INDONESIAN JOURNAL OF LIFE SCIENCES

Volume 5, No. 2, September 2023

RESEARCH ARTICLE

Comparative Study of k-Mean, k-Medoid, and Hierarchical Clustering Using Data of Tuberculosis Indicators in Indonesia

Nanda Rizqia Pradana Ratnasari*

Department of Bioinformatics, Institut Bio Scientia Internasional Indonesia, Jakarta, Indonesia
**Corresponding author: nanda.ratnasari@i3l.ac.id*

ABSTRACT

Cluster analysis is an important topic in which the ultimate goal is to classify data into several groups based on a similar basis. The most applied cluster methods or algorithms to classify data are k-means, k-medoids, and hierarchical clustering methods. Therefore, this study aimed to compare methods in cluster analysis employing healthcare data on attributes related to tuberculosis. The best method was assigned based on the level of accuracy for each algorithm and the number of clusters. There were four main steps in the clustering analysis used in this study, which were feature selection, clustering algorithm, cluster validation, and interpretation. The clustering algorithm used were k-means, k-medoids and hierarchical clustering, with cluster sizes of two, three, and four, respectively. The results showed that k-medoids had a higher accuracy than other clustering algorithms or methods for both training and testing data. K-medoid was better than the other two algorithms as it was more robust to noise and outliers which were found in the datasets. This outcome was consistent between the training and testing datasets. In terms of the number of clusters, the two-cluster model was better than the three-cluster or the four-cluster model, as this model could classify the groups more vividly. The results were consistent between k-mean, k-medoid, and hierarchical clustering methods, with the smallest sum of squares value being 24.7% for the k-mean. The smallest diameters and the average dissimilarities of k-medoid models were found in group 1. This result explained that group 1, in all algorithms, was more compact and more similar than other groups.

KEYWORDS

k-mean, k-medoid, Hierarchical clustering

HIGHLIGHTS

❖ Comparing clustering algorithm.

❖ Utilization of R studio packages.

❖ Application of clustering for Tuberculosis indicators.

❖ K-medoid has the highest accuracy compared to k-mean and hierarchical clustering.

INTRODUCTION

Cluster analysis is an important topic and analysis in data mining, machine learning, and statistic for decision-making. The ultimate goal of cluster analysis is to classify data into several groups based on a similar basis. The analysis is developed from various fields; such as mathematics, statistics, and computer science; and utilized in many different applications (Zhao & Zhou, 2021). Cluster assigned from cluster analysis can capture natural characters in a data (Punithavalli et al., 2010). Cluster analysis is frequently used to reveal hidden clusters and composition in large data; however, it was not widely applied in healthcare and medical applications (Liao et al., 2016). Some of the utilization of cluster analysis in medical contexts could be found in studies, such as identifying patient groups based on targeted intervention (Clatworthy, 2005), classifying psychiatric patients based on the symptoms of their diseases (Blashfield, 1984), categorizing groups of genes based on biological functions (Eisen et al., 1998), and categorizing pulmonary diseases based on the phenotypes (Koo, et al., 2022). The application of cluster analysis could help find the true features or attributes in a data to generate a hypothesis, make predictions based on groups, and even elaborate on data exploration (Liao et al., 2016).

There are several models of cluster analysis based on the techniques and methods applied to the data for creating the groups. In general, the models can be classified as the connectivity model, centroid model, distribution model, and density model. The connectivity model is applied for hierarchical clustering, which provides a tree of clusters. The tree categorizes the cluster from higher levels to lower levels, with the latter being a branch of the prior one. Objects in lower-level clusters normally have higher similarity than the ones in higher-level clusters. The centroid model determines the groups based on the geometric center (centroids), and partitions objects based on the nearest centers to the entities. The distribution model identifies clusters based on the distributions of objects in each cluster. Initially, objects or entities are considered as data with a mixture of probability distributions. Each cluster is defined as a set of objects or entities with the same distribution (Novoselsky & Kagan, 2021).

Cluster analysis has been employed to analyze diverse diseases, including chronic obstructive pulmonary diseases such as Tuberculosis (TB), as it can identify groups and subgroups of patients considered according to their features (Castaldi et al., 2014). Tuberculosis is one of the diseases with the highest rate of death worldwide. There were about 10 million people suffering from TB disease in 2019, including 1.2 million children and approximately 1.4 million death cases. Most TB cases were caused by poverty and vulnerability in the community (World Health Organization [WHO], 2020). In developing countries, TB remained a major health problem (Mahendradhata et al., 2003), accounting for approximately 95% of TB-related deaths (Mohajan, 2015). Tuberculosis (TB) remains a major public health and is known by the heterogenous phenotype of the disease, which makes it hard to diagnose. Besides, an efficient classification tool for TB has not been well-developed yet (Cadena et al., 2017). TB has also significantly contributed to disease burden, economic loss, and social problems (Noviyani et al., 2021). Numerous strategies have been implemented to prevent the increasing number of TB cases; however, the decline in TB infection trend is relatively slow (WHO, 2020). These conditions were supported by significant challenges such as gaps in funding, lack of access and treatment services, and limitation in resources (Lakoh et al., 2020). The World Health Organization (WHO) stated in their 2022 annual report that Indonesia was the second highest TB-burdened country in the world (WHO, 2021). The high rate and incidence of TB in Indonesia indicate that the prevention of the disease should be prioritized and controlled (Erawati & Andriany, 2020). As the characteristics of Indonesian people and geography are heterogenous, indicating regions with specific features related to TB cases can help the government provide appropriate treatments for each area. Therefore, cluster analysis becomes an important tool to classify Indonesian regions in regards to TB cases.

