

## **A District/City Profiling Based on Poverty Indicators in East Nusa Tenggara Using the Centroid Linkage Algorithm**

**Andre Tri Rian Dani<sup>1\*</sup>, Yossy Candra<sup>2</sup>, Fachrian Bimantoro Putra<sup>3</sup>,  
Meirinda Fauziyah<sup>4</sup>**

<sup>1,4</sup> *Program Studi Statistika, Jurusan Matematika, FMIPA, Universitas Mulawarman, Samarinda, Indonesia*

<sup>2</sup> *Statistisi, Sekretariat Jenderal, Kementerian Agama, Jakarta, Indonesia*

<sup>3</sup> *BNI Staff Banking, Future Relationship Manager, BNI, Jakarta, Indonesia*

*\*Corresponding Author's Email: [andreatririandani@fmipa.unmul.ac.id](mailto:andreatririandani@fmipa.unmul.ac.id)*

### **ABSTRACT**

Poverty is a complex multidimensional phenomenon that significantly impacts human life. Poverty has always been a problem that the government has discussed regionally, centrally, and internationally. The issue of poverty is interesting to approach and analyze using a statistical approach, namely cluster analysis. Cluster analysis is used to group objects based on their level of similarity. In this research, the algorithm used is the Centroid Linkage Algorithm. The Centroid Linkage algorithm was chosen based on its advantages in the grouping process. Distance similarity measurement uses Squared Euclidean. The data used are district/city poverty indicators in East Nusa Tenggara Province. The analysis results show that two optimal clusters were obtained with their distinguishing characteristics. Hopefully, the results of this analysis can be used as a reference in formulating policies for alleviating poverty

**Keyword:** *Cluster Analysis, Centroid Linkage, Poverty, Squared Euclidean*

---

#### **Article info:**

*Received: May 3, 2025*

*Revised: June 6, 2025*

*Accepted: October 12, 2025*

*Available online: October 20, 2025*

---

#### **How to cite this article:**

Dani, A. T. R., Candra, Y., Putra, F. B., & Fauziyah, M. (2025). A District/City Profiling Based on Poverty Indicators in East Nusa Tenggara Using the Centroid Linkage Algorithm. *Zeta - Math Journal*, 10(2), 81-91. <https://doi.org/10.31102/zeta.2025.10.2.81-91>

---

© 2025 The Author. Published by the Mathematics Study Program, Universitas Islam Madura, Indonesia. This article is an open access article under the CC BY-NC-ND license (<https://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial, and no modifications or adaptations are made.

## 1. Introduction

Cluster analysis is a technique employed to categorize data according to their resemblances. This method employs the data's information to group objects together, resulting in the formation of a single cluster with a high degree of similarity (homogeneity) within the cluster (Hikmah, et.al., 2022). On the other hand, objects in different clusters will show apparent differences (Han, J., Kamber, M. & Pei, J., 2012)

This article employs cluster analysis to categorize districts/cities in East Nusa Tenggara Province according to poverty indicators. The employed algorithm is the Centroid Linkage algorithm. The selection of this algorithm was based on its merits, as it yields superior cluster outcomes compared to the Single Linkage and Complete Linkage algorithms. This is represented because, in the process, the Centroid Linkage algorithm combines clusters based on the average between all points in the cluster (Widodo, E., et.al., 2018) so that the resulting cluster will be more centralized and have lower variation. Apart from that, the Centroid Linkage algorithm is more stable in the presence of outlier data, efficient in the grouping process, and produces dendograms that are easier to interpret.

Various prior studies have been conducted by researchers using the Centroid Linkage algorithm.

