Fuzzy C-Means for Regional Clustering in East Java Province Based on Human Development Index Indicators

Marita Qori'atunnadyah [(1)]

[1]Program Studi Informatika, Institut Teknologi dan Bisnis Widya Gama Lumajang

Jalan Gatot Subroto No. 4 Lumajang
e-mail: maritaqori@gmail.com[(1)]

ABSTRAK

Indeks Pembangunan Manusia (IPM) adalah tolok ukur utama PBB untuk mengukur kemajuan manusia di suatu negara. IPM menggabungkan aspek penting dalam kehidupan manusia, termasuk pendapatan per kapita, harapan hidup, dan pendidikan. Di Indonesia, IPM digunakan untuk menilai kesejahteraan masyarakat, dan meskipun beberapa upaya telah dilakukan untuk meningkatkannya, IPM di Jawa Timur masih berada di bawah nilai rata-rata nasional dan target pemerintah. Untuk mengatasi masalah ini, penelitian ini menggunakan metode Fuzzy C-Means untuk mengelompokkan wilayah di Jawa Timur berdasarkan indikator IPM. Hasil penelitian menunjukkan bahwa ada lima kelompok yang optimal berdasarkan uji pseudo F-statistic. Hasil analisis One-Way MANOVA menunjukkan adanya variasi dalam karakteristik antara berbagai kelompok, sementara uji One-Way ANOVA mengonfirmasi bahwa keempat variabel indikator IPM berperan dalam pengelompokkan ini. Hasil pengelompokkan berdasarkan indikator IPM menunjukkan bahwa kelompok 3 berstatus tinggi, sementara kelompok 1 berstatus rendah. Kelompok 2 memiliki status cukup tinggi, kelompok 4 memiliki status sedang, dan kelompok 5 memiliki status cukup rendah. Oleh karena itu, dianjurkan kepada pemerintah untuk lebih berfokus pada upaya perbaikan kelompok dengan indikator IPM yang rendah, dengan tujuan meningkatkan kesejahteraan masyarakat di Jawa Timur. Penelitian ini dapat menjadi dasar bagi pemerintah dan pemangku kepentingan lainnya dalam merancang kebijakan untuk meningkatkan IPM di wilayah ini.

Kata kunci: *Fuzzy C-Means*; Indikator; IPM, Klaster

ABSTRACT

The Human Development Index (HDI) is the UN's key metric for gauging human advancement within a country, blending vital elements like per capita income, life expectancy, and education. In Indonesia, the HDI assesses societal well-being, with East Java's HDI lagging behind national and governmental targets despite mitigation efforts. To address this, the study utilizes Fuzzy C-Means clustering to classify East Java's regions based on HDI indicators, revealing five optimal groups via pseudo-F-statistic analysis. One-way MANOVA confirms variations among these groups, while One-Way ANOVA validates the significance of the four HDI indicators in categorization. The HDI-based categorization denotes Group 3 as high-status, Group 1 as low-status, Group 2 as moderately high-status, Group 4 as moderate, and Group 5 as moderately low-status. Consequently, it's advised that the government concentrates on improving low-HDI groups to uplift East Java's populace. This research can serve as a cornerstone for policymakers and stakeholders in their efforts to enhance the HDI in this region.

Keywords: *Cluster, Fuzzy C-Means, HDI, Indicator*

INTRODUCTION

The Human Development Index (HDI) is a parameter used by the United Nations to assess the social progress of humans in a country. The HDI combines several vital dimensions of human life, such as life expectancy, education, and per capita income [1]. HDI is used to measure the level of well-being of the population in a country, and one of the countries that uses HDI as a tool to gauge its social and economic development is Indonesia. In 1990, the Badan Pusat Statistik (BPS) introduced the Human Development Index (HDI) for the first time in Indonesia. HDI measures three primary dimensions: life expectancy, education, and per capita income. Life expectancy reflects the health and quality of life of the population and is calculated based on the average age of life expectancy at birth. Education comprises two components: mean years of schooling and expected years of schooling. Per capita income reflects the average income level of the population in a country. Improving HDI is usually a crucial indicator for governments in their efforts to enhance the well-being of the population.

