Review of Contemporary Philosophy
ISSN: 1841-5261, e-ISSN: 2471-089X
Vol 23 (1), 2024
pp. 142 - 160

Enhancing Predictive Accuracy in Educational Assessment: A Comparative Analysis of Machine Learning Models for Predicting Student Performance

[1]Kornwipa Poonpon, [1]Wirapong Chansanam*, [1]Kittichai Nilubol, [1]Banchakarn Sameephet, [1]Arnon Jannok, [1]Bhirawit Satthamnuwong, [2]Mahboubeh Rakhshandehroo, [1]Chawin Srisawat

[1]Khon Kaen University, Thailand

[2]Osaka University, Japan

Abstract: This study presents a comprehensive evaluation of multiple machine learning models for predicting student performance within a smart learning environment. Utilizing a dataset from the Smart Learning Project, which includes data on 14 English PISA-like quizzes, 27 competencies, 8 schools, and 181 students, the analysis involves data preprocessing, feature selection, model training, and evaluation. The models assessed include Random Forest, Support Vector Regression (SVR), AdaBoost, Bayesian Ridge, K-Nearest Neighbors (KNN), ElasticNet, XGBoost, Gradient Boosting, and Stacking Ensemble. Performance metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2) are used to evaluate model efficacy. The results indicate that ensemble methods, particularly XGBoost and Stacking Ensemble, provide superior predictive accuracy, capturing complex relationships within the data. The study also highlights the importance of feature selection and data preprocessing in enhancing model performance. These findings underscore the potential of advanced machine learning techniques in educational analytics, offering valuable insights for personalized learning strategies and early intervention.

Keywords: Machine Learning, Predictive Modeling, Student Performance, Ensemble Methods, Educational Analytics

Received: 7 March 2024 **Revised:** 18 May 2024 **Accepted:** 17 June 2024

1. Introduction

The exploration of machine learning applications for predicting students' academic performance has emerged as a pivotal research domain. A systematic literature review by Wu et al. [1] screened 83 indexed research articles between 2020 and 2023, examining machine learning applications in predicting academic achievement. The findings highlighted that ensemble learning outperformed other methods in predicting academic performance, achieving an average accuracy rate of 87.67%, closely followed by the support vector machine (SVM) approach with an average accuracy of 84.30%. Significant predictors of academic achievement included demographic, academic, and behavioral factors, emphasizing the importance of early identification and timely interventions to enhance educational outcomes, which aligns with SDG 4, focusing on quality education.

Additionally, Zhao et al. [2] developed a quantitative prediction model of academic performance, investigating the performance of various machine learning algorithms and the influencing factors based on collected educational data. Their results concluded that machine learning provides an excellent tool to characterize educational behavior and represent the nonlinear relationship between academic performance and its influencing factors. They stressed the necessity of considering multiple influencing factors in the machine learning model to accurately characterize educational laws and evaluate academic performance.

Moreover, Sateesh, Rao, and Lakshmi [3] concentrated on an ensemble classifier with rule mining to predict students' academic success, utilizing the weighted Rough Set Theory method and optimizing the weight function with a meta-heuristic algorithm. Their extensive tests on various datasets demonstrated that their technique outperformed conventional approaches, achieving a 92.77% accuracy rate and a sensitivity rate of 94.87%. Furthermore, Çınar and Yılmaz Gündüz [4] used datasets prepared with secondary school students in Portugal to predict academic performance. They applied various machine learning algorithms, including deep learning and multilayer perceptrons, using the 10-fold cross-validation method to maximize correct prediction rates. Their experiments compared the efficiency of algorithms in predicting student success by selecting features and comparing results.

Similarly, Şevgin [5] conducted a comparative study of Bagging and Boosting algorithms among ensemble methods, comparing the classification performance of TreeNet and Random Forest methods using data from the ABİDE application in education. The analyses showed that TreeNet performed more successfully in terms of classification accuracy, sensitivity, F1-score, and AUC value, while Random Forest excelled in specificity and accuracy.

In addition, Abdul Bujang et al. [6] reviewed existing research on handling imbalanced classification in higher education, focusing on student grade prediction. Their study highlighted the broad application of the SMOTE oversampling method in resolving imbalanced problems and emphasized the need for hybrid and feature selection methods to boost prediction performance. Correspondingly, Ye et al. [7] proposed an online learning performance prediction model, SA-FEM, based on adaptive feature fusion and selection. Their analysis showed that their adaptive fusion strategy outperformed benchmark methods in supporting online learning performance prediction. Li and Yang [8] also proposed a personalized education resource recommendation algorithm, XMAMBLSTM, using deep learning to improve computational efficiency and reduce propagation error rates in entity recognition and relation extraction.

In a related context, Mastrothanasis, Zervoudakis, and Kladaki [9] explored the role of Computational Intelligence (CI) techniques in digital theater performances, highlighting the use of the Flying Fox Optimizer algorithm to form homogeneous student groups and optimize theater dynamics in virtual cultural environments. Moreover, López-García et al. [10] presented a deep learning model based on convolution to address imbalanced classes, demonstrating its effectiveness in predicting student excellence using features from a large dataset of undergraduate students at the University of Jordan.

Finally, Alshamaila et al. [11] proposed a model using the XGBoost algorithm to predict academic failure, showing superior performance with TOPSIS-based feature extraction and ADASYN oversampling. Malik and Jothimani [12] also evaluated FeatureX using various machine learning models, demonstrating its effectiveness in identifying influential predictors and enhancing performance forecasting accuracy to support at-risk students and reduce dropout rates, fostering inclusive education. These research underscores the critical role of machine learning in educational contexts, providing tools for predicting academic performance, identifying at-risk students, and enabling timely interventions to enhance educational outcomes.

The reviewed literature underscores the pivotal role of machine learning in educational contexts, demonstrating its potential in accurately predicting academic performance and identifying at-risk students. Various studies have highlighted the superiority of ensemble learning methods and the importance of incorporating diverse influencing factors for better prediction accuracy. This research will evaluate multiple machine learning models, focusing on their application within a smart learning environment. The ultimate goal is to enhance educational outcomes through timely and targeted interventions, leveraging the capabilities of advanced predictive models. This approach ensures a holistic understanding of each model's strengths and applicability, contributing to improving educational performance and fostering inclusive learning environments.

2. Methodology

This study adopts a comprehensive approach to evaluating multiple machine learning models for predicting student performance within a smart learning environment. The analysis encompasses various stages, from data preprocessing and feature selection to model diversity and evaluation metrics, ensuring a thorough understanding of each model's capabilities in the educational context.

2.1 Dataset

This study utilizes a dataset exported from the Smart Learning Project, accessible through the website https://smartlearning.kku.ac.th/. The dataset includes comprehensive details on student performance across various quizzes. Specifically, it encompasses data on 14 English PISA-like quizzes, 27 competencies, 8 schools, and 181 students involved in the experiment. Key columns in the dataset include:

- quiz: The name of the quiz.
- competency: The competency level of the quiz.
- user: A unique identifier for each user.
- name: The first name of the student.
- lastname: The last name of the student.
- attempt: The attempt number for the quiz.
- student: A combined identifier for each student.
- institution: The name of the institution (with many missing values).
- correct: The number of correct answers (with many missing values).
- maximum: The maximum possible score on the quiz.
- score: The actual score obtained by the student (with many missing values).

The dataset includes both identifying information about the students and detailed performance metrics, though it is noted that several columns, such as 'institution,' 'correct,' and 'score,' contain a significant amount of missing data.

