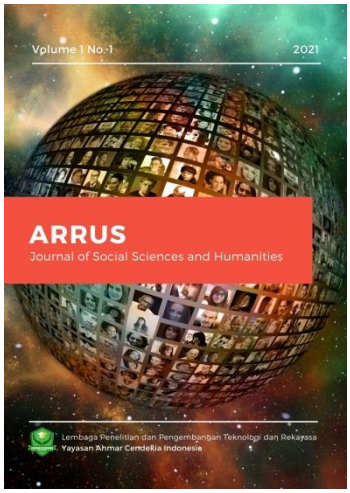
https://doi.org/10.35877/soshum2587

OPEN ACCESS

ISSN: 2776-7930 (Print) / 2807-3010 (Online)



RESEARCH ARTICLE

K Means Cluster for Grouping Regencies/Cities in South Sulawesi Province Based Human Development Index on the 2023

Isma Muthahharah[1,*], Sitti Masyitah Meliyana[1], Ansari Saleh Ahmar[1], & Abdul Rahman[2]

[1]Department of Statistics, Universitas Negeri Makassar, Makasssar, 90223, Indonesia
[2]Department of Mathematics, Universitas Negeri Makassar, Makasssar, 90223, Indonesia

*Corresponding author: Isma Muthahharah, Departement of Statistics, Faculty of Mathematics and Natural Sciences,Universitas Negeri Makassar, Makasssar, South Sulawesi, Indonesia

E-mail: isma.muthahharah@unm.ac.id

Abstract: The grouping of districts/cities in South Sulawesi province should be carried out as a material for planning and evaluating government program objectives. The aim is to increase human development figures based on the indicators that form the HDI including Life Expectancy (UHH) at birth, Expected Years of Schooling (HLS), Average Years of Schooling (RLS). Therefore, cluster analysis was used using the K-means method, which is a type of statistical analysis. This type of research is applied research with a quantitative approach by collecting and analyzing data using the K-Means Cluster method. The data used in the determination was sourced from the South Sulawesi Provincial Central Statistics Agency in 2023. The results of grouping using K-Means clusters showed that there were 3 clusters, where cluster 1 had 23 regencies/cities in the Southern Province consisting of the Selayar Islands, Bantaeng, Jeneponto, Takalar, Sinjai, Maros, Pangkep, Barru, Bone, Soppeng, Wajo, Luwu and North Luwu with low criteria. Cluster 2 has 3 regencies/cities, namely Makassar, Pare Pare and Palopo with high criteria. Cluster 3 has 8 regencies/cities, namely Bulukumba, Gowa, Sidrap, Pinrang, Enrekang, Tana Toraja, East Luwu and North Toraja with medium criteria. The suggestion in this research is that you can use indices and other methods to group districts/cities in South Sulawesi Province.

Keywords: HDI, cluster, K-Means, Regency/City, South Sulawesi

1. Introduction

National development is the main goal of a country's development. In Indonesia, people's welfare is one of the national goals as stated in the fourth paragraph of the Preamble to the 1945 Constitution. The goal of national development is to make people's lives smarter and promote general welfare. Human happiness is a situation that is dynamic and never ends in quantity because it is eternal. This changes along with the development of human life needs. Welfare Certain aspects, namely population, health and nutrition, education, employment, consumption levels and patterns, housing and the environment, poverty, and other social criteria in an effort to improve the quality of life (Musa & Fallo, 2023). Indonesia has experienced an increase from year to year. In Indonesia, HDI is categorized into 4 categories, namely low, medium, high, and very high with cut points at 60, 70, and 80. The cut point for low HDI determined by the UNDP is slightly different, namely lower than 55.

https://doi.org/10.35877/soshum2587

ISSN: 2776-7930 (Print) / 2807-3010 (Online)

Regency or city HDI in Indonesia is medium HDI. Lampung Province, Central Sulawesi Province, and Maluku changed from medium HDI to high HDI with HDI growth of 0.79%, 0.70%, and 0.73% in 2022. If the categorization of human development in an area only uses a composite index, so consideration of the cut-off point for each category becomes very important (Fahmiyah & Ningrum, 2023) present the material simply and concisely.

