

1.3 Thesis Contribution

This thesis makes the following contributions:

1. Developed an AI-Based Early Warning System for predicting students' academic performance using educational engagement and assessment data extracted from the Learning Management System (LMS). The proposed system enables early identification of students who are academically at risk, allowing timely intervention before final assessments.
2. Conducted a comprehensive comparative analysis of five traditional machine learning algorithms (Decision Tree, Logistic Regression, Support Vector Machine, K-Nearest Neighbors, and Naïve Bayes) and three ensemble learning algorithms (Random Forest, AdaBoost, and XGBoost) to determine the most effective predictive model for academic performance classification.
3. Improved data quality and model performance through an effective preprocessing pipeline that included data transformation, stratified train-test splitting, and class balancing using the Synthetic Minority Oversampling Technique (SMOTE), reducing the impact of class imbalance on prediction accuracy.
4. Applied feature importance analysis using the Random Forest algorithm to identify the most influential educational indicators affecting students' academic performance. Low-impact features were eliminated, resulting in a simplified and more efficient predictive model.
5. Enhanced model performance through model optimization by applying Grid Search to optimize the parameters of the best-performing classifiers. The optimization process improved the Decision Tree classifier from an accuracy of **79.93% to 82.80%**, making it the highest-performing model in this study.
6. Provided a practical decision-support framework that assists educational institutions in identifying academically at-risk students and supports educators in implementing timely interventions based on data-driven predictions, thereby improving student success and reducing the likelihood of academic failure.

Chapter 3

Methodology and Proposed Approach

3.1 Introduction

In this chapter, we describe our methodology adopted to design, implement, and evaluate the proposed **AI-Based Early Warning System for Academic Risk Detection**. The proposed methodology utilizes educational data collected from the Learning Management System (LMS) to predict students' academic performance and identify those who are at risk of poor academic achievement. The methodology follows a systematic workflow consisting of data preprocessing, feature selection, machine learning model development, model optimization, and performance evaluation.

Our methodology begins with data collection and preprocessing, where the dataset is cleaned, transformed, and balanced using the Synthetic Minority Oversampling Technique (SMOTE). Next, feature selection is performed using the Random Forest Feature Importance method to identify the most influential educational indicators. The processed dataset is then divided into training and testing subsets before training multiple traditional and ensemble machine learning models.

To further improve prediction performance, **Model Optimization** using **Grid Search** is applied to the best-performing models. Finally, all developed models are evaluated using standard classification metrics, including Accuracy, Precision, Recall, and F1-Score. The best-performing model is then selected as the prediction engine of the proposed Early Warning System.

3.2 Dataset Description

The experiments were conducted using a real educational dataset extracted from the Learning Management System (LMS). The dataset consists of **1,395 student records** represented by **18 educational attributes** describing students' academic performance, participation, assessments, virtual classroom activities, homework completion, discussion participation, and other engagement indicators.

	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Exams Participation %	Homeworks	Homeworks Participation %	Interactive contents	Contents Participation %	Virtual Classrooms	Virtual Classrooms Participation %	AVG Participations %	Discussion Rooms	Discussion Rooms Participation %	Teachers messages	Mailbox Participation %	Student login %	Interactive contents Pdet	Final Mark	Absence
2	0	1	100%	3	0%	0	0%	25%	0	0%	0	0%	100%	0	60	70
3	0	1	0%	3	0%	0	25%	0%	0	0%	0	0%	100%	0	70	60
4	0	1	100%	3	0%	0	25%	25%	0	0%	0	0%	100%	0	90	50
5	0	1	100%	3	33%	0	25%	33%	0	0%	0	0%	100%	0	100	20
6	0	1	0%	3	33%	0	25%	8%	0	0%	0	0%	100%	0	100	100
7	0	1	0%	3	0%	0	25%	0%	0	0%	0	25%	100%	0	56	60
8	0	1	100%	3	0%	0	25%	25%	0	0%	0	25%	100%	0	70	60
9	0	1	100%	3	100%	0	25%	50%	0	0%	0	25%	100%	0	30	60
10	0	1	30%	3	33%	0	25%	8%	0	0%	0	25%	100%	0	100	60
11	0	1	100%	3	33%	0	25%	33%	0	0%	0	25%	100%	0	22	60
12	0	1	100%	3	0%	0	25%	25%	0	0%	0	25%	100%	0	25	60
13	0	1	30%	3	33%	0	25%	8%	0	0%	0	25%	100%	0	55	60
14	0	1	100%	3	0%	0	25%	25%	0	0%	0	25%	100%	0	66	100
15	0	1	100%	3	67%	0	25%	42%	0	0%	0	25%	100%	0	55	60
16	0	1	30%	3	67%	0	25%	17%	0	0%	0	25%	100%	0	23	60
17	0	1	30%	3	0%	0	25%	0%	0	0%	0	25%	100%	0	67	100
18	0	1	100%	3	100%	0	25%	50%	0	0%	0	25%	100%	0	60	100
19	0	1	100%	3	67%	0	25%	42%	0	0%	0	25%	100%	0	67	100
20	0	1	30%	3	0%	0	25%	0%	0	0%	0	25%	100%	0	56	100
21	0	1	100%	3	0%	0	25%	25%	0	0%	0	25%	100%	0	45	100
22	0	4	75%	6	33%	0	25%	27%	0	0%	0	25%	100%	0	34	100
23	0	4	100%	6	83%	0	25%	46%	0	0%	0	25%	100%	0	34	100
24	0	4	50%	6	67%	0	25%	29%	0	0%	0	25%	100%	0	32	100
25	0	4	25%	6	0%	0	25%	6%	0	0%	0	25%	100%	0	55	100
26	0	4	100%	6	50%	0	25%	38%	0	0%	0	25%	100%	0	55	100
27	0	4	30%	6	17%	0	25%	4%	0	0%	0	25%	100%	0	55	100
28	0	4	25%	6	0%	0	25%	6%	0	0%	0	25%	100%	0	55	100
29	0	4	30%	6	0%	0	25%	0%	0	0%	0	25%	100%	0	60	100

The target variable represents students' academic performance and classifies each student into one of four performance categories:

- Excellent
- Good
- Pass
- Fail

These categories represent different levels of academic achievement and form the basis of the proposed early warning system for identifying students who require academic intervention.

