BAREKENG: Jurnal Ilmu Matematika dan Terapan
June 2022 Volume 16Issue 2 Page 687–694
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol16iss2pp687–694

ROBUST CLUSTERING OF COVID-19 PANDEMIC WORLDWIDE

Rizki Agung Wibowo [1], Khoirin Nisa[2*], Hilda Venelia[3], Warsono[4]

[1,2,3,4]*Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Lampung*
Prof. Dr. Soemantri Brodjonegoro St., No. 1, Bandar Lampung, 35145, Indonesia

Corresponding author's e-mail: [2] khoirin.nisa@fmipa.unila.ac.id*

Abstract. *COVID-19 pandemic is described as the most challenging crisis that humans have faced* since *World War II. From December 2019 until August 2021 based on the dataset provided by WHO, globally 219 countries in the world are affected by this virus. There are 205.338.159 cases cumulative total and 4.333.094 death cumulative total caused by this virus. In this paper, the data of 219 countries are analyzed using a robust clustering method namely K-Medoids cluster analysis. Based on the result, 219 countries in the world can be divided into five clusters based on four COVID-19-related variables, i.e. the number of cases cumulative total, death cumulative total, positive cases per capita, and case fatality rate. The distribution of the countries in five clusters was as follows; the first cluster contained 48 countries, the second cluster contained 3 countries, the third and fourth clusters contained 16 and 89 countries respectively, and the last cluster contained 63 countries. The largest cluster is the fourth one, containing countries that form a cluster with a centroid below the world average, and the smallest cluster is the second cluster with the high cases in all attributes, consisting of the USA, India, and Brazil.*

Keywords: *cluster analysis, COVID-19, K-Medoids.*

Article info:

Submitted: 22nd March 2022 *Accepted: 2nd May 2022*

How to cite this article:
R. A. Wibowo, K. Nisa, H. Venelia and Warsono, "ROBUST CLUSTERING OF COVID-19 PANDEMIC WORLDWIDE", *BAREKENG: J. Il. Mat. & Ter.*, vol. 16, iss. 2, pp. 687-694, June, 2022.

1. INTRODUCTION

In December 2019, an outbreak of pneumonia of unknown origin was reported in Wuhan, Hubei, China, this case were epidemiologically linked to the Huanan Seafood Wholesale Market [1]. The pneumonia then was known as Coronavirus Disease 19 *(*COVID-19*)* and spread across the world becoming a lethal pandemic. The effects of the pandemic were operationalized in terms of mobility, economy, and healthcare system. The mobility affected includes restructions on travel by airplane, ship and land transportation. Pandemics affect the economy in terms of demand and supply, decelerating the economic growth of affected countries, leading to reduction in trade and increase in poverty [2]. From December 2019 until August 2021 based on the dataset provided by World Health Organization (WHO), globally there are 219 countries in the world affected by this virus. There are 205338159 case cumulative total and 4333094 death cumulative total caused by this virus. To observe the spread of the pandemic among countries around the world, it is necessary to group countries with homogeneous characteristics of COVID-19 related variables as a basis for the United Nations (UN), particularly WHO, to analysis COVID-19 situation worldwide.

Cluster analysis is a statistical technique for finding groups of objects from multivariate data. The aim of cluster analysis is to construct groups with homogeneous properties out of heterogeneous large samples [3]. An assumption that must be fulfilled in performing cluster analysis is the independencies between variables [4]. However, correlation between variables commonly occurs in research data. When the variables in the data are correlated, one should handle the problem by performing principal component analysis (PCA) to obtain new uncorrelated variables called "principal components" (PCs) [5]. The PCs are built as linear combinations of the original variables.

Clustering algorithms are designed to identify an underlying structure of data and use the detected relationships within the structure to group the objects into distinct groups. One of the most commonly used algorithms among the partitioning methods in cluster analysis is the K-means algorithm [6], [7]. K-means starts by assigning K initial cluster centroids, either randomly or by an initialization algorithm. All objects are distributed into each cluster based on their distance to the centroids. The solution is refined by first electing a new cluster centroid, based on the mean values of each object in the cluster, and then redistributing the object accordingly. K-means refines the solution until changes are no longer made or until a maximum limit of iterations has been reached [8]. However, K-means is very sensitive to outliers. Even one outlier can affect the result of K-means clustering [9], [10]. Therefore a robust cluster algorithm is needed when we deal with data containing outliers. One of the efficient robust clustering techniques is the K-medoids, also simply referred to as Partitioning Around Medoids (PAM) algorithm [11].

