Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 216 (2023) 356–363

Procedia Computer Science

www.elsevier.com/locate/procedia

7th International Conference on Computer Science and Computational Intelligence 2022

Clustering models for hospitals in Jakarta using fuzzy c-means and k-means

Karli Eka Setiawan[a,*], Afdhal Kurniawan[a], Andry Chowanda[a], Derwin Suhartono[a]

[a]*Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480*

Abstract

After facing the COVID-19 pandemic, national and local governments in Indonesia realized a gap in the distribution of health care and human health practitioners. This research proposes two unsupervised learning methods, K-Means and Fuzzy C-Means (FCM), for clustering a list of hospital data in Jakarta, Indonesia, which contains information about the number of its human health resources. The datasets used in this study were obtained from the website the Ministry of the Health Republic of Indonesia provided through the content scraping method. The result shows that implementing K-Means and FCM clustering results in the same number of clusters. Nevertheless, both results have different areas and proportions that can be observed by three distance metrics, such as Hamming, Euclidean, and Manhattan distance. By using the clustering result using the K-Means algorithm, the hospital list was separated into three clusters with a proportion of 84.82%, 14.66%, and 0.52% for clusters 0, 1, and 2, respectively. Meanwhile, using the FCM algorithm, the hospital list was separated into three clusters with a proportion of 17.80%, 73.82%, and 8.38% for clusters 0, 1, and 2, respectively. To the best of our knowledge, this is the first discussion of clustering healthcare facilities in Indonesia, especially hospitals, based on their health professionals.

Keywords: healthcare clustering; k-means; fuzzy c-means;

1. Introduction

The development of the health sector is a part of the national or local government's concern by providing adequate health facilities and increasing life expectancy because health is one of the fundamental rights for society to be

* Corresponding author. Tel.: +62-21-534-5830.
E-mail address: karli.setiawan@binus.ac.id

1877-0509
10.1016/j.procs.2022.12.146

actualized[1]. Data collection and performance analysis in healthcare facilities are essential for future government policy, especially in health policy. Indonesia, a nation with more than 200 million residents, faces a problem distributing healthcare facilities and human health resources such as doctors, nurses, midwives, and others, all of which are concentrated in metropolitan areas[2]. When facing the COVID-19 pandemic, the available medical staff in Indonesia was insufficient, and there was an increasing healthcare demand[3]. Currently, there is limited access to hospital data in Indonesia. Therefore, it is a daunting task for society to determine the level of the hospital in Indonesia. Moreover, the data regarding the hospital, such as the number of specialists, nurses, doctors, beds, ICUs, and emergency services in a particular hospital, are scattered all over the place. Hence, this research aims to collect and combine all the data and place it in one place as a dataset and use machine learning to automatically cluster the level of the hospital in Indonesia. As the first attempt to model the hospital cluster, we limited the hospital to only hospitals that are located in Indonesia's capital city, Jakarta. This study may help national and local governments make decisions related to public health policy by grouping the list of hospitals in Jakarta into some clusters[4]. The research contributions are the creation of a dataset containing a list of hospitals in Jakarta, the capital of Indonesia, which contain the number of various human health personnel and the implementation of unsupervised learning algorithms for clustering hospital data based on human health workers.

This research implemented two unsupervised algorithms, K-Means and FCM, for comparison purposes. K-Means was used because it is a widely applicable clustering technique that seeks to build close-fitting clusters and works well with low-dimensional data. When the number of variables is enormous, the K-means algorithm outperforms hierarchical clustering [5]. Meanwhile, the FCM can categorize data points belonging to a large number of clusters, which can benefit with precise and robust data recovery without any loss of data information because of the fuzzy algorithm, which is the advantage of FCM compared to other clustering methods [6]. The rest of the paper is organized as follows: the related work related to this research area is illustrated in the next section. The proposed research methodology is comprehensively shown in the Research Methodology section. The results are thoroughly discussed in the Experiment Results section. Finally, the conclusion and future work directions are presented in the last section.

