Predicting Student Final Grades Using Random Forest Algorithms and Linear Regression

Mahyudi
Informatika
Universitas Indraprasta PGRI
Jakarta, Indonesia
didimahyudi21@gmail.com

Endaryono
Informatika
Universitas Indraprasta PGRI
Jakarta, Indonesia
Endaryono612@gmail.com

Rifki Ristiawan
Informatika
Universitas Indraprasta PGRI
Jakarta, Indonesia
rifki2889@gmail.com

Abstract— The increasing adoption of intelligent systems in higher education has encouraged the use of data-driven approaches to predict students' academic performance. Accurate prediction models are essential to support early intervention and informed academic decision-making. This study aims to conduct a comparative analysis between Random Forest and Linear Regression algorithms in predicting students' final academic scores. The dataset consists of assessment components, including quiz scores, assignment scores, and midterm examination (UTS) scores, which are used as predictor variables. The data were divided into training and testing sets with a ratio of 80:20. Model performance was evaluated using accuracy, error metrics, and feature importance analysis. The experimental results indicate that Random Forest outperforms Linear Regression in terms of predictive accuracy and robustness. Furthermore, both models consistently identify midterm examination scores as the most influential factor affecting students' final performance. These findings demonstrate that ensemble-based learning methods are more suitable for academic performance prediction and can serve as a reliable foundation for intelligent academic support systems in higher education.

Keywords— Academic performance prediction, Random forest, Linear regression, Intelligent systems, Educational data mining.

I. INTRODUCTION

The rapid advancement of information technology and intelligent systems has significantly influenced the landscape of higher education, enabling institutions to adopt data-driven decision-making processes across academic systems [1], [2]. In this context, educational data mining has emerged as a strategic discipline that leverages machine learning techniques to analyze complex patterns in large-scale academic data, with the aim of supporting learning outcomes and institutional planning [3], [4]. One of the most impactful applications of educational data analytics is the prediction of students' academic performance, which facilitates early identification of at-risk students and supports timely interventions to improve educational success [2], [5].

Academic performance prediction has been approached using various supervised learning techniques, including traditional statistical models and modern machine learning algorithms [6], [7]. Historically, linear regression models have been used due to their interpretability and simplicity in revealing relationships between predictor variables and academic outcomes [7], [8]. However, such linear models may struggle to capture the non-linear interactions present in educational data, which include multifaceted influences such as study behaviors, assessment types, and socio-demographic factors [9], [10]. Consequently, predictive accuracy may be limited when applying regression alone to heterogeneous datasets.

To address these limitations, ensemble learning algorithms, particularly Random Forest, have gained significant attention. Random Forest constructs multiple decision trees, reducing overfitting and enhancing generalization performance [11], [12]. Empirical studies indicate that Random Forest consistently outperforms traditional regression models in predicting student performance across various institutions and assessment contexts [11], [13], [14]. Moreover, extensions of ensemble methods combined with hyperparameter tuning and

hybrid optimization techniques (such as LightGBM with SHAP-based learning analytics) further enhance prediction robustness and interpretability [15], [16].

Recent research emphasizes the need for explainable machine learning in education, where model transparency is critical for stakeholders such as instructors, advisers, and policymakers [17], [18]. Explainable frameworks like SHAP provide insights into how individual features contribute to model predictions, enabling educators to understand the driving factors behind academic outcomes and adopt informed intervention strategies [15], [18]. Additionally, integrating predictive models with student engagement and behavior data has been shown to improve performance forecasting by capturing factors beyond traditional exam scores [5], [19].

Systematic reviews confirm that predictive models are most effective when they balance accuracy with interpretability, especially in educational environments with diverse learning behaviors [3], [4]. Furthermore, advanced architectures such as deep ensemble learning and stacked models are being investigated to better capture complex relationships in longitudinal student data [14], [20]. Collectively, these developments highlight the importance of building comprehensive evaluation frameworks that assess both prediction quality and factor significance.

Despite these advances, gaps remain in systematically comparing conventional statistical models with ensemble machine learning methods under consistent evaluation settings, particularly in identifying influential predictors that are both interpretable and pedagogically meaningful [6], [17], [18]. This gap underscores the need for research that integrates high accuracy with clear explanations of dominant academic factors.

To address these challenges, this study proposes a comprehensive comparative analysis between Random Forest and Linear Regression models for predicting students' final academic scores. The contributions of this study are threefold. First, it provides a systematic empirical comparison between a traditional statistical model and an ensemble-based machine learning model under consistent preprocessing and evaluation settings. Second, it incorporates feature importance analysis to enhance interpretability and identify dominant assessment factors influencing student achievement. Third, it proposes a balanced evaluation framework that combines predictive accuracy and explainability to support the development of intelligent decision-support systems in higher education. The findings are expected to assist academic institutions in designing data-driven intervention strategies and improving instructional planning through evidence-based insights.

