Comparison of Average Linkage and K-Means Methods in Clustering Indonesia's Provinces Based on Welfare Indicators

Comparison of Average Linkage and K-Means Methods in Clustering Indonesia's Provinces Based on Welfare Indicators

A L Yusniyanti, F Virgantari*, and Y E Faridhan

Department of Mathematics , Pakuan University, Bogor 16143, Indonesia

*E-mail : fitriav12@gmail.com

Abstract. This study compares two clustering methods, i.e. average linkage and K-means, in grouping Indonesia's provinces based on welfare indicators in education, health, and income. Data from Statistics Indonesia (BPS) covering Indonesia's 34 provinces are used. Welfare variables exercised in this study are population, rate of population with government-assisted health covers, morbidity rate, human development index, expense rate per capita, and rate of the population aged 15 or over who graduated from junior high school (completed Year 9). Results show that the average linkage method generates three clusters; the first cluster of which consists of 32 provinces, while the second and third clusters each consist of only one province. On the other hand, the K-means method is set to generate equally three clusters. Unlike the first method, K-means's first cluster, in this case, consists of 14 provinces, while its second and thirds clusters consist of 13 and 7 provinces, respectively. Performances of both methods are measured using the variance ratio. The average linkage and k-means cluster methods yield variance ratios of 0.08275 and 0.28881, respectively. Based on these criteria, the average linkage method is shown to exercise a better performance due to its smaller variance ratio.

1. Introduction

Indonesia is the fourth most populated country in the world, equivalent to 3.51% of the total world population [1]. Currently its annual increase rate in population is 1.07%, which is an increase of almost 2.9 million people in one year.

The increase in population should be followed by the increase in their welfare. As a rule of thumb, people's basic well-being state is met when their life necessities are fulfilled. Primary life necessities are immediate basic needs of food, clothing and shelter; while secondary life necessities covers items such as sanitation, education, and health care [2]. However, there are gaps in people's welfare such as in education, health care, as well as income amongst provinces in Indonesia [3]. A welfare-based identification should be made to reduce the gap. As there are usually a number of welfare indicators, one way to tackle this problem is to group the provinces using cluster analysis [4, 5].

In the local scope, a number of studies look into grouping regions using cluster analysis which based on welfare indicators. Two of such examples are [6] who applies average linkage method in clustering regencies and cities in Central Java, while [7] uses K-means cluster in grouping Indonesia's provinces. However, comparisons on clustering methods in province level have not identified any preferred method in terms of grouping.

There are two broad categories of cluster analysis, i.e. hierarchical and non-hierarchical categories. In this paper, one method of each category will be exercised in grouping Indonesia's provinces based on welfare indicators, namely average linkage and K-means methods. These two methods will be

compared based on their performance in clustering the provinces. The structure of this paper is as follows. Section 2 describes the clustering methods employed here. In Section 3 attributes and data source are described. Section 4 reports the findings, and Section 5 concludes the paper.

2. Cluster Analysis

There have been many variants of clustering methods proposed in recent decades; yet for simple reference, each clustering method may be seen belonging into one of two broad categories, i.e. hierarchical and non-hierarchical. Hierarchical methods are used to structurally group observations based on their similarities; the number of desired clusters is not determined in advance. On the other hand, non-hierarchical methods start by determining the number of desired clusters [8, 9].

Hierarchical clustering category falls into two sub-categories, i.e. agglomerative and divisive [10]. The hierarchical method used in this paper, namely average linkage, belongs to the agglomerative category. This particular method will be discussed further in subsection 2.1.

Non-hierarchical clustering category consists of several sub-categories, i.e. partitioning, density-based, grid-based, and others. This "other" category covers methods such as machine learning methods as well as methods for categorical and high dimensional data [11]. Of these four categories, partitioning is the most widely used. K-Means method belongs to this category, and will be discussed further in subsection 2.2. Figure 1 summarize the clustering categories.