2. Related Works

In the medical field, the clustering approach, which is a part of unsupervised learning, has been proven worthwhile in revealing the subgroups from a vast amount of data[7]. The clustering approach has been implemented in medical areas such as diagnosing various disorders; breast cancer, Parkinson's, headaches, mental illness, heart disease, diabetes, and many others. The research was done by Alashwal et al., studying a clustering of Alzheimer's disease by implementing K-Means, K-Means-Mode, multi-layer clustering, and hierarchical agglomerative clustering. Their research can determine the stage of Alzheimer's disease, which medical practitioners cannot resolve. Another study about clustering in the medical field was done by Rochman et al., by clustering many groups of tuberculosis patients based on many attributes such as age, gender, the infection status of human immunodeficiency virus, the history of diabetes mellitus, x-ray results, and Molecular Rapid Test results with K-Means and Fuzzy C-Means (FCM) approach[8]. Combined unsupervised learning such as K-Means and FCM were also used for detecting brain tumors in the research done by Alam et al.[5]. The data they processed was the collection of human brain images containing brain tumors. The unsupervised models can be used to detect the existence of brain tumors.

Another implementation of both K-Means and FCM clustering in the medical field was done by Reddy et al. by implementing it to the big data containing health information such as universal health, E-medical record, societal health, clinical information, and picture information[6]. The results of their research show that FCM was more efficient than the K-Means approach, even though the scenario in which they compared them was untold. Based on the previous study, this research was inspired to cluster the collection data of many hospitals in Jakarta, Indonesia, containing information about the number of their human health resources by implementing and comparing hard clustering with K-Means and FCM. To the best of our knowledge, this is the first-time discussion of clustering a healthcare facility based on their health professionals, especially a hospital.

Another related research using K-means and FCM clustering was done by Hamur et al. by clustering all the provinces in Indonesia based on an indicator of quality healthcare in 2015 [9]. The dataset used by Hamur et al. was the data obtained from the National Socio-Economic Survey (Susenas) in 2015 contained the information health centers, hospitals, doctors, nurses, midwifes, health workers, dentists, public health personnel, environmental health workers, and financial information. The results showed that the FCM method outperformed the K-Means method with a smaller ratio value of standard deviation in groups and between groups. A different comparison study related to K-Means and FCM done by Dana et al. was about measuring the performance of healthcare services based on patient satisfaction [10]. The dataset used was questionnaires from 3,000 patients obtained from Cirebon City Health Office containing information about 21 healthcare centers and their indicators. The result of Dana et al. research showed that K-Means was superior to Fuzzy C-Means.

3. Research Methodology

3.1. Datasets

The datasets used in this research were obtained from web scraping processes using HTML parsing technique from the website named "Badan PPSDM Kesehatan Informasi SDM Kesehatan Kementerian Kesehatan Republik Indonesia versi 4.0", which is under and responsible to the Ministry of Health Republic Indonesia. The information about the number of human health resources in every hospital in Indonesia is publicly available. Web scraping is a legally effective way to collect and extract data automatically, which works like a bot scraper on the internet from the website, then it creates structured databases based on the results of the scaping process[11]. It is legal because it does not breach any privacy data but collects public data on websites like regular visitors. This research collected data from the Board for Development and Empowering Human Health Resources 4.0 version websites, which has 11 tables, such as specialized doctors, clinical psychology, nursing, midwifery, pharmacy, public health, environmental health, nutrition, physical therapy, medical technicians, and biomedical engineering. Then the result of web scraping was transformed into one table containing a column list described as the health human resources type list in Table 1.