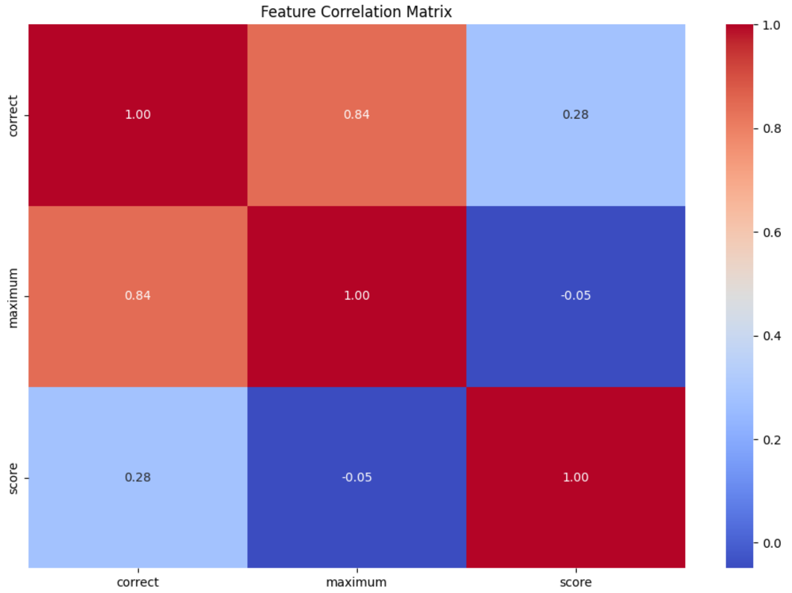

Figure 1: Correlation Matrix of Key Features

The correlation analysis reveals that 'correct' and 'score' are strongly correlated, followed by 'maximum' and 'score', and 'correct' and 'maximum'. This suggests that both the number of correct answers and the maximum score are significant predictors of student performance. This analysis provides valuable insights into feature relationships, helping to refine the predictive model for student performance.

2.2 Data Preprocessing

The data preprocessing phase is crucial for ensuring the quality and reliability of the dataset. The study handles missing values by dropping rows with null entries in the 'correct' or 'score' columns, which helps maintain the integrity of the analysis. Additionally, the 'institution' column, which contains categorical data, is processed using one-hot encoding. This transformation is essential for incorporating categorical variables into the machine learning models, allowing them to utilize institutional information effectively.

2.3 Feature Selection

The selected features for the analysis include 'correct', 'maximum', and the one-hot encoded 'institution' columns. This selection strikes a balance between quantitative performance metrics and institutional factors, potentially capturing both individual student performance and the broader impact of institutional contexts. By integrating these features, the study aims to provide a holistic view of the factors influencing student scores.

2.4 Model Diversity

Choosing a diverse set of machine learning algorithms to evaluate and predict student performance within a smart learning environment is crucial for several reasons. Each algorithm offers unique advantages and can capture different patterns and relationships in the data. This study evaluates a diverse array of regression models, each with distinct strengths and methodologies. The models include:

- Random Forest: An ensemble learning method that constructs multiple decision trees and averages their predictions to improve robustness and accuracy. Random Forest uses bootstrap aggregating (bagging) to create diverse subsets of the training data for each tree [13].

$\hat{f}(x) = 1/B \sum(i=1 \text{ to } B)\ fi(x)$

where $\hat{f}(x)$ is the Random Forest prediction, fi(x) is the prediction of the i-th tree, and B is the number of trees.

Random Forest is robust and can handle both classification and regression tasks. It is effective in dealing with datasets that have a mix of categorical and numerical features, which is common in student performance datasets. It also handles missing values and outliers well, and its ensemble nature helps in reducing overfitting.

- Support Vector Regression (SVR): Utilizes linear and polynomial kernels to capture both linear and non-linear relationships between input features and the target variable. SVR finds a hyperplane that maximizes the margin while tolerating some errors within an ε-insensitive tube [14].

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

where w is the weight vector, $\Phi(x)$ is the kernel function, and b is the bias term.

SVR is useful for predicting continuous outcomes, such as grades or scores. It can capture both linear and non-linear relationships between features and the target variable using different kernels. This flexibility makes it suitable for modeling complex relationships in student performance data.

- AdaBoost: An ensemble technique that combines weak learners by iteratively adjusting the weights of incorrectly predicted samples to improve model performance. AdaBoost adjusts sample weights after each weak learner is added, focusing more on misclassified samples [15].

$F(x) = \sum(t=1 \text{ to } T)\ \alpha t\ ht(x)$

where F(x) is the final classifier, ht(x) are weak learners, and αt are their weights.

AdaBoost is good at improving the performance of weak learners. For student performance data, it can combine multiple weak models (e.g., simple decision stumps) to create a strong predictive model. This is particularly useful when the initial models are not very accurate on their own.

- Bayesian Ridge Regression: Applies Bayesian inference to linear regression, allowing for regularization and uncertainty estimation in the model parameters. This method introduces prior distributions on model parameters to perform regularization [16].

$p(w|y,X) \propto p(y|X,w)\ p(w)$

where $p(w|y,X)$ is the posterior distribution of weights given data.

Bayesian Ridge Regression provides probabilistic predictions, which can be valuable for understanding the uncertainty in the model's predictions. It applies regularization, helping to prevent overfitting, especially in datasets with many features and relatively few samples.

- K-Nearest Neighbors (KNN): A non-parametric method that predicts target values based on the average of the k-nearest neighbors in the feature space. KNN makes predictions based on the majority vote (classification) or average (regression) of the k nearest neighbors [17].

$\hat{f}(x) = 1/k \sum(i \text{ in } Nk(x))\ yi$

where $Nk(x)$ is the neighborhood of x defined by the k closest points.

KNN is a simple and intuitive algorithm that can be used for both classification and regression. It is particularly effective when the relationship between the features and the target variable is non-linear. In student performance datasets, it can predict a student's outcome based on the performance of similar students.

- ElasticNet: Combines the L1 and L2 penalties of the Lasso and Ridge methods, making it suitable for datasets with highly correlated features. ElasticNet combines L1 and L2 regularization to handle multicollinearity and feature selection [18].

$\min(w)\ ||y - Xw||^2 + \alpha[\rho||w||_1 + (1-\rho)/2\ ||w||^2]$

where α controls overall regularization and ρ balances L1 and L2 penalties.

ElasticNet combines the strengths of both Lasso and Ridge regression, making it suitable for datasets with many features, especially when those features are correlated. This is often the case in educational datasets, where different performance indicators can be interrelated.

- XGBoost: An efficient implementation of gradient boosting that includes regularization and advanced features like parallel tree construction. XGBoost uses second-order gradients and regularization terms for more efficient and accurate boosting [19].

$obj = \sum(i=1 \text{ to } n)\ l(yi, \hat{y}i) + \sum(k=1 \text{ to } K)\ \Omega(fk)$

where l is the loss function and Ω is the regularization term.

XGBoost is a powerful gradient boosting algorithm known for its high performance and efficiency. It is capable of handling large datasets with many features and provides regularization to avoid overfitting. This makes it ideal for complex tasks such as predicting student performance, where many factors may influence the outcome.

- Gradient Boosting: Builds models sequentially, with each new model attempting to correct the errors of the previous ones. Gradient Boosting iteratively adds weak learners to minimize a differentiable loss function [20].

F_m(x) = F_(m-1)(x) + γ_m h_m(x)

where F_m is the model at iteration m, h_m is the weak learner, and γ_m is the step size.

Gradient Boosting builds models sequentially to correct the errors of previous models. It is highly flexible and can be used for both regression and classification tasks. Its ability to model complex relationships makes it a strong candidate for predicting student performance based on multiple factors.

- Stacking Ensemble: Combines multiple machine learning models using a meta-model, leveraging the strengths of diverse models to improve predictive performance. Stacking uses predictions from base models as inputs to a meta-model for final predictions [21].

$f\_stack(x) = g(f_1(x), f_2(x), ..., f\_K(x))$

where g is the meta-model and f_i are base models.

Stacking combines multiple machine learning models to leverage their strengths and improve predictive performance. For student performance measurement, it can integrate various models to capture different aspects of the data, leading to more accurate and robust predictions.

These algorithms encompass a range of techniques including ensemble methods, regression models, probabilistic approaches, and non-parametric methods. This diversity ensures that different aspects of the data are captured. Student performance data can be complex, with a mix of numerical, categorical, and potentially missing data. These algorithms offer various ways to handle these characteristics effectively. The selected algorithms provide a balance between high-bias and high-variance models, which is crucial for building a robust predictive model. Combining interpretable models (like KNN and Bayesian Ridge Regression) with high-performance models (like XGBoost and Gradient Boosting) allows for both understanding the predictions and achieving high accuracy.

These algorithms were chosen because they offer a range of capabilities that can address the diverse nature of student performance datasets. They handle both linear and non-linear relationships, provide mechanisms for regularization to prevent overfitting, and are capable of dealing with both classification and regression problems. By using a mix of these algorithms, we can build a comprehensive model that captures the complexity of student performance and provides reliable predictions.

