Optimization of Artificial Intelligence (AI) Dependency Level Classification Among Students Using Random Forest with SMOTE Oversampling Technique to Address Data Imbalance

Hermanto[1*], Riza Fahlapi[2], Antonius Yadi Kuntoro[3], Khoirul Rista Abidin[4]

[1]Department of Information Technology, Politeknik 'Aisyiyah Pontianak, Pontianak, West Kalimantan, Indonesia
[2]Department of Information Technology, Universitas Bina Sarana Informatika, Central Jakarta, DKI Jakarta, Indonesia
[3]Department of Information Systems, Universitas Nusa Mandiri, East Jakarta, DKI Jakarta, Indonesia
[4]Department of Medical Laboratory Technology, Politeknik 'Aisyiyah Pontianak, Pontianak, West Kalimantan, Indonesia
E-mail: [1*]hermanto.hm@polita.ac.id, [2]riza.rzf@bsi.ac.id, [3]antonius.aio@bsi.ac.id, [4]khoirulrista@polita.ac.id

(Received: 4 Jun 2026, revised: 27 Jun 2026, accepted: 29 Jun 2026)

Abstract

The development of generative artificial intelligence (AI) has significantly transformed the way students complete academic tasks. Excessive reliance on AI has the potential to diminish students' creativity and independent thinking abilities. This study aims to classify the level of students' dependency on AI in relation to academic creativity using the Random Forest algorithm optimized with the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance in the data. Data were collected through a structured questionnaire using a Likert scale of 1–5 consisting of 24 statement items from 74 student respondents at Politeknik 'Aisyiyah Pontianak, covering variables of AI usage intensity, dependency behavior, creativity, independent thinking, and academic motivation. Respondents were classified into three classes: Low (n=30), Moderate (n=19), and High (n=25). The results indicate that the Random Forest + SMOTE model evaluated with Cross Validation achieved an accuracy of 97.14% (±6.02%), a precision of 97.06% (micro average), and recall values of 100% for the Low class, 95.83% for the Moderate class, and 95.00% for the High class. The most dominant feature was the tendency to think of AI-generated answers when questioned by lecturers (B4, importance=0.1097), followed by AI usage intensity of more than 3 hours per day (A2, importance=0.0872). These findings contribute to the development of an early detection system for AI dependency among students in higher education institutions.

Keywords: Data Mining, AI Dependency, Data Classification, Student Creativity, Random Forest.

I. INTRODUCTION

The development of generative artificial intelligence (AI) technology, particularly large language models (LLMs) such as ChatGPT, Gemini, and Copilot, has brought fundamental changes to the world of higher education. Students now have easy access to tools capable of generating text, answering questions, and automatically completing academic tasks [1].

A study by Zhai et al. [2] found that excessive dependency on AI dialogue systems negatively impacts students' cognitive abilities, including critical thinking, creativity, and learning independence. In line with this, Kasneci et al. [3] emphasized that although AI has the potential to positively transform education, its negative effects on the development of higher-order thinking skills require serious attention. At the local level, students at Politeknik 'Aisyiyah Pontianak demonstrate a similar pattern, with many acknowledging their inability to complete assignments without the assistance of AI.

The primary challenge in classifying dependency levels is the presence of an imbalanced data distribution (imbalanced dataset), which can cause machine learning models to be biased toward the majority class [4]. This study proposes a combination of the Random Forest algorithm with the SMOTE oversampling technique to address this issue.

The research questions of this study are as follows: (1) How is the distribution of students' dependency levels on AI? (2) Is the SMOTE technique capable of improving the classification performance of the Random Forest model? (3) Which features are most influential in classifying students' AI dependency? The objectives of this study are to classify the level of dependency, compare model performance with and without SMOTE, and identify the dominant features.

II. LITERATURE REVIEW

A. AI Dependency in Education

Dependence on AI in academic contexts is defined as students' tendency to delegate their thinking process and task completion to AI systems without making any effort to understand or work through problems independently. Kasneci et al. [5] argue that generative AI opens up opportunities for transforming learning, while simultaneously introducing academic risks in the form of diminished independent thinking and the erosion of creativity. Farrokhnia et al. [6] employed a SWOT framework to analyze ChatGPT, finding that its primary threat lies in long-term cognitive dependency. Liao & Sundar [7] further contend that excessive trust in AI can lead to over-reliance if left unmanaged.

