Математические заметки СВФУ
Апрель—июнь, 2026. Том 33, № 2

UDC 534-8

Algorithmic Forecasting in Architectural Pedagogy: A Longitudinal, Explainable Machine Learning Framework for Architectural Design Studio Performance

Youness Hajjioui, Youssef Jedidi, Mohamed Benslimane, Abdelali Ibriz, Karim El Moutaouakil

Abstract: The architectural design studio serves as the core of architectural education, characterized by project based learning and subjective evaluation rubrics. Identifying struggling before the semester begins remains a critical challenge for educational institutions. Despite advancements in educational data mining, predicting performance specifically within architecture education is rarely tackled in the literature. This study presents a data driven early warning system tailored to architectural pedagogy, specifically aimed at predicting struggling students in the architectural design studio. Utilizing a longitudinal dataset of 670 student records collected from a higher architecture education institution in Morocco between 2021 and 2024, the research incorporates foundational demographics, pre-enrollment metrics, prior academic trajectories, and behavioral indicators. To capture the dynamic nature of student progress, we introduce extracted features representing the trend, volatility, and average of prior studio performances. The Random Forest classifier demonstrated superior performance based on Balanced Accuracy. Global interpretability analysis using SHAP reveals the overwhelming predictive weight of prior architectural design studio performance, theoretical and technical modules, and historic absenteeism, objectively validating the necessity of continuous engagement in the design process.

DOI: 10.5281/zenodo.20374261

Keywords: Educational Data Mining (EDM), Student Performance Prediction, Architectural Design Studio, Explainable AI (XAI), SHAP, Machine Learning, Early Warning System.

1. Introduction

The integration of Artificial Intelligence (AI) and Educational Data Mining (EDM) has altered how higher education institutions evaluate and support student cohorts [1]. By shifting institutional strategies from reactive grading to proactive forecasting, universities can deploy targeted interventions that enhance student performance [2], [3]. However, the vast majority of existing predictive frameworks have been calibrated for science, technology, engineering, and mathematics disciplines [4], where knowledge acquisition is frequently measured through standardized assessments.

This work was supported by the Ministry of National Education, Professional Training, Higher Education and Scientific Research (MENFPESRS), the Digital Development Agency (DDA) of Morocco (Nos. Alkhawarizmi/2020/23), and the National Scientific and Technical Research Centre (CNRST) under the ≪PhD-Associate Scholarship - PASS≫ program.

Architectural education deviates from this norm. The core of the architectural curriculum is the architectural design studio, a subjective, iterative, and peer-driven environment where students synthesize spatial reasoning, environmental logic, and creative expression [5]. Success in this setting relies heavily on sustained spatial reasoning, physical engagement in peer critiques, and the cumulative progression of design capabilities.

Despite the rapid expansion of EDM, predicting performance specifically within architecture education is rarely tackled in the literature. Furthermore, many studies related to predicting student performance do not tackle explainability [6], [7].

This research addresses these gaps by developing an architecture-specific machine learning pipeline designed to predict student performance in the core architectural design studio. The integration of Explainable AI (XAI), particularly SHapley Additive exPlanations (SHAP), represents a critical pedagogical opportunity to understand the predictive model’s internal logic and to derive concrete pedagogical insights. Leveraging a dataset of 670 longitudinal observations collected from a higher architecture education institution in Morocco between 2021 and 2024, this study utilizes foundational demographics, pre-enrollment metrics, entrance exams features, prior academic trajectories, and behavioral indicators. To quantify a student's academic momentum, we engineer historical performance variables measuring the trend, volatility, and average of their architectural design capabilities over time. Utilizing the pedagogical grading threshold of 13 out of 20, as defined by the institution’s assessment criteria, the objective is to provide educational administrators with a transparent early warning system that forecasts whether a student is safe or struggling.

The remaining sections of this paper are structured as follows: Section 2 provides a review of the relevant literature. Section 3 details the dataset characteristics. Section 4 outlines the methodology, including feature extraction, multicollinearity control, classification algorithms, model evaluation, and explainability. Section 5 presents and discusses the results, including classification performance, overfitting analysis, and global SHAP explanations. Section 6 concludes the study with the pedagogical implications.

2. Literature Review

Mehmood et al. [8] conducted a data-driven analysis of student performance based on demographics, socio-economic variables, and academic records collected via surveys. After testing Naive Bayes (NB), J48 Decision Trees (DT), Random Forest (RF), and Support Vector Machines (SVM), their experiments showed that SVM achieved the highest accuracy (62.50%).

Varsos et al. [9] sought to predict secondary-school student cognitive performance using demographic and early-year grade data. Evaluating algorithms including CHAID, Quest, C5, Bayesian Networks, and RF, the study found that the CHAID algorithm achieved the highest test accuracy of 82.14%.