Table 1. Previous Research

Article Title	Writer	Analysis Results
Cluster Analysis with Outlier Data using Centroid Linkage and K-Means Clustering for Grouping HIV/AIDS Indicators in Indonesia	Silvi, 2018	The research results show that the number of clusters formed was 7, with the Centroid Linkage algorithm producing more homogeneous groupings than K-Means.
Grouping Provinces Based on Internet Network Quality Using the Centroid Linkage Method	Dzikrullah, 2022	The cluster analysis research has yielded the conclusion that there are 4 distinct clusters. The provinces of Papua and West Papua are part of clusters that exhibit a tendency towards low internet network quality.
Implementation of the Centroid Linkage and K-Medoids Algorithms in Grouping Districts/Cities in South Sulawesi based on Education Indicators	Raja, Tinungki, Sirajang, 2024	The research results show that the Centroid Linkage algorithm produces 2 clusters with a minimum ratio of standard deviation within and between clusters, so the Centroid Linkage algorithm is the best algorithm.

Cluster analysis with the Centroid Linkage algorithm is applied to poverty problems. Poverty is a complex and crucial problem faced by various countries, including Indonesia. Poverty is a multidimensional problem measured by material deprivation and includes other aspects such as education, health, and access to essential services. This is because poverty is a crucial problem and must be handled seriously, so efforts are required to overcome it comprehensively. This study aims to move beyond single-dimension poverty metrics by clustering East Nusa Tenggara's districts and cities with the centroid-linkage method, thereby capturing the full spectrum of income, education, health, and infrastructure deprivation. The profiling of districts/cities based on poverty indicators in East Nusa Tenggara becomes reference material that can be used in formulating policies for alleviating the problem of poverty.

## 2. Literature Review

### 2.1 Data Mining

Data mining identifies exciting information hidden in large data sets stored in databases, warehouses, or other data storage sources (Tan, P. N., Steinbach, M. & Kumar, V., 2006). Data mining is commonly referred to as knowledge discovery in databases (KDD) (Nanang, A. & Susanti, D., 2018). KDD, or Knowledge Discovery in Databases, is a systematic process that involves gathering and examining past data in order to identify patterns, relationships, or regularities within large datasets, with the goal of generating valuable insights (Santosa, B., 2007). The process comprises three primary stages: initial data exploration, model construction and validation, and implementation of the created model (Prasetyo, E., 2012).

### 2.2 Cluster Analysis

Cluster analysis is the data mining technique employed to categorize a collection of objects into multiple clusters based on their similar attribute characteristics. The objective is to create clusters in which objects that share identical characteristics are grouped together, while objects that differ from each other are placed in separate clusters. The proximity of a cluster is directly proportional to the degree of similarity within it, and inversely proportional to the dissimilarity from other clusters.

According to Hair et al. (2014), cluster analysis is considered good if the clusters fulfill the following characteristics:

1. High level of homogeneity among members of the same cluster (within-cluster)
2. High level of heterogeneity between one cluster and another (between-cluster)

There are two main methods in cluster analysis, namely non-hierarchical methods and hierarchical methods (Dewi, I., Syaripuddin & Hayati, M. N., 2020).

### 2.3 Hierarchy Method

The grouping technique in the hierarchical method uses agglomerative techniques or divisive techniques. The hierarchical method in cluster analysis forms a multilevel or hierarchical structure like a tree because grouping is done in stages or levels. The results of grouping using the hierarchical method are usually presented as a dendrogram. A dendrogram is an advantage of the hierarchical method, as it visually represents the steps in cluster analysis. It shows how clusters are formed and the distance coefficient values at each step (Simamora, B., 2005).

### 2.4 Centroid Linkage Algorithm

Centroid Linkage is a cluster analysis method in which the cluster is represented by the average of all objects within it. The distance between two clusters is determined by calculating the distance between the centroids of each cluster. The cluster centroid is the middle value of the observations on the variables in each cluster. In this method, the centroid is recalculated every time a new cluster is formed until a stable cluster is formed (Sokal, R. R. & Michener, C. D., 1958). The advantage of this method is that outliers do not have a significant effect compared to other methods. Therefore, this method is suitable for data containing outliers (Karlita, T., 2006). Apart from that, this method also effectively reduces within-cluster variance in linkage centroids, which are based on the location of the central point formed in the previous stage (Xu, N., et.al., 2022). The distance between two clusters is defined in Equation (1).

$$d_{UV(W)} = d(\bar{x}_1, \bar{x}_2) \quad (1)$$

The new cluster centroid is formed using the formula in Equation (2).

$$\bar{x} = \frac{N_1 \bar{x}_1 + N_2 \bar{x}_2}{N_1 + N_2} \quad (2)$$

with,

$N_1 = N_2$  : Many objects

The centroid is the arithmetic mean of all individuals within the cluster. When objects are merged, a new centroid is computed such that the centroid will keep changing until a stable group is formed (Sokal, R. R. & Michener, C. D., 1958). The centroid linkage hierarchical clustering algorithm utilizes Euclidean distance (Xu, N., et.al., 2022) among all hierarchical clustering algorithms.

