Cyberspace: Jurnal Pendidikan Teknologi informasi
Volume 7, Nomor 2, Oktober 2023, hal. 118 - 125
ISSN 2598-2079 (print) | ISSN 2597-9671 (online)

PRINCIPAL COMPONENT K-MEANS SOFT CONSTRAINT BASED ON WELL-BEING INDICATORS IN ACEH PROVINCE

Winny Dian Safitri

Department of Economics, Faculty of Islamic Economics and Business, Ar-Raniry State Islamic University, Banda Aceh, 23111
E-mail: winny.diansafitri@ar-raniry.ac.id

Abstract

The success of government policies can be from the state of the well- being indicators. This research was conducted to obtain district/city groupings based on the similarity of characteristics of the well-being indicators of each district/city in Aceh Province in 2022. The data used in the Aceh well-being indicator data for 2022 consists of 29 variables. The analysis method used is the principal component k-means soft constrain method. The background information data can be used as a provision to streamline the clustering algorithm by creating soft constraints which is found as the most appropriate algorithm. The results of this study indicate there are four district/city clusters in Aceh Province. The characteristics of the first cluster are that kindergarten and elementary school facilities are adequate, while the school enrollment rate needs to be improved. The characteristics of the second cluster are superior to the Gross Enrollment Rate (GER) and the population of university graduates, but still very lacking in school facilities. The third cluster is the cluster that is the center of well-being in Aceh, so this cluster is the cluster with the best well-being level. The characteristic of the fourth cluster is that it is very good in the school participation rate indicator, but it must increase early childhood school participation.

Keywords: *Well-being, Clustering, Principal Component, K-means, Constrain*

Abstrak

Keberhasilan kebijakan pemerintah dapat digambarkan dari keberhasilan pembangunan yang tergambarkan dari kesejahteraan masyarakat. Penelitian ini dilakukan untuk memperoleh pengelompokan kabupaten/kota berdasarkan kesamaan karakteristik indikator kesejahteraan khususnya sektor pendidikan masing-masing kabupaten/kota di Provinsi Aceh pada tahun 2022. Data yang digunakan dalam penelitian ini yaitu data indikator kesejahteraan Aceh tahun 2022 terdiri dari 29 variabel. Metode analisis yang digunakan adalah analisis komponen utama K-means dengan batasan. Data informasi latar belakang dapat digunakan sebagai bekal untuk memproses algoritma clustering dengan membuat soft constraint yang ditemukan sebagai algoritma yang paling tepat. Hasil penelitian ini menunjukkan terdapat empat klaster kabupaten/kota di Provinsi Aceh. Karakteristik klaster pertama adalah fasilitas TK dan SD memadai, sedangkan tingkat pendaftaran sekolah perlu ditingkatkan. Karakteristik klaster kedua lebih unggul dibandingkan *Gross Enrollment Rate* (GER) dan populasi lulusan universitas, namun masih sangat kurang di fasilitas sekolah. Klaster ketiga adalah klaster yang menjadi pusat kesejhateraan di Aceh, sehingga klaster ini merupakan klaster dengan jenjang kesejahteraan terbaik. Karakteristik klaster keempat adalah sangat baik dalam

indikator tingkat partisipasi sekolah, tetapi harus meningkatkan partisipasi sekolah anak usia dini.

Kata Kunci: *Kesejahteraan, Klaster, Komponen Utama, K-mean, Batasan*

1. Introduction

Prosperity is a dream for all countries of the world. In the eyes of a country is said to be prosperous, if the education is good. Education is one aspect that becomes the benchmark in development success. In the education system, there is a transfer of knowledge between humans. According to Law No. 20 of 2003, the definition of education is an effort and planning to create a controlled teaching and learning process by including spiritual abilities to create independent, moral, and knowledgeable humans. Education will be successful if it has a transparent system and facilities that are met.

The United Nations Development Programs (UNDP) since 1990 have started to issue annual reports on human development in various countries, namely the Human Development Report. Some of the approaches used in measuring poverty levels include proper education. Several developed countries, such as Indonesia, should already have a sound education system. The Indonesian Central Bureau of Statistics formulates education indicators, which are present the development of Indonesian education over time and compare provinces and areas of residence.

Education in Indonesia, Aceh Province, is notably still inferior, from readiness to availability of facilities. Based on the National Examination results issued by the Ministry of Education and Culture of the Republic of Indonesia, Aceh is ranked 27th out of 34 Provinces in Indonesia. It very concerns about the education world in Aceh.

