IJTE International Journal of Technology in Education
www.ijte.net

Metamodeling for Learning Analytics in Higher Education

Olga Ovtšarenko

Vilnius Tech, Vilnius, LT-10223, Lithuania & TTK University of Applied Sciences, Tallinn, Estonia 0000-0003-2980-3678
Corresponding author: Olga Ovtšarenko (olga.ovtsarenko@tktk.ee)

Article Info

Article History

Received:
24 February 2026

Revised:
11 May 2026

Accepted:
4 June 2026

Published:
18 June 2026

Keywords

Adaptive education
Learning analytics
Metamodeling
Optimization
Weighted attributes
Educational data mining

Abstract

In modern higher education, extensive learning log data is generated, capturing students' behavioral, temporal, and cognitive activity in online courses. Transforming these diverse data streams into understandable, interpretable, and actionable insights for adaptive learning is an important research challenge. This study presents a metamodeling approach to learning analytics that applies principles of educational analytics engineering. The data model was created using learning logs from six CAD e-courses, which involved 155 students and utilized various teaching methods. Each online course was weighted based on ECTS credits and the number of students, ensuring hierarchical normalization and data comparability across courses. Six machine learning models were trained to predict dropout risk. Three models - Decision Tree, Random Forest, and Gradient Boosting - showed similarly high validation accuracy, but the Gradient Boosting model was selected for further analysis due to its stability and interpretability within the metamodeling framework. The results confirmed that metamodeling improves the interpretation and use of data to understand learning dynamics. By considering educational processes as design systems, this study proposes an approach to using surrogate models in adaptive learning environments. The proposed metamodel supports the development of teacher-centered early warning systems in learning management systems such as Moodle.

Citation: Ovtšarenko, O. (2026). Metamodeling for learning analytics in higher education. *International Journal of Technology in Education (IJTE), 9*(3), 790-811. https://doi.org/10.46328/ijte.7748

ISSN: 2689-2758 / © International Journal of Technology in Education (IJTE).

Introduction

The digital transformation of education has generated vast amounts of student-generated data on learning platforms—from log records and assessments data to behavioral and cognitive metrics - enabling data-driven educational design. Research in educational analytics (EA) and artificial intelligence (AI) has shifted from descriptive monitoring of learning activities to predictive and prescriptive frameworks capable of personalizing the educational experience in real time (Herodotou et al., 2019; Zawacki-Richter et al., 2019; López-Pernas et al., 2025). These methods facilitate a shift toward adaptive education, in which learning systems dynamically adjust content, pace, and student support to meet the changing needs of each student. Despite these advances, the challenge remains of transforming multidimensional, heterogeneous educational data into interpretable analytical models for both pedagogical decision-making and system design and optimization.

The field of design analytics for optimizing engineering design and computational modeling has evolved around similar challenges. Designers explore multivariate spaces and use data-driven methods to understand system behavior, explore alternatives, and develop innovations (Simpson et al., 2001; Negrín-Diaz et al., 2023). A key role in this field is played by metamodeling, also known as surrogate modeling, which approximates the behavior of complex systems through simplified mathematical or statistical representations (Wang & Shan, 2007). Metamodels are designed to predict outcomes, quantify uncertainty, and facilitate iterative improvement (Paulson &Tsau, 2024).

Educational systems are design systems: learners, learning resources, and pedagogical strategies form a complex adaptive network whose behavior can be modeled, analyzed, and optimized. Therefore, adaptive education becomes a task of design analytics, where the goal is to find optimal configurations of learning variables that improve engagement and learning outcomes and enable the adaptation of learning content. Metamodeling offers precisely the tools necessary for such analytical transformation: abstraction, interpretability, and prediction using actionable data.

Weighted and feature-based methods, such as the weighted attribute method (WAM) (Ovtsarenko, 2025), quantify cognitive complexity, time allocation, and engagement to assess student performance. These methods function similarly to surrogate models in engineering design, where multivariate relationships between design parameters and performance metrics are approximated for optimization. Similarly, logistic regression and ensemble learning approaches, widely used in educational data mining (Ahmed et al., 2025; Ersozlu et al., 2024), are analogous to the use of interpretable and hybrid surrogates in an engineering context (Jog et al., 2024; Liu et al., 2025). Together, they demonstrate a common methodological framework that links design analytics with learning analytics through the concept of metamodeling.

While the use of metamodeling in engineering has reached methodological maturity, its theoretical and practical integration into educational research remains largely unexplored. Existing learning analytics frameworks excel at predictive analysis but need improvements to optimize and interpretively model the learning process itself (Hernández-Leo et al., 2018; Tirado et al., 2024). As a result, educators gain insight into what is likely to happen,

but not into the reasons for it or how course design can be systematically improved. Adopting metamodeling principles can fill this gap by enabling project-based analytics that capture the relationships between pedagogical inputs (task complexity, sequence, time) and learning outcomes (engagement, learning, persistence) in a transparent and optimizable manner. Ethical and human-centered principles necessitate interpretable analytical frameworks in education. As Topali et al. (2024) emphasizes, adaptive systems must preserve human agency, privacy, and pedagogical values. Metamodeling's emphasis on simplification, traceable uncertainty, and model transparency creates a promising foundation for responsible, explainable educational AI that supports, rather than replaces, teacher autonomy.

This study proposed the integration of metamodeling methods for design analytics in adaptive education and examines how principles and methods derived from engineering design analytics can be adapted to learning system modeling, combining quantitative optimization with qualitative pedagogical interpretation. The study aims to:

- Explore metamodeling methodologies relevant to design analytics and assess their potential applicability to educational data modeling.
- Identify parallels between design space exploration and adaptive learning analytics, highlighting how surrogate modeling can enhance interpretability and effectiveness.
- Situate metamodeling within the human-centered and ethical framework of adaptive education.
- Propose a conceptual model of design analytics adaptive learning - a data-driven approach to educational system optimization.

By framing adaptive education as a metamodeling problem, this study contributes to the theoretical and methodological development of learning analytics by offering a path toward transparent, data-efficient, and ethically consistent adaptive learning systems.

Literature Review

In process design, metamodels serve as analytical surrogates that reflect input-output relationships derived from simulations or experimental data. Prominent approaches include Gaussian process regression, radial basis functions, artificial neural networks, and symbolic regression (Negrín-Diaz et al., 2023; Jog et al., 2024). These models support tasks such as design space exploration, uncertainty quantification, reliability analysis, and multi-objective optimization. In the field of design analytics, the use of metamodeling in workflows will enable information generation and support decision making. Paulson and Tsau (2024) described Bayesian optimization as a ’closed analytical loop’ in which surrogate models inform future best design estimates, improving the understanding of system behavior. Liu et al. (2025) demonstrated how multi-point surrogates integrate low- and high-resolution models, providing scalability and interpretability in energy system design. Beyond predictive performance, contemporary research focused on interpretable and hybrid metamodeling - the combination of symbolic and statistical models to provide transparency and explainability (Jog et al., 2024). This trend aligns with the needs of educational systems, where model interpretability and human control are critical ethical and pedagogical requirements (Topali et al., 2024; Pardo & Siemens, 2014).

The principles underlying metamodeling - abstraction, optimization, and decision support - are closely aligned with the goals of adaptive learning analytics. Both areas involve high-dimensional data, nonlinear dependencies, and the need to balance predictive accuracy and interpretability (Ahmed et al., 2025; Hernández-Leo et al., 2018). In educational research, predictive LA and AI methods have been used to predict student performance, identify at-risk students, and tailor educational interventions (Ersozlu et al., 2024). Integrating metamodeling can extend these systems from predictive analytics to prescriptive ones, helping to optimize the learning process itself. a weighted abstraction of engagement, difficulty, and time allocation. The weighted attributes method (WAM) (Ovtšarenko, 2025) like engineering methods functioned as an educational surrogate model - WAM approximated latent performance functions under limited data conditions. Such frameworks enabled dynamic exploration of ’learning design spaces,’ facilitating curriculum optimization and data-driven adaptation of learning resources.