Badrani et al. [10] developed a personalized guidance system for Moroccan students. The authors compared SVM, Neural Networks (NN), and classical methods, identifying a hybrid SVM-NN model as the superior framework, achieving 99.17% accuracy and 99.37% recall.

Al-Ahmad et al. [11] predicted early grade point averages using longitudinal university data. Comparing deep learning against traditional algorithms, their Long Short-Term Memory model outperformed the others with an R^2 score of 99%.

Bellaj et al. [12] evaluated deep learning, SVM, K-Nearest Neighbors (KNN), and tree-based ensembles to improve students’ academic performance forecasting. In their comparative analysis, RF reached approximately 91% accuracy, while Random Tree achieved 75.18% accuracy among decision tree variants.

Khairy et al. [13] aimed to predict the exam performance of undergraduate students in computer department statistics courses to reduce future failure rates. The researchers evaluated RF, DT, NB, NN, and KNN on a dataset comprising midterm and practical exam grades. The DT and RF classifiers emerged as the top performing algorithms with an accuracy of 98.70%.

Harif and Kassimi [14] evaluated predictive modeling of student performance in Moroccan universities. By integrating Recursive Feature Elimination with Cross-Validation, Lasso Regression emerged as the most stable predictive model with an R^2 of 0.86.

Badal and Sungkur [15] evaluated classifiers to predict final grades and engagement levels on an online learning platform. The quantitative analysis established that the RF algorithm yielded the highest accuracy, reaching 85% for grades and 83% for student engagement.

Alturki and Alturki [16] investigated the prediction of academic performance to facilitate early interventions. The study compared six data mining methods: C4.5, Simple CART, LADTree, NB, Bayes Net, and RF. NB achieved the highest accuracy (69.67%) for predicting overall graduation grades, while RF proved optimal for predicting honorary students with an accuracy of 92.6%.

Ouatik et al. [17] aimed to predict student success and failure rates in Moroccan university settings. By integrating the MapReduce framework with various algorithms to handle educational datasets, their SVM model achieved the highest recognition rate of 87.32%.

Hashim et al. [18] proposed a supervised machine learning model to predict university students' final outcomes using demographic and course-related features. Evaluating DT, NB, LR, SVM, KNN, and NN, they found LR to be the top performer, reaching 88.8% accuracy for pass/fail status.

Asif et al. [19] analyzed four-year undergraduate performance using admission marks and early-year course grades. Utilizing classifiers such as DT, rule-based learners, 1-NN, NB, NN, and RF, they reported NB as the best performing model with an accuracy of about 83.65%.

3. Dataset Description

3.1 Target variable distribution

The dataset comprises 670 discrete samples representing architecture students tracked longitudinally from their second to fourth semesters (S2 to S4). The target variable dictates the student's holistic performance in the current semester's architectural design studio. The target variable is dichotomized using the academic scale adopted by institutional pedagogy. It presents a relatively balanced distribution, with 48% of the cohort classified as Struggling and 52% classified as Not Struggling.

3.2 Feature description and data specification

Table 1 presents a structured overview of the variables used to model student performance in architectural design studios. The dataset integrates demographic attributes, pre-enrollment indicators, entrance exam scores, prior academic results, and behavioral factors. Each feature is defined along with its type and value range to ensure consistency in preprocessing and reproducibility of the experimental pipeline.

Table 1. Description of the features used for student performance prediction

Category	Feature Name	Description	Type / Values
Identifiers & Context	STUDENT_ID	Unique student identifier	Numerical
	SEMESTER	Current semester	Categorical (S2, S3, S4)
Demographics	AGE	Student age group	Categorical (17, 18, ≥19)
	SEX	Gender	Categorical (M, F)
	SCHOL	Scholarship / financial aid status	Categorical (Yes, No)
Pre-Enrollment Metrics	BAC_YEARS	Years since high school graduation	Categorical (0, 1, ≥2)
	BAC_GEN	General cumulative baccalaureate grade	Numerical (0–20)
	BAC_NAT	National baccalaureate grade	Numerical (0–20)
	BAC_REG	Regional baccalaureate grade	Numerical (0–20)
	BAC_PRES	Preselection grade	Numerical (0–20)
Entrance Exams	ENT_DRAW	Architectural drawing exam score	Numerical (0–20)
	ENT_QCM	Multiple-choice exam score	Numerical (0–20)
	ENT_GEN	Overall combined entrance grade	Numerical (0–20)
Prior Academic Trajectory	LAG_ARCHI	Previous architectural design studio grade	Numerical (0–20)
	LAG_ARTS	Previous plastic arts grade	Numerical (0–20)
	LAG_DESSIN	Previous technical drawing grade	Numerical (0–20)
	LAG_M_SS	Previous social sciences & communication module grade	Numerical (0–20)
	LAG_M_CB	Previous construction and building technology module grade	Numerical (0–20)
Behavioral Indicators	LAG_ABS_ARCHI	Absences in the previous architectural design studio	Numerical
Target Variable	ARCHI_TARGET	Final grade achieved in the current semester’s architectural design studio	Target (Binary)

4. Methodology

4.1 Feature extraction: quantifying academic momentum

Academic performance in a design studio is a dynamic trajectory reflecting cognitive adaptation and creative resilience. To encapsulate this momentum, we engineered three longitudinal features based on historical architectural design studio grades.

