Journal of Research in Mathematics Trends and Technology
Journal homepage: https://talenta.usu.ac.id/jormtt

Implementation of K-Means Clustering to Human Development Indicators in East Nusa Tenggara

Justin Eduardo Simarmata[*1], Debora Chrisinta[2], Miko Purnomo[3]
[1]*Mathematics Education Study Program, University of Timor, Kefamenanu, 85613, Indonesia*
[2]*Information Technology Study Program, University of Timor, Kefamenanu, 85613, Indonesia*
[3]*Mathematics Study Program, University of Timor, Kefamenanu, 85613, Indonesia*
[*]Corresponding Author: justinesimarmata@unimor.ac.id

ARTICLE INFO

Article history:
Received: 1 June 2024
Revised: 3 August 2024
Accepted: 5 September 2024
Available online: 30 September 2024

E-ISSN: 2656-1514
P-ISSN: -

How to cite:
Simarmata, J.E., Chrisinta, D., Purnomo, M, "Implementation of K-Means Clustering to Human Development Indicators in East Nusa Tenggara," Journal of Research in Mathematics Trends and Technology and, vol. V6, no. 2, Sep. 2024, doi: 10.32734/jormtt.v6i2.17066

ABSTRACT

K-Means has been adopted to group various cases related to the quality of human resources and economic growth. This study aims to apply K-Means to the characteristics based on selected Human Development Index (HDI) indicators, namely the average length of schooling, the expectation of length of schooling and life expectancy in Province of East Nusa Tenggara. The optimal cluster obtained is 6 clusters. The cluster with the highest average value of all variables is in cluster 1 which is Kupang City area. Meanwhile, condition cluster 6 provides the smallest life expectancy value compared to other clusters. The smallest average variable of school length is in cluster 2. The regions that provide the smallest number expectation of length of schooling is in cluster 5.

Keyword: K-Means, the average length of schooling, the expectation of length of schooling, the life expectancy

http://doi.org/10.32734/jormtt.v6i2.17066

1. INTRODUCTION

The success of a country or nation's development is highly dependent on the quality of its human resources, and one of the key factors that influence this is education. This is because education plays a crucial role in enabling individuals to attain a decent standard of living [1]. The importance of the government's role in improving the quality of human resource development is a benchmark for determining the success of the implementation of order in various fields, especially the economy of the community. According to [2] the Human Development Index (HDI) consists of three components: health, education, and economy. Given that two components, namely economy and education, are two references that are used simultaneously to see the HDI, then indirectly these two aspects influence each other.

Other aspects that also affect the economic growth of a region besides education can be seen based on the life expectancy of the community [3]. This is because life expectancy is also related to the quality of human

data sources. If people have a long life expectancy and do not have skills, it will lead to a decline in regional development, and this is supported by the limited availability of jobs for people who are in the elderly category but are still able to work [4]. Based on the problems that arise, in order to minimize the decline in economic growth and human resources, it is necessary to map the distribution of average length of schooling, expected length of schooling, and life expectancy in the community. Identification of the region is carried out by selecting provinces that show a high poverty rate, namely the East Nusa Tenggara Province. BPS noted that East Nusa Tenggara is in third place as the region with the poorest percentage of provinces after Papua and West Papua Provinces.

The mapping process of the distribution characteristics of the selected HDI indicators, namely the average length of schooling, the expected length of schooling, and life expectancy in the East Nusa Tenggara region involves observation objects, namely all regencies in East Nusa Tenggara. The mapping of characteristics is carried out by applying the concept of the clustering method [5]; [6]. This method is used to group objects based on their similarity measure. The objects that tend to have similarities will be in the same group [7]; [8].

The clustering concept has been widely adopted to group various cases related to the quality of human resources and economic growth. Based on [9], conducted a cluster analysis using the K-Means method to group HDI indicators in Maluku Province. conducted clustering on the level of basic dimension indicators in Banten and DKI Jakarta Provinces which also used the K-Means method [10]. [11] clustered poor areas in Riau Province using the K-Means method. [12] dan [13] in their research used the K-means method to cluster the Human Development Index components in regencies/cities in Central Java.

