Journal of Advanced Technology and Multidiscipline (JATM)
Vol. 02, No. 01, 2023, pp. 27-33
e-ISSN: 2964-6162

Human Development Clustering in Indonesia: Using K-Means Method and Based on Human Development Index Categories

Indah Fahmiyah[1], Ratih Ardiati Ningrum[1]
[1]*Faculty of Advanced Technology and Multidiscipline, Universitas Airlangga, Surabaya, Indonesia*

***Abstract*—The quality of life for Indonesia's population can be measured from the human development index in each province. People who have a good quality of life indicate a prosperous life. The government has the responsibility to advance the welfare of the nation under the mandate of the constitution. The clustering of the *Human Development Index* (HDI) in Indonesia is used to determine the distribution of quality of life or the distribution of social welfare. In this study, the K-Means method, which is a popular non-hierarchical clustering method, is used to classify human development in each province based on HDI indicators, namely *Life Expectancy at Birth*, *Expected Years of Schooling*, *Mean Years of Schooling*, and *Adjusted Expenditure Per Capita*. Provinces in Indonesia are clustered into 4 clusters. These results were also compared with the clustering based on HDI categories determined by Statistics Indonesia based on certain cut-off values. According to the HDI category, provinces in Indonesia fall into the medium, high, and very high categories. The results of the two groupings show that there is a trend toward appropriate characteristics for each group. Thus, K-Means can classify provinces in Indonesia according to the characteristics of the HDI indicators.**

***Keywords*— Human development, HDI, K-Means, Quality of life.**

I. INTRODUCTION

General welfare is a community right and has been regulated by the Indonesian state constitution. Thus, the Indonesian government is responsible for fulfilling the state's mandate. Country achievements in national development can be seen from various factors. Economic growth and the quality of human resources are several factors that support the success of a country. The human development index (HDI) is one measure to assess the quality of human resources [1]. HDI is a measure of the quality of human life as well as an indicator of development goals [2]. HDI explains how citizens can access development outcomes in terms of income, health, education, and other aspects of life [3]–[5]. United Nations Development Programme (UNDP) has used three dimensions to form HDI, namely long and healthy life, knowledge, and a decent standard of living [2], [6].

Indonesia is an archipelagic country consisting of 5 large island groups namely Sumatra, Kalimantan, Sulawesi, Java, Bali, Nusa Tenggara, Maluku, and Papua [2]. Each region has a diversity which is a challenge for the government in human development. The value of the human development index can be used as a reference for the government to make budget policies for each region and strategies for achieving national or regional development [2]. Regional grouping based on human development is used to determine the distribution of the quality of life of the population. By knowing this distribution, the government can formulate short-term and long-term development strategies to improve and increase the quality of life of the population.

HDI Indonesia has increased from year to year [2]. In Indonesia, HDI is categorized into 4 categories, i.e. low, medium, high, and very high with cutoff points at 60, 70, and 80 [1]. The cutoff point for low HDI that is determined by UNDP is slightly different, namely lower than 55 [6]. Most regencies or cities' HDI in Indonesia is medium HDI [2]. Lampung Province, Central Sulawesi Province, and Maluku Province changed from medium to high HDI with an HDI growth of 0.79%, 0.70%, and 0.73% in 2022 [2]. If the categorization of human development in an area only uses a composite index, then the consideration of the cutoff point for each category becomes very crucial.

Cluster analysis aims to group of observations into clusters depends on the similarities and dissimilarities in the characteristics of the dataset. The K-Means is a frequently used non-hierarchical clustering method [7]–[9]. The form of the K-Means algorithm is to assign each object to the cluster that has the closest centroid. Thus, the objective of K-Means is to obtain minimum between-cluster variation and maximum inter-cluster variation.

Much research has been carried out regarding the grouping of human development in Indonesia based on HDI indicators by using cluster analysis, such as hierarchical clustering [5], K-Means [4], [5], [10], K-Medoids [5], and Fuzzy C-Means [10]. The researchers found that the number of clusters is four clusters using K-Means [4], [5], [10] and Fuzzy C-Means [10], but two clusters for hierarchical clustering and five clusters for

Corresponding Author:
Indah Fahmiyah
Faculty of Advanced Technology and Multidiscipline, Universitas Airlangga, Campus C UNAIR Gedung Kuliah Bersama, Mulyorejo, Surabaya, Indonesia
indah.fahmiyah@ftmm.unair.ac.id

Vol. 02, No. 01, 2023, pp. 27-33
e-ISSN: 2964-6162

K-Medoids [5]. HDI grouping in a particular area in Indonesia was also carried out [1], [11], [12]. Grouping human development of countries in the world is also carried out using the K-means and Partitioning Around Medoids algorithms and offers cut-off values to classify countries with low (lower than 65), medium (65-85), and high (greater than 85) human development [13]. The study by [3] concludes that there is a spatial dependence so that the HDI in a region can be influenced by the HDI in the nearest area.

