Education as an important aspect to determine human development index by province in Indonesia

S Zahroh\textsuperscript{1} and R S Pontoh\textsuperscript{2*}

\textsuperscript{1}Basmalah Corporation, Krisik, Jatinunggal, Sumedang 45376, Indonesia
\textsuperscript{2}Department of Statistics, Universitas Padjadjaran, Jl Raya Bandung-Sumedang Km. 21, Sumedang 45363, Indonesia

*resa.septiani@unpad.ac.id

Abstract. The Human Development Index (HDI) is a comparative measure of life expectancy, education, and standard of living in all countries over the world to classify whether a country is a developed country, a developing country, or even an underdeveloped country. Currently, science and technology are developing rapidly. In order to compete with other countries, Indonesia needs to improve the quality of education, which is one of the dimensions of HDI. In this study, the HDI in Indonesia and several aspects of education are grouped by provinces then it is compared whether the cluster of HDI formed has the same members as the cluster of several variables related to education in Indonesia. The goal is to observe if the quality of education in Indonesia might represent HDI. The clustering method that will be used is k-means clustering with average silhouette method to determine the optimal number of cluster. The results show different cluster in two groups of dataset means that the high index of human development does not determine the high quality of education in Indonesia. It indicates that the assessment of the quality of education needs to be continued both for regions with high HDI conditions and vice versa because the quality of education does not appear to be reflected in the high HDI.

1. Introduction

Development of a country is not only seen in terms of per capita income, but also includes various aspects of society such as social affairs, politics, economy, security, and law. The development of the economy as a process of increasing per capita income goes hand in hand with developments in the institutional system that grows with the changes. When mapping economic development, it is therefore not sufficient to talk only about GDP growth per capita \cite{1}. Sarkar et al. \cite{2} found that the development of human is one of the main deliberations in determining the development level of a country.

The United Nations Development Program (UNDP) introduced a Human Development Index (HDI) in 1990. HDI is a development indicator with which the comparative status of the socio-economic development of a country can be analyzed and at the same time, the human development in a country can be described \cite{3}. One of the goals of HDI is to point out the citizens and their skills should be the main factors, along with economic growth, to describe a country's development. In addition, HDI is able to assess life expectancy, education and standard of living in all countries in order to characterize them as industrialized, developing or underdeveloped countries \cite{4}. Three dimensions as the basics of calculation of this index are health dimension, education dimension, and standard of living dimension.

Education is central to development as the aim is to improve the quality of human resources \cite{5}. Not only material and physical development but also spiritual development, such as human development, which is the main task of education. In order to improve the quality of human resources, high-quality education is required because education is one of the main characteristics of development that increases
the quality of human resources [6]. Consequently, education leads to people to produce human resources that support development, and development outcomes can support education. Hence, education and development of a nation therefore have a mutually supportive relationship [7].

The education and training of the workforce are essential for sustainable economic growth. With the support of health and education, everything is very possible. If the health of children improves with proper nutrition, moreover, a healthy child might improve their academic performance [8]. A decrease in absenteeism in school, a decrease in the school dropout rate, an increase in children's learning ability, and an increase in children's academic performance are instances of education indicators to determine how good they are. Good education certainly affects labour productivity, thereby, not only knowledge need to be improved but also the importance of family planning that has an effect on lowering birth rates. Overall, this offers advantages for the development process, with the aim of productive and high-quality generation can be achieved [9].

As we know, the current level of education in Indonesia has not fully enlightened society through the standards and prosperity of education itself. Confirmed that schooling in Indonesia has not developed ideal is the quality of the graduates [10]. The relevance of education to the needs of the individual is still viewed as low. Moreover, education is often used as a politician for local officials. The formation of the Indonesian nation, which is increasingly lagging behind compared to other nations, should motivate us to improve immediately. The number of educational problems that arise is a depiction of our educational quality [11].

As long as the education is assumed as one of the important things to calculate the HDI, in this study, we tried to make clusters of HDI in Indonesia by province and compared to the clusters of several criteria that define the education rate of each province. Furthermore, it can be expected that education is an important thing to calculate the HDI if they are in the same cluster as the cluster of HDI’s value. Therefore, the provinces belonging to the group with low HDI scores need special attention from the central government for quality improvement of education in these areas. The method that will be used is the K-means clustering because this is the most renowned method for clustering as unsupervised learning [12, 13]. It is common used because of its ability to group large data in a short time [14].

