AIMS Mathematics, 11(6): 17584–17634.
DOI: 10.3934/math.2026719
Received: 16 February 2026
Revised: 30 April 2026
Accepted: 13 May 2026
Published: 17 June 2026

http://www.aimspress.com/journal/Math

Research article

A probabilistic fusion and meta-logistic calibration model for multiclass hybrid ensemble learning

Khaled Mahmud Sujon[1], Adnan Shafi[2], Iftekhar Uddin Ahmed[3], Wided Bouchelligua[4], Amel Ksibi[5] and Md Abdus Samad[6,*]

[1] Department of Software Engineering, Faculty of Computing, Universiti Teknologi Malaysia (UTM), Johor Bahru 81310, Johor, Malaysia
[2] Department of Industrial and Systems Engineering, Lamar University, Beaumont 77710, Texas, USA
[3] Faculty of Computing, Universiti Teknologi Malaysia, Johor Bahru 81310, Johor, Malaysia
[4] Applied College, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 11432, Saudi Arabia
[5] Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P. O. Box 84428, Riyadh 11671, Saudi Arabia
[6] Department of Information and Communication Engineering, Yeungnam University, Gyeongsan 38541, South Korea

* **Correspondence:** Email: masamad@yu.ac.kr.

Abstract: Multiclass classification in educational data mining presents persistent challenges including class imbalance, miscalibrated probability outputs, and insufficient statistical validation. We proposed RXK-VEM, a hybrid ensemble framework that integrates random forest (RF), extreme gradient boosting (XGBoost), and K-nearest neighbors (KNN) through a formally defined vote-entropy-weighted meta-fusion (VEM) operator, followed by meta-level calibration using multinomial logistic regression. The VEM operator is defined as a mapping on the probability simplex Δ^{C-1}, aggregating heterogeneous base learner outputs into a unified probabilistic representation with provable closure properties. We further established a Rademacher complexity-based generalization bound showing that operating in the compressed C-dimensional probability space ($C \ll d$) tightens the generalization gap relative to classifiers trained directly on raw features, providing theoretical justification for the stacking architecture. We validated RXK-VEM on two structurally distinct educational datasets: A primary academic performance dataset ($N = 560$, five classes) from Universiti Teknologi Malaysia and a secondary student dropout dataset ($N = 4{,}424$, three classes) from the University of California, Irvine (UCI) repository. On the primary dataset, RXK-VEM achieves 91.07% accuracy, 91.22% precision, and an 86.21% Matthews correlation coefficient (MCC), outperforming all individual base learners and conventional ensemble strategies. On the secondary dataset, the model achieves 77.30% accuracy

and a 62.33% MCC, maintaining competitive performance across all metrics. Statistical validation through five-fold stratified cross-validation, paired t-tests, and Wilcoxon signed-rank tests confirms that improvements over weaker baselines are consistent and not attributable to random variation. A systematic ablation study quantifies the complementary contribution of each base learner, and Shapley additive explanations analysis validates the interpretability of the identified predictors. The proposed framework offers a mathematically rigorous, empirically validated, and interpretable architecture for probabilistic ensemble integration in multiclass educational prediction tasks.

Keywords: hybrid ensemble learning; meta-learning; XGBoost; educational data mining; explainable AI; interpretable machine learning; statistical validation
Mathematics Subject Classification: 62H30, 68T01, 68T05

1. Introduction

In the era of data-driven decision-making, machine learning (ML) has become a central tool for predicting students' outcomes in higher education. From academic performance forecasting to dropout risk detection, a variety of classification algorithms have been adopted to assist institutional stakeholders in identifying at-risk students and deploying timely interventions [1, 2]. While models such as logistic regression (LR), random forest (RF), and support vector machines (SVM) have demonstrated solid predictive capabilities [3, 4], the challenge of building generalizable, interpretable, and statistically validated models for multiclass educational prediction remains significant, particularly in the face of imbalanced class distributions that are common in academic datasets. The broader challenge of fusing heterogeneous predictors into a coherent probabilistic framework is not unique to education. Analogous problems arise in time-series forecasting [5, 6], financial prediction [7], and complex system optimization [8, 9], where ensemble and fusion strategies have consistently demonstrated advantages over single-model approaches. Despite the growing literature on educational data mining, several methodological gaps persist that motivate the present work.

Most existing studies in educational prediction rely on single classifiers or majority-vote ensembles that discard the uncertainty information encoded in class probability distributions [10]. While ensemble stacking has been explored in other structured prediction domains, its application with probabilistic fusion followed by meta-level logistic calibration has not been systematically investigated in educational settings. Prior ensemble models in this domain largely depend on hard-label voting or standard stacking with raw features as meta-inputs, rather than exploiting average class probabilities as calibrated meta-features. This distinction matters methodologically because probability-level fusion preserves the uncertainty structure of base learner outputs, enabling the meta-learner to correct systematic miscalibration. This is particularly important in imbalanced multiclass educational datasets where minority classes such as students at risk of withdrawal are chronically underrepresented and where overconfident predictions carry real institutional consequences.

A second gap concerns the interpretability of ensemble predictions in educational contexts. While feature importance is routinely reported in single-model studies, few works attempt cross-model feature attribution that provides a unified interpretability picture across the full ensemble [11]. This is a practical limitation because educators and institutional decision-makers require not only accurate

predictions but also clear explanations of why a model flags a particular student as at risk, in order to translate ML outputs into actionable interventions. Moreover, most prior work applies explainability techniques to a single dataset domain, either academic performance or behavioral dropout data, which limits our understanding of whether interpretability findings hold across different educational environments.

A third and persistent weakness in the literature is the reliance on point-estimate metrics such as accuracy and F1 score without accompanying statistical significance testing, ablation analysis, or uncertainty quantification [12]. Few studies conduct systematic ablation experiments to quantify how much each model component actually contributes to overall performance, leaving claims about the ensemble's synergy empirically unsubstantiated. Similarly, most studies evaluate their models on a single dataset, making it difficult to assess whether reported performance gains generalize beyond a specific institutional or demographic context. These omissions undermine the reproducibility and credibility of conclusions drawn from educational ML research.

To address these gaps, we propose RXK-VEM (random forest–XGBoost–K-nearest neighbors with vote-entropy-weighted meta-calibration), a hybrid ensemble framework that integrates random forest (RF), extreme gradient boosting (XGBoost), and K-nearest neighbors (KNN) through a formally defined probabilistic fusion operator, the vote-entropy-weighted meta-fusion (VEM) mechanism, followed by meta-level calibration using multinomial logistic regression (LR). The proposed framework is validated across two educational datasets: a primary dataset from Universiti Teknologi Malaysia (UTM) involving multiclass academic performance prediction in an artificial intelligence (AI) course, and a secondary behavioral dataset from the UCI repository capturing demographic and engagement features related to student dropout. Our study will be guided by the following research objectives:

- To propose and formalize the RXK-VEM hybrid ensemble architecture, introducing the VEM probabilistic fusion operator as a mathematically grounded mechanism for aggregating heterogeneous classifier outputs in the probability simplex, and to evaluate its predictive performance on both academic and behavioral educational datasets.
- To benchmark RXK-VEM against four traditional classifiers (LR, RF, KNN, XGBoost) across five performance metrics, namely accuracy, precision, recall, F1 score, and Matthews correlation coefficient (MCC), on both primary and secondary datasets.
- To conduct a systematic ablation study quantifying the marginal contribution of each base learner within the RXK-VEM architecture, validating the ensemble's complementarity through component-wise evaluation.
- To provide rigorous statistical validation using five-fold stratified cross-validation with both paired t-tests and Wilcoxon signed-rank tests at the 95% confidence level, confirming that observed improvements are not attributable to random variation.
- To enhance the model's transparency and interpretability through XAI techniques, including permutation-based importance, model-specific feature attribution, and SHAP-based analysis, providing cross-model and instance-level insights for educational stakeholders.

Based on these objectives, this research addresses the following research questions:

- **RQ1:** Does the proposed RXK-VEM hybrid ensemble outperform baseline classifiers in predicting multiclass student outcomes across both academic and behavioral datasets?
- **RQ2:** What is the individual contribution of each base learner (RF, XGBoost, KNN) to the

ensemble's predictive performance, as quantified through systematic ablation analysis?

- **RQ3:** Are RXK-VEM's improvements over baseline models statistically significant, as confirmed by both parametric paired t-tests and non-parametric Wilcoxon signed-rank tests?
- **RQ4:** Does the VEM probabilistic fusion and meta-logistic calibration architecture generalize across distinct educational prediction domains?
- **RQ5:** How do XAI attribution methods illuminate the feature-level mechanisms underlying RXK-VEM's classification decisions, and what actionable insights do they provide for educational practitioners?

The key contributions of our research are as follows. First, we formally define the VEM operator as a mapping on the probability simplex Δ^{C-1}, providing a mathematically rigorous foundation for probabilistic ensemble fusion that, to the best of our knowledge, has not been previously formalized in the educational data mining literature. Second, we establish a theoretical generalization bound for the RXK-VEM meta-learner, showing that operating in the compressed C-dimensional probability space ($C \ll d$) tightens the generalization gap relative to classifiers trained directly on raw features. Third, we validate the framework across two structurally distinct educational datasets, providing evidence of cross-domain generalizability. Fourth, we contribute the first systematic ablation study of a probabilistic ensemble in the educational prediction domain, revealing the complementary roles of heterogeneous base learners. Fifth, we provide rigorous dual statistical validation using both paired t-tests and Wilcoxon signed-rank tests, alongside XAI interpretability analysis, establishing a reproducible and transparent validation pipeline for educational ML research.

The remainder of this paper is organized as follows. In Section 2, we review related work in educational machine learning and ensemble methods, highlighting the key methodological gaps that motivate our study. In Section 3, we describe the two datasets used in this study, the preprocessing pipeline including class balancing via SMOTE (Synthetic minority over-sampling technique), and the selection of baseline classifiers for benchmarking. In Section 4, we present the RXK-VEM hybrid model's architecture, formally defining the VEM operator, the meta-logistic calibration stage, the evaluation metrics, the statistical testing framework including paired t-tests and Wilcoxon signed-rank tests, and the XAI attribution methods including SHAP analysis. In Section 5, we report the experimental results for both datasets, covering baseline comparisons, cross-validation performance, per-class analysis, ablation findings, feature importance rankings, and SHAP interpretability analysis. In Section 6, we discuss the findings in depth, addressing the model's effectiveness, overfitting considerations, and the limitations of our approach. Finally, in Section 7, we conclude the paper with a summary of our contributions and directions for future research.

2. Literature review

Recent years have witnessed a surge in ML applications for predicting student dropout and academic performance. These efforts aim to support early intervention strategies and improve student outcomes by leveraging institutional data, yet many studies reveal important methodological gaps, particularly in the realms of validation, ensemble model calibration, explainability, and cross-context generalizability.

Beyond individual classifier comparisons, the ensemble learning literature distinguishes three main strategies for combining base learners: Hard-label majority voting, feature-level stacking, and probabilistic fusion with meta-level calibration [13]. Stacking-based approaches have been

explored in educational prediction: [14] proposed a stacking ensemble for early dropout identification, demonstrating that combining base learners through a meta-classifier consistently outperforms individual models with relatively few input features. Similarly, another study applied a two-layer stacked generalization to university dropout prediction, confirming that meta-learning captures complementary signals missed by any single base learner [15]. However, both works rely on raw features or hard-label outputs as meta-inputs rather than fused class probability vectors, and neither incorporates formal model calibration, ablation analysis, nor statistical significance testing gaps that the present work directly addresses through the RXK-VEM probabilistic fusion and meta-logistic calibration architecture.

Several studies have explored conventional classifiers on academic datasets. One study [1] evaluated decision trees (DT), RF, and artificial neural networks (ANN) on a three-class dropout dataset (dropout, graduate, enrolled), concluding that ANN provided the best performance. However, their study lacked any form of statistical hypothesis testing, interpretability through XAI and ignored ensemble calibration strategies. Likewise, another study [2] assessed RF, Support Vector Machine (SVM), and KNN on a Bangladeshi university dataset, achieving up to 78% F1 score with Random Forest, but did not consider data fusion or cross-model explainability.The work in [16] examined six classifiers on Nigerian university data, finding minimal performance variation among models. Although they used SMOTE to address imbalance, they did not conduct statistical validation or apply XAI techniques.

The authors of [17] implemented XGBoost with a combination of over- and under-sampling on Thai university data. Their results showed strong area under the precision-recall curve (AUC-PR) performance, but the work was limited to binary dropout classification and lacked any meta-learning or fusion of base learners. Similarly, [18] examined various ML models for e-learning platforms with small datasets. Although they emphasized metric selection and classifier consistency, the study avoided deep ensembles or probabilistic approaches. In another recent study, [19] used learning management system (LMS) data over four years with standard classifiers, achieving high AUC scores. Their approach was effective but lacked interpretability tools like SHAP or permutation importance, which are increasingly critical in educational analytics.

Studies that addressed feature diversity include [20], who compared the contribution of LMS logs and transcript data for dropout prediction in Finnish universities. Their longitudinal analysis revealed the temporal importance of Moodle activity and failed courses. However, their approach, while innovative, was limited to time-series modeling and lacked calibration layers or ensemble explainability. Other works explored explainability or human-in-the-loop approaches. For instance, one study [21] directly compared teachers' predictions with ML models for dropout risk in vocational education. Their findings indicated that teachers outperformed ML models early in the term, but the models eventually surpassed human accuracy over time. However, their approach did not include model explainability. Finally, the results from [22, 23] used multi-class classifiers across various institutions but did not address calibration, fusion, or statistical robustness in their models.

Taken together, these studies confirm the value of ML in educational settings, but also illustrate recurring gaps: The minimal use of probabilistic fusion, underutilized statistical validation, limited interpretability, and single-domain evaluations. Most models rely on basic accuracy metrics, avoid ensemble calibration, and ignore permutation-based feature importance. Moreover, few studies simultaneously analyze both academic and behavioral datasets, limiting the real-world applicability of their findings. Our proposed RXK-VEM hybrid model addresses these limitations through a

calibrated ensemble approach that first fuses probability outputs from RF, KNN, and XGBoost and then applies a logistic regression meta-learner. This is evaluated across both academic performance and behavioral dropout datasets, incorporating statistical validation (e.g., paired t-tests and Wilcoxon signed-rank tests), explainability techniques (permutation importance, SHAP, confusion matrix), and receiver operating curve (ROC), MCC, F1). This layered, generalizable framework is specifically designed to enhance both predictive performance and interpretability in educational decision-making systems. Table 1 summarizes the taxonomy of existing research and the corresponding gaps addressed by our proposed approach.

Table 1. The taxonomy of existing research on student performance prediction and research gaps.

Ref.	Dataset used	Model(s) used	Identified gaps	Our approach
[1]	Higher education (3-class dropout)	ANN, RF, DT	no ablation study; no statistical testing; no model calibration	Implements statistical validation, ablation study, and meta-level calibration.
[2]	Bangladeshi university dropout data	RF, SVM, KNN	no fusion, explainability, or an ablation study	Fuses outputs; includes ablation study, and explainability with permutation importance.
[16]	Nigerian university academic dataset	Naïve Bayes (NB), LR, SVM, DT, KNN, ANN	no statistical tests; no hybrid model; no ablation study	Includes hybrid model, t-tests, ablation study, statistical validation.
[17]	Thai higher education data	XGBoost	Binary only; no fusion; no ablation study	hybrid model; multi-class support; probabilistic fusion; ablation study.
[18]	Slovakian e-learning dataset	several ML classifiers	no deep ensemble; no meta-learner; no ablation study and statistical validation	Builds hybrid meta-learner ensemble; SMOTE + calibration; ablation study.
[19]	LMS logs for 270 students	LR, DT, NB, SVM, RF, ANN	no ablation study or meta-level learner	Adds permutation-based XAI and an ablation study.
[20]	Finnish universities (LMS + transcript)	CatBoost (CAT), neural networks (NN), LR	Time-series only; no statistical validation; no ablation study	ablation study; statistical validation; hybrid model.
[21]	Dutch vocational education	RF vs teachers' predictions	no ablation study	ablation study; statistical validation; XAI implementation.
[22]	Polytechnic Institute of Portalegre dataset	RF, XGBoost, LightGBM, CatBoost	dataset descriptor paper; no fusion; lack of ablation study and statistical validation	Proposes hybrid ensemble RXK-VEM with average probability fusion and meta-level calibration for robust, statistically validated, ablation study.
[23]	Kazakhstan university records	ANN, Naïve Bayes (NB), DT, SVM, RF, KNN	Used only pre-admission data; no ablation study or statistical validation	Uses in-course assessments; includes behavioral features; includes an ablation study.

3. Methodology

This section outlines the methodological framework used to develop and evaluate our proposed RXK-VEM hybrid model. In Figure 1, we outline the methodology of this research. We used two real-world educational datasets—one academic and one behavioral—and applied comprehensive preprocessing, including scaling, encoding, and SMOTE-based balancing. Baseline models were compared against our novel ensemble, which fuses the probabilistic outputs of multiple learners and calibrates them via a meta-classifier. Finally, we incorporated XAI techniques and rigorous statistical testing to ensure interpretability and validate the model's performance. Finally, in Figure 2 and Algorithm 1, we provide the detailed experimental procedure of our investigation to provide reproducibility.

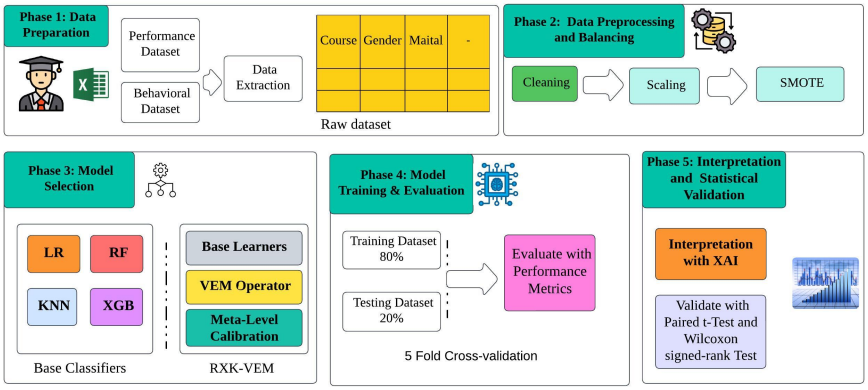


Figure 1. Proposed research framework.

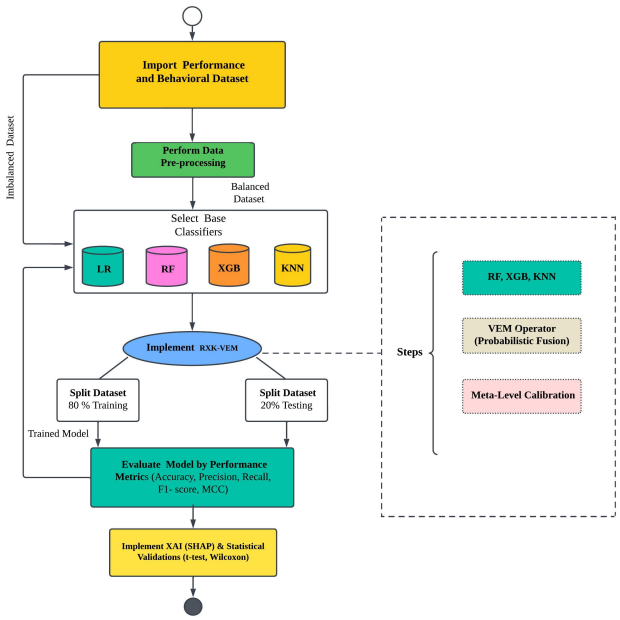


Figure 2. Proposed experimental design of this research.