## 2.5 Similarity Measures

Multiple distance measurements can be employed, among which is the Euclidean distance. Euclidean distance is a highly efficient method for calculating similarity or dissimilarity between objects based on their attributes, particularly when dealing with numerical data. It is commonly used for grouping objects. In addition, Euclidean distance exhibits a strong data clustering structure (Aditya, A., Sari, B. N. & Padilah, T. N., 2021), surpasses other distance measurement techniques in effectiveness, and is particularly well-suited for small datasets (Mohibullah, M., Hossain, M. Z. & Hasan, M., 2015). The assumption that must be met when using the Euclidean distance is that all observed variables have no correlation and the variables have the same units (Mongi, C. E., 2015).

In this research, the distance measure used is Square Euclidean distance. The Square Euclidean distance between two observations can be calculated using Equation (3).

$$d_{ij} = \sum_{k=1}^p (x_{ik} - x_{jk}) ; i, j = 1, 2, 3, \dots, n \quad (3)$$

with,

$d_{ij}$  : Square Euclidean distance from observations  $i$  and  $j$   
 $x_{ik}$  : The value of the  $i$ -th observation on the  $k$ -th variable  
 $x_{jk}$  : The value of the  $j$ -th observation on the  $k$ -th variable

## 2.6 Silhouette Coefficient (SC)

The silhouette coefficient is a technique employed to assess the quality and robustness of clusters, and it quantifies the extent to which an object is appropriately assigned to a cluster (Shoolihah, A., Furqon, M. T. & Widodo, A. W., 2017). The silhouette coefficient method is a combination of two methods: the cohesion method, which quantifies the proximity between objects within a cluster, and the separation method, which quantifies the distance between a cluster and other clusters. The steps involved in computing the silhouette coefficient are as follows (Handoyo, R., Mangkudjaja, R. & Nasution, S. M., 2014):

1. Calculate the average observation with all other observations that are in the same cluster using Equation (4).

$$a(i) = \frac{1}{[A] - 1} \sum_{j \in A, j \neq i} d(i, j) \quad (4)$$

2. Calculate the average distance of the observation to all other observations in other clusters, then take the minimum value using Equation (5).

$$d(i, C) = \frac{1}{[A]} \sum_{j \in C} d(i, j) \quad (5)$$

3. Calculate the silhouette coefficient value using Equation (6).

$$ds(i) = \frac{b(i) - a(i)}{\max \{a(i), b(i)\}} \quad (6)$$

The silhouette coefficient yields values ranging from -1 to 1. A higher value approaching 1 indicates a more optimal grouping of data within a single cluster. On the other hand, if the value is close to -1, it is considered detrimental to group the data into one cluster.

### 3. Research Methods

#### 3.1 Data Sources and Research Variables

The Central Statistics Agency (CSA) is an institution tasked with providing official statistical data in Indonesia. This research's data source comes from the Official Publication of the East Nusa Tenggara (NTT) Provincial Central Statistics Agency for 2022. The unit of observation in this research is the Regency/City. The variables used in this research are:

- $X_1$  : Human Development Index (HDI)
- $X_2$  : Percentage of Poor Population (%)
- $X_3$  : Open Unemployment Rate (%)
- $X_4$  : Average Years of Schooling (Year)
- $X_5$  : Percentage of population aged 15 years and over who are studying at university (%)
- $X_6$  : Percentage of population aged 15 years and over who are illiterate (%)
- $X_7$  : Economic Growth Rate (%)
- $X_8$  : Life Expectancy (Year)

The eight variables used are represented as poverty indicators that can be used for profiling.