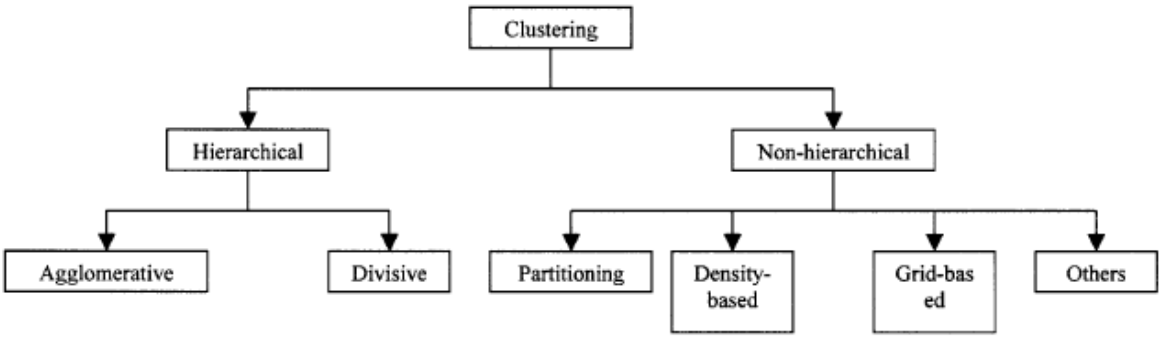

Figure 1. Categorization of clustering methods [8].

Agglomerative clustering starts with individual objects as clusters, and works its way bottom-up into a single cluster as similarities decrease. In contrast, divisive or top-down clustering starts with a single cluster, which is divided into at least two subgroups that are dissimilar from each other. Both types of hierarchical clustering use a matrix of similarities (also known as distance matrix), and result in a dendrogram. Dendograms display mergers or divisions at levels of clustering.

On the contrary, non-hierarchical clustering techniques do not store distance matrix and basic data, thus easier to apply on larger data sets. They are also faster in terms of computational time [8, 10].

2.1. Hierarchical method: Average Linkage

Agglomerative category is comprised of four sub-categories, namely linkage, Ward, centroid and median methods. Average linkage is one of three linkage methods, others being single linkage and complete linkage [12].

Average linkage calculates the distance between two clusters as the average distance between all pairs of objects, where one member of a pair fits into each cluster. [10] gives the details of general agglomerative clustering algorithm for grouping objects, starting with determining the distance matrix. Euclidean distance is employed here. The nearest, or most similar objects, are merged to form a cluster, say *(UV)*. The distances between *(UV)* and another cluster *W* is given by

$$d_{(UV)W} = \frac{\sum_i \sum_k d_{ik}}{N_{(UV)} N_W} \quad (1)$$

where d_{ik} is the distance between object *i* in the cluster *(UV)* and object *k* in the cluster *W*, while $N_{(UV)}$ and N_W are the number of items in clusters *(UV)* and *W*, respectively.

2.2. Non-hierarchical method: K-Means

K-Means is the most popular method in the partitioning category. Other methods such as K-Modes, K-Prototypes, and Fuzzy C-Means are its variants [8]. This method assigns each object to the cluster that has the nearest centroid, or mean. [10] elaborates the algorithm in its simplest version, starting with defining *K* initial centroids as seed points. Using Euclidean distance, the distance between object *i* and cluster *l* is defined as

$$d(i,l) = \left(\sum_{j=1}^{p} [x(i,j) - \bar{x}(i,j)]^2\right)^{\frac{1}{2}} \quad (2)$$

with

$$E[p(n,k)] = \sum_{i=1}^{n} D[i, l(i)]^2 \quad (3)$$

where $\bar{x}(i,j)$ denotes the average of variable *j* in cluster *l*; $E[p(n,k)]$ is the partition error, *1(i)* is the cluster containing object *i*; and *D[i,1(i)]* is the distance between object *i* and average of the cluster containing the object *i* [12].

This initial selection partly determines the final assignment of objects the final clusters. It is advisable to re-run the algorithm with different initial partition.

[10] cautions against pre-defining the number of clusters *K*, for this following reasons:

1. If two or more seed points happen to lie within a single cluster, the resulting clusters will be poorly differentiated.
2. Should there be an outlier, it might yield at least one group with dissimilar objects.
3. As the sampling method may not select data from the rarest group, forcing data into pre-defined *K* groups would result in nonsensical cluster.

2.3. Measuring cluster validity: The ideal cluster

Comparison on cluster methods includes measuring validity on the resulting clusters, which is whether the clusters are well separated. As proposed by [13], an ideal cluster should minimize the variance within clusters while maximizing the variance between clusters. [14] gives the formula as follows. The variance of cluster *k* can be determined using

$$V_k{}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}_k)^2 \quad (4)$$

Given N as the number of members in all clusters, the variance within clusters can be defined as

$$V_w{}^2 = \frac{1}{N-k} \sum_{i=1}^{k} (n_i - 1)\, V_i{}^2 \quad (5)$$

Variance between clusters ($V_b{}^2$) is defined as

$$V_b{}^2 = \frac{1}{(k-1)} \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2 \quad (6)$$

where $\bar{x}$ is the grand mean of all clusters.

