Proposed Modification of K-Means Clustering Algorithm with Distance Calculation Based on Correlation

Muhammad Ibnu Choldun Rachmatullah
Politeknik Pos Indonesia, Jl. Sariasih 54, Bandung 40151, Indonesia

ARTICLE INFO

Article history:

Received March 25, 2022
Revised April 14, 2022
Accepted April 30, 2022

Keywords:

Clustering;
K-Means;
Distance calculation;
Correlation;
Accuracy

ABSTRACT

Clustering is a technique in data mining that groups a set of data into groups (clusters) of similar data. In general, there are two methods of clustering, namely the hierarchical method and the partition method. One of the most commonly used partition clustering methods in clustering is K-Means. The use of K-means method has been widely used in various fields with various purposes. Many research has been carried out to improve the performance of the K-Means method, for example, by modifying the method of determining the initial centroid or determining the appropriate number of clusters. In this research, the modification of the K-Means algorithm was carried out in calculating the distance by considering the correlation value between attributes. Attributes that have a high correlation value are assumed to have similar characteristics so that they determine the location of data in a particular cluster. The steps of the proposed method are: calculating the correlation value between attributes, determining the cluster centroid, calculating the distance by considering the value of correlation, and determining the data into certain clusters. The first contribution of this research is to propose a new distance calculation technique in the K-Means algorithm by considering correlation and the second contribution is to apply the proposed algorithm to a specific dataset, namely Iris dataset. In this research, the performance calculation of the modified algorithm was also carried out. From the experimental results using the Iris dataset, the proposed modification of the K-Means algorithm has fewer iterations than the original K-Means method, so that it requires less processing time. The original K-Means method requires 8 iterations, while the proposed method requires only 6 iterations. The proposed method also produces a higher accuracy rate of 89.33% than the original K-Means method, which is 82.67%.

Corresponding Author:

Muhammad Ibnu Choldun Rachmatullah, Politeknik Pos Indonesia, Jl. Sariasih No. 54, Bandung, 40151, Indonesia
Email: ibnucholdun@poltekpos.ac.id

1. INTRODUCTION

The method of grouping data in data mining is known as clustering. Clustering is an attempt to group data into several clusters or classes based on the level of similarity, the more similar the value of a data point, the more likely it will be in the same class [1]. This clustering method uses two main approaches, namely, based on hierarchy and partition. The clustering method with a hierarchical approach is done by creating a hierarchy, usually in the form of a dendrogram, by placing data that have the same level of similarity in one hierarchy. As a result, data with a low level of similarity will occupy a hierarchy that is far apart [2][3]. Clustering with a partition-based clustering approach is carried out by grouping the data and sorting the analyzed data into existing clusters [4][5].

One of the commonly used clustering methods in the field of data mining is the K-Means clustering method [6]. K-Means is a method of analyzing data in data mining by carrying out a modeling process without supervision (unsupervised) and is one method of grouping data with a partition system. This method is done by grouping objects into k clusters or classes. To perform this clustering, the value of k must be determined first. The purpose of clustering data into k classes is to minimize variation within a class and maximize variation between classes.

The K-Means method has been widely used for clustering in various fields, for example, smart investment [7], sensors [8], Defect Detection [9], big data applications [10], Covid-19 risk [11], earthquake magnitude prediction [12], and multi-objective programming problems [13]. In addition to research using the K-Means method in various fields, research is also carried out to improve the performance of the K-Means method. Efforts to improve the performance of the K-Means algorithm have also been made to improve its performance. One way to improve performance is to combine the K-Means method with other methods, for example, semi-supervised learning [14], feature selection [15], hybrids with other methods [16], using Fuzzy metrics [17], with weighted K-NN [18], combining with Multi-Column Matrices Selection [19], One-Class SVM [20][21], particle swarm optimization algorithm [22]. There are also researchers who improve the performance in terms of the speed of the K-Means method by parallelizing iterations [23]. The research was also carried out with improvements to the k-means algorithm, for example, selecting the number of clusters based on the density characterization of objects [24], determining the initial centroid using the maximin method [25], determining the number of clusters using the Maximum Stable Set Problem and Continuous Hopfield Network [26]. Research with the K-Means algorithm is also carried out for special data, for example, those with outliers [27], associated with ensemble learning [3], [28].