#### 3.2 Data Analysis

The steps taken in data analysis are presented as follows:

1. Carry out descriptive statistical analysis to see a general overview of the research data.
2. Multicollinearity checking with Pearson correlation.
3. Grouping data using the centroid linkage algorithm with the following stages:
  - a. Standardize data with  $Z_{score}$ .
  - b. Calculate the Square Euclidean distance for all observations using Equation (3).
  - c. Combining observations that have the smallest distance, then calculating the average between the combined observations using Equation (2), thus forming a new data set.
  - d. Repeat steps (a) and (b) until the number of clusters formed is only 1 cluster.
4. Calculate the silhouette coefficient using Equations (4), (5), and (6) to find out the most optimal grouping. The clusters tested were 2, 3, 4 and 5 clusters.
5. Profiling and interpretation of cluster results.

### 4. Results and Discussion

In the sub-chapter, the results and discussion will be presented regarding descriptive statistics, multicollinearity checks, grouping using the Centroid Linkage algorithm, and the final results in profiling districts/cities in East Nusa Tenggara Province.

#### 4.1. Descriptive Statistics

Descriptive statistics is an analytical technique used to summarize, present, and describe data in a form that is easy to read and understand. The goal is to provide helpful information about the data and assist in decision-making. In this article, the descriptive statistics used are the Minimum, Maximum, and Average values, which are presented in Table 2.

Table 2. Descriptive Statistics

Variables	Minimum	Maximum	Mean
$x_1$	57.030	79.740	64.263
$x_2$	9.170	34.270	21.748
$x_3$	0.970	9.760	3.433
$x_4$	6.350	11.600	7.627

Variables	Minimum	Maximum	Mean
$x_5$	4.100	15.160	7.668
$x_6$	0.800	13.350	5.638
$x_7$	0.550	2.570	1.705
$x_8$	60.660	69.730	66.271

## 4.2. Multicollinearity Check

Multicollinearity is a condition with a robust linear relationship between two or more variables. In cluster analysis, multicollinearity can be a problem because it can cause distortion in the grouping results and make interpretation difficult. Identifying multicollinearity problems can use heatmap correlation visualization.

Pearson correlation results visualized through heatmap correlation visualization can be seen in Figure 1.

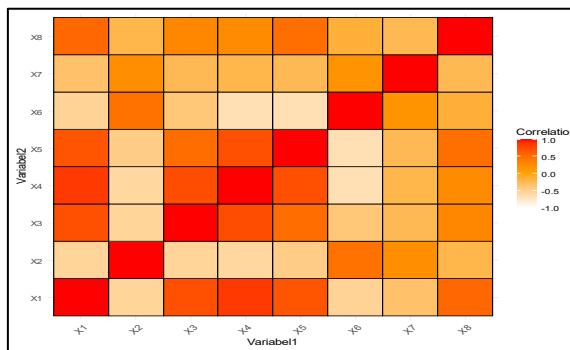


Figure 1. Heatmap correlation visualization

Based on Figure 1., the heatmap correlation does not show red cells; it can be concluded that there is no strong linear relationship between the research variables. This indicates that there is no significant multicollinearity between these variables.

## 4.3. Centroid Linkage Algorithm

Centroid linkage is a technique employed in hierarchical clustering. Hierarchical clustering is a method that seeks to construct a hierarchical structure of clusters. Centroid linkage is a method that determines whether two clusters should be merged based on the distance between their centroids, which are the central points of each cluster. The grouping results visualized through a dendrogram using the centroid linkage algorithm can be seen in Figure 2.