2. Method

2.1. Material

The variables that will be used for make clusters of HDI are the expected length of the school, adjusted per capita expenditure, the average length of the school, life expectancy at birth of each province in Indonesia in the year of 2019. Moreover, variables that will be used to make clusters for the education criteria are the participation rate of children in early childhood education, the rate of school enrollment in universities, the proportion of the population aged 15 to 24 and 15-59 with information and communication technology skills, the average length of school for the population in old age of 15 years, the ratio of the gross enrollment rate for women and men at the university and the net enrollment rate for women and men in each educational level. All data is collected from the official website of the Central Agency of Statistics (BPS) that published for the year 2019.

2.2. Method

Data mining is an interdisciplinary field of computer science. It is a technique for data analysis to discover the undetected relationship among data items [15]. Clustering is a comprehensive technique for finding observation groups in a data set. When observations are grouped, observations in the same group are expected to have similar characteristics and observations that have different characteristics supposed to be indifferent groups. Clustering is an unsupervised method because there is no response variable [16]. Therefore, this method tries to find the relationship between n observations without being trained on the response and allow us to determine which observations are similar and possibly classify them within the same group. One of the simple popular methods to characterize the set of data into k groups is K-means Clustering.

It provides classification by involving the distance between paired observations. The decision of this calculation is identified as a matrix of distance. The distance information, which defines how the
similarity of two elements \((x, y)\) is calculated and it affects the form of the cluster. In this study, Euclidean distance will be used as the Eq. 1.

\[
d_{\text{euc}}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}
\]  

(1)

K-means clustering which will be involved in this study consists of stating clusters, thereby, the total within-cluster variation is minimized. Various methods can be utilized to determine the optimal number of clusters, such as Silhouette method, Elbow method, and Gap Statistic method. Then, the Average Silhouette method is applied in this study. The average silhouette approach is used to determine the quality of the cluster, for that reason, the method can measure how well the object is in the cluster. A cluster with high average silhouette width is good. The optimal number of clusters is obtained by maximizing the average silhouette over a possible range of values for \(k\) [17]. Hartigan-Wong algorithm [18] explains the total within-cluster variation as the sum of squared Euclidean distances between corresponding centroid and items as Eq. 2.

\[
W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2
\]  

(2)

where \(x_i\) is a data units be appropriate to the cluster \(C_k\) and \(\mu_k\) is the average value of the units appointed to the cluster \(C_k\). The total within-cluster variation that determines the goodness of clustering follows the Eq. 3 and expected to be as small as possible.

\[
\sum_{k=1}^{K} W(C_k) = \sum_{k=1}^{K} \sum_{x_i \in C_k} (x_i - \mu_k)^2
\]  

(3)

In general, the K-means algorithm can be elaborated as this following steps [19]:

1. Determine the number of clusters to create \((K)\)
2. Randomly choose \(k\) objects from the dataset to be the centre of the initial cluster
3. Allocates each observation to its closest centroid according to the Euclidean distance between the object and the centroid
4. For each \(k\) cluster, apprise the cluster centre by computing the new average of all values in each cluster. A vector of length \(p\) that contains the averages of all variables for the values in the \(k^{th}\) cluster is the centroid of the \(K^{th}\) cluster while \(p\) is the number of variables.
5. Minimize repeatedly the Eq. 3. Repeat steps 3 and 4 until the cluster position no longer changes or the maximum number of repetitions is attained.

3. Result and Discussion

This study creates clusters of various variables used to determine the quality of education in Indonesia using the K-means clustering method. All analyzes using software R. Before conducting analysis using the clustering method, the first step is to determine the method of calculating the distance, using the Euclidean distance in this study. The distance matrix for each province in Indonesia is shown in Figure 1.

Figure 1 shows which provinces are very different (darker) and which appear very similar (lighter). Indonesia consists of 34 provinces marked with numbers 1-34 in the following order: Aceh, North Sumatra, West Sumatra, Riau, Jambi, South Sumatra, Bengkulu, Lampung, Bangka Belitung, Riau Islands, DKI Jakarta, West Java, Central Java, Yogyakarta, East Java, Banten, Bali, West Nusa Tenggara, East Nusa Tenggara, West Kalimantan, Central Kalimantan, South Kalimantan, East Kalimantan, North Kalimantan, North Sulawesi, Central Sulawesi, South Sulawesi, Southeast Sulawesi, Gorontalo, Sulawesi West, Maluku, North Maluku, West Papua and Papua.
Based on the calculation results, the optimal number of clusters for the education variable is 4 clusters, while the optimal cluster for the IPM variable is only 2 clusters. The optimal cluster diagram is shown in Figure 2.

