International Journal of Academic and Applied Research (IJAAR)
ISSN: 2643-9603
Vol. 8 Issue 2 February - 2024, Pages: 29-34

Clustering Countries in the World Based on Welfare Indicators Using Non-Hierarchical Analysis

[1]M. Fariz Fadillah Mardianto; [1]Dita Amelia; [1]Elly Ana; [1]Nur Azizah; [1]Firqa Aqila Hizbullah; [1]Michelle Adelia Suwarno; [1]Raja Van Den Bosch Sihotang; [1]Nuzulia Anida

[1]Statistics Program, Department of Mathematics, Faculty of Science and Technology, Airlangga University, Surabaya, Indonesia

**Corresponding author's e-mail: m.fariz.fadillah.m@fst.unair.ac.id*

Abstract: *This research focuses on clustering 167 countries based on welfare indicators using a non-hierarchical cluster analysis method, specifically the K-Medoids algorithm. Welfare, which includes economic, social, health, environmental, and security aspects, is the main goal of every country's development. In this study, welfare indicators such as health, exports, imports, and income are used as parameters for clustering. The K-Medoids method is used to cluster countries based on the similarity of welfare indicator characteristics. Validation of the cluster results was done using the Silhouette and Dunn indices. The results showed that there were two optimal clusters with a Silhouette index value of 0.2846391 and a Dunn index of 0.0666. This research is expected to provide new insights into how countries can be grouped based on welfare indicators and contribute to efforts to improve global welfare.*

Keywords: K-Medoids; Non-Hierarchical Cluster; Silhouette; Sustainable Development Goals; Welfare Indicator

1. INTRODUCTION

The relationship between globalization and the welfare state is complex, and understanding the implications of the ongoing process of economic internationalization for the long-term sustainability of the welfare state is a highly relevant topic. Certain aspects of economic internationalization (trade openness, financial liberalization, or foreign direct investment) may be more important than others in relation to welfare state policies. Similarly, the "welfare state" today encompasses and connects a large number of policy areas, and some parts of the welfare state are likely to be affected differently by globalization than others [2]. It is imperative that all nations comprehend global concerns that pose a danger to welfare and investigate potential remedies. The United Nations has established the Sustainable Development Goals as one of the practical measures to enhance global welfare. SDG No. 3 underscores the significance of optimal health and welfare for individuals of all ages. By paying attention to these aspects in the era of globalization, it is expected to create a holistic solution to improve welfare around the world.

The welfare of a country can be seen from various indicators, such as exports, imports, inflation, and others. For developing countries, analyzing the development based on these indicators is an important step as an evaluation to determine the direction of development and the focus of improvements needed to achieve developed country status [5]. For developed countries, this cluster analysis can be useful to see factors that can still be improved so that the welfare of the country can continue to increase [5].

Cluster analysis is a set of multivariate techniques whose main objective is to group items, objects or individuals (here: EU and VG countries) based on their characteristics. The basic criterion used to group objects is their similarity. In this way, objects belonging to one cluster are similar to each other in terms of the variables measured in it [13]. There are two types of cluster analysis, hierarchical and non-hierarchical. Hierarchical methods do not cluster data directly like non-hierarchical methods, but use grouping or division to gradually assemble or disassemble the data points into clusters [1]. The K- Medoids algorithm starts by randomly determining the initial cluster center (medoid), then calculates the distance of objects to the medoid and forms a new cluster. This process is repeated until there is no difference with the already formed medoid [4].

Some previous studies that have been carried out are grouping provinces in Indonesia based on education indicators using the K-Medoids method by Astrika et al. in 2021 [3], grouping cities or districts based on poverty indicators in East Java in 2020 using the K-Medoids method by Febiyanti et al. in 2022 [7], and grouping the spread of covid-19 in Indonesia using the K-Medoids method by Sukma et al. in 2020 [8]. Based on previous research and these problems, the author wants to analyze data on countries in the world with multivariate analysis methods, namely non-hierarchical clusters. This research aims to apply the K-Medoids method to 167 countries' welfare indicator data, hoping to provide new insights into how these countries can be grouped based on these indicators [4]. It is hoped that the results of this study can contribute to efforts to improve the welfare of countries in the world.

