JURNAL ECOTIPE
Electronic, Control, Telecommunication, Information, and Power Engineering

Volume 10, Issue 1, April 2023, pp. 86-94
ISSN 2355-5068 ; e-ISSN 2622-4852
DOI: 10.33019/jurnalecotipe.v10i1.3896

K-means and K-medoids Algorithm Comparison for Clustering Forest Fire Location in Indonesia

Ichwanul Muslim Karo Karo[1], Sri Dewi[2], Mardiana[3], Fanny Ramadhani[4], Putri Harliana[5]
[1,2,4,5]Computer Science, Medan State of University, Medan, 20221, Indonesia
[3]Electrical Engineering, Medan State Polytechnic, Medan, 20221, Indonesia

ARTICLE INFO

Article historys:

Received : 01/03/2023
Revised : 01/04/2023
Accepted : 19/04/2023

Keywords:

Clustering, K-Means, K-Medoids, Feature Importance, Silhouette Coefficient

ABSTRACT

Forest fires are the most common cause of deforestation in Indonesia. This condition hurts the survival of living things. Of course, this has received special attention from various parties. One effort that can be made for prevention is to group these points into areas with the potential for fire using the clustering method. In this research, a comparative study of the clustering algorithm between K-Means and K-Medoids was conducted on hotspot location data obtained from Global Forest Watch (GFW). Besides that, important variables that affect the clustering process are also analyzed in terms of feature importance. There are nine important variables used in the clustering process, of which the Acq_time variable is the most important. The clustering quality of both algorithms is evaluated using the silhouette coefficient (SC). Both algorithms are capable of producing strong clusters. The best number of clusters is six clusters. The K-medoids algorithm is better at grouping data than K-means.

Corresponding Author:

Ichwanul Muslim Karo Karo
Compute Science, Medan State University, Medan, 20221, Indonesia
Email: ichwanul@unimed.ac.id

1. INTRODUCTION

Indonesia is one of the world's lungs, with a total area of 94.1 million hectares [1]. However, Indonesia's deforestation threat is 17% annually [2]. Forest fires are the most common cause of deforestation. Global Forest Watch (GFW) is an organization in the environmental sector that has noted Indonesia as the country most frequently affected by forest fires. The impact is that air quality in Indonesia deteriorates over time, causing damage to the world's lungs due to global warming. In addition, the destruction of forests results in the instability of flora and fauna ecosystems, so many species are threatened with extinction.

The negative repercussions of forest and land fires encourage various parties to take early prevention measures. Given that the volume of forest and land fires will increase as the dry season approaches, one of the anticipatory steps that can be taken is to predict the distribution of hotspots and classify these points into areas with the potential for forest or land fires [3]. The distribution of hotspots indicates the likelihood of forest fires' occurrence in a given area [3,4].

Hotspots are areas with higher temperatures than the surrounding surface areas. Hotspots could be detected as locations for forest and land fires. It uses the MODIS sensors on the Terra/Aqua satellite and the SNP VIIRS satellite[5]. Hotspot datasets can be grouped depending on their information similarity using data mining techniques, and these techniques can process data on a large scale [6]. One approach to data mining is clustering. The clustering algorithm will group the data into the same cluster based on their similarity.

JURNAL ECTIPE
Electronic, Control, Telecommunication, Information, and Power Engineering
Volume 10, Issue 1, April 2023, pp. 86-94
ISSN 2355-5068 ; e-ISSN 2622-4852
DOI: 10.33019/jurnalecotipe.v10i1.3896

A study applied a data mining approach in its category of clustering to identify areas susceptiblto forest fires in West Kalimantan Province [3]. The K-means algorithm is used to cluster the forest fire dataset. The dataset was obtained from the Institute of Aeronautics and Space (LAPAN). Evaluation of clustering results using the Davies Bouldin Index. Based on the results of the grouping, there are three with information on fire-prone, non-prone, and fire areas. By using identical algorithm, study by [6] clustering points of forest fires on the island of Sumatra. Forest fire hotspot data was obtained from the EOSDIS website. In addition, the study also analyzes the comparison of the K-means algorithm with the Isodata algorithm. Cluster results were evaluated using the silhouette coefficient (SC). Based on the results of their study, the K-means and Isodata algorithms are able to cluster data with very strong cluster quality, with SC values above 0.9. The results of further analysis show that the K-means algorithm is better at producing clusters, with a greater SC value than the Isodata algorithm.

