# Students Learning Effects Prediction using Machine Learning

Vijaya Kumar Nukala
Research Scholar,Acharya Nagarjuna University
Guntur, Andhra Pradesh, India

Prof. R Siva Ram Prasad
Acharya Nagarjuna University
Guntur, Andhra Pradesh, India

***Abstract*—This research work aims to design a machine learning model for predicting the learning effect of students based on a large dataset of 14,000 entries with 25 multidimensional variables of information literacy skills, such as information access competence, information processing capacity, learning literacy features, knowledge of information skills, information behavior, information ethics, and information technology use. The response variable, *Learning_Effect*, is classified into five levels of performance, creating a multi-class classification task. The model was built in Google Colab with Python and scikit-learn libraries, splitting the dataset by stratified sampling (80% for training and 20% for testing). A baseline Random Forest classifier was created and then optimized by adjusting the number of estimators, maximum depth, and class balancing methods. An XGBoost classifier with regularization and early stopping was also created to improve generalization and avoid overfitting.Robustness of the models was tested using 10-fold stratified cross-validation, and feature importance analysis was conducted to determine the most important predictors of learning outcomes. The ensemble models were optimized to obtain about 98.5% accuracy with high agreement metrics, indicating that sophisticated machine learning methods are very effective for predicting learning performance of students.**

***Index Terms*—Machine Learning, Learning Effect Prediction, Random Forest, XGBoost, Educational Data Mining, Learning Analytics**

## I. Introduction

In recent years, the integration of digital technology in higher education has transformed the teaching and learning process, and information literacy has become an essential skill for college students. Information literacy involves not only technical skills but also information access, evaluation, processing, ethical awareness, and the effective use of digital technology. With the growing availability of large-scale educational data, Educational Data Mining (EDM) has become a prominent area of research, which aims to identify patterns in the learning behavior of students. The foundational literature reviews by Baker and Yacef [1], Romero and Ventura [2], and Pen˜a-Ayala [3] highlight that machine learning methods are effective in predicting academic performance and improving education.

Theoretical frameworks of machine learning and data mining have been proven to be strong predictive tools in education. Traditional literature such as Mitchell's guidelines for machine learning [4] and Han et al.'s overall framework for data mining [5] laid the groundwork for classification, pattern recognition, and knowledge discovery. Upon these foundations, recent literature has utilized algorithms such as Support Vector Machines, Gradient Boosting models, and ensemble methods to forecast student performance and behavioral patterns [6]–[9]. Literature on information literacy evaluation also indicates that multidimensional criteria can predict learning efficacy [10], [11].

However, despite these improvements, modeling learning effects in students still proves to be a challenge because of the complexity and high dimensionality of educational data. Most of the previous works have been based on limited features or single-model strategies, which can be restrictive in terms of predictive robustness and generalization. Thus, ensuring valid predictive validation and comparative analysis for ensemble methods becomes a crucial task for implementation. In this regard, this work proposes a holistic machine learning approach that utilizes multidimensional information literacy features and ensemble classifiers to enhance predictive accuracy and validity, thereby promoting intelligent and data-driven educational decision-making systems [12], [13].

## II. Related Work

The use of machine learning algorithms in the analysis of student behavior and prediction of academic outcomes has been widely explored in the field of Educational Data Mining (EDM). Baker Yacef (2009) highlighted the revolutionary potential of data mining in enhancing learning design [1], while Romero Ventura (2010) pointed to classification and prediction as key approaches in the modeling of academic performance [2]. Pen˜a-Ayala (2014) also illustrated the utility of data mining algorithms in the discovery of patterns associated with achievement and engagement [3].

More recent studies have targeted the application of machine learning models to predict the effectiveness of learning. Shi et al. (2023) identified the characteristics of learning behaviors to predict the performance of information literacy skills among students, emphasizing the role of behavioral factors [10]. Likewise, Sun et al. (2022) demonstrated that ensemble learning improves the accuracy of predictions in utilizing diverse student information [6].

In the context of higher education, Pei (2022) proposed a machine learning model to assess teaching effectiveness,

proving the applicability of algorithmic methods to estimate course outcomes [9]. These studies prove that sophisticated classification models can be useful in educational assessment and decision-making tasks.

