Jurnal Mantik, 5 (4) (2022) 2392-2399

Published by:Institute of Computer Science (IOCS)

Jurnal Mantik

Journal homepage: www.iocscience.org/ejournal/index.php/mantik/index

K-Means Clustering Gross Participation Rate Regency/City Area In North Sumatra

Sujarwo

Manajemen Informatika,
Politeknik Unggul LP3M, Indonesia

E-mail: sujarwo2268@gmail.com

A R T I C L E I N F O

Article history:

Received: Des 27, 2021
Revised: Jan 17, 2022
Accepted: Feb 28, 2022

Keywords:

Clustering,
K-Means,
Gross Participation Rate

A B S T R A C T

The percentage of school-age population in North Sumatra Province is 35.95%. Meanwhile, the average gross participation rate for districts/cities in North Sumatra is for the SD level of 110.71; the average for junior high school level is 93.73; for high school level 89.93; and for College 20.23. In this study, the data was grouped using the K-Means Clustering algorithm on the Regency/City Gross Participation Rate data. The application of the K-Means Clustering algorithm is carried out for up to 5 iterations with 4 clusters. The first cluster includes the areas of Mandailing Natal, Tapanuli Tengah, Asahan with an average Gross Participation Rate value, the smallest Gross Participation Rate, the largest Gross Participation Rate respectively 76.39; 73.40; and 81.14. For the Second Cluster covering the Medan City area, the average Gross Participation Rate value, the smallest Gross Participation Rate , and the largest Gross Participation Rate are the same, namely 87.50. The Third Cluster covers the areas of Nias, Tapanuli Selatan, Tapanuli Utara, Simalungun, Karo, Deli Serdang, Nias Selatan, Serdang Bedagai, Batu Bara, Nias Utara, Sibolga, Tanjung Balai, Pematang Siantar, Tebing Tinggi, Binjai, Padang Sidempuan, Gunungsitoli with average Gross Participation Rate value, smallest Gross Participation Rate, largest Gross Participation Rate respectively 78.02; 73.45; 85.62. The Fourth Cluster includes Toba, Labuhanbatu, Dairi, Langkat, Humbang Hasundutan, Pakpak Bharat, Samosir, Padang Lawas Utara, Padang Lawas, Labuhan Batu Utara, Nias Barat areas with the average Gross Participation Rate value, the smallest Gross Participation Rate , the largest Gross Participation Rate respectively 79,38; 76.84; 81.43

1. Introduction

The area of North Sumatra Province is 72,981.23 square km [1], with a population in 2020 of 14,799,361 people [2] with a population density of North Sumatra of approximately 200 people per square km. Meanwhile, the population aged 5 – 24 years is 5,320,299 people [2], which means the percentage of the school-age population is 35.95%. Based on data from the Central Statistics Agency of North Sumatra, the Gross Enrollment Rate in 2020 is on average for the Elementary School level of 110.71; the average for junior high school level is 93.73; for high school level 89.93; and for College 20.23.

The Gross Participation Rate is a comparison between the number of people who are still in school at a certain level of education (regardless of the age of the population) and the number of people who meet the official requirements of the school-age population at the same level of education [3]. Each region in the Regency/City of North Sumatra Province has a different Gross Participation Rate. These differences can occur due to several reasons, including the number of residents, the number of jobs, geographical conditions of the area, social conditions and others. The following is a graph of the Regency/City Gross Participation Rate in North Sumatra..

