Lisbon School of Economics & Management
Universidade de Lisboa

MASTER'S FINAL WORK

XAI-Based Dropout Prediction and Personalized Recommendation Model in Online Education

MONSERRAT ESQUIVEL LÓPEZ

Supervisor: Professor Doctor João Afonso Bastos

Document specially prepared for the attainment of the Master's degree in Data Business Analytics for Business.

ISEG, MARCH 2026

Abstract

This Master Final Work proposes an explainable, action-oriented framework that integrates machine learning–based dropout prediction with post hoc Explainable Artificial Intelligence (XAI) techniques to generate an automated personalized educational recommendation model.

Using data from the UK Open University Learning Analytics Dataset (OULAD), the work develops and compares multiple machine learning models for dropout risk estimation, after which XAI techniques are applied to produce both global and individual-level explanations.

By automatically operationalizing XAI outputs into personalized student recommendations, the findings demonstrate that incorporating explainability substantially enhances the practical value of dropout analytics by shifting the focus from risk prediction alone toward transparent, pedagogically meaningful interpretation and decision support in online learning contexts.

Key Words: Machine Learning, Explainable Artificial Intelligence (XAI), Student Dropout Prediction, Learning Analytics, Decision Support Systems, Online Education.

INDEX

FIGURE INDEX

TABLE INDEX

1. Introduction[1]

Machine learning models (ML) can now predict student dropout risk with reasonable accuracy; however, educational institutions often struggle to translate these predictions into pedagogically meaningful insights. This creates a central paradox in dropout analytics: while predictive performance has improved, the practical interpretability and educational usefulness of model outputs remain limited. This means that most existing approaches focus primarily on predictive performance, offering limited support for **understanding why students are at risk** and **how institutions should intervene in response to model outputs**.

Unlike other domains such as customer churn prediction, where general marketing campaigns may be sufficient, educational interventions require fast interpretability, pedagogical alignment, and personalization to succeed. Without clear explanations of the factors driving dropout risk, predictive models risk remaining descriptive tools that flag students as "at risk" without informing actionable and context-sensitive educational strategies.

Explainable artificial intelligence (XAI) has emerged as a promising approach to address this gap. By revealing the contribution of different drivers to predicted dropout probabilities, XAI techniques offers insights at both individual and aggregate levels that are used to design focalized strategies.

In this context, rather than focusing exclusively on dropout risk prediction, this work emphasizes the role of explanation in bridging predictive analytics and pedagogically informed decision-making in online education by **designing an automated model that integrates the outputs of the machine learning modeling and XAI main drivers.**

By moving beyond explanation toward recommendation, the proposed approach demonstrates how explainable machine learning can be operationalized to support structured, transparent, and actionable decision-making.

[1] This work employed AI-assisted language editing (ChatGPT) to improve grammar, clarity, and writing style.

1.1. Research objectives

To design a machine learning–based framework enhanced with XAI techniques for predicting student dropout risk, generating interpretable explanations, and producing automated personalized educational recommendations that support strategic decision-making in online education to reduce future dropout rates.

1.2. Expected contributions

This work contributes to literature in three distinct ways.

First, the work provides a practical and replicable machine learning–based approach for modeling student dropout risk in online higher education. Using the OULAD dataset, it compares and tunes multiple predictive models based on standard performance metrics, identifying a well-performing model that establishes a robust analytical baseline for subsequent explanatory analysis.

Second, the work demonstrates how XAI techniques can uncover heterogeneous risk profiles among students with similar predicted dropout probabilities. By moving beyond aggregate performance measures, this contribution shows that dropout risk is not homogeneous and that explainability reveals meaningful variation in the factors driving student outcomes, insights that remain obscured in purely black-box prediction approaches.

Third, by linking model explanations to the automated generation of personalized educational support recommendations, the framework illustrates how predictive analytics can be easily translated into pedagogically meaningful insights to inform reflective intervention planning and academic decision-making, shifting the focus from "how well we predict" to "how predictions are used."

1.3. Research questions

I. Can machine learning models reliably predict student dropouts in an online education, and which models provide the best predictive performance?

II. Does XAI techniques provide meaningful interpretability that improves understanding of the drivers of student dropout risk compared to black-box models with similar predictive performance?

III. Can XAI-based explanations be used to design distinct behavioral strategies among students with similar predicted dropout probabilities?
IV. To what extent would the average dropout rate be reduced if a learning strategy were applied in advance to students at high risk of dropping out?

1.4. Hypothesis

The integration of XAI techniques into machine learning–based dropout prediction models enhances risk interpretability by identifying key dropout drivers and heterogeneous risk profiles, enabling differentiated, data-informed educational support recommendations in online learning courses for educational planning and intervention design.

2. Literature Review

2.1. Related works

Within the fields of learning analytics and educational data mining, prior studies have shown that demographic, enrollment, and behavioral interaction data from virtual learning environments can be used by machine learning models (ML) to estimate student dropout risk with reasonable accuracy. Given the extensive literature on this topic, including comparative and review studies based on datasets such as OULAD and others, a consolidated summary of related work is provided in the Appendix.

Among this extensive body of literature on student dropout prediction using ML, **two studies are particularly closely aligned with the objectives of the present research**.

The first is *Interpretable Dropout Prediction: Towards XAI-Based Personalized Intervention* by Nagy and Molontay (2024). The authors propose an interpretable dropout prediction framework that combines high-performing ML with post-hoc XAI techniques. Using pre-enrollment academic data from approximately 8,500 students at Budapest University of Technology and Economics (BME), the study relies on pre-enrollment and early academic variables, including secondary school grades, admission exam scores, the time gap between secondary education and university enrollment, demographic attributes such as gender, and program-level information.

Several predictive models were evaluated, and the results demonstrate that complex ensemble methods, particularly CatBoost, outperform linear baselines while remaining interpretable through SHAP and LIME explanations. The analysis identifies *high school grade point average (GPA)*,

mathematics-related admission exam performance, the *number of years between secondary education* and *university enrollment*, and *program-specific subject* variables as the **most influential features for predicting dropout**. A key empirical finding of the study is that dropout risk is not driven by a single dominant factor, but rather by **heterogeneous combinations of academic and enrollment-related variables**.

The findings show that local explanations can support stakeholders in understanding individual dropout risk and conceptually inform personalized interventions. However, recommendations are not operationalized within an automated framework, and **the study focuses primarily on pre-entry characteristics.**

The second relevant study is *Predicting at-Risk Students at Different Percentages of Course Length for Early Intervention Using Machine Learning Models* by Adnan, M., Habib, A., Ashraf, J., Mussadiq, S., Raza, A.A., Abid, M., Bashir, M., & Khan, S. U. (2021). The authors propose an early warning framework aimed at **identifying at-risk students at different stages of course progression**. The study is also conducted using data from 32,593 students enrolled in 22 courses offered in distance-learning at the Open University Learning Analytics Dataset (OULAD).

The analysis evaluates predictive performance at multiple temporal checkpoints corresponding to different percentages of course completion (e.g., 10%, 20%, and 30%). The feature set primarily consists of **assessment-related and course progress indicators**, such as assignment submission status, partial grades, and course completion markers available at each stage. Although OULAD includes detailed Virtual Learning Environment (VLE) interaction logs, **these interaction variables are not explicitly used as explanatory behavioral features in the modeling process**.

Logistic Regression, Decision trees, Random Forest, K-Nearest Neighbors, Naïve Bayes, and Support Vector Machines were compared. **The results show that predictive performance consistently improves as the course progresses and more data becomes available**. Among the evaluated models, **Random Forest achieves the most stable and highest predictive performance** across different temporal checkpoints, outperforming simpler linear models, particularly in later stages of the course. Across experiments, the authors report moderate to strong classification performance, highlighting a clear trade-off between **early prediction accuracy and intervention timeliness**.

Risk identification is communicated through probability thresholds or categorical risk levels (e.g., at-risk vs. not at-risk). However, the models largely operate as **black-box predictors**, with no explicit explainability techniques applied to identify the drivers of individual predictions. Consequently, the proposed interventions remain **generic and risk-level–based**, offering limited personalization grounded in interpretable factors or individual student behavior patterns.

2.2. Research gap and expected contribution

Despite substantial progress, important gaps remain in how predictive models are interpreted, operationalized, and applied in fully online learning contexts.

On the one hand, explainability-oriented approaches such as Nagy et al. (2024) show that post-hoc XAI techniques can improve transparency and help stakeholders understand individual dropout predictions. However, the study primarily rely on static pre-enrollment and early academic variables and **do not consider dynamic interaction data generated within the learning management system throughout the course**, which is automatically recorded by the system and usually readily available for analysis. Moreover, explainability is used mainly as an interpretive layer, **without being systematically linked to automated or structured intervention mechanisms that increases the efficiency of the decision making** or improve in a faster way the design for the next courses strategies.

On the other hand, temporal early warning approaches by Adnan et al. (2021), focus on identifying at-risk students at different stages of course progression and despite the availability of rich virtual learning environment interaction data in OULAD, this study do not explicitly leverage lightweight interaction signals as interpretable behavioral indicators of dropout risk.

As a result, students with similar predicted dropout probabilities are often treated as homogeneous risk cases, even though prior evidence suggests that dropout risk arises from heterogeneous combinations of academic background, enrollment behavior, and online engagement patterns.

To address these limitations, this work proposes an explainable machine learning framework that combines dropout prediction with XAI techniques. By providing global and individual explanations, the framework reveals that students with similar dropout risk can be influenced by different factors that should **guide personalized recommendations for future online students**.

3. Methodology

3.1. Research design

Using a real-world online university learning dataset as an empirical application, this work (i) develops machine learning models for early dropout prediction based OULAD data, (ii) applies XAI techniques to identify individual and global level drivers of dropout risk, and (iii) implement a model that automatically generates personalized educational recommendations for student cohorts by linking predicted risk levels with their main explanatory factors derived from SHAP values.

3.2. Data source and preparation

The Open University (OU) is one of the world's leading institutions in the provision of high-quality distance education. It was established in the United Kingdom in 1969 and is widely recognized for its inclusive educational model, which offers flexible learning opportunities regardless of students' location, age, or background. With approximately 170,000 students enrolled in a wide range of programs, the OU is the largest academic institution in the UK.

In line with its commitment to educational research, the OU released the Open University Learning Analytics Dataset (OULAD), a publicly accessible and anonymized dataset designed to support learning analytics research. OULAD contains demographic, academic, and interaction data collected from the university's virtual learning environment for courses delivered during 2013 and 2014. The dataset was made publicly available in 2015 through the UCI Machine Learning Repository and it has been widely used as a benchmark in educational data mining studies.

As previously mentioned, the OULAD dataset includes records from 32,592 students across 22 module presentations and comprises diverse types of information, such as detailed clickstream data reflecting students' interactions with the VLE.

The dataset comes as 7 separate csv files and requires processing and transformation to extract features before building prediction models. For example, student demographic information is linked with other information segments of assessments and VLE interactions.

Table 1. Summary of the OULAD dataset before feature engineering

Data File	No. of Records	Description	Attributes
courses	22	Information about the courses	code_module, code_presentation, module_presentation_length
studentInfo	32,593	Contains demographic information about the student	code_module, code_presentation, id_student, gender, region, highest_education, imd_band, age_band, num_of_prev_attempts, studied_credits, disability, **final_result**
studentRegistration	32,593	Registration of the student for a course presentation	code_module, code_presentation, id_student, **date_registration, date_unregistration**
assessments	196	Assessments for every course presentation	code_module, code_presentation, id_assessment, assessment_type, date, weight
studentAssessments	173,740	Assessments submitted by the students	id_assessment, id_student, date_submitted, is_banked, score
vle	6,365	Online learning resources and materials	id_site, code_module, code_presentation, activity_type, week_from, week_to
studentVle	1,048,575	Student interaction with the VLE resources	code_module, code_presentation, id_student, id_site, date, sum_click

Source: Table adapted from Jha, N. (2019)

Accordingly, feature engineering was performed to merge multiple information sources into a structured student-level dataset suitable for predictive modeling, incorporating demographic, registration, and VLE interaction data.

