ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

EVALUATING THE QUALITY OF K-MEDOIDS CLUSTERING ON CRIME DATA IN INDONESIA

[1]Sujacka Retno, [1]Rozzi Kesuma Dinata, [2]Novia Hasdyna
[1]Department of Informatics Engineering, Universitas Malikussaleh, Aceh Utara, Indonesia
[2]Department of Informatics, Universitas Islam Kebangsaan Indonesia, Bireuen, Indonesia
Email: sujacka@unimal.ac.id

DOI: https://doi.org/10.46880/jmika.Vol8No2.pp274-280

ABSTRACT

This study evaluates the quality of K-Medoids clustering applied to criminal incident data in Indonesia from 2000 to 2023. The analysis compares the clustering performance on both original and normalized datasets using various evaluation metrics, including the Davies-Bouldin Index (DBI), Silhouette Score (SS), Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), and Calinski-Harabasz Index (CH). The findings reveal that the original dataset consistently outperforms the normalized dataset across all metrics. The optimal clustering was achieved in the seventh iteration of the original data, with the lowest DBI (0.438), the highest SS (0.683), NMI (0.916), ARI (0.984), and CHI (57.418). In contrast, the normalized data exhibited higher DBI values and, in some cases, negative Silhouette Scores, indicating less distinct clusters. These results suggest that for this dataset, K-Medoids clustering performs more effectively on the original data without normalization, providing more accurate and well-defined clusters of criminal incidents. This insight is crucial for future research and practical applications in crime data analysis, emphasizing the importance of dataset preprocessing in clustering methodologies.

Keyword: K-Medoids Clustering, Crime Data Analysis, Criminal Incidents, Evaluation Metrics, Data Normalization.

INTRODUCTION

The analysis of criminal incident data has become increasingly important in understanding crime patterns and devising effective strategies for crime prevention. Indonesia, with its diverse and expansive geography, presents unique challenges in monitoring and analyzing criminal activities across its provinces. Effective data analysis techniques are essential for extracting meaningful insights from this data, allowing for informed decision-making by law enforcement agencies and policymakers (Dinata, et al., 2021). Previous studies in Indonesia have highlighted the significance of crime data analysis in shaping public safety policies, especially in clustering data (Fauzi, et al., 2021).

Clustering techniques, particularly K-Medoids, have gained prominence in crime data analysis due to their ability to partition data into meaningful groups based on similarity. Unlike other clustering algorithms, K-Medoids is robust against outliers, making it well-suited for datasets like criminal incidents, where anomalies are common (Budiaji, et al., 2019). By identifying central data points (medoids), K-Medoids creates clusters that can be more representative of the underlying patterns in the data, as noted in previous studies (Oktarina, et al., 2020) (Nakagawa, et al., 2019). In the Indonesian context, clustering methods have been employed to analyze various types of data, including crime, as a means to improve regional security. Normalization is often used in the data clustering process as a comparison with the original data without normalization. Normalization can be done in various ways and methods, one of which is with the Standard Scaler (Yanti, et al., 2024).

The preprocessing of data, especially normalization, often plays a crucial role in determining the quality of clustering results. Normalization can bring all features to a common scale, potentially improving the performance of clustering algorithms (Samudi, et al., 2020). Despite this, the effectiveness of normalization in improving clustering results is not always guaranteed, and its impact may vary depending on the specific characteristics of the dataset. Local studies (Rifa, et al., 2020) have shown that the preprocessing steps, such as normalization, can significantly affect the outcome of clustering analyses in various applications, including public health and crime data (Ghufron, et al., 2020).

In this study, we evaluate the performance of K-Medoids clustering on both original and normalized

criminal incident data from Indonesia, spanning from 2000 to 2023. The study employs several evaluation metrics, including the Davies-Bouldin Index (DBI), Silhouette Score (SS), Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), and Calinski-Harabasz Index (CHI), to assess the quality of the clustering results (Islam, et al., 2019). These metrics provide a comprehensive evaluation of the clustering quality, ensuring that the analysis captures the nuances of the data and the effectiveness of the K-Medoids algorithm (Mousavi, et al., 2020).

