Indonesian Journal of Data and Science

Volume 7 Issue 2
ISSN 2715-9936
https://doi.org/10.56705/ijodas.v7i2.388

Research Article

Applicant Data Segmentation and Pattern Analytics for University Admissions Strategy Using Hybrid SOM and K-Means

Diyah Ruswanti [1]; Dahlan Susilo [2]
[1] Universitas Sahid Surakarta, Surakarta, Indonesia, dyahruswanti.@usahidsolo.ac.id
[2] Universitas Sahid Surakarta, Yogyakarta, Indonesia, dahlansusilo@usahidsolo@ac.id

Correspondence should be addressed to Diyah Ruswanti; dahlansusilo@usahidsolo@ac.id

Received 21 February 2026; Accepted 30 July 2026; Published 31 July 2026

Abstract:
Introduction: Effective university admissions strategies require a clear understanding of applicant demographics, academic characteristics, geographic origins, and information channels. This study applies a hybrid clustering approach to identify meaningful applicant segments that can support targeted recruitment and resource allocation. **Method**: Historical applicant records from Universitas Sahid Surakarta covering 2021–2025 were preprocessed through data cleaning, one-hot encoding of categorical variables, and Min-Max normalization. A hybrid Self-Organizing Map (SOM) and K-Means framework was employed, where SOM projected high-dimensional applicant characteristics into a lower-dimensional topological representation and K-Means partitioned the resulting prototypes into distinct clusters. Clustering quality was evaluated using the Davies–Bouldin Index (DBI) to determine the optimal number of segments. **Results and Discussion**: The lowest DBI was obtained for three clusters, indicating the most appropriate segmentation structure. The resulting groups were characterized as proximity-driven local applicants, regional career-oriented applicants dominated by vocational-school backgrounds, and high-achieving out-of-region applicants with stronger academic performance and greater reliance on institutional websites and search channels. These patterns provide actionable insight for differentiated recruitment strategies. **Conclusion**: The hybrid SOM–K-Means approach effectively identifies interpretable applicant segments and provides descriptive intelligence that can support more targeted marketing, channel selection, scholarship strategies, and admissions resource allocation in higher education.

Keywords: Davies-Bouldin Index, Data Mining, K-Means, Prediction, Segmentation, SOM.

1. Introduction

The rapid development of internet technology has led to the increasing popularity of electronic computers and smartphones [1]. Segmentation serves as a foundation for marketing processes, as it is an essential aspect of marketing strategy that requires identifying relevant customer groups using appropriate segmentation approaches [2]. The sustainability of private universities largely depends on tuition fees paid by students, as they do not receive operational funding like public universities [3], [4]. Therefore, for private universities, the recruitment of new students is the lifeblood of the institution. The newer students a university attracts, the more stable its operations become, provided there is effective management [5].

Marketing strategies play a crucial role in determining the number of new students each year [6], [7]. Universitas Sahid Surakarta, a private university in Central Java Province, relies heavily on tuition fees for its operations. Predicting the number of new students is a major challenge for educational institutions as they plan effective strategies for student admissions. One way to predict the number of new students is through data segmentation. With data segmentation, Universitas Sahid Surakarta can gain a better understanding of prospective students' characteristics based on existing data, such as demographics, educational background, interests, and more. This segmentation helps identify homogeneous groups within the data, which can then be used to map existing patterns [4].

This study aims to explore the potential use of data segmentation methods based on SOM (Self-Organizing Maps) and K-Means in the context of predicting new student enrollments in an educational institution. SOM are used in this study for their ability to organize and map complex data into a two-dimensional grid structure [8], [9]. This method enables clear visualization of data distribution and relationships between observed variable categories [10]. On the other hand, K-Means is utilized to cluster data into groups based on similar characteristics, facilitating further analysis of each formed cluster [11], [12].