ARCHI_TREND: This tracks the velocity and direction of the student's architectural abilities. Because the amount of historical data varies by the student's current semester, the calculation adapts:

For Semester 4: Calculated as the linear regression slope of the architectural design studio grades spanning S1 to S3, establishing a long-term trajectory.

For Semester 3: Calculated as the studio grade in S2 minus the grade in S1, measuring semester-over-semester improvement.

For Semester 2: To partially compensate for the lack of prior S2 studio observations, we use the entrance drawing exam (ENT_DRAW) as a proxy for pre-university artistic ability. For S2, ARCHI_TREND is calculated as the difference between the studio grade in S1 and entrance drawing exam (LAG_ARCHI - ENT_DRAW). This metric approximates the student's transition from innate artistic capability to structured university design practice.

ARCHI_VOLAT: This metric represents the stability of the student's performance. It is computed as the standard deviation of the previous architectural grades utilized in the ARCHI_TREND calculation for a specific semester.

ARCHI_AVG: Represents the baseline historical capability, calculated as the arithmetic average of all previous architectural design studio grades.

4.2 Correlation analysis

Hierarchical clustering is an unsupervised learning method that organizes data into nested groups according to similarity relationships. Its objective is to construct a hierarchy of clusters without requiring the number of clusters in advance. The method produces a dendrogram that represents successive merging or splitting operations among clusters.

Consider a dataset X={x_1,x_2,…,x_n }, where each observation belongs to an m-dimensional space. The first step consists of computing pairwise distances between observations. The most common metric is the Euclidean distance:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{m}(x_{ik} - x_{jk})^2}$$

Another frequently used metric is the Manhattan distance:

$$d(x_i, x_j) = \sum_{k=1}^{m}|x_{ik} - x_{jk}|$$

In agglomerative hierarchical clustering, each point initially forms an independent cluster:

$$\mathrm{C_i} = \{\mathrm{x_i}\}, \qquad \mathrm{i} = 1, \dots, \mathrm{n}$$

At each iteration, the two closest clusters are merged according to a linkage criterion. In single linkage clustering, the inter-cluster distance is defined by:

$$\mathrm{D(A,B)} = \min_{\mathrm{x \in A, y \in B}} d(x,y)$$

while complete linkage uses:

$$\mathrm{D(A,B)} = \max_{\mathrm{x \in A, y \in B}} d(x,y)$$

Average linkage computes the mean distance:

$$\mathrm{D(A,B)} = \frac{1}{|A||B|} \sum_{\mathrm{x \in A}} \sum_{y \in B} d(x,y)$$

For centroid linkage, each cluster is represented by its centroid:

$$\mu_A = \frac{1}{|A|} \sum_{x \in A} x$$

and the cluster distance becomes:

$$\mathrm{D(A,B)} = d(\mu_A, \mu_B)$$

One of the most important approaches is Ward's method, which minimizes the increase of intra-cluster variance after merging:

$$\Delta E = E(A \cup B) - E(A) - E(B)$$

where:

$$E(C) = \sum_{x \in C} ||x - \mu_C||^2$$

The hierarchy generated by the algorithm can be represented mathematically as a nested sequence of partitions:

$$\Pi_1 \preccurlyeq \Pi_2 \preccurlyeq \cdots \Pi_n$$

Hierarchical clustering is widely used because it provides an interpretable multilevel structure of the data. However, its computational complexity is generally high, often requiring $O(n^2)$ memory and up to $O(n^3)$ computational time for large datasets.

To ensure algorithm stability and valid global explanations, we implement this hierarchical clustering framework as a multicollinearity control mechanism prior to model training. High inter-feature correlations can distort learning weights and feature importance rankings. We deploy hierarchical clustering on Spearman rank-order correlations. A stringent distance threshold of 0.15 is applied to cut the dendrogram. This threshold dictates that any features exhibiting a Spearman correlation ≥0.85 are considered redundant. As illustrated in the resulting hierarchical clustering dendrogram (Fig. 1), only BAC_NAT and BAC_PRES exhibited a highly collinear relationship exceeding the 0.85 correlation threshold. Because BAC_PRES is a derived metric calculated directly from BAC_NAT and BAC_REG, we retained BAC_NAT to preserve the original assessment of national academic rigor and dropped the redundant variable.

