ClusteringThe Level Of Education Quality In The Central Java ProvinceUsing C-Means Method

Prizka Rismawati Arum[1], Indah Manfaati Nur[2]
{prizka.rismawati@gmail.com, indahmnur@unimus.ac.id}

[1]Universitas Muhammadiyah Semarang, Kedungmundu,Tembalang, Semarang City, Central Java, Indonesia.

Abstract. Education is one of the important things in building a nation to create knowledgeable people. The better the quality of education of a nation, the better the quality of the nation. In Indonesia, the level of education quality is prioritized because education has a very important role in improving the quality of human resources. In measuring the level of education quality, the components analyzed include elementary school dropout rates, junior high school dropout rates, high school dropout rates and an average length of schooling in a district/city. This study has a purpose to clustering the level ofeducation quality from 35 districts/cities in Central Java Province based on 3 categories. Therefore, researchers use the c-means method which is one of the methods in statistics that used to group data into several predetermined groups. From the result of the analysis, it was found that group 1 was a district /city with a good quality level of education consisting of 13 districts/cities, group 2 was a district/city level with a fairly good quality of education consisting of 9 districts/cities, and group 3 was a district/city with a poor quality level of education consisting of 13 districts/cities.

Keywords:C-means, Cluster Analysis, Central Java, City, District, EducationQuality

1. Introduction

Education is one of the important things in the development of the Indonesian nation to create people who are knowledgeable, pious and cultured to face challenges in the future that are so great. With education can create smart and skilled students in the community. Education is also one of the foundations in the progress of a nation, the better the quality of education held by a nation, the better the quality of the nation will be followed. In Indonesia the quality of education is prioritized, improving the quality of education is a major factor determining the success of nation building. In the field of education, Indonesia is currently facing serious problems such as a lack of adequate learning systems in schools and the quality of education which is still low because of the high dropout rates.

There are several methods in cluster analysis, but in general cluster analysis methods that are often used are hierarchical methods and non-hierarchical methods or c-means. The hierarchical method is used when information on the number of clusters is not known while in c-means it is used to group data into clusters where the number of clusters has been determined or previously known.

Based on this background, researchers will analyze the level of quality of education in districts/cities in Central Java with a grouping method to find out which districts/cities are included in the group with good, good enough, and not good quality education. In grouping the level of education quality, the components analyzed include elementary school dropout rates, junior high school dropout rates, high school dropout rates and average length of schooling in each district/city. This study aims to classify the quality of education from 35 districts /cities in Central Java Province based on 3 categories. Therefore, researchers use the c-means method which is one of the methods in statistics used to group data into several predetermined groups.

Clustering is a method in data mining where data that has the most similar characteristics will be grouped on the same cluster and data that has different characteristics will be grouped into different clusters [1] [10]. In the Clustering analysis system classifies each object where the most similar characteristics will be in the same cluster.There are two types of data clustering that are often used, namely hierarchical clustering and non-hierarchical clustering [2][3].Cluster is one of the statistical methods to divide several data sets into several groups or clusters based on the level of similarity [4]. This data grouping is based on the principle of maximizing data similarity and maximizing inequality in different clusters. Cluster analysis or commonly called clustering is the process of partitioning a set of

ICE 2019, September 28-30, Purwokerto, Indonesia

DOI 10.4108/eai.28-9-2019.2291056

data or observation objects into a subset [5].Agood cluster has several forms including homogeneous (Similarity), the similarity of internal characteristics of data in one cluster (Within-Cluster)and Heterogeneous (difference), the difference in external characteristics between clusters (Between-Clusters) [6].

2. Data and Method

Data. The data used in this study are 35 secondary data. The variables used are elementary school dropout rate, junior high school dropout rate, high school dropout rate and average length of schooling from each of the 35 districts/cities in Central Java.

