Available online at http://scik.org

Commun. Math. Biol. Neurosci. 2022, 2022:41

https://doi.org/10.28919/cmbn/7335

ISSN: 2052-2541

COMPARISON OF CLUSTERING IN TUBERCULOSIS USING FUZZY C-MEANS AND K-MEANS METHODS

EKA MALA SARI ROCHMAN[1,2], MISWANTO[1,*], HERRY SUPRAJITNO[1]

[1]Department of Mathematics, Faculty of Sciences and Technology, Airlangga University, Surabaya, Indonesia.

[2]Departemen of Informatics, Faculty of Engineering, University of Trunojoyo Madura, Bangkalan, Indonesia

Abstract: Tuberculosis (TB) is a health problem that has yet to be resolved in Indonesia. Based on WHO data, in 2021 Indonesia will still be in the third rank of the highest TB cases in the world. This study aims to determine how many groups of TB patients are based on age, gender, HIV status, history of diabetes mellitus, chest X-ray, and the results of the Molecular Rapid Test (TCM). The data used in this study were 985 from 2017 to 2020. The method used in this research is K-Nearest Neighbor (KNN) in carrying out the imputation process, as well as comparing the k-means and Fuzzy C-Means (FCM) methods in classifying TB data. Before doing the grouping, the data cleaning process is carried out by an imputation process which is useful for filling in the missing data in this case, using the KNN method. To produce maximum results of data grouping or clustering, it is necessary to determine the right number of clusters. For this reason, this study tries to compare the elbow, silhouette coefficient, and Davies Bouldin Index (DBI) methods. The application of the KNN method in the data imputation process in this study is to use k=5. The application of the K-Means algorithm is to form groups of TB patients based on six features. Determination of

*Corresponding author

E-mail address: miswanto@fst.unair.ac.id

Received March 07, 2022

the optimal number of clusters using the K-means and FCM methods shows the optimal number of clusters, namely K = 2 but with different values. The results of the clustering test using the elbow method with the K-means and FCM methods are 93288.49. The DBI value for the K-means and FCM methods is 0.4937. Meanwhile, the clustering trial with the silhouette coefficient on K-means yields a value of 0.6318 which is better than the FCM which produces a value of 0.6321. This shows that the results of clustering k-means with silhouette coefficients produce better cluster quality because they have a lower silhouette coefficient value than FCM.

Keywords: tuberculosis; imputation; cluster; k-means; FCM; elbow; silhouette coefficient; DBI.

2010 AMS Subject Classification: 62H30.

1. INTRODUCTION

One of the health problems that is very serious and can cause death is Tuberculosis (TB), this includes in Indonesia. However, TB is a potentially infectious disease that can be treated and cured [1]. TB is an infectious disease caused by a bacterial infection. TB generally attacks the lungs, but can also attack other organs of the body, such as the kidneys, spine, and brain [2] and [3].

According to WHO, Indonesia is the 3rd country with the highest TB cases in the world after India and China. In 2018, 10 million people contracted this disease, and 1.5 million lost their lives to this disease. As many as 251,000 of them are people with HIV/AIDS.[4]

In large data, defective data are often found, such as missing values or missing data. In this TB data, there are missing values in some of its features. The quality and quantity of data greatly determine the quality of the results of a study, because, from the character of the sample, the character of a population will be generalized as the result of a study. A missing value is a condition where there are empty values or incomplete values in the data [5]. Several imputation methods have been developed to minimize the negative impact of missing values, including the mean method which replaces the missing value with the average value of a variable.

Nearest neighbor (NN) is one of the supervised learning algorithms, where this algorithm has the aim of finding new patterns in the data. This is done by connecting existing data patterns with new

data patterns. KNN imputed the missing value based on information from the closest observation that had a similar level of characteristics (features) with the observation where the missing value was located. In dealing with missing values using the KNN method, it begins by determining the number of closest neighbors and then calculating the smallest distance from each observation that does not contain a missing value.

In data mining, mathematical processes, statistics, artificial intelligence and identification are used in finding information, as well as clustering. This is applied to the grouping of TB patients so as to produce useful information on mapping or data distribution. Grouping data that have a maximum or even minimum similarity between one another into the same cluster is the goal of clustering [6]. Several methods that can be used for grouping are K-means and Fuzzy C-Means (FCM) methods. K-Means is a method that is quite often used in grouping this because this algorithm is simple and quite easy to implement. K-Means can also group large amounts of data with efficient and relatively fast computation time [6]. The Fuzzy C-Means algorithm is often used for datasets with varied attributes, for the K-Means Cluster algorithm it is more used for datasets with a small number of attributes [7] and [8].

However, K-Means also has a drawback, namely that there are no definite provisions when determining the best initial center of the cluster, if the determination of the initial center of a different cluster will result in different memberships. Then in 2019 the same research was carried out using a combination of the AHC (Agglomerative Hierarchical Clustering) and K-Means methods on the nutritional status of toddlers with the aim of classifying the nutritional status of toddlers with better accuracy results than previous studies. AHC is a hierarchical clustering method. the AHC method is capable or good at identifying small groups. In this study, the accuracy produced by the combination of AHC and K-Means methods reached 90% and proved to be better than the K-Means method itself [9] and [10].

Another study that uses K-means is to classify TB cases in children using K-Means grouping and identify distribution patterns using GIS (Geographic Information Systems) [4]. This study groups

TB cases in children by region. Based on the analysis, the spatial pattern of the distribution of TB in children can be categorized into areas with high prevalence, areas with moderate prevalence, and areas with low prevalence. The results can be used to assist decision-making in controlling TB cases in children.

