RESEARCH ARTICLE • OPEN ACCESS

Identifying Digital Literacy Profiles in Distance Education: A K-Prototypes Clustering Approach

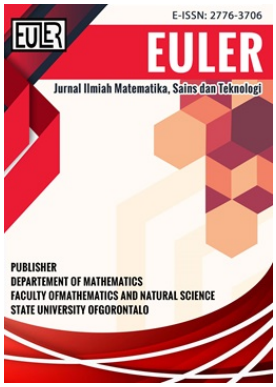
Arman Haqqi Anna Zili, Made Diyah Putri Martinasari, and Selly Anastassia Amellia Kharis

Volume 13, Issue 3, pp. 346–351, Dec. 2025

Received 25 September 2025, Revised 8 November 2025, Accepted 15 November 2025, Published 1 December 2025

To Cite this Article : A. H. A. Zili, M. D. P. Martinasari, and S. A. A. Kharis,"Identifying Digital Literacy Profiles in Distance Education: A K-Prototypes Clustering Approach", *Euler J. Ilm. Mat. Sains dan Teknol.*, vol. 13, no. 3, pp. 346–351, 2025, *https://doi.org/10.37905/euler.v13i3.34568*

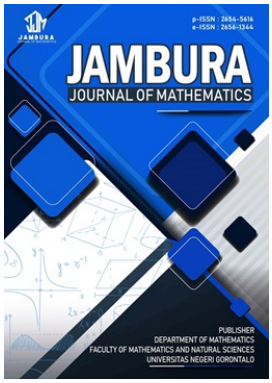
JOURNAL INFO • EULER : JURNAL ILMIAH MATEMATIKA, SAINS DAN TEKNOLOGI



Homepage	:	http://ejurnal.ung.ac.id/index.php/euler/index
Journal Abbreviation	:	Euler J. Ilm. Mat. Sains dan Teknol.
Frequency	:	Three times a year
Publication Language	:	English (preferable), Indonesia
DOI	:	https://doi.org/10.37905/euler
Online ISSN	:	2776-3706
Publisher	:	Department of Mathematics, Universitas Negeri Gorontalo
Country	:	Indonesia
OAI Address	:	http://ejurnal.ung.ac.id/index.php/euler/oai
Google Scholar ID	:	QF_r-_gAAAAJ
Email	:	euler@ung.ac.id

JAMBURA JOURNAL • FIND OUR OTHER JOURNALS

Jambura Journal of Biomathematics



Jambura Journal of Mathematics

Jambura Journal of Mathematics Education

Jambura Journal of Probability and Statistics

Euler : Jurnal Ilmiah Matematika, Sains dan Teknologi, Volume 13, Issue 3, Pages 346–351, Dec. 2025 *https://doi.org/10.37905/euler.v13i3.34568*

Research Article

Check for updates

Identifying Digital Literacy Profiles in Distance Education: A K-Prototypes Clustering Approach

Arman Haqqi Anna Zili, Made Diyah Putri[1] Martinasari[2], Selly Anastassia Amellia Kharis[2,*]

[1]*Department of Mathematics, Universitas Indonesia, Depok 16424, Indonesia*
[2]*Study Program of Mathematics, Universitas Terbuka, Tangerang Selatan 15418, Indonesia*

ARTICLE HISTORY
Received 25 September 2025
Revised 8 November 2025
Accepted 15 November 2025
Published 1 December 2025

KEYWORDS
Cluster Analysis
K-Prototypes
Machine Learning
Behavioural Segmentation

ABSTRACT. *Education quality is one of the main focuses of Indonesia's Sustainable Development Goals (SDGs), particularly in the goal that emphasizes equitable access and lifelong learning. Universitas Terbuka (UT) is a higher education institution that implements an open and distance learning system. This setting creates a diverse student body in terms of age, occupation, and digital literacy levels. Segmenting students based on their digital literacy is both essential and challenging, as it involves combining demographic data with daily digital behavior. This study aims to identify the digital literacy profiles of UT students using cluster analysis with the K-Prototypes algorithm. Data were obtained from a survey of 10,396 students with 42 variables. The Elbow Method analysis revealed three distinct clusters, each reflecting unique engagement profiles. The first cluster, the Engaged Evening Digital User, is active during the evening and balances work with social activities. The second cluster, the Hyper Connected Communicator, relies heavily on messaging applications for social interaction. The third cluster, the Balanced Digital Citizen, shows a more even distribution of digital use across academic, entertainment, and communication activities. These clusters predominantly comprise Generation Z individuals, many of whom are actively engaged in the private sector. The profound implications of these findings lie in their capacity to forge highly targeted strategies for digital learning, communication, and student support, thereby enhancing educational outcomes. Furthermore, this research significantly advances methodological literature by demonstrating a powerful, integrated approach to clustering mixed-type attributes, offering a more nuanced understanding of learner profiles in distance education.*

