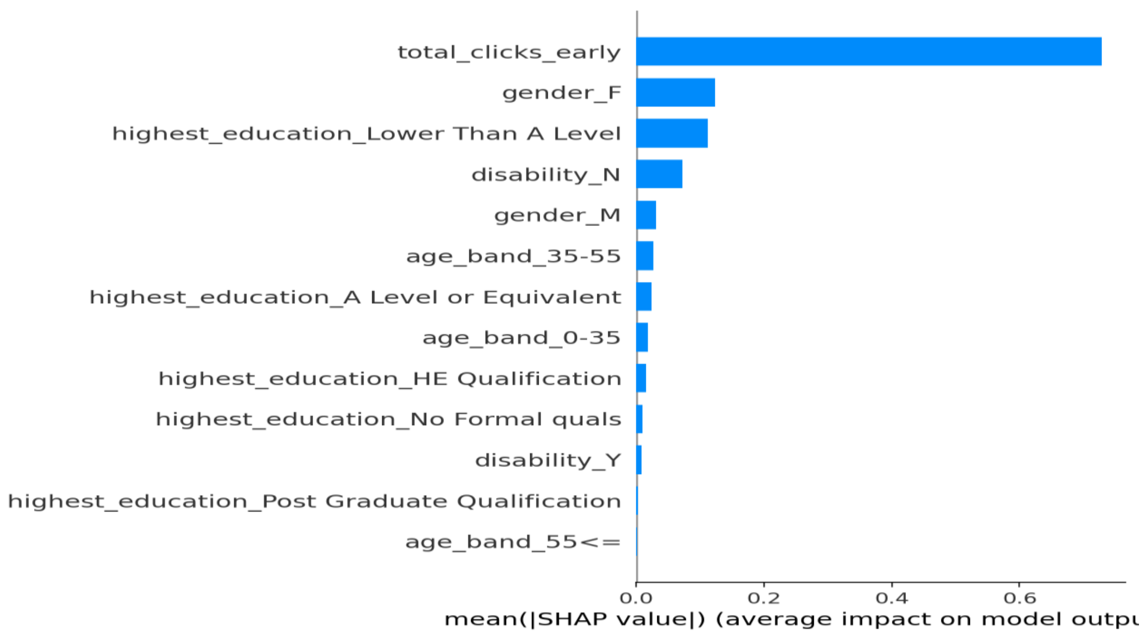

Figure 12 – SHAP Feature Importance for XGBoost

The SHAP feature importance plot indicates that the predictor with the greatest influence on early dropout is total_clicks_early, so it is true that the students who achieved very low early engagement are at a considerably higher risk. Demographic variables including highest education, age band and disability status had significant contributions to the decision made by the model but made them significant structural factors linked to disengagement.