The development of HDI in Indonesia has shown improvement over time, although challenges remain in enhancing access to quality education, achieving balanced development, and reducing social and economic inequalities. Efforts continue to be made to increase the HDI, which is a key factor in assessing human progress in Indonesia. In 2022, Indonesia's Human Development Index (HDI) reached a figure of 72.91, which is still below the target set by the government since 2018, ranging between 73.41 and 73.46 [2]. Out of the 38 provinces in Indonesia, East Java Province's HDI is still below the national average and has not yet reached the 2022 HDI target of 72.75 [3]. Therefore, it is necessary to establish groups of regencies/cities based on the HDI indicators in East Java to assist the government in designing effective policies to enhance the HDI in this region.

Cluster analysis is a multivariate analysis method used to group objects based on similar characteristics they possess. Within a specific cluster, there is a high level of similarity among objects, whereas the similarity between different clusters is low [4]. There are two types of clustering methods: hierarchy and non-hierarchy [5]. One example of a non-hierarchical clustering method is the fuzzy c-means and c-means method. C-means is an example of a non-hierarchical clustering method that partitions data into one or more clusters. This method partitions data by grouping similar data into one cluster, while data with different characteristics are placed in different clusters. The goal is to optimize a specified objective function in the clustering process, which aims to minimize the variation within a cluster and maximize the variation between clusters [6]. Fuzzy c-means is an extension of c-means that utilizes fuzzy weighting. Previous research has compared hierarchical and non-hierarchical clustering methods by simulating a dataset. The research results indicate that fuzzy c-means provide the best outcomes, especially in cases involving data with outliers and overlaps, compared to hierarchical clustering methods (such as single linkage, complete linkage, and average linkage), Self-Organizing Maps (SOM), and c-means [7].

Several previous studies related to clustering, such as regional clustering based on road conditions using K-Means [8], and regional clustering based on the teacher-student ratio at various educational levels using K-Means [9], [10]. The results of the previous research open up opportunities to develop methods for clustering teacher data, clustering based on undergraduate or non-graduate qualifications does not reflect the distribution of teachers well. There is a gap between provinces in eastern Indonesia and some outside Java in the ratio of undergraduate and between provinces in eastern Indonesia and some outside Java in the ratio of undergraduate to non-graduate teachers; and non-graduate teachers, provinces with student-teacher ratios and teacher-school ratios below the teacher-school ratios below the set standards. In addition, it shows a gap in the distribution of teachers between provinces in Java and outside Java. Thus, the use of the K-Means Clustering method in this study contributes to providing a better perspective of contributes to providing a better perspective on the distribution and quality of teachers in Indonesia as well as providing a basis for recommendations for improvement.

Furthermore, research on regional classification based on HDI indicators has been conducted using the c-means method [11]. The commonality in all three studies is the adoption of the c-means method. Therefore, in this research, a different approach is applied, which is the fuzzy c-means method. The use of the K-Means method in this study resulted in the determination of the optimum number of groups of 4, with statistical test results showing significant differences between groups. The four variables used influence the average differences between groups, and the analysis shows the characteristics and indicators of the Human Development Index (HDI) that need improvement. The findings guide policies and interventions appropriate to the characteristics of each group to improve the quality of human development in these regions. Some previous studies that have used the fuzzy c-means method include the clustering of Junior High Schools in Indonesia based on the National Education Standards [12], clustering of Indonesian provinces based on indicators of the well-being of the population [13], clustering of regions based on health indicators [14], clustering of tax revenue types in Makassar City [15], and clustering of COVID-19 cases in Indragiri Hilir Regency [16]. From several studies presented, it can be concluded that the Fuzzy C-Means (FCM) method provides advantages in overcoming various complex data clustering problems. The advantage of FCM lies in its ability to handle data uncertainty and complexity by providing a finer degree of membership in each cluster. FCM is also flexible in handling variations and complex patterns, providing more accurate and informative clustering results to support decision-making.