To ensure fair evaluation of the developed models, the dataset was divided into training and testing subsets using the **Stratified Train-Test Split** technique. Approximately **80%** of the dataset was allocated for model training, while the remaining **20%** was reserved for model testing. Stratified sampling was selected to preserve the original distribution of student performance classes in both subsets.

3.3 Data Preprocessing

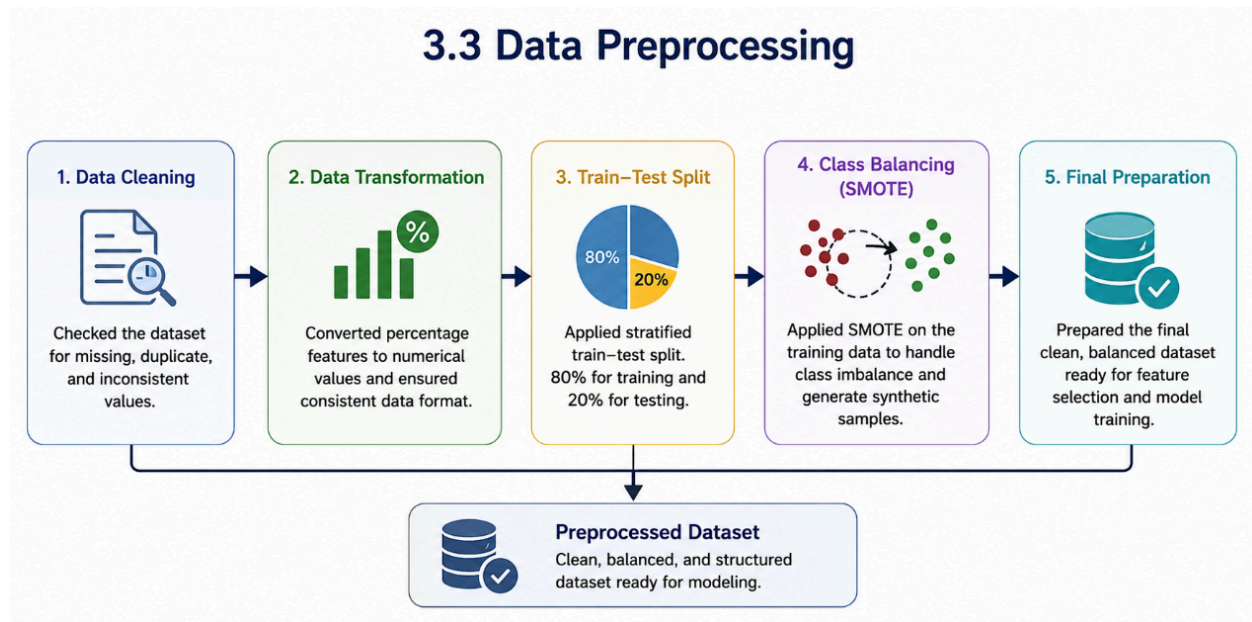
3.3 Data Preprocessing

Data preprocessing is an essential step in developing an accurate and reliable machine learning model. The preprocessing stage was conducted to improve data quality, prepare the dataset for model training, and reduce potential sources of bias. The following preprocessing steps were applied:

- **Data Inspection:** The dataset was examined to verify data consistency and identify any missing or inconsistent values.
- **Data Transformation:** Percentage-based variables related to student participation and engagement were converted into numerical values suitable for machine learning algorithms.

- **Train-Test Split:** The dataset was divided into **80% training data** and **20% testing data** using the **Stratified Train-Test Split** technique to preserve the original class distribution.
- **Class Balancing:** Since the dataset exhibited class imbalance, the **Synthetic Minority Oversampling Technique (SMOTE)** was applied to the training dataset to generate synthetic samples for minority classes and reduce prediction bias.
- **Final Dataset Preparation:** The balanced dataset was prepared and used as the input for feature selection and machine learning model development.

Figure 3.3 illustrates the preprocessing pipeline adopted in this research.



3.4 Feature Selection

Feature selection was performed to identify the most influential variables affecting student academic performance. The following steps were applied:

- **Feature Importance Analysis:** A Random Forest classifier was employed to rank all features according to their contribution to prediction accuracy.
- **Feature Ranking:** Features were ordered from the highest to the lowest importance score.
- **Feature Elimination:** Features with minimal contribution, including **Teachers Messages** and **Student Login Percentage**, were removed from the final dataset.
- **Final Feature Set:** The remaining high-importance features were used as input for all machine learning models.

3.5 Machine Learning Models

To determine the most suitable predictive model for the proposed early warning system, several supervised machine learning algorithms were implemented and evaluated under identical experimental conditions.

The evaluated models were divided into two categories.

3.5.1 Traditional Machine Learning Models

Five traditional classification algorithms were investigated:

- Decision Tree
- Logistic Regression
- Support Vector Machine (SVM)
- K-Nearest Neighbors (KNN)
- Naïve Bayes

These models represent different learning strategies and provide a comprehensive comparison between statistical, distance-based, probabilistic, and tree-based classifiers.

3.5.2 Ensemble Learning Models

In addition to the traditional classifiers, three ensemble learning algorithms were evaluated:

- Random Forest
- AdaBoost
- XGBoost

Ensemble learning combines multiple learners to improve prediction accuracy, increase model stability, and reduce overfitting. These algorithms have demonstrated strong performance in many educational data mining applications and, therefore, were included in the comparative analysis.

3.6 Model Optimization

After evaluating the baseline machine learning models, a model optimization stage was conducted to further improve predictive performance.

The optimization process employed **Grid Search** to automatically evaluate different combinations of model parameters and identify the configuration that produced the highest classification accuracy. This procedure was applied to the best-performing models, including Decision Tree, Random Forest, and XGBoost.

Model optimization improves the ability of classifiers to generalize unseen data while reducing the possibility of overfitting. The optimized models were subsequently re-evaluated using the testing dataset before selecting the final prediction model.

3.7 Proposed AI-Based Early Warning Framework

The proposed AI-Based Early Warning System follows a sequential workflow that transforms raw educational data into meaningful academic risk predictions.

The framework begins by collecting student academic and engagement records from the Learning Management System. After preprocessing and feature selection, the dataset is divided into training and testing subsets, followed by class balancing using SMOTE. Multiple traditional and ensemble learning algorithms are then trained and optimized using Grid Search.

