

Article

Design and Implementation of an Intelligent Diagnostic System for Academic Performance Analysis in Medical Education

Margarita Aucancela ^{1,2,*} , Alfonso González-Briones ¹ and Pablo Chamoso ¹ 

¹ BISITE Research Group, University of Salamanca, Edificio I+D+i, Calle Espejo 2, 37007 Salamanca, Castile and León, Spain; alfonso@usal.es (A.G.-B.); chamoso@usal.es (P.C.)

² School of Public Health, Escuela Superior Politécnica de Chimborazo, Panamericana Sur km 1 ½, Riobamba 060155, Chimborazo, Ecuador

* Correspondence: idu023958@usal.es or maucancela@epoch.edu.ec

Abstract

This study presents the design and implementation of a single-institution intelligent diagnostic system to identify low mid-period academic performance, aimed at activating proactive and preventive tutoring before a final assessment. The system features an integrated analytical architecture comprising an inferential framework, a predictive framework, an explainability framework, a validation framework, and a Streamlit-based web prototype. The sample uses 18,604 longitudinal academic records from 1264 unique students enrolled across 7 consecutive academic periods (2017–2020) at an Ecuadorian university. Results indicate that curricular level is the structural predictor with the greatest independent contribution (semi-partial $R^2 = 0.044$), followed by academic period (semi-partial $R^2 = 0.026$). Random Forest achieved the best overall performance (MAE = 1.267 ± 0.04 ; RMSE = 1.714 ± 0.05 ; $R^2 = 0.551 \pm 0.02$), outperforming other algorithms. SHAP explainability confirms the primacy of curricular level and academic period as individual-level risk-associated factors, enabling the generation of interpretable alerts for tutors. The equity analysis revealed that students aged 30–50 years (ratio = 1.375) and the province with code 18 (ratio = 1.395) constitute priority subgroups for data enrichment prior to institutional deployment. External validation with real users is identified as the next research stage.

Keywords: intelligent diagnostic system; academic performance; Random Forest; mixed linear model; SHAP; medical education; GroupKFold; explainable AI; Streamlit

1. Introduction

Academic performance in medical education is determined by multiple individual, curricular, and institutional factors. Academic risk is the estimated probability that academic performance will fall below a critical institutional threshold in a future period, jeopardizing a student's educational continuity [1]. Students enrolled in medical programs must maintain high academic standards for thirteen or more semesters while simultaneously developing cognitive, procedural, and attitudinal competencies that evolve from the basic sciences toward clinical and hospital practice [2,3]. Poor academic performance is directly associated with academic lag, course failure, and dropout, entailing high costs for students, the educational system, and society at large [1].

Intelligent diagnostic systems for academic performance analysis represent one of the most promising applications of artificial intelligence in higher education, as they transform large volumes of data into interpretable signals that enable the faculty tutor to take pedagogically grounded actions before academic lag consolidates [4].



Academic Editor: George Angelos Papadopoulos

Received: 1 May 2026

Revised: 13 June 2026

Accepted: 15 June 2026

Published: 25 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

In Ecuador, 34 universities located across 12 provinces offer medical training programs. A substantial proportion of enrolled students must relocate from their home provinces to access their education, adding a socioeconomic burden that compounds an already demanding academic workload [5].

Faculty tutors are formally responsible for individual academic monitoring under Ecuadorian higher education regulations [6], yet must simultaneously manage teaching, research, community outreach, and institutional administration, which constrains the time available for proactive tutorial follow-up. Furthermore, the absence of intelligent diagnostic systems that provide timely and accurate information on individual academic performance reduces the possibilities of timely tutorial intervention, increasing the risk of academic failure. Compounding this problem, four critical methodological gaps persist in the existing literature. First, most intelligent diagnostic systems operate on small or single-semester datasets without modeling the longitudinal hierarchical structure of multi-period educational data [7,8]; the absence of studies that simultaneously employ large-volume longitudinal records spanning multiple consecutive academic periods with explicit hierarchical modeling constitutes a fundamental methodological gap that limits the operational applicability of the models proposed in the literature [9]. Second, existing systems direct their output toward institutional managers or automated LMS platforms, not toward the individual faculty tutor as the operational recipient [7,10]. Third, the impact of preprocessing decisions on algorithm selection is rarely reported, limiting reproducibility [11]. Fourth, no existing diagnostic system for medical education integrates inferential and predictive SHAP layers into a unified explanatory attribution interface deployable as a web prototype.

This study proposes the design and implementation of a single-institution intelligent diagnostic system to identify low mid-period academic performance, with the aim of activating proactive and preventive tutoring before a final assessment. The system is intended to assist faculty tutors in analyzing academic performance and the following:

- (1) Uses the first two grades of the same academic period ($N1_C$ and $N2$), summing them ($SUM = N1_C + N2$) and using this result as an intermediate checkpoint to transform performance data into pedagogical inputs that activate proactive and preventive tutoring before the final assessment;
- (2) Integrates an inferential framework based on a three-level mixed linear model (MLM) with a predictive framework based on a supervised regression pipeline for individual academic performance scoring;
- (3) GroupKFold cross-validation applies to respect the longitudinal structure of student academic trajectories;
- (4) Implements an explainability and interpretability framework based on SHAP at both the population level (MLM) and the individual level (Random Forest) to produce interpretable low-performance profiles for tutors;
- (5) Implements a predictive uncertainty validation framework through a robustness and subgroup equity analysis to determine whether the supervised regression model is stable, reproducible, and applicable across the distinct groups composing the student population;
- (6) Explicitly reports the impact of encoding decisions on algorithm performance;
- (7) Demonstrates the operational feasibility of the system through a Streamlit web prototype that integrates a dual SHAP framework together with a tutor-facing explanatory attribution matrix.

Mixed linear models (MLMs) constitute the inferential framework of the present study because they allow modeling data with multilevel hierarchical structure by decomposing total variance into three interpretable components: variance attributable to observable

fixed effects (marginal R^2), variance attributable to persistent interindividual differences such as academic self-regulation or resilience (the difference between conditional R^2 and marginal R^2), and unexplained residual variance; semi-partial R^2 coefficients additionally isolate each predictor's unique contribution, enabling a rigorous distinction between statistical significance and practical effect size [12–14]. However, MLMs quantify population-level relationships without generating individual predictions with the precision required to guide tutoring decisions and are therefore complemented by supervised machine learning algorithms.

The predictive framework for this study uses four supervised machine learning algorithms: Random Forest, XGBoost, Ridge Regression, and Multilayer Perceptron (MLP). The selection of these algorithms is based on their capacity to capture both linear and nonlinear relationships, control multicollinearity, and evaluate different levels of model complexity, in line with prior studies in medical education [15].

Random Forest is a supervised machine learning algorithm that constructs a specified number of decision trees in parallel using bootstrap sampling and random predictor subsets per node, obtaining the final prediction as the mean across all trees to reduce variance and improve generalization [4,16,17].

XGBoost builds trees sequentially and additively, where each tree minimizes the residual error of the previous ensemble through gradient descent with L1 and L2 regularization, capturing nonlinear interactions and threshold effects with greater precision on tabular data [4,18,19].

Ridge Regression incorporates an L2 quadratic penalty on the coefficients that shrink estimators toward zero without eliminating variables, making it particularly useful in the presence of multicollinearity and serving as a regularized baseline to quantify the predictive gain of ensemble models [16,20].

The Multilayer Perceptron (MLP) learns hierarchical nonlinear representations through hidden layers with ReLU activation and the Adam optimizer, offering competitive performance when data volume is sufficient, although its higher computational cost and lower interpretability generally place it below ensemble models on moderate-size datasets [7,20].

All four algorithms are evaluated via GroupKFold cross-validation, a k-fold variant that ensures all observations from the same student appear exclusively in either the training or the test set within each fold, preventing data leakage and producing realistic generalization estimates for unseen students [16,21].

Interpretability of individual predictions is achieved through SHAP (SHapley Additive exPlanations), a post hoc framework grounded in cooperative game theory that quantifies the marginal contribution of each predictor by simulating all possible variable coalitions and producing attribution values satisfying the properties of efficiency, symmetry, and additivity [22,23]; specifically, TreeExplainer generates a global mean-importance bar chart and an individual-level waterfall plot that transforms a black-box prediction into an actionable model-derived explanation for the tutor [4].

The validation framework includes robustness and subgroup equity analysis. The robustness analysis verifies that the model's error does not exhibit a systematic structure over time or across application conditions, confirming that the metrics reported on the full dataset are representative of the expected performance in each operational subset; the RMSE/MAE ratio stratified by academic period provides the standard indicator of this property, as values consistently below $\sqrt{2} \approx 1.414$ across all periods confirm that uncertainty is predominantly random and does not introduce cumulative systematic bias [24].

The subgroup equity analysis serves to verify that the predictive performance of a machine learning model is not unevenly distributed across the groups composing the population of interest (students differentiated by gender, province of origin, age range,

curricular level, and academic period), and that predictive uncertainty is not significantly higher for any particular subgroup—a condition that, if unmet, would imply that the system systematically directs less accurate alerts toward students belonging to structurally more vulnerable groups, amplifying rather than reducing pre-existing educational inequalities [4]. From the algorithmic fairness perspective, the MAE parity ratio per subgroup, computed as $MAE_{subgroup}/MAE_{global}$, quantifies whether the model commits proportionally larger errors for specific groups; an operational threshold of ≤ 1.25 indicates that the subgroup error does not exceed the global error by more than 25%, a standard adopted from the artificial intelligence fairness literature [4], and its exceedance in subgroups such as students aged over 30 years or underrepresented provinces signals the need for data enrichment or model adjustment prior to institutional deployment, ensuring that the system complies with the algorithmic accountability principles required of tutoring support systems in higher education [6,25].

Predictive uncertainty is validated through MAE, RMSE, and the RMSE/MAE ratio: when this ratio approaches $\sqrt{2} \approx 1.414$, errors are predominantly random and follow an approximately normal distribution; when it exceeds this threshold, outlier errors dominate and indicate systematic bias [24]; stratified decomposition of this ratio by academic period and sociodemographic subgroup constitutes the standard robustness analysis for verifying that the model's predictive uncertainty does not compromise its operational applicability as a differentiated tutoring support tool [24,26].

In mixed linear models and machine learning algorithms, the way categorical variables are encoded directly influences both the interpretation of coefficients and the predictive performance of the model. Among the most widely used methods are dummy coding, effect coding, ordinal encoding, one-hot encoding, and target encoding.

Dummy coding, or reference coding, assigns a binary column (0/1) to each category using a baseline category; in this way, the coefficients represent differences relative to that reference. This scheme is the default in libraries such as statsmodels and lme4 [27]. Effect coding, on the other hand, centers the estimates with respect to the grand mean of the dependent variable, allowing each category to be interpreted as a deviation from the overall average, which is useful when the interest lies in main effects rather than in an arbitrary reference category [27].

Ordinal encoding assigns consecutive integer values while preserving the natural ordering of categories—for example, curricular levels, from the seventh to the thirteenth level. This approach is particularly appropriate for variables with a hierarchical structure and is efficiently exploited by tree-based algorithms such as Random Forest and XGBoost, which learn split points along the ordinal axis [11]. In contrast, one-hot encoding generates an independent binary column for each category without assuming any ordering relationship among them, making it the most methodologically appropriate option for nominal variables such as gender or province of origin, where imposing a numerical scale would introduce artificial hierarchies and potential biases in the models [8].

Target encoding replaces each category with the meaning of the outcome variable within that group. This method offers important advantages for high-cardinality variables, as it avoids the excessive dimensionality increase produced by one-hot encoding [27]. Pargent et al. [28] demonstrated that regularized target encoding consistently outperforms one-hot and ordinal encoding when the number of categories is large. In practical terms, the literature considers low cardinality when fewer than 10 categories exist, medium cardinality between 10 and 50, and high cardinality when more than 50 categories are present. As the number of levels increases, one-hot encoding generates highly sparse matrices, which negatively affects the stability of both linear and ensemble models [29].

The explanatory attribution matrix constitutes a tool that integrates multiple explainability metrics to estimate the relative importance of each predictor within the model [29]. Its primary utility lies in resolving discrepancies between different interpretation methods—for example, when the coefficients of a mixed linear model differ from the SHAP values obtained from ensemble models [4,19]. At the global level, this matrix allows identification of whether a variable exerts a protective or risk-associated effect on academic performance; while at the individual level, it facilitates recognition of which factors most strongly influence the specific performance of each student [29]. In intelligent diagnostic systems applied to education, this interpretive layer is fundamental because it translates the model's quantitative output into actionable information for faculty and tutors, strengthening evidence-based decision-making [4,24].