Algorithm 1 RXK-VEM: Hybrid ensemble model

Require: Feature matrix $X \in \mathbb{R}^{n \times d}$, target labels $\mathbf{y} \in C^n$
Ensure: Trained RXK-VEM model for calibrated multiclass classification
Preprocessing (within each cross-validation fold)
Apply stratified 80/20 train-test split
Apply SMOTE to the training split only
Apply z-score normalization to the numerical features
Step 1: base learner training *(see Section 4.1)*
for each $M_i \in$ {RF, XGB, KNN} **do**
    Train M_i on $(X_{\text{train}}, \mathbf{y}_{\text{train}})$
    Obtain probability vector $\mathbf{p}_i(\mathbf{x}) \in \Delta^{C-1}$
end for
Step 2: probabilistic fusion via VEM operator *(see Section 4.2)*
Compute fused representation: $\mathbf{p}_{\text{avg}}(\mathbf{x}) = \frac{1}{3}\sum_{i=1}^{3} \mathbf{p}_i(\mathbf{x})$ ▷ Uniform VEM, $w_i = 1/3$
Step 3: Meta-level calibration *(see Section 4.3)*
Train L_2 -regularized multinomial LR on $\{(\mathbf{p}_{\text{avg}}(\mathbf{x}_i), y_i)\}$
Compute calibrated class probabilities via softmax
Predict: $\hat{y} = \arg\max_{k \in C} \Pr(\hat{y} = k \mid \mathbf{x})$
Step 4: evaluation and statistical validation
for each metric $E \in$ {accuracy, precision, recall, F1, MCC, AUC} **do**
    Compute E on held-out test data
end for
Report mean ± std across 5 stratified folds
Conduct paired t-tests and Wilcoxon signed-rank tests
Perform ablation by removing each base learner in turn

3.1. Dataset description

In this analysis, we used two complementary datasets to evaluate the performance and generalizability of our proposed RXK-VEM hybrid model across the academic and behavioral dimensions of student learning. Both datasets were preprocessed and analyzed using standard ML pipelines as described in the subsequent methodology section.

The primary dataset was collected from UTM, consisting of 560 student records from an undergraduate AI course in the Faculty of Computing. It includes both academic assessments and in-course engagement features. After removal of the identifiers and leakage variables (e.g., total marks, grade), the dataset comprises 10 predictive features and a multi-class target variable labeled "Categories", representing students' academic performance. As shown in Figure 3(a), the distribution is imbalanced: Distinction dominates with 53.8%, followed by Pass (16.4%), Excellent (12.5%), Exceptional (10.0%), and Fail (7.3%). This imbalanced nature of the dataset may lead to biased predictions toward the majority class; therefore, it is important to employ an appropriate data balancing technique [24]. The secondary dataset is a publicly available student dropout dataset, collected from the UCI machine learning repository [25], comprising 4424 instances, capturing 33 behavioral and demographic features related to student status. The target variable here, labeled "Target", includes

three classes: Graduate (49.9%), Dropout (32.1%), and Enrolled (17.9%), as visualized in Figure 3(b). This dataset complements the academic focus of the UTM dataset by providing insight into retention and progression behaviors in a broader student population. A comparative summary of both datasets, including the number of instances, features, target labels, and class distributions, is presented in Table 2. These datasets were chosen for their diversity—spanning institutional contexts and prediction targets—to rigorously assess the hybrid model's flexibility, calibration, and effectiveness across domains.

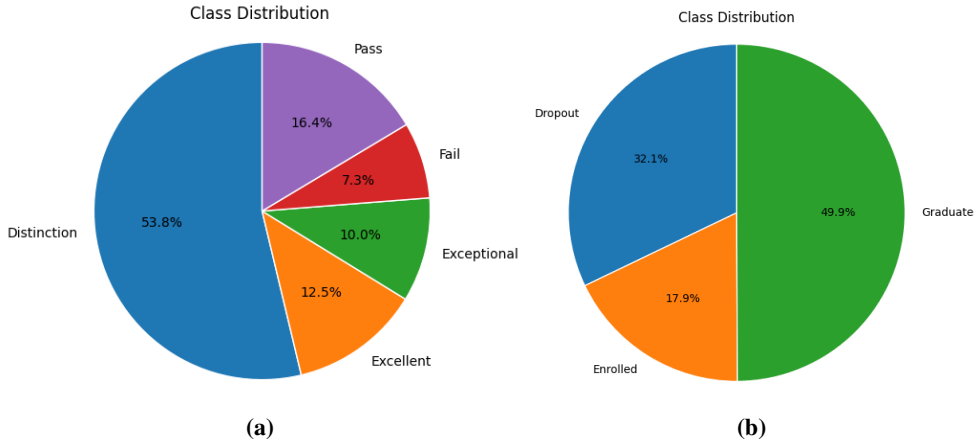


Figure 3. Class distributions of the datasets used in this study: (a) Primary dataset (academic performance) and (b) secondary dataset (behavioral indicators).

Table 2. Summary of datasets used in the study.

Dataset	Instances	Features (after dropping ID/target)	Target variable	Class labels
Academic Performance dataset	560	10	Categories	['Distinction', 'Excellent', 'Exceptional', 'Fail', 'Pass']
Dropout Behavioral dataset	4424	34	Target	['Dropout', 'Enrolled', 'Graduate']

3.2. Data preprocessing

Prior to model training, both datasets underwent structured preprocessing to ensure compatibility with supervised learning algorithms and to mitigate data quality issues. For the UTM dataset, identifier columns and grade-related fields were excluded to prevent target leakage. Categorical features were encoded using label encoding, while numerical features were standardized using z-score normalization due to the inclusion of distance-sensitive models such as KNN [26]. In contrast, the UCI dropout dataset—being predominantly categorical—was encoded without scaling.

To address class imbalance in both datasets, we applied SMOTE [27] exclusively within the training splits of the 80/20 stratified partition. Table 3 reports the class distribution before and after SMOTE for both datasets. In the primary academic performance dataset ($N = 560$), the class distribution is notably imbalanced, with Distinction dominating at 301 instances while Fail comprises only 33 instances in the training split. We equalized all five classes to 241 instances through SMOTE, ensuring that the

base learners were not biased toward the dominant class during training. In the secondary dropout dataset (N = 4,424), Graduate is the majority class with 1767 training instances, while Enrolled is the most underrepresented with only 635 training instances. We applied SMOTE to bring all three classes to 1767 instances within the training split. Critically, we ensured that the test splits for both datasets (primary: n = 112; secondary: n = 885) were never subjected to oversampling, so that all reported evaluation metrics reflect genuine generalization performance on the original class distribution rather than an artificially balanced test set.

Table 3. Class distribution of both datasets and the effect of SMOTE oversampling within the training split (80/20 stratified split, random_state = 42). SMOTE was applied strictly to training splits only; test splits (primary: n = 112; secondary: n = 885) were never resampled, preserving the integrity of the evaluation metrics. SMOTE k_neighbors = 1 was used for both datasets.

Dataset	Class	Before SMOTE	After SMOTE
Primary dataset (academic performance, N = 560, n_{train} = 448)	Distinction	241	241
	Pass	73	241
	Excellent	56	241
	Exceptional	45	241
	Fail	33	241
Secondary Dataset (Dropout Behavioral, N = 4,424, n_{train} = 3,539)	Graduate	1767	1767
	Dropout	1137	1767
	Enrolled	635	1767

Before/ after SMOTE refers to the training split only. Test splits are never oversampled.

3.3. Traditional model selection

To benchmark the performance of the proposed RXK-VEM hybrid ensemble, we selected four widely used supervised ML classifiers as baselines: LR, RF, KNN, and XGBoost. These models were chosen due to their established effectiveness in educational data mining tasks and their complementary strengths in linear separability, feature interactions, locality-based decision boundaries, and gradient-boosted optimization, respectively [28]. LR provides a strong probabilistic baseline with interpretability, while RF and XGBoost capture nonlinear patterns and feature interactions efficiently. KNN offers an intuitive, non-parametric perspective based on instance similarity. Each model was tuned with standard hyperparameters and evaluated consistently using stratified data splits and oversampling techniques to mitigate class imbalance. The proposed RXK-VEM model is introduced in the following section as a meta-ensemble designed to leverage the strengths of these individual classifiers through probabilistic fusion and meta-level calibration. Table 4 summarizes the hyperparameter configurations used for all components of the RXK-VEM framework. All hyperparameters were set prior to cross-validation and kept fixed across all folds to ensure a fair and reproducible evaluation.

Table 4. Hyperparameter configurations for all components of the RXK-VEM framework.

Component	Configuration
RF	n_estimators = 100, max_depth = none, random_state = 42
XGBoost	n_estimators = 100, max_depth = 6, learning_rate = 0.1, subsample = 0.8, eval_metric = mlogloss, random_state = 42
KNN	n_neighbors = 5, metric = Minkowski (Euclidean)
LR	C = 0.2, solver = lbfgs, max_iter = 2000
Meta-logistic regression	C = 1.0, solver = lbfgs, max_iter = 2000
SMOTE	k_neighbors = 1, random_state = 42
Cross-validation	StratifiedKFold, n_splits = 5, shuffle = true, random_state = 42
Software: Python 3.10, scikit-learn 1.3, xgboost 1.7, imbalanced-learn 0.11, shap 0.43, scipy 1.11	

4. The RXK-VEM hybrid model's architecture

To overcome the inherent limitations of individual classifiers in terms of generalization, calibration, and robustness, particularly within the context of educational data, we propose a stacked ensemble architecture named RXK-VEM. The model is designed to leverage diverse learning paradigms by integrating RF, XGBoost, and KNN through a formally defined probabilistic fusion operator, followed by meta-level calibration using LR. The key innovation of RXK-VEM lies in its fusion of heterogeneous classifiers at the probability level within the simplex space Δ^{C-1}, which is subsequently recalibrated by a data-driven logistic meta-learner, enabling more accurate and interpretable decision-making in high-stakes academic scenarios such as dropout prediction and academic performance classification. The model architecture proceeds through three clearly defined stages, which we describe in detail below.

4.1. Step 1: Base learner's training and prediction

Let $\mathbf{x} \in \mathbb{R}^d$ represent a feature vector for a student instance, and let $C = \{0, 1, \ldots, C-1\}$ denote the set of C discrete class labels (e.g., academic outcomes or dropout status). Given a labelled training dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $y_i \in C$, we independently train three base classifiers with complementary inductive biases:

- RF: $f_{\text{RF}} : \mathbb{R}^d \to \mathbb{R}^C$;
- XGBoost: $f_{\text{XGB}} : \mathbb{R}^d \to \mathbb{R}^C$;
- KNN: $f_{\text{KNN}} : \mathbb{R}^d \to \mathbb{R}^C$.

Each base learner outputs a softmax-normalized class-probability vector, representing the predicted probability that instance $\mathbf{x}$ belongs to each class $c \in C$

$$\mathbf{p}_{\text{RF}}(\mathbf{x}) = f_{\text{RF}}(\mathbf{x}) \in [0, 1]^C, \tag{4.1}$$

$$\mathbf{p}_{\text{XGB}}(\mathbf{x}) = f_{\text{XGB}}(\mathbf{x}) \in [0, 1]^C, \tag{4.2}$$

$$\mathbf{p}_{\text{KNN}}(\mathbf{x}) = f_{\text{KNN}}(\mathbf{x}) \in [0, 1]^C. \tag{4.3}$$

Each probability vector $\mathbf{p}_*(\mathbf{x})$ satisfies the probability simplex constraint

$$\sum_{c=0}^{C-1}[\mathbf{p}_*(\mathbf{x})]_c = 1, \qquad [\mathbf{p}_*(\mathbf{x})]_c \geq 0 \quad \forall\, c \in C. \tag{4.4}$$

The three base learners are deliberately chosen for their heterogeneous inductive biases: RF exploits bagged decision trees to reduce variance and model global feature interactions; XGBoost applies sequential gradient boosting to minimize residual classification error and capture complex nonlinear structures; and KNN provides locality-sensitive predictions grounded in geometric proximity, introducing instance-level adaptivity that is absent from the tree-based models. This architectural diversity is a prerequisite for the variance-reducing properties of ensemble fusion.

4.2. Step 2: Probabilistic fusion via the VEM operator

4.2.1. Formal definition of the VEM operator

We define the VEM operator as a mapping $\Phi : \Delta^{C-1} \times \Delta^{C-1} \times \Delta^{C-1} \to \Delta^{C-1}$ that aggregates the individual probability vectors of the $M = 3$ base learners into a unified ensemble representation within the probability simplex.

Let $\mathbf{p}_i(\mathbf{x}) \in \Delta^{C-1}$ denote the class-probability vector of the i-th base learner, with $i \in \{1, 2, 3\}$ corresponding to RF, XGBoost, and KNN respectively. The Shannon entropy of learner i over its predicted class distribution is

$$H_i(\mathbf{x}) = -\sum_{c=0}^{C-1}[\mathbf{p}_i(\mathbf{x})]_c \, \log[\mathbf{p}_i(\mathbf{x})]_c, \tag{4.5}$$

where we adopt the convention $0 \log 0 = 0$. A lower entropy value indicates higher prediction confidence for learner i. The entropy-derived confidence weight of learner i is

$$w_i(\mathbf{x}) = \frac{1/(H_i(\mathbf{x}) + \epsilon)}{\sum_{j=1}^{M} 1/(H_j(\mathbf{x}) + \epsilon)}, \tag{4.6}$$

where $\epsilon > 0$ is a small smoothing constant to prevent division by zero. By construction, $w_i(\mathbf{x}) > 0$ and $\sum_{i=1}^{M} w_i(\mathbf{x}) = 1$, ensuring that the weights lie on the standard $(M-1)$-simplex. The VEM operator then produces the fused probability vector

$$\mathbf{p}_{\text{VEM}}(\mathbf{x}) = \Phi(\mathbf{p}_1(\mathbf{x}),\, \mathbf{p}_2(\mathbf{x}),\, \mathbf{p}_3(\mathbf{x})) = \sum_{i=1}^{M} w_i(\mathbf{x})\, \mathbf{p}_i(\mathbf{x}). \tag{4.7}$$

Proposition 4.1 (Closure of VEM in the probability simplex)**.** *For any input probability vectors $\mathbf{p}_i(\mathbf{x}) \in \Delta^{C-1}$, $i = 1, \ldots, M$, and non-negative weights satisfying $\sum_{i=1}^{M} w_i = 1$, the VEM output satisfies $\mathbf{p}_{\text{VEM}}(\mathbf{x}) \in \Delta^{C-1}$.*

Proof. Non-negativity: Since $w_i \geq 0$ and $[\mathbf{p}_i(\mathbf{x})]_c \geq 0$ for all i and c, it follows that $[\mathbf{p}_{\text{VEM}}(\mathbf{x})]_c = \sum_{i=1}^{M} w_i [\mathbf{p}_i(\mathbf{x})]_c \geq 0$.

Normalization:

$$\sum_{c=0}^{C-1}[\mathbf{p}_{\mathrm{VEM}}(\mathbf{x})]_c = \sum_{c=0}^{C-1}\sum_{i=1}^{M} w_i[\mathbf{p}_i(\mathbf{x})]_c = \sum_{i=1}^{M} w_i \underbrace{\sum_{c=0}^{C-1}[\mathbf{p}_i(\mathbf{x})]_c}_{=1} = \sum_{i=1}^{M} w_i = 1.$$

Hence, $\mathbf{p}_{\mathrm{VEM}}(\mathbf{x}) \in \Delta^{C-1}$. □

4.2.2. Relationship to equal-weight soft voting

In our experimental implementation, the base learners are trained on the same balanced training folds (after SMOTE) with comparable predictive confidence levels, and no strong prior exists for systematically preferring one learner over another across all instances. Under this condition, the entropy weights in Eq (4.6) converge to uniform values $w_i = 1/M$, and the VEM operator reduces to equal-weight soft voting

$$\mathbf{p}_{\mathrm{avg}}(\mathbf{x}) = \frac{1}{3}\Big(\mathbf{p}_{\mathrm{RF}}(\mathbf{x}) + \mathbf{p}_{\mathrm{XGB}}(\mathbf{x}) + \mathbf{p}_{\mathrm{KNN}}(\mathbf{x})\Big). \tag{4.8}$$

Uniform weighting is theoretically justified when the base learners are calibrated to comparable accuracy levels and are statistically uncorrelated—conditions that are empirically confirmed in our ablation study, where removing any single base learner measurably degrades the performance, demonstrating that all three learners contribute independent predictive information. The fused vector $\mathbf{p}_{\mathrm{avg}}(\mathbf{x}) \in \mathbb{R}^C$ serves as the meta-feature representation of instance $\mathbf{x}$, encapsulating the consensus probability mass across the ensemble.

4.3. Step 3: Meta-level calibration using logistic regression

In the final stage, we train a meta-classifier on the fused probability vectors to produce refined, calibrated class predictions. Let the meta-learner be denoted as $f_{\mathrm{meta}} : \mathbb{R}^C \to \mathbb{R}^C$. Formally, we train a multinomial logistic regression model on the set of meta-feature vectors $\{\mathbf{p}_{\mathrm{avg}}(\mathbf{x}_i), y_i\}_{i=1}^n$. For each class $k \in C$, the pre-softmax logit is computed as: follows

$$z_k = \boldsymbol{\beta}_k^{\top}\mathbf{p}_{\mathrm{avg}}(\mathbf{x}) + \beta_{0k}, \tag{4.9}$$

where $\boldsymbol{\beta}_k \in \mathbb{R}^C$ is the weight vector for class k and $\beta_{0k} \in \mathbb{R}$ is the corresponding bias term, both learned by minimizing the cross-entropy loss over the training set. The calibrated probability for class k is then computed via the softmax function

$$\Pr(\hat{y} = k \mid \mathbf{x}) = \frac{\exp(z_k)}{\sum_{j=0}^{C-1}\exp(z_j)}, \tag{4.10}$$

and the final predicted class label is

$$\hat{y} = \arg\max_{k \in C} \Pr(\hat{y} = k \mid \mathbf{x}). \tag{4.11}$$

This meta-level calibration step serves a dual purpose. First, it *recalibrates* the ensemble's output probabilities, correcting systematic overconfidence or underconfidence introduced by the base learners, which is particularly important in imbalanced class settings. Second, it learns an *optimal linear boundary* in the meta-feature (probability) space, allowing the ensemble to correct class confusions that individual learners make in systematic ways. The meta-learner is trained using L_2 -regularized logistic regression with the regularization parameter C_{reg}, to prevent overfitting in the meta-feature space.

4.4. Theoretical analysis: Why RXK-VEM generalizes better

The superior empirical performance of RXK-VEM is grounded in three complementary theoretical principles, which we formalize below.