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

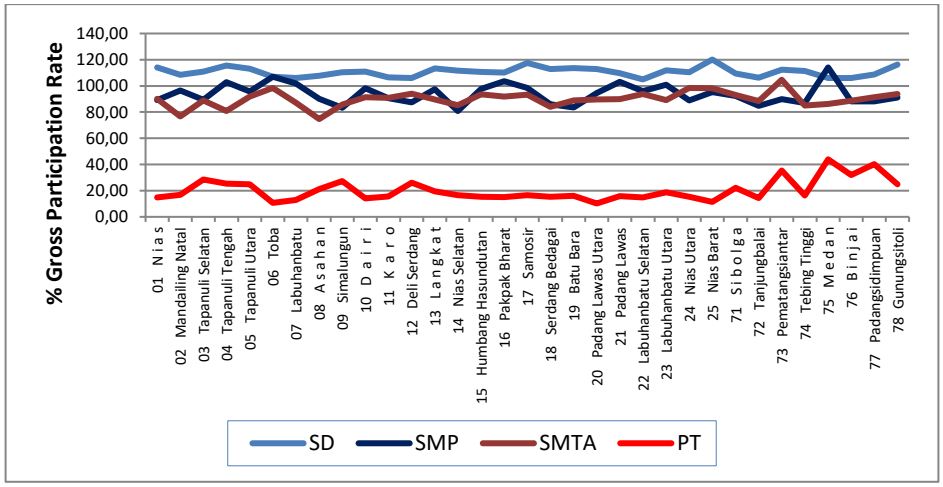

Fig 1. North Sumatra Regency/City Gross Participation Rate

Clustering algorithm with K-Means method is a method that can be used to group data. Based on previous research in classifying data, including grouping Covid-19 data in Indonesia in 2020 [4], determining the level of sales of Telkomsel data packages [5], mapping the number of train passengers [6], predicting student graduation times [7], Grouping the nutritional status of children under five in Kembang Songo Village [8] and others. Likewise, the district/city Gross Participation Rate in North Sumatra can be grouped using the K-Means method

By grouping the regions based on the Gross Participation Rate, to map the similarities between the regions based on the Gross Participation Rate, and to make it easier to take actions on the regions in one group. In grouping regions based on Gross Participation Rate, in this study the Spreadsheet application was used

2. Methods

2.1. Gross Participation Rate

The Gross Participation Rate is the ratio between the number of people who are still in school at a certain level of education (regardless of the age of the population) and the number of people who meet the official requirements for the population of school age at the same level of education. Since 2007, Non-Formal Education (Package A, Package B, and Package C) has been taken into account.

The Gross Participation Rate can be more than 100 percent because the population of students attending a certain level of education includes children outside the school age limit at that level of education. The reason is the registration of students at an early age, registration of students who are late for school, or repetition of classes. This can also show that the area is able to accommodate the school-age population more than the actual target. A high Gross Participation Rate indicates a high level of school participation, regardless of the accuracy of school age at the level of education [3].

2.2. Data Mining

Data mining is a method in computer science that is commonly used in the knowledge search process. The stages in it are useful for finding certain patterns from the data in the database. This method is widely found in the fields of machine learning and statistics. At first, data mining methods were developed due to the increasing complexity of computer work. However, this is where the advantage of data mining in the process of collecting and selecting data is more practical.

According to Larose (2006), interpreting data mining is a process of finding something meaningful by sorting data through a repository with the help of pattern socialization technology, statistics, and mathematics. According to Berry, data mining is an activity of analyzing large amounts of data in order to find useful patterns and rules. According to Pramudiono (2006), conveying that data mining is an analysis process that is carried out automatically on complex and large amounts of data to obtain a pattern or trend that is generally not realized..

Not a few people do not know what the difference between data warehouse and data mining is. In terms of its name, data mining is a combination of two English words "data" which means data and "mining" which means to mine. In other words, data mining is a data mining process. While the data "warehouse" is a warehouse

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

or data storage area. In addition to the differences in data warehousing and data mining, both have the same use of words intended to describe a process. But data warehousing means data collection. Despite the differences between data warehouse and data mining, the two are still interrelated. The data mining process requires a data warehouse to retrieve data to be processed and observed for patterns [9].

The data mining method is a method that is applied but needs to be adapted to the user's goals. There are several distributions of the following data mining methods that you can know [9].

a. Classification

Classification of data mining is a process of finding the definition of the similarity of characteristics in a group or class (class). Classification of data mining is one of the most common methods to use. This method aims to estimate the class of an object whose label is not known.

b. Association

This method aims to identify products that are often purchased together by customers. For example, some customers will buy packaged snacks and drinks at the same time. That way it's easier for companies to know if the two items are often purchased together.

c. Clustering

It is a segmentation method. The purpose of segmentation in the data mining method is to group a class into several segments based on the specified attributes.

d. Regression

A method that aims to look for patterns of numeric values, not classes. The result of the regression method is a functional relationship as a determinant of results based on the value of the input.

e. Forecasting

Forecasting data mining is a method used to predict the value to be achieved in one period. By using this technique, noise data and values in the previous period are used as the basis for predictions.

f. Sequencing

Sequence is a sequence of events that serves to find a pattern in a series of events or sequence

g. Descriptive

This data mining method aims to understand more deeply about the data included in the observation. The end result is knowing the behavior of the data itself.

