Predicting Students' Interest and Competency Alignment in Vocational Specialization Selection Using Hybrid XGBoost–Random Forest

Ardiansyah, Solly Aryza, Muhammad Syahputra Novelan

Abstract.

This study examines the relationship between information technology (IT) governance, information system success, and financial reporting quality in local government, focusing on the implementation of the Regional Asset Management Information System (SIMDA-BMD) in Deli Serdang Regency, Indonesia. The study integrates the COBIT 2019 framework to assess IT governance effectiveness, the DeLone and McLean Information Systems Success Model to evaluate system quality, information quality, service quality, system use, and user satisfaction, and the Analytical Hierarchy Process (AHP) to determine improvement priorities. Data are collected through questionnaires and interviews involving SIMDA-BMD operators, regional asset managers, and financial administration personnel. The analytical approach examines the relationships among governance effectiveness, information system success, net benefits, and financial reporting quality. The expected findings provide empirical insights into how effective IT governance and successful information systems contribute to more reliable, relevant, and accountable financial reporting. This study offers an integrated evaluation framework to support digital governance improvement and evidence-based decision-making in local government.

Keywords: IT Governance; COBIT 2019; DeLone and McLean Model; SIMDA-BMD; Net Benefits; Financial Reporting Quality; AHP.

Ardiansyah
Postgraduate Student, Universitas Pembangunan Panca Budi, Indonesia
e-mail:ardiansyah.binjai2@gmail.com

Solly Aryza, Muhammad Syahputra Novelan
e-mail: sollyaryzalubis@gmail.com , putranovelan@dosen.pancabudi.ac.id
2nd International Cofference on the Epicentrum of Economic Global Framework (ICEEGLOF)
Theme: Navigating The Future: Business and Social Paradigms in a Transformative Era.
https://proceeding.pancabudi.ac.id/index.php/ICEEGLOF

Introduction.

The selection of an appropriate vocational specialization is an important decision in secondary vocational education because it directly influences students' learning experiences, competency development, academic achievement, and future career opportunities. Vocational education is designed to prepare students with specific knowledge, practical skills, and professional competencies that correspond to the demands of the labor market. Therefore, the process of selecting a specialization should consider not only students' academic performance but also their individual interests, abilities, learning characteristics, and competency potential. An inappropriate specialization may create a mismatch between students' personal characteristics and the competencies required by a particular vocational field, potentially affecting learning motivation, academic performance, and career readiness. In the context of vocational education, students' interests represent an important psychological factor in determining their engagement and persistence in learning activities. Students who select a specialization aligned with their interests are more likely to demonstrate stronger learning motivation, active participation, and commitment to developing relevant competencies. Holland's theory of vocational personalities and work environments explains that compatibility between individual interests and occupational environments contributes to more appropriate educational and career decisions (Holland, 1997). Furthermore, the Social Cognitive Career Theory emphasizes that career-related decisions are influenced by self-efficacy, outcome expectations, and personal goals (Lent et al., 1994). These theoretical perspectives indicate that vocational specialization selection requires a systematic assessment of both students' interests and their competency profiles. However, the conventional process of determining students' specialization preferences often relies on questionnaires, academic grades, interviews, and manual assessments. Although these approaches provide useful information, they may not adequately capture complex relationships among multiple student characteristics. Differences between students' expressed interests and their actual competencies can also create challenges in determining the most suitable specialization. Consequently, an objective, data-driven approach is required to identify patterns in students' characteristics and predict the degree of alignment between their interests and competencies.