Table 1. Variables And Data Scales

Variables	Description	Scales
X_1	Percentage of elementary school dropout rates	Ratio
X_2	Percentage of junior high school dropout rates	Ratio
X_3	Percentage of high school dropout rates	Ratio
X_4	average length of schooling	Ratio

Methods. In this study, data were processed using SPSS software. The method used to analyze the data is C-means. The C-means method or non-hierarchical method is a clustering method or grouping data into pre-determined clusters. In this method the data will be partitioned into several clusters where data that has similar characters will be grouped on the same cluster and data that has different characteristics will be grouped on different clusters.

1. **C-means,**a non-hierarchical method that is often used in grouping because it has the nature of convenience and can group large amounts of data with fast and efficient computing processes. Basically the c-means work system by looking at the distance between objects and the value of the centroid cluster [5]. The workings of the c-means method is in the c-means algorithm defining the centroid of the cluster as the average value of the points in the cluster[7] [8], the following steps:
   a. Randomly select and determine k as the number of clusters to be formed
   b. Generating a random value that is used for the centroid center of the initial cluster of k (number of clusters).
   c. Calculate the distance of each data against each centroid using the Euclidean Distance formula:

      $d_{ik} = \sqrt{\sum_{j=1}^{m}(Xij - Cij)^2}$

      where :
      Xij = i[th] of object *x*
      Cij = i[th]of object*y*
      M = objects
   d. Group each data based on the closest distance between data and centroid.
   e. Renew centroid value. The renewal of the centroid value is obtained from the cluster average value [9], using the formula:

      $C_{ij} = \frac{\sum_{j=1}^{n} Xij}{n}$

      Where :
      Xij =k^{th}*cluster*
      n = number of cluster members

2. **Cluster Validation**. ANOVA is a statistical method as a technique to test hypotheses to find out the relationship between variables. In ANOVA testing, the aim is to find out whether the intergroup variables include the perception[6]. One way ANOVA test results can determine which variables affect the cluster results.The following are calculations in one way ANOVA test.

Hypothesis: $H_0 : \tau 1 = 0$

$H_1 : \tau 1 \neq 0$

Test of Statistics : $F_{\text{Count}} = \dfrac{\sum_{t=1}^{T} {}_{\text{nt}}\left(\overline{X}_{\text{t}-}\overline{X}\right)^2 / (t-1)}{\sum_{t=1}^{T}\sum_{k=0}^{n}\left(\overline{X}_{\text{tj}-}\overline{X}\right)^2 / \sum_{t=1}^{T} nt-t}$

Testing Criteria: Reject H_0 if F_{Count}greather than dariF_{Table}

Table 2. *One-Way* ANOVA

Source of Variation	Sum Square	Degree of freedom (df)
Treatment	$\sum_{t=1}^{T} {}_{\text{nt}}\left(\overline{X}_{\text{t}-}\overline{X}\right)^2$	T-1
Error	$\sum_{t=1}^{T}\sum_{k=0}^{n}\left(\overline{X}_{\text{tj}-}\overline{X}_{\text{j}}\right)^2$	$\sum_{t=1}^{T} nt - T$
Total	$\sum_{t=1}^{T}\sum_{k=0}^{n}\left(\overline{X}_{\text{tj}-}\overline{X}\right)^2$	$\sum_{t=1}^{T} nt - 1$

3. Results and Discussion

Descriptive analysis is used to assist in describing research variables. The following are the results of the descriptive analysis presented in the following table:

Table 3. Descriptive Statistics

Variables	Minimum	Maximum	Mean	Std. Deviation
Elementary school dropout rates	0.00	33.39	6.67	8.01
Junior high school dropout rates	0.00	43.27	15.67	14.56
High school dropout rates	37.11	91.82	71.71	11.17
Average length of schooling	0.29	10.50	5.45	3.70

The number of data as much as 35 includes districts/cities in Central Java Province. The average dropout rate is 6.67% with a minimum grade of 0 and a maximum grade of 33.39%. And the average junior high school dropout rate is 15.67% with a minimum grade of 0 and a maximum grade of 43.27%. While the average high school dropout rate is 71.71% with a minimum grade of 37.11% and a maximum value of 91.82%.

