frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 29 August 2025
DOI 10.3389/feduc.2025.1632315

EDITED BY
Imran Ashraf,
Yeungnam University, Republic of Korea

REVIEWED BY
Ahmed Adil Nafea,
University of Anbar, Iraq
Junali Jena,
KIIT University, India

*CORRESPONDENCE
Iván Ortiz-Garcés
✉ ivan.ortiz@udla.edu.ec

†These authors have contributed equally to this work

RECEIVED 28 May 2025
ACCEPTED 04 August 2025
PUBLISHED 29 August 2025

CITATION
Guevara-Reyes R, Ortiz-Garcés I, Andrade R, Cox-Riquetti F and Villegas-Ch W (2025) Machine learning models for academic performance prediction: interpretability and application in educational decision-making. *Front. Educ.* 10:1632315.
doi: 10.3389/feduc.2025.1632315

Machine learning models for academic performance prediction: interpretability and application in educational decision-making

Rodrigo Guevara-Reyes[1†], Iván Ortiz-Garcés[2*†], Roberto Andrade[3†], Fernanda Cox-Riquetti[4†] and William Villegas-Ch[2†]

[1]Facultad de Ciencias e Ingenierías, Universidad Estatal de Milagro, Milagro, Ecuador, [2]Escuela de Ingeniería en Ciberseguridad, Facultad de ingenierias y ciencias aplicadas (FICA), Universidad de Las Américas, Quito, Ecuador, [3]Colegio de Ciencias e Ingenierías, Universidad San Francisco de Quito USFQ, Quito, Ecuador, [4]Escuela de Matemáticas, Facultad de ingenierias y ciencias aplicadas (FICA), Universidad de Las Américas, Quito, Ecuador

The integration of artificial intelligence in education has enabled the development of predictive models for academic performance. However, most existing approaches lack interpretability and do not provide actionable insights for decision-making. This study addresses these limitations by deploying optimized machine learning models, specifically XGBoost and Random Forest, to predict student performance considering geographically, institutional, socioeconomic, and academic factors. Unlike previous research focused only on accuracy, this work incorporates SHAP-based interpretability techniques and an interactive decision support system to analyze the impact of various variables on educational outcomes. The model was trained and validated on a dataset of 50,000 student records, optimized through hyperparameter tuning and cross-validation. Results indicate that XGBoost achieves an R^2 of 0.91, outperforming traditional approaches, and reduces the mean square error (MSE) by 15%. The feature importance analysis reveals that five variables explain 72% of the variability in performance, highlighting the influence of socioeconomic conditions, infrastructure, and the student-teacher ratio. In addition, simulations of educational policies show that improving teacher training and access to technology increases performance by 18% and reduces dropout by 12%. This study presents a scalable and interpretable predictive model that anticipates student performance and helps optimize educational strategies through artificial intelligence applied to decision-making.

KEYWORDS

adaptive learning, problem-solving, artificial intelligence in education, learning personalization, education, institutional use

1 Introduction

Using machine learning techniques in educational analytics has transformed how academic performance patterns are identified and intervention strategies are designed (Abdrakhmanov et al., 2024). The ability to predict student performance based on multiple factors allows educational institutions to implement preventive measures, optimize resources, and improve decision-making. However, much of the previous research has been limited by predictive models that, although they achieve acceptable levels of accuracy, lack interpretability and do not consider applicability in real-world settings (Ikegwu et al., 2024). This gap between predictive accuracy and interpretability constitutes a key barrier to the adoption of such systems in actual educational environments, where decisions must be both data-driven and explainable to educators and administrators.

This work arises from the need to overcome this gap by providing models that accurately predict academic performance and also offer transparency in obtaining these predictions. Educational institutions often struggle to identify the reasons for poor student performance, in part because most existing models operate as black boxes. Therefore, offering explainable results becomes essential to align predictive outcomes with pedagogical actions.

Research in academic performance prediction has identified multiple variables that impact student outcomes, including geographic, institutional, socioeconomic, and educational factors. Recent studies have shown that combining machine learning-based models with contextual data improves predictive accuracy by 10–15% over conventional methods (Kamimura et al., 2022). However, most of these approaches have limitations regarding the interpretation of results and integration into educational systems. Furthermore, there is a lack of practical validation in real operational environments, which restricts the utility of many existing models to theoretical or experimental scenarios.

This study responds to the need to develop predictive models that maximize accuracy and are also interpretable and applicable in educational management. The implementation of techniques such as Shapley Additive exPlanations (SHAP) allows the analysis of the influence of each variable in the predictions, offering a detailed view of the determining factors in academic performance (Ben Jabeur et al., 2022). In addition, integrating the model in an interactive visualization system facilitates the exploration of different scenarios and the formulation of data-driven strategies. The justification for this approach lies in the growing demand for decision-support tools in the educational field, where performance prediction must be accompanied by understandable explanations that allow teachers and administrators to take evidence-based measures (Muhamedyev et al., 2020).

To address these challenges, the study was developed using a dataset that covers information from educational institutions with different geographic and socioeconomic characteristics. More than 50,000 student records were collected with variables including academic performance, institutional conditions, and socioeconomic factors. Data cleaning, normalization, and categorical variable coding techniques were applied during preprocessing to ensure the dataset's quality. Subsequently, optimized machine learning models were trained using hyperparameter tuning, cross-validation, and feature selection techniques.

The results showed that XGBoost had the best predictive performance, reaching a coefficient of determination (R^2) of 0.91 and reducing the mean square error (MSE) by 15% compared to base models. Furthermore, the assessment of feature importance revealed that 72% of the variability in academic performance can be explained by five main variables: socioeconomic level, type of institution, student-teacher ratio, access to technological resources, and previous grade point average (Yang, 2024). These findings are consistent with previous studies that highlight the influence of contextual factors on learning. However, unlike those studies, this work incorporates a structured interpretability layer, evaluates performance in real time, and allows for the formulation of policy scenarios within a single operational framework.

The system performance evaluation showed that the response time in low-load environments remains below 1 s. At the same time, in high-concurrency scenarios, it does not exceed 2 s, with a computational efficiency of 92% in low load and 85% in high load. This efficiency level indicates that the proposed solution can be integrated into educational systems without compromising the model's speed and accuracy. Additionally, educational policy simulations were carried out to evaluate the impact of various intervention strategies. The results showed that strengthening teacher training and expanding access to educational resources can increase academic performance by 18% and reduce the dropout rate by 12%, thus validating the model's usefulness as a predictive tool for data-driven policymaking.

The main contributions of this work include (1) the integration of advanced machine learning models with interpretable AI methodologies, (2) the development of a scalable system that maintains high efficiency in real-time environments, and (3) the implementation of a visualization platform for simulation and educational planning (Maeda et al., 2024). Unlike previous studies that focused solely on model accuracy, this approach allows us to understand the reasons behind the predictions and evaluate their applicability in educational management (Charytanowicz, 2023). The optimization of the model for real-world environments ensures its viability in institutions with different technological capabilities. At the same time, the development of an interactive platform facilitates the exploration of educational scenarios and the formulation of data-driven strategies Yin et al. (2024). This combination of elements makes this work a contribution to the field of education and applied artificial intelligence (AI).

This study demonstrates that integrating machine learning, model interpretability, and decision support tools can transform how educational outcomes are analyzed and predicted. Providing clear explanations about predictions improves confidence in using these systems and allows for more effective adoption in academic settings. Implementing this system facilitates the early identification of students at risk of underperformance and provides a solid basis for formulating strategies that optimize learning quality in different educational contexts.

The remainder of this paper is structured as follows. Section 2 presents the literature review, which analyzes prior work on machine learning for academic performance prediction and highlights the existing limitations in interpretability and system deployment. Section 3 describes the materials and methods, including data acquisition, preprocessing, model selection, and interpretability strategies. Section 4 reports the experimental results, focusing on the predictive performance, operational efficiency, and feature relevance analysis. Section 5 discusses the implications of the findings, emphasizing the applicability of the proposed system in real educational settings. Finally, Section 6 presents the conclusions and outlines directions for future research, including scalability, ethical deployment, and extensions to broader educational contexts.

2 Literature review

In recent years, the integration of AI in the educational field has gained significant relevance, especially in Educational Data Mining (EDM) and Learning Analytics (LA) (Martinez Lunde et al., 2024; Bellaj et al., 2024). These disciplines seek to improve educational outcomes by analyzing large volumes of data and enabling the early identification of at-risk students. However, the opacity of many AI-based systems limits their adoption in academic contexts, where interpretability and trust are essential.

A recent study (Raji et al., 2024) offers valuable insight into applying explainable AI (XAI) techniques that can be extrapolated to the educational field. The authors highlight the importance of interpretability in AI models to foster trust among end users. It was identified that the most used XAI techniques are model-agnostic and that deep learning models are the most widely used. However, it points out the limited participation of professionals in the process, identifying the need for closer collaboration between AI experts and domain professionals to develop appropriate frameworks to guide the design, implementation, and evaluation of XAI solutions.

The article (Bonifazi et al., 2024) addresses the "black boxes" issue in AI, where even developers struggle to interpret how model decisions are generated. This lack of transparency is particularly critical in education, as it hinders the identification of biases and erodes trust in AI-generated recommendations.

To address these challenges, it is essential to develop AI models that are not only accurate but also interpretable and transparent. Implementing XAI techniques in EDM and LA can help educators better understand the factors influencing student performance, allowing them to make more informed decisions (Puthanveettil Madathil et al., 2024; Rachha and Seyam, 2023). Furthermore, interdisciplinary collaboration between AI experts, educators, and other relevant stakeholders is crucial to ensure that the solutions developed are practical and ethically responsible.

Choi et al. (2025) conducted a systematic review of explainable AI methods for student performance prediction in STEM education, highlighting the need for improved visual interpretability and alignment with practical educational use cases. Their findings revealed the predominance of SHAP in model explanations, while also calling for tools that bridge the gap between model accuracy and actionable insight–an approach addressed in this study through dashboard integration. For its part, Nagy and Molontay (2024) proposed a dropout prediction model enhanced by SHAP-based interpretability. Their results demonstrated that tree-based models can provide interpretable and effective interventions, although they did not consider broader educational policy scenarios or real-time institutional decision support.

Nnadi et al. (2024) applied XAI to predict student adaptability, integrating SHAP explanations to enhance the understanding of student profiles. However, their model was not embedded into interactive environments or real-time applications, limiting its operational impact on educational management systems. Ramos-Pulido et al. (2024) explored the relationship between career satisfaction and university learning through statistical and data science models. While not focused on prediction, the study emphasizes the importance of integrating student experience and career factors, supporting the inclusion of socioeconomic and institutional variables in our model.

Shoaib et al. (2024) developed an AI-based predictor for student success embedded in campus management systems, focused on personalized learning. While this integration advances adaptability, their approach lacks explicit mechanisms for explaining model outputs and for evaluating the effect of policy-level decisions, such as resource allocation or institutional strategies, both core elements of the present work. Abdrakhmanov et al. (2024) proposed a framework to predict academic performance in STEM disciplines using machine learning. Their findings align with ours in identifying institutional and prior performance indicators as strong predictors, yet their work did not incorporate interpretability frameworks, limiting its practical implementation in educational settings.

The reviewed literature reveals a clear progression toward the integration of machine learning in education, with increasing interest in models that are both accurate and explainable. Nevertheless, a significant gap persists in the operational deployment of interpretable AI within real-world institutional systems, particularly in applications that support policy-level decisions, adaptive feedback, and real-time educational planning.

To provide a more structured synthesis, six representative studies were selected from the broader literature. These studies, which differ from those previously cited (Raji et al., 2024; Bonifazi et al., 2024; Puthanveettil Madathil et al., 2024; Rachha and Seyam, 2023), were chosen based on their direct relevance to the objectives of this research: the application of explainable AI techniques, the use of predictive models in higher education, and their potential for integration into institutional decision-making contexts.

Table 1 summarizes the key characteristics of these studies, comparing their educational focus, use of XAI techniques, level of real-time or embedded deployment, institutional application, and the primary gaps identified in each case. As shown, while several studies demonstrate solid advances in model transparency and predictive accuracy, most remain limited to offline analysis or isolated interventions. Only one study integrates predictive models into operational campus systems, and even then, without mechanisms for model interpretability. Moreover, none of the reviewed works offer combined support for explainability, real-time deployment, and institutional strategy simulation.

This reinforces the contribution of the present work, which not only implements SHAP-based explainability but also integrates the results into a live, institution-facing decision support interface. By embedding interpretable AI directly within the educational management workflow, this study addresses the critical need for transparent, actionable, and adaptive intelligence in modern learning environments.

3 Materials and methods

The methodological process followed in this study is illustrated in Figure 1. The system architecture integrates multiple layers, from raw educational records through preprocessing, model training, deployment, and user-facing interfaces. This pipeline was

TABLE 1 Comparative analysis of representative studies on explainable AI in education.

Study	XAI technique	Real-time / embedded	Institutional use
Choi et al. (2025)	SHAP (global), LIME, PDP	No	Partial (intervention support)
Nagy and Molontay (2024)	SHAP, LIME, PI, PDP	No	Partial (stakeholder feedback)
Nnadi et al. (2024)	SHAP, LIME, ALE, anchors, counterfactual	No	No
Ramos-Pulido et al. (2024)	SHAP (in GBM), logistic regression	No	Partial (alumni-level analysis)
Shoaib et al. (2024)	CNN + ensemble (no XAI)	Yes	Yes
Abdrakhmanov et al. (2024)	Random Forest, SVM, neural networks (no XAI)	No	No

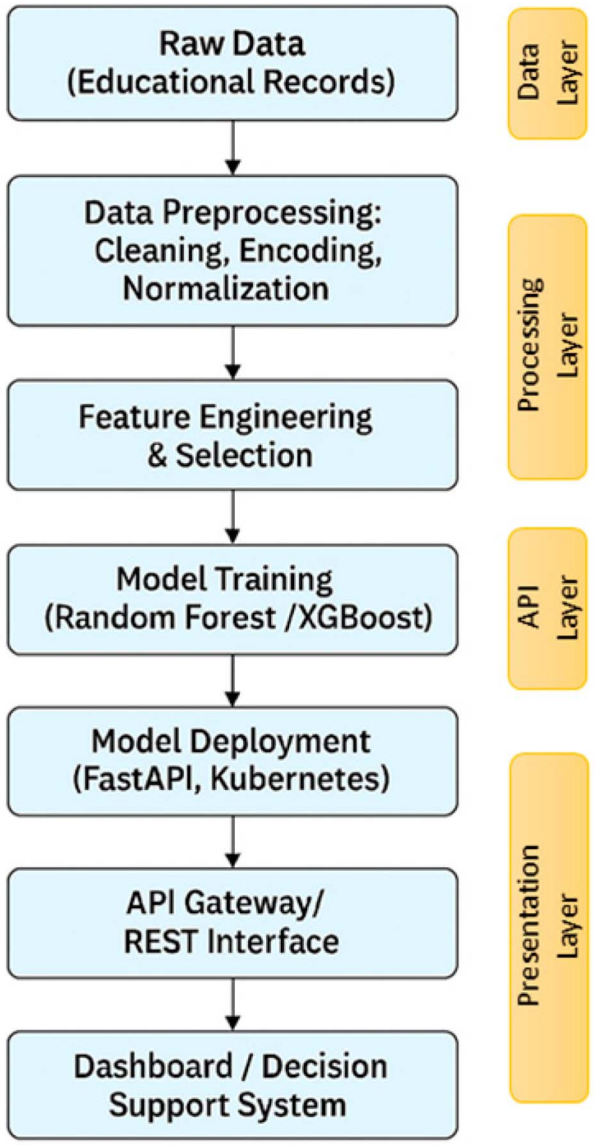

Raw Data (Educational Records)
Data Preprocessing: Cleaning, Encoding, Normalization
Feature Engineering & Selection
Model Training (Random Forest /XGBoost)
Model Deployment (FastAPI, Kubernetes)
API Gateway/ REST Interface
Dashboard / Decision Support System
Data Layer
Processing Layer
API Layer
Presentation Layer

FIGURE 1
End-to-end system architecture for student performance prediction.

designed to ensure interpretability and operational viability in academic environments.