B. Random Forest Algorithm

Random Forest is an ensemble learning method that constructs a number of decision trees and generates predictions based on the class mode across all individual trees [8]. This algorithm is robust against overfitting, capable of handling high-dimensional data, and produces feature importance scores that aid in model interpretability. In the context of classifying educational survey data, Random Forest has been shown to yield competitive performance [9].

C. SVM / Naïve Bayes Algorithm

The use of classification algorithms such as Support Vector Machine (SVM) and Naïve Bayes has been widely applied across various text mining studies, including in the classification of student complaints, where SVM demonstrated a higher accuracy of 84.45% compared to Naïve Bayes at 69.75% [10].

D. Text Mining / Classification

Sentiment analysis research comparing Naïve Bayes and SVM on application reviews has shown that Naïve Bayes can outperform SVM in certain cases, as demonstrated in the classification of user comments on the Getcontact application, where Naïve Bayes achieved an accuracy of 82.97% compared to SVM's 78.00% [11].

E. SMOTE Oversampling Technique

SMOTE (Synthetic Minority Over-sampling Technique) is a method for addressing class imbalance by generating new synthetic samples in the minority class through interpolation between existing samples [12]. Assert that SMOTE remains the most widely used baseline technique and frequently delivers significant improvements in recall metrics for the minority class. The combination of SMOTE with Random Forest has been reported to yield the best performance across various classification domains [9].

F. Students' Academic Creativity

Creativity in academic contexts encompasses the ability to generate original ideas, solve problems through novel approaches, and develop divergent thinking [13]. Grassini [14] asserts that excessive AI use can erode students' intrinsic creativity, as they lose opportunities to practice developing ideas independently. Wardat et al. [15] found a negative correlation between the intensity of ChatGPT usage and academic creativity in the context of science learning.

III. RESEARCH METHODOLOGY

A. Research Type and Design

This study employs a quantitative approach with a descriptive-experimental design grounded in data mining. Classification methods were applied to survey data in order to categorize the level of students' dependence on AI and its impact on academic creativity.

B. Population and Sample

The research population comprised all active students of Politeknik Aisyiyah Pontianak in the 2025/2026 academic year (Table 1). Sampling was conducted using purposive sampling with the following criteria: active students from the 2nd through 8th semester who had used at least one generative AI platform for academic purposes. Data collection was carried out via Google Form in May 2026.

Table 1. Details of Research Respondent Count

Description	Count	Percentage (%)
Total respondents	74	100%
Valid respondents	74	100%
Invalid respondents	0	0%
Female	63	85.1%
Male	10	13.5%
2nd Semester	14	18.9%
4th Semester	38	51.4%
6th Semester	11	14.9%
8th Semester and above	11	14.9%
Mean GPA	3.56	—

C. Research Instrument

The instrument used was a structured questionnaire with a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree), consisting of 24 statement items divided into five variables: (A) Frequency and duration of AI use (5 items: A1–A5); (B) AI-dependent behavior (5 items: B1–B5); (C) Creativity and originality of ideas (5 items: C1–C5); (D) Independent thinking and learning (5 items: D1–D5); and (E) Academic motivation (4 items: E1–E4).

D. Research Workflow

The research stages were as follows (Figure 1): (1) Data collection via Google Form; (2) Data preprocessing, comprising the encoding of Likert responses into a numerical scale of 1–5 and the imputation of missing values using the median; (3) Class label formation based on the mean scores of

DOI: 10.34148/teknika.v15i2.1509

variables A and B using tertile division: Low (≤2.44), Moderate (2.45–2.98), and High (>2.98); (4) Stratified data splitting at an 80:20 ratio; (5) 10-fold Cross Validation was applied to the training data, where SMOTE oversampling was performed exclusively within each training fold to prevent data leakage - synthetic samples were never included in the validation fold; (6) Training of the Random Forest model on each oversampled training fold; (7) Final evaluation using 10-fold Cross Validation results in RapidMiner.

