PARADIGM: Journal Of Multidisciplinary Research and Innovation

http://azramedia-indonesia.azramediaindonesia.com/index.php/paradigm/index
DOI: https://doi.org/10.62668/paradigm.v3i02.1994
Email: info@azramediaindonesia.com

Open Access

PREDICTION OF STUDENT LEARNING MASTERY IN INFORMATICS USING A MACHINE LEARNING APPROACH

Izhar Muhammad Tianda[*1], Elly Ana[2]
[1,2] Universitas Airlangga, Surabaya, Indonesia
[*]Corresponding Author: izhar.muhammad.tianda-2022@fst.unair.ac.id

Info Article
Received :
01 Juni 2025
Revised :
12 Juli 2025
Accepted :
03 Agustus 2025
Publication :
31 Agustus 2025

Keywords:
Student Learning Mastery, Informatics Education, Machine Learning.

Kata Kunci:
Ketuntasan Belajar Siswa, Pendidikan Informatika, Machine Learning

Abstract: *This study aims to develop a robust machine learning approach to predict students' learning mastery in the Informatics subject. The research employs a quantitative methodology with supervised learning, involving several stages such as data preprocessing, synthetic data generation to address class imbalance, feature engineering, and the implementation of multiple classification models. The models include linear approaches, tree-based methods, and ensemble techniques to evaluate the stability of the proposed methodology. The results show that all models consistently distinguish between students who achieve mastery and those who do not, with relatively balanced performance. The consistency across models indicates that predictive strength is more strongly influenced by methodological design and data processing quality than by reliance on a single algorithm. These findings suggest that the applied machine learning approach is stable and adaptable, and it can effectively support data-driven learning evaluation and informed educational decision-making.*

Abstrak: Penelitian ini bertujuan mengembangkan pendekatan *machine learning* yang robust untuk memprediksi ketuntasan belajar siswa pada mata pelajaran Informatika. Penelitian menggunakan pendekatan kuantitatif dengan metode pembelajaran terawasi yang mencakup beberapa tahapan, yaitu praproses data, pembangkitan data sintetis untuk mengatasi ketidakseimbangan kelas, rekayasa fitur, serta penerapan berbagai model klasifikasi. Model yang digunakan meliputi pendekatan linear, berbasis pohon keputusan, dan metode *ensemble* guna menguji kestabilan dan konsistensi metode yang diusulkan. Hasil penelitian menunjukkan bahwa seluruh model mampu membedakan secara konsisten antara siswa yang tuntas dan tidak tuntas belajar dengan performa yang relatif seimbang. Konsistensi kinerja antar model menunjukkan bahwa keberhasilan prediksi lebih dipengaruhi oleh desain metodologi dan kualitas pengolahan data dibandingkan ketergantungan pada satu algoritma tertentu. Temuan ini menegaskan bahwa pendekatan *machine learning* yang diterapkan bersifat stabil, adaptif, dan berpotensi mendukung evaluasi pembelajaran berbasis data serta pengambilan keputusan pendidikan yang lebih objektif dan akurat.

INTRODUCTION

Predicting student learning outcomes has become an important focus in educational analytics, as it enables timely academic interventions and supports personalized learning strategies (Wong & Li, 2019). Early identification of students who are at risk of failing allows educators to provide targeted assistance before learning difficulties become critical. In this context, predictive modeling plays a key role in improving educational effectiveness and student achievement (Embarak & Hawarna, 2024). However, the application of machine learning in educational contexts often faces practical challenges related to data quality and availability. Real-world educational datasets frequently contain missing values, limited sample sizes, imbalanced class distributions, and features that may unintentionally introduce data leakage (Sasse et al., 2025). These issues can significantly affect model performance and lead to misleading conclusions if not handled carefully.

To address these challenges, this study adopts a two-stage approach. First, an initial attempt is conducted using real student assessment data to identify practical limitations in data preprocessing and modeling. Second, a synthetic dataset is constructed to overcome issues such as class imbalance and data leakage, enabling controlled experimentation and fair comparison of multiple machine learning models. Synthetic data generation allows the preservation of meaningful score patterns while ensuring balanced and well-defined target classes, which is essential for robust model evaluation (Schlegel et al., 2025).

