Advance Sustainable Science, Engineering and Technology (ASSET)
Vol. 3, No.1, April 2021, pp. 0210107-1 ~ 0210107-10
ISSN: 2715-4211 DOI: https://doi.org/10.26877/asset.v3i1.8461

Implementation of *k*-means and K-Medians Clustering in Several Countries Based on Global Innovation Index (GII) 2018

Ade Famalika[1], Pardomuan Robinson Sihombing[2]

[1]Universitas Bina Insan, Jl. Raya Siliwangi No.6, RT.001/RW.004, Sepanjang Jaya, Kec. Rawalumbu, Kota Bekasi, Jawa Barat 17114 Indonesia

[2*]Badan Pusat Statistik, Jl. Dr. Sutomo 6-8 Jakarta 10710 Indonesia

*robinson@bps.go.id

Abstract. The Global Innovation Index (GII) is an instrument to assess the ranking of innovation capabilities of all countries. The sub-index of the GII has seven enabler pillars: Institutions, Human Capital and Research, Infrastructure, Market sophistication, Business Sophistication, Knowledge and Technology Outputs, and Creative Outputs. The *k*-means method and *k*-medians method are methods for cluster countries based on GII. Cluster 1 in *k*-means method consists of 48 Countries, Cluster 2 consists of 45 Countries and Cluster 3 consists of 33 Countries and has the average value of seven variables are the highest. Cluster 1 in *k*-medians method consists of 33 Countries and has the average value of seven variables are the highest., Cluster 2 consists of 53 Countries and Cluster 3 consists of 40 Countries. The result clustering with using *k*-means method and *k*-medians method showed that *k*-medians is better than *k*-means method because the variance value of *k*-medians is smaller than *k*-means.

Keywords: GII, *k*-means Cluster, K-Medians Cluster, Variance

(Received 2021-04-28, Accepted 2021-04-30, Available Online by 2021-04-30)

1. Introduction

The Global Innovation Index (GII) is an instrument to assess the ranking of innovation capabilities of 126 countries carried out by the World Intellectual Property Organization (meaning in the English World Intellectual Property Organization-WIPO) in coordination with INSEAD Institute (France) and Cornell University (United States). The Global Innovation Index (GII) project was launched by Professor Dutta at INSEAD in 2007 with the simple goal of determining how to find metrics and approaches that better capture the richness of innovation in society and go beyond such traditional measures of innovation as the number of research articles and the level of research and development (R&D) [1].

There were several motivations for setting this goal. First, innovation is important for driving economic progress and competitiveness-both for developed and developing economies. Many governments are putting innovation at the centre of their growth strategies. Second, the definition of

innovation has broadened, it is no longer restricted to R&D laboratories and to published scientific papers. Innovation could be and is more general and horizontal in nature, and includes social innovations and business model innovations as well as technical ones. Last but not least, recognizing and celebrating innovation in emerging markets is seen as critical for inspiring people, especially the next generation of entrepreneurs and innovators.

The sub-index of the GII has seven enabler pillars: Institutions, Human Capital and Research, Infrastructure, Market sophistication, Business Sophistication, Knowledge and Technology Outputs, and Creative Outputs. Seeing the importeance of the Global Innovation Index, this study aimed at finding out the clustering of countries based on GII data in 2018 with using *k*-means method and *k*-medians method, then compare the two methods [1].

2. Methods

2.1. Research Variables

This study used secondary data sourced from World Intellectual Property Organization-WIPO coordinating with INSEAD and Cornell University. The three institutions measured a country's level of global innovation based on seven variables, including [1]:

- Institutions Variable (X_1)
  The Institutions pillar captures the institutional framework of a country. Nurturing an institutional framework that attracts business and fosters growth by providing good governance and the correct levels of protection and incentives is essential to innovation.
- Human Capital and Research Variable (X_2)
  The level and standard of education and research are activities in a country are prime determinants of the innovation capacity of a nation. This pillar tries to gauge the human capital of countries.
- Infrastructure Variable (X_3)
  The infrastructure includes three sub-pillars: Information and communication technologies (ICTs), General infrastructure, and Ecological sustainability.
- Market Sophistication Variable (X_4)
  The availability of credit and an environment that supports investment, access to the international market, competition, and market scale are all critical for businesses to prosper and for innovation to occur.
- Business Sophistication Variable (X_5)
  The business sophistication tries to capture the level of business sophistication to assess how conducive firms are to innovation activity.
- Knowledge and Technology Outputs Variable (X_6)
  This pillar covers all those variables that are traditionally thought to be the fruits of inventions and/or innovations. The first subpillar refers to the creation of knowledge. The second sub-pillar, on knowledge impact, includes statistics representing the impact of innovation activities at the micro- and macroeconomic. The third sub-pillar, on knowledge diffusion.
- Creative OutputsVariable (X_7)
  The last pillar, on creative outputs, has three sub-pillars The first sub-pillar on intangible assets includes statistics on trademark applications by residents at the national office. The second sub-pillar on creative goods and services and the third sub-pillar on online creativity.

2.2. Stage of Research

2.2.1. Cluster Analysis

Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A cluster is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters [2].

Data clustering algorithms can be hierarchical or partitional. Hierarchical algorithms find successive clusters using previously established clusters, whereas partitional algorithms determine all

clusters at time. Hierarchical algorithms can be agglomerative (bottom-up) or divisive (top-down). Agglomerative algorithms begin with each element as a separate cluster and merge them in successively larger clusters. Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters.

There are two assumptions that must be fulfilled in cluster analysis, namely samples that are representative (representing the population) and there are no cases of multicollinearity between variables[3]. A representative sample can be seen from the Kaiser-Meyer-Olkin (KMO) value that is greater than 0.5[4].

$$KMO_j = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} u_{ij}^2} \quad (1)$$

The presence or absence of multicollinearity between variables can be seen from the value of Variance Inflation Factor (VIF) which is greater than 10 [5].

$$VIF_k = \frac{1}{1 - {R_k}^2} \quad (2)$$

Here ${R_k}^2$ is the $R^2 = value$ obtained by regressing the k^{th} predictor on the remaining predictors.

2.2.2. K-Means Method

Clustering is a classification of similar objects into several different groups, it is usually applied in the analysis of statistical data which can be utilized in various fields, for example, machine learning, data mining, pattern recognition, image analysis and bioinformatics [6]. In general, partitioning algorithms such as *k*-means and EM highly recommended for use in large-size data. This is different from a hierarchical clustering algorithm that has good performance when they are used in small size data [7]. The method of K-means algorithm as follows [8]:

1) Determine the number of clusters *k* as in shape. To determine the number of clusters K was done with some consideration as theoretical and conceptual considerations that may be proposed to determine how many clusters.
2) Generate K centroid (the center point of the cluster) beginning at random. Determination of initial centroid done at random from objects provided as K cluster, then to calculate the *i*-cluster centroid next, use the following formula:

$$v = \sum_{i=1}^{n} \frac{x_i \quad ; i = 1,2, \dots, n}{n} \quad (3)$$

where $\boldsymbol{v}$ is cluster centroid, $\boldsymbol{n}$ is the number of objects to members of the cluster and $\boldsymbol{x_i}$ is the object to-i.

3) Calculate the distance of each object to each centroid of each cluster. To calculate the distance between the object with the centroid author using Euclidian Distance.

$$\boldsymbol{d(x,y) = \|x - y\| = \sqrt{\sum_{j=1}^{n} (x_j - y_j)^2}} \quad (4)$$

4) Allocate each object into the nearest centroid. To perform the allocation of objects into each cluster during the iteration can generally be done in two ways, with a hard K-means, where it is explicitly

every object is declared as a member of the cluster by measuring the distance of the proximity of nature towards the center point of the cluster, another way to do with fuzzy *k*-means.