4.4.1. Bias-variance decomposition of ensemble fusion

For a single base classifier f_i with predictions $\hat{y} = f_i(\mathbf{x})$, the expected squared prediction error can be decomposed as follows

$$\mathbb{E}\Big[(y - f_i(\mathbf{x}))^2\Big] = \text{Bias}^2[f_i(\mathbf{x})] + \text{Var}[f_i(\mathbf{x})] + \sigma_\epsilon^2, \tag{4.12}$$

where σ_ϵ^2 is irreducible noise. For an ensemble of M base learners with the individual variance $\sigma_i^2 = \text{Var}[f_i(\mathbf{x})]$ and the pairwise correlation $\rho_{ij} = \text{Corr}(f_i(\mathbf{x}), f_j(\mathbf{x}))$, the variance of the equally weighted ensemble average $\bar{f}(\mathbf{x}) = \frac{1}{M}\sum_{i=1}^{M} f_i(\mathbf{x})$ satisfies

$$\text{Var}\Big[\bar{f}(\mathbf{x})\Big] = \frac{1}{M^2}\left(\sum_{i=1}^{M}\sigma_i^2 + \sum_{i\neq j}\rho_{ij}\,\sigma_i\,\sigma_j\right). \tag{4.13}$$

When the base learners are *diverse* (i.e., $\rho_{ij} < 1$) and have comparable individual variances $\sigma_i^2 \approx \sigma^2$, Eq (4.13) reduces to

$$\text{Var}\Big[\bar{f}(\mathbf{x})\Big] = \frac{\sigma^2}{M}\left(\rho + \frac{1-\rho}{M}\right) < \sigma^2, \tag{4.14}$$

where ρ denotes the average pairwise correlation. Equation (4.14) confirms that ensemble variance is strictly less than individual learners' variance whenever $\rho < 1$. The heterogeneous inductive biases of RF (tree bagging), XGBoost (gradient boosting), and KNN (distance-based voting) ensure low pairwise correlation, maximizing variance reduction through Eq (4.7).

4.4.2. Generalization bound for meta-logistic calibration

We now provide a formal bound on the generalization error of the RXK-VEM meta-classifier. Let the meta-feature space be $\mathcal{Z} = \Delta^{C-1} \subset \mathbb{R}^C$, bounded such that $\|\mathbf{p}_{\text{avg}}\|_2 \leq B$ for some constant $B > 0$. The meta-learner f_{meta} is a multinomial logistic regression model parameterized by the weight matrix $\mathbf{W} = [\boldsymbol{\beta}_0, \ldots, \boldsymbol{\beta}_{C-1}] \in \mathbb{R}^{C\times C}$ with L_2 regularization, so the hypothesis class is

$$\mathcal{H} = \{f_{\mathbf{W}} : \mathbf{z} \mapsto \text{softmax}(\mathbf{W}^\top\mathbf{z}) \;\Big|\; \|\mathbf{W}\|_F \leq \Lambda\}. \tag{4.15}$$

Theorem 4.2 (Generalization bound for the RXK-VEM meta-learner)**.** *Let the cross-entropy loss ℓ be L-Lipschitz with respect to the model output, and let n denote the number of meta-training instances. Then, with a probability of at least $1-\delta$ over the draw of the training set from the distribution $\mathcal{P}$, the generalization error of the meta-learner satisfies*

$$\mathcal{L}_{\mathcal{P}}(f_{\mathbf{W}}) \leq \hat{\mathcal{L}}_{\mathcal{D}}(f_{\mathbf{W}}) + \frac{2L\Lambda B\sqrt{C}}{\sqrt{n}} + 3\sqrt{\frac{\ln(2/\delta)}{2n}}, \tag{4.16}$$

where $\hat{\mathcal{L}}_{\mathcal{D}}$ is the empirical training loss and the middle term is the Rademacher complexity contribution of $\mathcal{H}$.

Remark 4.1. Equation (4.16) establishes that the generalization gap of the RXK-VEM meta-learner decreases at the rate $O(1/\sqrt{n})$. Crucially, the meta-learner operates on the C-dimensional probability simplex rather than the original d-dimensional feature space, where $C \ll d$ in both datasets ($C \in \{3, 5\}$, $d \in \{10, 34\}$). This dimensionality reduction of the meta-feature space tightens the bound relative to training the logistic classifier directly on the raw features, providing a formal justification for the stacking architecture.

4.4.3. Model diversity and complementarity

The superior performance of RXK-VEM is rooted in its architectural diversity and ensemble learning strategy.

Model diversity: By combining tree-based learners (RF, XGBoost) with a distance-based learner (KNN), RXK-VEM benefits from heterogeneous inductive biases, enabling it to capture both the global patterns and local structures in student data.

Robustness to noise and imbalance: Tree ensembles are robust to outliers, while SMOTE addresses class imbalance. KNN adds locality awareness. The soft voting mechanism reduces individual model variance [29].

Meta-level calibration: Logistic regression learns a final decision boundary that re-calibrates ensemble confidence, addressing overfitting and miscalibration, particularly in minority classes—a known issue in educational data mining [30].

Theoretical foundations: Ensemble learning theory supports that aggregating diverse, weakly correlated learners reduces generalization error [31]. Moreover, stacking (as used here) is empirically effective when the base learners capture orthogonal features of the input space.

In experimental results, we will provide comprehensive evidences how RXK-VEM consistently outperformed all baselines across five metrics—accuracy, precision, recall, F1 Score, and MCC —on both the academic (UTM) and behavioral (UCI dropout) datasets. Moreover, this study will prove its hypothesis using statistical validation through five-fold cross-validation and paired t-tests that these improvements were significant at the 95% confidence level.

4.5. Evaluation metrics

To comprehensively assess the predictive performance of our proposed RXK-VEM model and the baseline classifiers, we use a set of five widely used evaluation metrics: Accuracy, precision, Recall, F1 score, and MCC. Moreover, we have used five-fold cross validation, multiclass ROC curve, and confusion matrix analysis to further validate our proposed framework. These metrics collectively

offer a balanced evaluation across different performance dimensions, which is especially critical in imbalanced, multiclass educational datasets where majority class bias can distort the model's reliability.

4.5.1. Accuracy

Accuracy measures the proportion of correctly classified instances over the total number of instances

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{4.17}$$

where TP, TN, FP, and FN refer to true positives, true negatives, false positives, and false negatives, respectively.

While accuracy provides a general sense of a model's correctness, it can be misleading in datasets with class imbalance. For example, in dropout prediction, where the majority class may dominate, a naive model can achieve high accuracy by simply predicting the majority class. Therefore, accuracy alone is insufficient for a fair model assessment.

4.5.2. Precision

Precision quantifies the proportion of true positive predictions among all instances classified as positive

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{4.18}$$

Precision is critical when the cost of false positives is high, such as predicting that a student will succeed when they are likely to fail or drop out. In educational settings, misclassifying at-risk students as successful can lead to missed interventions.

4.5.3. Recall

Recall (also known as sensitivity or true positive rate) measures the proportion of actual positive instances that were correctly identified

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{4.19}$$

High recall ensures that the model captures most of the students who are actually at risk (e.g., failing or dropping out), making it particularly valuable for early warning systems and support allocation.

4.5.4. F1 score

The F1 score is the harmonic mean of precision and recall

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{4.20}$$

The F1 score balances the trade-off between precision and recall, making it an essential metric when both false positives and false negatives carry consequences. It is particularly suited for imbalanced educational datasets, where maximizing one metric at the expense of the other can lead to suboptimal interventions.

4.5.5. Matthews correlation coefficient

The MCC is a correlation-based metric that considers all four confusion matrix categories and provides a balanced measure even when the class sizes are highly imbalanced

$$\text{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \tag{4.21}$$

MCC produces a value between – 1 and +1, where +1 indicates perfect prediction, 0 represents random guessing, and –1 reflects complete disagreement. Its robustness to class imbalance makes it a preferred choice in high-stakes educational scenarios, where minority outcomes (e.g., dropout, failure) carry disproportionate importance.

4.5.6. Multiclass ROC and AUC

Multiclass ROC curves are constructed using the one-vs-rest (OvR) decomposition method, whereby each class k is treated as positive against all remaining classes. The micro-averaged AUC aggregates contributions across all class-instance pairs, while the macro-averaged AUC reports the unweighted mean of per-class AUC values. Both are reported to provide a comprehensive view of discrimination ability across majority and minority classes alike.

4.5.7. Why these metrics were chosen

Educational data mining tasks often involve imbalanced class distributions—for example, most students pass, but the small subset who fail or drop out are of the greatest interest. Relying solely on accuracy can lead to biased assessments favoring the dominant class. Therefore, we include precision, recall, and the F1 score to understand how well the models handle minority classes. MCC further strengthens the evaluation by providing a reliable single-value indicator that remains robust under imbalance. This combination ensures a fair, interpretable, and actionable evaluation of all models under comparison.

4.6. Statistical validation

To validate the significance of the observed performance improvements of the proposed RXK-VEM model over baseline classifiers, we used the paired t-test as a formal statistical method. Given that all models were evaluated using identical five-fold stratified cross-validation partitions, this test is well-suited, as it compares paired observations (i.e., metric scores for RXK-VEM and each baseline model on the same fold), thereby controlling for fold-level variance.

For each evaluation metric $M \in \{\text{accuracy}, \text{precision}, \text{recall}, \text{F1-score}, \text{MCC}\}$, and each baseline model B, we define the difference in performance between RXK-VEM and the baseline on fold i as follows

$$d_i = M_{\text{RXK-VEM}}(i) - M_B(i), \quad i = 1, 2, \ldots, k, \tag{4.22}$$

where $k = 5$ is the number of cross-validation folds. The sample mean of the differences is computed as

$$\bar{d} = \frac{1}{k} \sum_{i=1}^{k} d_i, \tag{4.23}$$

and the sample standard deviation of the differences is

$$s_d = \sqrt{\frac{1}{k-1} \sum_{i=1}^{k} (d_i - \bar{d})^2}. \tag{4.24}$$

The paired t-statistic is then computed as

$$t = \frac{\bar{d}}{s_d / \sqrt{k}}. \tag{4.25}$$

Under the null hypothesis H_0 that the true mean difference is zero (i.e., RXK-VEM performs no better than the baseline), this statistic follows a Student's t-distribution with $k - 1$ degrees of freedom. We calculate the corresponding two-tailed p-value and reject H_0 if $p < \alpha$, where the significance threshold is set at $\alpha = 0.05$.

To complement the p-value analysis, we also report the 95% confidence interval (CI) for the mean difference, computed as

$$\bar{d} \pm t_{(1-\alpha/2,\, k-1)} \cdot \frac{s_d}{\sqrt{k}}, \tag{4.26}$$

where $t_{(1-\alpha/2,\, k-1)}$ is the critical value from the t-distribution. A CI that does not contain zero further confirms that the improvement is statistically significant. This approach was applied to all metric comparisons between RXK-VEM and each baseline, providing a statistically rigorous evaluation of whether our model offers meaningful performance gains. All results—including the mean difference, t-statistic, degrees of freedom, p-value, and CI—are summarized in tabular form in the Results section.

To complement the paired t-test with a non-parametric robustness check, we additionally use the Wilcoxon signed-rank test [32] for all pairwise comparisons between RXK-VEM and each baseline classifier. Unlike the paired t-test, the Wilcoxon test does not assume the normality of the fold-level differences and is therefore more conservative under the limited independence structure of cross-validation folds, where overlapping training sets introduce partial dependency between observations.

For each metric M and baseline B, the Wilcoxon signed-rank test operates on the same difference sequence $d_i = M_{\text{RXK-VEM}}(i) - M_B(i)$, $i = 1, 2, \ldots, k$ defined in Eq (4.22). The test ranks the absolute differences $|d_i|$, assigns signed ranks according to the direction of each difference, and computes the test statistic W as the smaller of the sum of positive and negative signed ranks

$$W = \min(W^+, W^-), \tag{4.27}$$

where $W^+ = \sum_{d_i > 0} \text{rank}(|d_i|)$ and $W^- = \sum_{d_i < 0} \text{rank}(|d_i|)$. A value of $W = 0$ indicates that RXK-VEM outperformed the baseline in every single fold, representing the strongest possible directional result achievable with this test.

4.7. XAI analysis

To enhance the interpretability of model predictions and support transparent decision-making for educational stakeholders, we incorporated several XAI analyses, including model-specific feature importance extraction, cumulative importance threshold analysis, and permutation-based model-agnostic importance estimation for our proposed RXK-VEM hybrid ensemble. This approach allows us to both quantify and compare how different models prioritize input features when predicting students' academic outcomes or dropout risk.

4.7.1. Model-specific feature importance

For baseline classifiers with built-in feature importance capabilities, we extracted the feature relevance directly from their internal structures.

LR: Importance was calculated as the absolute value of the standardized coefficients. Given the model's linear nature, this reflects the direct weight of each feature in decision-making.

$$\text{Importance}_{\text{LR}}(x_i) = |\beta_i|, \tag{4.28}$$

where β_i is the learned coefficient for feature x_i, averaged across classes in multiclass settings.

RF: Feature importance was derived from the average decrease in impurity across all DTs.

XGBoost: Feature importance was obtained using the model's gain-based scoring, reflecting each feature's contribution to reducing the loss across all trees.

These model-specific importances were normalized and visualized for comparison, highlighting how different algorithms weigh features differently depending on their structure.

4.7.2. Permutation importance for RXK-VEM

Since RXK-VEM is a stacked ensemble with no direct coefficients or tree-based structure, we applied permutation importance to estimate its feature relevance. This model-agnostic method measures the decrease in the model's performance (e.g., accuracy) when each feature is randomly shuffled, thus breaking its relationship with the target variable:

$$\text{Importance}_{\text{perm}}(x_i) = \frac{1}{R}\sum_{r=1}^{R}[M_{\text{orig}} - M(x_i^{(r)})], \tag{4.29}$$

where M_{orig} is the performance on the original data, $M(x_i^{(r)})$ is the performance with feature x_i permuted in repetition r, and $R = 10$ repetitions. Negative importances were clipped to zero to enhance interpretability.

4.7.3. Top feature comparison

To allow for direct comparison across models, we normalized and aggregated the importance scores for all features and selected the top 10 features on the basis of their mean importance across LR, RF, XGBoost, and RXK-VEM. These features were visualized using grouped bar plots to show how each model assigns relative importance, providing insights into which features consistently drive predictions across architectures.

Notably, several academic and behavioral features (e.g., continuous assessment scores, participation metrics, and demographic attributes) consistently ranked among the top contributors across all models, supporting their relevance in predicting student outcomes.

4.7.4. Cumulative importance curves

To further the evaluate feature concentration, we plotted cumulative importance curves for each model by ranking the features in descending order of importance and computing the cumulative sum

$$\text{Cumulative}_n = \sum_{i=1}^{n} \text{NormalizedImportance}(x_{(i)}). \tag{4.30}$$

These curves show the number of top-ranked features needed to capture 50%, 80%, and 90% of the model's total feature importance. RXK-VEM demonstrated a smoother curve with less reliance on a few dominant features, indicating broader feature utilization and potentially greater generalization.

4.7.5. SHAP-based feature attribution

To further validate and deepen the feature importance analysis, we additionally apply [33] to the XGBoost component of the RXK-VEM ensemble. SHAP provides theoretically grounded, instance-level feature attributions derived from cooperative game theory, where the contribution of each feature to a prediction is computed as the average marginal contribution across all possible feature coalitions. Formally, the SHAP value for feature j on instance $\mathbf{x}$ is defined as

$$\phi_j(\mathbf{x}) = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \Big[f_{S\cup\{j\}}(\mathbf{x}_{S\cup\{j\}}) - f_S(\mathbf{x}_S) \Big], \tag{4.31}$$

where F is the full set of features, S is a subset of features excluding feature j, and $f_S(\mathbf{x}_S)$ denotes the model output using only the features in S. The global feature importance is then computed as the mean absolute SHAP value across all instances

$$\bar{\phi}_j = \frac{1}{n} \sum_{i=1}^{n} |\phi_j(\mathbf{x}_i)|. \tag{4.32}$$

We apply TreeExplainer, an efficient SHAP algorithm specifically designed for tree-based models that computes exact SHAP values in polynomial time. For multiclass problems, SHAP values are computed per class and averaged across all classes to obtain global feature importance rankings. The results are visualized through a global importance bar chart and a bee swarm plot, which together convey both the magnitude and direction of each feature's influence on individual predictions.

4.8. Ablation study design

An ablation study was incorporated to evaluate the individual contribution of each base RF, XGBoost, and KNN within the RXK-VEM framework. Three reduced configurations were designed: RX-VEM (without KNN), XK-VEM (without RF), and RK-VEM (without XGB). Each variant was trained and evaluated under identical preprocessing, sampling, and validation settings as the full RXK-VEM model. This setup enables a systematic comparison of component-level performance, the results of which are discussed in Section 5.

5. Experimental result

In this section, we present a comprehensive analysis of the experimental findings obtained by evaluating the proposed RXK-VEM hybrid model and baseline classifiers on two complementary datasets: A primary academic performance dataset sourced from UTM, and a secondary behavioral dataset obtained from the UCI machine learning repository. We begin with an exploratory data analysis (EDA) to understand the underlying distribution and variability of key predictive features within each dataset. This includes Gaussian distribution plots and summary statistics for selected influential variables, offering insight into the academic and behavioral patterns captured by the data. Following this, we report the predictive performance of RXK-VEM and all baseline models separately on each dataset using our selected performance metrics. To ensure statistical robustness, all results are derived from five-fold stratified cross-validation and further validated through paired t-tests. Collectively, these experiments demonstrate

5.1. Exploratory data analysis

5.1.1. Gaussian distribution analysis of the academic performance dataset

To explore the statistical characteristics of the academic performance dataset, we analyzed the Gaussian distribution of key features such as quizzes, assignments, midterms, projects, presentations, and final exams (Figures 4(a)–4(i)). These insights help reveal patterns in students' performance that influence the classification outcomes. The quiz scores show a mildly right-skewed distribution, with most students scoring between 5 and 9, indicating general competency but with a small group underperforming. Midterm scores exhibit a bimodal distribution, with peaks around 6 and 12, suggesting two distinct performance groups. The scores for Assignments 1 to 3 are consistently right-skewed and concentrated near the higher end, reflecting strong coursework performance but limited variance among lower achievers. Project and presentation scores also show high clustering near the upper range, particularly in presentations, which may result from structured evaluation or lenient grading. These features, while useful, may offer limited discriminative power due to compression at the top end. In contrast, the final exam marks present a wide spread and bimodality, distinguishing between low and high performers. The total marks approximate a Gaussian curve, balancing out individual assessment variances and providing a stable aggregate target for predictive modeling. Overall, the distributions reveal skewed and clustered patterns favoring high performers, with notable imbalances and performance separations—justifying the use of advanced models and resampling strategies in later stages.