Several methods can be used to determine the right number of clusters, including the elbow method, Partition Entropy (PE), GAP Statistics, cross-validation, silhouette coefficients [10],[11] and [12]. Each of these methods has its advantages and disadvantages, so accuracy is needed in integrating the clustering method used, the method for determining the right number of clusters, and the data structure and data size.

The elbow method is used to determine the best number of clusters that can be used to produce the best cluster results and maximize the quality of cluster results. Silhouette Coefficient is used to see the quality and strength of the cluster, how well or poorly an object is placed in a cluster [13]. In addition to the Elbow method and the SC method, the method used to test the cluster results is the Davies Bouldin Index (DBI) method. This method is one of the methods used to measure cluster evaluation based on the value of separation and cohesion. Cohesion is the amount of data proximity to the cluster center of the cluster being followed. Separation is the distance between the cluster centers of the cluster.

The contribution of this study is to map TB cases based on features of age, gender, history of diabetes, HIV status, chest X-ray, and Molecular Rapid Test (TCM) results using the K-Means method compared with FCM. To overcome the occurrence of missing values in the data, we propose using the KNN method. Meanwhile, to get the best number of clusters, each K-Means and FCM method will be combined with the Elbow, Silhouette Coefficient (SC), and Davies Bouldin Index (DBI) methods.

2. PRELIMINARIES

With data mining, important trends or patterns from the data will be obtained easily. Excavating

important information in data for some time requires data mining [14]. One of the data mining techniques used in this research is the clustering technique.

A. Imputation with K-Nearest Neighbor (KNN)

KNN (K-Nearest Neighbor) is an imputation method based on data that has the closest distance to the object data. The purpose of applying this method is to determine the value of the new object based on the attributes and training sample [5]. KNN is the simplest formula that is often used in implementing distance search. This method was chosen because the K-NN method is a form of decision support model that classifies data based on the closest distance [15]. The closest distance is calculated using the Euclidean formula with the following equation:

$$d\ (x_a x_b) = \sqrt{\sum_{j=1}^{m} \left(x_{aj} - x_{bj}\right)^2} \quad (1)$$

Where d(a,b) is the Euclidean distance, while x_{aj} is the sample data and x_{bj} is the test data. For parameter j is the j^{th} attribute and m is the number of attributes

B. K-Means

Clustering is a method of grouping data into several groups, where one group has the same characteristics as each other and has different characteristics from other groups [16] and [17]. Clustering cannot be equated with classification, because clustering itself does not have a target variable. This clustering algorithm is looking for a way to override the amount of data that is in a similar cluster uniformity into other clusters [13]. This technique is one method in solving problems regarding grouping. K-Means is a clustering method, which can classify large amounts of data quickly and efficiently. The K-means algorithm is an effective algorithm for finding clusters in the data stack. K-means is a method of analyzing data by carrying out a modeling process without a learning process and grouping with a partition system. This method seeks to minimize variations between data in a cluster and maximize variation with data in other clusters.

Euclidian Distance is a distance calculation mode that will be used to calculate the distance between two points on the Euclidian distance (2 Dimensions, 3 Dimensions, or more).

Some of the steps carried out in this method are as follows:

1. Determine the number of clusters that you want to form from a data set.

2. Determine the center location or centroid of the cluster randomly.

$$d(x,y) = \left|\left|x - y\right|\right|_2 = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2 \; ; i = 1,2,3, \ldots\ldots, n} \quad (2)$$

where x_i is the x-i object, while y_i is the i^{th} y centroid and n is the number of objects

3. For each data row, find the cluster closest to the cluster center (group).

4. The iteration is done by determining the new centroid using equation (2).

Work again until there are sufficient cluster members and do not move using the 3rd stage to the 4th stage

C. Fuzzy C-Means (FCM)

Fuzzy C-Means is an algorithm that works by assigning membership to each data point that corresponds to each cluster center based on the distance between the cluster center and data points. The more data that is close to the center of the cluster, the more membership to a particular cluster center [18]. Fuzzy C-Means uses a fuzzy grouping model with a fuzzy index using Euclidean Distance so that the data can be members of all classes or clusters formed with different membership degrees between 0 to 1 [4].

The basic thing in applying Fuzzy C-Means, is to determine the center of the cluster first, so that it can mark the average location for each cluster. This is because the center of the cluster cannot be said to be accurate. Each cluster that is formed will have a degree of membership at each data point. That is by repairing the center of the cluster and the degree of membership repeatedly, so that later the center of the cluster will shift to the right location. This iteration is based on the minimization of the objective function that describes the distance from a given data point to the center of the cluster which is weighted by the degree of membership of the data point. Here are the Steps in FCM

1. input data in the x matrix, where the matrix is m x n, where m is the number of data to be clustered and n is the attribute for each data. Example xij = i^{th} data (i=1,2,…m), j^{th} attribute

(j=1,2,…n).