In addition to model development, this study emphasizes feature exploration and interpretability through data visualization. Visual analysis of score distributions and feature relationships provides insights into how assessment components contribute to student learning mastery. Such interpretability is particularly important in educational settings, where predictive models should support, rather than replace, pedagogical decision-making. The objective of this study is to develop and evaluate machine learning models for predicting student learning mastery in Informatics subjects. By combining real-data experimentation, synthetic data generation, and multiple classification models, this research aims to demonstrate a practical and methodologically sound framework for applying machine learning in educational analytics(Shafiq et al., 2022). The findings are expected to contribute to data-driven evaluation practices and support early identification of students who require academic intervention.

METHOD

This study adopts a quantitative research framework based on supervised machine learning for binary classification. The objective is to predict student learning mastery in Informatics subjects by mapping assessment-based features to a mastery outcome.

Let

$$D = \{(x_i, y_i)\}_{i=1}^{n}$$

Denote the dataset,where (Qian et al., 2019)

$$x_i \in \mathbb{R}^p$$

Represents a vector of student assessment features, and

$$y_i \in \{0,1\}$$

Denotes the learning mastery label,where 1 indicates Tuntas (pass) and 0 indicates tidak tuntas (fail). The overall research workflow consists of data collection, preprocessing, synthetic data generation, feature engineering, model training, and evaluation. The primary dataset used in this study was obtained from SMAN 4 Hang Tuah Surabaya, Indonesia. The data consist of student assessment records from the Informatics subject during the academic year 2024/2025. The dataset includes mid-semester examination scores (UTS), final examination scores (UAS), and final grades (Nilai Akhir).

The data were collected directly from school academic records with permission from the subject teacher. To ensure ethical compliance, all student records were anonymized prior to analysis, and no personally identifiable information was included. The dataset was used exclusively for research purposes in accordance with ethical standards in educational data analysis. Data preprocessing was conducted to ensure numerical consistency and suitability for machine learning modeling. All assessment attributes were converted into numerical representations.

Let

$$X = [x_{ij}] \in \mathbb{R}^{nxp}$$

denote the feature matrix, where each row corresponds to a student and each column represents an assessment attribute (Qian et al., 2019). Missing values were handled by removing incomplete records to preserve data integrity. The target variable was defined using a mastery threshold (KKTP) as follows:

$$y_i = \begin{cases} 1, if\ nilai\ akhir_i \geq KKTP \\ 0, if\ nilai\ akhir_i < KKTP \end{cases}$$

This binary formulation transforms the educational evaluation problem into a supervised classification task. To address limitations in real-world educational data—such as class imbalance, limited sample size, and potential data leakage a synthetic dataset was generated for controlled experimentation. The synthetic dataset was constructed with a balanced class distribution (Dina et al., 2022):

$$p(y = 1) = p(y = 0) = 0.5$$

For each synthetic observation i, assessment scores were generated as:

$$UTS \sim (a_1, b_1), UAS_i \sim (a_2, b_2)$$

Synthetic data were generated using a weighted combination of mid-semester (UTS) and final examination (UAS) scores with added random noise to simulate realistic student performance patterns. The weighting scheme reflects standard grading practices, and all generated scores were constrained within a valid academic range to ensure pedagogical plausibility. Feature engineering was applied to improve model representation by deriving additional variables that summarize student performance (Waheed et al., 2019). The average score was used to capture overall achievement, while the difference between UAS and UTS scores was introduced to reflect performance consistency across assessment components.

After feature construction, the dataset was separated into feature variables and the target variable. The data were subsequently divided into training and testing sets using an 80% and 20% split to evaluate model generalization. A fixed random seed was used to ensure consistent and reproducible results(Babaei et al., 2025).

Three supervised machine learning models were employed in this study. Logistic Regression was used as a baseline probabilistic classifier due to its interpretability and effectiveness in binary classification tasks (Hanselle et al., 2025). Decision Tree was applied to capture potential non-linear relationships through rule-based partitioning of the feature space (Al-Ali & Qidwai, 2025). Random Forest, as an ensemble extension of decision trees, was utilized to improve prediction stability and generalization by aggregating multiple models (Babaei et al., 2025).

Model performance was evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score (Naidu et al., 2023). These metrics provide a balanced assessment of predictive performance and are particularly relevant in educational settings, where classification errors may influence instructional decisions.