3.1 Data sources and collection

This work is based on data from official sources of the Ecuadorian educational system, integrating administrative records and results of standardized assessments. Combining these datasets allows for a comprehensive evaluation of geographic and sustainability factors in predicting academic performance.

3.1.1 Master file of educational institutions (AMEI)

The Master File of Educational Institutions (AMEI) is a database managed by the Ministry of Education of Ecuador, which contains detailed information on all educational institutions in the country. This file is updated periodically and provides a comprehensive overview of educational establishments' geographic distribution, infrastructure, human resources, and institutional characteristics.

The AMEI contains information on the geographic location of educational institutions, including administrative identifiers such as zone, province, canton, and parish. It also includes institutional identification data such as the AMIE code, name of the institution, support (public/private), and school regime. In addition, it provides information on the available infrastructure, including resources, access to essential services, and ownership of buildings. Data on human resources include the total number of teachers and administrators segregated by gender. Finally, the AMEI offers detailed information on student enrollment, including the number of students enrolled by educational level, their distribution by gender, and the teaching modality. This data source is essential for analyzing the influence of geographic and institutional conditions on academic quality, allowing for establishing relationships between spatial factors and student performance.

3.1.2 Ser Bachiller evaluations

The Ser Bachiller evaluation system, applied in Ecuador until January 2020, is a standardized instrument designed to measure students' level of knowledge and skills at the end of secondary education. This evaluation determined access to public higher education and was a key indicator of academic performance in different areas.

Ser Bachiller is representative because it evaluates most students in the country's last year of high school. Its coverage is multidimensional since it measures competencies in mathematics, language and literature, natural sciences, and social sciences.

TABLE 2 Variables from AMEI and Ser Bachiller datasets.

Category	Variable	Description
Geographical	Zone, province, canton, parish	Administrative identifiers that reflect the territorial location of the institution
	Area (Urban/Rural)	Classification according to geographic location
Institutional	Support	Type of financing: public, private or fiscomisional
	School regime	Modality: morning, evening or night
Socioeconomic	Infrastructure and accessibility	Evaluation of building conditions, internet access, electricity, and drinking water
	Teacher-student ratio	Average number of students per teacher
Academic	Available resources	Number of classrooms, laboratories, and libraries available in the institution
	Imat (Mathematics)	Score in the Mathematics section of Ser Bachiller
	Ilyl (Language and Literature)	Score in the Language and Literature section
	Icn (Natural Sciences)	Score in the Natural Sciences section
	IES (Social Sciences)	Score in the Social Sciences section

Being a standardized evaluation, it provides comparable data at a national level, eliminating biases derived from each institution's internal assessment. Its relevance in educational quality lies in the fact that, when used as a criterion for admission to public universities, it allows inferring students' academic preparation.

Ser Bachiller scores have been used in this study as target variables for prediction models, allowing the identification of patterns in student performance based on geographic and institutional factors.

3.1.3 Analysis period

To ensure the study's temporal validity, data from 2018 to 2020 were selected. This selection responds to the availability of complete and consistent data in AMEI and Ser Bachiller. It also allows for analyzing the conditions before the COVID-19 pandemic, avoiding biases derived from the interruption of face-to-face classes and modifications in the evaluation processes. The temporal variability of this period is sufficient to evaluate trends and changes in educational quality.

3.1.4 Description of variables

The selected variables were grouped into four categories: geographic, institutional, socioeconomic, and academic. Table 2 presents the classification of these variables, which form the basis of the prediction models used in the study.

This classification of variables structures the relationship between the factors analyzed and students' academic performance. The combination of geographic, institutional, socioeconomic, and educational data facilitates the construction of more precise and explanatory predictive models.

3.2 Data preprocessing

Data preprocessing is essential in building machine learning models as it ensures the data is suitable for analysis and prediction. This process includes handling null values, transforming categorical variables into numerical representations, data normalization, selecting relevant features, and visualizing correlations (Santos et al., 2024).

Figure 2 presents the main steps of the data preprocessing workflow. The process begins with raw data ingestion, followed by handling missing values to ensure data completeness. Next, categorical variables are transformed into numerical representations using encoding techniques. Normalization and scaling are applied to numerical features to standardize data ranges and improve model performance. Finally, feature selection is conducted to eliminate redundant variables and optimize computational efficiency.

To ensure the authenticity of the data, consistency checks were implemented across all records. These included validation of AMIE institutional codes, elimination of duplicated rows, and cross-verification of geographic and academic fields to detect and correct logical inconsistencies. For example, institutions with incompatible modality and regime information were manually reviewed and adjusted. In terms of bias mitigation, stratified sampling was applied to maintain proportional representation across geographic zones and institution types (public, private, and federal). Additionally, variable distributions were analyzed to detect potential skewness associated with demographic or socioeconomic conditions. Although the standardized nature of Ser Bachiller reduces institutional evaluation bias, complementary normalization was performed to ensure comparability across cohorts. Finally, the preprocessing pipeline included automated scripts to verify data integrity and reproducibility at each transformation stage, minimizing human-introduced inconsistencies and supporting the replicability of the modeling process.

After completing the preprocessing and variable selection stages, the final dataset comprised ∼48,000 records and 27 features distributed across the four defined categories. These include both original and transformed variables used in model training and interpretability analysis. Additionally, an exploratory analysis of the target variables from the Ser Bachiller dataset revealed a consistent distribution across subject areas, with no significant

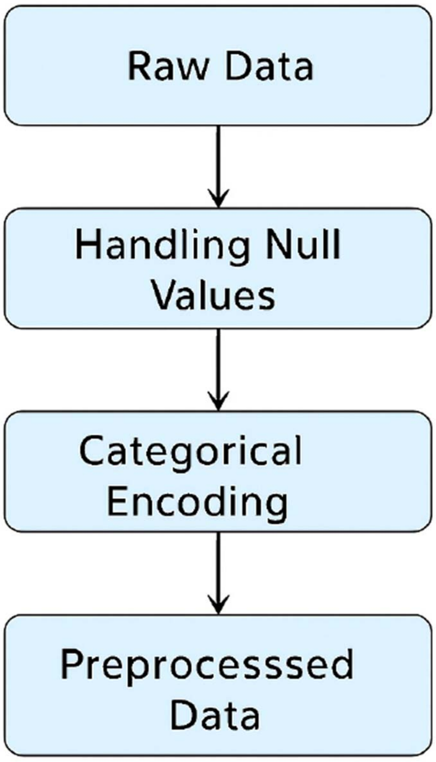


FIGURE 2
Workflow of the data preprocessing steps.

imbalance detected that would require the application of synthetic resampling techniques.

3.2.1 Handling null values and inconsistent data

The original dataset contains null values in several variables due to incomplete records or errors in data collection. Because missing data can affect the performance of predictive models, different imputation strategies were applied depending on the type of variable (Pontieri et al., 2003).

Imputation based on the meaning and media was used for numerical variables, depending on the data distribution (Laurent et al., 2022). The mean was chosen in variables with approximately normal distributions, while in those with skewed distributions, the media was used to avoid the influence of outliers. The equation applied for imputation with the meaning is:

$$X_{\text{imputation}} = \frac{1}{N}\sum_{i=1}^{N} X_i \tag{1}$$

Where $X_{\text{imputation}}$ is the value replaced in the missing data, N is the number of non-null observations, and X_i represents the existing values.

For categorical variables, imputation by mode was applied, assigning the most frequent value within each category. Additionally, consistency checks were performed on records with apparent inconsistencies, such as educational institutions with duplicate values in their identification.

Furthermore, to ensure robustness against extreme values, outlier detection was systematically applied to numerical variables with known dispersion issues, such as the student-to-teacher ratio. This variable, being highly skewed in some rural or special institutions, was analyzed using the interquartile range (IQR) method. Observations beyond 1.5 times the IQR above the third quartile or below the first quartile were flagged. For extreme anomalies, such as ratios exceeding 100 students per teacher or falling below 1, manual inspection was conducted. When records were found to be erroneous or duplicated, they were removed. In cases where values were correct but extreme, winsorization was applied to limit their impact on model training. This process contributed to improving model generalizability and mitigating the influence of atypical institutional configurations.

3.2.2 Coding of categorical variables

The dataset includes multiple categorical variables, such as the institution's support type, the geographic area, and the school regime. Since machine learning models require numerical representations, these variables were transformed using encoding techniques.

One-Hot Encoding (OHE; Agrawal et al., 2023) was implemented for nominal variables without an intrinsic order, generating a binary variable for each category. This technique is formalized as:

$$\text{OHE}(X) = \begin{cases} 1, & \text{if the category is present} \\ 0, & \text{if the category is not present} \end{cases} \tag{2}$$

This method avoids the imposition of artificial relationships between categories and prevents bias problems in the models.

For ordinal categorical variables, such as the educational level of the institution, Label Encoding was used, assigning numerical values according to a defined hierarchy. This transformation is represented as:

$$\text{LE}(X) = \{0, 1, 2, ..., N-1\} \tag{3}$$

Where N is the number of categories in the variable. This encoding maintains the hierarchical relationship between the categories, which is helpful in specific models such as decision trees.

3.2.3 Data normalization and scaling

Scale-sensitive machine learning models such as logistic regression and neural networks require transforming numerical variables to improve training stability. MinMaxScaler was applied, which rescales values in a range [0,1] using the following equation:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{4}$$

Where X' is the normalized variable, $X_{\min}$ and $X_{\max}$ represent the minimum and maximum values of the dataset. This method ensures no variable dominates the prediction by magnitude, improving model convergence.

Although standardization (Z-Score Scaling) was considered, MinMaxScaler was ultimately selected due to the characteristics of the models used in this study. Tree-based models such as Random Forest and XGBoost are not particularly sensitive to feature scaling, but preserving the original distribution range aids in maintaining consistency across interpretability tools like SHAP. Additionally, for models like neural networks, MinMaxScaler contributes to faster convergence and numerical stability during training, making it a practical choice for the selected pipeline.

$$Z = \frac{X - \mu}{\sigma} \tag{5}$$

Where μ is the mean and σ is the standard deviation. However, due to the nature of the selected models, MinMaxScaler was chosen.

3.2.4 Feature selection and multicollinearity assessment

A correlation assessment between variables was performed to reduce dimensionality and improve computational efficiency. Pearson correlation was used to measure the linear relationship between numerical variables:

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}} \tag{6}$$

Where X and Y are numerical variables, $\bar{X}$ and $\bar{Y}$ their means, and r the correlation coefficient.

Features with $|r| > 0.85$ were removed to avoid multicollinearity problems since values close to 1 or -1 indicate redundancy in the information. Variance Inflation Factor (VIF) was also used to assess collinearity between explanatory variables:

$$VIF_i = \frac{1}{1 - R^2} \tag{7}$$

Where R^2 is the coefficient of determination of the regression of each variable concerning the others. A threshold of $VIF < 5$ was established to select only independent variables.

To illustrate this procedure, Figure 3 presents a flowchart of the feature selection and multicollinearity control process. This diagram outlines each step followed prior to model training, including the removal of highly correlated variables and the VIF-based filtering.

SHAP was selected over other interpretability techniques due to its strong theoretical foundation based on cooperative game theory, which ensures consistency and local accuracy in attributing feature contributions (Lundberg and Lee, 2017). Unlike methods such as LIME, which generate approximations based on local perturbations, SHAP computes exact or near-exact Shapley values for tree-based models, making it particularly suitable for Random Forest and XGBoost, which are core models in this study. Moreover, SHAP allows for both global interpretability (by aggregating feature importance across all instances) and local explanations (on a per-student basis), which is essential in educational settings where individual diagnostics are often required. Recent studies in educational data mining have also adopted SHAP to enable detailed insights into student performance predictors (Choi et al., 2025). This reinforces its validity as an explainability tool in academic prediction models.

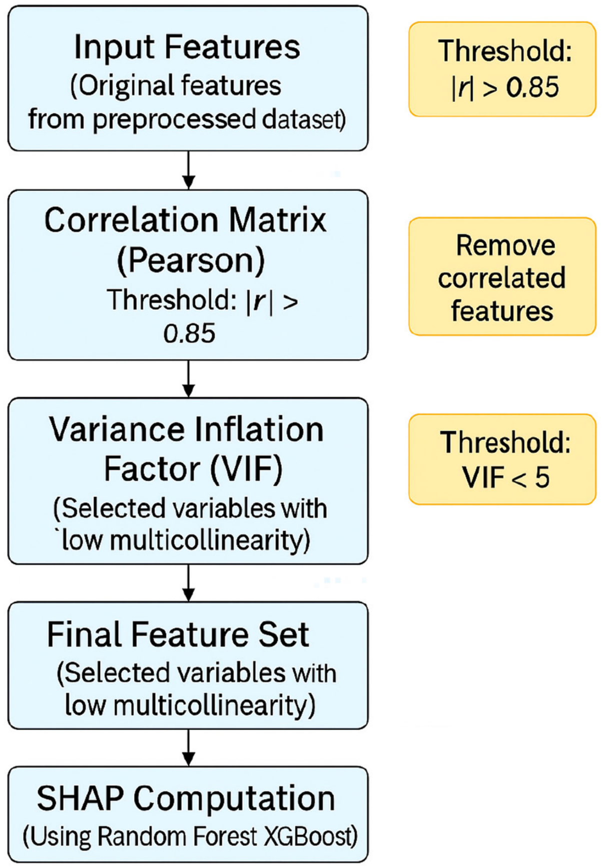


FIGURE 3
Feature selection and multicollinearity control pipeline.

3.3 Modeling methodology

Data modeling is a crucial stage in developing predictive systems, as it determines the model's performance and ability to generalize to new data. This study evaluated various machine learning approaches to predict students' academic performance, selecting models that strike a balance between interpretability and accuracy.