5) Do iteration, then specify a new centroid position using equation in step 2.
6) Repeat step 3 if the new centroid position is not the same.

2.2.3. K-Medians Method

The *k*-medians method is the development of the *k*-means method. Both produce *k*-cluster formed by measuring the distance between the center point and each object, then each object is grouped according to the nearest center point. Both of these methods have differences, one of which is at the center of the cluster. As the name implies, *k*-means uses the mean (mean) and *k*-medians using the median. Furthermore, the median is descriptive statistics which tend to be more resistant to outliers. Therefore, the use of the K-Medians method will minimize errors in the cluster [9]. The method of *k*-medians algorithm as follows:

1) Determine the number of clusters
   In the *k*-median method the number *k* must be determined in advance and there is no specific rule in determining the number of clusters k, because sometimes the determination of the number of clusters is based on the subjectivity of the researcher. In this study, cluster number *k* was determined using Silhouette Coefficient. Stages of silhouette coefficient calculation[10]:
   - Calculate the average distance of objects with other objects in the cluster with the equation:

$$\boldsymbol{a(i) = \frac{1}{[A] - 1} \sum j \in_{A, j \neq 1} d(i, j)} \quad (5)$$

     Where *a(i)* is average distance between group components, *I* is an object in *cluster* A
     J is other objects in *cluster* A, *d(i,j)* is distance betweenobject *i* and *j*.
   - Calculate the average distance of objects with all other objects in another cluster, then take the minimum value with the equation:

$$d(i, C) = \frac{1}{[A]} \sum j \in C \, d(i, j) \quad (6)$$

     Where, *d(i,*C*)* is the average distance between objects *i* with all objects in another cluster (C), where A ≠ C
   - Calculate the value of silhouette coefficient with the equation:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (7)$$

2) Determine the center point (centroid)
   Some opinions on choosing centroids for the *k*-medians method are as follows:
   - Based on Hartigan (1975), the selection of centroids can determined based on the interval of the number of each observation [11].
   - Based on Rencher (2002), the selection of centroids can be determined through the approach of one of the hierarchical methods [12].
   - Based on Teknomo (2007), the selection of centroids can be randomized from all observations [13].

   In this study, the centroid was chosen based on Teknomo's opinion in determining the centroid, which is to choose centroids randomly from all observation units.
3) Determine the distance of each observation unit to each centroid
   In this case, distance measurements are used to place observations into clusters based on the nearest centroid. The measure of distance used in the *k*-medians method is Manhattan's distance [14].

Manhattan distance is a measurement based on a grid system in which the points in question are placed. The concept is that in order to move from start to end point, one of four directions must be chosen for the point to advance: up, down, left, or right. Each decision will move the start point one unit in the chosen direction. The Manhattan distance is determined by the number of decisions required for the start point to reach the end point [15]. Manhattan distance can be written as follows:

$$\boldsymbol{d(x_{ij}, c_{ij}) = \sum_{j=1}^{p} |x_{ij} - c_{ij}| \; ; i = 1, 2, \dots, k} \tag{8}$$

2.3. Determining the Goodness of the Clustering Method with Standard Deviation

Varians can be calculated by

$$varians = \frac{S_w}{S_B} \tag{9}$$

To find out which method has the best result, we can use the standard deviation in the cluster (S_w) and the standard deviation between clusters (S_B)[16]. The average standard deviation formula in the cluster (S_w):

$$S_w = K^{-1} \sum_{k=1}^{K} S_k \tag{10}$$

Here K is number of clusters formed and S_k is standard cluster k^{th}. Standard deviation formula between clusters (S_B):

$$S_B = \left[(K-1)^{-1} \sum_{k=1}^{K} (\bar{X}_k - \bar{X})^2 \right]^{1/2} \tag{11}$$

Where $\bar{X}_k$ is kth cluster average and X is average of overall clusters.