Table 2. Summary of the dataset after feature engineering.

Attributes Category	Number of Attributes	Attributes	Type
Student Identification & Course Information	3	id_student, code_module, code_presentation	Categorical
Student Demographic Information	11	gender_, region_, highest_education_, imd_band_, age_band_, disability_, num_of_prev_attempts, studied_credits	Categorical / Numeric
Registration Information	1	date_registration	Numeric
VLE Interaction – Aggregate Clicks	4	total_clicks, avg_clicks, max_clicks, num_interactions	Numeric
Target Variable	1	dropout	Binary

The feature engineering process is described in the Appendix by detailing how each group of variables was constructed and prepared for modeling.

For the **final dataset construction**, all engineered features were merged into a unified student-level dataset using *id_student* as the primary key[2]. The final dataset represents a comprehensive combination of demographic, academic, and behavioral information, and was used as input for all subsequent modeling experiments.

[2] Missing values resulting from the merging process were filled with zeros to indicate the absence of recorded activity.

Table 3. Sample- Final dataset[3]

id_ student	code_ module	code_ presentation	num_prev _attempts	studied_ credits	drop out	date_ registration	total_ clicks	avg_ clicks	max_ clicks	num_ interactions
11391	AAA	2013J	0	240	0	-159	934	4.8	76	196
28400	AAA	2013J	0	60	0	-53	1435	3.3	23	430
30268	AAA	2013J	0	60	1	-92	281	3.7	23	76
31604	AAA	2013J	0	60	0	-52	2158	3.3	22	663
32885	AAA	2013J	0	60	0	-176	1034	2.9	22	352
38053	AAA	2013J	0	60	0	-110	2445	3.4	22	723
45462	AAA	2013J	0	60	0	-67	1492	4.2	48	355
45642	AAA	2013J	0	120	0	-29	1428	2.7	46	531
52130	AAA	2013J	0	90	0	-33	1894	3.2	34	593

3.3. Predictive modeling[4]

Machine learning models are techniques within the field of artificial intelligence that enable computers to "learn" with data without being explicitly programmed.

Predictive modeling refers to the set of statistical and machine learning methods that aim to learn a function f relating a feature vector $\mathbf{X}$ to an outcome Y, in order to predict future or unseen values of Y (James et al. 2021).

In binary classification settings, the objective is to learn a function that separates the two classes as effectively as possible based on the observed predictors. This can be achieved using parametric models such as logistic regression or nonparametric approaches such as tree-based ensembles.

James et al. (2021) describe the predictive modeling process as involving data splitting into training and test sets, model fitting, model selection based on appropriate performance metrics (e.g., accuracy, ROC-AUC), and final assessment on an independent test set, with cross-validation playing a central role in estimating generalization performance.

The following subsections describe the main algorithms considered in this work, outlining their key characteristics, mathematical formulations and **including their application in context of student dropout prediction**.

[3] The hidden cells correspond to the categorical variables. They are hidden for the restriction of the space.

[4] The development of this theoretical review (sections 3.3 to 3.5) was assisted by the Perplexity Pro for Education AI system, employed as a complementary tool for literature search, synthesis of information from scientific sources, and clarification of key concepts. Content was reviewed and adapted for the context of student dropout prediction.

3.3.1. Logistic Regression

Logistic regression is a widely used model for binary classification that approximates the probability of class membership through the logistic function applied to a linear combination of predictors (James et al. 2021). It is a natural baseline for classification problems, offering a good compromise between predictive performance and model transparency when the number of predictors is moderate.

Given $Y \in \{0,1\}$ and $\mathbf{X} = (X_1, \ldots, X_p)$, the model assumes that the log-odds of $P(Y = 1 \mid \mathbf{X})$ are linear in $\mathbf{X}$:

$$\log\left(\frac{P(Y = 1 \mid \mathbf{X})}{1 - P(Y = 1 \mid \mathbf{X})}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \tag{3.1}$$

Equivalently, the conditional probability can be written as:

$$P(Y = 1 \mid \mathbf{X}) = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)} \tag{3.2}$$

The coefficients are typically estimated via maximum likelihood, and $\exp(\beta_j)$ can be interpreted as an odds ratio, which makes logistic regression particularly attractive when interpretability is important.

In the context of student dropout prediction, the resulting coefficients can be interpreted to quantify how changes in these features are associated with an increase or decrease in dropout risk.

3.3.2. RF- Random Forest

Random Forest is an ensemble method that extends bagging for decision trees by introducing an additional layer of randomness. The algorithm works as follows:

- Create a set of samples by sampling with replacement the available observations.
- For each sample, grow a tree using a modified tree learning algorithm that selects, at each candidate split, a random subset of the features.

This additional randomness increases diversity among the trees and reduces correlation between them, leading to a lower variance ensemble compared to a single decision tree without substantially increasing bias.

If B denotes the number of trees and $\mathrm{F}_b(\mathbf{x})$ the prediction of the b-th tree, the Random Forest predicted value for a test observation in a regression problem is:

$$\mathrm{f}_{\mathrm{RF}}(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} \mathrm{F}_b\,(\mathbf{x}) \tag{3.3}$$

In a classification problem, the class predicted by the majority of the base models is the predicted class for a test observation.

For student dropout prediction, Random Forest capture complex nonlinear relationships and interactions between student characteristics and behavioral indicators that may be difficult to model with simpler methods.

3.3.3. XGBoost (eXtreme Gradient Boosting)

XGBoost is an efficient implementation of gradient boosting for decision trees, in which the model is built additively by fitting new trees that correct the errors of the existing ensemble. In a general boosting framework, starting from an initial model $\hat{f}_0(\mathbf{x})$, at each iteration m a new base learner $h_m(\mathbf{x})$ is added:

$$\hat{f}_m(\mathbf{x}) = \hat{f}_{m-1}(\mathbf{x}) + \nu\, h_m(\mathbf{x}) \tag{3.4}$$

where ν is the learning rate, and h_m is fitted to approximate the negative gradient of the loss function in function space. XGBoost includes explicit regularization terms to control tree complexity, efficient handling of missing values, parallelization capabilities, and support for various loss functions, which often lead to strong performance on real-world classification and regression tasks. In practice, hyperparameters such as the *number of trees*, *maximum depth*, *learning rate*, and $L1/L2$ regularization parameters play a crucial role in balancing fit to the training data and generalization to new observations.

For dropout prediction, XGBoost is well suited to modeling heterogeneous and nonlinear relationships between student behavior and dropout risk, as successive trees iteratively refine prediction errors.

3.3.4. LightGBM (Light Gradient Boosting Machine)

LightGBM is a gradient boosting framework based on decision trees that is designed to be highly efficient in terms of computation and memory, especially for large datasets with many features.

In particular, LightGBM can be viewed as a specific implementation of boosting that adopts a leaf-wise tree growth strategy with depth constraints and uses histogram-based splits to accelerate training. These design choices preserve the functional form of gradient boosting described by James et al. (2021), while allowing the method to scale to high-dimensional and large-sample problems.

Applied to student dropout prediction, LightGBM is especially attractive when working with large cohorts and many input features. Its efficiency and ability to handle high-dimensional tabular data make it suitable for real-time or frequently updated risk scoring pipelines in online education settings.

3.3.5. CatBoost (Categorical Boosting)

CatBoost is a gradient boosting algorithm over decision trees that is specifically designed to handle categorical features efficiently by using target statistics computed in an ordered manner to avoid target leakage. The algorithm employs symmetric trees where the same splitting condition is used at each level, which simplifies tree structure and enables highly optimized implementations.

It fits into the same gradient boosting paradigm, where successive trees are added to minimize a chosen loss function via functional gradient descent, with a learning rate and a fixed number of boosting iterations. Practical guides report that CatBoost typically achieves strong performance on heterogeneous tabular datasets with relatively modest hyperparameter tuning, especially in the presence of many categorical inputs.

In the dropout prediction setting, CatBoost is particularly useful because many relevant predictors are inherently categorical and can be processed natively by the algorithm. This allows

the model to leverage rich categorical information about students and courses without extensive manual encoding, while still providing competitive predictive performance for identifying at-risk students.

To ensure that such predictive models are not only accurate but also robust and generalizable, hyperparameter optimization is a key step in practical machine learning applications.

3.3.6. Hyperparameter optimization

In machine learning, a hyperparameter is a configuration value that is set before the training process begins and that controls aspects of the learning algorithm or model structure, such as model complexity or learning dynamics, but is not directly learned from the data.

Unlike model parameters, which are estimated during training (for example, the coefficients in logistic regression or the split points in decision trees), hyperparameters include choices such as the regularization strength in logistic regression, the number and depth of trees in Random Forests, or the learning rate and number of boosting iterations in gradient boosting methods.

Hyperparameter optimization or "tuning" is the process of selecting, within a predefined search space, the combination of hyperparameters that maximizes a chosen performance metric estimated on validation data.

James et al. (2021) describe Grid Search and Random Search, often combined with K-fold cross-validation, as standard strategies for hyperparameter tuning. In Grid Search, a finite set of candidate values is specified for each hyperparameter and all possible combinations in this grid are evaluated, typically by cross-validation, in order to identify the configuration that yields the best average performance on the training/validation data.

In the context of student dropout prediction in online courses, such **predictive modeling will be used to estimate the probability that a given student will drop out** based on demographic, academic, and interaction features collected from the learning platform.

However, most **classification algorithms initially produce numerical scores rather than calibrated probabilitie**s. In binary classification problems, these scores are associated with each observation and reflect the model's confidence about class membership. Their primary purpose

is to separate the two classes as effectively as possible, ranking students from higher to lower risk.

From a decision-theoretical perspective, the fundamental objective of a classifier is discrimination, that is, to distinguish between classes as accurately as possible. The numerical output is therefore first understood as a scoring or ranking mechanism, and it should not automatically be interpreted as a probability. **Its probabilistic meaning requires additional assumptions and empirical validation through calibration analysis**.

Recent methodological literature highlights that predictive scores should only be interpreted as probabilities if they are empirically well aligned with observed frequencies. This perspective underscores the distinction between predictive scores and calibrated probabilities. A model may exhibit strong discriminative performance (e.g., high ROC-AUC) while still producing poorly calibrated outputs. Thus, discrimination and calibration represent complementary but distinct properties of predictive models.

Accordingly, the predicted scores from the models presented in the previous section were subjected to a calibration analysis (section 4.2) to evaluate whether they are sufficiently aligned with observed dropout frequencies, justifying their interpretation as estimated probabilities of dropout.

3.4. Selected XAI techniques

Beyond global performance metrics, XAI techniques provide tools to understand how complex models arrive at their predictions. These methods can highlight which features drive the predictions, how changes in specific variables affect the output, and how individual instances are explained by the model, thereby supporting trust, debugging, and domain insight. As Molnar (2022) also states, "the more interpretable a machine learning model, the easier it is for someone to understand why certain decisions or predictions were made."

The relevance of interpretability becomes evident in applied predictive settings where high predictive accuracy alone is insufficient. Performance metrics such as accuracy or AUC provide only a partial description of real-world decision problems, particularly in high-stakes or complex domains. In such cases, understanding *why* a model produces a given prediction enables deeper insight into the data-generating process, supports error analysis, and facilitates informed

intervention design. This perspective motivates the selection of XAI techniques in this work, which are intended not only to explain model predictions but to support decision-making in an educational context where transparency and actionability are essential.

