Jurnal Mandiri IT ISSN 2301-8984 (Print)
Vol. 11 No. 1, July (2022), pp. 20-25

Application of the K-Means Clustering Algorithm in Grouping Regencies/Cities in North Sumatra Province Based on Human Development Index Indicators

Sonia Sihombing[1], Rahmat W. Sembiring[2], Eka Irawan[3]
[1,3]STIKOM Tunas Bangsa Pematangsiantar, North Sumatra, Indonesia
[2]AMIK Tunas Bangsa Pematangsiantar, North Sumatra, Indonesia

ARTICLE INFO

Article history:

Received May 10, 2022
Revised Jun 12, 2022
Accepted Jul 28, 2022

Keywords:

Data Mining
Human Development Index
Clustering
K-Means

ABSTRACT

This research aims to group districts/cities in the Province of North Sumatera based on the Human Development Index Indicator. To solve this problem, the researcher uses the K-Means Clustering Algorithm Method. Where the source of research data is collected based on documents describing the Human Development Index produced by the Central Bureau of Statistics. The data used in this researh is the Human Development Index data and the Human Development Index Indicator in 2020. The data will be processed by doing clustering in 3 clusters namely high, medium, and low level of development index. By doing this research, it is expected to be able to the provincial government of North Sumatera regarding the policies that need to be carried out to create equitable development in the Province of North Sumatera, as well as increase the Human Development Index in the Province of North Sumatera.

Corresponding Author:

Sonia Sihombing,
STIKOM Tunas Bangsa Pematangsiantar, North Sumatra, Indonesia,
Jl. Jend.Sudirman Blok A No.1-3 Pematangsiantar, North Sumatra
Email: soniashb08@gmail.com

1. INTRODUCTION

Humans are the real wealth of the nation so that the ultimate goal of development must be focused on humans. This condition will create an environment that allows people to enjoy a long, healthy and productive life. This condition became the forerunner of the emergence of the human development index (HDI). The Human Development Index was introduced by the United Nations Development Program (UNDP) in 1990 and is published regularly in the annual Human Development Report (HDR). The Human Development Index is an important indicator to measure success in efforts to build the quality of human life (community/population) and can determine the ranking or level of building a region/country. In Indonesia, the HDI is used as the basis for determining the central government's transfer budget, namely the General Allocation Fund (DAU) and the Regional Incentive Fund (DID) for provinces and districts/cities(Aprianto, 2018; Sapaat et al., 2020).

The Human Development Index is formed by three basic dimensions, namely longevity and healthy living, knowledge, and a decent standard of living (UARA, 2021) . In 2014 the HDI used a new calculation method with four indicators, namely life expectancy at birth, expected years of schooling, average length of schooling, andper capita expenditure adjusted(Sangga, 2018; Wicaksono & Yolanda, 2021). Life expectancy is defined as the estimated average number of years

taken by a person from birth. The average length of schooling is defined as the number of years the population spends in formal education. Expected length of schooling is defined as the length of schooling (in years) that is expected to be felt by children at a certain age in the future. While the adjusted per capita expenditure is determined from the value of the per capita expenditure of purchasing power parity funds (Mongan, 2019).

Development programs carried out by the Central Government and Regional Governments cannot be separated from the high and low values of the Human Development Index. One of them is North Sumatra Province. North Sumatra Province has 25 regencies and 8 cities. The North Sumatra Human Development Index from 2015-2019 experienced an average growth of 0.8 percent per year. In 2020 the HDI achievement in North Sumatra has reached 71.77. This figure increased by 0.03 points from the achievement in 2019 which was 71.74 (Simbolon, 2021) . This development shows the improvement in human development in general in North Sumatra. However, the problem that arises is that the increase in HDI in North Sumatra Province is not followed by equitable distribution of development and human development in each Regency/City in North Sumatra Province.The grouping of Regencies/Cities in North Sumatra Province needs to be done as material for program planning for the following year and as an evaluation of the Government's program targets to increase human development figures based on indicators forming the Human Development Index. The grouping also aims for equitable development in North Sumatra Province. There are several grouping algorithms that can be used, one of which is the K-Means Algorithm (Muningsih et al., 2021; Sinaga et al., 2021).