This study aimed to compare methods in cluster analysis that employ healthcare data on attributes related to TB. The best method was assigned based on the level of accuracy for each algorithm and the number of clusters. The best cluster method could be utilized to do appropriate predictions for tuberculosis. Therefore, it can later be used to execute improvements related to TB in the future.

Cluster Analysis

Cluster analysis is defined as a method to discover groups or clusters in data. It is aimed at uncovering previously unknown groups of objects and natural structures in a dataset. The developed clusters should be cohesive structures that are isolated from each other, yet have homogenous elements (Landau & Ster, 2010). Cluster analysis is also known as a technique for splitting a given dataset and variables into homogeneous groups or clusters, in which entities in one cluster have high similarities while entities in different clusters have significant differences. The classification in cluster analysis also discloses the hidden structures in the obtained data (Novoselsky & Kagan, 2021). As cluster analysis groups the objects based on the information describing the relationship of the data or variables, greater similarities within groups and greater differences between groups are produce better cluster analysis results (Gupta et al., 2012). Therefore, The main objective of cluster analysis is to determine the grouping process for unlabelled data (Punithavalli et al., 2010). As such, the criteria and number of groups are decided by the analysts to satisfy the requirements of the analysis (Strehl et al., 2000).

In terms of a multidimensional space, cluster analysis is assigned as a process of organization of patterns for a vector of measurements based on similarities. Patterns within valid clusters are more similar than patterns belonging to other clusters. As cluster analysis is an unsupervised classification, the problem of grouping a given collection of patterns into meaningful clusters may occur. Cluster analysis may be useful in explanatory analysis, such as decision-making, machine learning, and data mining. However, little availability of prior information regarding the patterns or initial clusters could lead to problems that require assumptions in the analysis (Jain et al., 1999).

Clusters or groups for classification in cluster analysis can be assigned based on the distance within and between the clusters. The distance between entities in one cluster, which is also known as intra-cluster, should be minimal to ensure that they are homogenous. Meanwhile, the inter-distance, addressing the distance between entities from different clusters should be at its maximum (Novoselsky & Kagan, 2021).

Cluster analysis as an unsupervised classification is different to discriminant analysis (supervised classification), which provides labelled patterns. The labels in supervised learning are used to learn the descriptions of clusters or classes in the data to label the patterns in the new given data (Jain et al., 1999). Cluster analysis is also different to factor analysis since cluster analysis has the main purpose of grouping the objects or entities in the data, while factor analysis aims to group the variables (Setyaningsih, 2012). Furthermore, the groups created by cluster analysis are made based on distances (proximity), while the groups assigned by factor analysis are developed by the pattern of variation (correlation) of the data (Hair et al., 2010).

There are no specific patterns and features of cluster analysis for certain conditions since the pattern-generation processes are frequently not controllable. Good data patterns can produce simple and understandable clustering. Similarity is a crucial aspect in the cluster analysis since it can measure the resemblances in clustering. Similarity between features and objects can be calculated using a distance measure accounted on the feature spaces (Jain et al., 1999). The most popular distance for continuous features is known as *Euclidean distance,* which can be formulated using the following formula:

$$d_p(x_i, x_j) = \left(\sum_{k=1}^{d} \left|x_{i,k} - x_{j,k}\right|^p\right)^{\frac{1}{2}} = \|x_i - x_j\|_p \qquad (1)$$

p = the dimensionality of the pattern spaces
x = a feature with individual scalar component x_i
p = the distance between two features
i, j = the index for objects or features

Euclidean distance is normally used to estimate objects' adjacency in two or three-dimensional spaces. This estimate works well for isolated clusters (Mao & Jain, 1995). Distortion of distance measures due to linear correlation can be alleviated by applying transformation to the data or using the squared Mahalanobis distance.

$$d_M(x_i, x_j) = (x_i, x_j)\Sigma^{-1}(x_i, x_j)^T \qquad (2)$$

where x_i and x_j are row vectors of the patterns or features, Σ is the sample covariance matrix of the pattern generation processes, and $d_M(. \ , \ .)$ is the different weights of different features based on the variance (Jain et al., 1999).