Several distance calculation methods that have been carried out by researchers are Euclidian, Manhattan, Chebyshev, Minkowski, etc. [29][30][31]. One modification of the distance calculation in K-Means that considers correlation is by calculating the Mahalanobis distance [32], but what is used is the covariance value, not the pure correlation value. Distance calculations by considering correlations using correlation coefficients have also been carried out but are only suitable for clustering data that have a large number of attributes, for example, for gene expression data [33]. In this research, we propose a modification of the K-Means algorithm by modifying the distance calculation. So the distance calculation is done by taking into account the correlation value between attributes, although, in fact, the correlation coefficient is actually only used to form a value based on its attributes and correlations, then used in calculating distances. Correlation shows the strength of the relationship between attributes [34][35]. So it is assumed that the attributes that have a high correlation value are assumed to have the same characteristics so that this correlation value plays a role in determining which cluster the data belongs to. So the main difference between the original K-Means algorithm with the proposed modification of the K-Means algorithm is in calculating the distance, whereas the original K-Means uses Euclidean distance, while the proposed modification considers the correlation between attributes. The main contribution of this research is to propose a new distance calculation technique in the K-Means algorithm by considering correlation. An additional contribution of this research is to apply the proposed algorithm to a specific dataset.

2. METHOD

The research method was carried out following the steps shown in Fig. 1, which consists of several steps that are Preparing the Dataset, Correlation Calculation, K-Means Algorithm Modification, Clustering Process, and Calculating Performance. Each step is explained in each subsequent section.

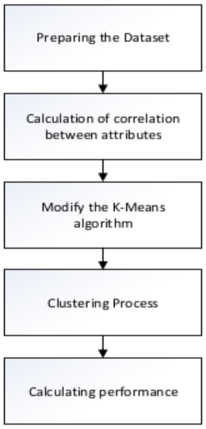
Preparing the Dataset
Calculation of correlation between attributes
Modify the K-Means algorithm
Clustering Process
Calculating performance

Fig. 1. Research steps

2.1. Preparing the Dataset

The dataset used was taken from the UCI Machine Learning Repository, which has a classification objective function so that it has a categorical label/output. The dataset used is the Iris dataset which consists of 150 data, has 4 input attributes, and one attribute as a label/output, which consists of three classes, as presented in Table 1. This class will be compared with the clustering as a result of clustering. These four input attributes are: sepal_length, sepal_width, petal_length, and petal_width. The three classes of output attributes are Iris-Setosa, Iris-Versicolour, and Iris-Virginica. The 1st to 50th data have the Iris-Setosa class, the 51st to 100th data have the Iris-Versicolour class, and the 101st to 150th data have the Iris-Virginica class. In Table 1, examples of 12 data from 150 data are given, and each class is represented by 4 data.

Table 1. The value of attributes

ith data	Sepal length(1)	Sepal width(2)	Petal length(3)	Petal width(4)	Class
1	5.1	3.5	1.4	0.2	Iris-setosa
5	5.0	3.6	1.4	0.2	Iris-setosa
20	5.1	3.8	1.5	0.3	Iris-setosa
50	5.0	3.3	1.4	0.2	Iris-setosa
51	7.0	3.2	4.7	1.4	Iris-versicolor
65	5.6	2.9	3.6	1.3	Iris-versicolor
80	5.7	2.6	3.5	1.0	Iris-versicolor
100	5.7	2.8	4.1	1.3	Iris-versicolor
101	6.3	3.3	6	2.5	Iris-virginica
125	6.7	3.3	5.7	2.1	Iris-virginica
140	6.9	3.1	5.4	2.1	Iris-virginica
150	5.9	3.0	5.1	1.8	Iris-virginica

2.2. Correlation Calculation

Correlation is a statistical term that expresses the degree of the linear relationship between two or more variables/attributes [36][37]. For attributes that have a correlation coefficient interval scale, used is Pearson's Product Moment. The correlation coefficient between two attributes can be calculated using equation (1).

$$r_{xy} = \frac{n\sum XY - \sum X \sum Y}{\sqrt{n\sum X^2 - (\sum X)^2}\sqrt{n\sum Y^2 - (\sum Y)^2}} \tag{1}$$

Where r_{xy} is correlation coefficient and n is the number of data.