K-Means is a method that is included in thealgorithm clustering distance-basedthat divides data into a number of clusters and this algorithm only works on numerical attributes (Nabila et al., 2021). Previous research studies have applied thetechnique K-Means Clustering including (Sinaga et al., 2021) from the results of the analysis carried out, it can be seen that the research can be completed by data mining techniques usingrules clustering to group villages/kelurahan according to the anticipation/mitigation of natural disasters in Indonesia. Indonesia. So that an assessment is obtained based on village/kelurahan indices according to natural disaster anticipation/mitigation efforts with 3 high-level provinces, namely West Java, East Java, 9 medium-level provinces, and 22 other provinces including low levels(Anggraeni et al., 2021; Sapaat et al., 2020).

Another research in this reserach explains that the K-Means Cluster Analysis method is quite effective in being applied in the process of classifying the characteristics of the research object(Alkhairi & Windarto, 2019). The K-Means algorithm is also not affected by the order of objects used, this is proven when the researcher tries to randomly determine the starting point of the cluster center of one of the objects at the beginning of the calculation. From research conducted is concluded that theAlgortima K-Means Clustering can be used to group data efficiently and effectively with the expected results(Fammaldo & Hakim, 2018).

2. RESEARCH METHOD

The research methodology provides an overview of the research design which includes, among others: details of all the sequences of conducting research and analyzing and designing the system used. The purpose of this research is to collect information and manage data to solve research problems and find solutions to the problems to be studied. The method used in this research is using themethod K-Means Clustering. The results of this study were conducted to determine the final result of the grouping of Regencies/Cities of North Sumatra Province based on the Human Development Index Indicator.

In this research, a research model is used which is presented in the form of a flow chart design in Figure 1.

ISSN 2301-8984 (Print)

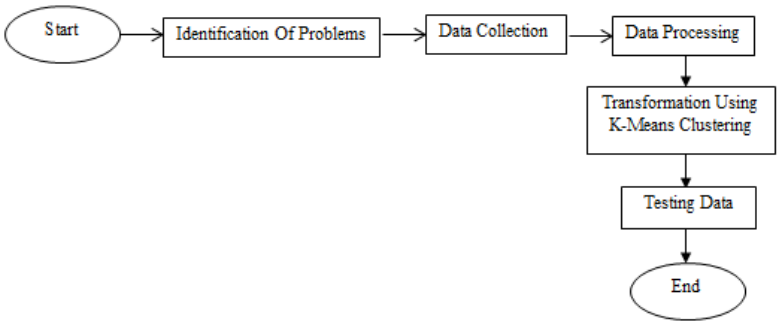

Figure 1. Flowchart Research Design

Process flowchart in Figure 1 are as follows: Problem Identification were analyzing a problem related to the Human Development Index (HDI) data along with indicators forming the 2020 HDI, Data Collection Techniques The data of this study were obtained from the Indonesian Central Statistics Agency (BPS) with the website https://bps.go.id regarding Human Development Index data. Data Processing; at this stage, the data will be processed using data mining techniques with the K-Means Clustering method. Transformation using theMethod. K-Means Clustering; Alphabet type data must first be initialized intoformnumerical. Then group the existing data into three clusters, namely high level, medium level and low level using the K-Means Clustering method. Data Testing With RapidMiner; this stage is testing the data using software RapidMiner, so the comparison results obtained in manual data processing with the results of data processing using a software.

