RESEARCH ARTICLE

KFWAdaBoost-Based Soft Label Learning Framework for Student Performance Prediction

Zhihong Yu[1,*]

[1] Fuzhou Technology and Business University, Fuzhou 350715, China

Abstract

Student performance prediction is a core task in educational data mining, as it enables early intervention, personalized learning support, and data-driven decision-making. Although existing machine learning models have shown promising results in this domain, challenges persist due to hard-to-classify samples—particularly students exhibiting borderline performance—and the discrete nature of hard labels, which together limit predictive effectiveness. To overcome these limitations, this paper proposes a KFWAdaBoost-based soft label learning framework that systematically enhances baseline model performance through a two-stage synergistic mechanism. In the first stage, K-means++ clustering is employed to generate similarity features, thereby providing structural awareness of underlying data patterns. In the second stage, probabilistic soft labels are derived from ensemble confidence scores to refine decision boundaries and better handle ambiguous cases. Experimental results on the widely used Mathematics and Portuguese Language course datasets demonstrate that the proposed framework consistently improves baseline performance across Accuracy, Precision, Recall, and F1-Score for models including LDA, Decision Tree, and SVM, with Decision Tree exhibiting the most substantial gains. This framework offers a reliable and effective approach for student performance prediction and holds strong potential for broader applications in educational data analytics.

Keywords: soft label learning, KFWAdaBoost, K-means++ clustering, student performance prediction, educational data mining.

1 Introduction

As the volume of educational data collected from learning management systems and student assessment platforms continues to grow, new opportunities for Educational Data Mining (EDM) have emerged [1, 2]. EDM refers to the application of data-driven techniques to analyze educational datasets, with the goal of understanding learning processes and optimizing outcomes for learners [3]. In EDM, student performance prediction has become an important application, underpinning early-warning and dropout-prevention systems [4–6]. For educators, accurate prediction of student performance enables timely interventions, substantially lowering dropout rates and boosting course engagement [7]. Moreover, identifying students needing extra support allows for better allocation of educational resources [8].

Submitted: 03 February 2026
Accepted: 25 February 2026
Published: 28 February 2026

Vol. 2, **No.** 1, 2026.
10.62762/TEDM.2026.459733

***Corresponding author:**
✉ Zhihong Yu
22014083067@hqu.stu.edu.cn

Citation
Yu, Z. (2026). KFWAdaBoost-Based Soft Label Learning Framework for Student Performance Prediction. *ICCK Transactions on Educational Data Mining*, 2(1), 1–13.

To accurately predict student performance, various machine learning models have been employed, ranging from classical classifiers to ensemble methods [9]. For instance, [10] demonstrated that Support Vector Machine (SVM) achieved competitive performance after hyper-parameter tuning. Linear Discriminant Analysis (LDA) has also been widely adopted for its simplicity and interpretability in educational settings [11]. However, single classifiers often struggle with complex data distributions and class imbalance inherent in educational datasets. To address these limitations, ensemble methods such as AdaBoost have been explored to enhance predictive robustness [12]. XGBoost also excels in imbalanced educational data [11].

Despite these advances, two critical challenges remain in student performance prediction. First, **hard-to-classify samples**—particularly students with borderline performance—often lead to unstable predictions, as traditional classifiers treat all misclassifications equally without considering the underlying data structure. Second, **hard labels** (i.e., discrete class assignments) fail to capture the nuanced, continuous nature of academic performance, where students may exhibit partial membership across multiple performance categories. These limitations motivate the need for a more sophisticated framework that leverages data structure awareness and probabilistic label representation.

Recently, clustering-guided ensemble methods have shown promise in enhancing classification performance by exploiting intrinsic data structures [13]. The K-means++ (KPP) algorithm, in particular, provides an effective mechanism for identifying representative prototypes that capture local data distributions [14]. Concurrently, soft label learning has emerged as a powerful paradigm that replaces crisp class assignments with continuous membership degrees, enabling finer-grained decision boundaries [15]. However, the integration of these two approaches—clustering-based structural enrichment and soft label refinement—remains underexplored in educational data mining.

Inspired by these developments, we propose a novel **KFWAdaBoost-based soft label learning framework** for student performance prediction. Our approach operates in two synergistic stages: (1) *clustering-guided feature enhancement*, where KPP-derived similarity features are integrated to provide structural context for difficult samples; and (2) *soft label learning*, where ensemble confidence scores generate probabilistic labels that refine decision boundaries.

The main contributions of this work are as follows:

- KPP Fuzzy Weighted AdaBoost (**KFWAdaBoost**) is proposed as a novel ensemble framework that integrates KPP clustering with FWAdaBoost. Unlike traditional AdaBoost variants that rely solely on sample weights, KFWAdaBoost leverages similarity-based features generated from cluster prototypes, providing explicit structural information to guide the classification of ambiguous samples.
- **Soft label embedding** is introduced based on ensemble confidence scores, capturing probabilistic class memberships and refining decision boundaries for students with borderline performance.
- This study presents a **two-stage hierarchical learning paradigm** that progressively transitions from coarse feature enrichment (Stage 1) to fine-grained label optimization (Stage 2), systematically addressing the challenges of hard-to-classify samples in educational data.

2 Methods

In this section, we first introduce the KFWAdaBoost-based Embedded Soft Label Learning, and then the Classification models, followed by the Soft Label Learning-based framework for student performance prediction.

2.1 The proposed method

The proposed methodology comprises two sequential stages: prototype-based feature augmentation and probabilistic label refinement. In the initial stage, clustering-driven feature enhancement is employed to strengthen model discriminability. Specifically, similarity measures between cluster centroids and individual samples are computed to construct supplementary feature representations, thereby improving the model's capacity to separate ambiguous instances located near decision boundaries. These structural features enable the model to capture underlying data patterns more effectively, consequently elevating the predictive performance of the base ensemble during training. In the subsequent stage, the feature-augmented dataset from the preceding phase serves as input for information enrichment. Leveraging the similarity matrix established earlier, the FWAdaBoost classifier