1. Introduction

Education quality is a central concern in the Sustainable Development Goals of Indonesia, particularly Goal 4, which emphasizes equitable access and lifelong learning opportunities [1]. In Indonesia, achieving this requires innovative approaches in both policy and practice. Universitas Terbuka (UT), the country's leading provider of open and distance learning, has played a significant role through digital innovations and flexible delivery modes. Yet, the success of distance education still depends heavily on students' ability to engage with learning materials independently [2].

The transition from the Fourth to the Fifth Industrial Revolution highlights the growing importance of digital literacy for students. It is not only a technical skill but also a determinant of learning success [3]. Several studies in Indonesia have examined digital literacy in relation to identity formation [4], media use [5], learning outcomes [6], and learning strategies within the framework of Society 5.0 [7]. At the same time, recent surveys in Indonesia indicate that the national digital literacy index still needs to be improved [8, 9]. For UT students, who are expected to regulate their own learning, identifying patterns of digital literacy is particularly important. Understanding these patterns can guide the development of targeted interventions that address the needs of different student groups.

Previous studies have emphasized that digital literacy is closely linked to students' readiness, motivation, and success in learning environments. Huba et al. [10] found that students at Universitas Tanjungpura possessed strong digital competencies, while Fadillah et al. [11] showed that students demonstrated an average score of 75 in digital literacy within mathematics courses. Furthermore, Akbar and Anggaraeni [12] confirmed a significant relationship between digital literacy and self-directed learning among thesis-writing students. These studies suggest that enhancing digital literacy not only supports general academic success but is also crucial in enabling students to complete their undergraduate programs effectively. For UT students, who rely heavily on self-regulated learning, identifying digital literacy clusters is essential to inform institutional strategies that support personalized interventions.

Cluster analysis provides a powerful approach to segmenting students based on their digital literacy levels. One of the most effective algorithms for categorical data is the k-modes algorithm, first introduced by Huang [13]. Unlike k-means, which is designed for numerical data, k-modes replaces means with modes and applies a frequency-based updating mechanism to minimize clustering costs [14]. Subsequent improvement have

*Corresponding Author.

Email : *selly@ecampus.ut.ac.id* (S. A. A. Kharis)
Homepage : http://ejurnal.ung.ac.id/index.php/euler/index / E-ISSN : 2776-3706

been made, such as weighted dissimilarity measures [15]. These enhancement underline the adaptability and robustness of k-modes algorithm in addressing categorical data clustering challenges.

The k-modes algorithm has been applied across diverse domains. In Indonesia, it has been used to classify fire cases in Jakarta [16], group elderly populations in South Sumatra [17], identify game character roles in Wild Rift [18], analyze hotel review sentiments [19], and cluster socioeconomic data in rural villages [20]. These examples demonstrate the versatility of k-modes in handling categorical data across public safety, demography, entertainment, sentiment analysis, and socioeconomic research.

Globally, researchers have also advanced the theoretical underpinnings and applications of categorical clustering. Ketchen and Shook [21] critically assessed the use of cluster analysis in management studies, highlighting methodological rigor. Building upon foundational work by Huang, various improvements and extensions to the k-modes algorithm have been proposed. For instance, Dorman and Maitra [22] introduced a computationally efficient k-modes implementation, while Xie et al. [23] discussed self-tuning approaches. Dinh et al. [24] provided a comprehensive synthesis of categorical data clustering, situating k-modes within the broader landscape of techniques. Moreover, the integration of metaheuristic approaches has led to significant advancements; Gan et al. [25] demonstrated the benefits of incorporating genetic algorithms into k-modes optimization, and the development of fuzzy k-modes [26], and further explored by Ng & Jing [27] and Oskouei et al. [28] has improved flexibility in handling overlapping clusters.