2. The next step is to determine: : a. Number of clusters = c;

b. Weight = w;

c. Maximum iteration = MaxIter;

d. Error value = ξ

e. Initial Objective Function = P0 = 0;

f. iteration: t = 1;

3. Then generate random numbers ik (with i=1,2, …m and k=1,2, …c) as elements of the initial partition matrix U, where Xi is the i[th] data

$$U = \begin{bmatrix} \mu_{11}(x_1) & \mu_{21}(x_1) \cdots & \mu_{c1}(x_1) \\ \vdots & \vdots & \vdots \\ \mu_{1i}(x_i) & \mu_{2i}(x_i) \cdots & \mu_{ci}(x_i) \end{bmatrix} \quad (3)$$

4. Calculate the distance to the center of the k[th] cluster: V_{kj} , with k=1,2,…,c and j = 1,2,…,n

$$V = \frac{\sum_{i=1}^{m} (\mu_{ik})^w * X_{ij}}{\sum_{i=1}^{m} (\mu_{ik})^w} \quad (4)$$

μ^w is the membership value raised to the power of the weight value (w) that has been selected

5. Finding the value of the objective function in the t[th] iteration, Pt:

$$P_t = \sum_{i=1}^{m} \sum_{k=1}^{c} \left(\left[\sum_{j=1}^{n} \left(X_{ij} - V_{ij} \right)^2 \right] (\mu_{ik})^w \right) \quad (5)$$

6. Then fix the value of the partition matrix by calculating the change in the degree of membership of each data cluster

7. Check the stop condition:

a. If: (|Pt – Pt-1| < ξ) or (t>MaksIter) then stop.

b. If Else: t = t+1, repeat step 4

Information: The last step is to check the difference of the objective function by $|P0-P1|$ = Absolute (Objective function of initial iteration - An objective function of iteration 1). If the difference value is below the smallest error value, the iteration stops, and to determine which cluster the data is in, use the new membership value and select the maximum value, if not, continue with the next iteration and use the new membership value.

D. Elbow Method

The Elbow method is a method used to generate information in determining the best number of clusters by looking at the percentage comparison of the number of clusters that will form an angle at a point. The purpose of the elbow method is to choose a K value that is small and still has a low SSE value [14]. This is done by selecting the cluster values and then adding them up to be used as a model to determine the best cluster. The comparison between the number of clusters added is a percentage of the resulting calculation [11]. Information about the difference in cluster values can use a graph so that it can display the results of a different percentage of each cluster value. The best cluster value is obtained if the value of the first cluster with the value of the next cluster gives an angle on the graph or the value has the largest decrease. SSE (Sum of Square Error) is done to get a comparison of each cluster value. Because the larger the number of K clusters, the smaller the SSE value will be. Figure 1 shows the optimal cluster search using the elbow method.

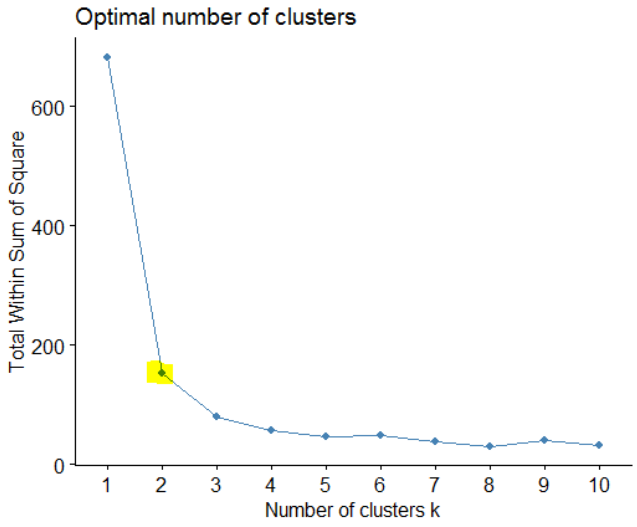

Figure 1. Determination of the Number of Clusters with the Elbow Method

The best cluster value in the elbow method is obtained from the Sum of Square Error (SSE) value which has a significant decrease and is elbow-shaped. To calculate SSE using the formula

$$SSE = \sum_{i=1}^{n}(d)^2 \quad (6)$$

Where, d is the distance between the data and the center of the cluster. Sum of Square Error (SSE)

is a formula used to measure the difference between the data obtained and the prediction model that has been done previously. Several studies often use SSE's in determining the optimal cluster.

E. Silhouette Coefficient

Silhouette Coefficient is a method where you want to know how well the capacity level tested in a cluster is. The method is a combination of cohesion and separation methods. Cohesion is a method for calculating how close the relationship between several objects in an identical cluster is. And the separation method is the opposite of cohesion, which is to estimate how far a cluster is, but the cluster is separated from other clusters. Below are the steps for calculating the silhouette coefficient [10]:

1. All objects in the same cluster are calculated using the average distance of the i^{th} object.

$$a(i) = \frac{1}{|A|-1}\sum_{j} \in_{A}, j \neq i\ d(i,j) \quad (7)$$

   where a(i) is the difference in the average object (i) of all other objects in A. While the value of d(i,j) is a measurement of the distance between data i and data j, and the value of A itself is a cluster

2. Next, all other objects in different clusters are calculated by the average distance of the i^{th} object.

$$d(i,C) = \frac{1}{|A|}\sum j\ \in d(i,j) \quad (8)$$

   the notion of d(i,C) is the average difference in the object (i) to all other objects that exist in C. The value of C itself is a cluster other than cluster A in other words cluster C is not the same as cluster A

3. Calculate d(i,C) for all C take the smallest value with the formula in Equation 9

$$b(i) = \min_{C \neq A} d(i,C). \quad (9)$$

   Cluster B that reaches its minimum i.e., d(i,B)) is called a neighbor of the object (i).

4. Finally, the calculation of the silhouette coefficient in equation 10

$$s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))} \quad (10)$$

According to Kaufman and Rousseuw [19], these criteria can be allocated as follows.

Table 1 Silhouette coefficient values.