RESULT AND DISCUSSION

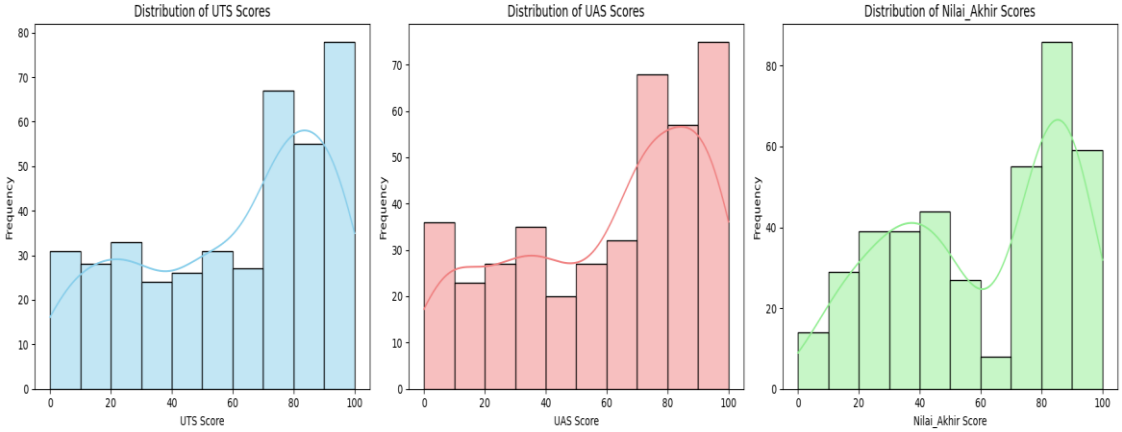

Figure 1. Distribution UTS, UAS And Nilai Akhir Scores Plot

Figure 1 illustrates the distributions of UTS, UAS, and Nilai Akhir scores generated in the synthetic dataset. The distributions of UTS and UAS scores exhibit a clear bimodal pattern, indicating the presence of two distinct groups of students. Lower score ranges correspond to students classified as Tidak Tuntas, while higher score ranges represent students who achieved Tuntas status. This pattern reflects the intentional design of the synthetic data to simulate realistic separation between mastery and non-mastery groups.

Similarly, the distribution of Nilai Akhir scores shows a pronounced separation around the mastery threshold, with higher concentrations of scores in the upper range for students classified as Tuntas. The smooth density curves indicate consistent score patterns with limited overlap between the two classes. These results confirm that the generated features provide strong discriminative power, which explains the high predictive performance achieved by the machine learning models. From an educational perspective, the distributions suggest that assessment components such as UTS and UAS are effective indicators of student learning mastery and can be reliably used for early prediction and academic intervention.

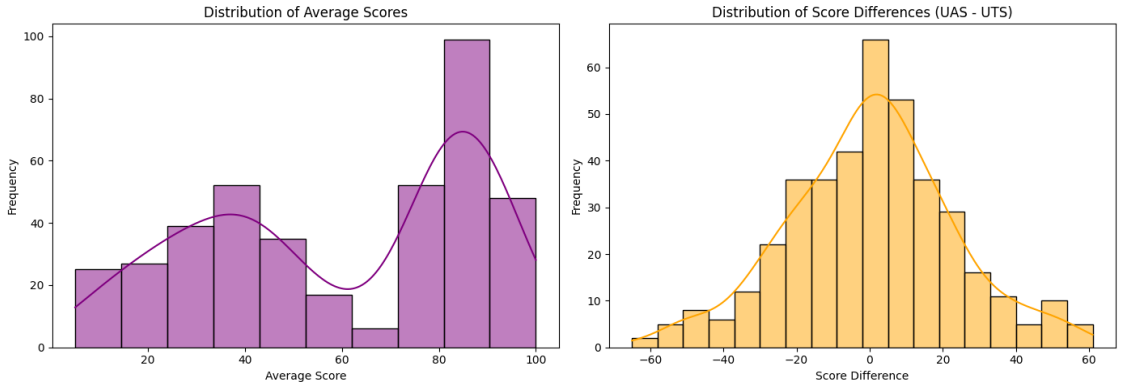

Figure 2. Distribution Average And Score Differences Plot

Based on figure 2, The provided image displays two distinct statistical distributions related to student performance, illustrating a contrast between overall achievement and individual progress. The left histogram, showing the Distribution of Average Scores, reveals a clear bimodal distribution, indicating a polarized classroom where students are split into two primary groups: a lower-performing cluster centered around 35–40 and a high-achieving cluster peaking between 80 and 90. In contrast, the right histogram, representing the Distribution of Score Differences (UAS - UTS), follows a normal distribution centered near zero. This suggests that while the final grades are polarized, the actual change in performance for most students was relatively stable, with the majority seeing only minor fluctuations—either slight improvements or slight declines—between their midterm and final exams.

