KOMTEKINFO JOURNAL
LPPM Universitas Putra Indonesia YPTK Padang
Jl. Raya Lubuk Begalung, Padang, West Sumatera, Indonesia, Zip Code: 25221
Volume: 13, Issue: 2, Page: 89 - 94, June 30th, 2026, e-ISSN: 2502-8758
Available online at website: https://jkomtekinfo.org/ojs/index.php/komtekinfo

Model for Identifying High-Achieving Students Using The K-Means Clustering Algorithm and C.4.5 Classification

Kalfinus Waruwu[1]✉, Gunadi Widi Nurcahyo[2], Sumijan[3]
[1,2,3] Fakultas Ilmu Komputer, Universitas Putra Indonesia YPTK, Padang, 25221, Indonesia
kalfinuswar7@gmail.com

Abstract

Student achievement refers to academic accomplishments or results obtained by students in the field of education, which can influence the process of determining student academic grades, class achievement, and accomplishments. This process plays a strategic role in supporting objective educational decision-making, especially in the preparation of coaching programs, the establishment of awards, and the continuous development of student potential. Based on this, the purpose of this study is to analyze data on high-achieving students using the K-Means and C4.5 algorithms. The research methods used include K-means, which functions to group student data into different groups. C4.5 classification is capable of analyzing data characteristic similarities, and the results of the decision tree are used as the basis for the decision-making process. The dataset in this study consisted of 345 students from SMK Negeri 3 Padangsidimpuan. Based on the results of this study, it was proven that the application of the K-Means and C4.5 algorithms could achieve an accuracy of 98.59%. This research contributes to identifying high-achieving students at SMKN 3 Padangsidimpuan using the K-Means and C4.5 algorithms, which can assist the school in formulating more effective and targeted guidance policies and presenting the results of identifying high-achieving students after clustering and decision tree analysis. This serves as a basis for decision-making in determining student development programs based on objectively identified academic and non-academic achievement clusters.

Keywords: Student Achievement, K-Means Clustering, C4.5 Algorithm, Classification, Decision Tree.

1. Introduction

In today's digital age, data has become a valuable asset in supporting effective decision-making in various sectors, including education. The education system in Indonesia requires a technology-based approach to optimize the use of data in improving the quality of education [1]. By introducing quality education, more experts are beginning to pay attention to the learning process. The main focus of educational evaluation reform has now shifted towards the concept of evaluating this process. Most teachers place greater emphasis on assessing student learning progress, both in terms of scientific aspects and completeness [2]. Technology and decision support systems play an important role in ensuring that the educational data obtained can be analyzed effectively and interpreted correctly, thereby supporting more effective and data-driven decision making. To that end, grouping based on Education Performance Reports is required, as well as the application of decision support system algorithms through clustering methods [3].

Student achievement refers to academic accomplishments or results obtained by students in the field of education. Given the large amount of data involved, this can influence the process of determining student academic grades, class achievement, and school achievement in assessing student achievement as a whole [4]. Collecting data and identifying patterns among student groups is important, but there are several limitations to this process. One of the main obstacles is the selection of variables used. The variables selected must have a significant correlation with student achievement. This means that not all variables selected will necessarily influence student learning [5].

In the context of educational technology and the current digital age, well-organized analysis and presentation of learning data is essential to improve student engagement and learning outcomes. This includes students' ability to plan, monitor, and evaluate their own learning process, which is the key to academic success [6]. Unsupervised machine learning can be applied to group students and provide personalized recommendations in education. This serves as a guide for integrating machine learning technology into education systems, and demonstrates how computational tools can support educators in improving engagement, equity, and academic success in increasingly digitized classrooms [7].

Cluster analysis can classify and process similar data. However, if the initial cluster centers in K-Means are chosen randomly, there is a possibility that some centers will be in the same cluster, thereby significantly reducing the validity of the clustering results [8]. Based on activity patterns and fitness levels, participants can be grouped using cluster analysis. This method is used

Submitted: 20-03-2026 | Revised: 19-04-2026 | Accepted: 20-05-2026 | Published: 30-06-2026

to reduce datasets with many variables, so that natural groups within the data can be identified [9].