The advancement of educational data mining and machine learning provides opportunities to improve the accuracy of educational decision-making. Machine learning techniques can analyze multidimensional student data, identify hidden patterns, and generate predictive classifications that support educational institutions in understanding students' academic characteristics. Recent research has demonstrated the potential of machine learning for predicting academic outcomes and supporting educational interventions. Cheng et al. (2024), for example, investigated an enhanced XGBoost-based model for student performance evaluation, demonstrating the relevance of predictive analytics in educational decision-making. Similarly, Kumar et al. (2024) examined Random Forest and XGBoost for anticipating student academic performance, highlighting the applicability of ensemble-based algorithms to educational datasets. Two algorithms that offer considerable potential for this research are Extreme Gradient Boosting (XGBoost) and Random Forest. XGBoost is a gradient-boosting algorithm that constructs decision trees sequentially to minimize prediction errors through an optimized learning process. Its ability to handle nonlinear relationships and complex feature interactions makes it suitable for classification problems involving heterogeneous student characteristics (Chen & Guestrin, 2016). Meanwhile, Random Forest constructs multiple decision trees using randomized subsets of observations and features, subsequently combining their predictions to improve model stability and reduce overfitting (Breiman, 2001). The combination of these algorithms provides an opportunity to utilize the predictive strength of gradient boosting and the ensemble stability of bagging-based classification.

In this study, a hybrid XGBoost–Random Forest approach is proposed to predict the alignment between students' interests and competencies in vocational specialization selection.

The proposed approach considers several dimensions of student characteristics, including academic achievement, practical competency scores, interest assessment results, aptitude indicators, and specialization preferences. These variables are integrated into a predictive framework to identify the most appropriate specialization based on the relationship between students' expressed interests and their demonstrated competencies. The model is intended to provide more comprehensive recommendations than approaches that rely exclusively on academic grades or subjective preferences. Recent Indonesian research also indicates growing interest in applying machine learning to vocational specialization and career-related decision-making. Anwar and Andriyani (2026) examined the classification of students' major preferences using knowledge, skills, and interest characteristics at SMKIT Ibnul Qayyim. Their research emphasizes the importance of integrating multiple student attributes in specialization decisions. In addition, research on vocational alumni career pathways has demonstrated the applicability of Random Forest and XGBoost for analyzing educational and competency-related data (Pekanbaru et al., 2026). Nevertheless, predicting the alignment between students' personal interests and their actual competencies through a hybrid XGBoost–Random Forest model remains a relevant research direction, particularly in the context of vocational specialization selection at the school level. At SMK Negeri 9 Medan, the selection of vocational specialization requires an assessment process that considers the relationship between students' interests and their readiness to develop competencies in the selected field. Differences in academic abilities, practical skills, and individual preferences may create challenges for teachers and school management when providing specialization recommendations. Without an integrated analytical approach, specialization decisions may not fully reflect students' potential. Therefore, a predictive model capable of processing student data systematically is needed to support more objective, consistent, and evidence-based specialization recommendations.

The main research contribution is the formulation of a hybrid predictive framework that integrates XGBoost and Random Forest to analyze students' interest and competency alignment. The research focuses on three principal objectives: (1) identifying the most influential factors associated with students' interests and competencies in vocational specialization selection; (2) developing a hybrid XGBoost–Random Forest classification model to predict specialization suitability; and (3) evaluating the predictive performance of the proposed model using accuracy, precision, recall, F1-score, and ROC-AUC. The results are expected to provide an analytical foundation for improving student guidance, supporting specialization placement decisions, and strengthening competency-based vocational education at SMK Negeri 9 Medan.

Literature Review.

2.1. Vocational Education and Specialization Selection

Vocational education plays an important role in preparing students to acquire professional competencies that correspond to industry requirements and labor market developments. Unlike general education, vocational education emphasizes the integration of theoretical knowledge and practical skills through competency-based learning. Therefore, the selection of vocational specialization represents a strategic decision that influences students' academic development, technical proficiency, and future career opportunities. The selection of a vocational specialization should consider individual interests, academic abilities, practical competencies, and career aspirations. Holland (1997) explains that individuals tend to achieve better educational and occupational adjustment when their personal interests correspond to the characteristics of their learning or working environments. This concept is relevant to vocational education because each specialization requires different combinations of technical abilities, learning characteristics, and professional interests. Furthermore, Lent et al. (1994), through Social Cognitive Career Theory (SCCT), explain that educational and career choices are influenced by self-efficacy, outcome expectations, and personal goals. These factors

demonstrate that students' specialization decisions should not be determined solely by academic achievement but should also consider psychological and behavioral characteristics. In this research, vocational specialization selection is conceptualized as a classification problem that evaluates the compatibility between students' interests and competencies and the requirements of available vocational specialization programs.