Finally, all developed models are evaluated using standard classification metrics, and the best-performing classifier is selected to predict students' academic risk levels. Based on these predictions, the proposed system generates early warning alerts that support educators in identifying students requiring academic assistance before poor academic performance occurs.

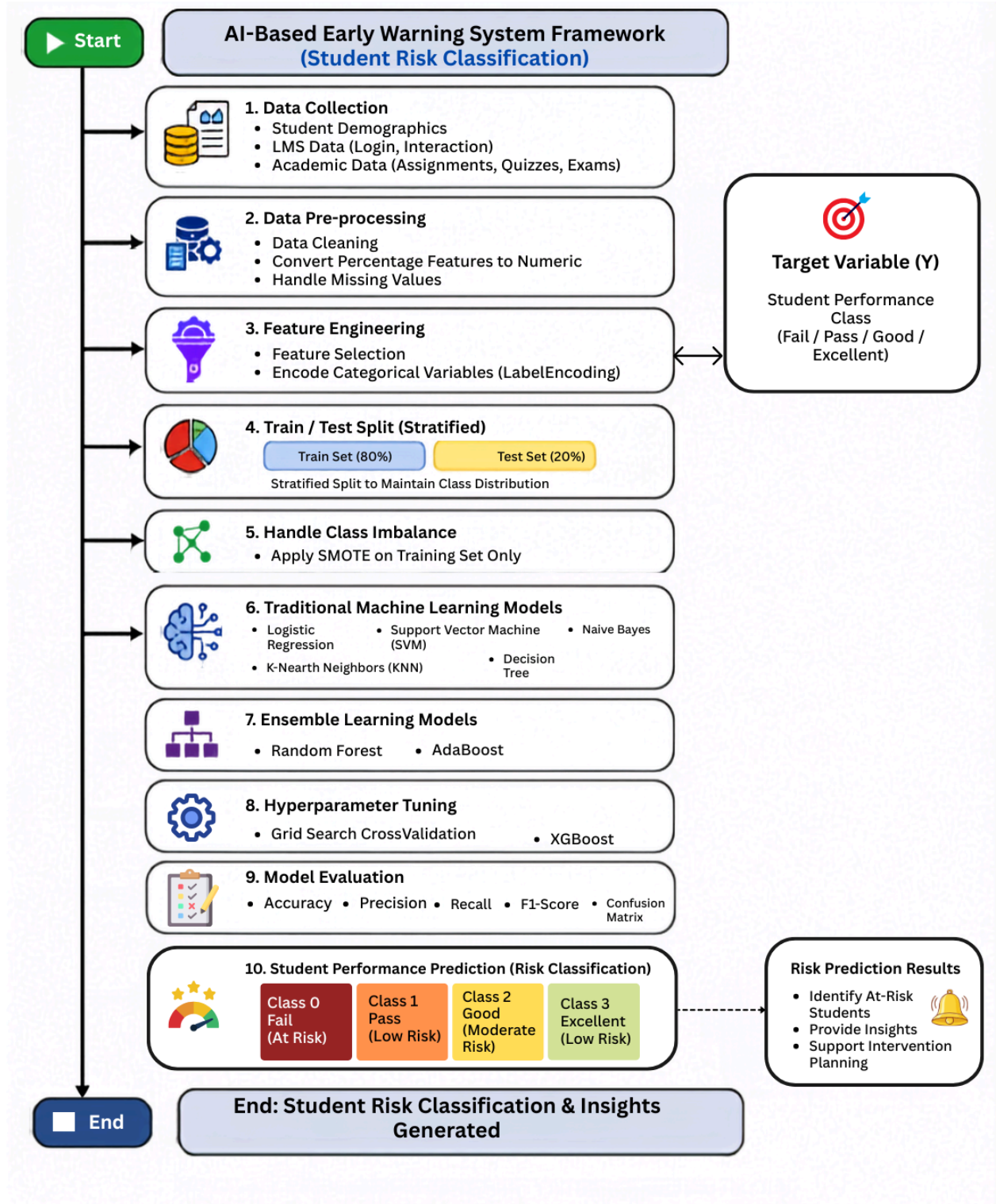


Figure 3.4 Proposed AI-Based Early Warning Framework

3.8 Summary

This chapter presented the proposed approach adopted for developing the AI-Based Early Warning System. It described the educational dataset, preprocessing procedures, feature selection method, machine learning algorithms, model optimization process, and the overall framework of the proposed prediction system. The following chapter presents the experimental results, compares the performance of the developed models, and discusses the effectiveness of the proposed approach in predicting students' academic performance.

8. Results

8.1 Experimental Setup

The experiments were conducted using Google Colab as the development environment and Python 3 as the programming language. The dataset consisted of 1,395 student records and 18 features related to student participation, engagement, assessments, and academic performance. Data preprocessing was performed using Pandas and NumPy libraries. Traditional machine learning algorithms were implemented using Scikit-learn, while XGBoost was used for ensemble learning experiments.

The dataset was divided into training and testing subsets using stratified sampling, with 80% of the data allocated for training and 20% for testing. To address class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was applied to the training dataset. Hyperparameter optimization was subsequently performed using Grid Search to identify the optimal configuration for the best-performing models.

Table 1. Experimental Environment

Component	Specification
Development Environment	Google Colab
Programming Language	Python 3
Dataset Size	1,395 records
Number of Features	18
Training Ratio	80%
Testing Ratio	20%
Class Balancing Method	SMOTE
Hyperparameter Optimization	Grid Search
Random Seed	42

Table 2. Software Libraries Used in the Experiments

Library	Purpose
Pandas	Data preprocessing
NumPy	Numerical computation
Matplotlib	Visualization
Seaborn	Statistical visualization
Scikit-learn	Machine learning algorithms
Imbalanced-learn	SMOTE implementation
XGBoost	Ensemble learning
GridSearchCV	Hyperparameter optimization

8.2 Class Distribution

Before applying SMOTE, the dataset exhibited an imbalanced distribution among the four academic performance classes.

Table 3. Distribution of Student Performance Classes

Class	Number of Students	Percentage (%)
Good	725	51.97
Excellent	322	23.08
Fail	241	17.28
Pass	107	7.67

As shown in Table 3, the Good class represented the majority of the dataset, accounting for approximately 52% of all records, whereas the Pass class represented only 7.67%. Therefore, SMOTE was employed to balance the training dataset and reduce classification bias toward the majority class.