In this research, the correlation value used is the absolute value because it only considers the strength of the relationship between two attributes without considering the direction. The larger the number, the stronger the relationship [38][39].

2.3. K-Means Algorithm Modification

The clustering process is basically grouping data that have the same characteristics in one cluster and those that have different characteristics in another cluster. In this research, data that have a high correlation are assumed to have similarities. In the original K-Means algorithm, the similarity is obtained by calculating the Euclidian distance [29]. In the proposed algorithm, the distance calculation is obtained by considering the correlation between attributes. To calculate the distance to the i^{th} data (x_i) at the center of the k^{th}-cluster (C_k) named (d_{ij}), the Euclidian formula can be used, namely equation (2).

$$d_{ij} = \sqrt{\sum_{j=1}^{m}(x_{ij} - c_{kj})^2} \tag{2}$$

In the proposed algorithm modification, the distance calculation is by subtracting the K value from the k^{th} cluster center (C_k). The value of K is calculated by equation (3), while the distance to the center is calculated by equation (4).

$$K_i = \sum r_{ij} x_i x_j \tag{3}$$

So the distance from a data point to the kth cluster center is:

$$d_i = |K - C_k| \tag{4}$$

2.4. Clustering Process

In the Iris dataset, the number of classes in the output attribute is three. Therefore, in the clustering process, the number of clusters is determined to be three. The differences between the K-Means algorithm and the modified K-Means algorithm are as follows, as presented in Table 2. The K-Means algorithm and the modified K-Means algorithm can be described with a flowchart, as shown in Fig. 2.

Table 2. Difference between Original K-Means Algorithm and modified K-Means

K-Means	Modified K-Means
1. Determine the number of clusters (k=3)	1. Determine the number of clusters (k=3)
2. Determine the cluster center randomly	2. Calculate the correlation between attributes
3. Calculate the Euclidian distance between the data and the center of each cluster	3. Determine the location of the cluster center randomly
4. Allocate data to clusters that have the closest distance	4. Calculate the distance of each data with the center of each cluster using the equation (3) and equation (4)
5. Calculate the new cluster center based on the data in the same cluster	5. Allocate data to clusters with the closest distance
6. Repeat steps 3 to 5 so that no data moves the cluster	6. Calculate the new cluster center based on the data in the same cluster
	7. Repeat steps 4 to 6 so that no data moves the cluster

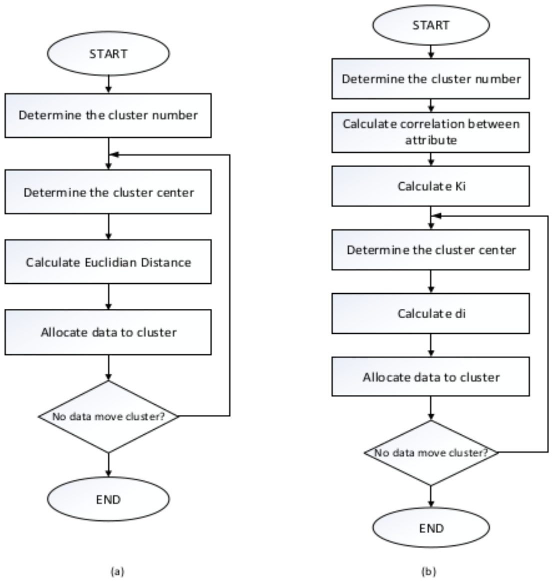
START
Determine the cluster number
Determine the cluster center
Calculate Euclidian Distance
Allocate data to cluster
No data move cluster?
END
(a)
START
Determine the cluster number
Calculate correlation between attribute
Calculate Ki
Determine the cluster center
Calculate di
Allocate data to cluster
No data move cluster?
END
(b)

Fig. 2. Flowchart: (a) Original K-Means (b) Proposed K-Means

So the main difference between the original K-Means algorithm and the modified K-Means algorithm is in the calculation of the distance. In the K-Means algorithm, in determining the center of the cluster, the attribute value of the selected data is used, while in the modified K-Means algorithm, the K value is used as in the equation (3).