Referring to the research that has been carried out in the use of the K-Means method for grouping in various problems around the HDI indicators, this study will group the selected HDI indicators, namely the average length of schooling, the expected length of schooling, and life expectancy in East Nusa Tenggara Province. This is done to determine the distribution of the level of human resource quality in East Nusa Tenggara Province in terms of education and life expectancy.

2. METHODS

2.1. Cluster Analysis

Cluster analysis is a technique used to group objects with multiple variables into one similar characteristic [14]. The determination of characteristics is based on the proximity or distance between objects. The object size used in this section is the Euclidean distance, the formula for which is as follows [15]:

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \left(x_{ik} - x_{jk}\right)^2} \quad (1)$$

where,

x_{ik} : Object i on variables 1, 2, ..., p

x_{jk} : Object j on variables 1, 2, ..., p

p : Number of variables

d_{ij} : Distance between objects i and j

2.2. K-Means Clustering

Hierarchical cluster analysis is a technique for clustering objects with a predetermined number of clusters and allocating objects to clusters based on the nearest distance. The advantage of this method is that it can solve clustering from larger data than hierarchical methods[16]-[19]. It has almost the same weaknesses as hierarchical methods, namely sensitivity to outliers, and the choice of distance measure also needs to be considered. One of the non-hierarchical methods is K-Means. The steps for cluster formation in K-Means are:

(a) Determining the number of clusters,
(b) Determining the center randomly,
(c) Placing objects into the closest cluster based on the nearest distance,
(d) Determining the new center, and
(e) Repeat steps 3 and 4 until there is no change in the objects in each cluster formed.

2.3. Data Source

The data used in this study is secondary data published by BPS (Badan Pusat Statistik) in the East Nusa Tenggara Province region in 2022. The research variables used come from 3 variables, namely:

(a) Expected length of schooling (X1)
(b) Average length of schooling (X2)
(c) Life expectancy (X3)

The definition of expected length of schooling is the duration of schooling in years that is expected to be experienced by children of a certain age in the future at various levels of education. The average length of schooling is the average number of years spent by people aged 15 and over in completing all types of education levels they have ever attended. Meanwhile, the definition of life expectancy is the average number of ages that are expected to occur for a person according to the death rate at a certain time and tends not to change in the future. These variables were selected based on the HDI indicators based on the education aspect and also to see the areas with the lowest quality of life expectancy in East Nusa Tenggara Province. Therefore, the research variable also includes life expectancy.

2.4. Data Analysis Stages

The analysis in this study applies the clustering concept, which is commonly known as clustering. The purpose of choosing cluster analysis is to determine the characteristics of the regions in East Nusa Tenggara Province based on the variables of expected length of schooling, average length of schooling, and life expectancy. The results of cluster formation are based on the KMeans clustering method. The concept of clustering is to group objects based on similarity measures (distance measures). The distance measure used in this study is the Euclidean distance. The assumption that must be met when applying the Euclidean distance is the similarity of the units of the variables used. Therefore, this study will also apply Principal Component Analysis (PCA) to overcome the assumption of the distance measure that must be met. The analysis stages are carried out using RStudio Software with the following steps:

(a) Data preprocessing: This stage is carried out to prepare the data for data analysis. It consists of missing data identification, data transformation, data distribution identification, and summary data.
(b) Applying cluster analysis: Applying cluster analysis to the data that has been preprocessed.
(c) Determining the optimal cluster: Determining the optimal number of clusters.
(d) Visualizing optimal clusters: Visualizing optimal clusters.
(e) Conducting evaluation and drawing conclusions: Conducting evaluation and drawing conclusions from the analysis results.

3. RESULTS AND DISCUSSION

3.1 Data Preprocessing

The initial stage of the clustering process for the selected HDI indicators, namely average length of schooling, expected length of schooling, and life expectancy in East Nusa Tenggara Province, was to check for missing data. The output results are shown in Table 1. Based on Table 1, it can be seen that all three variables do not show any missing data. This is indicated in the last column, which shows a value of 0% in the missing column.