The government's efforts to improve the education system are not yet optimal, so it is necessary to have an education strategy in studies with existing data. The study results were included in policymaking to achieve an intelligent community by the "Aceh Carong" program promoted by the Aceh government to know the obstacles in each district/city in the education system with different levels of difficulty.

Research related to factors that influence education, including planning, regulations, human resources, technical, coordination, and procurement of goods and services, affect the realization of the education budget in Aceh Province. The factors that influence the success of district/city education in Aceh must be different, given the region's geographic location and the various cultures, so it is necessary to have an analysis that can identify each region. One of the statistical methods used is cluster analysis with the primary condition that there must be no relationship between variables (multicollinearity).

When analyzing the data, there were several problems, including a high correlation between variables. This high correlation can cause a high bias value from the analysis results. Principal Component Analysis is a solution to overcome high correlation problems between variables by reducing the dimensions of large variables. The hope is to provide more accurate results from clusters. The use of PCA in k-means to reduce the high dimensional data showed that the PCA can be able to be a solution to produce the better grouping result by using k-means.

The other problem that could occur when analyzing the data is missing value. In order to continue the analysis, the missing values must be addressed first. One of the cluster analysis methods that can handle the problem of missing value without imputation is K-means Soft Constraint method. Hence, the Principal Component Analysis K-means Soft Constraint method is used in this research.

2. Literature Review

Principal Component Analysis

The principal component analysis is a reasonably good method for obtaining estimator coefficients in regression equations with multicollinearity problems. The independent variable in principal component regression is a linear combination of the original Z variable called the principal component. This method's estimation coefficient is obtained from the shrinkage of the main component dimensions, with the subset of the main components selected having to maintain the most remarkable diversity.

The method of eliminating the principal component starts from the procedure for selecting the root feature of an equation:

$$| AX - \lambda I | = 0 \quad (1)$$

If the root of the feature λ_j is ordered from the most massive value to the smallest value, then the effect of the main component W_j corresponds to λ_j. These components explain the proportion of diversity to the dependent variable Y, which is getting smaller and smaller. The main components of W_j are orthogonal to each other and are formed through a relationship:

$$W_j = v_{1j} Z_1 + v_{2j} Z_2 + v_{3j} Z_3 + \ldots + v_{p\,j} Z_p \quad (2)$$

where p is the number of variables used. The vector v_j is obtained from each feature root λ_j which satisfies a homogeneous system of equations:

$$| AX - \lambda_j I | v_j = 0 \quad (3)$$

where $v_j = (v_{1j}, v_{2j}, v_{3j}, \ldots, v_{pj})$.

There are three methods commonly used to determine the number of main components, namely:

1. If the number of main components produced is q where $q \leq p$, then what has been transformed (main component score data) has as many variables as q. Suppose the proportion for the root of trait i^{th} is:

$$\frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i} \quad (4)$$

   The determination of the number of principal components (q) is based on the cumulative proportion of its characteristic roots. If the cumulative proportion of q, the first feature root is 80% or more, then the number of principal components is q [4].
2. The main components' selection is based on the variety of the main components, which are none other than the root features. According to, after doing a study, the better cut off is 0.7.
3. The scree plot is a plot between the root features λ_k and k. Using this plot, the number of principal components selected is k. If at that point k, the plot is steep to the left but not steep to the right.

K-means Soft Constraints

The background information data can be used as a provision to streamline the clustering algorithm by creating soft constraints. Soft constraints are a function made as initial information from members of a group. The use of constraint becomes important for several clustering algorithms. Several researchers have shown that constraints can improve the results of a variety of clustering algorithms. Thus, the soft constraint addition in k-means algorithm becomes beneficial.

K-means soft constraint is a developed k-means algorithm that has robustness for grouping a set of data without any imputation process required. Missing value issue on a dataset may limit the use of clustering methods. Hence, a method that resistance for

clustering the dataset which contains missing values is required. It's obtained that k-means soft constraint on dataset Glass, Wine, Iris, and Breast Cancer outperformed k-means for all datasets. Beside the robustness for dealing with missing value, k-means soft constraint also showed its performance for dealing with dataset that contains multicollinearity. It's found that the use of k-means soft constraint results in high accuracy on dataset contains multicollinearity.