Moreover, the metamodeling workflow supports adaptive feedback loops in education. Bayesian-based update mechanisms, widely used in engineering optimization (Paulson & Tsau, 2024), can be adapted to continuously refining predictive models based on student interaction data, enabling real-time adjustments to instructional materials and strategies. This alignment positioned metamodeling as a theoretical foundation for design analytics in adaptive learning systems. A key challenge for both engineering and education is the tradeoff between model complexity and interpretability. In design optimization, transparent surrogate models such as symbolic regressions are often used to inform decisions (Han et al., 2017; Kianifar et al., 2020). Similarly, in education, interpretability ensures that educators can understand model outputs and act responsibly based on them (Topali et al., 2024; Tirado et al., 2024).

Human-centered AI frameworks (Allison et al., 2025) emphasize that analytics should complement, not replace, human expertise. For educators, visualizing and interpreting uncertainty (e.g., confidence intervals or success probabilities) enables evidence-based decision making. Multicriteria metamodeling (Negrín-Diaz et al., 2023; Liu et al., 2025) can be applied to multicriteria instructional design, balancing engagement, effectiveness, and cognitive load. Symbolic regression and hybrid surrogate models (Jog et al., 2024) offered interpretable pathways for identifying causal relationships in student behavior data.

In educational systems, these approaches can enable teachers to create metamodel-based dashboards that visualize how adjustments to task difficulty or resource duration impact predicted learning outcomes. Bayesian optimization principles can further automate the development of personalized curricula by adaptively selecting the most effective learning strategies based on pedagogical approaches. Therefore, metamodeling for design analytics offers a promising interdisciplinary methodology: an interpretable and adaptive approach to modeling learning processes as dynamic design systems, combining the predictive capabilities of AI with the ethical principles of pedagogy and ensuring students’ data protection (GDPR, 2018).

Method

The methodological foundation of this study was the application of metamodeling principles from engineering design analytics to adaptive learning systems. In adaptive education, learning environments produce large

volumes of diverse data detailing the relationships between learner behavior, task difficulty, time management, and learning outcomes. Modelling these relationships is difficult due to data sparsity, noise, and nonlinearity. A metamodeling approach, however, allowed the creation of surrogate representations of the learning process - mathematical models that encapsulated key behavioral and cognitive interactions without overfitting the data. Design analytics focused on how data affects decision-making in design processes, combining descriptive and predictive models to optimize performance while maintaining human interpretability (Paulson & Tsau, 2024). This helped to identify educational strategies that maximize engagement and achievement within pedagogical limits.

To construct interpretable behavioral indicators for the predictive metamodel, two core computational steps were applied. The weighted attribute function, adapted from response-surface metamodeling in engineering design analytics, was used to compute a normalized success indicator S_{norm} as a function of weighted learning-activity attributes (Equation 1). This formulation enabled the aggregation of heterogeneous LMS events into a single interpretable performance measure.

$$S_{norm,i} = \frac{\alpha \cdot W_{logs,i} + \beta \cdot D_{resource,i} + \gamma \cdot T_{resource,i}}{\sum_i S_{resource,i}}, \quad (1)$$

where

$S_{resource,i}$ - success indicator of a resource,

$W_{logs,i}$ - weight of e-course resource logs,

$D_{resource,i}$ - difficulty weight of e-course resource,

$T_{resource,i}$ – planned time weight of a resource,

α/ β/ γ – normalization factors of weights (0,2/ 0,3/ 0,5 accordingly), partitioning e-course resources weight was by forcing α+β+γ=1, and distributing 'importance' across the features: 20% of the success score was explained by logs, 30% by complexity, 50% by planned ac. hours, normalization factors can be selected by the course developer, offering flexibility in the application of the formula.

Course-level importance weightings were calculated using course credits and enrolment counts. These raw weights were then normalized so that their sum equaled 1 across all courses in the dataset (Equation 2). This hierarchical normalization ensured comparability between courses of different sizes, structures, and credit loads.

$$Ij = 0,6 \cdot \frac{ECTS\ j}{\sum ECTSj} + 0,4 \cdot (enrolled\ j), \quad (2)$$

where

I_j – weighted factor of a course importance,

enrolled j – an e-course students count,

$ECTS_i$ – an e-course credits.

Equation 1 and Equation 2 are referenced in the Data and Model Structure section to describe how these computations are applied within the full modelling workflow.

Research Design

The proposed methodological framework adopted a meta-analytical and design-based research structure, combining computational modelling with pedagogical interpretation (see Figure 1).

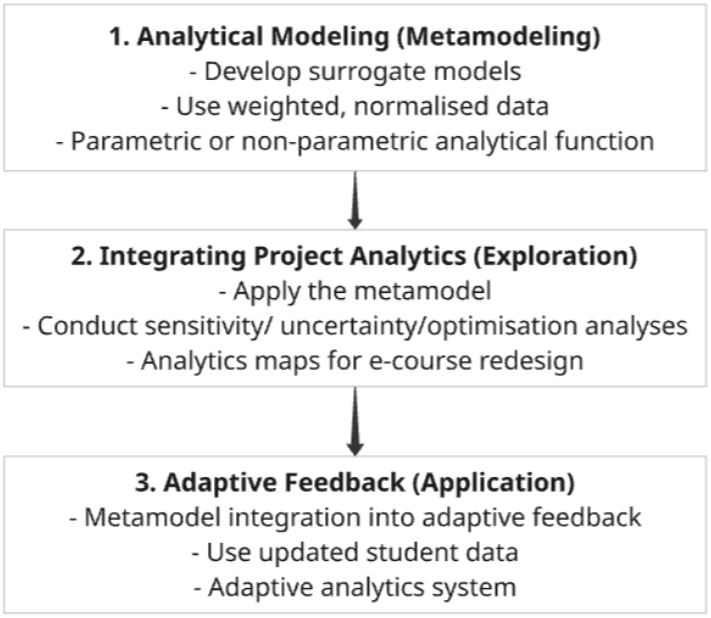

Figure 1. Research Design Phases

Design phases description:

- Phase 1- analytical modelling (metamodeling)

  Goal: developed surrogate models that approximate the relationship between learning project variables (inputs) and learning outcomes (outputs).

  Implementation: used weighted, normalized representations of cognitive and behavioral data (difficulty, engagement, task completion time) to build an interpretable metamodel.

  Expected result - parametric or non-parametric analytical function.

- Phase 2 - integrating project analytics (exploration)

  Goal: applied the metamodel to explore and visualize the multidimensional space of the learning project.

  Implementation: conducted sensitivity, uncertainty, and optimization analyses to determine which variables most significantly influence learning outcomes.

  Expected result - analytics maps that help instructors redesign or organize course content for maximum effectiveness.

- Phase 3 - adaptive feedback (application)

  Goal: integrated the metamodel into adaptive feedback loops for continuous improvement.

  Implementation: used updated student data to iteratively refine the metamodel parameters, similar to active learning.

  Expected result - an adaptive analytics system capable of predicting, intervening, and adjusting the curriculum in real time.

Results

Data and Model Structure

The metamodeling methodology was applied to multidimensional educational historical datasets obtained from the Moodle learning management system of TTK University of Applied Sciences (Estonia) for the 2024/2025 academic year, which included data of six CAD courses for different specialties - student activity logs available to each teacher without administrator rights and the necessary approval procedures (Ovtsarenko, 2025) and were presented:

- behavioral variables (number of logs, interaction frequency, completion rates),
- cognitive variables (task difficulty levels mapped to Bloom's taxonomy, assessed learning outcomes),
- temporal variables (time-on-task, engagement intervals).