3. Results and Discussion

3.1. Variables Description

Before clusters of countries using *k*-medians clustering, the average, median and standard deviations of each variable are calculated first. This calculation is done to calculate the confidence interval that will be used in classifying clusters. The calculation results can be seen in Table 1.

Table 1. Descriptive of variables

Variables	Mean	Median	Standard deviation
Institutions	64.07	62.2	15.11
Human Capital and Research	32.57	30.45	15.62
Infrastructure	45.19	45.15	12.66
Market sophistication	48.04	46.8	10.73
Business Sophistication	33.89	30.25	12.34
Knowledge and Technology Outputs	26.58	23.1	13.88
Creative Outputs	30.41	28.35	13.15

3.2. Silhouette coefficient value

The value of the silhouette coefficient is obtained by using software R and shown in Table 3.2 These values show how good the grouping process and the quality of the group formed.

Table 2. Silhouette coefficient value

K	Silhouette Coefficient (*k*-means)	Silhouette Coefficient (*k*-medians)
3	0,4485	0,4730

4	0,2859	0,2581
5	0,2806	0,2237
6	0,3087	0,3032

Based on Table 2 it can see that the highest value of silhouette coefficient on each cluster is K = 3. Therefore, the study uses 3 clusters

3.3. Cluster Analysis

3.3.1. Outlier Detection and Sample Representing the Population

Using R application obtained, the data has an outlier. The value of Kaiser-Meyer-Olkin Measure of Sampling Adequacy is 0.919. The KMO value of 0.919 ranges from 0.5 to 1, it can be concluded that the sample can represent the population and variables can be used for further analysis [17].

3.3.2. Multicollinearity Assumption

All values of VIF are less than five. Based on the results show that VIF in variables $\boldsymbol{X_1, X_2, X_3, X_4, X_5, X_6}$, and $\boldsymbol{X_7}$, there is no multicollinearity.

3.3.3. Cluster Results using K-Means Clustering

After cluster analysis using the *k*-means method, obtained 3 clusters of countries based on the Global Innovation Index.

Table 3. Cluster results using *k*-means clustering

Cluster 1	Albania, Argentina, Bosnia and Herzegovina, Bulgaria, Bahrain, Belarus, Brunei Darussalam, Brazil, Chile, Colombia, Costa Rica, Croatia, Georgia, Greece, Hungary, India, Iran, Jordan, Kazakhstan, Kuwait, Latvia, Lithuania, Moldova, Montenegro, Mongolia, Mauritius, Malaysia, Mexico, Morocco, Namibia, Oman, Panama, Peru, Poland, Qatar, Romania, Russia, Saudi Arabia, Serbia, Slovakia, South Africa, Thailand, Turkey, The former Yugoslav Republic of Macedonia, Tunisia, Ukraine, Uruguay, Vietnam
Cluster 2	Algeria, Armenia, Azerbaijan, Bangladesh, Benin, Bolivia, Botswana, Burkina Faso, Cambodia, Cameroon, Côte d'Ivoire, Dominican Republic, Ecuador, Egypt, El Savador, Ghana, Guinea, Guatemala, Honduras, Indonesia, Jamaica, Kenya, Kyrgyzstan, Lebanon, Madagascar, Malawi, Mali, Mozambique, Niger, Nigeria, Nepal, Philippines, Pakistan, Paraguay, Rwanda, Senegal, Sri Lanka, Tajikistan, Tanzania, Togo, Trinidad and Tobago, Uganda, Yemen, Zambia, Zimbabwe
Cluster 3	AE United Arab Emirates, Australia, Austria, Belgium, Canada, China, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Hong Kong (China), Ireland, Israel, Iceland, Italy, Japan, Korea, Luxembourg, Malta, Netherlands, New Zealand, Norway, Portugal, Singapore, Slovenia, Spain, Sweden, Switzerland, United Kingdom, United States of America

Based on Table 3, we can find out the results of grouping using the *k*-means algorithm using Euclidean distance, which is in Cluster 1 consists of 48 Countries, Cluster 2 consists of 45 Countries and Cluster 3 consists of 33 Countries. Then to differentiate the cluster results that is formed, it is necessary to do profilization by calculating the average value of each variable on Table 4. The result as follows:

Ta**ble 4.** Characteristic cluster *k*-means

Variable	Cluster 1	Cluster 2	Cluster 3
Institutions	62.960	50.551	84.112
Human Capital and Research	33.421	17.791	51.485
Infrastructure	46.385	32.296	61.030

Market sophistication	47.083	40.413	59.842
Business Sophistication	30.156	25.029	51.412
Knowledge and Technology Outputs	24.290	15.613	44.879
Creative Outputs	29.740	18.727	47.330

Based on Table 4, it can be known the characteristics of each cluster. Cluster 1 has the average value of seven variables are quite high. Cluster 2 has the average value of seven variables are low, whereas Cluster 3 has the average value of seven variables are the highest.

3.3.4. Cluster Results using K-Medians Clustering

Using R application to find *k*-medians cluster, obtained 3 clusters of countries based on the Global Innovation Index as shown in Table 5.

Table 5. Cluster results using K-Medians Clustering

Cluster 1	AE United Arab Emirates, Australia, Austria, Belgium, Canada, China, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Hong Kong (China), Ireland, Israel, Iceland, Italy, Japan, Korea, Luxembourg, Malta, Netherlands, New Zealand, Norway, Portugal, Singapore, Slovenia, Spain, Sweden, Switzerland, United Kingdom, United States Of America
Cluster 2	Albania, Argentina, Armenia, Azerbaijan, Bosnia And Herzegovina, Bulgaria, Bahrain, Belarus, Brunei Darussalam, Botswana, Brazil, Chile, Colombia, Costa Rica, Croatia, Georgia, Greece, Hungary, India, Iran, Jamaica, Jordan, Kazakhstan, Kuwait, Latvia, Lithuania, Moldova, Montenegro, Mongolia, Mauritius, Malaysia, Mexico, Morocco, Namibia, Oman, Philippines, Panama, Peru, Poland, Qatar, Romania, Russia, Saudi Arabia, Serbia, Slovakia, South Africa, Thailand, Turkey, The Former Yugoslav Republic Of Macedonia, Tunisia, Ukraine, Uruguay, Vietnam
Cluster 3	Algeria, Bangladesh, Benin, Bolivia, Burkina Faso, Cambodia, Cameroon, Côte D'ivoire, Dominican Republic, Ecuador, Egypt, El-Savador, Ghana, Guinea, Guatemala, Honduras, Indonesia, Kenya, Kyrgyzstan, Lebanon, Madagascar, Malawi, Mali, Mozambique, Niger, Nigeria, Nepal, Pakistan, Paraguay, Rwanda, Senegal, Sri Lanka, Tajikistan, Tanzania, Togo, Trinidad and Tobago, Uganda, Yemen, Zambia, Zimbabwe

Based on Table 5 we can find out the results of grouping using the *k*-means algorithm using Euclidean distance, which is in Cluster 1 consists of 33 Countries, Cluster 2 consists of 53 Countries and Cluster 3 consists of 40 Countries. Then to differentiate the cluster results that is formed, it is necessary to do profilization by calculating the average value of each variable on Table 6. The result as follows:

Table 6. Characteristic Cluster K-Median

Variable	cluster 1	cluster 2	cluster 3
Institutions	84.112	62.851	49.533
Human Capital and Research	51.485	32.364	17.218
Infrastructure	61.030	45.649	31.745
Market sophistication	59.842	46.789	40.060
Business Sophistication	51.412	30.008	24.590
Knowledge and Technology Outputs	44.879	23.874	15.090
Creative Outputs	47.330	29.249	18.098

Based on Table 6, it can be known the characteristics of each cluster. Cluster 1 has the average value of seven variables are the highest. Cluster 2 has the average value of seven variables are quite high, whereas Cluster 3 has the average value of seven variables are low.