In general, there are two main methods in cluster analysis: the partitioning method and the hierarchical method. A clustering method is a strategy applied to categorizing or clustering problems. The partitioning method creates clusters by dividing the data into subsets or partitions based on some evaluation criteria. Meanwhile, the hierarchical method decomposes the dataset into a group hierarchy (Kaur & Attwal, 2014). There are two major subcategories of the partitioning methods: the centroid algorithm and the medoids algorithm. The centroid algorithm defines the clusters or groups by using the gravity centre (Punithavalli et al., 2010).

K-means clustering

K-means clustering is one of the algorithms in partitioning-based clustering methods. K-means defines a K-class and categorize the dataset into classes by adopting the standard of square errors (Zhao & Zhou, 2021). K-mean clustering is one of the known algorithms used in the centroid model, which is a model that represents clusters as a convex shape drawn around their centre, known as a centroid (Novoselsky & Kagan, 2021). K-means assigns every point in the data into a cluster whose centre is the nearest (Punithavalli et al., 2010). The centre of the k-means clustering is calculated using the arithmetic mean for each dimension separately over all points in the cluster (Cheung, 2003). K-means algorithms will separate the data into several partitions or clusters relying on a given initial centre or centroid. All the points in the data will be classified into a group with the nearest centre. The partitions will be updated with every point included in the clusters until the cluster is stable (Kanungo et al., 2002).

The k-means clusters can be identified using this algorithm with the following steps: (1) choosing the number of clusters *k*, (2) randomly generating clusters *k* and the centre of each cluster or generating *k* random points as the centre of clusters, (3) assigning each point in the data to the nearest centre or centroid, (4) updating the cluster and the centre using points in each cluster, (5) re-computing the new centre, and (6) repeating the process until the clusters are stable (Kanungo et al., 2002).

Every point is assigned to the nearest partition based on the closeness of the points to the centroid using Euclidean distance. The main purpose of the k-means algorithm is to minimize the mean squared distance from each data point to the nearest centroid. Points classified in the same cluster are considered to have similar parameters and characters (Kanungo et al., 2002). The correct number of *k* will lead to the best distinction of clusters and separation (Punithavalli et al., 2010).

K-medoids clustering

K-medoids can classify the data points into clusters using the medians of the data as the centre. K-medoids is a clustering technique categorized as a partition method. Compared to k-means, k-medoids are more robust as the *k* in k-medoids can minimize the dissimilarities of the data and reduce noise and outliers. The k-medoids method is used to diminish the weaknesses of the k-means algorithm (Arora et al., 2016). In general, k-medoids has higher accuracy, execution time, and time complexity of algorithm (Nurhayati et al., 2018). The medoids used in this method is defined as the most centrally located points in the data. Despite calculating the mean as conducted in k-means, the medoid is the actual point in the data (Xin Jin & Jiawei Han, 2010).

K-medoid is developed by assigning *k* representative points as the initial clusters' centroids. The centroids are the data medians, or most central points. The *k* clusters are constructed by setting up each point to the nearest representative medoids. Each point in the data will be part of the clusters with the closest distance. Once points are assigned to a specific cluster, the new medoids will be defined and the distance between points and each medoid are calculated. The processes are updated until the points in each cluster are stable (Kaufman & Rousseeuw, 2005). The distance between each point to the medoids is calculated using Euclidean distance (Fialine et al., 2021). Therefore, the location of each medoid changes accordingly with each iteration (Arora et al., 2016).

Hierarchical clustering

Hierarchical clustering is a data exploratory method that classifies data into groups by building a binary merge tree. The data is divided into two-by-two sub-sets, in which each subset consists of points with the closest distance or similarities. The tree separates the data by creating roots or branches until roots or branches that contain all elements are reached. Points with similar characteristics, which are calculated using Euclidean distance, are included in the same clusters (Nielsen, 2016). The hierarchical method develops the clusters by partitioning the data points into a top-down or bottom-up model. The method can be split into an agglomerative hierarchical cluster or a divisive hierarchical cluster. In the agglomerative hierarchical cluster, each point represents a single cluster. Then, the clusters with similar characteristics are sequentially integrated until the expected structures are obtained. Meanwhile, the divisive hierarchical cluster starts with a single cluster, which consists of all the data, before splitting whole points until the intended structures are constructed. The hierarchical method produces a graph which informs groups of data with similarities at different levels (Gupta & Jain, 2014).

The main benefit of hierarchical clustering is that it can provide multiple resolutions of grouping and clear visualisation of the cluster structure. Therefore, it can be easily applied and interpreted (Sharara & Getoor, 2010).

MATERIAL AND METHODS

This study used clustering analysis, which consisted of k-mean, k-medoid and hierarchical clustering, to classify cities in Indonesia based on the levels of tuberculosis rate. Furthermore, the study also compared the accuracy of each clustering method for tuberculosis cases to find the best method or algorithm for such cases. The data used for this study was health information related to tuberculosis and consisted of twelve indicators or attributes representing the condition of each city that could affect the development of tuberculosis cases, such as number of TB adult patients, number of TB children's patients, cure rate, etc. The original data consisted of 1593 observations; however, some data were removed due to incompleteness. The data type for all attributes were discrete.