Each learning activity was treated as a design element, and each learner profile represents an element in the learning design space. The weighted attribute function (Equation 1), adapted from engineering response-surface formulation, is used to compute a normalized success indicator S_{norm} as a function of weighted attributes. The function can be used to automate calculations in an Excel spreadsheet or a Colab interactive notebook using a Python algorithm, and to store the results in an intermediate file with a specific name. Using an external resource for calculations was convenient since the Moodle platform database was not overloaded and did not require cleaning of used data. To code the e-course, the main course code (131/ 133/ 138) and the form of study were used: 0 - full-time study, 1 - correspondence study/elective subject, 2 - subject for Erasmus students in English.

In Figure 2 is presented using dataset data. The number of students who failed required or elective courses varies greatly, which is explained by students' more responsible attitude towards electives. Therefore, it was crucial to prevent students from dropping out of these courses and to intervene promptly to support them and provide study guidance.

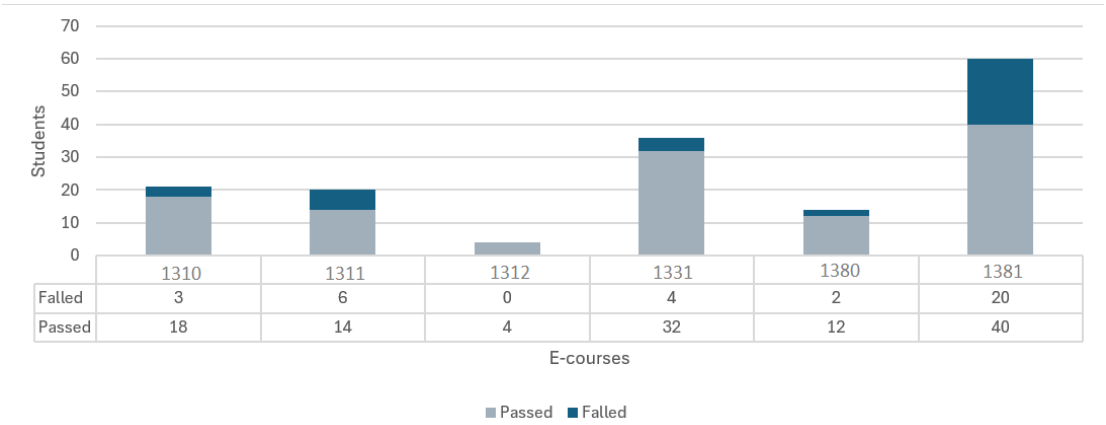

Figure 2. E-courses General Data Visualization

For metamodeling in adaptive educational systems, a hierarchical approach to assigning weighting coefficients was considered the most suitable. It involved normalizing resource weights within each course and then scaling

them across all courses in the data model. This approach ensured interpretability, comparability, and adaptability both at the course level and across the entire set of courses. Course importance weightings calculated based on credits and the number of enrolled students and then normalized to a total of 1 (Equation 2). Table 1 presents the formula components used and I_j calculated data.

Table 1. Courses Data Used for Metamodelling

E-course	1310	1311	1312	1331	1380	1381	Total
$ECTS_j$	3	3	3	6	3	3	21
$enrolled_{.j}$	21	20	4	36	14	60	155
I_j	0.140	0.137	0.096	0.264	0.122	0.241	1

The calculated coefficients I_j were utilized to aggregate the metamodel data with normalized global parameters (Ovtsarenko, 2025).

Data Model Use

Several algorithms of architecture can be effective for intelligent analysis of educational data and student dropout prediction. Six common and effective algorithm types (GeeksforGeeks, 2025): Logistic Regression, Decision Trees, Random Forests, Gradient Boosting, neural networks (Support Vector Machine and K-Nearest Neighbors) were used. Training the data model with the provided algorithms produced the following results, as shown in Table 2.

Table 2. Evaluation Results on Validation Set

Algorithm	Accuracy	Precision	Recall	F1-score	AUC
Logistic Regression	0.8696	0.5000	0.3333	0.4000	0.8833
Decision Tree	1.0000	1.0000	1.0000	1.0000	1.0000
Random Forest	1.0000	1.0000	1.0000	1.0000	1.0000
Gradient Boosting	1.0000	1.0000	1.0000	1.0000	1.0000
SVM	0.8696	0.0000	0.0000	0.0000	N/A
KNN	0.9130	1.0000	0.3333	0.5000	0.6167

Based on the evaluation results, the Decision Tree, Random Forest, and Gradient Boosting algorithms demonstrated the best performance on the validation set. The hyperparameter search space for these most effective algorithms was defined, and GridSearchCV was used to identify the optimal hyperparameters for each model using the training data, with the validation set utilized for evaluation. For these three top models, a range of potential hyperparameter values (such as the depth of a tree or the number of estimators) was established. GridSearchCV (Grid Search with Cross-Validation) systematically tested all possible hyperparameter combinations on the training data to find the most effective one.

After identifying the optimal hyperparameters, all algorithms were re-executed and assessed on the validation set

to gauge their performance, ensuring proper tuning and avoiding overfitting to the training data. The results obtained were significantly improved. The three chosen algorithms demonstrated consistent results from initial application without further tuning.

Algorithm Selection

Next, the most effective model was selected on an unknown test set. According to the validation results, the Decision Tree, Random Forest, and Gradient Boosting algorithms used for models training performed best after tuning. The tuned Gradient Boosting model was selected (Emami & Martínez-Muñoz, 2025). Its advantages were due to the ensemble nature of the model and the sequential learning process, which was aimed at minimizing the systematic error of the model and increasing the forecasting accuracy:

- Gradient Boosting integrated multiple weak learners into a sequential ensemble, where each subsequent tree aimed to correct the errors of the previous one (Friedman, 2001). This iterative process enabled the model to attain high predictive accuracy and effectively model nonlinear relationships within the data. In the educational context, where learning outcomes were shaped by a range of cognitive, behavioral, and contextual factors, such adaptability was essential for accurately modelling student performance and dropout risk.
- The method offered robustness against noise and missing values, particularly when employing stochastic sampling or regularization (Chen & Guestrin, 2016).
- From an analytical standpoint, gradient boosting models offered interpretability and transparency through feature importance metrics and explainability tools such as SHAP (SHapley Additive ExPlanations) scores (Lundberg & Lee, 2017), enabling researchers and educators to identify the most influential predictors of student achievement.

The combination of accuracy, flexibility, and interpretability make a finely tuned Gradient Boosting model especially suitable for data-driven educational analytics and the development of adaptive learning systems.

Metamodel Development

The metamodel used a machine learning model, specifically a developed and tuned Gradient Boosting model, to predict and understand student attrition:

- Prediction - took student data as input and produces a probability estimate of attrition. This enabled educators to proactively identify students at risk.
- Insight and prevention - analyzing the metamodel (for example, through feature importance) offered insight into the key factors that most significantly influence the risk of attrition. This understanding was vital for devising targeted and effective prevention strategies, and advisable to concentrate resources on the specific factors that the model identifies as most critical for individual students or groups.

Consequently, the metamodel converted raw student data into actionable insights and predictions that can help support and enhance interventions aimed at maintaining student engagement and academic success. Based on the chosen algorithm, a metamodel must be developed to predict and prevent student dropout. This included interpreting the model's predictions and identifying key factors that contribute to dropout.

Analysis of Features

Determining the feature importance values from the best model (configured with Gradient Boosting) was necessary to understand these key factors, and how they can be utilized in dropout prevention strategies, and to explain how the model can be used for prediction with new data presented in Table 3.

