Faktor Exacta

Vol. 15, No. 2, June 2022, pp. 109~116
eISSN: 2502-339X, pISSN: 1979-276X, DOI: https://doi.org/10.30998/faktorexacta.v15i2.12106

Algorithm Analysis of K-Means and Fuzzy C-Means for Clustering Countries Based on Economy and Health

Lily Wulandari[1], Bima Olga Yogantara[2]
[1,2] Information System Management,
Gunadarma University

Article Info

Article history:

Received Jan 3, 2022
Revised May 28, 2022
Accepted June 18, 2022

Keywords:

Clustering
K-Means
Fuzzy C-Means
Country
Data Mining

ABSTRAK

Teknik pembelajaran mesin tanpa pengawasan, yang membagi populasi menjadi beberapa kelompok atau klaster data ke dalam kelompok yang serupa, merupakan inti dari *clustering*. Algoritme *clustering* yang ada di antaranya algoritme K-Means dan Fuzzy C-Means. Proses *clustering* dilakukan untuk mengelompokkan negara-negara di dunia menjadi dua kategori utama, yaitu negara maju dan negara berkembang dengan kategori tingkat kesejahteraan masyarakatnya. Artikel ini membahas tentang perbandingan algoritme K-Means dan Fuzzy C-Means. Proses Algoritme K-Means menghasilkan 32 negara maju dan 135 negara berkembang, sedangkan proses Algoritme Fuzzy C-Means menghasilkan 33 negara maju dan 134 negara berkembang. Hasil analisis pengujian performa dengan parameter *Davies Bouldin Index* pada algoritme K-Means mendapatkan nilai paling kecil (lebih baik), yaitu sebesar 0.6606398 DB. Sementara itu, hasil pengujian dengan parameter *Silhouette Coefficient* pada Fuzzy C-Means adalah semakin besar nilainya (semakin baik) dan mendapatkan nilai sebesar 0.896 S. Di sisi lain, Pengujian yang cukup signifikan pada penelitian ini adalah hasil pengukuran parameter *Execution Time* pada algoritme *K-Meansi*, yakni sebesar 0.00199 detik dan prosesnya jauh lebih cepat.

Corresponding Author:
Bima Olga Yogantara
Information System Management,
Gunadarma University,
Jl. Margonda Raya No. 100, Pondok Cina, Depok 16424
Email: bimaolgayogantara19@gmail.com

1. INTRODUCTION

A country is a social group that occupies a specific area or area organized under effective political and government institutions, has political unity, is sovereign so that it has the right to determine national goals. A country classified as a developed or developing country can be seen from many factors such as population size, population growth rate, crime rate, percentage of corruption, birth and death rates, unemployment rate, inflation, number of visitors to the tourism sector, income per capita, and others [1].

The grouping of countries as developed and developing countries needs to be done with the aim of making it easier for these institutions to prioritize which countries need to be assisted first in the fields of finance, health, education, and other fields. Therefore, to make it easier to get the required data quickly and accurately, statistical analysis is needed to group countries from several indicators using clustering.

Clustering is a data mining technique that aims to group data that has similar characteristics between one data and another. Several clustering algorithms are available. This study uses the K-Means and Fuzzy C-Means (FCM) algorithms as algorithms that are compared in the process of grouping developed and developing countries. K-Means is one of the most widely implemented. This algorithm optimizes the cluster results and constantly redistributes the target dataset to each cluster center to get the optimal solution. The advantages of this algorithm lie in its simplicity, speed, and objectivity which are widely used in various research fields such as data processing, image recognition, market analysis, and risk evaluation [2]. Fuzzy C-Means is an algorithm clustering that reallocates data into each cluster by utilizing fuzzy theory. The Fuzzy C-Means method refers to how likely data can be a member of a cluster [3]. The Fuzzy C-Means method is also often used in clustering because this method gives smooth and quite effective results [4]. The advantages of Fuzzy C-Means are that it is more uncomplicated, easy to implement, can group more significant data, and is more robust against outliers [5].

There are many clustering algorithms that can be used, but each clustering algorithm has its own advantages and disadvantages. Therefore, it is necessary to compare the performance of the K-Means and FCM algorithms to determine which algorithm is better and which produces an optimal cluster. This study uses 4 parameters to compare the performance of the K-Means and FCM algorithms to measure the performance of the two algorithms. The parameters used are Davies Bouldin Index, Silhouette Coefficient, Calinski Harabasz, and Execution Time.