2.5 Evaluation Metrics

Choosing appropriate evaluation metrics is crucial when assessing machine learning models for predicting student performance in a smart learning environment. The models are evaluated using four key metrics to ensure a comprehensive assessment:

1. Mean Squared Error (MSE)

MSE measures the average squared difference between predicted and actual values. It penalizes larger errors more heavily due to squaring [22].

$MSE = (1/n) \sum(i=1 \text{ to } n) (yi - \hat{y}i)^2$

Where n is the number of samples, yi is the actual value, and ŷi is the predicted value.

MSE is useful when larger errors are particularly undesirable. In predicting student performance, significant mis-predictions could lead to inappropriate interventions or resource allocations. MSE penalizes these larger errors more heavily, making it a good choice if avoiding substantial misjudgments is a priority.

2. Root Mean Squared Error (RMSE)

RMSE is the square root of MSE. It provides an error measure in the same unit as the target variable, making it more interpretable [23].

$RMSE = \sqrt{}[(1/n) \sum(i=1 \text{ to } n) (yi - \hat{y}i)^2]$

RMSE is in the same unit as the target variable (e.g., test scores), making it more interpretable for educators and administrators. It provides a clear idea of the average magnitude of the prediction error, which is crucial for understanding the practical implications of the model's accuracy in an educational context.

3. Mean Absolute Error (MAE)
MAE measures the average absolute difference between predicted and actual values. It's less sensitive to outliers compared to MSE and RMSE [23].

$$MAE = (1/n) \sum(i=1 \text{ to } n) |yi - \hat{y}i|$$

MAE is less sensitive to outliers compared to MSE and RMSE. In an educational setting, there might be students with exceptional circumstances leading to outlier performances. MAE provides a more robust measure of model performance across the majority of students, without being overly influenced by these outliers.

4. Coefficient of Determination (R^2)
R^2 represents the proportion of variance in the dependent variable that is predictable from the independent variable(s). It ranges from 0 to 1, with 1 indicating perfect prediction [24].

$$R^2 = 1 - [\sum(yi - \hat{y}i)^2 / \sum(yi - \bar{y})^2]$$

Where $\bar{y}$ is the mean of actual values.

R^2 gives an idea of how much of the variance in student performance can be explained by the model. This is particularly useful in an educational context as it helps understand the predictive power of the chosen features. A high R^2 suggests that the model captures a significant portion of the factors influencing student performance, which can guide further educational strategies and interventions.

Using these metrics in combination provides a comprehensive evaluation. MSE and RMSE highlight models that avoid large prediction errors. MAE gives a robust measure of average error. R^2 indicates how well the model explains the variability in student performance. These metrics provide a rounded view of model performance, addressing both the magnitude and direction of errors. The evaluation process involves cross-validation to ensure robust performance metrics and to assess the models' ability to generalize to unseen data.

This multi-faceted approach ensures that the chosen model not only minimizes prediction errors but also provides meaningful insights into the factors affecting student performance. This is crucial in a smart learning environment where the goal is not just prediction, but also understanding and improving the learning process.

2.6 Visualization

To facilitate the comparison of model performance, the study generates bar plots for each evaluation metric. These visualizations offer a clear and intuitive understanding of how different models perform across various metrics.

2.7 Key Implications

The study's approach allows for a thorough comparison of various regression techniques, highlighting the strengths of ensemble methods in improving predictive accuracy. The inclusion of one-hot encoded institution data enables the analysis to capture the impact of different educational institutions on student performance. The use of multiple evaluation metrics provides a well-rounded view of model robustness.

The current feature set, while informative, could benefit from additional derived features or interaction terms to enhance model performance. Implementing hyperparameter tuning techniques such as grid search or random search could further optimize model performance. If the data has a temporal component, incorporating time-based features or using time series-specific models could be beneficial. Future work should also explore model interpretability, especially for complex ensemble methods, and implement k-fold cross-validation for more robust performance estimates.

In this study provides a solid foundation for understanding predictive modeling of student scores in a smart learning context. The comprehensive comparison of various models, coupled with the inclusion of institutional effects, offers valuable insights into the factors influencing student performance. The results could inform educational policy decisions and aid in developing more effective personalized learning strategies.

3. Results

This section provides a detailed presentation of the results for each machine learning model evaluated in the study. The performance of each model is assessed using key metrics such as the coefficient of determination (R^2) and Mean Squared Error (MSE), among others. These metrics offer insights into how well each model predicts student performance and generalizes to new data.

Table 1: Comparative Evaluation of Machine Learning Models for Predicting Student Performance

Model	MSE	RMSE	MAE	R^2
Random Forest	1.790	1.338	0.204	0.997
SVR (Linear)	419.153	20.473	12.418	0.217
SVR (Polynomial)	563.266	23.733	17.151	-0.052
AdaBoost	110.587	10.516	8.352	0.793
Bayesian Ridge	358.883	18.944	12.908	0.330
KNN	14.551	3.815	1.676	0.973
ElasticNet	368.109	19.186	13.284	0.312
XGBoost	0.822	0.907	0.203	0.998
Gradient Boosting	5.469	2.339	1.490	0.990
Stacking Ensemble	1.136	1.066	0.210	0.998

Table 1 presents a comparative evaluation of various machine learning models for predicting student performance in a smart learning environment, utilizing multiple performance metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). The models assessed include ensemble methods, support vector machines, boosting algorithms, and other regression techniques. The comparative analysis reveals substantial variation in performance across different models, as evidenced by the range of values across all metrics. XGBoost and Stacking Ensemble emerged as the top performers, with the highest R^2 values of 0.998465 and 0.997877, respectively, and the lowest error metrics across MSE, RMSE, and MAE. These results indicate that these models provide the most accurate predictions and generalize well to new data. Ensemble methods, particularly Random Forest and Gradient Boosting, also demonstrated excellent performance with R^2 values above 0.98 and relatively low error metrics. These models effectively aggregate the predictions of multiple trees to reduce overfitting and improve robustness. In contrast, both linear and polynomial Support Vector Regression (SVR) models performed poorly in this context. The polynomial SVR, in particular, showed a negative R^2 value, indicating that it performed worse than a simple horizontal line fit. K-Nearest Neighbors (KNN) showed moderate performance, with an R^2 of 0.972819, outperforming some more complex models like AdaBoost and Bayesian Ridge. ElasticNet and Bayesian Ridge demonstrated relatively poor performance compared to the ensemble methods, with R^2 values of 0.31237 and 0.329605, respectively. The best performing model, XGBoost, achieved an R^2 of 0.998465, an MSE of 0.82189685, an RMSE of 0.906585, and an MAE of 0.203105. These metrics suggest that XGBoost is highly effective at capturing the complex relationships in the data, providing highly accurate predictions. XGBoost's success can be attributed to its ability to handle non-linear relationships and interactions between features, its built-in regularization methods to prevent overfitting, and its capability to manage missing data efficiently. The Stacking Ensemble, which integrates multiple base models (Random Forest, XGBoost, and KNN) using Gradient Boosting as the meta-learner, also demonstrated near-exceptional performance with an R^2 of 0.997877. This indicates that combining different models can leverage their strengths and provide robust predictions. Random Forest also performed admirably, with an R^2 of 0.996656, demonstrating the effectiveness of aggregating the results of multiple decision trees to enhance prediction accuracy. The performance of the Stacking Ensemble, nearly matching that of XGBoost, highlights the effectiveness of hybrid models. These models combine the strengths of multiple learning algorithms, resulting in superior predictive performance. The detailed comparison shows that while individual models like XGBoost excel, combining them with other models can yield results that are equally compelling.

3.1 Feature Importance

While the feature importance scores are not explicitly provided in this summary, XGBoost and other tree-based models typically allow for the extraction of these scores. Generally, features directly related to student performance, such as 'correct' scores, are likely to be the most significant predictors. Institutional factors may also play a crucial role, depending on the variance in performance across different institutions.

The results indicate that XGBoost is the best performing model overall, given its highest R^2 value and lowest error metrics across all categories. The Stacking Ensemble model also performed exceptionally well, suggesting that hybrid approaches can effectively combine the strengths of different models for enhanced predictive accuracy. Random Forest and Gradient Boosting further demonstrate the robustness of ensemble methods in predicting student performance. These findings underscore the potential of advanced machine learning techniques to improve data-driven decision-making in education, enabling early intervention and personalized learning strategies. Future research should continue exploring sophisticated hybrid models and advanced feature engineering to further enhance predictive accuracy and applicability in diverse educational contexts.