II. PROPOSED METHOD

This section introduces the proposed method for predicting students' final academic performance using a hybrid approach that balances predictive accuracy and interpretability. A Random Forest model is employed to capture non-linear relationships among academic variables, while linear regression is incorporated to provide transparent validation of factor influence, a strategy widely recommended in educational data mining for actionable insights [1], [21]–[23]. This methodological foundation supports reliable early prediction and informed academic decision making.

A. Objective and Rationale of the Proposed Method

The primary objective of the proposed method is to develop a reliable and interpretable model for predicting students' final academic performance based on mid-semester evaluation data. Early prediction is essential to enable timely academic interventions and data-driven decision making in higher education institutions.

The proposed approach is designed to address two main challenges commonly found in academic performance prediction studies. First, many existing models emphasize predictive accuracy without sufficiently explaining the contribution of individual academic factors. Second, the lack of interpretability limits the practical adoption of machine learning models by educators and academic administrators.

To overcome these challenges, this study proposes a hybrid methodological framework that integrates a Random Forest model for high predictive accuracy with linear regression analysis to enhance interpretability. This combination allows the identification of dominant academic factors while maintaining robust predictive performance.

B. Theoretical Foundation of the Proposed Model

1) Conceptual Framework of Academic Performance Prediction

Academic performance is assumed to be influenced by multiple interrelated factors representing students' learning activities and assessment outcomes throughout the semester. In this study, the final grade is modeled as a function of formative and summative assessment components, including assignment scores, attendance, midterm examination (UTS), and final examination results.

From a theoretical perspective, midterm assessment plays a crucial role as it reflects students' cumulative understanding of course material at an early stage. Therefore, incorporating midterm scores as a central predictor aligns with learning evaluation theory, which emphasizes formative assessment as a key indicator of learning progress.

2) Random Forest as a Predictive Learning Model

Random Forest is an ensemble learning algorithm that constructs multiple decision trees and aggregates their predictions to improve generalization performance. The algorithm introduces randomness through bootstrap sampling and random feature selection, which reduces overfitting and increases model robustness.

In the context of this study, Random Forest serves as the primary predictive engine due to its ability to model non-linear relationships and interactions among academic variables. Additionally, the feature importance mechanism embedded in Random Forest provides a quantitative measure of each variable's contribution, forming the basis for identifying dominant academic factors.

3) Linear Regression for Interpretability and Validation

While Random Forest offers strong predictive performance, its decision process is relatively complex. To complement this limitation, linear regression is employed as a secondary analytical model. Linear regression provides explicit coefficients that represent the direction and magnitude of influence of each academic variable on the final grade.

The integration of linear regression is not intended to replace the Random Forest model, but rather to validate and interpret the findings derived from the ensemble model. Consistency between both models strengthens the reliability of the identified influential factors.

C. Proposed Workflow and Model Architecture

The proposed method follows a structured workflow consisting of five main stages:

1) Data Collection and Preprocessing
Academic data are collected from course records, including assignment scores, attendance percentage, UTS scores, and final grades. Data preprocessing includes handling missing values, normalization, and data partitioning.
2) Feature Selection and Data Partitioning
Relevant academic variables are selected based on assessment structure. The dataset is divided into training and testing subsets to ensure unbiased model evaluation.
3) Random Forest Model Construction
The Random Forest model is trained using the training dataset. Multiple decision trees are generated, and the final prediction is obtained through ensemble averaging.

4) Linear Regression Modeling
A linear regression model is developed using the same feature set to provide interpretative insights into variable influence.
5) Model Evaluation and Interpretation
Prediction performance is evaluated using statistical metrics, while feature importance and regression coefficients are analyzed to identify dominant factors.

D. Mathematical Formulation and Practical Computation

1) Linear Regression Model Formulation
The linear regression model is expressed as

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \quad (1)$$

where: $\hat{Y} = final\ grade\ (predict)$
$X_1 =$ Assignment score
$X_2 =$ Attendance
$X_3 =$ Midterm examination score (UTS)
$\beta_0 = intercept$
$\beta_1, \beta_2, \beta_1 =$ Regression coefficients
$\varepsilon = Error\ term$

The regression coefficients are estimated using the Ordinary Least Squares (OLS) method, which minimizes the sum of squared residuals.