Another popular clustering algorithm is K-medoids. Research by [7] clusters hotspots in the Saravan Forest, Iran. The dataset was obtained from the local Ministry of Forestry. The study also analyzes the comparison of K-medoids with the fuzzy c-means (FCM) algorithm. Clustering results are evaluated with the silhouette coefficient. Both algorithms are capable of clustering hotspots with quite a good cluster quality. By using the same algorithm, a work clusters areas with the potential for forest/land fires based on hotspots [8]. The distribution of hotspots covering the Southeast Asia region was obtained from the database of the NASA LANCE –FIRM MODIS Active Fire website. The data used contains information in the form of brightness temperature, FRP, latitude, longitude, and confidence values. To evaluate the cluster results, they use the silhouette coefficient. The K-medoids algorithm can cluster the data. However, the quality of the cluster results in the low-quality range is quite good.

Based on the conditions and facts above, this study analyzes two partitioning clustering algorithms, these are K-means and K-medoids. Both algorithms cluster forest fire hotspot data. There are two kinds of datasets, each obtained from a distinct source [1, 5]. Furthermore, this study also provides information on the most important variables in the research using the important feature.

2. RESEARCH METHOD

This section describes several research processes (Figure 1). The process begins with collecting data, followed by pre-processing. The next process entails that every dataset be clustered with the K-means and K-medoids algorithms. The silhouette coefficient is used to evaluate clustering results.

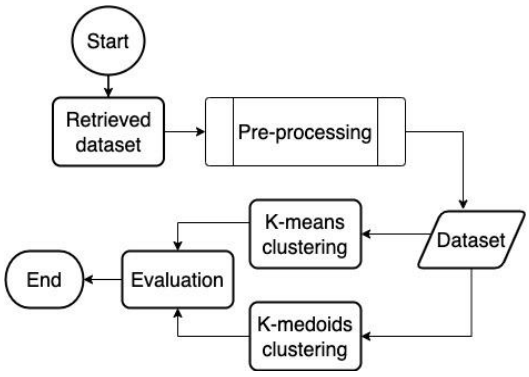


Figure 1. Research flowchart

2.2. Dataset

There are two types of datasets used in this study. Global Forest Watch (GFW) provided the first dataset (Dataset I). GFW is a data and tool platform available online to monitor forest fires [5]. The second dataset (Dataset II) is obtained from the research sampling [1]. The dataset I have 185 287 hotspot records and Dataset II consists of 25 000 records. Dataset II is a sampling from Dataset I. Both datasets have the same variable and there are twelve variables in dataset Table 1.

JURNAL ECOTIPE
Electronic, Control, Telecommunication, Information, and Power Engineering
Volume 10, Issue 1, April 2023, pp. 86-94
ISSN 2355-5068 ; e-ISSN 2622-4852
DOI: 10.33019/jurnalecotipe.v10i1.3896

Table 1. Variable description

Variable	Description
latitude	Latitude
longitude	Longitude
bright_ti4	The brightness temperature is I-4 in kelvins
scan, track	Scan size in pixels
acq_date	Track size in pixels
acq_time	Date of acquisition of VIIRS
instrument	Instrument
confidence	N= Suomi National Polar-orbiting Partnership (Suomi NPP), 1=NOAA-20 (designated JPSS-1 prior to launch)
satellite	satellite
version	Version
frp	Hot spot radiative strength
bright_ti5	The brightness temperature is I-5 in kelvins

2.3. Pre-processing

This is a vital phase of the data mining process[9]. Most of the energy of research is consumed by this process. In addition to preparing data, this process can also improve the performance of the resulting model [15]. Several procedures are executed at this stage to obtain a ready dataset. The processes are raw selection, normalization, and variable selection. Raw selection is selecting used records or deleting unused records. This study employs the principle of deleting unused records. This study removes invalid records that contain the value "X" on variable longitude and "Y" on latitude. Of course, the record deletion does not have a significant impact because the number is small in comparison to the raw dataset size.