This study extends the existing body of research on regional classification by employing the Fuzzy C-Means method to cluster East Java Province based on the Human Development Index (HDI) in 2022. In contrast to previous studies that utilized methods like K-means and c-means, the fuzzy c-means approach provides enhanced flexibility in addressing uncertainties and overlaps within the data, allowing for more nuanced cluster assignments. The researcher aims to determine optimal clusters and rigorously test the obtained results, contributing novel insights into the development characteristics of East Java's regencies/cities.

METHOD

This research utilized secondary data obtained from the source of the BPS-Statistics of Jawa Timur Province [17]. The data utilized encompassed 38 regencies and cities within that region and included various indicators within the Human Development Index (HDI). The HDI indicators used in this study involved Life Expectancy (LE), Expected Years of Schooling (EYS), Mean Years of Schooling (MYS), and adjusted per capita income.

A. Fuzzy C-Means

The fuzzy c-means clustering method reallocates data into groups using the concept of non-binary membership. In this method, membership function variables are employed, indicating the extent to which data can belong to a specific group. There is also a variable '*m*', referred to as the weighting exponent, which governs the extent to which data can be a member of a group. The value of '*m*' is typically greater than 1, with the standard value often being set at 2 [6].

In the fuzzy c-means method, the initial cluster centers are determined as the mean locations of each cluster. In this concept, each data point can belong to multiple clusters simultaneously, making the boundaries between clusters fuzzy. By iteratively refining the cluster centers and membership degrees through the fuzzy c-means algorithm, we can obtain cluster centers that approximate the appropriate locations. The iteration process is based on minimizing the objective function in the given equation. The value of 'm' used in this algorithm influences the extent to which the membership degrees of data in each cluster can change during the iteration. The iteration in this algorithm is based on minimizing the objective function in the equation (1).

$$J\left(U,c_1,\ldots,c_g\right)=\sum_{c=1}^{g}J_c=\sum_{c=1}^{g}\sum_{j}^{n}u_{ci}^{m}d_{ci}^{2} \quad (1)$$

Notation :

u_{ci} : Membership degree of object *i* to cluster *c*

c_g : The matrix of centroids for all clusters

n : Number of data points

c : Number of clusters

m : Weighting exponent

$d_{ci}=\| c_g - x_i \|$: Euclidean distance between cluster *c* and cluster center *i*

This objective function reflects the distance between the given data point and the cluster center, weighted by the membership degree of that data point. The membership degree of a data point in a particular cluster can be calculated using the following equation (2).

$$u_{ci}=\frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ci}}{d_{ki}}\right)^{2/(m-1)}} \quad (2)$$

Notation :

u_{ci} : Membership degree of data point *i* to cluster *c*

d_{ci} : Centroid value of data point *i* in cluster *c*

d_{ki} : Centroid value of data point *i* in cluster *k*

c : Number of clusters

m : Weighting exponent

The membership function has a range of values between 0 ≤ u_{ci} ≤ 1. To calculate the cluster centers, the following equation (3) can be used.

$$c_i=\frac{\sum_{i=1}^{n}u_{ci}^{m}x_i}{\sum_{i=1}^{n}u_{ci}^{m}} \quad (3)$$

where x_i represents the object or data point *i*.