No	Score SC	Criteria
1.	$0,7 < SC \leq 1$	Close Structure
2.	$0,5 < SC \leq 0,7$	Medium Structure
3.	$0,25 < SC \leq 0,5$	Tensile Structure
4.	$SC \leq 0,25$	Unstructured

F. Davies Bouldin Index (DBI)

Davies Bouldin Index (DBI) is a method for checking clustering results in addition to Elbow and Silhouette Coefficient. The DBI testing approach is in the form of separation and cohesion values. Cohesion is the sum of the similarity of existing data to the center of the cluster, while separation is the distance between the cluster centers of the cluster. Clusters that have high separation values and low cohesion values are optimal [6]. The Sum of a square within-cluster (SSW) is the equation to find out the optimal cluster can be seen in Equation 11.

$$DBI = \frac{1}{K} \sum_{i=1}^{k} max_{i \neq j}(R_{ij}) \quad (11)$$

Where K is the number of clusters. The Davies Bouldin Index (DBI) value which is getting closer to 0 indicates the better the cluster obtained.

3. MAIN RESULTS

A. Data Collection

The object under study is the object from the TB (Tuberculosis) dataset taken from Syarifah Ambami Rato Ebuh Hospital which is one of the hospitals located in Bangkalan Regency, East Java Province, Indonesia. The data used are 985 data from 2017-2020. In this study, 6 parameters will be used, namely age, gender, chest X-ray, HIV status, history of diabetes, TCM results.

B. System Design

System design is a description of a sequence of processes that exist in the system to make it easier

to understand the concept of algorithms. This flowchart describes a main system algorithm in the clustering process on TB data. An overview of the flowchart can be seen in Figure 2.

Based on Figure 2, the process of grouping TB data objects can be ordered as follows:

1. Input TB data along with the six features used.
2. Carry out the imputation process using the KNN method
3. Grouping using K-means and FCM methods.
4. Then the results of cluster membership in each method were tested using the Elbow, Silhouette Coefficient and DBI methods.

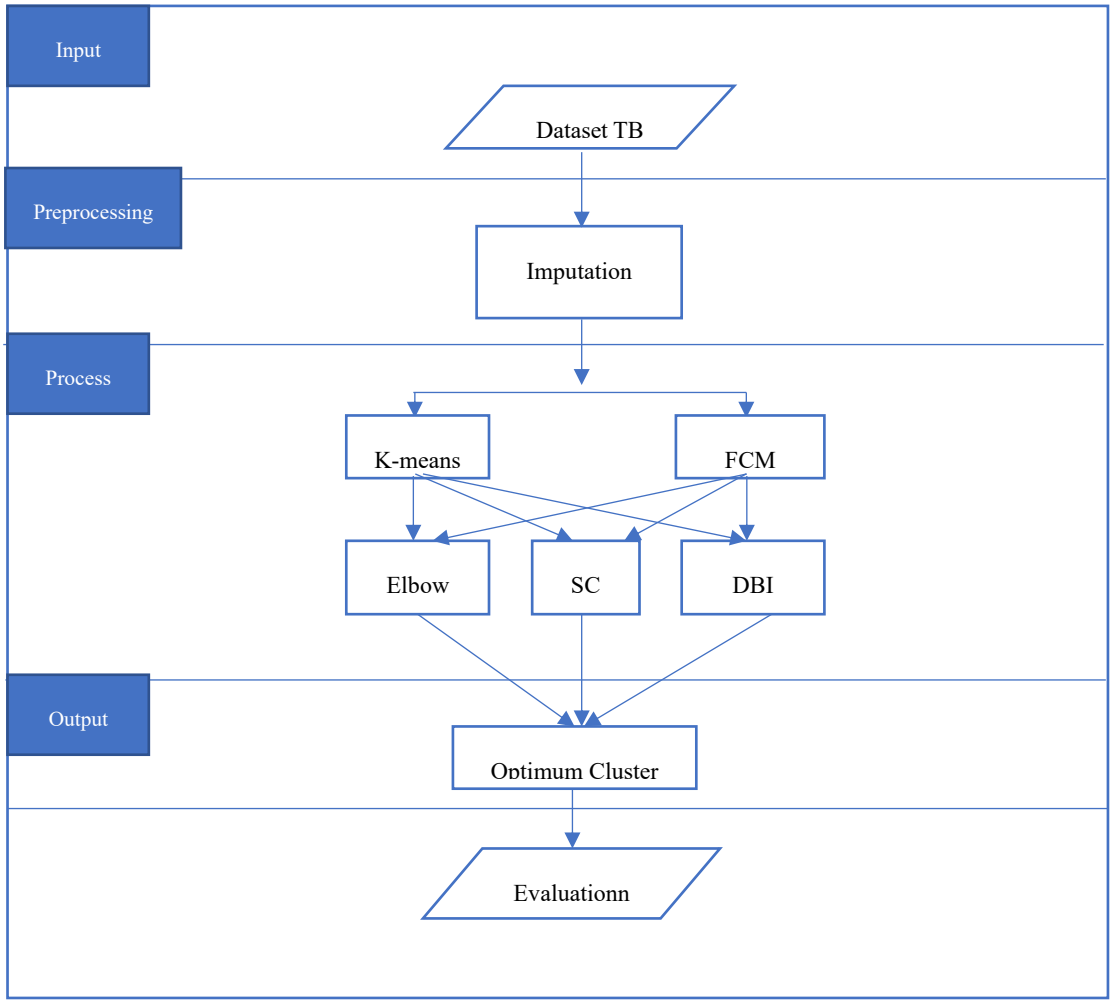

Figure 2. Flow of TB data grouping system

C. Imputation

In this KNN method, 985 records were entered which had 6 attributes, namely, age, gender, chest X-ray, HIV status, history of diabetes and TCM results. Then all data in the form of categorical converted into numeric data. If all the data is in the form of numeric data, then handling the missing value can be done using the experiment k=1-5. In this study, filling in the missing value in the tuberculosis data was carried out using a value of k=5.