2.5. Calculating Performance

The performance of the algorithm is determined by the amount of data that is grouped into the correct cluster based on the labels in the dataset. This performance is measured as accuracy by the equation (5).

$$Accuracy = \frac{the\ count\ of\ data\ grouped\ correctly}{the\ count\ all\ of\ data} \tag{5}$$

3. RESULTS AND DISCUSSION

The results and analysis section presents the results of applying the modified K-Means algorithm to the Iris dataset. As a comparison in several steps, the results of the two algorithms are presented, both the original K-Means algorithm and the modified K-Means algorithm.

3.1. Correlation Calculation

The four input attributes of the Iris dataset are (1) sepal length, (2) sepal width, (3) petal length, and (4) petal width. The results of calculating the correlation coefficient based on equation (1) between the two attributes are as in Table 3.

Table 3. Correlation between two attributes

r	Value	Absolute value
r_{12}	-0.109	0.109
r_{13}	0.872	0.872
r_{14}	0.818	0.818
r_{23}	-0.421	0.421
r_{24}	-0.357	0.357
r_{34}	0.963	0.963

From the calculation of the correlation coefficient, there are three correlations that are negative, but the absolute value is used because what is considered is only the strength of the relationship [38][39].

3.2. Clustering Process with Modified K-Means Algorithm

After obtaining the correlation value between attributes, as shown in Table 3, the next is determining the initial cluster center. For example, the 5th data, 65th data, and 125th data are taken from the cluster center. The selection of these three data with consideration to represent each class is presented in Table 1. For the calculation simulation, apart from the three data centers, three other data were also taken, for example, the 20th data, 80th data, and 140th data. The selection of these three data is also with consideration to represent each class, on the condition that other than the data that has been selected as the center. The three centers and three other data, along with the value of each attribute, are presented in Table 1. For example, the calculation of the K_i value for each of the six selected data is as in Table 4.

Table 4. The calculation of K_i

i^{th}	K_i	Centroid-k
5	r_{12}*5*3.6+r_{13}*5*1.4+r_{14}*5*0.2+r_{23}*3.6*1.4+r_{24}*3.6*0.2+r_{34}*1.4*0.2 =0.109*5*3.6+0.872*5*1.4+0.818*5*0.2+0.421*3.6*1.4+0.357*3.6*0.2+0.963*1.4*0.2= 6.782	Centroid-1
65	r_{12}*5.6*2.9+r_{13}*5.6*3.6+r_{14}*5.6*1.3+r_{23}*2.9*3.6+r_{24}*2.9*1.3+r_{34}*3.6*1.3=0.109*5.6*2.9+0.872*5.6*3.6+0.818*5.6*1.3+0.421*2.9*3.6+0.357*2.9*1.3+0.963*3.6*1.3= 24.077	Centroid-2
125	r_{12}*6.7*3.3+r_{13}*6.7*5.7+r_{14}*6.7*2.1+r_{23}*3.3*5.7+r_{24}*3.3*2.1+r_{34}*5.7*2.1=0.109*6.7*3.3+0.872*6.7*5.7+0.818*6.7*2.1+0.421*33.3*5.7+0.357*3.3*2.1+0.963*5.7*2.1= 48.363	Centroid-3
20	7.770	
80	22.292	
140	48.230	

In Table 4, for the 20th data, 80th data, and 140th data, only the final value is shown for the calculation of the K_i value. Table 5 shows the distance of the three data to each cluster center so that it can be concluded which cluster belongs to. Then the distance of the three data to each cluster center is calculated, and the data is entered into which cluster is based on the closest distance. The calculation of this distance can be presented in Table 5.