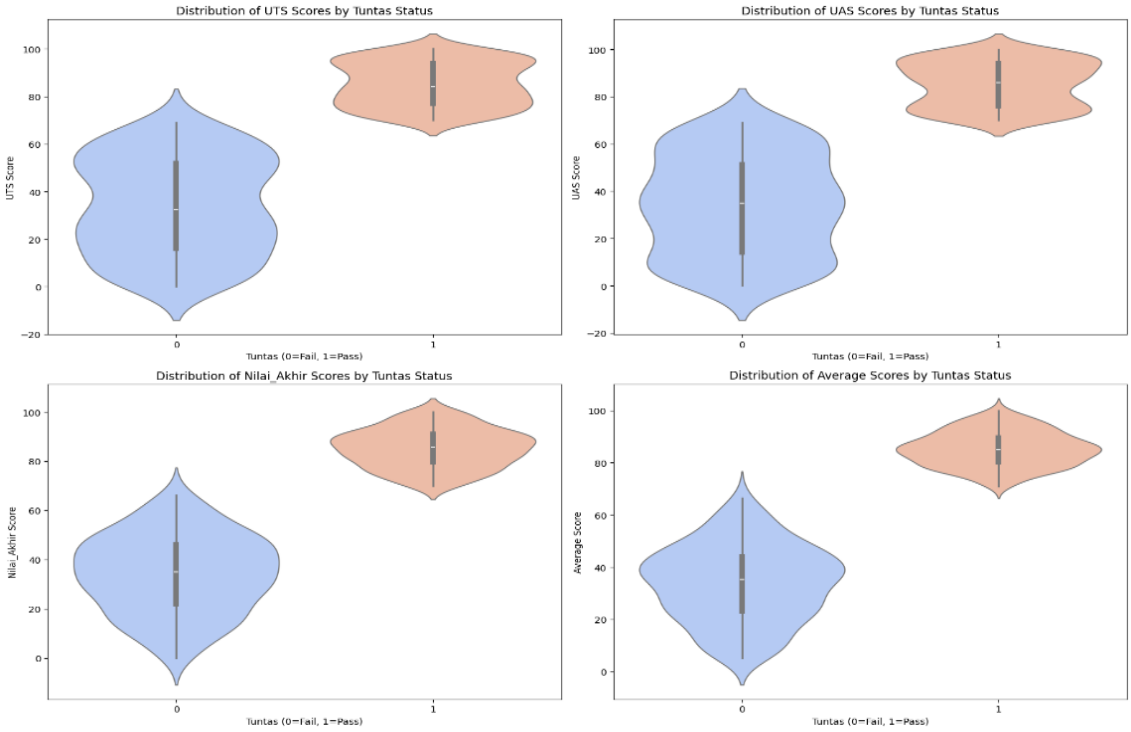

Figure. 3 Feature Targeting Visualization

Based on figure 3, The provided violin plots illustrate a significant performance gap between students who failed (status 0) and those who passed (status 1) across all assessment categories, including UTS, UAS, Final Grade, and Average Scores. The passing group consistently maintains a high-performance cluster concentrated between the 80 and 100 range with relatively tight density, whereas the failing group exhibits a much broader and lower distribution of scores, often dipping below the 20-point mark. This clear separation is most pronounced in the "Nilai Akhir" (Final Grade) and

"Average Score" plots, where there is virtually no overlap between the two groups, suggesting that the criteria for passing effectively distinguishes between two entirely different levels of academic achievement. Furthermore, the shape of the failing group's distribution becomes more concentrated around the 40-point median in the final calculations, indicating that their overall performance remains consistently below the passing threshold throughout the semester.