The following subsections present the definition and objective of the techniques used in the work.

3.4.1. Tree-based Feature Importance (impurity / gain)

Tree-based ensembles, such as Random Forests and Gradient Boosting Machines, provide built-in measures of global feature importance, which are based on how frequently and effectively each feature is used to split the data. In these models, importance is typically computed by aggregating, over all trees, the reduction in an impurity or loss criterion (e.g., Gini impurity, entropy, or training loss) that results from splits on a given feature.

For a single decision tree, the importance of feature X_j can be defined as the sum, over all internal nodes that split on X_j, of the impurity decrease at that node weighted by the proportion of training samples reaching it. In an ensemble, the global importance of X_j is obtained by averaging or summing these node-level contributions across all trees, and the resulting feature importances are usually normalized to sum to one.

In the context of student dropout prediction, tree-based feature importance is used to identify which variables contribute most to the model's ability to separate dropouts from non-dropouts.

3.4.2. SHapley Additive ExPlanations (SHAP)

SHapley Additive exPlanations (SHAP) are a family of explanation methods that attribute a model prediction for an individual instance to its input features based on concepts from cooperative game theory (Molnar, 2022).

SHAP values are grounded in Shapley values, which distribute the "payout" of a cooperative game fairly among players according to their contributions across all possible coalitions.

For a model prediction $f(\mathbf{x})$ and a baseline value ϕ_0 (often the expected prediction), SHAP represents the prediction as an additive decomposition:

$$f(\mathbf{x}) = \phi_0 + \sum_{j=1}^{p} \phi_j \tag{3.5}$$

where ϕ_j is the SHAP value for feature j, interpreted as the contribution of X_j to the deviation of $f(\mathbf{x})$ from the baseline (Molnar, 2022). The Shapley value ϕ_j is defined as the average marginal contribution of feature j across all subsets S of features that do not contain j:

$$\phi_j = \sum_{S \subseteq \{1,\ldots,p\} \setminus \{j\}} \frac{|S|!\,(p-|S|-1)!}{p!} \left[f_{S \cup \{j\}}(\mathbf{x}) - f_S(\mathbf{x}) \right] \tag{3.6}$$

where $f_S(\mathbf{x})$ denotes the model evaluated using only the features in subset S, with the remaining features integrated out or fixed according to a chosen background distribution (Molnar, 2022).

SHAP explanations satisfy desirable properties such as local accuracy, consistency, and missingness, which makes them attractive for explaining predictions of complex models such as gradient boosting machines (Molnar, 2022).

3.4.3. SHAP Feature importance

SHAP feature importance is a global interpretability technique that derives feature importance scores from local SHAP values. For each instance, SHAP decomposes the prediction into feature-wise contributions; global importance is then obtained by aggregating these contributions across all instances, typically by taking the mean absolute SHAP value for each feature (Molnar, 2022). Formally, if ϕ_{ij} denotes the SHAP value of feature j for instance i, and n is the number of instances, the global importance of feature j can be defined as

$$\mathrm{Imp}_{\mathrm{SHAP}}(X_j) = \frac{1}{n} \sum_{i=1}^{n} |\, \phi_{ij} \,| \tag{3.7}$$

Features with larger average absolute SHAP values are considered more important, since they contribute more, on average, to pushing predictions away from the baseline.

In the context of student dropout prediction, permutation feature importance can be used to identify which variables contribute most to the model's ability to distinguish between dropouts and non-dropouts.

3.4.4. Force Plot

Force plots are a visualization technique for local explanations that represent how individual feature contributions push a model prediction away from a reference value (Molnar, 2022). They are commonly used together with SHAP values, where the baseline ϕ_0 is the expected prediction and each feature's SHAP value ϕ_j is visualized as a force that either increases or decreases the prediction for a given instance.

In a typical SHAP force plot, features that increase the prediction relative to the baseline are shown in one color (e.g., pink) pushing the prediction to the right, while features that decrease it are shown in another color (e.g., blue) pushing to the left. (Molnar, 2022). The length of each bar is proportional to the magnitude of the corresponding SHAP value, and the final prediction $f(\mathbf{x})$ is given by the baseline plus the sum of all contributions:

$$f(\mathbf{x}) = \phi_0 + \sum_{j=1}^{p} \phi_j \tag{3.8}$$

Force plots provide an intuitive, instance-level explanation by making it visually clear which features are responsible for driving a prediction above or below the baseline, and by how much (Molnar, 2022). This makes it particularly appealing for communicating explanations of complex models to non-technical stakeholders.

In dropout prediction applications, SHAP-based force plots can visually show, for a given student, which factors push the predicted risk upwards and which factors push it downwards. Such visual explanations enhance the transparency of complex ensemble models, enabling educators and decision-makers to better understand the drivers behind predictions and increase confidence in the model's outputs.

3.4.5. Accumulated Local Effects (ALE)

Accumulated Local Effects (ALE) plots describe how a feature affects the prediction on average when the features are correlated. ALE is also based on the conditional distribution of the features differences in predictions instead of averages. ALE plots are centered at zero. This makes their interpretation nice because the value at each point of the ALE curve is the difference to the mean prediction (Molnar, 2022).

Instead of averaging predictions over the entire feature distribution, ALE focuses on local changes in the prediction when a feature varies within small intervals and then accumulates these local effects.

For a continuous feature X_j, its domain is partitioned into intervals $[z_{k-1}, z_k]$. Within each interval, the local effect is approximated by the average difference in predictions when X_j moves from z_{k-1} to z_k, keeping all other features fixed:

$$\Delta f_{j,k} = \frac{1}{n_k} \sum_{i:x_{i,j} \in [z_{k-1}, z_k]} \left[f(z_k, \mathbf{x}_{i-j}) - f(z_{k-1}, \mathbf{x}_{i-j}) \right] \tag{3.9}$$

where n_k is the number of instances with $x_{i,j}$ in that interval and $\mathbf{x}_{i-j}$ denotes all features except X_j. The ALE function for feature X_j is obtained by accumulating these local effects from a lower bound $z_{\min}$ up to a value z:

$$\text{ALE}_j(z) = \sum_{k:z_k \leq z} \Delta f_{j,k} \tag{3.10}$$

ALE plots illustrate how predictions change as X_j varies across their observed range, considering only feature combinations present in the data. This approach minimizes extrapolation artifacts and provides a more reliable interpretation of feature effects.

For student dropout models, ALE plots can be used to visualize how changes in a particular feature, such as *total clicks* or *number of interactions*, affect the predicted dropout risk on average, while considering only feature combinations that actually occur in the data to avoid misleading effects from correlated variables. These plots help identify ranges of feature values associated

with increased or decreased dropout risk, providing actionable insights for course design and support strategies.

3.5. Evaluation metrics

Evaluation metrics are quantitative measures used to assess the performance of a predictive model on a given task, typically computed on a test set or via cross-validation. For binary classification, many metrics are derived from the confusion matrix, which summarizes the counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

Different metrics capture different aspects of performance and are explained in Table 4:

- **Accuracy** focuses on overall correctness.
- **Precision** and **Recall** quantify performance in the positive class from different perspectives.
- **F1** score balances precision and recall.
- **ROC-AUC** evaluates the trade-off between true positive and false positive rates across all possible thresholds.

Table 4. Classification Performance Metrics

Metric	Definition	Formula	Interpretation
Accuracy	Proportion of correctly classified instances over the total number of instances.	$\frac{TP + TN}{TP + TN + FP + FN}$	Indicates overall correctness of the classifier.
Precision	Proportion of predicted positive cases that are actually positive.	$\frac{TP}{TP + FP}$	High precision means that when the model predicts the positive class, it is usually correct.
Recall (Sensitivity)	Proportion of actual positive cases correctly identified by the classifier.	$\frac{TP}{TP + FN}$	High recall indicates that the model successfully detects most positive instances.

F1 Score	Harmonic mean of precision and recall.	$2\frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$	Provides a balanced measure between precision and recall. It penalizes large discrepancies between them, yielding a low value if one metric is substantially lower than the other.
ROC-AUC	Area under the Receiver Operating Characteristic curve, which plots TPR against FPR across thresholds.	$\text{TPR}=\frac{\text{TP}}{\text{TP}+\text{FN}}$ $\text{FPR}=\frac{\text{FP}}{\text{FP}+\text{TN}}$	AUC $\in$ [0,1], measures the model's ability to discriminate between classes across all thresholds. A value of 0.5 indicates random performance, while 1 represents perfect discrimination.

Now, in the specific context of student dropout prediction, the entries of the confusion matrix can be interpreted as follows:

- True positives (TP) are students who actually drop out and are correctly identified as at risk by the model.
- True negatives (TN) are students who complete the course and are correctly classified as non-dropouts.
- False positives (FP) are students who complete the course but are incorrectly flagged as potential dropouts; and
- False negatives (FN) are students who drop out but are not identified by the model.

So,

- **Accuracy** reflects the overall proportion of correctly classified students, considering both dropouts and non-dropouts.
- **Recall** reflects the model's ability to capture at-risk students (minimizing FN).
- **Precision** indicates how reliable the dropout alerts are (controlling FP).
- **F1 score** balances both aspects.
- **ROC-AUC** summarizes the overall discriminative ability of the model across all possible thresholds.

4. Results and Analysis

We used the predictive models[5] described in section 3 to estimate the probability that a given student will drop out based on demographic, academic and interaction patterns collected from the learning platform.

4.1. ML model selection

This section show the results of implementing and evaluating the five machine learning models considered in this work. All models were trained and validated under identical experimental conditions to ensure a fair comparison. Specifically, the dataset was split into training and testing subsets using a **stratified 70/30 to** ensure a sufficiently large test set for reliable performance evaluation, while retaining enough data for model training.

Their performance was assessed using the standard evaluation metrics described also in Section 3, which are reported in Figure 1.

Figure 1. ML performance metrics (Optimized)

	Accuracy	Precision	Recall	F1 Score	ROC-AUC
LightGBM	0.811	0.742	0.603	0.665	0.874
CatBoost	0.813	0.750	0.602	0.668	0.874
XGBoost	0.812	0.744	0.603	0.666	0.873
Logistic_Regression	0.763	0.593	0.766	0.668	0.842
RandomForest	0.771	0.649	0.575	0.610	0.825

BEST MODEL SELECTED: CatBoost

Given that CatBoost and LightGBM achieved identical ROC-AUC values (0.874), while CatBoost outperformed LightGBM in three additional evaluation metrics, a statistical comparison of their discriminative performance was conducted using **DeLong's test** for two correlated ROC curves.

[5] The complete machine learning pipeline, including data preprocessing, model training procedures, hyperparameter tuning, calibration steps, and generated output datasets, is publicly available at: https://github.com/moonel-master/XAI-Dropout-Prediction.git. The repository is provided to ensure transparency, reproducibility, and accessibility of the experimental implementation described in this thesis.

The null hypothesis assumes that both models have equal areas under the curve (AUC), whereas the alternative hypothesis states a difference between their AUC values, formally expressed as:

$$H_0: \mathrm{AUC}_{\mathrm{CatBoost}} = \mathrm{AUC}_{\mathrm{LightGBM}}$$
$$H_1: \mathrm{AUC}_{\mathrm{CatBoost}} \neq \mathrm{AUC}_{\mathrm{LightGBM}}$$

This non-parametric test accounts for the dependency between ROC curves evaluated on the same dataset. The results indicated no statistically significant difference between CatBoost and LightGBM in terms of ROC-AUC ($p = 0.981$), suggesting comparable discriminative performance.