D. K-Means and FCM with Elbow

Determination of the optimal number of clusters in this study using one of the cluster analysis methods, namely the Elbow method, taking into account the comparison value (from the SSE calculation for each cluster value) between the number of clusters that will form an angle at a point, so the greater the number of clusters k, the SSE value. will get smaller. The experimental range of K values is from 1 to 9. The comparison of the best cluster results can be seen in Figures 2 and 3

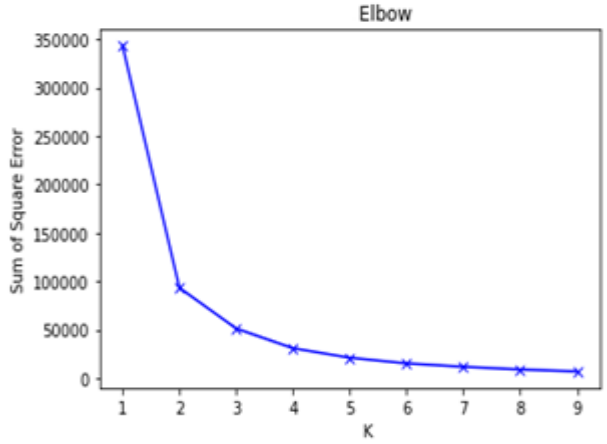

Figure 3. K-Means with Elbow

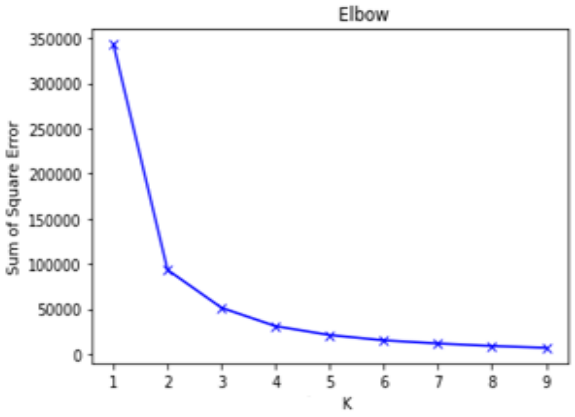

Figure 4. FCM with Elbow

Figures 3 and 4 illustrate that the X-axis represents the K value and the Y-axis represents the SSE value. SSE is the total distance from a centroid to data that are in the same cluster. The SSE used is SSE with a small/no longer significant decrease because the desired data is data that is closely spaced [3]. The selected K value is the K value which is very decreasing. In the graph above, both the SSE K-means and FCM decrease significantly when going to the value of K = 2, after that the

inertia decreases.

E. K-means and FCM with Silhouette Coefficient (SC)

Silhouette Score is a metric used to measure the level of goodness of the clustering technique. Silhouette values range from -1 to 1 with the following information [4]:

1: Clusters are separated from each other

0: The distance between groups is not significant

-1: Incorrect cluster used

When the Silhouette value is closer to 1, the better the grouping of objects into a cluster. On the other hand, if the silhouette value is close to -1, the data grouping method in the cluster will be worse [3].

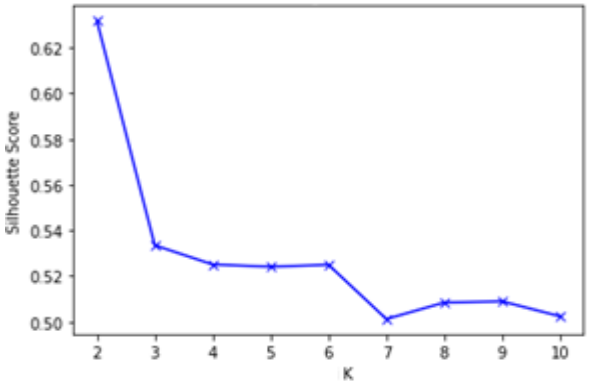

Figure 5. K-Means with SC

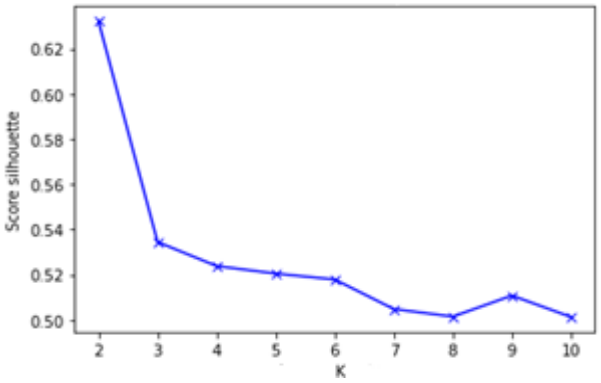

Figure 6. FCM with SC

Based on Figures 5 and 6, the X-axis shows the number of clusters and the Y-axis shows the SC score, it was found that the silhouette score on the K-Means and FCM methods was the best at K=2.