The remainder of the paper is structured as follows: Section 2 reviews related work. Section 3 describes the system architecture and computational flow and presents the web prototype implemented in Streamlit. Section 4 details the materials and methods. Section 5 presents the results. Section 6 discusses the findings. Section 7 presents the limitations and future work. Section 8 presents the conclusions.

2. Related Work

This section organizes related work around five thematic axes:

- a. Hierarchical and longitudinal modeling;
- b. Predictive models based on machine learning;
- c. SHAP interpretability framework in interactive web applications;
- d. Intelligent systems for academic performance diagnosis;
- e. Ecuadorian context.

2.1. Hierarchical and Longitudinal Modeling

Educational data inherently exhibit a nested hierarchical structure: students are enrolled within courses, courses are part of curricular levels, and levels are part of a curriculum or program. Mixed linear models (MLMs), also known as hierarchical linear models or multilevel models, decompose total variance into between-group components, yielding unbiased coefficient estimates [30]. Nakagawa and Schielzeth [13] established marginal R^2 (variance explained by fixed effects) and conditional R^2 (variance explained by the full model including random effects) as standard effect size metrics for MLMs. Semi-partial R^2 coefficients further isolate the unique contribution of each predictor after partialing out shared variance [12], enabling a rigorous distinction between statistical significance and practical effect size—essential in large samples where trivial differences reliably yield $p < 0.001$ [12,30].

Recent work has combined mixed linear models (MLMs) with machine learning to leverage the complementary strengths of both approaches [15]. The present study adopts this dual-framework strategy: the MLM addresses the inferential question (which factors matter and to what extent?), while the predictive models address the predictive question (what will this student's performance be?).

2.2. Predictive Models Based on Machine Learning

Monteverde-Suárez et al. [31] developed a binary classification model to identify academically irregular first-year medical students at UNAM ($n = 7976$, seven cohorts) using academic, sociodemographic, and family environment variables. The most relevant predictors of irregularity were gender and family income, while prior academic performance dominated the prediction of regular student status.

Mastour et al. [32] evaluated the discriminative capacity of classical classification models and ensemble models for predicting medical students' performance on high-stakes examinations, such as the Comprehensive Basic Medical Sciences Examination (CMBSE), using a five-year database. The Random Forest model achieved AUC-ROC and accuracy values of 0.813 and 0.803, respectively, confirming the feasibility of early detection from institutional academic records.

Bilal et al. [16] analyzed demographic and academic variables from first-semester higher education students using a binary classification model to predict academic performance at the tenth semester. A decision tree was used to identify the most relevant variables. The combined use of both variable categories consistently improved model performance compared to using each category individually. The best-performing model was SVM, achieving 92% accuracy.

Suaza-Medina et al. [19] proposed a performance prediction model for a standardized examination known as Saber 11 in lagging regions of Colombia, using sociodemographic, academic, psychographic, and behavioral variables, combining nine classification algorithms with SHAP values to identify the influence of each variable. The best-performing algorithms were XGBoost, LightGBM, and GBM.

The models implemented in the studies by Monteverde-Suárez et al. [31], Mastour et al. [4,32], Bilal et al. [16], and Suaza-Medina et al. [19] converge on binary or multiclass classification, in contrast to the present study, which introduces a continuous regression model evaluated using metrics specific to the prediction problem, such as MAE and RMSE. Furthermore, a sensitivity analysis was conducted to validate model robustness, which is not evidenced in the studies.

Regarding the identification of the most relevant model features, Bilal et al. [16] used a decision tree, while Mastour et al. [4] and Suaza-Medina et al. [19] applied SHAP for variable explainability, and Monteverde-Suárez et al. [31] used the Naïve Bayes algorithm to analyze the influence of each variable. However, none of these studies address data with a hierarchical structure resolved through a mixed linear model (MLM), nor do they conduct an effect size analysis to determine the practical impact of variables on academic performance.

2.3. SHAP Interpretability Framework in Interactive Web Applications

The incorporation of SHAP explainability into interactive web interfaces designed for educators represents the most recent frontier in the design of intelligent academic support systems. Recent studies have highlighted the need for tools that bridge the gap between model accuracy and actionable insight for educators, noting the predominance of SHAP in student performance prediction model explanations and the need for dashboard integration that makes results interpretable without technical training.

Mastour et al. [4] incorporated an explainable artificial intelligence (XAI) framework based on SHAP values, validated across three universities through two high-stakes assessments ($n = 997$ and $n = 777$ students). The combination of academic and non-academic metrics in an interpretable stacking meta-model achieved an accuracy of 0.98 and an AUC-ROC of 0.97–0.99, and the application of XAI improved institutional acceptance of the predictions by making the contribution of each variable transparent at the individual level.

Islam et al. [33] integrated SHAP, Shapash, ELI5, and LIME as XAI techniques over classification models—Decision Tree, Random Forest, Gradient Boosting, and XGBoost—for global and local explainability of academic performance, identifying XGBoost as the highest-accuracy model (83%) and SHAP as the technique with the greatest capacity to produce individually interpretable explanations. Bañeres et al. [34] implemented interactive dashboards for faculty displaying continuous assessment risk levels with color-coded distribution and the

historical percentage of students passing with a similar profile, constituting one of the first systems to translate ML predictions into real-time actionable visualizations for the tutor.

Unlike these works, the present study implements an interactive Streamlit web interface that simultaneously integrates a dual SHAP explanatory attribution matrix combining an inferential framework (MLM) with a predictive framework (Random Forest), featuring a Priority Index per variable and an operational model-derived summary in natural language—functionality that has not been documented in any prior intelligent diagnostic system for medical education.

2.4. Intelligent Systems for Academic Performance Diagnosis

The development of intelligent systems for academic performance diagnosis has advanced significantly toward the incorporation of explainability techniques that increase the transparency of predictive models and facilitate their adoption by educators [35].

Abukader et al. [35] presented an intelligent student performance prediction system based on LightGBM with metaheuristic hyperparameter optimization, incorporating SHAP as the interpretability layer and demonstrating that this combination systematically decomposes each prediction into the contribution of input variables, improving transparency and supporting evidence-based decision-making within educational institutions.

Mastour et al. [4] proposed a performance prediction system for comprehensive medical assessments in Iran through a stacking meta-model combining XGBoost, Random Forest, and Adaptive Boosting with 26 academic and 5 non-academic variables, validated across three universities with accuracy = 0.98 and AUC-ROC = 0.97–0.99; their architecture included SHAP for global explainability but oriented toward institutional managers and accreditation committees.

Namoun and Alshantqi [7] conducted a systematic literature review on academic performance prediction through data mining and learning analytics, covering 186 studies published between 2010 and 2020. Ensemble models such as XGBoost and Random Forest dominated the latter years of the decade, the average reported accuracy was 82%, and the main limitation identified was the absence of implementation and user validation studies. This implementation gap motivates the operational workflow proposed in the present study.

Guanin-Fajardo et al. [8] presented an academic success prediction system for Ecuadorian university students using multiple machine learning techniques on admission and first-semester data. While geographically closer to the present study, this work does not focus on medical education, which reinforces the originality of the present contribution.

Pozo-Burgos et al. [36] conducted a study on academic performance with undergraduate students using surveys and multivariate dependence techniques combined with logistic regression to determine the degree of incidence of sociocultural and demographic factors on academic success; however, this study is not related to medical education.

Al Hashmi et al. [10] developed a machine learning-based predictive study using midterm grades from health sciences students. They compared five classification algorithms—XGBoost, Gradient Boosting, Random Forest, Logistic Regression, and Naïve Bayes—within a pipeline architecture with ROSE balancing, achieving AUC = 0.87 with XGBoost as the best model; their system was directed toward program coordinators rather than the individual faculty tutor.

Unlike Guanin-Fajardo et al. [8], Pozo-Burgos et al. [36], and Mastour et al. [4,32], this study uses a mixed linear model (MLM) to determine the factors influencing academic performance. Furthermore, unlike Guanin-Fajardo et al. [8], Pozo-Burgos et al. [36], Abukader et al. [35], and Mastour et al. [4], this study conducts a longitudinal analysis with intra-individual variability modeling.

Unlike Mastour et al. [32], Monteverde-Suárez et al. [31] and Cannistrà [37] this study models the same student in a medical program across 7 consecutive academic periods—a design not implemented in any of the aforementioned studies: Mastour et al. [32] focused on the Comprehensive Basic Medical Sciences Examination (CMBSE) using repeated cross-sectional cuts over 5 years; Monteverde-Suárez et al. [31] used repeated cross-sectional data from the first year of 7 cohorts (2011–2017); and Cannistrà [37] addressed the hierarchical problem through the Generalized Mixed-Effects Random Forest (GMERF) algorithm, producing a single model to determine dropout among first-semester engineering students at an Italian institute.

Furthermore, this proposal is oriented toward supporting the activities of faculty tutors in accordance with the regulations established by Ecuador's Secretariat of Higher Education, Science, Technology, and Innovation (SENESCYT) [6], which governs the academic activities of higher education institutions in Ecuador—unlike other studies that direct their outputs toward managers, program coordinators, or LMS platforms.

2.5. Ecuadorian Context

In Latin America, countries such as Mexico, Brazil, and Argentina concentrate the largest scientific output on learning analytics and AI in medical education. However, Ecuador does not appear in the studies analyzed in regional bibliometric reviews, confirming the need to generate local research that considers the country's specific institutional, curricular, and social conditions [36].

Ecuador comprises 24 provinces distributed across four natural regions: Coast, Sierra, Amazon, and the Insular Region (Galápagos). The academic offerings of medical programs are concentrated in 13 public higher education institutions and 21 private universities, distributed across 12 provinces: Azuay, Chimborazo, El Oro, Esmeraldas, Guayas, Imbabura, Loja, Manabí, Pichincha, Santo Domingo de los Tsáchilas, Sucumbíos, and Tungurahua [8].

The university that is the subject of the present study is located in the province of Chimborazo, in the central Sierra region, and receives students from all 24 provinces of the country. This condition makes the university a national reference for medical training with a markedly heterogeneous student body in terms of geographic origin, socioeconomic context, and cultural background.

This inter-provincial flow generates a specific socioeconomic reality with a significant impact on academic persistence: students who migrate from their home provinces must independently—or with limited family support—cover the costs of housing, food, transportation, academic materials, and other resources necessary for their professional training. These sustenance costs can compromise program continuity, particularly in middle- and low-income households and in families with more than one member enrolled in higher education. This economic pressure acts as an additional risk factor that compounds the academic demands of a program as rigorous as medicine. Incorporating province of origin as an institutional monitoring variable in academic management systems is therefore not merely a descriptive exercise, but a tool that enables the early identification of differentiated vulnerability profiles and directs tutorial actions toward those who need them most [1].

The medical program in this study is offered in face-to-face modality and comprises 13 curricular levels organized into three stages: (1) Basic Sciences (levels 1–4, including anatomy, physiology, biochemistry, and microbiology); (2) Clinical Sciences (levels 5–9, encompassing internal medicine, surgery, pediatrics, gynecology, and other clinical specialties with both theoretical and practical components); and (3) Hospital Internship (levels 10–13, corresponding to full-time rotations in university teaching hospitals). The total duration of the program is 6.5 years, including the mandatory rural health service year required to obtain a medical degree.

Annual enrollment has historically averaged between 180 and 220 new students per cohort. Approximately 65 to 70% of enrolled students complete all curricular requirements within the normative timeframe of 13 semesters, while the remainder experience academic lag, requiring additional semesters to complete specific courses. The overall graduation rate within a 10-year window reaches approximately 78%, with the highest dropout concentrated in the first three levels (basic sciences), consistent with internationally reported patterns [38].

3. System Architecture and Computational Flow

For the design and implementation of a single-institution intelligent diagnostic system that identifies low mid-period academic performance with the aim of activating proactive and preventive tutoring before a final assessment, the architecture comprises seven functionally distinct modules together with an interactive web interface, as shown in Figure 1.