Table 3. Descriptive Statistics for Features (X DataFrame)

Feature	Count	Mean	Std	25%	50%	75%	Median	Skewness	Kurtosis
success indicator, $S_{recourse,i}$	153.0	5.408	23.810	1.350	1.521	2.336	1.521	10.052	111.548
time, $T_{resource,i}$	153.0	20.563	90.764	4.845	5.975	9.366	5.975	10.003	110.495
time weight norm, $WT_{resource,i}$	153.0	10.706	47.140	2.690	2.944	4.574	2.944	10.049	111.492
difficulty weight, $D_{resource,i}$	153.0	0.114	0.515	0.016	0.032	0.058	0.032	10.160	114.209
log weight, $WL_{resource,i}$	153.0	0.118	0.517	0.009	0.015	0.045	0.015	9.983	111.543
tool	153.0	2.843	12.431	1.000	1.000	1.000	1.000	10.179	114.366
importance weightings, I_j	153.0	0.020	0.087	0.005	0.006	0.009	0.006	9.978	110.130
success indicator, $S_{recourse,i}$	153.0	5.408	23.810	1.350	1.521	2.336	1.521	10.052	111.548
level_scaled, *li*	153.0	0.080	0.350	0.020	0.020	0.040	0.020	10.161	114.149
logs, L_i	153.0	802.902	3616.024	36.000	95.000	294.000	95.000	9.522	102.069

For this purpose, the previously calculated ’feature importances sorted’ parameter and a bar chart were utilized to visualize the importance of each feature (see Figure 3). Feature importance showed which parts of student data most affect the model’s prediction of dropout risk. To create effective prevention strategies, it was essential to concentrate on features with the highest scores of importance. If ’level scaled’ and ’time weight norm’ were highly significant, this indicated that student academic progress and time spent on resources were vital factors.

The feature importance analysis offered valuable insights for developing strategies to prevent student dropouts. Concentrating on enhancing the ’success indicator,’ tracking ’time on resources,’ and potentially providing targeted support based on ’level’ and ’resource difficulty’ were essential areas for intervention. The analysis of instances predicted as dropouts (which matched the actual dropouts in the test set) also revealed that even students with higher engagement and success indicators on specific resources might be at risk, suggesting that other underlying factors or a broader perspective on their learning path could be necessary for comprehensive prevention. Based on correlation, univariate, and bivariate analysis (Ovtsarenko, 2025), a synthesis of results regarding the relationship between the selected features and student dropout was presented.

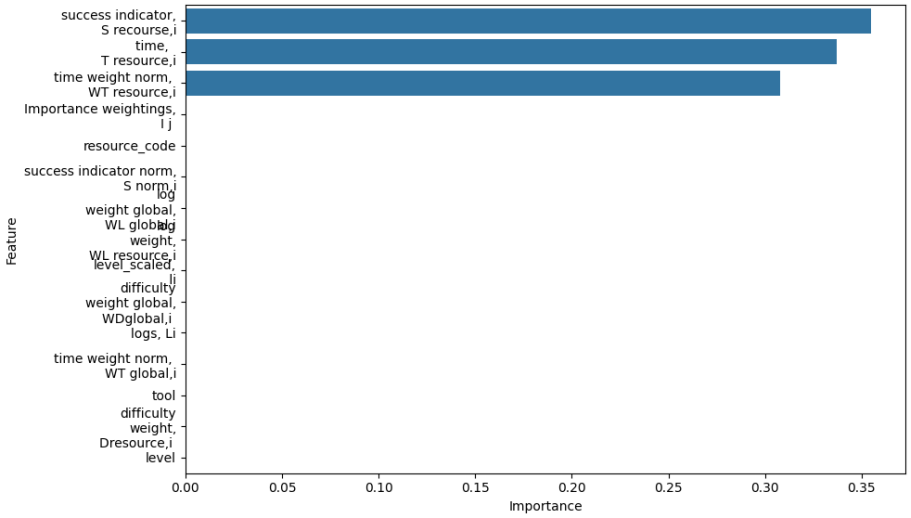

Figure 3. Feature Importances (Tuned Gradient Boosting)

Univariate Analysis for Selected Features

Univariate analysis was performed on the selected features to compare distributions between the outlier and non-outlier groups. The results are presented in Figure 4.

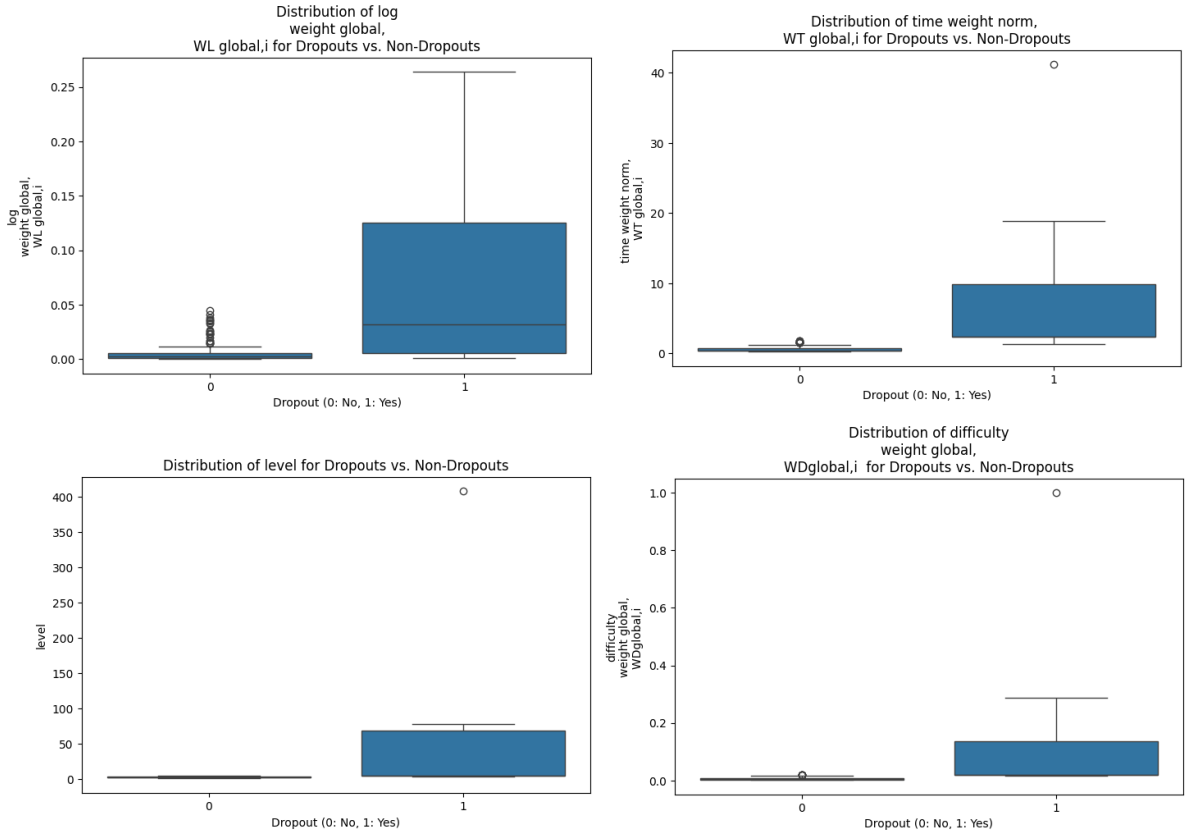

Figure 4. Univariate Analysis Box Plots

Box plots compared the distributions of the selected features for the dropout (1) and non-dropout (0) groups. The

median and overall distributions of these features were higher for the dropout group than for the non-dropout group, indicating that dropouts tended to have higher values of these global logs and time-based measures. Boxplots for level and difficulty weight global also demonstrated some differences in distribution between the groups, consistent with their correlations. The dropout group tended to have slightly higher median levels and difficulty weights globally, although the distributions partially overlap in the distributions.