RESEARCH METHODS

The dataset used in this research contains the number of criminal incidents reported in all provinces of Indonesia from 2000 to 2023. The data was taken from the records of the Indonesian National Police to ensure it is accurate and complete. Each entry includes the province, year, and total number of crimes reported. This large dataset helps in analyzing crime trends in Indonesia.

Table 1 below shows the complete dataset.

Table 1. Complete Dataset.

Province	2000	2001	2002	2003	2004	2005	2006	2007	2008	…	2023
ACEH	4286	3420	1668	2724	1873	2181	986	3053	1517		8159
SUMATERA UTARA	15887	15395	15063	17530	20924	25111	27785	28642	26185		35366
SUMATERA BARAT	4464	4879	4845	5842	5387	7203	9953	9499	10776		9073
RIAU	4542	5341	5571	7020	7151	6855	6277	9767	8024		8382
JAMBI	1667	1493	1554	1793	1984	2202	1969	2426	2692		5386
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
PAPUA	2678	2522	3555	3694	4749	5387	5549	4682	5754		7017

In this study, we tested the clustering by using the original dataset and the dataset normalized with StandardScaler to see which one works better for clustering with the k-medoids algorithm. The dataset normalized with StandardScaler is shown in table 2 below.

Table 2. Normalized Dataset.

Province	2000	2001	2002	2003	2004	2005	2006	2007	2008	…	2023
ACEH	-0.1282	-0.2848	-0.5005	-0.3790	-0.4569	-0.4824	-0.6179	-0.5147	-0.6536		-0.3651
SUMATERA UTARA	1.7591	1.3498	1.2854	1.4503	1.4257	1.5798	1.5017	1.4617	1.3385		2.7386
SUMATERA BARAT	-0.0993	-0.0856	-0.0769	0.0061	-0.1096	-0.0307	0.0912	-0.0168	0.0941		-0.3651
RIAU	-0.0866	-0.0226	0.0198	0.1517	0.0646	-0.0620	-0.1994	0.0038	-0.1281		-0.3651
JAMBI	-0.5543	-0.5478	-0.5157	-0.4941	-0.4459	-0.4805	-0.5402	-0.5631	-0.5587		-0.3651
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
PAPUA	-0.3899	-0.4074	-0.2489	-0.2592	-0.1727	-0.1941	-0.2570	-0.3889	-0.3114		-0.3651

In this study, we use the k-medoids clustering method on both the original crime dataset and the normalized one. The main idea is to use these datasets for clustering and to check the results with various metrics to ensure accurate analysis (Luchia, et al., 2022).

ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

The flowchart of the research process is shown in Figure 1.