The algorithm applied in the fuzzy c-means method to identify cluster centers c_i *with the membership matrix* **U** is as follows:

1. Determine the number of clusters or groups to be formed (c).
2. Choose the value of the weighting exponent (m > 1), with the common value being2.
3. Specify the tolerance threshold or stopping criteria for iterations.
4. Create the initial partition matrix **U** (membership degree matrix). Matrix U is filled with random numbers between 0 and 1 based on the following equation (4):

$$U=\begin{bmatrix}\mu_{11}(x_1) & \mu_{12}(x_2) & \cdots & \mu_{1i}(x_i)\\ \mu_{21}(x_1) & \mu_{22}(x_2) & \cdots & \mu_{2i}(x_i)\\ \vdots & \vdots & \ddots & \vdots\\ \mu_{c1}(x_1) & \mu_{c2}(x_2) & \cdots & \mu_{ci}(x_j)\end{bmatrix} \tag{4}$$

5. Calculate the fuzzy centroids c_i , i = 1 ,..., c using Equation (3) and create a new matrix for centroid values to calculate the objective function.
6. Compute the fuzzy c-means objective function value based on Equation (1). The objective function value is used to determine whether the iteration continues or stops. Iteration stops if the objective function value falls below the tolerance threshold.
7. If the objective function value is still above the threshold, update the calculation of the membership matrix or partition matrix **U** based on Equation (2), and repeat these steps starting from step 5. This matrix is used to determine the groups of each observation after the iteration stops.

B. Pseudo F-Statistic

The pseudo-F-statistic was first introduced by Calinski and Harabasz. This statistic is used to measure the validity of clustering to determine the optimum or best number of clusters [18]. The highest pseudo-F-statistic reflects an optimal clustering outcome, where the similarity within a cluster is high, while the differences between clusters are significant [19]. Here is the equation (5) used to calculate the pseudo F-statistic.

$$Pseudo\ F-statistic=\frac{SSB/(k-1)}{SSW/(n-k)} \tag{5}$$

Notation: :

SSB = variation between clusters, also known as sum of squares between.
SSW = variation within clusters, also known as sum of squares within.
k = the number of clusters generated by the clustering algorithm.
n = the number of data points.

C. One Way MANOVA and One Way ANOVA

The use of One-Way MANOVA is to compare means of two or more populations when there are multiple dependent variables or to assess the effects of a treatment on responses [5]. MANOVA is used to evaluate the similarity among the formed groups. Before conducting the One-Way MANOVA test, it is important to undergo multivariate normality testing and homogeneity testing. Multivariate normality testing is an extension of univariate normality testing that involves at least two observed variables. In multivariate analysis, multivariate normality testing is necessary to ensure that the observed data follows a multivariate normal distribution [20]. Furthermore, to check the homogeneity of covariance matrices, Box's M test is used. One-way ANOVA, on the other hand, is used to test differences between groups when only one dependent variable is used or to test differences between variables among group members [5].

RESULT AND DISCUSSION

In this study, a regional cluster analysis in East Java Province based on the 2022 Human Development Index indicators was conducted using the fuzzy c-means method. There are 38 regencies and cities in East Java that will be grouped into several clusters. Clustering was performed by trying different numbers of clusters, ranging from 2 to 5 clusters, and the most optimal cluster configuration was chosen using R software. The results of the clustering for each cluster are displayed in Table 1.

Table 1. Clustering using the Fuzzy C-Means Method

Cluster	**Number of Clusters**			
	2	**3**	**4**	**5**
1	26	18	12	7
2	12	16	9	7
3		3	3	3
4			14	11
5				10

Table 1 presents data on the number of members (regencies/cities) in each cluster resulting from the clustering of 38 regencies/cities in East Java Province. Clustering was performed using the fuzzy c-means method with varying numbers of clusters ranging from 2 to 5. The selection of the optimal number of clusters can be determined based on the highest pseudo F-statistic value among these variations. Here are the pseudo F-statistic values associated with each cluster.