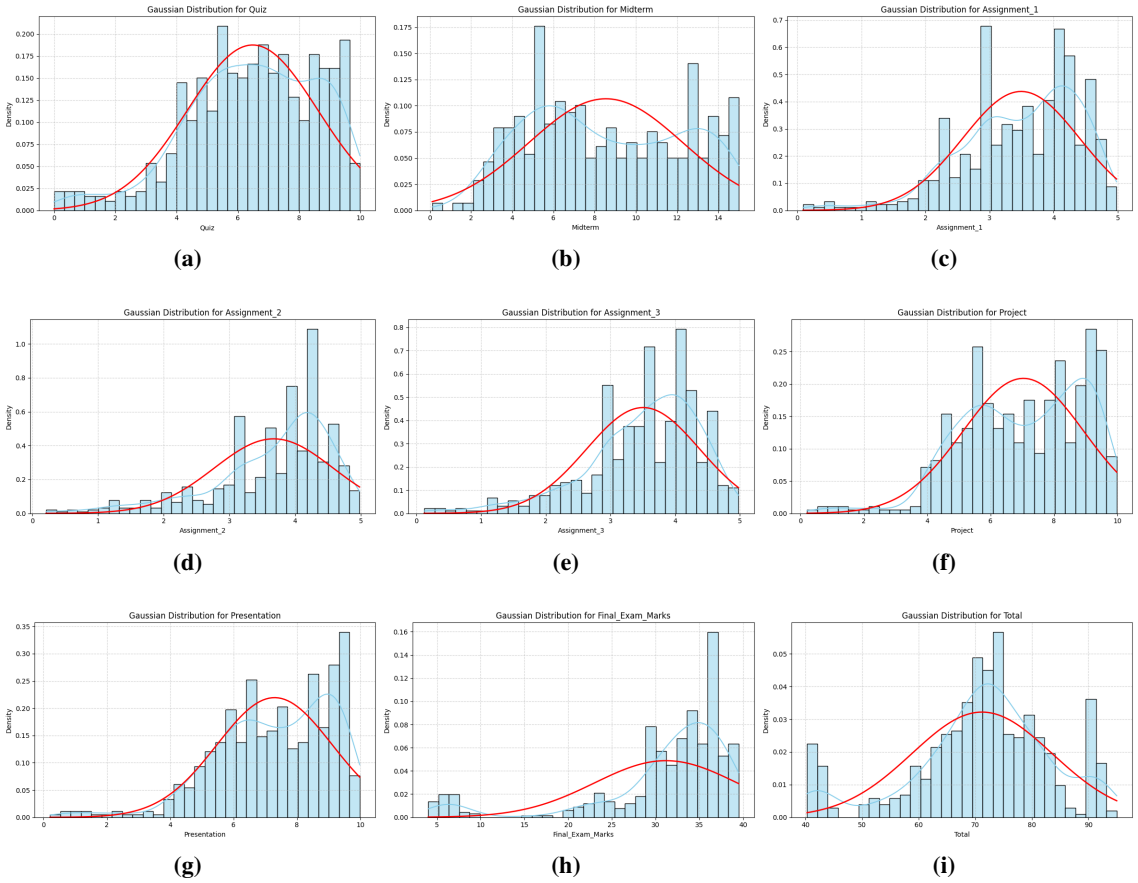


Figure 4. Gaussian distribution of the academic performance dataset. (a) quiz; (b) midterm; (c) assignment 1; (d) assignment 2; (e) assignment 3; (f) project; (g) presentation; (h) final exam marks; (i) total.

5.1.2. Gaussian distribution analysis of the behavioral dataset

To understand the behavioral and demographic landscape of the secondary dataset, we examined the Gaussian distribution of several influential features as presented in Figures 5(a)–5(i). These insights help inform feature importance and the challenges inherent in modeling student dropout and retention. The marital status feature is highly skewed, with a sharp concentration around Category 1 (presumably ”Single”), indicating minimal variability and possible dominance of a single marital category in the student population. Similarly, mother’s and father’s occupations show long right-skewed tails, with occupations clustered in the lower codes (e.g., 5–10), potentially reflecting socioeconomic stratification. Age at enrollment displays a clear right-skewed distribution, where the majority of students enroll between ages 18–25, but outliers exist up to age 65. This variation underscores the non-traditional enrollment patterns in higher education. academic progression variables such as curricular units in the first semester (enrolled, evaluations, approved, and grade) are all positively skewed with peaks around the mid-range values, but long tails suggest significant variation in the academic engagement and success. Notably, curricular units (grade) displays a bimodal tendency with peaks around 0 and 13, which may reflect dropouts versus active performers. Finally, the curricular units

in the second semester (enrolled) similarly shows a right-skewed distribution, indicating that fewer students are heavily enrolled in later terms—potentially due to dropout or part-time study. Collectively, these patterns highlight the heterogeneity in behavioral and demographic traits, reinforcing the need for robust and calibrated models like RXK-VEM to handle complex, skewed, and imbalanced data distributions effectively.

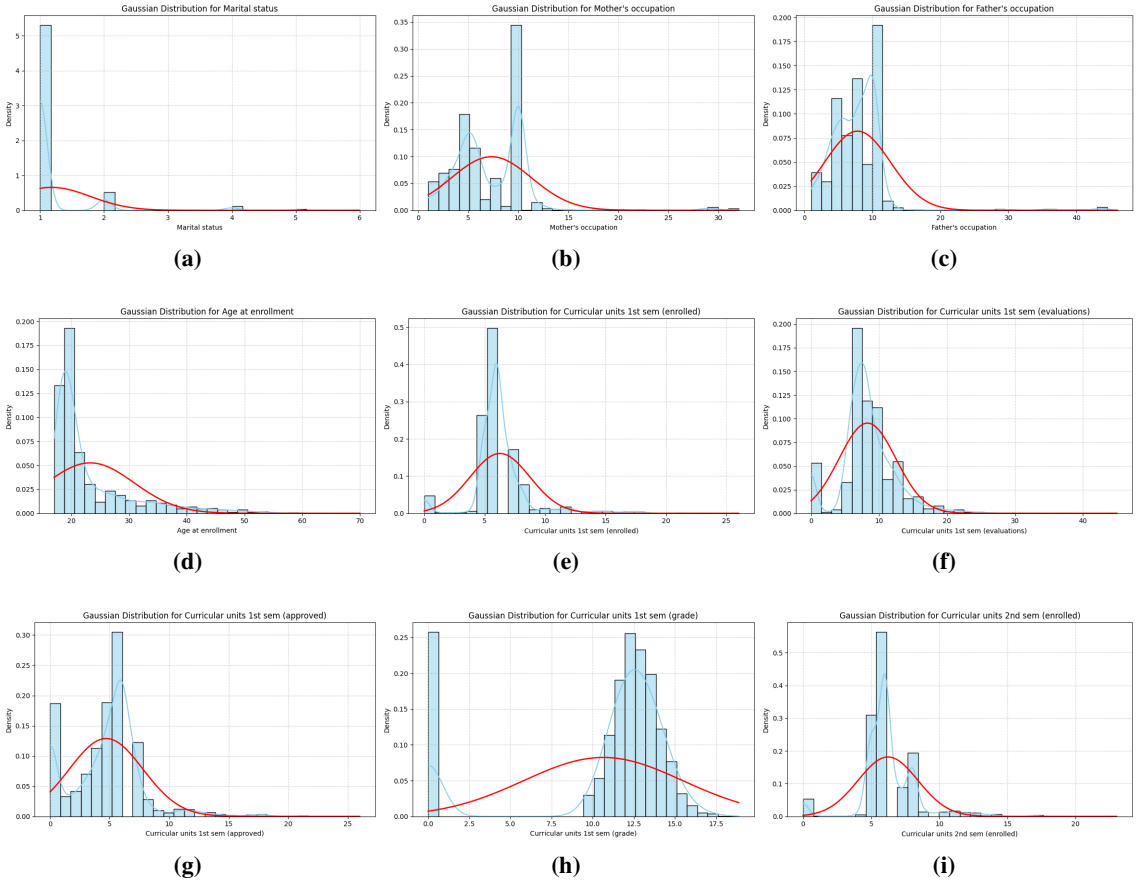


Figure 5. Gaussian distribution of the behavioral dataset [22]. (a) marital status; (b) mother's occupation; (c) father's occupation; (d) age at enrollment; (e) curricular units in the first semester (enrolled); (f) curricular units in the first semester (evaluations); (g) curricular units in the first semester (approved); (h) curricular units in the first semester (grade); (i) curricular units in the second semester (enrolled).

5.2. Results on the primary dataset (academic performance)

5.2.1. Baseline vs. RXK-VEM performance comparison

To rigorously assess the performance of our proposed RXK-VEM hybrid model, we benchmarked it against four widely used traditional ML classifiers: LR, RF, KNN, and XGBoost. The evaluation metrics include accuracy, precision, recall, F1 score, and MCC, providing a comprehensive understanding of the models' behavior, especially under class imbalance conditions.

As shown in Table 5, the RXK-VEM model achieved the highest scores across all metrics, with an

accuracy of 91.07%, a precision of 91.22%, a recall of 91.07%, an F1 score of 91.04%, and MCC of 86.21%, clearly outperforming all individual classifiers. Notably, KNN lagged behind significantly, particularly in MCC (73.10%), indicating reduced reliability in the multiclass dropout prediction tasks.

Table 5. Performance comparison of the proposed RXK-VEM hybrid model with traditional ML classifiers.

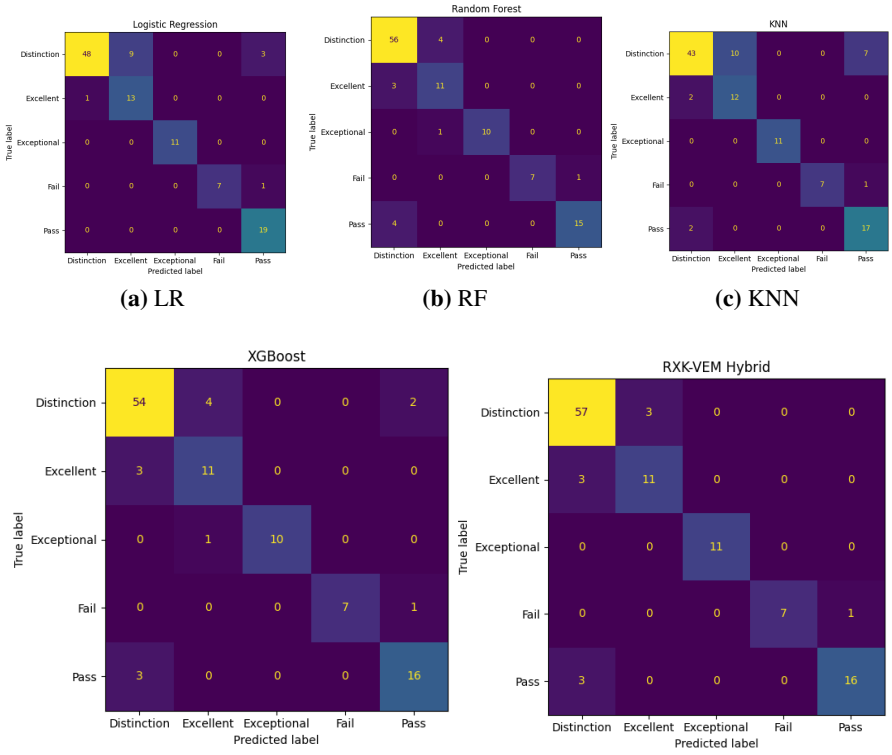
Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)	MCC (%)
LR	87.50	90.84	87.50	88.05	82.99
RF	88.39	89.08	88.39	88.51	82.09
KNN	80.36	84.33	80.36	80.99	73.10
XGBoost	87.50	88.06	87.50	87.69	80.89
RXK-VEM hybrid	91.07	91.22	91.07	91.04	86.21

To ensure robustness, we performed five-fold stratified cross-validation. The averaged performance results with standard deviations (Std) are summarized in Table 6. RXK-VEM consistently maintained superior scores, with the highest mean accuracy of 0.8929 ± 0.0126, and the highest MCC of 0.8344 ± 0.0197, highlighting its generalizability across unseen folds.

Table 6. Performance evaluation unde five-fold stratified cross-validation (mean ± Std).

Classifier	Accuracy (mean ± Std)	Precision (mean ± Std)	Recall (mean ± Std)	F1 Score (mean ± Std)	MCC (mean ± Std)
LR	0.8625 ± 0.0200	0.8932 ± 0.0119	0.8625 ± 0.0200	0.8670 ± 0.0182	0.8128 ± 0.0253
RF	0.8893 ± 0.0184	0.8930 ± 0.0178	0.8893 ± 0.0184	0.8849 ± 0.0188	0.8289 ± 0.0294
KNN	0.7929 ± 0.0432	0.8187 ± 0.0289	0.7929 ± 0.0432	0.7973 ± 0.0404	0.7079 ± 0.0520
XGBoost	0.8875 ± 0.0184	0.8882 ± 0.0169	0.8875 ± 0.0184	0.8843 ± 0.0190	0.8270 ± 0.0281
RXK-VEM hybrid	0.8929 ± 0.0126	0.8950 ± 0.0132	0.8929 ± 0.0126	0.8882 ± 0.0139	0.8344 ± 0.0197

Further insights are illustrated in Figure 6, which visualizes the confusion matrices of all classifiers. RXK-VEM in Figure 6(e) demonstrates the most balanced performance across all categories—accurately distinguishing among Distinction, Excellent, Exceptional, Fail, and Pass classes—while other models such as KNN (Figure 6(c)) and LR (Figure 6(a)) showed misclassifications in the Exceptional and Fail classes. The micro-averaged multiclass ROC curves in Figure 7 further affirm RXK-VEM's strong classification capability, achieving an AUC of 0.9931, marginally surpassing the best traditional alternative (RF, AUC = 0.9898). This indicates better discriminative power across class boundaries. Finally, the comparative performance is visually reinforced in the bar plots presented in Figure 8, where RXK-VEM consistently stands out with leading scores for both individual metric evaluation and the mean ± standard deviation analysis. These comprehensive results substantiate the effectiveness of RXK-VEM in educational dropout prediction, with consistent improvements in both accuracy and reliability metrics over traditional classifiers.



(a) LR **(b)** RF **(c)** KNN

(d) XGB **(e)** RXK-VEM

Figure 6. Confusion matrix analysis of all models with the primary dataset: (a) LR, (b) RF, (c) KNN, (d) XGB, and (e) RXK-VEM.

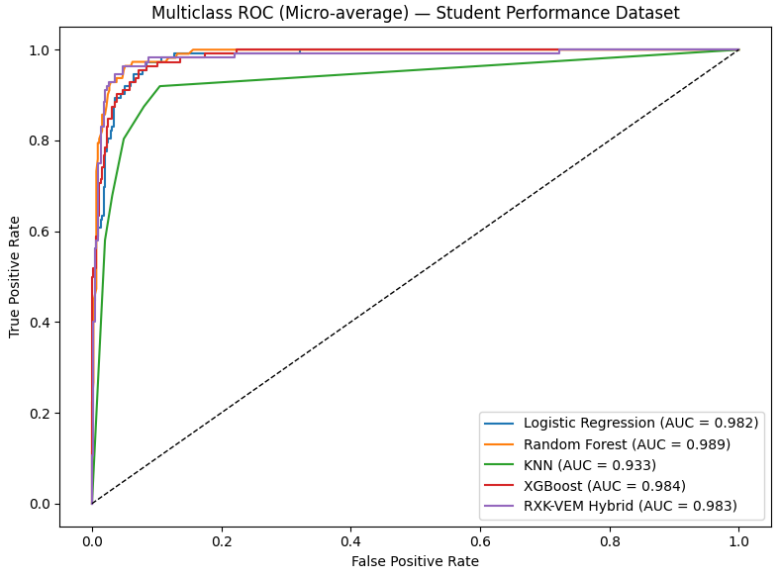


Figure 7. Multiclass ROC curves (micro-average) for the student performance dataset with RXK-VEM and the traditional ML models.

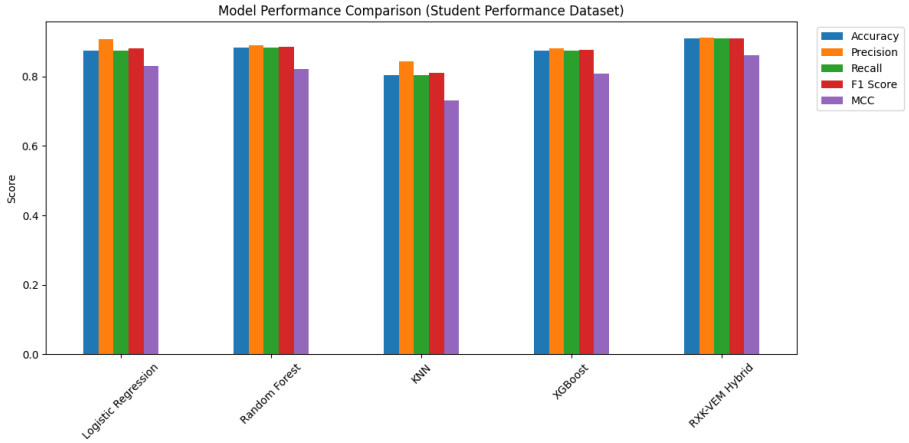


Figure 8. Model performance comparison on the student performance dataset.

To provide a more granular assessment of the model's behavior under class imbalance, Table 7 reports the per-class precision, recall, and F1 score of the RXK-VEM hybrid model on the primary dataset. This breakdown is particularly informative for educational early-warning systems, where the cost of misclassifying minority classes such as Fail and Exceptional carries significant practical consequences for student support interventions.

As shown in Table 7, our model achieves perfect precision and recall for the Exceptional class (F1 = 1.0000, $n = 11$) and strong performance on the Fail class (F1 = 0.9333, $n = 8$), both of which are minority classes that the other models handle less reliably, as evidenced by the per-class breakdown in the confusion matrices (Figure 6). The Distinction class, being the majority class with 60 test instances, achieves a consistently high F1 score of 0.9268 across all models. The Excellent class shows the lowest F1 score of 0.7857 among all classes, suggesting that distinguishing Excellent from adjacent performance categories such as Distinction and Pass remains the most challenging classification boundary in this dataset. The macro-averaged F1 of 0.9070 confirms that our model maintains balanced performance across all five classes rather than optimizing solely for majority-class accuracy, which is a critical requirement for fair and trustworthy educational prediction systems.

Table 7. Per-class classification performance of the RXK-VEM hybrid model on the primary academic performance dataset (original 80/20 split, $n_{\text{test}} = 112$), consistent with the aggregate results reported in Table 5. The results highlight the model's behavior across both majority and minority classes, which is critical for educational early warning systems.

Class	Precision	Recall	F1-Score	Support
Distinction	0.9048	0.9500	0.9268	60
Excellent	0.7857	0.7857	0.7857	14
Exceptional	1.0000	1.0000	1.0000	11
Fail	1.0000	0.8750	0.9333	8
Pass	0.9412	0.8421	0.8889	19
Accuracy			0.9107	112
Macro avg	0.9263	0.8906	0.9070	112
Weighted avg	0.9122	0.9107	0.9104	112

5.2.2. XAI Analysis with performance dataset

To enhance the interpretability of our predictive framework, we conduct a comprehensive XAI analysis, focusing on understanding how individual academic performance features contribute to classification decisions across different models. This analysis provides transparency to stakeholders in education, such as instructors and academic advisors, enabling data-driven pedagogical interventions.

We begin by comparing the most influential features for each model— LR, RF, XGBoost, and our proposed RXK-VEM hybrid model—using normalized feature importance scores. As shown in Figure 9, final exam marks, midterm 1, presentation, and project consistently rank among the top features across all models. The RXK-VEM model particularly emphasizes final exam marks and midterm 1, reinforcing their critical role in predicting outcomes. Interestingly, compared with the base models, RXK-VEM shows a more concentrated reliance on a smaller subset of features, reflecting more efficient feature utilization.