Standardization of data is done to avoid significant differences in data units between the variables studied. Because there are some values of zero and the variable drop out rate elementary and junior high school. Therefore the standardization of data process is carried out first.C-means or often called non-hierarchy with the work system partitioning into several clusters. This method has several work steps includingdetermination of Initial Cluster (c).

Table 4.Initial Cluster Centers

Variables	Cluster		
	1	2	3
Zscore(Elementary school dropout rates)	-0.83313	3.33592	-0.81690

Zscore(Junior high school dropout rates)	-0.66134	0.22855	-1.01427
Zscore(High school dropout rates)	-3.09726	1.06345	1.14312
Zscore(Average length of schooling)	1.33419	0.77406	-1.39070

Based on standardized research data, the distance between centroids can be calculated. The calculation of centroid euclidean with SPSS, the results are obtained:

Table 5.Calculation Of Centroid Euclidean

Case Number	District/City	Cluster	Distance
1	CilacapDistrict	3	1.539
2	BanyumasDistrict	3	1.924
3	PurbalinggaDistrict	2	1.125
4	BanjarnegaraDistrict	2	1.027
5	KebumenDistrict	3	1.043
6	PurworejoDistrict	3	1.301
7	WonosoboDistrict	1	1.444
8	MagelangDistrict	1	1.013
9	BoyolaliDistrict	1	0.924
10	KlatenDistrict	3	1.489
...	...	...	...
32	SalatigaCity	1	1.449
33	SemarangCity	1	1.532
34	PekalonganCity	2	0.561
35	TegalCity	1	0.906

On table 5 each district/city data will be a member of the cluster by having the closest distance from the center of the cluster, so that the following cluster members are obtained:

Table 1.Cluster Membership

Cluster	Membership
Cluster 1	Wonosobo District, Magelang District, Boyolali District, Jepara District, Pekalongan District, PemalangDistrict, Tegal District, Brebes District, Magelang City, Salatiga City, Semarang City, Tegal City
Cluster 2	Purbalingga District, Banjarnegara District, Grobogan District, Rembang District, Pati District, Kudus District, Batang District, Surakarta City, Pekalongan City
Cluster 3	Cilacap District, Banyumas District, Kebumen District, Purworejo District, Klaten District, Sukoharjo District, Wonogiri District, Sragen District, Blora District, Demak District, Semarang District, Temanggung District, Kendal District

The results of clustering using euclidean, it was found that in the division into three clusters, cluster 1 had 13 members, cluster 2 had 9 members, and cluster 3 had 13 members.In addition to providing specific characteristics in describing the contents of the formed clusters can be seen in the output of the SPSS software obtained as follows:

Table 7.Cluster Specifications

Variables	Cluster		
	1	2	3
Zscore(Elementary school dropout rates)	-0.58486	1.35621	-0.35406
Zscore(Junior high school dropout rates)	0.43766	0.53518	-0.80817

Zscore(High school dropout rates)	-0.57772	0.17575	0.45605
Zscore(Average length of schooling)	0.68268	0.37358	-0.94131

The average value of the variables in each cluster can be seen the characteristics of the three clusters formed with the following interpretation:

1. Cluster 1 in the Primary and Middle School Drop Out Rate variables has the lowest value than the other clusters, while in the average length of school variable in cluster 1 has the highest value compared to other clusters.
2. Cluster 2 on the variable Dropout Rate Elementary School has the highest value compared to other clusters.
3. Cluster 3 on the variable Middle School Dropout Rate and the average length of school has the lowest value of the other clusters, while the High School Dropout Rate variable has the highest value compared to other clusters.

There are three districts/cities clusters in Central Java Province with four variables. Cluster validation is used to determine whether the three clusters are different from each other, in testing using the ANOVA test with SPSS software.