Table 1. Dataset tables information

Table Names	Health human resources type	Unique type
specialized doctors	General practitioners, dentists, 42 different kinds of specialized doctors, 10 different kinds of specialized dentists, and 1 sub-specialist doctor.	55
clinical psychology	Clinical psychology.	1
nursing	Nurse, non-nurse, child health nurse, maternity nurse, medical surgical nurse, geriatric nurse for elderly, mental health nurse, community nurse, and SPK nurse.	10
midwifery	Clinical, village, teaching, general, and other midwives.	5
pharmacy	Pharmacist, non-pharmacist, pharmacist assistant, pharmaceutical analysis, and Pharmacy SMF.	5
public health	Health epidemiologist, health promotion, behavioral science, occupational health, health administration and policy, biostatistics and populations, reproduction and family, health informatics, and general public health.	9
environmental health	Environmental sanitation, health entomologist, health microbiologist, and environmental health.	4
nutrition	Nutritionist, dietitian, and nutritionist auxiliary.	3
physical therapy	Physiotherapist, occupational therapist, speech therapist, and acupuncture.	4
medical technicians	Medical recorder and health information, cardiovascular technician, blood service technician, optometric refractionist, dental technician, anesthetist, dental and oral therapist, audiologist, and Dental and oral therapist.	9
biomedical engineering	Radiographer, electromedical, medical laboratory technologist, medical physicist, radiotherapist, prosthetic orthotic, and health analyst	7

Table 1 shows the dataset containing information about the quantity of every type of human health resource in every hospital on the list. This research bagged 191 hospital ID lists in Jakarta City as the primary key for scraping the Board for Development and Empowering Human Health Resources 4.0 version websites. This research used a python library, BeautifulSoup, to collect the data from the website. The dataset also contained information about the number of beds in every hospital.

3.2. Principal Component Analysis (PCA)

It was challenging to illustrate the dataset when it contained enormous dimensions, primarily since this research studied a dataset containing hundreds of dimensions. Principal Component Analysis (PCA) is a mathematical method that can be implemented to decrease the number of dimensions in a dataset [12]. The transformed data with PCA can be clarified quickly for visualization and further analysis. This research reduced the dataset dimension from 55 specialized doctors into one dimension, which is expressed as the Doctors axis; 57 other human health resources included clinical psychology, nursing, midwifery, pharmacy, public health, environmental health, nutrition, physical therapy, medical technician, and biomedical engineering into one dimension which is expressed as Supports axis.

$$x_n = W^T(t_n - \bar{t}\,) \tag{1}$$

The result of dimensional reduction with PCA is illustrated in equation (1), denoted by $\mathrm{x_n}$ [13]. The symbol of W represents the principal axes with $\mathrm{W} = (w_1, w_2, \ldots, w_q)$ where q is orthonormal axes obtained when variance under projection is maximal, $\mathrm{t_n}$ represents a set of observed d-dimensional data vectors with n∈{1...N}, and $\bar{\mathrm{t}}$ represents sample mean.

The PCA result in the Doctors dimension has 132.13, -17.53, and 5.95×10^{-16} value for maximum, minimum, and mean values, respectively. The PCA result in the Support dimension has $1{,}456.77$, -111.82, and 4.6×10^{-15} for maximum, minimum, and mean values, respectively. The Bed dimension describes the actual number of bed amounts with 968, 0, and 97.14 for maximum, minimum, and mean values, respectively.

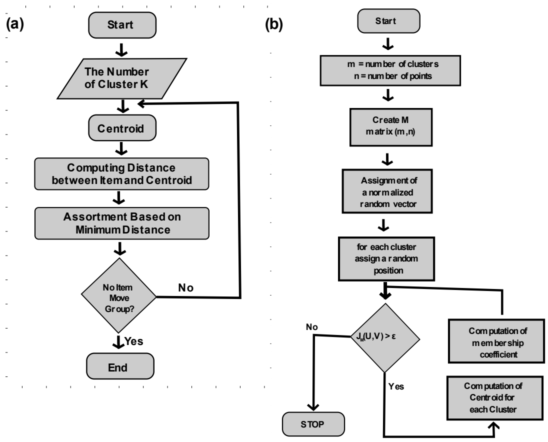

Fig 1. (a) K-Means Clustering Flowchart; (b) FCM Clustering Flowchart.