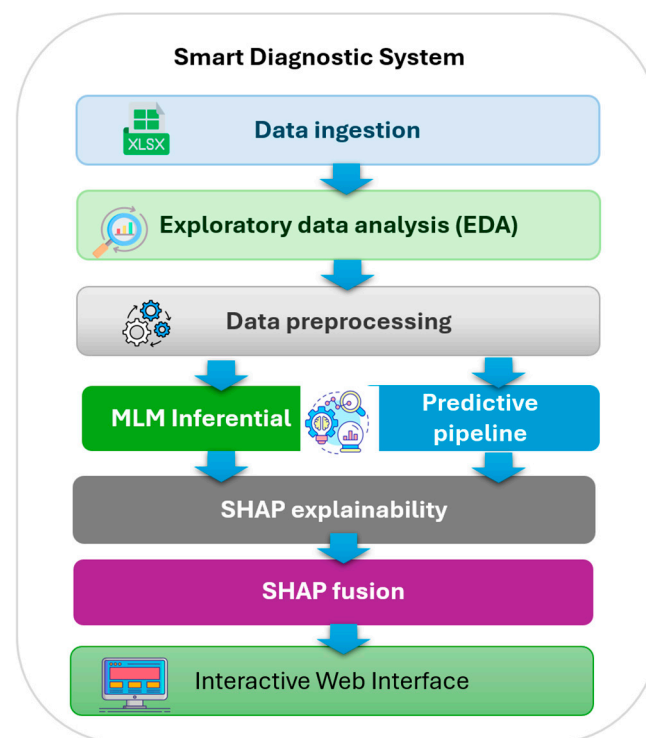


Figure 1. Comprehensive architecture of the intelligent diagnostic system for academic performance analysis.

3.1. Data Ingestion Module

This module handles data loading into the system using the pandas 2.0 library. The module receives as input a file in (.csv) format, which is read through the `pd.read_csv` function with a semicolon (;) as the column separator. This file is an anonymized secondary dataset. The output is a dataframe containing 3 academic variables: academic period (ID_PE), curricular level (COD_NI), and course (ID_MAT); 3 sociodemographic variables: gender (ID_SEX), province of origin (COD_PRO), and student age (ED_EST); 2 performance variables: first partial grade (N1) and second partial grade (N2); and one student identifier variable (COD_EST).

3.2. Exploratory Data Analysis (EDA) Module

This module is responsible for performing descriptive statistical analysis of the input variables to understand the data structure, academic context, and the presence of missing values. The libraries used in this module include matplotlib, seaborn, scipy.stats, pandas, and numpy. The EDA determined the following aspects:

- Data structure: variables, data types, cardinality, normality, behavior, and dispersion measures.
- Total number of records in the database.
- Number of academic periods, curricular levels, and courses.
- Number of curricular levels per academic period.
- Number of courses per level.
- Total number of unique enrolled students.
- Number of repeated observations per student.
- Number of female and male students.
- Number of students per province.
- Age: minimum, maximum, and mean.
- Grades: minimum, maximum, and mean—by period, by level, by course, and other breakdowns.

At this stage, the module receives as input a dataframe with 31,970 records and produces as output a dataframe with an analytic sample of 18,604 records.

3.3. Data Preprocessing Module

This module is responsible for performing feature engineering and variable transformation to ensure data compatibility in subsequent analyses. It receives as input a dataframe with 9 variables. The first partial grade (N1) is rescaled from 8 to 10 points (N1_C) in order to facilitate comparison with the second partial grade (N2) and obtain the academic performance variable ($SUM = N1_C + N2$), which represents the sum of the first two partial assessments within the same academic period on a 0–20 point scale. Age is log-transformed: $\text{np.log1p}(ED_EST)$. The variables PE_ORD and COD_NI are then transformed as ordinal variables, while COD_PRO, ID_SEX, and ID_MAT are transformed as nominal variables using the scikit-learn library. The output is a dataframe with 12 variables.

3.4. Inferential Analysis Module (MLM)

This module is responsible for implementing the inferential framework of the system, considering the data structure. The model formula is defined using the following variables: academic performance (SUM), age (ED_EST), province of origin (COD_PRO), gender (ID_SEX), academic period (PE_ORD), and curricular level (COD_NI) to implement the mixed linear model; the student identifier (COD_EST) is used as the grouping variable. The module was implemented using Python's statsmodels library. Additionally, custom functions were created to obtain marginal R^2 , conditional R^2 , the intraclass correlation coefficient (ICC), and semi-partial R^2 coefficients per predictor.

3.5. Predictive Analysis Module (Pipeline)

This module receives as input a dataframe with the 12 variables obtained from the data preprocessing module. A pipeline is then used for preprocessing and subsequent training of the predictive model together with four algorithms—Ridge, XGBoost, MLP, and Random Forest—via GroupKFold cross-validation ($k = 5$, grouped by student identifier COD_EST). The output is the best-performing predictive model based on MAE, RMSE, and R^2 across all folds.

A pipeline is a sequential computational structure that chains the preprocessing, feature engineering, and predictive model training stages in an ordered and reproducible manner into a single object that can be fitted, evaluated, and deployed as a coherent unit [11]. Its primary utility lies in ensuring that all transformations applied to the training data—OneHotEncoder, OrdinalEncoder, FunctionTransformer, and StandardScaler—are fitted exclusively on the training partition of each cross-validation fold and subsequently

applied to the test partition without accessing its distribution, thereby preventing data leakage that produces artificially optimistic performance estimates [8].

3.6. Validation Module

This module is responsible for performing the error analysis of the winning predictive model, along with the robustness and subgroup equity analysis, using the `GroupKFold` and `cross_val_predict` functions from the `sklearn` library. The outputs include error decomposition by academic period; MAE parity ratios by province of origin, age range, and gender; absolute error distribution plots by province of origin (COD_PRO) and gender; actual versus predicted performance plots; residual distribution; residuals versus fitted values; and model calibration plots.

3.7. SHAP Explainability Module

This module is responsible for implementing the SHAP explainability framework. It receives as input the MLM and the winning model—Random Forest.

At the global level, SHAP explainability ranks model variables according to their mean importance ($\text{mean } |\text{SHAP value}|$) across the entire sample, identifying which predictors have the greatest systematic influence on predictions; at the individual level, it decomposes the prediction for a specific case to quantify the marginal contribution of each predictor variable to an individual prediction generated by a machine learning model [22].

SHAP TreeExplainer was applied to the trained winning model to generate global feature importance rankings and individual-level Shapley value decompositions. The beeswarm plot and the mean absolute SHAP value bar chart constitute the interpretability layer for the tutor.

In the interactive web interface, the SHAP explainability framework generates waterfall plots and an explanatory attribution matrix to integrate the results of the inferential framework with the predictive framework, using the `shap` library (v0.42).

3.8. Interactive Web Interface

The interactive web interface was implemented using the Spyder 6.1 platform with Python 3.13 and Streamlit 1.54.0. The preliminary diagnostic prototype transfers the results of the inferential framework and the machine learning framework to the individual tutor in an interpretable and practical format that requires no technical expertise [7,10]. The preliminary diagnostic prototype integrates Module 4 (inferential layer) and Module 5 (Random Forest predictive layer) with Module 7 (SHAP explainability) into a three-tab interface, accessible at the following web address: <https://prototypeews-srcpuzg4c8otwfazefmj2q.streamlit.app/>, accessed on 1 June 2026.

The interactive web interface integrates six functional elements organized across a left sidebar and a main panel with three explanatory attribution analysis tabs.

The left sidebar contains a dropdown menu where the tutor can select a student identifier, which then displays four interface elements in the main panel. The first element is the student's predicted academic performance (SUM) on a 0–100 scale. The second is a color-coded risk semaphore (green: low risk, yellow: moderate risk, red: critical alert). The third element is a strategic action recommendation in natural language for the selected student. The fourth element is a set of 3 tabs. The first tab displays an individual-level waterfall plot presenting the results of the inferential SHAP (MLM layer). The second tab displays a waterfall plot presenting the results of the predictive SHAP (Random Forest model), which captures nonlinear interaction patterns that may protect performance in specific level–course–period combinations. The third tab presents the Explanatory Attribution Matrix, which maps the SHAP values from both frameworks (inferential and predictive) onto the study variables and computes a Priority Index, identifying the dominant risk-

associated factor and automatically generating the natural language clinical directive for the tutor. Figure 2 shows a screenshot of the web interface of the implemented platform.

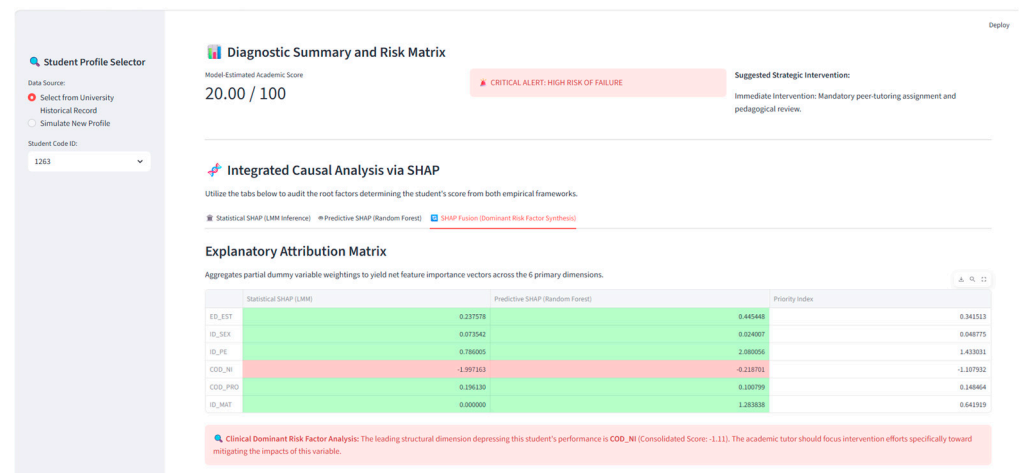


Figure 2. SHAP Fusion tab: Consolidated explanatory attribution matrix mapping SHAP values from both frameworks (MLM and Random Forest) onto 6 model variables, with Priority Index and operational model-derived summary. COD_NI was identified as the dominant risk-associated factor (Priority Index = -1.09).

3.9. Technical Implementation

The intelligent diagnostic system was implemented in two phases: a testing phase and a deployment phase. The testing phase was implemented using the Jupyter 7.5 platform and encompassed data ingestion, exploratory data analysis, data preprocessing, the inferential framework, the predictive framework, the SHAP explainability framework, and the validation framework. Within the predictive framework, a model comparison was conducted using the Ridge, XGBoost, MLP, and Random Forest algorithms, with the latter emerging as the winning model.

In the deployment phase, the preliminary diagnostic prototype was created through a project on the Spyder 6.1 platform using Python 3.13 and Streamlit (v1.54.0). The modules were developed in accordance with the system architecture, and each element of the prototype interface was implemented, including the alert messages and the tabs displaying the results of the inferential framework, the predictive framework, and the SHAP explanatory attribution matrix. The MLM was fitted using the MixedLM module from statsmodels; the Random Forest model was serialized using joblib for deployment within the Streamlit application. SHAP TreeExplainer and waterfall plots were generated using shap (v0.42). The application was first deployed locally, after which the source files were uploaded to GitHub Version 3.5.12 and the prototype was deployed on the Streamlit cloud at the following address: <https://prototypeews-srcpuzg4c8otwfazefmj2q.streamlit.app/>, accessed on 1 June 2026.

4. Materials and Methods

4.1. Data Ingestion

The intelligent diagnostic system designed to assist faculty tutors in analyzing academic performance first analyzes input variables and their effects through an MLM, then predicts students' academic performance using supervised learning algorithms, and uses these predictions to identify low-performance profiles that require early tutorial intervention.

The CSV file contains academic variables including: academic period (PE_ORD), curricular level (COD_NI), and course (ID_MAT); sociodemographic variables including: student age (ED_EST), gender (ID_SEX), and the student's province of origin code

(COD_PRO); and performance variables consisting of partial assessment grades: (N1) for the first partial assessment and (N2) for the second partial assessment of the same academic period, together with the student identifier (COD_EST).

4.2. Exploratory Data Analysis (EDA)

The exploratory data analysis was conducted by applying descriptive statistics to the dataset variables, grouping variables, using Python functions, and creating custom functions to understand the data structure and its academic context.