It is explained that k-means clustering is an algorithm that is most appropriate for grouping with soft constraints, so that soft constraints are used as information in the grouping. K-means with soft constraints is done by dividing the data set into two parts, namely the set with complete data variables (F_o) and the set with incomplete data variables (F_m). Suppose *sc* is the symbol for soft constraints, F_m is the set of the incomplete data variable, x_{im} is the item of the i^{th} object of the incomplete data variable *m*, x_{jm} is the item of the jth object of the missing data variable *m*, *f* is an incomplete variable member. The soft constraints of x_{im} and x_{jm} are:

$$sc = -\sqrt{\sum_{f \in F_m} (x_{im} f - x_{jm} f)^2} \quad (5)$$

where *sc* is always negative. It indicates that one object has different groups. The k-means soft constraints algorithm adopts the steps of the k-means algorithm in dividing k objects into c suitable groups. The stages of the k-means soft constraints algorithm are:

1. Determine the center of the c^{th} band.
2. Determine the member of the c^{th} band by calculating the minimum distance of an object to the k^{th} band to the center of the c^{th} band

$$C = \underset{c_c}{\arg\min} \left((1-w) \frac{v}{v_{\max}} + w \frac{cv}{cv_{\max}} \right) \quad (6)$$

by calculating the distance from the k^{th} object to the c^{th} center of the complete data variable *d* is as follows:

$$v = \sum_{d=1}^{D} (x_{kd} - c_{cd})^2 \quad (7)$$

where:

c_{cd}= center of the c-th band based on the-d variable.

w = weighting factor determined subjectively by value w ∈ [0,1], in this study using the value of w = 0.5.

v_{max}= the maximum distance from all objects to the center of the cluster on the complete data variable.

cv = sum of the squares of soft constraints containing the value of sc

cv_{max}= sum of squares of all soft constraints

3. Repeat steps 1 to 2 through $\max_{c,d} \left(| c_{cd}^{(r)} - c_{cd}^{(r-1)} | \leq 10^{-4} \right)$ (8)

3. Material and Method

The data used in this study is secondary data, namely well-being indicators based education statistics data for 2022 sourced from the Central Statistics Agency of Aceh Province. The software that used to conduct the analysis is R version 3.4.1. The research variables consisted of indicators of the participation rate of children aged 3-6 years in the Early Childhood Education program. It consisted of 3 variables, the School Participation Rate, which consisted of 3 variables, the Gross Enrollment Rate, which consisted of 3

variables, the net enrollment rate, which consists of 3 variables, the percentage of the population ten years and over is detailed according to the highest diploma held which consists of 6 variables and the number of schools consisting of 11 variables.

The step carried out in this research is started by presenting the overview of Aceh literacy rate. The next step is the implementation of PCA. This is aimed to reduce the dimensions of 29 variables into the smaller dimensions. The result of the dimension reduction is then grouped by using cluster analysis namely k-means soft constraints. The clusters obtained were identified based on the characteristics of each cluster.

4. Result and Discussion

Aceh's well-being-based education indicators have changed significantly from year to year. Along with these changes, an analysis of the education indicator variables in Aceh will be carried out and classifying the districts in Aceh Province based on their similar characteristics.

From the results of the correlation analysis, if it's looked at the correlation in the correlation matrix R measuring p x p = 29 x 29 (p is the number of observed variables), there are several high correlations between independent variables that indicate multicollinearity, which may be due to different units of measurement. The correlation matrix is shown in Figure 1.

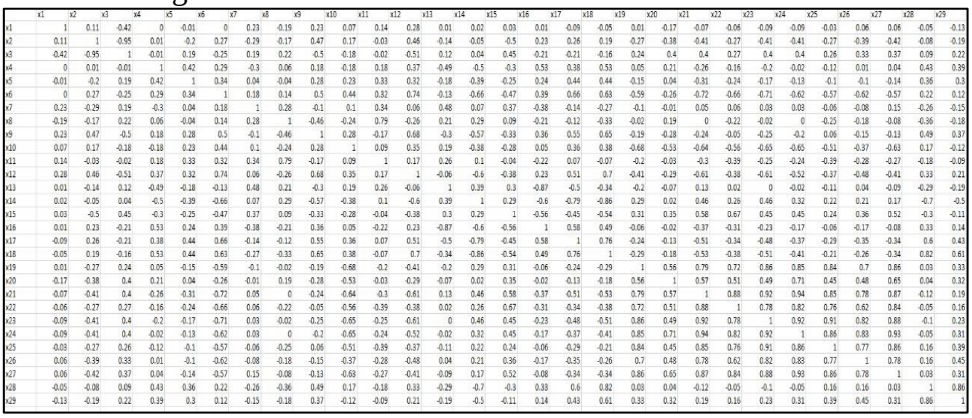

Figure 1. Correlation matrix between variables

Multicollinearity can be overcome by principal component analysis by first standardizing the X variables into Z variables and selecting the component dimensions, which must have a cumulative diversity of more than 70 percent.