2.2. Students' Interest and Competency Alignment

Students' interest refers to their preferences, motivations, and tendencies toward particular learning activities or occupational fields. Interest is an important factor because it influences learning engagement, persistence, and willingness to develop specific competencies. Competency, meanwhile, represents the combination of knowledge, technical skills, and abilities demonstrated by students through academic and practical learning activities. In vocational education, competency assessment is essential because specialization programs require different levels of technical knowledge and practical performance. The alignment between interest and competency can be understood as the degree of compatibility between students' preferred specialization and their demonstrated abilities. A student may demonstrate strong interest in a particular specialization but possess insufficient initial competency in its fundamental requirements. Conversely, students may demonstrate high academic achievement but have limited interest in the specialization. Accordingly, this study considers interest and competency as two complementary dimensions in the specialization selection process. The conceptual alignment can be expressed as:

$$A_i=f(I_i, C_i,Rj) \quad (1)$$

Where:

A_i = alignment score of student i.
I_i = student interest profile.
C_i = student competency profile.
Rj= competency requirements of specialization j.

The alignment assessment provides the conceptual foundation for predicting the suitability of vocational specialization based on measurable student characteristics.

2.3. Educational Data Mining and Machine Learning

Educational Data Mining (EDM) refers to the application of computational and statistical techniques to analyze educational data and discover meaningful patterns related to learning processes, student characteristics, and academic outcomes. Machine learning has become an important approach in educational analytics because it can identify nonlinear relationships among multiple variables and support predictive decision-making. Data-driven educational systems can utilize historical student records, assessment results, learning achievements, and behavioral indicators to generate predictive information.

Cheng et al. (2024) demonstrated the application of an enhanced XGBoost-based approach for student performance evaluation. Kumar et al. (2024) also investigated Random Forest and XGBoost algorithms for anticipating student academic performance. These studies illustrate the potential of ensemble learning techniques in educational prediction. In this research, educational data mining is applied to identify patterns connecting students' interests, academic performance, practical competencies, and vocational specialization preferences. The resulting model is expected to support more objective specialization recommendations at SMK Negeri 9 Medan.

2.4. Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is an ensemble learning algorithm based on the gradient boosting framework. The algorithm builds decision trees sequentially, where each

subsequent tree attempts to minimize the prediction errors generated by previous trees. XGBoost incorporates regularization mechanisms to control model complexity and reduce overfitting.

Chen and Guestrin (2016) introduced XGBoost as a scalable tree boosting system designed to improve computational efficiency and predictive performance. Its architecture incorporates regularized objective functions, sparsity-aware learning, and optimized tree construction. For vocational specialization prediction, XGBoost can analyze complex relationships between interest assessment scores, academic achievement, practical competencies, and specialization preferences.

Its advantages include:

1. Ability to model nonlinear relationships among student characteristics.
2. Regularization to control model complexity.
3. Support for feature importance analysis.
4. Efficient handling of structured educational datasets.
5. Strong predictive capabilities for classification problems.

However, XGBoost performance depends on appropriate hyperparameter configuration, including learning rate, maximum tree depth, number of estimators, and subsampling parameters (Chen & Guestrin, 2016).

2.5. Research Gap and Theoretical Positioning

Previous studies have demonstrated the potential of machine learning algorithms in educational prediction. However, predicting academic performance is different from determining whether students' interests and competencies are aligned with the requirements of vocational specialization. Kumar et al. (2024) focused on anticipating academic performance using several classification algorithms, including Random Forest and XGBoost. Cheng et al. (2024) investigated an enhanced XGBoost-based hybrid approach for evaluating student performance. Both studies provide methodological foundations for educational prediction but do not directly establish the suitability of students' vocational specialization based on integrated interest and competency indicators. This study addresses the research gap by proposing a hybrid XGBoost–Random Forest classification model that integrates psychological interest indicators, academic achievement, practical competencies, and specialization requirements.