The statistical analysis revealed that the initial dataset contains a total of 31,970 records, 7 academic periods, 13 curricular levels, 65 courses, and 3 partial grades per academic period: the first grade (N1) scored out of 8 points, the second (N2) out of 10 points, and the third out of 10 points. Given the proactive and preventive tutoring paradigm, only the first two partial grades (N1) and (N2) of the same academic period were considered for this study. The institutional threshold for low academic performance indicates that a partial assessment score must be less than or equal to 7 points; consequently, the sum of the first two partial assessments implies a threshold below 14. Student ages range from 18 to 44 years; students originate from 22 provinces of Ecuador and include both male and female students.

Variable grouping enabled the extraction of the number of students with unique records, the number of male students, the number of female students, the number of students enrolled per academic period and per level, the number of observations per student, and the number of students per academic period and curricular level. This last grouping revealed that the feature matrix exhibits an unbalanced hierarchical nesting configuration with sequential left truncation. The curricular level encompasses longitudinal trajectories that asymptotically converge at the terminal level (Level 13). This architecture comprises 7 sequential sub-cohorts (from cohort 1–13 through cohort 7–13), where each subsequent level naturally omits observations from prior periods, which motivated the adoption of a selective intentional sampling strategy to ensure longitudinal comparability between period 0 and period 6, controlling for the confounding effect arising from the structural reduction in levels across periods. Accordingly, for all subsequent analyses, the dataset was restricted to curricular levels 7 through 13. This decision is justified because these levels concentrate the largest proportion of students enrolled in courses belonging to the clinical sciences and hospital practice phases, which constitute the core training stage of the medical program and present an academic structure that differs substantially from that of the initial curricular levels.

Figure 3 illustrates the hierarchical data structure, where each academic period contains levels; each level includes courses; each course is divided into parallel sections; and each parallel section includes enrolled students.

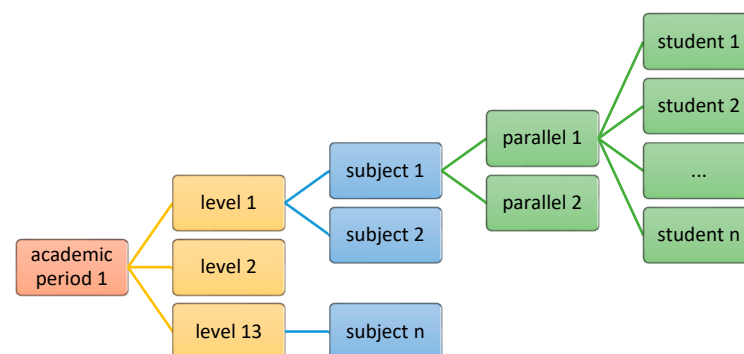


Figure 3. Hierarchical data structure.

The final analytic sample comprises 18,604 records corresponding to 1264 unique students across seven consecutive academic periods (2017–2020). The number of observations per student ranges from 1 to 30. The number of courses spanning levels 7 through 13 is 30. The records in the analytic sample were restricted to students enrolled in curricular levels 7 through 13, while records corresponding to other levels were excluded due to the structural reduction in levels across academic periods. There are no missing values. Furthermore, the dataset does not include course-level, section-level, or instructor-level identifiers; the student identifier (COD_EST) is the only grouping variable available. Table 1 presents a summary of the data selected for the final sample.

Table 1. Summary of data selected for the final sample.

PE_ORD	Period	Number of Records per Academic Period	Number of Students per Academic Period	Number of Male Students	Number of Female Students	Number of Underperforming Students
0	2017–2017	2623	677	256	421	212
1	2017–2018	2791	682	261	421	197
2	2018–2018	2617	632	231	401	207
3	2018–2019	2636	637	234	403	217
4	2019–2019	2633	641	239	402	185
5	2019–2020	2675	680	263	417	184
6	2020–2020	2629	683	267	416	52

Note 1: “Number of records per academic period” refers to database records restricted to levels 7–13. Note 2: The restriction to levels 7–13 is an analytic decision motivated by the structural reduction in levels across academic periods, not a missing-data exclusion; no records were removed due to missingness.

As shown in Table 1, the dataset includes the academic period (PE_ORD), the period description, the number of database records per academic period, the number of students enrolled per academic period from level 7 to level 13, the number of students enrolled by gender, and the number of underperforming students according to the institutional threshold defined for low performance ($SUM < 14$ points).

Figure 4 presents the distribution of academic performance by academic period, represented through box plots. In general, median academic performance levels are relatively consistent across periods, typically falling within a medium-to-high range, suggesting overall stability in academic performance over time.

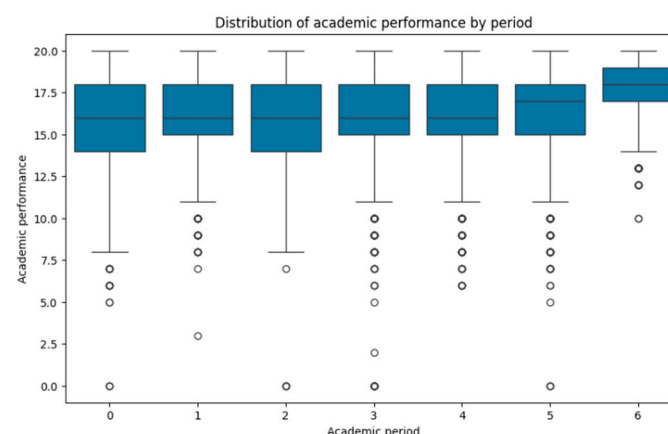


Figure 4. Distribution of academic performance by academic period.

4.3. Data Preprocessing

At this stage, feature engineering and variable transformation were performed to improve data quality and prepare the variables for inferential analysis and predictive modeling. Feature engineering addressed the creation of the following variables:

Province of origin (COD_PRO): This variable reflects the territorial context without requiring the collection of additional data. It is derived from the national identification number, whose first two digits indicate the student's province of origin.

Age (ED_EST): Calculated from the date of birth. A logarithmic transformation was applied because the variable initially exhibited a positive asymmetric distribution (Shapiro–Wilk $p < 0.001$). The logarithmic transformation reduces skewness and brings the variable closer to normality, thereby satisfying the regression model assumptions [36]. The variables (ED_EST) and (COD_PRO) correspond to pre-calculated institutional indicators and are part of the anonymized database.

First partial examination grade (N1_C): Rescaled to a ten-point scale using the linear transformation $G_{10} = (G_8 \times 10)/8$, which ensures comparability with N2 and enables the construction of the composite academic performance variable SUM.

Academic performance (SUM): The sum of the two partial assessments (N1_C and N2) within the same academic period, yielding a continuous variable on a 0–20-point scale. This variable serves as the dependent variable in the predictive model.

The selection of variable encoding techniques was grounded in each variable. Table 2 summarizes the techniques applied.

Table 2. Summary of techniques used in variable transformation.

Variable	Variable Type	Cardinality	Treatment in the Mixed Linear Model (MLM)	Treatment in Supervised Predictive Models
Academic period (PE_ORD)	Ordinal categorical with hierarchy	7	Numeric covariate (0–6); sequential temporal progression	OrdinalEncoder (consecutive integers: 0, 1, 2, ...)
Curricular level (COD_NI)	Ordinal categorical with hierarchy	7	Numeric covariate (7–13); linear progression justified by sequential curriculum	OrdinalEncoder (consecutive integers: 0, 1, 2, ...)
Age (ED_EST)	Continuous numeric (asymmetric)	25	Log-transformation to correct positive skewness	FunctionTransformer(np.log1p) and StandardScaler()
Course (ID_MAT)	Nominal categorical without hierarchy	30	Dummy coding C(ID_MAT); 30 binary indicator columns; discarded due to overfitting	OneHotEncoder or Target Encoding
Gender (ID_SEX)	Nominal categorical without hierarchy	2	Dummy coding C(ID_SEX); binary (0/1)	OneHotEncoder (explicit binary dummification)
Province of origin (COD_PRO)	Nominal categorical without hierarchy	22	Dummy coding C(COD_PRO); 22 binary indicator columns	OneHotEncoder (explicit binary dummification)
Student identifier (COD_EST)	Nominal categorical without hierarchy	1264	Grouping variable	Grouping variable
Academic performance (SUM)	Target variable	—	Dependent variable	Target variable

As shown in Table 2, the academic period (PE_ORD) and curricular level (COD_NI) variables possess a hierarchy, which motivated an ordinal categorical transformation using the Pandas function (.astype('category').cat.codes) for the MLM, while OrdinalEncoder was used for the predictive models in order to respect the ordinal scale and the curricular progression sequence in tree-based models.

The EDA revealed that the age variable (ED_EST) exhibits positive skewness and lacks a Gaussian distribution; therefore, to ensure numerical stability in distribution-sensitive

estimators, a logarithmic transformation was applied for both the MLM and the predictive models. Additionally, a StandardScaler() transformation was applied to this variable for use in the predictive models.

The course (ID_MAT), province of origin (COD_PRO), and gender (ID_SEX) variables, being nominal in nature, were subjected to indexed coding (Label Encoding) for the MLM, while OneHotEncoder was used for the predictive models, transforming each category into an independent binary column and eliminating any numerical hierarchy assumption that could introduce bias in the model's linear estimators [8,11].

The student identifier variable (COD_EST) is used as a grouping variable in the MLM, while in the predictive models it is excluded from the feature space to prevent overfitting through memorization due to its high cardinality.

The variable (SUM) is the system's metric output dimension, providing a continuous performance scale evaluated in common by all estimators.

The preprocessing for the predictive framework was encapsulated in a sklearn.Pipeline using ColumnTransformer. This architecture ensures that scaling and encoding are fitted exclusively on the training partition of each fold, preventing test set contamination [11].

4.4. Inferential Framework: Three-Level Mixed Linear Model

For the inferential analysis, the following variables were selected: academic period (PE_ORD), curricular level (COD_NI), age (ED_EST), province of origin (COD_PRO), and gender (ID_SEX), based on their nature and structure. The course variable (ID_MAT) was not considered due to the structural confounding between the course variance component and the fixed effect of curricular level, as well as the overparameterization of the random-effects structure. Accordingly, applying the parsimony criterion of Bates et al. [39]—which holds that models with lower structural complexity tend to exhibit greater generalization capacity and lower risk of overfitting [40]—the model was simplified to a single random effect (COD_EST), yielding more robust estimates.

The full model specification is:

$$Y_{ij} = \beta_0 + \beta_1 \cdot \text{Period}_i + \beta_2 \cdot \text{Level}_i + \beta_3 \cdot \text{Sex}_j + \beta_4 \cdot \text{Origin}_j + \beta_5 \cdot \log(\text{Age}_j) + u_j + \varepsilon_{ij} \quad (1)$$

where Y_{ij} is the academic performance of student j in period i , β_0 is the global intercept, β_1 – β_5 are fixed-effect coefficients for period, level, gender, province of origin, and log-transformed age respectively, $u_j \sim N(0, \sigma^2_u)$ is the random intercept capturing persistent interindividual differences, and $\varepsilon_{ij} \sim N(0, \sigma^2)$ is the residual error. Multicollinearity was assessed using Variance Inflation Factors (VIF threshold < 5). Residual diagnostics confirmed approximate normality (Q-Q plots and Shapiro–Wilk test) and the absence of systematic heteroscedasticity patterns. Effect size was estimated using marginal and conditional R^2 following Nakagawa and Schielzeth [13], and semi-partial R^2 coefficients were computed for each fixed predictor to quantify independent contributions to outcome variance.

4.5. Predictive Framework: Supervised Regression Process

4.5.1. Definition of the Prediction Problem

The EDA enabled the characterization of variables, from which a supervised regression problem was formulated. A variable mapping and segmentation strategy was defined for the preprocessing pipeline, structured according to the nature and cardinality of the data. The core data preprocessing architecture was then built using advanced Scikit-Learn tools version 1.8.0 (ColumnTransformer and Pipeline).

The predictor matrix X comprises the following variables: PE_ORD (ordinal), COD_NI (ordinal), log-age np.log1p(ED_EST) (continuous), ID_SEX (nominal, One-Hot encoding), and COD_PRO (nominal, One-Hot encoding). The course variable (ID_MAT) was included in the final predictive model as a nominal variable with 30 unique categories corresponding

to the courses spanning curricular levels 7 through 13 and was encoded using One-Hot encoding. The target variable y is SUM (continuous, 0–20 scale). The student identifier COD_EST is used exclusively as the grouping variable for GroupKFold cross-validation and is excluded from the predictor matrix X to prevent identity-based memorization and the overfitting associated with its high cardinality (1264 unique categories).