Multivariate Feature Interaction Analysis

Conducting a bivariate analysis of selected features to examine their relationships with the target variable (student dropout risk) helped understand how each feature (independent variable) connects to the target. A multivariate analysis was performed to visualize the interaction patterns among four core engagement-related indicators: the log weight global, the time weight norm, level, and difficulty weight global. The pairwise scatter matrix differentiated between dropout and non-dropout students to identify multidimensional behavioral structures that predicted persistence. Figure 5 showed a pair plot (bivariate scatter matrix) with relationships between several key features and the dropout indicator (where orange points represent dropout students and blue points represent non-dropouts).

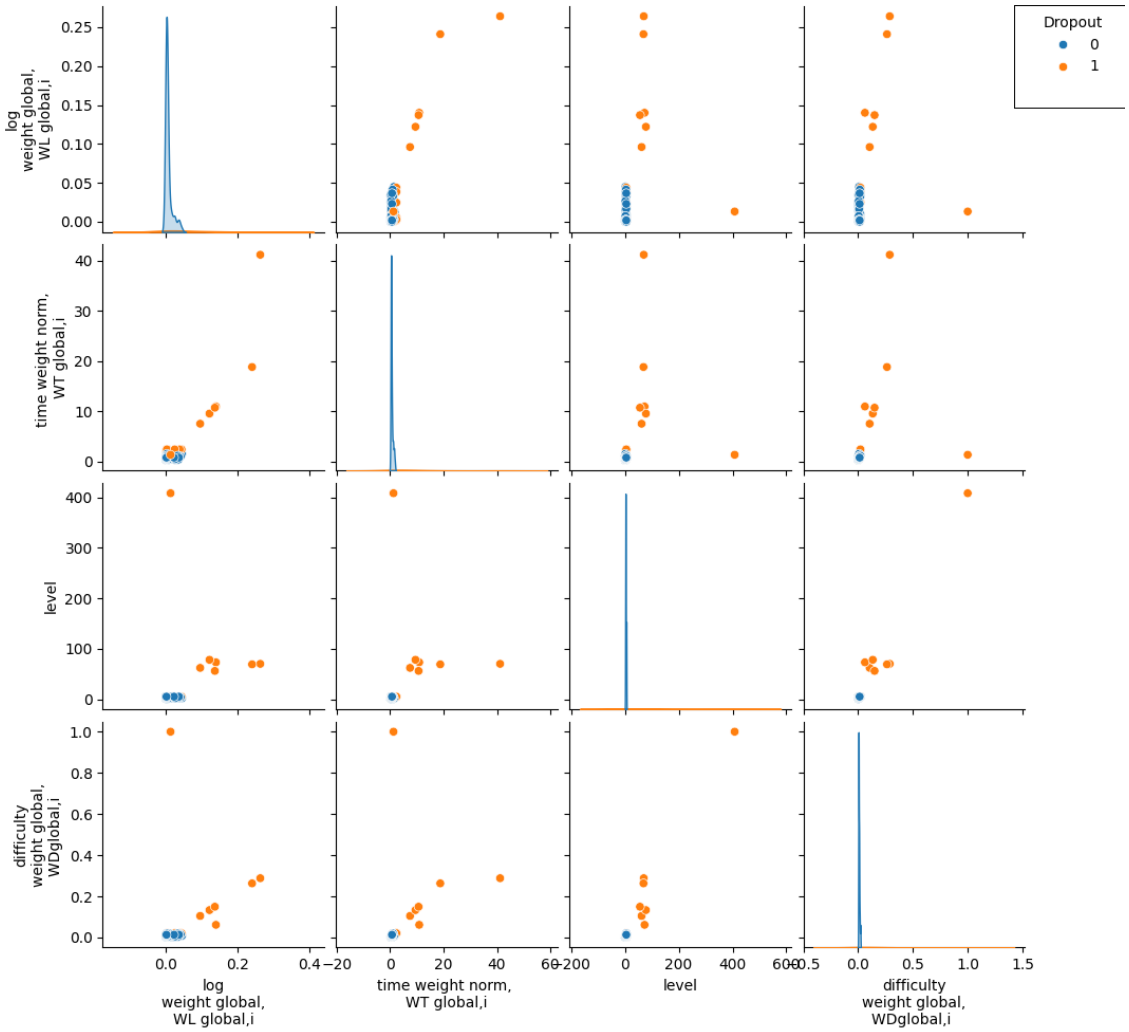

Figure 5. Bivariate Scatter Matrix of Key Features

The visualized relationships revealed distinct clustering patterns that categorized students based on engagement intensity. Non-dropout students formed concentrated clusters at higher levels of all weight indicators, signifying strong and consistent engagement with learning materials of varying difficulty. Conversely, dropout students clustered near the origin of each plot, indicating low interaction frequency, minimal time investment, and limited exposure to complex learning resources. This multidimensional separation supported the hypothesis that low engagement across behavioral and cognitive dimensions was a reliable precursor to withdrawal.

A notable positive relationship was observed between the log weight global and difficulty weight global indicators, showing that students who actively interacted with course resources also engaged with more challenging materials. Similarly, the time weight norm correlated positively with difficulty weighting, indicating that time investment in learning activities increases with content complexity among non-dropout students. The level variable followed similar trends - students who progressed to higher levels maintained higher engagement measures, whereas dropout students showed little variance across levels, suggesting early disengagement.

These interrelationships demonstrated that engagement behaviors in digital learning environments were highly correlated and mutually reinforcing. Consistent time investment, interaction with diverse resources, and willingness to engage with difficult materials together formed a multidimensional profile of persistence. Conversely, uniform inactivity across all engagement dimensions characterized dropout-prone students.

These findings supported multivariate engagement patterns as features in predictive analytics models for early risk detection. Monitoring changed in these joint indicators enabled teachers and learning systems to identify students transitioning towards low-engagement states and implement timely, personalized interventions. This is aligned with evidence from large-scale learning analytics studies, where combined behavioral and temporal variables outperformed univariate predictors in identifying dropout risk (Herodotou et al., 2019; Acosta et al., 2024).

Correlation Analysis

A correlation analysis was conducted between all features and the target variable using Spearman's rank correlation coefficient, and the correlation matrix was visualized in Figure 6. Based on the Spearman Correlation with success indicator, $S_{global,i}$ (Absolute Values), the most significant parameters for determining the risk of student dropout are:

- time weight norm, $WT_{global,i}$ (0.994) - this feature has the strongest correlation, indicating a very strong monotonic relationship with the global success indicator. Changes in this normalized global time weight are highly predictive of student success.
- time weight norm, $WT_{resource,i}$ (0.850) - like the global time weight, the normalized time weight for individual resources also plays a significant role.
- difficulty weight global, $WD_{global,i}$ (0.834) - the global difficulty weight is also highly correlated, implying that the perceived difficulty of the overall curriculum or tasks has a strong relationship with student dropout risk.

These features, with absolute correlation values close to 1, were the most influential in determining the risk of student dropout - $WT_{global,i}$ was the most influential predictors of student success and dropout, followed by $WT_{resource,i}$, and $WD_{global,i}$. Their very high absolute correlations indicate strong monotonic relationships, confirming that time-weight patterns, perceived importance, and difficulty exposure are central to understanding dropout risk.

