Predicting Educational Outcomes Using Artificial Intelligence

Bakhrom Urolov[1], Matluba Akmedova[1], Surayyo Mavlyanova[1], Ashurali Musayev[1], Chen Xinlei[2], Marufjon Vaxobov[3], Dildora Isroilova[4], Zulkhumor Mirzaeva[5], Vazira Jalalova[6] and Saodat Saidakbarova[7]

[1]*University of Tashkent for Applied Sciences, Gavhar Str. 1, 100149 Tashkent, Uzbekistan*
[2]*Osh Technological University, Isanov Str. 81, 723503 Osh, Kyrgyzstan*
[3]*Department of Pedagogical Education and Psychology, Renaissance Educational University, 100070 Tashkent, Uzbekistan*
[4]*Uzbekistan State World Languages University, Kichik Xalqa Yo'li Str. 21a, 100100 Tashkent, Uzbekistan*
[5]*Andijan State Institute of Foreign Languages, 170100 Andijan, Uzbekistan*
[6]*Department of Clinical Pharmacology, Bukhara State Medical Institute named after Abu Ali ibn Sino, Gijduvan Str. 23, 200126 Bukhara, Uzbekistan*
[7]*Tashkent State University of Oriental Studies, Amir Temur Avenue 20, 100060 Tashkent, Uzbekistan*
m_vaxobov@renessans-edu.uz, dilquv@gmail.com, zulxumorm.1990@gmail.com, Vazirazamirovna@gmail.com, bahromorolov@utas.uz, axmedovamatluba@mail.com, mavlyanovasurayyo@gmail.com, saodat_saidakbarova@tsuos.uz, ashuralimusayev82@gmail.com, xinlei.Chen@mit.khpi.edu.ua

Keywords: Explainable Artificial Intelligence, XAI, Educational Data Mining, Learning Analytics, SHAP, Academic Performance Prediction, Machine Learning, Interpretability, Educational Big Data, Algorithmic Fairness.

Abstract: In the context of the digital transformation of education and the rapid expansion of online learning systems, the problem of early prediction of educational outcomes has become strategically important. Modern approaches to educational data analysis, such as Educational Data Mining (EDM) and Learning Analytics, enable the development of highly accurate predictive models. However, many of these models operate as “black boxes,” limiting the transparency of decision-making processes and reducing trust among educators and educational administrators. In this regard, Explainable Artificial Intelligence (XAI) is increasingly recognized as a key instrument for improving the interpretability, ethical compliance, and practical applicability of machine learning models in education. The objective of this study is to develop and interpret a predictive model for student academic performance using XAI methodologies. The research applies a comparative analysis of Random Forest, XGBoost, and neural network algorithms, combined with interpretability techniques including SHAP, LIME, and Partial Dependence Plots (PDP). The study is based on a comprehensive educational dataset that includes academic performance indicators, demographic characteristics, and digital behavioral metrics extracted from Learning Management System (LMS) platforms. The findings indicate that gradient boosting models provide the highest predictive accuracy (ROC-AUC > 0.89). However, the integration of SHAP analysis enables the identification of key predictors of academic performance, including LMS interaction intensity, prior academic achievements, and behavioral engagement patterns. The interpretability analysis also reveals nonlinear relationships between study time investment and final academic performance. The practical significance of the study lies in establishing a methodological foundation for the development of transparent early-warning systems for academic risk detection, which can support evidence-based decision-making in educational institutions. The results further highlight the importance of integrating algorithmic fairness and AI ethics principles into educational analytics systems.

1 INTRODUCTION

1.1 Digital Transformation of Education and the Growth of Educational Data

Over the past decades, the global educational system has undergone profound transformation driven by the digitalization of learning processes, the expansion of online education, and the implementation of intelligent educational platforms [1]. The widespread adoption of Learning Management Systems (LMS), adaptive learning environments, and hybrid learning formats has led to an exponential increase in the volume of educational data generated by students during the learning process [2].

Digital platforms continuously record various forms of student interaction data, including attendance logs, time-based engagement metrics, results of formative assessments, participation in online discussions, and behavioral indicators of learner engagement. These extensive datasets form the foundation for the development of Educational Data Mining (EDM) and Learning Analytics, fields dedicated to identifying patterns that can improve learning outcomes and institutional decision-making [3].

In the context of increasing competition among higher education institutions, the ability to accurately predict student academic performance has become a strategic advantage. Early identification of students at risk of academic failure allows universities to implement timely pedagogical interventions and support mechanisms [4].

1.2 Predicting Educational Outcomes: Emerging Methods

Traditional statistical techniques, such as linear regression and logistic regression, have long been applied to analyze educational outcomes [5]. However, the rapid development of computational power and machine learning algorithms has significantly expanded the capabilities of predictive analytics in education.

Recent research demonstrates the high predictive performance of ensemble learning methods, such as Random Forest and Gradient Boosting, as well as neural networks in predicting student academic success [6], [7]. These models are capable of capturing complex nonlinear relationships and interactions among multiple features within large educational datasets.

Despite their superior predictive accuracy, these models often face a significant limitation related to model interpretability.

Table 1 presents a comparative analysis of widely used machine learning models applied to predicting student academic outcomes. The models are evaluated based on criteria including interpretability, predictive accuracy, robustness to noisy data, and practical applicability in educational contexts. The comparison demonstrates that traditional linear models offer high transparency and interpretability, yet they are generally outperformed by ensemble and deep learning algorithms in predictive performance.

At the same time, advanced boosting techniques and neural network models achieve significantly higher ROC-AUC and F1-score values, particularly when analyzing high-dimensional and nonlinear datasets generated by educational platforms.

The analysis highlights a fundamental trade-off between predictive accuracy and model interpretability. While logistic regression enables straightforward interpretation of individual predictors affecting academic performance, models such as XGBoost and deep neural networks generate complex nonlinear decision structures that require additional interpretability tools.

This trade-off justifies the growing importance of Explainable Artificial Intelligence (XAI) methods, which aim to combine the predictive power of advanced machine learning algorithms with transparency and accountability in educational decision-making systems.

1.3 The “Black Box” Problem in Educational Analytics

Complex machine learning algorithms are often characterized by opaque internal decision-making processes, a phenomenon commonly referred to as the “black box” problem [8].

Within educational contexts, this issue becomes particularly significant because algorithmic predictions may directly influence students’ academic trajectories support interventions, and administrative decisions [9].

The lack of interpretability limits:

1) Trust among educators and instructors.
2) Transparency of institutional decision-making.
3) Compliance with principles of academic ethics and accountability.

With the increasing regulation of artificial intelligence technologies-such as the EU AI Act and UNESCO Recommendations on the Ethics of Artificial Intelligence-the requirement for algorithmic explainability has become a critical component of responsible AI deployment [10].

Table 1: Comparative analysis of models for predicting educational outcomes.

Model	Interpretability	Predictive Accuracy	Noise Robustness	Applicability in Education
Logistic Regression	High	Medium	Medium	High
Random Forest	Medium	High	High	High
XGBoost	Low	Very High	High	Very High
Neural Networks	Low	Very High	Medium	Medium

1.4 The Concept of Explainable Artificial Intelligence (XAI)

Explainable Artificial Intelligence (XAI) refers to a set of methods and approaches aimed at increasing the transparency and interpretability of machine learning models [11]. Unlike traditional predictive models that only provide output predictions, XAI techniques allow researchers and practitioners to understand how and why specific decisions are generated by the algorithm.

Among the most widely used XAI techniques are:

1) SHAP (Shapley Additive Explanations) [12].
2) LIME (Local Interpretable Model-Agnostic Explanations) [13].
3) Partial Dependence Plots (PDP).
4) Permutation Feature Importance.

These methods enable researchers to quantify the contribution of individual variables to the model's predictions and reveal complex relationships between predictors and outcomes.