In machine learning, K-Means is used to find clusters and patterns in data. This algorithm begins with the random selection of a number of cluster centers (centroids), then each data point is placed on the nearest centroid. After that, the position of the centroid is updated based on the average of the data points in the cluster, and this step is repeated until the position of the centroid no longer changes [10]. The K-Means algorithm is used in the clustering process. The process is carried out through centroid initialization, assigning data points to the nearest centroid, updating the centroid position, and repeating iterations until convergence is achieved [11]. K-means serves to group student data into different clusters. K-means requires time for convergence and the structure of the data set being studied. K-means shows that the number of iterations is related to the quality of clustering and can help identify irrelevant features in the data, as well as improve the results of existing feature selection algorithms [12].

Classification techniques are efficient algorithms for constructing decision trees to handle both discrete and numeric attributes. These algorithms serve to describe or distinguish classes or concepts of data in the classification process. The main objective of classification is to determine relevant attributes, while in clustering, data is organized according to the required categories [13].

Classification is performed by analyzing the similarity of data characteristics, and the results of the decision tree are used as the basis for the decision-making process. This approach is part of data mining, which aims to predict the membership of a data set in a particular group [14]. Classification is a data analysis model that forms a decision model such as feasible or not feasible. It functions to read Excel files to access data, and a decision tree algorithm is applied to run the classification method [15].

Decision trees are widely recognized as an effective tool for classification. Decision trees have a structure consisting of root nodes, branches, and leaf nodes that are used to predict outcomes. Decision trees focus on explaining decision tree algorithms and their application in classifying student academic performance [16].

2. Methods

Research methodology is a series of systematic steps used to design, conduct, and evaluate scientific research in order to achieve predetermined objectives. This methodology includes the selection of research types and approaches, data collection techniques, data analysis methods, and procedures for testing and validating research results.

The research framework is a conceptual structure that describes the systematic flow of the entire research process, starting from problem identification, literature review, data collection, to the data analysis and pre-processing stages, which include cleaning, selection, and transformation processes. The processed data is then analyzed using the K-Means algorithm for student grouping, followed by the C4.5 algorithm to form a decision tree as the basis for classifying high-achieving students. The final stage includes model evaluation, system implementation, result testing, and presentation of results and discussion as the basis for drawing research conclusions.

2.1 K-Means Algorithm

The steps in the K-Means algorithm begin with determining the number of clusters and selecting initial centroids at random. After that, the distance of each data point is calculated relative to each centroid to determine the most appropriate cluster. This process continues by updating the position of the centroid based on cluster members until the algorithm reaches a convergent condition [17]:

a. Determining the K Value
b. Determine the Initial Centroid

$$v = \frac{\sum_{i=1}^{n} x}{n} : i = 1,2,3 \ldots . n \quad (1)$$

In this formula, *V* represents the new centroid value obtained from recalculating the cluster center, while n indicates the amount of data included in each cluster. The new centroid value is calculated based on the average of all data in the cluster, thereby reflecting a more accurate cluster center in each iteration of the clustering process.

c. Calculate Euclidean Distance

$$D(i,j) = \sqrt{(X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2 + \cdots + (X_{ki} - X_{kj})^2} \quad (2)$$

D(*i,j*) is the distance between data i and cluster center j, which is used to determine the proximity of data to a cluster. The value X*ki* represents data *i* in attribute *k*, while X*k*j represents the value of cluster center *j* in the same attribute. This distance calculation is fundamental to the data clustering process because it determines the placement of data into the most appropriate cluster.