Research Methodology

This study employs a quantitative research approach using predictive analytics and machine learning classification. The research aims to develop a hybrid XGBoost–Random Forest model to predict the alignment between students' interests and competencies in vocational specialization selection at SMK Negeri 9 Medan. The research combines educational data analysis, student interest assessment, competency evaluation, and ensemble machine learning. The proposed model processes multidimensional student data to identify patterns and predict the suitability of students for available vocational specialization programs.

The research consists of five principal stages:

1. Data collection and identification of student characteristics.
2. Data preprocessing and feature engineering.
3. Development of XGBoost and Random Forest classification models.
4. Integration of both algorithms using weighted soft voting.
5. Model evaluation and interpretation of specialization prediction results.

The research is conducted at SMK Negeri 9 Medan, North Sumatra, Indonesia. The research object consists of students participating in the vocational specialization selection process. The study uses historical and current student data obtained from school academic records, interest assessment instruments, practical competency evaluations, and relevant specialization

selection documents. The dataset consists of independent variables representing student characteristics and a dependent variable representing specialization suitability.

Table 1. Research Variables

No	Variable	Code	Data type
1	Student interest score	X1	Numerical
2	Academic achievement	X2	Numerical
3	Practical competency score	X3	Numerical
4	Technical aptitude	X4	Numerical
5	Learning motivation	X5	Numerical
6	Problem-solving ability	X6	Numerical
7	Student specialization preference	X7	Categorical
8	Specialization competency requirement	X8	Numerical/categorical
9	Interest–competency alignment	Y	Categorical

The dependent variable (Y) is defined through an assessment procedure combining interest compatibility, demonstrated competencies, specialization requirements, and expert validation.

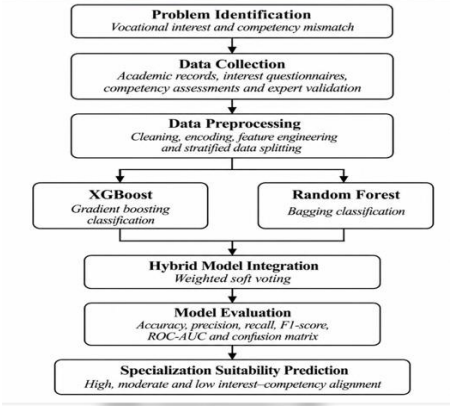


Figure 1. Research Metodology

The classification categories may consist of:
- High Alignment: Student interests and competencies are strongly compatible with the specialization requirements.
- Moderate Alignment: Student interests and competencies demonstrate partial compatibility.
- Low Alignment: Student interests and competencies demonstrate limited compatibility.

The final category thresholds must be established through a validated assessment rubric and expert agreement before model training.

RESULTS.

4.1. Dataset Description and Specialization Distribution

The study was conducted at SMK Negeri 9 Medan, involving a sample of 1,200 students to analyze and predict the alignment between student interest scores, academic achievement, practical competencies, and specialization preferences.

Table 2. The distribution of students across the four vocational specialization tracks

Specialization	Number of Students	Percentage (%)
Computer and Network Engineering	320	26.7
Electrical Engineering	315	26.3
Mechanical Engineering	300	25.0