The K-Means algorithm is widely adopted for information management because it allows for more efficient organization and management of information by clustering similar products [13], [14]. By integrating these two methods, this study is expected to contribute significantly to understanding the dynamics of new student admissions. The segmentation results are expected to support more accurate decision-making in admission strategies and improve the efficiency and effectiveness of the recruitment process in educational institutions. The research consists of five sections. The first provides a general introduction, followed by a comprehensive literature review. The third section thoroughly examines SOM and K-Means algorithm techniques, while the fourth evaluates their performance through a series of tests. Finally, the fifth section summarizes the findings and limitations of this study. The data for this study is sourced from the Admissions Unit of Universitas Sahid Surakarta, including demographic data of prospective students, academic data, extracurricular activity data, and other information influencing applicants' decisions to enroll. The data will first be cleaned, especially for missing values or outliers [15]. Once cleaned, data normalization will ensure all features are on the same scale, especially when data comes from various sources with different scales. The data will then be split into training and testing datasets for model validation [16], [17].

SOM method will map data onto a 2D grid. SOM will identify patterns in the data and group it based on feature similarities [18]. SOM will be trained with the training data. This training process adjusts neurons in the grid to reflect the data distribution. The SOM results will then be visualized to identify patterns and clusters using heatmaps or U-Matrix [9]. Advanced segmentation with K-Means will use SOM visualization results as input for the K-Means algorithm. SOM mapping can help determine the optimal number of clusters [18]. The K-Means algorithm will then cluster the data mapped by SOM and determine the appropriate number of clusters based on SOM results. Clustering results will be evaluated using the Davies-Bouldin Index metric to ensure segmentation quality [19], [20].

A feature for predicting new student enrollment will be developed using classification to map whether prospective students will register. The predictive model will then be trained using training data and validated with test data, employing cross-validation to prevent overfitting [21], [22]. Finally, the results will be interpreted to understand the factors most influencing prospective students' decisions. Recommendations based on these findings will be made for more effective recruitment strategies. The model can be implemented in the recruitment system to predict new student enrollments in real time. This approach combines SOM's strength in visualizing and identifying data patterns with K-Means' ability to perform more specific clustering [23], [24]. With these steps, more accurate data segmentation and effective predictive models for new student enrollment are expected.

The study begins with hypothesis testing regarding the interest of applicants at Universitas Sahid Surakarta. The initial hypothesis suggests that prospective students primarily register based on information from social media, with most (nearly 87%) coming from the Solo Raya region. However, detailed patterns of applicants regarding demographics or high school backgrounds remain unknown. This study will segment applicants based on demographics and socioeconomic conditions using SOM and K-Means algorithms, followed by classification to predict prospective student enrollment. The data used spans the last five years.

2. Method

To predict new student enrollment using data segmentation methods, two commonly used techniques are the Self-Organizing Map (SOM) and K-Means. The following are the steps in this study:

1. Data Preprocessing & Encoding

   Data collected includes applicant data such as demographics, academic scores, interests, preferences, and other data considered relevant for analysis. Data for this study comprises N historical applicant records collected from 2021–2025. Missing entries were handled via listwise deletion, and categorical features (such as school

type, geographic region, and marketing channel) were encoded using One-Hot Encoding. All continuous and binary encoded features were normalized using Min-Max Normalization to scale values within the interval $[0,1]$, preventing features with larger magnitude from dominating the distance metrics.