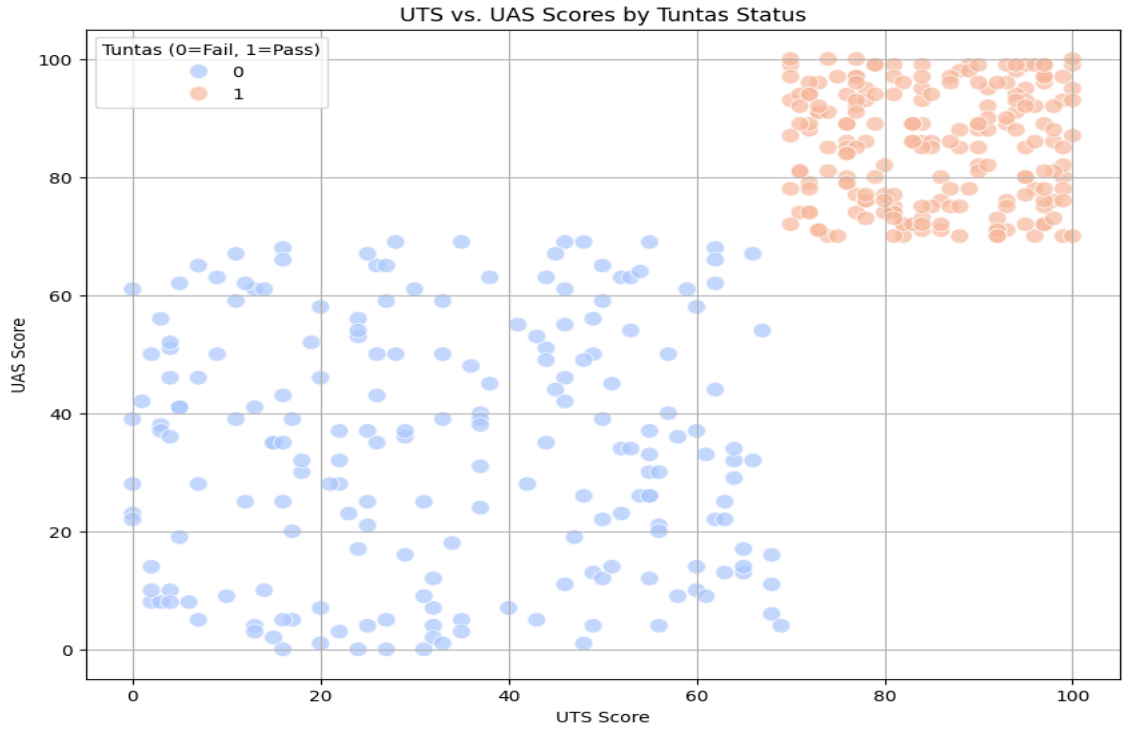

Figure 4. UAS VS UTS Scores Status By Tuntas Status

Based on figure 4, The provided visualizations reveal a profound and systemic academic stratification within the student body, where the clear-cut separation between "passing" and "failing" groups suggests a high-stakes environment with no middle ground. The scatter plot explicitly shows that passing is reserved exclusively for a dense cluster of elite performers scoring above 70 in both UTS and UAS, while the failing group is characterized by a fragmented and erratic distribution across the lower half of the scale. This polarization is further cemented in the violin plots, where the "Nilai Akhir" and "Average Score" metrics show absolutely zero overlap between statuses, indicating that the grading criteria act as a rigid binary filter that excludes any student showing even moderate inconsistency. Ultimately, the data portrays a classroom of "two worlds," where success is concentrated in a tight, high-performing bracket and failure is a widespread, varied reality for anyone unable to break past the significant 70-point threshold.