2. RESEARCH METHODS

This research was conducted in August 2021. The data used in this research was sourced from WHO's website, with the variables used being case cumulative total (CCT), death cumulative total (DCT), positive case per capita (PCC), and case fatality rate (CFR) caused by coronavirus. The dataset was collected from December 2019 until August 2021.

Algebraically, principal components are particular linear combination of the p random variables $X_1, X_2, \dots, X_p$ [4]. Let $\boldsymbol{X}' = [X_1, X_2, \dots, X_p]$ be random vector, and $\boldsymbol{\Sigma}$ is the covariance matrix of $\boldsymbol{X}$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \dots \geq \lambda_p \geq 0$ and eigenvectors $\boldsymbol{a'_i}$, i = 1, 2, …, p. Consider the linear combinations:

$$Y_i = \boldsymbol{a_i'X} = a_{i1}X_1 + a_{i2}X_2 + a_{i3}X_3 + \dots + a_{ip}X_p \; ; \quad i = 1,2,\dots,p \tag{1}$$

with variance and covariance:

$$Var(Y_i) = \boldsymbol{a_i' \Sigma a_i} \quad ; i = 1,2,\dots,p \tag{2}$$

$$Cov(Y_i, Y_k) = \boldsymbol{a_i' \Sigma a_k} \quad ; i,k = 1,2,\dots,p \tag{3}$$

The principal components are those uncorrelated linear combinations $Y_1, Y_2, \dots, Y_p$ whose variances in (2) are as large as possible [4]. In general, the *i*-th principal component is a linear combination $\boldsymbol{a_i'X}$ which

maximizes $Var(\boldsymbol{a_i'X})$ subject to $\boldsymbol{a_i'a_i} = \mathbf{1}$ and $Cov(\boldsymbol{a_i'X}, \boldsymbol{a_k'X}) = 0$ for $k < i$. If the variables in the data have different units, the correlation matrix is used instead of the covariance matrix.

The K-Medoids clustering algorithm is also a partition-based clustering algorithm [12]. Many studies to improve K-Medoids algorithm have been done in decades (see e.g. [13]–[15]). The uses of the K-medoids have been also applied in various research fields, one can see e.g. [10], [11], [16]–[19]. The procedure of PAM algorithm can be summarized as follows:

1. Determine initial medoids
   - Calculate the distance between every pair (i, j) of all objects using Euclidean distance:

$$d(i,j) = d_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2} \tag{4}$$

   - Calculate v_i for object i as follows

$$v_i = \sum_{j=1}^{n} \frac{d_{ij}}{\sum_{l=1}^{n} d_{jl}} \qquad ; j = 1,2, \dots, n \tag{5}$$

   - Sort v_i in ascending order, then select k objects having the first k smallest values as initial medoids.
   - Obtain the initial cluster result by assigning each object to the nearest medoid.
   - Calculate the sum of distance from all objects to their medoids.
2. Update medoids
   - Find a new medoid from each cluster by minimizing the total distance to other objects in cluster.
   - Update new medoid in each cluster.
3. Assign objects to medoids
   - Set each object to the nearest medoid and obtain the cluster result.
   - Calculate the sum distance of all objects to their medoids.
   - Repeat from step 2, if the sum is equal to the previous iteration then stop algorithm [20].

The performance of cluster result is necessarily evaluated to see the level of homogenity of each cluster and determine the optimal number of clusters underlying the data. One of the most widely used statistics for cluster evaluation is the R-Squared (RS). R-Squared is computed as:

$$RS = \frac{SS_B}{SS_T} = \frac{SS_T - SS_W}{SS_T} = \frac{\left\{\sum_{j=1}^{n}(x_j - \bar{x})^2\right\} - \left\{\sum_{i=1}^{n_c}\sum_{j=1}^{r_i}(x_{ij} - \bar{x})^2\right\}}{\left\{\sum_{j=1}^{n}(x_j - \bar{x})^2\right\}} \tag{6}$$

The value of RS ranges from 0 to 1, with 0 indicating no differences among cluster and 1 indicating maximum differences among cluster.