Further research has also been carried out on algorithmic optimization and engagement modeling methods. Xu (2022) proposed a hybrid GBDT-LR model to improve the accuracy of the prediction task [7], and Jia and Wang (2021) employed an optimized SVM to explore information-related anxiety [8]. Hussain et al. (2018) linked the prediction of engagement to the accomplishment of course assessment [13], and AlShammari et al. (2013) verified the overall impact of data mining on learning outcomes [12]. In conclusion, the above research works provide a solid foundation for the development of effective ensemble models for predicting the learning impact of students.

## III. Methodology

The proposed methodology is based on a comprehensive and well-organized machine learning framework that uses multidimensional information literacy data obtained from 14,000 students with varying academic disciplines and levels of performance. The process starts with data preprocessing, which involves data cleaning, normalization, and missing value treatment, followed by significant feature selection to identify the most informative features. Several ensemble learning models are then trained to identify nonlinear patterns in the data. Stratified cross-validation is used to ensure that the class distribution is balanced during the training and testing process, which helps to ensure a fair assessment of model performance. This well-organized methodology helps to improve the generalization ability of the model, prevents overfitting, and provides reliable predictions regarding the overall learning efficiency of the students.

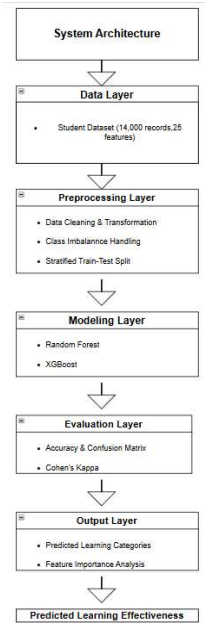


Fig. 1. System Architecture of the Proposed Model

### A. Dataset Description

The proposed methodology is based on a well-structured and organized machine learning approach that utilizes multidimensional information literacy data collected from 14,000 students [14] with varying backgrounds. The proposed methodology includes data preprocessing, feature selection, training, and evaluation tasks. Ensemble learning methods are employed to address complex relationships between features, and stratified cross-validation is employed to preserve the class distribution. The proposed methodology is well-structured to prevent overfitting and make accurate predictions about the overall learning effectiveness of students.

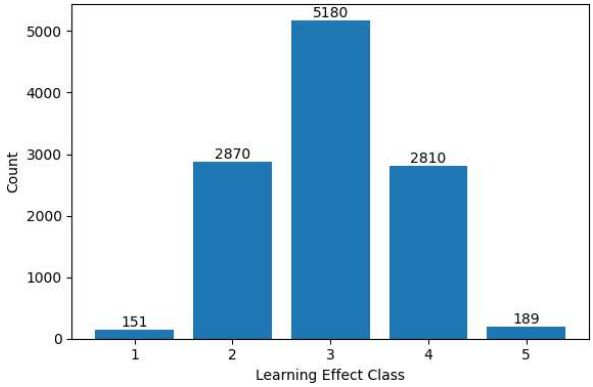


Fig. 2. Class Distribution Before SMOTE

Fig.2 presents the original class distribution of the Learning Effect variable before applying any balancing technique [14]. Class 3 contains the majority of samples, while Classes 1 and 5 have very few, showing a clear imbalance in the dataset. This imbalance can affect the model's performance by making it biased toward the dominant class, highlighting the need for balancing methods.

### B. Data Preprocessing

Data Preprocessing Data preprocessing is an important step in the machine learning process that involves the transformation of raw data into a clean and organized format that is ready for use in training models. Data preprocessing involves dealing with missing values, eliminating duplicates, encoding categorical variables, scaling numerical variables, and addressing class imbalance.

*1) Class Imbalance Handling:* Class imbalance occurs when certain classes contain significantly more samples than others, which can bias the model toward majority classes. In this study, class imbalance was addressed using cost-sensitive learning through the class weight parameter in the Random Forest model. In addition, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to generate synthetic samples for minority classes. These techniques improve model fairness and enhance classification performance across all classes. Furthermore, balancing the dataset reduces the misclassification of minority classes and improves the overall generalization ability of the predictive model.