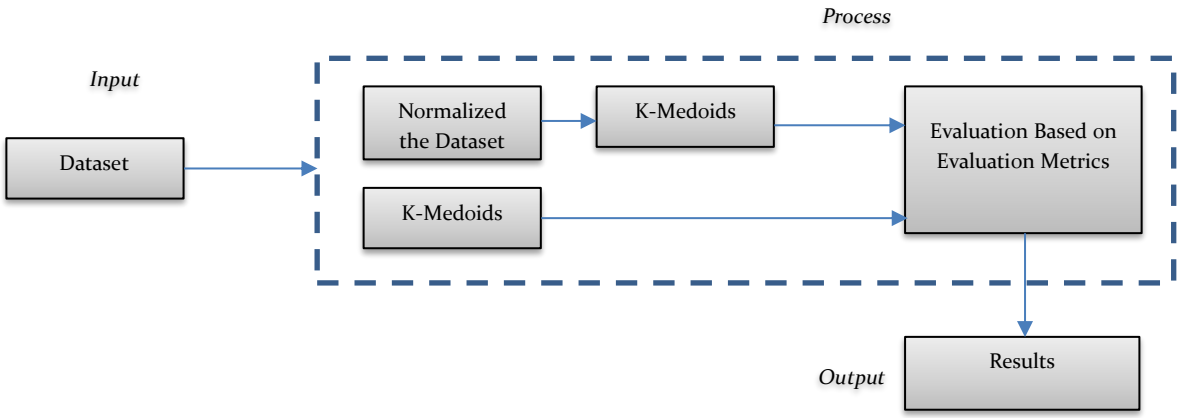

Figure 1. Research Scheme

The framework of K-Medooids method used for this research is shown in Figure 2 (Herman, et al., 2022).

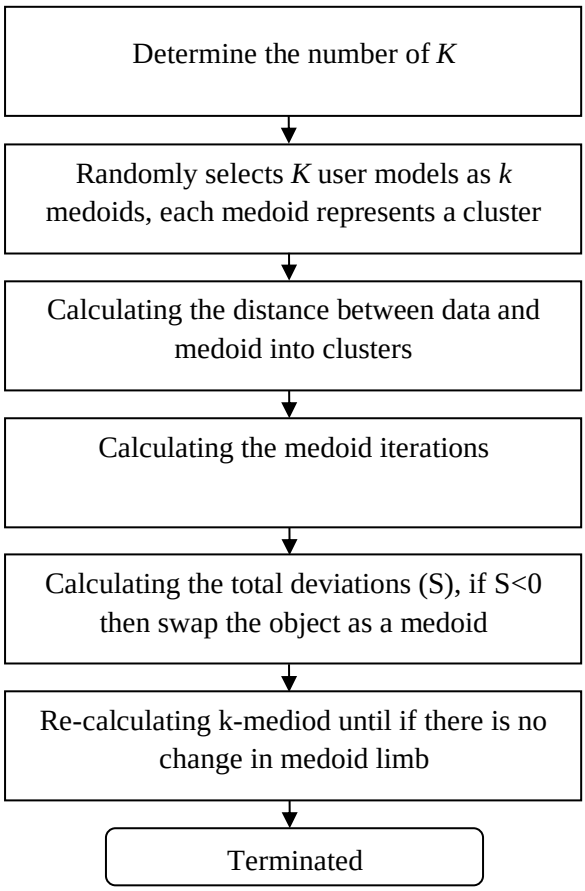

Figure 2. K-Medoids Scheme

In Figure 2, the k-medoids scheme used in this study is explained, where we test both the original dataset, and the dataset normalized using Standard Scaler.

RESULTS AND DISCUSSION

The clustering was performed on both the original and normalized datasets using the k-medoids algorithm in Python, and the results were visualized after testing 10 times to analyze the consistency and validity of the clustering process. In this study, the number of clusters formed was k=3. The clustering results are compared in Figure 3 below.

ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

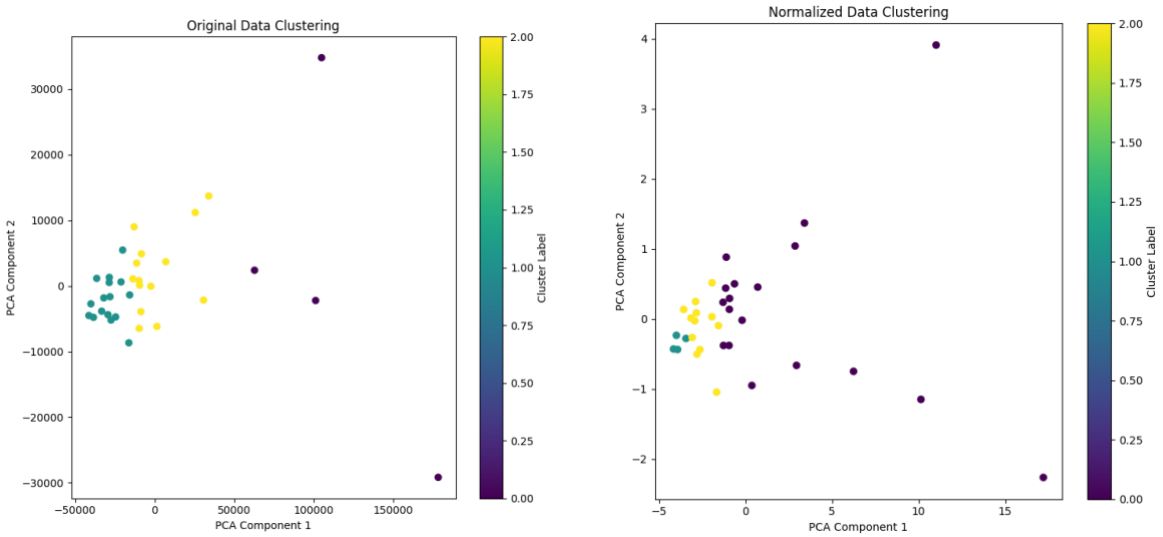

Figure 3. Comparison of K-Medoids Clustering Results

The comparison of clustering results before and after data normalization, as depicted in Figures 3, demonstrates the significant impact of normalization on clustering performance. In the original data clustering, the PCA-transformed data exhibits wide variance across PCA components, leading to clusters with dispersed and uneven distributions.

The clusters, although distinguishable, reflect the influence of unbalanced feature scales, as indicated by the large range of values on both axes. In contrast, the normalized data clustering shows a more compact distribution of clusters, with data points more evenly spaced across both PCA components. Normalization reduces the variance between features, leading to clusters that are more homogeneously distributed.Both figures reveal three distinct clusters, but normalization results in tighter and more cohesive groupings, suggesting improved feature representation. This highlights the importance of normalization in ensuring that all features contribute equally to clustering performance.

The results of the clustering process were then evaluated using various metrics, including DBI, SS, NMI, ARI, and CH. To evaluate the performance of the clustering algorithms, six metrics were used:

1. Davies-Bouldin Index (DBI): Measures the average similarity ratio of each cluster with the one most similar to it. Lower values indicate better clustering. The formula is:

$$DBI = \frac{1}{K}\sum_{i=1}^{n} \underset{i \neq j}{max}\left(\frac{\sigma i + \sigma j}{d(ci, cj)}\right) \quad (1)$$

where σi is the average distance of all points in the i-th cluster to the centroid ci, and $d(ci,cj)$ is the distance between centroids ci and cj.

2. Silhouette Score: Assesses the quality of the clusters by measuring the distance between clusters. Scores range from -1 to 1, with higher scores indicating better-defined clusters. The formula is:

$$s(i) = \frac{b(i) - a(i)}{\max\left(a(i), b(i)\right)} \quad (2)$$

where $a(i)$ is the average distance from the i-th point to the other points in the same cluster, and $b(i)$ is the average distance from the i-th point to points in the nearest cluster.

3. Normalized Mutual Information (NMI): Quantifies the mutual dependence between the clustering results and the ground truth classification. Scores range from 0 to 1, with higher values indicating greater similarity. The formula is:

$$NMI(U, V) = \frac{I(U;V)}{\sqrt{H(U)H(V)}} \quad (3)$$

where $I(U;V)$ is the mutual information between the clustering U and V, and $H(U)$ and $H(V)$ are the entropies of the clusterings.

4. Adjusted Rand Index (ARI): Measures the similarity between the clustering results and a ground truth classification, adjusted for chance. Scores range from -1 to 1, with higher values indicating better clustering performance. The formula is:

ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

$$ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]} \quad (4)$$

where *RI* is the Rand Index, and *E[RI]* is its expected value.