Processed resulted are analyzed using the help of Rstudio software. The analysis procedure can be described as follow:

a) Outliers detection by using the Mahalanobis distance:

$$d_i(\boldsymbol{x_i}, \bar{x}) = \sqrt{(\boldsymbol{x_i} - \bar{x})'\boldsymbol{S}^{-1}(\boldsymbol{x_i} - \bar{x})},$$

$\boldsymbol{x_i}$ is an outlier if $d_i^2(\boldsymbol{x_i}, \bar{x}) > \chi^2_{p,1-\alpha}$, where p is the number of variables and α is a significance level with the default cut off commonly used is α=0.05.

b) Calculate the correlations among variables CCT, DCT, PCC, and CFR

c) If the variables are correlated, PCA is performed.

d) Clustering the countries using PAM algorithms based on the principal components scores.

e) Evaluate the cluster result

3. RESULTS AND DISCUSSION

The correlations between the variables are presented in Table 1. Based on Table 1, it has been shown that the correlation value between CCT and DCT is very high, as much as 0,932, although other correlations are mild, we have to perform PCA to obtain independent new variables.

Table 1. Correlation between four variables

	CCT	DCT	PCC	CFR
CCT	1	0.932	0.156	0.035
DCT	0.932	1	0.164	0.156
PCC	0.156	0.164	1	-0.126
CFR	0.035	0.156	-0.126	1

The following table shows the eigenvalues of the sample correlation matrix above. The percentage of variances contained in each eigenvalue is described graphically in Figure 1.

Table 2. Eigen value

Eigen	Eigenvalue	Variance Percent	Cumulative Variance Percent
λ_1	1.995	49.868 %	49.868 %
λ_2	1.126	28.138 %	78.006 %
λ_3	0.820	20.490 %	98.496 %
λ_4	0.060	1.504 %	100 %

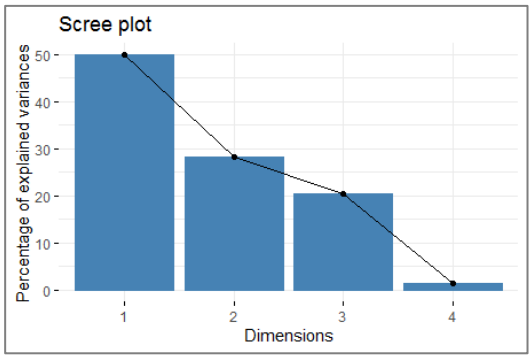

Figure 1. Scree plot of eigen values

Based on Table 2, we only use two initial eigenvalues for further analysis since both eigenvalues already contain 78 % of cumulative variance, and scree plot in Figure 1 also suggests that two components be retained. The scatter plot of objects using the first and the second Principal Component Scores (PC_1 and PC_2) is presented in Figure 2.

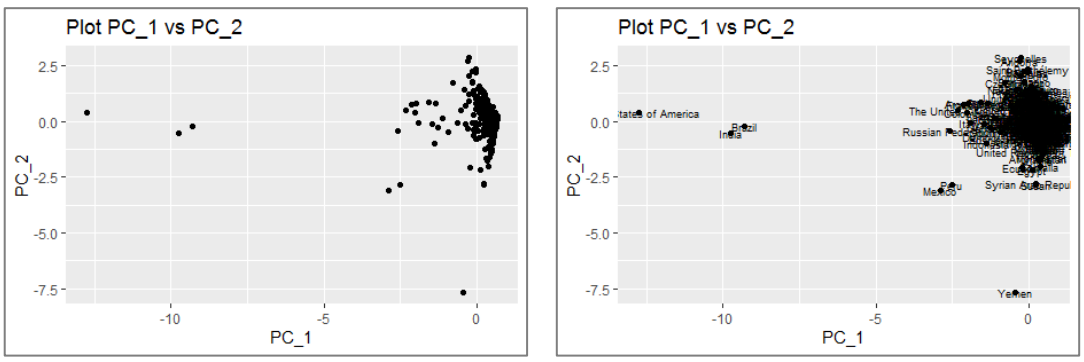

Figure 2. Plot PC_1 vs PC_2

Based on Figure 2, both variables form a stationary pattern, which means the two variables are uncorrelated, then the assumption of mutually uncorrelated variables is satisfied. However, it is indicated that there are outliers in the data. We conducted outlier detection using robust squared of Mahalanobis distance and the result is shown in Figure 3.