d. Recalculate the new centroid
e. K-Means Clustering Results

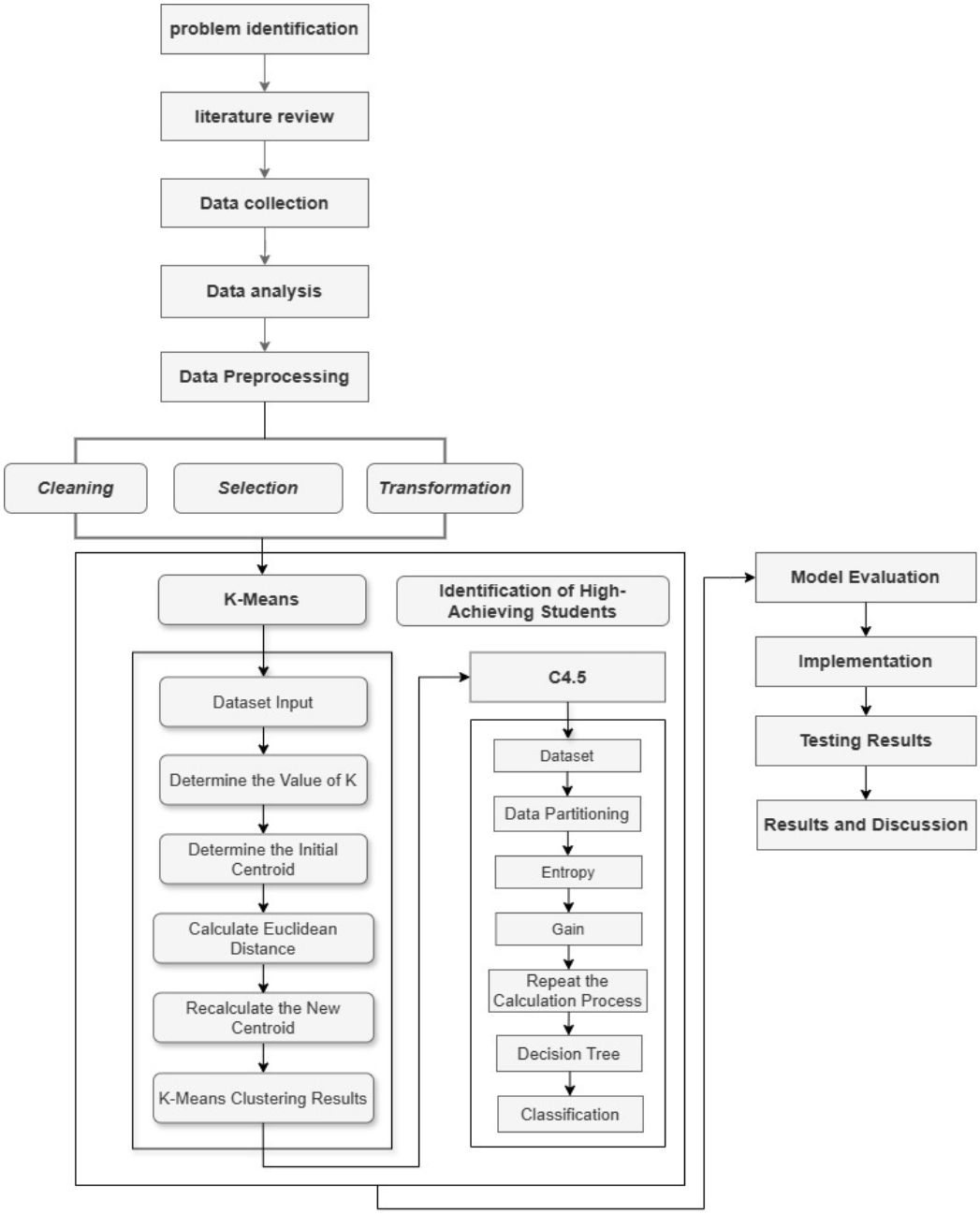

Figure 1. Research framework

2.2 C4.5 Algorithm

After obtaining the data grouping results, the next step is to apply the C4.5 method to build a classification model. This method works by selecting attributes that have the highest information value as the basis for data separation. In general, the C4.5 algorithm involves calculating gains, forming decision nodes, and creating branches until a complete decision tree structure is formed as follows [12]:

a. Entropy

$$Entropy(S) = \sum_{i=1}^{n} - pi * log_2 pi \quad (3)$$

Entropy (S) represents the total entropy value used to measure the level of uncertainty in data set S. Variable n indicates the number of partitions formed in set S, while p_i represents the proportion of each partition S_i to the entire data set S. This entropy value forms the basis of the attribute evaluation process because it reflects how homogeneous or heterogeneous the data distribution is in each partition.

b. Gain

$$Gain(S, A)\ Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i) \quad (4)$$

S represents the set of all cases analyzed, while A indicates the attributes used in the calculation process.

The Entropy(S) value indicates the total uncertainty level in the case set S, with n as the number of partitions generated by attribute A. Meanwhile, $|S_i|$ indicates the number of cases in the i-th partition and $|S|$ indicates the total cases in set S, which are used to calculate the weight of each partition in the attribute evaluation process.

3. Results and Discussions

This study integrates the K-Means and C4.5 algorithms to comprehensively analyze student data. The K-Means algorithm is used in the initial stage to group data based on similarity of characteristics through an iterative process until homogeneous clusters are obtained. Next, the clustering results are classified using the C4.5 algorithm to form an entropy-based decision tree and information gain as the basis for objective decision making.