3.3. K-Means Clustering

K-Means clustering is one of unsupervised learning, which is centroid-based by partitioning a set of data into cluster K[14] [15]. This clustering algorithm needs an appropriate number of clusters k, since the initial centroid may change, impacting inconsistent grouping data results. The Sum Squared Error (SSE) can be used to decide the number of appropriate centroids by implementing the elbow method[16]. The set C of K Clusters Cj is the primary goal of K-means clustering, which is explained in equation (2) with the calculation of SSE.

$$E = \sum_i^k = 1 \sum xi \in cj \; \|cj - xi\|^2 \tag{2}$$

E is the alteration of SSE, where cj represents the cluster mean, $\|\ldots\|$ refers to the distance metrics between the mean and data point. The flowchart of K-means clustering can be seen in Fig. 1(a)[16]. This research used KMeans on the sklearn library to perform the clustering process.

3.4. Fuzzy C-Means (FCM)

The fuzzy C-Means clustering algorithm, which Bezdek introduced in 1981, is a type of unsupervised learning with the concept that the more data is near the cluster center, the more the membership value of data will become part of that cluster[5]. The main benefit of FCM implementation is that each data point can have a degree in [0,1] of cluster membership, allowing the data to be expressed in more than one cluster[17].

$$J_m(U,V) = \sum_{j=1}^{N}\sum_{i=1}^{C}\left(u_{i,j}\right)^m\left(\left\|x_j - v_i\right\|\right) \quad \text{Where } 1 \leq m \leq \infty \tag{3}$$

The FCM algorithm repetitively optimizes $J_M(U,V)$ with the continuously updated value of U, the membership function, and V, the cluster feature center. Where $\left\|x_j - v_i\right\|$ describes the Euclidean distance between x_j as a data point and v_i as the data centroid; N describes the number of feature vectors; C describes the number of clusters, and m is any actual number greater than 1[5] [18]. The FCM algorithm is depicted in Figure 1(b). This research implemented FCM clustering by using FCM on the fcmeans library.

3.5. Distance Measurement

This study implements three different distance measurements such as Hamming, Euclidean, and Manhattan distances. The hamming distance is used in this research because it can represent the number of different digits of locations[19]. The Hamming distance calculation is described in equation (4), which ⊕ is a summation of mod two from two data, centroid and data points.

$$D(x,y) = \sum_{k=1}^{n} x_k \oplus y_k \tag{4}$$

Another distance measurement is Euclidean distance, denoted in equation (5), where the calculation is based on the distance of two data in Euclidean space[20]. Euclidean distance is capable of being implemented for more than two dimensions. For example, equation (5) describes that d_{ij} is the distance between i and j, where in this research i and j represent the centroid and data points, and n represents the amount of data.

$$d_{ij} = \sqrt{\sum_{k=1}^{n}(x_{ik} - y_{jk})^2} \tag{5}$$

The last distance metric was used as the Manhattan distance. This distance measurement calculates the absolute differences between the two data[20]. Equation (6) describes the calculation of the Manhattan distance, where the Manhattan distance is symbolized as d_{ij}, the symbols of i and j illustrate the centroid and data point, and the symbol of n represents the total amount of data.

$$d_{ij} = \sum_{k=1}^{n}\left|x_{ij} - y_{ik}\right| \tag{6}$$

4. Experiment Results

4.1. Clustering Results with K-Means

K-Means and FCM clustering models implement the same default hyperparameter for comparing apple-to-apple, such as random init, n_init is defined by 10, 300 maximum iterations, and random_state is defined by 42. This study searched by testing SSE values ranging from one to ten clusters. The best number of clusters can be decided by implementing the elbow method, which can be seen in Fig. 2(a), where the elbow is located in 3 clusters. This research also used KneeLocator from the kneed library to ensure the best number of clusters. The KneeLocator also showed that 3 clusters are the appropriate number of clusters.