The grouping of districts/cities in South Sulawesi province should be carried out as a material for planning and evaluating government program objectives. The aim is to increase human development figures based on the indicators that make up the HDI. So it is necessary to group regencies/cities in South Sulawesi Province based on similar characteristics in terms of the indicators that make up the HDI which include Life Expectancy at Birth, Expected Years of Schooling, Average Years of Schooling. Therefore, cluster analysis was used using the K-means method, which is a type of statistical analysis. The analysis was carried out by grouping districts/cities in Maluku Province based on HDI indicators. The existing indicators are represented as variables, and the district/city HDI in Maluku is represented as objects grouped based on similar characteristics. All other objects are in the same cluster (Talakua et al., 2017). K-Means is a Non-hierarchical data grouping (subdivision) method that allows data to be divided into two or more people. This method will divide the data into several groups of places, data that has the same characteristics is grouped into one data and those that have different but the same characteristics are divided into another group. The goal of grouping is to minimize the objective function determined by the grouping process and seeks to minimize and maximize the variation between groups (Gustientiedina et al., 2019). Based on research (Nnk et al., 2023) states that the formation of 2 groups (clusters) through K-Means Cluster Analysis. Cluster 1 has the characteristics of provinces with high to very high scores on UHH, HLS, RLS, and adjusted expenditure. Meanwhile, Cluster 2 consists of provinces with medium to high scores on UHH, HLS, RLS and adjusted expenditure. Meanwhile in (Sianipar & Gunawan, 2021) cluster 1 has 18 data, cluster 2 has 11 data, cluster 3 has 4 data. The aim of this research is to find out districts/cities that have a Human Development Index that is classified as high, medium and low.

From the description above, this research discusses the grouping of districts/cities based on the Human Development Index which is seen from 3 categories, namely life expectancy, expected length of schooling and average length of schooling using the K-Means Clustering method.

2. Literature Review

2.1. Cluster Analysis

Cluster analysis is a multivariate technique that has the main objective of classifying objects based on their characteristics. Thus, the purpose of grouping is so that the objects included in a group are objects that are similar (or related) to each other and different (not related) to objects in other groups. Thus, the similarity between groups between individual classes (intraclass) should be small and high between different groups (interclass), similarity is considered as a measure of distance (Muthahharah & Juhari, 2021).

The advantages of cluster analysis are (1) it can group large observation data and a relatively large number of variables. (2) can be used on ordinal, interval and ratio data scales. However, the weaknesses of this cluster analysis are (1) the grouping is subjective for the researcher, because only dendogram images are observed. (2) In the case of heterogeneous data between one research object and another, researchers find it difficult to determine the number of groups to be formed. (3) there are significant differences in the methods used, so calculations usually compare each method. (4) the larger the observation, usually the greater the error rate.

https://doi.org/10.35877/soshum2587

ISSN: 2776-7930 (Print) / 2807-3010 (Online)

The basic process of clustering is grouping data which is usually done in two ways, namely (Andarini et al., 2023):

- Hierarchical method, where this method groups two or more objects that have the closest similarity. The process is then continued to other nearby objects. And so on, so that the cluster forms a kind of tree in which there is a hierarchy (clear levels) between objects, a dendogram is usually used to explain the hierarchical process.
- Non-hierarchical methods start by first determining the desired number of clusters. The selected cluster center is a temporary cluster center which is updated every iteration until it meets the criteria, so that objects can move from one cluster to another. A well-known non-hierarchical method is K-Means Cluster.

2.2. K Means

K Means is a non-hierarchical data grouping method that tries to divide existing data into one or more clusters/groups. This method divides data into clusters/groups so that data with the same characteristics is grouped into the same cluster (Hutabarat, 2021). K-Means is a clustering algorithm with an iterative process. The letter K is defined as the number of groups to be created. Next, the K value is determined randomly. Meanwhile, resources have a temporary value equal to the cluster center or also called centroid. The distance to each centroid is calculated for each existing datum using Euclid's formula to obtain the closest distance to the centroid for each datum (Zaki et al., 2022).