In this study, provinces in Indonesia were grouped based on HDI indicators (*Life Expectancy at Birth*, *Expected Years of Schooling*, *Mean Years of Schooling*, and *Adjusted Expenditure Per Capita*) using K-Means and according to HDI categories (low, medium, high, and very high). Next, the characteristics of each cluster or category of the two grouping schemes are analyzed. This study also evaluates the distribution of HDI from each grouping result.

II. Fundamental Theory

A. Cluster Analysis

Cluster analysis is an analysis to group similar elements as research objects into different and mutually exclusive groups [1], [14]. This analysis is useful for summarizing data by grouping objects based on certain characteristic similarities among the objects to be studied. Cluster analysis is a tool for grouping several *n* objects based on variables that have relatively similar characteristics between these objects so that the variance within groups is smaller than the variance between groups. Objects will be classified into one or more clusters until objects in one cluster will have similar characters.

Cluster analysis consists of two methods, i.e. hierarchical and non-hierarchical methods [15]. The hierarchical method starts by grouping data that has the closest similarity [16]. Then proceed to other objects that have the second closest proximity and so on. Hence the groups will form a kind of tree called the dendrogram, where there is a clear hierarchy between objects, from the most similar to the least similar. The non-hierarchical clustering method is used more for group objects than for variables to collect *k* clusters. The number of clusters, *k*, is determined by the researcher or determined by the part of the clustering procedure. Non-hierarchical clustering can be applied better in large samples than hierarchical methods.

A.1. K-Means Method

A popular non-hierarchical method or partitioning method is the K-Means method [7]–[9]. The algorithm of this method groups objects based on the distance between the objects and the centroid cluster [17] where the distance is obtained by an iterative process. The analysis needs to determine the number of *k* as input to the algorithm [12], [18]. The goal of K-Means clustering is to get clusters of objects by maximizing the similarity of objects within clusters (or minimizing variance within clusters) and maximizing the differences between clusters.

The different units and outlier values in the dataset can be avoided with normalization data [19]. The original data was transformed to make the same scale in the dataset, such as using min-max normalization. It performs a linear transformation on the original data. Suppose we want to transform values in variable X, its formula is shown in (1)

$$x_i' = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{1}$$

where x_i' is the normalized value from the original value x_i, x_{min} is the minimum value of X, and x_{max} is the maximum value of X. After min-max normalization, the values become between 0 (minimum value) and 1 (maximum value).

Suppose that *n* observations (objects) from a dataset consisting of *p* variables are partitioned into *k* clusters, namely $C_1, C_2, \ldots,$ and C_k. The K-Means method groups the objects into *k* clusters such that to get a minimum within-cluster sum of squares (WCSS) with the formula (2).

$$WCSS = \sum_{j=1}^{k} \sum_{x_i \in C_j} \|x_i - m_j\|^2 \tag{2}$$

WCSS is the objective function for this algorithm [5], [17], x_i is the *i*th observation, m_j is the centroid of cluster *j*, C_j is the *j*th cluster, $i = 1, 2, \ldots, n,$ and $j = 1, 2, \ldots, k$. To classify each observation into *k* clusters, the K-Means algorithm is divided into the following steps [17]:

1. Determine the number of clusters, *k*, and select *k* points as initial centroids (consists of *p* coordinates).
2. Put the observations into the cluster with the closest centroid value. The closeness distance between observations to the centroid is measured by the Euclidean distance, d_{ij}, presented in (3)

$$d_{ij} = \sqrt{\sum_{\ell=1}^{p} \left(x_{i\ell} - m_{j\ell}\right)^2} \tag{3}$$

where d_{ij} is the distance between the *i*th observation to the *j*th centroid, $x_{i\ell}$ is the *i*th observation on variable ℓ, $m_{j\ell}$ is the ℓth coordinate of centroid *j*, and $\ell = 1, 2, \ldots, p$.