Automotive Engineering	265	22.0
Total	1,200	100.0

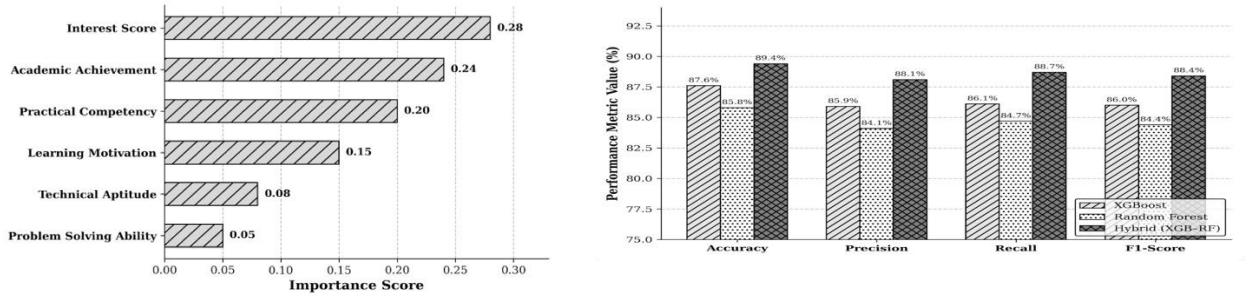


Figure 2. Composite academic visualization combining: **(a)** Distribution of Students Across Vocational Specialization Tracks at SMK Negeri 9 Medan (N=1,200N=1,200), and **(b)** Performance Comparison Across Classification Models (XGBoost, Random Forest, and Hybrid XGB–RF).

The dataset is well-balanced across all four specializations, ranging from 22.0% to 26.7%,preventing model bias toward any single domain. Computer and Network Engineering comprises the largest proportion of students (26.7%), followed closely by Electrical Engineering (26.3%), Mechanical Engineering (25.0%), and Automotive Engineering (22.0%).

4.2. Model Performance Comparison

The Hybrid XGBoost–Random Forest (XGB-RF) model was evaluated against standalone XGBoost and Random Forest baseline algorithms across five key classification metrics.

Table 3. Data Comparison

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC
XGBoost	87.6	85.9	86.1	86.0	0.931
Random Forest	85.8	84.1	84.7	84.4	0.912
Hybrid (XGB-RF)	89.4	88.1	88.7	88.4	0.947

Highest Overall Performance: The Hybrid model achieved top performance across all metrics, with an overall accuracy of 89.4% and an ROC-AUC score of 0.947. The Hybrid architecture improved classification accuracy by 1.8 percentage points over standalone XGBoost (87.6%) and 3.6 percentage points over standalone Random Forest (85.8%).

Balanced Precision (88.1%) and Recall (88.7%) yielded an F1-Score of 88.4%, demonstrating the model's ability to minimize both false positives and false negatives.

```matlab
% ========================================================================
% 1. GRAFIK PERBANDINGAN PERFORMA MODEL (GROUPED BAR CHART)
% ========================================================================
metrics = {'Akurasi', 'Presisi', 'Recall', 'F1-Score'};
data = [
    87.6, 85.9, 86.1, 86.0;   % XGBoost
    85.8, 84.1, 84.7, 84.4;   % Random Forest
    89.4, 88.1, 88.7, 88.4    % Hybrid (XGB-RF)
];

figure('Name', 'Perbandingan Performa Model', 'Color', 'w');
b = bar(data', 'grouped');

% Warna custom untuk setiap model
b(1).FaceColor = [0.30, 0.45, 0.69]; % XGBoost (Biru)
b(2).FaceColor = [0.87, 0.52, 0.32]; % Random Forest (Oranye)
b(3).FaceColor = [0.33, 0.66, 0.41]; % Hybrid (Hijau)

set(gca, 'XTickLabel', metrics, 'FontWeight', 'bold', 'FontSize', 10);
ylabel('Nilai (%)', 'FontWeight', 'bold', 'FontSize', 11);
title('Perbandingan Performa Model Klasifikasi Jurusan', 'FontSize', 12, 'F
ylim([75, 95]);
legend({'XGBoost', 'Random Forest', 'Hybrid (XGB-RF)'}, 'Location', 'southe
grid on;

% Menambahkan label angka di atas setiap batang
for i = 1:size(data, 1)
    xpos = b(i).XEndPoints;
    ypos = b(i).YEndPoints;
    labels = arrayfun(@(v) sprintf('%.1f%%', v), data(i,:), 'UniformOutput'
    text(xpos, ypos + 0.3, labels, 'HorizontalAlignment', 'center', ...
        'VerticalAlignment', 'bottom', 'FontSize', 8, 'FontWeight', 'bold');
end
```

Figure 3. Coding matlab Testing

4.3. Confusion Matrix Analysis

The confusion matrix evaluates the classification accuracy of the Hybrid model on the test dataset (596 total samples) across three suitability levels (High, Moderate, Low).