Table 1. Main component selection

Dimension	Eigen value	Variance percent	Cumulative variance percent
Dim.1	10.55463796	36.3953033	36.3953
Dim.2	5.496869248	18.95472154	55.35002
Dim.3	2.847149847	9.817758094	65.16778
Dim.4	2.131925502	7.351467249	72.51925
Dim.5	1.787848096	6.164993436	78.68424
Dim.6	1.164694421	4.01618766	82.70043
Dim.7	1.036020135	3.572483225	86.27291
Dim.8	0.827721401	2.854211729	89.12713

Grouping objects using the grouping method. Principal Component Analysis of K-Means Soft Constraint is to see the distance between objects, but the initial process is to deal with multicollinearity problems first. If the distance value for each object is small, it will be grouped into one cluster. The following are the district/city clusters' results in Aceh based on the 2022 education indicators.

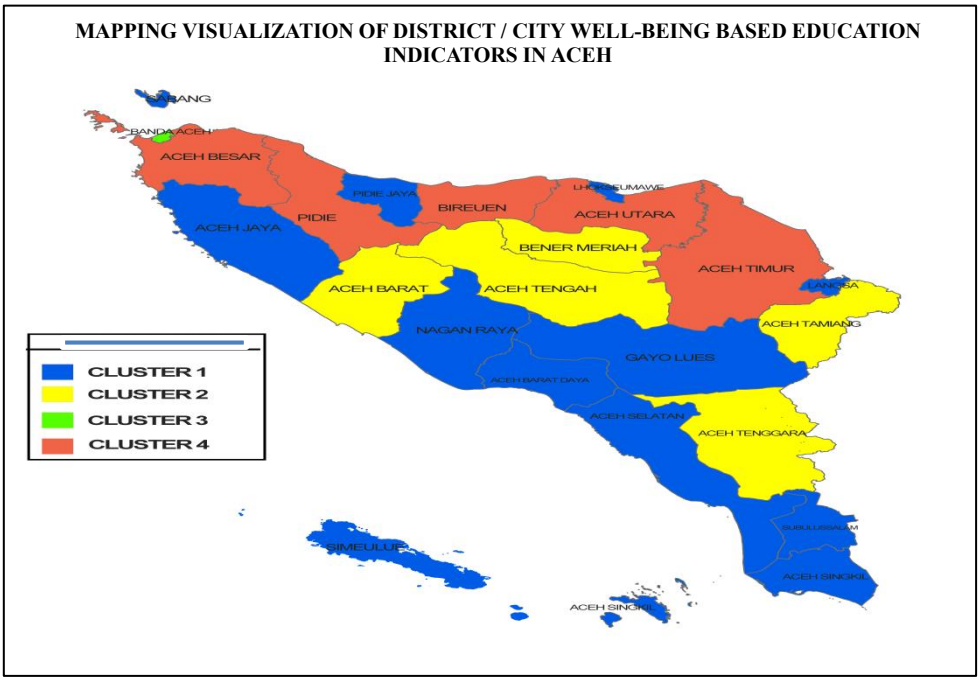

Figure 2. Results of the Aceh education cluster in 2022

Based on Figure 2, it was obtained four district/city clusters in Aceh Province based on well-being-based education indicator data for 2022, namely:

1. Members of the first cluster are Aceh Jaya, Pidie Jaya, Nagan Raya, Aceh Barat Daya, Aceh Selatan, Gayo Lues, Subulussalam, Aceh Singkil, Sabang, Lhokseumawe, Langsa, and Simeulue. The first cluster's characteristics are that kindergarten and elementary school facilities are adequate, while the school enrollment rate needs improvement.
2. Members of the second cluster are Aceh Barat, Aceh Tengah, Bener Meriah, Aceh Tamiang, and Aceh Tenggara. The second cluster characteristics are superior in the Gross Enrollment Rate and university graduates' population, but still lacking in school facilities.
3. Members of the third cluster, namely Banda Aceh. Banda Aceh is the center of education in Aceh Province. The third cluster characteristic is the cluster that is the center of education in Aceh, so that this cluster is the cluster with the best education level.
4. Members of the fourth cluster are Aceh Besar, Pidie, Bireuen, Aceh Utara, and Aceh Timur. The fourth cluster characteristic is that it is very good in the School Participation Rate indicator, but it must increase early childhood school participation.