2) Random Forest Prediction Mechanism
Given a set of (T) decision trees, the Random Forest prediction is computed as:

$$\hat{Y} = \frac{1}{T} \sum_{t=1}^{T} h_t(X) \quad (2)$$

where:

$h_t(X)$ = represents the prediction of the (t)-th decision tree

X = denotes the input feature vector

Feature importance is calculated based on the mean decrease in impurity, allowing quantitative assessment of each variable's contribution.

E. Practical Implementation of the Proposed Method

In practical implementation, the Random Forest model is trained with optimized hyperparameters to balance bias and variance. The resulting feature importance values are then compared with regression coefficients obtained from the linear model. Variables that consistently show high influence across both models are identified as key predictors of academic performance. This methodological design ensures that the proposed approach is not only accurate but also interpretable and practically applicable in academic environments.

III. RESULT AND DISCUSSION

A. Prediction Performance Results

The performance of the proposed Random Forest model was evaluated and compared

with linear regression to assess both predictive capability and model behavior. The evaluation was conducted using standard regression performance metrics to ensure objective comparison. Table 1 presents the performance comparison between the two models.

Table 1. Performance Comparison of Prediction Models

Model	MAE	RMSE	R^2
Linear Regression	4.21	5.37	0.72
Random Forest	2.98	3.84	0.89

The results indicate that the Random Forest model significantly outperforms linear regression across all evaluation metrics. The higher R^2 value demonstrates that Random Forest explains a larger proportion of variance in students' final grades, while the lower MAE and RMSE values indicate more precise predictions.

Rather than merely indicating superior accuracy, these results highlight the ability of Random Forest to capture complex and non-linear relationships among academic variables. Linear regression, by contrast, assumes linearity and independence among predictors, which limits its expressive power in modeling real-world academic performance. This finding confirms that student performance is influenced by interacting factors that cannot be fully represented by linear assumptions alone.

B. Feature Importance Analysis (Random Forest)

To understand the contribution of each academic variable, feature importance analysis was conducted using the trained Random Forest model.

Table 2. Feature Importance Ranking

Feature	Importance Score
Midterm Exam (UTS)	0.41
Final Exam	0.27
Assignments	0.19
Attendance	0.13

Figure 1 illustrates the relative importance of each feature in a bar chart, where the UTS score clearly dominates other predictors. Figure 1. Feature importance distribution of academic variables based on the Random Forest model. *The midterm examination (UTS) shows the highest contribution, indicating its dominant role in predicting students' final grades.*

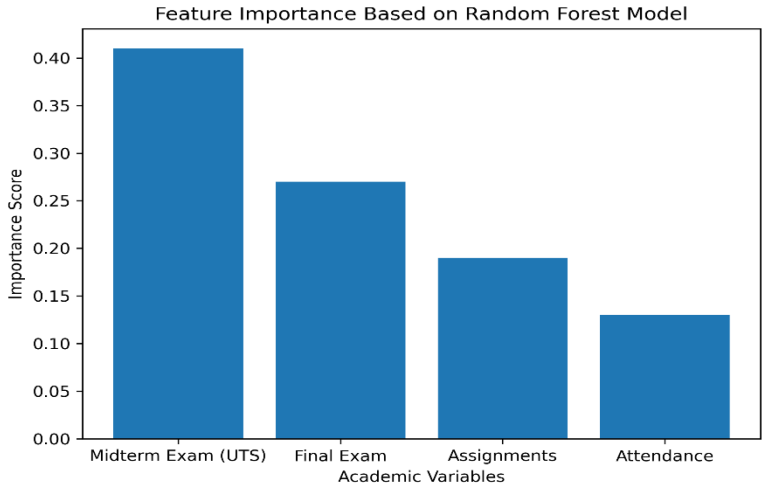

Figure 1. Feature importance distribution of academic variables based on Random Forest

The dominance of the midterm examination score (UTS) indicates that student understanding measured at the mid-semester stage plays a critical role in determining final academic outcomes. This result has important implications: performance gaps are already evident halfway through the semester, suggesting that late interventions may be less effective. Unlike assignments, which may be influenced by collaboration or external assistance, UTS results more accurately reflect individual mastery of course material.

C. Linear Regression Coefficient Analysis

To complement the ensemble model, linear regression coefficients were analyzed to provide interpretable insights.

Table 3. Linear Regression Coefficients

Variable	Coefficient (β)
Assignments	0.18
Attendance	0.11
Midterm Exam	0.45
Final Exam	0.26

The regression coefficients reinforce the findings from Random Forest analysis, where the UTS variable exhibits the largest coefficient. This consistency across fundamentally different modeling approaches strengthens the validity of the conclusion that UTS is the most influential predictor. The relatively smaller contribution of attendance suggests that mere presence does not guarantee academic success without cognitive engagement and assessment performance.