Tabel 1. Output of Missing Data Identification

No	Variable	Stats. (Value)	Freqs (% of Valid)	Valid	Missing
1	Expected Length of Schooling [Numeric]	• Mean (sd): 13.1 (0.9) • min < med < max : 12.3 < 13 < 16.4 • IQR (CV) : 0.8 (0.1)	21	22 (100%)	0 (0%)
2	Average Length of Schooling [Numeric]	• Mean (sd): 7.7 (1.1) • min < med < max : 6.4 < 7.5 < 11.6 • IQR (CV) : 1 (0.1)	22	22 (100%)	0 (0%)
3	Life Expectancy [Numeric]	• Mean (sd): 66.6 (2.2) • min < med < max : 60.9 < 67.3 < 70.1 • IQR (CV) : 2.5 (0.1)	20	22 (100%)	0 (0%)

Furthermore, Table 1 shows the number of different values for each variable in the Freqs (% of Valid) column, while the Valid column shows that all values are in the numeric category and there are no errors in the data input. To see the distribution of the data, refer to the Stats. (Value) column, which provides an overview of the range of average values for each variable along with the standard deviation, and the data centering value used is the median value flanked by the minimum and maximum values. It also provides the IQR (Interquartile Range) and CV (Coefficient of Variation) values. For example, the expected length of schooling variable has an average value of 13.1 and a standard deviation of 0.9. This means that the average duration of schooling in years that is expected to be experienced by children of a certain age in the future at various levels of education

is around 13.1 years, and the level of spread of the collected expected length of schooling values is 0.9. Furthermore, the median value of expected schooling is obtained as 13, while the smallest and largest values obtained are 12.3 and 16.4, respectively. In addition, the data range that is at 75% of the data and 25% of the data is obtained as 0.8 (IQR) and the coefficient of variation of the data (CV) is obtained as 0.1, which means that the data collected has a fairly small variation.

The next process is data transformation and normalization to ensure that the units on the variables have the same units. This is done because the similarity measure used in the clustering process uses the Euclidean distance, which has the assumption that the units of the variables must be the same to avoid bias in the resulting similarity measure. Data normalization is performed simultaneously based on the results of data transformation. The output of the data transformation is presented in Table 2.

Table 2. Data Transformation

No	Before Transformation			After Transformation		
	Expected Length of Schooling	Average Length of Schooling	Life Expectancy	Expected Length of Schooling	Average Length of Schooling	Life Expectancy
1	13.15	6.85	67.35	0.04540	-0.77942	0.36321
2	12.85	7.33	65.38	-0.29126	-0.32644	-0.54934
3	13.88	7.41	65.28	0.86461	-0.25094	-0.59566
4	12.6	6.76	66.68	-0.57182	-0.86435	0.05285
5	13.34	7.97	67.35	0.25862	0.27753	0.36321
⋮	⋮	⋮	⋮	⋮	⋮	⋮
22	16.43	11.61	70.11	3.72624	3.71262	1.64169

The transformation results presented in Table 2 show the changes in the values of the variables observed. These values show that there is no difference in the observation units and the numbers are already in the same condition based on the standard normal transformation formula used. The transformation formula and calculation process used are as follows:

$$Z = \frac{x - \bar{x}}{\sigma} = \frac{13.15 - 13.1}{0.9} \cong 0.04540$$

Furthermore, the difference between the results obtained and manual calculations is due to rounding of numbers. The calculation process when conducting further analysis is carried out with the help of Software, which reduces calculation errors and the actual analysis results.

The transformation results that have been carried out are then applied to the QQ-Plot to ensure that the data has shown a condition approaching normal distribution. Based on Figures 1, 2 and 3, it can be seen that the data distribution points are around the red linear line. This has shown that the data tends to have approached a normal distribution.