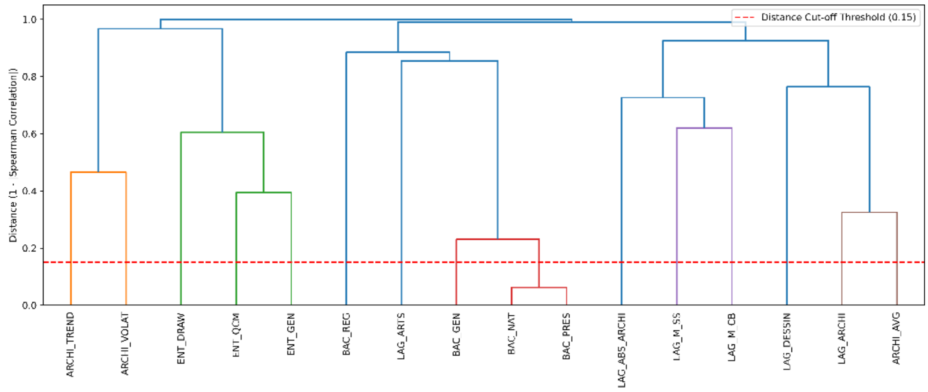

Fig. 1. Hierarchical clustering dendrogram of feature correlations

4.3 Classification algorithms and evaluation metrics

The experimental setup evaluates five classification algorithms to identify the optimal framework for predicting student outcomes. These models range from linear probabilistic approaches to complex non-linear ensembles.

Naive Bayes is a probabilistic classifier based on Bayes' theorem with the strong (naive) assumption of conditional independence between features given the class [20]. For a class C_k and feature vector $x = (x_1, \dots, x_n)$, the posterior is computed as:

$$\mathrm{P}(\mathrm{C_k} \mid \mathrm{x}) = \frac{P(C_k) \prod_{i=1}^{n} P(x_i \mid C_k)}{P(x)}$$

In practice, classification is done by maximizing the numerator:

$$\hat{C} = \arg\max_{\mathrm{C_k}} P(C_k) \prod_{i=1}^{n} P(x_i \mid C_k)$$

Despite its simplicity, it performs well in text classification and high-dimensional sparse data.

Logistic Regression is a linear probabilistic model used for binary classification [21]. It models the probability of class membership using the sigmoid function:

$$P(y = 1 \mid x) = \sigma(w^T x + b) = \frac{1}{1 + e^{-(w^T x + b)}}$$

Parameters w,b are learned by minimizing the cross-entropy loss:

$$\mathcal{L} = -\sum_{i=1}^{N} [y_i log\hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)]$$

It is widely used due to its interpretability and strong performance on linearly separable data.

Decision Trees are non-parametric models that split the feature space into regions using hierarchical if-then rules. At each node, the model selects a feature and threshold that best separates the data, typically by minimizing impurity [22]. For regression trees, a common criterion is variance reduction:

$$Var(D) = \frac{1}{|D|} \sum_{i \in D} (y_i - \bar{y})^2$$

The split is chosen to minimize:

$$\Delta = Var(D) - \left(\frac{|D_L|}{|D|} Var(D_L) + \frac{|D_R|}{|D|} Var(D_R) \right)$$

For classification trees, impurity is often measured using Gini index or entropy. The Gini impurity is:

$$G(D) = 1 - \sum_{k=1}^{K} p_k^2$$

Support Vector Machines aim to find the optimal separating hyperplane that maximizes the margin between classes [23]. The decision function is:

$$f(x) = w^T x + b$$

The optimization problem (hard-margin case) is:

$$\min_{w,b} \frac{1}{2} ||w||^2 \text{ subject to } y_i \left(w^T x_i + b \right) \geq 1$$

For non-linearly separable data, slack variables and kernels (e.g., $K(x_i, x_j)$) are introduced, allowing SVMs to construct nonlinear decision boundaries in high-dimensional feature spaces.

Random Forest is an advanced ensemble classifier that constructs a multitude of decision trees via bootstrap aggregating (bagging), providing a robust, low-variance classifier. It trains each tree on a bootstrap sample of the training data and, at each split, considers only a random subset of features, which reduces correlation between trees and improves generalization [24].