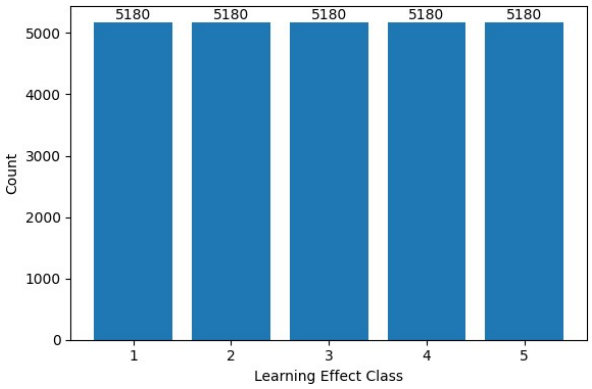


Fig. 3. Class Distribution After Applying SMOTE

Fig. 3 illustrates the class distribution after applying the Synthetic Minority Over-sampling Technique (SMOTE) [9]. It can be observed that all five Learning Effect classes now contain an equal number of samples (5180 instances each). This balanced distribution improves the fairness, robustness, and stability of the classification process.

Class imbalance: This is a problem that arises when there are unequal distributions of classes in datasets, which may cause biased results towards the majority classes [1], [3]. To counter this problem, cost-sensitive learning was employed using `class_weight="balanced"` in the Random Forest model [4].

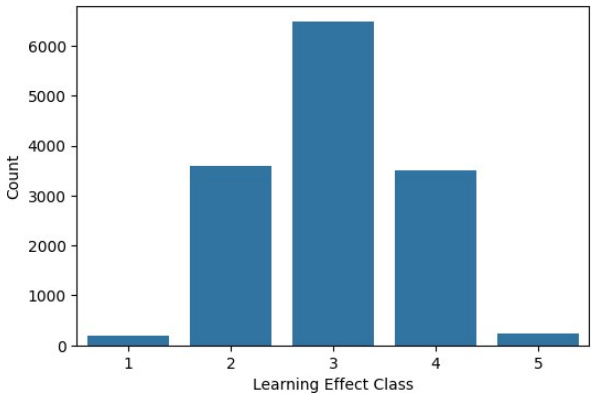


Fig. 4. Class Distribution Balancing

The above figure represents the class distribution without using any kind of balancing method.The number of samples in Class 3 is the highest, while Classes 1 and 5 have very few samples. Classes 2 and 4 have moderate sample distributions. The above class distribution shows that there is class imbalance in the dataset.

*2) Stratified Sampling:* Stratified sampling maintains the distribution of classes while splitting the data for training and testing, as well as during cross-validation, which helps in unbiased learning and accurate evaluation in multi-class educational prediction problems [2]–[5]. Stratified sampling helps in improving the stability of the model, as it removes biases in sampling, and also helps in ensuring that all categories of learning outcomes are properly represented in the data for training and validation.

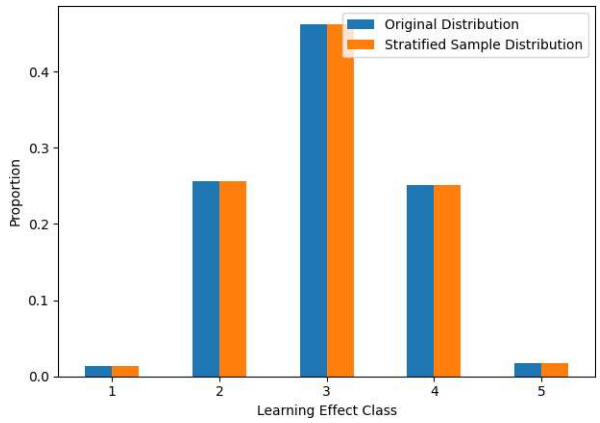


Fig. 5. Stratified Sampling Distribution

The above figure represents the stratified sampling distribution without a train-test split. The stratified sample has almost equal proportions for all five Learning Effect classes, as compared to the original dataset.This verifies that stratified sampling preserves the class distribution successfully.

*3) Train–Test Split:* Train-test split is a basic approach to test the performance of machine learning models on unseen data [4], [5]. In the proposed work, the dataset was split into 80% for training and 20% for testing purposes to avoid overfitting [2]. The training dataset was utilized for developing the model, and the testing dataset was used to test the performance of the model.