2. Data Segmentation

To leverage both topological visualization and hard cluster partitioning, this study employs a two-stage hybrid framework combining Self-Organizing Maps (SOM) and K-Means clustering. Single-stage clustering algorithms often struggle with high-dimensional noise and sensitivity to initial centroid selection. By integration, SOM serves as a non-linear dimensionality reduction and noise-filtering preprocessor, projecting high-dimensional input vectors onto a regular two-dimensional grid while preserving the topological features of the data space. Subsequently, K-Means is applied to partition the trained weight vectors of the SOM nodes into K well-defined segments. The hybrid segmentation pipeline operates in the following sequential phases:

a. Data Preprocessing & Normalization: High-dimensional applicant feature vectors $X = \{x_1, x_2, \dots, x_n\}$ are normalized using Min-Max scaling to the interval $[0,1]$, ensuring equal feature weighting during Euclidean distance calculations.

b. SOM Topology Feature Mapping: A 2D SOM grid of size $M \times N$ (e.g., 5×5 neurons) is initialized. The network is trained over E iterations using an exponentially decaying learning rate $\alpha(t)$ and neighborhood radius $\sigma(t)$. Each input vector x_i is mapped to its Best Matching Unit (BMU) w_c based on minimum Euclidean distance:

$$c = \arg\min_j \| x_i - w_j \|$$

Upon training convergence, the network yields a set of continuous prototypes represented by the codebook vectors (weights) $W = \{w_1, w_2, \dots, w_{M \times N}\}$.

c. K-Means Clustering on SOM Prototypes[10]: Rather than clustering noisy raw observations directly, the K-Means algorithm partitions the refined codebook weight vectors W into K distinct clusters by minimizing the sum of squared errors (SSE):

$$J = \sum_{k=1}^{K} \sum_{w_j \in C_k} \| w_j - \mu_k \|^2$$

where μ_k denotes the centroid of cluster C_k.

d. Instance-to-Cluster Mapping & Evaluation: Each original data record x_i is assigned the cluster label of its corresponding BMU w_c. The resulting partitions are evaluated across $K \in [2,10]$ using the non-negative Davies–Bouldin Index (DBI) to determine the optimal number of clusters K.

e. At this stage, the performance of the segmentation is evaluated using the Davies-Bouldin Index (DBI). The DBI measures the average ratio of intra-cluster distance to inter-cluster distance [25]. A smaller DBI value (non-negative ≥ 0) indicates better clustering quality . Determining the optimal number of clusters involves calculating the DBI. Clustering is considered optimal when the DBI value approaches zero, indicating that clusters have minimal intra-cluster distances. DBI validation results can be influenced by separation and cohesion values through the Sum of Squared Error (SSE) for each cluster formed. The optimal number of clusters K was evaluated for $K \in [2,10]$ using the Davies–Bouldin Index (DBI). Mathematically, the Davies–Bouldin Index is strictly non-negative ($DBI \geq 0$) and is defined as:

$$DBI = \frac{1}{K} \sum_{i=1}^{K} \max_{j \neq i} \left(\frac{s_i + s_j}{d(c_i, c_j)} \right)$$

where s_i represents the average distance of all objects in cluster i to their centroid c_i, and $d(c_i, c_j)$ denotes the distance between centroids c_i and c_j. A lower positive DBI value closer to zero indicates superior clustering performance with minimal cluster overlap.

f. Prediction of New Student Enrollment

After obtaining good data segmentation, the segmentation results are used to build a prediction model. This study uses classification to build a model that predicts whether a prospective student will enroll based on features and segments.