3. Calculate the new centroid for each cluster using the mean value of all objects in each cluster. Each centroid coordinate for cluster *j* is calculated with (4)

$$m_{j\ell} = \frac{1}{|C_j|} \sum_{x_{i\ell} \in C_j} x_{i\ell} \tag{4}$$

where $|C_j| = n_j$ is the number of objects in the *j*th cluster or the size of C_j.

4. Repeat steps 2 and 3 until no rearrangements are possible.

A.2. The Optimal Number of Clusters

The K-Means method is non-hierarchical clustering so researchers determine the number of clusters themselves [1]. The selection of the number of clusters is very important for the clustering results obtained. Therefore we need a method to determine the optimal number of clusters, one of which is the Elbow method [1], [9], [13], [17]. The graph of the Elbow method, namely the plot between the number of clusters (x-axis) and the total variation within the clusters (y-axis) can be calculated from (2). The elbow point of the curve determines the number of clusters built [13], [17].

Validation of how well the grouping of observations based on the variables in the dataset is an important part of the clustering algorithm [20]. In this study, it is measured by the Calinski-Harabasz (CH) index, which is also called the Variance Ratio Criterion (VRC), presented in (5) as follows [5], [20]–[22]

$$CH = \frac{BCSS/(k-1)}{WCSS/(n-k)} \quad (5)$$

where $WCSS$ is shown in (2) and $BCSS = \sum_{j=1}^{k} n_j \left\| m_j - m \right\|^2$ is the between-cluster sum of squares with m is the centroid of the dataset (global centroid). The greater the $BCSS$ value, the greater the degree of dispersion between clusters (increasing differences between clusters), while the smaller the WCSS value, the lower the degree of dispersion within a cluster (higher the similarity between objects in a cluster). The greater the value of the CH index indicates the better the grouping effect [22]. Therefore, the CH index can also suggest the optimal number of clusters in cluster analysis.

B. Human Development Index

The Human Development Index (HDI) is a measure of progress in efforts to improve the quality of life of the community by taking into account development outcomes such as income, health, and education. The term Human Development Index (HDI), was popularized by the United Nations Development Programme (UNDP) in 1900 [2], [6]. The UNDP defines human development as a process of expanding choices for the population, in the sense that humans are given the freedom to choose more choices in terms of meeting life's needs.

UNDP uses three dimensions to construct HDI. This was also followed by Indonesia. These dimensions are long and healthy life, knowledge, and a decent standard of living [1], [2], [12], [14]. These three dimensions are the chosen approach in describing the quality of human life and have not changed until now. The HDI calculation carried out in Indonesia refers to these 3 dimensions. Each dimension is represented by an index value obtained through data normalization, generally using (1) to prevent unit differences [13], [19].

The dimensions of a long and healthy life are represented by life expectancy at birth. The use of life expectancy as an indicator is based on the common belief that longevity is worth living and the fact that several factors are closely related to life expectancy, such as adequate nutrition and good health. *Life Expectancy at Birth* (LEB) is the average estimated time (in years) a person can live over their lifetime. LEB values are normalized by converting them into a life expectancy (health) index, which is calculated based on the maximum and minimum LEB values according to UNDP standards, i.e. the maximum value is 85 years and the minimum value is 20 years [2] as follows.

$$health_{index} = \frac{LEB - LEB_{min}}{LEB_{max} - LEB_{min}} \quad (6)$$

The knowledge or education dimensions are represented by *Expected Years of Schooling* (EYS) and *Mean Years of Schooling* (MYS). These two indicators reflect the community's ability to access education, particularly formal education. The EYS describes the community's approach to formal education, while the MYS describes the human resources in the region. *Expected Years of Schooling* is the number of years of schooling (in years) that a 7-year-old is expected to complete, while *Mean Years of Schooling* (MYS) is the number of years of formal education for those aged 25 and over. The calculation of the education index is based on the average of the EYS index and the MYS index which are equally weighted. The EYS and AYS indices are calculated using the same maximum and minimum value limits as UNDP standards. The maximum and minimum values for EYS are 18 years and 0 years, while the maximum and minimum values for MYS are 15 years and 0 years [2] as follows.

$$edu_{index} = \frac{EYS_{index} + MYS_{index}}{2} \quad (7)$$

$$EYS_{index} = \frac{EYS - EYS_{min}}{EYS_{max} - EYS_{min}} \quad (8)$$

$$MYS_{index} = \frac{MYS - MYS_{min}}{MYS_{max} - MYS_{min}} \quad (9)$$

UNDP uses gross national income (GNI) per capita as an indicator of the extent of a decent standard of living. However, this information is not available at the regional level, so an adjusted real expenditure per person is used as an alternative. This indicator can be scaled down to the regional or city level. Indicators of real expenditure per capita can also reflect indicators of people's income and describe the wealth of the population as a result of economic activity. In calculating the expenditure index, the maximum (26,572.352 thousand rupiahs) and minimum (1,007.436 thousand rupiahs) limits are used [2] as follows.