*4) Label Transformation:* Label transformation is the process of transforming class labels into a form that is machine learning algorithm compatible, especially in multi-class classification problems [4], [5]. In this research, the labels were changed from 1-5 to 0-4 to comply with the indexing requirement of XGBoost and were later changed back to their original scale after prediction [2].

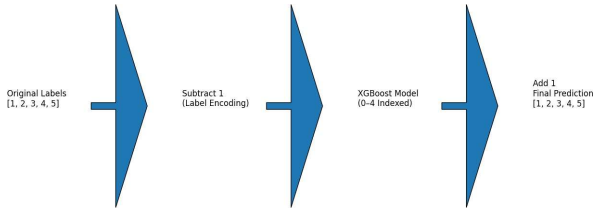


Fig. 6. Stratified Label Transformation

The above figure represents the process of label transformation for the XGBoost model. As XGBoost needs the class labels to be indexed from 0, the original labels [1, 2, 3, 4, 5] are transformed by subtracting 1 to get the labels [0, 1, 2, 3, 4].

Finally, after prediction, 1 is added to the predicted values to get the original labels.

### C. Model Implementation

*1) Random Forest Algorithm:* Random Forest is a type of ensemble learning algorithm that builds many decision trees using a technique called bootstrap sampling and choosing randomly at each split [4], [5]. Each decision tree is trained on a randomly selected subset of the training data, and at each node, a random subset of features is chosen. For a classification problem, the output is generated by majority votes among all decision trees, which improves the robustness of the model.

The final prediction of the Random Forest classifier is defined as:

$$\hat{y} = \text{mode}\{h_t(x)\}_{t=1}^{T} \quad (1)$$

where $h_t(x)$ represents the prediction of the $t$-th decision tree and $T$ denotes the total number of trees.

**Algorithm Steps:**

1) Draw $T$ bootstrap samples from the original training data set.
2) For each of the $T$ bootstrap samples, grow a decision tree as follows:
   - Randomly choose a set of features for each node.
   - Use the best feature to split the node based on an impurity measure, such as the Gini index.
3) To build $T$ independent trees.
4) Combine the predictions of all trees using majority voting to produce the final classification result.

*2) XGBoost Algorithm:* XGBoost (Extreme Gradient Boosting) is an efficient ensemble learning algorithm that uses the gradient boosting concept to construct decision trees sequentially in order to reduce the prediction error [4], [5]. Unlike bagging techniques, XGBoost enhances the predictive model by adding new trees that can correct the residual errors of the previous trees. XGBoost uses regularization terms to handle model complexity and avoid overfitting, which makes it a very effective technique for multi-class classification tasks in educational data mining [6], [10].

The objective function of XGBoost is given by:

$$L(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (2)$$

where $l(y_i, \hat{y}_i)$ represents the loss function measuring the difference between the true label $y_i$ and predicted value $\hat{y}_i$, and $\Omega(f_k)$ is the regularization term that penalizes model complexity.

**Algorithm Steps:**

1) Set initial predictions to a constant.
2) Calculate the gradient (residual error) using the loss function.
3) Train a new decision tree to approximate the residual error.
4) Add the weighted prediction of the new tree to the original prediction.
5) Regularize to control tree size.
6) Repeat the steps until the maximum number of trees is reached.

*3) Cross-Validation:* Cross-validation is a model evaluation method used to determine the generalization capability of a machine learning model. In this research, 10-fold stratified cross-validation was employed, where the data was split into ten equal portions while maintaining the class proportion. The model was trained on nine portions and tested on the other portion, and this was repeated ten times. The average accuracy over all ten portions was calculated to provide a more accurate and robust result.