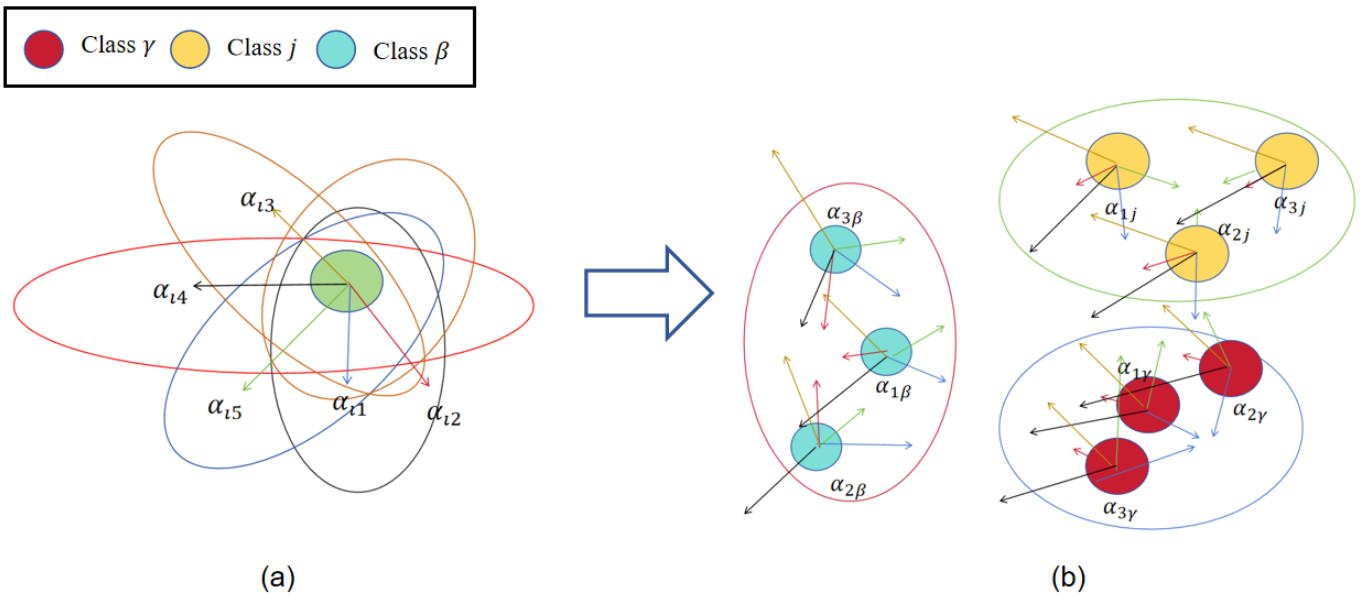

Figure 1. Example of cluster features aiding in sample classification.

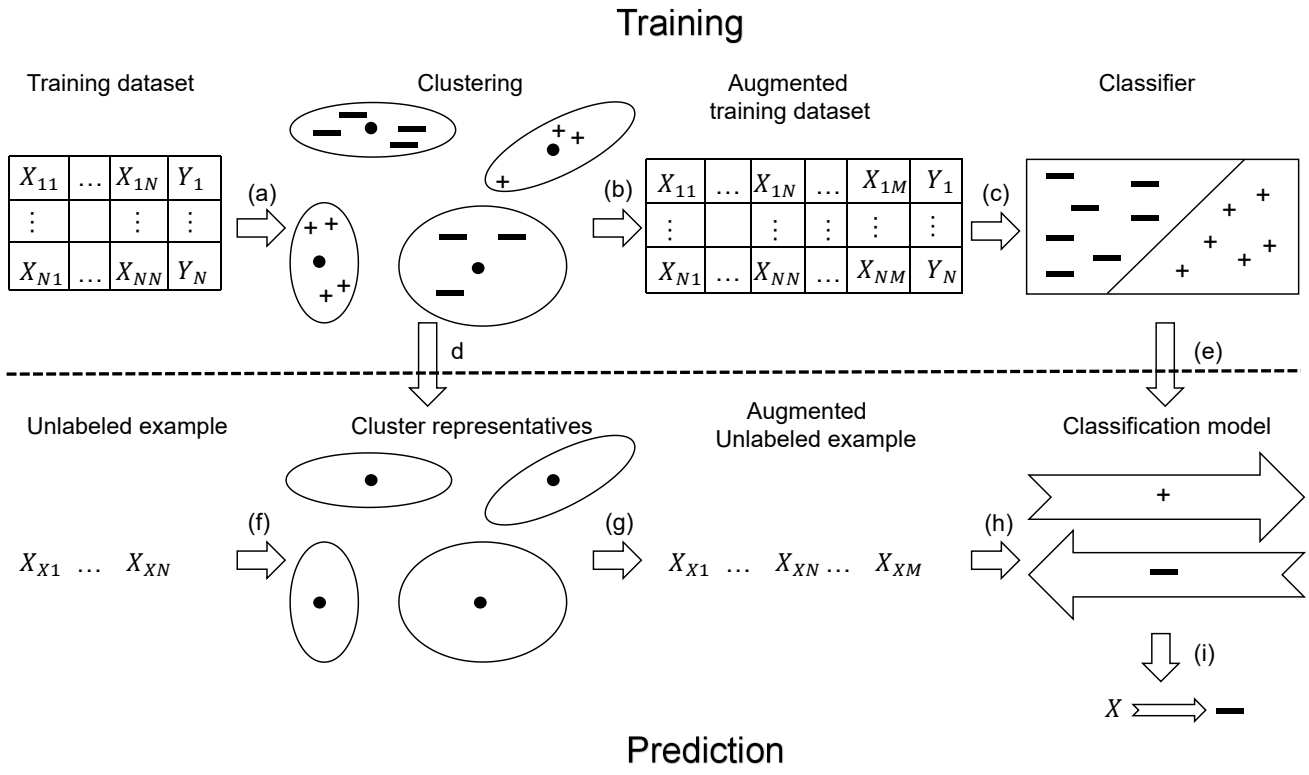

Figure 2. The workflow diagram for generating features using the KPP method is as follows : (a) Using the KPP method to divide the training samples into m clusters; (b) New features are created based on similarity with cluster centers and merged with the existing features of the samples; (c) The new set of samples is used for classifier training; (d) The calculated clusters are saved; (e) A classification model is obtained through training; (f) Unclassified samples are put into learning; (g) The similarity of this sample with each cluster center is calculated and encoded as new features; (h) The new sample is input into the model for prediction; (i) The model makes a prediction for the sample [13].

generates probabilistic soft labels for each instance. These continuous-valued labels are then concatenated with the original feature vectors, creating an information-rich representation that facilitates more nuanced classification. This augmentation allows the learning algorithm to discern finer-grained distinctions among samples that would otherwise remain indistinguishable under hard-label supervision.

The hierarchical dependency between these two stages operates as follows: the initial clustering phase establishes an improved feature foundation by injecting structural awareness into the representation space. Building upon this enhanced substrate, the second phase implements soft label propagation to refine the supervisory signal. This cascading design enables a transition from macroscopic feature engineering to microscopic label optimization, ultimately yielding superior classification accuracy. The overall workflow of this two-stage feature enhancement process is illustrated in Figure 2.