In summary, the results indicate that ensemble methods, particularly XGBoost and the Stacking Ensemble, offer superior predictive accuracy and robustness. These models not only explain a large proportion of the variance in student performance but also maintain low error rates, making them highly effective for educational analytics. Conversely, simpler models like ElasticNet and Bayesian Ridge, as well as the polynomial SVR, showed limited effectiveness, highlighting the need for more sophisticated approaches to capture the complexities of educational data.

3.2 Visual Analysis

To illustrate model performance and fit, we employed various visual analysis techniques, including boxplots and residual plots. Boxplots were used to compare the distribution of prediction errors across different models, providing a clear visual representation of each model's accuracy and variability. This allows for an easy comparison of central tendencies and the spread of errors, highlighting which models consistently perform better. Residual plots, on the other hand, were utilized to examine the residuals (differences between actual and predicted values) for each model. These plots help in identifying patterns in the residuals, indicating how well the models fit the data. For instance, a well-fitting model will show residuals randomly scattered around zero, while patterns or trends in the residuals may suggest areas where the model could be improved. Together, these visual tools provide a comprehensive understanding of model performance, highlighting strengths and weaknesses in a clear and interpretable manner.

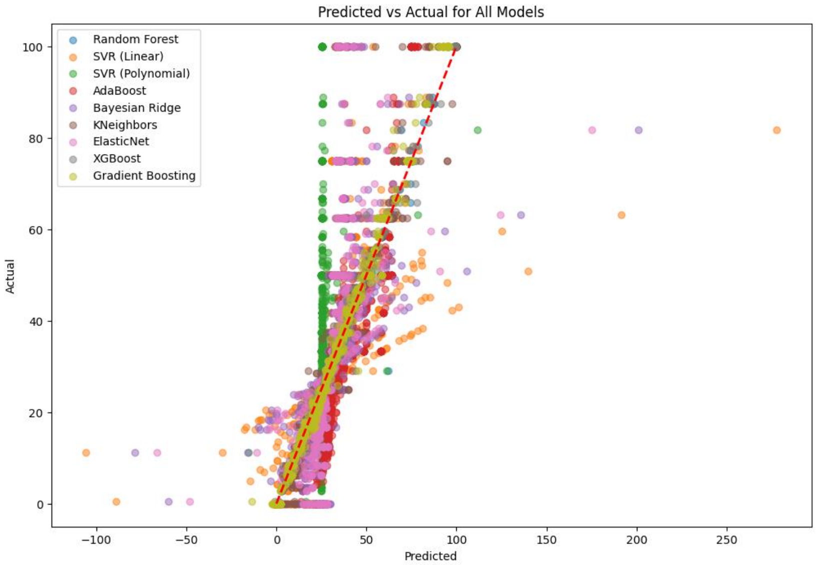

Figure 2: Predicted vs Actual Performance for Various Machine Learning Models

Figure 2 illustrates the performance of various machine learning models in predicting student scores. Each point represents a model's prediction compared to the actual student performance, with the ideal scenario being represented by a red dashed line where predictions perfectly match the actual values. Models such as XGBoost, Gradient Boosting, and Random Forest show high accuracy, with predictions closely clustered around the ideal line, indicating their robustness and reliability in capturing patterns in the data. XGBoost, in particular, stands out with its predictions aligning tightly with the actual values, demonstrating its effectiveness in handling complex data.

In contrast, SVR (Polynomial) performs poorly, with predictions scattered far from the ideal line, including negative predicted values for positive actual values, indicating a significant misfit. KNeighbors and Bayesian Ridge show moderate performance, capturing general trends but lacking the precision of top-performing models. The overall trend reveals that ensemble methods, especially XGBoost and Gradient Boosting, provide the most accurate and reliable predictions, highlighting their capability to handle the variability in student performance data effectively.

The plot underscores the importance of model selection, demonstrating that while some models excel in predictive accuracy, others struggle to capture the underlying patterns. Ensemble methods, in particular, emerge as superior in managing the complexities of educational data, making them valuable tools for accurate student performance prediction.

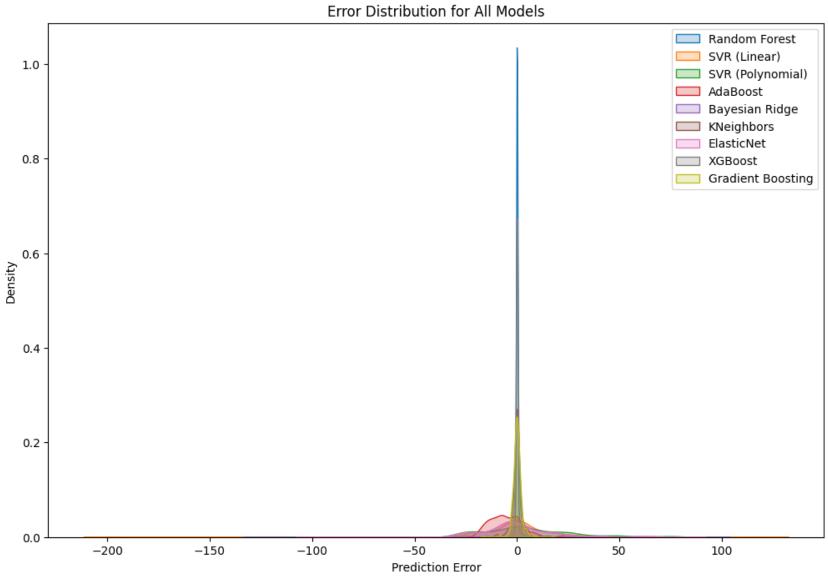

Figure 3: Error Distribution of Predictions for Various Machine Learning Models

Figure 3 displays the distribution of prediction errors for various machine learning models used to predict student performance. The models included are Random Forest, SVR (Linear and Polynomial), AdaBoost, Bayesian Ridge, KNeighbors, ElasticNet, XGBoost, and Gradient Boosting, each represented by different colors. The plot shows that the majority of models have their error distributions tightly centered around zero, indicating that their predictions are generally close to the actual values. Notably, models like XGBoost, Gradient Boosting, and Random Forest exhibit a very high peak at zero error, demonstrating their high accuracy and minimal deviation from the true scores. These models are effective in capturing the underlying patterns in the data, leading to precise predictions with little error. In contrast, the SVR (Polynomial) model displays a wider error distribution, with errors spreading far from zero, indicating poor performance and significant deviation in its predictions. This suggests that the polynomial SVR struggles to fit the data accurately, resulting in higher prediction errors. Similarly, models like AdaBoost and Bayesian Ridge show wider error distributions compared to the top-performing models, indicating moderate accuracy but less reliability in their predictions. Overall, the density plot underscores the superior performance of ensemble methods, particularly XGBoost and Gradient Boosting, which demonstrate the smallest errors and highest accuracy in predicting student performance. The tight clustering of errors around zero for these models highlights their robustness and effectiveness in handling educational data. Conversely, models with wider error distributions, such as SVR (Polynomial), indicate the need for more sophisticated approaches to achieve better predictive accuracy.

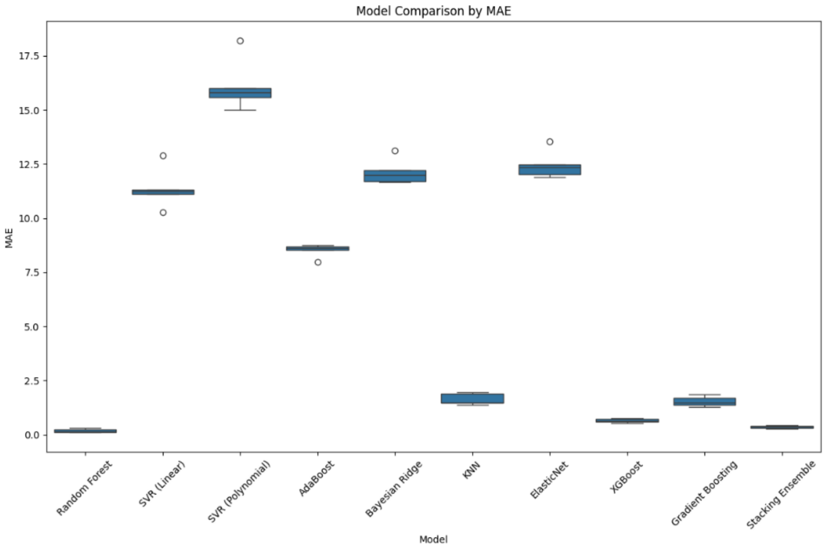

Figure 4: Model Comparison by Mean Absolute Error (MAE)

Figure 4 illustrates the Mean Absolute Error (MAE) for various machine learning models in predicting student performance. XGBoost, Gradient Boosting, and the Stacking Ensemble exhibit the lowest MAE, indicating high predictive accuracy. Conversely, models like SVR (Polynomial) and Bayesian Ridge display higher MAE values, reflecting less accurate predictions. Random Forest and KNeighbors also perform moderately well, with lower MAE compared to simpler models. This plot underscores the superior accuracy of ensemble methods, particularly XGBoost and Gradient Boosting, in minimizing prediction errors and effectively capturing the underlying patterns in the data.