The selection of an appropriate clustering algorithm is critically dependent on the nature of the data being analyzed. For instance, the foundational K-Means algorithm is a widely used method designed specifically for numerical data, partitioning observations by minimizing the squared Euclidean distances to a cluster's mean, or centroid. In contrast, when dealing with categorical data, an adaptation called K-Modes is more suitable [29]. This algorithm replaces numerical means with modes (the most frequent attribute values) and utilizes a simple matching dissimilarity metric to handle non-numeric features.

While the k-modes algorithm effectively handles categorical attributes, many real-worls datasets, including those describing students' digital literacy, consist of both numerical and categorical variables. To address this, the K-Prototypes algorithm extends k-means and k-modes by combining numerical and categorical distance measures within a single framework. This hybrid approach allows for more comprehensive clustering of mixed-type data, making it particularly suitable for educational datasets that capture both quantitative indicators and qualitative attributes preferred digital tools or self-assessed competencies.

The fundamental goal is to group objects such that those within the same cluster are highly similar, while objects in different clusters are dissimilar. The dataset for this research, which aims to segment students based on their digital literacy profiles, is composed of mixed data types, containing both numerical and categorical variables. The presence of these heterogeneous attributes renders algorithms like K-Means or K-Modes unsuitable in isolation. To address this challenge, the K-Prototypes algorithm was selected as the primary analytical method. This hybrid algorithm effectively integrates the principles of both K-Means and K-Modes, employing a combined dissimilarity measure that can handle heterogeneous data simultaneously [30]. Therefore, the K-Prototypes algorithm provides a robust framework for identifying meaningful and accurate student segments within our mixed-attribute dataset.

The K-Prototypes algorithm, introduced by Huang, combines the K-Means and K-Modes algorithms, applying the squared Euclidean distance for numeric attributes and a simple matching dissimilarity for categorical attributes [31]. The algorithm was chosen in this study because it can handle mixed data types by combining Euclidean distance for numerical variables and simple matching dissimilarity for categorical variables. This approach has become a popular choice for segmentation tasks in fields such as customer profiling, healthcare, and behavioral studies [32–34].

This body of research emphasizes that clustering, particularly using k-modes, offers a reliable framework for addressing categorical data. For the context of UT students' digital literacy, clustering provides actionable insights for policymakers. By identifying groups of students with similar digital competencies, institutions can design tailored interventions, ensuring equitable opportunities for students to thrive in open and distance learning environments.

2. Methods

This study employs a quantitative approach using cluster analysis, a statistical technique designed to partition a dataset into distinct groups, or clusters. The objective function for the K-Prototypes algorithm is to minimize the cost function, which is defined as:

$$E = \sum_{l=1}^{k} \sum_{i=1}^{n} w_{i,l}\, d(X_i, Q_l), \tag{1}$$

where:

- E is the sum of dissimilarities of all objects to their cluster prototypes.
- k is the number of clusters.
- n is the number of objects.
- $w_{i,l}$ is an element of the partition matrix, indicating whether object X_i belongs to cluster l.
- $d(X_i, Q_l)$ is the dissimilarity measure between object X_i and prototype Q_l.

The dissimilarity measure between object X_i and prototype Q_l for mixed data is defined as:

$$d(X_i, Q_l) = \sum_{j=1}^{p} \left(x_{i,j}^{(r)} - q_{l,j}^{(r)}\right)^2 + \gamma_l \sum_{j=p+1}^{m} \delta\left(x_{i,j}^{(c)},\, q_{l,j}^{(c)}\right) \tag{2}$$

where

- The first term is the squared Euclidean distance for the p numeric attributes.
- The second term is the simple matching dissimilarity for the $m-p$ categorical attributes, where $\delta(a,b) = 0$ if $a = b$ and $\delta(a,b) = 1$ if $a \neq b$.
- γ_l is a weight to balance the contribution of categorical attributes.

A crucial step in cluster analysis is determining the optimal number of clusters, k number [35]. In this study, we utilized the Elbow Method, a widely used heuristic for this purpose. The method involves running the clustering algorithm for a range of k values and plotting the cost function (or within-cluster sum of squares) for each k. The plot typically shows the cost decreasing as k increases. The "elbow" on the curve represents the point of diminishing returns, where adding more clusters does not significantly reduce the cost [36]. This point is considered to be the optimal number of clusters.

To illustrate the steps of the research in this study, the methodological process is summarized step by step in the Figure 1.