SHAP Summary Plot

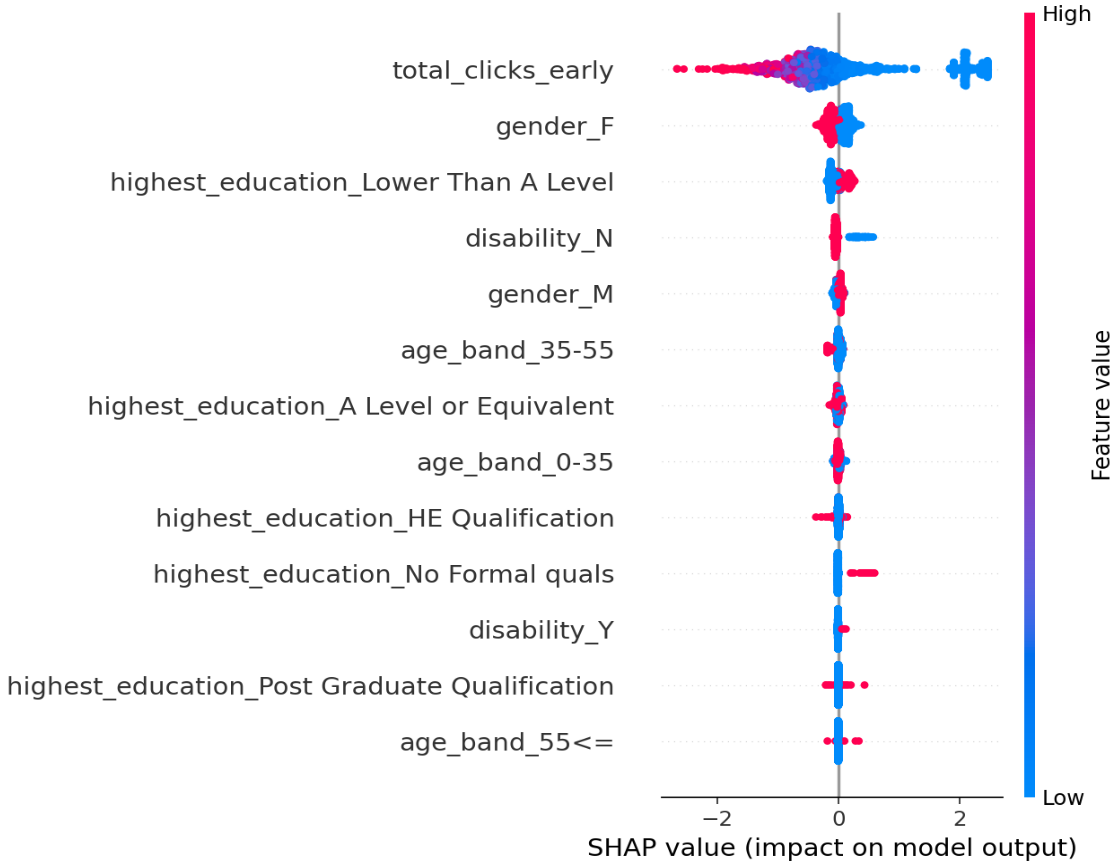

Figure 13 - SHAP Summary Plot for XGBoost

Key observations include

The SHAP analysis indicated the existence of some significant trends that can be used in the predictions of dropouts. Low early engagement proved to be the best predictor of dropout with low total_clicks_early students being and repeatedly placed in the dropout category. The level of education was also very significant with students who had lower levels of previous education showing the likelihood of being disengaged and dropping out. Furthermore, there was a presence of recorded disability which added moderately to higher dropout probability. Some age groups especially students between 0-35 recorded greater dropout rates, which could be attributed to early career expectations or lack of readiness to do self-directed online studies. The combined effects of these insights demonstrate the impact of behavioural and demographic factors on the risk of early dropout.

Interpretation and Practical Observation

The analysis which is based on SHAP offers a number of useful implications towards academic staff as it brings out the factors that are most significantly related with early dropout. The findings indicate that the most vital indicator is the initial engagement meaning that students who are actively involved in VLE in the first four weeks need to be immediately observed in case they can be intervened. Educational background also becomes a significant variable and students who come with lower qualifications might need extra academic scaffolding in order to stay involved. The results concerning disability substantiate the necessity to provide timely and specific support services to guarantee equal learning opportunities. Besides, SHAP provides an individual-level explanation which specifies why a particular student was labelled as high risk, allowing individual and reasonable interventions. Not only does SHAP make XGBoost more approachable, which translates the more complex decision-making of XGBoost into a more comprehensible system, but it also guarantees that the system contributes to evidence-based and personalised retention strategies and not just the identification of at-risk learners.

6.7 Experiment 7 – High /Medium/Low Risk Stratification

This section will show the end product of the early-warning system where the students in the test dataset would be classified into the High, Medium, and Low risk groups based on the generated dropout probabilities of the chosen model (XGBoost).The role of this stage is to render the numerical predictions into actionable insights to the tutors to allow timely academic interventions that can take place in the first few weeks of the course.

The prediction outputs of the predictive models were divided into three levels of distinct risk according to preset thresholds to aid in early intervention. Students whose probability was predicted to be more than 0.70 were defined as High Risk which implies high chances of dropping out. Individuals having probabilities of 0.50 to 0.69 were put in Medium Risk category, which showed a moderate level of concern. The students with the predicted probability less than 0.50 were regarded as Low Risk, meaning that there was a smaller tendency of disengagement. The following table gives a summary of the distribution of these risk levels within the entire test data, which gives us a general picture of the risk patterns among the students.

Serial Number	Risk	Count	Percentage
0	High	950	14.57%
1	Medium	1072	16.44%
2	Low	4497	68.98%

Figure 14 – Overall Risk Distribution in the Test Set

The analysis of the distribution of risks shows that the number of students who are in the High-Risk group is 14.57% (950 people), which means that there is a high probability of early disengagement, and the highest necessity of prompt academic assistance. The behavioural and demographic patterns of these students place them at a considerably high risk of dropping out of the course unless they are supported in good time. Medium-Risk group is 16.44 (1,072 students) percent of the dataset and consists of learners that show moderate warning signs; without specific monitoring or early intervention, such students can advance to the high-risk group in the long term. The largest proportion is Low-Risk students, 68.98% (4,497 learners), which is characteristic of real learning records with the vast majority of students enrolled in education and only a small percentage becoming disengaged. Despite the fact that the High-Risk group constitutes the least size of the population, it is the most critical in terms of the intervention perspective because such students need proactive and individual assistance to avoid the imminent dropout. This stratification thus offers a very practical structure to the educators so that they can be precise in resource distribution, and their intervention strategies can be taken in time.