For regression, suppose the Random Forest contains T trees, and each tree t outputs a prediction $\hat{y}^t(x)$ for an input x. The final prediction is the average of the individual tree predictions:

$$\hat{y}(x) = \frac{1}{T} \sum_{t=1}^{T} \hat{y}^t(x)$$

For classification, each tree produces a class prediction C ̂^((t)) (x). The Random Forest prediction is obtained by majority vote:

$$C(x) = \arg\max \sum_{t=1}^{T} 1(C^{(t)}(x) = c)$$

Equivalently, if each tree outputs a class probability estimate $\hat{p}_c^{(t)}(x)$ for class c, the forest's class probability is the average of tree probabilities:

$$\hat{p}_c(x) = \frac{1}{T}\sum_{t=1}^{T}\hat{p}_c^{(t)}(x)$$

and the predicted class is:

$$C(x) = \arg\max_{c} \hat{p}_c(x)$$

Models are evaluated primarily using Balanced Accuracy, alongside Precision, Recall, F1-Score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). Balanced Accuracy calculates the arithmetic mean of sensitivity (true positive rate) and specificity (true negative rate), ensuring that the model's performance is rigorously assessed across both the Safe class and the Struggling class.

4.4 Model selection and evaluation

A constraint of this dataset is its longitudinal nature, containing up to three repeated measures per individual student. Applying a standard randomized cross-validation split would result in data leakage. The algorithm could inadvertently learn an individual student's static profile across different folds rather than generalizable pedagogical patterns [25]. To prevent this, the framework employs GroupKFold cross-validation utilizing 10 splits. By designating STUDENT_ID as the grouping vector, the process ensures that all records belonging to a specific student are isolated entirely within either the training or the validation fold for any given iteration.

Within this evaluation loop, hyperparameter tuning is executed using RandomizedSearchCV with 20 iterations. This approach randomly samples from a defined hyperparameter grid, locating the optimal configuration for each algorithm while remaining computationally efficient [26].

4.5 Explainability

To clarify the decision boundaries, the framework incorporates SHAP, applied to the best optimized model. Rooted in cooperative game theory, SHAP provides a unified measure of global feature importance [27]. By deploying SHAP over the non-collinear feature set, we utilize both SHAP Bar plots and Beeswarm plots to explicitly rank and directionally map the pedagogical drivers of student success.

5. Results and Discussion

5.1 Hyperparameter tuning results

The application of RandomizedSearchCV (20 iterations) within the 10-fold GroupKFold cross-validation loop identified the optimal configurations to maximize the models' generalization capabilities. The process established fixed constraints to handle inherent class dynamics and tuned parameters to optimize the decision boundaries. Table 2 details these configurations.

Table 2. Fixed and tuned hyperparameters for the evaluated machine learning models

Model	Fixed Parameters	Tuned Hyperparams

NB	-	var_smoothing = 1e-09
LR	class_weight = balanced	penalty = l2, C = 0.1
DT	class_weight = balanced	max_depth = 3 min_samples_split = 5 min_samples_leaf = 5
SVM	class_weight = balanced probability = True	kernel = rbf, C = 1 gamma = 0.01
RF	class_weight = balanced	n_estimators = 100 max_depth = 3 min_samples_split = 20 min_samples_leaf = 5

5.2 Overfitting analysis

To determine the generalization capacity of each algorithm under the 10-fold GroupKFold strategy, the Balanced Accuracy on the training set was directly compared against the testing set. Fig. 2 provides a visual representation of this performance gap.

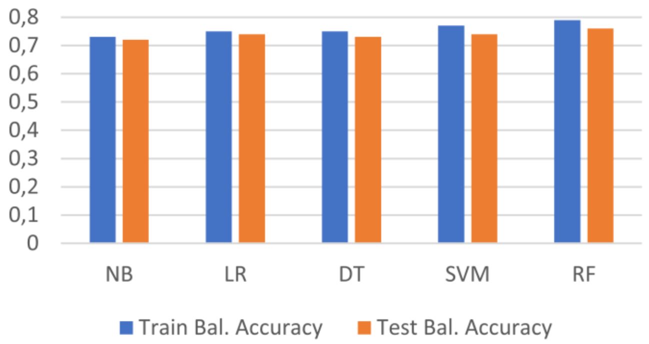

Fig. 2. Train vs. Test balanced accuracy across classification models

The analysis illustrates varying degrees of variance across the models. Linear and simpler probabilistic models (LR, NB) demonstrated minimal overfitting, exhibiting narrow gaps between training and testing performance, but failed to capture the upper bounds of predictive accuracy. SVM presented a medium stable gap of 0.04 (0.77 to 0.73), demonstrating good generalization to unseen student groups without sacrificing boundary complexity. The ensemble method, RF, achieved the highest overall training score (0.79) and successfully preserved a strong test score (0.76), indicating that its bootstrap aggregation mechanism and tuned parameters effectively managed variance while maximizing predictive capacity.

5.3 Classification performance

Table 3 details the final evaluation metrics for the models’ performance.