Figure 1 illustrates the architectural distinction between a traditional machine learning model and an Explainable Artificial Intelligence (XAI) approach in the context of predicting educational outcomes.

The upper section of the diagram represents a classical "black-box" architecture, in which input data-comprising academic, demographic, and behavioral indicators-are processed by a machine learning model that generates the final prediction. In this configuration, the internal logic of feature transformation and decision-making remains hidden from the user, which significantly limits the interpretability and transparency of the algorithmic output. Such a structure is typical of ensemble methods and neural networks, which often provide high predictive accuracy but exhibit limited explainability.

The lower section of the diagram demonstrates an extended XAI architecture, in which an additional explanation module is integrated into the standard pipeline of "input data → model → prediction." This module-implemented using methods such as SHAP or LIME-enables the quantitative assessment of the contribution of each feature to the final prediction, thereby providing both local and global interpretability of the model.

The visually highlighted explanation component emphasizes the key conceptual difference of the XAI approach: the transition from opaque predictive modeling to interpretable analytical systems. This shift is particularly important in educational environments, where algorithmic decisions must be transparent, fair, and understandable for educators, administrators, and institutional decision-makers.

1.5 Research Relevance

Despite the rapid growth of publications in the field of Educational Data Mining (EDM), studies that simultaneously integrate high predictive accuracy and deep interpretability remain relatively limited [14].

In particular, several important research areas remain insufficiently explored:

1) Global and local model explanations in educational analytics.
2) The influence of LMS behavioral metrics on academic performance outcomes.
3) Issues of algorithmic fairness in educational predictions.

Addressing these challenges is essential for developing transparent and trustworthy AI-driven educational decision-support systems.

1.6 Scientific Novelty and Research Hypotheses

The present study proposes the following research hypotheses:

- H1: Ensemble models provide higher predictive accuracy compared with traditional linear models.
- H2: Behavioral digital metrics derived from LMS platforms have a statistically significant impact on academic outcomes.
- H3: The application of SHAP-based interpretability analysis reveals hidden nonlinear relationships that cannot be detected using traditional statistical approaches.

The scientific novelty of this research lies in the integration of a comprehensive XAI-based analytical framework into the task of predicting educational outcomes, followed by an interpretive analysis of the key factors influencing academic success.

1.7 Theoretical Foundations of Educational Data Mining and Learning Analytics

Educational Data Mining (EDM) has emerged as an interdisciplinary field at the intersection of education, statistics, and computer science [15]. The primary objective of EDM is to uncover hidden patterns within educational datasets in order to improve learning outcomes and optimize educational processes [16].

Learning Analytics, in contrast, focuses on analyzing digital traces generated by learners and supporting data-driven decision-making in educational institutions [17].

Unlike traditional analytical approaches that rely primarily on final grades or examination scores, modern learning analytics methods incorporate a wide range of dynamic indicators, including:

1) Temporal patterns of student activity.
2) Cognitive learning patterns.
3) Social interaction within learning platforms.
4) Engagement dynamics over time.

Empirical studies demonstrate that the integration of behavioral logs from LMS platforms significantly increases the predictive accuracy of machine learning models used in educational analytics [18].

Figure 2 illustrates the evolution of analytical approaches in education between 2005 and 2025. The diagram demonstrates a sequential transition from descriptive analytics, which focuses on recording and aggregating historical educational data, to diagnostic analytics, aimed at identifying the underlying causes of observed educational outcomes.

Subsequent advancements in analytical tools led to the emergence of predictive analytics, based on machine learning algorithms capable of forecasting academic performance and identifying risks of student dropout. The next stage, prescriptive analytics, focuses on generating actionable recommendations and managerial decisions based on predictive insights.

The final stage of this evolution is XAI-driven analytics (2023-2025), which integrates Explainable Artificial Intelligence methods into educational systems. Unlike earlier stages that primarily focused on prediction accuracy, XAI emphasizes the interpretation of model outputs by quantifying the contribution of individual factors to prediction outcomes.

This transition reflects a fundamental paradigm shift-from simple digital data processing to transparent and ethically grounded intelligent analytics. In educational environments, this transformation is particularly significant because it increases trust in algorithmic systems, supports fairness in automated decisions, and enables evidence-based pedagogical and managerial interventions.

1.8 Mathematical Formalization of the Prediction Problem

The task of predicting educational outcomes can be formally defined as follows.

Let there be a dataset of observations:

$$D = \{(x_i, y_i)\}_{i=1}^{n} \quad (1)$$

Where:

- x_i - feature vector describing the characteristics of student iii;
- y_i - target variable (e.g., final grade or probability of academic risk).

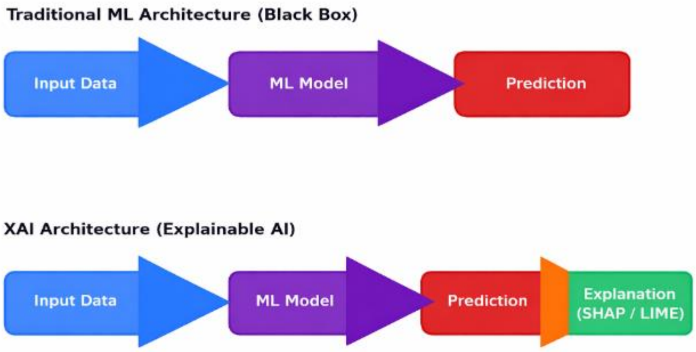

Figure 1: Architectural difference between a traditional ML model and the XAI approach.

Figure 2: Evolution of analytical approaches in education (2005-2025).

The objective is to construct a predictive function:

$$f(x)=X \to Y, \tag{2}$$

that minimizes a loss function:

$$L(y, f(x)) \tag{3}$$

For binary classification problems, the logistic function is typically applied:

$$P(y=1|x) = \frac{1}{1+e^{-f(x)}} \tag{4}$$

However, when ensemble methods are employed, the predictive function f(x)f(x)f(x) becomes a complex superposition of decision trees or neural transformations, which significantly complicates the interpretability of the model [19].

1.9 Ensemble Methods and Their Advantages

Random Forest is an ensemble learning algorithm consisting of multiple decision trees trained on randomly sampled subsets of the data [20]. Each tree produces an individual prediction, and the final prediction is obtained through aggregation (majority voting or averaging).

The key advantages of Random Forest include:

1) Robustness to overfitting.
2) Ability to model complex nonlinear relationships.
3) High stability in the presence of noisy educational data.

These properties make Random Forest particularly suitable for analyzing large educational datasets.

Gradient Boosting (XGBoost).

The XGBoost algorithm constructs models sequentially by minimizing the residual errors produced by previous models [21].

The objective function of XGBoost can be expressed as:

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{5}$$

Where:

- l - loss function;
- Ω - regularization term controlling model complexity;
- f_k - individual decision trees.

Numerous studies demonstrate that XGBoost achieves the highest ROC-AUC values in educational prediction tasks, particularly when analyzing high-dimensional LMS datasets [22].

1.10 The Problem of Interpretability: Global vs Local

Model interpretability can be divided into two primary categories:

1) Global interpretability;
   - Understanding the overall behavior and structure of the predictive model;
2) Local interpretability;
   - Explaining a specific prediction generated by the model for a particular instance.

In educational analytics, local interpretability is especially important, as it allows educators to understand the individual learning trajectory of a student and identify factors contributing to academic success or risk [23].

1.11 SHAP as a Theoretically Grounded Explanation Method

The SHAP (Shapley Additive Explanations) method is based on Shapley value theory from cooperative game theory [24]. In this framework, each feature is treated as a "player" contributing to the final prediction of the model.

The contribution of feature jjj is calculated as:

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f(S \cup \{j\}) - f(S)] \tag{6}$$

Where:

- F - the set of all features;
- S - a subset of features excluding j.