$$expend_{index} = \frac{\ln(expend) - \ln(expend_{min})}{\ln(expend_{max}) - \ln(expend_{min})} \quad (10)$$

HDI is calculated as the geometric mean (11) of the life expectancy (health), education, and expenditure indices as shown in (6), (7), and (10) respectively. HDI is divided into low, medium, high, and very high. In Indonesia, low HDI if $HDI < 60$, medium HDI if $60 \leq HDI < 70$, high HDI if $70 \leq IPM < 80$, and very high HDI if $HDI \geq 80$ [2].

$$HDI = \sqrt[3]{health_{index} \cdot edu_{index} \cdot expend_{index}} \cdot 100 \quad (11)$$

III. METHODOLOGY

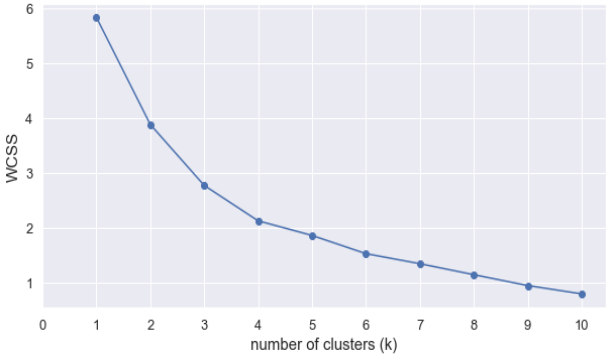
In this study, the data was taken from Statistics Indonesia (BPS) regarding the human development index in Indonesia in 2022. The research unit used was 34 provinces in Indonesia. The provinces are grouped using the K-Means method based on HDI indicators, namely *Life Expectancy at Birth* (LEB), *Expected Years of Schooling* (EYS), *Mean Years of Schooling* (MYS), and *Adjusted Expenditure Per Capita* (EPC). These indicators have different units so that the data is normalized using (1). This data normalization is used to avoid outliers and obtain a more homogeneous relative contribution between variables [19] so that the grouping results are accurate because the grouping of objects is based on the distance between the data points and the center point (centroid).

The normalized HDI indicators data is used to group provinces in Indonesia using the K-Means. After normalizing data, provinces in Indonesia grouped using the K-Means method based on these HDI indicators. The optimal number of clusters is determined based on the Elbow method and Calinski-Harabasz index with (2) and (5) respectively. From K-Means method results, the characteristics of each cluster are analyzed. And then, those results compared to the grouping HDI values become a low, medium, high, or very high category. The HDI characteristics of each group were also discussed, both those using the K-Means method and the HDI categories.

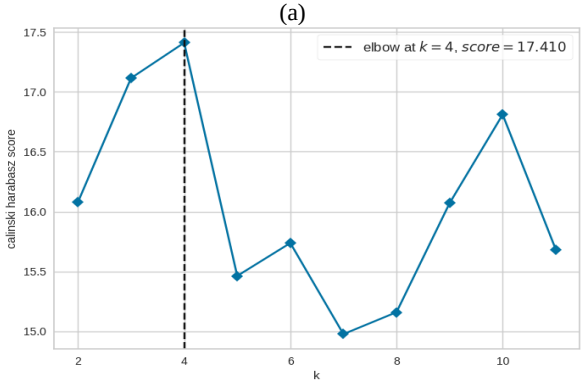
IV. Results And Discussion

A. Clustering Human Development Using K-Means Method

The first stage in the K-Means algorithm is to determine the number of clusters formed. For this reason, the Elbow method is used to determine the optimal *k* clusters. Figure 1 shows that the optimal number of clusters is four clusters for grouping provinces in Indonesia based on the 2022 HDI indicators (normalized data). This can be seen from the differences in the WCSS values in Figure 1(a) which are not too different between $k = 4$ and $k = 5$ and so on, so the number of clusters formed is 4. The Calinski-Harabasz (CH) index in Figure 1(b) is used to evaluate the clustering results and determined the optimal number of clusters, which is $k = 4$. The results of grouping data without normalization present that the CH index value is much larger, so it tends to be inaccurate.