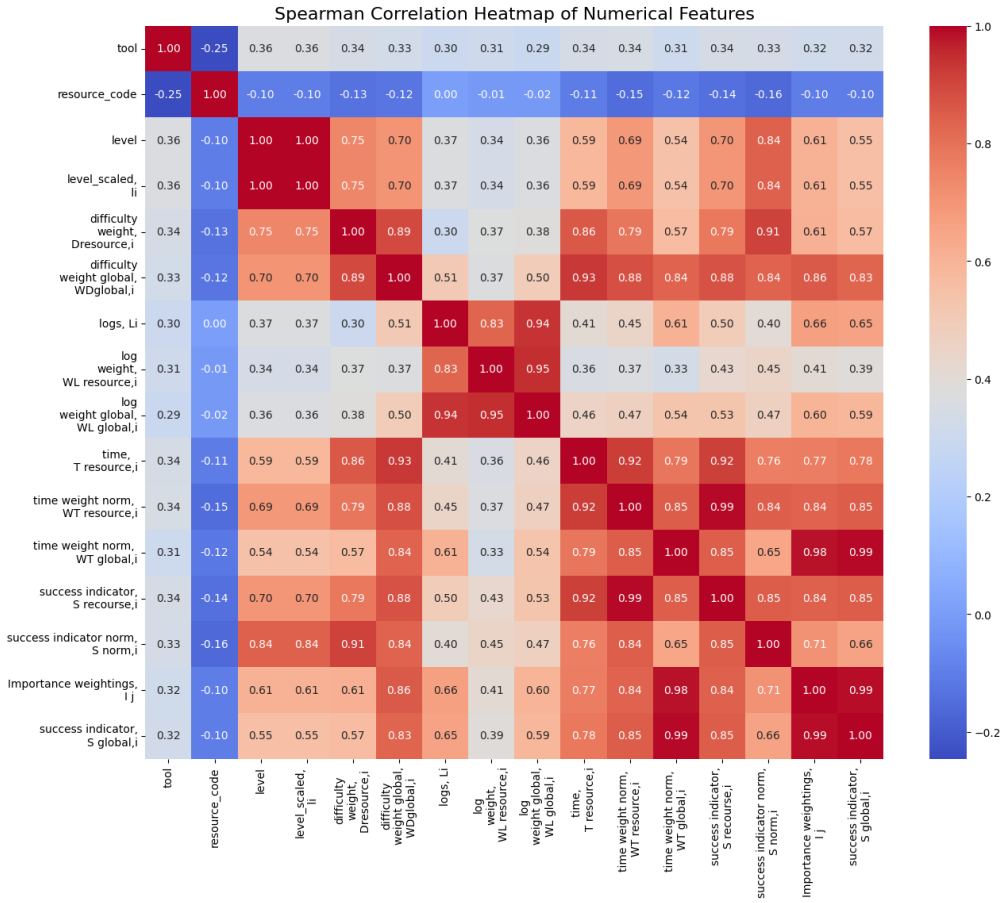

Figure 6. Heatmap of Pairwise Correlations

These analyses confirmed that, beyond the top features identified by the model, features related to overall log activity, time spent on resources, as well as academic level and resource difficulty, demonstrated notable associations with student dropout. These findings supported recognizing these features as important for further model development and for informing targeted intervention strategies, especially when combined with insights from feature importance and the analysis of predicted dropout cases.

Summary

The developed metamodel was evaluated using the fine-tuned Gradient Boosting algorithm, which exhibited strong predictive performance in student dropout detection (Yağcı, 2022), achieving perfect accuracy and F1 scores of 1.0000 on the validation set. The model's robustness and interpretability were validated through feature importance, univariate, bivariate, and correlation analyses, which collectively demonstrated how weighted

engagement indicators forecast dropout behavior. The feature importance ranking (see Figure 3) showed that time weight normalized, difficulty weight global, and level were the most influential variables for predicting student dropout. These findings supported the weighted attribute function structure, confirming that time allocation, task difficulty, and progression level are the key dimensions of engagement that influence persistence or disengagement.

The analysis integrated univariate distributions, bivariate relationships, and feature correlations to examine the factors influencing student dropout risk using the proposed metamodel indicators. The data comprised four main engagement weights - log weight, normalized time weight, difficulty weight, and level - and success and importance indicators. These metrics were derived from normalized learning logs and represent multiple behavioral attributes of student engagement.

Univariate Analysis: as shown in Figure 4, box plot comparisons revealed distinct distributional differences between dropout (1) and non-dropout (0) groups. Dropouts exhibited higher medians and greater variance across all global weight indicators. This reflected irregular engagement and cognitive overload among students who eventually withdrew. Conversely, non-dropouts demonstrated lower, more stable distributions, suggesting consistent and efficient learning patterns.

Bivariate Analysis: the pairwise scatter matrix (see Figure 5) illustrated strong positive relationships between all global indicators. Non-dropout students clustered at higher engagement levels, while dropouts formed low-value clusters, confirming a multidimensional differentiation in engagement behavior. This validated the use of composite, multivariate features for early dropout prediction, consistent with findings by Acosta et al. (2024).

The correlation heatmap (see Figure 6) confirmed several clear patterns. Time-related features - particularly the global time-weight $WT_{global,i}$ - showed the strongest positive associations with dropout risk (0.994) while global difficulty weight $WD_{global,i}$ (0.834) demonstrated a similar but slightly weaker pattern. Log-weight features and level indicators showed only mild positive correlations with dropout (r≈0.10-0.20). In contrast, both the global success indicator $S_{global,i}$ and the normalized success indicator $S_{norm,i}$ were negatively correlated with dropout (r≈−0.20), indicating that higher normalized success was associated with greater persistence. Overall, these relationships suggest that attrition is not driven solely by low activity but is more strongly linked to unstable, inefficient, or poorly aligned engagement patterns.

Table 4 presented interpretation and educational insight within metamodeling framework for early dropout prediction. These results confirmed that $WT_{global,i}$, $S_{norm,i}$ and $WD_{global,i}$ were the dominant predictors of dropout. Together, they show that dropout risk is shaped by a combination of time-related behavior, performance stability, and perceived difficulty. High global time-weight indicates inefficient or excessive time investment, normalized success reflects consistent performance (with lower values signaling risk), and global difficulty weight captures the cognitive load imposed by course materials. Their combined influence demonstrates that dropout is not driven by a single factor but emerges from the interaction of time pressure, difficulty exposure, and unstable learning success.

Table 4. Analysis Results Interpretation and Comparing

Feature	Distribution Pattern	Correlation with Dropout	Interpretation	Dropout Risk
$WT_{global,i}$	High variance; long time use	Very strong (≈0.99)	Strong monotonic relationship: global time-weight is highly predictive of success/dropout	Very High
$S_{recourse,i}$	Higher and more stable for persistent learners; lower and more variable for dropouts	Strong (≈ 0.85)	Strong monotonic relationship: consistent success at the resource level closely reflects overall learning performance and predicts persistence	High
$WD_{global,i}$	Widespread; high difficulty exposure	Strong (≈0.83)	Global difficulty strongly affects success and dropout	High

Discussion

The results confirmed that metamodeling offered an analytical framework for modelling complex educational systems. Using available student activity data as interpretable weights, the metamodel established relationships between engagement intensity, cognitive effort, and learning outcomes. Integrating univariate, bivariate, and correlational analyses revealed consistent trends:

- Students who spent excessive time and experienced high levels of difficulty without corresponding success were at the highest risk of dropping out.
- Stable, moderate engagement combined with high normalized performance indicated effective learning and persistence.

These findings aligned with previous research showing that irregular behavior and cognitive overload were strong indicators of dropout (Cerezo et al., 2020; López-Pernas et al., 2025). The interpretability of the metamodel aligned with ethical principles of AI in LA (Topali et al., 2024; Tirado et al., 2024), enabling educators to transform data into pedagogically meaningful actions.