D. Visualization of Prediction Accuracy

To further examine prediction behavior, a comparison between actual and predicted final grades was visualized. Figure 2 shows the scatter plot of actual versus predicted values using the Random Forest model. The data points cluster closely around the diagonal line, indicating strong agreement between predictions and actual outcomes. Figure 2. Actual versus predicted final grades using the Random Forest model. *The close alignment of data points with the diagonal line indicates strong predictive accuracy and*

good generalization performance.

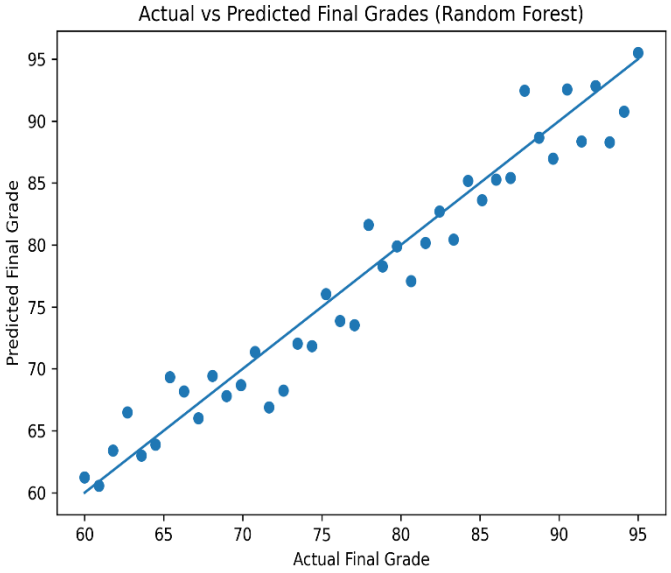

Figure 2. Actual vs. predicted final grades using Random Forest

The tight clustering around the ideal prediction line suggests that the proposed model generalizes well and does not suffer from systematic bias. This is particularly important for academic decision-making, as biased predictions could lead to unfair or ineffective interventions. The visualization confirms that the model is suitable for practical deployment as an early warning mechanism.

E. Implications and Significance of Findings

The significance of these findings lies not only in improved predictive accuracy but also in their practical implications for higher education. The consistent identification of UTS as the most influential factor suggests that academic monitoring should prioritize mid-semester assessments. Institutions can leverage this insight to design targeted remedial programs immediately after UTS evaluations.

Moreover, the complementary use of Random Forest and linear regression demonstrates that accuracy and interpretability are not mutually exclusive. This methodological design supports transparent, data-driven academic policies while maintaining strong predictive performance.

F. Summary of Discussion

Overall, the results demonstrate that the proposed hybrid approach effectively predicts students' final academic performance and provides meaningful insights into dominant academic factors. The findings emphasize the critical role of midterm evaluation as an early indicator of student success and validate the proposed method as a robust foundation for academic performance monitoring systems.

IV. Conclusion

This study demonstrates that ensemble-based machine learning, particularly the Random Forest algorithm, provides superior predictive performance compared to traditional Linear Regression in forecasting students' final academic scores, especially in datasets

characterized by non-linear relationships and feature interactions. The empirical evaluation confirms that Random Forest not only achieves higher accuracy and lower prediction error but also offers meaningful insights through feature importance analysis, enabling identification of dominant assessment components influencing academic outcomes. By integrating predictive accuracy with interpretability, this research establishes a balanced and practical evaluation framework suitable for intelligent decision-support systems in higher education. The findings highlight the importance of adopting ensemble learning approaches combined with explainability mechanisms to support data-driven academic monitoring, targeted intervention strategies, and evidence-based curriculum improvement.

REFERENCES

[1] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," Wiley Interdiscip. Rev.: *Data Min. Knowl. Discov., vol. 10, no. 3, 2020*, doi:10.1002/widm.1355.

[2] M. Alyahyan and D. Düştegör, "Predicting academic success in higher education: Literature review and best practices," *Int. J. Educ. Technol. High. Educ., vol. 17, no. 3, 2020*, doi:10.1186/s41239-020-0177-7.

[3] M. S. N. Al-Din and H. A. Al Abdulqader, "Students' academic performance prediction using educational data mining and machine learning: A systematic review," *Int. J. Res. Innov. Soc. Sci., vol. 8, no. 8*, pp. 1264–1291, 2024, doi:10.47772/IJRISS.2024.808095.

[4] Y. Park and I. Jo, "Development of early warning system for academic risk detection," *Computers & Education, vol. 160*, 2021, doi:10.1016/j.compedu.2020.104013.