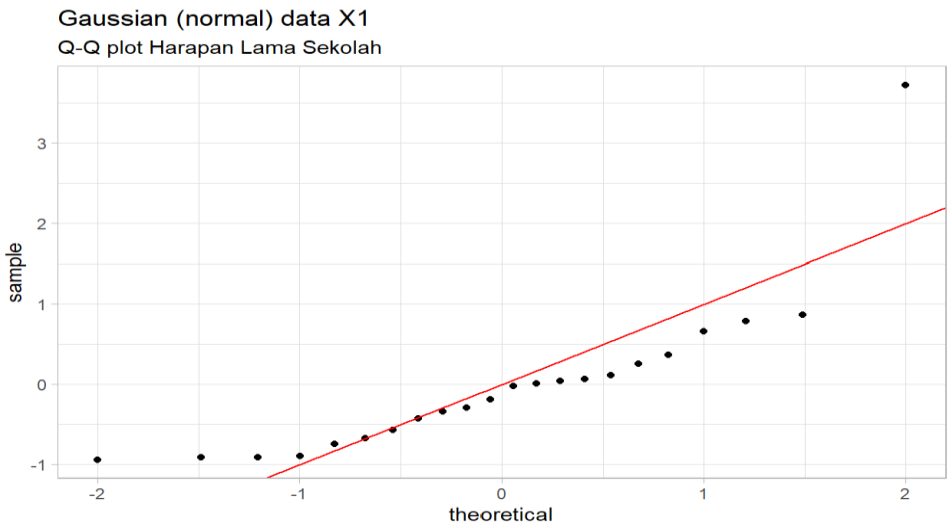

Figure 1. QQ-Plot of X1 After Data Transformation

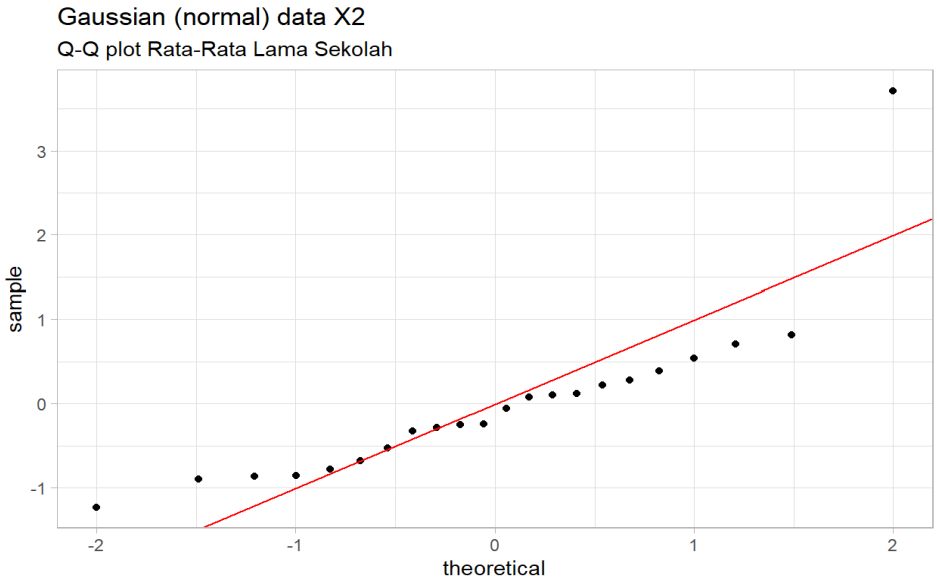

Figure 2. QQ-Plot of X2 After Data Transformation

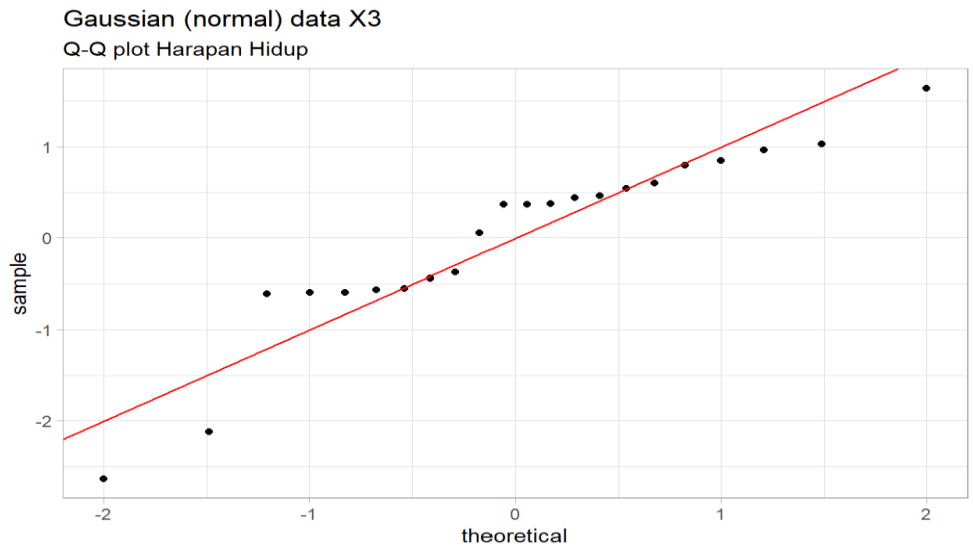

Figure 3. QQ-Plot of X3 Before Data Transformation

After ensuring that the variables to be clustered have met the requirements, the first step is to look at a general overview of the condition of the variables used based on descriptive statistics. The output results of descriptive statistics are presented in Table 3.

Table 3. Descriptive Statistics

Variabel	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Expected Length of Schooling	22	13.11	0.891	12.27	12.533	13.307	16.43
Average Length of Schooling	22	7.676	1.06	6.37	7	7.955	11.61
Life Expectancy	22	66.566	2.159	60.87	65.35	67.828	70.11

Based on Table 3, it can be seen that there are 22 districts in East Nusa Tenggara Province with an average expected length of schooling of 13.11, an average length of schooling of 7.67 and a life expectancy of 66.56. These figures are close to the figures obtained at the East Nusa Tenggara Provincial level, namely an expected length of schooling of 13.21, an average length of schooling of 7.70 and a life expectancy of 67.47. This shows a relatively normal condition for the life expectancy variable because it does not have a range that is far from the national life expectancy according to BPS (2022) of 