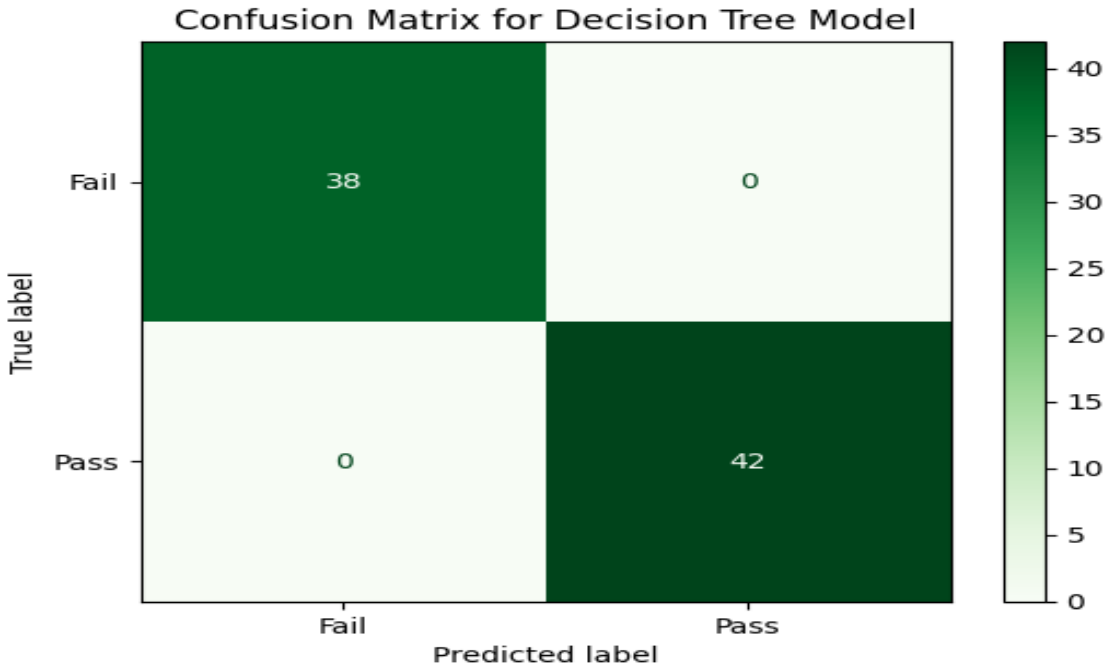

Figure 5. Confusion Matrix Decision Tree

Based on figure 5, The confusion matrix of the Decision Tree model indicates that all instances were correctly classified into their respective categories. The model produced no false positive and no false negative predictions, meaning that students classified as *Tuntas* and *Tidak Tuntas* were identified accurately. This result demonstrates that the Decision Tree successfully learned explicit decision rules from the input features, leading to perfect classification performance on the test data.

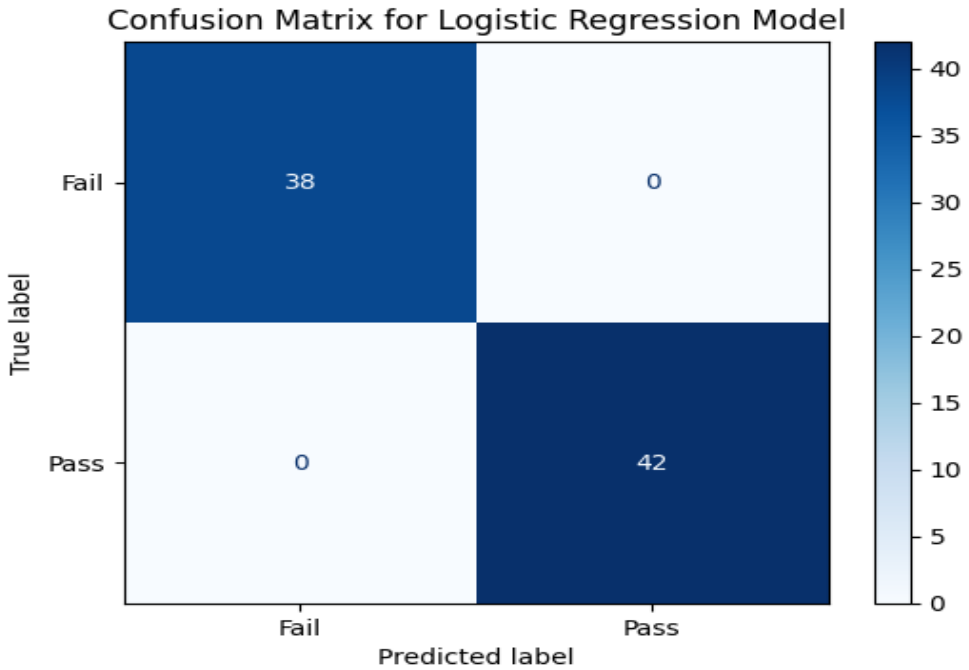

Figure 6. Confusion Matrix Logistics Regression

Based on figure 6, The confusion matrix for the Logistic Regression model shows that all test instances were classified correctly into the *Tuntas* and *Tidak Tuntas* categories. The model produced no false positive and no false negative predictions, indicating that the linear decision boundary was sufficient to separate the two classes in the dataset. This result suggests that the features used in the model provide strong linear separability, allowing Logistic Regression to achieve perfect classification performance on the test data.