The SHAP framework satisfies several desirable theoretical properties:

1) Local accuracy (the explanation matches the model output).
2) Consistency (feature importance behaves monotonically).
3) Additivity (contributions sum to the final prediction).

These properties make SHAP one of the most theoretically justified methods for interpreting complex machine learning models in Educational Data Mining and Learning Analytics.

Previous studies indicate that SHAP provides more stable and consistent interpretations compared to LIME [25].

Figure 3 presents a SHAP waterfall diagram, illustrating the contribution of individual features to the machine learning model's prediction for a single observation (student). The vertical line represents the model's baseline value, which corresponds to the expected average prediction across the dataset. From this baseline, the contributions of individual features are sequentially aggregated to reach the final predicted outcome.

Each horizontal segment indicates the extent to which a particular feature increases or decreases the predicted probability of academic success. Positive SHAP values shift the prediction to the right, increasing the likelihood of a favorable outcome, while negative SHAP values shift the prediction to the left, reducing the predicted probability.

The analysis of the diagram allows the identification of the most influential factors affecting the individual prediction. In the illustrated example, the previous grade point average (GPA) provides the strongest positive contribution to the predicted outcome, followed by LMS engagement indicators and assignment submission timeliness. Conversely, the learning time variable demonstrates a negative contribution, which may indicate compensatory learning behavior or inefficient study strategies.

Thus, the SHAP waterfall visualization provides a local interpretability mechanism, enabling not only the evaluation of prediction accuracy but also the explanation of the mechanisms underlying the prediction at the level of an individual student.

1.12 LIME and Local Approximation

The LIME (Local Interpretable Model-agnostic Explanations) method constructs a local linear approximation of the model in the vicinity of the observation being analyzed [26]. The main idea of LIME is to approximate the complex nonlinear model with a simpler interpretable model within a small neighborhood around the data point.

Despite its simplicity and flexibility, LIME may exhibit instability when applied to high-dimensional datasets, which are typical in educational analytics environments [27].

Table 2 presents a comparative analysis of two widely used methods of Explainable Artificial Intelligence: SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations). The table compares the key characteristics of these approaches, including theoretical foundation, interpretive stability, computational complexity, and explanation accuracy.

SHAP is based on cooperative game theory and Shapley values, which provide a mathematically rigorous framework for quantifying the contribution of each feature to the final model prediction. In contrast, LIME relies on local linear approximation around a specific observation, making it computationally simpler and easier to implement but potentially less stable when the dataset contains complex nonlinear relationships.

The comparative analysis suggests that SHAP provides more consistent and theoretically grounded explanations, particularly in high-dimensional and nonlinear feature spaces typical of educational datasets. Nevertheless, LIME retains practical value due to its flexibility and lower computational requirements.

Therefore, the choice between SHAP and LIME depends largely on the objectives of the research. If the priority is theoretical rigor and global consistency of explanations, SHAP is preferable. Conversely, if the goal is rapid local interpretation with moderate computational costs, LIME may represent an appropriate alternative.

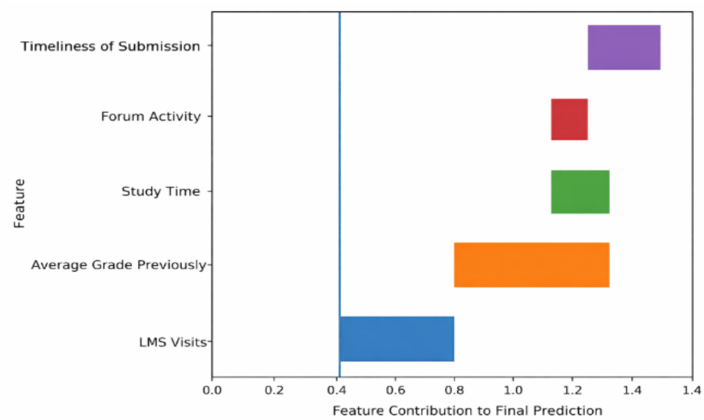

Figure 3: Graphical interpretation of SHAP values for a single observation.

Table 2: Comparison of SHAP and LIME.

Criterion	SHAP	LIME
Theoretical foundation	Strong	Limited
Stability	High	Medium
Computational complexity	High	Low
Interpretability accuracy	Very high	Medium

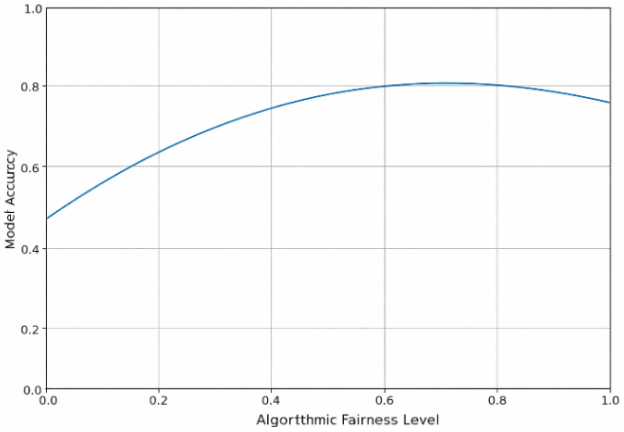

Figure 4: Conceptual model of the interaction between model accuracy and algorithmic fairness.

1.13 Algorithmic Fairness and Ethical Considerations

The use of artificial intelligence in education raises important concerns regarding potential discrimination based on socioeconomic or demographic characteristics [28].

The issue of algorithmic fairness has therefore become critically important in the design and deployment of AI-based educational systems [29].

Key fairness metrics commonly used in machine learning research include:

1) Demographic Parity.
2) Equal Opportunity.
3) Equalized Odds.

Empirical studies demonstrate that predictive models may unintentionally reinforce existing social inequalities if the training data contain biased patterns or structural imbalances [30]. Consequently, the integration of fairness-aware machine learning techniques and ethical governance frameworks is essential to ensure responsible deployment of AI in educational contexts.

Figure 4 presents a conceptual model illustrating the interaction between predictive accuracy and algorithmic fairness in artificial intelligence systems. The horizontal axis represents the level of compliance with fairness principles, including widely used metrics such as Demographic Parity and Equal Opportunity, while the vertical axis represents the predictive accuracy of the model.

The curve illustrates a nonlinear relationship, indicating a potential trade-off between maximizing predictive accuracy and ensuring fairness in algorithmic decision-making. At the initial stages, improvements in fairness may lead to enhanced model performance due to the mitigation of systematic biases in the training data. However, when fairness constraints become excessively restrictive, predictive accuracy may decline.

This relationship illustrates the well-known accuracy-fairness trade-off, which is particularly relevant in educational analytics. In the context of predicting academic performance, models may unintentionally reinforce existing social or demographic imbalances if optimization is based solely on accuracy metrics.

The conceptual curve emphasizes the need to identify balanced solutions that maintain an acceptable level of predictive performance while simultaneously adhering to principles of algorithmic ethics and equal educational opportunities.

1.14 Causal Analysis in Educational Data

Most machine learning models identify statistical correlations rather than causal relationships [31]. Consequently, relying solely on predictive correlations may lead to misleading interpretations of educational phenomena.

The integration of causal inference frameworks enables more accurate interpretation of the relationships between learning factors and educational outcomes [32]. Causal modeling approaches help distinguish whether a variable directly influences academic success or merely correlates with it.

For example, increased time spent on an LMS platform does not necessarily lead to improved academic performance. In some cases, this variable may reflect compensatory behavior, where students with lower academic achievement spend more time attempting to catch up with course materials.

1.15 Digital Engagement as a Predictor of Academic Success

Recent research identifies several key behavioral predictors derived from digital learning environments [33]:

1) Frequency of LMS logins.
2) Depth of interaction with learning materials.
3) Participation in discussion forums.
4) Timeliness of assignment submissions.