Therefore, **CatBoost was selected as the final model** based on its superior overall performance across multiple evaluation metrics.

4.2. Calibration Analysis

After selecting the best-performing model based on discriminative metrics, a calibration analysis was conducted to assess whether the predicted scores are aligned with observed dropout frequencies and can therefore be interpreted as estimated probabilities.

Calibration refers to the agreement between predicted probabilities and actual observed outcomes. A well-calibrated model assigns probabilities that match empirical frequencies. For example, among students predicted to have a 70% dropout probability, approximately 70% should actually drop out.

To evaluate calibration, both graphical and quantitative tools were employed. Detailed results are also reported in the Appendix.

Taken together, the **calibration curve**, **Q–Q plot**, **distribution analysis**, and a **Brier score of 0.128686** indicate that the selected CatBoost model is not only discriminative but also well calibrated. Therefore, **the predicted scores can be reliably interpreted as estimated probabilities of dropout.**

This validation supports their subsequent use for risk stratification, threshold-based categorization, and personalized recommendation generation. An extract of these predicted probabilities is presented in Table 5.

Table 5. CatBoost prediction for a sample of students

id_student	prob_dropout
11391	55.8%
28400	10.6%
30268	55.0%
31604	7.7%
32885	21.4%
38053	10.2%

Consistent with the objective of this work, the next step is to complement these quantitative results by moving beyond aggregate performance measures. The subsequent section therefore presents the XAI analysis, aimed at identifying and interpreting the key factors underlying the predicted dropout risk in online learning contexts.

4.3. XAI results and risk drivers

Figure 2 shows the global feature importance across Tree-based Ensemble models where it is possible to highlight by model the engagement variables and academic characteristics that most strongly influence whether a student is classified as at risk of dropping out or likely to persist.

Figure 2. Global tree-based top feature importance across ensemble models

Feature	RF	XGB	LGBM	CatBoost
Days from start	0.0302	0.0156	798	6.2178
Number of Interactions	0.2293	0.1358	746	22.2636
Average Clicks	0.1483	0.0187	692	6.3338
Maximum Clicks	0.1712	0.0161	624	5.9269
Total Clicks	0.2172	0.0220	602	13.0664
Credits Studied	0.0776	0.0245	400	6.3425

Feature importance values are not directly comparable across models, as each algorithm computes importance using different criteria. Random Forest (RF) estimates importance based on mean impurity reduction, while gradient boosting models such as XGBoost (XGB),

LightGBM (LGBM), and CatBoost derive importance from model-specific measures related to split gain or contribution to loss reduction. Therefore, values should be interpreted within each model rather than compared in absolute magnitude across models.

Despite the differences in importance scales across models, consistent patterns can be observed. The results indicate that variables related to **student interaction** with the Virtual Learning Environment account for the highest importance across all models. Features such as *Number of Interactions*, *Total Clicks*, *Average Clicks*, and *Maximum Clicks* consistently rank among the most relevant, suggesting that the frequency and intensity of student activity are closely associated with dropout risk.

In particular, the **CatBoost** model assigns the highest importance to *Number of Interactions* (22.2636) and *Total Clicks* (13.0664), which means that, within this model, changes in these variables have a strong impact on the predicted dropout probability compared to other features, reinforcing the relevance of behavioral engagement metrics for predicting the dropout rate.

While the previous table report impurity-based feature importance obtained from the internal mechanisms of each tree-based ensemble model, Figure 3 shows the SHAP Global feature importance for the final CatBoost model, computed as the mean absolute SHAP value of each feature across all students, reflecting the magnitude of each feature's contribution to the predictions, so features with large absolute Shapley values are important.

Figure 3. Top 10 SHAP Global features importance for CatBoost model

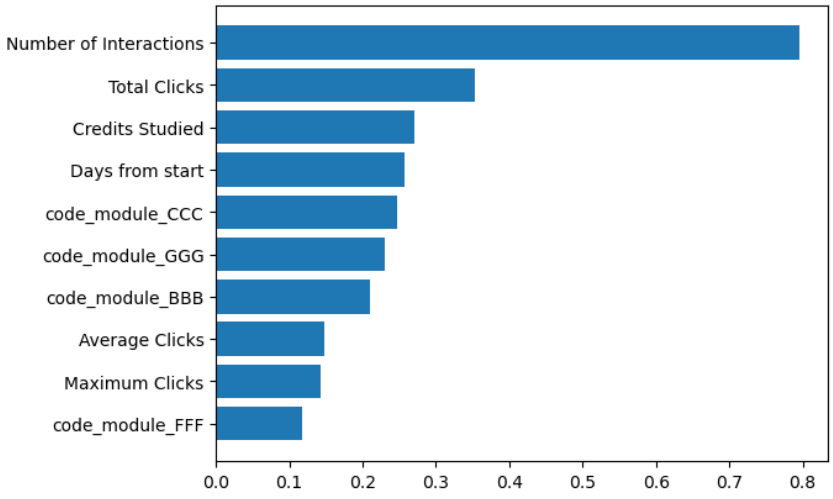


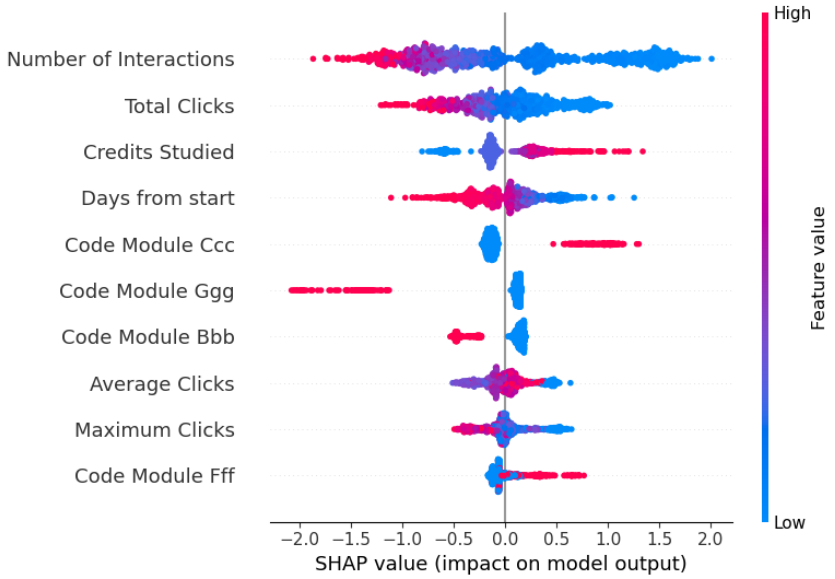
Figure 3 confirms the central role of engagement variables: *Number of Interactions* and *Total Clicks* appear as the two most influential predictors, followed *by Credits Studied*, which indicates that both interaction activity and academic load are key drivers of the estimated dropout probability.

In addition, categorical variables representing the modules in which students are enrolled provide complementary information. In Section 4.5, we analyze how enrollment in specific modules, or in any module in general, is associated with an increased probability of dropout, highlighting how course selection interacts with engagement and academic load to influence student retention.

Building on the global SHAP feature importance results, Figure 4 provides a summary plot that combines feature importance with feature effects. The features are ordered according to their importance and illustrate how individual feature values influence the CatBoost model's predicted dropout probability for each student.

This plot shows one point per student and feature, where the horizontal axis represents the SHAP value (the impact on the model output), and the color encodes whether the feature value is low (blue) or high (pink).

Figure 4. SHAP global (Beeswarm)



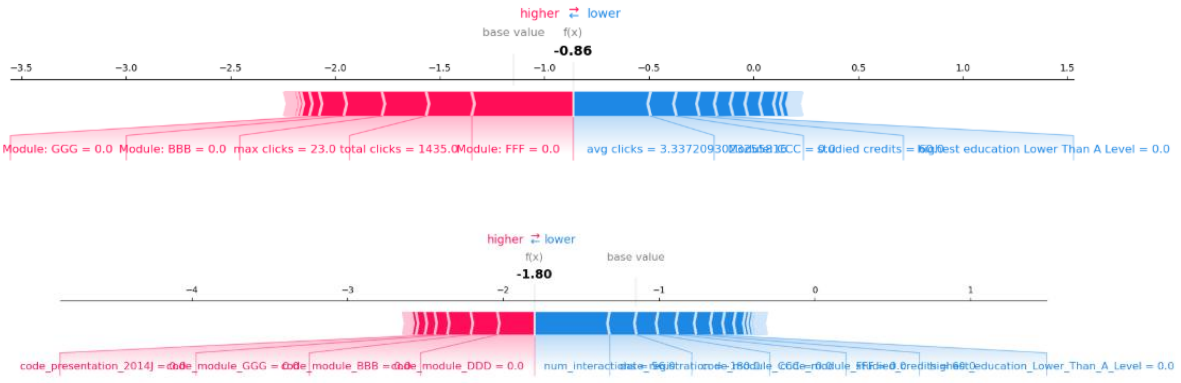
For key engagement variables such as *Number of Interactions* and *Total Clicks*, low values (blue) appear on the left with negative SHAP values, while high values (pink) cluster on the right with positive SHAP values, **indicating that limited activity in the virtual campus pushes predictions toward higher dropout risk**, whereas sustained interaction shifts predictions toward persistence.

Similarly, patterns for *Credits Studied* and *Registration Days from start (*first day student registered relative to course start) show specific combinations of academic load and enrollment timing can either increase or decrease the estimated dropout risk, adding an instance-level perspective that complements the purely global rankings from the previous figures.

Another powerful tool are the SHAP Force Plots, which connect the global interpretability results with the core objective of this work: moving from accurate dropout predictions to actionable, student-level recommendations.

While previous figures identified the main engagement and academic variables driving dropout risk at the cohort level, the Force Plot decomposes the prediction for an individual student into positive and negative SHAP contributions, showing exactly which factors push the estimated dropout probability up or down relative to the baseline.

Figure 5. SHAP local explanation- Student 1 and Student 15



The pink bars on the left correspond to features whose SHAP values push the prediction towards higher dropout risk for these students; the longer the bar, the stronger that

variable increases the predicted probability of dropping out, while the blue bars on the right correspond to features whose SHAP values push the prediction towards lower dropout risk.

Force Plots are used to illustrate how individual predictions are composed from multiple contributing factors. However, due to their fine-grained and non-aggregated nature, they are not directly suitable for decision-making.

Therefore, in this framework, the automated Recommendation Model is derived from the features with the top largest positive SHAP values for each student, as these variables represent the strongest or **main drivers** of their predicted dropout outcome.

Finally, from the global perspective, these two panels (Figure 6 and Figure 7) present ALE plots that describe how platform engagement shapes the CatBoost model's predicted dropout probability.