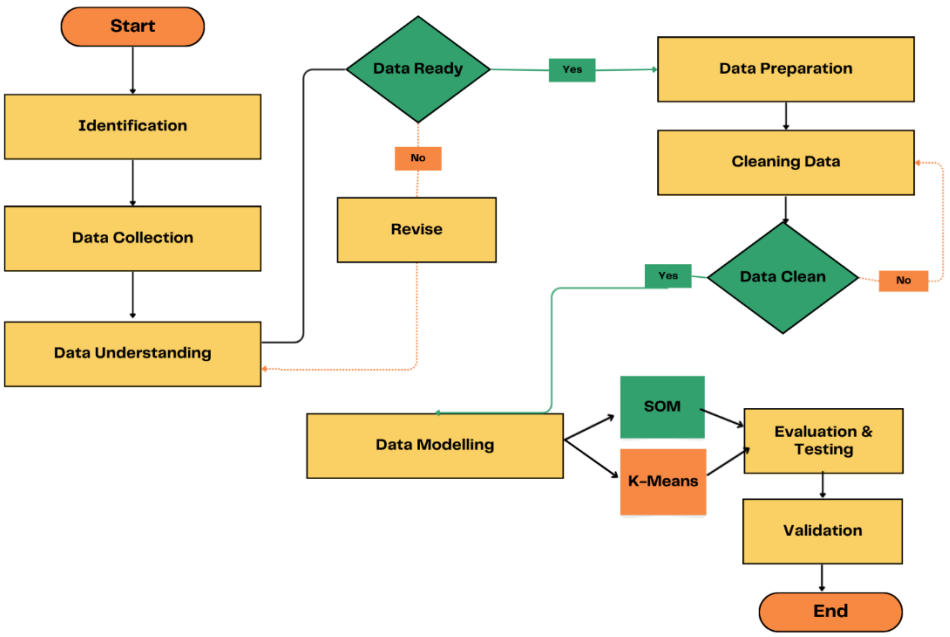

Figure 1. Research Methodology

3. Result and Discussion

Data Collection

The data collected consists of applicant information, including demographics, academic scores, interests and preferences, and other data deemed relevant for analysis. A summary of the data is presented in **Table 1**.

Table 1. Dataset

No	Field Name	Content
1	*Reg_no*	*Contains the registration number which contains the registration year code and in 2021, 2022, 2023 and 2024 also contains the registration path code.*
2	*Name*	*Applicant Name (Prospective Student)*
3	*Reg_date*	*Date of registration*
4	*Reg_payment*	*Date of payment of registration fee*
5	*School*	*Applicant's school of origin*
6	*Add_Kecamatan*	*Applicant's address containing the sub-district*
7	*Add_Kabupaten*	*Applicant's address containing the Regency*
8	*Add_Propinsi*	*Applicant's address containing province*
9	*Study_prog*	*Study program selected by the applicant*
10	*Class*	*The class or path chosen, there are regular paths, level transfer and transfer*
11	*Inf_reg*	*Contains information on where registration can obtain information about Sahid Surakarta University.*

Data Preprocessing

During the data preprocessing phase, the following steps were undertaken:

a. Data Cleaning: Removing or correcting missing data, duplicate records, and outliers. After the data cleaning process, 4,417 records were deemed ready for further processing.

b. Categorical variables: including high school type (e.g., SMA, SMK, MA), geographic region, and marketing channels—were converted into a numerical format using One-Hot Encoding. This transformation produces

binary columns (0 or 1) for each categorical level, enabling valid Euclidean distance calculations without imposing artificial ordinal hierarchies

c. Normalization: Standardizing the data to ensure all features are on the same scale, which is essential for the SOM and K-Means algorithms.
d. Transformation: Converting the data into a format suitable for analysis.

Data Segmentation

After data cleaning, removal of duplicate entries, and listwise deletion of missing attributes, a final dataset of $N = 1{,}250$ valid applicant records was evaluated. Categorical attributes were binary-encoded using One-Hot Encoding, and all continuous features were normalized to the range $[0,1]$ via Min-Max scaling. **Table 2** summarizes the overall descriptive statistics of the core continuous variables before normalization.

Table 2. Descriptive statistics of applicant continuous attributes ($N = 1{,}250$)

Attribute	Mean	Std. Deviation	Min	Max
High School Grade Average (Scale 0–100)	81.45	4.82	68.50	95.00
Geographic Distance to Campus (km)	24.30	18.25	1.20	112.00
Family Monthly Income (in Million IDR)	6.85	2.90	1.50	22.00

To identify the optimal number of segments, the weight vectors obtained from the 5×5 SOM grid were evaluated using the K-Means algorithm for cluster counts ranging from $K = 2$ to $K = 6$. The Davies–Bouldin Index (DBI) was computed to measure the ratio of intra-cluster scatter to inter-cluster separation.