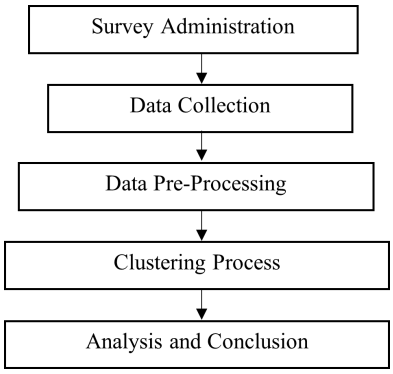


Figure 1. Sequential steps in research methodology

The research was conducted following a systematic procedure. The steps included:

1. Survey Administration: Development of survey tools.
2. Data Collection: A survey was administered to students of Universitas Terbuka to collect data on their digital literacy and demographic characteristics.
3. Data Pre-processing: The collected data was cleaned to handle missing values and inconsistencies. Categorical variables were converted to the appropriate data type for the K-Prototypes algorithm.
4. Clustering: The K-Prototypes algorithm was implemented in Python. The optimal number of clusters was determined using the Elbow Method.
5. Analysis and Interpretation: The resulting clusters were analyzed to understand the characteristics of each group. This involved examining the cluster prototypes and the distribution of variables within each cluster.
6. Conclusion: Conclusions were drawn based on the analysis of the clusters, providing insights into the different digital literacy profiles of the students.

3. Results and Discussion

This section presents the outcomes of the six methodological steps outlined in the research method: survey administration, data collection, data pre-processing, clustering, analysis and interpretation, and conclusion formulation. The survey instrument was developed based on established digital literacy frameworks, encompassing domains such as digital competencies, information ethics, and online behaviour. The questionnaire was designed to capture both categorical and numerical variables to enable mixed-type data analysis. The dataset utilized in this study is composed of survey responses collected to evaluate the digital literacy profiles of students. The data captures a wide array of student attributes, comprising a total of 42 variables across 10,396 student responses. This rich dataset contains a mix of both categorical (9 variables) and numerical data (33 variables), making it well-suited for a hybrid analytical approach. The variables can be broadly categorized into several key areas:

1. Demographics: This includes background information such as the student's regional location (UT Daerah), generational cohort (Profile Generation), and employment status (Job Profile).
2. Digital Behavior: This category details student habits related to internet and social media usage. It includes variables such as the typical time of day for internet access (Internet access hours), the primary online activities performed, the most frequently used social media platforms, and the average daily time spent on these platforms.
3. Digital Competencies: A significant portion of the dataset consists of self-reported skill levels across various digital tasks and platforms. These are captured using a Likert-type scale, where students rated their ability to use specific distance learning tools (Microsoft Teams, elearning.ut.ac.id), access digital libraries, manage digital information, and utilize online services.
4. Information Literacy and Ethics: The data also includes responses related to students' practices in evaluating and comparing information sources and their understanding of academic integrity in a digital context.
5. Overall Literacy Level: The dataset concludes with a pre-calculated categorical variable that classifies each student into a specific digital literacy level, such as 'Satisfactory', 'High', or 'Very High'.

Given the heterogeneous nature of the data, which combines descriptive categories with ordinal skill ratings, it provides a comprehensive foundation for segmenting students into distinct digital literacy profiles.

After cleaning the data, next step is determine the k value, which is the number of cluster. To determine the optimal number of clusters, we used the Elbow Method using several k values as illustrated in Figure 2.

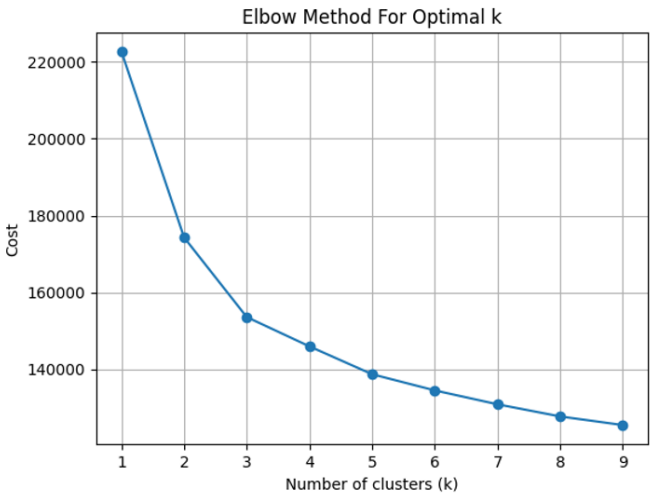


Figure 2. Cost values using number of cluster $k = 1$ to $k = 9$

The cost function decreased sharply between $k = 1$ and $k = 3$ and then showed only gradual improvement for higher k values. This "elbow" pattern suggested that $k = 3$ was the most appropriate choice [36]. Selecting $k = 3$ allowed us to capture key behavioral differences without overfitting or creating overly fragmented groups [37].