Computer science and artificial intelligence are inseparable from the technical use of data mining. Its benefits can be felt in various other fields, including business and marketing. Here are a number of benefits of data mining.

a. Knowing the trend
b. Methods for predicting future business decisions
c. Knowing the products purchased together
d. Observing consumer behavior
e. Model as a means to develop sales increase strategy

2.3. Clustering

Clustering is a common technique for statistical data analysis, used in many fields, including machine learning, data mining, pattern recognition, image analysis, and bioinformatics. Clustering is the process of grouping similar objects into different groups, or more precisely, dividing a data set into subsets, so that the data in each subset is according to some specified distance measure [10].

2.4. K-Means Clustering

K-Means Clustering is a data analysis method or Data Mining method that performs the modeling process without supervision (unsupervised) and is one method that performs data grouping with a partition system. There are two types of data clustering that are often used in the data grouping process, namely Hierarchical and Non-Hierarchical, and K-Means is one of the non-hierarchical or Partitional Clustering data clustering methods. The K-Means Clustering method tries to group the existing data into several groups, where the data in one group has the same characteristics as each other and has different characteristics from the data in other groups. In other words, the K-Means Clustering method aims to minimize the objective function set in the clustering process by minimizing variations between data in a cluster and maximizing variations with data in other clusters [10].

Algorithm with K-Means [8]:

a. Determine the number of Clusters

To determine the number of clusters, you can determine yourself according to your needs

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

b. Allocate data into groups randomly
To allocate the data, based on a data center that can be taken randomly from existing data

c. Calculate the center of the group (centroid/mean) from the data in each group

There are several ways that can be used to measure the distance of the data to the center of the group, including Euclidean. The measurement of distance in Euclidean distance space uses the formula:

$$\mathbf{D(X_1, X_2) = \left|\left| X_1 - X_2 \right|\right| = \sqrt{\sum_{j=1}^{p}(X_{2j} - X_{1j})^2}} \quad (1)$$

d. Allocate each data to the nearest central point

$$\mathbf{a_{il} = \begin{cases} 1 & d = \min\{D(X_i, C_1)\} \\ 0 & \text{others} \end{cases}} \quad (2)$$

a_{ik} is the point membership value X_i to group center C_l, d is the shortest distance from the data X_i to K group after comparison, and C_l is the lth centroid (group center) to-*l*.

e. Return to step three, when there is still data moving around the cluster, or when there is a change in the centroid value above the specified threshold value, or when the change in the value on the objective function used is still above the specified threshold value.

3. Result and Analysis

3.1 Result

a. Data collection

The data is taken from the web www.sumut.bps.go.id, namely data on the Gross Partition Rate in Regencies/Cities in North Sumatra. The data taken is the Gross Participation Rate data for elementary, junior high, high school and university levels. The following is the Gross Enrollment Rate Data in North Sumatra [3]