(a)



(b)

Figure. 1 Optimal number of clusters: (a) Elbow method and (b) Calinski Harabasz index

Of the 34 provinces in Indonesia, each province is grouped into Cluster 1 (8 provinces), Cluster 2 (15 provinces), Cluster 3 (6 provinces), or Cluster 4 (5 clusters). The distribution of each cluster based on indicators of *Mean Years of Schooling*, *Life Expectancy*, and *Adjusted Expenditure Per Capita* is shown with a 3-dimensional plot in Figure 2. Cluster 4 tends to have high *Mean Years of Schooling* (MYS), *Life Expectancy at Birth* (LEB), and *Adjusted Expenditure Per Capita* (EPC), while the three indicators of Cluster 3 tend to have low values. This is also in line with the values in Table 1. Table 1 shows the original data for ease of interpretation, but the data used to create clusters is normalized data.

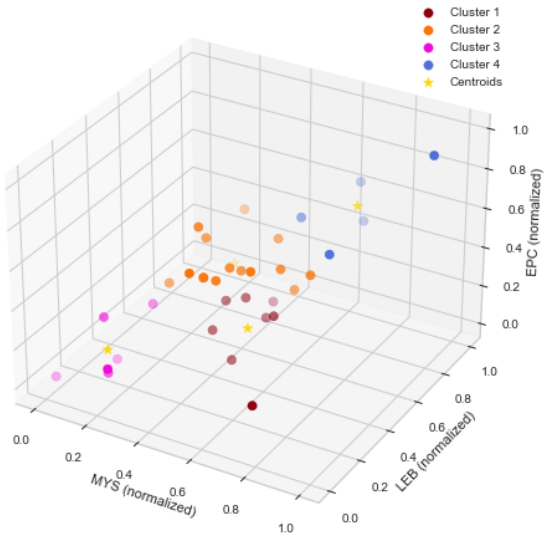


Figure. 2 Three-dimensional plot K-Means method

Table 1 shows that the standard deviation of each indicator in Cluster 4 is the largest compared to other clusters. The *life expectancy at birth* of the Indonesian population is between 65-75 years with a standard deviation of about 1-2 years. The average *expected years of schooling* is around 12-14 years, which illustrates that most Indonesians have the opportunity to pursue formal education up to high school. However, the *mean years of schooling* is lower than the *expected years of schooling*, that is, not all people complete their education up to high school (12 years), even only elementary and junior high schools. *Adjusted expenditure per capita* of the population varies from around 7 million rupiahs to 19 million rupiahs.

Table 1. Characteristics of HDI indicators for each cluster

Indicators	Cluster			
	1 (N = 8)	2 (N = 15)	3 (N = 6)	4 (N = 5)
Life Expectancy at Birth				
Mean	69.36	71.46	66.90	73.22
Std. Deviation	1.42	1.39	1.02	1.82
Minimum	66.45	69.13	65.63	70.50
Maximum	71.37	74.57	68.51	75.08
Expected Years of Schooling				
Mean	13.78	12.90	12.92	13.81
Std. Deviation	0.37	0.35	0.94	1.08
Minimum	13.31	12.18	11.14	12.99
Maximum	14.37	13.53	13.96	15.65
Mean Years of Schooling				
Mean	9.35	8.58	7.71	10.15
Std. Deviation	0.43	0.57	0.38	0.74
Minimum	8.89	7.59	7.02	9.39
Maximum	10.19	9.68	8.08	11.31
Adjusted Expenditure Per Capita				
Mean	9,932.38	11,262.33	8,975.00	14,892.20
Std. Deviation	976.43	1,059.55	1,503.52	2,376.85

Vol. 02, No. 01, 2023, pp. 27-33
e-ISSN: 2964-6162

Minimum	8,398.00	9,350.00	7,146.00	12,641.00
Maximum	11,130.00	13,358.00	10,687.00	18,927.00

The grouping obtained from the K-Means algorithm is illustrated with a map of Indonesia in Figure 3. The provinces on Sumatera Island are grouped into Cluster 1 (4 provinces), Cluster 2 (5 provinces), and Cluster 4 (1 province). Provinces on Kalimantan Island are included in Cluster 2, except for East Kalimantan Province (Cluster 4). Provinces on Java Island are included in Cluster 2, except for the Special Region of Yogyakarta (Cluster 4). Each of the 2 provinces in Sulawesi Island is included in Cluster 1 and Cluster 2, and 3 other provinces are included in Cluster 3. Provinces in Papua Island are included in Cluster 3. Provinces of North Maluku and Maluku are included in Cluster 1, the Province of Bali is grouped in Cluster 4, and Provinces in Nusa Tenggara are included in Cluster 3.