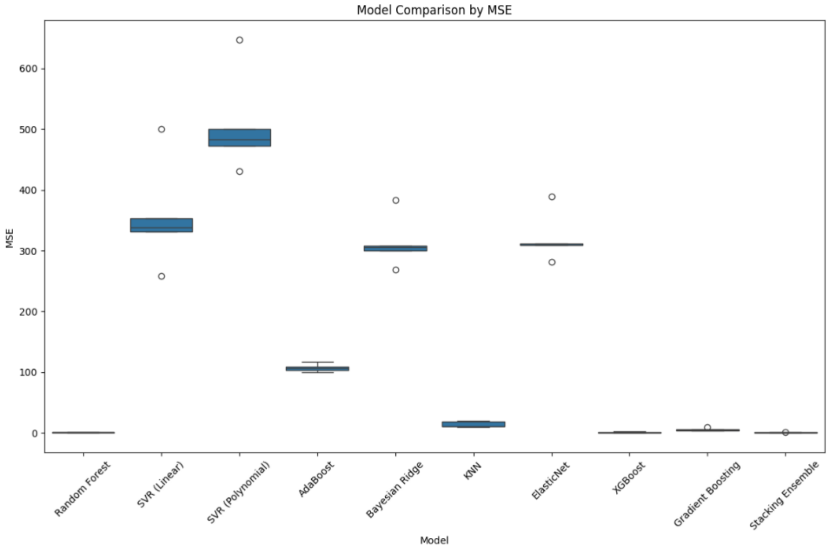

Figure 5: Model Comparison by Mean Squared Error (MSE)

Figure 5 displays the Mean Squared Error (MSE) for various machine learning models predicting student performance. XGBoost, Gradient Boosting, and the Stacking Ensemble exhibit the lowest MSE values, indicating high accuracy with minimal prediction errors. Conversely, models like SVR (Polynomial) and Bayesian Ridge show higher MSE values, reflecting poorer performance. Random Forest and KNeighbors demonstrate moderate performance with relatively lower MSE compared to simpler models. This plot highlights the superior accuracy of ensemble methods, particularly XGBoost and Gradient Boosting, in effectively capturing and predicting student performance with minimal errors.

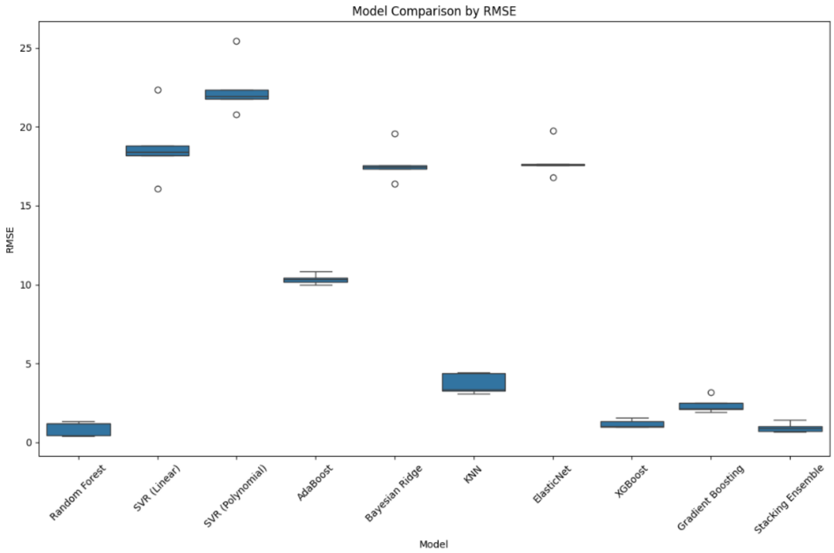

Figure 6: Model Comparison by Root Mean Squared Error (RMSE)

Figure 6 displays the Root Mean Squared Error (RMSE) for various machine learning models predicting student performance. XGBoost, Gradient Boosting, and the Stacking Ensemble show the lowest RMSE values, indicating high accuracy and minimal prediction errors. Conversely, SVR (Polynomial) and Bayesian Ridge have higher RMSE values, reflecting poorer performance. Random Forest and KNeighbors demonstrate moderate performance with relatively lower RMSE compared to simpler models. This plot highlights the superior accuracy of ensemble methods, particularly XGBoost and Gradient Boosting, in effectively capturing and predicting student performance with minimal errors.

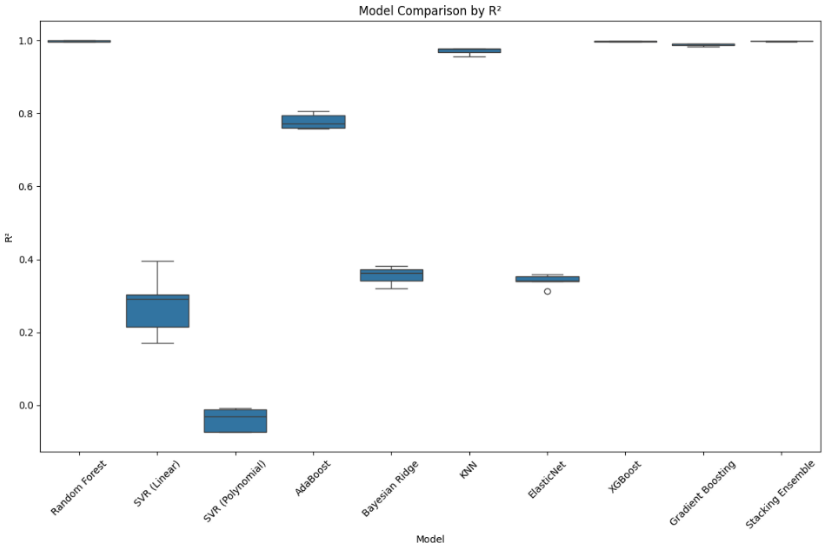

Figure 7: Model Comparison by R-squared (R^2) Value

Figure 7 displays the coefficient of determination (R^2) for various machine learning models predicting student performance. XGBoost, Gradient Boosting, and the Stacking Ensemble exhibit the highest R^2 values, indicating that these models explain the largest proportion of variance in the data. Conversely, SVR (Polynomial) and Bayesian Ridge show lower R^2 values, reflecting poorer performance in capturing the underlying patterns. Random Forest and KNeighbors demonstrate moderate performance with relatively higher R^2 compared to simpler models. This plot highlights the superior predictive power of ensemble methods, particularly XGBoost and Gradient Boosting, in accurately modeling student performance.