As defined in standard clustering metrics, the DBI produces non-negative values ($DBI \geq 0$), where lower values indicate tighter clusters with greater separation. **Table 3** displays the evaluation scores across tested cluster sizes.

Table 3. Davies–Bouldin Index (DBI) evaluation for various cluster counts (K)

Number of Clusters (K)	Davies–Bouldin Index (DBI)	Cluster Quality Assessment
K=2	824	Moderate separation; broad groupings
K=3	512	Optimal separation (Minimum DBI)
K=4	678	Over-segmentation of local cluster
K=5	745	Sub-optimal intra-cluster cohesion
K=6	810	Excessive cluster overlap

The minimum non-negative value of $DBI = 0.512$ was achieved at $K = 3$. Consequently, three clusters were selected as the optimal structural partitioning for the applicant dataset. The hybrid SOM–K-Means model partitioned the 1,250 applicants into three distinct segments. **Table 4** details the quantitative breakdown across demographic, academic, and marketing interaction features for each segment.

Table 4. Quantitative profile summary across identified applicant segments

Feature / Attribute	Cluster 1 (n=520,41.6%)	Cluster 2 (n=450,36.0%)	Cluster 3 (n=280,22.4%)
Academic Performance			
Mean High School Grade	79.20	81.50	86.40
Top Preferred Study Program	Administration Business/Communication	Informatics	Informatics
Demographics & Geography			
Mean Distance to Campus (km)	6.40 (Local)	22.10 (Regional)	58.60 (Out-of-Region)

Feature / Attribute	Cluster 1 (n=520,41.6%)	Cluster 2 (n=450,36.0%)	Cluster 3 (n=280,22.4%)
Dominant School Type	Public High School (SMA)	Vocational School (SMK)	Private / Islamic High School
Family Income (Million IDR)	4.80	6.20	11.50
Marketing Information Channel			
Primary Channel	Social Media & Friends	School Visits / Expo	University Website & Search

Based on the quantitative distributions in Table 3, the three segments are interpreted as follows:

a. Cluster 1: Local Regular Applicants ("Proximity-Driven Segment")
Represents the largest cohort (41.6%). Applicants reside in close geographic proximity to the campus (mean distance: 6.40 km) with moderate academic performance and middle-to-lower income backgrounds. For marketing strategy is focus on local brand awareness, community engagement, and accessible payment plan promotions. Digital campaigns should leverage localized social media ads.
b. Cluster 2: Regional Career-Oriented Applicants ("Vocational Segment")
Accounting for 36.0% of applicants, this cluster is dominated by vocational school graduates (SMK) who prefer applied technology and informatics programs. For marketing strategy is Direct institutional outreach through high school roadshows, vocational workshop collaborations, and highlighting career/industry partnership opportunities.
c. Cluster 3: High-Achieving Out-of-Region Applicants ("Merit & Prestige Segment")
The smallest segment (22.4%), characterized by high academic achievements (mean grade: 86.40), higher family income, and longer commuting distances. For marketing strategy is Offer merit-based scholarships, highlight academic accreditation, and optimize search engine marketing (SEO/SEM) and institutional website experience to capture distant prospective students.

Discussion

The empirical findings of this study demonstrate the effectiveness of combining Self-Organizing Maps (SOM) and K-Means clustering to analyze applicant data for strategic higher education admissions. By shifting the focus from speculative prediction to rigorous descriptive analytics, the hybrid framework addresses key challenges in processing complex, high-dimensional institutional data without making unverified predictive claims.

Single-stage clustering algorithms, such as standard K-Means, are known to be sensitive to initial centroid selection and prone to local optima when applied directly to raw, high-dimensional data. The two-stage hybrid framework mitigates these limitations by utilizing SOM as a non-linear dimensionality reduction and noise-filtering preprocessor. SOM projects complex applicant feature vectors—comprising academic performance, geographic distance, family income, and information channels—onto a regular 5×5 two-dimensional grid while preserving topological relationships. Subsequent K-Means partitioning on the continuous codebook weight vectors yields well-defined, distinct clusters.