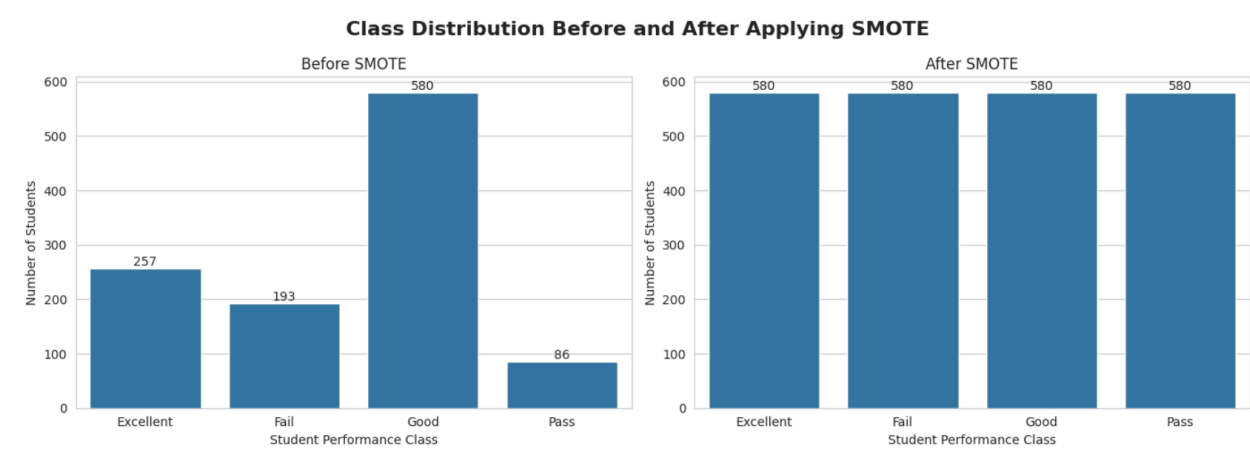


Figure 3. Distribution of Student Performance Classes Before SMOTE.

8.3 Feature Importance Analysis

A Random Forest classifier was employed to evaluate the importance of the input features. The analysis identified the most influential variables affecting student performance prediction.

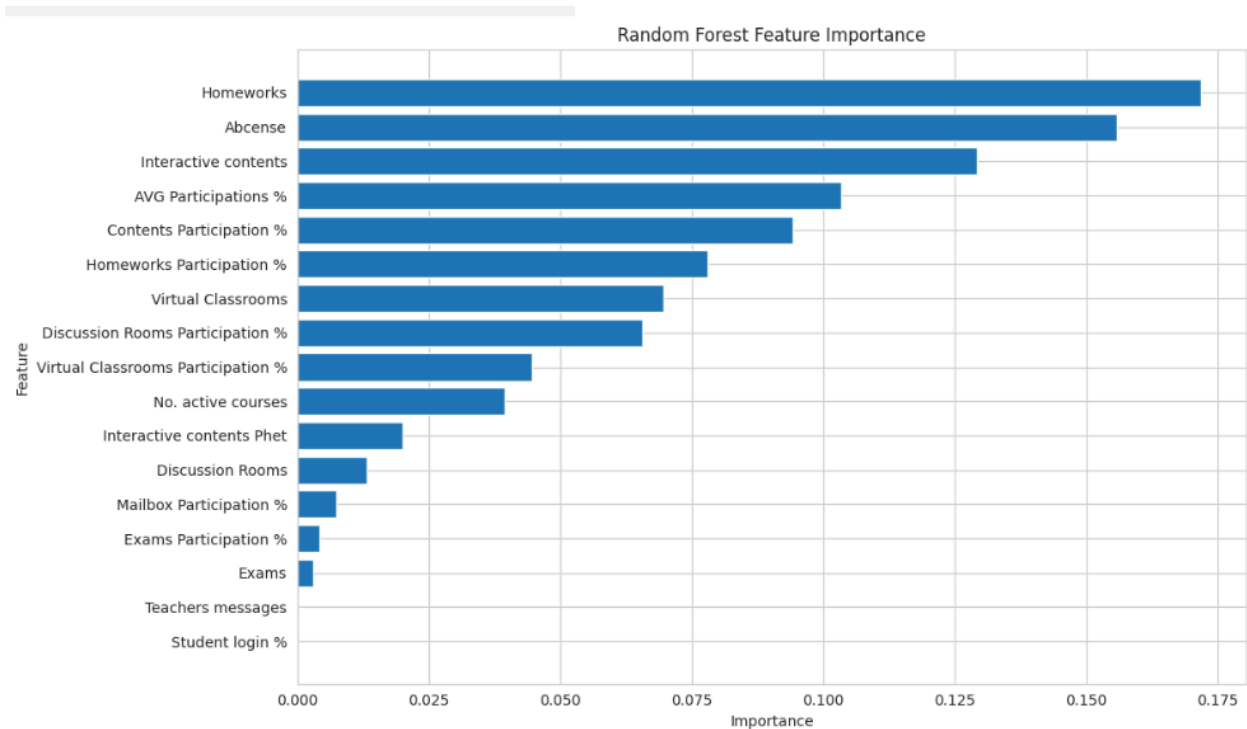


Figure 4. Random Forest Feature Importance Ranking.

The feature importance analysis revealed that participation-related variables and assessment indicators were among the most influential predictors. Conversely, Teacher Messages and Student Login Percentage contributed minimally to the prediction process and were therefore removed from the final feature set.

8.4 Performance of Traditional Machine Learning Models

Five traditional machine learning algorithms were evaluated using Accuracy, Precision, Recall, and F1-Score metrics.

Table 4. Performance Comparison of Traditional Machine Learning Models

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.7993	0.8056	0.7993	0.7984
Logistic Regression	0.6667	0.7098	0.6667	0.6674
Support Vector Machine	0.6452	0.6816	0.6452	0.6403

K-Nearest Neighbors	0.6057	0.6511	0.6057	0.6155
Naïve Bayes	0.1613	0.4396	0.1613	0.1605

The results indicate that the Decision Tree classifier achieved the highest performance among all traditional machine learning algorithms, obtaining an accuracy of 79.93%. Logistic Regression and Support Vector Machine achieved moderate performance levels, while Naïve Bayes demonstrated poor predictive capability due to its strong probabilistic assumptions, which were not suitable for the characteristics of the dataset.