Several previous studies that used data mining with the K-Means and Fuzzy C-Means algorithms, namely Rio et al. compared the performance of the K-Means and Fuzzy C-Means algorithms based on the processing speed. They tracked the RMSE parameter values of each clustering to measure the level of audience satisfaction. The result is that Fuzzy C-Means can produce a more precise cluster accuracy than the K-Means modeling cluster [6]. Anissa and Yessica grouped data on employee performance values. The results were that the FCM algorithm had a better accuracy rate with a result of 76% compared to the K-Means algorithm, which had an accuracy value of 44% [7]. Aina and Nurkaromah performed a comparative analysis of the K-Means and Fuzzy C-Means algorithms with 2 cluster validity test parameters, namely Dunn-Index and Davies Bouldin-Index, to obtain optimal cluster results. The results showed that testing the K-Means algorithm resulted in a more excellent accuracy than the Fuzzy C-Means [8]. Many data clustering studies have been carried out using the K-Means and Fuzzy C-Means algorithms. So this research is based on previous research by developing some of the best methods in data clustering. Therefore, it is still very necessary to develop more broadly regarding data clustering with the topic of clustering countries based on economic and health factors using the K-Means and Fuzzy C-Means algorithms. Based on the need for information about which algorithm is better in grouping countries, this study is interested in researching and studying further about the comparison of the performance of the K-Means and Fuzzy C-Means algorithms.

2. RESEARCH METHOD

This research was conducted through several stages, namely planning, data collection, implementation, comparison test, and result analysis. The stages of the research can be seen in Figure 1.

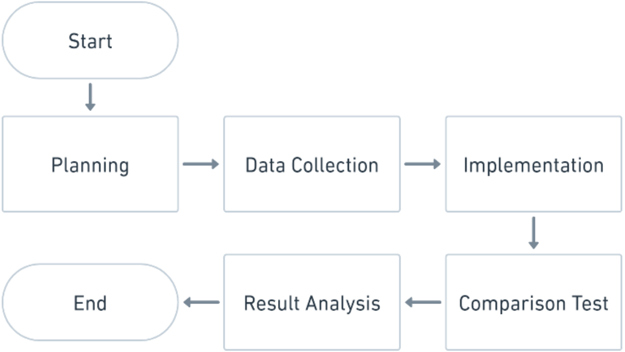

Figure 1. Research Stages

2.1. Planning

At the planning stage, it is necessary to determine the system requirements and the scope of the system being developed. In this study, a system was developed that performs data clustering analysis. This clustering analysis performs grouping of countries in the world into groups of developed or developing countries. The scope of the analysis is carried out on child_mort, exports, health, imports, income, inflation, life_expec, total_fer, and gdpp data in 167 countries in the world. This clustering analysis uses the K-Means and Fuzzy C-Means algorithms while at the same time a comparison of the performance of the two algorithms is carried out using RStudio with the R programming language.

2.2. Data Collection

The data is obtained from Kaggle which was created in 2020. The data contains 9 attributes including child_mort, exports, health, imports, income, inflation, life_expec, total_fer, and gdpp in 167 countries in the world with the extension .csv (comma-separated-value).

2.3. Implementation

At the implementation stage, it begins with data collection. The data is first processed in the data cleaning stage. After that, determining the number of clusters, implementing the K-Means and Fuzzy C-Means algorithms, then testing the performance comparison between the two algorithms. An overview of the implementation stages is shown in Figure 2.

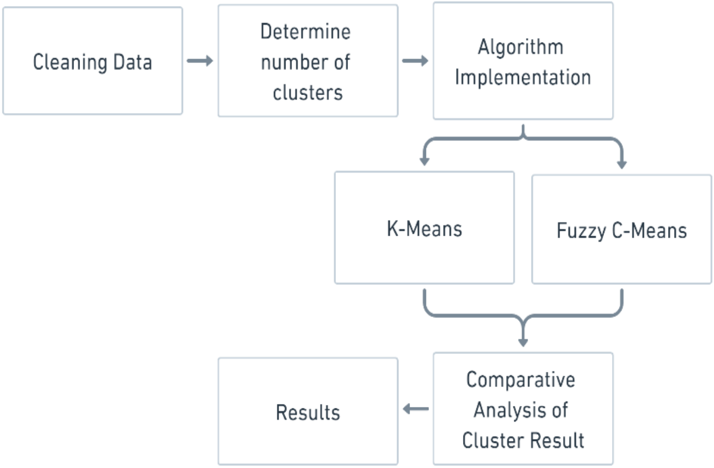
Cleaning Data
Determine number of clusters
Algorithm Implementation
K-Means
Fuzzy C-Means
Comparative Analysis of Cluster Result
Results