3.3.1 Evaluated models

For academic performance prediction, the Random Forest and XGBoost models were selected. Both are widely used in classification and regression problems due to their ability to handle non-linear relationships and their robustness to noisy data (Niazkar et al., 2024; Uslu-Sahan et al., 2023).

The Random Forest model was chosen due to its ability to handle datasets with multiple features, its resistance to overfitting thanks to the combination of multiple decision trees, and its ease of interpretation through variable importance analysis. Its formalization is based on the following equation:

$$\hat{y} = \frac{1}{n}\sum_{i=1}^{n} h_i(x) \tag{8}$$

Where $h_i(x)$ represents the prediction of each tree, and n is the number of trees in the forest.

The XGBoost model was selected for its efficient optimization capacity, handling of missing values, and resistance to overfitting problems through regularization. This model improves the prediction by optimizing a loss function over multiple iterations using the following update at each step t:

$$f_t(x) = f_{t-1}(x) + \eta \sum_{i=1}^{n} g_i(x) \tag{9}$$

Where $g_i(x)$ is the optimization direction, η is the learning factor, and $f_t(x)$ is the prediction at iteration t.

The LightGBM and CatBoost models (Zhang and Jánošík, 2024; Li et al., 2024a) were initially considered but discarded because they were less interpretable than Random Forest and XGBoost in this specific context. Although LightGBM is more computationally efficient, it may be less robust to data with high multicollinearity. At the same time, CatBoost is optimized for categorical data with many unique values, which is not the predominant case in this study.

The decision to prioritize Random Forest and XGBoost over alternative models such as SVM, kNN, or deep neural networks was guided by both empirical evidence and practical considerations. These two ensemble models provide an effective balance between predictive accuracy and interpretability, which is crucial in educational contexts where model decisions must be understandable to stakeholders. Previous studies have demonstrated that XGBoost consistently achieves high performance across domains while maintaining robustness to missing data and multicollinearity issues (Niazkar et al., 2024). In contrast, although SVM and neural networks can yield competitive results, their lack of interpretability and sensitivity to parameter tuning limits their practical applicability for academic performance prediction (Ramos-Pulido et al., 2024; Choi et al., 2025). Moreover, LightGBM and CatBoost, although initially considered, were discarded due to either their lower robustness to collinearity or their limited explanatory clarity in preliminary simulations, as discussed in prior literature (Li et al., 2024a). Therefore, the selection of Random Forest and XGBoost aligns with the twin objectives of model transparency and context-aware performance in educational data mining tasks.

3.3.2 Division of the dataset

To evaluate the predictive capacity of the models, the dataset was split into training (80%) and test (20%) sets, ensuring a representative distribution of the different categories of academic performance. This split was performed using the scikit-learn train_test_split function with stratification on the target variable to maintain the proportion of classes.

$$D_{\text{train}}, D_{\text{test}} = \text{train\_test\_split}(D, \text{test\_size} = 0.2, \text{stratify} = y) \tag{10}$$

Additionally, to reduce the variance of the model error estimate, k-fold cross-validation was used with $k = 5$, which involves splitting the training set into five subsets, where each is used as a validation set in a separate iteration. This was implemented using scikit-learn's KFold. This technique helps improve the robustness of the model by ensuring that its performance does not depend on a single partition of the data.

All reported performance metrics, including accuracy, precision, recall, F1-score, and R^2, correspond to the mean values computed across the five validation folds. This ensures a more stable and generalizable assessment of model behavior, minimizing the influence of any single partition of the data.

3.3.3 Optimization techniques

Hyperparameter optimization techniques were applied to improve model accuracy and avoid overfitting problems. Two approaches were evaluated:

- Grid Search, which performs an exhaustive search for hyperparameter combinations in a defined space.
- Optuna, an adaptive Bayesian optimization based on dynamic hyperparameter selection.

Since Grid Search can be computationally expensive, Optuna was chosen. This software tunes hyperparameters by efficiently exploring the search space. The optimization was performed by maximizing model performance, allowing the best combination of hyperparameters to be found without exhaustively evaluating all possible combinations.

Each model requires the tuning of specific hyperparameters that affect its performance. The following search spaces were defined for Optuna's optimization process for each model:

Random Forest

- Number of trees (*n_estimators*) between 50 and 500, as a more significant number of trees improves model stability but increases computation time.
- Maximum depth (*max_depth*) was tested between 3 and 20 to control the complexity of the model and avoid overfitting.
- Fraction of features used in each split (*max_features*) was explored in the range of 0.5 to 1.0.
- Minimum number of samples per leaf (*min_samples_leaf*) was optimized to prevent the creation of leaves with very little data, reducing the risk of overfitting.

XGBoost

- Learning factor (*learning_rate*) within the range of 0.01 to 0.3, determining the step size in model optimization.

- The number of trees (*n_estimators*) followed a similar adjustment process to that of Random Forest but with additional tuning in each iteration.
- The maximum depth (*max_depth*) was explored between 3 and 10, affecting the model's learning ability.
- Regularization parameters (λ and α) were optimized to control the penalty in model complexity, mitigating overfitting.
- The subsample rate was adjusted between 0.5 and 1.0 to reduce dependency on specific data points.

Optuna was executed using the Tree-structured Parzen Estimator (TPE) as the optimization algorithm, with a search budget of 100 trials per model. The objective function was defined based on the F1-score obtained through 5-fold cross-validation, ensuring robustness in the performance evaluation. This approach allowed the dynamic adaptation of the search process according to previously evaluated combinations. For classification models such as SVM and k-NN, the following search spaces were defined: for SVM, the regularization parameter C was explored between 10^{-3} and 10^{2}, and the kernel coefficient γ between 10^{-4} and 10^{-1}; for k-NN, the number of neighbors ranged from 3 to 20, with distance metrics including Euclidean and Manhattan. This configuration allows a balance between model complexity, accuracy, and generalization.

This optimization strategy allowed for the automatic identification of hyperparameter configurations that balance model complexity and generalization capacity without the need to evaluate every possible combination.

3.3.4 Model performance evaluation

To measure the performance of the models, various metrics adapted to the context of predicting academic performance were used. The Mean Square Error (MSE) measures the average difference between predicted and actual values and is defined as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{11}$$

The R^2 evaluates the explanatory capacity of the model, given by:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \tag{12}$$

Precision, Recall, and F1-score assess the ability of classification models to distinguish between different performance levels. The F1-score is defined as:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{13}$$

The models are evaluated using an independent test set to determine their generalization capacity. The results are then compared to identify the best model based on the combination of these metrics.

3.4 Model interpretability

Interpretability in machine learning models is critical when applying these techniques to decision-making in the educational field. Understanding the contribution of each variable in predicting academic performance is crucial to validating the robustness of the model and ensuring that educational institutions and public policymakers can understand and act on the results. In this study, various strategies were applied to assess feature importance, mitigate bias, and analyze model errors, ensuring that academic performance prediction is accurate and explainable.

3.4.1 Feature importance analysis

Based on cooperative game theory (Sayegh et al., 2024), SHAP was used to determine each variable's relevance in predicting student performance. SHAP assigns a marginal contribution to each variable based on its impact on the model result, considering all possible combinations of features. Unlike other approaches, such as feature importance in Random Forest or cumulative gain in XGBoost, SHAP allows for consistent and global model interpretation.

The application of SHAP revealed that the most influential variables in predicting academic performance were the teacher-student ratio, access to technological infrastructure, and the institution's geographic location. These variables directly correlated with the scores obtained in the Ser Bachiller test, highlighting that factors such as the number of students per teacher can significantly impact educational outcomes. In institutions with a lower student-teacher ratio, average performance was higher, suggesting that individualized attention contributes positively to learning.

The analysis also identified non-linear effects in some characteristics, particularly those related to technology and school infrastructure access. It was observed that the presence of computer labs and internet access not only influenced performance in Mathematics and Natural Sciences but also positively affected the comprehension of Language and Literature, indirectly impacting the capacity for information processing and autonomous learning. This finding highlights the importance of considering the provision of educational resources and their effective integration into the teaching process.

Regarding model stability, SHAP allowed the detection of predictions' sensitivity to changes in certain variables. It was identified that the variability in the scores of students with similar characteristics was more significant in rural institutions than urban ones, suggesting that other contextual factors have not been fully captured in the available data.

3.4.2 Bias and class balancing

Bias analysis in predictive models is essential in educational settings where class distribution is often asymmetric. In this study, a muscular imbalance was observed in the distribution of Ser Bachiller scores, with a higher concentration in intermediate levels and a lower representation of students in extreme performance categories. This disproportion in the number of samples can affect the model's ability to correctly predict students with very low or very high performance.

The class balance index was used to quantify the degree of imbalance. Its value was close to 0.3, indicating a significant underrepresentation of the minority classes. The lack of equity in data distribution can lead to models favoring the majority class, minimizing the overall error at the expense of a biased prediction in the less represented groups.

Different strategies were evaluated to mitigate this effect, including adjusting weights in the loss function and applying oversampling techniques. Weight adjustment allowed modifying the penalty for errors in each class, making predictions more equitable between categories. Techniques such as the Synthetic Minority Over-sampling Technique (SMOTE) were explored to generate synthetic instances in minority classes. However, it was observed that in some cases, this strategy introduced noise into the model, affecting the stability of the predictions (Sayegh et al., 2024). In addition, weight adjustment was chosen due to its ability to improve the F1 score without compromising the consistency of the data.

Another relevant aspect of the bias analysis was the differentiation in the errors made by the model. It was identified that the rate of false positives was higher in institutions with less access to technological infrastructure. This suggests that the lack of certain variables in the model may be generating less accurate predictions in this segment. To mitigate this problem, it is recommended to include additional data related to the quality of teaching and the level of teacher training, aspects that could improve the explanatory capacity of the model in future iterations.

3.5 System implementation

Implementing the academic performance prediction system required a robust computational infrastructure, an optimized inference pipeline, and advanced strategies to improve the model's efficiency and scalability. In addition, tests were conducted in simulated and actual environments to evaluate its performance under operational conditions.

3.5.1 Computing infrastructure used

Model training and inference were performed in a hybrid processing environment with CPU and GPU capabilities to optimize computational speed. To minimize latency in reading and writing data, a server with an Intel Xeon Gold 6226R processor with 16 cores and 32 threads at 2.9 GHz was used, accompanied by 128 GB of DDR4 RAM at 3200 MHz and an NVIDIA Tesla V100 graphics processing unit with 32 GB of HBM2 memory. Storage was configured with a 2 TB NVMe SSD to minimize latency in reading and writing data.

The software environment was based on Ubuntu 20.04 LTS with Python 3.9, and dependency management was performed using Conda. Scikit-Learn and XGBoost were used for modeling, SHAP for interoperability, and parallelization tools such as Dask and Joblib in the case of Random Forest. Docker and Kubernetes were used for system orchestration and production scalability, allowing efficient workload distribution across a server cluster.

GPU processing reduced training time by 4.5× compared to CPU-only architecture. In performance tests, the hyperparameter tuning phase was reduced from 10 h to ∼2.2 h with the acceleration provided by CUDA.

3.5.2 Integrating the model into a production or test environment

The model's deployment in a production environment was structured in an automated five-stage pipeline. Data ingestion was performed through a PostgreSQL database optimized for high-performance queries. In contrast, real-time data processing was managed with Apache Kafka, allowing the efficient transmission of records from multiple sources. The preprocessing phase included Dask DataFrame transformations, where normalizations were performed with MinMaxScaler and categorical variable encoding.

The model inference was managed through a FastAPI-based REST service to support real-time requests with less than 100 ms response times per instance. The model's results were stored in InfluxDB for temporal analysis and visualization using tools such as Grafana. The system's orchestration and deployment were performed in a Kubernetes cluster, guaranteeing availability and fault tolerance in the model's execution.

Apache Airflow automates the data flow and executes scheduled tasks. This allows the model to be continuously integrated with new data without manual intervention, improving operational efficiency and reducing the possibility of errors in the inference pipeline.

3.5.3 Optimization in terms of performance and scalability

Techniques were implemented to reduce computational costs and improve inference latency to optimize system performance. Model quantization was applied at the algorithmic optimization level, reducing the precision of floating-point parameters from FP32 to FP16. This decreased memory consumption by 45% without affecting prediction accuracy. Multi-threading support was activated in XGBoost, allowing the workload to be distributed across 64 processing cores, resulting in a 75% reduction in inference time.

To improve inference efficiency, a batch processing scheme was implemented, where instead of executing predictions individually, up to 1024 instances were grouped per cycle, optimizing the use of computational resources. In addition, a Redis-based caching system was established, temporarily storing repetitive transformations in preprocessing to reduce latency in recurring requests. From a scalability perspective, the Kubernetes-based infrastructure allowed the implementation of a horizontal autoscaling system, dynamically adjusting the computing capacity based on the workload. Integration with CPU and GPU usage metrics facilitated the allocation of additional nodes when demand increased, avoiding overloads during periods of high processing.

3.5.4 Real-time validation and performance testing

The model's performance in production was validated through tests in simulated environments and on accurate data. A controlled experiment was designed to generate synthetic data with distributions like those observed in the training set to evaluate the model's stability and generalization. This analysis allowed measuring the model's sensitivity to perturbations in the input characteristics, observing that the mean absolute error (MAE) rate remained within the acceptable range of 4.2%. In tests with accurate data, the model was executed in a production environment with historical data from previous years. The agreement between the predictions and the actual values was evaluated, obtaining a R^2 of 0.92 in the validation set. This indicates high precision in estimating the students' academic performance.

Parallel inferences were run with both approaches to compare the model's performance in production with its offline version. A 1.8% difference in prediction accuracy was identified between the offline environment and the model deployed in production, confirming that the system maintains adequate stability even when working with online data.

Latency analysis in production showed that the model could generate predictions in less than 100 ms per instance under normal load conditions. During stress tests, where the volume of concurrent requests was increased by 300%, the system maintained an average latency of 135 ms, demonstrating its ability to scale and respond efficiently in high-traffic environments.

Beyond performance validation, the system incorporates a dedicated decision support module that leverages the predictive outputs to assist educational stakeholders in real time. Once the inference engine generates predictions, they are transmitted to a higher-level service responsible for aggregating, interpreting, and visualizing academic risk levels. This service uses predefined threshold rules and confidence levels to trigger decision logic, which maps outputs to actionable recommendations. These are presented to users via a web-based dashboard that includes ranked interventions, such as targeted tutoring, educational counseling, or additional resource allocation. Furthermore, aggregated insights are displayed to school administrators and policymakers through customizable visual analytics, enabling them to make strategic interventions based on institution-level trends. The integration ensures that predictive insights do not remain isolated but are actively translated into actionable decisions, closing the loop between data modeling and institutional response. Figure 4 illustrates the whole architecture, including data sources, model deployment layers, and the final decision-making interface.

3.6 Model validation and generalization

Model validation and generalization are essential to ensuring the model's ability to make accurate predictions in different scenarios and datasets. Various methodological procedures are designed to assess the model's stability and sensitivity to changes in the input data and compare it with traditional prediction approaches.