71.85 years. Meanwhile, the education aspect seems to give figures in the category of relatively low because it is smaller and tends to be almost the same when compared to the national average length of schooling and expected length of schooling which are at 8.69 and 13.10 respectively. Therefore, this situation needs to be grouped to see the districts/cities that cause the low expected and average length of schooling.

3.2 K-Means Clustering Analysis

K-Means Clustering is one of the methods in clustering that belongs to the non-hierarchical technique and works on variables that come from numerical variables. The process of applying K-Means is done by first determining the optimal cluster. The determination of the optimal cluster is based on the average value of the Silhouette index by carrying out the clustering process with the number of groups (k) formed starting from 1 to 10. The optimal cluster is selected at the value k=6 which has shown consistency in the change in the movement of the average value of the Silhouette index. The overall results of the optimal cluster determination can be seen in Figure 4.

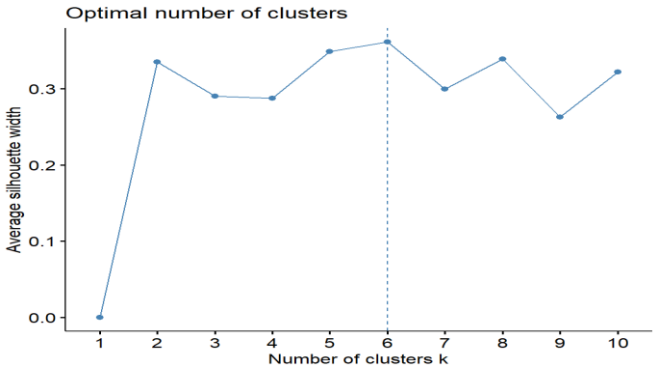

Figure 4. Determination of Optimal Cluster

After obtaining the optimal cluster, the next step is to visualize the 6 cluster groups that have been formed. Visualization uses the average value of the variables in each cluster that is formed and is shown in Figure 5.