Figure 2. Implementation Phase

2.3.1. Cleaning Data

At this stage, data cleaning is carried out to clean data from incomplete data, incorrect formats, and remove attributes that do not affect the calculation so as to produce high-quality clustering. The countrywide economic and health data obtained from Kaggle could not be used directly for this study. Data mining provides optimal results if the data processed is good. Good means data without noise, clean from empty values, and does not contain errors.

2.3.2. Determine Number of Clusters

Determination of the number of clusters in this study using the Elbow and Silhouette Coefficient methods. The search for the optimum k value is done by comparing the SSE (Sum of Square Error) values which are presented in graphical form.

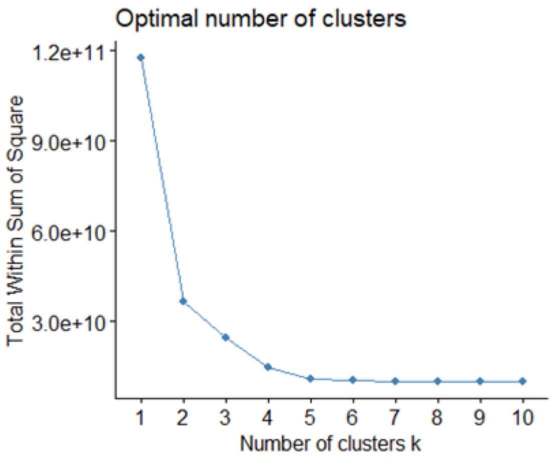

Figure 3. Elbow Method

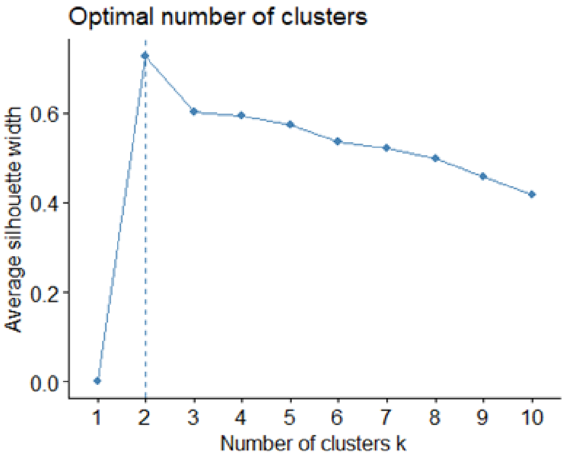

Figure 4. Silhouette Coefficient

The optimum k value is obtained when fault conditions are found on the graph. Based on Figure 3 and figure 4, it can be seen that the drastic decrease and increase are located at the value of k = 2, where the point shows the number 2, which means that the results of the implementation of the Elbow and Silhouette Coefficient methods produce 2 clusters. The number of clusters generated by the two methods is then implemented in clustering as the number of clusters used.

2.3.3. Implementation of K-Means Algorithm

At this stage, a clustering model is made using the K-Means algorithm. The application of K-Means is carried out in several stages as shown in Figure .

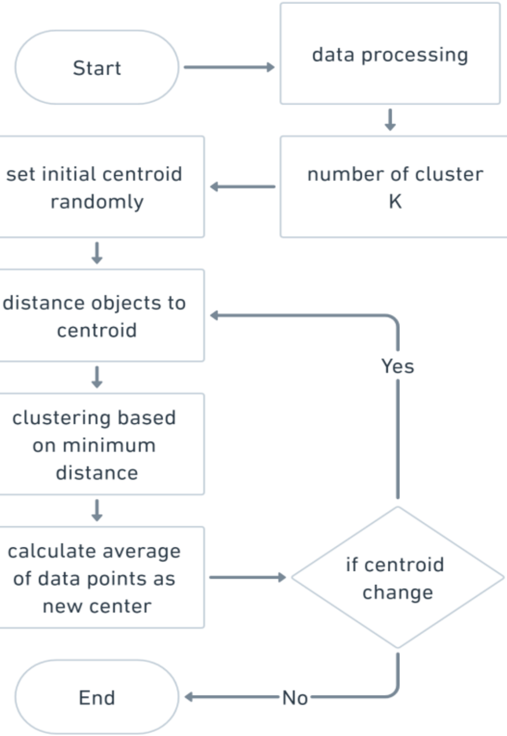
Start
data processing
number of cluster K
set initial centroid randomly
distance objects to centroid
clustering based on minimum distance
calculate average of data points as new center
if centroid change
Yes
No
End