Regarding the illustrative mechanism depicted in Figure 1, panel (a) demonstrates how the proposed framework sculpts more discriminative decision boundaries during the learning process. The visualization depicts sample displacement driven by centroid-oriented similarity metrics, where individual instances (represented as markers) undergo spatial transformation within the embedding space. The surrounding ellipses denote affinity-based groupings. Conceptually, the similarity coefficients function as attractive potentials that draw samples toward their respective cluster prototypes. This gravitational effect reorganizes the feature space topology, compressing intra-cluster distances while expanding inter-cluster separability, thereby facilitating boundary formation. Panel (b) of Figure 1 illustrates how differential attraction strengths between samples promote improved cluster allocation. The attractive force metaphor characterizes the quantitative influence of prototype similarity on sample positioning. Through iterative adjustment of these attraction fields, the ensemble mechanism progressively reallocates ambiguous instances toward their most compatible clusters. Notably, the dimensionality of these attraction fields exceeds the categorical label space, as the clustering procedure generates prototype counts that equal or surpass the number of ground-truth classes. This overcomplete representation captures latent similarities that transcend nominal category boundaries, revealing subtle affinities among instances belonging to different semantic classes.

2.2 Clustering with membership degree encoder

In the pipeline of deploying diverse datasets for predictive modeling, preparatory data transformation constitutes a fundamental necessity. This phase encompasses categorical variable quantization and numerical standardization to condition the dataset for subsequent analysis.

Categorical Quantization: The raw data encompasses qualitative attributes, such as operational service classifications, communication protocols, and operational states. To facilitate pattern recognition algorithms in processing these attributes, we employ Integer Coding to transform categorical values into numerical representations. This mapping scheme assigns a distinct integer identifier to each category level (e.g., chromatic values [crimson, emerald, azure] mapped to $[0, 1, 2]$). Relative to Binary Vector Encoding, this approach preserves the original feature space dimensionality without expansion.

Numerical Standardization: This procedure harmonizes the magnitude disparities across heterogeneous variables by projecting them onto a unified scale or statistical distribution, ensuring equitable contribution from all predictors. The current investigation adopts Standard Score transformation for continuous attributes. This technique maintains the empirical distribution shape while relocating values to a normalized interval, thereby facilitating the discovery of authentic data regularities. The computational formulation appears in Eq. (1).

$$z_i = \frac{v_i - \mu_v}{\delta_v} \tag{1}$$

where μ_v denotes the arithmetic mean of attribute v, and δ_v represents its standard deviation.

Optimal Partition Cardinality Identification: The analytical procedure commences with ascertaining the ideal cluster count m using internal validation metrics that evaluate clustering quality based solely on the intrinsic structure of the data. Two commonly employed approaches are the elbow method and the silhouette coefficient.

The elbow method examines the within-cluster sum of squares (WCSS) as a function of the number of clusters k:

$$\text{WCSS}(k) = \sum_{i=1}^{k} \sum_{\boldsymbol{u} \in C_i} |\boldsymbol{u} - \boldsymbol{c}_i|^2 \tag{2}$$

The optimal k is identified at the "elbow point" where the marginal decrease in WCSS begins to diminish, indicating a balance between cluster compactness and model complexity.

Alternatively, the silhouette coefficient $s(\boldsymbol{u})$ for a sample $\boldsymbol{u}$ measures both intra-cluster cohesion and inter-cluster separation:

$$s(\boldsymbol{u}) = \frac{b(\boldsymbol{u}) - a(\boldsymbol{u})}{\max a(\boldsymbol{u}), b(\boldsymbol{u})} \tag{3}$$

where $a(\boldsymbol{u})$ is the mean distance to other samples in the same cluster, and $b(\boldsymbol{u})$ is the mean distance to samples in the nearest neighboring cluster. The average silhouette width across all samples is computed for each candidate k, with higher values indicating better-defined clusters. The cluster count m that maximizes the average silhouette width is selected for subsequent analysis.

Upon establishing the requisite cluster quantity m via these internal validation criteria, we execute the K-means++ procedure to derive the representative point matrix $\mathcal{C}$ encompassing all designated prototypes:

$$\mathcal{C} = [c_1, c_2, \ldots, c_m] \tag{4}$$

For every observation $\boldsymbol{u}_j$ and representative point $\boldsymbol{c}_h$, we quantify the affinity $\phi^h(\boldsymbol{u}_j)$ via exponential distance decay:

$$\phi^h(\boldsymbol{u}_j) = \exp\left(-\|\boldsymbol{u}_j - \boldsymbol{c}_h\|_2^2\right) \tag{5}$$

We assemble an affinity characteristic matrix $\mathcal{M}$ of size $N \times m$, where each entry $\mathcal{M}_{j,h}$ encodes the proximity between observation $\boldsymbol{u}_j$ and prototype $\boldsymbol{c}_h$:

$$\mathcal{M}_{j,h} = \phi^h(\boldsymbol{u}_j) \tag{6}$$

The affinity matrix $\mathcal{M}$ is fused with the primitive characteristic matrix $\mathcal{X}$ to constitute an expanded characteristic matrix $\mathcal{X}^*$:

$$\mathcal{X}^* = [\mathcal{X} \oplus \mathcal{M}] \tag{7}$$

The holistic structure of the reference inference system is illustrated in Figure 3. Examination of this figure reveals that the system comprises K zero-order conditional rules (one per target category) instantiated through representative samples. Although isolated zero-order inference modules function as linear discriminators due to fixed output parameters, their collective orchestration enables intricate nonlinear discrimination, with each constituent specializing in particular local linear configurations. Through adaptive weighting of region-specific module outputs, these localized linear separators amalgamate into a globally nonlinear separation surface. This architecture achieves complex pattern discrimination without explicit high-order terms or kernel mappings, balancing computational efficacy with explanatory transparency. The conditional rule adopts this structure:

$$\begin{aligned}\mathbf{Q}^k : &\text{IF } (s \approx r_1^k) \text{ OR } (s \approx r_2^k) \text{ OR } \ldots \text{ OR } (s \approx r_{M^k}^k) \\ &\text{THEN (category } k)\end{aligned} \tag{8}$$

where $\boldsymbol{s}$ denotes an input instance from the augmented characteristic set $\mathcal{X}^*$, the symbol "$\approx$" signifies proximity, $\boldsymbol{r}_\ell^k$ identifies the ℓ-th representative of category k, and M_k indicates the total representatives for that category. These representatives constitute maximally informative exemplars that assist in maintaining structural fidelity.

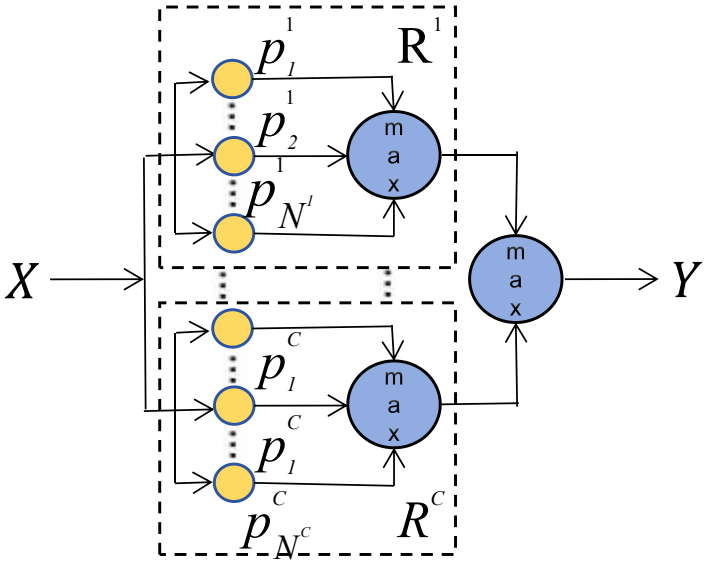

Figure 3. Architecture of the SOFIS[15][17].