Table 2. Pseudo F-Statistic Values of the Fuzzy C-Means Method

Number of Clusters	***Psudo F-Statistic***
2	64,80039
3	95,53214
4	144,1018
5	**170,1298**

Table 2 shows the calculation of pseudo F-statistics using the fuzzy c-means method for the number of clusters ranging from 2 to 5. The optimal number of clusters for grouping regencies/cities in East Java based on the Human Development Index indicators is 5 clusters. This can be observed from the highest pseudo F-statistic value, which is 170.1298, found in the clustering with 5 clusters. The fuzzy c-means clustering process involves several initial steps, including determining the desired number of clusters (in this example, 5 clusters), setting the initial weighting exponent (m) to 2, and setting the tolerance threshold or iteration stopping criterion to approximately 10^{-6}. Next, the initial partition matrix U is created using random numbers between 0 and 1, with the number of rows corresponding to the number of regencies/cities in East Java Province (38 rows) and the number of columns corresponding to the number of clusters (5 columns) to be formed.

The next step is to determine the centroid (center point) for each cluster and calculate the objective function value. The objective function value is used to determine whether the iteration continues or stops based on a comparison with the stopping criterion value. If the objective function value is greater than the tolerance threshold of approximately 10^{-6}, then the calculation of the new partition matrix U (membership function) is performed. After that, the centroid values and objective function are recalculated until the objective function value reaches less than 10^{-6}. When the objective function value reaches this threshold, the iteration is stopped, and the membership for each of the five clusters is determined based on the membership degree matrix in the last iteration. For example, if the first regency has the highest membership degree in cluster three, it will be a member of cluster three, and the same applies to the other regencies up to the 38th regency. The results of clustering using 5 clusters can be seen in Table 3.

Table 3. List of Regencies/Cities in 5 Clusters

Cluster	**Regencies/Cities**			
1	Pacitan Bangkalan	Lumajang Sampang	Jember Pamekasan	Sumenep
2	Sidoarjo Blitar City	Mojokerto Kota Batu	Gresik Mojokerto City	Pasuruan City
3	Malang City	Madiun City	Surabaya City	
4	Kediri Jombang Magetan	Banyuwangi Nganjuk Ngawi	Probolinggo Madiun Lamongan	Probolinggo City Kediri City
5	Ponorogo Malang Situbondo	Trenggalek Bondowoso Blitar	Tulungagung Pasuruan	Tuban Bojonegoro

Table 3 contains the members of each cluster. In clustering the regencies/cities in East Java Province based on the Human Development Index indicators using the fuzzy c-means method, it is expected that there are differences in characteristics within each group related to all Human Development Index indicators. To assess whether there are characteristic differences in the formed groups, this can be done through one-way MANOVA and one-way ANOVA methods. Before conducting these tests, the initial steps are to test whether the data is normally distributed in multivariate form and to test the homogeneity of variances among groups.

The multivariate normality distribution test is used to evaluate whether the data follows a multivariate normal distribution or not. The result of the multivariate normality distribution test shows a correlation value of 0.986 as indicated above.

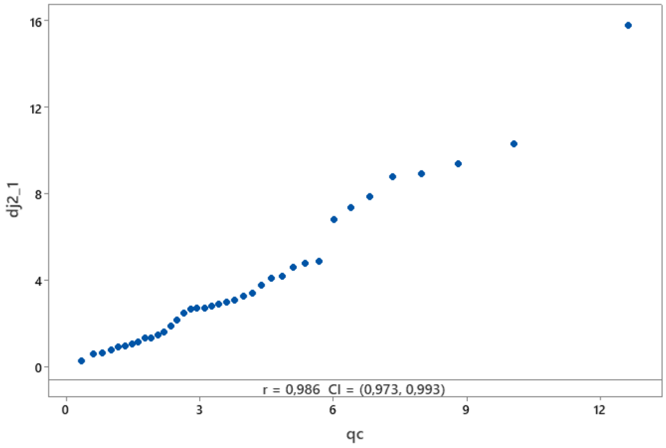

Figure 1. Multivariate Normal Probability Plot

These values will be compared with the critical point from the table of the probability plot of the correlation coefficient for normality (PPCC). The critical point obtained at a 5% significance level is 0.9700. The result indicates that there is not enough evidence to reject H0, which means that the data follows a multivariate normal distribution. To test the homogeneity of the covariance matrix, Box's M test is used at a 5% significance level. The test result shows that the value of Box's M is 46.847.