Figure 5. K-Means Algorithm Stages

The following is an explanation of each stage of the application of the K-Means algorithm :

1. Determine the optimal number of clusters using the Elbow and Silhouette Coefficient methods.
2. Determine the cluster center point or centroid randomly.
3. Calculate the distance of each data with the center of the cluster using the Euclidean Distance formula.

$$D_{(ij)} = \sqrt{\sum_{i=1}^{p} | X_{ki} - X_{kj} |^2}$$

4. Grouping data based on the distance to the nearest cluster.
5. Calculate the center point of the new cluster using the average data distance from the cluster center point.

$$C_k = \frac{\Sigma\, d_i}{n_k}$$

6. Compare the new cluster with the initial cluster. If the newly formed cluster has a different centroid from the initial cluster, repeat the step 3 process again. If the new cluster centroid is the same as the previous cluster, the process can be stopped, and the final cluster result is obtained.

2.3.4. Implementation of Fuzzy C-Means Algorithm

At this stage, a clustering model is made using the Fuzzy C-Means algorithm. The application of Fuzzy C-Means is carried out in several stages as shown in Figure 4.

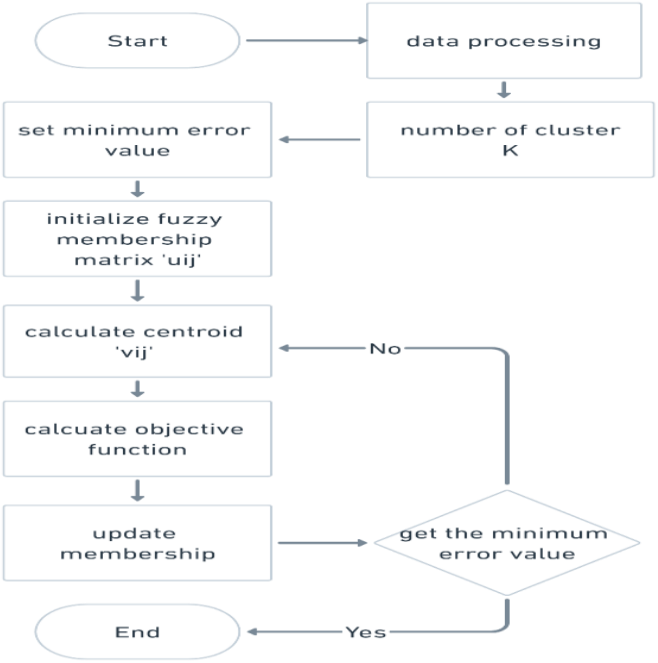

Figure 4. Fuzzy C-Means Algorithm Stages

The following is an explanation of each stage of the application of the Fuzzy C-Means algorithm :

1. Determine the optimal number of clusters using the Elbow and Silhouette Coefficient methods.
2. Specifies the minimum error value as the value constraint at which the loop ends.
3. Initialize the membership matrix U = $\{u_{ij}\}$ randomly as the elements of the initial membership matrix.
4. Calculating the cluster centre (V) where the matrix element V is v_{kj} can be calculated using the following equation.

$$v_{kj} = \frac{\Sigma_{j=1}^{n}(\mu_{ik})^{m}\, X_{ij}}{\Sigma_{j=1}^{n}(\mu_{ik})^{m}}$$

5. Calculating the value of the objective function used to get the error value, can be calculated using the following equation.

$$FO = \sum_{i=1}^{n}\sum_{k=1}^{c}([\sum_{j=1}^{m}(X_{ij} - V_{kj})^{2}]^{(\mu_{ik})^{w}})$$

6. Calculate the change in the membership matrix using the following equation.

$$\mu_{ik} = \frac{[\Sigma_{j=1}^{m}(X_{ij} - V_{kj})^{2} \quad]^{\overline{w-1}^{\,-1}}}{\Sigma_{k=1}^{c}(X_{ij} - V_{kj})^{2} \quad]^{\overline{m-1}^{\,-1}}}$$

7. If certain conditions have not been reached, repeat the process of step 4 and so on. If the minimum error value has been obtained, the process can be stopped, and the final cluster result is obtained.