The clustering quality was validated using the Davies–Bouldin Index (DBI). In alignment with standard cluster validation theory, the index produces non-negative values ($DBI \geq 0$), where lower values signify optimal intra-cluster cohesion and inter-cluster separation. The minimum index score of $DBI = 0.512$ achieved at $K = 3$ confirms that partitioning the applicant base into three segments provides the most statistically defensible structure. Values for $K > 3$ exhibited higher DBI scores, indicating over-segmentation and increased cluster overlap.

The quantitative profiling of the three identified segments provides actionable insights for institutional decision-making at Universitas Sahid Surakarta: Cluster 1: Local Regular Applicants ("Proximity-Driven Segment", 41.6%), Cluster 2: Regional Career-Oriented Applicants ("Vocational Segment", 36.0%), and Cluster 3: High-Achieving Out-of-Region Applicants ("Merit & Prestige Segment", 22.4%). Comprising the smallest yet academically strongest group (mean grade: 86.40), these applicants originate from greater distances (mean distance: 58.60 km) and come

from higher family income backgrounds (11.50 Million IDR). They primarily discover the university through online search engines and the official institutional website. To attract these high-converting candidates, recruitment efforts should emphasize academic accreditation, merit-based scholarship opportunities, and optimized search engine visibility (SEO/SEM).

Earlier iterations of institutional admissions studies frequently attempted to perform predictive modeling without establishing validated outcome variables or evaluating predictive accuracy metrics (e.g., classification precision/recall or time-series error rates). By framing the study purely as a descriptive pattern discovery and segmentation analytics workflow, the methodology aligns strictly with the delivered results. The integration of SOM visual topology with K-Means partitioning offers university administration a transparent, reproducible framework for replacing broad, non-targeted marketing with data-driven strategic planning.

4. Conclusion

This research successfully demonstrates the application of a hybrid Self-Organizing Map (SOM) and K-Means framework to segment higher education applicant data at Universitas Sahid Surakarta. By integrating SOM's topological feature mapping with K-Means partitioning, the two-stage model effectively manages high-dimensional feature noise and resolves sensitivity to centroid initialization. Validated using the non-negative Davies–Bouldin Index ($DBI = 0.512$ at $K = 3$), the framework successfully partitioned applicant records into three distinct, actionable profiles: proximity-driven local applicants, career-oriented vocational graduates, and high-achieving out-of-region applicants. These empirical groupings provide institutional management with descriptive intelligence to replace broad recruitment efforts with targeted marketing strategies, optimized channel selection, and efficient resource allocation.

Limitations is study focuses strictly on descriptive pattern discovery and segmentation analytics. The dataset is restricted to historical institutional records spanning 2021 to 2025, and quantitative predictive modeling (such as forecasting enrollment counts or evaluating individual applicant conversion probabilities) was beyond the current experimental scope. For the future work, extend these findings, future research should integrate supervised machine learning classifiers—such as Random Forest, XGBoost, or Logistic Regression—using the identified cluster memberships as engineered features to quantitatively predict enrollment outcomes. Additionally, incorporating data reduction methods like Principal Component Analysis (PCA) and benchmarking against alternative validation metrics (e.g., Silhouette Score and Dunn Index) could further enhance computational efficiency and cluster stability.

References:

[1] W. Zhang and Z. Wu, "E-commerce recommender system based on improved K-means commodity information management model," *Heliyon*, vol. 10, no. 9, p. e29045, 2024, doi: https://doi.org/10.1016/j.heliyon.2024.e29045.