Table 4. Box's M Test Result

	Value
Box's M	46,847
F	1,155
df1	30
df2	1884,898
Sig.	0,258

At a significance level of 5% (alpha 0.05) and with 40 degrees of freedom, the value is obtained as 55,758. The found Box's M value is smaller than the expected critical value, which is $\chi^2_{\frac{1}{2}(g-1)p(p+1)}$. Furthermore, considering the significance value of 0.258, which exceeds the significance level alpha (0.05), the conclusion is that H_0 fails to be rejected. Therefore, it can be concluded that the covariance matrix is homogeneous.

After undergoing multivariate normality testing and checking for data homogeneity, the results indicate that the data follows a multivariate normal distribution and has a homogeneous covariance matrix. Therefore, the test for differences in characteristics using One-Way MANOVA adopts Pillai's Trace test statistic. In this analysis, we want to assess the factors or treatments that are suspected to have a significant impact on the response variable, which in this case is the formed clusters. The response variable in the One-Way MANOVA test is the Human Development Index indicators.

Table 5. One-Way MANOVA Test Results

Pillai's Trace Value	*F*	Degrees of Freedom for Hypothesis	Degrees of Freedom for Error	Sig.	*Partial Eta Squared*
1,243	3,718	16	132	0,000	0,311

Based on TABLE 5, it was found that the One-Way MANOVA test results showed a Pillai's Trace test statistic value of $F = 3.718$. The critical value $F_{64;528;0.05}$, which is the significance threshold, is 1.334. When comparing these two values, the found F value is greater than the critical value $F_{64;528;0.05}$, resulting in the rejection of the null hypothesis (H_0). This means there is a significant difference among the formed clusters.

One-way ANOVA testing is used to evaluate differences among variables among group members. The results of the One-Way ANOVA test are as follows.

Table 6. One-Way ANOVA Test Results

Variable	F	Sig
Life Expectancy	3,552	0,016
Expected Years of Schooling	8,413	0,000
Mean Years of Schooling	24,843	0,000
Adjusted Gross National Income per Capita	170,132	0,000

From Table 6, the F-values for each variable were found. These values will be compared to the $F_{4;33;0,05}$ value, which is equivalent to 2.659. When comparing the F-values of the variables with the $F_{4;33;0,05}$ value, it is evident that all four variables have higher F-values. Therefore, the null hypothesis (H_0) is rejected, indicating that there are significant differences in the characteristics of the four variables among the formed clusters. This suggests that these four variables have a significant impact on the cluster formation.

By applying the fuzzy c-means method, the districts/cities in East Java Province were successfully grouped into 5 clusters based on the IPM indicators. The results of the one-way MANOVA test indicate significant differences among the five formed clusters, and the four variables have varying effects on these cluster differences, as revealed through the one-way ANOVA results. Here is a description of each of the formed clusters.

Table 7. List of Regencies/Cities in 5 Clusters

Cluster	**1**	**2**	**3**	**4**	**5**
n	7	7	3	11	10
Life Expectancy	70,24	73,32	73,78	72,02	72,05
Expected Years of Schooling	12,80	14,07	15,01	13,56	12,99
Mean Years of Schooling	6,43	10,03	10,96	8,29	7,49
Adjusted Gross National Income per Capita	9251,43	13750,29	17248,33	11953,64	10559,60