In the initial data collection process using the K-Means algorithm, the initial centroid point cj is formed. Usually, the source center formation is generated randomly. The number of centroids created by cj corresponds to the number of clusters determined at the beginning. After k centroids are formed, the distance between xi and the jth centroid of each datum is calculated by k, denoted by d (xi, cj). Several distance measures are used to measure the similarity between data examples, one of which is Euclidean distance. Calculating the Euclidean distance in Equation (Asroni & Adrian, 2016)

$$d(X_i, C_j) = \sqrt{\sum_{i=1}^{N} (X_i - C_j)^2}$$

If the value is d(Xi,Cj), then the similarity between the two observation units is closer. The condition for using Euclidean distance is that all features in the dataset are uncorrelated. If it has a correlation function, then it uses the Mahalanobis distance concept.

2.3. Human Development Index

Development Index (HDI) is a combined indicator that cannot measure all aspects of human development, but measures three main dimensions of humanity which are considered to reflect the basic skills (abilities) of a population. Longevity and health, knowledge and skills, and access to resources necessary to maintain an acceptable standard of living are three basic skills (Sibarani et al., 2022).

The United Nation Development Program (UNDP) defines human development as a process where the population's choices are expanded in the sense that people are given the freedom to choose more ways to meet their living needs. The Human Development Index (HDI) is a measure that can describe the progress of human development in a measurable and representative manner. HDI explains how residents can access development outcomes in terms of income, health, education and more. Basically, HDI is expected to be able to represent performance over time (Anggraeni & Arum R, 2022).

The indicators chosen to measure the HDI dimensions are as follows (Yektiningsih, 2018):

- Longevity, measured by the variables life expectancy at birth and infant mortality rate per thousand population.

https://doi.org/10.35877/soshum2587

ISSN: 2776-7930 (Print) / 2807-3010 (Online)

- Education is measured by two indicators, namely the literacy ability of the population aged 15 years and over (adult literacy) and the average school attendance of the population aged 25 years and over (average years of schooling).
- Resource availability can be measured at the macro level with real GDP per capita, using purchasing power parity terminology in US dollars, and can be added to labor.

The Human Development Index (HDI) of South Sulawesi Province has progressed in a decade. South Sulawesi's HDI experienced growth from 67.26 in 2012 to 78.82 in 2022. In this period, South Sulawesi's HDI experienced growth at an average rate per year. amounted to 0.81 percent and increased from "medium" to "high" since 2017. Post-Covis-19 South Sulawesi, HDI will grow by 0.80 percent in 2022 compared to last year, faster than HDI growth of 0, 43 percent in 2021. The achievement of HDI in 2022, which increased by 0.58 points to 2022, is supported by the growth of all its components (Rusdi, 2023).

3. Research Method and Materials

3.1. Types of Research

This type of research is applied research with a quantitative approach by collecting and analyzing data using the K-Means Cluster method.

3.2. Data Source

The data used in the research comes from the South Sulawesi Provincial Central Statistics Agency for 2023.

3.3. Research Variable

The variables used in this research are as follows:

x_1 = Life Expectancy (Years)

x_2 = Expected length of school (Years)

x_3 = Average length of school (Years)

3.4. Research Stages

The stages in the grouping process using the K-Means cluster method are as follows:

Inspire data used in R applications.

- Create new variables and calculate descriptive statistics.
- Calculate the distance from each data to other data.
- Determine the number of clusters based on the elbow method with the formula:

$$SSE = \sum_{k=1}^{k} \sum_{xi\varepsilon} \|X_i - C_k\|^2$$

- Where k is number of groups used in the K-means algorithm, X_i is amount of data, C_k is number of clusters in the kth cluster.
- Create clustering of data based on the number of characters that have been determined
- Create a data frame for cluster results and determine districts/cities based on clusters.