Inspection of Eq. (8) reveals that conditional premises are compressed into an economical, data-driven, nonparametric prototype formulation. The disjunctive connection of prototypes within $\mathbf{Q}^k$ permits its interpretation as a concurrent aggregation of elementary rules ($\ell = 1, 2, \ldots, M^k$):

$$\mathbf{Q}^k : \text{IF } (s \approx r_\ell^k) \text{ THEN (category } k) \tag{9}$$

The incorporation of partition-derived characteristics furnishes supplementary signals that enhance the model's structural comprehension [13]. Within adaptive boosting frameworks, misclassified instances receive elevated emphasis in successive iterations, potentially inducing excessive sensitivity to anomalous cases and degrading overall performance. The injection of proximity-based characteristics furnishes additional relational context, enabling more robust discrimination when handling ambiguous instances.

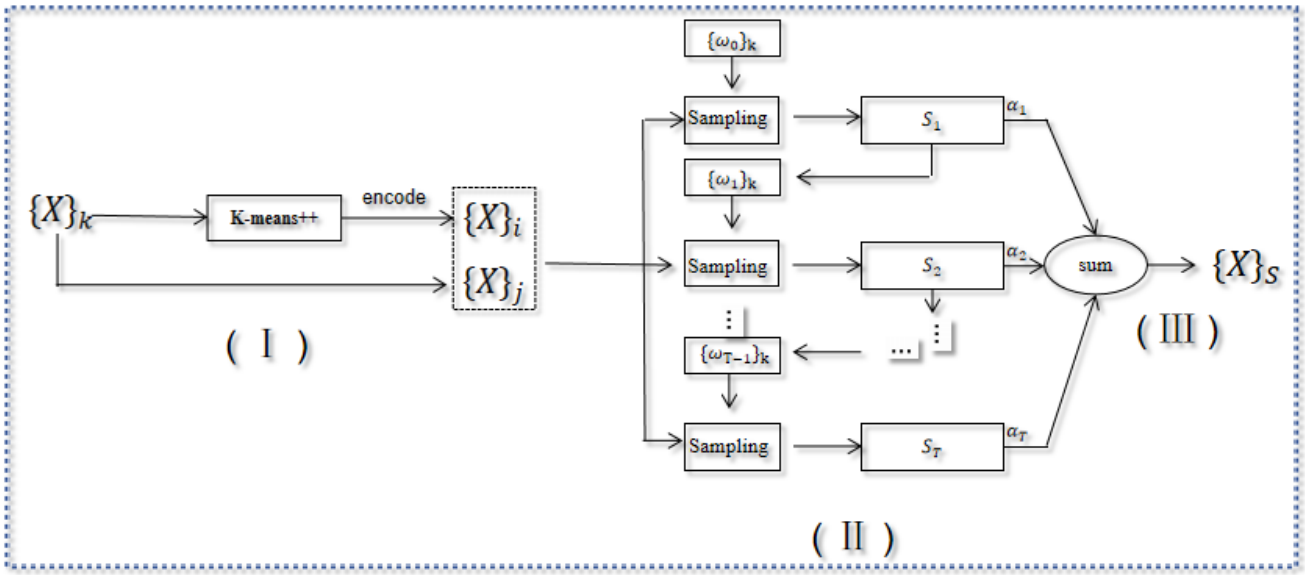

Figure 4. The KFWAdaBoost framework operates as follows: (I) KPP-based feature extraction enriches the original feature space; (II) The FWAdaBoost ensemble learner is trained on the resulting enhanced feature representations; (III) Soft label aggregation is performed by weighting outputs from all base classifiers to obtain per-class weighted soft label assignments.

2.3 Soft label learning

Within the adaptive ensemble framework, each constituent learner computes a proximity-based reliability measure $\psi(u)$ for every instance u, quantifying the certainty of that instance's assignment to its anticipated category. This reliability index is formulated as:

$$\psi^k(u_j) = \max_{r \in \{r\}^k_{M^k}} \left(\exp\left(- \|u_j - r\|_2^2 \right) \right). \tag{10}$$

where r_k denotes the representative prototype for class k; elevated values of this metric indicate diminished distance between instance u and prototype r_k, thereby signifying heightened membership confidence for class k.

Each base learner generates reliability estimates for unannotated evaluation instances, with these estimates subsequently contributing to the weighted aggregation of probabilistic labels. For a specific unannotated instance u_j, the t-th learner yields a predicted category designation $\hat{\omega}_{t,j}$. This prediction is then transformed into a K-dimensional encoding vector $\hat{\Omega}_{t,j} = [\hat{\Omega}^1_{t,j}, \hat{\Omega}^2_{t,j}, \ldots, \hat{\Omega}^K_{t,j}]^{\mathrm{T}}$ through the following mapping, where $k = 1, 2, \ldots, K$:

$$\hat{\Omega}^k_{t,j} = \begin{cases} 1, & \text{if } k = \hat{\omega}_{t,j}, \\ -\frac{1}{K-1}, & \text{otherwise.} \end{cases} \tag{11}$$

The ultimate output (probabilistic labels) emergent from the learning procedure is computed through:

$$g^k(\beta_j) = \sum_{t=1}^{T} \beta_t \hat{\theta}_{t,j} \hat{\Omega}^k_{t,j} \tag{12}$$

$$\hat{\theta}_{t,j} = \psi_t^{\hat{\omega}_{t,j}}(u_j) - \psi_t^{\hat{k}^*}(u_j) \tag{13}$$

$$\psi_t^{\hat{k}^*}(u_j) = \max_{\substack{k=1,2,\ldots,K; \\ k \neq \hat{\omega}_j}} \left(\psi_t^k(u_j) \right) \tag{14}$$

Here, $\hat{\theta}_{t,j}$ represents the margin between the highest and second-highest reliability scores, while $\psi_t^{\hat{k}^*}$ denotes the maximal reliability score among alternative categories for the t-th base learner.