Top – 10 At High-Risk Students (Early Warning List)

The system produces a list of the Top-10 students who have highest probability of dropping out which will be prioritised allowing tutors to concentrate on the most urgent cases. These students are the ones that are at most risk and hence need immediate academic support or intervention. The Top-10 riskiest students are as follows; a complete list of Top-50 students was also produced and used at the evaluation stage to prove the ability of the system to detect and rank students by their tendency of disengagement.

	gender	age_band	highest_education	disability	total_clicks_early	actual	dropout_proba	risk_level
0	F	0-35	A Level or Equivalent	Y	0	1	0.938	High
1	F	0-35	A Level or Equivalent	Y	0	1	0.938	High
2	F	0-35	A Level or Equivalent	Y	0	1	0.938	High
3	F	0-35	A Level or Equivalent	Y	0	1	0.938	High
4	F	0-35	A Level or Equivalent	Y	0	1	0.938	High
5	F	0-35	A Level or Equivalent	Y	0	1	0.938	High
6	F	0-35	A Level or Equivalent	Y	0	0	0.938	High
7	F	0-35	A Level or Equivalent	Y	0	1	0.938	High
8	F	0-35	A Level or Equivalent	Y	0	0	0.938	High
9	F	0-35	A Level or Equivalent	Y	0	1	0.938	High
10	F	0-35	A Level or Equivalent	Y	0	1	0.938	High

Figure 15 – Top 10 High Risk student (Early Warning List)

The Top-10 list is a proactive dashboard allowing tutors to assist the students that are most likely to drop out in time. It enables educators to put importance on interventions to the learners who demonstrate very low engagement in the early stages and also to those who have known risk factors like low educational backgrounds or recorded disabilities. The system will assist in early intervention by identifying these students in the first four weeks of the course before their failure to engage is permanent. This directly corresponds to the main objective of early dropout prediction, which is to detect and intervene on indicators of less interest at a point where academic intervention can be most effective.

Risk stratification is a process that translates the predicted probabilities of the model into a systematic decision-making model. The high-risk students are given the priority to get immediate academic follow-up because they have a high tendency of disengagement. The medium-risk category members can be supervised or provided with support of light intensity to ensure that their risk level is not increased. The low-risk students usually do not need any direct intervention, however, monitoring their further activity is still observed. This step-by-step stratification system increases the utility and explainability of the predictive model and hence it can be used in practical contexts of real academic environments where informed and timely decision making is a necessity.

6.8 Discussion

The discussion indicates different model results with 28-day engagement data. The Logistic Regression was very high in recalls but had excessive false positives, hence cannot be used in practise. The ANN was close but had the drawback of the small feature space. Decision Tree was explainable but less generalisable and unstable. XGBoost always provided the most favourable balance of all the metrics, and it is able to capture delicate early-engagement trends. SHAP findings were consistent with intuition with low education level, low early clicks, and disability indicators noted as the main predictors of dropouts. On the whole, XGBoost has the best balance between accuracy and interpretability, which makes it the most appropriate model to use in identifying early dropouts and risk-based intervention.

7 Conclusion and Future Work

This paper demonstrates that early engagement behaviour with demographic attributes can accurately predict student dropout in online learning settings. XGBoost showed the best and most consistent results, having the highest Accuracy (0.7191), AUC (0.7303) and F1-score (0.5430), of all the assessed models. The model most effective in identifying potential dropouts was Logistic Regression, which has the highest Recall (0.7673), but the lower precision and generalisability of this model meant that it was not the most appropriate to use as the main predictive model. Explainability by SHAP also provided the system with extra strength since the early click activity and previous educational background emerged as the most impactful ones, delivering straightforward and understandable evidence to teachers.

The study, although promising, has a number of limitations. The fact that one dataset was used (OULAD) restricts the generalisability of the findings and the space of features was limited to the demographic variables and aggregated early engagement. More complicated behavioural indicators, like weekly activity patterns or forum activity, or assignment patterns, were not considered. In addition, the early activity was reduced to one metric thus the model failed to capture the temporal engagements. The deep learning model performance might have been limited also because of the class imbalance in addition to the basic ANN architecture.