Figure 6. ALE of student interaction frequency on Dropout Risk

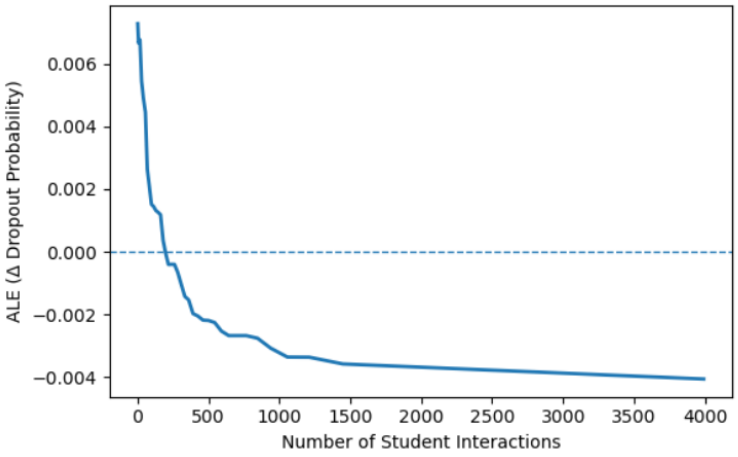
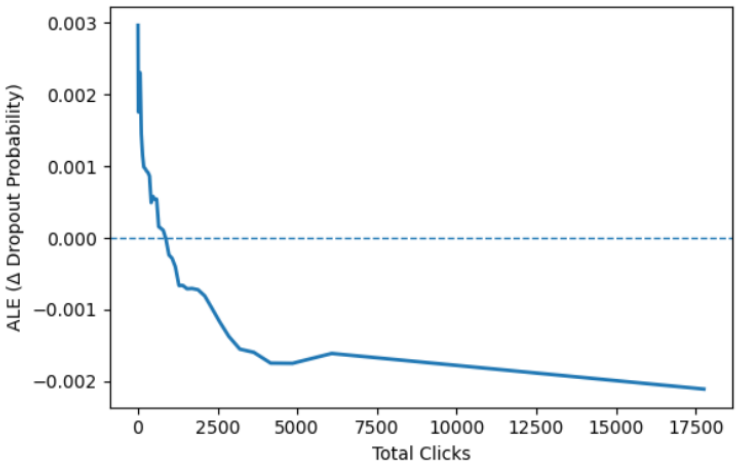


Figure 7. ALE of platform activity volume on Dropout Risk



In both cases, very low levels of activity (few interactions or clicks) are associated with positive ALE values, meaning that the **model's predicted dropout probability increases when students hardly engage with the platform**. As interactions increase, the ALE curves quickly fall below zero and then gradually flatten, indicating that higher engagement substantially reduces dropout risk at the beginning, but that additional activity beyond moderate levels leads to only marginal improvements in the model's predictions.

4.4. Implementation of automated Recommendation Model

This section describes the complete implementation of the personalized Recommendation Model developed in this work. The objective is to integrate the predictive dropout probabilities generated by the CatBoost model with XAI results for OULAD students, **providing a concrete case study designed to serve as a reference for other institutions that collect data with similar characteristics and variable structures to those used in this dataset**.

Step 1: Extraction of individual SHAP explanations

The first step consists of computing local SHAP values for all students using the full set of predictor variables employed by the predictive model. This ensures full alignment between prediction, explanation, and downstream recommendations.

For each student, the model identifies the top SHAP drivers based on the absolute magnitude of their contribution to the predicted dropout probability. These drivers represent the most influential factors shaping each individual prediction.

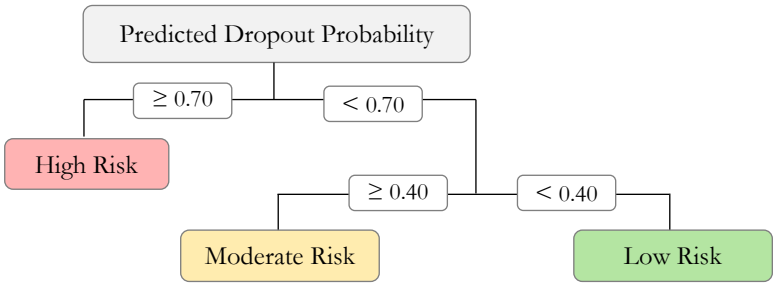
The output of this step is summarized in the variable *Main_Risk_Drivers*, which stores the three most influential features per student, including both the feature name and its signed SHAP value. This information serves as the analytical foundation for all subsequent explanation and recommendation components.

Step 2: Risk level stratification

To enable differentiated institutional responses, predicted dropout probabilities are grouped into three interpretable risk levels: High Risk, Moderate Risk, and Low Risk.

This stratification (Figure 8) allows the framework to scale the intensity of recommendations according to the severity of the predicted risk, ensuring proportionality and avoiding over-intervention for low-risk students. The resulting categorical variable *Risk Level* is used consistently across explanation narratives, action mappings, and evaluation tables.

Figure 8. Model-based Risk Level thresholds



Step 3: Generation of local explanatory narratives ("WHY explanation")

Before generating recommendations, a concise explanatory text is produced for each prediction. This narrative combines the predicted probability of dropout with the corresponding dominant SHAP drivers.

The purpose of this step is to provide transparency and interpretability for academic stakeholders, allowing them to understand *why* a student has been classified at a given risk level before considering any intervention. These explanations are descriptive rather than prescriptive and are not used directly to trigger actions yet.

Step 4: Definition of Driver categories

To enhance interpretability and enable structured recommendation generation, individual SHAP drivers **were grouped into categories**. This categorization allows the system to translate feature-level contributions into meaningful **intervention dimensions aligned with student engagement patterns, academic background, and institutional context**.

For example, engagement-related variables were explicitly separated into three complementary dimensions: *engagement volume* (how much the student interacts with the platform), *engagement regularity* (how consistent interactions are over time), and *engagement intensity* (how concentrated interactions are in short periods). This separation allows the recommendation model to distinguish between different engagement patterns that may have distinct pedagogical implications. Additional categories capture academic history, educational background, course load, enrollment timing, and declared disability. Table 6[6] summarizes the mapping between predictive variables and their corresponding driver categories.

Table 6. Driver Categories for the Personalized Recommendation Engine

Variable	Category	Interpretation
total_clicks	Engagement_Volume	Overall amount of activity generated by the student within the platform.
num_interactions		Frequency of recorded interactions across learning resources and activities.
avg_clicks	Engagement_Regularity	Average level of activity over time, reflecting how consistently the student participates.
max_clicks	Engagement_Intensity	Peak interaction level, indicating bursts of concentrated engagement.
num_of_prev_attempts	Academic_History	Number of previous enrollment attempts, reflecting prior academic persistence or difficulty.
highest_education	Educational_Background	Highest formal education level attained prior to enrollment.
studied_credits	Course_Load	Number of credits currently undertaken, indicating study intensity.
date_registration	Enrollment_Timing	Timing of course registration, potentially reflecting planning or last-minute enrollment patterns.
disability	Disability_Declared	Self-declared disability status, relevant for contextual and support considerations.

Step 5: Risk-level–dependent strategy mapping

[6] Note: Enrollment_Timing and Disability_Declared variables are treated as post hoc interpretative signals rather than causal predictors and are used exclusively to contextualize recommendations rather than to infer early dropout risk. This categorization step provides an abstraction layer that bridges raw model features and higher-level educational reasoning.

For each combination of predicted dropout risk level (High, Moderate, Low) and driver category, **a corresponding set of action-oriented educational strategies** is specified. These strategies are formulated as **concrete actions** (Table 7) that academic coordinators or support teams could realistically implement by running the model before the end of the semester.

Table 7. Design of Institutional actions by Risk Levels and Drivers Categories

	Institutional Action (part 1)			
Risk Level	**Engagement_ volume**	**Engagement_ regularity**	**Engagement_ intensity**	**Academic history**
High Risk	Initiate immediate outreach to increase overall participation, including mandatory academic follow-up, structured use of learning resources, and close monitoring of weekly activity levels.	Implement structured study planning support by defining weekly routines, establishing fixed checkpoints, and providing time-management guidance.	Address reactive or last-minute study behavior by providing pacing guidance, workload distribution strategies, and early assessment planning support.	Provide proactive academic advising focused on prior learning difficulties, including tutoring referrals and individualized learning plans.
Moderate Risk	Encourage increased participation through targeted reminders, guided use of course materials, and optional academic support sessions.	Support the development of more consistent study habits through light-touch mentoring and periodic progress reminders.	Encourage smoother distribution of study effort across the course timeline and promote early interaction with learning materials.	Offer targeted academic guidance and monitor for early signs of recurring difficulties.
Low Risk	Maintain current participation levels through standard academic follow-up and periodic engagement validation.	Reinforce existing engagement patterns through positive feedback and confirmation of effective study routines.	Monitor engagement peaks to confirm sustainable study intensity and prevent future overload.	Continue standard academic monitoring to ensure previous difficulties do not re-emerge.

	Institutional Action (part 2)			
Risk Level	**Educational_ background**	**Course_ Load**	**Enrollment_timing**	**Disability_ Declared**
High Risk	Provide additional academic orientation and foundational learning support tailored to the student's prior educational background.	Review and adjust course load where possible, prioritizing essential modules and aligning workload with student capacity.	Strengthen early-course academic orientation and clarify expectations to support smoother adaptation.	Review whether digital learning materials, assessments, and platform interactions are accessible and sufficiently flexible.
Moderate Risk	Monitor academic adaptation and offer support if learning gaps emerge.	Assess workload balance and provide guidance on managing study commitments effectively.	Provide additional guidance during initial course phases to support adjustment.	Ensure learning resources are clearly organized, navigation is intuitive, and support channels are visible if needed.
Low Risk	Ensure continued access to standard academic orientation resources and self-support materials.	Maintain current course load and periodically validate workload sustainability.	Confirm that standard academic orientation was sufficient, and no additional onboarding support is required.	Maintain inclusive digital design and accessible learning materials.

The content and intensity of the recommended actions vary systematically with the student's risk level. For example, high-risk students receive proactive and structured interventions, whereas low-risk students are primarily subject to monitoring and reinforcement. This design ensures that recommendations are not only personalized in terms of content but also proportionate in terms of intervention intensity.

Step 6: Selection of actionable drivers using local SHAP values

For each student, the recommendation model processes the ranked local SHAP explanations generated by the final predictive model. Only features with positive SHAP values are considered, as these represent factors that actively increase the predicted probability of dropout.

This restriction is intentional: features with negative SHAP values are interpreted as protective factors and therefore do not require corrective intervention. By focusing exclusively on risk-increasing drivers, the model avoids generating unnecessary or contradictory recommendations.

The positive SHAP drivers are then ranked by magnitude, ensuring that the most influential contributors to dropout risk are prioritized in the recommendation process.

Step 7: Prioritized recommendation generation

In the final step, the model maps the top-ranked positive SHAP drivers to their corresponding driver categories and retrieves the appropriate strategies based on the student's risk level. To avoid redundancy and cognitive overload, only one action per driver category is selected, and the total number of recommended actions is capped.

If no dominant positive drivers are detected for a given student, the model defaults to a risk-level–appropriate fallback recommendation focused on monitoring and general academic support.

The output of this process (Table 8) is a concise, prioritized set of educational support actions tailored to each student's specific risk profile and underlying risk drivers.

Table 8. Sample - Automated Personalized Recommendation Model

Student _ID	Dropout_ Probability	Main_Risk_ Drivers	Risk_Level	WHY_Explanation	Primary_ Driver_ Category	Personalized_ Recommendation
11391	56%	studied_credits (+1.287); date_registration (+0.502); code_module_DDD (-0.181)	Moderate Risk	The model estimates a 55.8% probability of dropout. This prediction is mainly influenced by: studied_credits (+1.287); date_registration (+0.502); code_module_DDD (-0.181).	Course Load	Assess workload balance and provide guidance on managing study commitments effectively. Provide additional guidance during initial course phases to support adjustment.
28400	11%	num_interactions (-0.610); total_clicks (-0.206); studied_credits (-0.173)	Low Risk	The model estimates a 10.6% probability of dropout. This prediction is mainly influenced by: num_interactions (-0.610); total_clicks (-0.206); studied_credits (-0.173).	Engagement Volume	Low predicted dropout risk. Continue routine academic monitoring.
135335	81%	num_interactions (+1.185); studied_credits (+0.919); total_clicks (+0.446)	High Risk	The model estimates a 80.7% probability of dropout. This prediction is mainly influenced by: num_interactions (+1.185); studied_credits (+0.919); total_clicks (+0.446).	Engagement Volume	Initiate immediate outreach to increase overall participation, including mandatory academic follow-up, structured use of learning resources, and close monitoring of weekly activity levels. Review and adjust course load where possible, prioritizing essential modules and aligning workload with student capacity.