4.5.2. Hyperparameters and Software Configuration

Hyperparameter configuration was performed through a manual grid search guided by values documented in the reference literature for comparable regression tasks [4,18,40]. Table 3 presents the complete technical specification.

Table 3. Algorithm hyperparameter configurations. Software: Python 3.13.11; scikit-learn 1.8.0; xgboost 3.1.3; numpy 2.4.1; pandas 2.3.3; shap 0.42. All stochastic components: random_state = 42.

Algorithm	Hyperparameters	Seed	Tuning Strategy
Ridge	alpha = 1.0; StandardScaler + OneHotEncoder	42	Manual; L2 regularization range from literature
Random Forest	n_estimators = 300; max_depth = 15; min_samples_leaf = 20; n_jobs = -1	42	Manual grid; leaf = 20 prevents overfitting on longitudinal data
XGBoost	n_estimators = 300; max_depth = 6; learning_rate = 0.05; subsample = 0.8; colsample_bytree = 0.8; objective = reg:squarederror	42	Manual grid following the ranges recommended by Chen and Guestrin [18]
MLP	hidden_layer_sizes = (64,32); activation = relu; solver = adam; max_iter = 500	42	Architecture sized to feature dimensionality; Adam adaptive learning rate

All pipelines were encapsulated within sklearn.Pipeline objects with a preprocessing step (StandardScaler for numeric features; OneHotEncoder for categorical features) applied before model fitting. The random_state = 42 seed was set consistently across all stochastic components.

4.5.3. Validation Strategy

GroupKFold cross-validation ($k = 5$) was used, with the student identifier (COD_EST) as the grouping variable to prevent data leakage. This strategy ensured that records corresponding to the same student were not included simultaneously in the training and test sets [41], providing a realistic estimate of model performance on previously unseen students. MAE, RMSE, and R^2 are recorded at the fold level; means and standard deviations across the five folds are reported.

4.5.4. Risk Alert Threshold and Classification Metrics

Continuous Random Forest algorithm predictions are converted into binary risk alerts using the institutional threshold of SUM < 14 points. Sensitivity, specificity, and AUC-ROC are computed from cross-validated predictions alongside the regression metrics.

4.6. SHAP Explainability Framework

In the testing phase, inferential SHAP extracts the fixed-effect parameters, which are the equivalents of SHAP values in linear modeling, and produces a bar chart showing the direct impact of variables on the mixed linear model.

Predictive SHAP extracts the estimators and preprocessors from the pipeline and processes the data matrix to obtain a strategic sample of 500 rows, computes the SHAP attribution values, and reconstructs the SHAP explanation object. The results are presented in a SHAP beeswarm plot, which provides a global explanation, consolidating the impact of all variables across the entire student population of the analytic sample into a single view, ranking features from top to bottom according to their importance.

In the deployment phase, SHAP waterfall plots are used to show the academic tutor which specific factor is pushing a student toward pedagogical risk. SHAP TreeExplainer [22] was applied to the trained Random Forest model and the mixed linear model. Global feature importance is quantified using mean absolute SHAP values.

4.7. Validation Framework: Robustness and Subgroup Equity Analysis

This framework performs robustness and subgroup equity analysis to ensure that the winning model is reliable across the entire student population. Predictive robustness was assessed through error decomposition stratified by period and by demographic and geographic subgroups. The RMSE/MAE ratio serves as an indicator of error structure: values below $\sqrt{2} \approx 1.414$ indicate predominantly random error consistent with a normal residual distribution; values substantially above $\sqrt{2}$ indicate the presence of systematic bias [24].

5. Results

5.1. Inferential Results: Mixed Linear Model

The results of the mixed linear model (MLM) are presented in Table 4. The analysis of fixed effects reveals that both academic period (PE_ORD) and curricular level (COD_NI) are critical predictors of academic performance, even after controlling for intra-individual variability through random effects (student identifier).

Table 4. Mixed linear model specification, fixed-effect coefficients, significance tests, and semi-partial R^2 effect sizes.

Panel A—Fixed Effects							
Variable	Coeff.	SE	z	p-Value	95% CI	R ² sp	Interpretation
Curricular level (COD_NI)	0.577	0.016	36.754	<0.001	[0.546, 0.608]	0.0440	Moderate
Academic period (PE_ORD)	0.255	0.012	22.089	<0.001	[0.233, 0.278]	0.0257	Small–moderate
Gender (ID_SEX) [T.1]	−0.184	0.078	−2.368	0.018	[−0.336, −0.032]	0.0002	Statistically significant; no practical effect
Log-age (np.log1p(ED_EST))	−2.706	0.413	−6.555	<0.001	[−3.515, −1.897]	0.0013	Statistically significant; no practical effect
Province (COD_PRO) omnibus	Varied	Varied	Varied	<0.001	See Supplementary Material	0.0009	Statistically significant; no practical effect
Intercept	24.060	1.617	14.880	<0.001	[20.891, 27.229]	—	—
Panel B—Effect size [13]							
Metric	Value	Interpretation					
Marginal R ²	0.297	Variance explained by observable fixed predictors					
Conditional R ²	0.473	Full model variance including between-student differences					
Panel C—Random effects and model fit.							
Component	Value						
Random-effect variance (σ ² _u)	1.337	0.039					
Residual variance (σ ² ε)—Scale	4.058	—					
ICC	0.176	Confirms hierarchical structure; MLM justified					

Note 1: Model specification: $SUM \sim C(ID_SEX) + C(COD_PRO) + PE_ORD + COD_NI + np.log1p(ED_EST)$. Method: REML. $n = 18,604$ observations; 1264 student groups. Note 2: Province coefficients (COD_PRO [T.1]–[T.21]): all negative relative to the reference province (COD_PRO = 5, university province); full coefficient table in Supplementary Material. Omnibus significance: $p < 0.001$. Note 3: R^2_{sp} = semi-partial R^2 (unique variance explained by each predictor after partialing shared variance). Shaded cells highlight dominant predictors. SE = standard error. Note 4: Reference categories: ID_SEX [T.1] = Male (reference = Female, coded 0); COD_PRO [T.k] = all provinces relative to COD_PRO = 5. Note 5: ICC = 0.176 confirms that 17.6% of total SUM variance is attributable to persistent between-student differences, empirically justifying the hierarchical modeling framework.

The marginal $R^2 = 0.297$ indicates that fixed effects explain 29.7% of the total SUM variance. The conditional $R^2 = 0.473$ indicates that the full model explains 47.3%, with random effects capturing an additional 17.6% attributable to persistent individual differences. Table 4 presents the semi-partial R^2 decomposition.

Curricular level (COD_NI) is the dominant structural predictor ($R^2_{sp} = 0.044$, moderate effect; $\beta = 0.577$, 95% CI: [0.546, 0.608]), indicating that each additional curricular level is associated with a 0.577-point increase in SUM. Academic period (PE_ORD) contributes a small-to-moderate unique effect ($R^2_{sp} = 0.026$; $\beta = 0.255$, 95% CI: [0.233, 0.278]). Gender, province, and log-age exhibit statistically significant coefficients ($p \leq 0.018$) but negligible practical effect sizes ($R^2_{sp} < 0.002$ each) [1,37].

5.2. Predictive Results: Algorithm Comparison

Regarding the supervised learning framework, results indicate that ensemble-based models consistently outperformed linear and neural approaches. Cross-validated performance metrics with fold-level standard deviations are presented in Table 5.

Table 5. Cross-validated predictive performance.

Algorithm	MAE (Mean $\pm$ SD)	RMSE (Mean $\pm$ SD)	R^2 (Mean $\pm$ SD)	Key Hyperparameters
Mean baseline (null)	2.41 $\pm$ 0.09	3.02 $\pm$ 0.11	0.000 $\pm$ 0.01	—
Ridge Regression	1.501 $\pm$ 0.06	2.009 $\pm$ 0.08	0.382 $\pm$ 0.03	alpha = 1.0
MLP	1.397 $\pm$ 0.05	1.873 $\pm$ 0.07	0.463 $\pm$ 0.03	(64, 32), relu, adam
XGBoost	1.276 $\pm$ 0.04	1.739 $\pm$ 0.06	0.537 $\pm$ 0.02	n = 300, depth = 6, lr = 0.05
Random Forest	1.267 $\pm$ 0.04	1.714 $\pm$ 0.05	0.551 $\pm$ 0.02	n = 300, depth = 15, leaf = 20

Random Forest achieved the best results (MAE = 1.267 $\pm$ 0.04, RMSE = 1.714 $\pm$ 0.05, $R^2 = 0.551 \pm 0.02$), outperforming XGBoost ($R^2 = 0.537$). Applying the institutional low-performance threshold—which defines any assessment below 14/20 as underperforming—the model achieved sensitivity = 0.532, specificity = 0.940, and AUC-ROC = 0.903.

The sensitivity of 0.532 indicates that approximately half of the students who will ultimately fall below the SUM threshold are not prospectively identified by the system from sociodemographic and curricular variables alone.

The high specificity indicates very few false alarms, a critical operational property for tutoring support tools where alert fatigue is a risk. The observed versus predicted scatter and the residual distribution are presented in Figure 5a,b.

The distribution of the residuals (errors) exhibits a bell-shaped curve centered on zero, indicating that the uncertainty is random rather than systematic. This suggests that the model maintains high accuracy across the majority of samples, minimizing the presence of significant outlier errors in the prediction of academic performance.

Figure 6 presents two plots. Figure 6a corresponds to the residuals versus fitted values, which evaluates the assumptions of homoscedasticity (constant error variance) and model linearity. Figure 6b is a model calibration plot, which assesses how well the estimated grades align with actual performance across different performance ranges.

The residuals plot confirms the absence of systematic bias in the central tendency (mean error close to zero). Likewise, the calibration plot evidences a high correspondence between predicted scores and observed performance, showing a robust and linear fit across the entire grade spectrum, particularly near the institutional passing threshold (SUM < 14). Notable deviations are isolated at actual values of zero, associated with atypical phenomena or dropout that fall outside the variability explained by the structural factors evaluated.

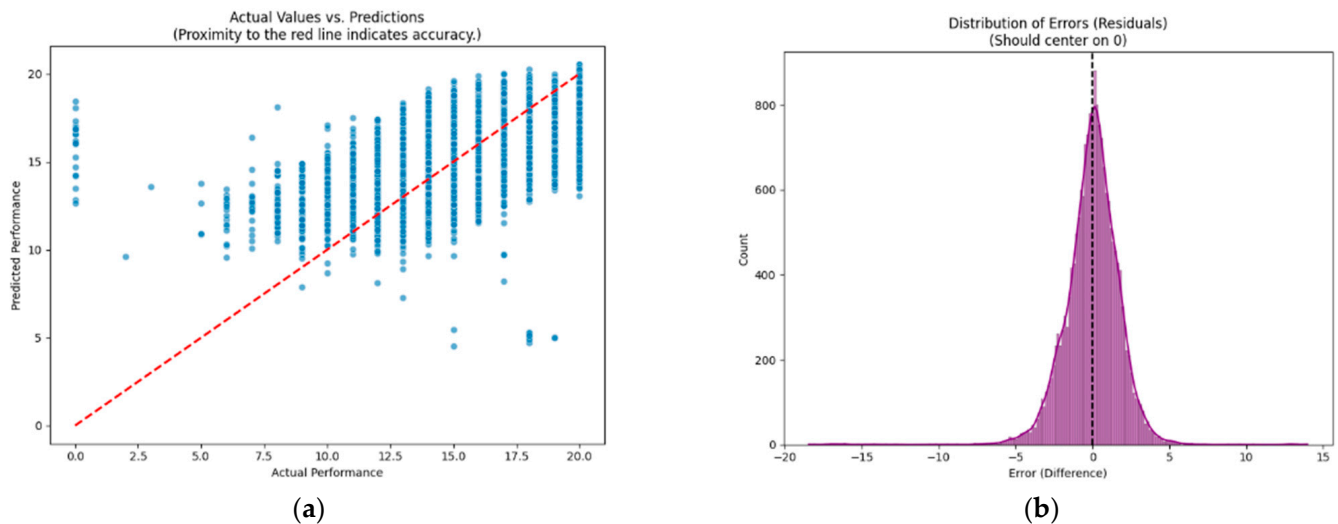


Figure 5. Empirical Fit and Residual Error Symmetry Charts (a) Scatter plot: Actual performance versus predicted performance (blue dots; red dashed line = perfect prediction). Vertical scatter reflects the discrete nature of SUM scores. (b) Residual distribution: histogram centered on 0 with near-normal distribution, confirming the absence of systematic directional bias.