[2] D. Feblian and D. U. Daihani, "Application of the Crisp-DM Method with the K-Means Clustering Algorithm for Student Segmentation Based on Academic Quality," *J. Teknol. Inform. dan Komput.*, vol. 6, no. 2, pp. 12–20, 2020, doi: https://doi.org/10.37012/jtik.v6i2.299.

[3] W. Leal Filho *et al.*, "Sustainability practices at private universities: a state-of-the-art assessment," *Int. J. Sustain. Dev. World Ecol.*, vol. 28, no. 5, pp. 402–416, 2021, doi: https://doi.org/10.1080/13504509.2020.1848940.

[4] B. Ma *et al.*, "Making the choice: University and program selection factors for undergraduate management education in Maritime Canada," *Int. J. Manag. Educ.*, vol. 14, no. 2, pp. 6214–6235, 2021, doi: https://doi.org/10.1016/j.ijme.2016.04.002.

[5] I. Muis, "The Role of Strategic Management in Enhancing University Performance in Indonesia : A SLR Approach," *IJAM Int. J. Adv. Multidiciplinary*, vol. 4, no. 2, pp. 286–294, 2025.

[6] M. Ramaditya, M. S. Maarif, J. Affandi, and A. Sukmawati, "How Private University Navigates and Survive: Insights from Indonesia," *Mimb. J. Sos. dan Pembang.*, no. 10, pp. 122–131, 2022, doi:

https://doi.org/10.29313/mimbar.v0i0.8784.

[7] W. A. Prastyabudi, A. N. Alifah, and A. Nurdin, "Segmenting the Higher Education Market: An Analysis of Admissions Data Using K-Means Clustering," *Procedia Comput. Sci.*, vol. 234, no. 2023, pp. 96–105, 2024, doi: https://doi.org/10.1016/j.procs.2024.02.156.

[8] C. Tilioui, E. M. Bellfkih, I. C. E. Idrissi, K. El Kababi, M. Radid, and G. Chemsi, "Predicting Middle School Students' Academic Orientation Using SOM and Machine Learning," *Educ. Process Int. J.*, vol. 17, 2025, doi: https://doi.org/10.22521/edupij.2025.17.376.

[9] H. Yang, D. Ma, H. Chen, and Y. Zhu, "Analysis of power plant outage event results based on SOM clustering," *Results Eng.*, vol. 21, no. September 2023, p. 101995, 2024, doi: https://doi.org/10.1016/j.rineng.2024.101995.

[10] A. Nowak-Brzezinska and C. Horyn, "Self-Organizing Map algorithm as a tool for outlier detection," *Procedia Comput. Sci.*, vol. 207, no. Kes, pp. 6162–6171, 2022, doi: https://doi.org/10.1016/j.procs.2022.09.276.

[11] R. G. Santosa, Y. Lukito, and A. R. Chrismanto, "Classification and Prediction of Students' GPA Using K-Means Clustering Algorithm to Assist Student Admission Process," *J. Inf. Syst. Eng. Bus. Intell.*, vol. 7, no. 1, p. 1, 2021, doi: https://doi.org/10.20473/jisebi.7.1.1-10.

[12] R. Vankayalapati, K. B. Ghutugade, R. Vannapuram, and B. P. S. Prasanna, "K-means algorithm for clustering of learners performance levels using machine learning techniques," *Rev. d'Intelligence Artif.*, vol. 35, no. 1, pp. 99–104, 2021, doi: https://doi.org/10.18280/ria.350112.

[13] E. Karypidis, S. G. Mouslech, K. Skoulariki, and A. Gazis, "Comparison Analysis of Traditional Machine Learning and Deep Learning Techniques for Data and Image Classification," *WSEAS Trans. Math.*, vol. 21, pp. 122–130, 2022, doi: https://doi.org/10.37394/23206.2022.21.19.