An ideal cluster should have minimum $V_w{}^2$ and maximum $V_b{}^2$; in other words, a minimum *v*, where

$$v = \frac{V_w{}^2}{V_b{}^2} x100\% \quad (7)$$

3. Data and Attributes

The methods presented in sub-sections 2.1 and 2.2 are employed to data of Indonesia's 34 provinces in 2019, provided by Statistics Indonesia [3, 15, 16, 17]. There are six welfare indicators exercised in this paper, divided into four types of attributes:

a. Demographic attribute: (1) population,
b. Economic attribute: (2) expense rate per capita,
c. Health attribute: (3) rate of population with government-assisted health cover and (4) morbidity rate,
d. Education attribute: (5) human development index and (6) rate of population aged 15 or over who graduated from junior high school (completed Year 9).

Indicators 2, 3, 5 and 6 implies people's well-being when the measures are high. Indicator 4 is an exception, where people's well-being indicated by low morbidity rate.

Table 1 shows basic descriptive statistics results for each of six observed welfare indicators in Indonesia's 34 provinces.

Table 1. Basic descriptive statistics of welfare indicators in Indonesia in 2019, N = 34 provinces

Welfare indicator	Mean	Standard deviation	Minimum	Maximum
Population	7,562,368	11,165,277	716,400	48,683,700
Population with government-assisted health cover	32.33%	12.35%	18.49%	72.38%
Human development index	70.39	39.94	60.06	80.47
Morbidity rate	13.51	25.81	9.37	19.64
Expense rate per capita	Rp 1,142,416.79	Rp 260,160.18	Rp 704,754.00	Rp 2,039,157.00
Population aged ≥ 15 who completed Year 9	20.14%	2.33%	14.54%	25.64%

These observed indicators use different metrics and are given in different measures. Therefore, all data are standardized.

Figure 2 displays the multiple box-plots of selected welfare indicators in Indonesia's 34 provinces in 2019, based on respective z-scores. Only one of six indicators i.e. morbidity rate that appears to distribute normally without outlier. The rest of the box-plots show gaps amongst some of Indonesia's provinces. In terms of population, the three top outliers are the most populated provinces, namely West Java (12), East Java (15) and Central Java (13). Shifting to the economic attribute, it is obvious that

being the nation's capital, DKI Jakarta is the outlier – the province with the highest expense rate per capita.