3.1 Preprocessing

Data pre-processing is the initial stage that serves to prepare data so that it can be processed optimally in the subsequent analysis process. The data used in this study came from students at SMK Negeri 3 Padangsidimpuan. This stage is very important to ensure that the data used is of good quality, free from errors, and consistent. The pre-processing process includes three main steps, namely cleaning, data selection, and data transformation.

Table 1. Dataset

No	X1	X2	X3	X4	X5
1	82,9	83	82,9	92	80
2	83,6	83	83,4	92	92
3	87,7	86,3	87,3	92	94
4	84,3	83	83,9	92	95
5	89,3	87,3	88,7	92	92

Table 1 presents the average calculation results of a number of variables used in the study. Each value displayed represents the quality level of students based on predetermined indicators. The data that has undergone this transformation process is then ready to be used in the statistical analysis stage to identify high-achieving students in a more in-depth and objective manner.

3.2 K-Means Method Analysis

The data analysis process begins with the application of the K-Means algorithm. This algorithm is used to find patterns or groups of data that have similar characteristics, thereby revealing hidden structures in the dataset. This method plays an important role in identifying student performance trends and grouping them based on similar academic characteristics. The K-Means stages begin with determining the number of clusters, setting the initial centroid, then grouping each data point to the nearest centroid, until the iteration process produces stable and representative groups.

No	X1	X2	X3	X4	X5	Y
1	82,9	83	82,9	92	80	Moderate
2	83,6	83	83,4	92	92	High
3	87,7	86,3	87,3	92	94	High
4	84,3	83	83,9	92	95	High
5	89,3	87,3	88,7	92	92	High
6	84	83	83,7	92	94	High
7	88,3	88,3	88,3	92	92	High
8	89	88,3	88,8	92	86	Moderate
9	84,3	83,7	84,1	92	80	Moderate
10	10,7	26,7	15,5	92	82	Low
11	83,9	84	83,9	92	82	Moderate
12	85,4	84	85	92	90	High
13	85,7	84	85,2	92	85	Moderate
14	86,3	84,7	85,8	92	85	Moderate
15	84,3	83,7	84,1	92	87	Moderate
16	86,3	84,7	85,8	92	96	High
17	23	54	32,3	92	80	Low

Table 2. K-Means Clustering Results

Figure 3. shows the final results of the K-Means clustering process for each data point based on the five attributes analyzed. It shows the mapping of each data point to the centroid with the smallest distance based on the five distance values compared. The information in the table provides an overview of the data trends for each cluster.

3.3 C4.5 Analysis

After the clustering process with K-Means produces groups of students based on similarity of characteristics, the next step is to build a classification model using the C4.5 algorithm. At this stage, the cluster results are combined with other supporting attributes to calculate the gain and gain ratio values in order to determine the most influential attributes.

Figure 2 this is based on the tree structure above, the visualization results show that the decision trees formed using the C4.5 method in both models have identical structures. This indicates that differences in the dataset division scenarios do not affect the attribute selection or decision rules generated by the model.

3.4 Evaluation Process

Model evaluation is performed by comparing the identification results produced by the system with validation values that serve as a benchmark for accuracy. This conformity indicates that the model is capable of effectively learning data patterns and producing classifications that are consistent with actual conditions. This evaluation matrix displays the actual values and predicted values for each class.