4.5. Evaluation of the Recommendation Model Outputs

This final section of the chapter evaluates the outputs of the proposed automated recommendation model.

The objective is to present complementary metrics that institutions can leverage to obtain a comprehensive overview of the analyzed student groups, as well as to examine the distribution of main drivers across different risk levels. These aggregated metrics are intended to serve as reference indicators, acknowledging that additional global metrics may be derived from the available data.

Figure 9 summarizes the **global reach of the recommendation engine** across the full student population.

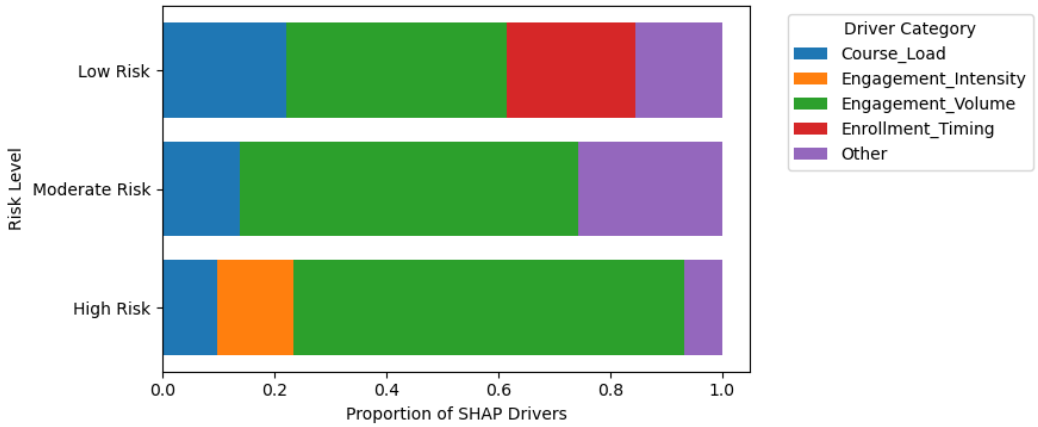
Figure 9. Summary of recommendations model

Metric	Value
Total students analyzed	32593
Average dropout risk	31.2%
High-risk students (≥70%)	4570
Medium-risk students (40–70%)	5731
Low-risk students (<40%)	22292

Out of 32,593 students analyzed, the average predicted dropout risk is 31.2%. Approximately 14% of students are classified as high risk, 18% as moderate risk, and the remaining 68% as low risk. This distribution reflects an expected pattern, as the majority of students are expected to successfully complete the courses on which they enroll.

Figure 10 presents the relative composition of SHAP risk drivers by category and risk level, expressed as proportions. Clear structural differences emerge across risk groups.

Figure 10. Aggregated Risk by Driver Category



For High-Risk students, the model identifies engagement-related factors, particularly *Engagement Volume* as the dominant contributor to dropout risk. This indicates that students with the highest predicted risk are primarily characterized by insufficient or highly concentrated interaction patterns, supporting intervention strategies centered on increasing sustained participation and improving study pacing.

In contrast, Moderate-Risk students display a more diversified driver composition. While *Engagement Volume* remains relevant, a substantial share of risk drivers falls into the "Other" category, reflecting heterogeneous contributing factors. This distribution justifies the use of lighter, more adaptive interventions focused on monitoring and selective support rather than intensive academic actions.

For Low-Risk students, the composition of drivers is more balanced across categories such as *Course Load*, *Enrollment Timing*, and engagement-related factors. This pattern suggests that dropout risk in this group is not driven by acute behavioral signals but rather by structural or contextual characteristics. As a result, the recommendation model appropriately defaults to maintenance-oriented strategies and routine monitoring, avoiding unnecessary intervention.

Taken together, the figure and table demonstrate that the recommendation model is **selective, interpretable, and behavior-aware**.

To go deeper into the behavior of categorical features, Figure 11 presents an aggregated view of dropout risk by module and by semester module. These summaries highlight noticeable differences in both average predicted risk and the concentration of high-risk students across categories. In particular, certain modules and academic periods concentrate a larger share of high-risk students, suggesting potential priority areas for institutional monitoring.

Figure 11. Dropout risk distribution across categorical features (Module and Semester)

Module	Students	Avg_Risk	High_Risk
CCC	4434	0.442773	848
DDD	6272	0.347138	953
FFF	7762	0.307788	1130
BBB	7909	0.307533	1280
EEE	2934	0.247470	315
AAA	748	0.221153	23
GGG	2534	0.120200	21

Semester Module	Students	Avg_Risk	High_Risk
2014J	11260	0.338978	1991
2014B	7804	0.336550	1097
2013B	4684	0.281404	514
2013J	8845	0.270945	968

Finally, one of the research questions addressed in this work explores the potential institutional impact of acting proactively on high-risk students: *to what extent could the average dropout rate be reduced if a learning support strategy were applied in advance to students identified as high risk?*

To answer this, a simulation was conducted by summing individual predicted dropout probabilities to estimate the expected number of dropouts. A hypothetical intervention scenario is then defined by applying a conservative **10% relative reduction to the predicted dropout probability of students classified as High-Risk** (Dropout Probability ≥ 0.7), while keeping the predictions for all other students unchanged.

The difference between the baseline and simulated scenarios provides an estimate of the potential institutional impact, **suggesting that even a modest reduction in risk could prevent approximately 392 student dropouts per semester**.

This approximation is proposed as a **business-oriented model output**, intended to support institutional adoption by **translating predictive analytics into measurable impact indicators** that can be **directly used to justify and prioritize data-driven decision-making initiatives**.

5. Conclusions

The central contribution of this work lies in addressing the long-standing gap between knowing which students are at risk and understanding how institutions should respond. While prior research has emphasized improving predictive performance, this framework reorients the analytical pipeline toward explanation-driven action. By extracting local SHAP-based drivers, organizing them into behaviorally grounded categories, and mapping them to risk-sensitive institutional responses, the proposed automated model operationalizes interpretability as a mechanism for intervention design rather than as a purely diagnostic tool.

Empirically, **the results show that dropout risk is not driven by uniform factors across students**. Instead, different risk levels are driven by distinct engagement patterns, academic context, and other factors. High-risk students are primarily characterized by insufficient or poorly distributed engagement, moderate-risk students exhibit more heterogeneous drivers, and low-risk students are influenced mainly by contextual and workload-related factors. The automated recommendation model reflects these differences by generating differentiated and category-consistent actions.

From a theoretical perspective, this work contributes to learning analytics and XAI literature by **demonstrating that explainability serves as a bridge construct between prediction and intervention**. From a practical standpoint, **this thesis offers institutions a replicable approach for transforming complex model outputs into structured recommendations**.

The findings should be interpreted within clear boundary conditions. The proposed recommendation logic is designed for a specific online learning environment where student interaction data is available. The recommended actions reflect institutional heuristics rather than experimentally validated causal effects. As such, **the model is intended to support informed decision-making rather than replace human judgment or pedagogical expertise**.

In closing, this thesis shows that the purpose of learning analytics is not fulfilled by predicting dropout more accurately, but by **making those predictions *usable***. By demonstrating how explainable models can be systematically translated into personalized, behavior-aware recommendations, this work identifies a concrete pathway through which **data and predictive analytics can meaningfully inform** educational action and support transparent, structured institutional **decision-making** in online learning contexts.

References

Adnan, M., Habib, A., Ashraf, J., Mussadiq, S., Raza, A. A., Abid, M., Bashir, M., & Khan, S. U. (2021). *Predicting at-risk students at different percentages of course length for early intervention using machine learning models.* IEEE Access, 9, 7519–7539. https://doi.org/10.1109/ACCESS.2021.3049446

Alhakbani, H. A., & Alnassar, F. M. (2022). *Open learning analytics: A systematic review of benchmark studies using Open University Learning Analytics Dataset (OULAD).* Proceedings of the 2022 7th International Conference on Machine Learning Technologies (ICMLT '22), 81–86. https://doi.org/10.1145/3529399.3529413

Al-Shabandar, R., Hussain, A., Liatsis, P., Keight, R., & Anderton, R. (2023). *Predicting student dropout and academic success using machine learning.* Education Sciences, 13(1), Article xx. https://doi.org/10.3390/educsci130100xx

Anvhane, V. (2025). *A comparative analysis of machine learning models for dropout prediction on the OULAD dataset*. Preprint. https://www.researchgate.net/publication/392323080_A_Comparative_Analysis_of_Machine_Learning_Models_for_Dropout_Prediction_on_the_OULAD_Dataset

Anvhane, V. (2025). *Student dropout analysis: Descriptive study for Open University Learning Analytics Dataset (OULAD).* Preprint. https://www.researchgate.net/publication/392322789_Student_Dropout_Analysis_A_Descriptive_Study_for_Open_University_Learning_Analytics_Dataset_OULAD

Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system.* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://arxiv.org/abs/1603.02754

Dorogush, A. V., Ershov, V., & Gulin, A. (2018). *CatBoost: Gradient boosting with categorical features support:* https://arxiv.org/abs/1810.11363

Esri. (2025). *How CatBoost algorithm works*. ArcGIS Pro Documentation. https://pro.arcgis.com/en/pro-app/latest/tool-reference/geoai/how-catboost-works.htm

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media.

Gupta, S., Sabitha, A. S., & Chowdhary, S. K. (2021). *Predicting the efficiency and success rate of programming courses in MOOC using machine learning approach for future employment in the IT industry*. Journal of Information Technology Research, 14(2), 30–47. https://doi.org/10.4018/JITR.2021040102

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R* (2nd ed.). Springer.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer: https://vuquangnguyen2016.files.wordpress.com/2018/03/applied-predictive-modeling-max-kuhn-kjell-johnson_1518.pdf

Kuhn, M., & Johnson, K. (2019). *Feature engineering and selection: A practical approach for predictive models*. Chapman & Hall/CRC: http://www.feat.engineering/

Martins, M. V., Baptista, L., Machado, J., & Realinho, V. (2023). *Multi-class phased prediction of academic performance and dropout in higher education*. Applied Sciences, 13(8), 4702. https://doi.org/10.3390/app13084702

Martins, M. V., Tolledo, D., Machado, J., Baptista, L. M. T., & Realinho, V. (2021). *Early prediction of student's performance in higher education: A case study. In Advances in Intelligent Systems and Computing* (pp. 166–175). Springer. https://doi.org/10.1007/978-3-030-72657-7_16

Molnar, C. (2022). *Interpretable machine learning*. https://christophm.github.io/interpretable-ml-book/

Murphy, K. P. (2022). *Probabilistic machine learning: An introduction*. MIT Press: https://dokumen.pub/probabilistic-machine-learning-an-introduction-1-9780262046824.html

Nagy, M., & Molontay, R. (2024). *Interpretable dropout prediction: Towards XAI-based personalized intervention*. International Journal of Artificial Intelligence in Education, 34, 274–300. https://doi.org/10.1007/s40593-023-00331-8

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). *Scikit-learn: Machine learning in Python*. Journal of Machine Learning Research, 12, 2825–2830. https://www.jmlr.org/papers/v12/pedregosa11a.html

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). *CatBoost: Unbiased boosting with categorical features*. Advances in Neural Information Processing Systems, 31, 6638–6648. https://doi.org/10.48550/arXiv.1706.09516

Waheed, H., et al. (2022). Student dropout analysis: *A descriptive study for the Open University Learning Analytics Dataset (OULAD)*. ResearchGate. https://www.researchgate.net/

Yandex. (2019). *Categorical features in CatBoost*. CatBoost Documentation. https://catboost.ai/docs/en/features/categorical-features

Appendix

I. Feature Engineering

- **Demographic Features**

Student demographic information was extracted from the *studentInfo* dataset. The selected variables included gender, region, highest level of education, IMD band, age band, disability status, number of previous attempts, and studied credits. Additionally, course identifiers (*code_module* and *code_presentation*) and the student identifier (*id_student*) were retained to ensure proper data alignment. The target variable (*dropout*) was also included at this stage.