Some values of the variables in this study dataset have non-standard ranges of values and high gaps. Therefore, process normalization is required. The data normalization process is carried out to standardize the data scale for each variable. This study uses the z-score method because the method is proven reliable for numeric and integer data types[10], [11]. Equation (1) is a z-score formula, Z notation represents normalized values, x is data, μ is the mean of data and σ is the standard deviation.

$$Z = \frac{x - \mu}{\sigma} \quad (1)$$

This study dataset has many features. These conditions make it difficult to build a model [12], [13]. This research performs feature selection using feature importance. There have been many studies using this method to solve dataset variety problems [1,12,13]. This study also used the feature importance method by threshold 0.5, which is refer to [1].

2.4. *Clustering*

Clustering is known as unsupervised learning to partition data into several clusters [14]. A cluster is a group of objects with a high degree of similarity. The quality of the clustering results is dependent on the method used. Generally, clustering algorithms are divided into three categories[4]: hierarchical, density, and partitioning. Partitioning clustering is a clustering algorithm that partitions data into k clusters [15]. The k clusters by the partitioning clustering method are frequently of higher quality than the k clusters by the hierarchical or density methods [3,14,16,17]. The popular partitioning clustering algorithms are K-means and K-Medoids [18].

2.5. K-Means Algorithm

The K-means algorithm is a non-hierarchical cluster analysis method that partitions objects into one or more groups based on similar characteristics, objects that have characteristics that are closer to being grouped in the same cluster are grouped into that cluster [19]. In other words, the K-Means algorithm goal is to minimize variation within a cluster while increasing variation with existing data in other clusters. Without knowing the target class, the learning algorithm divides the data into k clusters based on the similarity value closest to the cluster's center (centroid). This learning is included in

Volume 10, Issue 1, April 2023, pp. 86-94
ISSN 2355-5068 ; e-ISSN 2622-4852
DOI: 10.33019/jurnalecotipe.v10i1.3896

unsupervised learning [18]. The number of clusters (*k*) is assigned manually at the beginning of the clustering process. This algorithm has been widely used in any case because it is simple to implement and has a minimum computational complexity [14].

K-means algorithm
1. Define *k* as the number of clusters
2. Determine the *k* initial centroid
3. Calculated similarity between centroid to each data
4. Allocate each object to the nearest centroid based
5. Update the centroid by finding the average value of the cluster members
6. Repeat steps 3 to 5 until there is no change in the centroid

2.6. K-Medoids Algorithm

Leonard Kaufman and Peter J. Rousseeuw provided the K-Medoids algorithm, which is an enhanced version of the K-Means algorithm [11] byproduct, the two algorithms are very similar. The distinction between the K-Means and the K-Medoids algorithm is in identifying the centroid; the K-Means algorithm applies the mean value of each cluster as the centroid, whilst the K-Medoids algorithm uses data objects as representatives (medoids) as the centroid.

K-medoids algorithm
1. Initialize *k* (number of clusters) centroid
2. Allocate each data (object) to the closest centroid based on similarity
3. Choose an object randomly from the members of each cluster as a new centroid candidate
4. Calculate distance of each object to the new centroid candidate in each cluster
5. Calculate the cost (*S*) by calculating the new total distance – the old total distance
6. If *S*<0, then swap centroid by new centroid
7. Repeat step 2-7 until there is no change in centroid

2.7. Evaluation Silhouette Coefficient

This study uses the silhouette coefficient (*SC*) to evaluate clustering results. SC is a technique for determining the quality and strength of clusters. The Silhouette Coefficient Method is a hybrid of the Cohesion and Separation Methods [13]. The cohesion method measures the closeness of relationships between objects in a cluster. While the separation method determines how far or close a cluster is to other clusters. Let *A* be a cluster, randomly select an object *i* from a member of cluster A, and then the steps for calculating the silhouette coefficient are as follows [19]:

1. Calculate the mean distance from an object ($a(i)$) to all other objects in a cluster using equation (2). Where *j* is another object in cluster A and $d(i,j)$ is distance object *i* to object *j*

$$a(i) = \frac{1}{|A|-1}\sum_{j\in A, j\neq i} d(i,j) \quad (2)$$

2. Calculate the mean distance from object *i* to all data in other clusters using equation (3), and take the smallest value *b(i)*. Where $d(i,C)$ is the object's distance to all objects in cluster C, where C is not the same as *A*.

$$d(i,C) = \frac{1}{|A|}\sum_{j\in C,} d(i,j) \quad (3)$$

3. Calculate the Silhouette Coefficient value by using equation (4).

$$S(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))} \quad (4)$$

The range of SC values is 0 to 1. SC values represent cluster quality. Table 2 shows the SC description.