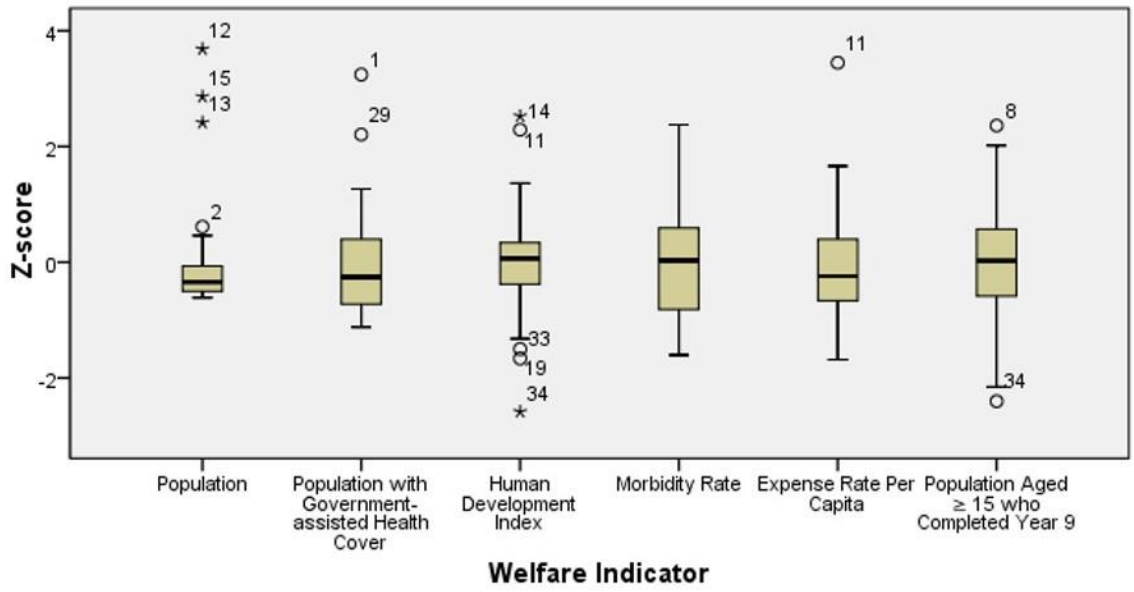

Figure 2. Box-plots of welfare indicators in Indonesia in 2019, n = 34 provinces (standardized values).

Looking at the health attribute, there are two provinces that have very high rate of population with government-assisted health cover, i.e. DI Aceh and Gorontalo. Furthermore, the last two welfare indicators tell another story. These two education attributes show that the province Papua has the lowest human development index as well as the lowest rate of population of aged 15 or over who graduated from junior high school (completed Year 9). On the other hand, provinces with the highest human development index are DI Yogyakarta and DKI Jakarta, while Lampung has the highest rate of population of aged 15 or over who completed Year 9.

4. Clustering Results and Discussion

4.1. Average linkage results

Figure 3 depicts the average linkage clustering process and results in form of a dendrogram. It appears that two provinces, DKI Jakarta and Papua, are separated from the others. This agrees with the descriptive results discussed in the previous section. Likewise, the three clusters formed by average linkage method are shown in Table 2.

According to the average linkage results, there are 32 members of cluster 1, whereas each of clusters 2 and 3 has only one member. DKI Jakarta as the only province in cluster 2 appears to have the highest expense rate per capita, the second highest human development index, and above the national average of population who completed Year 9. This nation's capital province also has low morbidity rate and almost 50% of population with government-assisted health cover. These statistics show that cluster 2 has the highest welfare compared to the other clusters.