Figure 4 illustrates the architectural mechanism for obtaining weighted probabilistic labels through the proposed methodology. For each instance u, we compute its aggregated reliability scores $g^k(\beta)$ across all categories k, subsequently normalizing these values to derive probabilistic labels $\pi(u)$:

$$\pi_k(u) = \frac{g^k(\beta)}{\sum_{k=1}^{K} g^k(\beta)} \tag{15}$$

Here, $\pi_k(u)$ signifies the membership probability of instance u for class k. These probabilistic labels encode the degree of category affiliation for each instance. This probabilistic label generation benefits from fuzzy-weighted ensemble ideas, akin to recent fuzzy Apriori stacking classifiers that incorporate

adversarial enhancements for better handling of uncertainty [16].

Finally, the primitive feature vector $\boldsymbol{u}$ is concatenated with the computed probabilistic labels $\pi(u)$ to construct an augmented representation $\boldsymbol{u}^+$:

$$\boldsymbol{u}^+ = [u_1, u_2, \ldots, u_d, \pi_1(u), \pi_2(u), \ldots, \pi_K(u)]^{\mathrm{T}} \quad (16)$$

where $u_1, u_2, \ldots, u_d$ constitute the original attributes, and $\pi_1(u), \pi_2(u), \ldots, \pi_K(u)$ represent the corresponding probabilistic label components.

2.4 Summary of Innovations

This paper introduces several key innovations to address the challenges of classification tasks, particularly in the handling of hard-to-classify samples and improving the classification accuracy. The main contributions and innovations are as follows:

- By combining the KPP clustering algorithm with the FWAdaBoost ensemble framework, the proposed KFWAdaBoost method enhances the model's ability to distinguish hard-to-classify samples. The clustering step generates similarity features that guide the ensemble classifier to make more informed decisions, particularly for ambiguous or challenging samples.
- The proposed method adopts a two-stage learning process, starting with clustering-based feature enhancement and followed by soft label learning. This progressive approach allows the model to move from coarse feature enhancement to fine-grained soft label learning, resulting in improved classification performance.
- The KFWAdaBoost framework generates weighted soft labels that reflect the confidence of each sample's membership in different classes. These labels are integrated into the input features, providing additional information for classification and enhancing the model's generalization capability.

2.5 Prediction models

In this section, we introduce Three machine learning models, including the SVM, Decision Tree, LDA.

2.5.1 SVM

SVM is a widely used machine learning model. The target of SVM is to find the optimal hyperplane that best separates data points of different classes in a high-dimensional feature space [18]. The support vectors are the critical data points that lie closest to this hyperplane. As a result, they are essential to define the decision boundary. By maximizing the margin (the distance between the hyperplane and the nearest support vectors), it aims to achieve strong generalization while reducing the risk of overfitting. This margin-maximization principle makes it especially effective in scenarios where the data is high-dimensional or where only a small number of training samples are available.

2.5.2 Decision Tree

A Decision Tree is a tree-based supervised learning method commonly used for classification and regression tasks [19]. It recursively partitions the dataset into increasingly homogeneous subsets based on feature values, forming a set of interpretable decision rules. Each internal node represents a test on a feature, each branch corresponds to an outcome of the test, and each leaf node holds a class label (for classification) or a predicted value (for regression). Decision Trees are intuitive, easy to interpret, and require minimal data preprocessing. However, a single Decision Tree is prone to overfitting and can be highly sensitive to small variations in the training data.

2.5.3 LDA

Linear Discriminant Analysis (LDA) is a classical supervised method for dimensionality reduction and classification. LDA seeks linear combinations of features that maximize the separation between different classes by maximizing the between-class scatter while minimizing the within-class scatter. Unlike unsupervised techniques such as PCA, LDA explicitly uses class labels to guide the projection. Under assumptions of normally distributed classes with equal covariance matrices, LDA produces optimal linear decision boundaries. In addition to classification, LDA is widely used for feature extraction and visualization, particularly in problems with a small number of classes and relatively high-dimensional input features.

2.6 Soft Label Learning-based Framework for Student Performance Prediction

Figure 4 shows the process of the proposed soft label learning framework for student performance prediction. The process begins with inputting a student performance dataset. As the educational dataset contain different types of values, data pre-processing is required to prepare the data for analysis. Specifically, feature encoding is

applied to convert non-numeric categorical data into numeric representations, and Z-score normalization is performed to scale numeric features to a standard distribution. After that, data shuffling is conducted to introduce randomness and eliminate potential ordering biases. Subsequently, the KPP method is applied to divide the training samples into m clusters and calculate the similarity between each sample and cluster centers. These similarity features are then merged with the original features to form an enhanced feature set. Next, the enhanced dataset is input into the FWAdaBoost ensemble classifier for training. During this process, each base classifier generates confidence scores for the samples, which are weighted and aggregated to produce soft labels representing the probabilistic membership of each sample to different categories. These soft labels are then combined with the original features to create a further enriched dataset. Finally, the model is trained on this enriched dataset to obtain the final prediction.

3 Experimental results and analysis

In this section, the description of the datasets is first introduced. Next, the evaluation metrics are provided. Finally, the experimental results and discussions are presented.

3.1 Data description

To validate the performance of our proposed student performance prediction model, we downloaded the educational dataset from the UCI Repository [21]. This dataset covers Mathematics and Portuguese Language subjects. Each dataset includes student grades alongside demographic, social, and school-related features. Specifically, the Mathematics dataset comprises 395 samples. The Portuguese Language datasets contains 649 samples. Both datasets have 30 input features and one output feature for final exam grades. Detailed statistical descriptions of the dataset are provided in Table 1.

3.2 Evaluation metrics

To evaluate the effectiveness of our proposed framework, we adopted four standard classification metrics: Accuracy (Acc), Precision (Pre), Recall (Rec), and F1-Score (F1). For a comparative analysis of model efficiency, we also recorded the computation time (in seconds) required for hyper-parameter tuning via grid search with 10-fold cross-validation [20]. Our objective is to maximize all four metrics, as higher values indicate better predictive performance. Their equations are as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{17}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{18}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{19}$$

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{20}$$

where TP , TN , FP , and FN denote the number of true positives, true negatives, false positives, and false negatives, respectively. These metrics collectively assess both the correctness and completeness of positive class predictions, with F1-Score providing a harmonic balance between Precision and Recall.

3.3 Experimental results

Figure 5 visualizes the experimental results reported in Table 2, comparing the performance of three baseline models (LDA, Decision Tree, and SVM) and their enhanced counterparts (LDA_P, Tree_P, and SVM_P) across four evaluation metrics. For all models, the enhanced versions consistently improve Accuracy, Precision, Recall, and F1-Score compared to their baselines, indicating enhanced classification performance and greater prediction stability.