Table 5. Distance to centroid

ith	d_i (centroid-1)	d_i (centroid-2)	d_i (centroid-3)	kth Cluster
20	|7.770-6.782|= 0.988	|7.770-24.077|= 16.307	|7.770-48.363|= 40.593	1
80	|22.292-6.782|= 15.51	|22.292-24.077|= 1.785	|22.292-48.363|= 26.071	2
140	|48.230-6.782|= 41.448	|48.230-24.077|= 24.153	|48.230-48.363|= 0.133	3

Calculations, as presented in Table 4 and Table 5, are carried out on all existing data so that each data is entered in the appropriate cluster based on the closest distance from the cluster center. Then the new center is determined from each cluster by averaging the results from the data included in a particular cluster. After that, the distance for each data with the new cluster center is calculated again. The calculation of the distance and the determination of the new cluster center is carried out repeatedly until no data moves clusters.

3.3. Calculating Performance

By using the modified K-Means algorithm, after repeating the distance calculation and determining the new cluster center, it turns out that after the 6th iteration, there is no more data moving clusters. If using the original K-Means algorithm, after the 8th iteration, there are no more data moving clusters. So based on the number of iterations performed, the modified K-Means algorithm requires fewer iterations. Or in other words, the total processing time becomes faster. By using this modified K-Means algorithm, the K_i calculation is only done once, not in every iteration, while in the original K-Means algorithm, the calculation of the Euclidian distance is carried out in each iteration which automatically also increases processing time.

At the end of the iteration using the modified K-Means algorithm, the results of the clustering turned out to be 134 data entered in the cluster according to the label on the dataset, while using the original K-means algorithm at the end of the iteration, there were 124 data entered into the appropriate cluster with dataset labels. Comparison of accuracy when iteration stops of the two algorithms are presented in Table 6. Accuracy is calculated using the equation (5).

Table 6. Performance comparison

Algorithm	Iteration number	Dataset number	Appropriate Class/Label	Accuracy
K-Means	8	150	124	82.67%
Modified K-Means	6	150	134	89.33%

The accuracy of the original K-Means algorithm is 82.67%, while the modified K-Means algorithm is 89.33%. A previous study conducted by Sakthi and Thanamani [40], using the original K-Means with the number of clusters being three, gave a slightly different accuracy of 81.25% compared to the experimental results conducted by the researcher. This difference could be due to differences in the selection of the initial centroid. While the application of K-Means with distance calculations that consider correlation, namely the Mahalanobis distance and the number of clusters is 3, which was carried out by previous researchers, giving an accuracy of 85.1% [32]. So by modifying the original K-Means, namely by considering the correlation between attributes, this research can increase the accuracy by 6.66%.

4. CONCLUSION

This research proposes to modify the K-Means algorithm to improve performance by modifying the distance calculation. Whereas the original K-Means method used Euclidian distance calculation, in the proposed algorithm, the distance calculation is carried out by considering the correlation between attributes. The experimental results show that the number of iterations required until no data moves clusters, the original algorithm requires more iterations than the proposed algorithm. So the proposed algorithm requires less processing time. The proposed algorithm has also succeeded in increasing the accuracy of clustering. For future research, before conducting clustering, it can be preceded by checking outliers because the presence of outliers also affects the quality of clustering.

REFERENCES

[1] X. Wang, A. S. Id, X. Hou, and L. Tan, "Research on cluster system distribution of traditional fort-type settlements in Shaanxi based on K-means clustering algorithm," *PLoS ONE*, vol. 17, 2022, https://doi.org/10.1371/journal.pone.0264238.

[2] A. A. Amer, "On K-means clustering-based approach for DDBSs design," *J. Big Data*, vol. 7, 2020, https://doi.org/10.1186/s40537-020-00306-9.

[3] R. Guo and J. Chen, "Hierarchical K-means clustering for registration of multi-view point sets," *Computers & Electrical Engineering*, vol. 94, 2021, https://doi.org/10.1016/j.compeleceng.2021.107321.

[4] Z. Hossain, N. Akhtar, R. B. Ahmad, and M. Rahman, "A dynamic K-means clustering for data mining," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 13, no. 2, pp. 521–526, 2019, https://doi.org/10.11591/ijeecs.v13.i2.pp521-526.