In general, there were four main steps in the clustering analysis, which were: feature selection, clustering algorithm, cluster validation, and interpretation. (1) Feature selection, which was also known as extraction, referred to a process of choosing and picking unique features from the data set. It included the process of selecting appropriate attributes for the intended case. Feature selection also employed transformation to obtain valuable features. This process also defined the type of data in each feature and discovered the impact of each attribute on the case, (2) Clustering algorithm design was the process of applying algorithms to construct clusters by optimizing the measures, (3) Cluster validation was conducted to determine the accuracy of cluster algorithm. Cluster validation could help measure an adequate degree of confidence in the results, and (4) Interpretation referred to the explanation needed for the partitions constructed from the previous processes (Shamitha & Ilango, 2019).

The cluster algorithms compared in this study were partitional, which consisted of centroid and medoid, and hierarchical methods, which consisted of agglomerative and divisive methods. The centroid method represented each cluster using the instance's gravity centre. The medoid method represented each cluster by the mean of the instance close to the gravity centre. The agglomerative method formed the cluster in a bottom-up model until all points belong to the same clusters. On the contrary, the divisive method splits the data into smaller clusters in a top-down model until each cluster contains one data point.

RESULTS AND DISCUSSION

After removing incomplete data and conducting the scaling or transformation, the number of data used for the analysis was 1451, which was later split into training and testing data with a share of 70% and 30%, respectively. The training data consisted of 1029 observations and testing data had 422 observations. Each training and testing data was treated using three different clustering methods and three different numbers of clusters. The iteration was repeated 25 times for each algorithm in order to get stable clusters. The feature selection process was conducted by inputting all the variables provided in the dataset which were related to tuberculosis. The numbers of features obtained in the analysis were 9 variables.

The best number of clusters *k* was generally decided based on the values of the total within-clustering sum of squares from each clustering model. Figure 1 showed that the total within-clustering sum of square values decreases corresponding to the increasing number of clusters. The graph represented that a significant reduction occurred from *k=1* to *k=2*, which meant that the suggested *k* was 2. This was corroborated by the fact that there were no other significant reductions in the graph. Therefore, it was expected that model with two clusters might be better than other models. Comparing the accuracy of the models could also provide a better decision.

K-mean with k=2 produced two clusters that contained 972 observation in group 1 and 57 observations in group 2. Cluster or group 2 constructed from transformed data generally had higher means compared to cluster or group 1 for all indicators. For instance, the mean of the indicator of *total TB cases* in cluster 1 was -0.0032, while in group 2 it was 0.0551. This difference was observed in all indicators. The mean values informed the centre of each indicator belonged to each cluster. The within cluster sum of square for k=2 model was 24.7%, which informed the variability within each cluster. Meanwhile, k-means with k=3 yielded three clusters which consisted of 947, 6 and 76 observations respectively for clusters 1, 2 and 3. The mean values for most indicators or attributes escalated gradually from group 1, group 2, to group 3. The within-cluster sum of squares for the k=3 model was 33.1%. Furthermore, the k=4 k-mean model produced 4 groups which had 38, 803, 2 and 186 observations respectively for groups 1, 2, 3 and 4. The variability within each cluster was 43.7%. Figure 2 clarifies the information about the variability of each model, in which it showed that model one (k=2) was more compact than other models. This meant that the *k=2* model could divide the group better compared to other *k* since the distances of the observations in one group were closer to each other. It was supported by the value of the within-cluster sum of squares, where the *k=2* model had the smallest value (24.7%). In general, a cluster that had a small sum of squares was more compact than a cluster that has a large sum of squares. Figure 2 could also explain that the model with *k=2* could divide the data vividly without notable intersections.

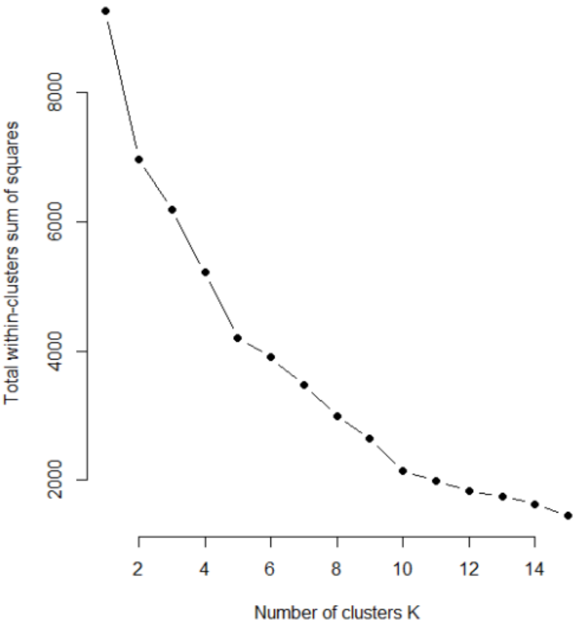

Figure 1. The values of the total within-clustering sum of squares for different number of clusters *k.*