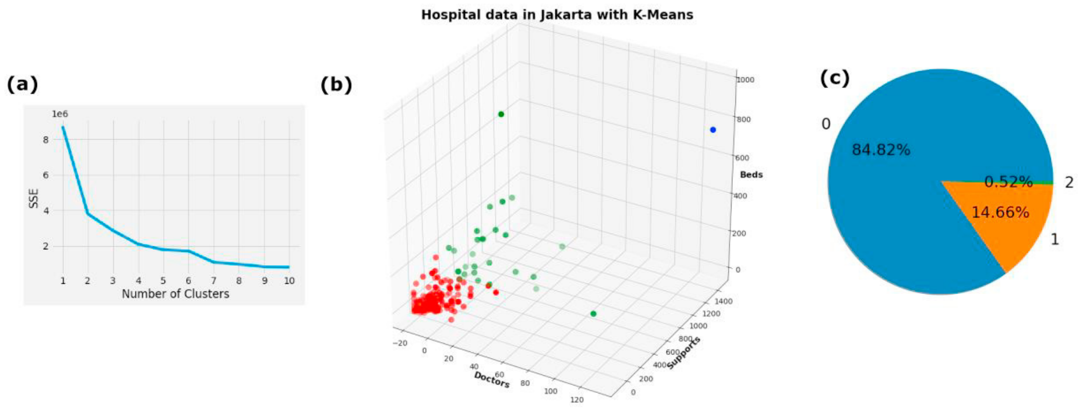

Fig 2. (a) SSE values from 1 cluster until 10 clusters; (b) Clustering Results with K-Means; (c) The Proportion of Data Distribution with K-Means.

Fig. 2(c) illustrates that 28 hospitals in Jakarta were grouped into cluster 0. As many as 162 hospitals in Jakarta are grouped into cluster 1, and only one hospital is exclusively clustered into cluster 2. The only hospital clustered into cluster 2 was the best in Jakarta, which has many beds, doctors, and other human health resources.

4.2. Clustering Results with Fuzzy C-Means

This study also searched by testing Partition Coefficient (PC) values from one to ten clusters. The PC was used to measure the overlap between clusters[21]. However, the PC is only applicable to fuzzy algorithms. The drawback of using a PC is that the PC cannot contain information about the direct connection of properties between data. The results of the PC are depicted in Fig. 3(a). KneeLocator from the kneed library is also used to find the best number of clusters based on PC values in Fig. 3(a) by showing that the best is 3 clusters. Fig. 3(c) depicted that 141 hospitals in Jakarta were categorized as cluster 2, colored blue, 34 hospitals were classified as cluster 1, and 16 hospitals were grouped into cluster 0.

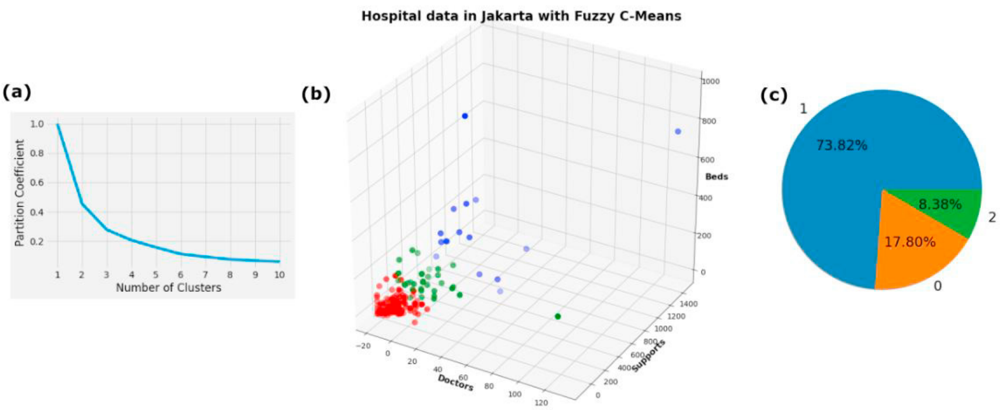

Fig 3. (a) SSE values from 1 cluster until 10 clusters; (b) Clustering Results with FCM; (c) The Proportion of Data Distribution with FCM.