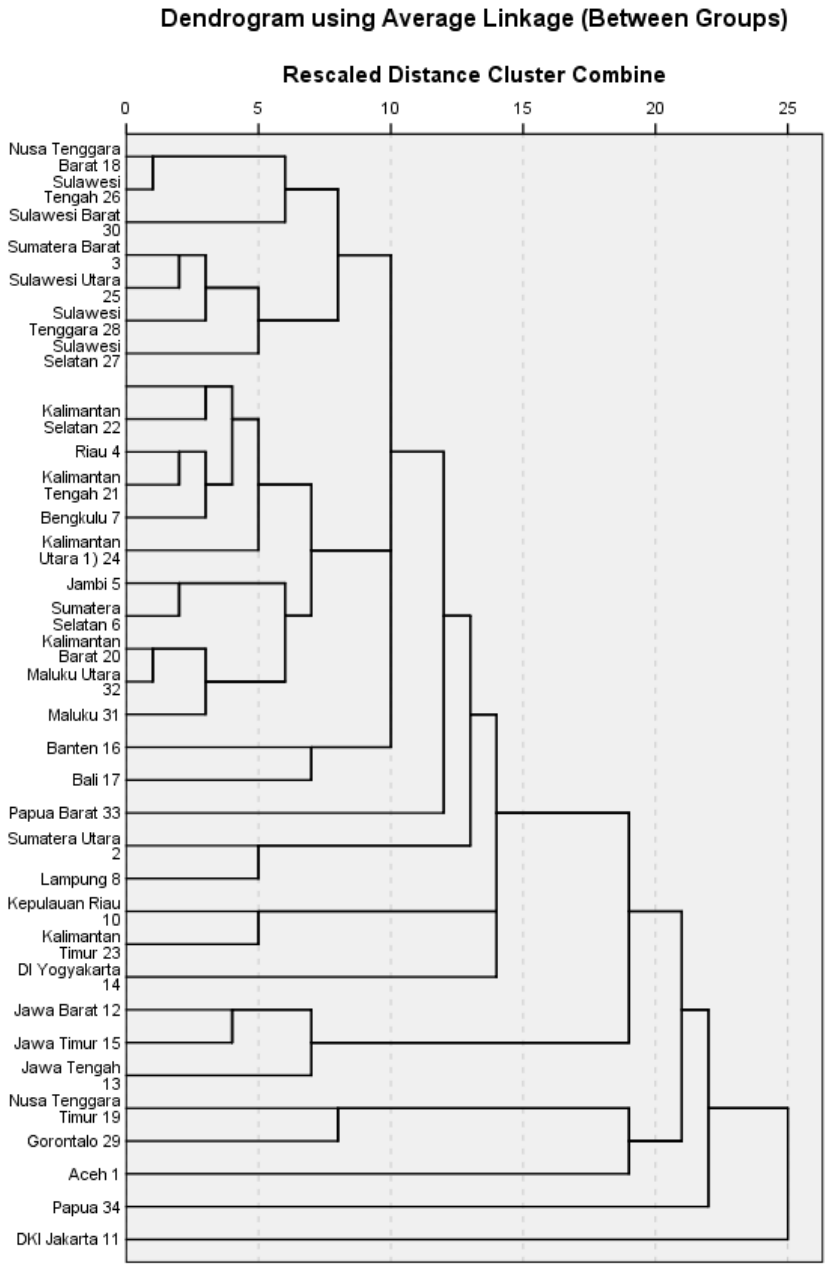

Figure 3. Dendrogram of welfare indicators in Indonesia's 34 provinces using average linkage.

On the other hand, Papua as the only province in cluster 3 has the lowest human development index, the lowest population who completed Year 9, and low population rate with government-assisted health cover. This confirms that cluster 3 has the lowest welfare compared to the other clusters.

As for 32 provinces in cluster 1, it can be said that their welfare are in the middle between clusters 2 and 3. However, to elaborate this further, it might need additional welfare indicators.

4.2. K-means results

K-means algorithm require pre-defined number of clusters that are going to be formed. In this study, the number of clusters is set to be three. This number is set based on Indonesian government's welfare level, i.e. high, medium, and low [3]. Table 3 shows members of clusters formed by K-means method.

Table 2. Members of clusters formed by average linkage.

Cluster	Member
1	Aceh, Sumatera Utara, Sumatera Barat, Riau, Jambi, Sumatera Selatan, Bengkulu, Lampung, Kepulauan Bangka Belitung, Kepulauan Riau, Jawa Barat, Jawa Tengah, DI Yogyakarta, Jawa Timur, Banten, Bali, Nusa Tenggara Barat, Nusa Tenggara Timur, Kalimantan Barat, Kalimantan Tengah, Kalimantan Selatan, Kalimantan Timur, Kalimantan Utara, Sulawesi Utara, Sulawesi Tengah, Sulawesi Selatan, Sulawesi Tenggara, Gorontalo, Sulawesi Barat, Maluku, Maluku Utara, Papua Barat.
2	DKI Jakarta
3	Papua

Table 3. Members of clusters formed by K-means.

Cluster	Member
1	Aceh, Nusa Tenggara Barat, Nusa Tenggara Timur, Kalimantan Barat, Sulawesi Utara, Sulawesi Tengah, Sulawesi Selatan, Sulawesi Tenggara, Gorontalo, Sulawesi Barat, Maluku, Maluku Utara, Papua Barat, Papua.
2	Sumatera Utara, Sumatera Barat, Riau, Jambi, Sumatera Selatan, Bengkulu, Lampung, Jawa Barat, Jawa Tengah, Jawa Timur, Banten, Kalimantan Tengah, Kalimantan Selatan.
3	Kepulauan Bangka Belitung, Kepulauan Riau, DKI Jakarta, DI Yogyakarta, Bali, Kalimantan Timur, Kalimantan Utara.

There are 14 provinces added to cluster 1. This cluster has the highest average of morbidity rate, as well as the lowest averages of human development index, expense rate per capita and population rate with government-assisted health cover. This shows that provinces in this cluster has the lowest welfare compared to other clusters.

Cluster 3 consists of seven provinces where the highest human development index and the highest expense rate per capita are found, as well as the lowest average on morbidity rate. This confirms that provinces in this cluster has the highest welfare compared to other clusters. As for cluster 2, which consists of 13 provinces, their welfare are in the middle between clusters 1 and 3.

4.3. Comparison of resulting clusters

Table 4 displays variance calculation on resulting clusters formed by average linkage and K-means methods. The variance ratio of the clusters formed by average linkage is smaller than that by K-means. Consequently, in this study, average linkage forms more ideal clusters compared to K-means. In other words, average linkage performs better than K-means in grouping Indonesia's provinces based on the six observed welfare indicators.