Figure. 3 Grouping human development with K-Means

B. Clustering Human Development Based on HDI Categories

The HDI value is obtained from the calculation of the geometric mean as in (11). Based on BPS cutoff points, the HDI values of the provinces in Indonesia in 2022 fall into the medium category (8 provinces), high category (24 provinces), very high category (2 provinces), and none are included in the low category. The characteristics of these categories based on HDI indicators are described in Table 2.

Table 2. Characteristics of HDI indicators for each HDI category

Indicators	Category		
	Medium (N = 8)	High (N = 24)	Very High (N = 2)
Life Expectancy at Birth			
Mean	67.65	71.03	74.20
Std. Deviation	1.74	1.82	1.24
Minimum	65.63	66.45	73.32
Maximum	71.02	74.62	75.08
Expected Years of Schooling			
Mean	12.99	13.23	14.37
Std. Deviation	0.86	0.54	1.82
Minimum	11.14	12.18	13.08
Maximum	13.96	14.37	15.65
Mean Years of Schooling			
Mean	7.89	9.02	10.53
Std. Deviation	0.64	0.67	1.10
Minimum	7.02	7.93	9.75
Maximum	9.24	10.37	11.31
Adjusted Expenditure Per Capita			
Mean	8,950.38	11,320.54	16,704.50
Std. Deviation	1,296.99	1,382.84	3,143.09
Minimum	7,146.00	8,876.00	14,482.00
Maximum	10,687.00	14,469.00	18,927.00

For all indicators, the highest average and the lowest average are in the very high and medium category respectively. This also corresponds to the maximum and minimum values. The minimum values of LEB, EYS, MYS, and EPC are 65.63 years, 11.14 years, 7.02 years, and 7.146 million rupiahs respectively as shown in Table 1 and Table 2. The lowest LEB is in West Sulawesi Province and the lowest EYS, MYS, and EPC are in Papua Province. Figure 4 presents Papua, West Sulawesi, West Papua, West Kalimantan, East Nusa Tenggara, West Nusa Tenggara, North Maluku, and Gorontalo fall into the medium (M) category. The Special Region of Yogyakarta has the highest LEB (75.08 years) and EYS (15.65 years). The Special Capital Region of Jakarta has the highest MYS (11.31 years) and EPC (18.927 million rupiahs). The Special Capital Region of Jakarta and the Special Region of Yogyakarta have been included in the very high (VH) category, but 24 other provinces are included in the high (H) category as shown in Figure 4.

Figure. 4 Grouping human development based on HDI categories

C. Comparison K-Means Clustering and HDI Categories

The tabulation between grouping by K-Means and HDI categories in Table 3 is used to compare the results of grouping the two. Provinces that are grouped in Cluster 1 and Cluster 2 are included in the medium and high categories. Provinces that are members of Cluster 4 are included in the high and very high categories, while the 6 provinces that are included in Cluster 3 are categorized in the medium HDI.

Table 3. Comparison K-Means Clustering and HDI Categories

Categories	**Cluster**				**Total**
	1	**2**	**3**	**4**	
Medium	1	1	6	0	8
High	7	14	0	3	24
Very High	0	0	0	2	2
Total	8	15	6	5	34

After grouping the provinces in Indonesia (Figure 3 and Figure 4), we compare the distribution of HDI values for each group. Table 4 shows that the range of HDI values for each cluster is not mutually exclusive because there is overlap, that is one value enters another cluster. For example, a score of 70 is included in Cluster 1 and Cluster 2. This is different from the grouping of HDI based on HDI categories (Table 5) which are already categorized as mutually exclusive and exhaustive. Cluster 3 has a similar HDI range as the medium category, which is between 61.39 to 69.81. The HDI in the very high category tends to be similar to Cluster 4, although the ranges of values for the two are different. This similarity can also be seen from the grouping results in Figure 3 and Figure 4. Cluster 1

and Cluster 2 have almost the same range, namely between 68-74. However, when compared to Table 2, Cluster 1 tends to have higher *Expected Years of Schooling* and *Mean Years of Schooling* scores than Cluster 2, while Cluster 2 tends to have higher *Life Expectancy at Birth* and *Adjusted Expenditure Per Capita* than Cluster 1. With characteristics that these differences, the K-Means method makes them into different clusters.