3.1. Cluster Characterization

Using $k = 3$ means we produced three distinct groups: Cluster 1 $(n = 3,198)$, Cluster 2 $(n = 3,337)$, and Cluster 3 $(n = 3,861)$. We examined descriptive statistics and frequency distributions to interpret the profiles of each cluster.

3.1.1. Cluster 1: The Engaged Evening Digital User

Cluster 1 showed moderate values for most numerical features (mean: 2.82). A majority of respondents were Generation Z (57%), and many worked as private employees (42%). Their online activity peaked between 18:01 and midnight (52%), and instant messaging dominated their digital habits, with WhatsApp as the preferred platform (53%). Most reported screen time between 3–5 hours daily (46%). Taken together, these patterns suggest a group of young professionals who use digital platforms intensively during leisure hours, perhaps balancing work responsibilities with evening social interaction.

3.1.2. Cluster 2: The Hyper-Connected Communicator

Cluster 2 stood out as the most communication-oriented group. Its numerical averages were slightly lower (mean: 2.11), but the share of Generation Z was even higher (60%). This group had the strongest reliance on instant messaging (67%) and WhatsApp (62%), which indicates that these platforms function primarily as social lifelines. One interpretation is that this cluster represents students who maintain frequent digital contact with peers, potentially using messaging apps as their main means of staying socially connected.

3.1.3. Cluster 3: The Balanced Digital Citizen

Cluster 3 was the largest and most "balanced" group, with characteristics similar to Cluster 1 (mean: 2.85). Generation Z: 55%, private employees: 41%). Evening activity (47%) and WhatsApp usage (58%) were still high, but this group appeared to have a more even distribution of activities across learning, socializing, and entertainment. This could mean that these students are better at integrating online resources into both academic and personal life. A deeper feature-level comparison might help clarify whether their balance reflects deliberate time management or simply different lifestyle patterns.

Across all three clusters, a consistent picture emerges: most students are young, digitally active in the evenings, and rely heavily on instant messaging, particularly WhatsApp, for communication. These results align with global trends showing increased digital immersion among younger generations. These findings have important implications for educators and policymakers. Since students are most active online during the evening, this period may be the most effective window for delivering learning content, announcements, or interactive activities. Leveraging widely used platforms such as WhatsApp could also improve participation and encourage peer collaboration.

Moreover, the K-Prototypes algorithm captured subtle behavioral differences, especially in the numerical dimensions, allowing us to separate the "hyper-connected communicator" subgroup from the more balanced digital users. This demonstrates the value of mixed-data clustering for uncovering patterns that might be missed by analyzing only categorical or numerical data [38]. These insights can inform targeted interventions, encouraging communication-focused students to broaden their digital engagement beyond messaging and to adopt more productive learning behaviors. Similar findings were report that K-Prototypes has become a widely used and powerful technique in the context of Educational Data Mining (EDM) [39]. The article notes that K-Prototypes achieve a balance between computational efficiency and interpretability, making it more practical than density-based on hierarchical approaches for large-scale educational datasets.

3.2. Limitations

While the findings are encouraging, several limitations should be considered. First, the analysis relied on self-reported survey data, which may be subject to response bias or inaccuracies in recall (for example, reported screen time may differ from actual usage). Second, the clustering results depend on the selected features and preprocessing choices; using additional variables or alternative dissimilarity measures might yield slightly different groupings. Third, the dataset represents a cross-sectional snapshot, so it does not capture how digital behaviors may change over time or across academic terms. Future studies could use longitudinal data and combine clustering with qualitative interviews to explore the reasons behind the observed patterns in greater depth.