4.3. Discussion

K-Means and FCM clustering methods resulted in the same number of clusters for their best outcome of three clusters. A glance at Fig. 3(b) shows that FCM clustering splits three into clusters based on the supports axis, whereas

K-Means clustering in Fig. 2(b) is based on data adjacency. To compare both clustering methods, this investigation used Hamming, Euclidean, and Manhattan distances to calculate the average distance of every piece of data with its centroid in each cluster and measure the distance between the farthest data to its centroid. A glance at Table 2 shows that the average distance between the data and its centroid with K-means methods was closer than the average distance between the data and its centroid with FCM. Table 2 noted the distance between the centroid and the furthest data.

Table 2. Distance between centroid and data with K-Means and FCM Clustering.

Algorithm	Cluster	Average Distance			Longest Distance		
		Hamming Distance	Euclidean Distance	Manhattan Distance	Hamming Distance	Euclidean Distance	Manhattan Distance
K-Means	Cluster 0	3.0	59.3299	84.1922	3.0	216.6310	297.8398
	Cluster 1	3.0	187.2088	251.8398	3.0	698.2779	751.6567
	Cluster 2	0	0	0	0	0	0
Fuzzy C-Means	Cluster 0	3.0	85.3380	117.7903	3.0	201.7648	274.2183
	Cluster 1	3.0	42.3850	59.1896	3.0	181.0295	207.0324
	Cluster 2	3.0	276.5828	358.9199	3.0	1114.7667	1517.7325

5. Conclusion and Future Work

Implementing K-Means and FCM forms a noticeably different number of clusters and cluster areas. The number of clusters significantly affects the breadth area of a cluster. The hospital list was divided into three clusters using the K-Means technique, with a proportion of 84.82%, 14.66%, and 0.52% for clusters 0, 1, and 2, respectively. Meanwhile, the hospital list was divided into three clusters using the FCM method, with proportions of 17.80%, 73.82%, and 8.38% for clusters 0, 1, and 2, respectively. This study concludes that for grouping hospital lists with Jakarta's hospital dataset based on human health resources, the K-Means method is more appropriate than the FCM method. However, the performance of clustering methods depends on the dataset's characteristics, and there is no correct answer for discovering clustering in unsupervised learning. The number of clusters with the K-Means method was decided by calculating the sum of squares method, and the number of sets with FCM was determined by calculating the partition coefficient. In future research, many approaches, such as the gap statistic, silhouette score, and cluster, can be used to determine the number of appropriate clusters. Besides K-means and clustering, other clustering approaches can be explored in the future, such as connectivity-based with hierarchical clustering, density-based with Density-Based Spatial Clustering of Application with Noise (DBScan), and distribution-based with Gaussian Mixture Model. The limitation of this work is that the models only trained with the dataset of Jakarta's hospital. The algorithms should be able to handle all of Indonesia's hospital data. First, the hospital data should be collected in one place as a dataset. The dataset should then be trained using the proposed algorithms. This will be the future work direction of this research.

Acknowledgements

This study is funded by the Directorate General of Higher Education, Ministry of Education, Culture, Research, and Technology (research project title: "SMAPHA, Smart Mapping for Public Health Application") as a part of the 2022 Matching Fund Program.

References

[1] Bakri H. The planning of community health center in Indonesia. Eur J Res Reflect Manag Sci [Internet]. 2018;6(3):12–8. Available from: http://www.idpublications.org/wp-content/uploads/2018/09/Full-Paper-THE-PLANNING-OF-COMMUNITY-HEALTH-CENTER-IN-INDONESIA.pdf

[2] Basri MH, Farmaciawaty DA, Rachmania IN, Widjaja FB, Wiyana AF. Healthcare Ecosystem Mapping in Primary Care: A Case Study in West Java, Indonesia. Proc Int Conf Bus Eng Manag (ICONBEM 2021). 2021;177:184–90.

[3] Mahendradhata Y, Andayani NLPE, Hasri ET, Arifi MD, Siahaan RGM, Solikha DA, et al. The Capacity of the Indonesian Healthcare System

to Respond to COVID-19. Front Public Heal. 2021;9(July):1–9.

[4] Rakmawati T, Hinchcliff R, Pardosi JF. District-level impacts of health system decentralization in Indonesia: A systematic review. Int J Health Plann Manage. 2019;34(2):e1026–53.