4. Results and Discussion

4.1. Descriptive Analysis

Descriptive analysis aims to determine the minimum value. maximum, average. Middle values and frequently occurring values (median).

https://doi.org/10.35877/soshum2587

ISSN: 2776-7930 (Print) / 2807-3010 (Online)

Table 1. Statistic Descriptive

Correlation	Minimum	Maximum	Mean	Median
x_1	72.57	74.32	73.75	73.69
x_2	12.12	13.64	13.39	13.29
x_3	7.00	8.63	8.52	8.31

Based on Table 1, it shows that Life Expectancy (x_1) has a minimum data of 72.57, a maximum of 74.32, a mean of 73.75 and a median of 73.69. Meanwhile, Life Expectancy (x_2) has a minimum data of 12.12, a maximum of 13.64, a mean of 13.39 and a median of 13.29. Meanwhile, the average length of schooling (x_3) has a minimum data of 7.00, a maximum of 13.64, a mean of 13.39 and a median of 8.31. Visualization of Regencies/Cities in South Sulawesi province.

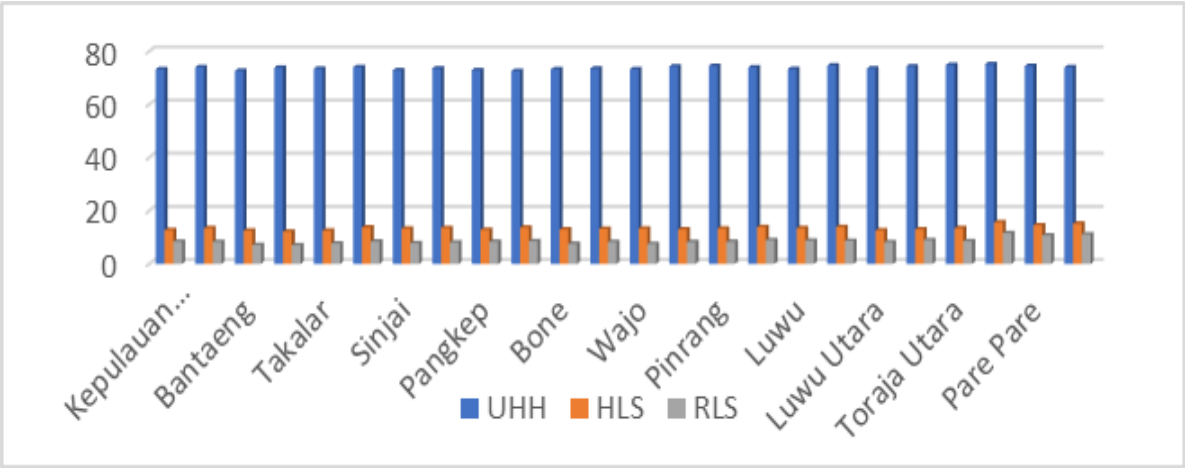


Figure 1. Diagram human Development Index

Based on Figure 1, it can be seen that the Life Expectancy Rate has the largest figure for each district/city in the province of South Sulawesi. Meanwhile, the expected length of schooling has a moderate figure for all districts/cities in South Sulawesi Province and the average length of schooling has the lowest figure for all districts/cities in South Sulawesi Province.

4.2. Research Stages

Determining the number of clusters based on the Elbow method can be seen in Figure 2.