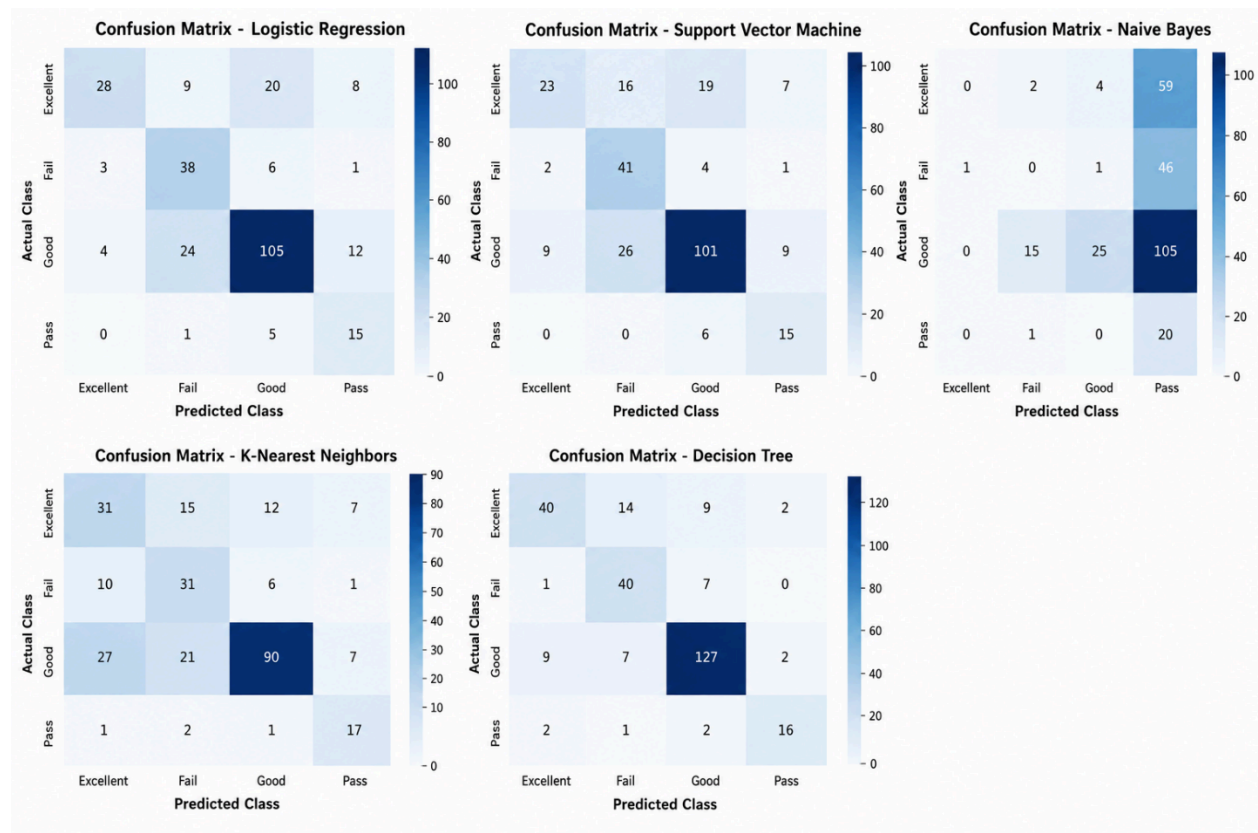


Figure 5. Confusion Matrices of Traditional Machine Learning Models

Figure 5 presents the confusion matrices of the five traditional machine learning models: Logistic Regression, Support Vector Machine (SVM), Naïve Bayes, K-Nearest Neighbors (KNN), and Decision Tree. Overall, the Decision Tree classifier achieved the best classification performance, with the highest number of correctly classified instances concentrated along the main diagonal. Logistic Regression and SVM produced moderate classification results, while KNN showed more misclassifications, particularly for the Good class. Naïve Bayes exhibited the

weakest performance, with a considerable number of samples incorrectly classified into the Pass class. These observations are consistent with the evaluation metrics, confirming that the Decision Tree model outperformed the other traditional machine learning algorithms.

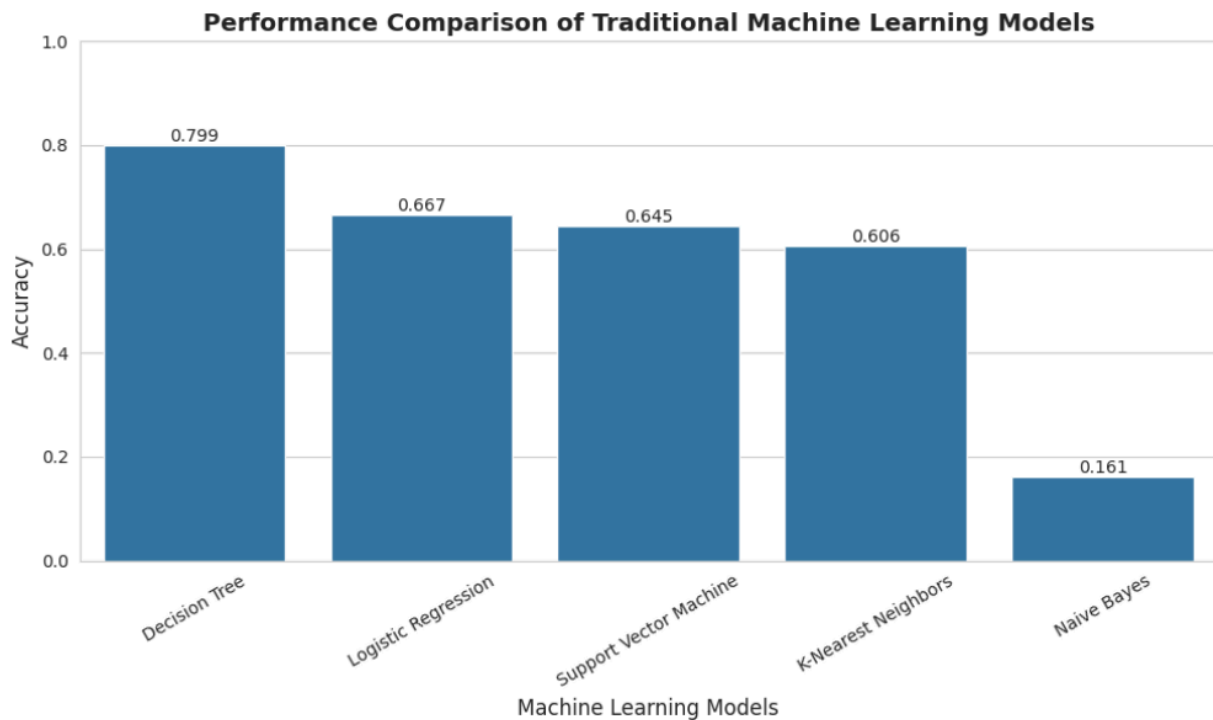


Figure 6. Accuracy Comparison of Traditional Machine Learning Models.