Despite these limitations, the research has good practical implications. The system provides interpretable risk stratification which can assist tutors to give priority to students who are early signs of disengagement. SHAP explanations create confidence in the decisions made by the model since they clearly show why certain students are considered as being high-risk. The findings indicate that even the modest amounts of early-stage data could significantly separate at-risk learners, which proves the larger worthiness of early-intervention systems in online learning.

Further development of work in this direction should consider more detailed behavioural and temporal characteristics that allow more complex sequential deep-learning models like LSTM or Transformer-based models. Applicability: It could be important to test the model in various institutions to confirm its suitability in practical situations. Further developments of the deep-learning interpretability techniques and fairness analysis would also help intensify the system so that early-warning tools can be transparent, fair, and meaningful.

REFERENCE

Jha, N.I., Ghergulescu, I. and Moldovan, A.N., 2019, May. OULAD MOOC Dropout and Result Prediction using Ensemble, Deep Learning and Regression Techniques. In CSEDU (2) (pp.154-164).https://pdfs.semanticscholar.org/4d82/e06071af59bf3d74068c89f49024faa24848.pdf

Torkhani¹, W. and Rezgui, K., 2025, February. OULAD MOOC Student Performance Prediction using Machine and Deep Learning. In Proceedings of International Conference on Decision Aid and Artificial Intelligence (ICODAI 2024) (Vol. 12, p. 228). Springer Nature.OULAD - Dataset

Alghamdi, S., Soh, B. and Li, A., 2025. A Comprehensive Review of Dropout Prediction Methods Based on Multivariate Analysed Features of MOOC Platforms. Multimodal Technologies and Interaction, 9(1), p.3.https://www.mdpi.com/2414-4088/9/1/3

Villar, A. and de Andrade, C.R.V., 2024. Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study. Discover Artificial Intelligence, 4(1), p.2.https://link.springer.com/article/10.1007/s44163-023-00079-z/

Basnet, R.B., Johnson, C. and Doleck, T., 2022. Dropout prediction in Moocs using deep learning and machine learning. Education and Information Technologies, 27(8), pp.11499-11513.https://link.springer.com/article/10.1007/s10639-022-11068-7

Cho, C.H., Yu, Y.W. and Kim, H.G., 2023. A study on dropout prediction for university students using machine learning. Applied Sciences, 13(21), p.12004.https://www.mdpi.com/2076-3417/13/21/12004

Zerkouk, M., Mihoubi, M., Chikhaoui, B. and Wang, S., 2024. A machine learning based model for student's dropout prediction in online training. Education and Information Technologies, 29(12), pp.15793-15812.https://link.springer.com/article/10.1007/s10639-024-12500-w

Er, E., 2023. An explainable machine learning approach to predicting and understanding dropouts in MOOCs. Kastamonu Education Journal, 31(1), pp.143-154.https://dergipark.org.tr/en/pub/kefdergi/article/1246458

Park, H.S. and Yoo, S.J., 2021. Early dropout prediction in online learning of university using machine learning. JOIV: International Journal on Informatics Visualization, 5(4), pp.347-353.https://www.joiv.org/index.php/joiv/article/view/732

Jin, C., 2023. MOOC student dropout prediction model based on learning behavior features and parameter optimization. Interactive Learning Environments, 31(2), pp.714-732.https://www.tandfonline.com/doi/abs/10.1080/10494820.2020.1802300

Rebelo Marcolino, M., Reis Porto, T., Thompsen Primo, T., Targino, R., Ramos, V., Marques Queiroga, E., Munoz, R. and Cechinel, C., 2025. Student dropout prediction through machine learning optimization: insights from moodle log data. Scientific Reports, 15(1), pp.1-16.https://www.nature.com/articles/s41598-025-93918-1

Magalhães, E.B., Santos, G.A., Molina, F.C.D., da Costa, J.P.J., de Mendonça, F.L. and de Sousa, R.T., 2021, November. Student dropout prediction in MOOC using machine learning algorithms. In 2021 Workshop on Communication Networks and Power Systems (WCNPS) (pp. 1-6). IEEE.https://ieeexplore.ieee.org/abstract/document/9626227

Dass, S., Gary, K. and Cunningham, J., 2021. Predicting Student Dropout in Self-Paced MOOC Course Using Random Forest Model. Information 2021, 12, 476 [online]**Research Paper**

Vaarma, M. and Li, H., 2024. Predicting student dropouts with machine learning: An empirical study in Finnish higher education. Technology in Society, 76, p.102474.https://www.sciencedirect.com/science/article/pii/S0160791X24000228

Rabelo, A.M. and Zárate, L.E., 2025. A model for predicting dropout of higher education students. Data Science and Management, 8(1), pp.72-85.https://www.sciencedirect.com/science/article/pii/S2666764924000341