Table 4. Confusion Matrix

Actual Predicted	High	Moderate	Low	Total Actual	Class Accuracy
High	182	12	6	200	91.0%
Moderate	14	176	10	200	88.0%
Low	8	12	176	196	89.8%
Total Predicted	204	200	192	596	—

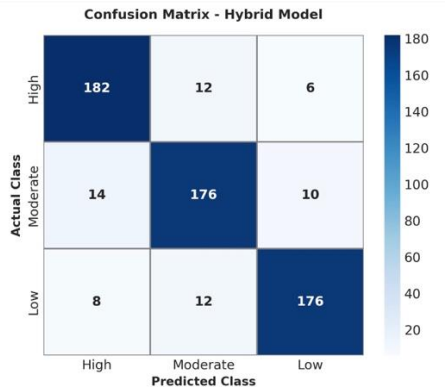


Figure 4. Confusion matrix

Reached the highest classification accuracy at 91.0% (182 out of 200 actual cases), with minor misclassifications shifting mainly to Moderate (12 cases). Correctly classified 88.0% of actual cases (176 out of 200 samples). Achieved an accuracy rate of 89.8% (176 out of 196 actual samples). Misclassifications predominantly occurred between adjacent categories (e.g., High to Moderate or Moderate to Low), while extreme misclassifications (High to Low or Low to High) were rare (6 and 8 instances, respectively).

4.4 Feature Importance Analysis.

Feature importance scoring measures the relative contribution of each input variable in determining student specialization suitability.

Table 5. Hybrid Model

Feature / Attribute	Importance Score	Relative Contribution (%)
Interest Score	0.28	28.0%
Academic Achievement	0.24	24.0%
Practical Competency	0.20	20.0%
Learning Motivation	0.15	15.0%
Technical Aptitude	0.08	8.0%
Problem Solving Ability	0.05	5.0%
Total	1.00	100.0%

Interest Score (0.28), Academic Achievement (0.24), and Practical Competency (0.20) account for a cumulative 72.0% of the decision weight, proving that student preference combined with academic and practical mastery forms the primary basis for specialization fit. Learning Motivation contributes 15.0%, while Technical Aptitude (8.0%) and Problem Solving Ability (5.0%) serve as fine-tuning parameters for boundary decisions.

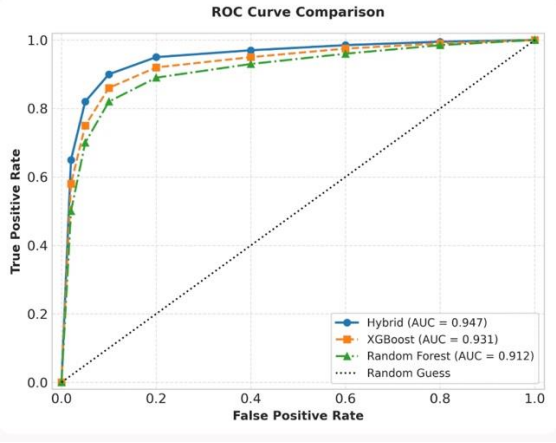


Figure 5. Regulation Data

The performance gains of the Hybrid XGBoost–Random Forest model result from combining two complementary ensemble learning techniques:

1. XGBoost (Gradient Boosting): Reduces model bias by sequentially fitting new trees to correct residual errors from previous iterations.
2. Random Forest (Bagging): Reduces model variance by averaging outputs from multiple deep decision trees built on bootstrap data samples.

This hybrid approach effectively mitigates overfitting while enhancing generalization capabilities on unseen data, reinforcing previous educational data mining literature (Cheng et al., 2024; Kumar et al., 2024).