From the cluster above, it can be seen that there should be more programs to improve the quality and quality of teaching staff as well as the distribution of teachers with national education standards evenly in every district/city in Aceh. These efforts were made

primarily to reduce the gap in education level and quality between districts/cities in Aceh. The level of disparities in district/city education development in Aceh must be minimized so that every Acehnese has the same opportunity to get proper education up to the highest level. Expectations from the high level of public education will automatically increase the standard of living so that just and equitable welfare can be realized in people's lives. However, the initial process is to deal with multicollinearity problems first. If the distance value for each object is small, it will be grouped into one cluster. The following are the district/city clusters' results in Aceh based on the 2022 education indicators.

5. Conclusion

Based on the results of the clustering of districts/cities in Aceh Province using the Principal Component K-Means Soft Constraint method, it shows that the members of each cluster are strongly influenced by geographic location, the proximity between districts/cities in Aceh so that the similar characteristics of the Aceh education indicators in 2022 are formed. The results of this study indicate there are four district/city clusters in Aceh Province. The first cluster's characteristics are that the kindergarten and elementary school facilities are adequate, while the school participation rate needs to be improved. The second cluster's characteristics are superior in the Gross Enrollment Rate and the population of university graduates, but still lacking in school facilities. Characteristics of the third cluster is the cluster that is the center of well-being based education in Aceh so that this cluster is the cluster with the best education level. The fourth cluster's characteristic is that it is very good in the School Participation Rate indicator, but it must increase early childhood school participation.

References

1. Aini E N, Isnaini I, Sukamti S & Amalia L N. 2018. Pengaruh Tingkat Pendidikan Terhadap Tingkat Kesejahteraan Masyarakat di Kelurahan Kesatrian Kota Malang. *Tecnomedia Journal* 3(1) 1-15.
2. BPS. 2022. Aceh Dalam Angka 2022. *Publication Of BPS Provinsi Aceh*.
3. Ding, C dan He, X. 2004. K -means Clustering via Principal Component Analysis. *ICML 2004: Proceedings of the Twenty-First International Conference on Machine Learning*.
4. Johnson R, A & Wichern D, W. 2014. *Applied Multivariate Statistical Analisys.* New Jersey, Prentice – Hall International.
5. Jolliffe I ,T. 2016. *Principal Component Analysis.* New York: Springer.
6. Majid M S, A. 2014. Analisis Tingkat Pendidikan dan Kemiskinan di Aceh. *Jurnal Pencerahan, vol 8 no 1. Jurnal Pencerahan,* **8**(1) 15-27.
7. MacQueen J, B. 1967. Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of the Fifth Symposium on Math, Statistics and Probability: 281–297.* University of California Press.
8. Mesquita, D, Gomes, J dan Rodrigues, L. (2016). K-means for Datasets with MissingAttributes Building Soft Constraints with Observed and Imputed Values. *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, Bruges.
9. Ruwaida, Darwanis, S.Abdullah. 2015.Faktor-faktor yang Mempengaruhi Realisasi Belanja Pendidikan di Provinsi Aceh. *Jurnal Akutansi, vol 4 no 4*.

10. Safitri W D, Nurhasanah and Rusyana A. 2012. *Pengelompokan Kabupaten/ kota di Provinsi Aceh Berdasarkan Tingkat Perubahan Kesejahteraan Rakyat*. Skripsi Universitas Syiah Kuala, Banda Aceh.
11. Safitri W D. Sartono B dan Erfiani. 2016. *Kajian Penggerombolan Data Tidak Lengkap dengan Algoritma Khusus Tanpa Imputasi.* Tesis. Institut Pertanian Bogor, Bogor.
12. Sambandam, R. 2003. *Collinearity Is a Natural Problem in Clustering*. *15*(1), 16–21.
13. Wagstaff K, Cardie C, Rogers S & Schroed S. 2001. Constrained K-means Clustering With Background Knowledge. *Proceedings. of the 18th Intl. Conf. on Machine Learning : 577–584.* USA.
14. Wagstaff, K. 2004. Clustering with Missing Values: No Imputation Required. *Proceedings of the Meeting of the International Federation of Classification Societies : 649–658.* California.
15. Wagstaff, K. L., Basu, S. dan Davidson, I. 2006. *When Is Constrained Clustering Beneficial, And Why?* Ionosphere, 58(60.1), 62-63.