Table 1. K-mean summary.

k	Group	Size	Sum Square by Cluster
2	1	972	24.7%
	2	57	
3	1	947	33.1%
	2	6	
	3	76	
4	1	38	41.6%
	2	803	
	3	2	
	4	186	

K-medoid with two groups (*k=2*) divided 803 observations into group 1 and 226 observations into group 2. Average dissimilarity in the k-medoid informed the average distance of the medoid point to all points in the cluster. The average dissimilarity of group 2 in this model was 3.2238 and 0.8649 in group 1. The medoid points (centres) for all attributes in group 1 was generally greater than the medoid points in group 2. Figure 3 showed that diameter of group 2 was greater than the diameter of group 1. This was in accordance with the value of average dissimilarities, which explained the average distance of each point in group 1 being closer than the average distance of points in group 2. This condition elucidated that those points or observations in group 1 had higher similarity than points in group 2.

K-medoid with 3 groups (*k=3*) split 430 observations to group 1, 393 observations to group 2, and 206 observations to group 3. The average dissimilarity value of group 1 was the smallest among other groups, which was 0.6171, while the average dissimilarity of groups 2 and 3 were 1.0091 and 3.2962, respectively. These values indicated that group 1 was the most compact cluster among all, wherein observation belonging to it had high similarities, while group 3 had less similar characters compared to group 1 and group 2. This was also explained in Figure 3 where group 3 had the highest diameter (38.7842), while groups 1 and 2 had diameters of 8.9835 and 26.6354, respectively. Meanwhile, k-medoid with *k=4* divided the data into 4 groups with sizes of 430, 393, 157 and 49 consecutively for groups 1, 2, 3, and 4. Group 1 tended to have smaller dissimilarity and diameter than other groups. This information corresponded to clusters 1 and 2 in the *k=2* model, in which the preceding cluster had a smaller distance or average dissimilarity and diameter compared to the latter cluster. Groups 1 and 2 in the *k=4* model might come from group 1 in the *k=2 model*, while groups 3 and 4 in the *k=4* model came from group 2 in the *k=2* model.

Table 2. K-medoid summary.

k	Group	Size	Average Dissimilarity	Diameter
2	1	803	0.8649	26.6354
	2	226	3.2238	38.7842
3	1	430	0.6171	8.9349
	2	393	1.0091	26.6354
	3	206	3.2962	38.7842
4	1	430	0.6171	8.9834
	2	393	1.0091	26.6354
	3	157	1.9659	33.5999
	4	49	5.9359	26.7368

Table 3 provided information regarding the cluster size for each group in the hierarchical clustering. Regardless of the number of cluster trees, most of the data gathered in group or cluster 1. This explained why most of the data had close distances and high similarities based on the hierarchical algorithm. The analysis also revealed that the mean and median of the data were in the cluster 1, hence the clustering in said cluster.

Nurhayati (2018) elaborated the accuracy between k-means and k-medoids algorithms applied in the big data, and it resulted that in general, the accuracy of k-medoid was higher than k-mean, whether the data was filtered or not. K-medoids also had better scalability for larger datasets due to it being more efficient than k-means (Soni & Patel, 2017). Table 4 denoted a result that confirmed the previous studies, in which generally, k-medoid had a higher accuracy compared to k-means for any *k*. This result showed that k-medoid was more robust to noise and outliers since, as represented on Figure 2 and Figure 3, there were points spreading far from other observations. Furthermore, de Souza et al. (2017) also stated that k-means and k-medoids had better accuracy compared to hierarchical clustering (de Souza et al., 2017). Results represented in Table 4 denoted the same information, in which on average, the accuracy of hierarchical clustering was the least among other methods, both for training and testing datasets. Moreover, there was no significant difference in the accuracy values between the training and testing data set for all methods based on two-mean different testing. It indicated that the accuracy values were consistent in both the training and testing dataset.

Table 3. Hierarchical clustering summary.

Cluster Size	Group	Size
2	1	1027
	2	2
3	1	1023
	2	2
	3	4
4	1	1023
	2	1
	3	1
	4	4

Table 4. Comparison of the algorithms' accuracy.