Table 1
Gross Participation Rate in North Sumatra

Kabupaten/Kota	Angka Partisipasi Kasar.			
	SD	SMP	SMA	Perg. Tinggi
01 N i a s	114,10	89,08	90,07	14,95
02 Mandailing Natal	108,54	96,45	76,77	16,74
03 Tapanuli Selatan	110,96	89,50	88,99	28,50
04 Tapanuli Tengah	115,62	102,75	80,91	25,27
05 Tapanuli Utara	113,17	95,67	91,91	24,77
06 Toba	106,97	107,01	98,51	10,65
07 Labuhanbatu	106,06	101,79	87,26	12,77
08 A s a h a n	107,76	90,01	74,65	21,19
09 Simalungun	110,34	83,37	85,66	27,18
10 D a i r i	111,00	98,24	91,44	14,18
11 K a r o	106,58	90,55	90,88	15,53
12 Deli Serdang	106,03	87,42	93,92	26,12
13 L a n g k a t	113,34	97,39	89,63	19,58
14 Nias Selatan	111,59	80,78	85,29	16,50
15 Humbang Hasundutan	110,70	97,60	93,64	15,26
16 Pakpak Bharat	110,04	103,67	91,80	14,99
17 Samosir	117,57	98,31	93,39	16,43
18 Serdang Bedagai	112,93	86,00	83,94	15,21
19 Batu Bara	113,70	83,57	88,95	16,06
20 Padang Lawas Utara	112,93	94,49	89,72	10,20
21 Padang Lawas	109,67	102,96	89,83	15,85
22 Labuhanbatu Selatan	104,95	95,93	93,73	14,92
23 Labuhanbatu Utara	111,92	100,95	89,21	18,78
24 Nias Utara	110,32	88,95	98,44	15,36
25 Nias Barat	120,12	95,27	98,17	11,49
71 S i b o l g a	109,54	92,30	93,11	22,10
72 Tanjungbalai	106,35	84,68	88,32	14,46
73 Pematangsiantar	112,44	89,90	104,65	35,48
74 Tebing Tinggi	111,43	87,05	85,10	16,18

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Kabupaten/Kota	Angka Partisipasi Kasar.			
	SD	SMP	SMA	Perg. Tinggi
75 M e d a n	105,90	114,07	86,13	43,89
76 B i n j a i	105,93	88,07	88,61	32,00
77 Padangsidimpuan	108,77	88,21	91,41	40,17
78 Gunungsitoli	116,21	91,14	93,79	24,91

b. Clustering K-Means

1) Number of Clusters

Determining the number of clusters can be done randomly. In this study, it was carried out based on the smallest average Gross Partition Rate, the largest average Gross Partition Rate and randomly

a) The data with the smallest average Gross Partition Rate is Asahan Regency
b) The data with the largest average Gross Partition Rate is Medan City
c) Data on Gross Partition Rate from Batu Bara District
d) Data on Gross Partition Rate from North Labuhan Batu Regency

Center $C_1(107{,}76\ ;\ 90{,}01\ ;\ 74{,}65\ ;\ 21{,}19)$
Center $C_2(105{,}93\ ;\ 114{,}07\ ;\ 86{,}13\ ;\ 43{,}89)$
Center $C_3(113{,}70\ ;\ 83{,}57\ ;\ 88{,}95\ ;\ 16{,}06)$
Center $C_4(111{,}92\ ;\ 100{,}95\ ;\ 89{,}21\ ;\ 18{,}78)$

2) First Iteration

Based on formula (1), calculated the value of the distance data from the center
First data group C1 (Nias Regency)

$$C_{11} = \sqrt{(114{,}10 - 107{,}76)^2 + (89{,}08 - 90{,}01)^2 + (90{,}07 - 74{,}65)^2 + (14{,}95 - 21{,}19)^2}$$

Second data Group C_1 (Mandailing Natal Regency)

$$C_{12} = \sqrt{(108{,}54 - 107{,}76)^2 + (96{,}45 - 90{,}01)^2 + (76{,}77 - 74{,}65)^2 + (16{,}74 - 21{,}19)^2}$$

The same goes for Clusters C_1. Continued with Cluster C_2
First data group C2 (Nias Regency)

$$C_{21} = \sqrt{(114{,}10 - 105{,}90)^2 + (89{,}08 - 114{,}07)^2 + (90{,}07 - 86{,}13)^2 + (14{,}95 - 43{,}89)^2}$$

First data group C3 (Nias Regency)

$$C_{31} = \sqrt{(114{,}10 - 113{,}70)^2 + (89{,}08 - 83{,}57)^2 + (90{,}07 - 88{,}95)^2 + (14{,}95 - 16{,}06)^2}$$

First data group C4 (Nias Regency)

$$C_{41} = \sqrt{(114{,}10 - 111{,}92)^2 + (89{,}08 - 100{,}95)^2 + (90{,}07 - 89{,}21)^2 + (14{,}95 - 18{,}78)^2}$$