F. K-means and FCM with DBI

The Davies Bouldin Index (DBI) value which is closer to 0 indicates the better the cluster obtained. The lower the DBI value indicates the optimal cluster result, it can be seen in Figures 7 and 8 the difference using the k-means and FCM methods

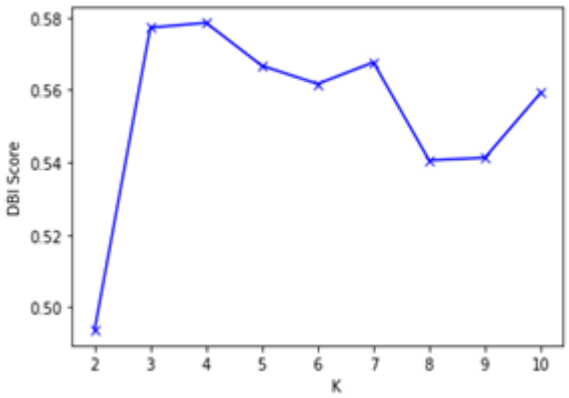

Figure 7. K-Means with DBI

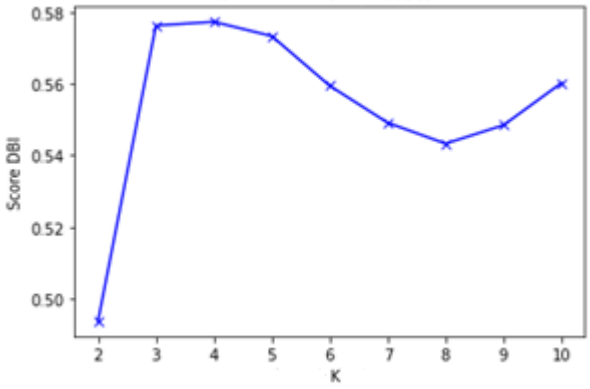

Figure 8. FCM with DBI

Based on Figures 7 and 8, the X-axis shows the number of clusters and the Y-axis shows the DBI score, it is found that the DBI score in the K-Means and FCM methods is the lowest at K=2.

G. Analysis

Based on the trials on K-means and FCM, the results of the comparison of cluster values using the Elbow, Silhouette coefficient and DBI are obtained as set out in Table 2.

Table 2. Test results of K-Means and FCM methods

K	SSE		SC		DBI	
	K-Means	FCM	K-Means	FCM	K-Means	FCM
2	93288,49545	93288.49544	0,631886	0.6321365	0,493737	0.493731
3	51226,4505	51226.45049	0,533944	0.5344387	0,576321	0.576233
4	30883,72951	30883.72950	0,525064	0.5237853	0,582335	0.577261
5	21065,21326	21065.21325	0,524076	0.5204611	0,569461	0.573284
6	15229,58753	15229.58752	0,525401	0.5178423	0,559634	0.559415
7	11737,925	11737.92500	0,505463	0.5049575	0,575289	0.549105
8	8969,160091	8969.160090	0,505731	0.5028249	0,545531	0.543274
9	6947,131016	6947.131016	0,509782	0.5107412	0,546169	0.5484369

Table 2 shows that in testing TB data for clustering using K-Means and FCM, the optimal cluster at K=2. The SC value of the K-Means algorithm is higher at 0.631886 and the FCM algorithm is 0.6321365. Meanwhile, in testing for Elbow K-Means clustering and FCM algorithms, the SSE results are relatively the same, namely 93288,49545. And the last test is the DBI of the two methods to get the same relative value of 0.493737. The discrepancy between the results of the Silhouette Coefficient test is influenced by the measurement of the Euclidean distance which can give unequal weight to the underlying factors.