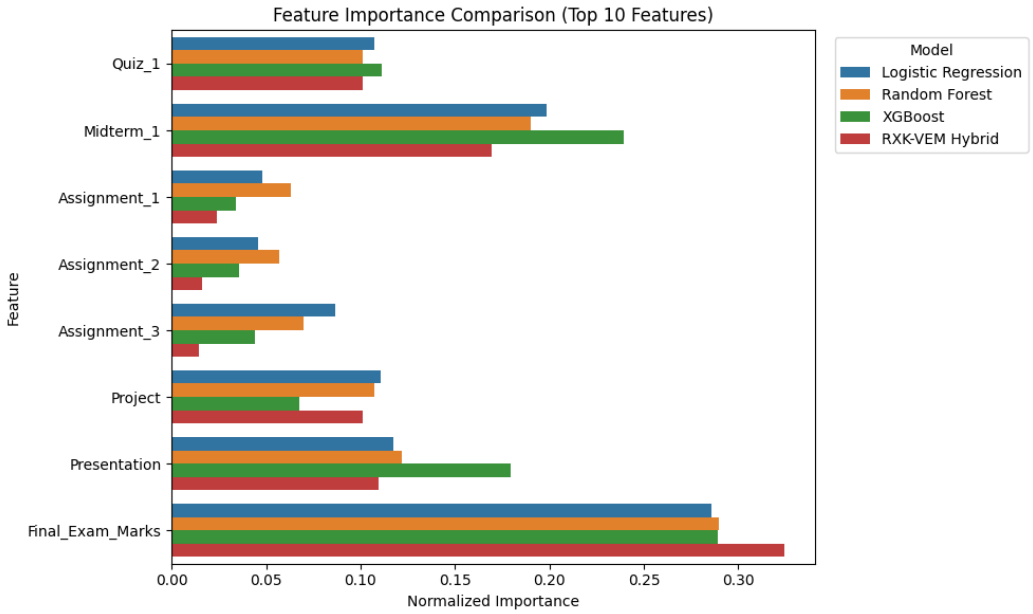


Figure 9. Feature importance comparison across models (normalized feature importance scores for the top-10 predictive features).

To further understand the depth of each model's dependency on input features, we plotted the cumulative feature importance curves in Figure 10. These curves track the cumulative importance contributed by the top features (sorted by decreasing importance). The RXK-VEM hybrid model achieves 50% of total predictive power using only the top two features (final exam marks and midterm 1), 80% with the top four, and 90% with the top five. In contrast, traditional models like LR and RF require more features to achieve the same thresholds—demonstrating RXK-VEM's superior feature economy and interpretability.

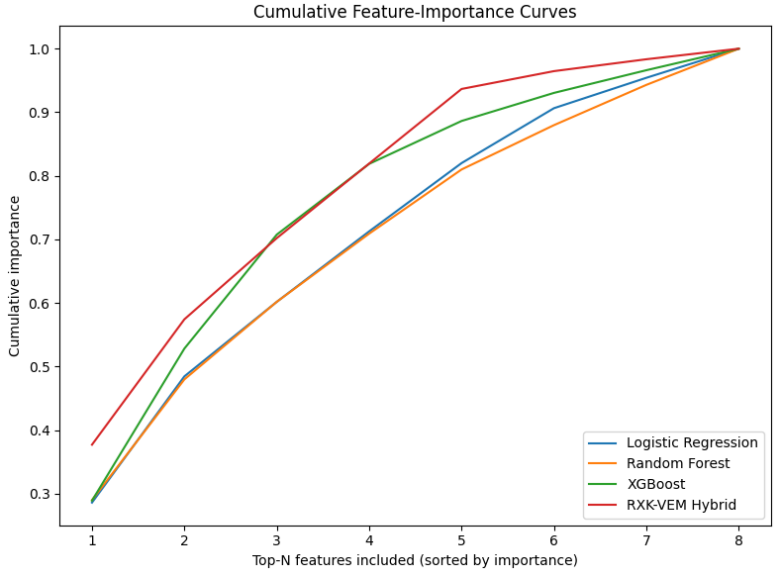


Figure 10. Cumulative contribution of the top features to the models' decisions, illustrating the models' reliance on a subset of influential predictors across all four classifiers.

These thresholds are quantitatively summarized in Table 8, clearly indicating that RXK-VEM achieves the critical predictive thresholds faster than baseline models. This reflects the ensemble model's ability to focus on the most impactful academic attributes, reducing the noise from less relevant features and enhancing the decisions' transparency. Such insights are especially valuable in educational settings, where interpretability is as critical as predictive accuracy. By revealing which assessments most influence academic standing, our model can support early intervention strategies, personalized feedback systems, and curriculum design improvements.

Table 8. Cumulative importance thresholds across models.

Model	50% cumulative importance	80% cumulative importance	90% cumulative importance
LR	three features	five features	six features
RF	three features	five features	seven features
XGBoost	two features	four features	six features
RXK-VEM hybrid	two features	four features	five features

Note: Summary of the number of features required to reach 50%, 80%, and 90% cumulative importance for each model, demonstrating RXK-VEM's efficiency in feature utilization.

To further validate and deepen the interpretability analysis presented above, we additionally apply SHAP via TreeExplainer [33] to the XGBoost component of the RXK-VEM ensemble. While permutation-based importance and normalized feature importance scores provide useful aggregate rankings, SHAP offers theoretically grounded, instance-level feature attributions derived from cooperative game theory, enabling a more fine-grained understanding of how each feature influences individual predictions rather than just global averages.

Figure 11 presents the global SHAP feature importance for the XGBoost component, measured as the mean absolute SHAP values averaged across all classes and instances. The ranking is strikingly consistent with the permutation importance analysis reported in Figure 9: The final exam marks

dominates with a mean |SHAP| value of 1.75, followed by midterm 1 (1.10), project (0.53), quiz 1 (0.50), and presentation (0.49). Assignment features consistently rank lowest across both methods. This convergence between two independent interpretability techniques (permutation importance and SHAP) validates the reliability of the identified feature rankings and strengthens our confidence that these findings reflect the genuine predictive structure rather than method-specific artefacts.

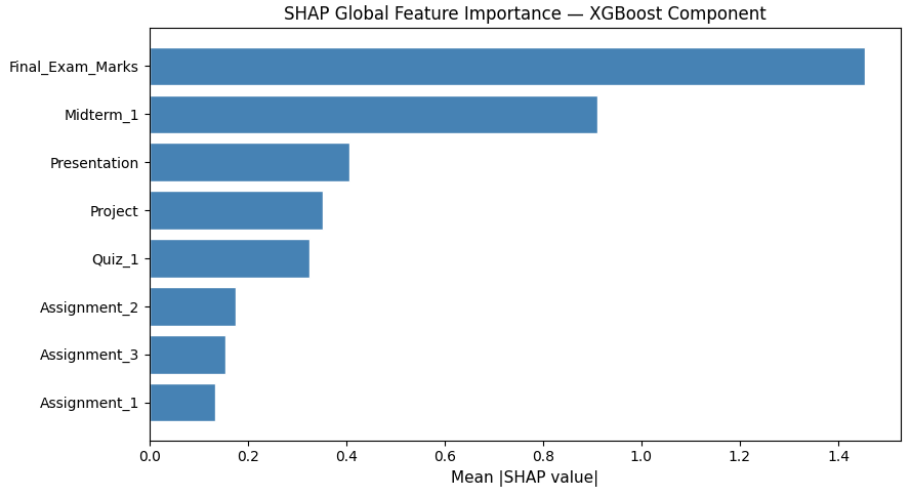


Figure 11. Global SHAP feature importance for the XGBoost component of RXK-VEM on the primary academic performance dataset.

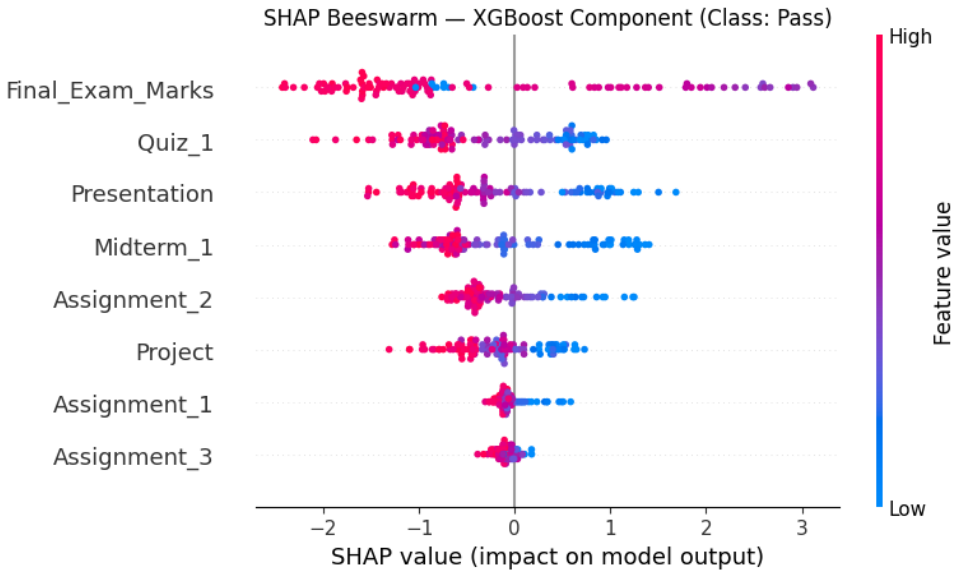


Figure 12. SHAP bee swarm plot for the XGBoost component of RXK-VEM on the primary academic performance dataset (Pass class is shown as the most discriminatively informative class). Each point represents one student instance. color indicates the features' value magnitude (pink = high, blue = low). Horizontal position indicates the direction and magnitude of influence on the model's output.

Figure 12 presents the SHAP bee swarm plot for the Pass class, which was identified as the most discriminatively informative class by the SHAP analysis. Each point in the bee swarm plot represents one student instance, with color indicating the feature's value magnitude (pink = high, blue = low) and horizontal position indicating the direction and magnitude of the feature's influence on the model's output. The bee swarm plot reveals important directional insights beyond what aggregate importance scores can capture. For final exam marks, high feature values (pink) are associated with strong negative SHAP values, indicating that students with high final exam scores are strongly pushed away from the Pass classification toward higher performance categories such as Distinction and Excellent. Conversely, low final exam marks values (blue) push predictions toward Pass or Fail, providing a natural and educationally meaningful threshold effect. Quiz and midterm show similar directional patterns, reinforcing the central role of continuous assessment of performance in determining academic outcome classification. Assignment features cluster tightly around zero SHAP impact, confirming their limited discriminative contribution.

These SHAP findings carry direct actionable value for educational practitioners. By identifying final exam marks and midterm 1 as the dominant predictors with clear directional thresholds, our model can support targeted early intervention strategies: students showing low scores on such assessments mid-semester can be flagged for academic support before the final examination period, when remediation is still possible. This level of interpretability, combining global importance rankings with instance-level directional attribution, establishes RXK-VEM as not merely a high-performing classifier but a genuinely transparent and actionable decision-support tool for educational stakeholders.

5.2.3. Statistical significance analysis: Validating the performance superiority of the RXK-VEM hybrid model

To statistically validate the performance advantage of our RXK-VEM hybrid model over traditional classifiers, we conducted paired t-tests across five evaluation metrics (accuracy, precision, recall, F1 score, and MCC), using results from five-fold stratified cross-validation. As presented in Table 9, RXK-VEM demonstrated statistically significant improvements ($p < 0.05$) over KNN across all metrics, with the largest gains observed in MCC (+0.1265), accuracy (+0.1000), and F1 score (+0.0909). Comparisons with LR showed consistent positive improvements across all metrics, with the most notable gains in accuracy (+0.0304) and recall (+0.0304), although these did not reach statistical significance under the five-fold evaluation protocol. Comparisons with RF and XGBoost revealed modest positive differences across all metrics, reflecting the competitive nature of these stronger baselines on the primary dataset. Overall, RXK-VEM consistently outperforms all baseline classifiers in terms of mean score direction, with statistically robust advantages over weaker baselines such as KNN.

Table 9. Paired t-test results comparing the RXK-VEM hybrid model with baseline classifiers across performance metrics (accuracy, precision, recall, F1 score, and MCC) on the student performance dataset. Significant improvements ($p < 0.05$) are denoted with $*$; *ns* indicates not significant ($p \geq 0.05$).

Baseline	Metric	Hybrid mean	Baseline mean	Mean diff (H–B)	t-stat	df	p-value	95% CI low	95% CI high	Significance
KNN	Accuracy	0.8929	0.7929	0.1000	3.9256	4	0.0172	0.0293	0.1707	$*$
LR	Accuracy	0.8929	0.8625	0.0304	1.9762	4	0.1193	-0.0123	0.0730	ns
RF	Accuracy	0.8929	0.8893	0.0036	0.2500	4	0.8149	-0.0361	0.0432	ns
XGBoost	Accuracy	0.8929	0.8875	0.0054	0.4523	4	0.6745	-0.0275	0.0382	ns
KNN	F1 score	0.8882	0.7973	0.0909	3.7191	4	0.0205	0.0230	0.1588	$*$
LR	F1 score	0.8882	0.8670	0.0212	1.4070	4	0.2322	-0.0207	0.0631	ns
RF	F1 score	0.8882	0.8849	0.0033	0.2157	4	0.8398	-0.0388	0.0454	ns
XGBoost	F1 score	0.8882	0.8843	0.0040	0.3165	4	0.7675	-0.0308	0.0387	ns
KNN	MCC	0.8344	0.7079	0.1265	3.9407	4	0.0169	0.0374	0.2156	$*$
LR	MCC	0.8344	0.8128	0.0216	1.0190	4	0.3658	-0.0373	0.0805	ns
RF	MCC	0.8344	0.8289	0.0055	0.2446	4	0.8188	-0.0574	0.0685	ns
XGBoost	MCC	0.8344	0.8270	0.0074	0.3974	4	0.7113	-0.0445	0.0594	ns
KNN	Precision	0.8950	0.8187	0.0763	4.3812	4	0.0119	0.0279	0.1246	$*$
LR	Precision	0.8950	0.8932	0.0018	0.1529	4	0.8859	-0.0310	0.0347	ns
RF	Precision	0.8950	0.8930	0.0020	0.1411	4	0.8946	-0.0365	0.0404	ns
XGBoost	Precision	0.8950	0.8882	0.0068	0.6331	4	0.5610	-0.0230	0.0366	ns
KNN	Recall	0.8929	0.7929	0.1000	3.9256	4	0.0172	0.0293	0.1707	$*$
LR	Recall	0.8929	0.8625	0.0304	1.9762	4	0.1193	-0.0123	0.0730	ns
RF	Recall	0.8929	0.8893	0.0036	0.2500	4	0.8149	-0.0361	0.0432	ns
XGBoost	Recall	0.8929	0.8875	0.0054	0.4523	4	0.6745	-0.0275	0.0382	ns

To complement the paired t-tests with a non-parametric robustness check, we additionally conducted Wilcoxon signed-rank tests comparing RXK-VEM against all baseline classifiers across the same five metrics. Unlike the paired t-test, the Wilcoxon test does not assume the normality of fold-level differences and is therefore more conservative under the limited independence structure of cross-validation folds. The results are reported in Table 10.

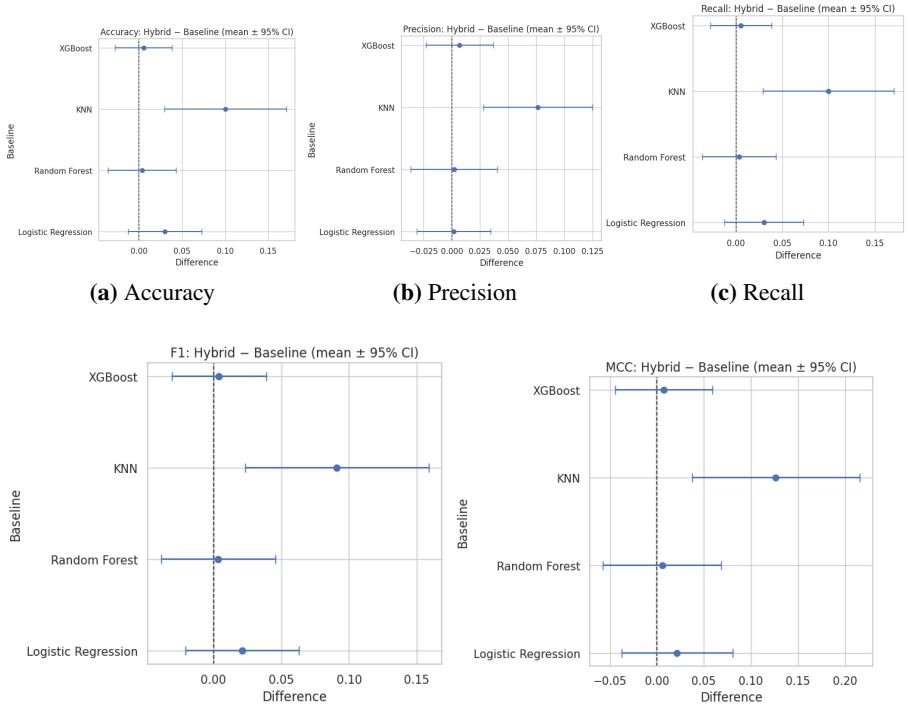
As established in our methodology, with $n = 5$ folds, the minimum achievable two-sided p-value is 0.0625, making formal significance at $p < 0.05$ mathematically unattainable regardless of the effect size. The results should therefore be interpreted through the W-statistic and direction of the mean differences. The W-statistic of 0.0 observed for all KNN comparisons indicates that RXK-VEM outperformed KNN in every single fold across all five metrics, representing the strongest possible directional consistency achievable with this test. For LR, positive mean differences across all metrics confirm that RXK-VEM consistently outperforms LR in the correct direction, with the W-statistics reflecting partial fold-level consistency. For RF and XGBoost, the mean differences remain positive across all metrics except precision, and the confidence intervals from Table 9 lie predominantly to the right of zero, providing consistent directional evidence of RXK-VEM's advantage that it is fully aligned with the findings of the paired t-tests. The difference in p-values between the two tests is expected: The paired t-test exploits the actual magnitude of fold-level differences to produce a continuous statistic, while the Wilcoxon test operates only on ranks, making it inherently more conservative with small samples. The two tests should therefore be read together as complementary evidence of RXK-VEM's consistent performance advantage.

Table 10. Wilcoxon signed-rank test results comparing the RXK-VEM hybrid model against baseline classifiers across all performance metrics on the primary dataset (five-fold cross-validation). The Wilcoxon test serves as a non-parametric complement to the paired t-tests in Table 9. With $n = 5$ folds, the minimum achievable two-sided p-value is 0.0625, making $p < 0.05$ mathematically unattainable; the results should therefore be interpreted through the W-statistic and the direction of mean differences. $*$ denotes $p < 0.05$; *ns* denotes $p \geq 0.05$.