First Iteration Results
Based on formula (2), the data allocation for each group is obtained

Table 2
First Iteration Data Allocation

Regency / City	C1	C2	C3	C4
01 N i a s			1	
02 Mandailing Natal	1			
03 Tapanuli Selatan			1	
04 Tapanuli Tengah	1			
05 Tapanuli Utara				1
06 Toba				1
07 Labuhanbatu				1
08 A s a h a n	1			
09 Simalungun			1	
10 D a i r i				1
11 K a r o			1	
12 Deli Serdang			1	
13 L a n g k a t				1
14 Nias Selatan			1	
15 Humbang Hasundutan				1
16 Pakpak Bharat				1
17 Samosir				1
18 Serdang Bedagai			1	
19 Batu Bara			1	

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

	Regency / City	C1	C2	C3	C4
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
71	S i b o l g a			1	
72	Tanjungbalai			1	
73	Pematangsiantar			1	
74	Tebing Tinggi			1	
75	M e d a n		1		
76	B i n j a i			1	
77	Padangsidimpuan			1	
78	Gunungsitoli			1	

3) Second Iteration

Based on the results of the first iteration, followed by the second iteration, by taking the center point of the data allocation results in the first iteration, namely taking the average data value of 1 from each group as a benchmark in taking the new center point.

Here's for the First Cluster center

a) The amount (Mandailing Natal Data + Tapanuli Tengah Data + Asahan Data) divided into three, respectively obtained

$(108{,}54 + 115{,}62 + 107{,}76)/3 = 107{,}41.$

$(96{,}54 + 102{,}75 + 90{,}01)/3 = 91{,}51.$

$(76{,}77 + 80{,}91 + 74{,}65)/3 = 80{,}01.$

$(16{,}74 + 25{,}27 + 21{,}12)/3 = 23{,}31.$

Pusat $C_1(107{,}41\ ;\ 91{,}51\ ;\ 80{,}01\ ;\ 23{,}31)$

b) In the same way, then search for C_2, C_3, C_4 as the new Cluster center.

Results of new cluster centers in a row

Center $C_2(105{,}90;\ 114{,}07;\ 86{,}13;\ 43{,}89)$

Center $C_3(110{,}72;\ 86{,}95;\ 89{,}60;\ 20{,}86)$

Center $C_4(111{,}63;\ 98{,}39;\ 92{,}31;\ 17{,}67)$

The next step is to find the data distance from each center point, as in the first iteration. And so on until the same data distance value is found. In this study the iteration ended until the fifth iteration. With the following results

Table 3

Fifth Iteration Data Allocation

	Regency / City	C1	C2	C3	C4
01	N i a s			1	
02	Mandailing Natal	1			
03	Tapanuli Selatan			1	
04	Tapanuli Tengah	1			
05	Tapanuli Utara			1	
06	Toba				1
07	Labuhanbatu				1
08	A s a h a n	1			
09	Simalungun			1	
10	D a i r i				1
11	K a r o			1	
12	Deli Serdang			1	
13	L a n g k a t				1
14	Nias Selatan			1	
15	Humbang Hasundutan				1
16	Pakpak Bharat				1
17	Samosir				1
18	Serdang Bedagai			1	
19	Batu Bara			1	

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

	Regency / City	C1	C2	C3	C4
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
71	S i b o l g a			1	
72	Tanjungbalai			1	
73	Pematangsiantar			1	
74	Tebing Tinggi			1	
75	M e d a n		1		
76	B i n j a i			1	
77	Padangsidimpuan			1	
78	Gunungsitoli			1	

3.2 Discussion

Based on the results of the cluster, it is obtained descriptive of the Regency/City Gross Participation Rate in North Sumatra. There is one area of Medan City that is not in a group with other regions, so that the Average, Smallest, and Largest Gross Partition Rate are the same.