Based on Table 7, you can observe the average characteristics of each cluster formed from the 2022 IPM indicator data in East Java Province. It's noted that Cluster 3 has the highest average values among the groups for each indicator. This indicates that Cluster 3 has a high IPM indicator, reflecting good quality in those indicators. Cluster 2 has the second-highest average values for each indicator after Cluster 3, but it's important to notice that there's a significant difference in the per capita expenditure indicator when compared to Cluster 3. Therefore, improvement is needed in this indicator. Following Cluster 2, the next-highest averages are found in Cluster 4, although the life expectancy indicator has a slightly lower average than Cluster 5. Hence, improvement should focus on the life expectancy indicator. Cluster 1 has the lowest average values for all indicators, thus requiring comprehensive improvements in all indicators. Meanwhile, Cluster 5 has the second-lowest averages, particularly in the indicators of expected years of schooling, mean years of schooling, and per capita expenditure. Although their averages are not significantly different from Cluster 1, improvements should be focused on these three indicators. Additionally, when examining the number of districts/cities in each group, Cluster 4 and 5 have a large number of members, even though their average indicators fall into the "fair" category. Nonetheless, more targeted improvements are still needed in both of these groups. Based on the average values of each indicator, the ranking status can be assigned to the groups of districts/cities as follows.

Table 8. Status of Each Cluster

Cluster	Status
1	Low HDI Indicators
2	Moderately High HDI Indicators
3	High HDI Indicators
4	Moderate HDI Indicators
5	Moderately Low HDI Indicators

CONCLUSION

Based on the results and discussion, the regional clustering in East Java based on the Human Development Index (HDI) indicators using the fuzzy c-means method yielded an optimal clustering result of 5 clusters. This result was obtained based on the largest pseudo-F-statistic value. In the one-way MANOVA test using Pillai's Trace statistic, significant differences were observed among the formed clusters. Additionally, the one-way ANOVA test indicated that all four variables had a significant impact on the differences in characteristics among these clusters. Cluster 3 showed high HDI indicators. Cluster 2, while having moderately high HDI indicators, requires adjustment, particularly in the per capita expenditure indicator. Cluster 4 exhibited moderate HDI indicators, thus necessitating improvements in life expectancy indicators. Cluster 5 displayed moderately low HDI indicators, requiring more focused efforts to enhance indicators related to expected years of schooling, mean years of schooling, and per capita expenditure. Cluster 1 represented the category with low HDI indicators, necessitating comprehensive improvements across all indicators. The Human Development Index (HDI) is a crucial parameter in assessing the success of efforts to improve the quality of human life. Therefore, the recommendation to the government is to concentrate improvement efforts on indicators with low values within each cluster. Other clustering methods like Revised Fuzzy C-Means (RFCM) can be applied to data with uneven cluster sizes and contamination from noise and outliers, as demonstrated in the study [21].

REFERENCE

[1] Badan Pusat Statistik, “Statistik Indonesia,” Jakarta, 2019.

[2] Kementerian PPN/Bappenas, “Pemuktahiran Rencana Kerja Pemerintah (RKP) Tahun 2022,” 2021. Accessed: Mar. 29, 2023. [Online]. Available: https://www.bappenas.go.id/show-result-satudata?name=publikasi&key=rkp&tahun=

[3] Badan Pusat Statistik, “ Indeks Pembangunan Manusia menurut Provinsi 2020-2022,” 2023.

[4] J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, “MULTIVARIATE DATA ANALYSIS EIGHTH EDITION,” 2019. [Online]. Available: www.cengage.com/highered

[5] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis 6th Edition*, 6th ed. United States of America: Pearson Prentice Hall, 2007.

[6] Y. Agusta, “K-Means-Penerapan, Permasalahan dan Metode Terkait,” 2007.

[7] S. A. Mingoti and J. O. Lima, “Comparing SOM neural network with Fuzzy c-means, K-means, and traditional hierarchical clustering algorithms,” *Eur J Oper Res*, vol. 174, no. 3, pp. 1742–1759, Nov. 2006, doi: 10.1016/j.ejor.2005.03.039.

[8] M. Qori'atunnadyah and F. D. Rahmawati, "Pengelompokkan Kabupaten dan Kota Berdasarkan Kondisi Infrastruktur Jalan Menggunakan Hierarchical Clustering," *Journal of Informatics Development*, vol. 1, no. 1, pp. 1–5, 2022.