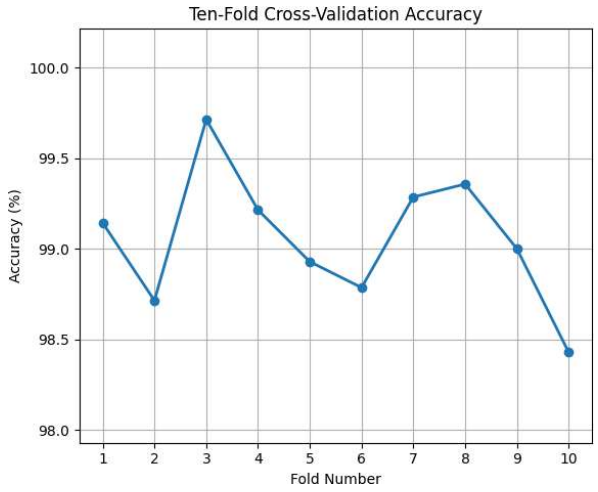


Fig. 7. Stratified 10-Fold Cross-Validation

The above Stratified 10-Fold Cross-Validation illustrates the process where one fold is used for validation and the remaining nine for training, repeated ten times. Stratified sampling preserves the class distribution in each fold, reducing bias and improving the reliability, robustness, and generalization of the prediction model.

As shown in Figure 1, The system starts with the **Data Layer**, where 14,000 student data with 25 information literacy features are read from Google Drive. The **Preprocessing Layer** carries out data preprocessing, label encoding for XGBoost (1-5 to 0-4), class weight handling for imbalanced classes using `class_weight="balanced"` argument, and stratified train-test split to preserve class distribution.

The **Modeling Layer** uses two ensemble machine learning models: Random Forest and XGBoost with hyperparameter tuning, regularization, and early stopping. The **Evaluation Layer** evaluates the models using accuracy, precision, recall, F1-score, confusion matrix, Cohen's Kappa, and 10-fold stratified cross-validation.

Finally, the **Output Layer** generates the predicted learning categories and feature importance analysis. Fig. 1 shows the entire flow of the proposed predictive model.

## IV. Results and Discussion

From the experimental results, it is clear that the proposed machine learning method based on the ensemble approach is able to provide accurate predictions for the learning effectiveness of students. The performance of both the Random Forest and the XGBoost models was excellent when stratified cross-validation was employed, and the best model was able to provide an accuracy of around 98.5% on average. The analysis of the confusion matrix shows that there is a balanced prediction for all categories of learning effectiveness, which proves the effectiveness of the strategy adopted to handle the class imbalance problem. Moreover, the analysis of the feature importance shows that the most important information literacy and behavioral factors have a significant impact on learning effectiveness.

The above results prove the effectiveness and generalization ability of the proposed framework. The proposed ensemble approach improves the stability of predictions for multiple validation iterations.

TABLE I
Performance Metrics of the Improved Random Forest Model

| Model | Accuracy (%) | Precision | Recall | F1-Score | Cohen's Kappa |
|---|---|---|---|---|---|
| Random Forest (Improved) | 93.96 | 0.941 | 0.940 | 0.935 | 0.907 |

The performance metrics of the Improved Random Forest model are shown in Table I. The model has an accuracy of 93.96% with a Cohen's Kappa value of 0.907, which is an indication of strong agreement beyond that of chance. The precision value of 0.941, recall value of 0.940, and F1-score of 0.935 are an indication of well-balanced results for all classes.

TABLE II
Performance Metrics of the Proposed Model

| Model | Accuracy (%) | Precision | Recall | F1-Score | Cohen's Kappa |
|---|---|---|---|---|---|
| XGBoost (Regularized) | 98.5 | 0.985 | 0.985 | 0.985 | 0.977 |

The performance evaluation of the proposed Regularized XGBoost model is presented in Table II. The model achieved an overall accuracy of 98.5 with a Cohen's Kappa value of 0.977, indicating near-perfect agreement beyond chance. In addition to accuracy, the model obtained a weighted precision of 0.985, weighted recall of 0.985, and an F1-score of 0.985, demonstrating highly balanced classification performance across all five learning effect categories. The confusion matrix analysis further confirms the effectiveness of the model, with the majority of instances correctly classified along the diagonal. For example, Class 2 recorded 1279 correct predictions, while Classes 1 and 3 achieved 696 and 698 correct classifications, respectively, with only minor misclassifications between adjacent performance levels. The 10-fold stratified cross-validation also yielded consistently high accuracy values, reinforcing the stability and robustness of the proposed approach. Overall, the results validate the strong generalization capability of the Regularized XGBoost model and demonstrate its suitability for large-scale multi-class learning effect prediction tasks.