Furthermore, for starters, based on the previous variables, clustering for education variables was performed with many clusters defined in 4 clusters. K-means clustering produced 4 clusters of sizes 10 in the first cluster, sizes 16 in the second cluster, size 1 in the third cluster, and sizes 7 in the fourth cluster. Members of the first cluster are Riau Island, Special Capital District of Jakarta, West Java, Central Java, East Java, Banten, Bali, West Nusa Tenggara, Central Kalimantan, and Gorontalo. Then, Aceh, North Sumatera, West Sumatera, Riau, Jambi, South Sumatera, Bengkulu, Lampung, Bangka Belitung Island, West Kalimantan, South Kalimantan, East Kalimantan, North Kalimantan, North Sulawesi, South Sulawesi, and Southeast Sulawesi are the members of the third cluster. Special Region of Yogyakarta became the only province in the third cluster. The last cluster consists of East Nusa
Tenggara, Central Sulawesi, West Sulawesi, Maluku, North Maluku, West Papua, Papua. The plot of this grouping can be seen in Figure 3 (the provinces symbolized by numbers).

Figure 3. The cluster plot of education variables

Thereafter, clustering analysis is conducted for the HDI variables. By use the method to find the optimal number of clusters, K-means clustering with 2 clusters of sizes 8 in the first cluster and sizes 26 in the second cluster are formed. Bangka Belitung Island, Riau Island, Special Capital District of Jakarta, Special Region of Yogyakarta, Banten, Bali, South Kalimantan, and East Kalimantan are grouped in the first cluster. Moreover, 26 other provinces included in cluster 2. The cluster plot of HDI displayed in Figure 4.

Figure 4. The cluster plot of HDI variables
According to Figure 2 and Figure 3, it can be seen that the member of each cluster of HDI variables and education variables are different. Moreover, the number of clusters is also different. Corresponding to [20], HDI does expose a country's performance in the educational and health aspect. It might be happened due to the change in the formula in HDI calculation. Calculating HDI values with a new method uses an aggregation method that is changed from the arithmetic mean to a geometric mean. According to the BPS, some indicators are also unsuitable for calculating the HDI. The literacy rate is no longer relevant for measuring education as a whole, as it cannot describe the quality of education. Besides, since the literacy rate is already high in most regions, it is not possible to clearly distinguish the level of education between the regions. GDP per capita cannot reflect the income of the people in a region either. By taking into account the average length of the school and the expected number of school years, a more relevant picture of education and changes can be obtained. Additionally, gross national income can replace gross domestic income as it reflects the income of the people in an area.

Based on data published in 2013 by the Central Statistics Agency (BPS), Indonesia's HDI value ranks 108th at an HDI value of 0.68. The Indonesian HDI value is compared to Norway which is a country with the highest HDI with the value of 0.94, and the difference is quite large. Based on the HDI published in 2016, the HDI value of Indonesia for 2014 was 0.684, which is Indonesia included in the category of middle human development in the order of 110 from 188 territories and countries. At intervals 1980 and 2014, Indonesia's HDI value rose 0.21 point to 0.684, which corresponds to an increase of 44.3 per cent with an average annual increase of around 1.08 per cent [21]. However, that number has increased quite well after calculating the HDI value in 2019 using a new method, the value is 71.92.

In line with that, by the middle of 2020, Indonesia is in the 70th position in the rank of world education while Norway that has highest HDI value for years positioned 16th in the world education ranking. Additionally, this ranking was calculated using the quality index and opportunity index (https://ceoworld.biz/2020/05/10/ranked-worlds-best-countries-for-education-system-2020/). However, Indonesia is still categorized in high human development.

The cluster means of two group datasets are mentioned in the following table. Table 1 shows the cluster means of each education variable, consisting of participation rate of children in early childhood education (A), the rate of school enrollment in universities (B), the proportion of the population in the age of 15 to 24 with information and communication technology skills (C), the proportion of the population in the age of 15 to 59 with information and communication technology skills (D), the average length of school for the population in old age of 15 years (E), the ratio of the gross enrollment rate for women and men at the university (F), the net enrollment rate for women and men at senior high school (G), and the net enrollment rate for women and men at junior high school (H).

|   | A     | B     | C     | D     | E     | F     | G     | H     |
|---|-------|-------|-------|-------|-------|-------|-------|-------|
| 1 | 39.93 | 30.27 | 87.35 | 62.91 | 8.93  | 100.34| 105.11| 101.00|
| 2 | 29.87 | 32.43 | 80.60 | 53.84 | 9.01  | 122.58| 105.62| 101.60|
| 3 | 67.75 | 73.14 | 97.91 | 75.04 | 9.83  | 96.79 | 113.18| 113.91|
| 4 | 31.02 | 35.33 | 57.03 | 40.34 | 8.76  | 102.84| 108.19| 101.12|