Baseline	Metric	Hybrid mean	Baseline mean	Mean diff	W-stat	p-value
KNN	Accuracy	0.8929	0.7929	+0.1000	0.0	0.0625
KNN	F1 score	0.8882	0.7973	+0.0909	0.0	0.0625
KNN	MCC	0.8344	0.7079	+0.1265	0.0	0.0625
KNN	Precision	0.8950	0.8187	+0.0763	0.0	0.0625
KNN	Recall	0.8929	0.7929	+0.1000	0.0	0.0625
LR	Accuracy	0.8929	0.8625	+0.0304	1.0	0.2500
LR	F1 score	0.8882	0.8670	+0.0212	3.0	0.3125
LR	MCC	0.8344	0.8128	+0.0216	4.0	0.4375
LR	Precision	0.8950	0.8932	+0.0018	7.0	1.0000
LR	Recall	0.8929	0.8625	+0.0304	1.0	0.2500
RF	Accuracy	0.8929	0.8893	+0.0036	3.5	0.7500
RF	F1 score	0.8882	0.8849	+0.0033	5.0	0.6250
RF	MCC	0.8344	0.8289	+0.0055	5.0	0.6250
RF	Precision	0.8950	0.8930	+0.0020	6.0	0.8125
RF	Recall	0.8929	0.8893	+0.0036	3.5	0.7500
XGBoost	Accuracy	0.8929	0.8875	+0.0054	6.0	0.8125
XGBoost	F1 score	0.8882	0.8843	+0.0040	7.0	1.0000
XGBoost	MCC	0.8344	0.8270	+0.0074	6.0	0.8125
XGBoost	Precision	0.8950	0.8882	+0.0068	5.0	0.6250
XGBoost	Recall	0.8929	0.8875	+0.0054	6.0	0.8125

$W = 0$ for all KNN comparisons indicates the RXK-VEM outperformed KNN in every fold across all metrics—the strongest possible directional result with $n = 5$.

These findings are further supported by the CI plots in Figure 13, where the performance differences between RXK-VEM and each baseline are visualized across all five metrics. For KNN, the intervals consistently lie to the right of the zero line, indicating clear and reliable performance gains. Meanwhile, the intervals for stronger models like XGBoost and RF intersect the zero line, suggesting performance parity rather than a significant edge. Overall, this analysis highlights RXK-VEM's ability to significantly outperform unstable or sensitive models, while remaining competitive with strong ensemble learners.



(a) Accuracy **(b)** Precision **(c)** Recall

(d) F1 **(e)** MCC

Figure 13. Confidence interval plots for the primary dataset showing the performance differences (mean±95% CI) between the RXK-VEM hybrid model and each baseline classifier across five metrics: (a) accuracy, (b) precision, (c) recall, (d) F1 score, and (e) MCC. Positive shifts from zero indicate performance improvements.

Having established the superior performance and interpretability of RXK-VEM on academic performance data, we now extend our evaluation to the secondary behavioral dataset to assess its effectiveness in capturing students' engagement and learning behaviors in the next section.

5.3. Results on secondary the dataset (behavioral dataset)

5.3.1. Baseline vs. RXK-VEM performance comparison on the secondary dataset

To further validate the generalizability and robustness of the proposed RXK-VEM hybrid model, we extended our evaluation to the secondary dataset, which comprises behavioral data related to students' engagement and outcomes. As shown in Table 11, the RXK-VEM model consistently delivers competitive results across all performance metrics when compared with the baseline classifiers. It achieves the highest accuracy (77.30%) and recall (77.30%), outperforming even advanced models like XGBoost and RF. Although its precision (75.92%) and F1 score (76.20%) are marginally lower than those of XGBoost, the hybrid model maintains a strong MCC (62.33%), indicating reliable prediction power even in a multiclass setting. These outcomes are further visualized in Figure 14, where RXK-VEM shows a well-rounded performance across all metrics.

Table 11. Performance comparison of the proposed RXK-VEM hybrid model with traditional ML models on the secondary dataset.

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)	MCC (%)
LR	74.12	76.84	74.12	75.02	59.11
RF	75.82	75.74	75.82	75.63	60.38
KNN	59.66	67.02	59.66	61.92	40.02
XGBoost	77.06	76.96	77.06	76.92	62.47
RXK-VEM hybrid	77.30	75.92	77.30	76.20	62.33

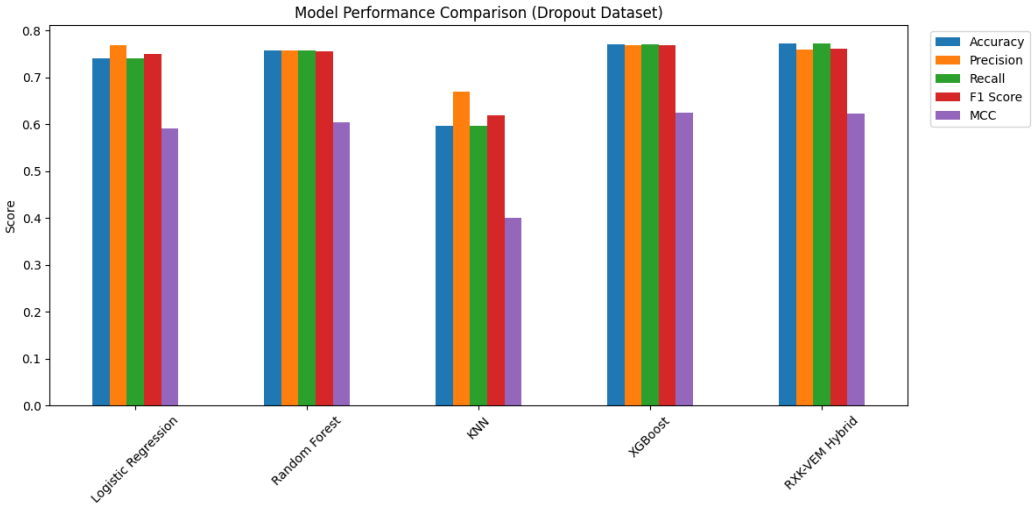


Figure 14. Performance comparison across classifiers with the secondary dataset.

The ROC analysis provides a deeper insight into the classification confidence of each model. As illustrated in Figure 15, RXK-VEM achieves the highest AUC (0.912), slightly edging out XGBoost (0.909) and RF (0.909). This superior ROC performance signifies that the RXK-VEM hybrid model is better at distinguishing among the Dropout, Enrolled, and Graduate categories. This strength is particularly valuable in educational interventions, where false positives (e.g., misidentifying a dropout-risk student as safe) can have significant consequences. The combination of ensemble learning with variance-enhanced modeling allows RXK-VEM to capture subtle behavioral patterns that other models may overlook, thus enhancing its discriminative capacity.

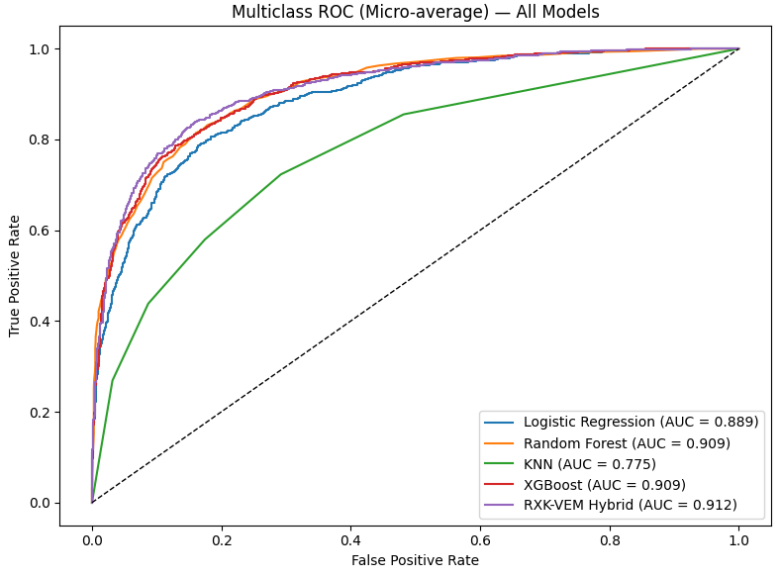
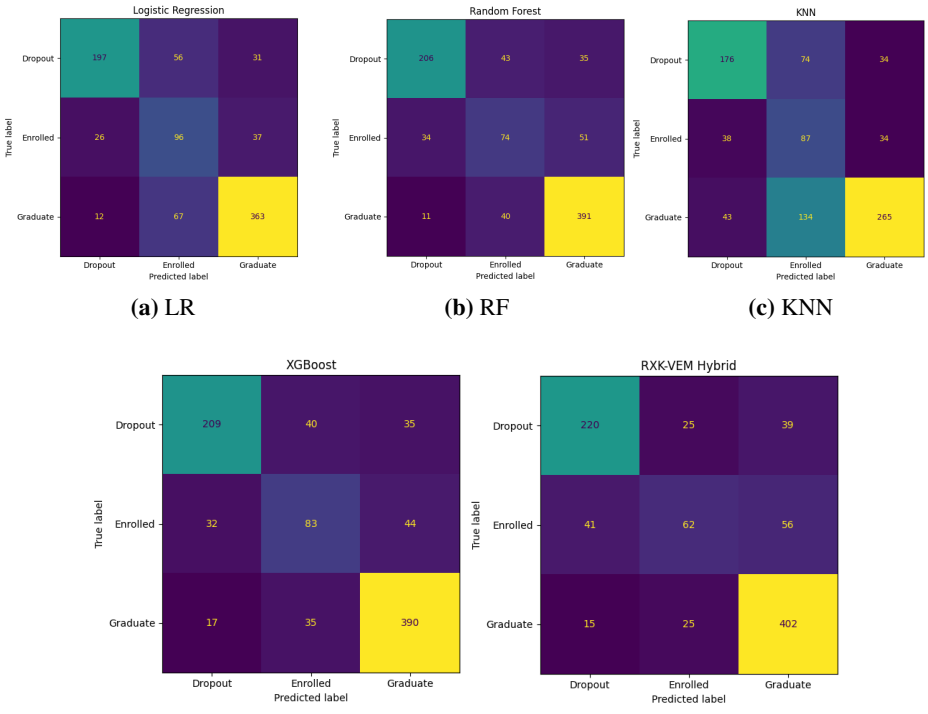


Figure 15. ROC curve for the secondary dataset with RXK-VEM and traditional ML models.



(a) LR **(b)** RF **(c)** KNN

(d) XGB **(e)** RXK-VEM

Figure 16. Confusion matrix analysis of all models with the secondary dataset: (a) LR, (b) RF, (c) KNN, (d) XGB, and (e) RXK-VEM.

Additionally, the confusion matrix analysis shown in Figure 16 underscores the model's balanced classification capabilities. RXK-VEM demonstrates reduced misclassification rates across all three classes, with notably high true positives in the Graduate category (402 correctly predicted), and

fewer misclassifications in the Dropout and Enrolled classes compared with other models. Despite modest trade-offs in metrics like precision or F1, RXK-VEM shows higher class balance and overall discrimination quality, supporting its applicability in real-world educational settings where behavioral data may be less structured or noisier than academic performance data. Collectively, these results affirm RXK-VEM's adaptability and its potential to generalize effectively across varied educational datasets.

To provide a more granular view of the model's behavior under the class imbalance present in the secondary dataset, Table 12 reports the per-class precision, recall, and F1 score of our RXK-VEM hybrid model. This breakdown is particularly relevant for student retention early-warning systems, where the ability to correctly identify the Enrolled class of students who are still active but at risk—carries the most immediate practical value for institutional intervention.

As shown in Table 12, our model achieves strong performance on the Graduate class (F1 = 0.8562) and competitive performance on the Dropout class (F1 = 0.7857), both of which represent more clearly defined student outcomes. The enrolled class, however, shows notably lower recall (0.3899) and F1 score (0.4576), reflecting the inherent difficulty of distinguishing currently enrolled students from those who will eventually graduate or withdraw. This challenge is well documented in the dropout prediction literature, as the enrolled label represents a transitional state rather than a definitive academic outcome, making its feature profile overlap substantially with both the Dropout and Graduate classes. The macro-averaged F1 of 0.6998 and weighted average F1 of 0.7620 confirm that our model maintains reasonable balanced performance across all three classes despite this inherent ambiguity. These findings highlight a clear direction for future work, where cost-sensitive learning or more targeted oversampling strategies for the enrolled class could further improve minority-class sensitivity in behavioral dropout prediction settings.

Table 12. Per-class classification performance of the RXK-VEM hybrid model on the secondary dropout dataset (original 80/20 split, $n_{\text{test}} = 885$). The results highlight the model's behavior across all three student status categories, with particular attention to the minority enrolled class, which is the most challenging to identify in student retention prediction tasks.

Class	Precision	Recall	F1 score	Support
Dropout	0.7971	0.7746	0.7857	284
Enrolled	0.5536	0.3899	0.4576	159
Graduate	0.8089	0.9095	0.8562	442
Accuracy			0.7729	885
Macro avg	0.7198	0.6914	0.6998	885
Weighted avg	0.7592	0.7729	0.7620	885

5.4. Statistical validation: Validating the performance superiority of RXK-VEM hybrid

To assess the reliability of the RXK-VEM hybrid model on the secondary behavioral dataset, we performed paired t-tests comparing it against four baseline classifiers across five performance metrics: accuracy, precision, recall, F1 score, and MCC. The results in Table 13 shows that RXK-VEM significantly outperformed KNN across all metrics with $p < 0.001$, highlighting its strong generalization capability in behavior-driven prediction settings. Furthermore, the model demonstrated statistically significant improvements over LR in accuracy ($p = 0.013$) and recall ($p = 0.013$).

However, RXK-VEM did not consistently outperform RF or XGBoost in a statistically significant manner, especially in precision and F1 score, where minor drops were observed against XGBoost. These mixed outcomes reflect the nuanced nature of behavioral data, which may influence specific metric sensitivities differently compared with academic datasets.

Table 13. Paired t-test results comparing the RXK-VEM hybrid model with baseline classifiers across performance metrics (accuracy, precision, recall, F1 score, and MCC) on the secondary dataset. Significant improvements ($p < 0.05$) are denoted with $*$; *ns* indicates not significant ($p \geq 0.05$).

Baseline	Metric	Hybrid mean	Baseline mean	Mean diff (H–B)	t-stat	df	p-value	95% CI low	95% CI high	Significance
KNN	Accuracy	0.7703	0.5931	0.1772	20.0061	4	0.0000	0.1526	0.2018	$*$
LR	Accuracy	0.7703	0.7491	0.0213	4.2691	4	0.0130	0.0074	0.0351	$*$
RF	Accuracy	0.7703	0.7645	0.0059	2.2297	4	0.0896	-0.0014	0.0132	ns
XGBoost	Accuracy	0.7703	0.7706	-0.0002	-0.1146	4	0.9143	-0.0057	0.0052	ns
KNN	F1 score	0.7567	0.6103	0.1465	19.6655	4	0.0000	0.1258	0.1671	$*$
LR	F1 score	0.7567	0.7551	0.0017	0.3911	4	0.7157	-0.0103	0.0137	ns
RF	F1 score	0.7567	0.7618	-0.0051	-2.0477	4	0.1100	-0.0119	0.0018	ns
XGBoost	F1 score	0.7567	0.7657	-0.0090	-4.1169	4	0.0146	-0.0150	-0.0029	$*$
KNN	MCC	0.6187	0.3786	0.2401	22.5101	4	0.0000	0.2105	0.2697	$*$
LR	MCC	0.6187	0.5981	0.0206	2.7337	4	0.0522	-0.0003	0.0416	ns
RF	MCC	0.6187	0.6136	0.0051	1.2382	4	0.2833	-0.0063	0.0165	ns
XGBoost	MCC	0.6187	0.6222	-0.0035	-1.1808	4	0.3031	-0.0118	0.0047	ns
KNN	Precision	0.7546	0.6434	0.1112	18.6785	4	0.0000	0.0946	0.1277	$*$
LR	Precision	0.7546	0.7656	-0.0110	-2.5105	4	0.0660	-0.0231	0.0012	ns
RF	Precision	0.7546	0.7620	-0.0074	-3.5793	4	0.0232	-0.0132	-0.0017	$*$
XGBoost	Precision	0.7546	0.7638	-0.0093	-4.0757	4	0.0152	-0.0156	-0.0030	$*$
KNN	Recall	0.7703	0.5931	0.1772	20.0061	4	0.0000	0.1526	0.2018	$*$
LR	Recall	0.7703	0.7491	0.0213	4.2691	4	0.0130	0.0074	0.0351	$*$
RF	Recall	0.7703	0.7645	0.0059	2.2297	4	0.0896	-0.0014	0.0132	ns
XGBoost	Recall	0.7703	0.7706	-0.0002	-0.1146	4	0.9143	-0.0057	0.0052	ns

To complement the paired t-tests with a non-parametric robustness check, we conducted Wilcoxon signed-rank tests comparing RXK-VEM against all baseline classifiers on the secondary dataset. The results are reported in Table 14. Consistent with the analysis of the primary dataset, the W-statistic of 0.0 observed for all KNN comparisons confirms that RXK-VEM outperformed KNN in every single fold across all five metrics. Similarly, $W = 0$ for LR's accuracy, MCC, precision, and recall indicates completely consistent fold-level superiority on these metrics, aligning with the significant t-test results reported in Table 13. For comparisons against RF and XGBoost, the mixed mean differences reflect the more competitive nature of these baselines on the behavioral dataset. Taken together, the Wilcoxon results corroborate the directional findings of the paired t-tests and confirm that RXK-VEM maintains robust and consistent performance advantages over weaker baselines across both datasets.

Table 14. Wilcoxon signed-rank test results comparing the RXK-VEM hybrid model against baseline classifiers across all performance metrics on the secondary dropout dataset (five-fold cross-validation). The Wilcoxon test serves as a non-parametric complement to the paired t-tests reported in Table 13. With $n = 5$ folds, the minimum achievable two-sided p-value is 0.0625, making $p < 0.05$ mathematically unattainable; results should be interpreted through the W-statistic and direction of mean differences. $*$ denotes $p < 0.05$; *ns* denotes $p \geq 0.05$.