[9] M. Qori'atunnadyah, "Pengelompokkan Wilayah Berdasarkan Rasio Guru-Murid Pada Jenjang Pendidikan Menggunakan Algoritma K-Means," *Journal of Informatics Development*, vol. 1, no. 2, pp. 33–38, 2022.

[10] F. Idris, F. Azmi, and P. S. Daru Kusuma, "PENGELOMPOKAN DATA GURU DI INDONESIA MENGGUNAKAN K-MEANS CLUSTERING TEACHER DATA GROUPING IN INDONESIA USING K-MEANS CLUSTERING," *eProceedings of Engineering*, vol. 6, no. 2, pp. 5648–5653, 2019.

[11] M. Qori'atunnadyah, "Metode C-Means untuk Pengelompokkan Kabupaten/Kota Provinsi Jawa Timur berdasarkan Indikator Indeks Pembangunan Manusia (IPM)," *Journal of Informatics Development*, vol. 1, no. 2, pp. 51–58, 2023, doi: 10.30741/jid.v2i2.1013.

[12] H. A. Chusna and A. T. Rumiati, "Penerapan Metode K-Means dan Fuzzy C-Means untuk Pengelompoan Sekolah Menengah Pertama (SMP) di Indonesia Berdasarkan Standar Nasional Pendidikan (SNP)," *Jurnal Sains dan Seni ITS*, vol. 9, no. 2, Feb. 2021, doi: 10.12962/j23373520.v9i2.58349.

[13] N. Dwitiyanti, N. Selvia, and F. R. Andrari, "Penerapan Fuzzy C-Means Cluster dalam Pengelompokkan Provinsi Indonesia Menurut Indikator Kesejahteraan Rakyat," *Faktor Exacta*, vol. 12, no. 3, p. 201, Nov. 2019, doi: 10.30998/faktorexacta.v12i3.4526.

[14] G. S. Nugraha and B. A. Riyandari, "IMPLEMENTASI FUZZY C-MEANS UNTUK PENGELOMPOKAN DAERAH BERDASARKAN INDIKATOR KESEHATAN," *Jurnal Teknologi Informasi*, vol. 4, no. 1, pp. 52–62, Jun. 2020, doi: 10.36294/jurti.v4i1.1222.

[15] I. Irwan, S. Sidjara, and A. P. Aryati, "Pengelompokan Jenis Penerimaan Pajak di Kota Makassar Menggunakan Fuzzy Clustering," *Euler : Jurnal Ilmiah Matematika, Sains dan Teknologi*, vol. 10, no. 1, pp. 98–102, May 2022, doi: 10.34312/euler.v10i1.14225.

[16] S. F. Octavia and M. Mustakim, "Penerapan K-Means dan Fuzzy C-Means untuk Pengelompokan Data Kasus Covid-19 di Kabupaten Indragiri Hilir," *Building of Informatics, Technology and Science (BITS)*, vol. 3, no. 2, pp. 88–94, Sep. 2021, doi: 10.47065/bits.v3i2.1005.

[17] BPS-Statistics of Jawa Timur Province, "Provinsi Jawa Timur Dalam Angka 2023," Surabaya, 2023.

[18] T. Calinski and J. Harabasz, "A dendrite method for cluster analysis," *Commun Stat Theory Methods*, vol. 3, no. 1, pp. 1–27, 1974, doi: 10.1080/03610927408827101.

[19] T. Hochin *et al.*, "Quasi-optimality under pseudo F statistic in clustering data Quasi-optimality under pseudo F statistic in clustering data Quasi-optimality under pseudo f statistic in clustering data," 2018. [Online]. Available: www.sciencepubco.com/index.php/IJET

[20] A. C. Rencher and W. F. Christensen, *Methods of Multivariate Analysis*, 3rd ed. Wiley, 2012.

[21] S. Askari, "Fuzzy C-Means clustering algorithm for data with unequal cluster sizes and contaminated with noise and outliers: Review and development," *Expert Syst Appl*, vol. 165, p. 113856, Mar. 2021, doi: 10.1016/j.eswa.2020.113856.