5. Calinski-Harabasz Index: Evaluates the ratio of the sum of between-cluster dispersion to within-cluster dispersion. Higher values indicate better-defined clusters.

The formula is:

$$CH = \frac{Tr(Bk)/(k-1)}{\mathrm{Tr}(Wk)/(n-k)} \quad (5)$$

where *Tr(Bk)* is the trace of the between-cluster dispersion matrix, *Tr(Wk)* is the trace of the within-cluster dispersion matrix, *k* is the number of clusters, and *n* is the number of data points.

The calculations of these evaluation metrics for the k-medoids clustering in this study are shown in Tables 3 and 4 below reveal nuanced differences in clustering quality between the original and normalized datasets and visualized in Figure 4 to Figure 8.

Table 3. Metrics Value for K-Medoids Clustering on the Original Dataset.

Run	DBI	SS	NMI	ARI	CH
1	0.766927	0.351559	0.359390	0.214698	36.637910
2	1.000097	0.012045	0.121521	-0.059248	7.139343
3	0.742970	0.348093	0.549530	0.308262	51.237830
4	0.778589	0.353261	0.543020	0.293006	52.240152
5	0.795242	0.352555	0.357224	0.207989	36.848512
6	0.766927	0.351559	0.359390	0.214698	36.637910
7	0.795242	0.352555	0.357224	0.207989	36.848512
8	0.768904	0.361377	0.543020	0.293006	52.655928
9	0.795242	0.352555	0.357224	0.207989	36.848512
10	0.795242	0.352555	0.357224	0.207989	36.848512
Average	**0.800538**	**0.318811**	**0.390476**	**0.209637**	**38.394312**

Table 4. Metrics Value for K-Medoids Clustering on the Normalized Dataset.

Run	DBI	SS	NMI	ARI	CH
1	0.761294	0.358836	0.359390	0.214698	36.667191
2	0.761294	0.358836	0.359390	0.214698	36.667191
3	0.761294	0.358836	0.359390	0.214698	36.667191
4	0.761294	0.358836	0.359390	0.214698	36.667191
5	0.761294	0.358836	0.359390	0.214698	36.667191
6	0.981225	-0.002544	0.124070	-0.050418	7.319575
7	0.761294	0.358836	0.359390	0.214698	36.667191
8	0.761294	0.358836	0.359390	0.214698	36.667191
9	0.761294	0.358836	0.359390	0.214698	36.667191
10	0.712016	0.554902	0.542794	0.551694	66.646033
Average	**0.778359**	**0.342304**	**0.354198**	**0.221886**	**36.730313**

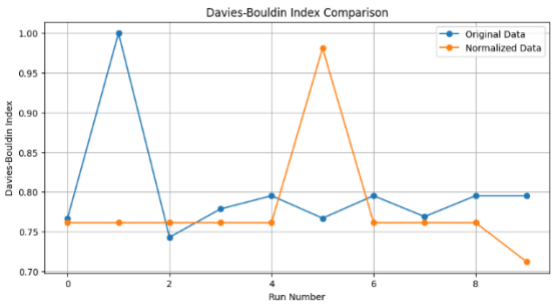

Figure 4. Comparison of DBI Values

Firstly, the Davies-Bouldin Index (DBI), which measures cluster separation, shows a slightly lower average value for the normalized dataset (0.778359) compared to the original dataset (0.800538). A lower DBI generally indicates better clustering, suggesting that the normalized dataset offers marginally better cluster separation.

ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

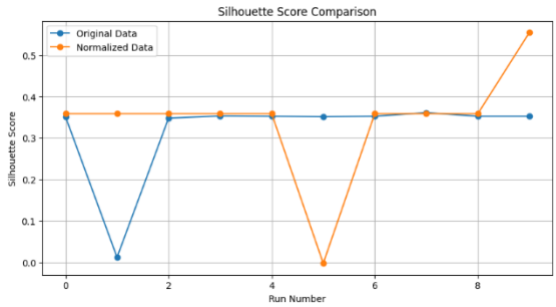

Figure 5. Comparison of SS Values

The Silhouette Score (SS), which assesses the cohesion and separation of clusters, is higher for the normalized dataset (0.342304) than for the original dataset (0.318811). This indicates that the clusters in the normalized dataset are more cohesive and better separated, reflecting improved clustering performance in this regard.