8.5 Performance of Ensemble Learning Models

Three ensemble learning techniques were evaluated to determine whether combining multiple learners could improve prediction performance.

Table 5. Performance Comparison of Ensemble Learning Models

Model	Accuracy	Precision	Recall	F1-Score
XGBoost	0.8208	0.8196	0.8208	0.8175
Random Forest	0.7921	0.7952	0.7921	0.7827
AdaBoost	0.7097	0.7394	0.7097	0.7071

Among the ensemble models, XGBoost achieved the highest classification performance with an accuracy of 82.08%, followed by Random Forest with an accuracy of 79.21%. AdaBoost achieved comparatively lower performance with an accuracy of 70.97%.

The results demonstrate that ensemble learning techniques can effectively model complex educational data and generally outperform most traditional machine learning approaches.

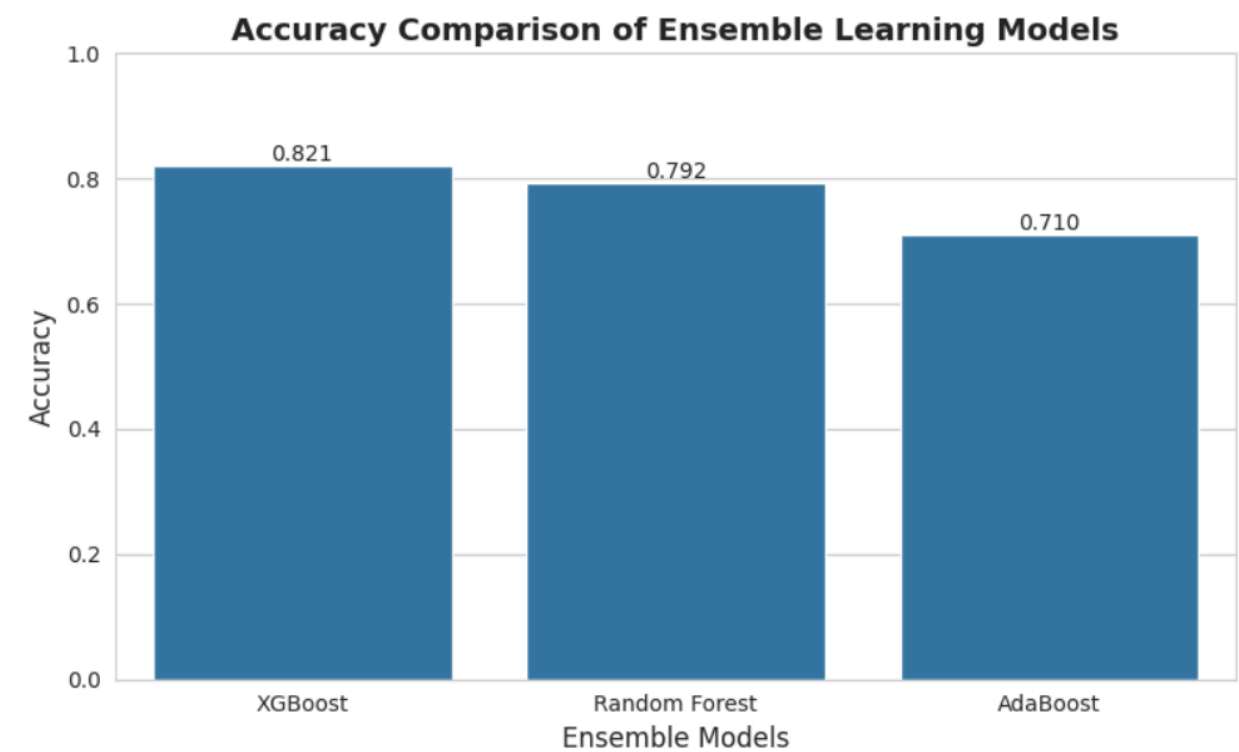


Figure 7. Accuracy Comparison of Ensemble Learning Models.