Table 2 presents a comprehensive comparison of the baseline classifiers—LDA, SVM, and Decision Tree—and their enhanced counterparts (denoted as LDA_P, SVM_P, and Tree_P) on the Mathematics dataset, evaluated across four standard classification metrics. The results reveal a consistent and statistically meaningful improvement across all models after integration with the proposed KFWAdaBoost-based soft label learning framework. Notably, the Decision Tree exhibits the most dramatic gains: its Recall surges from 45.45% to 63.04%, while F1-Score increases by 2.38 percentage points (from 47.62% to 50.00%). This substantial boost in Recall—without a significant drop in Precision (50.00% → 51.43%)—suggests that the framework effectively mitigates the tree model's historical tendency to under-predict minority or borderline classes, a common issue in educational data where failing students often form a small but critical subgroup.

The improvements observed in LDA and SVM, though more modest, are equally instructive. LDA_P achieves a balanced enhancement across all metrics,

Table 1. Statistical properties of the student performance dataset.

Feature	Symbol	Type	Values
School	school	Binary	"GP" (Gabriel Pereira); "MS" (Mousinho da Silveira)
Sex	sex	Binary	"1" (male); "2" (female)
Age	age	Numeric	15–22 (years)
Address	address	Binary	"1" (urban); "2" (rural)
Family Size	famsize	Binary	"1" (≤ 3 members); "2" (> 3 members)
Parent Status	Pstatus	Binary	"2" (living together); "1" (living apart)
Mother's Education	Medu	Numeric	0 (none), 1 (primary), 2 (5th–9th grade), 3 (secondary), 4 (higher education)
Father's Education	Fedu	Numeric	0 (none), 1 (primary), 2 (5th–9th grade), 3 (secondary), 4 (higher education)
Mother's Job	Mjob	Nominal	"1" (teacher), "2" (health care), "3" (civil services), "4" (at home), "5" (other)
Father's Job	Fjob	Nominal	"1" (teacher), "2" (health care), "3" (civil services), "4" (at home), "5" (other)
School Choice Reason	reason	Nominal	"home" (close to home), "reputation", "course" (course preference), "other"
Guardian	guardian	Nominal	"mother", "father", "other"
Travel Time	traveltime	Numeric	1 (< 15 min), 2 (15–30 min), 3 (30 min–1 h), 4 (> 1 h)
Study Time	studytime	Numeric	1 (< 2 h/week), 2 (2–5 h), 3 (5–10 h), 4 (> 10 h)
Past Failures	failures	Numeric	0, 1, 2, or 4 (≥ 3 failures)
School Support	schoolsup	Binary	"yes"; "no"
Family Support	famsup	Binary	"yes"; "no"
Paid Extra Classes	paid	Binary	"yes"; "no"
Extracurricular Activities	activities	Binary	"yes"; "no"
Nursery Attendance	nursery	Binary	"yes"; "no"
Pursue Higher Education	higher	Binary	"yes"; "no"
Home Internet Access	internet	Binary	"yes"; "no"
Romantic Relationship	romantic	Binary	"yes"; "no"
Family Relationship Quality	famrel	Numeric	1 (very bad) – 5 (excellent)
Post-School Free Time	freetime	Numeric	1 (very low) – 5 (very high)
Socializing with Friends	goout	Numeric	1 (very low) – 5 (very high)
Weekday Alcohol Consumption	Dalc	Numeric	1 (very low) – 5 (very high)
Weekend Alcohol Consumption	Walc	Numeric	1 (very low) – 5 (very high)
Current Health Status	health	Numeric	1 (very bad) – 5 (very good)
School Absences	absences	Numeric	0–93 (days)
First-Period Grade (Target)	G1	Numeric	0–20 (course-specific: Math or Portuguese)
Second-Period Grade	G2	Numeric	0–20 (course-specific: Math or Portuguese)
Final Grade (Output Target)	G3	Numeric	0–20 (course-specific: Math or Portuguese)

All features follow the original encoding in the UCI Student Performance datasets. Numeric codes for categorical variables (e.g., sex, address, jobs) are preserved as defined in the source. The target variable for prediction is typically G3 (final grade), though G1 and G2 may also be used in sequential modeling.

reflecting the benefit of incorporating similarity-aware features into a linear discriminant space that otherwise assumes Gaussian class distributions. SVM_P, meanwhile, shows notable gains in Recall (53.19% → 56.82%) and F1 (42.74% → 43.86%), indicating that the soft labels help the kernel-based model better navigate ambiguous regions near decision boundaries—precisely where hard-label training often leads to overconfident yet incorrect classifications.

These trends are robustly replicated on the Portuguese Language dataset (Table 2), further validating the generalizability of our approach. Intriguingly, the magnitude of improvement is even more pronounced in the Portuguese course, particularly for Tree_P, which nearly doubles its Recall (18.68% → 36.17%) and F1-Score (25.56% → 39.25%). This disparity between subjects likely stems from differences in grade distribution: the Portuguese dataset exhibits greater class imbalance and higher variance in student performance, making it more susceptible to