2. METHODS

2.1 Literature Review

Cluster analysis is a statistical analysis technique used to place a collection of objects into two or more groups based on similar characteristics of the objects [10]. There are two types of cluster analysis methods: hierarchical and non-hierarchical. Cluster analysis with the non-hierarchical method is one of the

cluster analysis methods used for grouping objects, where the number of clusters to be formed can be specified in advance. There are several types of cluster analysis with non-hierarchical methods, one of which is the K-Medoids method. K-Medoids works with representative values for each cluster. K-Medoids works with representative values for each cluster. In the K-Medoids technique, the representative value is selected based on the cluster's center value [4]. Medoids are items inside a cluster-representing set of things. This method is also known as Partitioning Around Medoids (PAM). This method overcomes perhaps the biggest disadvantage of the K-Mean method, i.e., the sensitivity to extreme values (outlier) [4]. Cluster analysis using the K-Medoids method can be done to group countries based on their welfare levels. The purpose of this clustering is to find out the right policy to be carried out by a group of countries that have low welfare.

Welfare is something that every country in the world wants to achieve. The goal of a country is to develop in a positive direction so that the welfare of the country and its people can be more secure. The welfare of a country can be measured and interpreted by various indicators. These indicators can be in the form of birth, death, health, export, import, and other indicators. According to Fajar, the welfare level of a country's population is generally measured using GDP per capita. This research supports the idea that an increase in GDP per capita indicates an increase in the welfare level of a country's population. GDP per capita can be used to compare welfare levels between countries in the world [11]. Another study by Ahmad explains that welfare levels can be measured by per capita income [7]. Research by Kate and Richard explains that the measure of child welfare depends on the average income of the country. Countries with high income disparities tend to have high infant mortality rates [6]. This research shows that indicators of state welfare are not only limited to the economy, but also include other aspects such as health and social.

2.2 Data Source

The data source utilized in this research is secondary data acquired from Kaggle with the last update in June 2023. The data obtained includes welfare indicators for 167 countries in various continents, namely Asia, Africa, Australia, South America, and North America, under the title 'Country Grouping Data'.

2.3 Research Variables

The variables in the grouping of 167 countries in the world with the welfare indicators used consist of 9 indicators as research variables. Table 1 displays the research variables.

Table 1: Research Variables

No.	Variables	Description
1	X_1	Child Mortality
2	X_2	Exports
3	X_3	Health
4	X_4	Imports
5	X_5	Income
6	X_6	Inflation
7	X_7	Life Expectancy
8	X_8	Total Fertility Rate
9	X_9	GDP

2.4 Analysis Procedure

This research was conducted using K-Medoids Clustering to group 167 countries in the world based on their welfare indicators. Figure 1 is a flowchart of the research steps.

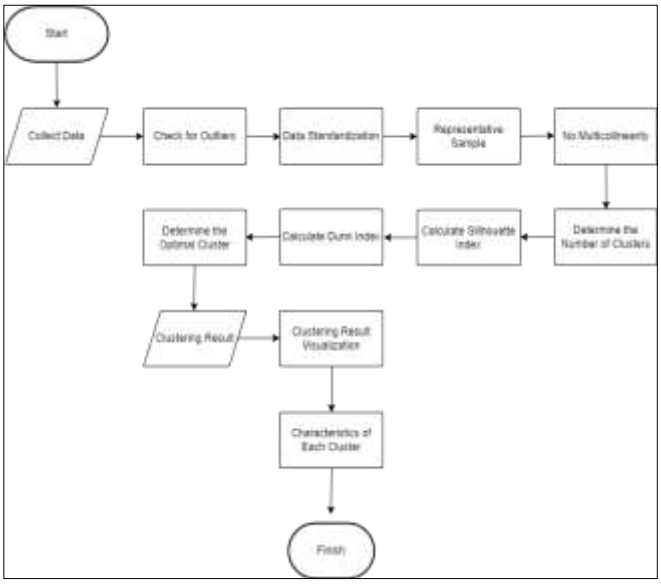

Fig. 1. *Research Flowchart*

The steps of using the K-Medoids algorithm are as follows:

- Perform outlier detection.
- Data standardization to equalize data units
- Assumption tests in cluster analysis are
  a. Test the assumption that the sample represents the population (representative) with the Kaiser Mayer Olkin (KMO) test.
  b. Test the assumption of non-multicollinearity using the Variance Inflation Factor (VIF) value.
- Determining the number of clusters to be created (k), the value of k used in this study is k = 2, 3 and 4.
- Find the distance of each object to the nearest cluster using the euclidean distance measure. Euclidean Distance is the distance between points in a straight line. This distance uses the Pythagorean theorem [9].