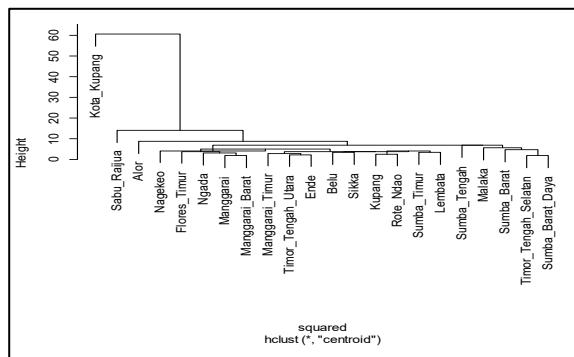


Figure 2. Dendrogram of Grouping Results

Based on Figure 2, it can be seen that Kupang City is spread far from other regencies/cities. The geographical separation of Kupang City from different towns/regencies can be influenced by various factors, including its status as the dominant government or economic center, better infrastructure, quality education and health facilities, and cultural and entertainment attractions. Geographic conditions and government policies can also be necessary in determining population distribution.

#### 4.4. Silhouette Coefficient Value

The method for measuring how optimal the groupings are is using the silhouette coefficient and the clusters tried, namely 2, 3, 4 and 5 clusters which can be seen in Table 3.

Table 3. Silhouette Test

Number of Clusters	Silhouette Score
2	<b>0.785</b>
3	0.363
4	0.139
5	0.264

Table 3. reveals that the maximum silhouette coefficient value is 0.785, indicating the presence of 2 clusters. Based on the variables analyzed using the centroid linkage algorithm, it is determined that there are 2 clusters formed when grouping districts/cities in East Nusa Tenggara Province.

#### 4.5. Profiling Grouping Results

To illustrate how the 22 districts/cities in East Nusa Tenggara coalesce into the optimal number of clusters determined earlier, we plotted the hierarchical tree generated by the centroid-linkage algorithm. The resulting dendrogram in Figure 3 arranges the units along the horizontal axis, while the vertical axis indicates the rescaled distance (or “height”) at which successive agglomerations occur. By reading the branch cut at the chosen height, one can observe which districts merge first, which remain isolated the longest, and how the final cluster structure is formed.

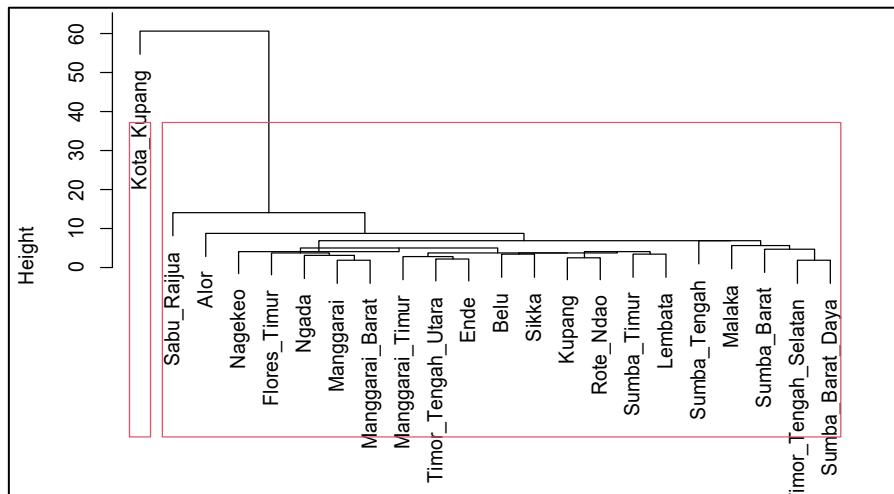


Figure 3. Dendrogram of Optimal Grouping Results

Furthermore, the grouping results obtained based on optimal clusters can be seen through the cluster plot in Figure 4.