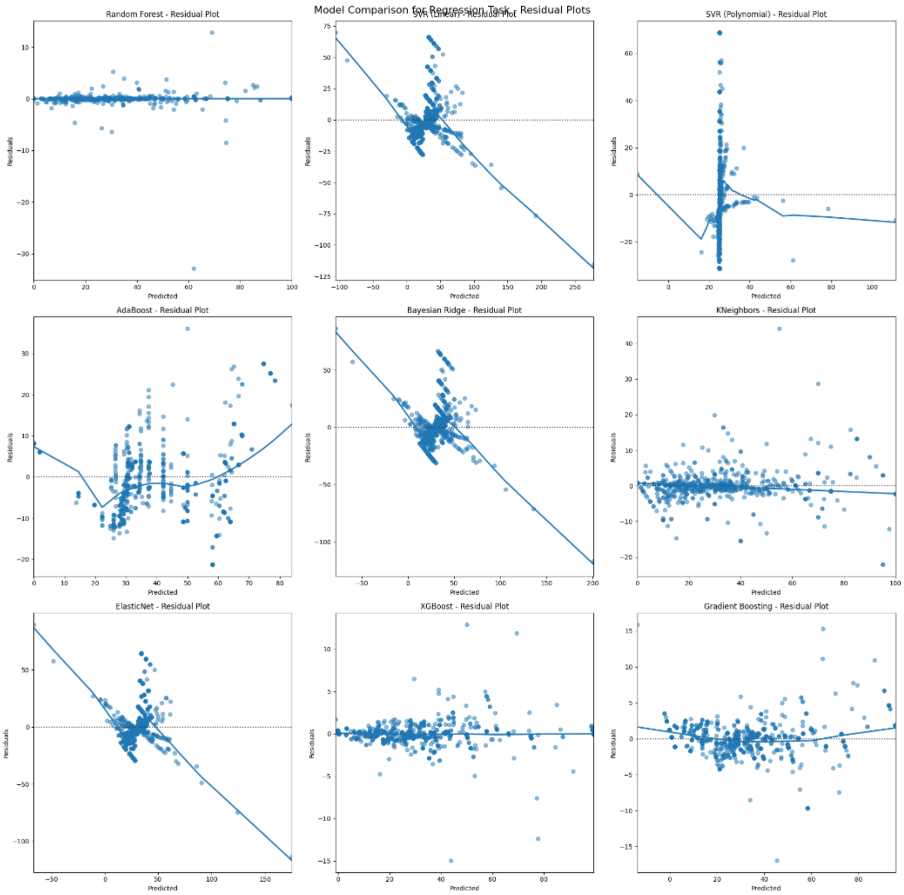

Figure 8: Residual Plots for Various Machine Learning Models

Figure 8 is the residual plots for various machine learning models predicting student performance reveal key insights into model accuracy. XGBoost and Gradient Boosting show residuals tightly clustered around zero, indicating high predictive accuracy. Random Forest and KNeighbors also display minimal residuals, suggesting moderate accuracy. Conversely, SVR (Polynomial) and Bayesian Ridge exhibit large, scattered residuals, reflecting significant prediction errors and poor model fit. AdaBoost and ElasticNet show patterns of systematic errors, indicating potential biases. These plots highlight the superior performance of ensemble methods, particularly XGBoost and Gradient Boosting, in accurately predicting student performance with minimal residual errors.

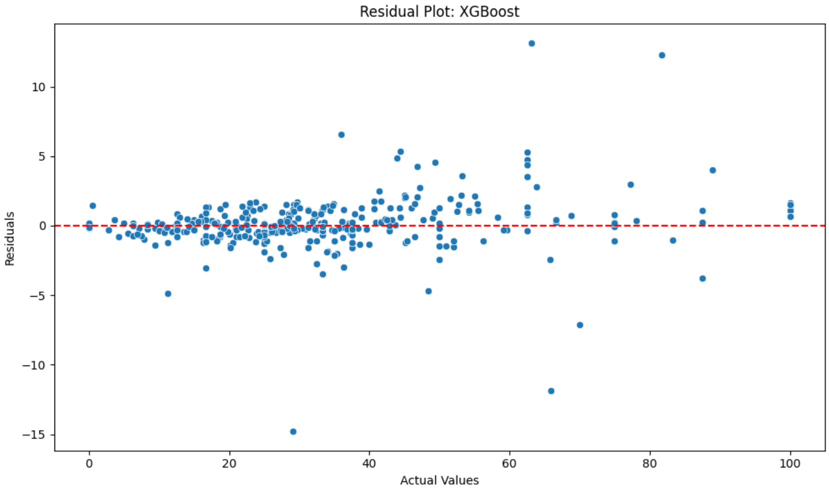

Figure 9: Residual Plot for XGBoost Model

Figure 9 is the residual plot for XGBoost in predicting student performance shows that the residuals are tightly clustered around the zero line, indicating high accuracy. Most points are near the red dashed line, reflecting minimal deviation between the actual and predicted values. The spread of residuals remains consistent across the range of actual values, suggesting that XGBoost maintains its accuracy throughout the dataset. Few outliers are present, which indicates occasional prediction errors but overall reliable performance. This plot highlights XGBoost's effectiveness in providing precise predictions with minimal error, making it a robust model for predicting student performance.

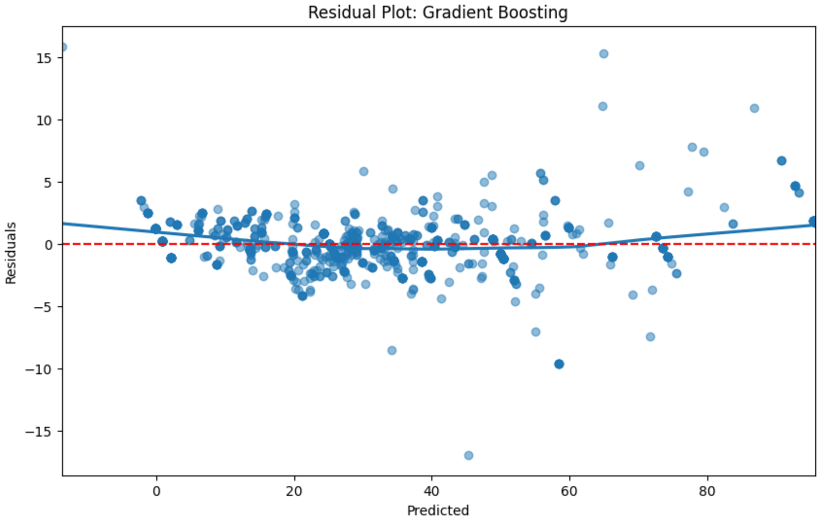

Figure 10: Residual Plot for Gradient Boosting Model

Figure 10 is the residual plot for Gradient Boosting in predicting student performance shows that the residuals are generally close to the zero line, indicating high accuracy. The points are scattered around the red dashed line with minimal deviation, suggesting reliable predictions. However, a slight curve in the residuals indicates a small systematic error, particularly for higher predicted values. Despite this, the model performs well overall, with most residuals within a narrow range, reflecting its effectiveness in capturing the underlying patterns in the data. Gradient Boosting proves to be a robust model, providing accurate predictions with minor inconsistencies.

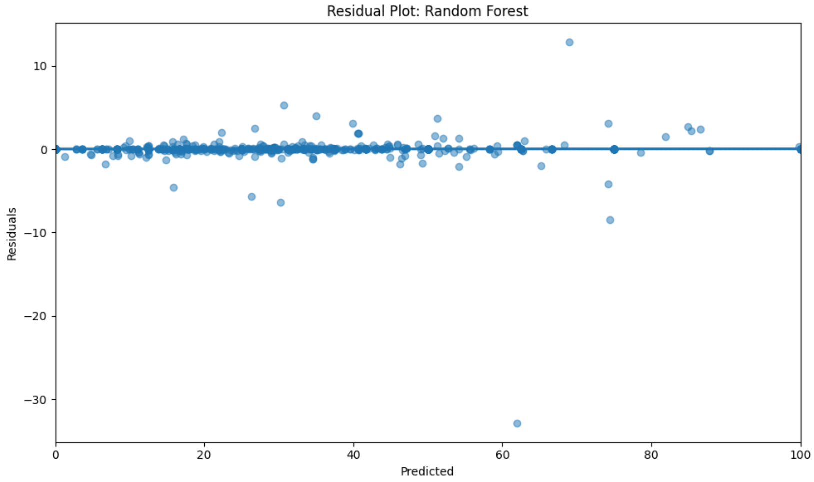

Figure 11: Residual Plot for Random Forest Model

Figure 11 is the residual plot for Random Forest in predicting student performance shows residuals closely clustered around the zero line, indicating high accuracy. Most points are near the red dashed line, reflecting minimal prediction errors. There are a few outliers, but overall, the residuals remain consistently low across the range of predicted values. This indicates that Random Forest effectively captures the underlying patterns in the data, providing reliable predictions with minimal error. The slight spread at the higher values suggests minor deviations, but the overall performance is robust and dependable.