Table 3. Performance evaluation of machine learning models

Model	Bal. Acc.	Precision	Recall	F1-score	AUC-ROC
NB	0.71	0.69	0.73	0.71	0.79
LR	0.73	0.72	0.72	0.72	0.81
DT	0.72	0.73	0.71	0.71	0.78
SVM	0.73	0.72	0.71	0.72	0.81
RF	0.76	0.75	0.75	0.75	0.82

The evaluation demonstrates a clear progression in predictive capability. While traditional algorithms such as LR and SVM established a reliable baseline (Balanced Accuracy of 0.73), they were outperformed by the tree-based ensemble. RF provided the most operationally effective performance across all evaluated metrics. It secured the highest Test Balanced Accuracy (0.76), F1-score (0.75), and AUC-ROC (0.82). In the context of an early warning system, RF’s balanced Precision (0.75) and Recall (0.75) are vital; this ensures that struggling students are correctly flagged for academic intervention while minimizing false alarms that could drain institutional resources.

5.4 Global Explainability and Pedagogical Insights

The global SHAP analysis revealed clear insights into the mechanics of the optimized RF model. To provide a comprehensive interpretation, we analyze both the SHAP Bar plot and the Beeswarm plot.

The SHAP Bar plot (Fig. 3) quantifies the mean absolute impact of each feature on the model’s predictions. The analysis unequivocally demonstrates that prior university-level academic performance dictates future studio success. The previous semester’s architecture grade, LAG_ARCHI, is the most dominant predictor by a significant margin (mean|SHAP|>+0.1). This is followed by the historical average, ARCHI_AVG, at +0.04, and crucially, performance in theoretical and technical modules: social sciences and communication, LAG_M_SS at +0.03, and construction and building technology, LAG_M_CB at +0.02. Notably, university entrance metrics like the drawing exam, ENT_DRAW, and the multiple-choice exam, ENT_QCM, along with demographic factors, AGE and SCHOL, exert minimal predictive influence (mean|SHAP| near zero). This highlights that while entrance exams serve as initial admission filters, sustained academic rigor and cross-disciplinary knowledge acquired at the university level are the true determinants of ongoing architectural design studio performance.

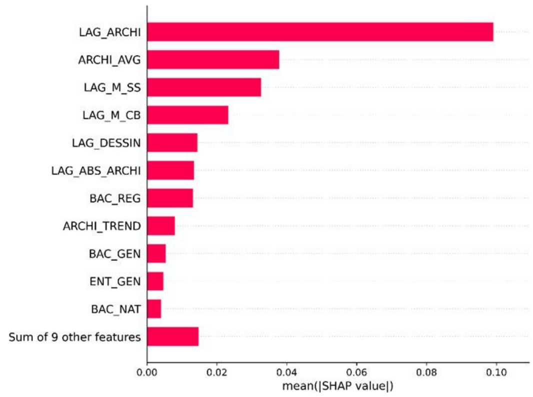

Fig. 3. SHAP Bar plot: Global feature importance estimated by SHAP values

The SHAP Beeswarm plot (Fig. 4) provides a detailed visualization of how specific feature values drive individual predictions. Each point represents a student, with colors indicating the feature's value (red for high, blue for low). The model predicts the probability of struggling students. For the top features, including LAG_ARCHI, ARCHI_AVG, LAG_M_SS, and LAG_M_CB, a strong negative relationship is observed regarding the struggling class. High grades (red points) stretch far into the negative SHAP values, drastically decreasing the probability that a student will struggle. Conversely, low grades (blue points) push the model output toward the right (positive SHAP), increasing the likelihood of identifying the student as at-risk. The extracted ARCHI_TREND feature also demonstrates that a positive trajectory lowers the risk of student struggling, whereas high volatility in prior grades pushes the model output toward the right, increasing the predicted probability of difficulty. This pattern validates the need to monitor academic momentum rather than isolated grades. Furthermore, behavioral indicators like previous absenteeism (LAG_ABS_ARCHI) demonstrate a strict positive impact on risk. High absenteeism consistently yields positive SHAP values. This provides empirical evidence that a lack of physical engagement with the peer-to-peer critique culture causes tangible damage to a student's design trajectory. The integration of SHAP with the RF classifier ensures these findings are highly interpretable and directly applicable for educational policymakers.