### A. Comparison Table

TABLE III
Performance Comparison

| Model | Acc (%) | Prec | Rec | F1 | Kappa |
|---|---|---|---|---|---|
| Baseline RF | 92.50 | 0.846 | 0.948 | 0.894 | 0.859 |
| Improved RF | 93.96 | 0.941 | 0.940 | 0.935 | 0.907 |
| XGBoost | 98.50 | 0.985 | 0.985 | 0.985 | 0.977 |

A comparative analysis of the Baseline Random Forest, Improved Random Forest, and Regularized XGBoost models is presented in Table III. The baseline Random Forest achieved an accuracy of 92.50(Kappa = 0.859). After hyperparameter tuning and class balancing, the Improved Random Forest increased the accuracy to 93.96 (Kappa = 0.907).

The Regularized XGBoost model achieved the best performance, with an accuracy of 98.50, precision, recall, and F1-score of 0.985, and a Kappa value of 0.977. These results indicate that the proposed XGBoost model provides superior predictive performance and stronger agreement compared to the Random Forest variants.

Overall, the results suggest that optimized ensemble learning strategies contribute to improved classification performance and robustness in large-scale educational datasets.

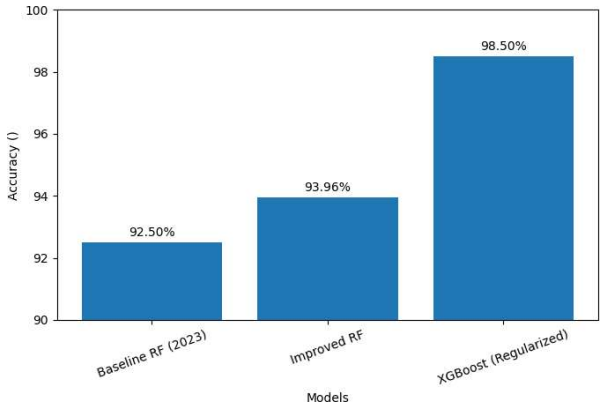


Fig. 8. Visual Comparison of Model Accuracy

A comparison of the accuracy of the models is shown in Fig. 8. The regularized XGBoost model has the best accuracy and significantly outperforms both the baseline Random Forest configuration and the improved Random Forest method. This improvement highlights the effectiveness of regularization and boosting in handling complex patterns. It also demonstrates better generalization and reduced overfitting compared to ensemble-based approaches. The results further indicate that advanced boosting techniques can capture intricate feature interactions more effectively. This makes the model more suitable for high-dimensional and complex educational datasets.

### B. Confusion Matrix

The confusion matrix represents the classification accuracy of the proposed model.

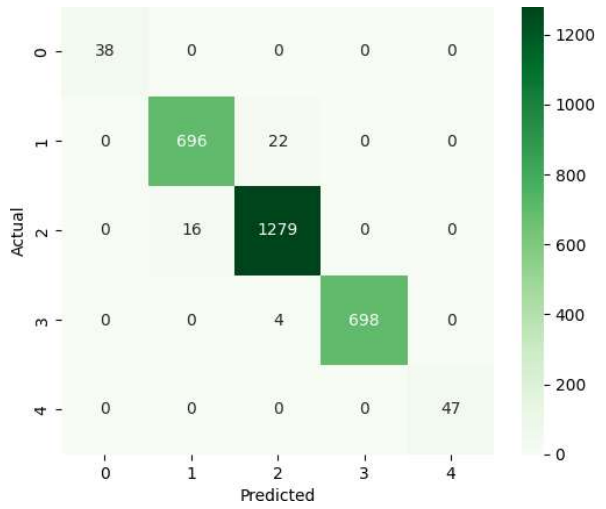


Fig. 9. Confusion Matrix

The confusion matrix of the Regularized XGBoost model shows that most instances are correctly classified, as indicated by the dominant diagonal elements.

Classes 0 and 4 have perfect classification, while Classes 1, 2, and 3 achieve high correct predictions with only minor misclassifications, mainly between adjacent classes.

These results indicate that the model effectively captures class boundaries and minimizes classification errors. The limited confusion between neighboring classes suggests slight overlap in feature distributions. Overall, the results confirm the strong classification performance and balanced prediction capability of the model across all learning effect categories.