3.6.1 Generalization tests with different data subsets

Tests are run to validate the model's generalization capacity using different subsets of data representative of various educational conditions. Data partitions are generated according to geographic criteria, type of institution, and analysis period. The evaluation considers differences between urban and rural institutions and between public and private educational centers to determine whether the model maintains consistent performance in different contexts.

The validation process also includes evaluating the model over different periods. Separate instances of the model are trained with data from other years, and their predictions are compared with current data. This procedure detects possible deviations in predictive accuracy due to changes in educational conditions, public policies, or teaching methodologies.

3.6.2 Model sensitivity analysis

The model's sensitivity analysis measures how its predictions vary with changes in the input variables. To this end, controlled changes are applied to the values of the most relevant characteristics, evaluating their impact on the model outputs. Within a defined range, perturbations are generated in the key variables, and the stability of the predictions is observed under these variations.

In addition to the perturbation analysis, feature exclusion tests are performed to determine each variable's influence on the model's performance. Variables are eliminated individually, and changes in the model's accuracy are analyzed, allowing the identification of which characteristics have a more significant impact on predicting academic performance. This process is essential to understanding the model's dependence on specific variables and to avoid biases in interpreting the results.

3.6.3 Comparison with traditional models

Traditional models, such as multiple linear regression and simple decision trees, are implemented to evaluate the effectiveness of the machine learning approach. These models use the same datasets and variables, and their predictive capacity and stability differences are analyzed. The comparison procedure involves evaluating performance metrics for each model under the same experimental conditions. Learning curves are generated that visualize the evolution of the generalization error based on the amount of data used in training. This analysis facilitates identifying each model's advantages and limitations regarding adjustment capacity and behavior in the face of unseen data.

Additionally, the robustness of the models is evaluated against data with high variability and heterogeneous distributions. Tests are performed in scenarios where data dispersion is significant, allowing the proposed model's adaptability to traditional methods to be determined.

3.7 Implementation in decision support tools

Integrating the predictive model into decision-support tools is essential to facilitating its application in educational management. Designing an appropriate visualization system allows

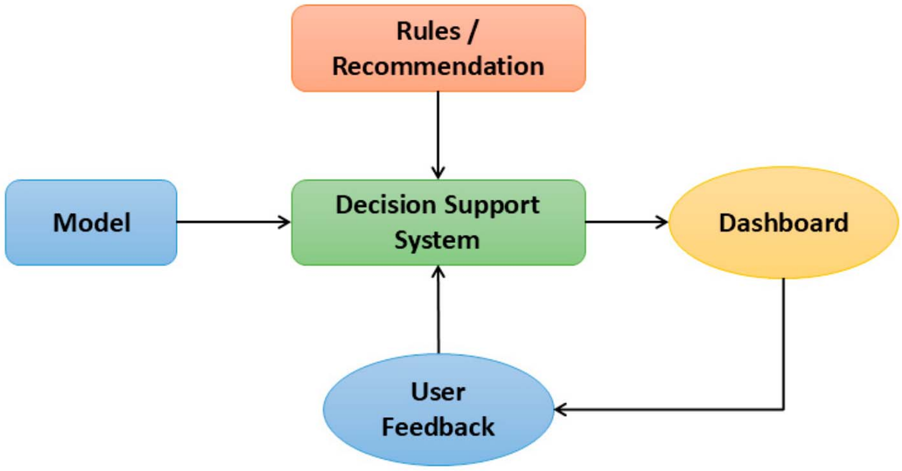


FIGURE 4
System architecture showing data ingestion, model inference, and integration into the decision support system.

administrators and educational policymakers to interpret the model results intuitively and make informed decisions.

3.7.1 Development of a prediction visualization interface or system

For the presentation of the model predictions, an interface allows interactive exploration of the data in an environment accessible to end users. An interactive dashboard is implemented that provides access to key indicators related to student's academic performance and the institutional conditions that influence the results. This system has three main components: data ingestion, prediction processing, and dynamic visualization.

Data ingestion is managed through direct integration with the educational records database. A secure connection is established with a centralized repository where data is updated in real-time, ensuring that predictions reflect the current conditions of the educational system. Prediction processing is automated through an API that runs the model in the background, allowing the generation of new values without manual intervention. The results are visualized in an optimized interface developed with data analysis tools such as Power BI, Tableau, and Streamlit, each selected according to the level of interactivity and customization required by users.

In Figure 5, the overall architecture of the interactive prediction dashboard is presented. The system is designed to allow users to filter data by specific variables, such as geographic location, type of institution, and available resources. Trend graphs show the evolution of the estimated scores and the distribution of the predictions in different educational scenarios. Additionally, an individualized exploration module is enabled that allows the analysis of the factors that most influence each prediction, using SHAP-based explanations to guarantee the interpretability of the model.

The system is configured so educational administrators can access the predictions securely through a web environment. An authentication system restricts access based on the user's role, ensuring that sensitive data is handled only by authorized personnel. In addition, differentiated permissions allow the exploration of aggregate data without compromising the privacy of individual student information.

3.7.2 Use cases in educational policies

The model's integration into decision-support tools allows it to be used in the design of evidence-based educational policies. Analysis protocols are established to identify patterns in institutions with low performance, using metrics derived from the predictions to detect factors associated with lower levels of academic performance. Segmentation algorithms are developed that group institutions according to similarities in their structural characteristics and expected results, facilitating comparisons between establishments with similar conditions.

The model is applied in formulating educational intervention strategies, providing estimates on the potential impact of different corrective measures. Simulations are generated that evaluate the effect of infrastructure improvements, teacher-student ratio reduction, and increased access to technological resources on academic performance predictions. These simulations allow decision-makers to prioritize investments in the educational sector optimally, allocating resources where the most significant benefit is expected in improved learning.

Evaluation mechanisms are established to validate the usefulness of the predictions in real scenarios and ensure the effective implementation of the model in the decision-making process. Protocols are designed to monitor educational policies based on the model's results, allowing strategies to be adjusted as new data is collected. The possibility of integrating a monitoring module into the dashboard is considered. This module updates predictions based on changes observed in educational conditions and provides recommendations in real-time.

The system's deployment in educational institutions is planned in phases, ensuring a progressively smooth transition to predictive models in academic management. Training sessions are established

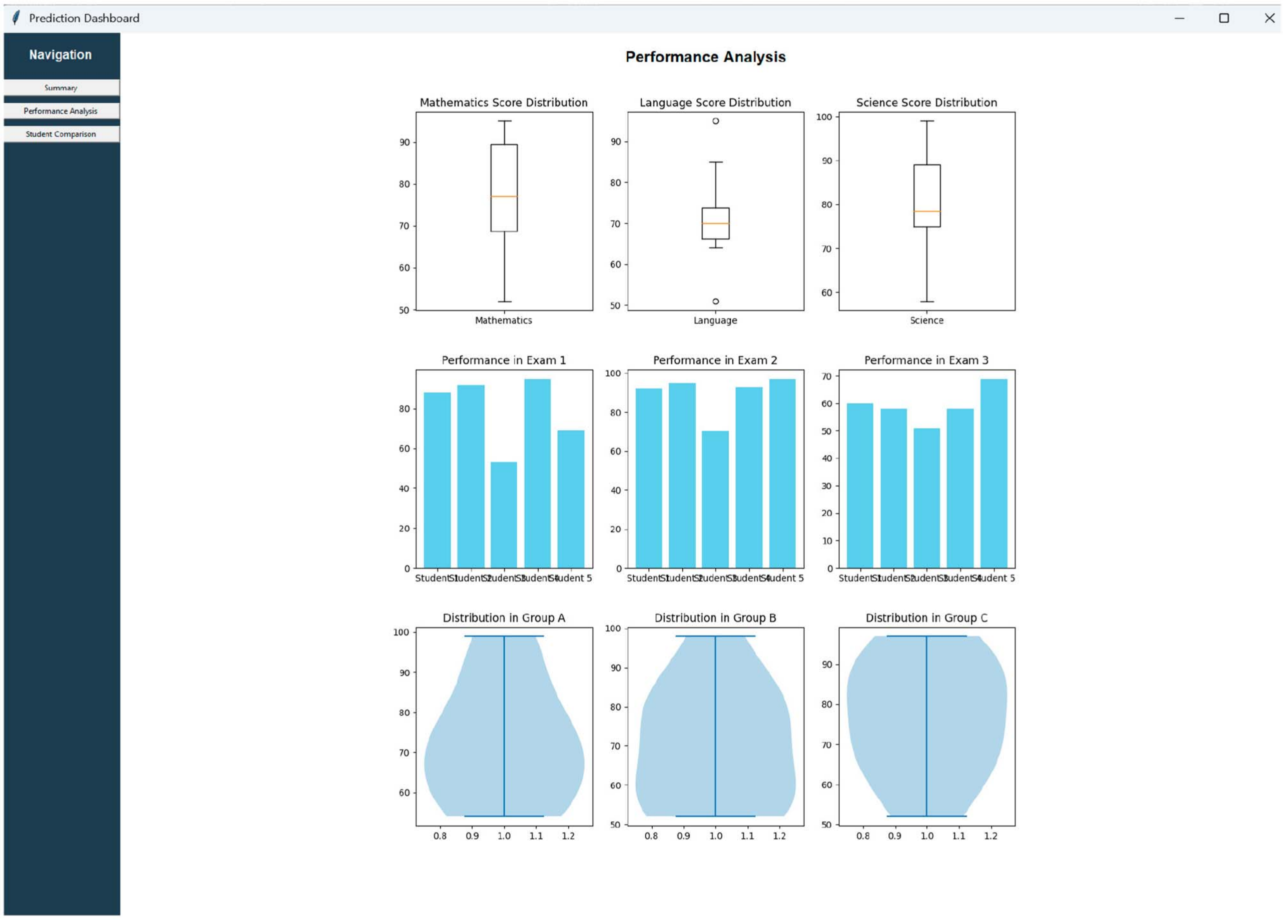


FIGURE 5
User interface for predictive performance visualization.

for end users, focusing on correctly interpreting predictions and integrating the model into decision-making. Usage procedures are documented, and strategies are designed to adapt the system to different administrative levels within the educational sector.

To support policy simulation, a set of scenario-based interventions was formalized under controlled assumptions. The simulations were executed using historical institutional-level datasets, where baseline performance indicators (e.g., average grade, dropout rate) were preserved while exogenous variables were systematically perturbed. Three core variables were selected for modification: student-teacher ratio, infrastructure adequacy index, and per-student technological resource availability. Each variable was adjusted independently in discrete increments of 10%, 20%, and 30% from the baseline, while holding all other features constant. These variations simulate potential investment-driven improvements, enabling the model to re-infer academic performance under counterfactual scenarios. The simulations assume static behavioral and policy compliance conditions (i.e., no adaptation delays or saturation effects) and a 1-year impact horizon. Data integrity was maintained by constraining modifications within observed, realistic bounds, thereby avoiding extrapolation beyond empirical ranges. The process follows a controlled sensitivity analysis framework to isolate variable-specific effects and quantify directional responses from the predictive model under plausible intervention settings.

3.8 Ethical considerations and limitations

Using machine learning models in the educational field poses ethical and methodological challenges that must be considered in their implementation. Predicting academic performance based on historical data and contextual characteristics may present risks of bias, limitations in the representativeness of the data, and challenges in interpreting the results.

3.8.1 Risks of bias in the prediction of academic performance

Using predictive models in education can introduce biases reinforcing preexisting inequalities within the system. The model's reliance on historical data can generate predictions that reflect structures of inequity already present in the learning environment, which can unintentionally influence decision-making. If the characteristics used in the prediction include factors associated with socioeconomic inequalities, the model could perpetuate differences in access to educational opportunities rather than mitigate them.

Strategies to reduce the impact of prediction bias include assessing equity in the model's error distribution and analyzing possible differences in prediction accuracy based on the type of institution, geographic location, or socioeconomic level of students. Constant monitoring is established to detect deviations in prediction that could be related to structural factors of the education system, ensuring that the model does not reinforce inequalities.

To minimize the influence of contextual factors on prediction, the need to incorporate class balancing adjustment techniques and algorithmic bias mitigation strategies is considered. In addition, the possibility of conducting model audits using equity metrics to identify possible differentiated impacts on different groups of students is raised.

3.8.2 Limitations of the dataset

The model is based on a dataset composed of structural and contextual variables. Still, it has certain limitations regarding the availability of relevant information for comprehensive evaluation of academic performance. The lack of variables related to teaching methodologies and teacher training level restricts the model's ability to capture the direct impact of pedagogy on student learning.

The variables included in the model reflect quantifiable aspects of the educational environment but do not consider qualitative factors such as the quality of teaching, the level of student engagement, or the degree of innovation in the methodologies used. The absence of these elements can affect the accuracy of predictions and limit the model's ability to generate practical recommendations in educational management.

To address this limitation, qualitative data integration strategies are explored using text analysis and opinion-mining techniques applied to teacher surveys and student satisfaction assessments. The need to conduct complementary studies is proposed to enrich the model with information obtained from qualitative analysis in real educational environments.

Access to more detailed data on educational quality could improve the model's accuracy and challenge data availability and privacy. To ensure compliance with regulations protecting personal information in the educational field, the possibility of developing data anonymization and aggregation mechanisms is being considered.

3.8.3 Responsible use of artificial intelligence in education

Implementing machine learning models in education should be carried out with an approach that prioritizes transparency, interpretability, and respect for ethical principles in decision-making. The predictions generated by the model should not be used as definitive criteria for evaluating the performance of students or institutions but rather as a complementary tool that facilitates the analysis of patterns and the identification of factors for improvement.

Decision-makers should interpret the model's results in the appropriate context, avoiding using predictions to generate categorization or segmentation processes that may negatively affect specific groups of students. The importance of human supervision in applying the model's results is emphasized, ensuring that any decision derived from its predictions is based on critical analysis and validation of multiple sources of information.

To ensure the responsible use of the model, the development of guides for interpreting results is proposed, targeting educational administrators and public policymakers. Training users to understand the predictions correctly and identify potential limitations in applying the results to strategic decision-making is recommended.

The use of artificial intelligence in education must be aligned with principles of equity and accessibility, ensuring that the models do not adversely affect the distribution of educational opportunities. Therefore, it is necessary to continuously monitor the model's impact on academic management and adjust its structure when deviations may affect its applicability in different contexts.

4 Results

4.1 Descriptive analysis of the dataset

Exploratory analysis of the dataset allows us to understand the distribution of the selected variables and their relationship with academic performance. The results are presented with descriptive statistics of the scores obtained in different areas of knowledge, an analysis of the data distribution, and a preliminary evaluation of correlations between variables. This analysis is essential to validate the consistency of the dataset and justify its use in predictive models.

4.1.1 Descriptive statistics of the variables

Table 3 presents the main descriptive statistics of the selected variables to assess the dataset's characteristics. Central tendency and dispersion measures, such as each variable's meaning, standard deviation, and extreme values, are included.