Table 8. ANOVA Test

Variables	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Zscore(Elementary school dropout rates)	11.315	2	0.355	32	31.845	0,000
Zscore(Junior high school dropout rates)	6.779	2	0.639	32	10.613	0,000
Zscore(High school dropout rates)	3.660	2	0.834	32	4.390	0,021
Zscore(Average length of schooling)	9.417	2	0.474	32	19.868	0,000

Based on the results of ANOVA test, the value of each variable is at the level (Sig) <0.05, which shows that in the three districts/cities clusters there are significant differences in characteristics between clusters related to variables of Primary Dropout Rate, Middle Dropout Rate, High School Dropout Rate and Average Length of School.

4. Conclusions

After analyzing and discussing the grouping of districts/cities in Central Java Province based on the level of quality of education, the following conclusions can be obtained, there are three clusters obtained in the analysis conducted, based on the results of the discussion obtained the division of clusters as follows:

1. Cluster 1 consists of 13 members namely Wonosobo District, Magelang District, Boyolali District, Jepara District, Pekalongan District, PemalangDistrict, Tegal District, Brebes District, Magelang City, Salatiga City, Semarang City, Tegal City. In this cluster are categorized as districts/cities with good quality education.
2. Cluster 2 consists of 9 members namely Purbalingga District, Banjarnegara District, Grobogan District, Rembang District, Pati District, Kudus District, Batang District, Surakarta City, Pekalongan City. In this cluster are categorized as districts / cities with a fairly good level of quality education.
3. Cluster 3 consists of 13 members namely Cilacap District, Banyumas District, Kebumen District, Purworejo District, Klaten District, Sukoharjo District, Wonogiri District, Sragen District, Blora District, Demak District, Semarang District, Temanggung District, Kendal District. In this cluster are categorized as districts/cities with poor quality education.

References

[1] Hair JF Jr, Anderson RE, Tatham RL, Black WC. 1995. Multivariate Data Analysis 4thEdition. New Jersey : Prentice Hall

[2] Johnson RA, Winchern DW. 1998. AppliedMultivariate Statistical Analisys 4thEdition. London : Prentice-Hall.

[3] Kaufman L, Rousseeuw PJ. 1990. FindingGroups in Data: An Introduction toCluster Analysis. New York : John Wiley.

[4] Darmi, Y., & Setiawan, A. (2016). Penerapan Metode Clustering K-Means Dalam. Y. Darmi, A. Setiawan, 12(2), 148–157.

[5] Han, J. (2011). Data Mining: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems). Retrieved from http://www.amazon.co.uk/Data-Mining-Concepts-Techniques-Management/dp/0123814790

[6] Seer, N. C. I.Analisis Clustering Untuk Mengelompokan Tingkat Kesejahteraan Kabupaten Kota Berdasarkan Sosial Ekonomi Rumah Tangga di Wilayah Provinsi Sulawesi Selatan., Dds, D. D. L., Affairs, R., Affairs, R., Except, M., … Hofferkamp, J. (2018). https://doi.org/.1037//0033-2909.I26.1.78

[7] Skripsi, B., Multimedia, M., Teknik, P., Dan, I., Universitas, K., & Jakarta, N. (2018). Implementasi algoritma k-means clustering untuk mengetahui bidang skripsi mahasiswa multimedia pendidikan teknik informatika dan komputer universitas negeri jakarta. (December 2017). https://doi.org/10.21009/pinter.1.2.10

[8] Tan, M. Steinbach, and V. Kumar, (2005) Introduction to Data Mining, (First Edition). Boston, MA, USA: Addison-Wesley Longman Publishing CO., Inc.

[9] Maimon, O., &Rokach, L. (Eds.). (2005). Data mining and knowledge discoveryhandbook (Vol. 2). New York: Springer.

[10] Yu, H., Fan, J., & La, R. (2019). Suppressed possibilistic c-means clustering algorithm. Applied Soft Computing Journal. https://doi.org/10.1016/j.asoc.2019.02.027