3.3.5. Determining the Goodness of The Clustering Method with Standard Deviation on K-Means
Standard deviation of cluster 1, for the variable mean in each country, where the value is $\bar{x}_I$ = 39.15.

$$S_1 = \sqrt{\frac{(\bar{x}_1 - \bar{x}_I)^2 + \ldots + (\bar{x}_{48} - \bar{x}_I)}{K-1}}$$
$$S_1 = 3.719234$$

Standard deviation of cluster 2, for the variable mean in each country, where the value is $\bar{x}_{II}$ = 28.63.

$$S_2 = \sqrt{\frac{(\bar{x}_1 - \bar{x}_{II})^2 + \ldots + (\bar{x}_{45} - \bar{x}_{II})}{K-1}}$$
$$S_2 = 3.785309$$

Standard deviation of cluster 2, for the variable mean in each country, where the value is $\bar{x}_{III}$ = 57.15.

$$S_3 = \sqrt{\frac{(\bar{x}_1 - \bar{x}_{III})^2 + \ldots + (\bar{x}_{33} - \bar{x}_{III})}{K-1}}$$
$$S_3 = 5.687154$$

So, the standard deviation value in a cluster using the *k*-means method is

$$S_w = \frac{3.719234 + 3.785309 + 5.687154}{3}$$
$$S_w = 4.397232$$

$$\bar{X} = \frac{\bar{x}_I + \bar{x}_{II} + \bar{x}_{III}}{3} = 41.64508$$
$$S_B = \left[(3-1)^{-1} \sum_{k=1}^{K} (\bar{X}_k - \bar{X})^2\right]^{1/2}$$
$$S_B = 14.42526$$

So, the ratio value of standard deviation in cluster and between clusters using *k*-means method is:

$$varians = \frac{S_w}{S_B} = 0.3048$$

3.3.6. Determining the goodness of the clustering method with standard deviation on k-medians
Standard deviation of cluster 1, for the variable mean in each country, where the value is $\bar{x}_I$ = 28.05.

$$S_1 = \sqrt{\frac{(\bar{x}_1 - \bar{x}_I)^2 + \ldots + (\bar{x}_{33} - \bar{x}_I)}{K-1}}$$
$$S_1 = 5.687154$$

Standard deviation of cluster 2, for the variable mean in each country, where the value is $\bar{x}_{II}$ = 38.68.

$$S_2 = \sqrt{\frac{(\bar{x}_1 - \bar{x}_{II})^2 + \ldots + (\bar{x}_{53} - \bar{x}_{II})}{K-1}}$$
$$S_2 = 3.825788$$

Standard deviation of cluster 2, for the variable mean in each country, where the value is $\bar{x}_{III}$ = 57.15.

$$S_3 = \sqrt{\frac{(\bar{x}_1 - \bar{x}_{III})^2 + \ldots + (\bar{x}_{40} - \bar{x}_{III})}{K-1}}$$
$$S_3 = 3.622979$$

So, the standard deviation value in a cluster using the *k*-means method is

$$S_w = \frac{5.687154 + 3.825788 + 3.622979}{3}$$
$$S_w = 4.37864$$

$$\bar{X} = \frac{\bar{x}_I + \bar{x}_{II} + \bar{x}_{III}}{3} = 41.29556$$
$$S_B = \left[(3-1)^{-1} \sum_{k=1}^{K} (\bar{X}_k - \bar{X})^2\right]^{1/2}$$
$$S_B = 14.72897$$

So, the ratio value of standard deviation in cluster and between clusters using *k*-medians method is:

$$varians = \frac{S_w}{S_B} = 0.2973$$

From the results of all clusters using the K-means and *k*-medians methods, cluster validation is sought for both methods using cluster variance values. the cluster variance value will get better when the value gets smaller.