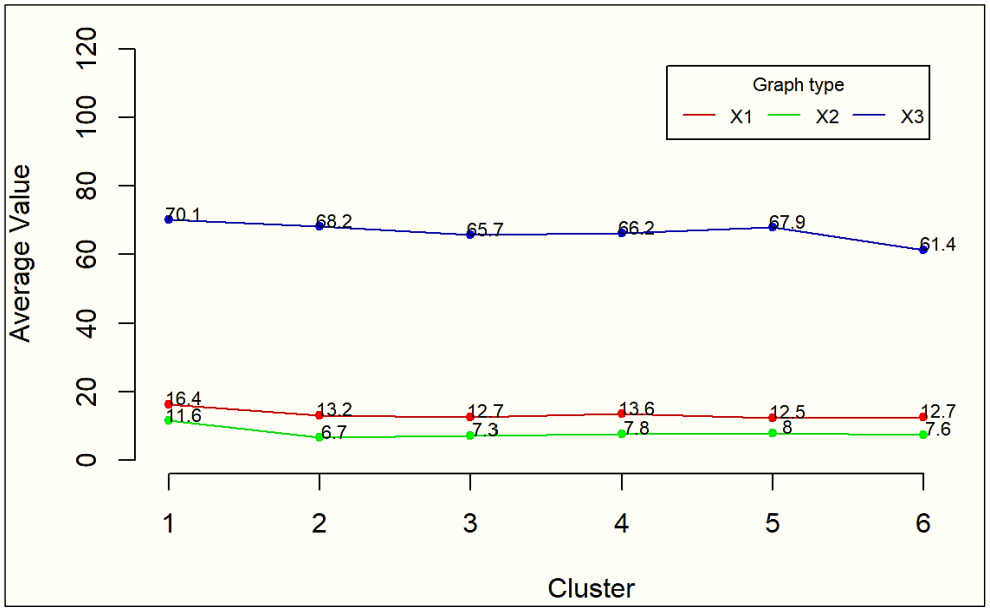

Figure 5. Visualization of Optimal Cluster Based on Average Value

Based on Figure 5, cluster 1 shows a higher average value than other clusters and refers to the Kupang City area. This indicates that Kupang City has a higher life expectancy, expected length of schooling and average length of schooling compared to other areas in East Nusa Tenggara Province. However, for the condition of Cluster 6, it gives the smallest life expectancy value compared to other clusters, consisting of the Kupang, North Central Timor, Ende, Manggarai, and Rote Ndao areas. The area that gives the lowest expected length of schooling is in cluster 5 which includes the West Sumba, Sikka, Central Sumba, and Southwest Sumba areas. Meanwhile, the lowest average length of schooling is in cluster 2 which includes the Lembata, Ngada, West Manggarai, Nageko, and East Manggarai areas.

The areas that are in other clusters are in a condition where the life expectancy, expected length of schooling and average length of schooling are not much different from the areas that are included in the smallest category. The conditions obtained based on the characteristics of each optimal cluster that has been formed can be used as a reference to improve the quality of life expectancy, expected length of schooling and average length of schooling, especially in areas that provide lower conditions compared to other areas based on the clusters that have been identified. In detail, the descriptive statistics for each cluster are presented in Table 1.

Table 4. Descriptive Statistics of Optimal Cluster

Cluster 1: Kota Kupang

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
X1	1	16.43	16.43	16.43	16.43	16.43	16.43
X2	1	11.61	11.61	11.61	11.61	11.61	11.61

X3	1	70.11	70.11	70.11	70.11	70.11	70.11

Cluster 2: Lembata, Ngada, Manggarai Barat, Nageko, Manggarai Timur

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
X1	5	12.46	0.176	12.3	12.31	12.51	12.73
X2	5	7.984	0.429	7.42	7.8	8.25	8.54
X3	5	67.9	0.416	67.52	67.56	68.29	68.4

Cluster 3: Alor, Sabu Raijua

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
X1	2	12.72	0..636	12.27	12.495	12.945	13.17
X2	2	7.6	1.174	6.77	7.185	8.015	8.43
X3	2	61.43	0.792	60.87	61.15	61.71	61.99