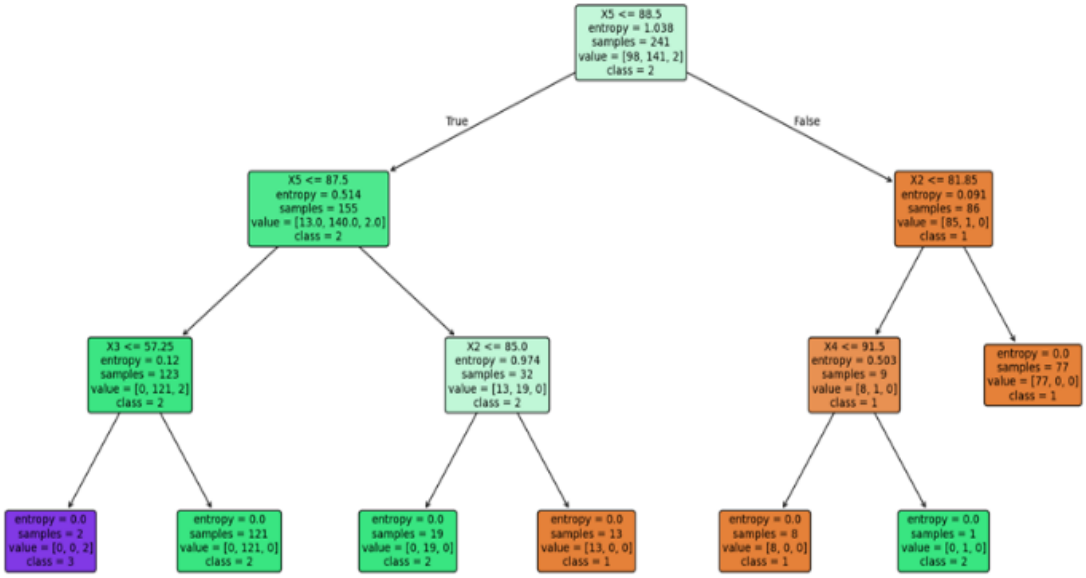

Figure 2. Decision Tree

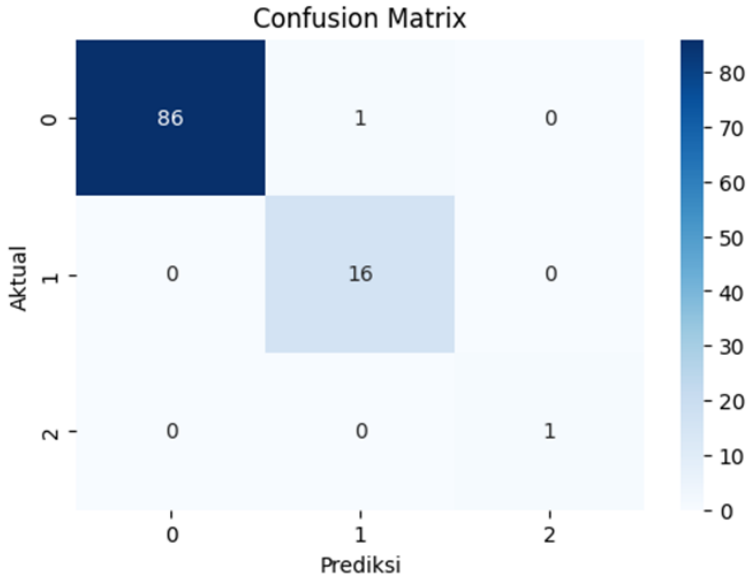

Figure 3. Evaluation Results

Figure 3. shows that although the training and test datasets are different, they perform well with *accuracy*, *precision, recall*, and F1-*score* values reaching 98% to 99%. This indicates that the model is able to classify all data correctly without prediction errors. Both datasets have almost the same high level of accuracy and reliability.

4. Conclusions

The model for identifying high-achieving students, which combines the K-Means algorithm for clustering and C4.5 for classification, has proven effective in analyzing student characteristics. The implementation of the model shows high accuracy, indicating that the combination of these two algorithms can be used as a reliable tool for data-driven decision making in education. These results support the formulation of objective and targeted student coaching and development programs.

References

[1] R. S. Mentari and M. Iqbal, "Student Data Analysis for Decision Making Using Decision Tree," *Journal of*