Categorical variables were transformed using one-hot encoding, while numerical variables were retained in their original form.

- **Registration Features**

Registration-related information was obtained from the *studentRegistration* dataset. For each student, the earliest registration date was extracted and used as a proxy for enrollment timing. This aggregation ensured that registration information was represented by a single feature per student.

- **VLE Interaction Features**

Student interaction with the virtual learning environment was derived from the *studentVle* dataset. Clickstream data were aggregated at the student level to capture overall engagement patterns. Four interaction features were generated: total number of clicks, average clicks per interaction, maximum clicks recorded, and the total number of interactions. These features summarize both the intensity and frequency of student engagement with online learning resources.

- **Target Variable Definition**

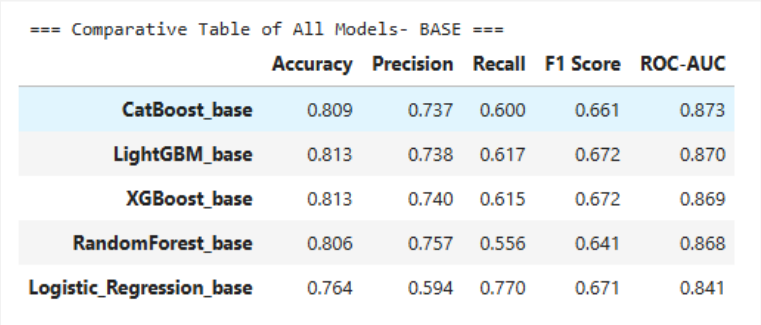
The target variable was derived from the *final_result* attribute. Students whose final result was classified as *Withdrawn* were labeled as dropouts, while all other outcomes (*Fail*, *Pass*, and *Distinction*) were grouped as non-dropouts. The resulting binary target variable (*dropout*) was used in all subsequent modeling experiments.

II. Hyperparameter process

This appendix provides a detailed description of the hyperparameter tuning process implemented in this study, as well as the baseline performance metrics obtained prior to model optimization.

Figure 12 illustrates the baseline model performance metrics.

Figure 12. Model performance metrics (Base Line)



=== Comparative Table of All Models- BASE ===

	Accuracy	Precision	Recall	F1 Score	ROC-AUC
CatBoost_base	0.809	0.737	0.600	0.661	0.873
LightGBM_base	0.813	0.738	0.617	0.672	0.870
XGBoost_base	0.813	0.740	0.615	0.672	0.869
RandomForest_base	0.806	0.757	0.556	0.641	0.868
Logistic_Regression_base	0.764	0.594	0.770	0.671	0.841

After performing Grid Search (cv=5) for hyperparameter optimization to improve model performance, Table 9 summarizes the impact of hyperparameter tuning on each model.

Table 9. Hyperparameter optimization

Model	Version	Principals Hyperparameters and Tuning	Accuracy	Precision	Recall	F1 Score	ROC-AUC
LightGBM	Base	n_estimators=300, random_state=42	0.813	0.738	0.617	0.672	0.87
	Optimized	n_estimators=200, max_depth=-1, num_leaves=31, learning_rate=0.05, random_state=42	**0.811** ↓	**0.742** ↑	**0.603** ↓	**0.665** ↓	**0.874** ↑
CatBoost	Base	iterations=300, random_state=42, verbose=0	0.809	0.737	0.6	0.661	0.873
	Optimized	iterations=300, depth=4, learning_rate=0.1, loss_function="Logloss", eval_metric="AUC", random_seed=42, verbose=0	**0.813** ↑	**0.750** ↑	**0.602** ↑	**0.668** ↑	**0.874** ↑
XGBoost	Base	n_estimators=300, max_depth=6, learning_rate=0.1, subsample=0.8, colsample_bytree=0.8, random_state=42,	0.813	0.74	0.615	0.672	0.869

		use_label_encoder=False, eval_metric="logloss"					
	Optimized	n_estimators=300, max_depth=6, learning_rate=0.05, subsample=0.8, colsample_bytree=1, random_state=42, use_label_encoder=False, eval_metric="logloss"	**0.812** ↑	**0.744** ↑	**0.603** ↓	**0.666** ↓	**0.873** ↑
Logistic Regression	Base	penalty=l2, C=1.0, solver="liblinear", max_iter=1000, class_weight="balanced"	0.764	0.594	0.77	0.671	0.841
	Optimized	C=0.1, penalty="l1", solver="liblinear", max_iter=1000, class_weight="balanced"	**0.763** ↓	**0.593** ↓	**0.766** ↓	**0.668** ↓	**0.842** ↑
Random Forest	Base	n_estimators=300, class_weight="balanced", random_state=42	0.806	0.757	0.556	0.641	0.868
	Optimized	n_estimators=300, max_depth=2, min_samples_split=2, min_samples_leaf=2, class_weight="balanced", random_state=42	**0.771** ↓	**0.649** ↓	**0.575** ↑	**0.610** ↓	**0.825** ↓

While accuracy and F1-score remained relatively stable, optimization introduced trade-offs between precision and recall. For boosting-based models, adjustments to learning rate and tree complexity resulted in higher discriminative performance, as reflected by ROC-AUC.

By analyzing closer the performance variations resulting from the optimization process, there are parameters that showed noticeable influence on model performance:

- For LightGBM, the most influential hyperparameters were the *learning rate* and the *number of leaves*, resulting in a higher ROC-AUC with a slight reduction in recall.
- For CatBoost, tuning the *tree depth* and *learning rate* resulted in consistent improvements across all evaluation metrics.
- For XGBoost, model performance was mainly driven by the *learning rate* and the column subsampling ratio, leading to higher ROC-AUC and precision, with a slight decrease in recall.
- For Logistic Regression, performance was largely determined by the *regularization parameter C*, showing minimal sensitivity to tuning.

- For Random Forest, the *maximum tree depth* was identified as the dominant hyperparameter, where increased regularization was associated with degraded overall performance.

III. Calibration process

This section describes the probability calibration procedure applied to the predictive model to improve the reliability of its probability estimates

- **Calibration Curve (Reliability Plot)**

The calibration curve compares predicted probabilities with observed dropout frequencies across bins of similar predicted values. The x-axis represents predicted probabilities, while the y-axis represents the empirical proportion of dropouts within each bin. The 45-degree diagonal line represents perfect calibration. Deviations from this line indicate miscalibration:

- Points above the diagonal indicate underestimation of risk.
- Points below the diagonal indicate overestimation of risk.

So, calibration is considered satisfactory when deviations from the diagonal are small and no systematic over- or underestimation pattern is observed.

Figure 13. Calibration Curve (CatBoost)

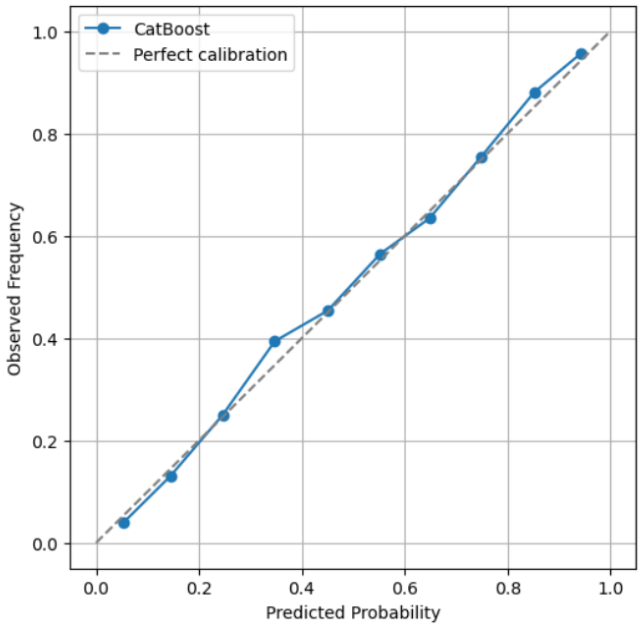
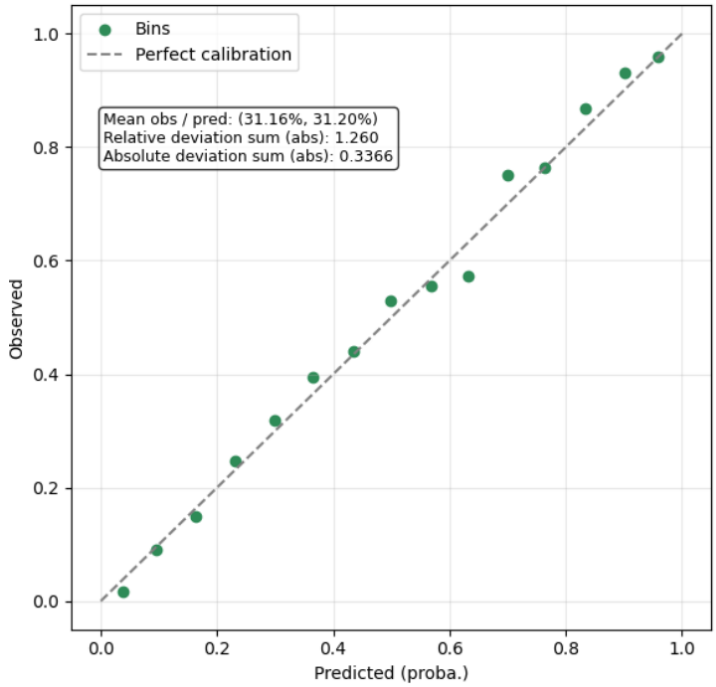


Figure 13 shows the calibration curve closely follows the reference diagonal, suggesting good agreement between predicted and observed dropout frequencies. This indicates that the model's probability estimates are well aligned with empirical outcomes.

- **Q–Q Plot of Predicted vs Observed Probabilities**

The Q–Q plot evaluates the agreement between the distribution of predicted probabilities and the empirical distribution of outcomes. If the predicted probabilities are well calibrated, the points should approximately follow a straight reference line. Substantial curvature or systematic deviations would indicate distributional mismatch and potential miscalibration.

Figure 14. QQ- Plot: Dropout Prediction



The results shown in Figure 14 that predicted and observed quantiles exhibit strong alignment, reinforcing the conclusion that the model's outputs are probabilistically consistent.

- **Distribution of Predicted Probabilities**

The distribution plot shows the separation between predicted probabilities for dropout and non-dropout students.