Table 4. HDI of each cluster

Cluster	Average	Min	Max
1	71.64	69.47	73.26
2	72.12	68.63	73.81
3	66.56	61.39	69.81
4	78.53	76.44	81.65

Table 5. HDI of each category

Category	Average	Min	Max
Medium	67.18	61.39	69.81
High	72.80	70.22	77.44
Very High	81.15	80.64	81.65

V.CONCLUSION

Provinces in Indonesia are clustered into 4 clusters based on the Human Development Index indicators in 2022 using the K-Means method. Based on BPS criteria, the HDI values of provinces in Indonesia fall into 3 categories, namely medium, high, and very high. A comparison of the two groupings shows that there is compatibility, namely the medium category tends to be similar to Cluster 3, while the very high category is similar to Cluster 4. Provinces that are included in Cluster 3 need to get attention to improving all HDI indicators. The members of Cluster 4 also need equity in human development in each regency or city. Provinces that are included in Cluster 1 and Cluster 2 have almost the same HDI ranges, but the characteristics of HDI indicators are different so K-Means makes them in different clusters. Therefore, provinces that are members of Cluster 1 have to improve on *Life Expectancy at Birth* and *Adjusted Expenditure Per Capita*, while Cluster 2 focuses on *Expected Years of Schooling* and *Mean Years of Schooling*.

REFERENCES

[1] L. Anggraini and P. R. Arum, "Analisis Cluster Menggunakan Algoritma K-Means Pada Provinsi Sumatera Barat Berdasarkan Indeks Pembangunan Manusia Tahun 2021," *Prosiding Seminar Nasional UNIMUS*, 2022, pp. 636–646. [Online]. Available: https://prosiding.unimus.ac.id/index.php/semnas/article/view/1214.

[2] Badan Pusat Statistik, *Indeks Pembangunan Manusia 2021*. Jakarta: BPS, 2022.

[3] D. Rahmawati and H. Bimanto, "Perbandingan Spatial Autoregressive Model dan Spatial Error Model dalam Pemodelan Indeks Pembangunan Manusia di Provinsi Jawa Timur," *J. Stat. Dan Apl.*, vol. 5, no. 1, pp. 41–50, Jun. 2021, doi: 10.21009/JSA.05104.

[4] N. Nurhasanah, N. Salwa, L. Ornila, A. Hasan, and M. Mardhani, "Classifying regencies and cities on human development index dimensions: Application of K-Means cluster analysis," *J. Sains Sosio Hum.*, vol. 5, no. 2, pp. 913–918, Dec. 2021, doi: 10.22437/jssh.v5i2.15801.

[5] E. Luthfi and A. W. Wijayanto, "Analisis perbandingan metode hirearchical, k-means, dan k-medoids clustering dalam pengelompokkan indeks pembangunan manusia Indonesia," *INOVASI*, vol. 17, no. 4, pp. 761–773, Dec. 2021, doi: 10.30872/jinv.v17i4.10106.

[6] United Nations Development Programme, *Human Development Report 2021/2022 Uncertain Times, Unsettled Lives: Shapping our Future in a Transforming Word*. New York: UNDP, 2022.

[7] Q. Zhou, J. Guo, Y. Yan, M. Wen, and S. Xia, "Cluster Analysis of Talent Portraits Based on K-means Algorithm," *2022 3rd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, Xi'an, China: IEEE, 2022, pp. 1–4. doi: 10.1109/ICBAIE56435.2022.9985923.

[8] Y. Li and Y. Yang, "K-means algorithm for cluster analysis in human resource information management system (HRIMS)," *The International Conference on Forthcoming Networks and Sustainability (FoNeS 2022)*, Hybrid Conference, Nicosia, Cyprus: IET, 2022, pp. 629–635. doi: 10.1049/icp.2022.2523.

[9] A. A. Arifiyanti, F. S. Darusman, and B. W. Trenggono, "Population Density Cluster Analysis in DKI Jakarta Province Using K-Means Algorithm," *J. Inf. Syst. Inform.*, vol. 4, no. 3, pp. 772–783, Sep. 2022, doi: 10.51519/journalisi.v4i3.315.

[10] Belia Mailien, A. Salma, Syafriandi, and D. Fitria, "Comparison K-Means and Fuzzy C-Means Methods to Grouping Human Development Index Indicators in Indonesia," *UNP J. Stat. Data Sci.*, vol. 1, no. 1, pp. 23–30, Feb. 2023, doi: 10.24036/ujsds/vol1-iss1/4.