Cluster 4: Sumba Timur, TTS, Belu, Flores Timur, Malaka

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
X1	5	12.7	0.256	12.6	12.6	12.85	12.94
X2	5	7.276	0.377	7.12	7.12	7.38	7.79
X3	5	65.66	0.585	65.62	65.62	65.62	66.68

Cluster 5: Sumba Barat, Sikka, Sumba Tengah, Sumba Barat Daya

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
X1	4	13.2	0.162	13.09	13.113	13.222	13.44
X2	4	6.728	0.256	6.37	6.64	6.877	6.96
X3	4	68.162	0.679	67.35	67.732	68.685	68.79

Cluster 6: Kupang, Timor Tangah Utara, Ende, Manggarai, Rote Ndao

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
X1	5	13.588	0.296	13.21	13.34	13.81	13.88
X2	5	7.77	0.271	7.41	7.62	7.97	8.09
X3	5	66.206	1.077	65.26	65.28	67.35	67.38

4. CONCLUSIONS

The mapping of characteristics by applying the concept of the K-Means clustering method based on the selected HDI indicators, namely the average length of schooling, expected length of schooling and life expectancy in the East Nusa Tenggara region involves observation objects, namely all districts in East Nusa Tenggara. The optimal cluster obtained is as many as 6 clusters. The cluster with the highest average value for all variables is in cluster 1, which only contains the Kupang City area. Meanwhile, for the condition of cluster 6, it gives the smallest life expectancy value compared to other clusters. The variable with the lowest average length of schooling is in cluster 2. The area that gives the lowest expected length of schooling is in cluster 5. The results of the areas that have been identified based on the optimal cluster can be a reference for the government to implement appropriate policies in order to increase the needs in areas that tend to provide lower average length of schooling, expected length of schooling and life expectancy compared to other areas.

REFERENCES

[1] R. Y. Sari, H. Oktavianto, and H. W. Sulistyo, "Algoritma K-Means Dengan Metode Elbow Untuk Mengelompokkan Kabupaten/Kota di Jawa Tengah Berdasarkan Komponen Pembentuk Indeks Pembangunan Manusia," Jurnal Smart Teknologi, vol. 3, no. 2, pp. 104–108, 2022.

[2] B. P. Statistik, "Indeks pembangunan manusia," Academic press, 2020.

[3] R. Muda, R. A. Koleangan, and J. B. Kalangi, "Pengaruh angka harapan hidup, tingkat pendidikan dan pengeluaran perkapita terhadap pertumbuhan ekonomi di sulawesi utara pada tahun 2003-2017," Jurnal Berkala Ilmiah Efisiensi, vol. 19, no. 01, pp. 44–55, 2019.

[4] I. Arofah and S. Rohimah, "Analisis Jalur Untuk Pengaruh Angka Harapan Hidup, Harapan Lama Sekolah, Rata-Rata Lama Sekolah Terhadap Indeks Pembangunan Manusia Melalui Pengeluaran Riil Per Kapita di Provinsi Nusa Tenggara Timur," Jurnal Saintika Unpam: Jurnal Sains dan Matematika Unpam, vol. 2, no. 1, pp. 76–87, 2019, doi: 10.32493/jsmu.v2i1.2920.

[5] D. Chrisinta, L. P. Gelu, and B. Baso, "Identifikasi Sebaran Karakteristik Kriminal di Indonesia Tahun 2021 Menggunakan Model-Based Clustering," Journal of Mathematics, Computations and Statistics, vol. 5, no. 2, pp. 98–105, 2022, doi: 10.35580/jmathcos.v5i2.36956.

[6] J. E. , Simarmata and S. Sutarman, "Object Classification with Classical Linear Discriminant Analysis and Robust Linear Discriminant Analysis," International Journal of Research Applied Science & Engineering Technology, vol. 6, no. 5, pp. 84–91, 2018, doi: 10.22214/ijraset.2018.5012.

[7] T. R. Mayasari, "Pengelompokkan Provinsi berdasarkan variabel kesehatan lingkungan dan pengaruhnya terhadap kemiskinan di Indonesia tahun 2018," Jurnal Siger Matematika, vol. 1, no. 1, pp. 24–30, 2020, doi: 10.23960%2Fjsm.v1i1.2471.