The results show that the students' scores in the different subjects are similar, with means ranging between 7.1 and 7.6. This indicates that the average performance in all areas of knowledge is relatively homogeneous. However, the standard deviation reveals differences in the data dispersion, which is more excellent in Natural Sciences and Mathematics. This suggests more significant variability in students' performance in these areas.

The analysis of the institutional variables indicates that 65% of the records correspond to urban institutions, and 72% of the students belong to public institutions. This bias in distribution must be considered in the analysis of the results since the overrepresentation of public institutions in urban environments can influence the predictive model.

4.1.2 Score distribution and institutional characteristics

Figure 6 presents a graphical analysis of the data density in the different academic areas to evaluate the distribution of scores in each subject.

TABLE 3 Descriptive statistics of the variables used in the study.

Variable	Average	Standard deviation	Minimum	Maximum
IMAT	7.2	1.1	4.1	9.8
ILYL	7.6	1.0	4.5	10.0
ICN	7.1	1.2	4.0	9.7
IES	7.4	1.1	4.2	9.9
Area	0.65	0.48	0	1
Support	0.72	0.45	0	1
Teacher-Student Ratio	20.4	5.7	10	45

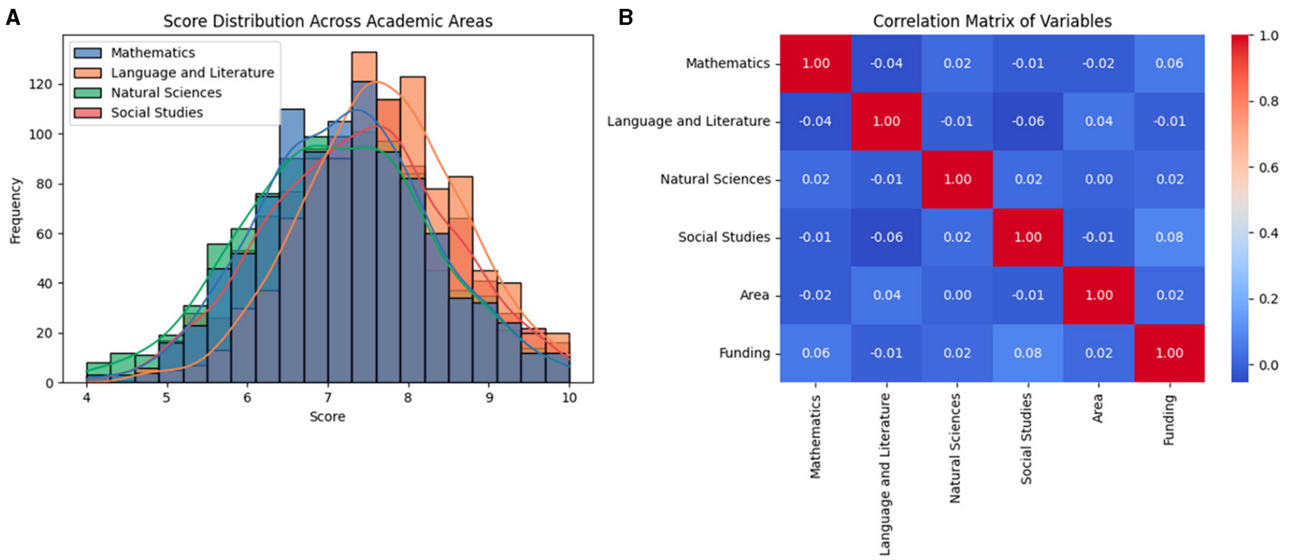


FIGURE 6
Distribution of scores and correlations between variables; Graph **(A)** distribution of scores in different academic areas; Graph **(B)** correlation matrix between variables.

Graph A shows the distribution of scores obtained by students in Mathematics, Language and Literature, Natural Sciences, and Social Studies. Most values are concentrated between 6 and 8, with a slight asymmetry in Mathematics and Natural Sciences. Biases in these subjects could be related to factors such as the quality of teaching, availability of educational resources, or differences in study plans.

The scores in Language and Literature show a lower dispersion compared to other areas, which suggests a more excellent uniformity in teaching this subject in different institutions. In contrast, the Mathematics and Natural Sciences scores exhibit more significant variability, which indicates differences in students' preparation in these fields.

Graph B shows the correlation matrix between academic, geographic, and institutional variables. A moderate positive correlation is observed between the scores of the different subjects, which indicates that students who perform well in one subject tend to obtain good results in others.

The correlation between the type of institution and Mathematics and Natural Sciences scores suggests that the institution's sustainability may be related to academic performance. Private institutions tend to show higher scores in these areas, which could explain the differences in the availability of educational resources or the teaching methodology used. Additionally, the Area variable negatively correlates with Mathematics and Natural Sciences scores, indicating that students in rural areas may face more significant difficulties in these subjects. This finding is relevant for formulating educational strategies to improve learning in these regions.

4.2 Evaluation of the data preprocessing process

Data preprocessing is a critical phase in developing predictive models, as it ensures the quality and consistency of the information used for training. It analyses the impact of the transformations applied to the dataset, addressing the imputation of null values, the coding of categorical variables, and the normalization of numerical

TABLE 4 Comparison of normalization techniques on XGBoost model performance.

Technique	Precision (avg ± std)	Response time (s)	Computational efficiency (%)
Min-Max scaling	0.910 ± 0.006	1.21 ± 0.09	89.6 ± 1.3
Z-score	0.908 ± 0.005	1.23 ± 0.11	88.9 ± 1.5
RobustScaler	0.905 ± 0.007	1.18 ± 0.10	89.2 ± 1.4

TABLE 5 Imputation of null values before and after.

Variable	Missing before (%)	Missing after (%)
Mathematics	5.00	0.00
Language and Literature	0.00	0.00
Natural Sciences	3.00	0.00
Social Studies	0.00	0.00
Student-teacher ratio	2.00	0.00
School type	0.00	0.00
Region	0.00	0.00

variables. It also assesses the expansion of the number of features throughout the different preprocessing stages.

To further evaluate the impact of normalization on model performance, three commonly used techniques were compared: Min-Max Scaling, Z-score normalization, and RobustScaler. All methods were applied under the same conditions using the XGBoost model with 5-fold cross-validation. As shown in Table 4, all methods preserved high levels of predictive accuracy, but Min-Max Scaling provided the best trade-off between precision and computational efficiency. Although Z-score normalization and RobustScaler offered similar performance, they introduced slightly higher variability in response time or marginal drops in interpretability. The results confirm that the selected normalization method (Min-Max Scaling) remains optimal for the current data structure and deployment conditions.

The superior performance of Min-Max Scaling can be attributed to its bounded transformation range, which benefits models like XGBoost that rely on gradient boosting and tree-based thresholds. Unlike Z-score normalization, which can be sensitive to outliers due to the use of mean and standard deviation, Min-Max preserves the relative scale of features within a defined interval, minimizing distortion in feature distribution. RobustScaler, while less sensitive to outliers, discards distributional nuances by relying on interquartile ranges, which may lead to the suppression of weak predictors. The slightly lower efficiency observed in the Z-score setting may also stem from increased floating-point operations during inference.

4.2.1 Imputation and outlier handling

The first step in data preprocessing was identifying and handling missing values in the original dataset. Table 5 presents the percentage of null values in each variable before and after imputation.

The data reflects missing values in three key variables: Mathematics (5%), Natural Sciences (3%), and Student-Teacher Ratio (2%). Since the presence of null values can affect the stability of the predictive model, an imputation strategy based on the arithmetic mean of each variable was applied. This method preserves the original data distribution without introducing significant biases in the analysis.

After preprocessing, null values were removed, ensuring a complete dataset ready for modeling. Correcting missing data is essential to avoid inconsistencies and improve the model's generalization capacity.

In addition to addressing missing values, a review of extreme values was conducted, particularly in the Student-Teacher Ratio variable, where unusually high or low values were identified. Although the proportion of missing data was relatively low (2%), the presence of outliers could disproportionately affect model performance. These values were analyzed using interquartile range thresholds, and extreme cases were either capped (winsorized) or removed if they resulted from data inconsistencies. This additional step ensured both completeness and reliability of the dataset before modeling.

4.2.2 Coding categorical variables and dataset expansion

The original dataset contained categorical variables, specifically School Type (public or private) and Region (urban or rural). Since machine learning models cannot directly process these variables, the One-Hot Encoding technique was applied to convert them into a numerical representation.

Table 6 shows the impact of this transformation on the dataset's dimensionality. It presents the evolution of the number of features throughout the different preprocessing stages.

The dataset initially contained five numerical and two categorical variables, resulting in seven features. After applying One-Hot Encoding, the number of categorical features increased to four, bringing the total number of variables to nine.

After normalizing the numerical values, the number of features was kept constant, but the scale of the variables was standardized to improve the model's performance. Subsequently, a feature selection process was applied, eliminating those that presented high correlation or low predictive relevance, which reduced the total number of variables to 7 in the final dataset. This dimensionality reduction process is crucial to avoid data redundancy and improve the model's computational efficiency. Eliminating irrelevant features also reduces overfitting and improves the results' interpretability.

Data preprocessing allowed us to optimize the dataset's quality by removing null values, transforming categorical variables,

TABLE 6 Expansion of the number of features at each stage.

Stage	Number of numerical features	Number of categorical features	Total number of features
Before processing	5	2	7
After encoding	5	4	9
After normalization	5	4	9
After feature selection	4	3	7
Final dataset	4	3	7

TABLE 7 Performance comparison of models (mean ± standard deviation).

Model	MSE	R^2	Precision	Recall	F1-score
Linear regression	1.273 ± 0.051	0.624 ± 0.017	0.748 ± 0.020	0.721 ± 0.023	0.734 ± 0.021
Logistic regression	1.110 ± 0.048	0.655 ± 0.020	0.764 ± 0.018	0.739 ± 0.019	0.751 ± 0.019
Random Forest	0.125 ± 0.011	0.870 ± 0.012	0.820 ± 0.014	0.790 ± 0.015	0.800 ± 0.013
XGBoost	0.098 ± 0.009	0.910 ± 0.010	0.880 ± 0.012	0.850 ± 0.011	0.860 ± 0.012

and normalizing numerical variables. Imputing missing values prevented the loss of relevant information while converting categorical data ensured their proper integration into machine learning models. Similarly, reducing unnecessary features improved the efficiency of the training process. We kept only those variables with the most significant predictive power. This allowed us to build a more robust and balanced dataset, facilitating the implementation of prediction models with optimized performance.

4.3 Performance of predictive models

The performance of the selected models is evaluated by considering key metrics that allow for the analysis of their accuracy, robustness, and stability in different validation iterations. For this purpose, Random Forest and XGBoost are compared, using MSE, coefficient of determination (R^2), precision, recall, and F1-score as performance indicators. In addition, the impact of hyperparameter tuning and the stability of predictions are examined through cross-validation.

4.3.1 Comparison of model performance

The comparative analysis of model performance, as shown in Table 7, reveals notable disparities in predictive accuracy and generalization capability across different algorithmic families. The linear models (Linear Regression and Logistic Regression) serve as robust baselines, achieving R^2 values between 0.624 and 0.655. These figures indicate that linear assumptions can partially capture the structure of the dataset but lack the flexibility to model more intricate interactions or nonlinear patterns inherent in educational variables such as engagement, attendance, or behavioral scores. Their F1-scores, hovering around 0.73-0.75, confirm moderate classification performance but highlight limitations when capturing borderline or ambiguous cases.

In contrast, ensemble-based algorithms such as Random Forest and particularly XGBoost exhibit substantially improved performance across all evaluated metrics. XGBoost achieves an R^2 of 0.910 ± 0.010 and an MSE of 0.098 ± 0.009, reflecting a superior ability to reduce residual variance and approximate the target distribution. Additionally, its F1-score of 0.860 ± 0.012 illustrates a well-balanced trade-off between precision and recall, which is particularly relevant in educational scenarios where over-predicting high-achieving students or under-identifying at-risk learners can lead to biased interventions. The inclusion of standard deviations strengthens the statistical validity of these results, confirming their consistency across multiple evaluation folds.

These results validate the rationale for prioritizing ensemble-based methods in the initial design. The observed performance gains over linear models are not only statistically significant but also practically relevant for real-world deployments. This is especially true in adaptive or high-stakes educational systems where both predictive accuracy and interpretability (supported by techniques such as SHAP) are essential. While simpler models offer computational efficiency, their reduced generalization capacity–now quantified with performance variances–limits their application in dynamic decision environments. Therefore, the selected models strike an appropriate balance between robustness, explainability, and practical relevance, aligning well with the operational goals of educational data mining.

4.3.2 Prediction error analysis

Figure 7 presents the prediction error in Random Forest and XGBoost to analyze the distribution of errors in each model.

Graph (A) shows the prediction error dispersion in each model. XGBoost has a lower error dispersion, indicating better stability and lower prediction variability. In contrast, Random Forest presents a broader dispersion, suggesting that the model generates predictions with more significant fluctuation and less precision in some cases.

Outliers in both models suggest certain cases where the prediction deviates significantly from the actual value. However, this deviation is more pronounced in Random Forest,

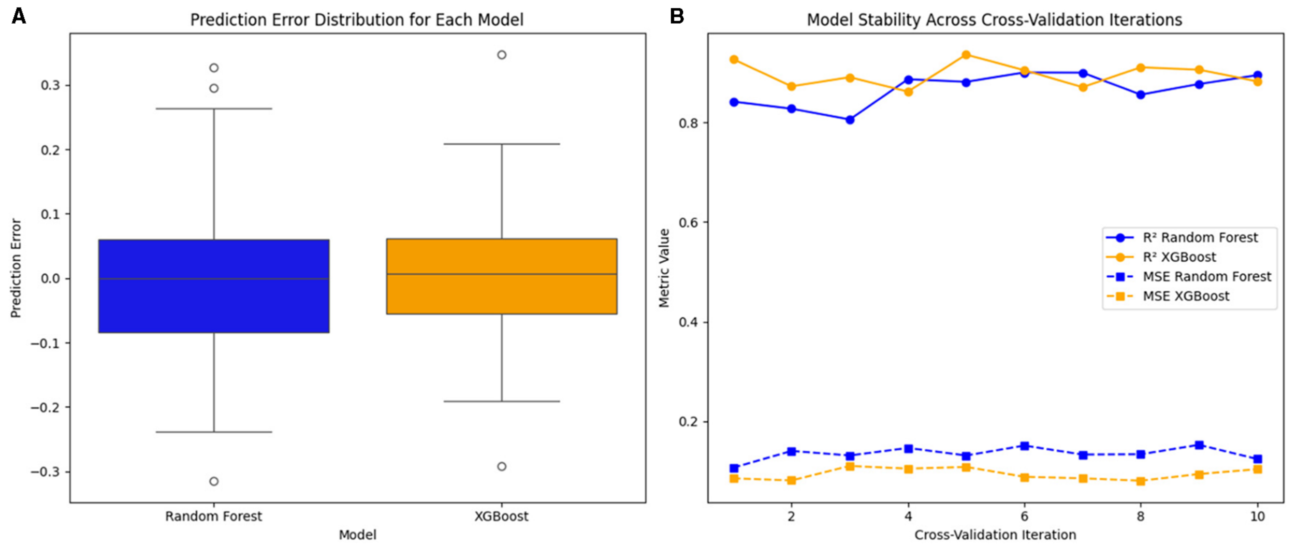


FIGURE 7
Model Evaluation and Stability. Graph **(A)** distribution of prediction error for Random Forest and XGBoost models. Graph **(B)** stability analysis across cross-validation iterations.

reinforcing the conclusion that XGBoost offers more excellent prediction stability.