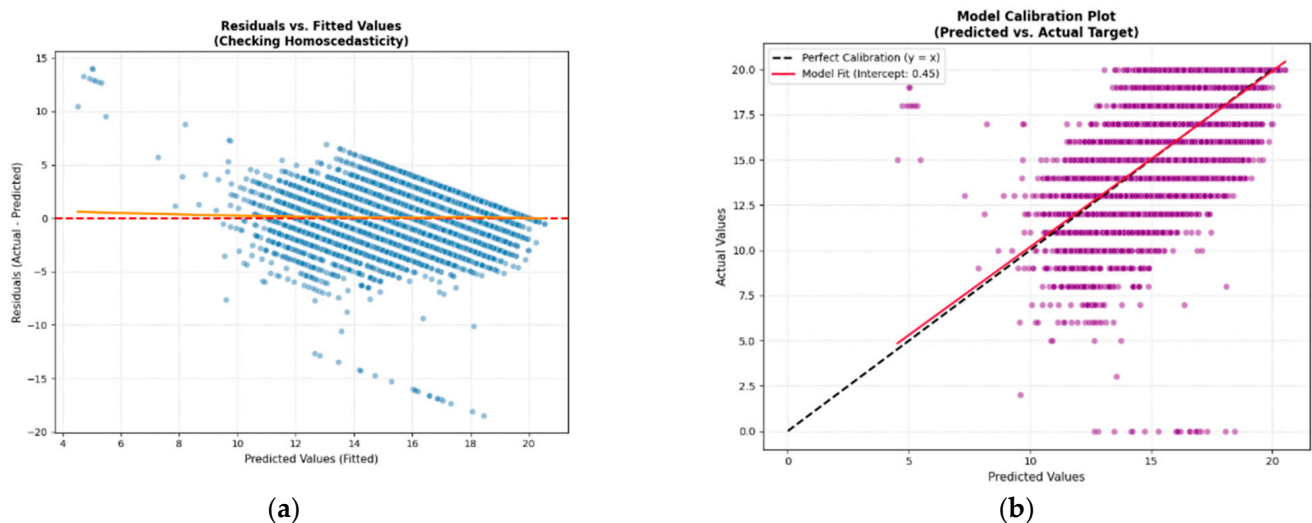


Figure 6. Error Diagnostic and Alert Reliability Charts. (a) Residuals versus fitted values. (b) Model calibration plot.

5.3. SHAP Explainability Analysis Results

The SHAP applied to the mixed linear model indicates the impact of variables on academic performance. This is shown in Figure 7, corroborating the results obtained from MLM.

The Predictive SHAP results are shown in Figures 8 and 9. Figure 8 presents the SHAP beeswarm plot and Figure 9 presents the global feature importance bar chart for the implemented Random Forest model. Curricular level (ord_COD_NI) is the dominant predictor variable with a mean $|SHAP| = 1.28$, followed by academic period (ord_PE_ORD, mean $|SHAP| = 0.53$) and log-age (num_ED_EST, mean $|SHAP| = 0.33$). High values of COD_NI generate positive SHAP contributions (higher predicted SUM), while low values generate negative contributions. Province-specific features (cat_mid_ID_MAT_*) contribute heterogeneously with regional variation consistent with the MLM province coefficients. Gender (bin_ID_SEX_1, mean $|SHAP| = 0.06$) has minimal individual-level

impact. The SHAP ranking corroborates the MLM semi-partial R^2 ranking, providing cross-framework validation of predictor importance.

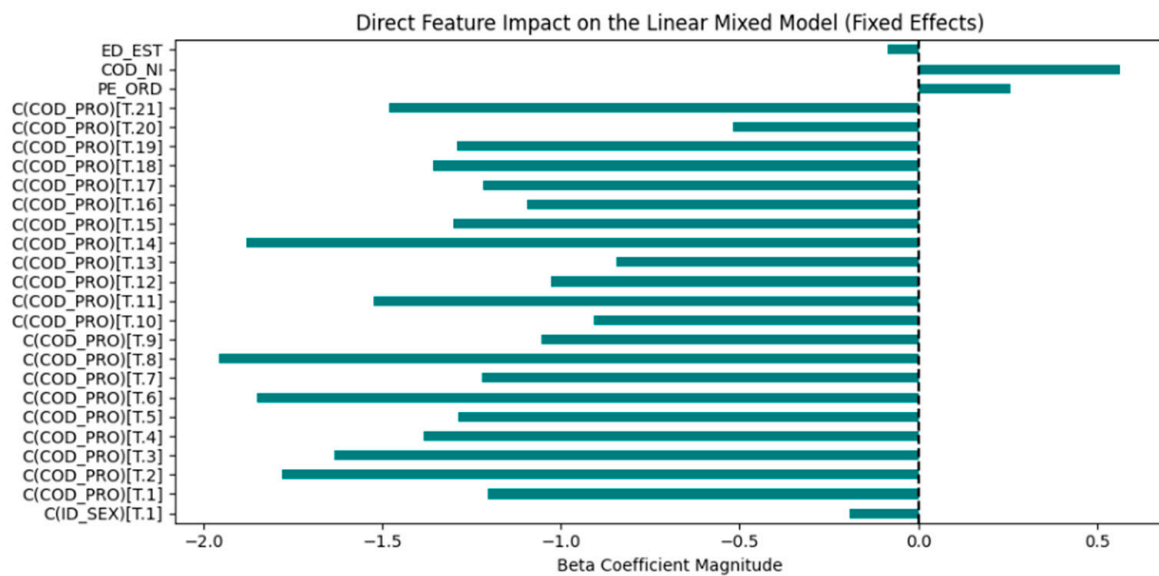


Figure 7. Direct impact of MLM variables on academic performance.

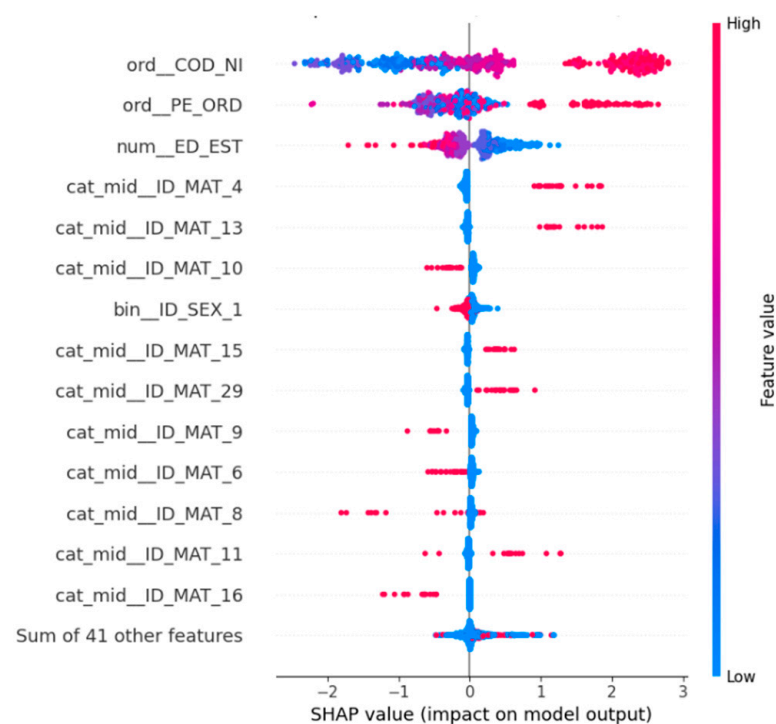


Figure 8. SHAP global explainability analysis: ord_COD_NI dominant (mean |SHAP| = 1.28). Red = high feature value; blue = low feature value.

Global Explainability Analysis (SHAP Beeswarm)—Random Forest.

Table 6 presents the results of the SHAP explanatory attribution matrix obtained for a student with identifier code 1263.

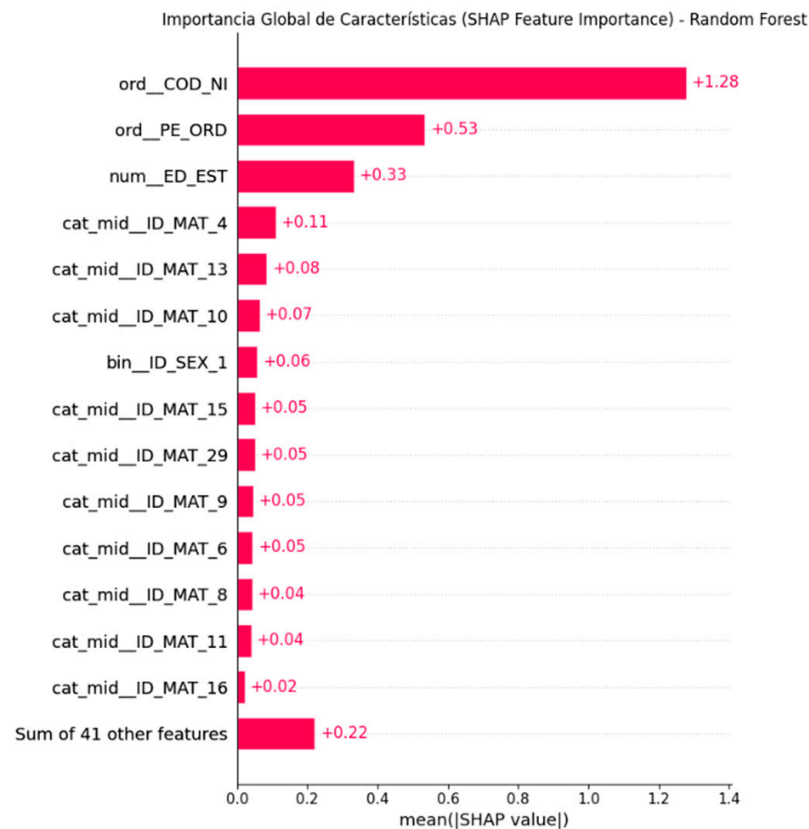


Figure 9. SHAP global feature importance: ord_COD_NI = 1.28; ord_PE_ORD = 0.53; num_ED_EST = 0.33.

Table 6. SHAP explanatory attribution matrix for the illustrative student (COD_EST = 1263). Priority Index = mean (SHAP MLM, SHAP RF). A negative Priority Index indicates a dominant risk-associated factor in the predicted SUM.

Variable	SHAP MLM (Inferential)	SHAP RF (Predictive)	Priority Index (Mean)	Role
ID_PE (period)	0.791	2.080	1.435	Protective
ID_MAT (course)	0.000	1.284	0.642	Protective
ED_EST (log-age)	0.199	0.445	0.322	Protective
COD_PRO (province)	0.181	0.101	0.141	Protective
ID_SEX (gender)	0.077	0.024	0.051	Neutral
COD_NI (level)	−1.952	−0.219	−1.086	High-priority risk factor

The data in Table 6 indicate that curricular level (COD_NI) yields a Priority Index of -1.086 , making it the dominant risk-associated factor. The system automatically generates an operational model-derived summary in natural language at the bottom of the panel: the primary factor depressing this student's grades is COD_NI (Consolidated Score: -1.09), indicating to the tutor that efforts should be concentrated on mitigating the impact of this variable. This directive identifies the dominant risk-associated factor and recommends a specific focus for the tutorial effort, constituting the operationalized output of the entire dual-framework process and representing the primary interface between the intelligent system and the faculty tutor.

5.4. Robustness and Subgroup Equity Analysis Results

5.4.1. Period-Stratified Error Decomposition

The results of the period-stratified error decomposition are presented in Table 7.

Table 7. Error decomposition by academic period—Random Forest (5-fold GroupKFold). RMSE/MAE < $\sqrt{2} = 1.414$ indicates predominantly random error [31].