JURNAL ECTIPE
Electronic, Control, Telecommunication, Information, and Power Engineering
Volume 10, Issue 1, April 2023, pp. 86-94
ISSN 2355-5068 ; e-ISSN 2622-4852
DOI: 10.33019/jurnalecotipe.v10i1.3896

Table 2. Interval Silhouette Coefficient

Interval SC	Description
0.71 - 1	Strong cluster
0.51 – 0.71	Reasonable cluster
0.26 – 0.5	Weak cluster
< 0.25	Wrong cluster

3. RESULT AND DISCUSS

This chapter describes the results and discusses four main points of discussion; the results of selecting variables with feature importance, clustering using the K-Means algorithm, clustering using the K-Medoids algorithm, and studying the comparison of the two algorithms.

3.1. Feature Selection

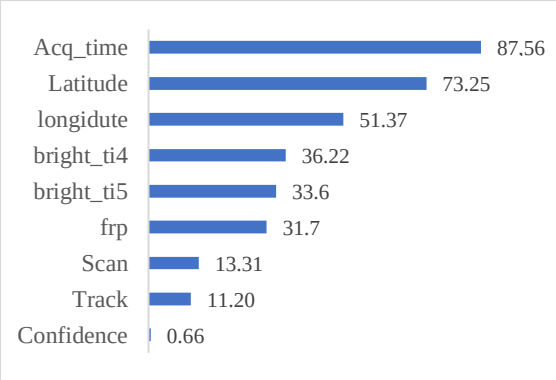


Figure 2. Feature importance value of variables

There are twelve features in this study dataset. According to the findings of this study, nine features fulfill the threshold requirements Figure 2. These are acq_time, latitude, longitude, bright_ti4, bright_ti5, FRP, scan, track, and confidence. The acq_time variable has the highest value, and the confidence variable has the lowest feature importance value. The higher the value, the greater the influence on building the model later, and vice versa. These nine variables are used to build the clustering model.

3.1. Clustering using K-Means

This section presents the results of implementing the K-means algorithm on two datasets. The K-means algorithm tests a number of clusters (k), including k = 2, 3, 4, 5, and 6, respectively. The clustering results were evaluated using the silhouette coefficient (*SC*). Evaluation results can be seen in Table 3.

Table 3. Cluster quality by K-means algorithm

k	SC Dataset I	SC Dataset II
2	0.729	0.7286
3	0.729	0.7287
4	0.777	0.776
5	0.776	0.742
6	0.794	0.754

Based on the information in the table, the K-means algorithm is able to produce strong clusters for both datasets. It means that the algorithm is capable to clusters a small or large number of records. Even though all the k tests produced strong clusters, six and four are the best number of clusters for dataset I and dataset II. In addition, k = 2 has lowest SC for both datasets, the number of cluster does not recommended as an insight of problem.

JURNAL ECTIPE
Electronic, Control, Telecommunication, Information, and Power Engineering
Volume 10, Issue 1, April 2023, pp. 86-94
ISSN 2355-5068 ; e-ISSN 2622-4852
DOI: 10.33019/jurnalecotipe.v10i1.3896