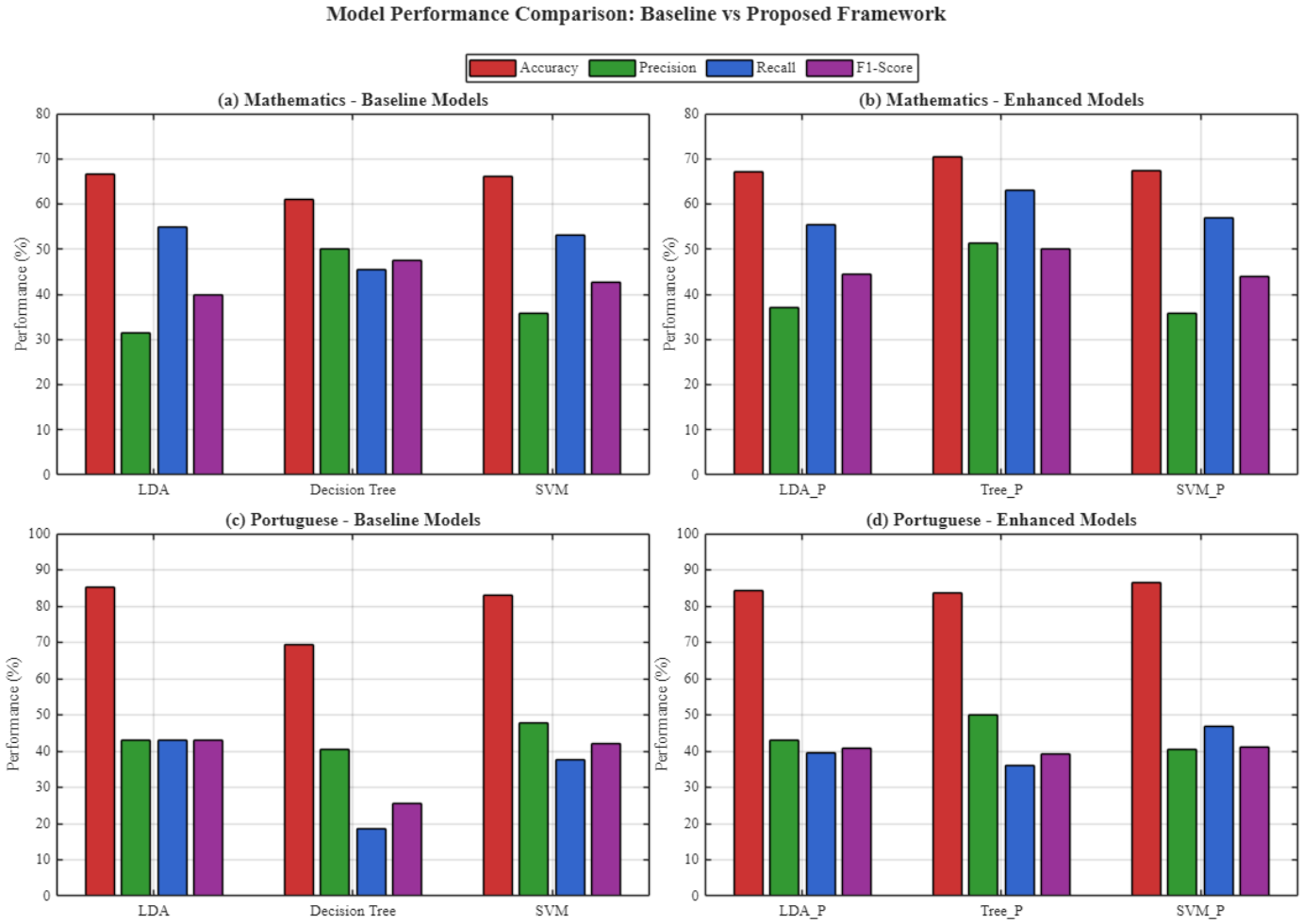

Figure 5. Experimental results based on the results in Table 2.

Table 2. Classification performance (%) for early prediction of student performance in Mathematics and Portuguese courses using base models (LDA, Decision Tree, SVM) and their enhanced versions (LDA_P, Tree_P, SVM_P) based on the proposed soft-label learning framework. The best result for each metric and dataset is highlighted in **bold**. All values are percentages (%) rounded to two decimal places.

	LDA / LDA_P				**Decision Tree / Tree_P**				**SVM / SVM_P**			
Dataset	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
Mathematics Course												
`math_g1`	66.50	31.43	55.00	40.00	60.91	50.00	45.45	47.62	65.99	35.71	53.19	42.74
	67.01	**37.14**	**55.32**	**44.44**	**70.56**	**51.43**	**63.04**	**50.00**	**67.51**	35.71	**56.82**	**43.86**
`math_g12`	81.73	76.47	72.22	74.29	64.98	69.12	49.47	57.67	81.22	76.47	71.23	73.76
	82.23	**77.94**	**74.60**	**75.18**	**76.14**	66.18	**66.10**	**65.69**	**82.23**	76.47	**74.24**	**74.45**
`math_g123`	91.37	90.91	84.51	87.59	76.14	80.30	60.92	69.28	85.28	71.21	82.46	76.42
	91.37	90.91	84.51	87.59	**88.83**	**83.33**	**85.00**	**83.08**	**88.83**	**77.27**	**87.93**	**82.26**
Portuguese Course												
`por_g1`	85.19	42.86	42.86	42.86	69.44	40.48	18.68	25.56	83.03	47.62	37.74	42.11
	84.26	42.86	39.53	40.91	**83.64**	**50.00**	**36.17**	**39.25**	**86.42**	40.48	**46.67**	**40.96**
`por_g12`	88.89	**84.21**	51.61	64.00	77.47	57.89	27.85	37.61	88.89	81.58	51.67	63.27
	88.89	84.21	51.61	64.00	**89.51**	**73.68**	**55.88**	**58.95**	**89.20**	78.95	**52.73**	62.50
`por_g123`	91.36	84.09	63.79	72.55	79.63	47.73	32.81	38.89	89.20	75.00	57.90	65.35
	92.28	84.09	**67.27**	**74.75**	**91.67**	**86.36**	**71.80**	**73.79**	**90.43**	70.45	**63.27**	**66.67**

Note: For each dataset (e.g., `math_g1`), the best value across all six models (LDA, LDA_P, DTree, Tree_P, SVM, SVM_P) is bolded for each evaluation metric (Accuracy, Precision, Recall, F1-Score). All values are percentages (%) rounded to two decimal places.

misclassification under hard-label supervision. The soft label framework, by encoding graded membership degrees rather than binary outcomes, provides a more nuanced supervisory signal that better captures the continuum of academic performance—especially valuable in early prediction scenarios where G1/G2 grades alone offer limited discriminative power.

Figure 5 visually reinforces these findings, showing that the enhanced models uniformly dominate their baselines across all four metrics. The consistent upward shift in both Precision and Recall for Tree_P indicates a genuine improvement in predictive quality, not merely a trade-off shift along the Precision–Recall curve. This is a crucial distinction: many techniques improve one metric at the expense of another, but our framework enhances both, suggesting a fundamental refinement of the decision boundary through enriched feature representation and probabilistic labeling.

The success of Tree_P can be attributed to a synergistic alignment between the model's inductive bias and the proposed framework's design. Decision trees inherently rely on local feature splits and are highly sensitive to label noise; the clustering stage of KFWAdaBoost introduces global structural priors (via similarity to centroids), while the soft label stage replaces brittle 0/1 assignments with continuous confidence scores. Together, these components transform the learning task from "classifying isolated points" to "navigating a similarity-guided probability landscape," enabling the tree to make more context-aware decisions. In contrast, LDA—a global linear method—benefits less because it cannot easily exploit the localized cluster structures, while SVM—though kernelized—still operates under hard-margin assumptions unless explicitly modified.

Nevertheless, limitations remain. For instance, LDA_P shows minimal improvement in Accuracy on the Portuguese dataset, suggesting that linear separability may already be near-optimal in that setting, leaving little room for enhancement. Moreover, the framework's reliance on clustering implies sensitivity to the choice of k (number of clusters); future work could explore adaptive cluster selection or density-based alternatives. Despite these caveats, the consistent cross-dataset, cross-model improvements underscore the framework's practical value: by embedding soft, similarity-informed labels into the learning process, we enable early-warning systems to identify at-risk students with greater sensitivity and reliability—ultimately supporting timely pedagogical interventions.