Method/Algorithm	Number of Clusters	Accuracy Training Data	Accuracy Testing Data
K-Mean	2	0.5578	0.5592
	3	0.4743	0.4810
	4	0.1788	0.2038
K-Medoid	2	0.7162	0.7133
	3	0.5306	0.5142
	4	0.2731	0.2607
Hierarchical	2	0.5044	0.3886
	3	0.4082	0.3910
	4	0.2566	0.2441

Table 4 represented the accuracy of each method for different cluster size. The accuracy value was obtained from a confusion matrix by summing the number of correctly classified values and dividing them by the total numbers of values. All methods provided a similar result of accuracy, in which a cluster size of 2 had a higher accuracy value than sizes 3 or 4. This was supported by the condition where cluster size 2 could separate the data more vividly, especially in the k-means method shown in Figure 2. This figure showed that 2 clusters-model separated the data into a group collected closely around the centre and a group that drifted far. However, k-means with k=2 still included some diffused data as part of group one, whereas the two-clusters-model in Figure 3 (*k-medoid*) could explain the outliers even clearer than the k-means method, as all outliers and dispersed data were classified into group 2 while gathered data were classified into group 1. Furthermore, the four-cluster model was the breakdown of the two-cluster model, for both k-mean and k-medoid.

Figure 2 showed the clustering graph. They informed how the data spread in two dimensions graph. The dimensions (dim 1 and dim 2) represented new variables produced from the dimensionality

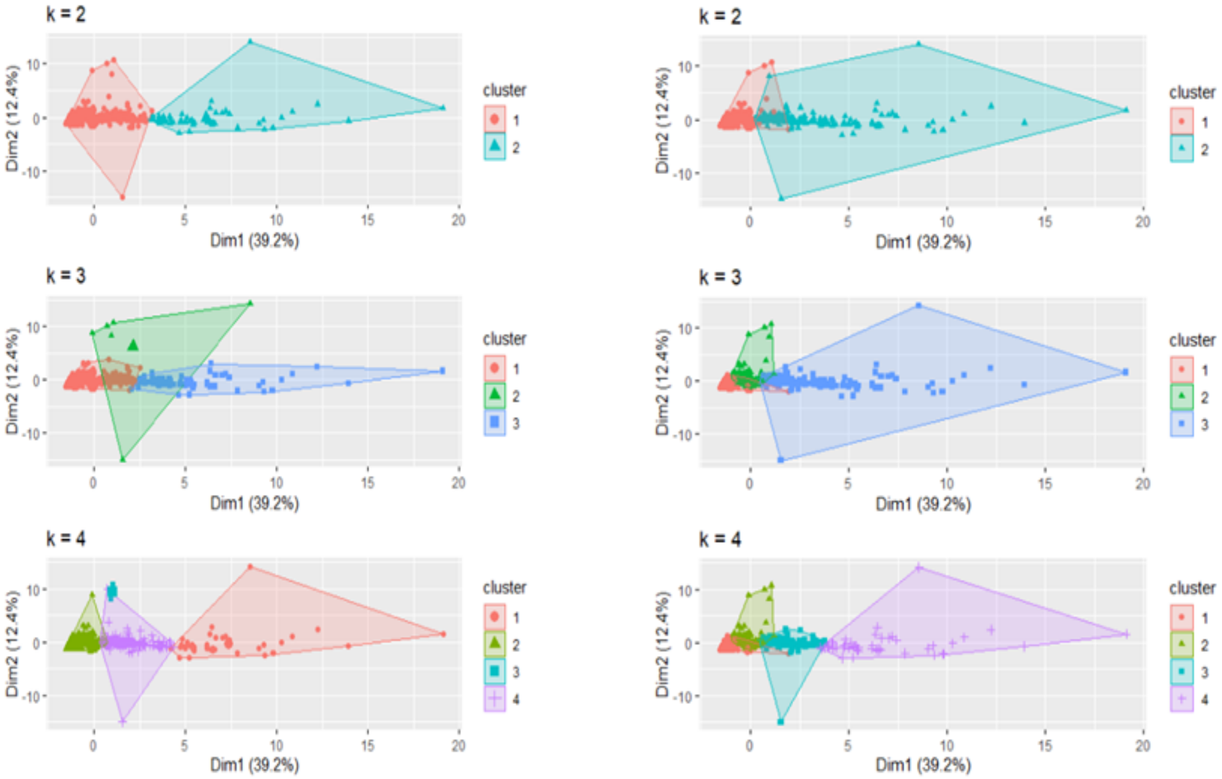

Figure 2. k-mean plot (left) and k-medioid plot (right).

reduction algorithm through principal component analysis (PCA). Each dimension described a certain amount of variation contained in the original dataset.

CONCLUSION

This study corroborated on previous studies that claimed k-medoids had a higher accuracy than other clustering algorithms or methods. This study explained that compared to k-means and hierarchical clustering, k-medoid had the highest accuracy for both training and testing data. K-medoid was better than the other two algorithms as it was more robust to noise and outliers that were found in the datasets. This outcome was applicable for the training and testing datasets. In terms of the number of clusters, the two-cluster-model was preferable for the dataset used in this study, as this model could classify the groups more vividly. The results were consistent in k-means, k-medoids, and hierarchical clustering methods.

Furthermore, in the *k=2* model of k-mean, it was also reported that the *k=2* model had the smallest sum of squares value. In general, a cluster that had a small sum of squares was more compact than a cluster that had a large sum of squares. This meant that the k=2 model could create more compact groups compared to other k models.