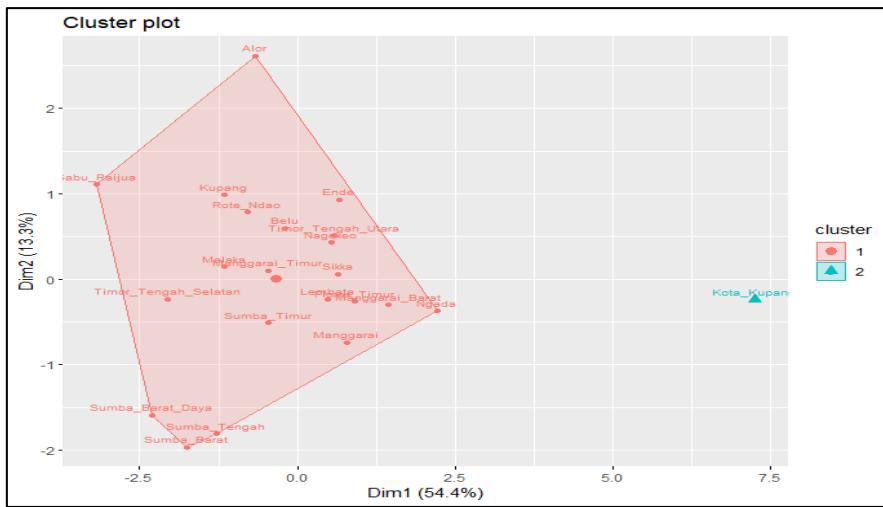


Figure 4. Cluster Plot Optimal Grouping Results

Complete details of each Regency/City in Figures 3. and 4. are detailed in Table 4.

Table 4. Optimal Cluster Grouping

Clusters	Number of Members	Regency/City
Cluster 1	1	Kota Kupang
Cluster 2	21	Flores Timur, Ngada, Manggarai, Manggarai Barat, Nagekeo, Timor Tengah Utara, Ende, Alor, Manggarai Timur, Sumba Timur, Lembata, Rote Ndao, Kupang, Malaka, Belu, Sikka, Sumba Tengah, Sumba Barat, Sabu Raijua, Timor Tengah Selatan, Sumba Barat Daya

The analysis of the two clusters reveals distinct characteristics for each cluster, as determined by the variables examined. Table 5. displays the attributes of these clusters, using mean values.

Table 5. Characteristics of Cluster Results

Variables	Mean	
	Cluster 1	Cluster 2
$x_1$	<b>79.740</b>	63.526
$x_2$	9.170	<b>22.347</b>
$x_3$	<b>9.760</b>	3.132
$x_4$	<b>11.600</b>	7.438
$x_5$	<b>15.160</b>	7.311
$x_6$	0.800	<b>5.868</b>
$x_7$	1.100	<b>1.734</b>
$x_8$	<b>69.730</b>	66.107

Table 5. reveals that cluster 1 represents a city characterized by a high Human Development Index (HDI), low open unemployment rate, significant average years of schooling, a notable percentage of the population aged 15 years and above pursuing higher education, and a long-life expectancy. On the other hand, cluster 2

represents a Regency/City cluster with a high percentage of poor individuals, a significant percentage of people aged 15 years and above who are illiterate, and a high rate of economic growth. Writing the results and discussion can be separated into different sub-sections or can also be combined into one sub. The summary of the results can be presented in the form of graphs and images. The results and discussion sections must be free from multiple interpretations. The discussion must answer the research problem, support and defend the answer with the results, compare with relevant research results, state the limitations of the research, and find novelties.

## 5. Conclusion

The analysis of the grouping of districts/cities in East Nusa Tenggara Province based on poverty indicators yielded the following conclusions:

1. Silhouette-coefficient diagnostics confirm that the 22 districts/cities of East Nusa Tenggara are best separated into two clusters when centroid-linkage agglomeration is applied to the eight poverty-related indicators.
2. The districts/cities in East Nusa Tenggara Province were categorized into 2 clusters based on poverty indicators.
  - Cluster 1 (High-HDI)  
Represented solely by Kota Kupang, this cluster is distinguished by the province's highest Human Development Index ( $\approx 79.7$ ), very low open-unemployment ( $\approx 0.8\%$ ), long life expectancy ( $\approx 69.7$  years), and the best educational metrics—average years of schooling  $\approx 11.6$  and tertiary-enrolment share  $\approx 15.2\%$ .
  - Cluster 2 (Lagging Regencies)  
Comprising the remaining 21 districts, this group shows markedly lower HDI ( $\approx 63.5$ ), shorter schooling ( $\approx 7.4$  years), higher illiteracy among adults, and more than double the unemployment rate ( $\approx 5.9\%$ ). Despite a relatively high economic-growth rate, the average poverty headcount remains elevated ( $\approx 22.3\%$ ).