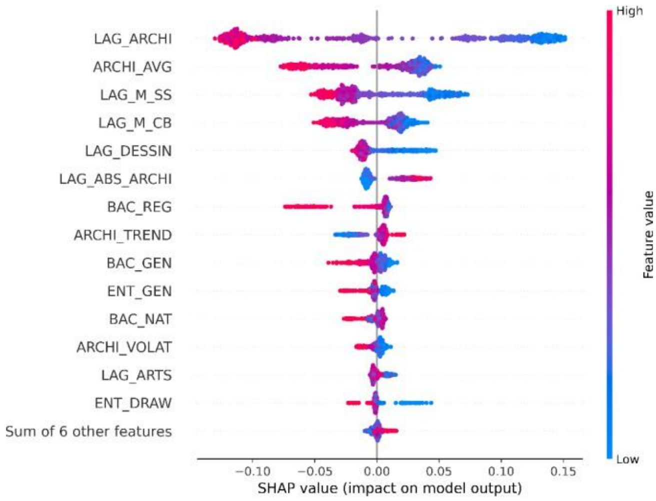

Fig. 4. SHAP Beeswarm plot: Distribution of feature impacts on the model output

6. Conclusions

This research successfully constructs and validates an early warning system explicitly tailored to the architectural design studio, an educational domain rarely tackled by predictive analytics literature. The integration of novel historical features, especially the performance trend and average of a student's past performance, enhanced the model's ability to assess academic momentum rather than relying on static data points.

The methodology's strict adherence to 10-fold GroupKFold cross-validation successfully mitigated the risk of longitudinal data leakage. Furthermore, addressing multicollinearity through hierarchical clustering on Spearman rank-order correlations ensured a robust, independent feature set. Among the evaluated models, RF yielded the most reliable forecasts, outperforming others in crucial metrics like Recall and Balanced Accuracy while demonstrating excellent resistance to overfitting.

The integration of global SHAP explainability allowed for the extraction of transparent pedagogical insights. The analysis indicates that prior studio performance and cumulative architectural competence overwhelmingly dictate future success. Furthermore, the significant weight of theoretical and technical modules, such as social sciences and communication as well as construction and building technology, empirically validates the multidisciplinary nature of architectural education.

The predictive dominance of deteriorating grade trends, theoretical and technical academic performance, and historic absenteeism points a clear policy direction: institutions should actively monitor continuous performance and attendance data to enable proactive, data informed interventions. Publishing these findings contributes directly to the discourse in engineering and architectural pedagogy, providing a framework for architecture institutions to transition from reactive grading to proactive student support.

While this study provides a strong and highly interpretable framework, the data is currently limited to a single institution in Morocco. Future work needs to consider cross-institutional validation to establish true generalizability across diverse educational contexts. We recommend future researchers to test this framework in other geographical and cultural contexts. This validation will examine whether base metrics (such as the historical average of grades of architectural design studios and previous studio grades) continue to be the strongest predictors and, critically, whether the extracted momentum features of academic “Volatility” and “Trend” remain useful as finer-grained risk signals for different pedagogies in different regions.

REFERENCES

1. *C. Romero and S. Ventura, “Educational data mining and learning analytics: An updated survey,” Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 10, no. 3, May 2020, doi: 10.1002/widm.1355.*

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

3. *J. Pan, Z. Zhao, and D. Han, “Academic Performance Prediction Using Machine Learning Approaches: A Survey,” IEEE Trans. Learn. Technol., vol. 18, pp. 351–368, 2025, doi: 10.1109/TLT.2025.3554174.*

4. *A. Turkmenbayev, E. Abdykerimova, S. Nurgozhayev, G. Karabassova, and D. Baigozhanova, “The application of machine learning in predicting student performance in university engineering programs: a rapid review,” Front. Educ., vol. 10, p. 1562586, Sep. 2025, doi: 10.3389/FEDUC.2025.1562586/BIBTEX.*

5. *A. Iranmanesh and Z. Onur, “Generation gap, learning from the experience of compulsory remote architectural design studio,” Int. J. Educ. Technol. High. Educ., vol. 19, no. 1, Dec. 2022, doi: 10.1186/S41239-022-00345-7.*

6. *S. Boujmiraz, H. Darhmaoui, and A. Drissi el maliani, “Predicting student performance: A comprehensive review of machine learning, deep learning, and explainable AI approaches,” Comput. Educ. Artif. Intell., vol. 10, p. 100548, Jun. 2026, doi: 10.1016/J.CAEAI.2026.100548.*

7. *F. T. Johora, M. N. Hasan, A. Rajbongshi, M. Ashrafuzzaman, and F. Akter, “An explainable AI-based approach for predicting undergraduate students academic performance,” Array, vol. 26, Jul. 2025, doi: 10.1016/j.array.2025.100384.*

8. *A. Mehmood, B. Mehmood, M. Ammar, M. Latif, A. B. Abro, and H. Nawaz, “Data-Driven Analysis of Student Performance Using Machine Learning:Insights from Demographics and Educational Records,” VFAST Trans. Softw. Eng., vol. 13, no. 1, pp. 111–120, Jan. 2025, doi: 10.21015/vtse.v13i1.2062.*

9. D. A. Varsos, N. C. Zygouris, K. M. Kolomvatsos, A. N. Dadaliaris, and G. Dimitriou, "Prediction of Secondary-School Student Cognitive Performance in Informatics," Int. J. Eng. Pedagog., vol. 16, no. 1, pp. 66–86, Mar. 2026, doi: 10.3991/ijep.v16i1.56637.