$$d_{ab} = \sqrt{\sum_{k=1}^{p} (x_{ak} - x_{bk})^2}$$

With

p : number of variables
d_{ab} : distance between object a and object b
x_{ak} : object value a of variable k
x_{bk} : object value b of variable k

- Validation of clustering results
  a. Calculate the silhouette coefficient and dunn index on each cluster formed with a distance of euclidean.
  b. Compare the silhouette coefficient and dunn index values of k = 2, 3 and 4 with the euclidean distance. The results of the silhouette coefficient calculation are in the range -1 to 1. The greater the silhouette coefficient value, the better the group value. And the higher the Dunn Index value, the more optimal the number of clusters produced.
- Perform K-Medoid clustering to determine each cluster member and visualize cluster results.
- Interpretation and profiling of regional characteristics from the best clustering results

3. RESULTS AND DISCUSSION

3.1 Descriptive Statistics

Descriptive statistics are used to provide an overview related to research variables such as seeing the minimum value, maximum value, and average value. Table 2 displays the subsequent descriptive statistics findings.

Table 2: Descriptive Statistics

Var	Minimum		Maximum	
	Value	Country	Value	Country
X_1	2.60	Iceland	208.00	Haiti
X_2	0.11	Myanmar	200.00	Singapore
X_4	0.07	Myanmar	174.00	Singapore
X_5	609.00	Congo, Dem. Rep.	125000.00	Qatar
X_6	-4.21	Seychelles	104.00	Nigeria
X_7	32.10	Haiti	82.80	Japan
X_8	1.15	Singapore	7.49	Niger
X_9	231.00	Burundi	105000.00	Luxembourg

3.2 Outlier Detection

Outlier detection uses a boxplot to determine whether the data used has outlier data. If there is data or points that come out of the boxplot, then the data used has outlier data.

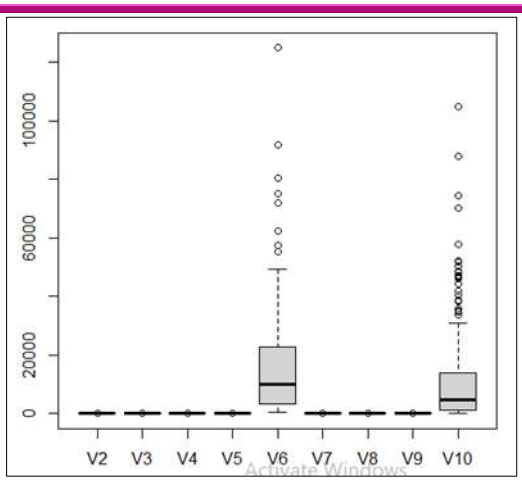

Fig. 2. *Output of outlier checking*

From Figure 2, it is known that there are points that come out of the boxplot in V6 and V10 which indicates the presence of outliers. So the K-Medoids Clustering method is appropriate for this analysis because it is not sensitive to the presence of outliers.

3.3 Data Standardization

Data standardization is carried out because the units of each variable in the data are not the same. The purpose of data standardization is to avoid bias in the model. Table 3 displays the outcomes of the data standardization process.

Table 3: Data Standardization

Country	Afganistan	Albenia	…	Yemen	Zembia
Childmort	1.29	-0.54	…	0.45	1.11
Export	-1.14	-0.48	…	-0.41	-0.15
Health	0.28	-0.10	…	-0.60	-0.34
Import	-0.08	0.07		-0.52	-0.66
Income	-0.81	-0.37	…	-0.66	-0.72
Inflation	0.16	-0.31	…	1.50	0.59
Life expect	-1.61	0.65	…	-0.34	-2.09
Total fer	1.90	-0.86	…	1.14	1.62
GDP	-0.68	-0.48	…	-0.64	-0.63

3.4 Assumption Test

There is an assumption test before conducting cluster analysis, which is as follows:
1. Representative Sample

Table 4: KMO and Barlett's Test Results

Kaiser – Meyer – Olkin Measure of Sampling Adequacy	0.678

From Table 4, KMO score of 0.678 indicates that there is enough data to meet the sample adequacy requirements.