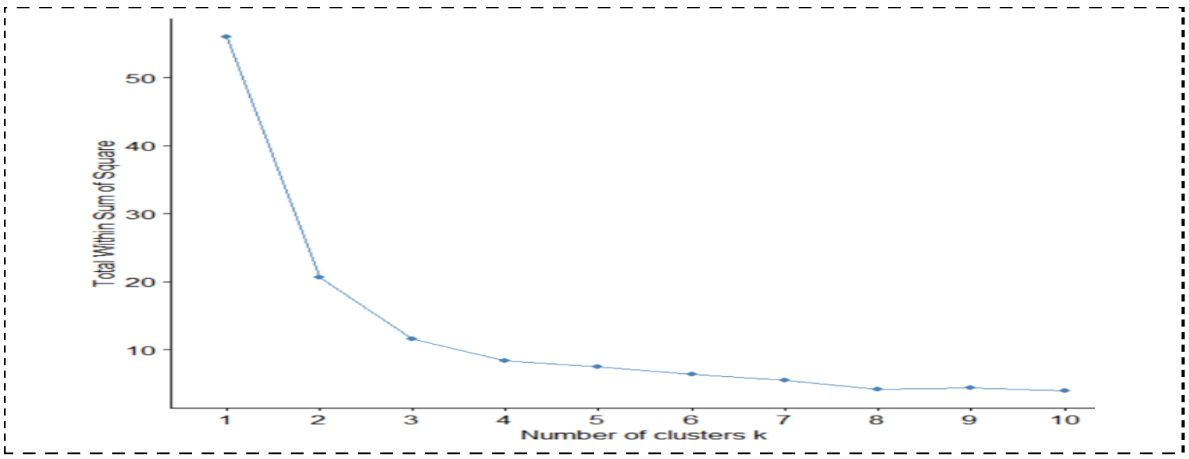


Figure 2. Diagram human Development Index

https://doi.org/10.35877/soshum2587

ISSN: 2776-7930 (Print) / 2807-3010 (Online)

Based on Figure 2, the Elbow method looks at the movement of the sloping graph after the steep graph which is at k = 3. Based on this method, supported by theory from (Aditya et al., 2020) the optimal number of clusters formed is 3. From the number of clusters formed, a vector cluster:

[1] 1 3 1 1 1 3 1 1 1 1 1 1 1 3 3 3 1 3 1 3 3 2 2 2

From the cluster vector, SSE is produced, namely [1] 7.784323 1.617000 2.189837 with a total SSE of 79%. From the optimal number of clusters, they will be clustered as in Figure 3.