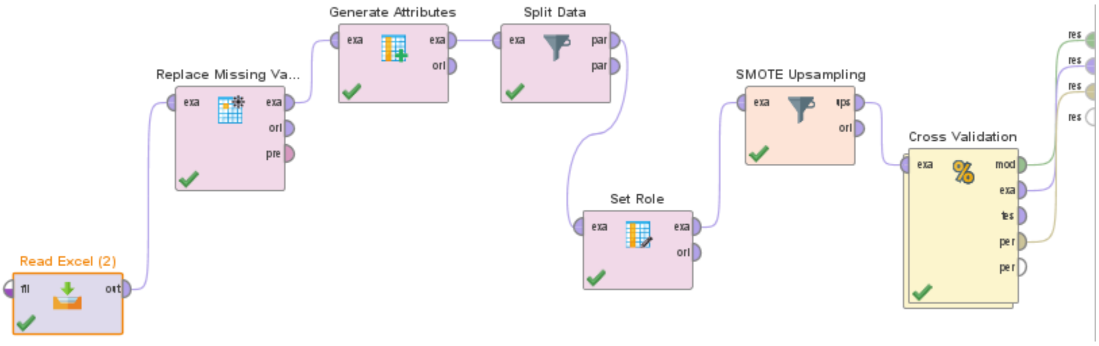

Figure 1. Research Process Workflow in RapidMiner

E. Model Parameter

Model parameters used in this research are shown in Table 2 below.

Table 2. Random Forest and SMOTE Model Parameters

Parameter	Value	Description
n_estimators	100	Number of decision trees
max_depth	10	Maximum tree depth
min_samples_split	2	Minimum samples required for a split
k_neighbors (SMOTE)	5	Number of neighbors for synthesis
random_state	42	Reproducibility seed
test_size	20%	Testing data proportion (15 samples)

IV. RESEARCH FINDINGS

A. Respondent Characteristics

This study involved 74 student respondents from Politeknik Aisyiyah Pontianak, with a predominance of female students at 85.1% (63 students) and male students at 13.5% (10 students). The majority of respondents were from the 4th semester (51.4%), followed by the 2nd semester (18.9%), the 6th semester (14.9%), and the 8th semester and above (14.9%). The respondents' mean GPA was 3.56 (SD = 0.24), ranging from 2.50 to 4.00. In terms of study programs, respondents were predominantly from the D3 Midwifery program (±33%) and the D4/D3 Medical Laboratory Technology program (±53%), with a small proportion from the Information Technology program.

B. Respondent Characteristics

Class label formation was carried out based on the mean scores of the dependency variables (A and B), categorized using a tertile approach, yielding three classes as presented in Table 3.

Table 3. Class Distribution Before and After SMOTE

Class	Original	SMOTE	Description
Low	30	24	Score ≤2.44
Moderate	19	24	Score 2.45–2.98
High	25	24	Score >2.98
Total	74	72	Train:Test = 59:15

Prior to the application of SMOTE, the training data exhibited an imbalanced distribution (Low = 24, Moderate = 15, High = 20 samples). Following the application of SMOTE with k_neighbors = 5, all classes became balanced at 24 samples each (total of 72 training samples). The application of SMOTE was considered essential for improving classification quality [4].

C. Classification Model Performance

Model performance evaluation was conducted using 10-fold cross-validation on the SMOTE-processed data in RapidMiner (Figure 2). A comparison of model performance is presented in Table 4. To ensure unbiased evaluation, SMOTE was applied only within each training fold of the cross-validation process, preventing any synthetic samples from contaminating the validation folds.

The Random Forest + SMOTE model achieved an accuracy of 97.14% (±6.02%) with a micro average of 97.06%, representing a significant improvement over Random Forest without SMOTE (81.33%). The application of SMOTE demonstrably enhanced the model's ability to recognize all classes evenly, consistent with the findings of Ndiaye et al. [9] that the combination of SMOTE with ensemble methods yields the best classification performance.

In addition to cross-validation evaluation, the final model was tested on the independent hold-out test set of 15 original samples (20% of the total data) that were entirely excluded from training and SMOTE oversampling. Table 4 presents the classification results on this real, unseen data.

DOI: 10.34148/teknika.v15i2.1509

Table 4. Performance Comparison of the Random Forest Model (10-Fold Cross Validation).