2. No Multicollinearity

Multicollinearity can be seen from the VIF value of each variable. A VIF value of ≥10 indicates that there is multicollinearity.

Table 5: VIF Value for Each Variable

Variables	VIF Value
X_1	7.366745
X_2	4.940232
X_3	1.769698
X_4	3.730589
X_5	7.585582
X_6	1.285318
X_7	5.942367
X_8	3.739482
X_9	7.466174

3.5 Optimal Cluster Determination

Based on the results of the analysis of clustering countries with the K-Medoids method and validation test of cluster results with silhouette index and dunn index for k = 2, 3, and 4 and euclidean distance. Then a comparison of validation values is made to get the best clustering, which can be seen in Table 6.

Table 6: Comparative Value of K-Medoids Clustering Validation

Clusters	Silhouette Index	Dunn Index
2	0.2846391	0.0666
3	0.2810418	0.0600
4	0.2054321	0.0633

Based on Table 6, it is found that clustering countries in the world with the K-Medoids method using two validations, the best clustering is obtained, namely :

1. For Silhouette Index validation, the maximum validation value is obtained at k = 2 using the euclidean distance. which is 0.2846391
2. For Dunn Index validation, the maximum validation value is obtained at k = 2 clusters with an Euclidean distance of 0.0666.

Based on clustering with two validations of silhouette index and dunn index, it can be concluded that the best clustering results with these two validations based on country clustering data with the K-Medoids method are clustering at k = 2 and euclidean distance measurement, because it has the most optimal validation index value in each validation.

3.6 Cluster Mapping

By using the optimal k of 2, the clustering results of the standardized data are obtained and the results of the medoids center are obtained as in Table 7.

Table 7: Center Value of Medoids

Country	Ghana	Poland
Cluster	Cluster 1	Cluster 2
X_1	0.903	-0.800
X_2	-0.424	-0.037
X_3	-0.581	0.235
X_4	-0.041	-0.198
X_5	-0.731	0.241
X_6	0.834	-0.579
X_7	-0.940	0.646
X_8	0.873	-1.016
X_9	-0.636	-0.020

From the clustering results, a visualization is made in the form of cluster mapping using a map consisting of two clusters as shown in Figure 3.

Fig. 3. *Map of Clustering Results of 167 Countries in the World*

Based on Figure 3, the result of clustering with optimal cluster 2 produces cluster 1 members with gray color as many as 69 countries and cluster 2 with blue color as many as 98 countries.

3.7 Characteristics of Cluster

The next step will be cluster profiling to determine the characteristics of each cluster formed, so that the tendency of each cluster can be determined. The purpose of identifying the characteristics of each cluster is as an effort to overcome the welfare problems faced by each country. The characteristics of the clusters formed can be represented by the centroids or average value of each variable in each cluster. The results of the calculation of the centroids or average value of each variable in each cluster are shown in Table 8.

Table 8: Centroid Value or Average of Variables in Each Cluster

Variables	Cluster 1	Cluster 2
X_1	0.906	-0.638
X_2	-0.391	0.275

Variables	Cluster 1	Cluster 2
X_3	-0.343	0.242
X_4	-0.186	0.131
X_5	-0.647	0.455
X_6	0.407	-0.287
X_7	-0.940	0.662
X_8	0.930	-0.655
X_9	-0.585	0.412

Based on Table 8, the characteristics of each cluster are presented in Table 9:

Table 9: Cluster Characteristics

Cluster	Characteristics
Cluster 1 (Low-welfare countries)	1. High child mortality rate 2. Low exports 3. Low health 4. Low imports 5. Low income 6. High inflation 7. Low life expectancy 8. High birth rate 9. Low amount of GDP
Cluster 2 (High welfare countries)	1. Low child mortality rate 2. High exports 3. High health 4. High imports 5. High income 6. Low inflation 7. High life expectancy 8. Low birth rate 9. High amount of GDP

Based on the cluster characteristics in Table 9, each cluster can be classified according to its welfare level. The country categories based on their welfare levels are presented in Table 10.