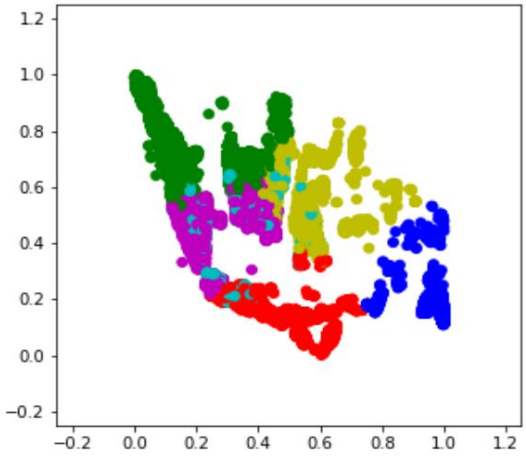


Figure 3 K-means visualization result for dataset I

The dataset I have a higher quality of clusters than Dataset II. This study argues that dataset sampling affects cluster quality. This study presents a visualization of the results of cluster dataset I depending on the attributes of latitude and longitude (Figure 3). There are six different colors, each representing a cluster. Even though the K-means algorithm produces strong clusters, there is one Tosca blue cluster with disjoint conditions between its members.

3.2. Clustering using K-Medoids

This section discusses the results of implementing the K-Medoids algorithm on both datasets. The K-medoids algorithm was also tested and evaluated using the same number of clusters and techniques that were used in the previous experiment. The evaluation results can be found in Table 4. Based on the table, $k = 2$ has the lowest SC for both datasets, while $k = 6$ has the highest SC for both datasets. It means that the number of clusters with the lowest SC is not a recommendation for grouping hotspots. On the other hand, the number of clusters with the highest SC could provide insight into grouping hotspots.

Table 4. Cluster quality by K-medoids algorithm

k	SC Dataset I	SC Dataset II
2	0.729	0.723
3	0.776	0.776
4	0.794	0.777
5	0.794	0.806
6	0.813	0.810

Based on all the numbers of k tested, K-medoid is also able to produce strong clusters for both datasets. Even though both clustering results show the same number of best clusters, the SC of the best number of clusters from dataset I is greater than the SC of the best number of clusters from dataset II. On other hand, the cluster quality of dataset I is still more reliable than dataset II. This condition reinforces the argument from this study that sampling datasets affect cluster quality. This argument is reinforced by the results of the visualization in Figure 4, the figure shows that the disjoint member of the cluster is less than the previous clustering results.

JURNAL ECTIPE
Electronic, Control, Telecommunication, Information, and Power Engineering
Volume 10, Issue 1, April 2023, pp. 86-94
ISSN 2355-5068 ; e-ISSN 2622-4852
DOI: 10.33019/jurnalecotipe.v10i1.3896

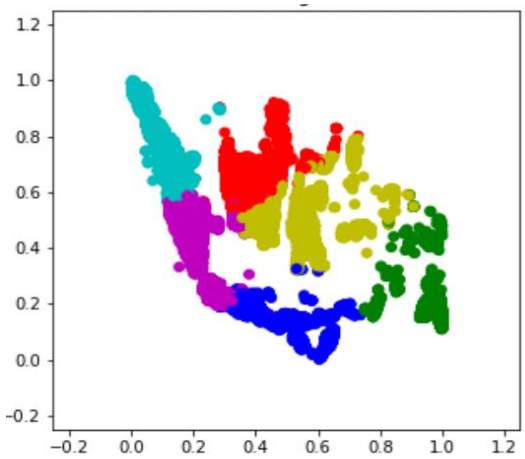
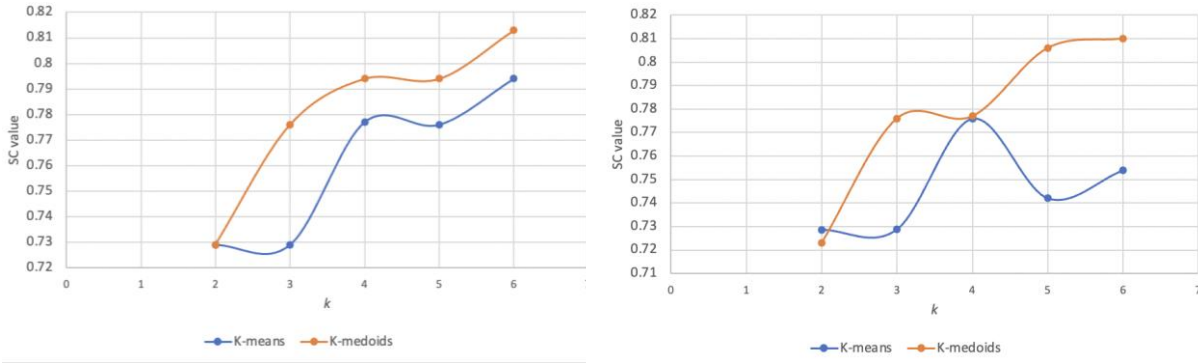