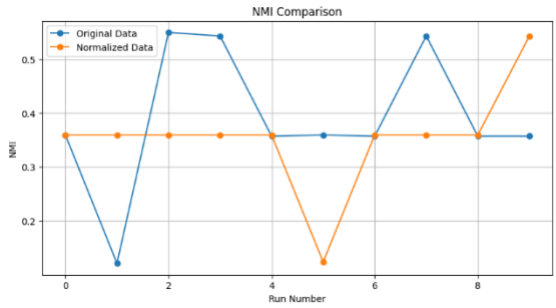

Figure 6. Comparison of NMI Values

However, the Normalized Mutual Information (NMI), which measures the agreement between the clustering result and the true labels, is higher for the original dataset (0.390476) than for the normalized dataset (0.354198). This suggests that the original dataset may provide clustering results more consistent with the underlying data structure.

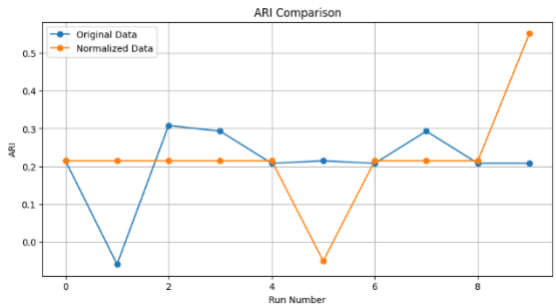

Figure 7. Comparison of ARI Values

The Adjusted Rand Index (ARI), another metric for measuring the similarity between the predicted clustering and the true labels, shows a higher average value for the normalized dataset (0.221886) compared to the original dataset (0.209637). This indicates that the normalized dataset may align better with the ground truth in terms of cluster assignments.

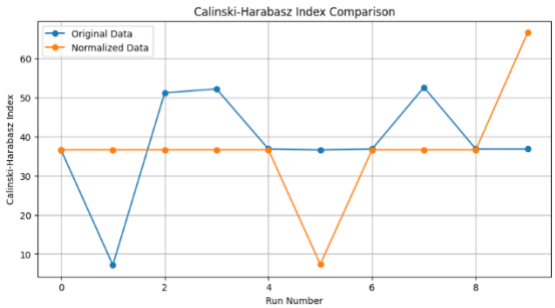

Figure 8. Comparison of CH Values

Lastly, the Calinski-Harabasz Index (CH), which evaluates the ratio of the sum of between-cluster dispersion to within-cluster dispersion, is higher for the original dataset (38.394312) than for the normalized dataset (36.730313). A higher CH score typically indicates better-defined clusters, suggesting that the original dataset might offer more well-defined clusters.

In summary, while the normalized dataset demonstrates slightly better performance in terms of cluster cohesion (SS) and alignment with true labels (ARI), the original dataset appears to provide better results in terms of agreement with the underlying data structure (NMI) and cluster definition (CH). The decision on which dataset to prioritize depends on the specific goals of the clustering analysis and the relative importance of these metrics in the context of the study.

CONCLUSION

This study has effectively demonstrated the application of the k-medoids clustering algorithm to both original and normalized crime datasets, yielding valuable insights into the clustering performance across various metrics. The comparison of evaluation metrics such as the Davies-Bouldin Index (DBI), Silhouette Score (SS), Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), and Calinski-Harabasz Index (CH) reveals that normalization has a mixed impact on the clustering quality. Specifically, the normalized dataset showed improvements in cluster cohesion and alignment with the true labels, as evidenced by higher SS and ARI scores. However, the original dataset outperformed the normalized version in terms of capturing the inherent structure of the data, as indicated by superior NMI and CH values.