The k-medoid method could explain the characteristics of each cluster clearly. Group 1 in all algorithms and models had the smallest diameters and was observed to display the average dissimilarities. This meant that observations in group 1 had the lowest distance to its centre and had similar characteristics to each other. On the other hand, the hierarchical clustering put most of the data into group 1. This condition might be caused by the fact that observations in group 1 had the highest compactness and smallest diameters or average dissimilarities, while other groups might include outliers or dispersed observations.

REFERENCES

de Souza, V. A., Rossi, R., Batista, G., & Rezende, S. (2017). Unsupervised active learning techniques for labeling training sets: An experimental evaluation on sequential data. *Intelligent Data Analysis 21(5)*, 1061-1095. 10.3233/IDA-163075

Arora, P., Deepali, D., & Varshney, S. (2016). Analysis of K-Means and K-Medoids Algorithm For Big Data. *Procedia Computer Science 78*, 507 – 512. 10.1016/j.procs.2016.02.095

Blashfield, R. (1984). The Classification of Psychopathology: Neo-Kraepelinian and Quantitative Approaches. *Springer*.

Cadena, A., Fortune, S., & Flynn, J. (2017). Heterogeneity in Tuberculosis. *Nat Rev Immunol; 17(11)*, 691–702. 10.1038/nri.2017.69.

Castaldi, P., Dy, J., Ross, J., Chang, Y., Wshko, G., Curran-Everett, D., . . . Cho, M. (2014). Cluster analysis in the COPDGene study identifies subtypes of smokers with distinct patterns of airway disease and emphysema. *Thorax: 69*, 415-422. 10.1136/thoraxjnl-2013-203601

Cheung, Y.-M. (2003). K*-Means: A new generalized k-means clustering algorithm. *Pattern Recognition Letters 24*, 2883 - 2893. 10.1016/S0167-8655(03)00146-6

Clatworthy, J., Buick, D., Hankins, M., Weinman, J., & Horne, R. (2005). The Use and reporting of cluster analysis in health psychology: A review. *British Journal of Health Psychology*. 10.1348/135910705X25697

Eisen, M., Spellman, P., Brown, P., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA, Vol: 9*, 14863–14868. 10.1073/pnas.95.25.14863

Erawati, M., & Andriany, M. (2020). The Prevalence and Demographic Risk Factors for Latent Tuberculosis Infection (LTBI) Among Healthcare Workers in Semarang, Indonesia. *J Multidiscip Healthc*, 197-206. 10.2147/JMDH.S241972

Fialine, A., Alodia, D., Endriani, D., & Widodo, E. (2021). Implementasi Metode K-Medoids Clustering untuk Pengelompokan Provinsi di Indonesia Berdasarkan Indikator Pendidikan. *Journal of Mathematics Education and Applied, Vol:2, No:2*. https://doi.org/10.36655/sepren.v2i2.606

Gupta, A., Gupta, A., & Mishra, A. (2012). Research Paper on Cluster Techniques of Data Variations. *International Journal of Advance Technology & Engineering Research (IJATER)*, 39-47.

Gupta, M., & Jain, R. (2014). A Performance Evaluation of SMCA Using Similarity Association & Proximity Coefficient Relation For Hierarchical Clustering. *International Journal of Engineering Trends and Technology (IJETT), Vol:15*, 354-359.

Hair, J. F., Black, W., Babin, B., & Anderson, R. (2010). *Multivariate Data Analysis A Global Perspective 7th Edition.* Pearson: Prentice Hall.

Jain, A., Murty, M., & Flynn, P. (1999). Data Clustering: Review. *ACM Computing Surveys, Vol. 31, No. 3*, 264-322.

Kanungo, T., Mount, D., Netanyahu, N., Piatko, C., Silverman, R., & Wu, A. (2002). An Efficient k-Means Clustering Algorithm: Analysis and Implementation. *IEEE Transactions On Pattern Analysis and Machine Intelligence, No: 24, Vol: 7*, 881 - 892. 10.1109/TPAMI.2002.1017616

Kaufman, L., & Rousseeuw, P. (2005). *Finding Groups in Data: An Introduction to Cluster Analysis s (Wiley series in probability and statistics).* New York: Wiley-Interscience.

Kaur, K., & Attwal, K. (2014). A Review Paper on Clustering in Data Mining. *Research Cell: An International Journal of Engineering Sciences, Vol: 3*, 144-151.