Figure 4 K-medoids visualization result for dataset I

3.3. Comparison of Two Algorithms

Based on the best cluster visualization results for both datasets in the previous section, it can be seen that the K-medoids clustering results are more compact than the K-means results. Further investigation results can be seen in Figure 5 (a) and (b). Figure 5 (a) is the SC comparison graph of algorithms for dataset I and Figure 5 (b) is the SC comparison graph of algorithms for dataset II. SC of K-medoids ≥ SC of K-means for a dataset I. The highest SC value of K-medoids is 0.813, while the K-means algorithm is less than 0.8. Similar conditions exist for dataset II, with the K-means algorithm only being superior when $k = 2$.



(a) (b)

Figure 5 Comparison SC value of algorithms

Lastly, both algorithms can produce strong clusters for both datasets. This can be seen from the SC value, which is always above 0.71. The facts show that the SC value of the K-medoids algorithm is greater than the SC value of the K-means algorithm for both datasets. It means that the performance of the K-medoids algorithm is superior to that of the K-means algorithm.

4. CONCLUSION

Based on the results and discussion, nine attributes are most influential in building a fire forest clustering model, these are Acq_time, longitude, latitude, bright_ti4, bright_ti5, FRP, scan, track, and confidence. Both algorithms can be implemented for grouping data on the location of forest/land fires in Indonesia based on the distribution of hotspots. Both algorithms produce strong clusters for large and small data sizes. Although the size of the dataset also affects the quality of the cluster, six clusters are the most numerous in both datasets. The quality of the clustering results of the K-medoids algorithm is superior to K-means.

Volume 10, Issue 1, April 2023, pp. 86-94
ISSN 2355-5068 ; e-ISSN 2622-4852
DOI: 10.33019/jurnalecotipe.v10i1.3896

REFERENCES

[1] I. M. K. Karo, "Implementasi Metode XGBoost dan Feature Importance untuk Klasifikasi pada Kebakaran Hutan dan Lahan," *Journal of Software Engineering, Information and Communication Technology*, vol. 1, no. 1, pp. 10–16, 2020.

[2] H. Wahyuni and S. Suranto, "Dampak Deforestasi Hutan Skala Besar terhadap Pemanasan Global di Indonesia," *JIIP: Jurnal Ilmiah Ilmu Pemerintahan*, vol. 6, no. 1, 2021, doi: 10.14710/jiip.v6i1.10083.

[3] N. A. Khairani and E. Sutoyo, "Application of K-Means Clustering Algorithm for Determination of Fire-Prone Areas Utilizing Hotspots in West Kalimantan Province," *International Journal of Advances in Data and Information Systems*, vol. 1, no. 1, pp. 9–16, Apr. 2020, doi: 10.25008/ijadis.v1i1.13.

[4] K. Pratama Simanjuntak and U. Khaira, "MALCOM: Indonesian Journal of Machine Learning and Computer Science Hotspot Clustering in Jambi Province Using Agglomerative Hierarchical Clustering Algorithm Pengelompokkan Titik Api di Provinsi Jambi dengan Algoritma Agglomerative Hierarchical Clustering," vol. 1, pp. 7–16, 2021.

[5] World Resources Institute, "Forest Monitoring, Land Use & Deforestation Trends | Global Forest Watch," *Global Forest Watch*. 2021.

[6] E. F. Sirat, B. D. Setiawan, and F. Ramdani, "Comparative Analysis of K-Means and Isodata Algorithms for Clustering of Fire Point Data in Sumatra Region," in *2018 4th International Symposium on Geoinformatics, ISyG 2018*, 2019. doi: 10.1109/ISYG.2018.8611879.

[7] M. Kurniawan, R. R. Muhima, and S. Agustini, "Comparison of Clustering K-Means, Fuzzy C-Means, and Linkage for Nasa Active Fire Dataset," *International Journal of Artificial Intelligence & Robotics (IJAIR)*, vol. 2, no. 2, 2020, doi: 10.25139/ijair.v2i2.3030.