## 5. Declaration of AI and AI assisted technologies in the writing process

The authors used several AI-assisted tools to improve the quality and accuracy of the manuscript. Specifically, Google Translate was used to support translation between Bahasa Indonesia and English; QuillBot and Grammarly were used for paraphrasing, language refinement, and grammatical improvement; and Mendeley Reference Manager was utilized for citation organization and reference formatting. All outputs generated or suggested by these tools were thoroughly reviewed, revised, and verified by the authors. The authors take full responsibility for the content of this manuscript.

## 6. CRediT Authorship Contribution Statement

Author contributions (CRediT): Andrea Tri Rian Dani, Meirinda Fauziyah—Conceptualization; Methodology; Data curation; Formal analysis; Software; Visualization; Project administration; Validation; Writing – original draft; Writing – review & editing; Supervision. Yossy Candra—Conceptualization; Validation; Writing – review & editing. Fachrian Bimantoro Putra—Data curation; Investigation; Resources; Software; Formal analysis; Project administration.

## 7. Declaration of Competing Interest

The authors declare no competing financial or personal interests.

## 8. Acknowledgments

This research was funded in part by DIPA BLU-PNBP FMIPA, Mulawarman University, Samarinda, Indonesia [No: 1700 / UN17.7 / LT/ 2023].

## 9. Data Availability

Primary data come from official BPS publications and are publicly accessible; derived data and analysis scripts are available from the corresponding author upon reasonable request.

## 10. Funding

This work was supported by Mulawarman University

## 11. Ethical Approval

Not applicable. The study did not involve human or animal subjects and did not access personal data.

## 12. References

Aditya, A., Sari, B. N., & Padilah, T. N. (2021). Perbandingan pengukuran jarak Euclidean dan Gower pada klaster k-medoids. *Jurnal Teknologi dan Sistem Komputer*, 9(1), 1–7.  
Doi: <https://doi.org/10.14710/jtsiskom.2021.13747>

Dewi, I., Syaripuddin, & Hayati, M. N. (2020). Pengelompokan Data Kategorik Dengan Algoritma Robust Clustering Using Links (Studi Kasus: PT. Prudential Life Jalan MT. Haryono Samarinda). *Jurnal EKSPONENSIAL*, 11(2), 133–138.  
<https://jurnal.fmipa.unmul.ac.id/index.php/exponensial/article/view/655>

Dzikrullah, A. A. (2022). Pengelompokan Provinsi Berdasarkan Kualitas Jaringan Internet Dengan Metode Centroid Linkage. *Journal of Mathematics, Computations, and Statistics*, 5(1), 48–57.  
<https://doi.org/10.35580/jmathcos.v5i1.32363>

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014) *Multivariate Data Analysis 7th Edition*, Person Education, Inc, USA.

Han, J., Kamber, M., & Pei, J. (2012) *Data Mining Concepts and Techniques Third Edition*, Elsevier, USA.

Handoyo, R., Mangkudjaja, R. & Nasution, S. M. (2014). Perbandingan Metode Clustering Menggunakan Metode Single Linkage dan K-Means pada Pengelompokan Dokumen. *Jurnal SIFO Mikrosil*, 15(2), 73–82, 2014.  
<https://doi.org/10.55601/jsm.v15i2.161>

Hikmah, Fardinah, Qadrini, L., & Tande, E. (2022). Analisis Klaster Pengelompokan Kecamatan di Sulawesi Barat Berdasarkan Indikator Pendidikan. *Jurnal Matematika, Sains, dan Pembelajarannya*, 8(2), 188–196. Doi: <https://doi.org/10.31605/saintifik.v8i2.383>

Karlita, T. (2006) *Algoritma Perbaikan Penentuan Titik Pusat Awal Berbasis Hirarki untuk Klasterisasi Data Kategorikal*, Institut Teknologi Sepuluh Nopember, Surabaya.