[5] A. M. Ikotun, M. S. Almutari, and A. E. Ezugwu, "K-Means-Based Nature-Inspired Metaheuristic Algorithms for Automatic Data Clustering Problems: Recent Advances and Future Directions," *Applied Science*, vol. 11, no. 23, p. 11246, 2021, https://doi.org/10.3390/app112311246.

[6] A. Fahim, "K and starting means for k-means algorithm," *Journal of Computational Science*, vol. 55, p. 101445, 2021, https://doi.org/10.1016/j.jocs.2021.101445.

[7] L. Li and J. I. A. Wang, "Efficiency Analysis of Machine Learning Intelligent Investment Based on K-Means Algorithm," *IEEE Access*, vol. 8, pp. 147463–147470, 2020, https://doi.org/10.1109/ACCESS.2020.3011366.

[8] Y. Fan, Y. Liu, H. Qi, F. Liu, and X. Ji, "Anti-Interference Technology of Surface Acoustic Wave Sensor Based on K-Means," *IEEE Sensors Journal*, vol. 21, no. 7, pp. 8998–9007, 2021, https://doi.org/10.1109/JSEN.2021.3052957.

[9] N. H. M. M. Shrifan, G. N. Jawad, N. O. R. Ashidi, M. A. T. Isa, and M. F. Akbar, "Microwave Nondestructive Testing for Defect Detection in Composites Based on K-Means Clustering Algorithm," *IEEE Access*, vol. 9, 2021, https://doi.org/10.1109/ACCESS.2020.3048147.

[10] F. Moodi and H. Saadatfar, "An improved K-means algorithm for big data," *IET Software*, vol. 16, no. 1, 2021, https://doi.org/10.1049/sfw2.12032.

[11] D. Abdullah, S. S. Ansari, S. Ahmar, and R. R. Rahmat, "The application of K - means clustering for province clustering in Indonesia of the risk of the COVID - 19 pandemic based," *Quality & Quantity*, vol. 56, no. 3, pp. 1283-1291, 2021, https://doi.org/10.1007/s11135-021-01176-w.

[12] R. Yuan, "An improved K-means clustering algorithm for global earthquake catalogs and earthquake magnitude prediction," *Journal of Seismology*, vol. 25, pp. 1005-1020, 2021, https://doi.org/10.1007/s10950-021-09999-8.

[13] Y. Abo-Elnaga and S. Nasr, "K-means cluster interactive algorithm-based evolutionary approach for solving bilevel multi-objective programming problems," *Alexandria Eng. J.*, vol. 61, no. 1, pp. 811–827, 2022, https://doi.org/10.1016/j.aej.2021.04.098.

[14] Q. Ning, Z. Ma, X. Zhao, and M. Yin, "SSKM _ Succ : A novel succinylation sites prediction method incoprating K - means clustering with a new semi-supervised learning algorithm," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 19, no. 1, pp. 643-652, 2022, https://doi.org/10.1109/TCBB.2020.3006144.

[15] M. Capó, A. Pérez, J. A. Lozano, and S. Member, "A Cheap Feature Selection Approach for the K -Means Algorithm," *IEEE Transactions On Neural Networks and Learning Systems*, vol. 32, no. 5, pp. 2195-2208, 2021, https://doi.org/10.1109/TNNLS.2020.3002576.

[16] S. Kumar and M. Singh, "A Novel Clustering Technique for Efficient Clustering of Big Data in Hadoop Ecosystem," Big Data Mining and Analytics, vol. 2, no. 4, pp. 240–247, 2019, https://doi.org/10.26599/BDMA.2018.9020037.

[17] X. Geng, Y. Mu, S. Mao, J. Ye, and L. Zhu, "An Improved K-Means Algorithm Based on Fuzzy Metrics," *IEEE Access*, vol. 8, 2020, https://doi.org/10.1109/ACCESS.2020.3040745.

[18] A. K. Gangwar, O. P. Mahela, B. Rathore, B. Khan, H. H. Alhelou, and P. Siano, "A Novel K-Means Clustering and Weighted K-NN Regression Based Fast Transmission Line Protection," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 9, pp. 6034-6043, 2021, https://doi.org/10.1109/TII.2020.3037869.