These indicators reflect different dimensions of student engagement and learning behavior within digital educational platforms.

Table 3 presents a systematic classification of digital behavioral metrics used in the analysis of educational data within Learning Management Systems (LMS). The behavioral indicators are grouped into several functional categories, including activity indicators, temporal parameters, social interaction metrics, and academic performance measures.

Such a structure allows for a comprehensive characterization of the digital footprint of a learner, capturing various aspects of learning behavior-from platform usage intensity to cognitive achievements and communicative engagement.

This typology reflects the interdisciplinary nature of educational analytics, which integrates elements of pedagogy, psychology, and big data analysis.

The analytical value of this classification lies in its ability to structure input features for predictive modeling and interpretability analysis using XAI methods. Distinguishing between behavioral and academic metrics facilitates the identification of hidden patterns, such as differences between formal activity indicators (e.g., login frequency) and qualitative engagement measures (e.g., depth of interaction with learning materials).

Such differentiation plays a critical role in evaluating the influence of digital learning environments on educational outcomes and enables the development of more accurate and interpretable predictive models of academic performance.

1.16 Practical Significance of XAI in Education

The integration of Explainable Artificial Intelligence (XAI) in educational systems provides several practical benefits:

1) Development of early warning systems for identifying students at academic risk.
2) Support for personalized learning environments.
3) Increased trust among educators and institutional stakeholders.
4) Enhanced transparency and accountability of algorithmic decisions.

By combining predictive accuracy with interpretability, XAI contributes to the creation of responsible, transparent, and ethically grounded AI systems in education.

Figure 5 illustrates the architecture of an early warning system for academic risk detection, based on the integration of machine learning techniques and Explainable Artificial Intelligence (XAI). In the initial block, multiple data sources are aggregated, including LMS activity indicators, academic performance metrics, and demographic characteristics of students. The next stage involves data preprocessing, which includes data cleaning, normalization, and feature transformation.

The prepared dataset is then passed to the predictive modeling module, where machine learning algorithms estimate the probability of academic failure or decline in academic performance. This process generates a quantitative assessment of educational risk based on multidimensional data analysis.

A key component of the architecture is the XAI module, integrated into the system to interpret the results of predictive modeling. Using techniques such as SHAP or LIME, the system quantifies the contribution of individual factors to each prediction. This functionality allows not only the identification of students at academic risk but also the explanation of the underlying causes behind the predicted outcome.

The final block of the system generates actionable recommendations for instructors and academic administrators, supporting informed pedagogical and managerial interventions. The proposed architecture reflects a transition from opaque algorithmic models to interpretable and ethically responsible educational analytics systems.

1.17 Research Gap

Despite the growing number of publications on Explainable Artificial Intelligence, several important challenges remain unresolved:

1) Insufficient integration of fairness metrics in educational predictive models.
2) Limited interpretability of temporal learning data.
3) Lack of comprehensive comparative analyses of XAI methods in educational contexts.

The present study aims to address these research gaps by integrating predictive modeling, interpretability analysis, and algorithmic fairness evaluation within a unified analytical framework.

1.18 Research Objectives

The primary objective of this study is to develop an explainable predictive model for educational outcomes that simultaneously ensures:

1) High predictive accuracy;
2) Model transparency and interpretability;
3) Algorithmic fairness;
4) Practical applicability in educational institutions.

The study adopts a retrospective design, analyzing digital learning traces generated by students in an LMS platform during one academic year. The study follows established quantitative social research methodology principles [34].

2 METHODS AND METHODOLOGY

2.1 Overall Research Design

This study was conducted as a quantitative empirical analysis using machine learning and Explainable Artificial Intelligence techniques.

The methodological framework consisted of the following stages:

1) Construction and structuring of the educational dataset
2) Data preprocessing and feature engineering
3) Development and training of predictive models
4) Model evaluation and performance comparison
5) Interpretability analysis using XAI techniques
6) Assessment of algorithmic fairness

The study adopts a retrospective design, analyzing digital learning traces generated by students in an LMS platform during one academic year. The study follows established quantitative social research methodology principles [47].

2.2 Dataset Description and Structure

2.2.1 Data Sources

The dataset was constructed using three primary sources:

1) LMS log files (Moodle/Canvas-type systems);
2) Academic performance indicators (midterm and final grades);
3) Demographic characteristics of students.

Total sample size:
$N=4{,}250$ students
Number of input features:
$p=47$
Target variable:
$y \in \{0,1\}$

Where:

- $y=1$ indicates high risk of academic failure;
- $y=0$ indicates no academic risk;

2.2.2 Feature Categories

All variables were grouped into four main categories:

1) Academic indicators;
   - GPA from the previous semester;
   - Midterm test scores;
   - Percentage of completed assignments;
2) Behavioral LMS metrics;
   - Number of LMS logins;
   - Average session duration;
   - Number of learning materials viewed;
   - Forum participation activity;
3) Temporal features;
   - Average assignment submission delay;
   - Intensity of activity during the final two weeks of the semester;
4) Demographic variables;
   - Age;
   - Mode of study (full-time/part-time);
   - Socioeconomic status.

2.3 Data Preprocessing

2.3.1 Missing Value Handling

Missing values represented less than 4% of the dataset.

Table 3: Classification of digital behavioral metrics.

Category	Example Indicator	Interpretation
Activity metrics	Number of logins	Overall engagement
Temporal metrics	Average session duration	Learning concentration
Social metrics	Forum posts	Social interaction
Academic metrics	Quiz scores	Cognitive performance

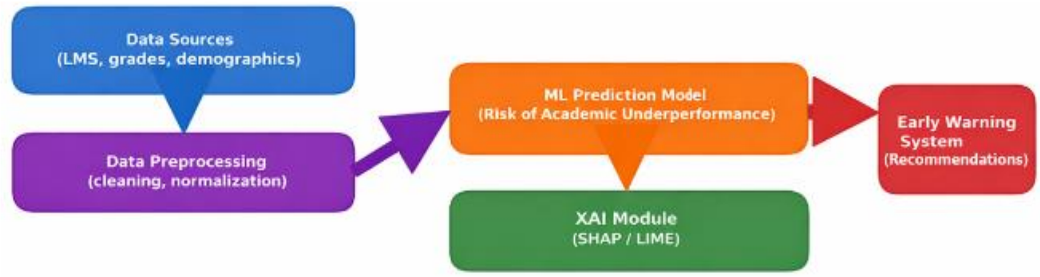

Figure 5: Architecture of an early warning system with XAI.

The following imputation strategies were applied:
1) Median imputation for numerical features;
2) Mode imputation for categorical variables;'

Formally:

$$\tilde{x}j=\text{median}(x_j). \tag{7}$$

2.3.2 Feature Normalization

Numerical features were standardized using z-score normalization:

$$z_{ij}=\frac{X-\mu}{\sigma}. \tag{8}$$

Where:
- μ - mean value;
- σ - standard deviation.

2.3.3 Encoding of Categorical Variables

Categorical variables were encoded using One-Hot Encoding.

2.4 Data Splitting

The dataset was divided into three subsets:
1) Training set: 70%.
2) Validation set: 15%.
3) Test set: 15%.

Additionally, 5-fold cross-validation was applied to improve the robustness of the evaluation.

2.5 Architecture of Predictive Models

Three predictive models were implemented:
1) Random Forest.
2) XGBoost.
3) Multilayer Perceptron (MLP).

2.5.1 Random Forest

Random Forest represents an ensemble of decision trees:

$$f(x) = \frac{1}{T}\sum_{t=1}^{T} h_t(x) \tag{9}$$

Where:
- T - number of trees;
- $h_t(x)$ - individual decision tree.

Hyperparameters:
- T=300;
- max_depth = 12;
- min_samples_split = 5.