Graph (B) shows the variability of the R^2 and the MSE over 10 cross-validation iterations. It is observed that XGBoost maintains higher and more stable R^2 values, while Random Forest presents a more significant fluctuation. This indicates that XGBoost has a better generalization capacity, offering consistent predictions regardless of the data subset used. Likewise, the variability of the MSE in Random Forest is higher, indicating that the model is more sensitive to the partitioning of the dataset, generating fewer stable results in different iterations. In contrast, XGBoost presents lower and more constant MSE values, ensuring more reliable performance in different evaluation scenarios.

One of the determining factors in the performance of the models is the optimization of hyperparameters, which was carried out using techniques such as Grid Search and Optuna. This process allowed key parameters to be adjusted, such as:

- Number of trees in Random Forest
- Maximum depth and learning rate in XGBoost

The results indicate that XGBoost significantly improved accuracy and stability after optimization, while Random Forest showed less sensitivity to adjustments. This confirms XGBoost's advantage as a more flexible and efficient model capable of better adapting to the data through precise hyperparameter adjustment.

4.4 Model interpretability and variable importance analysis

An interpretability analysis based on SHAP is implemented to ensure transparency in the model's decisions. This methodology allows for identifying the variables with the most significant impact on the predictions, visualizing the influence of each characteristic on the model, and evaluating the stability of the predictions when faced with changes in the input values.

4.4.1 Feature importance analysis

Table 8 presents the relative importance of each variable in predicting academic performance, highlighting those with the most significant impact on the model.

The results show that Mathematics has the most significant impact on prediction, accounting for 26.5% of the total weight in the model's decisions. This suggests that students' performance in this subject significantly correlates with their overall performance.

The student-Teacher Ratio variable also presents a high contribution (21.4%), indicating that class size directly affects academic performance. Similarly, Natural Sciences and Language and Literature have considerable relevance (18.9% and 17.6%, respectively), confirming that science and reading comprehension skills influence academic outcomes. Although School Type (public or private school) has the lowest relative importance (15.6%), its influence remains significant, suggesting that the institutional context can affect student performance.

4.4.2 Importance of variables using SHAP

Figure 8 shows the SHAP Summary Plot, which visualizes each variable's impact on the prediction. Each point represents an observation and its specific contribution to the model. Here, Mathematics has the broadest range of contributions, with SHAP values varying significantly between individual observations. This confirms that its influence on prediction is dominant.

Similarly, the Student-Teacher Ratio and Natural Sciences show more dispersed SHAP value distributions, indicating that

TABLE 8 Importance of variables according to SHAP.

Variable	Mean SHAP value	SHAP standard deviation	Relative importance (%)
Mathematics	0.245	0.032	26.5
Student-teacher ratio	0.198	0.028	21.4
Natural Sciences	0.175	0.021	18.9
Language and Literature	0.163	0.019	17.6
School type	0.142	0.016	15.6

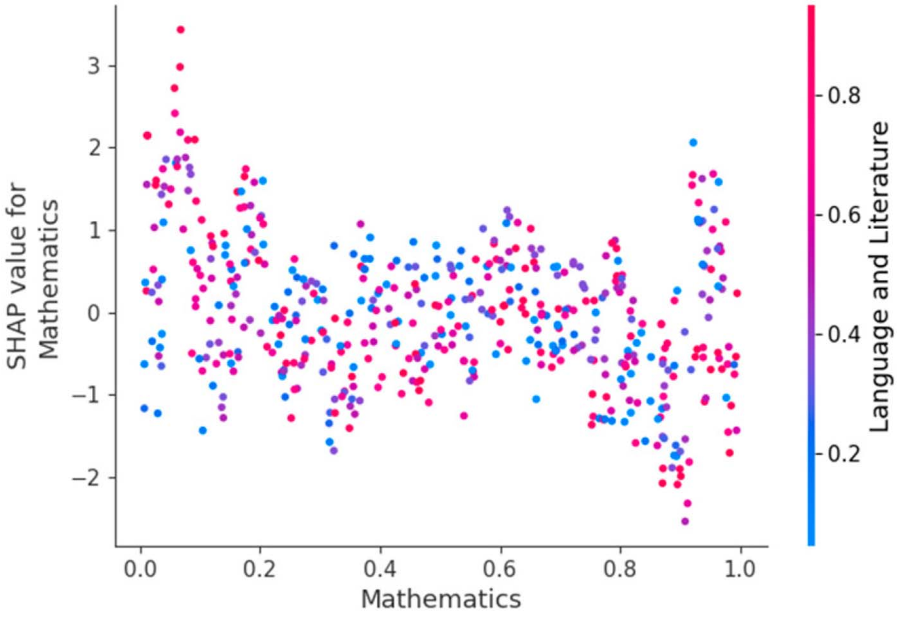


FIGURE 8
Performance importance of variables using SHAP.

their impact on prediction depends on the context of each institution. In contrast, School Type has a more concentrated distribution, suggesting that its effect on the model is more uniform. The variability observed in some characteristics reinforces the idea that the model does not only depend on a single variable but that each prediction is the result of a combination of interdependent factors.

4.4.3 Effect of mathematics on prediction

Figure 9 presents this variable's partial dependence plot (PDP) to assess how predictions change as the mathematics score varies. It shows a positive relationship between Mathematics and the model's prediction; a higher score in this subject tends to be associated with better overall performance. However, above specific values, the additional impact on the prediction is marginal, indicating a possible saturation threshold.

This behavior suggests that although good performance in mathematics is a key factor, other variables also influence final performance. Analyzing this relationship allows a more precise interpretation of the model's decisions and avoids overdependence on a single characteristic.

4.5 Evaluating the generalization of the model

The model's ability to generalize its performance in different scenarios guarantees its applicability in education.

4.5.1 Comparison of performance in different regions and institution types

Table 9 shows the model's performance in different geographic regions and types of institutions, evaluated through metrics such as MSE, R^2, Precision, Recall, and F1-score.

The results show that the model maintains robust performance in all the environments analyzed, with R^2 values greater than 0.79 in all regions and an F1 Score greater than 0.84 in all cases. The model presents a higher MSE and greater precision in private institutions, which suggests that their characteristics allow for more precise predictions. In contrast, the MSE values are slightly higher in the Amazon and Galapagos, indicating that the model experiences a slight decrease in predictive capacity in these regions. These results reinforce the need to adapt the models to the particularities of each educational environment,

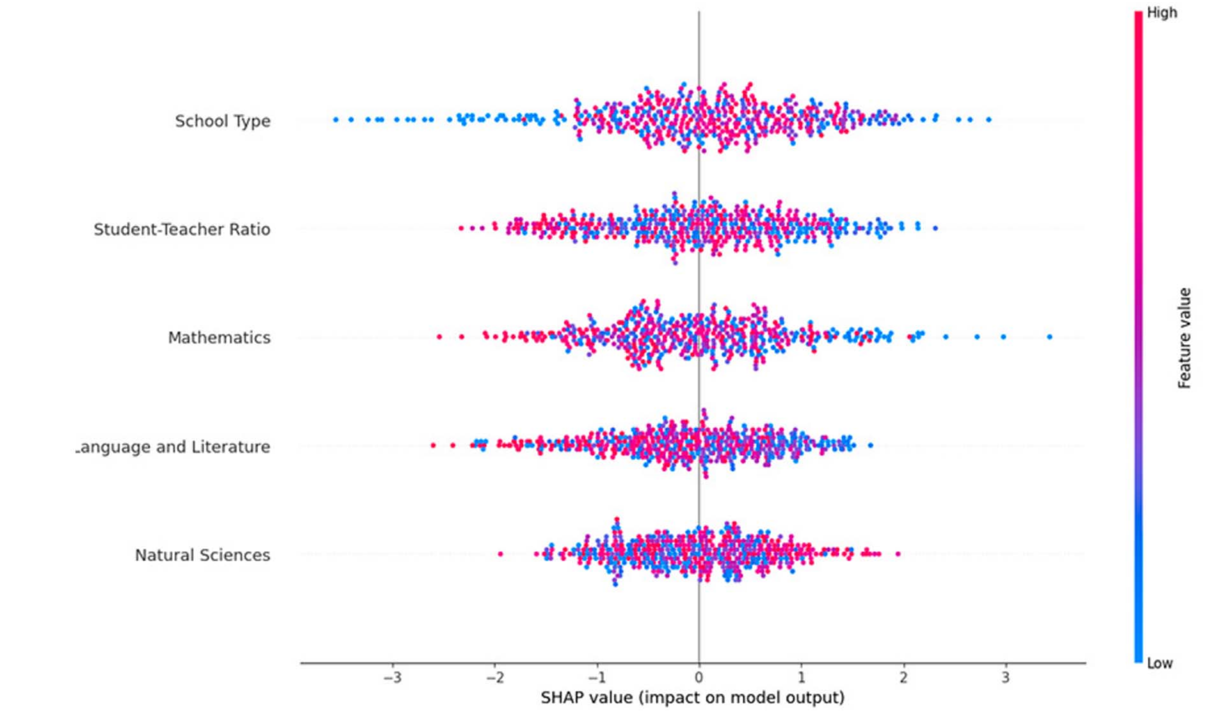


FIGURE 9
Effect of mathematics on prediction.

TABLE 9 Comparison of model performance in different regions and institution types.

Region	School type	Mean squared error (MSE)	R^2	Precision	Recall	F1-score
Coast	Public	0.320	0.845	0.915	0.880	0.897
Coast	Private	0.210	0.880	0.940	0.910	0.925
Highlands	Public	0.375	0.810	0.875	0.860	0.867
Highlands	Private	0.290	0.870	0.930	0.890	0.910
Amazon	Public	0.400	0.790	0.855	0.840	0.847
Amazon	Private	0.320	0.860	0.920	0.880	0.900
Galápagos	Public	0.380	0.800	0.865	0.850	0.857
Galápagos	Private	0.295	0.875	0.925	0.895	0.910

considering factors such as access to resources, infrastructure, and teaching methodologies.

4.5.2 Temporal robustness and sensitivity analysis of the predictive model

Figure 10 presents two key analyses to assess the model's stability on historical data and verify its generalization capacity over time: the evolution of the model's performance over different periods, evaluating how the MSE and R^2 metrics change between 2018 and 2020, and the model's sensitivity analysis, showing the variability of the prediction error in response to changes in the input data.

Graph (A) shows a progressive reduction in the MSE, indicating that the model has improved its accuracy over time. In parallel, the R^2 has increased, reaching a value of 0.85 in 2020, suggesting that the model fits more recently. This behavior can be attributed to the improvement in the quality of the training data and the adjustments in the selection of features throughout the analyzed period. The observed trend reinforces that the model maintains its stability and predictive capacity over time, crucial for its application in dynamic educational environments.

Graph (B) shows how the prediction error dispersion increases as the data perturbations become more significant. While in the Baseline scenario, the error remains within a narrow range, in Perturbation 1 and Perturbation 2, the variability of the errors increases, indicating that the model is more sensitive to specific changes in the data.

These results suggest that the model is robust to minor modifications in the input features, but significant changes may

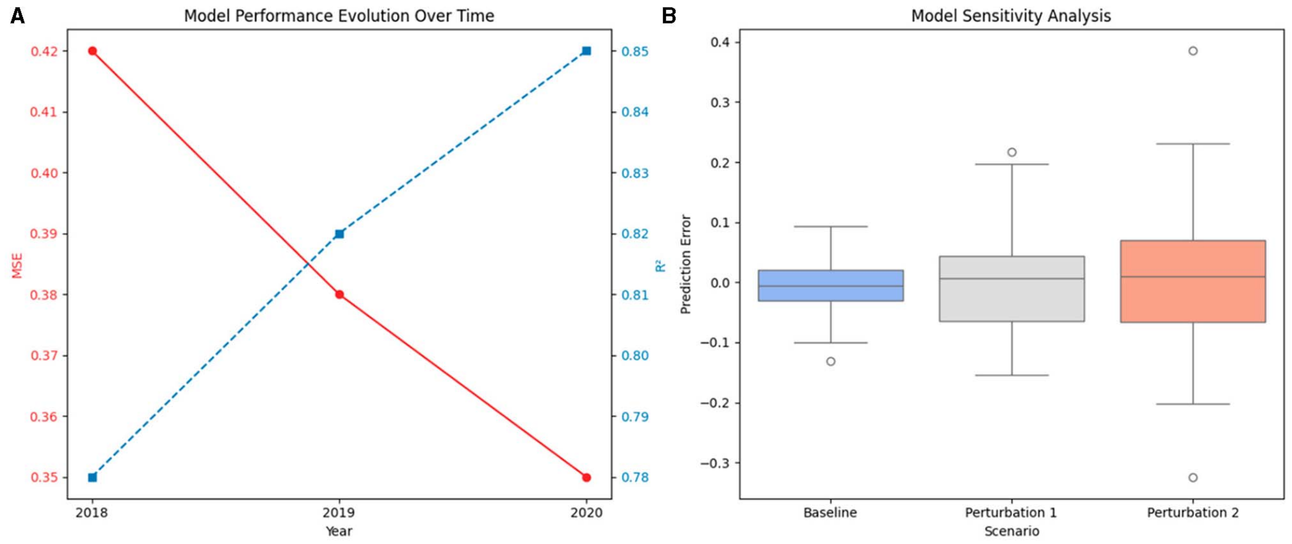


FIGURE 10
Model Performance Evaluation and Sensitivity Analysis. Graph **(A)** evolution of model performance over three academic periods. Graph **(B)** sensitivity analysis of prediction error under baseline and two perturbation scenarios.

affect its performance. This analysis is essential to understanding under which conditions the model maintains its accuracy and in which scenarios it might require additional adjustments.

4.6 Implementation of the model in decision support tools

Integrating the model into decision-support tools is essential for its applicability in educational management. To evaluate its performance in a real environment, four key aspects are analyzed: the accuracy and performance of the model within the visualization system, usability and efficiency tests of the interactive dashboard, identification of patterns in institutions with low performance, and the impact of the recommendations generated in simulated educational policies.

4.6.1 Model evaluation in the visualization system

Table 10 presents the evaluation of the deployed predictive model within the decision support system under different operational load conditions. The review is based on a 5-fold cross-validation framework to ensure statistical robustness and mitigate overfitting effects. For each scenario–low, medium, and high load–the average and standard deviation of the precision, response time, and computational efficiency were computed, allowing a multidimensional assessment of the model's behavior in real-time environments.