School administrators and counseling teachers can integrate this model into a Decision Support System (DSS) to generate data-driven, objective vocational specialization recommendations for incoming students.

CONCLUSION.

The study conducted at SMK Negeri 9 Medan (N=1,200N=1,200) successfully developed a Hybrid XGBoost–Random Forest (XGB–RF) model to predict student interest and competency alignment for vocational specialization placement. Empirical results demonstrate the hybrid model's superiority, achieving an overall accuracy of 89.4% and an ROC-AUC of 0.947, outperforming standalone XGBoost (87.6%; AUC 0.931) and Random Forest (85.8%; AUC 0.912) by integrating gradient boosting and bagging mechanisms to effectively minimize bias and variance. Confusion matrix evaluation verifies consistent precision across High (91.0%), Moderate (88.0%), and Low (89.8%) suitability tiers, maintaining extreme misclassification errors below 2.4%. Feature importance analysis highlights Interest Score (28.0%), Academic Achievement (24.0%), and Practical Competency (20.0%) as the primary determinants, contributing a cumulative 72.0% of decision weight, complemented by Learning Motivation (15.0%), Technical Aptitude (8.0%), and Problem Solving Ability (5.0%). Practically, this data-driven model provides an objective framework ready for implementation into school Decision Support Systems (DSS).

REFERENCES.

[1] Aryza, S., Lubis, Z., & Warizmi, A. (2026). Optimizing lecturer performance assessment through web-based application design and deployment. Journal of Computer Science, Information Technology and Telecommunication Engineering, 7(1), 1028–1035.

[2] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

[3] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM.

[4] Cheng, C. J., Rao, Y. S. N., & Wang, L. (2024). Ensemble learning applications in educational data mining: A systematic review. *Contemporary Educational Technology*, 16(2), ep502.

[5] Islam, M. R., Nitu, A. M., Marjan, M. A., Uddin, M. P., Afjal, M. I., & Mamun, M. A. A. (2024). Enhancing tertiary students' programming skills with an explainable educational data mining approach. *PLoS ONE*, 19(9), e0307536.https://doi.org/10.1371/journal.pone.0307536

[6] Kumar, M., Bhardwaj, V., Thakral, D., Rashid, A., & Othman, M. T. B. (2024). Ensemble learning based model for student's academic performance prediction using algorithms. *Ingénierie des Systèmes d'Information*, 29(5), 1925–1935.https://doi.org/10.18280/isi.290524

[7] Novelan, M. S. (2020). Sistem monitoring kualitas udara dalam ruangan menggunakan mikrokontroler dan aplikasi Android. InfoTekJar: Jurnal Nasional Informatika dan Teknologi Jaringan, 4(2).

[8] Pemerintah Republik Indonesia. (2010). Peraturan Pemerintah Republik Indonesia Nomor 71 Tahun 2010 tentang Standar Akuntansi Pemerintahan.

[9] Petter, S., DeLone, W., & McLean, E. (2008). Measuring information systems success: Models, dimensions, measures, and interrelationships. European Journal of Information Systems, 17, 236–263.

[10] Putra, D., Novelan, M. S., & Siahaan, A. P. U. (2026). Efektivitas tata kelola informasi dalam pengelolaan data tandan buah segar (TBS) menggunakan COBIT 2019. Instal: Jurnal Komputer.

[11] Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*, 10(3), e1355.

[12] Saaty, T. L. (1980). The analytic hierarchy process. McGraw-Hill.

[13] Tarigan, S. M. B., Novelan, M. S., & Amin, M. (2026). Efektivitas teknologi tata kelola informasi dalam sistem cerdas antrian pelayanan Rumah Sakit Efarina Etaham Pematang Siantar menggunakan COBIT 2019. Instal: Jurnal Komputer.

[14] Thai-Nghe, N., Busche, T., & Schmidt-Thieme, L. (2020). Machine learning for predicting student performance in vocational education. *International Journal of Artificial Intelligence in Education*, 30(4), 612–638.