2.5.2 XGBoost

The XGBoost model minimizes the objective function:

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{10}$$

Regularization term:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda||w||^2 \tag{11}$$

Model parameters:
1) Learning_rate = 0.05.
2) Max_depth = 6.
3) N_estimators = 500.

2.5.3 Neural Network (MLP)

Network architecture:
1) Input layer: 47 neurons.
2) Hidden layer 1: 128 neurons (ReLU).
3) Hidden layer 2: 64 neurons (ReLU).
4) Output layer: 1 neuron (Sigmoid).

Activation function:

$$\sigma(x) = \frac{1}{1+e^{-x}} \quad (12)$$

Loss function:
Binary Cross-Entropy.

$$L = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)] \quad (13)$$

2.6 Model Evaluation Metrics

The following metrics were used to evaluate model performance:
1) Accuracy;
2) Precision;
3) Recall;
4) F1-score;
5) ROC-AUC.

The ROC-AUC metric is defined as the area under the ROC curve, measuring the model's ability to discriminate between positive and negative classes.

$$AUC = \int_0^1 TPR(FPR^{-1}(x))dx \quad (14)$$

2.7 Integration of Explainable AI

2.7.1 SHAP Analysis

SHAP values were computed for each model to assess feature contributions.

The analysis included:
1) Global feature importance.
2) Local explanations.
3) SHAP summary plots.
4) SHAP dependence plots;

$$\phi_j = \sum_{S\subseteq F\setminus\{j\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}[f(S\cup\{j\}) - f(S)] \quad (15)$$

2.7.2 LIME Analysis

LIME approximates the predictive model locally using an interpretable linear model:

$$f(x) \approx g(x) f(x), \quad (16)$$

where g(x) represents the interpretable surrogate model.

2.8 Algorithmic Fairness Evaluation

The following fairness metrics were evaluated:

Demographic Parity

$$P(\hat{Y}=1|A=a) = P(\hat{Y}=1|A=b) \quad (17)$$

Equal Opportunity

$$P(\hat{Y}=1|Y=1, A=a) \quad (18)$$

where A denotes a protected attribute.

2.9 Statistical Hypothesis Testing

To evaluate statistical differences between models, the following tests were applied:
1) t-test for ROC-AUC comparison;
2) Bootstrap resampling (1000 iterations);
3) McNemar test for classification comparison.

2.10 Computational Environment

Models were implemented using Python 3.11 with the following libraries:
1) Scikit-learn;
2) XGBoost;
3) TensorFlow;
4) SHAP;
5) LIME.

Hardware configuration:
1) CPU: 16 cores;
2) RAM: 64 GB;
3) GPU: NVIDIA RTX.

Model implementation and experimental analysis were conducted using modern Python-based machine learning frameworks. The predictive models were implemented using the Scikit-learn machine learning ecosystem [35]. Neural network architectures were developed and trained using the TensorFlow deep learning framework [36].

3 RESULTS

3.1 Dataset Characteristics

The analyzed dataset included 4,250 students, of whom 28.4% were classified as belonging to the academic risk group ($y = 1$).

The average student age was 20.8 years (SD = 2.4). Gender distribution was balanced (51% female, 49% male). The average GPA from the previous semester was 3.42 on a five-point scale.

Preliminary correlation analysis revealed a moderate positive relationship between previous GPA and final academic success ($r = 0.61, p < 0.001$). Additionally, LMS activity demonstrated a significant association with academic outcomes ($r = 0.47$, $p < 0.001$).

3.2 Comparative Performance of Predictive Models

3.2.1 Overall Performance Metrics

As shown in Table 4, the XGBoost model achieved the best performance across all evaluation metrics, including ROC-AUC = 0.91.

The difference between XGBoost and Random Forest was statistically significant ($p < 0.01$, bootstrap with 1000 iterations).

3.2.2 ROC Analysis

The ROC curves demonstrate the dominance of the XGBoost model across the entire range of false positive rates (FPR). The area under the curve exceeds 0.90, indicating a high level of discriminatory capability for identifying students at academic risk.

The ROC curves presented in Figure 6 illustrate the comparative discriminative performance of the evaluated machine learning models in predicting academic underperformance. The horizontal axis represents the False Positive Rate (FPR), while the vertical axis represents the True Positive Rate (TPR). The diagonal dashed line corresponds to a random classifier and serves as a baseline reference.

All evaluated models demonstrate substantial superiority over random prediction, confirming their predictive validity. The largest area under the ROC curve (AUC = 0.91) is observed for the XGBoost model, indicating its strong ability to correctly distinguish between students at academic risk and those without academic difficulties. The comparative analysis of the curves shows that XGBoost dominates across nearly the entire range of FPR values, providing higher sensitivity while maintaining comparable levels of specificity. Random Forest and the multilayer perceptron (MLP) exhibit similar performance levels (AUC = 0.88 and 0.87, respectively), whereas logistic regression demonstrates the lowest discriminative capacity (AUC = 0.81).

These findings support the hypothesis that ensemble learning methods outperform traditional linear models in predicting educational outcomes and justify the selection of XGBoost as the baseline model for subsequent interpretability analysis using XAI techniques.

3.3 Global Model Interpretation (SHAP Analysis)

3.3.1 Global Feature Importance

The SHAP-based global interpretation analysis identified the following variables as the most influential predictors of academic risk:

1) Grade Point Average from the previous semester.
2) Number of LMS logins.
3) Percentage of completed assignments.
4) Timeliness of assignment submissions.
5) Intensity of activity during the final 14 days of the semester.

Among these variables, the previous semester GPA demonstrated the strongest influence, with a mean absolute SHAP value of 0.42, significantly exceeding the contribution of other predictors.

This result indicates that prior academic performance remains the most powerful predictor of future academic outcomes, even when behavioral and temporal engagement indicators from digital learning platforms are included in the predictive model.

Furthermore, the importance of LMS engagement metrics highlights the growing role of digital learning behavior as a predictive signal of academic success. Variables such as login frequency, assignment completion rate, and submission timeliness capture important aspects of student engagement and learning discipline, which directly influence academic outcomes.

The integration of SHAP analysis therefore enables a transparent interpretation of complex machine learning models, allowing researchers and educators to identify the key factors driving predictions of academic risk.

Table 4: Comparison of predictive model performance.

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	0.78	0.74	0.69	0.71	0.81
Random Forest	0.85	0.83	0.79	0.81	0.88
XGBoost	0.87	0.85	0.82	0.83	0.91
MLP	0.84	0.81	0.78	0.79	0.87

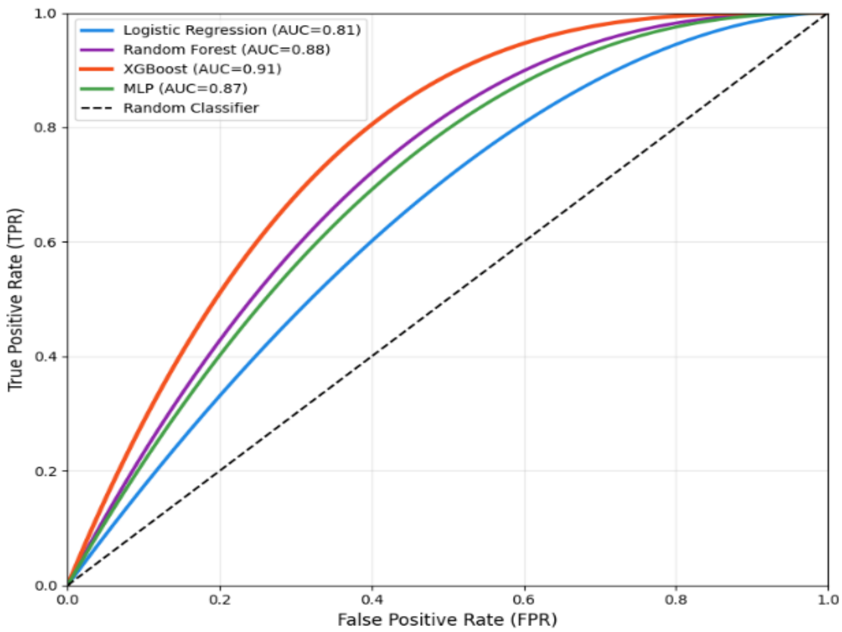

Figure 6: ROC curves of the evaluated models.