Based on Table 1, it can be said that cluster 3 has the highest mean since this cluster has fewest members. Afterwards, the cluster means for each HDI variable mentioned in Table 2. Five variables determine the HDI value in each province with a new method. The variables are expected length of school in a year (P), adjusted per capita expenditure in thousand Rupiah per year per person (Q), the average length of school in a year (R), life expectancy at birth in a year (S).

|   | P     | Q     | R     | S     |
|---|-------|-------|-------|-------|
| 1 | 13.21 | 13921.37 | 9.24 | 71.57 |
| 2 | 13.03 | 9990.73  | 8.33 | 69.37 |
As shown in previous Table 1, the cluster means value for each variable in cluster 1 is greater than that of cluster 2 because the number of members for cluster 1 is smaller than for cluster 2 with a ratio of 4/13.

4. Conclusion
This study included two sets of variables. The first group is the variable that is used to determine the quality of education in Indonesia, while the second group is the variable that is used to calculate the HDI score using a new, adapted method. From the two groups, it can be concluded that the two groups did not form any clusters with the same members between the first and the second group. The first group formed 4 clusters and the second group formed only 2 clusters.

For instance, the education variables of the island of Riau, the special capital region Jakarta, West Java, Central Java, East Java, Banten, Bali, West Nusa Tenggara, Central Kalimantan and Gorontalo are grouped in the first cluster. This first cluster shows that the regions within the cluster have a similar educational status. On the other hand, Bangka Belitung Island, Riau Island, Jakarta Special Capital Region, Yogyakarta Special Region, Banten, Bali, South Kalimantan and East Kalimantan belong to the same group in the first variable HDI cluster. These two things suggest that a high HDI does not mean that the region has a better level of education. The results of this study indicate that the assessment of the quality of education needs to be continued both for regions with high HDI conditions and vice versa because the quality of education does not appear to be reflected in the high HDI.

Acknowledgments
This research is fully supported by the Department of Statistics, Universitas Padjadjaran. We would thank the reviewer for all the suggestions.

References
[1] Arisman 2018 J. Ilmu Ekonomi 7 113-122
[2] Sarkar Md S K, Sadeka S and Sikdar Md M H 2012 Asian J. Appl. Sci. Eng. 1 23-34
[3] Soubbotina, T P 2004 Beyond Economic Growth: An Introduction to Sustainable Development (Washington, D.C.: World Bank Group)
[4] Thomas V, Dailimi M, Dhareshwar A, Kaufmann D, Kishor N, Lopes R and Yan W 2000 The Quality of Growth (Washington, D. C.: World Bank Group)
[5] Rafiei N and Davari F 2015 Mater. Sociomed. 27 35-38
[6] Jan H 2017 Research on Humanities and Social Science 7 50-54
[7] Seleim A and Bontis N 2013 Knowl. Process. Manag. 20 131-140
[8] Rampersaud G C, Pereira M A, Girard B L, Adams J, and Metzl J D 2005 J. Am. Diet. Assoc. 105 743-760
[9] Hess P and Ross C 1997 Economic Development: Theories, Evidence, and Policies (Forth Worth: Harcourt Brace College Publishers)
[10] Sulisworo D 2016 Journal of Education and Learning 10 127-138
[11] Sukasni A and Efendy H 2017 Int. J. Educ. 9 183-199
[12] Madadipouya K 2017 JOIV: Int. J. Inform. Visualization. 1 61–71.
[13] Sambour M and Othman Z 2016 Int. J. Adv. Sci. Eng. Inf. Technol. 6 1127–1133.
[14] Ahmar A S, Napitupulu D, Rahim R, Hidayat R, Sonatha Y and Azmi M 2018 Proc. Int. Conf. on Statistics, Mathematics, Teaching, and Research (Makassar) vol 1028 (Bristol: IOP Publishing) 012006
[15] Kaur N, Sahiwal J K, Kaur N and Punjab P 2012 Int. J. Adv. Res. Comput. Eng. Technol. 1 85-91
[16] Yadav A and Dhingra S 2016 Int. J. Eng. Dev. Res. 4 545-548
[17] Kaufmann L and Rousseeuw P J 1990 Finding Groups in Data: An Introduction to Cluster Analysis (New Jersey: John Wiley & Sons, Inc)
[18] Hartigan J and Wong M 1979 J. R. Stat. Soc. C-Appl. 28 100–108
[19] Jain A K and Dubes R C 1988 Algorithms for Clustering Data (New Jersey: Prentice-Hall, Inc)
[20] Al-Hilana H 2012 IOSRJBM 2 24-28
[21] Human Development Report (HDR) 2015 Work for Human Development Briefing note for countries on the 2015 HDR (Nepal: United Nations Development Programme)