The data analysis process can be carried out after collecting valid data. In conducting this research, the authors analyze secondary statistical data in which the data is not obtained from the direct source, but has been collected and processed in detail, where the data relates to the problem being studied. The following is data obtained from the Central Statistics Agency (BPS):

Table 1. HDI Data and HDI Indicators for North Sumatra Province Regency/City in 2020
Source: (Central Bureau of Statistics)

District/City	Components Of The Human Development Index				Human Development Index
	Life Expectancy	Expected Length Of Schooling	Average Length Of School	Per Capita Expenditure Adjusted *(Rp.000,-)*	
Districts					
01. N i a s	69,75	12,57	5,36	6898	61,93
02. Mandailing Natal	62,6	13,32	8,62	9684	66,79
03. Tapanuli Selatan	64,91	13,24	9,28	11236	70,12
04. Tapanuli Tengah	67,15	13,06	8,62	10071	69,23
05. Tapanuli Utara	68,63	13,69	9,85	11648	73,47
06. Toba	70,08	13,45	10,52	12154	75,16
07. Labuhanbatu	69,93	12,73	9,24	11150	72,01
08. A s a h a n	68,26	12,60	8,79	10890	70,29
09. Simalungun	71,22	12,78	9,60	11308	73,25
10. D a i r i	69,00	13,10	9,58	10350	71,57
11. K a r o	71,40	12,76	9,79	12349	74,43
…..	……	…..	……	…….	…..
33. Gunungsitoli	71,19	13,74	8,61	7980	69,31

Based on table 1 above, it shows the Human Development Index data along with the Human Development Index Indicators by Regency/City of North Sumatra Province in 2020.

3. RESULTS AND DISCUSSIONS

The research results are presented in accordance with the research that the author has done. The data used in this study is the Human Development Index data and the District/City Human Development Index Indicators of North Sumatra Province in 2020. In this study, they are grouped into 3 parts, namely High, medium and low level Human Development Index Indicator Areas.

The collection of data that the author obtained is used as input data in making a rule model using the K-Means algorithm and using Rapidminer software. To get the results of the research conducted, the following is a description of the manual calculation of the Human Development Index clustering process using the K-Means algorithm. The clustering process is done by determining the data you want to cluster. In this case, the data variables to be clustered are Life Expectancy Age, Expected Length of Schooling, Average Years of Schooling, Per Capita Expenditure and the Human Development Index.

3.1 Determining the Centroid Value (Cluster Center)

The determination of the initial cluster center is determined randomly taken from the data available in the range. The value for high cluster (C1) is taken from the highest value found in table 4.1 for medium cluster (C2) is taken from the average value of each variable found in table 4.1, and for low cluster (C3) is taken from the lowest value found in table 2 below:

Table 2. Initial Data Centroid Iteration 1

	Score				
Cluster	X1	X2	X3	X4	X5
C1	73,55	14,74	11,39	14890	80,98
C2	69,02	13,19	9,1	10314	70,75
C3	62,6	12,23	5,36	5830	61,51

3.2 Calculating the Distance of Each Data to the Centroid

After the initial cluster center value data is determined, the next step is to calculate the distance of each data to the cluster center using Euclidean Distance. To calculate the distance of each Human Development Index data and Human Development Index Indicator to the center of the cluster, we can use a formula whose calculations up to centroid iteration 5 can be seen as follows:

Table 3. Initial Data Centroid Iteration 5

	Score				
Cluster	X1	X2	X3	X4	X5
C1	71,80	13,58	10,54	12811	76,66
C2	68,09	13,09	9,28	10915,9	71,25
C3	69,05	13,13	7,77	7443,89	65,82

The results of the calculation of the data distance with the center point of the cluster in iteration 5 using Eulidean Distance.