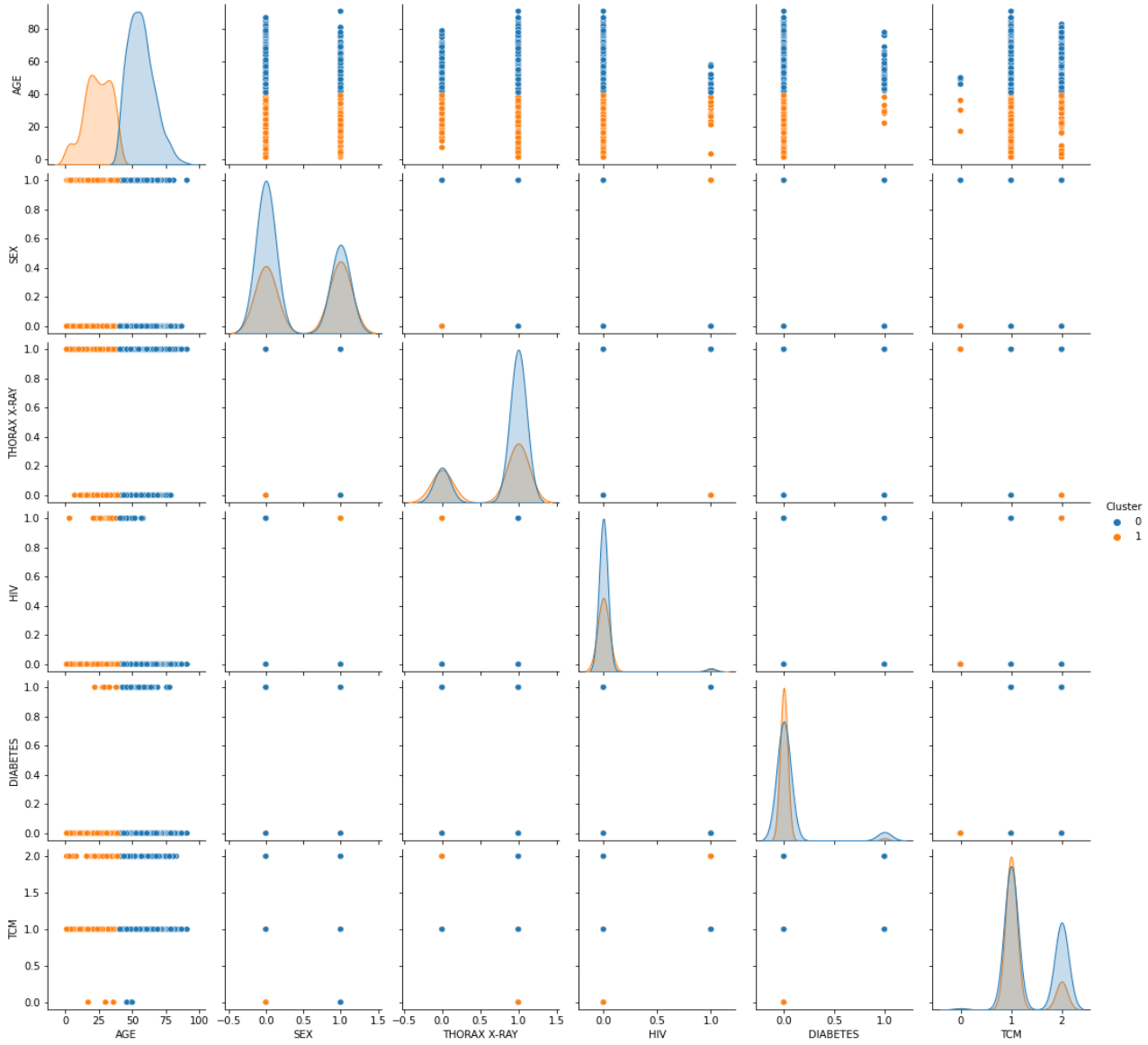

Figure 9. Results of Cluster K=2 on K-Means

Figure 8 shows the cluster on TB is 2, with the following explanation:

Cluster 0

- Age under 40 years
- Male gender: HIV positive, TCM rif resistant
- Female gender: negative chest x-ray
- Negative chest X-ray: HIV positive, TCM negative
- Positive chest X-ray: TCM rif resistant
- HIV negative: TCM rif resistant
- Negative diabetes: TCM rif resistant

Cluster 1

- Age above 40 years
- Male gender: chest X-ray positive, HIV positive, HIV negative
- Female gender: chest x-ray negative, chest x-ray positive, HIV negative, TCM rif resistant, TCM negative
- Negative chest X-ray: negative HIV, positive diabetes, negative diabetes
- Positive chest X-ray: HIV negative, HIV positive, TCM negative
- HIV negative: TCM negative
- Both positive and negative diabetes enter cluster 1, unless diabetes is negative with TCM rif resistance to enter cluster 0
- All sensitive TCM rifs go to cluster 1

So it can be concluded that the number of members in each cluster can be seen in Figure 9 below, where Cluster 0 has 596 data and Cluster 1 has 389 data.

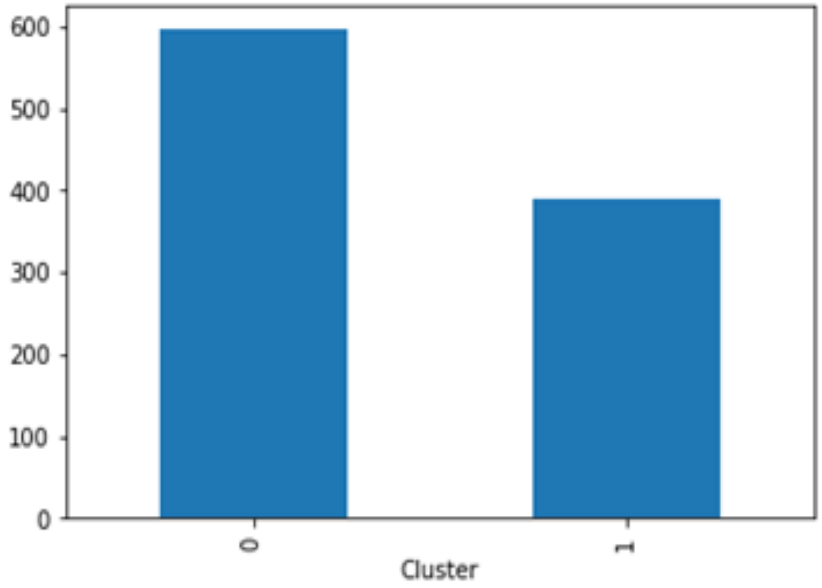

Figure 9. Number of members in the distribution of Cluster k=2 K-means method

CONCLUSION

Based on the research and testing carried out on the k-means and FCM methods, the conclusions obtained include:

1. To overcome missing values in TB data with features of age, gender, HIV status, DM history, chest X-ray, and TCM results, machine learning algorithms, namely KNN with *k*=5.
2. For TB data, grouping using K-means and FCM methods produces the optimum cluster at K=2.
3. The K-means and FCM methods with Elbow produce the same SSE value of 93288,49545
4. The K-means method with SC produces a value of 0.631886, this is better than FCM with SC which is 0.6321365.
5. The K-means and FCM methods with DBI resulted in the same DBI value of 0.493737.

ACKNOWLEDGMENT

The researcher would like to thank the promoters and co-promoters who have helped and directed this research. Thanks also to Universitas Airlangga for allowing researchers to be able to develop

their work and knowledge.

CONFLICT OF INTERESTS

The author(s) declare that there is no conflict of interests.