[19] J. He, K. Yu, Z. Huang, and Z. Chen, "Multi-Column Matrices Selection Combined With k-Means Scheme for Mobile OCC System With Multi LEDs," *IEEE Photonics Technology Letters*, vol. 33, no. 12, pp. 623–626, 2021, https://doi.org/10.1109/LPT.2021.3079264.

[20] T. M. Hoang, N. M. Nguyen, and T. Q. Duong, "Detection of Eavesdropping Attack in UAV-aided Wireless Systems: Unsupervised Learning with One-Class SVM and K-means Clustering," *IEEE Wirel. Commun. Lett.*, vol. 9, no. 2, pp. 139-142, 2020, https://doi.org/10.1109/LWC.2019.2945022.

[21] T. Y. Wen, S. Member, S. Armiza, M. Aris, and S. Member, "Hybrid Approach of EEG Stress Level Classification Using K-Means Clustering and Support Vector Machine," *IEEE Access*, vol. 10, 2022, https://doi.org/10.1109/ACCESS.2022.3148380.

[22] Y. Li, X. Chu, D. Tian, J.Feng, and W. Mu, "Customer segmentation using K-means clustering and the adaptive particle swarm optimization algorithm," *Applied Soft Computing*, vol. 113, 2021, https://doi.org/10.1016/j.asoc.2021.107924.

[23] W. Kwedlo and M. Łubowicz, "Accelerated K -Means Algorithms for Low-Dimensional Data on Parallel Shared-Memory Systems," *IEEE Access*, vol. 9, 2021, https://doi.org/10.1109/ACCESS.2021.3080821.

[24] J. Ye, J. Zou, J. Gao, G. Zhang, M. Kong, Z. Pei, and K. Cui, "A New Frequency Hopping Signal Detection of Civil UAV Based on Improved K -Means Clustering Algorithm," *IEEE Access*, vol. 9, 2021, https://doi.org/10.1109/ACCESS.2021.3070491.

[25] H. Rong and A. Ramirez-serrano, “Image Object Extraction Based on Semantic Detection and Improved K-Means Algorithm,” *IEEE Access*, vol. 8, pp. 171129–171139, 2020, https://doi.org/10.1109/ACCESS.2020.3025193.

[26] K. Kandali, L. Bennis, and H. Bennis, “A New Hybrid Routing Protocol Using a Modified K-Means Clustering Algorithm and Continuous Hopfield Network for VANET,” *IEEE Access*, vol. 9, pp. 47169–47183, 2021, https://doi.org/10.1109/ACCESS.2021.3068074.

[27] Z. Zhang, Q. Feng, J. Huang, Y. Guo, J. Xu, and J. Wang, “Neurocomputing A local search algorithm for k -means with outliers q,” *Neurocomputing*, vol. 450, pp. 230–241, 202, https://doi.org/10.1016/j.neucom.2021.04.028.

[28] L. Bai, J. Liang, and F. Cao, “A multiple k-means clustering ensemble algorithm to find nonlinearly separable clusters,” *Information Fusion*, vol. 61, pp. 36-47, 2020, https://doi.org/10.1016/j.inffus.2020.03.009.

[29] J. Eliyanto and Sugiyarto Surono, “Distance Functions Study in Fuzzy C-Means Core and Reduct,” *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)*, vol. 7, no. 1, pp. 118–130, 2021, https://doi.org/10.26555/jiteki.v7i1.20516.

[30] T. M. Ghazal, M. Z. Hussain, R. A. Said, and A. Nadeem, “Performances of K-Means Clustering Algorithm with Different Distance Metrics,” *Intelligent Automation & Soft Computing*, vol. 30, no. 2, 2021, https://doi.org/10.32604/iasc.2021.019067.

[31] A. Chakraborty, A. Punhani, N. Faujdar, and S. Saraswat, “Comparative Study ofK-Means Clustering Using Iris Data Set for Various Distances,” *2020 10th Int. Conf. Cloud Comput. Data Sci. Eng.*, pp. 332–335, 2020, https://doi.org/10.1109/Confluence47617.2020.9058328.