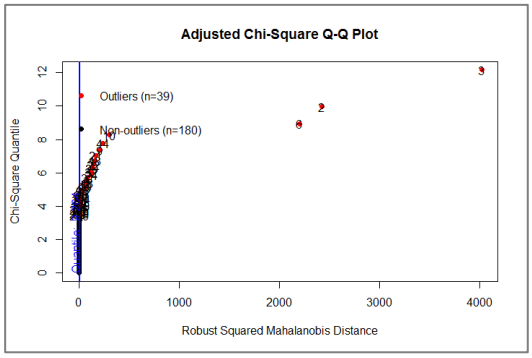

Figure 3. Outlier detection

The result shows that the number of outliers in the data is 39. Then a robust clustering algorithm is required. The robust cluster analysis used is based on the principal component scores that have been obtained. Using the K-Medoids algorithm, we obtained the optimal number of clusters according to silhouette width, which is equal to 5, as presented in the following figure 4.

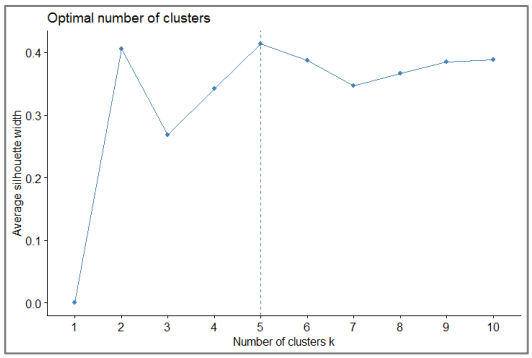

Figure 4. Optimal number of clusters

The result of the K-medoids algorithm for five clusters yields an R-Square value of as much as 0.8105, which means that 81,05 % of the characteristics between clusters are different from each other. The resulted clusters are graphically shown in Figure 5, and the members of each cluster are presented in Table 3.

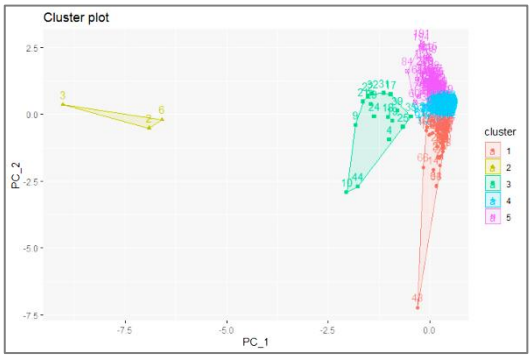

Figure 5. Cluster plot result

To determine cluster characteristics based on each variables and paid more attention to in each cluster, it is necessary to carry out a descriptive analysis. From all countries in the data, the average value of each variables are $\bar{X}_{CCT} = 933380{,}584$, $\bar{X}_{DCT} = 19755{,}571$, $\bar{X}_{PCC} = 0{,}042$ and $\bar{X}_{CFR} = 0{,}019$. These average values are compared to the cluster centers $(\bar{X}_{c,j})$, if $\bar{X}_{c,j} \leq \bar{X}_j$ (where j is the variable: CCT, DCT, PCC,

CFR), it is interpreted sequentially as "medium", "low", "very low", if $\bar{X}_{c,j} > \bar{X}_j$, it is interpreted sequentially as "high", "very high". Details of the result are described in Table 4 and Table 5.