The Gradient Boosting approach proved suitable for the metamodel structure. Its sequential training process captured nonlinear dependencies between weighted features, producing interpretable and accurate predictions (Friedman, 2001; Lundberg & Lee, 2017). Using feature importance allowed educators and researchers to visualize the impact of features on dropout risk, supporting the application of AI in ES (Islam et al., 2025).

These results demonstrated the feasibility of implementing the metamodel in adaptive learning management systems (LMS). Embedding this model as an analysis tool in Moodle or similar platforms enabled the automatic identification of at-risk students and the initiation of early interventions, such as adaptive learning pacing, provision of templated content, or instructor support. This resonated the workflows of design analytics in engineering optimization, where surrogate models guide decision-making in real time (Negrín-Diaz et al., 2023; Paulson & Tsau, 2024). Examples include:

- A trained model can be used to predict the attrition risk of current students periodically (at the beginning

or middle of a course) or in real time as new data becomes available.

- Based on the model's output (the predicted probability of being in a risk group or the risk group itself), students at high risk of attrition can be identified.
- Once high-risk students were identified, interventions can be implemented using an early warning system (informing faculty, counsellors, or support staff about high-risk students).
- Engaging with high-risk students to provide academic support, counselling, or other resources. Personalized learning recommendations can suggest specific resources or activities that address factors increasing risk (if the model identifies difficulties with certain materials, additional resources are recommended).
- Mentoring programs can pair high-risk students with successful peers. Adjusting the curriculum if the model consistently is identified issues with certain resources or course elements may indicate a need to revise the curriculum or teaching methods.

Feature importance derived from the Gradient Boosting model can offer valuable insights into the reasons behind a student's predicted high risk. For example, if a low achievement weight and a high difficulty weight were key factors contributing to a student's high-risk prediction, this suggested the student struggles with challenging resources. This information can inform tailored interventions. It was vital to monitor the effectiveness of interventions and utilize this feedback to improve both the interventions and, potentially, the model itself. The model functions as an early warning system, helping proactively identify students needing support before they drop out, enabling timely and targeted interventions to prevent it.

Implication for Practice

Option 1 - high $WT_{global,i}$ but low $S_{norm,i}$

A student with a high predicted dropout risk demonstrates ineffective performance: they spend a lot of time studying but make little progress. This indicates unstable learning strategies or difficulties managing workload.

Intervention would focus on study skills support and targeted guidance to improve efficiency and understanding of core materials.

Option 2 - high $WD_{global,i}$, high $WT_{resource,i}$ and moderate $S_{norm,i}$

A student with a high predicted dropout risk likely puts in a lot of effort but struggles with complex materials.

Intervention would focus on providing support with specific resources, simplified explanations, step-by-step instruction, or alternative learning paths to reduce cognitive overload.

Option 3 - high $WT_{global,i}$ score and tool-related difficulties

If a student has a high risk of dropping out, this indicates difficulties using the learning platform or specific tools. The risk may also be related to technical barriers rather than academic ability.

Intervention may involve offering practical assistance, training on the tool, or assessing whether the tool itself creates unnecessary barriers for students.

If a model repeatedly identified certain features as significant risk indicators for many students, this may highlight broader issues in course design, curriculum, or available resources that require systemic attention. Using characteristic importance can help identify and understand risk factors, enabling more effective and targeted preventative measures. Responsible, pedagogical sound use of AI in learning analytics is achieved by ensuring:

- *Transparency and interpretability*
  The metamodeling framework created weighted, human-readable indicators rather than opaque model outputs. This ensured that instructors could understand why a student is labeled as at-risk.
- *Proportionality and minimal intervention*
  The system relied entirely on LMS behavior logs and avoids the use of personal or demographic data. This is aligned with ethical principles of data minimization and reduces the risk of stigma or bias.
- *Human decision making*
  The workflow explicitly required instructors to analyze model outputs and decide on the need and method of intervention.
- *Practical applicability for teaching*
  The model's event-based structure (aligned with assignment due dates) provides timely, context-specific data that instructors can transform into feedback, reminders, or support.

Limitations

1. The dataset used to develop the model was limited to six CAD e-courses from a learning management system (LMS) at the Moodle platform. While this dataset suffices for proof-of-concept modelling, generalization of different disciplines and student populations is necessary. Future research should utilize larger and more diverse datasets to verify the scalability and applicability of the model to external settings.
2. The metamodel captured correlations between engagement and attrition patterns but did not establish causal relationships or temporal progression. Student engagement was inherently dynamic, and future research should employ sequential modelling (e.g., Bayesian updating, recurrent neural networks) to infer learning trajectories and identify critical transition points preceding attrition.
3. While Gradient Boosting offered robust performance and explainability, its ensemble complexity may still obscure more subtle causal relationships. The use of symbolic or hybrid surrogate models can further improve transparency, especially for educators who prefer intuitive, rule-based interpretations.

These limitations defined the scope of the current study while also outlining clear directions for future research. Increasing the diversity of datasets, incorporating affective data, and testing the metamodel in realistic settings will enhance both scalability and pedagogical impact. Addressing these challenges will strengthen the position of metamodeling as a methodological bridge between computational optimization and human-centered adaptive learning.

Conclusion

This study demonstrated that metamodeling principles borrowed from engineering design analytics can be

effectively transferred to the field of learning analytics to support adaptive online education. By structuring LMS behavioral data using weighted indicators and integrating them into an interpretable metamodel based on Gradient Boosting, the proposed framework ensured both high predictive accuracy and pedagogical transparency and interpretability. Results showed that normalized time weights and a global difficulty weight were the strongest predictors of student attrition, while a normalized performance weight functions as a stabilizing indicator of learning persistence and task engagement. These results directly answered the research questions, confirming that metamodeling can improve both the interpretability and effectiveness of predictive models in educational institutions.

Beyond predictive effectiveness, the study utilized key pedagogical and ethical principles, prioritizing the use of accessible data, interpretability, and human-informed decision making. The metamodel structure allowed educators to understand the rationale behind risk classification, ensuring that analytics serve as decision support rather than automated inference. This alignment between the methodological approach and pedagogical values underpinned the theoretical underpinning of the study and its contribution to the responsible use of AI in education.

In practice, the results laid the foundation for developing educator-focused analytics and event-based early warning systems that can improve student retention and support timely, targeted instructional interventions. The emphasis on transparency and behavioral indicators made this framework suitable for scalable deployment across various online learning environments and across various subject areas. Future studies will expand on this work by validating the metamodel over the course of an academic semester using updated data and exploring its integration with multimodal learning data. By combining data efficiency, interpretability, and pedagogical relevance, metamodeling offered a robust path to the next generation of responsible, adaptive AI systems in education.

Novelty and Contribution

This article presented results, methodological contributions, and conceptual ideas in the fields of learning analytics and adaptive learning systems that were novel and relevant.

1. The study presented a novel application of metamodeling principles traditionally used in engineering design analytics to be modelling student behavior in online learning environments. This interdisciplinary transfer provided a new methodological approach for constructing interpretable weighted indicators based on LMS log data.
2. The study utilized new empirical data on predictors of student attrition. The analysis showed that normalized time weighting and global difficulty weighting were the strongest risk indicators, while normalized performance weighting acts as a stabilizing factor for maintaining motivation. These relationships offered new insights into how temporal and task-related dynamics influence student engagement.
3. The study proposed a unified metamodeling framework that combined predictive accuracy with pedagogical interpretability, leveraging ethical and pedagogical principles for transparency, data minimization, and human participation in decision making. This contributed to the development of

responsible AI in education.

4. The study made a practical contribution by describing how the metamodel can support teacher-centered analytics and early warning systems. These results provided practical tools for educational practice.

These contributions demonstrated that this study made both conceptual and practical contributions to the field of learning analytics, supporting the development of interpretable, ethically sound AI-based educational systems.