Table 4. Iteration Centroid Distance 5

No	Districts/City	C1	C2	C3	Shortest Distance
1	N i a s	5913,02	4017,96	545,91	545,91
2	Mandailing Natal	3127,03	1231,96	2240,12	1231,96
3	Tapanuli Selatan	1575,03	320,07	3792,12	320,07
4	Tapanuli Tengah	2740,01	844,95	2627,11	844,95
5	Tapanuli Utara	1163,01	732,06	4204,12	732,06
6	Toba	657,00	1238,06	4710,12	657,00
7	Labuhan Batu	1661,01	234,06	3706,12	234,06
8	A s a h a n	1921,01	25,97	3446,11	25,97
9	Simalungun	1503,00	392,07	3864,12	392,07
10	D a i r i	2461,01	565,95	2906,12	565,95
11	K a r o	462,01	1433,06	4905,12	462,01
….	……..	……	……..	…….	……..
33	Gunung Sitoli	4831,01	2935,95	536,13	536,13

ISSN 2301-8984 (Print)

Table 5. Iteration Cluster Position Grouping

No	Districts/City	C1	C2	C3
1	N i a s			1
2	Mandailing Natal		1	
3	Tapanuli Selatan		1	
4	Tapanuli Tengah		1	
5	Tapanuli Utara		1	
6	Toba	1		
7	Labuhan Batu		1	
8	A s a h a n		1	
9	Simalungun		1	
10	D a i r i		1	
11	K a r o	1		
12	Deli Serdang	1		
13	Langkat		1	
14	Nias Selatan			1
15	Humbang Hasundutan			1
16	Pakpak Bharat			1
17	Samosir			1
18	Serdang Bedagai		1	
19	Batu bara		1	
20	Padang Lawas Utara		1	
21	Padang Lawas			1
22	Labuhanbatu Selatan		1	
23	Labuhanbatu Utara		1	
24	Nias Utara			1
25	Nias Barat			1
26	Sibolga		1	
27	Tanjungbalai		1	
28	Pematangsiantar	1		
29	Tebing Tinggi	1		
30	M e d a n	1		
31	B i n j a i		1	
32	Padangsidimpuan		1	
33	Gunung Sitoli			1

Manual calculations on the human development index data and the human development index indicator data above obtained the final results where in iteration 5 the data grouping carried out on 3 clusters with iteration 3 obtained the same results. The results of the two iterations are C1 = 6, C2 = 18, and C3 = 9 in the data position of each cluster. So that the position of the cluster in the data does not change again, the iteration process stops. Based on theposition cluster of each human development index data as well as the human development index indicator data and thevalue cluster of iteration 5, it can be concluded are:

Cluster High (C1) with the number of human development index data and human development index indicator data as many as 3 Districts and 3 cities, namely: Toba, Karo, Deli Serdang, Pematangsiantar, Tebing Tinggi, and Medan.

Cluster Medium (C2) with the number of human development index data and human development index indicator data as many as districts and 4 cities, namely: Mandailing Natal, South Tapanuli, Central Tapanuli, North Tapanuli, Labuhanbatu , Asahan, Simalungun, Dairi, Langkat, Serdang Bedagai, Batu Bara, North Padang Lawas, South Labuhanbatu, North Labuhanbatu, Sibolga, Tanjungbalai, Binjai, Padangsidimpuan.

Cluster Low (C3) with the number of human development index data and human development index indicator data as many as 8 districts and 1 city, namely: Nias, South Nias, Humbang Hasundutan, Pakpak Barat, Samosir, Padang Lawas, North Nias, West Nias, Gunung Sitoli.

4. CONCLUSION

TheCluster First(C1) is ancluster areathat has a High Level Human Development Index Indicator number with a total of 3 Regencies and 3 Cities, the Second Cluster (C2) is ancluster areathat has a Medium Level Development Index Indicator number with a total of 14 Districts and 4 Cities, and the Third Cluster (C3) is ancluster areathat has a Low Level Human Development Index Indicator number with a total of 8 Regencies and 1 City.The Grouping of Regencies/Cities in North Sumatra

Province Based on the Human Development Index Indicators can be a recommendation to the North Sumatra Provincial Government and related institutions in determining policies and strategies for the development of the Districts and Cities of North Sumatra Province.

REFERENCES

Alkhairi, P., & Windarto, A. P. (2019). Penerapan K-Means Cluster Pada Daerah Potensi Pertanian Karet Produktif di Sumatera Utara. *Seminar Nasional Teknologi Komputer & Sains (SAINTEKS)*, *1*(1).