Mohibullah, M., Hossain, M. Z., & Hasan, M. (2015). Comparison of Euclidean Distance Function and Manhattan Distance Function Using K-Medoids. *International Journal of Computer Science and Information Security*, 13(10), 61–71.  
[https://www.researchgate.net/publication/291970199\\_Comparison\\_of\\_Euclidean\\_Distance\\_Function\\_and\\_Manhattan\\_Distance\\_Function\\_Using\\_K-Medoids](https://www.researchgate.net/publication/291970199_Comparison_of_Euclidean_Distance_Function_and_Manhattan_Distance_Function_Using_K-Medoids)

Mongi, C. E. (2015). Penggunaan Analisis Two Step Clustering untuk Data Campuran. *d'Cartesian*, 4(1), 9–19.  
<https://doi.org/10.35799/dc.4.1.2015.7251>

Nanang, A. & Susanti, D. (2018). Penerapan Data Mining Menggunakan Algoritma K-Means untuk Mengelompokkan Konsumen pada CV Mizan Cirebon. Prosiding Seminar Nasional Energi & Teknologi (SINERGI), 3, 93–102.  
<https://jurnal.unismabekasi.ac.id/index.php/sinergi/article/view/1148/1016>

Prasetyo, E. (2012) *Data mining : konsep dan aplikasi menggunakan MATLAB*, CV Andi Offset, Yogyakarta.

Raja, N. A., Tinungki, G. M., & Sirajang, N. (2024). Implementasi Algoritma Centroid Linkage dan K-Medoids dalam Mengelompokkan Kabupaten/Kota di Sulawesi Selatan Berdasarkan Indikator Pendidikan. *ESTIMASI: Journal of Statistics and Its Application*, 5(1), 61–74.  
<https://doi.org/10.20956/ejsa.v5i1.13605>

Santosa, B. (2007) *Data Mining, Teknik Pemanfaatan Data untuk Keperluan Bisnis : Teori dan Aplikasi*, Graha Ilmu, Yogyakarta.

Shoolihah, A., Furqon, M. T., & Widodo, A. W. (2017). Implementasi Metode Improved K-Means untuk Mengelompokkan Titik Panas Bumi. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 1(11), 1270–1276.  
<https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/453>

Silvi, R. (2018). Analisis Cluster dengan Data Outlier Menggunakan Centroid Linkage dan K-Means Clustering untuk Pengelompokan Indikator HIV/AIDS di Indonesia. *Jurnal Matematika MANTIK*, 4(1), 22–31.  
<https://doi.org/10.15642/mantik.2018.4.1.22-31>

Simamora, B. (2005) *Analisis Multivariat Pemasaran*, Gramedia Pustaka Utama, Jakarta.

Sokal, R. R. & Michener, C. D. (1958). A Statistical Method for Evaluating Systematic Relationship. *University of Kansas Scientific Bulletin*, 28, 1409–1438.  
[https://ia800703.us.archive.org/5/items/cbarchive\\_33927\\_a statistical method for evaluation in 1902/a statistical method for evaluation in 1902.pdf](https://ia800703.us.archive.org/5/items/cbarchive_33927_a statistical method for evaluation in 1902/a statistical method for evaluation in 1902.pdf)

Tan, P. N., Steinbach, M., & Kumar, V. (2006) *Introduction to Data Mining*, Person Education, Inc, USA.

Widodo, E., Sari, N. N., Hidayati, I., Yubinas, F., Yuniarti, M., & Novyantika, R. D. (2018). Analisis Cluster Penderita Disabilitas Mental di Provinsi Daerah Istimewa Yogyakarta Tahun 2016. Prosiding Konferensi Nasional Penelitian Matematika dan Pembelajarannya (KNPMP) III, 577–586.  
<http://hdl.handle.net/11617/10140>

Xu, N., Xu, C., Finkelman, R. B., Engle, M. A., Li, Q., Peng, M., He, L., Huang, B., & Yang Y. (2022). Coal elemental (compositional) data analysis with hierarchical clustering algorithms. *International Journal of Coal Geology*, 249.  
<https://doi.org/10.1016/j.coal.2021.103892>