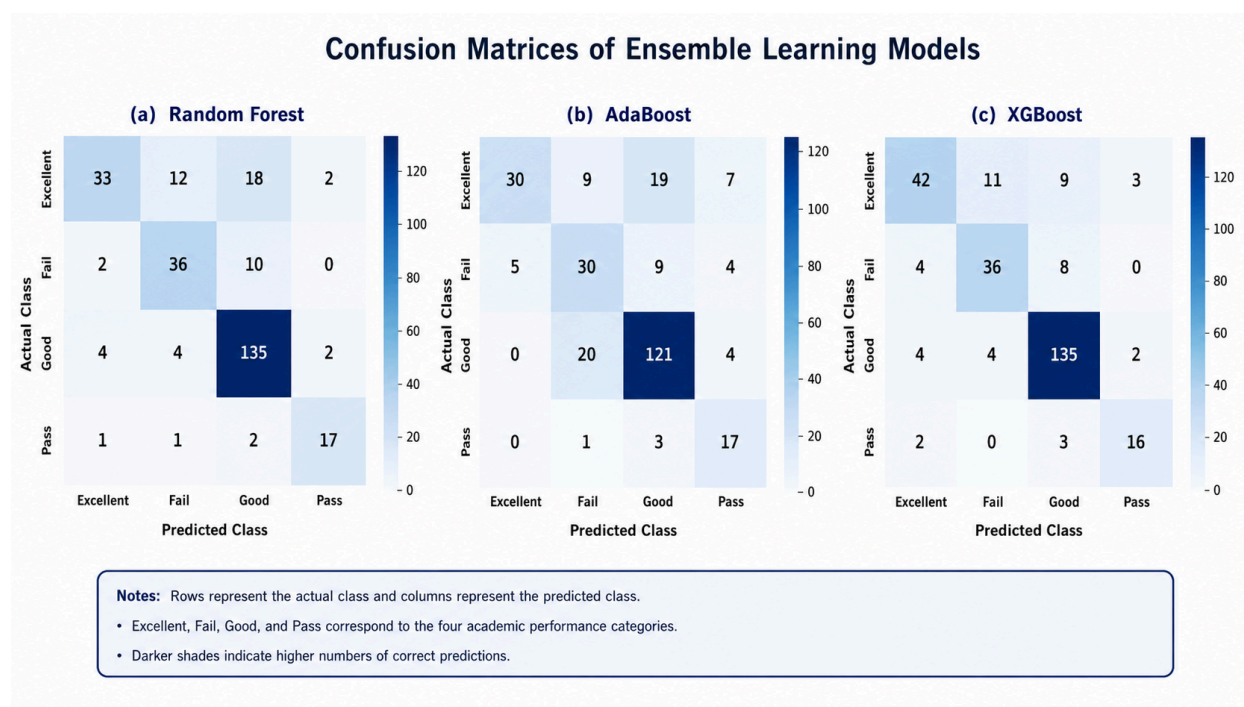


Figure 8. Accuracy Comparison of Traditional Machine Learning Models.

Figure 8 presents the confusion matrices of the three ensemble learning models: Random Forest, AdaBoost, and XGBoost. The majority of observations are concentrated along the main diagonal, indicating correct classifications. Among the three models, XGBoost achieved the best overall classification performance, particularly in identifying the Excellent, Fail, and Good classes. Random Forest produced comparable results, while AdaBoost showed a higher number of misclassifications, especially between the Good and Fail classes. Overall, the confusion matrices confirm the superiority of XGBoost among the evaluated ensemble learning models.

8.6 Comparison Between Traditional and Ensemble Learning Models

A comparison was conducted between the best-performing traditional and ensemble learning models.

Table 6. Comparison Between Traditional and Ensemble Learning Models

Category	Best Model	Accuracy	Precision	Recall	F1-Score
Traditional Machine Learning	Decision Tree	0.7993	0.8056	0.7993	0.7984
Ensemble Learning	XGBoost	0.8208	0.8196	0.8208	0.8175

The comparison indicates that XGBoost outperformed the best traditional machine learning model by approximately 2.15% in classification accuracy. This improvement demonstrates the effectiveness of ensemble learning methods in enhancing predictive performance for educational data mining applications.

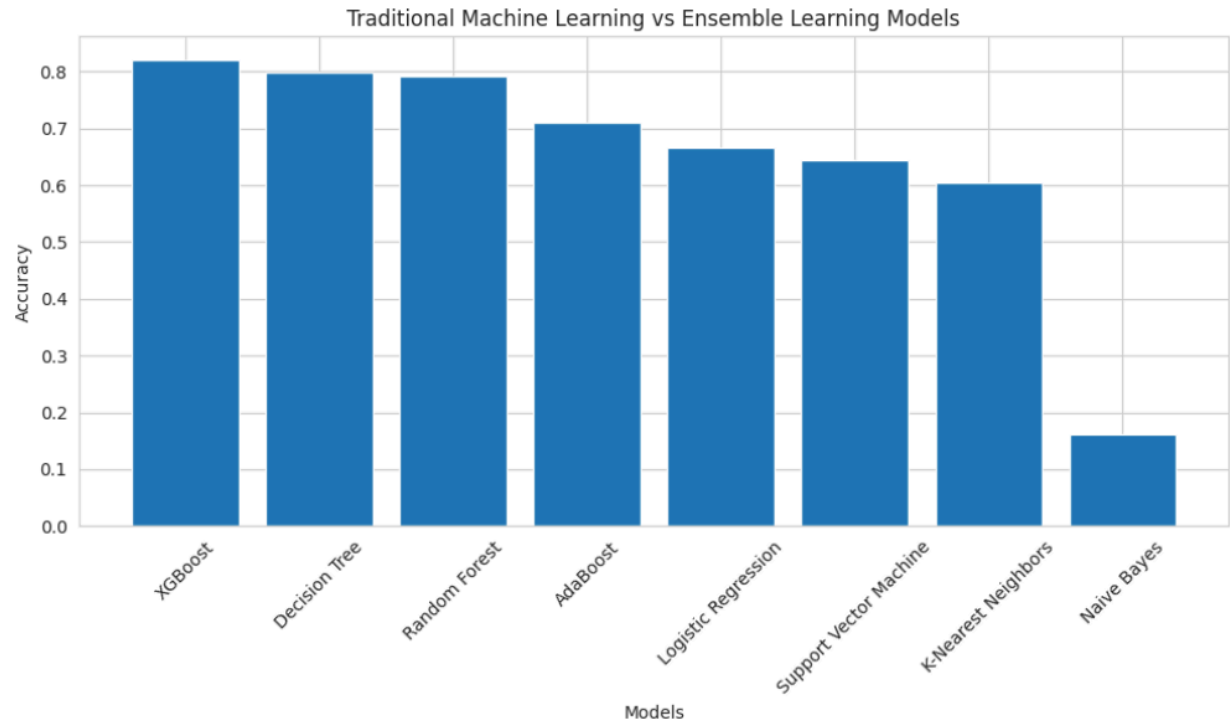


Figure 9. Comparison Between Traditional and Ensemble Learning Models.

8.7 Hyperparameter Optimization Results

To further improve prediction performance, Grid Search optimization was applied to the top-performing models.