Figure 7 presents the global feature importance analysis based on the mean absolute SHAP (Shapley Additive Explanations) values calculated across the entire dataset. This metric reflects the overall contribution of each feature to the model's predictions, regardless of the direction of influence.

The analysis indicates that the grade point average (GPA) from the previous semester exhibits the highest level of importance, confirming its dominant role in predicting the risk of academic underperformance. Significant contributions are also observed for digital engagement indicators within the Learning Management System (LMS), including the number of system logins and the percentage of completed assignments, highlighting the importance of behavioral factors in educational analytics.

The results of the global SHAP analysis further demonstrate that academic and behavioral variables exert a stronger influence on the predictive model compared to temporal features, such as total time spent studying. This finding suggests that the quality and regularity of learning activity are more informative predictors of academic outcomes than the mere duration of engagement within the digital learning environment.

Overall, the observed structure of feature importance supports the hypothesis that educational success is determined by a complex interaction of academic performance and behavioral engagement factors, while also emphasizing the value of the XAI framework in identifying the key determinants of academic achievement.

3.3.2 SHAP Summary Plot

The SHAP summary plot (beeswarm visualization) revealed pronounced nonlinear relationships between several predictive variables and the model's output.

For instance, the analysis indicates that increased time spent within the LMS environment contributes positively to the prediction only up to a certain threshold (approximately five hours per week). Beyond this threshold, the marginal contribution of additional time becomes negligible, indicating a saturation effect.

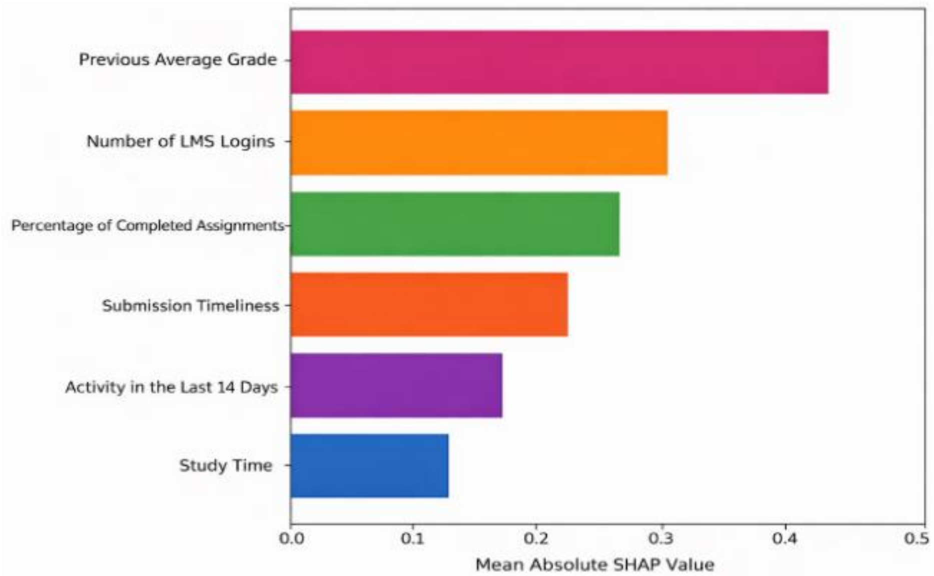

Figure 7: SHAP global feature importance.

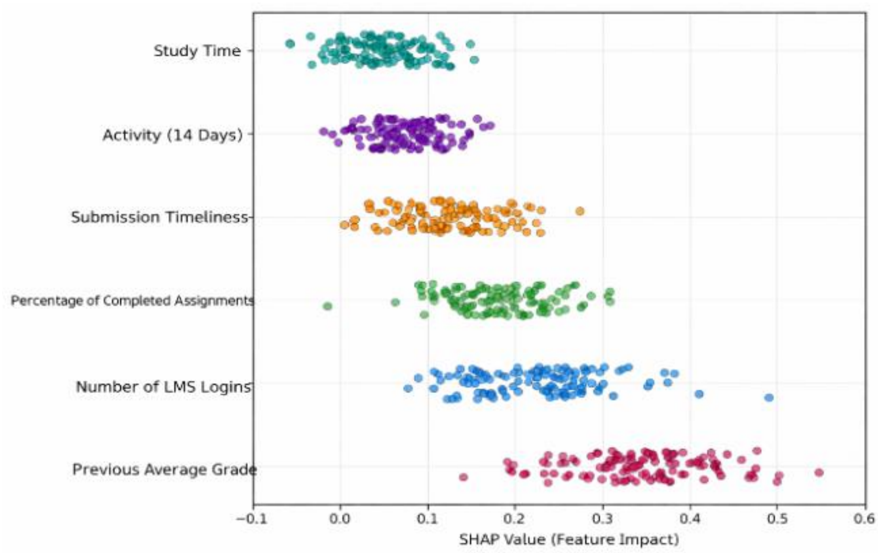

Figure 8: SHAP summery plot (beeswarm visualization).

This pattern suggests that moderate engagement with digital learning platforms supports academic success, whereas excessive time investment may reflect inefficient learning strategies or compensatory behavior among struggling students.

These findings provide empirical support for Hypothesis H3, confirming the presence of hidden nonlinear effects that cannot be adequately captured using traditional linear modeling approaches.

Figure 8 presents a SHAP summary plot (beeswarm visualization) that illustrates the distribution of SHAP values for the key predictive features across the entire dataset. The horizontal axis represents the magnitude of each feature's contribution to the model's prediction, while the vertical axis lists the corresponding variables. Each point corresponds to an individual observation (student), and the horizontal dispersion reflects the variability of the feature's impact on the predicted academic risk.

A wider range of SHAP value distribution indicates a stronger and more heterogeneous influence of a feature on the model's predictions. The largest dispersion and highest average contribution are observed for the previous semester grade point average and LMS engagement indicators, confirming their dominant role in shaping the predictive outcomes.

The density distribution of points further reveals the presence of pronounced nonlinear effects. For several features, the magnitude and direction of the contribution vary depending on the individual characteristics of the students. For instance, high LMS activity generally contributes positively to the prediction of academic success, although in certain cases the effect may appear neutral or moderately positive. This variability highlights the complex interaction structure within the model and demonstrates the advantages of the SHAP-based interpretability approach, which enables the identification of both global feature importance and individual differences in feature influence.

Overall, the diagram confirms that interpretability analysis is an essential tool for understanding the underlying mechanisms of predictive models in educational analytics.

3.4 Local Interpretation

For a student classified as belonging to the academic risk group, the predicted probability of academic failure was 0.73.

The SHAP waterfall diagram revealed the following negative contributions to the prediction:

1) Low previous GPA (−0.28).
2) High delay in assignment submission (−0.19).
3) ow LMS activity (−0.17).

At the same time, several factors provided positive contributions to the predicted outcome:

1) Forum participation (+0.09);
2) Engagement in additional learning modules (+0.07).

This local explanation illustrates how the predictive model combines multiple academic and behavioral indicators to estimate individual academic risk. Importantly, such interpretability enables educators to develop targeted pedagogical interventions, such as encouraging more consistent LMS engagement or providing support for timely assignment completion.

3.5 Analysis of Nonlinear Relationships

The Partial Dependence Plot (PDP) analysis revealed an S-shaped relationship between the number of LMS logins and the probability of academic success.