REFERENCES

[1] A. Rachmad, N. Chamidah, R. Rulaningtyas, Mycobacterium tuberculosis images classification based on combining of convolutional neural network and support vector machine, Commun. Math. Biol. Neurosci. 2020 (2020), Article ID 85. https://doi.org/10.28919/cmbn/5035.

[2] A. Rachmad1, N. Chamidah, R. Rulaningtyas, Mycobacterium tuberculosis identification based on colour feature extraction using expert system, Ann. Biol. 36 (2020), 196-202.

[3] A. Tostmann, S.V. Kik, N.A. Kalisvaart, M.M. Sebek, S. Verver, M.J. Boeree, D. van Soolingen, Tuberculosis transmission by patients with smear-negative pulmonary tuberculosis in a large cohort in the Netherlands, Clin. Infect. Dis. 47 (2008), 1135-1142. https://doi.org/10.1086/591974.

[4] R.S. Wardani, Purwanto, Sayono, A. Paramananda, Clustering tuberculosis in children using K-Means based on geographic information system, AIP Conf. Proc. 2114 (2019), 060012. https://doi.org/10.1063/1.5112483.

[5] T. Mahboob, A. Ijaz, A. Shahzad, M. Kalsoom, Handling missing values in chronic kidney disease datasets using KNN, K-means and K-medoids algorithms, in: 2018 12th International Conference on Open Source Systems and Technologies (ICOSST), IEEE, Lahore, Pakistan, 2018: pp. 76–81. https://doi.org/10.1109/ICOSST.2018.8632179.

[6] B.N. Sari, Identification of tuberculosis patient characteristics using K-means clustering, Sci. J. Inform. 3 (2016), 31-40.

[7] B. Al Kindhi, T.A. Sardjono, M.H. Purnomo, G.J. Verkerke, Hybrid K-means, fuzzy C-means, and hierarchical clustering for DNA hepatitis C virus trend mutation analysis, Expert Syst. Appl. 121 (2019), 373–381. https://doi.org/10.1016/j.eswa.2018.12.019.

[8] S.S. Sundari, N. Ariani, Penerapan Data Mining Untuk Pengelompokan Penyakit Dengan Algoritma Fuzzy C-

Means (Studi Kasus: UPT Puskesmas Salawu), Jurnal VOI (Voice of Informatics), 8 (2019), 63-76.

[9] Intan Alpiana, Lilik Anifah, Penerapan Metode KnA (Kombinasi K-Means dan Agglomerative Hierarchical Clustering) dengan Pendekatan Single Linkage untuk Menentukan Status Gizi pada Balita, Indones. J. Eng. Technol. 1 (2019), 61-68.

[10] E.M.S. Rochman, A. Khozaimi, I.O. Suzanti, et al. A combination of algorithm agglomerative hierarchical cluster (AHC) and K-means for clustering tourism in Madura-Indonesia, J. Math. Comput. Sci. 12 (2022), Article ID 62. https://doi.org/10.28919/jmcs/7086.

[11] M.A. Syakur, B.K. Khotimah, E.M.S. Rochman, B.D. Satoto, Integration K-means clustering method and elbow method for identification of the best customer profile cluster, IOP Conf. Ser.: Mater. Sci. Eng. 336 (2018), 012017. https://doi.org/10.1088/1757-899X/336/1/012017.

[12] D.R. Agustian, B.A. Darmawan, Analisis Clustering Demam Berdarah Dengue Dengan Algoritma K-Medoids (Studi Kasus Kabupaten Karawang). Jiko (Jurnal Informatika dan Komputer), 6 (2022), 18-26.

[13] M.F. Fahmi, Y.K. Suprapto, Wirawan, Segmentation and distribution of watershed using K-modes clustering algorithm and Davies-Bouldin index based on geographic information system (GIS), in: 2016 International Seminar on Application for Technology of Information and Communication (ISemantic), IEEE, Semarang, Indonesia, 2016: pp. 235–240. https://doi.org/10.1109/ISEMANTIC.2016.7873844.

[14] N.T. Hartanti, Metode Elbow dan K-Means Guna Mengukur Kesiapan Siswa SMK Dalam Ujian Nasional, Jurnal Nasional Teknologi dan Sistem Informasi, 6 (2020), 82–89. https://doi.org/10.25077/TEKNOSI.v6i2.2020.82-89.

[15] Susanti, S. Martha, E. Sulistianingsih, K Nearest neighbor dalam Imputasi Missing value, Bimaster: Buletin Ilmiah Matematika, Statistika dan Terapannya, 07 (2018), 9–14.

[16] E.M.S. Rochman, A. Rachmad, Clustering tourist destinations based on number of visitors using the K-mean method, in: Proceedings of the 1st International Multidisciplinary Conference on Education, Technology, and Engineering (IMCETE 2019), Atlantis Press, Banten, Indonesia, 2020. https://doi.org/10.2991/assehr.k.200303.075.

[17] M. Fauzi, Yudi, Penerapan Algoritma K-Means Clustering Untuk Mendeteksi Penyebaran Penyakit TBC (Studi

Kasus: Di Kabupaten Deli Serdang), JTIK (Jurnal Teknik Informatika Kaputama), 1 (2017), 1-7.

[18] S. Kusumadewi, H. Purnomo, Aplikasi Logika Fuzzy untuk pendukung keputusan, Graha Ilmu: Yogyakarta (2004).

[19] L. Kaufman, P. J. Rousseuw, Finding groups in data: An introduction to cluster analysis, vol. 344. John Wiley & Sons, Inc., Hoboken, New Jersey, 1990.