While this plot primarily reflects discrimination rather than calibration, it provides additional context. A well-performing model should produce:

- Higher predicted probabilities concentrated among dropout cases
- Lower predicted probabilities concentrated among non-dropout cases

Figure 15. Distribution of Predicted Probabilities

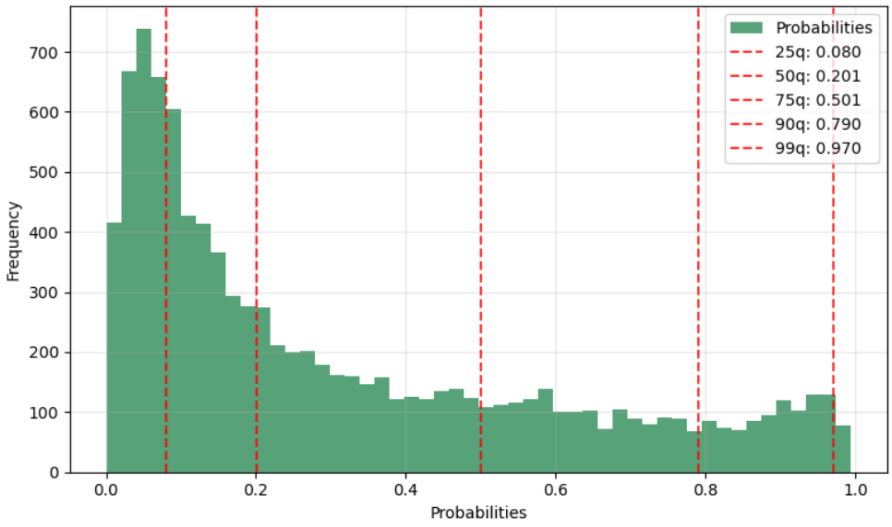


Figure 15 shows clear separation between the two distributions indicates strong ranking capability, which complements calibration assessment.

- **Brier Score**

The Brier score provides a quantitative measure of probabilistic accuracy. It is defined as the mean squared difference between predicted probabilities and actual binary outcomes:

$$\text{Brier Score} = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2 \tag{A.3}$$

where $p_i$is the predicted probability and $y_i \in \{0,1\}$.

The Brier score ranges between 0 and 1, with lower values indicating better probabilistic accuracy. It captures both discrimination and calibration components.

For the selected CatBoost model, the Brier score was:

Brier Score = 0.128686

This relatively low value indicates good overall probabilistic performance and supports the graphical evidence of adequate calibration; values substantially below 0.25 in binary settings generally indicate meaningful predictive power.

IV. Summary of Related works

This appendix presents complementary tables summarizing prior research related to student dropout prediction and machine learning applications in education. **The purpose of these tables is to provide contextual support for the present study by illustrating existing approaches, datasets, and modeling techniques reported in the literature.**

Table 1 includes recent studies (2025-2021) based on publicly available educational datasets to contextualize the current work within open-data educational analytics research.

Table 2 focuses specifically on benchmark studies that used the Open University Learning Analytics Dataset (OULAD), which is also employed in the present research.

Table 3 summarizes representative research applying machine learning and deep learning techniques for student dropout prediction, demonstrating methodological trends in the field.

It is important to note that some studies listed in these tables were compiled from previously published reviews and are presented here for contextual comparison purposes.

Table 1. Related Works on **Public sources**:

Author	Title	Variables	Objective	Techniques	Results	Conclusions
Anvhane (2025)	A Comparative Analysis of Machine Learning Models for Dropout Prediction on the OULAD Dataset	Academic performance, demographics, assessments, VLE interactions	Predict student dropout and compare ML models	Logistic Regression, SVM, Random Forest, XGBoost	XGBoost outperformed other models	ML models effectively detect at-risk students; XGBoost supports timely intervention
Anvhane (2025)	Student Dropout Analysis: Descriptive Study for OULAD	Study area, age, grades, clicks,	Descriptive dropout analysis	Descriptive statistics	Identified dropout trends by	Early interaction metrics key to understanding dropout

		educational background			engagement and age	
Tertulino & Almeida (2025)	Evaluating Federated Learning for At-Risk Student Prediction: A Comparative Analysis of Model Complexity and Data Balancing	Early academic performance, engagement volume & quality, demographics	Evaluate centralized vs federated models for at-risk prediction	Logistic Regression, Deep Neural Network	Federated model achieved ≈85% ROC-AUC; similar to centralized	Federated Learning viable for privacy-preserving early-warning systems
Sturmer et al. (2024)	Analysis of MOOC Dataset	Activity logs, participation, certification metrics	Analyze engagement–certification relationship	Descriptive analysis	Patterns between activity and certification	Engagement metrics useful for certification prediction
Villar & Andrade (2023)	Supervised ML Algorithms for Predicting Student Dropout	Enrollment, demographics, socio-economic factors, semester performance	Compare ML algorithms for dropout	RF, SVM, KNN, Logistic Regression	Random Forest performed best	Algorithm choice and class imbalance handling are critical
Gupta et al. (2021)	Predicting the Efficiency and Success Rate of Programming Courses in MOOC	Programming performance, course activity, assessments	Predict efficiency and success in programming MOOCs	Supervised ML	Activity and contest performance predicted success	Early participation strong predictor
Realinho et al. (2021)	Early Prediction of Student's Performance in Higher Education	Enrollment, demographics, socio-economic data	Early prediction of performance	Supervised classification, RF, Gradient Boosting	High predictive accuracy	Boosting methods improve early detection

Source: Authors' own elaboration.

Table 2: Summary of benchmark studies **using the OULAD dataset**.

Author	**Technology Used**	**Dataset Features**	**Tasks**	**Results/Remarks**
Hlosta et al., 2017	Ouroboros, Classical ML	Engagement and Demographic	At-risk students prediction	XGBoost model was reported best for prediction when data from the only ongoing course was used.
Alshabandar et al., 2018	GMM, k-NN and LR	Engagement and Demographic	Early dropout prediction	GMM achieved an F1 score of 0.835 and reported best. Latent engagement features proved effective in prediction.

Heuer and Breiter, 2018	Classical ML	Engagement, Demographic and Performance	Final exam success prediction	SVM performed best with 87% accuracy. Binarized daily activity has a similar impact as exact clickstreams on performance prediction.
Aljohani et al., 2019	LSTM and Conventional ML	Engagement (click-stream)	Early dropout and final success	LSTM outperformed baseline models with 93.46% accuracy. Student's final success can be predicted with high accuracy as early as in weeks.
Jha et al., 2019	Various Ensemble ML	Engagement, Demographic and Performance	Early dropout and final success	GBM approach reported best with AUC of 0.91 and 0.93 for dropout and final result prediction, respectively.
Chui et al., 2020	RTV-SVM	Engagement, Performance and Demographic	At-risk/marginal students prediction	RTV-SVM achieved approximately 94% prediction accuracy. RTV-SVM reduced the training time by up to 59%.
He et al., 2020	RNN-GRU	Engagement and Demographic	At-risk students prediction	RNN-GRU achieved around 80% prediction accuracy. Simpler RNN algorithms performed better than complex algorithms.
Karimi et al., 2020	Deep Learning, SVM and LR	Engagement and Demographic	Performance prediction	The proposed deep learning model achieved up to 85% for binary classification and 57% for four-class classification.
Mubarak et al., 2020	SELOR and IOHMM	Engagement	At-risk students prediction	SELOR and IOHMM achieved an accuracy of 84%. Proposed models outperformed the baseline machine learning models.
Nazif et al., 2020	PNN and Conventional ML	Engagement, Demographic and Performance	Final success prediction	PNN with FSCNCA feature selection achieved the best accuracy of 93.4%.
Tomasevic et al., 2020	Classical ML	Engagement, Demographic and Performance	At-risk students and final success	ANN performed best with an F1 score of 96% using engagement and performance. No impact of demographic features on the final exam score prediction.
Waheed et al., 2020	Deep Learning	Engagement, Demographic and Performance	At-risk students prediction	Deep learning model outperformed conventional machine learning models with accuracy of 93%.
Ahmad et al., 2021	ANN and RF	Engagement, Demographic and Performance	Performance prediction	ANN performed best with an accuracy of 91%. Feature engineering was applied to achieve improved performance.
Fatema et al., 2021	k-NN, SVC and ANN	Engagement, Demographic and Performance	Final success prediction	k-NN and ANN achieved a classification accuracy of 99%. Improved results are potentially due to feature scaling.

Hidalgo et al., 2021	ANOVA, RFECV and Meta-Learning	Engagement	Students performance prediction	The meta-learning approach was able to achieve 67.2% accuracy on the test dataset.
Hlioui et al., 2021	Various Conventional ML	Engagement, Demographic and Performance	At-risk students prediction	RF performed best with 84% accuracy using a balanced dataset with discrete behavioural indicators.
Waheed et al., 2021	SC-GAN	Engagement, Demographic and Performance	At-risk students prediction	SC-GAN achieved 7% improvement in comparison to conventional up-sampling approaches for data balancing.
Yuan et al., 2021	Classical ML and LSTM	Behavioural	At-risk students prediction	SVM and CART for aggregate data. LSTM for sequence data.

Source: Compiled from Alhakbani et al. (2022)

Table 3: Summary of research studies **using ML/DL techniques for student dropout**.

Research Study	**Algorithm Used and Performance Achieved**	**Problem Addressed**	**Limitations**
Chung and Lee (2019)	Random Forest (RF) with 95% accuracy	Students' binary classification	Potential inaccuracy in calculating feature weights
Gray and Perkins (2019)	1-Nearest Neighbour with 97% accuracy	Identifying failing students at week 3	Model only applicable to Bangor University students
Al-Shabandar et al. (2019)	Gradient Boosting Model with 95% accuracy	Identification of at-risk students for early intervention	Temporal features not considered; not validated with additional datasets
Lee and Chung (2019)	Random Forest and Boosted Decision Trees (BDT), BDT = 99% accuracy	Improving dropout prediction using ML-based early warning system	Limited NEIS database; not all features included
Behr et al. (2020)	Random Forest with AUC = 0.86	Binary classification of student dropout	Students' satisfaction features not considered
Martins et al. (2017)	Gradient Boosting, Deep Learning, Distributed RF; DL TPR = 71.1%	Early prediction of student attrition	First semester data not included
Hussain et al. (2019)	ANN and SVM, highest accuracy = 75%	Predicting students' difficulties in online learning	Dropout prediction not performed; low accuracy
Mduma et al. (2019)	Linear Regression with ROC = 0.88	Identification of at-risk students	Under-sampling with penalized model not applied

Figueroa-Cañas and Sancho-Vinuesa (2019)	Decision Trees >90% accuracy	Early identification of dropout-prone students	Training and validation sets from same academic year
Ortigosa et al. (2019)	C5.0 algorithm >85% accuracy	Early dropout prevention system challenges	Real-world validation and retention effectiveness needed
Imran et al. (2019)	Feed-forward Deep Neural Network >90% accuracy	Predicting and explaining student dropout	Prediction performed after course completion
Wu et al. (2019)	CLMS-Net (CNN + LSTM + SVM), 91.55% accuracy	MOOC dropout prediction	Validation on additional datasets needed
Rosé et al. (2019)	Traditional ML models	Understanding limits of ML-only approaches	Models lack interpretability and actionability
Mubarak et al. (2020)	LSTM, ANN, SVM, LR; LSTM = 93% accuracy	Predicting learning analytics in MOOCs	Only video interaction patterns considered
Liao et al. (2019)	SVM with AUC = 0.70	Identifying low-performing students	Early prediction not addressed
Sekeroglu et al. (2019)	SVR, LSTM, SVM, GBC, ANN; ANN = 87.78%	Student performance classification	Small dataset; earliest detection not addressed
Mao (2018)	BKT, IBKT, LSTM; LSTM = 74%	Student intervention modeling	Manual hyperparameter tuning; late intervention
Iqbal et al. (2017)	CF, MF, RBM; RBM best (RMSE=0.3, MSE=0.09, MAE=0.23)	Grade prediction	Motivation features not included; limited dataset
Fwa and Marshall (2018)	HMM with LOOCV	Modeling programming students	Engagement features not considered; no early prediction
Xu et al. (2017)	LR, Logistic Regression, RF, kNN, EPP (lowest MSE)	Tracking and predicting student performance	No intervention strategy implemented

Source: Compiled from Adnan et al. (2021)