[5] Alam MS, Rahman MM, Hossain MA, Islam MK, Ahmed KM, Ahmed KT, et al. Automatic human brain tumor detection in mri image using template-based k means and improved fuzzy c means clustering algorithm. Big Data Cogn Comput. 2019;3(2):1–18.

[6] Reddy BR, Vijay Kumar Y, Prabhakar M. Clustering large amounts of healthcare datasets using fuzzy c-means algorithm. 2019 5th Int Conf Adv Comput Commun Syst ICACCS 2019. 2019;93–7.

[7] Alashwal H, El Halaby M, Crouse JJ, Abdalla A, Moustafa AA. The application of unsupervised clustering methods to Alzheimer's disease. Front Comput Neurosci. 2019;13(May):1–9.

[8] Rochman EM, Miswanto, Suprajitno H. COMPARISON OF CLUSTERING IN TUBERCULOSIS USING FUZZY C-MEANS AND K-MEANS METHODS. 2022;1–20.

[9] Hamur AS, Susetyo B, Indahwati. Comparing K-Means and Fuzzy C-Means Clustering (Case : Clustering of Provinces in Indonesia Based on the Indicator of the Health Service in 2015). Int J Eng Manag Res. 2017;7(3):387–90.

[10] Dana RD, Dikananda AR, Sudrajat D, Wanto A, Fasya F. Measurement of health service performance through machine learning using clustering techniques. J Phys Conf Ser. 2019;1360(1).

[11] Muehlethaler C, Albert R. Collecting data on textiles from the internet using web crawling and web scraping tools. Forensic Sci Int [Internet]. 2021;322:110753. Available from: https://doi.org/10.1016/j.forsciint.2021.110753

[12] Beattie JR, Esmonde-White FWL. Exploration of Principal Component Analysis: Deriving Principal Component Analysis Visually Using Spectra. Appl Spectrosc. 2021;75(4):361–75.

[13] Tipping ME, Bishop CM. Probabilistic principal component analysis. J R Stat Soc Ser B (Statistical Methodol. 2002;61(2):611–22.

[14] Syakur MA, Khotimah BK, Rochman EMS, Satoto BD. Integration K-Means Clustering Method and Elbow Method for Identification of the Best Customer Profile Cluster. IOP Conf Ser Mater Sci Eng. 2018;336(1).

[15] Yuan C, Yang H. Research on K-Value Selection Method of K-Means Clustering Algorithm. J. 2019;2(2):226–35.

[16] Jahwar A. Meta-Heuristic Algorithms for K-means Clustering: A Review. PalArch's J Archaeol Egypt/Egyptology [Internet]. 2021;17(7):7–9. Available from: https://archives.palarch.nl/index.php/jae/article/view/4630

[17] Bora DJ, Gupta DAK. A Comparative study Between Fuzzy Clustering Algorithm and Hard Clustering Algorithm. Int J Comput Trends Technol. 2014;10(2):108–13.

[18] Taherpour A, Cheshmeh Sefidi A, Bemani A, Hamule T. Application of Fuzzy c-means algorithm for the estimation of Asphaltene precipitation. Pet Sci Technol [Internet]. 2018;36(3):239–43. Available from: https://doi.org/10.1080/10916466.2017.1416632

[19] Yang K, Miao R. Research on Improvement of Text Processing and Clustering Algorithms in Public Opinion Early Warning System. 2018 5th Int Conf Syst Informatics, ICSAI 2018. 2019;(Icsai):333–7.

[20] Faisal M, Zamzami EM, Sutarman. Comparative Analysis of Inter-Centroid K-Means Performance using Euclidean Distance, Canberra Distance and Manhattan Distance. J Phys Conf Ser. 2020;1566(1).

[21] Mota VC, Damasceno FA, Leite DF. Fuzzy clustering and fuzzy validity measures for knowledge discovery and decision making in agricultural engineering. Comput Electron Agric [Internet]. 2018;150(January):118–24. Available from: https://doi.org/10.1016/j.compag.2018.04.011