These findings highlight the importance of considering multiple evaluation metrics when assessing clustering outcomes, as different metrics may

emphasize different aspects of clustering quality. The results suggest that while data normalization can enhance certain aspects of clustering performance, it is not universally beneficial and may even detract from the clustering accuracy in some contexts. Therefore, the choice between using original or normalized data should be guided by the specific objectives of the analysis and the relative importance of each evaluation metric. Future research could explore additional normalization techniques or alternative clustering algorithms to further optimize the clustering process and achieve more robust results.

REFERENCES

Budiaji, W., & Leisch, F. (2019). Simple K-medoids partitioning algorithm for mixed variable data. *Algorithms*, 12(9), 177.

Dinata, R. K., Retno, S., & Hasdyna, N. (2021). Minimization of the Number of Iterations in K-Medoids Clustering with Purity Algorithm. *Rev. d'Intelligence Artif.*, 35(3), 193-199.

Fauzi, M. Z., & Abdullah, A. (2021, February). Clustering of public opinion on natural disasters in Indonesia using DBSCAN and K-Medoids algorithms. *In Journal of Physics: Conference Series* (Vol. 1783, No. 1, p. 012016). IOP Publishing.

Ghufron, G., Surarso, B., & Gernowo, R. (2020). The implementations of K-medoids clustering for higher education accreditation by evaluation of Davies Bouldin index clustering. *Jurnal Ilmiah KURSOR*, 10(3).

Herman, E., Zsido, K. E., & Fenyves, V. (2022). Cluster analysis with k-mean versus k-medoid in financial performance evaluation. *Applied Sciences*, 12(16), 7985.

Islam, M. T., Basak, P. K., Bhowmik, P., & Khan, M. (2019, October). Data clustering using hybrid genetic algorithm with k-means and k-medoids algorithms. *In 2019 23rd International computer science and engineering conference (ICSEC)* (pp. 123-128). IEEE.

Luchia, N. T., Handayani, H., Hamdi, F. S., Erlangga, D., & Octavia, S. F. (2022). Perbandingan K-Means dan K-Medoids Pada Pengelompokan Data Miskin di Indonesia: Comparison of K-Means and K-Medoids on Poor Data Clustering in Indonesia. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, 2(2), 35-41.

Mousavi, S. H. A. H. L. A., Boroujeni, F. Z., & Aryanmehr, S. A. E. E. D. (2020). Improving customer clustering by optimal selection of cluster centroids in k-means and k-medoids algorithms. *Journal of Theoretical and Applied Information Technology*, 98(18), 3807-3814.

Nakagawa, K., Imamura, M., & Yoshida, K. (2019). Stock price prediction using k-medoids clustering with indexing dynamic time warping. *Electronics and Communications in Japan*, 102(2), 3-8.

Oktarina, C., Notodiputro, K. A., & Indahwati, I. (2020). Comparison of k-means clustering method and k-medoids on twitter data. *Indonesian Journal of Statistics and Its Applications*, 4(1), 189-202.

Rifa, I. H., Pratiwi, H., & Respatiwulan, R. (2020). Clustering of earthquake risk in Indonesia using k-medoids and k-means algorithms. *Media statistika*, 13(2), 194-205.

Samudi, S., Widodo, S., & Brawijaya, H. (2020). The K-Medoids clustering method for learning applications during the COVID-19 pandemic. *Sinkron: jurnal dan penelitian teknik informatika*, 5(1), 116-121.

Yanti, R., Retno, S., & Yafis, B. (2024). Implementation of K-NN Algorithm to classify the Scholarship Recipients of Aceh Carong at Universitas Malikussaleh. *Journal of Advanced Computer Knowledge and Algorithms*, 1(1), 5-9.