However, it is worth to note that K-means's performance in this paper might be affected by some factors. One might be related to the pre-defined number of clusters that [10] cautions. The second would be adding a step of attempting different initial partitions and subsequently choosing the most ideal clusters based on [13, 14]'s cluster validity measurement.

Table 4. Cluster validity comparison

Variance	Average linkage	K-means
Within clusters (V_w^2)	0.124	0.178
Between clusters (V_b^2)	1.5012	0.616
Ratio (v)	0.083	0.289

5. Conclusions and Future Works

This paper was intended to compare two clustering methods, each represents hierarchical and non-hierarchical analysis. The average linkage method seems to perform better than K-means in this study; however, more cautions should be given when employing K-means method. One should also examine the cluster shapes, as K-means assumes spherical shapes of clusters [18]. Partitioning algorithms generally do not work well with clusters of arbitrary shapes [8]. In terms of outliers, K-medoid method works better than K-means. However, amongst the advantages of K-means over K-medoid are its clear geometric and statistical meaning [11].

As a future work, other clustering methods across hierarchical and non-hierarchical analysis could be added to the comparison. In terms of clustering Indonesia's provinces, the most immediate would be to consider adding more welfare indicators to the analysis. By using either or both approaches, it is hoped that it would result in better suggestions for Indonesia's future policies concerning welfare increase in each province.

References

[1] Worldometers 2020 *Indonesian Population* https://www.worldometers.info/world-population/indonesia-population/ [Accessed on 15th October 2020]
[2] Stiglitz J E, Sen A and Fitoussi J -P 2009 *Report by the Commission on the measurement of economic performance and social progress* Commission on the Measurement of Economic Performance and Social Progress, Mimeo
[3] BPS 2018 *Statistik Kesejahteraan Rakyat 2018* (Jakarta:Badan Pusat Statistik)
[4] Hirschberg J G, Maasoumi E and Slottje D J 1991 Cluster analysis for measuring welfare and quality of life across countries *J. Econometrics* **50** 131-50
[5] Luzzi G F, Flückiger Y and Weber S 2008 A cluster analysis of multidimensional poverty in Switzerland *Quantitative Approaches to Multidimensional Poverty Measurement*, ed Kakwani N and Silber J (London:Palgrave Macmillan) chapter 4 pp 63-79
[6] Yulianto S and Hidayatullah K H 2014 Analisis klaster untuk pengelompokan kabupaten/kota di provinsi Jawa Tengah berdasarkan indikator kesejahteraan rakyat *Statistika* **2** 56-63
[7] Ramdhani F 2015 Pengelompokan provinsi di Indonesia berdasarkan karakteristik kesejahteraan rakyat menggunakan metode K-Means Cluster *Gaussian* **4** 875-84
[8] Ma E W M and Chow T W S 2004 A new shifting grid clustering algorithm *Pattern Recognition* **37** 503-514
[9] Mattjik A A and Sumertajaya I M 2011 *Sidik Peubah Ganda* (Bogor:IPB Press)
[10] Johnson R A and Wichern D W 2007 *Applied Multivariate Statistical Analysis* 6e (New Jersey:Prentice-Hall International)
[11] Berkhin P 2006 A survey of clustering data mining techniques *Grouping Multidimensional Data* (Berlin, Heidelberg: Springer) pp 25-71
[12] Nugroho S 2008 *Statistik Multivariat Terapan* (Bengkulu:UNIB Press)
[13] Ray S and Turi R H 1999 Determination of number of clusters in K-means clustering and application in colour image segmentation *Proc. 4th Int. Conf. on Advances in Pattern Recognition and Digital Techniques* pp 137-43
[14] Barakbah A R and Arai K 2004 Determining constraints of moving variance to find global

optimum and make automatic clustering *Proc. Industrial Electronics Seminar (IES)* pp 409-13
[15] BPS 2019a *Indeks Pembangunan Manusia* (Jakarta:Badan Pusat Statistik)
[16] BPS 2019b *Rata-Rata Pengeluaran per Kapita Sebulan di Daerah Perkotaan dan Pedesaan Menurut Provinsi dan Kelompok Barang (rupiah)* (Jakarta:Badan Pusat Statistik)
[17] BPS 2019c *Statistik Indonesia 2019* (Jakarta:Badan Pusat Statistik)
[18] Dabbura I 2018 *K-Means Clustering: Algorithms, Applications, Evaluations Methods, and Drawbacks* https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a [Accessed on 17th October 2020]