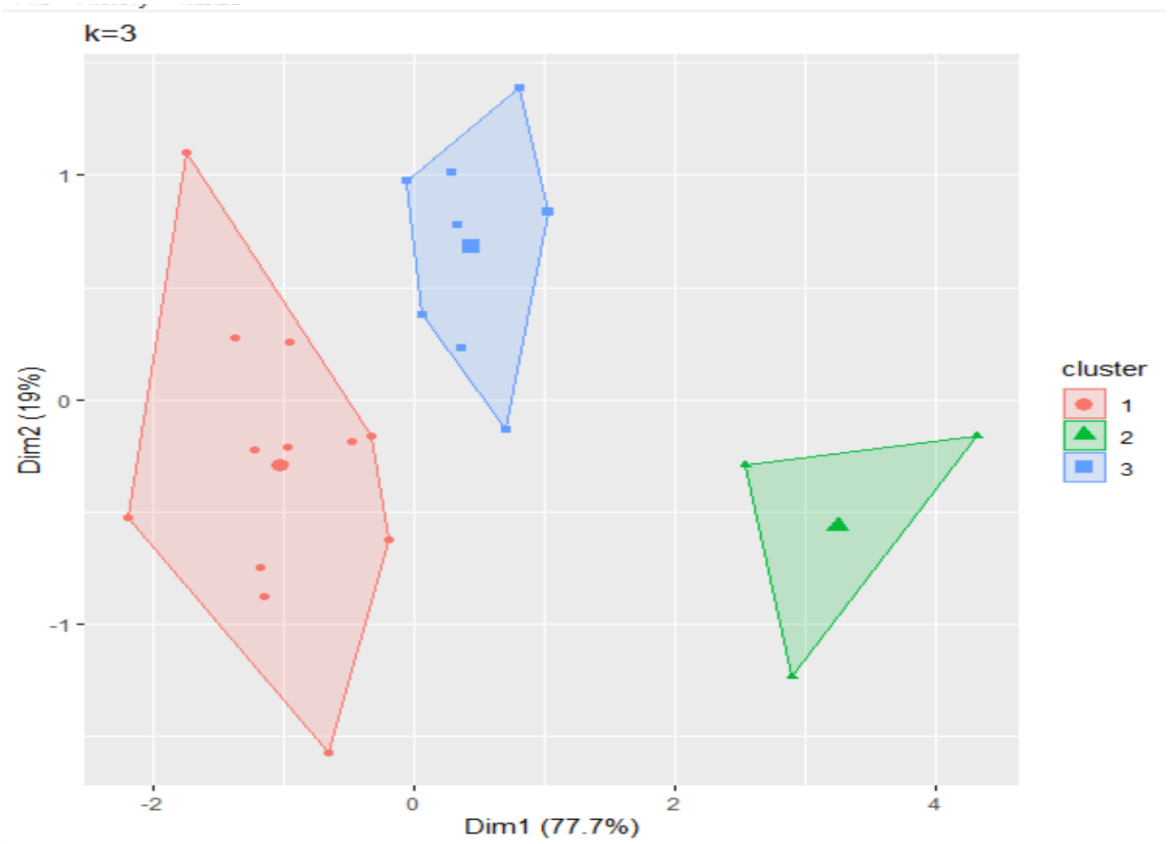


Figure 3. District/City Clustering

Based on Figure 3, it can be analyzed by visualizing the distribution of data in each Cluster which is differentiated based on color, in the method Cluster 1 is depicted in red, Cluster 2 is depicted in green, and Cluster 3 is depicted in blue. This graph can also be analyzed for the proximity of one data to other data collected in a cluster. From the cluster visualization, the clustering results are displayed in Table 2.

Table 2. Regency/City Grouping Results

Cluster	Member	Amount	Mean	Criteria
1	Kepulauan Selayar, Bantaeng, Bantaeng, Jeneponto, Takalar, Sinjai, Maros, Pangkep Barru, Bone, Soppeng, Wajo, Luwu dan Luwu Utara	13	31.36	Low
2	Makassar, Pare-Pare, Palopo	3	33.58	High
3	Bulukumba, Gowa, Sidrap, Pinrang, Enrekang, Tana Toraja, Luwu Timur, Toraja Utara	8	32.10	Middle

https://doi.org/10.35877/soshum2587

ISSN: 2776-7930 (Print) / 2807-3010 (Online)

5. Conclusion

The results of grouping using K-means clusters show that there are 3 clusters, where cluster 1 has 23 regencies/cities in the Southern Province consisting of Selayar Islands, Bantaeng, Jeneponto, Takalar, Sinjai, Maros, Pangkep, Barru, Bone, Soppeng, Wajo, Luwu and North Luwu with low criteria. Cluster 2 has 3 regencies/cities, namely Makassar, Pare Pare and Palopo with high criteria. Cluster 3 has 8 regencies/cities, namely Bulukumba, Gowa, Sidrap, Pinrang, Enrekang, Tana Toraja, East Luwu and North Toraja with medium criteria. The suggestion in this research is that you can use indices and other methods to group districts/cities in South Sulawesi Province.

References

Aditya, A., Jovian, I., & Sari, B. N. (2020). Implementasi K-Means Clustering Ujian Nasional Sekolah Menengah Pertama di Indonesia Tahun 2018/2019. *Jurnal Media Informatika Budidarma*, *4*(1), 51. https://doi.org/10.30865/mib.v4i1.1784

Andarini, R., Imanni, H., Sulistianingsih, E., & Perdana, H. (2023). Analisis Cluster Menggunakan Algoritma K-Means Berdasarkan Faktor Penyebab Stunting Pada Provinsi Kalimantan Barat. *Bimaster*, *12*(3), 301–308.

Anggraeni, L., & Arum R, P. (2022). Analisis Cluster Menggunakan Algoritma K-Means Pada Provinsi Sumatera Barat Berdasarkan Indeks Pembangunan Manusia Tahun 2021. *Prosiding Seminar Nasional UNIMUS*, *5(1)*, 636–646.