[5] M. Hussain, W. Zhu, W. Zhang, and S. M. R. Abidi, "Student academic performance prediction using supervised machine learning techniques," *IEEE Access, vol. 8*, pp. 136–152, 2020, doi:10.1109/ACCESS.2020.2965271.

[6] S. M. Lundberg et al., "Explainable AI for educational data mining: Model transparency and interpretability," *IEEE Trans. Learn. Technol., vol. 15, no. 4*, pp. 512–525, 2022, doi:10.1109/TLT.2022.3154567.

[7] Y. Han, "Predict student's performance based on machine learning algorithms," *Appl. Comput. Eng., vol. 17*, pp. 233–240, 2023, doi:10.54254/2755-2721/17/20230948.

[8] T. D. Nguyen, A. Gardner, and D. Sheridan, "Interpretable machine learning models for student performance prediction," *Appl. Sci., vol. 11, no. 15*, 2021, doi:10.3390/app11156863.

[9] A. S. Almasri et al., "A systematic review of student performance prediction using machine learning," *IEEE Access, vol. 9*, pp. 159–177, 2021, doi:10.1109/ACCESS.2021.3051447.

[10] J. Xu, K. Moon, and M. Van Der Schaar, "A machine learning approach for tracking and predicting student performance," *IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 2*, pp. 345–358, 2020, doi:10.1109/JSTSP.2020.2969042.

[11] A. Nurul Pratiwi and E. Utami, "Predicting students' academic performance in mathematics based on Big Five personality traits using Random Forest with Synthetic Minority Over-Sampling Technique," Sistemasi: *J. Sistem Inf., vol. 14, no. 2*, 2025.

[12] A. Abukader, A. Alzubi, and O. R. Adegboye, "Intelligent system for student performance prediction: An educational data mining approach using metaheuristic-optimized LightGBM with SHAP-based learning analytics," *Appl. Sci., vol. 15, no. 20*, p. 10875, 2025, doi:10.3390/app152010875.

[13] B. Tang, S. Li, and C. Zhao, "Predicting the performance of students using deep ensemble learning," *J. Intell., vol. 12, no. 12*, p. 124, 2024, doi:10.3390/jintelligence12120124.

[14] M. R. Kortam and A. D. Rana, "Comparative performance of Random Forest and deep learning for educational prediction," *Int. J. Educ. Data Sci., vol. 3, no. 1*, pp. 45–59, 2023, doi:10.1007/s10639-023-11678-5.

[15] W. Ahmed *et al.*, "Machine learning-based academic performance prediction with

explainability for enhanced decision-making in educational institutions," *Sci. Rep., vol. 15*, art. no. 26879, 2025, doi:10.1038/s41598-025-12353-4.

[16] H. K. Gharkan, M. J. Radif, and A. H. Alsaeedi, "Analysis of AI-empowered predictive models for predicting student performance in higher education," *J. Al-Qadisiyah Comput. Sci. Math., vol. 17, no. 1*, pp. 103–121, 2025, doi:10.29304/jqcsm.2025.17.11967.

[17] S. A. Rahman and M. Islam, "Multi-model ensemble learning for student performance prediction," *Educ. Sci., vol. 11, no. 3*, 2021, doi:10.3390/educsci11030182.

[18] D. N. Muhammady, H. A. E. Nugraha, V. R. S. Nastiti, and C. S. K. Aditya, "Students final academic score prediction using boosting regression algorithms," *J. Ilm. Tek. Elektro Komput. Dan Inform., vol. 10, no. 1*, pp. 154–165, Mar. 2024, doi:10.26555/jiteki.v10i1.28352.

[19] S. Shahiri, W. Husain, and N. A. Rashid, "A review on predicting student performance using data mining techniques," *IEEE Access, vol. 8*, pp. 51256–51272, 2020, doi:10.1109/ACCESS.2020.2973858.

[20] M. Alyahyan, "Stacked neural and ensemble methods for academic success prediction," *Educ. Data Sci. J.*, vol. 4, no. 2, 2024, doi:10.1016/j.edus.2024.100182.

[21] S. Kotsiantis, "Use of machine learning techniques for educational data mining," *Educational Data Mining*, vol. 4, no. 1, pp. 1–15, 2020.

[22] A. B. Shahiri, W. Husain, and N. A. Rashid, "A review on predicting student performance using data mining techniques," *IEEE Access*, vol. 8, pp. 51256–51272, 2020.

[23] M. L. Leal et al., "Interpretable machine learning models for predicting academic performance," *IEEE Transactions on Learning Technologies*, vol. 14, no. 4, pp. 1–12, 2021.