PE_ORD	Period	MAE	RMSE	Mean Error	RMSE/MAE	Error Status
0	2017–2017	1.277	1.716	0.001	1.343	Random
1	2017–2018	1.179	1.524	0.007	1.292	Random
2	2018–2018	1.325	1.727	0.002	1.303	Random
3	2018–2019	1.327	1.837	−0.039	1.384	Random
4	2019–2019	1.147	1.528	0.039	1.332	Random
5	2019–2020	1.169	1.602	−0.018	1.370	Random
6	2020–2020	0.976	1.342	0.005	1.375	Random

The RMSE/MAE ratio ranges from 1.29 to 1.38 across the seven periods, remaining consistently below $\sqrt{2} = 1.414$, confirming predominantly random predictive uncertainty throughout the 2017–2020 period [24]. Mean error values remain close to zero (range: −0.039 to +0.039), confirming the absence of directional bias. MAE decreases monotonically from 1.277 (PE_ORD 0) to 0.976 (PE_ORD 6), reflecting progressive cohort homogenization.

5.4.2. Equity Analysis by Province of Origin

Table 8 presents the MAE parity ratios for the 22 province categories relative to the global MAE of 1.267.

Table 8. MAE parity ratio by province of origin (COD_PRO). A threshold ≤ 1.25 indicates equitable predictive performance [29].

COD_PRO	MAE	Parity Ratio	Equity Status (≤ 1.25)
Global	1.267	1.000 (reference)	—
0	1.073	0.847	Equitable
1	1.201	0.948	Equitable
2	1.189	0.938	Equitable
3	1.324	1.045	Equitable
4	1.231	0.971	Equitable
5	1.257	0.992	Equitable
6	1.201	0.948	Equitable
7	1.237	0.976	Equitable
8	1.453	1.147	Equitable
9	1.295	1.022	Equitable
10	1.509	1.191	Equitable
11	1.353	1.068	Equitable
12	1.067	0.842	Equitable
13	1.403	1.107	Equitable
14	1.411	1.114	Equitable
15	1.444	1.140	Equitable
16	1.288	1.017	Equitable
17	1.321	1.043	Equitable
18	1.768	1.395	Elevated uncertainty
19	1.260	0.994	Equitable
20	1.176	0.928	Equitable
21	1.270	1.003	Equitable

As shown in Table 8, 21 of the 22 province categories achieve MAE parity ratios ≤ 1.25 , confirming equitable predictive performance across the majority of geographic subgroups. Province COD_PRO 18 exhibits an elevated ratio of 1.395, approaching but not exceeding the systematic bias threshold of $\sqrt{2}$. Figure 10 shows that the disparity in COD_PRO 18 is driven by outlier predictions rather than systematic bias across all students from that province.

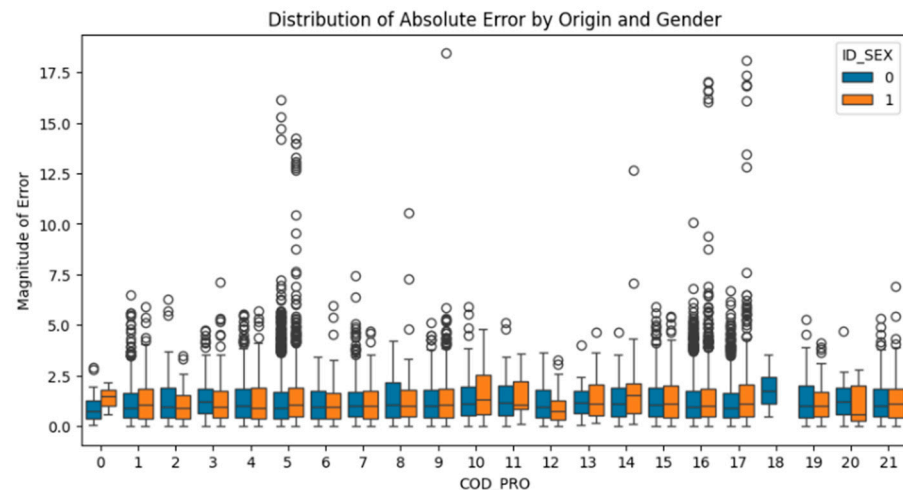


Figure 10. Box plot: Distribution of absolute error by province of origin (COD_PRO) and gender. Provinces COD_PRO 5, 16, and 17 exhibit elevated outliers; gender parity is maintained across all provinces.

5.4.3. Equity Analysis by Age Range

Students aged 18–22 years (parity ratio = 1.090) and 22–30 years (0.969) fall within the equitable range. Students aged 30–50 years exhibit a ratio of 1.375, close to $\sqrt{2} = 1.414$ but not exceeding it. This group likely includes non-traditional students with less predictable administrative records, justifying the future integration of LMS behavioral data. The results are presented in Table 9.

Table 9. MAE parity ratio by age range. Threshold ≤ 1.25 [30].

Age Range	MAE	Parity Ratio	Equity Status
18–22 years	1.381	1.090	Equitable
22–30 years	1.227	0.969	Equitable
30–50 years	1.742	1.375	Elevated uncertainty (near threshold)
.	1.267	1.000	Reference

5.4.4. Gender Equity Analysis

Gender MAE parity ratios of 1.03 (female) and 1.06 (male) fall within the ≤ 1.25 threshold [25], confirming equitable model performance across genders, consistent with the negligible MLM semi-partial R^2 for gender ($R^2_{sp} = 0.0002$) and the SHAP contribution (mean $|SHAP| = 0.06$).

6. Discussion

This study presents the design and implementation of a single-institution intelligent diagnostic system to identify low mid-period academic performance. This system integrates an inferential analysis framework based on a mixed linear model, a predictive analysis framework based on Random Forest, a validation framework based on robustness and subgroup equity analysis, a dual SHAP explainability framework, and a Streamlit-based web prototype. The system is intended to assist faculty tutors in analyzing academic performance by using the first two grades of the same academic period, summing them ($SUM = N1_C + N2$), and taking this result as an intermediate checkpoint to transform performance data into pedagogical inputs that activate proactive and preventive tutoring before the final assessment of an academic period. The system addresses four gaps identified in the literature: the absence of longitudinal hierarchical modeling in intelligent diagnostic systems; the misdirection of these systems' outputs toward managers rather than

tutors; the lack of transparent preprocessing reporting; and the absence of a dual SHAP explanatory interface in any prior intelligent diagnostic system for medical education.

The most methodologically significant analytical finding is the material impact that appropriate categorical variable encoding has on the performance of both inferential and predictive models, empirically demonstrating that exploratory data analysis and preprocessing principles directly affect the reliability and reproducibility of a research study.

The predominance of curricular level as the structural predictor—confirmed by both the MLM ($R^2_{sp} = 0.044$; $\beta = 0.577$) and SHAP explainability (mean $|SHAP| = 1.28$)—reflects the progressive academic specialization of the clinical curriculum. Crucially, this finding is corroborated by the SHAP explanatory attribution matrix in the Streamlit prototype, where curricular level (COD_NI) yields the most negative Priority Index (-1.086) among all selected variables, confirming that the population-level structural finding translates directly into individual-level clinical relevance for the tutor. This cross-framework, cross-scale corroboration—from population inference (MLM semi-partial R^2) through individual prediction (RF SHAP) to operational model-derived summary (Explanatory Attribution Matrix)—is a methodological contribution that is not present in any prior intelligent diagnostic system study.

The formal equity analysis of subgroups addresses a gap in the current literature on intelligent diagnostic systems [9,25]. Of the 22 provincial subgroups, 21 achieve MAE parity ratios ≤ 1.25 . The near-threshold ratio (1.395) corresponding to province code 18 (COD_PRO = 18) and the ratio (1.375) corresponding to students aged between 30 and 50 years indicate elevated predictive uncertainty for these groups, likely reflecting latent structural factors not captured by the variables analyzed in this study. These findings warrant targeted qualitative follow-up rather than disqualification of the model.

The Streamlit-based web prototype constitutes the primary technological contribution that goes beyond standard medical education analytics. The SHAP Explanatory Attribution Matrix computes a Priority Index per variable that synthesizes population-level structural effects (MLM) with individual nonlinear prediction effects (RF) and generates an operational model-derived summary in natural language—making it the most operationally original component of this work, one that is not present in any prior intelligent diagnostic system study in the literature.

7. Limitations and Future Work

Several limitations bound the current contribution. First, data originate from a single institution in Ecuador; multi-institutional validation is the primary next research step. Second, the dataset lacks course-level, section-level, and instructor-level identifiers, preventing more complete hierarchical modeling. Third, temporal validation—training on earlier periods and testing on later periods—was not implemented, and constitutes a priority for the next development phase. Fourth, formal automated hyperparameter optimization was not conducted. Fifth, the sensitivity of 0.532 indicates that approximately half of the students who will ultimately fall below the institutional SUM threshold are not prospectively identified by the system from sociodemographic and curricular variables alone. This reflects an inherent limitation due to the lack of behavioral data such as LMS interaction data or psychometric instruments. The system is therefore designed to complement, not replace, the faculty tutor's direct academic monitoring, and should not be interpreted as a comprehensive safety net. Sixth, the Streamlit prototype was deployed locally; production deployment within an institutional network has not been conducted, constituting the next validation step. Seventh, while the prototype was implemented in a local development environment sufficient for prototype validation, a future institutional deployment will demand computational resources associated with data processing and the

computation of SHAP attribution values, particularly if the system is intended to operate with multiple concurrent users across the entire institutional student population. Eighth, the predictive model included the course as a predictor variable; however, a systematic comparison of model performance with and without the course variable was not conducted and constitutes a line of future work.

Future work will: (1) implement temporal cross-validation (train on 2017–2018, test on 2019–2020); (2) conduct multi-institutional validation across Ecuadorian medical programs; (3) integrate LMS behavioral and psychometric data to improve sensitivity and reduce parity ratios for near-threshold subgroups; (4) deploy the Streamlit interface within the institutional server and conduct a prospective controlled study measuring its impact on early intervention rates; (5) apply formal automated hyperparameter tuning within the GroupKFold structure; (6) extend the SHAP Explanatory Attribution Matrix to support real-time streaming updates as new partial grade data become available during the academic period; and (7) implement caching mechanisms for predictions and pre-computed per-student SHAP values for potential migration toward a cloud computing infrastructure with horizontal scaling capacity.

8. Conclusions

This paper presents the design and implementation of a single-institution intelligent diagnostic system to identify low mid-period academic performance with the aim of activating proactive and preventive tutoring. The system is intended to assist faculty tutors in analyzing academic performance in medical education. The contribution of this work is articulated across seven dimensions with direct empirical support.

First, the system uses the first two partial grades of the same academic period as an intermediate checkpoint, transforming academic, sociodemographic, and academic performance data into pedagogical inputs that activate proactive and preventive tutoring before academic lag consolidates at the end of the period, constituting an intervention window with greater corrective potential than the reactive mechanisms documented in the literature.

Second, the integration of a mixed linear model (marginal $R^2 = 0.297$; conditional $R^2 = 0.473$; ICC = 0.176) with a Random Forest regression pipeline that produces individual risk scores ($MAE = 1.267 \pm 0.04$; $R^2 = 0.551 \pm 0.02$) simultaneously addresses the inferential question of which factors matter and by how much, and the predictive question of how this specific student will perform—a combination not documented in any prior intelligent diagnostic system for medical education.

Third, GroupKFold cross-validation grouped by student identifier (COD_EST) ensures that records from the same student never appear simultaneously in training and test sets, producing realistic generalization estimates for unseen students and avoiding the performance overestimation that affects studies by omitting this grouping strategy.

Fourth, the implementation of the dual SHAP explainability framework—at the population level through MLM coefficients and at the individual level through SHAP TreeExplainer on Random Forest—produces interpretable low-performance profiles that the faculty tutor can directly use: curricular level emerges as the dominant predictor in both the inferential analysis ($R^2_{sp} = 0.044$; $\beta = 0.577$) and the predictive analysis (mean $|SHAP| = 1.28$), providing cross-framework coherence that strengthens the validity of the findings.

Fifth, the robustness and subgroup equity analysis confirms that predictive uncertainty is predominantly random across all academic periods (RMSE/MAE between 1.29 and 1.38, consistently below $\sqrt{2} = 1.414$) and that the system produces equitable estimates across 21 of 22 provinces and both genders (parity ratios ≤ 1.25), with the exception of province

code 18 and students aged over 30 years, which identify priority areas for data enrichment prior to institutional deployment.

Sixth, the exploratory data analysis enables adequate characterization of variables, revealing that this analytical decision materially affects the results and conclusions of the study, with direct implications for reproducibility in educational machine learning research.