3.3. Practical Implications

The clustering results provide actionable insights for educators and administrators at Universitas Terbuka and similar institutions. Knowing that students are most digitally active during the evening suggests that online tutorials, announcements, or interactive activities could achieve higher participation rates if scheduled later in the day. The strong preference for WhatsApp and instant messaging platforms highlights an opportunity to integrate these channels into official communication strategies, for instance, by using them to share reminders, learning tips, or peer collaboration opportunities. Moreover, the identification of a highly communication-focused group (Cluster 2) suggests that targeted interventions could encourage these students to use digital tools for academic purposes, potentially improving their learning outcomes. Finally, recognizing a "balanced" cluster provides a model for designing programs that help students manage their time online more effectively, promoting healthy digital habits.

4. Conclusion

This study demonstrates that the K-Prototypes algorithm is a practical approach for segmenting datasets with mixed data types and offers valuable insights into the digital habits of Universitas Terbuka students. The analysis highlights the strong presence of Generation Z, their preference for instant messaging platforms, especially WhatsApp, and their peak online activity dur-

ing the evening. These findings can guide the development of targeted communication strategies, platform improvements, and initiatives to support students' digital literacy. Future research could examine more detailed behavioral drivers behind the differences between Cluster 1 and 3, perhaps using additional features or qualitative follow-up studies. Employing validation measures such as the silhouette score or gap statistic would also strengthen the robustness of the results. Longitudinal analyses could further reveal whether these digital engagement patterns remain consistent over time.

Author Contributions. **Arman Haqqi Anna Zili**: Conceptualization, methodology, programming, validation, analysis, data processing, writing, editing, visualization, and supervision. **Made Diyah Putri Martinasari**: Conceptualization, methodology, validation, writing, and editing. **Selly Anastassia Amellia Kharis**: Conceptualization, methodology, validation, writing, editing, and supervision. All authors discussed the results and contributed to the final manuscript.

Acknowledgement. The author would like to express sincere gratitude to the Lembaga Penelitian dan Pengabdian kepada Masyarakat Universitas Terbuka (LPPM UT) for their support and facilities provided throughout the implementation of this research.

Funding. This research was financially supported by Lembaga Penelitian dan Pengabdian kepada Masyarakat Universitas Terbuka (LPPM UT).

Conflict of interest. The authors declare that there is no conflict of interest related to this article.

Data availability. Not applicable.

References

[1] Badan Perencanaan Pembangunan Nasional, "SDGs Knowledge Hub," 2025. [Online]. Available: https://sdgs.bappenas.go.id/17-goals/goal-4/

[2] S. A. A. Kharis, N. Mahin, H. Lubis, A. H. A. Zili, and A. Robiansyah, "Kecemasan Matematika dan Permasalahannya dalam Pembelajaran Jarak Jauh," *EDUKATIF: Jurnal Ilmu Pendidikan*, vol. 5, no. 1, pp. 508–518, Mar. 2023, doi: 10.31004/edukatif.v5i1.4735.

[3] R. Setyaningsih, E. Prihantoro, U. Darussalam Gontor, U. Gunadarma, and J. Raya Siman, "Model Penguatan Literasi Digital Melalui Pemanfaatan E-learning," *Jurnal ASPIKOM*, vol. 3, no. 6, pp. 1200–1214, 2019.

[4] F. Nurfauziyanti and F. Alwan Bahrudin, "Pengaruh Literasi Digital Terhadap Perkembangan Wawasan Kebangsaan Mahasiswa," *Jurnal Pendidikan Kewarganegaraan Undiksha*, vol. 10, no. 3, 2022. [Online]. Available: https://ejournal.undiksha.ac.id/index.php/JJPP

[5] S. Jan, "Investigating the relationship between student digital literacy and their attitude towards using ICT," *International Journal of Educational Technology*, vol. 5, no. 2, pp. 26–34, 2018. [Online]. Available: http://educationaltechnology.net/ijet/

[6] W. W. W. Brata, R. Y. Padang, C. Suriani, E. Prasetya, and N. Pratiwi, "Student's Digital Literacy Based on Students' Interest in Digital Technology, Internet Costs, Gender, and Learning Outcomes," *Int. J. Emerg. Technol. Learn.*, vol. 17, no. 3, pp. 138–151, 2022, doi: 10.3991/ijet.v17i03.27151.

[7] I. Wati, M. Ernita, Ristiliana, and M. I. Lubis, "Peran Literasi Digital dalam Pembelajaran di Era Society 5.0 pada Mahasiswa Pendidikan Ekonomi UIN Suska Riau," *EKLEKTIK*, vol. 6, no. 1, pp. 21–35, 2023.