Asroni, A., & Adrian, R. (2016). Penerapan Metode K-Means Untuk Clustering Mahasiswa Berdasarkan Nilai Akademik Dengan Weka Interface Studi Kasus Pada Jurusan Teknik Informatika UMM Magelang. *Semesta Teknika*, *18*(1), 76–82. https://doi.org/10.18196/st.v18i1.708

Fahmiyah, I., & Ningrum, R. A. (2023). Human Development Clustering in Indonesia: Using K-Means Method and Based on Human Development Index Categories. *Journal of Advanced Technology and Multidiscipline*, *2*(1), 27–33. https://doi.org/10.20473/jatm.v2i1.45070

Gustientiedina, G., Adiya, M. H., & Desnelita, Y. (2019). Penerapan Algoritma K-Means Untuk Clustering Data Obat-Obatan. *Jurnal Nasional Teknologi Dan Sistem Informasi*, *5*(1), 17–24. https://doi.org/10.25077/teknosi.v5i1.2019.17-24

Hutabarat, L. Y. (2021). Penerapan Algoritma K-Means dalam Pengelompokan Jumlah Penduduk Berdasarkan Kelurahan di Kota Pematangsiantar. *Jurnal Ilmu Komputer Dan Teknologi*, *2*(2), 20–26.

Musa, M., & Fallo, S. I. (2023). Hierarchical Cluster Analysis on People'S Welfare in Southeast Sulawesi Province. *BAREKENG: Jurnal Ilmu Matematika Dan Terapan*, *17*(2), 1163–1172. https://doi.org/10.30598/barekengvol17iss2pp1163-1172

Muthahharah, I., & Juhari, A. (2021). A Cluster Analysis with Complete Linkage and Ward's Method for Health Service Data in Makassar City. *Jurnal Varian*, *4*(2), 109–116. https://doi.org/10.30812/varian.v4i2.883

Nnk, K., Khoirunnisaa, N., Viewianti ENF, G., & Yunizar Pratama Yusuf, A. (2023). Implementasi Algoritma K-Means Clustering Menggunakan Aplikasi Orange Untuk Mengetahui Pola Indeks Pembangunan Manusia Tahun 2022. *Journal of Information and Information Security (JIFORTY)*, *4*(1), 65–76.

http://ejurnal.ubharajaya.ac.id/index.php/jiforty

Rusdi, M. (2023). Economics and Digital Business Review Pengaruh Index Pembangunan Manusia Terhadap Kemiskinan di Sulawesi Selatan. *Pengaruh Index Pembangunan Manusia Terhadap Kemiskinan Di Sulawesi Selatan…*, *4*(1), 971–981.

Sianipar, K. D. R., & Gunawan, I. (2021). Algoritma K-Means Dalam Pengelompokan Kabupaten/Kota Berdasarkan Indeks Pembengunan Manusia Di Sumatera Utara. *Jurnal Infomedia*, *6*(2), 57. https://doi.org/10.30811/jim.v6i2.2426

ISSN: 2776-7930 (Print) / 2807-3010 (Online)

Sibarani, H., Solikhun, Saputra, W., Gunawan, I., & Nasution, Z. M. (2022). Penerapan Metode K-Means Untuk Pengelompokkan Kabupaten/Kota Di Provinsi Sumatera Utara Berdasarkan Indikator Indeks Pembangunan Manusia. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *6*(1), 154–161. https://doi.org/10.36040/jati.v6i1.4590

Talakua, M. W., Leleury, Z. A., & Talluta, A. W. (2017). Analisis Cluster Dengan Menggunakan Metode Provinsi Maluku Berdasarkan Indikator Indeks Pembangunan Manusia Tahun 2014. *Jurnal Ilmu Matematika Dan Terapan*, *11*(2), 119–128.

Yektiningsih, E. (2018). Analisis Indeks Pembangunan Manusia (Ipm) Kabupaten Pacitan Tahun 2018. *Jurnal Ilmiah Sosio Agribis*, *18*(2), 32–50.\]

https://doi.org/10.30742/jisa1822018528

Zaki, A., Irwan, I., & Sembe, I. A. (2022). Penerapan K-Means Clustering dalam Pengelompokan Data (Studi Kasus Profil Mahasiswa Matematika FMIPA UNM). *Journal of Mathematics Computations and Statistics*, *5*(2), 163.

https://doi.org/10.35580/jmathcos.v5i2.38820