Metode	Accuracy	Precision	Recall	F1-Score
RF (without SMOTE)	81.33%	80.00%	81.33%	80.64%
RF + SMOTE	**97.14% (±6,02%)**	**97.06% (micro)**	**95.83%** (Moderate)	**97.06%**

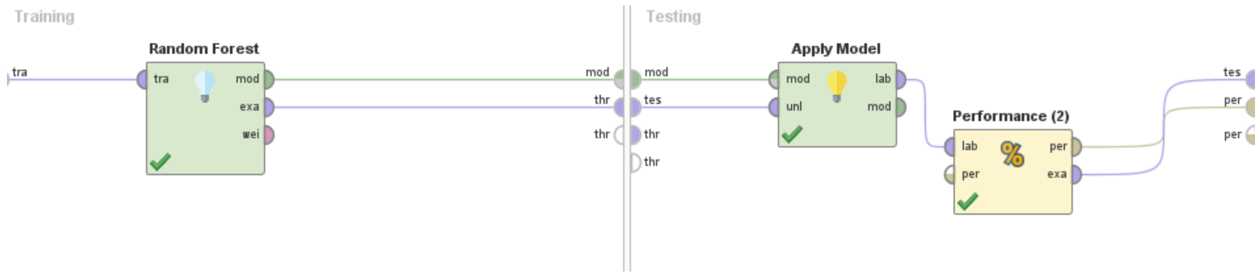

Figure 2. Cross-Validation Structure in RapidMiner

accuracy: 97.14% +/- 6.02% (micro average: 97.06%)

	true Sedang	true Rendah	true Tinggi	class precision
pred. Sedang	23	0	1	95.83%
pred. Rendah	1	24	0	96.00%
pred. Tinggi	0	0	19	100.00%
class recall	95.83%	100.00%	95.00%	

Figure 3. Confusion Matrix of Cross Validation Results for RF + SMOTE (Accuracy 97.14% ±6.02%)

D. Confusion Matrix and Per-Class Performance

The confusion matrix resulting from Cross Validation, along with the recall and precision values per class, are presented in Figure 3 and Table 5.

Table 5. Confusion Matrix and Per-Class Performance of RF + SMOTE

Actual \ Pred.	Low	Moderate	High	Recall
Low	24	1	0	100%
Moderate	0	23	1	95.83%
High	0	1	19	95,00%
Precision	**96.00%**	**95.83%**	**100%**	**97.06%**

The Low class was classified perfectly (recall = 100%, precision = 96.00%). The Moderate class demonstrated a recall of 95.83% with a precision of 95.83%, in which only 1 sample was incorrectly predicted as High. The High class yielded a recall of 95.00% and a precision of 100%. The excellent balance across all three classes represents a positive outcome of SMOTE application, which successfully addressed the imbalanced data distribution [4].

E. Feature Importance

Table 6 presents the ranking of features based on their level of importance in the Random Forest + SMOTE model.

Table 6. Feature Importance of the Random Forest Model (Top 10 Features)

Rank	Code	Feature Name	Score	Variable
1	B4	Tendency to recall AI answers when questioned by lecturers	0.1097	Dependency
2	A2	AI used for more than 3 hours per day	0.0872	Frequency
3	B2	Difficulty completing assignments without internet/AI	0.0839	Dependency
4	A4	Unable to complete assignments without AI	0.0764	Frequency
5	B3	No need for deep thinking due to AI	0.0724	Dependency
6	A5	AI used for almost all types of assignments	0.0690	Frequency
7	B5	Reduced thinking ability due to AI	0.0582	Dependency
8	A3	Prefers asking AI over consulting references	0.0451	Frequency
9	B1	Directly submits AI output without analysis	0.0418	Dependency
10	E2	AI motivates more diligent studying	0.0416	Motivation