Information Technology, computer science and Electrical Engineering, vol. 1, no. 3, pp. 517–521, 2024.
[2] X. Li and L. Yuan, "Using Multiple Data Mining Technologies to Analyze Process Evaluation in the Blended-Teaching Environment," *Sustainability*, vol. 15, no. 5, p. 4075, 2023.
[3] L. R. Ilmi and M. Sumunaringtyas, "Clustering of High School Quality Using Fuzzy C-Means in the Special Region of Yogyakarta Province," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, 2025.
[4] N. R. Jevintya, U. Darussalam, and S. Abdullah, "Application of the K-Means and Decision Tree Algorithms in Determining Student Achievement," *JIKO (Jurnal Informatika Dan Komputer)*, vol. 7, no. 1, pp. 13–18, 2024.
[5] H. Shi, Y. Zhou, V. P. Dennen, and J. Hur, "From unsuccessful to successful learning: profiling behavior patterns and student clusters in Massive Open Online Courses," *Educ. Inf. Technol. (Dordr).*, vol. 29, no. 5, pp. 5509–5540, 2024.
[6] B. Pansri *et al.*, "Understanding Student Learning Behavior: Integrating the Self-Regulated Learning Approach and K-Means Clustering.," *Educ. Sci. (Basel).*, vol. 14, no. 12, 2024.
[7] K. Ouassif, B. Ziani, J. Herrera-Tapia, and C. A. Kerrache, "Empowering Education: Leveraging Clustering and Recommendations for Enhanced Student Insights," *Educ. Sci. (Basel).*, 2025.
[8] J. Zhang, Y. Kuang, and J. Zhou, "Data Mining Technology for Smart Campus in Behavior Association Analysis of College Students," *Scalable Computing: Practice and Experience*, vol. 25, no. 6, pp. 5701–5713, 2024.
[9] A. Korcz *et al.*, "Cluster analysis of physical activity and physical fitness and their associations with components of school skills in children aged 8–9 years," *Sci. Rep.*, vol. 15, no. 1, p. 5053, 2025.
[10] S. Kumar, R. Rani, S. K. Pippal, and R. Agrawal, "Customer segmentation in e-commerce: K-means vs hierarchical clustering," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 23, no. 1, pp. 119–128, 2025.
[11] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Inf. Sci. (N Y).*, vol. 622, pp. 178–210, 2023.
[12] H. Hendri *et al.*, "A hybrid data mining for predicting scholarship recipient students by combining K-means and C4. 5 methods," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 33, no. 3, pp. 1726–1735, 2024.
[13] E. V. Astuti, A. Afandi, and D. M. Efendi, "Classification and Clustering of Internet Quota Sales Data Using C4. 5 Algorithm and K-Means," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)*, vol. 9, no. 2, pp. 268–283, 2023.
[14] R. Wandri, Y. Arta, A. Hanafiah, and R. Oktaviaani, "Prediction of Student Scholarship Recipients Using the K-Means Algorithm and C4. 5," *The Indonesian Journal of Computer Science*, vol. 12, no. 1, 2023.
[15] S. Sukriadi and S. Suherman, "Penerapan Data Mining Untuk Klasifikasi Penerima Beasiswa Di SMK Negeri 3 Soppeng Menggunakan Metode Decision Tree," *Jurnal RISTER: Riset Sistem Cerdas*, vol. 1, no. 2, pp. 68–73, 2024.
[16] Y. Xie, "Efficiency of Hybrid Decision Tree Algorithms in Evaluating the Academic Performance of Students.," *International Journal of Advanced Computer Science & Applications*, vol. 15, no. 2, 2024.
[17] R. Ishak and A. Bengnga, "Clustering Prestasi Akademik Lulusan Menggunakan Metode K-Means," *Jambura Journal of Electrical and Electronics Engineering*, vol. 6, no. 1, pp. 76–81, 2024.
[18] A. Ramadhanu, H. Hendri, F. Hadi, D. Guswandi, D. M. Putra, R. Hardianto, and S. D. Rizki, "Hybrid decision support system and image processing for classifying priority applications in the Padang Government," CSRID (Computer Science Research and Its Development Journal), vol. 18, no. 1, pp. 178–191, 2026.
[19] A. Ramadhanu, H. Hendri, and F. Hadi, "Organic fertilizer content detection based on image segmentation and texture analysis," in Proc. 2025 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), Dec. 2025, pp. 50–55.

Biographies of Authors

Kalfinus Waruwu
I was born on September 9, 2002, in Sironcitan. My most recent education is a Bachelor's degree (S1) in Computer Science from Universitas Putra Indonesia YPTK Padang. Currently, I am pursuing a Master's degree (S2) in Computer Science at Universitas Putra Indonesia YPTK Padang. I can be contacted via email: kalfinuswar@gmail.com

Gunadi Widi Nurcahyo was born in Temanggung, 14 March 1969. He was graduated bachelor's degree in informatics management at Universitas Putra Indonesia YPTK Padang in 1992. He completed his Master and PhD in Computer Science at Universiti Teknologi Malaysia in 2003. I can be contacted via email: gunadiwidi@yahoo.co.id

Dr. Ir. Sumijan, M.Sc Dr. Ir. Sumijan, M.Sc., born in Nganjuk on May 7, 1966, is an academic, researcher, and educator who is active in the development of science, especially in the field of Information Technology. He serves as a lecturer at Universitas Putra Indonesia "YPTK" Padang, with NIDN 0005076607 and can be contacted via email at sumijan@upiyptk.ac.id.