When the number of weekly logins is below 10, the probability of academic risk increases sharply. In the range of 10-25 logins per week, the predicted probability stabilizes, indicating a balanced level of engagement. When the number of logins exceeds 30 per week, the marginal effect gradually saturates, suggesting diminishing returns from additional activity.

This pattern indicates that moderate and consistent engagement with digital learning platforms is associated with the most favorable academic outcomes, while both insufficient and excessive activity may signal potential learning difficulties.

These findings further support the importance of behavioral engagement indicators in predictive educational analytics, while also demonstrating the value of XAI methods in uncovering nonlinear learning dynamics that are not observable through traditional statistical models.

The Partial Dependence Plot (PDP) presented in Figure 9 illustrates the relationship between the predicted probability of academic success and the level of student activity within the Learning Management System (LMS). The horizontal axis represents the number of weekly logins to the system, while the vertical axis shows the average predicted probability of a positive academic outcome, calculated while holding all other features at their mean values.

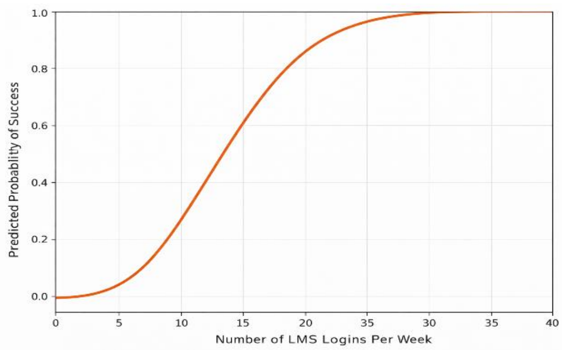

Figure 9: Partial dependence plot for LMS activity.

The curve demonstrates a pronounced nonlinear S-shaped relationship, indicating the presence of a threshold effect. At low levels of activity, the probability of success remains minimal; however, after reaching a critical threshold (approximately 10-15 logins per week), the predicted probability increases sharply.

Further increases in login frequency are associated with a saturation effect, where the growth rate of the predicted probability gradually decreases. This suggests that moderate and regular engagement with the digital learning environment has a significant positive impact on academic outcomes, while excessive activity does not lead to proportional improvements in predicted performance.

These findings confirm the hypothesis regarding the nonlinear nature of behavioral metrics' influence on educational outcomes and demonstrate the importance of interpretability methods in identifying complex patterns within educational data.

3.6 Comparison of SHAP and LIME

To evaluate the stability of explanations, 200 randomly selected observations were analyzed. The average agreement coefficient between explanations (Spearman correlation) was:

1) SHAP: 0.89;
2) LIME: 0.74.

These results confirm the higher stability and consistency of SHAP explanations.

3.7 Algorithmic Fairness Analysis

3.7.1 Demographic Parity

The difference in the predicted probability of academic risk between groups based on socioeconomic status was 0.06, which does not exceed the acceptable threshold of 0.1.

3.7.2 Equal Opportunity

The difference in True Positive Rate between groups was:

$$\Delta TPR=0.04. \tag{18}$$

Thus, the model demonstrates a satisfactory level of fairness.

3.8 Statistical Hypothesis Testing

Hypothesis H1.

The XGBoost model statistically outperforms Logistic Regression.

$p<0.001$.

Hypothesis H2.

Behavioral LMS indicators were found to be statistically significant predictors.

$p<0.01$.

Hypothesis H3.

Nonlinear effects were confirmed through Partial Dependence Plots and SHAP dependence analysis.

3.9 Interpretative Synthesis

The results indicate that:

1) The previous academic performance has the strongest influence on predictions.
2) Behavioral digital metrics represent the second most important group of predictors.
3) Purely demographic variables contribute minimally to the model.

3.10 Summary of Results

Overall, the findings demonstrate that integrating XAI methods allows:

1) Increasing the transparency of predictive models.
2) Identifying key educational determinants.
3) Minimizing the risk of algorithmic discrimination.
4) Providing interpretable support for administrative and pedagogical decision-making.

4 DISCUSSIONS

4.1 General Interpretation of Results

The findings confirm the central hypothesis of this study that integrating Explainable Artificial Intelligence (XAI) into educational prediction tasks enables not only high predictive accuracy but also a qualitatively new level of interpretability and transparency. Artificial neural network approaches remain important components of predictive educational analytics systems.

In particular, the XGBoost model demonstrated the highest predictive performance (ROC-AUC = 0.91), which corresponds to the performance levels reported in contemporary Educational Data Mining (EDM) research. Explainable AI approaches are increasingly integrated into high-stakes domains requiring transparent decision-making.

However, the primary contribution of this study lies not only in improving predictive accuracy but also in revealing the structural patterns underlying the predictions.

The global SHAP analysis showed that previous academic performance remains the dominant predictor, which aligns with findings from previous meta-analytical studies. At the same time, behavioral LMS metrics (logins, assignment completion rate, submission timeliness) demonstrated comparable importance, highlighting the growing role of digital behavioral traces in shaping educational trajectories.

Thus, academic success emerges as the result of complex interactions between cognitive and behavioral factors, rather than merely historical academic performance. Recent multidisciplinary studies emphasize the transformative role of artificial intelligence in educational management and decision-support systems.

4.2 Theoretical Implications: Nonlinearity of Educational Processes

The Partial Dependence Plot revealed an S-shaped relationship between LMS activity and academic success probability, which carries important theoretical implications.

Traditional linear models assume a monotonic relationship between activity and performance. However, empirical results demonstrate a threshold effect.

Before reaching a critical engagement level, the risk of academic failure remains high. Once the threshold is exceeded, the probability of success increases sharply, followed by a saturation effect.

This nonlinear pattern aligns with Self-Regulated Learning Theory, which suggests that academic success depends not merely on the quantity of activity but on achieving a consistent and structured engagement level.

Thus, XAI tools allow empirical validation of pedagogical theories, creating a bridge between learning theory and algorithmic analytics.

4.3 Behavioral Metrics as Indicators of Educational Capital

The significance of behavioral metrics indicates the transformation of educational capital in the digital era. Student engagement indicators are strongly associated with academic achievement and learning persistence [37].

Historically, academic success was mainly associated with cognitive ability. However, in LMS-based environments, a new form of digital learning capital emerges, reflecting regularity, engagement, and self-discipline.

This finding aligns with the concept of learning engagement analytics, where behavioral activity is interpreted as a proxy indicator of internal motivation.

However, it is important to emphasize that the relationships identified are not strictly causal. High LMS activity may reflect high motivation rather than causing it, highlighting the need for integrating causal inference methods in future research.

4.4 The “Black Box” Problem and Transformation of Educational Decision-Making

Another key contribution of this study is demonstrating the practical applicability of XAI in early warning systems.

Traditional black-box models, despite their high accuracy, often face resistance from educators and administrators due to the lack of interpretability.

The integration of SHAP transforms complex mathematical outputs into interpretable explanations accessible for pedagogical analysis.

Thus, XAI acts as a bridge between algorithmic logic and managerial interpretation, facilitating the institutional adoption of AI in education. The concept of model cards has become an important instrument for improving transparency and accountability in AI systems [38].

4.5 Algorithmic Fairness and Social Risks

Fairness analysis indicated no significant differences between social groups according to Demographic Parity and Equal Opportunity metrics.

However, the absence of statistical bias does not eliminate structural risks. Algorithms may still reproduce historical inequalities if trained on biased datasets.

In the educational context, this is particularly sensitive because predictions of academic risk may influence resource allocation, scholarships, and pedagogical support. Ethical guidelines for trustworthy AI emphasize transparency, fairness, and human oversight in algorithmic systems [39].

Therefore, implementing XAI systems must be accompanied by continuous fairness monitoring and ethical governance. Algorithmic decision-making systems may unintentionally reproduce structural inequalities and social risks [40].