2.4. Comparative Analysis of Two Algorithm

The performance measurement of the K-Means and Fuzzy C-Means algorithms is carried out using several parameters, namely Davies Bouldin Index, Silhouette Coefficient, Calinski Harabasz, and Execution Time. Measurements using RStudio.

The Davies Bouldin Index value describes how well the cluster is formed. The smaller the Davies Bouldin Index value or the closer the value to 0 indicates how good the cluster is [9]. The Silhouette Coefficient value is in the range of -1 to 1. The higher the value, the better the quality of the algorithm's performance [10]. Calinski Harabasz parameter which indicates the optimal number of clusters if the resulting value is greater or the highest value [11]. Furthermore, execution time testing where the smaller the execution time in the clustering process, the better the algorithm's performance.

3. RESULTS

This section discusses the results of the implementation of the K-Means and Fuzzy C-Means algorithms using four assessment parameters, namely the Davies Bouldin Index, Silhouette Coefficient, Calinski Harabasz, and Execution Time. This clustering process uses the K-Means and FCM library. The library generates 2 cluster center points with the value of each attribute, then generates clusters for each country. Then the number of countries in each cluster is calculated. The last stage is a visualization of the results of the clustering in the form of a graph that has 2 central points.

3.1. K-Means Clustering Results

After getting the optimal number of clusters, the next step is to process data clustering using K-Means. The results of data clustering on the K-Means algorithm are visualized in Figure 5.

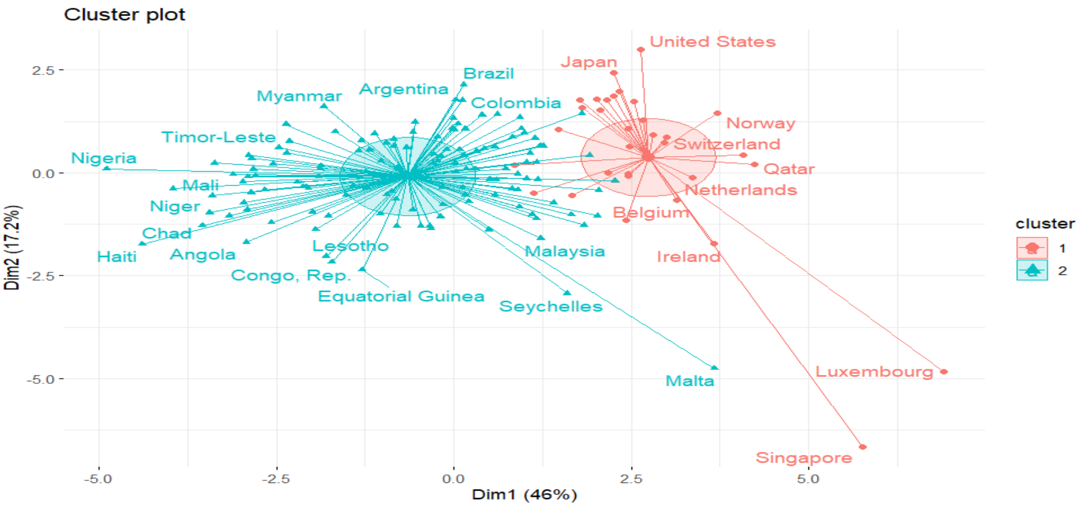

Figure 5. K-Means Clustering Result

Based on Figure 5, it is known that the results of clustering visualization using the K-Means algorithm are in the form of a graph that has 2 central points where the green data are countries that are included in cluster 2, namely developing countries, and red data are included in cluster 1, namely developed countries. So that the data obtained as in Table 1.

Table 1. K-Means Clustering Result

Variable	Cluster 1 (Developed Country)	Cluster 2 (Developing Country)
Number of Countries	32	135

The results of the K-Means clustering are 32 country data connected with centroid 1 which has a high economic and health index which is grouped into developed countries. And 135 country data is connected to centroid 2 which has low economic and health index values which are grouped into developing countries.