Table 2 shows the confusion matrix for XGBoost, which shows a high overall accuracy of 95% in predicting student performance. The model exhibits perfect precision (1.00) for below-threshold predictions and perfect recall (1.00) for above-threshold predictions. The F1-scores are 0.94 and 0.95 for below and above thresholds, respectively, indicating a balanced performance. The macro and weighted averages for precision, recall, and F1-score are all 0.95, demonstrating the model's robustness and reliability in classification tasks. This high accuracy and consistency across metrics highlights XGBoost's effectiveness in accurately distinguishing student performance levels.

Table 2: Confusion Matrix for XGBoost Model

	precision	recall	f1-score	support
Below Threshold	1.00	0.89	0.94	464
Above Threshold	0.90	1.00	0.95	493
accuracy		0.95	957	
macro avg	0.95	0.94	0.95	957
weighted avg	0.95	0.95	0.95	957

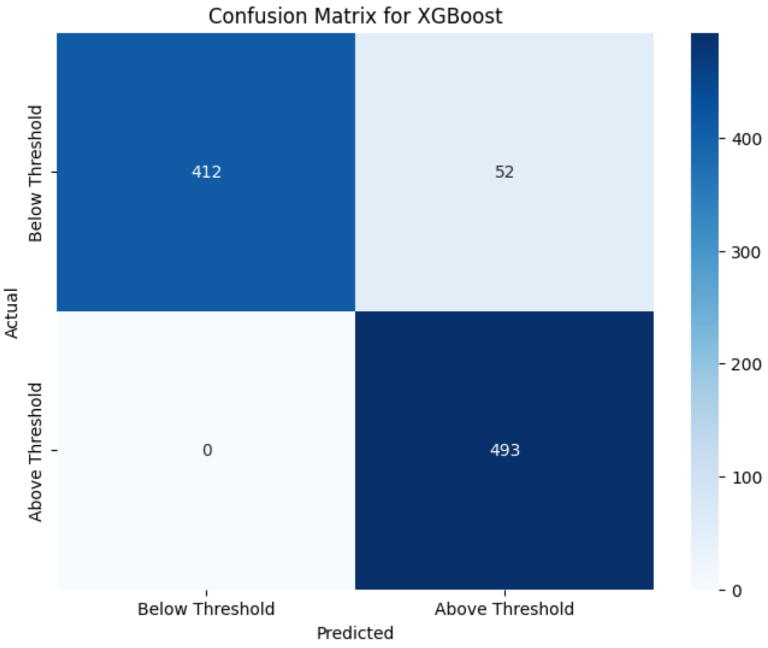

Figure 12: Confusion Matrix for XGBoost Model

Figure 12 displays the confusion matrix for the XGBoost model used in predicting student performance, categorized by a threshold. The matrix shows the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The values indicate that the model correctly predicted 412 instances below the threshold (TN), 493 instances above the threshold (TP), while there were 52 false positives (FP) and 0 false negatives (FN). The color intensity represents the density of the predictions, providing a clear visualization of the model's performance in classification tasks. This figure provides classification performance metrics useful for binary classification or threshold-based analysis in regression tasks. The most important features in the XGBoost model are 'correct', with an importance score of 0.533694, and 'maximum', with an importance score of 0.452255. These features are significantly more predictive than others, with 'correct' being the most critical.

Institutional features have much lower importance scores, with 'institution_SKKU' being the highest at 0.013349. This indicates that the number of correct answers and the maximum possible score are the strongest predictors of the overall score, while the institution has a minor impact. This finding aligns with the expectation that direct performance metrics are the most predictive of student performance, while institutional effects are less significant.

The results demonstrate that ensemble and boosting methods, particularly XGBoost and Stacking Ensemble, are most effective for this regression task. These models capture complex relationships in the data effectively. Conversely, the poor performance of SVR models suggests that the data structure does not suit the assumptions of these algorithms.

The performance gap between the best and worst models highlights the importance of model selection in machine learning tasks and the benefits of ensemble methods in handling complex, non-linear relationships. Future research should investigate the importance of feature in top-performing models, hyperparameter tuning to enhance performance, and exploring data characteristics that favor ensemble methods. These insights are valuable for model selection and optimization in similar regression tasks, especially in complex or non-linear data contexts.

4. Discussion

Our comprehensive evaluation of various regression models for predicting student performance reveals important insights into model performance and suitability for this task. Consistent with findings from prior studies, ensemble methods, particularly tree-based algorithms, demonstrated a clear advantage in capturing the underlying patterns in our dataset [1], [2]. Random Forest emerged as the top performer, exhibiting the highest R^2 score and lowest error metrics (MSE, RMSE, MAE). This superior performance can be attributed to its ensemble nature, which mitigates overfitting by averaging predictions from multiple decision trees, each trained on bootstrap samples with random feature selection. This approach enhances the model's ability to generalize across diverse data points, aligning with Zhao et al.'s findings on the effectiveness of ensemble learning [2]. Similarly, XGBoost and Gradient Boosting showed strong predictive capabilities, closely following Random Forest in performance metrics. The success of these boosting algorithms lies in their sequential tree-building process, where each subsequent tree aims to correct the errors of its predecessors. XGBoost's slightly better performance over traditional Gradient Boosting can be ascribed to its optimized implementation, which includes built-in regularization and advanced tree-pruning strategies [2].

In contrast, linear models such as Bayesian Ridge and ElasticNet demonstrated moderate performance. While these models offer good interpretability and handle multicollinearity well, their relatively lower performance suggests the presence of non-linear relationships in our data that these models couldn't fully capture [3]. The Bayesian Ridge's probabilistic approach and ElasticNet's balanced regularization provided stable, albeit not outstanding, predictions. Support Vector Regression (SVR) with different kernels yielded mixed results. The linear kernel's performance indicates that simple linear relationships are insufficient to model our data accurately. The polynomial kernel's higher error rates suggest potential overfitting, highlighting the challenge of selecting an appropriate kernel and hyperparameters for SVR in complex datasets [4]. The K-Nearest Neighbors (KNN) algorithm showed competitive performance for a non-parametric method but fell short of the ensemble methods. This outcome underscores KNN's limitations in handling high-dimensional data and its sensitivity to the local structure of the dataset [5]. Interestingly, the Stacking Ensemble, which combined predictions from multiple base models, demonstrated robust performance. This approach leverages the strengths of diverse algorithms, potentially capturing different aspects of the underlying data structure [6].

These findings emphasize the importance of model selection in regression tasks. The superior performance of ensemble methods, particularly Random Forest and boosting algorithms, suggests that our dataset contains complex, possibly non-linear relationships that these models are well-suited to capture. However, the computational cost and reduced interpretability of these complex models should be considered in practical applications. The findings from our machine learning model comparison provide valuable insights for educators and policymakers working in smart learning environments. The superior performance of models like Random Forest and XGBoost underscores the potential of leveraging learning analytics data for predicting student performance [1], [2]. This capability empowers educators and administrators to make well-informed, data-driven decisions regarding curriculum design, resource allocation, and targeted intervention strategies. The high accuracy of our predictive models enables early identification of at-risk students. This early warning system allows educators to implement timely interventions, potentially mitigating academic challenges before they escalate [4].

The complex interactions captured by ensemble methods such as Random Forest highlight the potential for creating personalized learning pathways. These models can guide the customization of educational experiences to better meet individual student needs and learning styles. Insights into the factors that most significantly predict student success, as revealed by feature importance analysis in top-performing models, can inform the optimization of resource allocation [5]. Educational institutions can focus on enhancing specific aspects of the learning environment or investing in educational technologies that align with these key predictors.

Policymakers can leverage these findings to develop evidence-based educational policies. For example, if institutional factors are identified as significant predictors, policies aimed at standardizing best practices across institutions could be prioritized. The effectiveness of sophisticated predictive models emphasizes the need for data literacy among educators. This finding supports the implementation of professional development programs focused on data interpretation and the application of predictive analytics in education [12].

While these predictive models offer substantial benefits, they also raise important ethical concerns regarding data privacy, potential biases, and the responsible use of predictive analytics in education. Policymakers must establish guidelines to ensure the ethical application of these technologies in educational settings. The success of our models in predicting short-term outcomes suggests the value of conducting longitudinal studies. Such studies could examine how early performance indicators predict long-term educational and career outcomes, thereby informing long-term educational strategies [10].