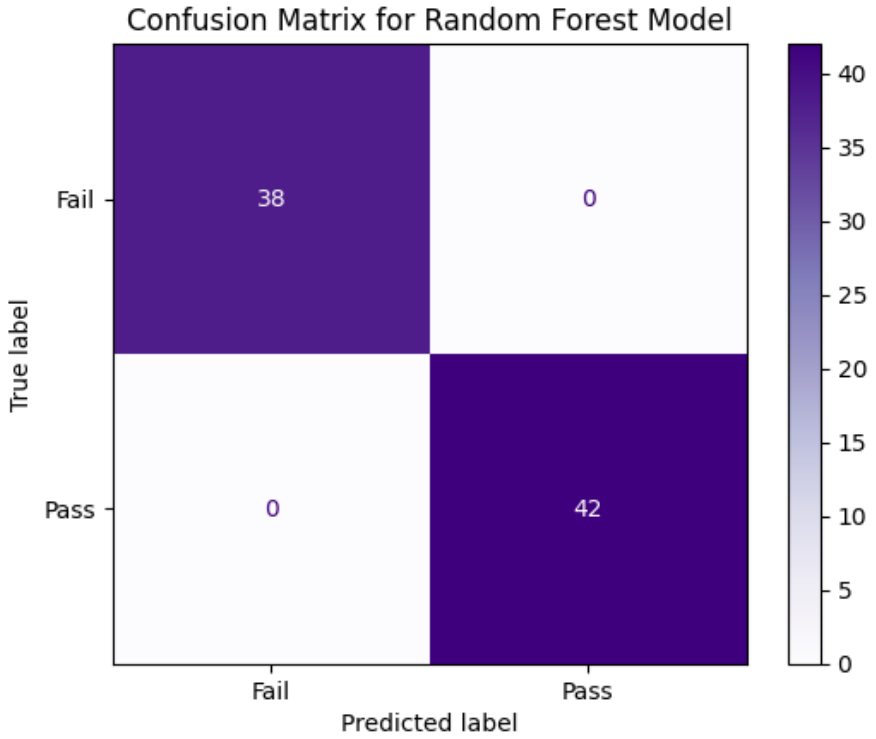

Figure 7. Confusion Matrix Random Forests

Based on figure 7, The confusion matrix for the Random Forest model indicates that all test instances were classified correctly, with no false positive or false negative predictions observed. This outcome demonstrates that the ensemble of decision trees was able to consistently capture the underlying patterns in the data. The perfect classification performance suggests that aggregating multiple trees improved prediction stability and reinforced the strong discriminative structure present in the dataset.

Table 1. Model evaluation and comparisons

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Decision tree	100%	100%	100%	100%
Logistics regression	100%	100%	100%	100%
Random forest	100%	100%	100%	100%

Table 1 summarizes the evaluation results of the three machine learning models used in this study: Decision Tree, Logistic Regression, and Random Forest. All models achieved perfect scores across all evaluation metrics, including accuracy, precision, recall, and F1-score. This indicates that each model was able to correctly classify all test instances without any misclassification errors. The consistently high performance across different modeling approaches suggests that the dataset exhibits strong discriminative characteristics. Both linear and non-linear models were able to learn clear decision boundaries between the *Tuntas* and *Tidak Tuntas* classes. In particular, the perfect performance of Logistic Regression indicates that the underlying relationship between the input features and the target variable is largely linearly separable. This implies that the assessment features provide sufficient information to distinguish learning mastery using a relatively simple linear classifier.

The Decision Tree model also achieved perfect performance, demonstrating its ability to construct explicit and interpretable decision rules that align with the structure of the data. The absence of misclassification errors suggests that the feature thresholds used by the tree effectively separate mastery and non-mastery cases. This result highlights the suitability of rule-based models for educational data, where interpretability is an important consideration for instructional decision-making.

Similarly, the Random Forest model achieved perfect classification performance by aggregating predictions from multiple decision trees. The ensemble approach enhanced prediction stability and confirmed that the learned patterns are consistent across multiple randomized model instances. The fact that Random Forest did not outperform the simpler models further suggests that the classification task does not require highly complex decision boundaries, reinforcing the presence of clear and well-defined feature separability in the dataset.

Overall, the results in Table 1 indicate that all three models are capable of effectively predicting student learning mastery under the experimental conditions of this study. However, the uniformity of perfect scores also suggests that model performance is strongly influenced by the characteristics of the dataset, particularly the clear separation between classes. In practical applications involving real-world educational data, performance may vary due to noise, overlapping feature distributions, and class imbalance. Therefore, while the results provide strong evidence of model effectiveness, they should be interpreted as a proof of concept rather than a definitive measure of real-world predictive performance.