3.2. Fuzzy C-Means Clustering Results

The results of data clustering on the Fuzzy C-Means algorithm are visualized in Figure 6.

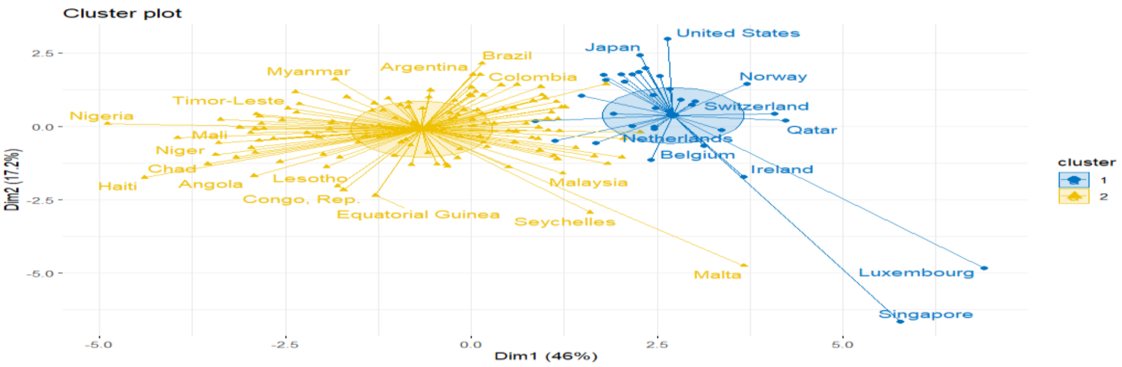

Figure 6. Fuzzy C-Means Clustering Result

Based on Figure 6, it is known that the results of clustering visualization using the Fuzzy C-Means algorithm are in the form of a graph with 2 central points where the blue data are countries that are included in cluster 2, namely developing countries, and yellow data are included in cluster 1, namely developed countries. So that the data obtained as in Table 2.

Table 2. Fuzzy C-Means Clustering Result

Variable	Cluster 1 (Developed Country)	Cluster 2 (Developing Country)
Number of Countries	33	134

The results of the Fuzzy C-Means clustering are 33 country data connected with centroid 1 which has a high economic and health index which is grouped into developed countries. And 134 country data is connected to centroid 2 which has low economic and health index values which are grouped into developing countries.

3.3. Clustering Algorithm Performance Test

After getting the results of clustering, at this stage the performance test of the K-Means and Fuzzy C-Means algorithms is carried out using 4 parameters, namely Davies Bouldin Index, Silhouette Coefficient, Calinski Harabasz, and Execution Time, the data obtained are as follows.

Table 3. Clustering Algorithm Performance Result

Parameter Performance Test	K-Means Algorithm	Fuzzy C-Means Algorithm
Davies Bouldin Index	0.6606398	0.6656474
Silhouette Coefficient	0.726	0.896
Calinski Harabasz	354.423142946443	354.423142958424
Execution Time	0.001996	0.363101

It can be concluded from Table 3, that there is no significant difference in results between the K-Means and Fuzzy C-Means algorithms. The smaller the Davies Bouldin Index value generated in the data clustering process, the better the performance of the clustering algorithm used. The Silhouette Coefficient value is in the range -1 to 1. The higher the value, the better the performance of the algorithm used. The Calinski Harabasz value indicates the optimal number of clusters if the resulting value is greater or the highest value. The last parameter used to measure the performance of the clustering algorithm is Execution Time. The faster the execution time, the better the performance of the algorithm used.

In the performance test, K-Means has advantages in measuring the Davies Bouldin Index and Execution Time parameters, while Fuzzy C-Means has advantages in Silhouette Coefficient parameters. The value of Calinski Harabasz's parameter testing of the K-Means and Fuzzy C-Means algorithms has a value that does not show a significant difference. Because the two algorithms have similar values at a glance, it means that both of them perform equally well in clustering.

3.4. Summary of Comparison Results

Based on the results of the analysis carried out in sub-chapter 3.3, it can be summarized as shown in Table 4. In Table 4 the contents of the K-Means and Fuzzy C-Means columns are 1 or 0. A value of 1 means that the algorithm is better than the others.