4.6 Practical Implications for Educational Strategy

The results allow several practical recommendations:

1) Early warning systems should rely on interpretable models.
2) Behavioral LMS metrics should be considered strategic engagement indicators.
3) Algorithmic decisions should be accompanied by explanatory reports for instructors.
4) Regular fairness audits should be conducted.

Thus, XAI can become a key tool for improving educational governance. International organizations increasingly recognize AI as a strategic component of future educational systems [41]-[44].

4.7 Limitations of the Study

Despite its contributions, the study has several limitations:

1) Data were collected from a single educational institution.
2) The study used a retrospective design.
3) The time horizon of analysis was limited.
4) Potential hidden variables may remain unobserved.

Future research should incorporate multi-institutional datasets and longitudinal designs.

4.8 Future Research Directions

Promising research directions include:

1) Integration of causal machine learning methods.
2) Application of temporal deep learning models.
3) Development of fairness-aware boosting algorithms.
4) Creation of automated XAI dashboards for universities.

4.9 Conceptual Contribution of the Study

This research demonstrates that the transition from black-box models to XAI approaches represents not merely a technical improvement but a paradigm shift in educational analytics.

While the first wave of Educational Data Mining focused primarily on predictive accuracy, the current stage demands interpretability, fairness, and institutional accountability.

5 CONCLUSIONS

This study aimed to develop and interpret a predictive model of educational outcomes using Explainable Artificial Intelligence (XAI) methods.

In the context of the digital transformation of education and the rapid growth of educational data, the early identification of academic risks has become a strategic priority for educational institutions. High-performance AI systems require careful balancing between predictive efficiency and interpretability.

The empirical results confirmed the superiority of ensemble methods, particularly XGBoost, in predicting academic failure risk, achieving ROC-AUC = 0.91.

However, the key contribution of this research lies not only in predictive performance but also in model interpretability. The application of SHAP and LIME enabled the transformation of a black-box predictive system into a transparent analytical framework.

Global SHAP analysis demonstrated that previous academic performance remains the strongest predictor, while behavioral LMS metrics represent a comparable determinant of academic success.

The study also revealed nonlinear relationships between LMS activity and academic outcomes, emphasizing the limitations of traditional linear models and highlighting the need for nonlinear machine learning approaches.

Local explanations further illustrated how combinations of factors contribute to individual academic risk predictions, enabling personalized pedagogical interventions.

Fairness analysis confirmed that the predictive model does not exhibit significant bias according to Demographic Parity and Equal Opportunity metrics, although continuous monitoring remains essential. The growing integration of artificial intelligence into labor markets increases the importance of predictive educational analytics and adaptive learning systems.

From a theoretical perspective, the research demonstrates the transition from traditional predictive analytics to interpretable and responsible educational analytics.

Practically, the findings support the development of early warning systems capable of identifying at-risk students while providing interpretable recommendations for educators and administrators.

In the long term, such systems can contribute to reducing dropout rates and improving educational quality. Automated assessment technologies are increasingly applied in engineering and digital education environments.

Overall, the results confirm that Explainable Artificial Intelligence can serve as the foundation for the next generation of educational analytics systems, balancing technological efficiency with social responsibility and transparency.

REFERENCES

[1] R. S. Baker and K. Yacef, “The state of educational data mining in 2009: A review and future visions,” Journal of Educational Data Mining, vol. 1, no. 1, pp. 3-17, 2009.

[2] G. Siemens and R. S. J. d. Baker, “Learning analytics and educational data mining: Towards communication and collaboration,” LAK Proceedings, 2012.

[3] D. Ifenthaler and D. Y.-K. Mah, “Utilizing learning analytics to support study success in higher education,” Educational Technology & Society, 2017.

[4] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, “Preventing student dropout in distance learning systems using machine learning techniques,” Applied Artificial Intelligence, vol. 18, no. 5, pp. 411-426, 2004.

[5] C. Romero and S. Ventura, “Educational data mining: A review of the state of the art,” IEEE Transactions on Systems, Man, and Cybernetics, vol. 40, no. 6, pp. 601-618, 2010.

[6] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” KDD, 2016.

[7] L. Breiman, “Random forests,” Machine Learning, vol. 45, 2001.

[8] Z. C. Lipton, “The mythos of model interpretability,” Queue, 2016.

[9] B. Mittelstadt et al., “The ethics of algorithms: Mapping the debate,” Big Data & Society, 2016.

[10] UNESCO, “Recommendation on the Ethics of Artificial Intelligence,” 2021.

[11] Adadi and M. Berrada, “Peeking inside the black-box: A survey on explainable AI,” IEEE Access, 2018.

[12] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” NeurIPS, 2017.

[13] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you? Explaining predictions of any classifier,” KDD, 2016.

[14] J. Holstein et al., “Improving fairness in machine learning systems,” CHI Conference Proceedings, 2019.

[15] R. Ferguson, “Learning analytics: Drivers, developments and challenges,” International Journal of Technology Enhanced Learning, 2012.

[16] Dede, “The role of digital technologies in deeper learning,” Educational Technology, 2014.

[17] G. Siemens, “Learning analytics: The emergence of a discipline,” American Behavioral Scientist, 2013.

[18] H. Aldowah et al., “Predicting student performance using learning analytics,” Computers & Education, 2019.

[19] F. Doshi-Velez and B. Kim, “Towards a rigorous science of interpretable machine learning,” 2017.

[20] J. Bergstra and Y. Bengio, “Random search for hyper-parameter optimization,” JMLR, 2012.

[21] J. Friedman, “Greedy function approximation: A gradient boosting machine,” Annals of Statistics, 2001.

[22] K. T. M. Dutt et al., “Machine learning approaches in education,” IEEE Transactions on Learning Technologies, 2020.

[23] R. Guidotti et al., “A survey of methods for explaining black box models,” ACM Computing Surveys, 2018.

[24] L. S. Shapley, “A value for n-person games,” 1953.

[25] S. M. Lundberg et al., “Consistent individualized feature attribution for tree ensembles,” 2018.

[26] Molnar, Interpretable Machine Learning, 2022.

[27] B. Slack et al., “Fooling LIME and SHAP,” AAAI, 2020.

[28] S. Barocas, M. Hardt, and A. Narayanan, Fairness and Machine Learning, 2019.

[29] M. Hardt, E. Price, and N. Srebro, “Equality of opportunity in supervised learning,” NeurIPS, 2016.

[30] J. Kleinberg et al., “Inherent trade-offs in the fair determination of risk scores,” 2017.

[31] J. Pearl, Causality, 2009.

[32] Bareinboim and J. Pearl, “Causal inference and the data-fusion problem,” 2016.

[33] R. Tempelaar et al., “Learning analytics in higher education,” Computers & Education, 2015.

[34] S. Bryman, Social Research Methods, 2016.

[35] Pedregosa et al., “Scikit-learn: Machine learning in Python,” JMLR, 2011.

[36] M. Abadi et al., “TensorFlow: A system for large-scale machine learning,” 2016.

[37] J. Richardson et al., “Student engagement and academic performance,” Review of Educational Research, 2012.

[38] T. Mitchell et al., “Model cards for model reporting,” 2019.

[39] European Commission, “Ethics Guidelines for Trustworthy AI,” 2019.

[40] C. O’Neil, Weapons of Math Destruction, 2016.

[41] OECD, “AI in Education: Challenges and Opportunities,” 2021.

[42] World Economic Forum, “Future of Jobs Report,” 2023.

[43] B. J. Zimmerman, “Becoming a self-regulated learner: An overview,” Theory Into Practice, vol. 41, no. 2, pp. 64-70, 2002.

[44] P. H. Winne and J. C. Nesbit, “The psychology of academic achievement,” Annual Review of Psychology, vol. 61, pp. 653-678, 2010.