4 Conclusion

This study proposes a KFWAdaBoost-based soft label learning framework for student performance prediction, which enhances baseline machine learning models by integrating K-means++ clustering with FWAdaBoost and introducing a soft label learning mechanism. The framework operates in two synergistic stages: the first stage generates similarity features through KPP clustering to provide structural awareness of data patterns; the second stage utilizes ensemble confidence scores to produce probabilistic soft labels, refining decision boundaries for students with borderline performance. Experimental results on Mathematics and Portuguese Language course datasets demonstrate that the proposed framework consistently improves the predictive performance of baseline models including LDA, Decision Tree, and SVM across all evaluation metrics (Accuracy, Precision, Recall, and F1-Score). Notably, Decision Tree benefits most significantly from the enhancement, with Recall and F1-Score improving by 17.59 and 2.38 percentage points respectively on the Mathematics dataset, and Recall nearly doubling on the Portuguese dataset. These improvements indicate that the framework effectively mitigates the tendency of traditional classifiers to under-predict hard-to-classify samples, particularly minority or borderline students. LDA and SVM also exhibit stable performance gains, validating the general applicability of clustering-guided feature enhancement and soft label embedding. These findings highlight the potential of the proposed framework as a reliable tool for student performance prediction, particularly suitable for educational scenarios requiring early identification of at-risk students. Beyond student performance prediction, the framework can be extended to broader educational data mining tasks (e.g., early warning systems for at-risk students, personalized curriculum and learning path recommendations) and other classification application domains facing challenges of class imbalance and ambiguous decision boundaries. While the framework demonstrates strong performance, several directions merit further exploration: (1) investigating adaptive cluster selection or density-based alternatives to replace the preset k; (2) integrating deep learning models to explore synergies with soft label learning; (3) incorporating multi-source heterogeneous data (e.g., sequential learning behaviors, social media sentiment) for richer predictions; (4) extending

to continual learning settings to adapt to evolving student behaviors; and (5) enhancing interpretability of cluster prototypes and soft labels to provide more actionable insights for educators.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

This study uses a public, anonymized UCI dataset with no direct human involvement. Ethics approval and informed consent are not required.

References

[1] Bai, X., Zhang, F., Li, J., Guo, T., Aziz, A., Jin, A., & Xia, F. (2021). Educational big data: Predictions, applications and challenges. *Big Data Research, 26*, 100270. [CrossRef]

[2] Rabelo, A., Rodrigues, M. W., Nobre, C., Isotani, S., & Zárate, L. (2024). Educational data mining and learning analytics: A review of educational management in e-learning. *Information Discovery and Delivery, 52*(2), 149–163. [CrossRef]

[3] Kalita, E., Oyelere, S. S., Gaftandzhieva, S., Rajesh, K. N., Jagatheesaperumal, S. K., Mohamed, A., ... & Ali, T. (2025). Educational data mining: a 10-year review. *Discover Computing, 28*(1), 81. [CrossRef]

[4] Hemdanou, A. L., Sefian, M. L., Achtoun, Y., & Tahiri, I. (2024). Comparative analysis of feature selection and extraction methods for student performance prediction across different machine learning models. *Computers and Education: Artificial Intelligence, 7*, 100301. [CrossRef]

[5] Öz, E., Bulut, O., Cellat, Z. F., & Yürekli, H. (2025). Stacking: An ensemble learning approach to predict student performance in PISA 2022. *Education and Information Technologies, 30*(6), 7753-7779. [CrossRef]

[6] Cao, W., & Mai, N. (2025). Predictive analytics for student success: AI-driven early warning systems and intervention strategies for educational risk management. *Educational Research and Human Development, 2*(2), 36-48.

[7] Bañeres, D., Rodríguez-González, M. E., Guerrero-Roldán, A. E., & Cortadas, P. (2023). An early warning system to identify and intervene online dropout learners. *International Journal of Educational Technology in Higher Education, 20*(1), 3. [CrossRef]

[8] Maiya, A. K., & Aithal, P. S. (2023). A review-based research topic identification on how to improve the quality services of higher education institutions in academic, administrative, and research areas. Maiya, AK, & Aithal, PS,(2023). A Review based Research Topic Identification on How to Improve the Quality Services of Higher Education Institutions in Academic, Administrative, and Research Areas. *International Journal of Management, Technology, and Social Sciences (IJMTS), 8*(3), 103-153.

[9] Fan, Z., Gou, J., & Wang, C. (2025). An error complementarity-based iterative learning approach via categorical boosting for student performance prediction. *Engineering Applications of Artificial Intelligence, 161*, 112192. [CrossRef]

[10] Ahmed, E. (2024). Student performance prediction using machine learning algorithms. *Applied computational intelligence and soft computing, 2024*(1), 4067721. [CrossRef]

[11] Zhang, P., Jia, Y., & Shang, Y. (2022). Research and application of XGBoost in imbalanced data. *International Journal of Distributed Sensor Networks, 18*(6), 15501329221106935. [CrossRef]

[12] Arslan, E., Gaftandzhieva, S., Gorgani Firouzjaei, A., Hassannataj Joloudari, J., & Doneva, R. (2025). Ex-ADA: a SHAP-based explainable AdaBoost framework for predicting at-risk students. *Frontiers in Education, 10*, 1728070. [CrossRef]

[13] Piernik, M., & Morzy, T. (2021). A study on using data clustering for feature extraction to improve the quality of classification. *Knowledge and Information Systems, 63*(7), 1771–1805. [CrossRef]

[14] Kapoor, A., & Singhal, A. (2017, February). A comparative study of K-Means, K-Means++ and Fuzzy C-Means clustering algorithms. In *2017 3rd international conference on computational intelligence & communication technology (CICT)* (pp. 1-6). IEEE. [CrossRef]

[15] Gu, X., Angelov, P., & Rong, H. J. (2019). Local optimality of self-organising neuro-fuzzy inference systems. *Information Sciences, 503*, 351-380. [CrossRef]

[16] Xie, R., Chung, F. L., & Wang, S. (2026). Fuzzy Apriori classifier enhanced by stacking and adversarial knowledge assistance. *Information Fusion, 125*, 103483. [CrossRef]

[17] Gu, X., & Angelov, P. P. (2018). Self-organising fuzzy logic classifier. *Information Sciences, 447*, 36-51. [CrossRef]

[18] Tanveer, M., Tiwari, A., Akhtar, M., & Lin, C. T. (2025). Enhancing imbalance learning: A novel slack-factor fuzzy SVM approach. *IEEE Transactions on Emerging Topics in Computational Intelligence, 9*(4), 3112-3121. [CrossRef]

[19] Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. *Expert Systems with Applications, 134*, 93–101. [CrossRef]

[20] Wong, T. T., & Yeh, P. Y. (2019). Reliable accuracy estimates from k-fold cross validation. *IEEE Transactions on Knowledge and Data Engineering, 32*(8), 1586-1594. [CrossRef]

[21] Cortez, P., & Silva, A. M. G. (2008). Using data mining to predict secondary school student performance. In *Proceedings of 5th Annual Future Business Technology Conference*.

Zhihong Yu received the master's degree from the College of Computer Science and Technology, Huaqiao University, Xiamen, China, in 2022. He is currently a lecturer with the Fuzhou Technology and Business University.His main research interests include data mining analysis and ensemble learning. (Email: 22014083067@hqu.stu.edu.cn)