Table 4. Summary of Comparison Results

Parameter Performance Test	K-Means Algorithm	Fuzzy C-Means Algorithm
Davies Bouldin Index	1	0
Silhouette Coefficient	0	1
Calinski Harabasz	1	1
Execution Time	1	0
Total	**3**	**2**

Table 4 shows the performance test values for the K-Means and Fuzzy C-Means algorithms using the Davies Bouldin Index, Silhouette Coefficient, Calinski Harabasz, and Execution Time parameters. Based on these four parameters, it can be concluded that the K-Means algorithm gets a better total test value than the Fuzzy C-Means algorithm in conducting clustering in this study.

4. CONCLUSION

Based on the results of the research conducted, the following conclusions can be drawn:

1. The K-Means and Fuzzy C-Means methods have succeeded in classifying developed and developing countries based on economic and health factors properly. Determination of the optimal number of clusters using the Elbow and Silhouette Coefficient methods resulted in 2 clusters. The K-Means algorithm produces 32 developed countries and 135 developing countries. While the Fuzzy C-Means algorithm produces 33 developed countries and 134 developing countries.
2. The results of the analysis of performance testing using the Davies Bouldin Index and Execution Time parameters on the K-Means algorithm have better results than the Fuzzy C-Means algorithm. While the Silhouette Coefficient parameter on Fuzzy C-Means has optimal results and is better than K-Means. However, the performance test using the Calinski Harabasz parameters of the K-Means and Fuzzy C-Means algorithms has the same value at a glance, meaning that both of them perform equally well in clustering. The K-Means algorithm is far superior to the Fuzzy C-Means as seen from the very fast Execution Time parameter measurement results of 0.001996 seconds. Based on the results of this study, the K-Means algorithm is a very effective algorithm used in clustering data.
3. The results of this clustering research can still be developed into a knowledge base for mapping decision support systems in developed and developing countries based on economic and health data.

REFERENCES

[1] U. A. Gani, S. R, R. Bambang, and K. Umam, “Analisis Diskriminan untuk Mengelompokkan Negara Maju dan ANALISIS Diskriminan Untuk Mengelompokkan Negara Maju Dan Negara Berkembang Dengan Metode Fishers Discriminant,” *Negara Berkembang dengan Metod. Fish.*, vol. 01, no. 01, pp. 1–12, 2018, [Online]. Available: http://www.journal.geutheeinstitute.com.

[2] P. M. Shakeel, S. Baskar, V. R. S. Dhulipala, and M. M. Jaber, “Cloud based framework for diagnosis of diabetes mellitus using K-means clustering,” *Heal. Inf. Sci. Syst.*, vol. 6, no. 1, pp. 1–7, 2018, doi: 10.1007/s13755-018-0054-0.

[3] A. F. Lestari and M. Hafiz, “Penerapan Algoritme Apriori Pada Data Penjualan Barbar Warehouse,” *INOVTEK Polbeng - Seri Inform.*, vol. 5, no. 1, p. 96, 2020, doi: 10.35314/isi.v5i1.1317.

[4] P. Mcbrien, “Department Of Computing AutoPig,” no. January 2001, 2013.

[5] A. S. Rizal and R. F. Hakim, “Metode K-Means Cluster Dan Fuzzy C-Means Cluster (Studi Kasus: Indeks Pembangunan Manusia Di Kawasan Indonesia Timur Tahun 2012),” *Pros. Semin. Nas. Mat. dan Pendidik. Mat. UMS 2015*, pp. 643–657, 2015, [Online]. Available: https://publikasiilmiah.ums.ac.id/xmlui/handle/11617/5803.

[6] D. S. Tv, “Rio Andika Malik, 2) Sarjon Defit , 3) Yuhandri,” vol. 3, no. 1, pp. 10–21, 2018.

[7] A. E. Pramitasari and Y. Nataliani, “Perbandingan Clustering Karyawan Berdasarkan Nilai Kinerja Dengan Algoritme K-Means Dan Fuzzy C-Means,” *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 8, no. 3, pp. 1119–1132, 2021, doi: 10.35957/jatisi.v8i3.957.

[8] A. L. R. Putri and N. Dwidayati, “Analisa Perbandingan K-Means Dan Fuzzy C-Means Dalam Pengelompokan Daerah Penyebaran Covid-19 Indonesia,” *UNNES J. Math.*, vol. 10, no. 2, pp. 4–7, 2021, [Online]. Available: http://journal.unnes.ac.id/sju/index.php/ujme.