Table 4
Data descriptive of the Gross Partition Rate for each Cluster

Cluster	Data Angka Partisiasi Kasar		
	Average	Min	Max
Cluster 1	76,39	73,40	81,14
Cluster 2	87,50	87,50	87,50
Cluster 3	78,02	73,45	85,62
Cluster 4	79,38	76,84	81,43

In accordance with the usefulness of the Gross Participation Rate, according to research conducted in the journal The Effect of Educational Participation on the percentage of the poor, that the increase/decrease in the value of the Gross Participation Rate can affect the rise/fall of poverty in the Central Java region [11]. Likewise with this research that with the grouping of the Gross Participation Rate, uniform policies can be taken for regions in one group.

4. Conclution

Based on the research that has been carried out, it is obtained that the clustering of the Gross Participation Rate of North Sumatra Regency / City if divided into 4 clusters the results are as follows:

a. The first cluster covers the Mandailing Natal area, Central Tapanuli, Asahan
b. The Second Cluster covers the Medan City area
c. The third cluster covers the areas of Nias, South Tapanuli, North Tapanuli, Simalungun, Karo, Deli Serdang, South Nias, Serdang Bedagai, Batu Bara, North Nias, Sibolga, Tanjung Balai, Pematang Siantar, Tebing Tinggi, Binjai, Padang Sidempuan, Gunungsitoli
d. The Fourth Cluster covers the districts of Toba, Labuhanbatu, Dairi, Langkat, Humbang Hasundutan, Pakpak Bharat, Samosir, Padang Lawas Utara, Padang Lawas, Labuhan Batu Utara, Nias Barat.

Other researchers can classify the Gross Enrollment Rate in North Sumatra with more clusters. Because with four clusters there are cities that have nothing in common with other regions

References

[1] BPSSumut, "Luas Daerah Sumatera Utara," *Web*, 2020. [Online]. Available: https://sumut.bps.go.id/statictable/2021/04/19/2066/luas-daerah-dan-jumlah-pulau-menurut-kabupaten-kota-di-provinsi-sumatera-utara-2020.html.
[2] BPSSumut, "Jumlah Penduduk Menurut Kelompok Umur dan Jenis Kelamin," *Web*, 2021. [Online]. Available: https://sumut.bps.go.id/statictable/2021/04/19/2100/jumlah-penduduk-menurut-kelompok-umur-dan-jenis-kelamin-2020.html.
[3] BPS, "Angka Partisipasi Kasar," *Web*, 2021. [Online]. Available: https://sirusa.bps.go.id/sirusa/index.php/indikator/565.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

[4] R. A. Indraputra and R. Fitriana, "K-Means Clustering Data COVID-19," vol. 10, no. 3, pp. 275–282, 2020.
[5] S. Handoko, E. Tri, and E. Handayani, "Implementasi Data Mining Untuk Menentukan Tingkat Penjualan Paket Data Telkomsel Menggunakan Metode K-Means Clustering," vol. 25, no. 1.
[6] A. Dan *et al.*, "Analisa dan Pemetaan Jumlah Penumpang Kereta Api di Indonesia Menggunakan Metode Statistik Deskriftif dan K-Means Clustering," vol. 3, no. 2, pp. 1–9, 2019.
[7] H. Priyatman, F. Sajid, and D. Haldivany, "Klasterisasi Menggunakan Algoritma K-Means Clustering untuk Memprediksi Waktu Kelulusan Mahasiswa," pp. 62–66, 2019.
[8] J. Informatika, W. Mega, and P. Dhuhita, "Clustering Menggunakan Metode K-Means Untuk Menentukan Status Gizi Balita," vol. 15, no. 2, 2015.
[9] Populix, "Apa itu Data Mining? Pengertian, Metode, Tahapan, dan Contoh Terbaru," *Web*. [Online]. Available: https://www.info.populix.co/post/data-mining-adalah.
[10] M. G. H. Omran, A. P. Engelbrecht, and A. Salman, "An overview of clustering methods," *Intell. Data Anal.*, vol. 11, no. 6, pp. 583–605, 2007.
[11] A. Ratih, A. Indrayani, F. Ekonomi, and U. Janabadra, "Pengaruh Partisipasi Pendidika terhadap Tingkat Kemiskinan Di Provinsi Jawa Tengah," vol. I, no. 2, pp. 123–134, 2010.