[8] B. Bulya and S. Izzati, "Indonesia's Digital Literacy as a Challenge for Democracy in the Digital Age," *J. Society Media*, vol. 8, no. 2, pp. 640–661, Oct. 2024, doi: 10.26740/jsm.v8n2.p640-661.

[9] I. M. Mujtahid, M. Berlian, R. Vebrianto, M. Thahir, and D. Irawan, "The Development of Digital Age Literacy: A Case Study in Indonesia," *J. Asian Finance Econ. Bus.*, vol. 8, no. 2, pp. 1169–1179, 2021, doi: 10.13106/jafeb.2021.vol8.no2.1169.

[10] N. H. Huba et al., "Analisis Kemampuan Literasi Digital Mahasiswa," *Jurnal Dunia Pendidikan*, vol. 4, no. 4, pp. 1450–1462, 2024.

[11] A. Fadillah, R. Sukmawati, and S. Rahardjo, "Analysis of Student Digital Literacy in Linear Algebra Courses During the Covid-19 Pandemic," *AKSIOMA*, vol. 10, no. 2, p. 1206, Jul. 2021, doi: 10.24127/ajpm.v10i2.3704.

[12] M. F. Akbar and F. D. Anggaraeni, "Teknologi dalam Pendidikan: Literasi Digital dan Self-Directed Learning pada Mahasiswa Skripsi," *Indigenous*, vol. 2, no. 1, pp. 28–38, 2017. doi: 10.23917/indigenous.v1i1.4458

[13] Z. Huang, "A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining," *DMKD*, vol. 3, no. 8, pp. 34–39, 1997.

[14] Z. Huang, "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values," *Data Mining and Knowledge Discovery*, vol. 2, pp. 283–304, 1998, doi: 10.1023/A:1009769707641.

[15] S. Aranganayagi and K. Thangavel, "Improved K-Modes for Categorical Clustering Using Weighted Dissimilarity Measure," *Int. J. Comput. Inf. Eng.*, vol. 3, no. 3, 2009.

[16] W. H. Riska, D. Permana, A. A. Putra, and Zilrahmi, "Categorical Data Clustering with K-Modes Method on Fire Cases in DKI Jakarta Province," *UNP J. Statistics Data Sci.*, vol. 2, no. 1, pp. 56–63, Feb. 2024, doi: 10.24036/ujsds/vol2-iss1/115.

[17] F. S. Jumeilah and D. Pratama, "Klasterisasi Penduduk Lanjut Usia Sumatera Selatan Menggunakan Algoritma K-Modes," *Technology Acceptance Model*, vol. 8, no. 2, pp. 85–89, 2017.

[18] D. R. Quinthara, A. C. Fauzan, and M. M. Huda, "Penerapan Algoritma K-Modes Menggunakan Validasi Davies Bouldin Index Untuk Klasterisasi Karakter Pada Game Wild Rift," *JSCE*, vol. 4, no. 2, pp. 123–135, 2023.

[19] Y. A. Sari, "Analisis Sentimen pada Ulasan Hotel dengan Fitur Score Representation dan Identifikasi Aspek pada Ulasan Menggunakan K-Modes," *JPTIIK*, vol. 2, no. 9, pp. 2777–2782, 2018.

[20] I. B. Syamsi and M. L. A. Muharrom, "Klasterisasi Data Penduduk Berdasarkan Status Ekonomi di Desa Gebang Menggunakan Metode K-Means Clustering." Universitas Muhammadiyah Jember, 2015.

[21] D. J. Ketchen and C. L. Shook, "The Application of Cluster Analysis in Strategic Management Research: An Analysis and Critique," *Strategic Management Journal*, vol. 17, pp. 441–458, 1996, doi: 10.1002/(SICI)1097-0266(199606)17:6<441::AID-SMJ819>3.0.CO;2-G.

[22] K. S. Dorman and R. Maitra, "An efficient k-modes algorithm for clustering categorical datasets," *Stat. Anal. Data Min.*, vol. 15, no. 1, pp. 83–97, Feb. 2022, doi: 10.1002/sam.11546.