10. M. Badrani, A. Marouan, N. Kannouf, and A. Chetouani, "Personalized Guidance for Moroccan Students: An Approach Based on Machine Learning and Big Data," Int. J. Eng. Pedagog., vol. 15, no. 1, pp. 125–136, Jan. 2025, doi: 10.3991/IJEP.V15I1.51985.

11. B. I. Al-Ahmad, A. Alzaqebah, R. Alkhawaldeh, A. M. Al-Zoubi, H. Lo, and A. Ali, "Predicting academic performance for students' university: case study from Saint Cloud State University," PeerJ Comput. Sci., vol. 9, Sep. 2025, doi: 10.7717/PEERJ-CS.3087.

12. M. Bellaj, A. Ben Dahmane, S. Boudra, and M. Lamarti Sefian, "Educational Data Mining: Employing Machine Learning Techniques and Hyperparameter Optimization to Improve Students' Academic Performance," Int. J. online Biomed. Eng., vol. 20, no. 3, pp. 55–74, 2024, doi: 10.3991/IJOE.V20I03.46287.

13. D. Khairy, N. Alharbi, M. A. Amasha, M. F. Areed, S. Alkhalaf, and R. A. Abougalala, "Prediction of student exam performance using data mining classification algorithms," Educ. Inf. Technol., vol. 29, no. 16, pp. 21621–21645, Nov. 2024, doi: 10.1007/s10639-024-12619-w.

14. A. Harif and M. A. Kassimi, "Predictive Modeling of Student Performance Using RFECV-RF for Feature Selection and Machine Learning Techniques," Int. J. Adv. Comput. Sci. Appl., vol. 15, no. 7, pp. 231–240, 2024, doi: 10.14569/IJACSA.2024.0150723.

15. Y. T. Badal and R. K. Sungkur, "Predictive modelling and analytics of students' grades using machine learning algorithms," Educ. Inf. Technol., vol. 28, no. 3, pp. 3027–3057, Mar. 2023, doi: 10.1007/S10639-022-11299-8.

16. S. Alturki and N. Alturki, "Using Educational Data Mining To Predict Students' Academic Performance For Applying Early Interventions," J. Inf. Technol. Educ. Innov. Pract., vol. 20, pp. 121–137, 2021, doi: 10.28945/4835.

17. F. Ouatik, M. Erritali, F. Ouatik, and M. Jourhmane, "Students' Orientation Using Machine Learning and Big Data," Int. J. online Biomed. Eng., vol. 17, no. 1, pp. 111–119, 2021, doi: 10.3991/ijoe.v17i01.18037.

18. A. S. Hashim, W. A. Awadh, and A. K. Hamoud, "Student Performance Prediction Model based on Supervised Machine Learning Algorithms," in IOP Conference Series: Materials Science and Engineering, Nov. 2020, vol. 928, no. 3. doi: 10.1088/1757-899X/928/3/032019.

19. R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, "Analyzing undergraduate students' performance using educational data mining," Comput. Educ., vol. 113, pp. 177–194, Oct. 2017, doi: 10.1016/j.compedu.2017.05.007.

20. I. Rish, "An empirical study of the naive Bayes classifier," in IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, 2001, pp. 41–46.

21. S. Sperandei, "Understanding logistic regression analysis," Biochem. Medica, vol. 24, no. 1, pp. 12–18, 2014, doi: 10.11613/BM.2014.003.

22. Y. Y. Song and Y. Lu, "Decision tree methods: applications for classification and prediction," Shanghai Arch. Psychiatry, vol. 27, no. 2, pp. 130–135, Apr. 2015, doi: 10.11919/J.ISSN.1002-0829.215044.

23. C. Cortes and V. Vapnik, "Support-Vector Networks," Mach. Learn., vol. 20, no. 3, pp. 273–297, 1995, doi: 10.1023/A:1022627411411.

24. L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324/METRICS.

25. D. R. Roberts et al., "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure," Ecography, vol. 40, no. 8. Blackwell Publishing Ltd, pp. 913–929, Aug. 01, 2017. doi: 10.1111/ecog.02881.

26. *J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," J. Mach. Learn. Res., vol. 13, pp. 281–305, 2012.*

27. *S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, 2017, vol. 2017-Decem, pp. 4766–4775.*

Youness Hajjioui, Youssef Jedidi, Mohamed Benslimane and Abdelali Ibriz
Laboratory of Sciences, Engineering and Management, Higher School of Technology of Fez, Sidi Mohamed Ben Abdellah University, Fez, Morocco
Karim El Moutaouakil
Laboratory of Mathematics and Data Science, Multidisciplinary Faculty of Taza, Sidi Mohamed Ben Abdellah University, Fez, Morocco