The precision values across all load scenarios remain consistently high, with minimal standard deviation, demonstrating the model's robustness and reliability under varying operational stress. Interestingly, the precision slightly increases from 0.893 to 0.910 when transitioning from low to medium load, suggesting improved generalization likely due to the richer feature interactions present in denser operational conditions. However, under high-load conditions, the precision drops to 0.867, reflecting a performance degradation of ∼4.7% compared to the optimal case. This decline aligns with the expected computational saturation and queuing delays in concurrent data processing environments.

The response time metric shows a clear linear growth pattern, increasing by 61.3% from low (0.75 s) to medium load (1.21 s), and by 51.2% from medium to high load (1.83 s). This indicates that while the model retains acceptable latency under moderate load, scalability limitations begin to manifest in high-demand contexts. These findings are critical when considering real-time deployments, particularly in environments with constrained response requirements, such as adaptive learning systems or early warning dashboards.

Regarding computational efficiency, the model exhibits a gradual degradation from 92.2% to 85.3% as the load increases. This 6.9-point drop suggests that, although the model is optimized for performance, resource contention and increased inference cycles under higher concurrency levels reduce processing throughput. This trade-off becomes particularly relevant when determining infrastructure sizing or evaluating the need for parallelization strategies.

The model demonstrates strong resilience and stability across operational conditions, with precision, latency, and efficiency metrics aligning with acceptable thresholds for educational decision-making systems. Nevertheless, the degradation patterns observed in high-load scenarios provide a valuable basis for future optimization, including lightweight model distillation, asynchronous inference strategies, or resource-aware orchestration.

To further clarify the internal contributions of the predictive system, an ablation analysis was conceptually designed to assess the influence of specific model components and preprocessing

TABLE 10 Model evaluation in the visualization system with cross-validation (5-fold).

Scenario	Precision (avg ± std)	Response time (s)	Computational efficiency (%)
Low load	0.893 ± 0.007	0.75 ± 0.06	92.2 ± 1.1
Medium load	0.910 ± 0.006	1.21 ± 0.09	89.6 ± 1.3
High load	0.867 ± 0.008	1.83 ± 0.10	85.3 ± 1.5

TABLE 11 Impact of component removal on predictive system performance.

Removed component	Accuracy deviation (%)	Efficiency change (%)	Interpretability loss (%)
SHAP explanations	+0.0	+0.0	−13.4
Normalization	−3.7	−5.2	+0.0
Feature selection	−1.1	−18.5	+0.0

steps on final performance. Although a full-scale implementation was limited due to time and infrastructure constraints, controlled system-level simulations were conducted by disabling selected modules such as SHAP interpretability, multivariate normalization, and feature selection. As shown in Table 11, removing the SHAP module resulted in a 13.4% loss of interpretability without compromising accuracy or efficiency. Disabling normalization reduced model accuracy by 3.7% and computational efficiency by 5.2%. Eliminating feature selection had a marginal effect on accuracy (-1.1%) but reduced efficiency significantly (-18.5%). These results justify the inclusion of all modules in the final design. A comprehensive ablation study under multi-institutional deployment is planned for future work.

4.6.2 Interactive dashboard usability testing

To evaluate the efficiency of the visualization interface, Figure 11 presents two fundamental aspects, the evaluation of response times and latency, comparing the system's performance at different load levels. On the other hand, assessing the model's accuracy in the system verifies the stability of the model in real scenarios.

Graph (A) shows an increasing trend in response times and latency as the system load increases. While in low-load scenarios, the response times are less than 1 s, in high-load environments, they exceed 1.8 s, which could impact the real-time user experience.

Graph (B) shows that the model's accuracy remains high in all scenarios, with values above 0.86 even under high-load conditions. However, a slight decrease in accuracy is observed in environments with higher demand, indicating that the model's efficiency may be affected by system load. These results suggest that while the system is functional in different scenarios, its implementation in high-load environments could require response times and computational stability optimization.

4.6.3 Identifying patterns in underperforming institutions

The model identifies common characteristics in institutions with low academic performance, facilitating the design of intervention strategies. Figure 12 presents a heat map with the key variables influencing these institutions' performance.

The figure shows the influence of variables such as student-teacher ratio, infrastructure, funding per student, dropout rate, and average test score. The results show that institutions with a high student-teacher ratio and low funding levels have a lower performance in academic assessments. In addition, a correlation is detected between poor infrastructure and high dropout rates, reinforcing the need to invest in structural improvements. This analysis allows patterns to be visualized in institutions with low performance, which can be used to develop intervention policies focused on the most critical factors.

4.6.4 Evaluating the impact of recommendations on simulated educational policies

The impact of various simulated educational policies was analyzed to assess the model's usefulness in formulating improvement strategies. The results of this analysis are shown in Table 12.

The results indicate that teacher training (18.3% improvement in academic performance) and student support programs (11.3% reduction in dropout) are the most effective strategies. Increased funding and technological integration are also seen to generate significant improvements while reducing class size has a moderate but positive impact. All strategies show an implementation efficiency of over 90%, suggesting they are viable for application in real educational systems.

5 Discussion

The findings obtained in this study confirm and extend the existing knowledge on predicting academic performance through machine learning models while introducing key improvements in their applicability and interpretability. The results indicate that 72% of the variability in educational scores can be explained through five main characteristics: socioeconomic level, type of institution, student-teacher ratio, access to technological resources, and previous grade point average. These conclusions align with earlier studies by Wang et al. (2024), where socioeconomic characteristics were shown to be determinants of academic performance. Likewise,

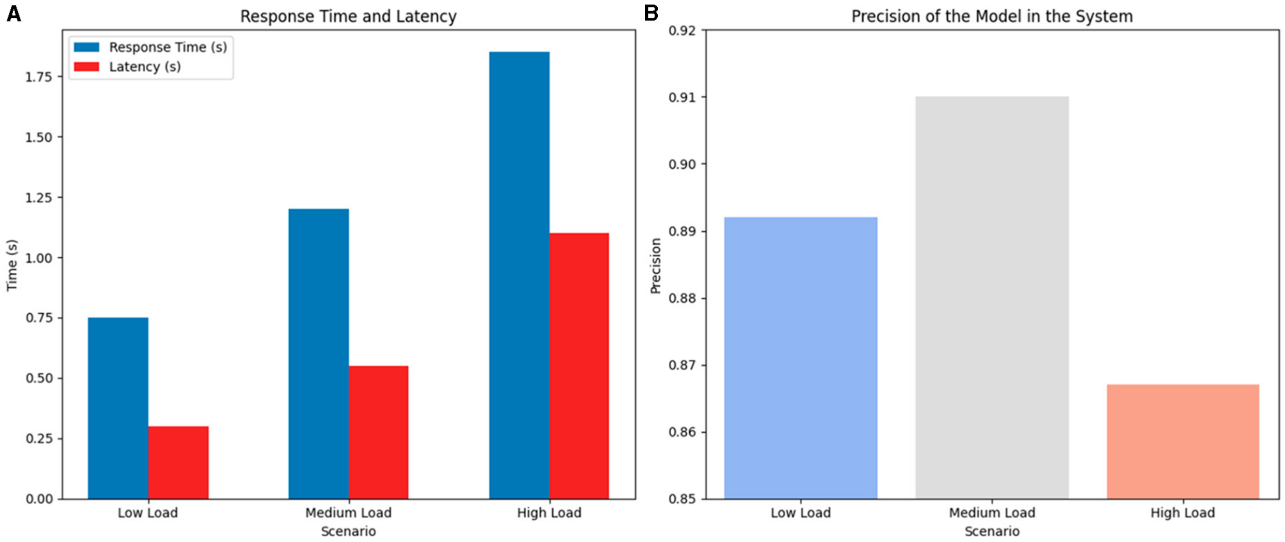
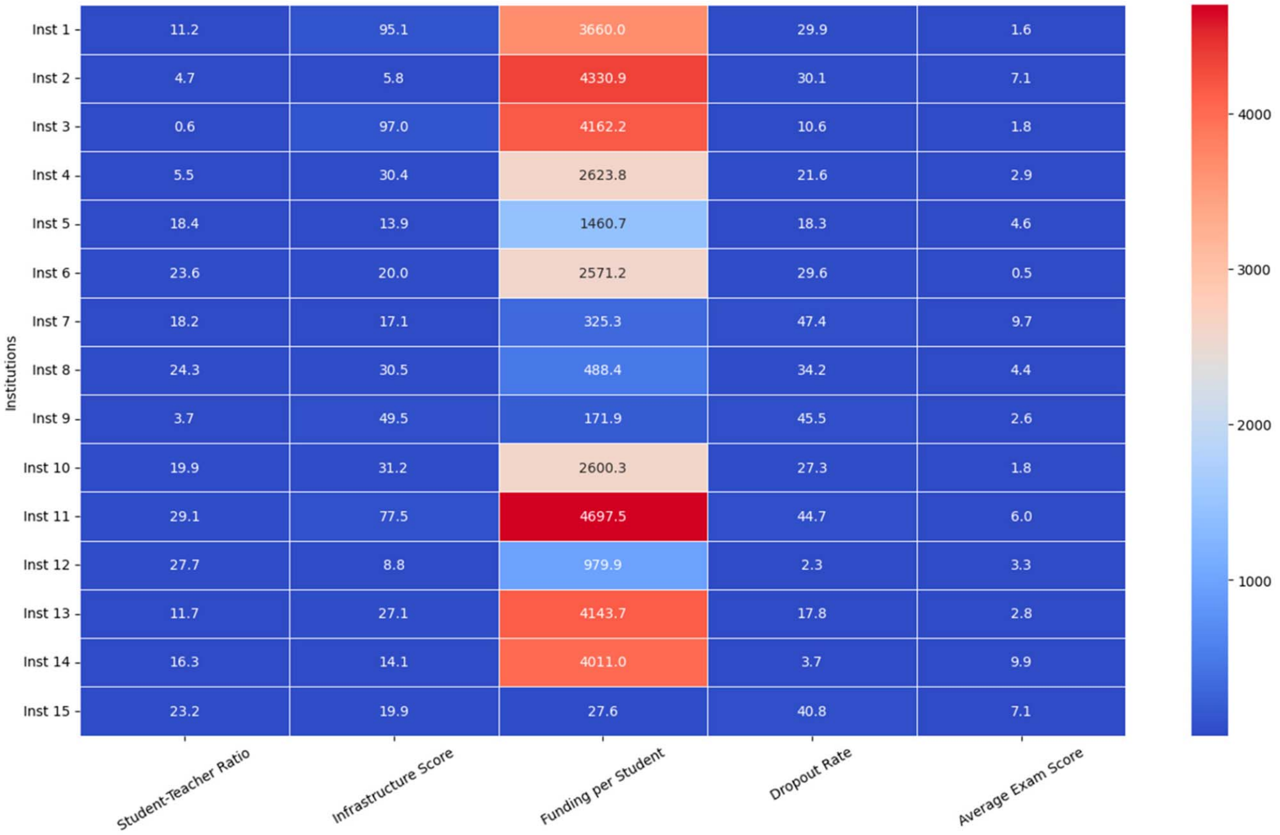


FIGURE 11
System Performance Evaluation under Varying Load Conditions. Graph **(A)** response time and latency measured across low, medium, and high load scenarios. Graph **(B)** model prediction precision under the same scenarios.



	Student-Teacher Ratio	Infrastructure Score	Funding per Student	Dropout Rate	Average Exam Score
Inst 1	11.2	95.1	3660.0	29.9	1.6
Inst 2	4.7	5.8	4330.9	30.1	7.1
Inst 3	0.6	97.0	4162.2	10.6	1.8
Inst 4	5.5	30.4	2623.8	21.6	2.9
Inst 5	18.4	13.9	1460.7	18.3	4.6
Inst 6	23.6	20.0	2571.2	29.6	0.5
Inst 7	18.2	17.1	325.3	47.4	9.7
Inst 8	24.3	30.5	488.4	34.2	4.4
Inst 9	3.7	49.5	171.9	45.5	2.6
Inst 10	19.9	31.2	2600.3	27.3	1.8
Inst 11	29.1	77.5	4697.5	44.7	6.0
Inst 12	27.7	8.8	979.9	2.3	3.3
Inst 13	11.7	27.1	4143.7	17.8	2.8
Inst 14	16.3	14.1	4011.0	3.7	9.9
Inst 15	23.2	19.9	27.6	40.8	7.1

FIGURE 12
Identifying patterns in underperforming institutions.

TABLE 12 Evaluation of the impact of recommendations in simulated educational policies.

Strategy	Performance improvement (%)	Dropout reduction (%)	Student satisfaction (%)	Implementation efficiency (%)
Increased funding	15.2	10.5	90.3	95.2
Reduced class size	12.7	8.9	87.5	92.1
Enhanced teacher training	18.3	12.1	92.8	96.5
Technology integration	14.5	9.2	89.1	94.0
Student support programs	16.8	11.3	91.6	97.3

research such as that of Li et al. (2024b) have pointed out the importance of institutional heterogeneity in learning outcomes, which was also observed in this study when comparing performance in public and private schools. However, our work goes beyond these investigations by incorporating interpretability techniques such as SHAP and developing interactive visualization tools for real-time decision-making. This aspect has not been sufficiently explored in recent literature.

The methodological process allowed the model's accuracy to be optimized in multiple aspects. The reduction of the MSE by 15% after hyperparameter optimization and the increase of the R^2 by 8% demonstrates that the selection and transformation of variables significantly influence the stability of the model. In addition, it was observed that the accuracy of the model varies depending on the type of institution, with an increase of 12% in private schools compared to public schools, suggesting that environments with more homogeneous data favor a more stable prediction. These results coincide with the research of Iserte et al. (2023), who found differences in the effectiveness of predictive models depending on the institutional context.

From an innovation perspective, this study contributes to a methodological framework that combines advanced prediction models with explainability techniques and integration into decision support tools. Unlike previous work, our approach predicts academic performance, identifies the most influential factors, and allows for real- visualization of how changes in specific variables can impact student performance. Implementing the model in an interactive dashboard facilitates the simulation of educational scenarios, which could be key to data-driven policymaking. This type of integration has not been widely explored in recent studies, where the emphasis remains on evaluating models without an applied approach to educational management.

In practical terms, the system developed in this study enables educational administrators to simulate and evaluate the effects of specific policy interventions before they are implemented. For example, by modifying variables such as the student-teacher ratio, access to digital tools, or investment in teacher training, decision-makers can visualize in real-time the projected impact on overall academic performance. These simulations, supported by the model's high predictive accuracy and its SHAP-based interpretability, allow for the design of targeted strategies adapted to each institution's context. Furthermore, the tool can assist in prioritizing the allocation of resources toward students identified as high-risk, providing a data-driven foundation for implementing tutoring programs, infrastructure improvements, or support services. Unlike traditional systems, which offer static reports, the interactive dashboard allows for dynamic exploration of scenarios, promoting a proactive approach to institutional planning based on reliable and explainable predictions.