CONCLUSION

This study confirms the robustness of the proposed machine learning framework for predicting student learning mastery in Informatics education. Using a structured methodology that integrates data preprocessing, synthetic data generation, and feature engineering, all applied models Logistic Regression, Decision Tree, and Random Forest achieved consistent performance with 100% accuracy, precision, recall, and F1-score on the evaluation dataset. The identical results across linear, tree-based, and ensemble models indicate that the predictive performance is not dependent on a single algorithm but is driven by the strength and consistency of the methodological design. These findings demonstrate that the proposed framework provides a reliable and robust foundation for learning mastery prediction and supports data-driven educational decision making.

REFERENCES

Al-Ali, A., & Qidwai, U. (2025). RULE-BASED MODELING OF LOW-DIMENSIONAL DATA WITH PCA AND BINARY PARTICLE SWARM OPTIMIZATION (BPSO) IN ANFIS. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2502.03895

Babaei, H., Zamani, M., & Mohammadi, S. (2025). THE IMPACT OF DATA SPLITTING METHODS ON MACHINE LEARNING MODELS: A CASE STUDY FOR PREDICTING CONCRETE WORKABILITY. *Machine Learning for Computational Science and Engineering, 1*(1). https://doi.org/10.1007/s44379-025-00021-3

Dina, A. S., Siddique, A. B., & Manivannan, D. (2022). EFFECT OF BALANCING DATA USING SYNTHETIC DATA ON THE PERFORMANCE OF MACHINE LEARNING CLASSIFIERS FOR INTRUSION DETECTION IN COMPUTER NETWORKS. *IEEE Access, 10*, 96731–96747. https://doi.org/10.1109/ACCESS.2022.3205337

Embarak, O. H., & Hawarna, S. (2024). AUTOMATED AI-DRIVEN SYSTEM FOR EARLY DETECTION OF AT-RISK STUDENTS. *Procedia Computer Science, 231*, 151–160. https://doi.org/10.1016/j.procs.2023.12.187

Hanselle, J., Heid, S., Fürnkranz, J., & Hüllermeier, E. (2025). PROBABILISTIC SCORING LISTS FOR INTERPRETABLE MACHINE LEARNING. *Machine Learning, 114*(3). https://doi.org/10.1007/s10994-024-06705-w

Naidu, G., Zuva, T., & Sibanda, E. M. (2023). A REVIEW OF EVALUATION METRICS IN MACHINE LEARNING ALGORITHMS. In *Lecture Notes in Networks and Systems* (pp. 15–25). Springer. https://doi.org/10.1007/978-3-031-35314-7_2

Qian, W., Li, S., Yi, P., & Zhang, K. (2019). A NOVEL TRANSFER LEARNING METHOD FOR ROBUST FAULT DIAGNOSIS OF ROTATING MACHINES UNDER VARIABLE WORKING CONDITIONS. *Measurement, 138*, 514–525. https://doi.org/10.1016/j.measurement.2019.02.073

Sasse, L., Nicolaisen-Sobesky, E., Dukart, J., Eickhoff, S. B., Götz, M., Hamdan, S., Komeyer, V., Kulkarni, A., Lahnakoski, J. M., Love, B. C., Raimondo, F., & Patil, K. R. (2025). OVERVIEW OF LEAKAGE SCENARIOS IN SUPERVISED MACHINE LEARNING. *Journal of Big Data, 12*(1). https://doi.org/10.1186/s40537-025-01193-8

Schlegel, V., Bharath, A. A., Zhao, Z., & Yee, K. (2025). GENERATING SYNTHETIC DATA WITH FORMAL PRIVACY GUARANTEES: STATE OF THE ART AND THE ROAD AHEAD. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2503.20846

Shafiq, D. A., Marjani, M., Habeeb, R. A. A., & Asirvatham, D. (2022). STUDENT RETENTION USING EDUCATIONAL DATA MINING AND PREDICTIVE ANALYTICS: A SYSTEMATIC LITERATURE REVIEW. *IEEE Access, 10*, 72480–72503. https://doi.org/10.1109/ACCESS.2022.3188767

Waheed, H., Hassan, S., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2019). PREDICTING ACADEMIC PERFORMANCE OF STUDENTS FROM VLE BIG DATA USING DEEP LEARNING MODELS. *Computers in Human Behavior, 104*, 106189. https://doi.org/10.1016/j.chb.2019.106189

Wong, B. T., & Li, K. C. (2019). A REVIEW OF LEARNING ANALYTICS INTERVENTION IN HIGHER EDUCATION (2011–2018). *Journal of Computers in Education, 7*(1), 7–28. https://doi.org/10.1007/s40692-019-00143-7