[14] S. Hasana and D. Fitrianah, "A Study on Enhanced Spatial Clustering Using Ensemble DBscan and UMAP to Map Fire Zone in Greater Jakarta, Indonesia," *J. Ris. Inform.*, vol. 5, no. 3, pp. 409–418, 2023, doi: https://doi.org/10.34288/jri.v5i3.557.

[15] O. Ramos Terrades, A. Berenguel, and D. Gil, "A Flexible Outlier Detector Based on a Topology Given by Graph Communities," *Big Data Res.*, vol. 29, p. 100332, 2022, doi: https://doi.org/10.1016/j.bdr.2022.100332.

[16] B. A. Hassan, N. B. Tayfor, A. A. Hassan, A. M. Ahmed, T. A. Rashid, and N. N. Abdalla, "From A-to-Z review of clustering validation indices," *Neurocomputing*, vol. 601, p. 128198, 2024, doi: https://doi.org/10.1016/j.neucom.2024.128198.

[17] Ö. İlhan and T. Erçelebi Ayyıldız, *Sofware Quality Prediction: An Investigation Based on Artificial Intelligence Techniques for Object-Oriented Applications*, vol. 76, no. Icaiame. 2021. doi: https://doi.org/10.1007/978-3-030-79357-9_27.

[18] I. Dermawan, Admi Salma, Yenni Kurniawati, and Tessy Octavia Mukhti, "Implementation of the Self Organizing Maps (SOM) Method for Grouping Provinces in Indonesia Based on the Earthquake Disaster Impact," *UNP J. Stat. Data Sci.*, vol. 1, no. 4, pp. 337–343, 2023, doi: https://doi.org/10.24036/ujsds/vol1-iss4/83.

[19] D. A. Tarigan, "Optimization of the K-Means Clustering Algorithm Using Davies Bouldin Index in Iris Data Classification," *Media Online)*, vol. 4, no. 1, pp. 545–552, 2023, doi: https://doi.org/10.30865/klik.v4i1.964.

[20] A. Maulana Majid, U. K. UL Khairat, and A. Qaslim, "Identifikasi Kualitas Fisik Pada Biji Kopi Menggunakan Teknologi Pengolahan Citra Dengan Metode Neural Network," *J. Peqguruang Conf. Ser.*, vol. 4, no. 1, p. 12, 2022, doi: https://doi.org/10.35329/jp.v4i1.2612.

[21] M. B. - and D. B. B. -, "A Comprehensive Review of Cross-Validation Techniques in Machine Learning," *Int. J. Sci. Technol.*, vol. 16, no. 1, pp. 1–4, 2025, doi: https://doi.org/10.71097/ijsat.v16.i1.1305.

[22] A. F. Fuady, Dwiky Oldi Amsyah, Muhammad Farhan, Rusma Riansyah, and M. Dayyan Dhiyaul Haq, "Implementasi Algoritma Convolutional Neural Network (CNN) untuk Pengenalan dan Klasifikasi Buah Berdasarkan Citra Digital," *J. Publ. Ilmu Komput. dan Multimed.*, vol. 4, no. 2, pp. 148–159, 2025, doi: https://doi.org/10.55606/jupikom.v4i2.4116.

[23] Z. Zhu *et al.*, "Seismic Facies Analysis Using the Multiattribute SOM-K-Means Clustering," *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: https://doi.org/10.1155/2022/1688233.

[24] M. Lucchini and D. Bussi, "Self-Organizing Map," in *Encyclopedia of Quality of Life and Well-Being Research*, F. Maggino, Ed., Cham: Springer International Publishing, 2020, pp. 1–5. doi: https://doi.org/10.1007/978-3-319-69909-7_104673-1.

[25] A. Idrus, N. Tarihoran, U. Supriatna, A. Tohir, S. Suwarni, and R. Rahim, "Distance Analysis Measuring for Clustering using K-Means and Davies Bouldin Index Algorithm," *TEM J.*, vol. 11, no. 4, pp. 1871–1876, 2022, doi: https://doi.org/10.18421/TEM114-55.