Finally, the Streamlit-based web prototype demonstrates the operational feasibility of the system by integrating the dual MLM-Random Forest framework into an interface accessible to faculty tutors without technical training, whose most innovative element is the SHAP Explanatory Attribution Matrix, which synthesizes the Shapley values from both analytical layers into a Priority Index per variable and automatically generates a model-derived explanation in natural language that transforms the numerical prediction into a concrete tutorial action—functionality absent from all intelligent diagnostic systems for medical education documented in the literature, and representing the central technological contribution of this work.

Supplementary Materials: A data dictionary and a synthetic dataset sufficient to reproduce the presented results can be found in the following repository: <https://github.com/magyta/PrototypeEWS.git>. Computing environment: Python 3.13.11; Pandas 2.3.3; NumPy 2.4.1; Scikit-learn 1.8.0; XGBoost 3.1.3; SHAP 0.42; Streamlit 1.54.0; Statsmodels: 0.14.6.

Author Contributions: Conceptualization, M.A., A.G.-B. and P.C.; methodology, M.A. and A.G.-B.; software, M.A.; validation, A.G.-B. and P.C.; formal analysis, M.A.; investigation, M.A.; resources, A.G.-B.; data curation, M.A.; writing—original draft preparation, M.A.; writing—review and editing, A.G.-B. and P.C.; visualization, M.A.; supervision, A.G.-B. and P.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset consists of anonymized secondary records provided under a formal institutional data confidentiality agreement. No personal information was accessed or processed. The study is exempt from research ethics committee approval under Article 2 of Ecuador's Organic Law on Personal Data Protection (2021), which states that analyses of anonymized data do not require approval by a Human Research Ethics Committee (CEISH). The risk alerts generated by the system are intended solely to support faculty mentoring decisions and are not designed to automate, replace, or justify adverse academic decisions, such as course suspension, academic probation, or program expulsion. The analysis code, source files, and implemented prototype are publicly available at: GitHub repository: <https://github.com/magyta/PrototypeEWS.git>, accessed on 1 June 2026. Implemented Streamlit prototype: <https://prototypeews-srcpuzg4c8otwfazefmj2q.streamlit.app/>, accessed on 1 June 2026.

Acknowledgments: During the preparation of this manuscript, the authors used ChatGPT version 5.5 and Gemini version 1.5 (free versions) to improve the writing quality. The authors have reviewed and edited the output and assumed full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Eames, D.; Thomas, S.; Norman, K.; Simanton, E.; Weisman, A. Sociodemographic disadvantage in the burden of stress and academic performance in medical school: Implications for diversity in medicine. *BMC Med. Educ.* **2024**, *24*, 348. [CrossRef] [PubMed]
2. Mabizela, S.E.; Bruce, J. Investigating the risk factors for academic difficulties in the medical programme at a South African university. *BMC Med. Educ.* **2022**, *22*, 208. [CrossRef] [PubMed]
3. Nawa, N.; Numasawa, M.; Nakagawa, M.; Sunaga, M.; Fujiwara, T.; Tanaka, Y.; Kinoshita, A. Associations between demographic factors and the academic trajectories of medical students in Japan. *PLoS ONE* **2020**, *15*, e0233371. [CrossRef] [PubMed]

4. Mastour, H.; Dehghani, T.; Moradi, E.; Eslami, S. Explainable artificial intelligence for predicting medical students' performance in comprehensive assessments. *Sci. Rep.* **2025**, *15*, 23752. [CrossRef] [PubMed]
5. SENESCYT. "Oferta Académica UEP-ISTT," Servicios Senescyt. Available online: <https://siau.senescyt.gob.ec/oferta-academica-uep-istt/> (accessed on 24 February 2025).
6. Consejo de Educación Superior. Reglamento de Régimen Académico. Available online: <https://www.ces.gob.ec/wp-content/uploads/2022/08/Reglamento-de-Re%CC%81gimen-Acade%CC%81mico-vigente-a-partir-del-16-de-septiembre-de-2022.pdf> (accessed on 1 June 2026).
7. Namoun, A.; Alshanqiti, A. Predicting Student Performance Using Data Mining and Learning Analytics Techniques: A Systematic Literature Review. *Appl. Sci.* **2020**, *11*, 237. [CrossRef]
8. Guanin-Fajardo, J.H.; Gúaña-Moya, J.; Casillas, J. Predicting Academic Success of College Students Using Machine Learning Techniques. *Data* **2024**, *9*, 60. [CrossRef]
9. Asosega, K.; Adebajji, A.O.; Aidoo, E.N.; Owusu-Dabo, E. Application of Hierarchical/Multilevel Models and Quality of Reporting (2010–2020): A Systematic Review. *Sci. World J.* **2024**, *2024*, 1–9. [CrossRef] [PubMed]
10. Al Hashmi, R.A.M.; Ozturk, I.; Elmehdi, H.M. Early detection of at-risk health sciences students: A machine learning-based predictive study using midterm grades. *BMC Med. Educ.* **2025**, *25*, 1651. [CrossRef] [PubMed]
11. Géron, A. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*; O'Reilly Media, Inc.: Tokyo, Japan, 2019.
12. Sullivan, G.M.; Feinn, R. Using Effect Size—Or Why the *p* Value is not Enough. *J. Grad. Med. Educ.* **2012**, *4*, 279–282. [CrossRef] [PubMed]
13. Nakagawa, S.; Schielzeth, H. A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods Ecol. Evol.* **2013**, *4*, 133–142. [CrossRef]
14. Lin, Y.; Kang, Y.J.; Lee, H.J.; Kim, D.-H. Pre-medical students' perceptions of educational environment and their subjective happiness: A comparative study before and after the COVID-19 pandemic. *BMC Med. Educ.* **2021**, *21*, 619. [CrossRef] [PubMed]
15. Zeng, B.; Sun, J.; Wen, H. Analyzing factors associated with student achievement in large-scale educational assessments: A two-stage machine learning approach. *Int. J. Educ. Res.* **2026**, *136*, 102886. [CrossRef]
16. Bilal, M.; Omar, M.; Anwar, W.; Bokhari, R.H.; Choi, G.S. The role of demographic and academic features in a student performance prediction. *Sci. Rep.* **2022**, *12*, 12508. [CrossRef] [PubMed]
17. Albriki, S.; Eid, H.F. Utilizing random forest algorithm for early detection of academic underperformance in open learning environments. *PeerJ Comput. Sci.* **2023**, *9*, e1708. [CrossRef] [PubMed]
18. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM: San Francisco, CA, USA, 2016; pp. 785–794.
19. Suaza-Medina, M.; Peñaña-Niebles, R.; Jubiz-Díaz, M. A model for predicting academic performance on standardised tests for lagging regions based on machine learning and Shapley additive explanations. *Sci. Rep.* **2024**, *14*, 25306. [CrossRef] [PubMed]
20. Ahmed, W.; Wani, M.A.; Plawiak, P.; Meshoul, S.; Mahmoud, A.; Hammad, M. Machine learning-based academic performance prediction with explainability for enhanced decision-making in educational institutions. *Sci. Rep.* **2025**, *15*, 26879. [CrossRef] [PubMed]
21. Vasilopoulos, A.; Matthews, G. Cross-validation Optimal Fold-Number for Model Selection. *Am. J. Undergrad. Res.* **2024**, *21*, 15–29. [CrossRef]
22. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: New York, NY, USA, 2017. Available online: <https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html> (accessed on 27 May 2026).
23. Bifarin, O.O. Interpretable machine learning with tree-based shapley additive explanations: Application to metabolomics datasets for binary classification. *PLoS ONE* **2023**, *18*, e0284315. [CrossRef] [PubMed]
24. Karunasingha, D.S.K. Root mean square error or mean absolute error? Use their ratio as well. *Inf. Sci.* **2022**, *585*, 609–629. [CrossRef]
25. Bellamy, R.K.E.; Dey, K.; Hind, M.; Hoffman, S.C.; Houde, S.; Kannan, K.; Lohia, P.; Martino, J.; Mehta, S.; Mojsilović, A.; et al. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM J. Res. Dev.* **2019**, *63*, 4:1–4:15. [CrossRef]
26. Chicco, D.; Warrens, M.J.; Jurman, G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Comput. Sci.* **2021**, *7*, e623. [CrossRef] [PubMed]
27. Brehm, L.; Alday, P.M. Contrast coding choices in a decade of mixed models. *J. Mem. Lang.* **2022**, *125*, 104334. [CrossRef]
28. Pargent, F.; Pfisterer, F.; Thomas, J.; Bischl, B. Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features. *Comput. Stat.* **2022**, *37*, 2671–2692. [CrossRef]
29. Ng, W.Y.; Wang, L.R.; Liu, S.; Fan, X. Causal SHAP: Feature Attribution with Dependency Awareness through Causal Discovery. In *Proceedings of the 2025 International Joint Conference on Neural Networks (IJCNN)*, Rome, Italy, 30 June–5 July 2025; pp. 1–8.

30. Clarke, P.; Crawford, C.; Steele, F.; Vignoles, A.F. The Choice Between Fixed and Random Effects Models: Some Considerations for Educational Research. *SSRN Electron. J.* **2010**, *36*. [\[CrossRef\]](#)
31. Monteverde-Suárez, D.; González-Flores, P.; Santos-Solórzano, R.; García-Minjares, M.; Zavala-Sierra, I.; De La Luz, V.L.; Sánchez-Mendiola, M. Predicting students' academic progress and related attributes in first-year medical students: An analysis with artificial neural networks and Naïve Bayes. *BMC Med. Educ.* **2024**, *24*, 74. [\[CrossRef\]](#) [\[PubMed\]](#)
32. Mastour, H.; Dehghani, T.; Moradi, E.; Eslami, S. Early prediction of medical students' performance in high-stakes examinations using machine learning approaches. *Heliyon* **2023**, *9*, e18248. [\[CrossRef\]](#) [\[PubMed\]](#)
33. Islam, M.M.; Sojib, F.H.; Mihad, M.F.; Hasan, M.; Rahman, M. The integration of explainable AI in Educational Data Mining for student academic performance prediction and support system. *Telemat. Inform. Rep.* **2025**, *18*, 100203. [\[CrossRef\]](#)
34. Bañeres, D.; Rodríguez-González, M.E.; Guerrero-Roldán, A.-E.; Cortadas, P. An early warning system to identify and intervene online dropout learners. *Int. J. Educ. Technol. High. Educ.* **2023**, *20*, 3. [\[CrossRef\]](#)
35. Abukader, A.; Alzubi, A.; Adegboye, O.R. Intelligent System for Student Performance Prediction: An Educational Data Mining Approach Using Metaheuristic-Optimized LightGBM with SHAP-Based Learning Analytics. *Appl. Sci.* **2025**, *15*, 10875. [\[CrossRef\]](#)
36. Pozo Burgos, E.J.; Burbano Pulles, M.R.; Vidal Chica, J.I.; Revelo Salgado, G.E. Sociocultural and demographic factors that influence academic performance: The pre-university case of the Universidad Politécnica Estatal del Carchi. *J. Technol. Sci. Educ.* **2022**, *12*, 147. [\[CrossRef\]](#)
37. Cannistrà, M.; Masci, C.; Ieva, F.; Agasisti, T.; Paganoni, A.M. Early-Predicting Dropout of University Students: An Application of Innovative Multilevel Machine Learning and Statistical Techniques. Available online: <https://www.tandfonline.com/doi/abs/10.1080/03075079.2021.2018415> (accessed on 3 April 2026).
38. Hefny, A.F.; Fathi, M.A.; Mansour, N.A.; Al-Ali, M.A. Early Student Attrition from Medical Schools: A Scoping Review. *Health Prof. Educ.* **2024**, *10*, 7. [\[CrossRef\]](#)
39. Bates, D.; Kliegl, R.; Vasishth, S.; Baayen, H. Parsimonious Mixed Models. *arXiv* **2015**, arXiv:1506.04967. Available online: <https://arxiv.org/abs/1506.04967v2> (accessed on 27 May 2026).
40. Hastie, T.; Tibshirani, R.; Friedman, J. Model Assessment and Selection. In *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Hastie, T., Tibshirani, R., Friedman, J., Eds.; Springer: New York, NY, USA, 2009; pp. 219–259.
41. Bertolini, R.; Finch, S.J.; Nehm, R.H. Enhancing data pipelines for forecasting student performance: Integrating feature selection with cross-validation. *Int. J. Educ. Technol. High. Educ.* **2021**, *18*, 44. [\[CrossRef\]](#) [\[PubMed\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.