4. Conclusion

The results of the study provided the conclusion based on the analysis that had been carried out, 3 clusters were formed on each method. Cluster 1 in *k*-means method consists of 48 Countries, Cluster 2 consists of 45 Countries and Cluster 3 consists of 33 Countries. Based on the average value, Cluster 1 has the average value of seven variables are quite high. Cluster 2 has the average value of seven variables are low, whereas Cluster 3 has the average value of seven variables are the highest. Furthermore, Cluster 1 in *k*-medians method consists of 33 Countries, Cluster 2 consists of 53 Countries and Cluster 3 consists of 40 Countries, and Cluster 1 has the average value of seven variables are the highest. Cluster 2 has the average value of seven variables are quite high, whereas Cluster 3 has the average value of seven variables are low. From the research that had been done, the result clustering with using *k*-means method and *k*-medians method showed that *k*-medians are better than *k*-means method because the varians value of *k*-medians = 0.297 is smaller than *k*-means = 0.305.

References

[1]. S. Dutta, B. Lanvin, and S. Wunsch-Vincent, eds. The global innovation index 2018: Energizing the world with innovation. WIPO, 2018.

[2]. T. S. Madhulatha. "An overview on clustering methods." IOSR Journal of Engineering: Vol 2 No. 4, 719- 725, 2012.
[3]. J. F. Hair, W. C. Black, B. J. Babin, R. E. Anderson, and R. L. Tatham. "Multivariate data analysis 6th Edition." Pearson Prentice Hall. New Jersey. humans: Critique and reformulation. Journal of Abnormal Psychology 87, 49-74, 2006.
[4]. S. Santoso, "Aplikasi SPSS pada statistik multivariat." Jakarta: PT Elex Komputindo (2012).
[5]. D.N. Gujarati. Basic Econometrics" fourth edition McGraw-Hill. New York. 2003
[6]. A. Chauhan, G. Mishra, and G. Kumar. "Survey on data mining techniques in intrusion detection." International Journal of Scientific & Engineering Research 2, no. 7, 1-4, 2011.
[7]. I. Riadi, "Internet forensics framework based-on clustering." International Journal of Advanced Computer Science and Applications 4, no. 12, 115-123, 2013.
[8]. N. S. Ediyanto, and M. N. Mara. "Characteristics classification by Method K-Means Cluster Analysis." Bul. Ilm 2, no. 2, 133-136, 2013.
[9]. L. Kaufman, and P. J. Rousseeuw. Finding groups in data: an introduction to cluster analysis. John Wiley & Sons, 2009.
[10]. R. Handoyo, R. Mangkudjaja, and S. M. Nasution. "Perbandingan metode clustering menggunakan metode Single Linkage dan K-means pada Pengelompokan Dokumen." Jurnal Sifo Mikroskil 15, no. 2, 73-82, 2014.
[11]. J. A. Hartigan, Clustering algorithms. John Wiley & Sons, Inc., 1975.
[12]. A. C. Rencher, "Methods of multivariate analysis. Canada: John Willey & Sons." Inc. Publications 2002.
[13]. K. Teknomo, "K-means clustering tutorial." Medicine 100, no. 4, 3, 2006.
[14]. B. J. Anderson, D. S. Gross, D. R. Musicant, A. M. Ritz, T. G. Smith, and L. E. Steinberg. "Adapting k-medians to generate normalized cluster centers." In Proceedings of the 2006 SIAM International Conference on Data Mining, pp. 165-175. Society for Industrial and Applied Mathematics, 2006.
[15]. M. R. Ackermann, J. Blomer, and C. Sohler. "Clustering for Metric and Non-Metric Distance Measures (full version)." 2009.
[16]. M. J. Bunkers, J. R. Miller, and A. T. DeGaetano. "Definition of climate regions in the Northern Plains using an objective cluster modification technique." Journal of Climate 9, no. 1, 130-146, 1996.
[17]. A. N. Fathia, R. Rahmawati, and T. Tarno. "Analisis klaster kecamatan di kabupaten semarang berdasarkan potensi desa menggunakan metode ward dan single linkage." Jurnal Gaussian 5, no. 4, 801-810, 2016.