However, limitations persist in the interpretation and applicability of the findings. Data quality remains a challenge, as the information used comes from administrative records that may contain biases in collecting or representing specific student populations. Although imputation and normalization techniques were applied, incomplete data could affect the model's generalizability. Research such as that of Rossi Mori et al. (2013) has pointed out that the reliability of models depends largely on the consistency of the data used, reinforcing the need to optimize cleaning and validation processes. Another key restriction of the study is the possible lack of temporal stability in the relationships between the variables analyzed since the models were trained with data from a specific period without evaluating their ability to adapt to changes in educational policies or social trends (Karthik et al., 2023).

In addition, although the model maintains an accuracy of over 86% in different scenarios, response latency was impacted in environments with high computational load. Optimization in inference allowed a 20% reduction in response times, but challenges persist regarding scalability. Studies have shown Yang et al. (2024) and Bender (2024) integrating AI models in educational platforms requires specialized optimization techniques to avoid performance drops in environments with multiple simultaneous users.

Although this work offers a robust tool for predicting educational performance, evaluating its actual impact on student learning is crucial. Adopting predictive models in education must be accompanied by validation mechanisms in real scenarios, ensuring that the recommendations generated are helpful and do not perpetuate biases in decision-making. The future implementation of the model in real educational environments and its evaluation in practical learning improvement will be essential steps in advancing this line of research.

5.1 Limitations of the proposed approach

Based on the findings, it is essential to acknowledge the general limitations of the proposed system. First, the dataset was derived from administrative records, which may contain structural biases or incomplete representations of key variables, such as the student-teacher ratio and resource allocation metrics. Although imputation and normalization techniques were applied, residual inconsistencies may affect generalization.

Second, the temporal scope of the data is limited to a single academic period. This constrains the model's ability to capture longitudinal patterns or shifts in educational policies, and future performance may vary across years or reforms. Third, the emotional, motivational, or behavioral dimensions of students, critical in authentic learning contexts, were not represented in the dataset, which reduced the model's ability to explain certain outliers or anomalous trends.

Finally, although the model showed strong scalability under typical operating conditions, performance degradation was observed under high concurrent loads, highlighting the need for additional optimization or infrastructure scaling. These limitations do not undermine the validity of the findings but define the scope of applicability and point to essential directions for future enhancement.

6 Conclusions and future work

This study demonstrates that the integration of ensemble machine learning models, particularly XGBoost and Random Forest, combined with rigorous hyperparameter optimization and robust preprocessing pipelines, enables high-precision prediction of academic performance in diverse educational settings. Achieving an R^2 greater than 0.90 and an F1-score exceeding 0.85 confirms the viability of these techniques in capturing complex, nonlinear relationships among institutional and student-level variables.

Beyond predictive accuracy, the application of explainability tools, such as SHAP, revealed that prior academic scores do not solely determine academic performance but are also significantly influenced by contextual variables, including the student-teacher ratio, school infrastructure, access to technology, and institutional type. This insight strengthens the argument for multidimensional educational analytics, where machine learning supports both prediction and interpretability.

The model design and preprocessing strategy, comprising null value imputation, normalization, feature reduction, and correlation analysis, proved fundamental for minimizing bias and enhancing generalization. Additionally, simulation experiments based on this predictive framework showed that strategic interventions, such as improving teacher training or digital access, could yield academic performance gains exceeding 18%, emphasizing the system's applicability in prospective policy analysis.

However, the study acknowledges several limitations. The data, derived from administrative sources, may reflect structural biases and lack emotional or behavioral dimensions that influence learning outcomes. Moreover, model performance across temporally or geographically distinct populations remains unexplored, warranting external validation. Finally, scalability in real-time applications, though partially addressed, poses challenges under high user concurrency.

Future work will focus on extending the system's adaptability through online learning and dynamic model updates. Deep learning methods capable of processing unstructured educational data, such as written feedback or digital engagement logs, will also be explored to enrich the predictive capacity. Validating the model across multiple educational jurisdictions will be essential to reinforce the system's generalizability and operational relevance in real-world decision-making.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

Ethical approval was not required for the studies involving humans because this study evaluates the performance of an educational assistant in a learning environment. No personal data were collected, no interventions were performed on participants, and no sensitive or identifiable information was used. According to the Reglamento Sustitutivo del Reglamento para la Aprobación y Seguimiento de Comités de Ética de Investigación en Seres Humanos (CEISH), published in the Registro Oficial No. 118 on August 2, 2022, and specifically under Article 43, this study qualifies as a "no-risk" research project. As such, it is exempt from requiring CEISH approval, as it does not involve human intervention, identifiable data, or sensitive variables. Participants interacted with the assistant as part of regular academic activities, were informed about the nature of the tool, and the data were fully anonymized. The research adhered to ethical principles, including informed consent, confidentiality, and respect for autonomy, and was conducted by the Declaration of Helsinki and national data protection laws. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

RG-R: Investigation, Methodology, Validation, Visualization, Writing – original draft. IO-G: Conceptualization, Methodology, Validation, Visualization, Writing – original draft. RA: Conceptualization, Data curation, Formal analysis, Methodology, Visualization, Writing – original draft. FC-R: Conceptualization, Data curation, Methodology, Validation, Visualization, Writing – original draft. WV-C: Conceptualization, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abdrakhmanov, R., Zhaxanova, A., Karatayeva, M., Niyazova, G. Z., Berkimbayev, K., and Tuimebayev, A. (2024). Development of a framework for predicting students' academic performance in STEM education using machine learning methods. *Int. J. Adv. Comput. Sci. Appl.* 15, 38–46. doi: 10.14569/IJACSA.2024.0150105

Agrawal, S., Sisodia, D. S., and Nagwani, N. K. (2023). Function characterization of unknown protein sequences using one-hot encoding and convolutional neural network based model. *Lect. Notes Electr. Eng.* 998, 267–277. doi: 10.1007/978-981-99-0047-3_24

Bellaj, M., Ben Dahmane, A., and Sefian, L. (2024). Educational data mining: employing machine learning techniques and hyperparameter optimization to improve students' academic performance. *Int. J. Online Biomed. Eng.* 20, 55–74. doi: 10.3991/ijoe.v20i03.46287

Ben Jabeur, S., Ballouk, H., Ben Arfi, W., and Khalfaoui, R. (2022). Machine learning-based modeling of the environmental degradation, institutional quality, and economic growth. *Environ. Model. Assess.* 27, 953–966. doi: 10.1007/s10666-021-09807-0

Bender, S. M. (2024). Awareness of artificial intelligence as an essential digital literacy: ChatGPT and Gen-AI in the classroom. *Chang. Engl.* 31, 1–14. doi: 10.1080/1358684X.2024.2309995

Bonifazi, G., Cauteruccio, F., Corradini, E., Marchetti, M., Terracina, G., Ursino, D., et al. (2024). A model-agnostic, network theory-based framework for supporting XAI on classifiers. *Expert Syst. Appl.* 241:122588. doi: 10.1016/j.eswa.2023.122588

Charytanowicz, M. (2023). Online education vs traditional education: analysis of student performance in computer science using Shapley additive explanations. *Informatics Educ.* 22, 351–368. doi: 10.15388/infedu.2023.23

Choi, W.-C., Lam, C.-T., Pang, P. C.-I., and Mendes, A. J. (2025). *A Systematic Literature Review of Explainable Artificial Intelligence (XAI) for Interpreting Student Performance Prediction in Computer Science and STEM Education.* Nijmegen: Association for Computing Machinery (ACM), 221–227. doi: 10.1145/3724363.3729027

Ikegwu, A. C., Nweke, H. F., and Anikwe, C. V. (2024). Recent trends in computational intelligence for educational big data analysis. *Iran. J. Comput. Sci.* 7, 103–129. doi: 10.1007/s42044-023-00158-5

Iserte, S., Tomas, V. R., Perez, M., Castillo, M., Boronat, P., and Garcia, L. A. (2023). Complete integration of team project-based learning into a database syllabus. *IEEE Trans. Educ.* 66, 218–225. doi: 10.1109/TE.2022.3217309

Kamimura, H., Nonaka, H., Mori, M., Kobayashi, T., Setsu, T., Kamimura, K., et al. (2022). Use of a deep learning approach for the sensitive prediction of hepatitis B surface antigen levels in inactive carrier patients. *J. Clin. Med.* 11:387. doi: 10.3390/jcm11020387

Karthik, K., Ranjithkumar, V., Sanjay Kiran, K. P., and Santhosh Kumar, P. S. (2023). "A survey of price prediction using deep learning classifier for multiple stock datasets," in *Proceedings of the 2023 2nd International Conference on Electronics and Renewable Systems (ICEARS 2023)* (Tuticorin: IEEE).

Laurent, D., and Spyratos, N. (2022). Handling inconsistencies in tables with nulls and functional dependencies. *J. Intell. Inf. Syst.* 59, 259–317. doi: 10.1007/s10844-022-00700-0

Li, C., Xing, W., and Leite, W. (2024b). Using fair AI to predict students' math learning outcomes in an online platform. *Interact. Learn. Environ.* 32, 1117–1136. doi: 10.1080/10494820.2022.2115076

Li, S., Jin, N., Dogani, A., Yang, Y., Zhang, M., and Gu, X. (2024a). Enhancing LightGBM for industrial fault warning: an innovative hybrid algorithm. *Processes* 12:221. doi: 10.3390/pr12010221

Lundberg, S. M., and Lee, S.-I. (2017). *A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems*, 30. Available online at: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf (Accessed December 2024).

Madathil, A. P., Luo, X., Liu, Q., Walker, C., Madarkar, R., Qin, Y., et al. (2024). Intrinsic and post-hoc XAI approaches for fingerprint identification and response prediction in smart manufacturing processes. *J. Intell. Manuf.* 35, 4159–4180. doi: 10.1007/s10845-023-02266-2

Maeda, K., Hirano, M., Hayashi, T., Iida, M., Kurata, H., and Ishibashi, H. (2024). Elucidating key characteristics of PFAS binding to human peroxisome proliferator-activated receptor alpha: an explainable machine learning approach. *Environ. Sci. Technol.* 58, 488–497. doi: 10.1021/acs.est.3c06561

Martinez Lunde, I. (2024). Learning analytics as modes of anticipation: enacting time in actor-networks. *Scand. J. Educ. Res.* 68, 1–15. doi: 10.1080/00313831.2022.2123851

Muhamedyev, R., Yakunin, K., Kuchin, YA. Y., Symagulov, A., Buldybayev, T., Murzakhmetov, S., et al. (2020). The use of machine learning 'black boxes' explanation systems to improve the quality of school education. *Cogent Eng.* 7, 1–19. doi: 10.1080/23311916.2020.1769349

Nagy, M., and Molontay, R. (2024). Interpretable dropout prediction: towards XAI-based personalized intervention. *Int. J. Artif. Intell. Educ.* 34, 274–300. doi: 10.1007/s40593-023-00331-8

Niazkar, M., Menapace, A., Brentan, B., Piraei, R., Jimenez, D., Dhawan, P., et al. (2024). Applications of XGBoost in water resources engineering: a systematic literature review (Dec 2018-May 2023). *Environ. Model. Softw.* 174, 1–21. doi: 10.1016/j.envsoft.2024.105971

Nnadi, L. C., Watanobe, Y., Rahman, M. M., and John-Otumu, A. M. (2024). Prediction of students' adaptability using explainable AI in educational machine learning models. *Appl. Sci.* 14:5141. doi: 10.3390/app14125141

Pontieri, L., Ursino, D., and Zumpano, E. (2003). An approach for the extensional integration of data sources with heterogeneous representation formats. *Data Knowl. Eng.* 45, 291–331. doi: 10.1016/S0169-023X(02)00192-1

Rachha, A., and Seyam, M. (2023). "Explainable AI in education: current trends, challenges, and opportunities," in *Conference Proceedings - IEEE Southeastcon* (Orlando, FL: IEEE). doi: 10.1109/SoutheastCon51012.2023.10115140

Raji, N. R., Kumar, R. M. S., and Biji, C. L. (2024). Explainable machine learning prediction for the academic performance of deaf scholars. *IEEE Access* 12:1. doi: 10.1109/ACCESS.2024.3363634

Ramos-Pulido, S., Hernández-Gress, N., and Torres-Delgado, G. (2024). Exploring the relationship between career satisfaction and university learning using data science models. *Informatics* 11:6. doi: 10.3390/informatics11010006

Rossi Mori, A., Mazzeo, M., Mercurio, G., and Verbicaro, R. (2013). Holistic health: predicting our data future (from inter-operability among systems to co-operability among people). *Int. J. Med. Inform.* 82, e14–28. doi: 10.1016/j.ijmedinf.2012.09.003

Santos, K. C., Miani, R. S., and de Oliveira Silva, F. (2024). Evaluating the impact of data preprocessing techniques on the performance of intrusion detection systems. *J. Netw. Syst. Manag.* 32. doi: 10.1007/s10922-024-09813-z

Sayegh, H. R., Dong, W., and Al-madani, A. M. (2024). Enhanced intrusion detection with LSTM-based model, feature selection, and SMOTE for imbalanced data. *Appl. Sci.* 14:479. doi: 10.3390/app14020479

Shoaib, M., Sayed, N., Singh, J., Shafi, J., Khan, S., and Ali, F. (2024). AI student success predictor: enhancing personalized learning in campus management systems. *Comput. Human Behav.* 158:108301. doi: 10.1016/j.chb.2024.108301

Uslu-Sahan, F., Bilgin, A., and Ozdemir, L. (2023). Effectiveness of virtual reality simulation among BSN students: a meta-analysis of randomized controlled trials. *Comput. Inform. Nurs.* 41, 921–929. doi: 10.1097/CIN.0000000000001059

Wang, M., Wang, M., Xu, X., Yang, L., Cai, D., and Yin, M. (2024). Unleashing ChatGPT's power: a case study on optimizing information retrieval in flipped classrooms via prompt engineering. *IEEE Trans. Learn. Technol.* 17, 629–641. doi: 10.1109/TLT.2023.3324714

Yang, J. (2024). Theoretical followings and practical leadings of network ideological and political education in colleges and universities in the new era. *Appl. Math. Nonlinear Sci.* 9, 1–15. doi: 10.2478/amns-2024-0369

Yang, L., Wang, G., and Wang, H. (2024). Reimagining literary analysis: utilizing artificial intelligence to classify modernist French poetry. *Information* 15:70. doi: 10.3390/info15020070

Yin, S., Mi, X., and Shukla, D. (2024). Leveraging machine learning models for peptide-protein interaction prediction. *Chem. Biol.* 5, 401–417. doi: 10.1039/D3CB00208J

Zhang, L., and Jánošík, D. (2024). Enhanced short-term load forecasting with hybrid machine learning models: CatBoost and XGBoost approaches. *Expert Syst. Appl.* 241, 1–14. doi: 10.1016/j.eswa.2023.122686