The high accuracy of ensemble methods supports the development of advanced adaptive learning systems [25]. These systems could utilize real-time data to adjust content difficulty, pacing, and learning activities, optimizing the learning experience for each student [26]. The robust performance of these models across various data subsets encourages increased data sharing and collaboration between educational institutions. Such collaboration could lead to more comprehensive and generalizable insights into factors affecting student success [11].

In conclusion, the predictive power of our machine learning models offers a valuable tool for enhancing educational outcomes. However, it is essential to balance the use of these technologies with ethical considerations and a holistic understanding of the learning process. Moving forward, the integration of these predictive models into educational practice should be approached thoughtfully, with ongoing evaluation of their impact and effectiveness in improving student outcomes.

5. Conclusions

This study presents a comprehensive evaluation of machine learning models for predicting student performance in smart learning environments, showcasing a range of strengths and limitations. By comparing diverse models such as Random Forest, Support Vector Regression, AdaBoost, Bayesian Ridge, K-Nearest Neighbors, ElasticNet, XGBoost, Gradient Boosting, and Stacking Ensemble, the research provides a thorough understanding of each model's capabilities in the educational context. Utilizing learning analytics data, the study generates valuable insights for data-driven decision-making in curriculum design and resource allocation. One key strength lies in the high accuracy of the models, particularly ensemble methods like Random Forest and XGBoost, which enable early identification of at-risk students and allow for timely interventions. These models also show potential for creating personalized learning pathways, tailoring educational experiences to individual student needs. The robust cross-validation techniques employed ensure reliable performance metrics and assess the models' generalizability, adding credibility to the results. Feature importance analysis in top-performing models offers actionable insights for optimizing the learning environment. However, the study faces limitations related to data quality and availability, which may not be representative of all educational contexts, and the generalizability of the models to other settings remains uncertain. Ethical concerns regarding data privacy and potential biases are acknowledged but not fully addressed, and the interpretability of some complex ensemble methods could challenge educators and policymakers. The focus on short-term predictions without empirical evidence on the long-term impact of interventions and the computational resources required for training and validating multiple complex models also pose limitations. Effective feature engineering is crucial for model performance, which may not be easily replicable without domain expertise. While the study supports the development of adaptive learning systems, implementing such systems requires significant technological investment, which may not be feasible for all institutions. In conclusion, this research provides valuable insights into the application of machine learning for predicting student performance, offering potential benefits for personalized learning and early intervention strategies. However, it underscores the need for careful consideration of data quality, ethical implications, model interpretability, and practical implementation challenges. Future research should explore more sophisticated hybrid models and advance feature engineering to further enhance predictive accuracy and robustness, incorporating diverse data sources and conducting longitudinal studies to examine the long-term impact of early interventions and personalized learning paths, ultimately guiding the development of long-term strategies to support student success.

6. Acknowledgements

This research paper is financially supported by the Program Unit for Human Resources & Institutional Development (PMU-B), Office of National Higher Education Science Research and Innovation Policy Council, Thailand.

References

[1] M. Wu et al., "Using Machine Learning-based Algorithms to Predict Academic Performance - A Systematic Literature Review," in 2024 4th Int. Conf. Innovative Practices in Technology and Management (ICIPTM), 2024, pp. 1-8. DOI: 10.1109/ICIPTM59628.2024.10563566

[2] L. Zhao, J. Ren, L. Zhang, and H. Zhao, "Quantitative Analysis and Prediction of Academic Performance of Students Using Machine Learning," Sustainability, vol. 15, no. 16, p. 12531, 2023. DOI: 10.3390/su151612531

[3] N. Sateesh, P. S. Rao, and D. R. Lakshmi, "Optimized ensemble learning-based student's performance prediction with weighted rough set theory enabled feature mining," Concurrency and Computation: Practice and Experience, vol. 35, 2023. DOI: 10.1002/cpe.7601

[4] D. Çınar and S. Yılmaz Gündüz, "CLASSIFICATION OF STUDENTS' ACADEMIC SUCCESS USING ENSEMBLE LEARNING AND ATTRIBUTE SELECTION," Eskişehir Technical University Journal of Science and Technology A - Applied Sciences and Engineering, 2024. DOI: 10.18038/estubtda.1394885

[5] H. Şevgin, "A comparative study of ensemble methods in the field of education: Bagging and Boosting algorithms," International Journal of Assessment Tools in Education, vol. 10, no. 3, pp. 544-562, 2023. DOI: 10.21449/ijate.1167705

[6] A. Bujang et al., "Imbalanced Classification Methods for Student Grade Prediction: A Systematic Literature Review," IEEE Access, vol. 11, pp. 1970-1989, 2023. DOI: 10.1109/ACCESS.2022.3225404

[7] M. Ye et al., "SA-FEM: Combined Feature Selection and Feature Fusion for Students' Performance Prediction," Sensors, vol. 22, no. 22, 2022. DOI: 10.3390/s22228838

[8] S. Li and B. Yang, "Personalized Education Resource Recommendation Method Based on Deep Learning in Intelligent Educational Robot Environments," Int. J. Inf. Technol. Syst. Approach, vol. 16, pp. 1-15, 2023. DOI: 10.4018/IJITSA.321133

[9] K. Mastrothanasis, K. Zervoudakis, and M. Kladaki, "An application of Computational Intelligence in group formation for digital drama education," Iran Journal of Computer Science, pp. 1-13, 2024. DOI: 10.1007/s42044-024-00186-9

[10] A. López-García, O. Blasco-Blasco, M. Liern-García, and S. Parada-Rico, "Early detection of students' failure using Machine Learning techniques," Operations Research Perspectives, vol. 11, p. 100292, 2023. DOI: 10.1016/j.orp.2023.100292

[11] Y. Alshamaila et al., "An automatic prediction of students' performance to support the university education system: a deep learning approach," Multim. Tools Appl., vol. 83, pp. 46369-46396, 2024.

DOI: 10.1007/s11042-024-18262-4

[12] S. Malik and K. Jothimani, "Enhancing Student Success Prediction with FeatureX: A Fusion Voting Classifier Algorithm with Hybrid Feature Selection," Educ. Inf. Technol., vol. 29, pp. 8741-8791, 2023. DOI: 10.1007/s10639-023-12139-z

[13] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[14] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," Statistics and Computing, vol. 14, no. 3, pp. 199-222, 2004. DOI: 10.1023/B:STCO.0000035301.49549.88

[15] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," Journal of Computer and System Sciences, vol. 55, no. 1, pp. 119-139, 1997. DOI: 10.1006/jcss.1997.1504

[16] D. J. MacKay, "Bayesian interpolation," Neural Computation, vol. 4, no. 3, pp. 415-447, 1992.

[17] E. Fix and J. L. Hodges Jr, "Discriminatory analysis-nonparametric discrimination: consistency properties," California Univ Berkeley, 1951.

[18] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," Journal of the Royal Statistical Society: Series B, vol. 67, no. 2, pp. 301-320, 2005. DOI: 10.1111/j.1467-9868.2005.00503.x

[19] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785-794. DOI: 10.1145/2939672.2939785

[20] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," Annals of Statistics, pp. 1189-1232, 2001. DOI: 10.1214/aos/1013203451

[21] D. H. Wolpert, "Stacked generalization," Neural Networks, vol. 5, no. 2, pp. 241-259, 1992.

[22] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)?– Arguments against avoiding RMSE in the literature," Geoscientific Model Development, vol. 7, no. 3, pp. 1247-1250, 2014. DOI: 10.5194/gmd-7-1247-2014

[23] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," Climate Research, vol. 30, no. 1, pp. 79-82, 2005. DOI: 10.3354/cr030079

[24] A. C. Cameron and F. A. Windmeijer, "An R-squared measure of goodness of fit for some common nonlinear regression models," Journal of Econometrics, vol. 77, no. 2, pp. 329-342, 1997. DOI: 10.1016/S0304-4076(96)01818-0

[25] L. K. Smirani, H. A. Yamani, L. J. Menzli, and J. A. Boulahia, "Using ensemble learning algorithms to predict student failure and enabling customized educational paths," Scientific Programming, vol. 2022, no. 1, p. 3805235, 2022. DOI: 10.1155/2022/3805235

[26] S. B. Keser and S. Aghalarova, "HELA: A novel hybrid ensemble learning algorithm for predicting academic performance of students," Education and Information Technologies, vol. 27, pp. 4521-4552, 2021. DOI: 10.1007/s10639-021-10780-0