Koo, H., Min, J., Kim, H., Ko, Y., Oh, J., Jeong, Y., Park, J. (2022). Cluster Analysis Categorizes Five Phenotypes of Pulmonary Tuberculosis. *Scientific Report; 12:10084*. 10.1038/s41598-022-13526-1

Lakoh, S., Jiba, D., Adekanmbi, O., Poveda, E., Shar, F., Deen, G., Yendewa, G. (2020). Diagnosis and treatment outcomes of adult tuberculosis in an urban setting with high HIV prevalence in Sierra Leone: A retrospective study. *International Journal of Infectious Diseases, Vol:96*. 10.1016/j.ijid.2020.04.038

Landau, S., & Ster, I. C. (2010). Cluster Analysis: Overview. *Elsevier*, 72-83. 10.1016/B978-0-08-044894-7.01315-4

Liao, M., Li, Y., Kianifard, F., Obi, E., & Arcona, S. (2016). Cluster Analysis and Its Application to Healthcare Claims Data: A Study of End-Stage Renal Disease Patients Who Initiated Hemodialysis. *BMC Nephrology*. doi.org/10.1186/s12882-016-0238-2

Mahendradhata, Y., Lambert, M., Deun, A., Matthys, F., Boelaert, M., & Stuyft, P. (2003). Strong general health care systems: a prerequisite to reach global tuberculosis control targets. *Int J Health Plann Manage*. 10.1002/hpm.724

Mao, J., & Jain, A. (1995). A self-organizing network for hyperellipsoidal clustering (HEC). *IEEE Trans. Neural Netw. 6*, 296–317. 10.1109/72.478389

Mohajan, H. (2015). Tuberculosis is a Fatal Disease among Some Developing Countries of the World. *American Journal of Infectious Diseases and Microbiology*, 18-31. 10.12691/ajidm-3-1-4

Nielsen, F. (2016). *Introduction to HPC with MPI for Data Science.* Switzerland: Springer Cham.

Noviyani, A., Nopsopon, T., & Pongpirul, K. (2021). Variation of tuberculosis prevalence across diagnostic approaches and geographical areas of Indonesia. *PLOS ONE 16(10)*. https://doi.org/10.1371/journal.pone.0258809

Novoselsky, A., & Kagan, E. (2021). An Introduction to Cluster Analysis. *Weizmann Institute of Science*. 10.13140/RG.2.2.25993.57448/1.

Nurhayati, Sinatrya, N., Wardhani, L., & Busman. (2018). Analysis of K-Means and K-Medoids's Performance Using Big Data Technology. *Journal Proceeding of The 6th International Conference on Cyber and IT Service Management (CITSM 2018)*. 10.1109/CITSM.2018.8674251

Punithavalli, M., Nathiya, G., & Punitha, S. (2010). An Analytical Study on Behavior of Clusters Using K-Means, EM and K* Means Algorithm. *(IJCSIS) International Journal of Computer Science and Information Security Vol. 7, No. 3*, 185-190. https://doi.org/10.48550/arXiv.1004.1743

Salman, R., Kecman, V., Li, Q., Strack, R., & Test, E. (1998). Fast K-Means Algorithm Clustering. *Proceedings of the Fifteenth International Conference on*, 91 - 99. 10.5121/ijcnc.2011.3402

Setyaningsih, S. (2012). Using Cluster Analysis Study to Examine the Successful Performance Entrepreneur in Indonesia. *Elsiever: Procedia Economics and Finance 4*, 286 – 298. https://doi.org/10.1016/S2212-5671(12)00343-7

Shamitha, S., & Ilango, V. (2019). A Roadmap For Intelligent Data Analysis Using Clustering Algorithms And Implementation On Health Insurance Data. *International Journal of Scientific and Technology Research, Vol: 8*, 2008-2018.

Sharara, H., & Getoor, L. (2010). Group Detection. In C. Sammut, & G. Webb, *Encyclopedia of Machine Learning* (pp. 489-492). New York: Springer.

Soni, K., & Patel, A. (2017). Comparative Analysis of K-means and K-medoids Algorithm on IRIS Data. *International Journal of Computational Intelligence Research, Vol: 13, No: 5*, 899 - 906.

Strehl, A., Ghosh, J., & Mooney, R. (2000). Impact of Similarity Measures on Web-page Clustering. *AAAI-2000: Workshop of Artificial Intelligence for Web Search*.

Wierzchoń, S., & Kłopotek, M. (2018). *Modern Algorithm of Cluster Analysis.* Warsaw, Polland: Springer.

World Health Organization.(2020). *Global Tuberculosis Report 2020.* Geneva.

World Health Organization. (2021, November 28). *World Health Organization*. Retrieved March 29, 2023, from https://www.who.int/indonesia/news/detail/28-11-2021-indonesia-commitment-to-eliminate-tb-by-2030-supported-by-the-highest-level-government#:~:text=Based%20on%20WHO%20Global%20TB,South%2DEast%20Asia%20Region%E2%80%9D.

Xin Jin, & Jiawei Han. (2010). K-Medoids Clustering. In C. Sammut, & G. Webb, *Encyclopedia of Machine Learning* (pp. 564–565). Boston, MA: Springer.

Zhao, Y., & Zhou, X. (2021). K-means Clustering Algorithm and Its Improvement Reserach. *Journal of Physics: Conference Series*.