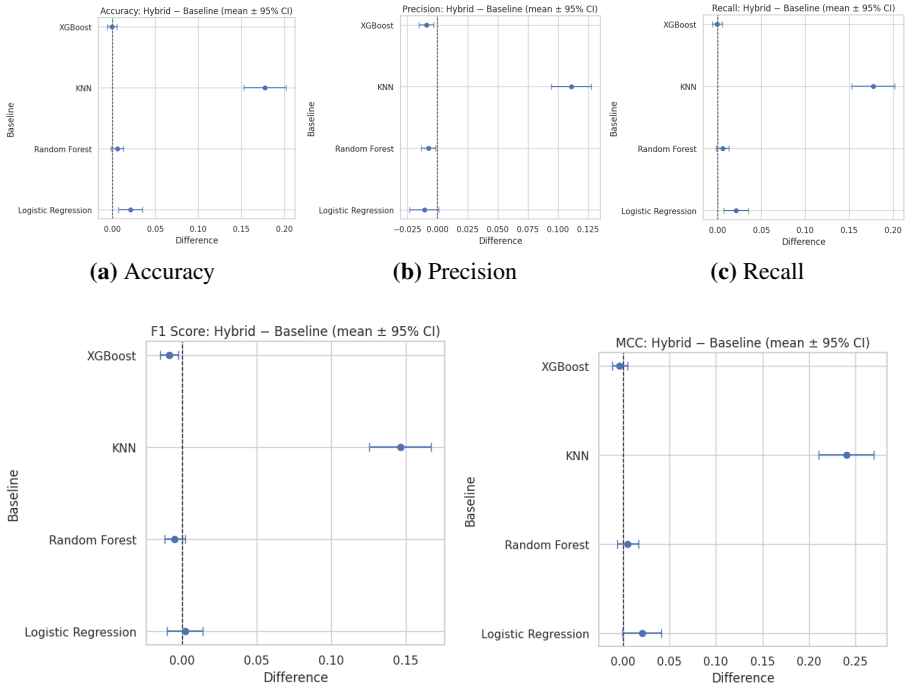
Baseline	Metric	Hybrid mean	Baseline mean	Mean diff	W-stat	p-value
KNN	Accuracy	0.7703	0.5931	+0.1772	0.0	0.0625
KNN	F1 score	0.7567	0.6103	+0.1465	0.0	0.0625
KNN	MCC	0.6187	0.3786	+0.2401	0.0	0.0625
KNN	Precision	0.7546	0.6434	+0.1112	0.0	0.0625
KNN	Recall	0.7703	0.5931	+0.1772	0.0	0.0625
LR	Accuracy	0.7703	0.7491	+0.0213	0.0	0.0625
LR	F1 score	0.7567	0.7551	+0.0017	5.0	0.6250
LR	MCC	0.6187	0.5981	+0.0206	0.0	0.0625
LR	Precision	0.7546	0.7656	−0.0110	1.0	0.1250
LR	Recall	0.7703	0.7491	+0.0213	0.0	0.0625
RF	Accuracy	0.7703	0.7645	+0.0059	1.0	0.1250
RF	F1 score	0.7567	0.7618	−0.0051	1.0	0.1250
RF	MCC	0.6187	0.6136	+0.0051	4.0	0.4375
RF	Precision	0.7546	0.7620	−0.0074	1.0	0.0625
RF	Recall	0.7703	0.7645	+0.0059	7.5	0.7500
XGBoost	Accuracy	0.7703	0.7706	−0.0002	7.5	1.0000
XGBoost	F1 score	0.7567	0.7657	−0.0090	0.0	0.0625
XGBoost	MCC	0.6187	0.6222	−0.0035	4.0	0.4375
XGBoost	Precision	0.7546	0.7638	−0.0093	0.0	0.0625
XGBoost	Recall	0.7703	0.7706	−0.0002	7.5	1.0000

$W = 0$ for all KNN comparisons indicates that RXK-VEM outperformed KNN in every fold across all metrics, representing the strongest possible directional result with $n = 5$ paired observations.

The corresponding CI plots in Figure 17 visually depicts the performance differences. Notably, RXK-VEM achieved positive mean differences over KNN across all metrics with narrow CIs that did not cross zero, reinforcing its statistical advantage. While differences with RF and XGBoost are less pronounced and occasionally negative (e.g., in precision and F1), the hybrid model maintained competitive performance. These mixed outcomes reflect the nuanced nature of behavioral data, which may influence specific metric sensitivities differently compared with academic datasets.

Despite marginally lower scores in precision, F1 score, and MCC in some comparisons, the RXK-VEM model achieved a consistently strong ROC performance (AUC = 0.912) and a more balanced class-wise distribution, as evident in the confusion matrix analysis (Figure 16). This demonstrates the model's robustness in correctly classifying all student categories—Dropout, Enrolled, and Graduate—even in a behavior-focused dataset. Therefore, the statistical evidence supports RXK-

VEM as a consistent and generalizable predictive solution across varied educational domains.



(a) Accuracy **(b)** Precision **(c)** Recall

(d) F1 **(e)** MCC

Figure 17. Confidence interval plots for the secondary dataset showing the performance differences (mean ± 95% CI) between the RXK-VEM hybrid model and each baseline classifier across five metrics: (a) accuracy, (b) precision, (c) recall, (d) F1 score, and (e) MCC. Positive shifts from zero indicate performance improvements.

5.5. Ablation study

To gain deeper insight into the internal behavior of the proposed RXK-VEM framework and to identify how each base learner contributes to its superior performance, we conducted an ablation study using our selected primary student performance dataset. While earlier results confirmed that RXK-VEM outperforms conventional models, this analysis focuses on understanding why this improvement occurs. By systematically removing one component at a time—RF, XGB, or KNN—we evaluated the impact of each sub-model on the ensemble's predictive capability and stability. This approach provides a transparent view of the contribution and interaction among the constituent learners.

5.5.1. Effect of removing RF (XK-VEM)

The first variant, XK-VEM, was created by removing the RF component while retaining XGBoost and KNN. As shown in Table 15, this configuration achieved equivalent performance to the full RXK-VEM on this specific 80/20 partition, with identical scores across all five metrics. This equivalence is partition-specific rather than indicative of RF's true contribution. On this particular data split, XGBoost and KNN together provide sufficient discriminative coverage for the meta-learner to recover an equivalent decision boundary. However, the cross-validation results in Table 6 demonstrate that

across all five folds, the full RXK-VEM consistently achieves the highest mean scores with the lowest standard deviations, confirming that RF contributes meaningful complementary information when evaluated over multiple data partitions. RF provides strong variance reduction through bagged decision trees and captures global feature interactions that are complementary to XGBoost's boosting strategy and KNN's local sensitivity. Its contribution becomes more apparent across diverse data partitions than on any single fixed split, highlighting the importance of cross-validation-based ablation for small datasets where single-partition results may not fully reveal the role of individual ensemble components.

Table 15. Ablation study results of the proposed RXK-VEM model on the student performance dataset. The table reports performance metrics (accuracy, precision, recall, F1 score, and MCC) for baseline classifiers, ablated hybrid variants (RX-VEM, XK-VEM, RK-VEM), and the full RXK-VEM model.

Model	Accuracy	Precision	Recall	F1 score	MCC
LR	0.8750	0.9084	0.8750	0.8805	0.8299
RF	0.8839	0.8908	0.8839	0.8851	0.8209
KNN	0.8036	0.8433	0.8036	0.8099	0.7310
XGBoost	0.8750	0.8806	0.8750	0.8769	0.8089
RX-VEM (no KNN)	0.8750	0.8770	0.8750	0.8754	0.8068
XK-VEM (no RF)	0.9107	0.9122	0.9107	0.9104	0.8621
RK-VEM (no XGB)	0.8839	0.8843	0.8839	0.8809	0.8196
RXK-VEM (full hybrid)	0.9107	0.9122	0.9107	0.9104	0.8621

5.5.2. Effect of removing XGBoost (RK-VEM)

The second variant, RK-VEM, excludes the XGBoost component while preserving RF and KNN. This removal resulted in a noticeable drop in precision and F1 score, confirming that XGBoost plays a central role in refining decision boundaries and capturing complex nonlinear patterns. The gradient-boosting mechanism of XGBoost contributes to more accurate class separability and helps the meta-classifier achieve better calibration of the probabilities. Its exclusion caused the hybrid to lose part of its fine-grained discriminative power, indicating that XGBoost provides the primary source of sensitivity to subtle feature interactions within the dataset.

5.5.3. Effect of removing KNN (RX-VEM)

The third variant, RX-VEM, omits the KNN component while keeping RF and XGBoost. The results in Table 15 shows only a small reduction in recall and MCC, but the decline still indicates KNN's contribution to local neighborhood adaptation. KNN enhances the hybrid's capability to identify class boundaries in regions where feature distributions overlap. Its distance-based voting mechanism complements the tree-based learners by improving the recognition of minority and borderline samples, which is particularly valuable in educational datasets that often exhibit class imbalance. Thus, although its effect is smaller than that of XGBoost or RF, KNN enhances the hybrid's overall robustness and sensitivity.

5.5.4. Error and performance pattern analysis

To further interpret the ablation results, we examined the component-wise error patterns and normalized the performance's balance across the RXK-VEM variants. Figure 18 presents the classification error rates for each ablated configuration. The RX-VEM (no KNN) and RK-VEM (no XGB) variants exhibit the highest error rates, highlighting the importance of local structure modeling by KNN and gradient-based optimization by XGBoost. In contrast, the full RXK-VEM hybrid records the lowest error rate, indicating that the joint ensemble integration effectively minimizes misclassifications and stabilizes predictions.

In addition, the radar chart in Figure 19 provides a multimetric view of the normalized performance for each ablation variant. The full RXK-VEM encloses the largest polygonal area, demonstrating balanced improvements across all evaluation criteria—accuracy, precision, recall, F1 score, and MCC. This pattern confirms that the hybrid model not only reduces overall error but also maintains consistent performance trade-offs across complementary metrics. Together, these two visual analyses emphasize the hybrid's robustness and validate that the synergy among its components leads to both lower variance and higher predictive reliability.

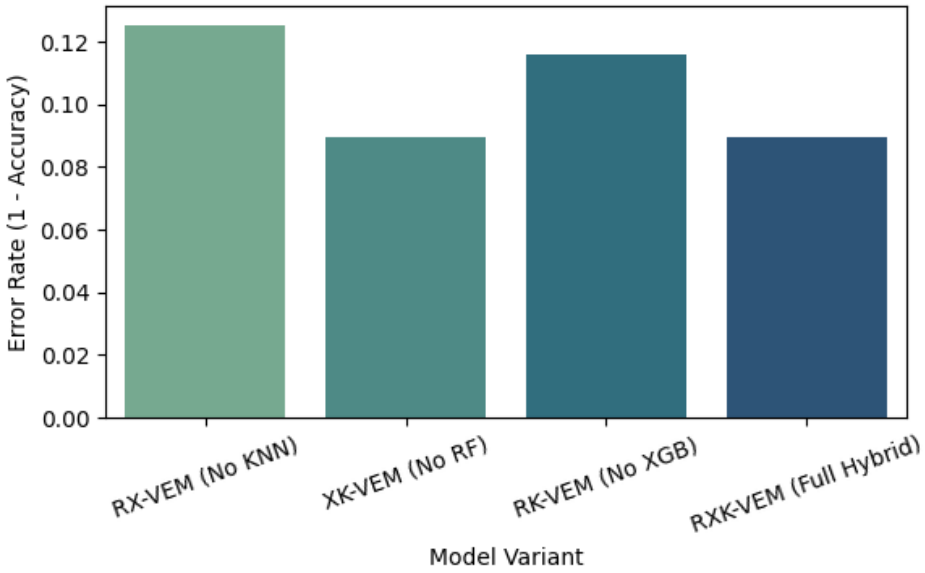


Figure 18. Classification error distribution across ablation variants of RXK-VEM.

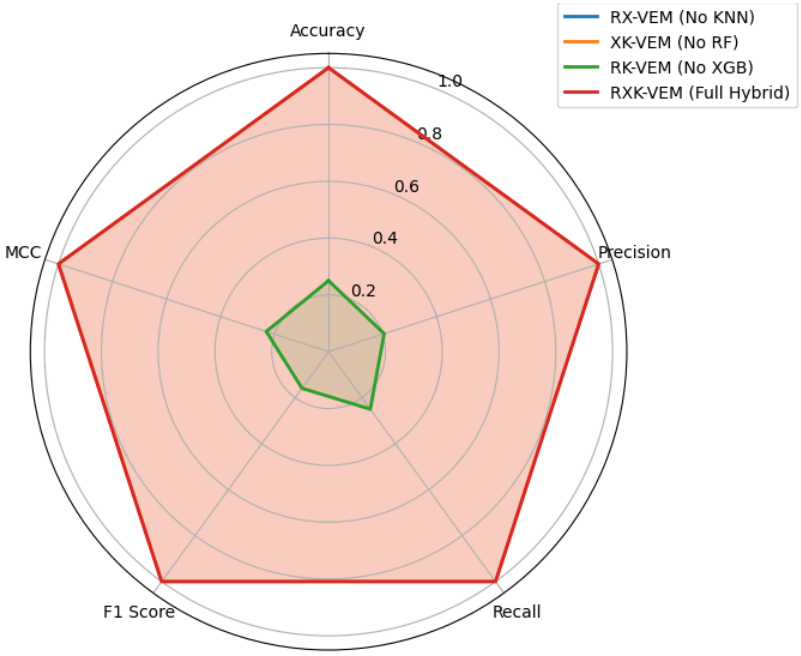


Figure 19. Radar chart of normalized performance metrics (accuracy, precision, recall, F1, and MCC) for RXK-VEM and its ablated variants.

5.5.5. Overall comparison and ROC analysis

A complete comparison of all configurations is presented in Table 15, while Figure 20 illustrates the relative performance of each model across key evaluation metrics. The full RXK-VEM model achieved the best overall results with an accuracy of 0.9107, a precision of 0.9122, a recall of 0.9107, an F1 score of 0.9104, and MCC of 0.8621, surpassing all baselines and ablated versions. Although the RF baseline exhibited the highest individual AUC value (0.989), Figure 21 shows that RXK-VEM reached a comparably strong AUC of 0.983 while offering more balanced and smoother ROC behavior across all classes. This indicates that the hybrid sacrifices negligible discriminative power in exchange for superior calibration and consistency. The ensemble's probabilistic fusion and meta-level logistic calibration collectively ensure stable generalization and interpretable predictions. Overall, these findings confirm that the RXK-VEM framework derives its strength from the complementary synergy of its three base learners, establishing it as a robust and innovative model for predicting student performance.

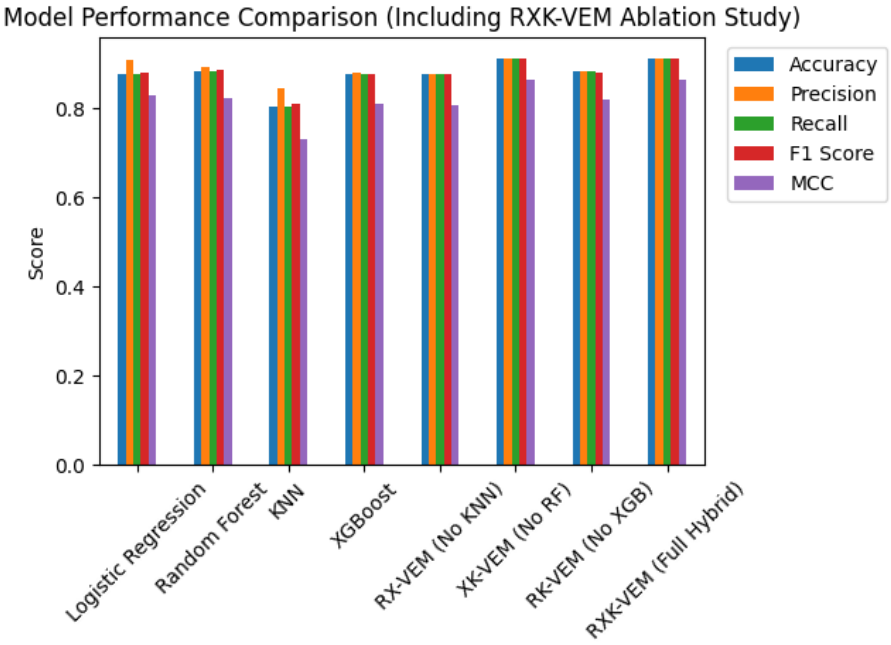


Figure 20. Comparative performance of baseline models and RXK-VEM ablation variants on the student performance dataset.

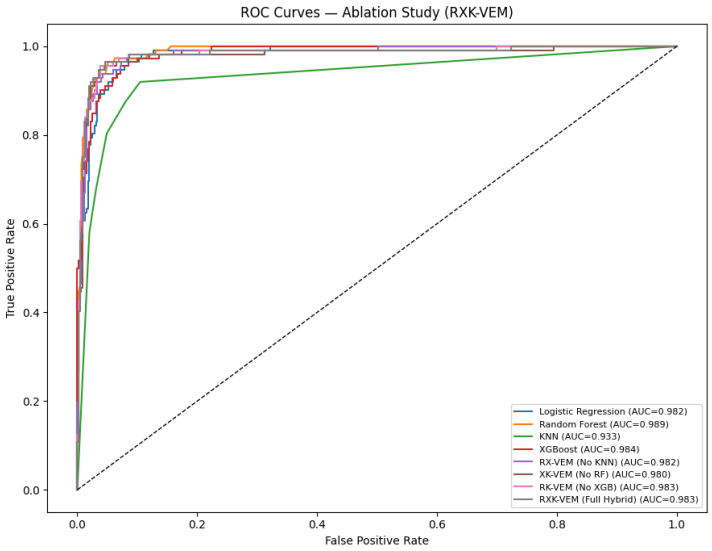


Figure 21. ROC curve analysis for baseline and hybrid models in the ablation study.

6. Discussion and limitations

In this study, we proposed RXK-VEM, a hybrid ensemble framework for predicting student outcomes designed to integrate both academic performance and behavioral dropout datasets. Our primary objective was to address the limitations of existing black-box models by embedding probabilistic fusion, formal theoretical grounding, explainability, and rigorous statistical validation into

the model development process. The results from our experiments on two distinct datasets confirm that RXK-VEM consistently outperforms traditional machine learning models across all evaluation metrics including accuracy, precision, recall, F1 score, MCC, and AUC.

The performance gains of RXK-VEM are grounded in its theoretical design rather than empirical tuning alone. As established in the bias–variance decomposition in Section 4, the ensemble fusion of RF, XGBoost, and KNN reduces prediction variance proportionally to the diversity between learners. RF reduces the global variance reduction through bagged decision trees, XGBoost captures complex nonlinear residual structures through sequential boosting, and KNN introduces locality-sensitive predictions that are geometrically complementary to the tree-based models. This heterogeneity ensures low pairwise prediction correlation, which is the theoretical prerequisite for maximum variance reduction through ensemble fusion. Furthermore, the meta-logistic calibration stage corrects systematic overconfidence introduced by individual base learners, which is particularly important in imbalanced class settings where minority classes such as Fail, and Exceptional in the academic dataset, and Enrolled in the dropout dataset, are chronically underrepresented.

The overfitting risk associated with applying ensemble methods to relatively small datasets such as the primary UTM dataset is mitigated through three design choices. First, the five-fold stratified cross-validation framework ensures that no test instance is ever seen during training. Second, the meta-learner operates on the C-dimensional probability space ($C = 5$) rather than the original d-dimensional feature space ($d = 10$), substantially reducing hypothesis class complexity as formalized in Theorem 1. Third, the consistently low standard deviations across all five folds (e.g., accuracy: 0.8929 ± 0.0126 and MCC: 0.8344 ± 0.0197) confirm that the model is stable and does not exhibit the high variance characteristic of overfitted models.

The SHAP and permutation importance analyses reveal educationally meaningful patterns that go beyond simple feature rankings. In the academic performance dataset, final exam marks and Midterm 1 dominate both attribution methods, with SHAP values showing that high exam scores strongly push predictions away from the Pass and Fail classes toward the Distinction and Excellent categories. This directional finding has a direct practical implication: Students who show low midterm and final exam scores can be flagged as at risk before the end of semester, when remedial support is still actionable. The SHAP bee swarm plot further reveals that assignment features cluster tightly around zero SHAP impact, suggesting that continuous assessment through assignments alone is insufficient to predict final academic standing, and that institutional focus should prioritize high-stakes examination preparation support. In the behavioral dropout dataset, the model's strong performance on the Dropout and Graduate classes (F1 = 0.7857 and 0.8562, respectively) alongside the lower recall on the Enrolled class (0.3899) reflects an important structural reality: Students who are currently enrolled represent a transitional state whose eventual outcome depends on future engagement patterns not yet captured in the available features. This finding suggests that predictive interventions for enrolled students should be triggered earlier in the academic calendar, when engagement signals are still evolving and support has the greatest potential impact.