Table 3. Cluster result

Cluster	Countries	Cluster size
Cluster 1	Madagascar, Australia, Nigeria, Mali, Malawi, Syrian Arab Republic, Guatemala, Ecuador, Senegal, Cambodia, Chad, Somalia, Zimbabwe, Tunisia, Bolivia, Haiti, Honduras, Sierra Leone, Bulgaria, Nicaragua, El Salvador, Liberia, Mauritania, Bosnia and Herzegovina, Jamaica, Gambia, Lesotho, Guinea-Bissau, Trinidad and Tobago, Eswatini, Comoros, Antigua and Barbuda, Montserrat	48
Cluster 2	India, United States of America (USA), Brazil	3
Cluster 3	Indonesia, Russian Federation, Mexico, Turkey, Iran , Germany, The United Kingdom, France, Italy, South Africa, Colombia, Spain, Argentina, Ukraine, Poland, Peru	16
Cluster 4	Nigeria, Japan, Ethiopia, Philippines, Thailand, Republic of Korea, Canada, Morocco, Saudi Arabia, Uzbekistan, Mozambique, Ghana, Nepal, Venezuela, Côte d'Ivoire, Cameroon, Sri Lanka, Burkina Faso, Romania, Kazakhstan, Zambia, Guinea, Rwanda, Benin, Burundi, South Sudan, Dominican Republic, Greece, Azerbaijan, Tajikistan, Hungary, Jersey, Papua New Guinea, Togo, Lao People's Democratic Republic, Paraguay, Libya, Kyrgyzstan, Singapore, Congo, Finland, Norway, Slovakia, Central African Republic, New Zealand, Republic of Moldova, Eritrea, Albania, Puerto Rico, Namibia, Botswana, Gabon, North Macedonia, Kosovo, Equatorial Guinea, Timor-Leste, Mauritius, Djibouti, Fiji, Guyana, Bhutan, Solomon Islands, Suriname, Brunei Darussalam, Belize, Bahamas, Iceland, Vanuatu, New Caledonia, Barbados, Sao Tome and Principe, Samoa, Saint Lucia, Guam, Grenada, Saint Vincent and the Grenadines, Dominica, Cayman Islands, Guernsey, Bermuda, Marshall Islands, Northern Mariana Islands, Greenland, Saint Kitts and Nevis, Faroe Islands, Anguilla, Wallis and Futuna, Saint Pierre and Miquelon, Falkland Islands	89
Cluster 5	Iraq, Malaysia, Chile, Netherlands, Belgium, Cuba, Czechia, Jordan, Portugal, Sweden, United Arab Emirates, Belarus, Austria, Switzerland, Serbia, Lebanon, Denmark, Oman, Palestine, Costa Rica, Ireland, Panama, Kuwait, Croatia, Georgia, Uruguay, Mongolia, Armenia, Qatar, Lithuania, Slovenia, Latvia, Bahrain, Estonia, Cyprus, Réunion, Luxembourg, Montenegro, Cabo Verde, Maldives, Malta, Guadeloupe, Martinique, French Guiana, French Polynesia, Mayotte, Curaçao, Aruba, United States Virgin Islands, Seychelles, Isle of Man, Andorra, Sint Maarten, Monaco, Saint Martin, Turks and Caicos Islands, Liechtenstein, San Marino, Gibraltar, British Virgin Islands, Bonaire, Saint Barthélemy, Holy See.	63

Table 4. Cluster Centre $(\bar{X}_{c,j})$

Cluster	CCT	DCT	PCC	CFR
1	167364.94	5315.25	0.011	0.037
2	29487418.33	536756.33	0.075	0.019
3	4300364.50	116061.13	0.066	0.031
4	182017.34	3350.18	0.020	0.012
5	363638.11	4855.95	0.091	0.011

Table 5. Cluster Characteristics

Cluster	CCT	DCT	PCC	CFR
1	Very low	Medium	Low	Very high
2	Very high	Very high	High	Medium
3	High	High	High	High
4	Low	Very low	Medium	Low
5	Medium	Low	Very high	Very low

Table 5 shows that the members in Cluster 1 have a very low COVID-19 number of cases, medium COVID-19 death cases, a low spread (postive case per capita) of COVID-19 and a very high COVID-19 fatality rate. Cluster 2 has very high confirmed and death cases, a high number of positive cases per capita and a medium fatality rate. Cluster 3 has high characteristics in all variables. Cluster 4 has a low COVID-19 cases and fatality rate, very low death cases, and medium spread. The last cluster

(Cluster 5) has medium COVID-19 cases, low death cases, very high spread, and a very low fatality rate.

4. CONCLUSIONS

In this paper, we applied a combination of robust clustering using the K-Medoids algorithm and principal component analysis to group 219 countries in the world based on the COVID-19 pandemic case. Based on the results and discussion, it can be concluded that 219 countries in the world can be divided into five clusters by using the k-medoid algorithm. Each cluster has unique characteristics and is different from the other clusters. Most countries in the world have COVID-19 related conditions below the world average. However, Madagascar, Australia, Nigeria, Mali, and 44 other countries in Cluster 1 have a very low COVID-19 number of cases but a very high fatality rate. Indonesia, Mexico, Turkey, Iran, Germany, and 11 other countries in Cluster 3 have high COVID-19 related conditions. While India, the USA, and Brazil have very high COVID-19 confirmed and death cases, they also have a high COVID-19 spread and a medium fatality rate.

AKNOWLEDGEMENT

The abstract of this paper has been presented in **16th APRU Multi-Hazards Symposium 2021** hosted by Disaster Risk Reduction Center Universitas Indonesia in collaboration with Association of Pacific Rim Universities.

REFERENCES

[1] M. Ciotti, M. Ciccozzi, A. Terrinoni, W. C. Jiang, C. Bin Wang, and S. Bernardini, "The COVID-19 pandemic," *Crit. Rev. Clin. Lab. Sci.*, vol. 57, no. 6, pp. 365–388, 2020.