[8] Laboratoire lorrain de recherche en informatique et ses applications and Institute of Electrical and Electronics Engineers, *1st IEEE International Workshop on Arabic Script Analysis & Recognition : April 3-5, 2017, LARIA, Nancy, France.*

[9] M. U. Salur and I. Aydin, "The Impact of Preprocessing on Classification Performance in Convolutional Neural Networks for Turkish Text," in *2018 International Conference on Artificial Intelligence and Data Processing, IDAP 2018*, 2019. doi: 10.1109/IDAP.2018.8620722.

[10] H. Henderi, "Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer," *IJIIS: International Journal of Informatics and Information Systems*, vol. 4, no. 1, 2021, doi: 10.47738/ijiis.v4i1.73.

[11] I. M. Karo Karo and H. Hendriyana, "Klasifikasi Penderita Diabetes menggunakan Algoritma Machine Learning dan Z-Score," *Jurnal Teknologi Terpadu*, vol. 8, no. 2, pp. 94–99, 2022.

[12] S. Khalid, T. Khalil, and S. Nasreen, "A survey of feature selection and feature extraction techniques in machine learning," in *Proceedings of 2014 Science and Information Conference, SAI 2014*, 2014. doi: 10.1109/SAI.2014.6918213.

[13] I. M. Karo Karo, S. Nadia Amalia, and D. Septiana, "Klasifikasi Kebakaran Hutan Menggunakan Feature Selection dengan Algoritma K-NN, Naive Bayes dan ID3," *Journal of Software Engineering, Information and Communication Technology*, vol. 3, no. 1, pp. 121–126, 2022.

[14] E. Schubert and P. J. Rousseeuw, "Fast and eager k-medoids clustering: O(k) runtime improvement of the PAM, CLARA, and CLARANS algorithms," *Inf Syst*, vol. 101, 2021, doi: 10.1016/j.is.2021.101804.

Volume 10, Issue 1, April 2023, pp. 86-94
ISSN 2355-5068 ; e-ISSN 2622-4852
DOI: 10.33019/jurnalecotipe.v10i1.3896

[15] M. A. Ahmed, H. Baharin, and P. N. E. Nohuddin, "Analysis of K-means, DBSCAN and OPTICS Cluster algorithms on Al-Quran verses," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 8, 2020, doi: 10.14569/IJACSA.2020.0110832.

[16] E. Lidrawati, S. Bahri, U. F. Zubaedi, V. P. Carolina, K. Kusrini, and D. Maulina, "Kebakaran Hutan Implementasi Metode CLARA Clustering Untuk Pengelompokan Data Potensi Kebakaran Hutan/Lahan Berdasarkan Persebaran Titik Panas (Hotspot)," *Journal of Computer System and Informatics (JoSYC)*, vol. 3, no. 4, 2022, doi: 10.47065/josyc.v3i4.2006.

[17] I. M. K. Karo and A. F. Huda, "Spatial clustering for determining rescue shelter of flood disaster in South Bandung using CLARANS Algorithm with Polygon Dissimilarity Function," in *Proceedings - 2016 12th International Conference on Mathematics, Statistics, and Their Applications, ICMSA 2016: In Conjunction with the 6th Annual International Conference of Syiah Kuala University*, 2017. doi: 10.1109/ICMSA.2016.7954311.

[18] S. Gultom, S. Sriadhi, M. Martiano, and J. Simarmata, "Comparison analysis of K-Means and K-Medoid with Ecluidience Distance Algorithm, Chanberra Distance, and Chebyshev Distance for Big Data Clustering," in *IOP Conference Series: Materials Science and Engineering*, 2018, vol. 420, no. 1. doi: 10.1088/1757-899X/420/1/012092.

[19] I. M. Karo Karo, A. Yusmanto, and R. Setiawan, "Segmentasi Nasabah Kartu Kredit Berdasarkan Perilaku Penggunaan Kartu Kreditnya Menggunakan Algoritma K-Means," *Journal of Software Engineering, Information and Communication Technology*, vol. 2, no. 2, pp. 101–107, 2021, [Online]. Available: https://www.kaggle.com/arjunbhasin2013/ccdata.