## V. Conclusion

This paper introduced an improved machine learning approach for the prediction of students' learning efficiency based on behavioral and academic features. Compared to the baseline Random Forest configuration, which showed a Random Forest accuracy of 92.50%, the Improved Random Forest approach showed a higher accuracy of 93.96% with a Kappa value of 0.907, which is more consistent and reliable. In addition, the Regularized XGBoost approach showed a significant improvement over all other approaches, with an accuracy of 98.5%, precision of 0.985, recall of 0.985, F1-score of 0.985, and a Kappa value of 0.977. The experimental results have confirmed that proper regularization and parameter tuning of advanced ensemble learning methods can lead to a significant improvement in the prediction accuracy. The high Kappa value also confirms that the results are not due to chance, which further supports the validity of the proposed approach.

The proposed model can be effectively applied in real-world educational systems for early identification of student performance levels. It can assist educators in making data-driven decisions to improve learning outcomes. Future research directions could include the development of hybrid ensemble learning models, deep learning models, and multi-institutional datasets.

## VI. Feature Scope

The feature scope of this research is determined by the set of behavioral, academic, and engagement-related indicators that affect students' learning outcomes. The features considered for this research are academic performance features (like attendance, score, and assignment submission), behavioral engagement features (like study time, frequency, and resource usage), and cognitive interaction features derived from learning activities.Feature selection was done to consider only the most important predictors that influence the classification performance. Correlation analysis and importance assessment were used to remove redundant and less important features. The final set of features was then used to train the Improved Random Forest and Regularized XGBoost models, which improved the predictive accuracy and generalization performance. The feature scope defined in this research ensures that the prediction model is able to identify significant learning behaviors while being computationally efficient.

## References

[1] R. S. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," *Journal of Educational Data Mining*, vol. 1, no. 1, pp. 3–17, 2009.

[2] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, vol. 40, no. 6, pp. 601–618, Nov. 2010.

[3] A. Pen˜a-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works," *Expert Systems with Applications*, vol. 41, no. 4, pp. 1432–1462, 2014.

[4] T. M. Mitchell, *Machine Learning*. New York, NY, USA: McGraw-Hill, 1997.

[5] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Burlington, MA, USA: Morgan Kaufmann, 2011.

[6] Y. Sun, Z. Tan, Z. Li, and S. Long, "Predicting and analyzing college students' performance based on multifaceted data using machine learning," in *Proc. 4th Int. Conf. Adv. Comput. Technol., Inf. Sci. Commun. (CTISC)*, 2022, pp. 1–6.

[7] H. Xu, "Gbdt-lr: A willingness data analysis and prediction model based on machine learning," in *Proc. IEEE Int. Conf. Adv. Electr. Eng. Comput. Appl. (AEECA)*, 2022, pp. 396–401.

[8] Y. Jia and E. Wang, "Research on information anxiety of college students based on support vector machine optimization algorithm," in *Proc. 2nd Int. Conf. Information Science and Education*, 2021, pp. 484–487.

[9] C. Pei, "The construction of a prediction model for the teaching effect of courses in colleges and universities based on machine learning algorithms," *Wireless Communications and Mobile Computing*, vol. 2022, p. 1167454, 2022.

[10] Y. Shi, F. Sun, H. Zuo, and F. Peng, "Analysis of learning behavior characteristics and prediction of learning effect for improving college students' information literacy based on machine learning," *IEEE Access*, vol. 11, pp. 50 447–50 459, 2023.

[11] J. Li, "Machine learning-based evaluation of information literacy enhancement among college teachers," *International Journal of Emerging Technologies in Learning*, vol. 17, no. 22, pp. 116–131, 2022.

[12] I. A. AlShammari, M. Aldhafiri, and Z. Al-Shammari, "A meta-analysis of educational data mining on improvements in learning outcomes," *College Student Journal*, vol. 47, no. 2, pp. 326–333, 2013.

[13] S. Hussain, W. Zhu, W. Zhang, and S. M. R. Abidi, "Student engagement predictions in an e-learning system and their impact on student course assessment scores," *Computational Intelligence and Neuroscience*, vol. 2018, p. 6347186, 2018.

[14] S. Banka, "Information literacy learning behaviour dataset," Kaggle dataset, 2023, https://www.kaggle.com/datasets/sivasankarbanka/information-literacy-learning-behaviour-dataset (accessed Feb. 26, 2026).