[2] N. Shrestha *et al.*, "The impact of COVID-19 on globalization," *One Heal.*, vol. 11, p. 100180, 2020.

[3] J. F. Hair, W. C. Black, and R. E. Anderson, *Multivariate Data Analysis: Pearson New International Edition*, 7th ed. England: Pearson Education Limited, 2014.

[4] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, 6th ed. Pearson Education Limited, 2014.

[5] S. D. A. Larasati, K. Nisa, and N. Herawati, "Robust Principal Component Trimmed Clustering of Indonesian Provinces Based on Human Development Index Indicators," *J. Phys. Conf. Ser.*, vol. 1751, no. 1, pp. 0–8, 2021.

[6] R. Shang, B. Ara, I. Zada, S. Nazir, Z. Ullah, and S. U. Khan, "Analysis of Simple K- Mean and Parallel K- Mean Clustering for Software Products and Organizational Performance Using Education Sector Dataset," *Sci. Program.*, vol. 2021, 2021.

[7] C. Wu *et al.*, "K -Means Clustering Algorithm and Its Simulation Based on Distributed Computing Platform," *Complexity*, vol. 2021, 2021.

[8] B. Suharjo and M. S. U. Utama, "K-Means Cluster Analysis of Sex, Age, and Comorbidities in the Mortalities of Covid-19 Patients of Indonesian Navy Personnel," *JISA(Jurnal Inform. dan Sains)*, vol. 4, no. 1, pp. 17–21, Jun 2021.

[9] X. Jin and J. Han, "K-Means Clustering," in *Encyclopedia of Machine Learning and Data Mining*, C. Sammut and G. I. Webb, Eds. Boston, MA: Springer US, 2017, pp. 695–697.

[10] P. Devi and K. Kaur, "A Robust Cluster Head Selection Method Based on K-Medoids Algorithm to Maximize Network Life Time and Energy Efficiency for Large WSNs," *Int. J. Eng. Res. Technol.*, vol. 3, no. 5, pp. 1430–1432, 2014.

[11] M. A. Ramdani and S. Abdullah, "Application of partitioning around medoids cluster for analysis of stunting in 100 priority regencies in Indonesia," *J. Phys. Conf. Ser.*, vol. 1722, no. 1, p. 012097, Jan 2021.

[12] S. Vishwakarma, P. S. Nair, and D. S. Rao, "A Comparative Study of K-means and K-medoid Clustering for Social Media Text Mining," *Int. J.*, vol. 2, no. 11, pp. 297–302, 2017.

[13] R. P. A, K. S. Vani, J. R. Devi, and D. . N. Rao, "An Efficient Density based Improved K- Medoids Clustering algorithm," *Int. J. Adv. Comput. Sci. Appl.*, vol. 2, no. 6, 2012.

[14] E. Schubert and P. J. Rousseeuw, "Faster k-Medoids Clustering: Improving the PAM, CLARA, and CLARANS Algorithms," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Oct 2019, vol. 11807 LNCS, pp. 171–187.

[15] R. K. Dinata, S. Retno, and N. Hasdyna, "Minimization of the Number of Iterations in K-Medoids Clustering with Purity Algorithm," *Rev. d'Intelligence Artif.*, vol. 35, pp. 193–199, 2021.

[16] G. Ghufron, B. Surarso, and R. Gernowo, "The Implementations of K-medoids Clustering for Higher Education Accreditation by Evaluation of Davies Bouldin Index Clustering," *J. Ilm. Kursor*, vol. 10, no. 3, pp. 119–128, Jul 2020.

[17] K. Nakagawa, M. Imamura, and K. Yoshida, "Stock price prediction using k-medoids clustering with indexing dynamic time warping," *Electron. Commun. Japan*, vol. 102, no. 2, pp. 3–8, Feb 2019.

[18] R. Hajlaoui, E. Alsolami, T. Moulahi, and H. Guyennet, "An adjusted K-medoids clustering algorithm for effective stability in vehicular ad hoc networks," *Int. J. Commun. Syst.*, vol. 32, no. 12, p. e3995, Aug 2019.

[19] I. H. Rifa, H. Pratiwi, and R. Respatiwulan, "Clustering Of Eartquake Risk in Indonesia Using K-Medoids and K-Means Algorithms," *MEDIA Stat.*, vol. 13, no. 2, pp. 194–205, Dec. 2020.

[20] 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," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 420, no. 1, 2018.