[23] J. Xie, M. Wang, X. Lu, X. Liu, and P. W. Grant, "DP-k-modes: A self-tuning k-modes clustering algorithm," *Pattern Recognit. Lett.*, vol. 158, pp. 117–124, Jun. 2022, doi: 10.1016/j.patrec.2022.04.026.

[24] T. Dinh, H. Wong, P. Fournier-Viger, D. Lisik, M.-Q. Ha, H.-C. Dam, and V.-N. Huynh, "Categorical data clustering: 25 years beyond K-modes," *Expert Systems with Applications*, vol. 272, 2025, Art. no. 126608, doi: 10.1016/j.eswa.2025.126608.

[25] G. Gan, Z. Yang, and J. Wu, "A Genetic k-Modes Algorithm for Clustering Categorical Data," *ADMA*, pp. 195–202, 2005.

[26] G. Gan, J. Wu, and Z. Yang, "A genetic fuzzy k-Modes algorithm for clustering categorical data," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 1615–1620, 2009, doi: 10.1016/j.eswa.2007.11.045.

[27] M. K. Ng and L. Jing, "A new fuzzy k-modes clustering algorithm for categorical data," *International Journal of Granular Computing, Rough Sets and Intelligent Systems*, vol. 1, no. 1, pp. 105–119, 2009, doi: 10.1504/IJGCRSIS.2009.026727.

[28] A. G. Oskouei, M. A. Balafar, and C. Motamed, "FKMAWCW: Categorical fuzzy k-modes clustering with automated attribute-weight and cluster-weight learning," *Chaos Solitons Fractals*, vol. 153, 2021, doi: 10.1016/j.chaos.2021.111494.

[29] F. Cao, J. Liang, and L. Bai, "A new initialization method for categorical data clustering," *Expert Syst. Appl.*, vol. 36, no. 7, pp. 10223–10228, 2009, doi: 10.1016/j.eswa.2009.01.060.

[30] J. Ji, T. Bai, C. Zhou, C. Ma, and Z. Wang, "An improved k-prototypes clustering algorithm for mixed numeric and categorical data," *Neurocomputing*, vol. 120, pp. 590–596, Nov. 2013, doi: 10.1016/j.neucom.2013.04.011.

[31] A. Kalogeratos and A. Likas, "Document clustering using synthetic cluster prototypes," *Data Knowl. Eng.*, vol. 70, no. 3, pp. 284–306, 2011, doi: 10.1016/j.datak.2010.12.002.

[32] A. Ahmad and S. S. Khan, "Survey of State-of-the-Art Mixed Data Clustering Algorithms," *IEEE Access*, vol. 7, pp. 31883–31902, 2019, doi: 10.1109/ACCESS.2019.2903568.

[33] N. B. Ellison, C. Steinfield, and C. Lampe, "The benefits of facebook 'friends:' Social capital and college students' use of online social network sites," *J. Comput.-Mediat. Commun.*, vol. 12, no. 4, pp. 1143–1168, Jul. 2007, doi: 10.1111/j.1083-6101.2007.00367.x.

[34] H. Hernández, E. Alberdi, A. Goti, and A. Oyarbide-Zubillaga, "Applica-

tion of the k-Prototype Clustering Approach for the Definition of Geostatistical Estimation Domains," *Mathematics*, vol. 11, no. 3, Feb. 2023, doi: 10.3390/math11030740.

[35] T. M. Kodinariya and P. R. Makwana, "Review on determining number of cluster in K-means clustering," in *Proceedings of the Conference*, 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:10090179.

[36] I. K. Khan et al., "Determining the optimal number of clusters by Enhanced Gap Statistic in K-mean algorithm," *Egyptian Informatics Journal*, vol. 27, 2024, doi: 10.1016/j.eij.2024.100504.

[37] T. Kuo and K. J. Wang, "A hybrid k-prototypes clustering approach with improved sine-cosine algorithm for mixed-data classification," *Comput. Ind. Eng.*, vol. 169, Jul. 2022, doi: 10.1016/j.cie.2022.108164.

[38] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987, doi: 10.1016/0377-0427(87)90125-7.

[39] A. Dutt, M. A. Ismail, T. Herawan, and I. A. Hashem, "Partition-Based Clustering Algorithms Applied to Mixed Data for Educational Data Mining: A Survey From 1971 to 2024," *IEEE Access*, vol. 12, pp. 172923–172942, 2024, doi: 10.1109/ACCESS.2024.3496929.