Table 7. Hyperparameter Optimization Results

Model	Default Accuracy	Optimized Accuracy	Improvement
Decision Tree	0.7993	0.8280	+0.0287
Random Forest	0.7921	0.8029	+0.0108
XGBoost	0.8208	0.8172	-0.0036

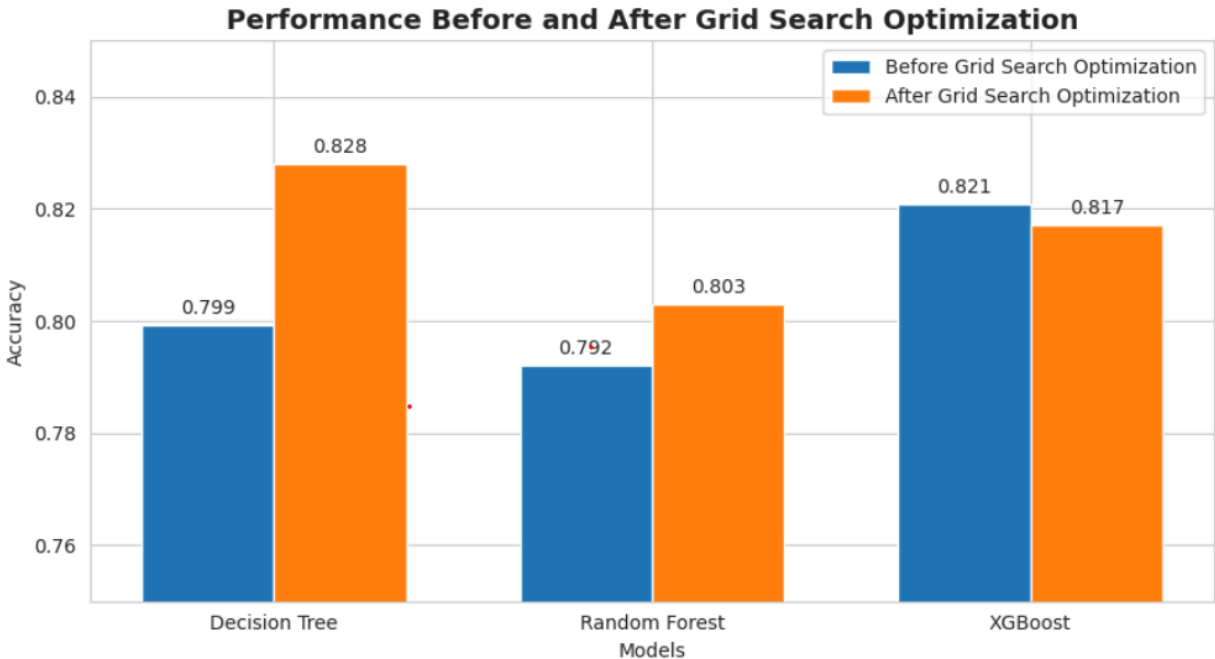


Figure 10. Comparison Between Before and After Grid Search Optimization.

The optimization process significantly improved the performance of the Decision Tree classifier. The model accuracy increased from 79.93% to 82.80%, representing the highest improvement among all evaluated models. Random Forest also benefited from optimization, while XGBoost exhibited a slight decrease in performance after tuning.

8.8 Final Model Selection

Based on the evaluation metrics and optimization results, the Decision Tree classifier was selected as the final predictive model.

Table 8. Final Performance of the Selected Model

Model	Accuracy	Precision	Recall	F1-Score
Optimized Decision Tree	0.8280	0.8256	0.8280	0.8237

The optimized Decision Tree model achieved the highest overall classification performance, with an accuracy of 82.80%, precision of 82.56%, recall of 82.80%, and F1-score of 82.37%. These results indicate that the model can accurately classify students into the four academic performance categories.

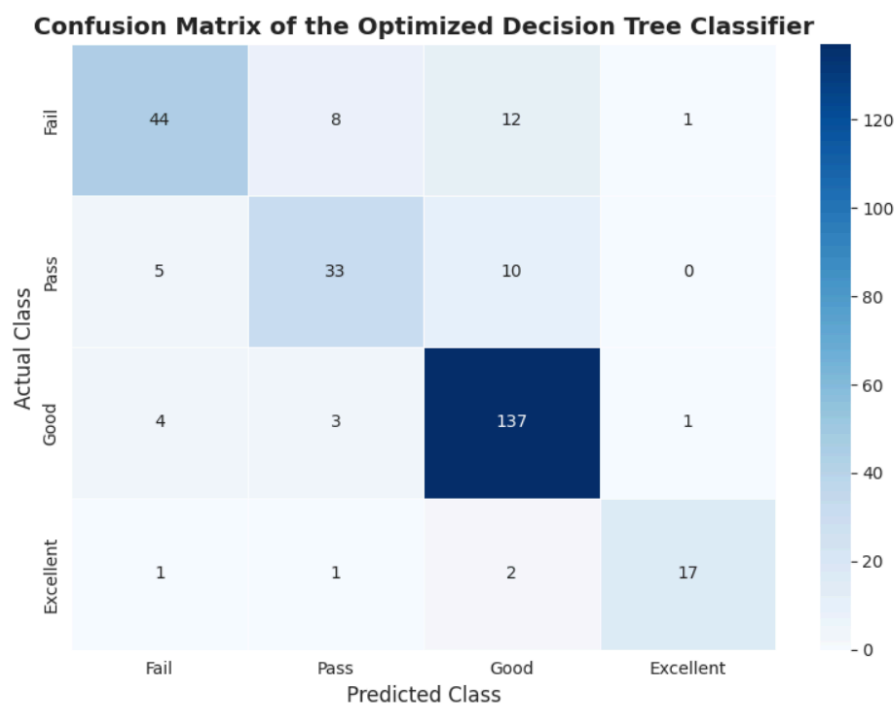


Figure 11. Confusion Matrix of the Optimized Decision Tree Classifier

Figure 11 illustrates the classification performance of the optimized Decision Tree model. Most observations are concentrated along the diagonal, indicating that the majority of students were correctly classified. The model achieved its best performance in the **Good** class, with 137 correctly classified instances, while the **Fail**, **Pass**, and **Excellent** classes also showed satisfactory classification results. Misclassifications mainly occurred between adjacent performance categories, reflecting similarities in students' academic and engagement

characteristics. Overall, the confusion matrix confirms the effectiveness of the optimized Decision Tree classifier in predicting student academic performance.

8.9 Answering the Research Question and Hypothesis

The primary research question of this study was:

"Which machine learning model provides the most accurate prediction of student academic performance?"

The experimental results demonstrate that the optimized Decision Tree classifier achieved the highest predictive performance among all evaluated models. Therefore, the research question is answered by identifying the optimized Decision Tree model as the most suitable classifier for predicting student academic performance within the studied dataset.

Hypothesis H1 stated that machine learning techniques can accurately predict student academic performance based on educational engagement and participation indicators.

The obtained results support H1, as several machine learning models achieved classification accuracies exceeding 79%, while the optimized Decision Tree model achieved an accuracy of 82.80%. Consequently, the hypothesis is accepted, indicating that student performance can be effectively predicted using machine learning approaches.