[8] D. Chrisinta, I. M. Sumertajaya, and Indahwati I., "Evaluasi Kinerja Metode Cluster Ensemble dan Latent Class Clustering Pada Peubah Campuran," Indonesian Journal of Statistics and Its Applications, vol. 4, no. 3, pp. 448–461, 2020, doi: 10.29244/ijsa.v4i3.630.

[9] M. W. Talakua, Z. A. Leleury, and A. W. Taluta, "Analisis cluster dengan menggunakan metode k-means untuk pengelompokkan Kabupaten/Kota di provinsi maluku berdasarkan indikator indeks pembangunan manusia tahun 2014," BAREKENG: Jurnal Ilmu Matematika Dan Terapan, vol. 11, no. 2, pp. 119–128, 2017, doi: 10.30598/barekengvol11iss2pp119-128.

[10] R. M. Reviansyah, "Klasterisasi Tingkat Indikator Dimensi Dasar Manusia Menggunakan Algoritma K-Means (Studi Kasus Provinsi Banten Dan Provinsi DKI Jakarta)," Applied Mathematics and Computation, vol. 219, no. 9, 2018, doi: 10.1016/j.amc.2012.10.053.

[11] F. Isyarah, M. A. Hasan, and F. Wiza, "Clustering Daerah Miskin di Provinsi Riau Menggunakan Metode K-Means," Jurnal Teknologi Informasi: Teori, Konsep dan Implementasi, vol. 9, no. 1, pp. 1–12, 2018, doi: 10.31849/semaster.v1i1.5487.

[12] G. Z. Shafa, "Analisis K-Means Clustering Komponen Indeks Pembangunan Manusia 2021 Berdasarkan Kabupaten/Kota Di Jawa Tengah," 2022. Accessed: Jun. 27, 2024. [Online]. Available: https://dspace.uii.ac.id/handle/123456789/38880

[13] N. Lucyana, P. Sari, D. Kurniawan, and M. A. Buchari, "Analisis Penyebab Kecelakaan Pesawat di Indonesia Menggunakan Metode K-Means," Jusikom: Jurnal Sistem Komputer Musirawas, vol. 7, no. 2, pp. 89–105, 2022, doi: 10.32767/JUSIKOM.V7I2.1730.

[14] A. K. Jain, "Data clustering: 50 years beyond K-means," Pattern Recognit Lett, vol. 31, no. 8, pp. 651–666, 2010.

[15] R. R. B. B. Balaji and R. B. Bapat, "On Euclidean distance matrices," Linear Algebra Appl, vol. 424, no. 1, pp. 108–117, 2007, doi: 10.1016/j.laa.2006.05.013.

[16] A. S. Lubis, Zulfan, M. F. Chania, I. M. Adha, and F. Kumalasari, "Analysis of the Use and Application of Mathematics in Economics: Demand and Supply Functions," J. Res. Math. Trends Technol., vol. 6, no. 1, pp. 16–23, 2024, doi: 10.32734/jormtt.v6i1.17603.

[17] Y. B. P. Siringoringo, L. Sitinjak, and E. D. Tarigan, "Determining the Location of Nenas Processing Factories in North Sumatra Using Dijkstra Algorithm," J. Res. Math. Trends Technol., vol. 6, no. 1, pp. 1–7, 2024, doi: 10.32734/jormtt.v6i1.16978.

[18] P. Gultom, Miranda, E. S. M. Nababan, Mardiningsih, and Suyanto, "Integration of AHP and VIKOR Method to Select the Optimum Destination Route," J. Res. Math. Trends Technol., vol. 6, no. 1, pp. 24–34, 2024, doi: 10.32734/jormtt.v6i1.17717.

[19] P. Gultom, R. Widyasari, Suyanto, and J. L. Marpaung, "Model for Working Capital Management of Micro, Small and Medium Enterprises in Indonesia by Using Multiple Objective Stochastic Programming," J. Res. Math. Trends Technol., vol. 5, no. 2, pp. 1–11, 2023, doi: 10.32734/jormtt.v5i2.15937.