Abbreviations

AI	artificial intelligence
EA	educational analytics
LMS	learning management systems
TTK UAS	TTK University of Applied Sciences
WAM	weighted attribute method

Statements and Declarations

Acknowledgments/Notes: During the preparation of this article, the author used Grammarly to provide language editing and proofreading support. After using Grammarly, the author reviewed and edited the content as needed and took full responsibility for the content of the publication.

Supplementary Materials: Not applicable.

Author Contributions: The author planned the study, designed the data collection tools, collected and analyzed the data for the study. The author is the corresponding author. The author has read and agreed to the published version of the manuscript.

Funding: Not applicable.

Data Availability: The datasets generated and analyzed during the current study are available from [https://doi.org/10.5281/zenodo.17481678].

Ethics Approval: All procedures performed in studies involving human participants were performed following the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed Consent: Not applicable.

Conflicts of Interest: Not applicable.

References

Acosta, H., Lee, S., Mott, B., Bae, H., Glazewski, K., Hmelo-Silver, C., & Lester, J. (2024). Multimodal Learning Analytics for Predicting Student Collaboration Satisfaction in Collaborative Game-Based Learning. *Proceedings of the 17th International Conference on Educational Data Mining*, 224-235. https://doi.org/10.5281/zenodo.12729802

Ahmed, S., Khan, M. A., Zameer, S., & Iqbal, J. (2025). Student academic performance prediction using ensemble learning models. *Scientific Reports, 15*, 12353.

Allison, J., Hwang, J., Mayer, R. E., Pellas, N., Karnalim, O., Ng, L., Huang, M., Hooshyar, D., Seidman, R. H., Al-Emran, M., Mikropoulos, T. A., Schroeder, N. L., Roscoe, R. D., & Sanusi, I. (2025). From Generative AI to Extended Reality: Multidisciplinary Perspectives on Challenges, Opportunities, and Future of Educational Computing. *Journal of Educational Computing Research, 63*(6). DOI: 10.1177/07356331251359964

Cerezo, R., Sánchez-Santillán, M., Paule-Ruiz, M. P., & Núñez, J. C. (2020). Students' LMS interaction patterns and their relationship with academic performance. *Computers & Education, 159*, 104023. https://doi.org/10.1016/j.compedu.2016.02.006

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://doi.org/10.1145/2939672.2939785

Emami, S., & Martínez-Muñoz, G. (2025). Condensed-gradient boosting. *Int. J. Mach. Learn. & Cyber*, *16*, 687–701. https://doi.org/10.1007/s13042-024-02279-0

Ersozlu, Z., Taheri, S., & Koch, I. (2024). A review of machine learning methods used for educational data. *Educ Inf Technol, 29*, 22125–22145. https://doi.org/10.1007/s10639-024-12704-0

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

GeeksforGeeks (2025). Machine learning algorithms. *GeeksforGeeks*. https://www.geeksforgeeks.org/machine-learning/machine-learning-algorithms/

General Data Protection Regulation (2018). *GDPR*. https://gdpr-info.eu/

Han, Z. H., Chen, X., & Li, Y. (2017). Weighted Gradient-Enhanced Kriging for High-Dimensional Design. *ArXiv*. https://arxiv.org/abs/1708.02663

Hernández-Leo, D., Martinez-Maldonado, R., Pardo, A., Muñoz-Cristóbal, J. A., & Rodríguez-Triana, M. J. (2018). Analytics for learning design: a layered framework and tools. *British Journal of Educational Technology, 50*(1), 139-152. https://doi.org/10.1111/bjet.12645

Herodotou, C., Hlosta, M., Boroowa, A., & Rienties, B. (2019). The role of learning analytics in supporting student learning in higher education. *The Internet and Higher Education, 42*, 100–111.

Herodotou, C., Rienties, B., Boroowa, A., Zdrahal, Z., & Hlosta, M. (2019). A large-scale implementation of predictive learning analytics in higher education: The relationship between engagement and performance. *Computers in Human Behavior, 92*, 493–505.

Islam, M. M., Sojib, F. H., Mihad, M. F. H., Hasan, M., Rahman, M. (2025). The integration of explainable AI in Educational Data Mining for student academic performance prediction and support system. *Telematics*

and Informatics Reports, 18, 100203. https://doi.org/10.1016/j.teler.2025.100203

Jog, S., Vázquez, D., Santos, L. F., Caballero, J. A., Guillén-Gosálbez, G. (2024). Hybrid analytical surrogate-based process optimization via Bayesian symbolic regression. *Computers & Chemical Engineering, 182,* 108563. https://doi.org/10.1016/j.compchemeng.2023.108563

Kianifar, M. R., Campean, F. (2020). Performance evaluation of metamodeling methods for engineering problems: towards a practitioner guide. *Struct Multidisc Optim, 61,* 159–186. https://doi.org/10.1007/s00158-019-02352-1

Liu, L., Li, Z., Kang, H., Xiao, Y., Sun, L., Zhao, H., Zhu, Z., & Ma, Y. (2025). Review of surrogate model assisted multi-objective design optimization of electrical machines: New opportunities and challenges. *Renewable and Sustainable Energy Reviews, 215,* 115609. https://doi.org/10.1016/j.rser.2025.115609

López-Pernas, S., Oliveira, E., Song, Y., & Saqr, M. (2025). AI, Explainable AI and Evaluative AI: Informed Data-Driven Decision-Making in Education. *In: Saqr M, López-Pernas S (eds) Advanced Learning Analytics Methods*. Springer Cham. https://doi.org/10.1007/978-3-031-95365-1_2

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *ArXiv.* https://doi.org/10.48550/arXiv.1705.07874

Negrín-Díaz, I. A., Kripka, M., & Yepes, V. (2023). Metamodel-assisted design optimization in the field of structural engineering: A literature review. *Structures, 52,* 609-631. https://doi.org/10.1016/j.istruc.2023.04.006

Ovtsarenko, O. (2025). Meta-model_1313338 development, training, results evaluation. *Zenodo.* https://doi.org/10.5281/zenodo.17481678

Ovtšarenko, O. (2025). Innovative techniques for e-learning log data processing: Trends and methods. *Journal of Innovation & Knowledge, 10*(5), 100765. https://doi.org/10.1016/j.jik.2025.100765

Pardo, A., & Siemens, G. (2014). Ethical and privacy principles for learning analytics. *British Journal of Educational Technology, 45(*3), 438–450. https://doi.org/10.1111/bjet.12152

Paulson, J. A., & Tsay, C. (2024). Bayesian optimization as a flexible and efficient design framework for sustainable process systems. *ArXiv*. https://arxiv.org/abs/2401.16373

Simpson, T., Poplinski, J., Koch, P. N., & Allen, J. K. (2001). Metamodels for Computer-based Engineering Design: Survey and recommendations. *EWC, 17*, 129–150. https://doi.org/10.1007/PL00007198

Tirado, A. M., Mulholland, P., & Fernandez, M. (2024). Towards an Operational Responsible AI Framework for Learning Analytics in Higher Education. *ArXiv*. https://arxiv.org/abs/2410.05827

Topali, P., Ortega-Arranz, A., Rodríguez-Triana, M. J., Er, E., Khalil, M., & Akçapınar, G. (2024). Designing human-centered learning analytics and artificial intelligence in education solutions: a systematic literature review. *Behaviour & Information Technology, 44*(5), 1071–1098. https://doi.org/10.1080/0144929X.2024.2345295

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education. *International Journal of Educational Technology in Higher Education, 16*(39), 1–27. https://doi.org/10.1186/s41239-019-0171-0

Yağcı, M. (2022). Educational data mining: Prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments, 9*(1), 1-19. https://doi.org/10.1186/s40561-022-00192-z