DOI: 10.34148/teknika.v15i2.1509

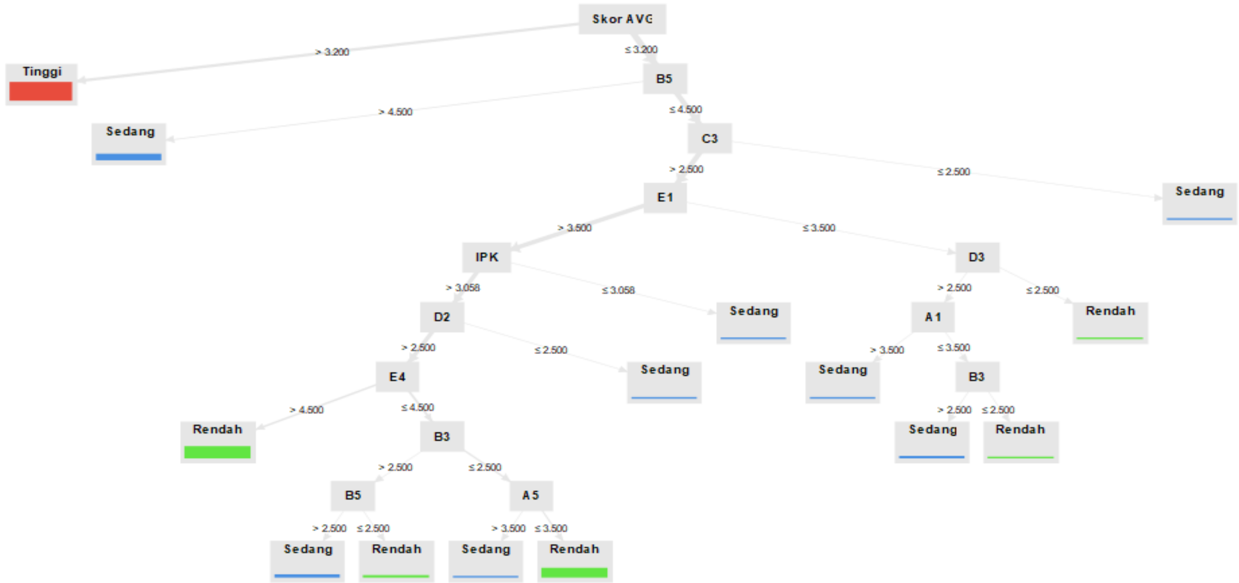

Figure 4. Example of a Decision Tree within the Random Forest Model

The most dominant feature was B4 (the tendency to recall AI-generated answers when questioned by lecturers, importance = 0.1097), reflecting a profound form of cognitive dependency, in line with the findings of [13]. Regarding cognitive offloading. Feature A2 (intensity of AI use exceeding 3 hours per day, importance = 0.0872) and B2 (inability to complete assignments without AI, importance = 0.0839) ranked second and third, respectively. A motivation variable (E2) also appeared among the top 10 features, indicating that not all AI use carries a negative impact on academic motivation [6].

The application of optimization-based classification methods such as Support Vector Machine combined with Particle Swarm Optimization (PSO) has also been shown to yield high accuracy of up to 93.0% with an AUC of 0.977 in sentiment analysis of application reviews [16].

F. Discussion

The research findings revealed that 33.8% of students (25 individuals) exhibited a high level of AI dependency, demonstrating a consistent pattern of using AI for more than 3 hours per day, being unable to complete assignments without AI, and tending to submit AI-generated outputs without further analysis. These findings are consistent with [1], who identified over-reliance as the primary threat of AI integration in higher education.

A negative trend was observed between the level of AI dependency and creativity scores: the Low group (mean = 3.43), Moderate group (mean = 3.30), and High group (mean = 3.15). This pattern is consistent with the hypothesis that excessive AI use can suppress students' academic creative expression [15].

The model accuracy of 97.14% (±6.02%) achieved through 10-fold Cross Validation represents an excellent result for a dataset of 74 samples with three target classes. This high level of accuracy demonstrates that the combination of Random Forest with SMOTE is capable of effectively addressing class imbalance (Figure 4) [9]. The practical implications of these findings include the need for balanced AI usage policies, the introduction of AI literacy as a compulsory module, and the implementation of a data mining-based early warning system at Politeknik Aisyiyah Pontianak [7].

This study acknowledges several limitations. The total sample size of N=74 is relatively small for training ensemble-based models such as Random Forest and applying SMOTE with k=5 nearest neighbors. With only 59 training samples, the synthetic boundaries generated by SMOTE may not fully represent the true population distribution, potentially affecting model robustness. Future studies are recommended to collect larger and more diverse datasets across multiple institutions to improve the generalizability of the classification model.

V. CONCLUSION

Based on the findings and discussion, the following conclusions can be drawn: (1) The distribution of AI dependency levels among students of Politeknik Aisyiyah Pontianak is divided into three classes: Low (40.5%), Moderate (25.7%), and High (33.8%), with feature B4 (the tendency to recall AI-generated answers when questioned by lecturers) as the most dominant feature (importance = 0.1097); (2) The Random Forest + SMOTE model achieved an accuracy of 97.14% (±6.02%) through 10-fold Cross Validation with a micro average of 97.06%, representing a significant improvement over the model without SMOTE (81.33%); (3) A negative trend exists between the level of AI dependency and students' academic creativity scores; (4) The dependency behavior variable (Var B) and frequency of use variable (Var A) serve as the primary predictors in the classification.