Table 10: Category of Welfare Level

Cluster	Welfare Level
Cluster 1	Low Welfare
Cluster 2	High Welfare

4. RECOMMENDATIONS

The welfare of a country can be realized if the government and its citizens participate in realizing the welfare of the country. Based on the characteristics of the two clusters, the policy recommendations for cluster 1 to overcome the existing welfare problems are as follows: The government is expected to be active in designing comprehensive policies, prioritizing the handling of urgent problems such as high child mortality rates, low life expectancy, and poor health. The government can make health programs affordable and easily accessible Meanwhile, the community has a crucial role in shaping change. Communities are expected to control high birth rates such as participating in family planning programs. Collaboration between the community and the government in identifying local needs and formulating relevant solutions can create a more significant impact.

On the other hand, the policy recommendations for cluster 2 with high welfare levels are as follows: The government is expected to continue to maintain and improve existing positive conditions such as improvements in sectors that have been successful, such as exports and health, and focus on developing human resources through higher education and training to support sustainable economic growth.

5. CONCLUSION

Clustering of countries data using the k-medoids method with euclidean distance for k = 2, 3, and 4 obtained the optimum cluster at k = 2, where the Silhouette Index (SI) value is 0.2846391 and the Dunn Index is 0.0666. Based on the results of cluster profiling, it can be seen that cluster 1 consists mostly of countries in Central America, Africa, Central Asia, South Asia, and Southeast Asia that have low welfare, with characteristics of high child mortality, low exports, low health, low imports, low income, high inflation, low life expectancy, high birth rate, and low total GDP. Cluster 2 consists mostly of countries in North America, South America, Europe, East Asia, and Australia that have high welfare, with characteristics of low child mortality, high exports, high health, high imports, high income, low inflation, high life expectancy, low birth rate, and high total GDP.

6. REFERENCES

[1] Hendricks, RM & Khasawneh, MT. (2021). A Systematic Review of Parkinson's Disease Cluster Analysis Research. Aging Dis. doi: 10.14336/AD.2021.0519. PMID: 34631208; PMCID: PMC8460306.

[2] "Globalization and the Welfare State - Political Science". (2024). Oxford Bibliographies. Oxford University Press. doi: 10.1093/obo/9780199756223-0164.

[3] Dini, S. K, & Fauzan, A. (2020). Clustering Provinces in Indonesia based on Community Welfare Indicators. *EKSAKTA:Journal of Sciences and Data Analysis*, 20(1), 56-63.

[4] Herman, E., Zsido, K.-E., & Fenyves, V. (2022). Cluster Analysis with K-Mean versus K-Medoid in Financial Performance Evaluation. Appl. Sci, 12(16).

[5] Tamura, Y., & Miyamoto, S. (2014). Two-Stage Clustering Using One-Pass K-Medoids and Medoid-Based Agglomerative Hierarchical Algorithms. In IEEE International Conference on Soft Computing and Intelligent Systems and 15th International Symposium on Advanced Intelligent System (pp. 484–488). Kitakyushu.

[6] Pickett, K., & Wilkinson, R. (2007). Child wellbeing and income inequality in rich societies: ecological cross sectional study.

[7] Rizal, M. (2021). Cluster analysis using hierarchical method for classification of district / city of North Kalimantan Province based on human development indicators (HDI). *International Journal of Scientific and Engineering Research*, *4*, 235-239.

[8] Gusmantoni, N. (2022). Application of Data Mining Clustering the Development of Covid-19 Using K-Medoids Method. *Journal of Computer Science and Information Technology*.

[9] Suwanda, R., Syahputa, Z., & Zamzami, E. (2020). Analysis of Euclidean Distance and Manhattan Distance in the K-Means Algorithm for Variations Number of Centroid K. Journal of Physics: Conference Series.

[10] Kaufman, L., & Rousseeuw, P. J. (2009). Finding groups in data: an introduction to cluster analysis. John Wiley & Sons.

[11] Fajar, M. (2023). A Simple Indicator to Measure Welfare. Euclid, 10(2).

[12] Zolfaghari, F., Khosravi, H., Shahriyari, A., Jabbari, M., & Abolhasani, A. (2019). Hierarchical cluster analysis to identify the homogeneous desertification management units. PLoS ONE.

[13] Kacperska, E., Łukasiewicz, K., & Pietrzak, P. (2021). "Use of Renewable Energy Sources in the European Union and the Visegrad Group Countries—Results of Cluster Analysis" *Energies* 14, no. 18: 5680. https://doi.org/10.3390/en14185680