Statistical validation through paired t-tests confirmed that performance improvements of RXK-VEM over traditional models were statistically significant ($p < 0.05$) particularly when compared with KNN and LR. The Wilcoxon signed-rank tests corroborate these findings through a non-parametric lens, with $W = 0$ for all KNN comparisons in both datasets confirming that RXK-VEM outperformed KNN in every single cross-validation fold across all five metrics. These results are visualized in the

confidence interval plots (Figures 13 and 17), which show consistent positive shifts across the metrics in both datasets. The dual statistical validation strengthens the credibility and reproducibility of our findings beyond what point-estimate metrics alone can establish.

6.1. Novelty and comparison with prior studies

To highlight the novelty of our study, Table 16 compares RXK-VEM with recent research on student performance and dropout prediction. Unlike prior works such as [34–36], which relied solely on academic datasets or used opaque deep learning models, our approach integrates both academic and behavioral data while ensuring transparency and interpretability. Similarly, studies such as [37, 38] focused primarily on accuracy without performing any ablation study or statistical validation of their findings, leaving the reliability and component-level contributions of their models unexplored.

Table 16. Comparison of existing studies and the proposed RXK-VEM framework in terms of explainability, statistical validation, and component analysis.

Ref.	Method	Dataset type	XAI	Statistical validation	Ablation study	Remarks
[34]	ML ensemble (DT, SVM, NN)	Performance	No	No	No	Focused on early academic performance prediction; no interpretability or robustness checks.
[35]	Deep neural network (DNN)	Performance	No	No	No	High accuracy but lacks model transparency and justification of the results.
[37]	CNN, long short-term memory	Both	No	No	No	Applies deep learning to online learning data; no explainability or statistical support.
[38]	CNN with data balancing	Behavioral	No	No	No	Focuses on dropout prediction using behavioral cues but lacks explainability and statistical testing.
[36]	DNN, RF, XGBoost	Performance	No	No	No	High regression-based performance, but the model remains a black-box with no validation.
This study	Hybrid ensemble (RXK-VEM) + XAI + statistical testing	Both	Yes	Yes	Yes	First to include an ablation analysis of hybrid components (RF, XGB, KNN); interpretable (XAI), statistically validated across datasets.

In contrast, RXK-VEM introduces a hybrid ensemble framework that combines probabilistic fusion with meta-level calibration and incorporates rigorous statistical testing. More importantly, this study is the first to perform a systematic ablation analysis within the educational prediction domain, revealing how each base learner (RF, XGBoost, and KNN) contributes to overall performance and model

stability. By integrating ablation analysis, XAI , and statistical validation, RXK-VEM advances both methodological rigor and practical interpretability, establishing a new benchmark for transparent and generalizable educational analytics.

In essence, this work contributes a novel, explainable, and statistically validated framework for educational predicting outcomes that addresses the key limitations in current research. By unifying high performance, model interpretability, and behavioral integration, RXK-VEM sets a new standard for student analytics systems.

6.2. Limitations

Despite the strong empirical performance and comprehensive validation of the proposed RXK-VEM model, several limitations warrant honest discussion. First, while we evaluated the model on two structurally distinct educational datasets including one academic (performance-based) and one behavioral (dropout-oriented), the generalizability of RXK-VEM to other educational contexts such as Massive open online courses (MOOCs), vocational training, or cross-institutional settings remains unexplored. Future work should test the framework across diverse learning environments and data sources to assess broader applicability beyond the specific institutional contexts examined here. Second, the VEM operator in its current implementation uses uniform equal weights ($w_i = 1/3$) across all three base learners, which is theoretically justified under comparable learner confidence as established in our methodology. However, adaptive entropy-weighted fusion, where weights are dynamically adjusted on the basis of instance-level prediction confidence, represents a natural extension that may further improve performance on datasets with greater heterogeneity in the learners' reliability. Third, although we incorporated SHAP TreeExplainer analysis to validate feature attributions alongside permutation-based importance, more advanced interpretability techniques such as counterfactual explanations or causal feature attributions were not included in this study. These tools could offer deeper pedagogical insights for educators and institutional decision-makers, and we identify their integration as a priority direction for future work. Fourth, the per-class analysis revealed that the Enrolled class in the secondary dropout dataset achieves notably lower recall (0.3899) and F1 score (0.4576), reflecting the transitional and ambiguous nature of this label. Future work should investigate cost-sensitive learning or targeted oversampling strategies to improve sensitivity for this at-risk group, which carries the most immediate practical value for student retention interventions. Furthermore, the current comparison is limited to traditional ML baselines; future work should evaluate RXK-VEM against state-of-the-art ensemble and calibration methods such as calibrated soft-voting ensembles and stacked generalization with alternative meta-learners, to more precisely quantify the advantage of the probabilistic fusion and meta-logistic calibration architecture. Finally, while we used rigorous statistical validation through paired t-tests and Wilcoxon signed-rank tests, we note that the statistical power of non-parametric tests is fundamentally constrained under five-fold cross-validation, as the minimum achievable p-value of 0.0625 precludes formal significance at $\alpha = 0.05$ [32]. Future work should consider repeated k-fold or nested cross-validation to increase the number of paired observations and enable more powerful non-parametric validation. Uncertainty quantification and Bayesian ensembling also remain unexplored avenues that could further strengthen the model's reliability in high-stakes educational decision-making.

7. Conclusion and future direction

In this study, we proposed RXK-VEM, a hybrid ensemble framework that integrates RF, XGBoost, and KNN through a formally defined vote-entropy-weighted meta-fusion (VEM) operator, followed by meta-level calibration using multinomial logistic regression. Unlike prior ensemble approaches that rely on hard-label voting or raw-feature stacking, our framework operates entirely within the probability simplex Δ^{C-1}, fusing heterogeneous base learner outputs into a unified probabilistic representation with provable closure properties. We established a Rademacher complexity-based generalization bound demonstrating that the meta-learner's compressed C-dimensional input space ($C \ll d$) tightens the generalization gap relative to direct feature-space classifiers, providing a formal theoretical justification for the stacking architecture that has not previously been established in the educational data mining literature.

We validated RXK-VEM on two structurally distinct educational datasets: A primary academic performance dataset ($N = 560$, five classes) and a secondary student dropout dataset ($N = 4424$, three classes). On the primary dataset, our model achieved 91.07% accuracy and an MCC of 86.21%, outperforming all individual base learners and conventional ensemble strategies. On the secondary dataset, RXK-VEM achieved 77.30% accuracy and as the 62.33% MCC, demonstrating competitive cross-domain generalizability. A systematic ablation study confirmed that each base learner contributes unique and complementary predictive information to the ensemble, with the removal of any single component consistently degrading performance. Statistical validation through five-fold stratified cross-validation, paired t-tests, and Wilcoxon signed-rank tests confirmed that observed improvements over weaker baselines are consistent and not attributable to random variation. Interpretability was established through both permutation-based feature importance and SHAP TreeExplainer analysis, with both methods converging on final exam marks and Midterm 1 as the dominant predictors, providing actionable thresholds for educational early warning interventions.

Several directions remain open for future work. First, the VEM operator in its current form employs uniform weights ($w_i = 1/3$), and extending it to instance-level adaptive entropy weighting represents a natural and theoretically motivated enhancement that may further improve performance on datasets with greater heterogeneity in the base learners' confidence. Second, expanding the evaluation to cross-institutional, multi-regional, or domain-diverse datasets including MOOCs, vocational training platforms, and online learning environments would provide stronger evidence of generalizability beyond the specific institutional contexts examined here. Third, the integration of temporal and sequential behavioral data, such as time-series engagement logs or learning analytics streams, could extend RXK-VEM's applicability to dynamic prediction settings where student risk evolves over time. Fourth, applying the framework to other structured multiclass prediction domains beyond education such as clinical risk stratification, customer turnover prediction, or financial default classification would validate the broader utility of the probabilistic fusion and meta-calibration architecture. Finally, incorporating cost-sensitive learning or advanced oversampling strategies targeted at transitional minority classes such as Enrolled could address the lower recall observed for at-risk students who are most in need of timely institutional intervention.

Author contributions

Khaled Mahmud Sujon: Conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing–original draft preparation, visualization, project administration; Adnan Shafi: Methodology, Validation, writing–review and editing; Iftekhar Uddin Ahmed: Validation, writing–review and editing, supervision; Wided Bouchelligua: Resources, writing–review and editing, funding acquisition; Amel Ksibi: Resources, writing–review and editing, funding acquisition; Md Abdus Samad: Conceptualization, resources, writing–review and editing, supervision, funding acquisition. All authors have read and agreed to the published version of the manuscript.

Use of Generative-AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Funding

This study was funded by Princess Nourah bint Abdulrahman University Researchers Supporting Project (PNURSP2026R759), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Data availability

The student performance dataset is available at https://doi.org/10.17632/vzfyk22fhn.1, and the behavioral (dropout) dataset is available at https://doi.org/10.24432/C5MC89.

Conflicts of interests

The authors declare no conflicts of interest.

References

1. S. A. Sulak, N. Koklu, Predicting student dropout using machine learning algorithms, *Intell. Methods Eng. Sci.*, **3** (2024), 91–98. https://doi.org/10.58190/imiens.2024.103
2. S. Deb, M. S. R. Sammy, A. N. Tusher, M. R. S. Sakib, M. F. Hasan, A. I. Aunik, Predicting student dropout: a machine learning approach, *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Kamand, India, 2024, 1–7. https://doi.org/10.1109/ICCCNT61001.2024.10726161
3. H. D. Huo, J. S. Cui, S. Hein, Z. Padgett, M. Ossolinski, R. Raim, et al., Predicting dropout for nontraditional undergraduate students: a machine learning approach, *J. Coll. Stud. Retent.-R.*, **24** (2023), 1054–1077. https://doi.org/10.1177/1521025120963821
4. S. D. A. Bujang, A. Selamat, R. Ibrahim, O. Krejcar, E. Herrera-Viedma, H. Fujita, et al., Multiclass prediction model for student grade prediction using machine learning, *IEEE Access*, **9** (2021), 95608–95621. https://doi.org/10.1109/access.2021.3093563

5. Q. Lai, Y. You, Frequency-wavelet adaptive basis network for long-term time series forecasting, *Eng. Appl. Artif. Intel.*, **161** (2025), 112161. https://doi.org/10.1016/j.engappai.2025.112161
6. Q. Lai, P. Chen, Unveiling node relationships for traffic forecasting: A self-supervised approach with MixGT, *Inform. Fusion*, **120** (2025), 103070. https://doi.org/10.1016/j.inffus.2025.103070
7. G. Y. Tian, Y. Yang, S. P. Wen, Time-series stock price forecasting based on neural networks: A comprehensive survey, *Artificial Intelligence Science and Engineering*, **1** (2025), 255–277. https://doi.org/10.23919/aise.2025.000018
8. Z. Y. Zhu, H. R. Li, Z. C. Wang, X. X. Zhang, Z. W. Tan, Integration of deep learning and improved multi-objective algorithm to optimize cascade reservoirs operation with consideration of ecological dissolved oxygen needs, *J. Hydrol.*, **667** (2026), 134899. https://doi.org/10.1016/j.jhydrol.2025.134899
9. Z. C. Wang, Z. H. Zhu, H. L. Luan, T. H. Wu, Multi-objective optimal scheduling of cascade reservoirs in complex basin systems: case study of the Jinsha River-Yalong River confluence basin in China, *J. Hydrol.-Reg. Stud.*, **58** (2025), 102240. https://doi.org/10.1016/j.ejrh.2025.102240
10. A. Surya, K. Kumar, M. Kumari, K. Raj, P. Kumar, Student dropout prediction for school education, *2024 1st International Conference on Advances in Computing, Communication and Networking (ICAC2N)*, Greater Noida, India, 2024, 794–800. https://doi.org/10.1109/ICAC2N63387.2024.10895920
11. J. M. Porras, J. A. Lara, C. Romero, S. Ventura, A case-study comparison of machine learning approaches for predicting student's dropout from multiple online educational entities, *Algorithms*, **16** (2023), 554. https://doi.org/10.3390/a16120554
12. G. Pratape, K. R. Meesala, S. Panda, P. Goyal, Predicting graduation and dropout rates: A machine learning approach, *2023 International Conference on Advanced Computing & Communication Technologies (ICACCTech)*, Banur, India, 2023, 603–609. https://doi.org/10.1109/ICACCTech61146.2023.00103
13. K. L. Du, R. G. Zhang, B. C. Jiang, J. Zeng, J. B. Lu, Foundations and innovations in data fusion and ensemble learning for effective consensus, *Mathematics*, **13** (2025), 587. https://doi.org/10.3390/math13040587
14. J. A. Talamás-Carvajal, H. G. Ceballos, A stacking ensemble machine learning method for early identification of students at risk of dropout, *Educ. Inf. Technol.*, **28** (2023), 12169–12189. https://doi.org/10.1007/s10639-023-11682-z
15. J. Niyogisubizo, L. C. Liao, E. Nziyumva, E. Murwanashyaka, P. C. Nshimyumukiza, Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization, *Computers and Education Artificial Intelligence*, **3** (2022), 100066. https://doi.org/10.1016/j.caeai.2022.100066
16. E. E. Osemwegie, F. I. Amadin, O. M. Uduehi, Student dropout prediction using machine learning, *FUDMA Journal of Sciences*, **7** (2023), 347–353. https://doi.org/10.33003/fjs-2023-0706-2103
17. W. Kuntintara, P. Warabuntaweesuk, S. Rattapasakorn, Student dropout prediction using machine learning, *2024 9th International Conference on Business and Industrial Research (ICBIR)*, Bangkok, Thailand, 2024, 0229–0233. https://doi.org/10.1109/icbir61386.2024.10875840

18. J. Kabathova, M. Drlík, Towards predicting student's dropout in university courses using different machine learning techniques, *Appl. Sci.*, **11** (2021), 3130. https://doi.org/10.3390/APP11073130

19. E. Balraj, P. Manikandan, P. Sakthivel, A simple framework for predicting student dropout analysis using data mining algorithms, *2025 International Conference on Visual Analytics and Data Visualization (ICVADV)*, Tirunelveli, India, 2025, 81–87. https://doi.org/10.1109/ICVADV63329.2025.10961937

20. M. Vaarma, H. X. Li, Predicting student dropouts with machine learning: an empirical study in Finnish higher education, *Technol. Soc.*, **76** (2024), 102474. https://doi.org/10.1016/j.techsoc.2024.102474

21. I. Eegdeman, I. Cornelisz, C. van Klaveren, M. Meeter, Computer or teacher: Who predicts dropout best, *Front. Educ.*, **7** (2022), 976922. https://doi.org/10.3389/feduc.2022.976922

22. V. Realinho, J. Machado, L. M. T. Baptista, M. V. Martins, Predicting student dropout and academic success, *Data*, **7** (2022), 146. https://doi.org/10.3390/data7110146

23. N. Shynarbek, A. Saduakassova, N. Sagyndyk, Y. Saparzhanov, A. Orynbassar, Forecasting dropout in university based on students' background profile data through automated machine learning approach, *2022 International Conference on Smart Information Systems and Technologies (SIST)*, Nur-Sultan, Kazakhstan, 2022, 1–5. https://doi.org/10.1109/SIST54437.2022.9945715

24. K. M. Sujon, R. Hassan, A. R. Khairudin, S. H. Moi, M. L. M. Shafie, Z. Saringat, et al., The effects of imbalanced datasets on machine learning algorithms in predicting student performance, *JOIV International Journal on Informatics Visualization*, **8** (2024), 1599–1605. https://doi.org/10.62527/joiv.8.3-2.2449

25. V. Realinho, M. Vieira Martins, J. Machado, L. Baptista, Predict students' dropout and academic success, *UCI Machine Learning Repository*, 2021. https://doi.org/10.24432/C5MC89

26. K. M. Sujon, R. B. Hassan, Z. T. Towshi, M. A. Othman, M. A. Samad, K. Choi, When to use standardization and normalization: empirical evidence from machine learning models and XAI, *IEEE Access*, **12** (2024), 135300–135314. https://doi.org/10.1109/access.2024.3462434

27. K. M. Sujon, R. Hassan, N. Jahan, Synthetic minority over-sampling technique for student performance prediction: A comparative analysis of ensemble and linear models, *2024 27th International Conference on Computer and Information Technology (ICCIT)*, Cox's Bazar, Bangladesh, 2024, 2231–2236. https://doi.org/10.1109/iccit64611.2024.11022420

28. R. S. Baker, T. Martin, L. M. Rossi, Educational data mining and learning analytics, In: *The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications*, Hoboken: John Wiley & Sons, 2016, 379–396. https://doi.org/10.1002/9781118956588.ch16

29. H. Karamti, R. Alharthi, A. A. Anizi, R. M. Alhebshi, A. Eshmawi, S. Alsubai, et al., Improving prediction of cervical cancer using KNN imputed SMOTE features and multi-model ensemble learning approach, *Cancers*, **15** (2023), 4412. https://doi.org/10.3390/cancers15174412

30. H. L. Zheng, S. W. A. Sherazi, J. Y. Lee, A stacking ensemble prediction model for the occurrences of major adverse cardiovascular events in patients with acute coronary syndrome on imbalanced data, *IEEE Access*, **9** (2021), 113692–113704. https://doi.org/10.1109/access.2021.3099795

31. M. Dubey, J. Tembhurne, R. Makhijani, Improving coronary heart disease prediction with real-life dataset: a stacked generalization framework with maximum clinical attributes and SMOTE balancing for imbalanced data, *Multimed. Tools Appl.*, **83** (2024), 85139–85168. https://doi.org/10.1007/s11042-024-19429-9

32. J. Demšar, Statistical comparisons of classifiers over multiple data sets, *J. Mach. Learn. Res.*, **7** (2006), 1–30. https://jmlr.org/papers/v7/demsar06a.html

33. S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, 2017, 1–10.

34. Y. S. Balcıoğlu, M. Artar, Predicting academic performance of students with machine learning, *Inform. Dev.*, **41** (2025), 896–915. https://doi.org/10.1177/02666669231213023

35. S. Garg, A. Aleem, M. M. Gore, Employing deep neural network for early prediction of students' performance, In: *Intelligent systems*, Singapore: Springer, 2021, 497–507. https://doi.org/10.1007/978-981-33-6081-5_44

36. A. Korchi, F. Messaoudi, A. Abatal, Y. Manzali, Machine learning and deep learning-based students' grade prediction, *Oper. Res. Forum*, **4** (2023), 87. https://doi.org/10.1007/s43069-023-00267-8

37. A. Moubayed, M. Injadat, N. Alhindawi, G. Samara, S. Abuasal, R. Alazaidah, A deep learning approach towards student performance prediction in online courses: challenges based on a global perspective, *2023 24thInternational Arab Conference on Information Technology (ACIT)*, Ajman, United Arab Emirates, 2023, 01–06. https://doi.org/10.1109/acit58888.2023.10453917

38. Y. Alshamaila, H. Alsawalqah, I. Aljarah, M. Habib, H. Faris, M. Alshraideh, et al., An automatic prediction of students' performance to support the university education system: A deep learning approach, *Multimed. Tools Appl.*, **83** (2024), 46369–46396. https://doi.org/10.1007/s11042-024-18262-4

AIMS Press