Anggraeni, D., Rizaldi, R., & Putra, G. M. (2021). Penerapan K-Means Clustering Untuk Pengelompokan Kelas Pada Taman Kanak-Kanak. *Building of Informatics, Technology and Science (BITS)*, *3*(3), 400–404.

Aprianto, K. (2018). Optimasi Kernel K-Means dalam Pengelompokan Kabupaten/Kota Berdasarkan Indeks Pembangunan Manusia di Indonesia. *Limits: Journal of Mathematics and Its Applications*, *15*(1), 1–15.

Fammaldo, E., & Hakim, L. (2018). Penerapan Algoritma K-Means Clustering Untuk Pengelompokan Tingkat Kesejahteraan Keluarga Untuk Program Kartu Indonesia Pintar. *Jurnal Ilmiah Teknologi Infomasi Terapan*, *5*(1), 23–31.

Mongan, J. J. S. (2019). Pengaruh pengeluaran pemerintah bidang pendidikan dan kesehatan terhadap indeks pembangunan manusia di Indonesia. *Indonesian Treasury Review: Jurnal Perbendaharaan, Keuangan Negara Dan Kebijakan Publik*, *4*(2), 163–176.

Muningsih, E., Maryani, I., & Handayani, V. R. (2021). Penerapan Metode K-Means dan Optimasi Jumlah Cluster dengan Index Davies Bouldin untuk Clustering Propinsi Berdasarkan Potensi Desa. *EVOLUSI: Jurnal Sains Dan Manajemen*, *9*(1).

Nabila, Z., Isnain, A. R., Permata, P., & Abidin, Z. (2021). Analisis Data Mining Untuk Clustering Kasus Covid-19 Di Provinsi Lampung Dengan Algoritma K-Means. *Jurnal Teknologi Dan Sistem Informasi*, *2*(2), 100–108.

Sangga, V. A. P. (2018). *Perbandingan algoritma K-Means dan algoritma K-Medoids dalam pengelompokan komoditas peternakan di provinsi Jawa Tengah tahun 2015*.

Sapaat, T. M., Lapian, A. L. C. P., & Tumangkeng, S. Y. L. (2020). Analisis faktor-faktor yang mempengaruhi indeks pembangunan manusia di Provinsi Sulawesi Utara tahun (2005-2019). *Jurnal Berkala Ilmiah Efisiensi*, *20*(03).

Simbolon, T. R. (2021). *Analisis Pengaruh Kemandirian Keuangan Daerah, Pendapatan Perkapita dan Jumlah Penduduk Miskin Terhadap Indeks Pembangunan Manusia di Indonesia*. UNIMED.

Sinaga, J. L. S., Solikhun, S., & Suhendro, D. (2021). Penerapan Algoritma K-Means Dalam Mengelompokkan Rata-Rata Konsumsi Kalori Menurut Provinsi. *Jurasik (Jurnal Riset Sistem Informasi Dan Teknik Informatika)*, *6*(1), 75–88.

UARA, A. (2021). *Analisis Faktor Yang Mempengaruhi Indeks Pembangunan Manusia (Ipm) Provinsi Jawa Tengah Tahun 2017-2019 Menggunakan Regresi Data Panel (Studi Kasus: Indeks Pembangunan Manusia Provinsi Jawa Tengah Tahun 20172019)*.

Wicaksono, A. S., & Yolanda, A. M. (2021). Pengelompokkan Kabupaten/Kota di Provinsi Nusa Tenggara Timur Berdasarkan Indikator Indeks Pembangunan Manusia Menggunakan K-Medoids Clustering. *Jurnal Statistika Terapan (ISSN 2807-6214)*, *1*(1), 79–90.

Witten, I.H and Frank, E. (2005). Data Mining : Practical Machine Learning Tools and Techniques Second Edition. Morgan Kauffman : San Francisco.