DOI: 10.34148/teknika.v15i2.1509

Based on the research findings, the following recommendations are proposed: (1) Expansion of the dataset across multiple universities to improve model generalizability; (2) Exploration of alternative algorithms such as XGBoost, LightGBM, or SVM for comparative purposes; (3) Development of a web-based prediction system for early detection of AI dependency; (4) Incorporation of demographic and contextual variables in future research.

REFERENCES

[1] M. Farrokhnia, S. K. Banihashem, O. Noroozi, and A. Wals, "A SWOT analysis of ChatGPT: Implications for educational practice and research," Innovations Educ. Teach. Int., vol. 61, no. 3, pp. 460–474, 2023. https://doi.org/10.1080/14703297.2023.2195846

[2] C. Zhai, S. Wibowo, and L. D. Li, "The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review," Smart Learning Environments, vol. 11, no. 1, p. 28, 2024. https://doi.org/10.1186/s40561-024-00316-7

[3] E. Kasneci et al., "ChatGPT for good? On opportunities and challenges of large language models for education," Learning and Individual Differences, vol. 103, p. 102274, 2023. https://doi.org/10.1016/j.lindif.2023.102274

[4] H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263–1284, 2009. https://doi.org/10.1109/TKDE.2008.239

[5] B. Charbuty and A. Abdulazeez, "Classification based on decision tree algorithm for machine learning," J. Appl. Sci. Technol. Trends, vol. 2, no. 01, pp. 20–28, 2021. https://doi.org/10.38094/jastt20165

[6] T. M. Amabile, Creativity in Context: Update to the Social Psychology of Creativity. Westview Press, 1996

[7] Q. V. Liao and S. S. Sundar, "Designing for responsible trust in AI systems: A communication perspective," in Proc. 2022 ACM Conf. Fairness, Accountability, Transparency (FAccT '22), pp. 1257–1268, 2022. https://doi.org/10.1145/3531146.3533182

[8] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001. https://doi.org/10.1023/A:1010933404324

[9] M. Ndiaye, O. Sarr, and D. Seck, "Oversampling methods for imbalanced datasets in classification: A systematic review," Artif. Intell. Rev., vol. 56, pp. 13057–13081, 2023. https://doi.org/10.1007/s10462-023-10492-8

[10] H. Hermanto, A. Mustopa, and A. Y. Kuntoro, "Algoritma Klasifikasi Naive Bayes Dan Support Vector Machine Dalam Layanan Komplain Mahasiswa," *JITK (Jurnal Ilmu Pengetahuan Dan Teknologi Komputer)*, vol. 5, no. 2, pp. 211–220, Feb. 2020. https://doi.org/10.33480/jitk.v5i2.1181'

[11] Hermanto, R. Fahlapi, A. Y. Kuntoro, and T. Asra, "Perbandingan Algoritma Klasifikasi Analisis Sentimen Pengguna Aplikasi Getcontact Dalam Pencegahan Penipuan Online," *J-INTECH (Journal of Information and Technology)*, pp. 158–167, 2024

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002. https://doi.org/10.1613/jair.953

[13] E. P. Torrance, Torrance Tests of Creative Thinking: Norms-Technical Manual. Personnel Press, 1966

[14] S. Grassini, "Shaping the future of education: Exploring the potential and consequences of AI and ChatGPT in educational settings," Education Sciences, vol. 13, no. 7, p. 692, 2023. https://doi.org/10.3390/educsci13070692

[15] Y. Wardat, M. A. Tashtoush, R. AlAli, and A. M. Jarrah, "ChatGPT: A revolutionary tool for teaching and learning mathematics," EURASIA J. Math. Sci. Technol. Educ., vol. 19, no. 7, p. em2286, 2023. https://doi.org/10.29333/ejmste/13272

[16] A. Noviriandini, Hermanto, and Yudhistira, "Klasifikasi Support Vector Machine Berbasis Particle Swarm Optimization Untuk Analisa Sentimen Pengguna Aplikasi Pedulilindungi," *JIKA (Jurnal Informatika)*, vol. 6, no. 1, pp. 50–56, Apr. 2022. https://doi.org/10.31000/jika.v6i1.5681