[32] R. Deepana, “On Sample Weighted Clustering Algorithm using Euclidean and Mahalanobis Distances,” *International Journal of Statistics and Systems*, vol. 12, no. 3, pp. 421–430, 2017, https://www.ripublication.com/ijss17/ijssv12n3_01.pdf.

[33] A. S. Shirkhorshidi, S. Aghabozorgi, and T. Y. Wah, “A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data,” *PLOS ONE*, vol. 10, no. 12, 2015, https://doi.org/10.1371/journal.pone.0144059.

[34] C. Sourav, “A new coefficient of correlation,” *Journal of the American Statistical Association*, vol. 116, no. 536, 2020, https://doi.org/10.1080/01621459.2020.1758115.

[35] J. Deng, Y. Deng, and K. H. Cheong, “Combining conflicting evidence based on Pearson correlation coefficient and weighted graph,” *International Journal of Intelligent Systems*, vol. 36, no. 12, 2021, https://doi.org/10.1002/int.22593.

[36] S. Kim, J. Lee, S. Jeon, M. Lee, H. An, K. Jung, S. Kim, and D. Park, “Correlation Analysis between Hydrologic Flow Metrics and Benthic Macroinvertebrates Index (BMI) in the Han River Basin, South Korea,” *Sustainability*, 2021, https://doi.org/10.3390/su132011477.

[37] E. Saccenti, M. H. W. B. Hendriks, and A. K. Smilde, “Corruption of the Pearson correlation coefficient by measurement error and its estimation, bias, and correction under different error models,” Scientific Reports, vol. 10, p. 438, 2020, https://doi.org/10.1038/s41598-019-57247-4.

[38] P. Schober, C. Boer, and L. A. Schwarte, “Correlation Coefficients: Appropriate Use and Interpretation,” *Anesthesia & Analgesia*, vol. 126, no. 5, pp. 1763–1768, 2018, https://doi.org/10.1213/ANE.0000000000002864.

[39] M. H. Hasnul Hadi, P. J. Ker, H. J. Lee, Y. S. Leong, M. A. Hannan, M. Z. Jamaludin, and M. A. Mahdi, “Color Index of Transformer Oil : A Low-Cost Measurement Approach Using Ultraviolet-Blue Laser,” *Sensors*, vol. 21, 2021, https://doi.org/10.3390/s21217292.

[40] M. Sakthi and A. S. Thanamani, “An Effective Determination of Initial Centroids in K-Means Clustering Using Kernel PCA,” *International Journal of Computer Science and Information Technologies*, vol. 2, no. 3, pp. 955–959, 2011, http://citeseerx.ist.psu.edu/viewdoc/download?rep=rep1&type=pdf&doi=10.1.1.206.2170.

BIOGRAPHY OF AUTHORS

Muhammad Ibnu Choldun Rachmatullah is a lecturer in Information Systems at Politeknik Pos Indonesia. In 1995, he completed a B. Eng in Informatics, Institut Teknologi Bandung (ITB), followed by a Master's degree in Industrial Engineering, ITB in 2001, and a Ph.D.(Informatics) in 2021 from the School of Electrical Engineering and Informatics, ITB. His research areas include Neural Networks, Machine Learning, and Soft Computing. Email: ibnucholdun@poltekpos.ac.id

paperIbnuJiteki23696-63305-1-PB.pdf

ORIGINALITY REPORT

15%
SIMILARITY INDEX

PRIMARY SOURCES

1	Submitted to Telkom University Your Indexed Documents	116 words — 3%
2	www.journaltocs.ac.uk Internet	108 words — 3%
3	Submitted to Telkom University Your Indexed Documents	84 words — 2%
4	ugoproto.github.io Internet	66 words — 2%
5	www.jatit.org Internet	34 words — 1%
6	umpir.ump.edu.my Internet	31 words — 1%
7	pen.ius.edu.ba Internet	30 words — 1%
8	trace.tennessee.edu Internet	29 words — 1%
9	static-aliyun-doc.oss-cn-hangzhou.aliyuncs.com Internet	22 words — 1%
10	www.questjournals.org Internet	

20 words — 1%

11	www.hindawi.com Internet	19 words — 1%

EXCLUDE QUOTES	ON	EXCLUDE SOURCES	< 1%
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	OFF