[11] A. N. Ambarwati, "Latent Class Cluster Analysis untuk Pengelompokan Kabupaten/Kota di Provinsi Jawa Tengah Berdasarkan Indikator Indeks Pembangunan Manusia 2017," *Var. J. Stat. Its Appl.*, vol. 1, no. 2, pp. 46–54, Jan. 2020, doi: 10.30598/variancevol1iss2page46-54.

[12] M. F. J. Muttaqin, "Cluster Analysis Using K-Means Method to Classify Sumatera Regency and City Based on Human Development Index Indicator," *Semin. Nas. Off. Stat.*, vol. 2022, no. 1, pp. 967–976, Nov. 2022, doi: 10.34123/semnasoffstat.v2022i1.1299.

[13] H. Wang, J.-H. Feil, and X. Yu, "Let the data speak about the cut-off values for multidimensional index: Classification of human development index with machine learning," *Socioecon. Plann. Sci.*, vol. 87, p. 101523, Jun. 2023, doi: 10.1016/j.seps.2023.101523.

[14] N. Imani, A. I. Alfassa, and A. M. Yolanda, "Self Organizing Map (SOM) Clustering untuk Analisis Data Indikator Sosial di Provinsi Nusa Tenggara Timur," *J. Gaussian*, vol. 11, no. 3, pp. 458–467, Jan. 2023, doi: 10.14710/j.gauss.11.3.458-467.

[15] T. C. Mayvani, R. Afin, A. R. Idialis, and Sariyani, "Analysis of Growth and Tourism Clusters in Madura," *J. Ekon. Dan Studi Pembang.*, vol. 14, no. 1, p. 59, Mar. 2022, doi: 10.17977/um002v14i12022p059.

[16] R. Steinmann, L. Seydoux, É. Beaucé, and M. Campillo, "Hierarchical Exploration of Continuous Seismograms with Unsupervised Learning," *J. Geophys. Res. Solid Earth*, vol. 127, no. 1, Jan. 2022, doi: 10.1029/2021JB022455.

[17] X. Wang, A. Shen, X. Hou, and L. Tan, "Research on cluster system distribution of traditional fort-type settlements in Shaanxi based on K-means clustering algorithm," *PLOS ONE*, vol. 17, no. 3, p. e0264238, Mar. 2022, doi: 10.1371/journal.pone.0264238.

[18] H. Liu, "Evaluation of Logistics Competitiveness of Node Cities along the New Land-Sea Corridor in Western China—Based on Factor Analysis and Cluster Analysis," *Int. J. Front. Sociol.*, vol. 5, no. 3, 2023, doi: 10.25236/IJFS.2023.050306.

[19] E. B. Mariano, D. Ferraz, and S. C. De Oliveira Gobbo, "The Human Development Index with Multiple Data Envelopment Analysis Approaches: A Comparative Evaluation Using Social Network Analysis," *Soc. Indic. Res.*, vol. 157, no. 2, pp. 443–500, Sep. 2021, doi: 10.1007/s11205-021-02660-4.

[20] M. Hassani and T. Seidl, "Using internal evaluation measures to validate the quality of diverse stream clustering algorithms," *Vietnam J. Comput. Sci.*, vol. 4, no. 3, pp. 171–183, Aug. 2017, doi: 10.1007/s40595-016-0086-9.

[21] 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.

[22] X. Wang and Y. Xu, "An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 569, no. 5, p. 052024, Jul. 2019, doi: 10.1088/1757-899X/569/5/052024.

Vol. 02, No. 01, 2023, pp. 27-33
e-ISSN: 2964-6162

Indah Fahmiyah graduated with a bachelor's and master's degree in Statistics from the Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia in 2016 and 2018 respectively. She is a lecturer of the Data Science Technology Study Program, Faculty of Advanced Technology and Multidiscipline, Universitas Airlangga, Indonesia. His research interests are predictive analytics, regression, multivariate analysis, time series analysis, and data mining.

Ratih Ardiati Ningrum received her bachelor's degree in Statistics from the Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia in 2017. In the same year, she took a master's program in Statistics at the same institute. She had a joint degree with the Institute of Statistics at National Chiao Tung University, Hsinchu, Taiwan in the second year while undergoing a master's program. She received her master's degree in 2019. She is a lecturer of the Data Science Technology, Faculty of Advanced Technology and Multidiscipline, Universitas Airlangga, Indonesia. Her research interests are survival analysis, machine learning, and multivariate analysis.