The 2nd International Conference on Multidisciplinary Engineering (2nd ICOMDEN 2024)

The 2nd International Conference on Multidisciplinary Engineering: 00040 (2024) https://doi.org/10.29103/icomden.v2.xxxx
Volume 2, 2024
eISSN: xxxx-xxxx
Research Original Article

Clustering the Spread of Tuberculosis Disease in Aceh Tamiang Regency Using the K-Medoids Method

Rania Sofia Salsabila Hrp✉[1], **Eva Darnila**[2], **Yesy Afrillia**[3]
[1,2,3] Informatics Engineering Study Program, Department of Informatics, Malikussaleh University Jl. Unimal Bukit Indah Campus, Lhokseumawe City, 24351, Indonesia
✉Corresponding Author: rania.200170252@mhs.unimal.ac.id, eva.darnila@unimal.ac.id, yesy.afrillia@unimal.ac.id

Abstract
This research aims to develop a web-based system using the K-Medoids method to classify Tuberculosis (TB) disease spread in Aceh Tamiang Regency. TB is a severe health problem in Indonesia that requires effective handling, especially in identifying areas with high levels of spread. The data used in this study includes the number of TB patients from 2019 to 2023 in 12 sub-districts, with five types of TB, namely Pulmonary TB, Extra Pulmonary TB, Latent TB, Billiary TB, and Drug Resistant TB. This study used the K-Medoids algorithm with a value of K=3. This clustering shows that the clustering for Pulmonary TB disease type obtained C1 by 16.67% or two sub-districts, C2 by 8.33% or one sub-district, and C3 by 75% or nine sub-districts. In the Extra Pulmonary TB disease type, C1 is 25%, C2 is 8.33% and C3 is 66.67%. C1 8.33%, C2 58.33% and C3 33.33% were obtained for the type of Latent TB disease. For the type of Billiary TB disease, there are C1 8.33%, C2 58.33% and C3 33.33%. For the drug-resistant TB, C1 16.67%, C2 16.67%, and C3 66.67% are used. The average deviation for the 5 types of disease clusters is 2.01. The results of this study are expected to be a reference for local governments in efforts to prevent and manage the spread of TB disease in Aceh Tamiang District.
Keywords: Clustering, K-Medoids, Tuberculosis

Introduction
Indonesia ranks third globally in the number of Tuberculosis (TB) cases. According to the WHO, by 2020, there will be 9.9 million Indonesians with TB and 1.5 million deaths[1]. TB cases are spread in almost all regions, including Aceh Tamiang District. The Aceh Tamiang District Health Office recorded 610 TB cases from 2019 to 2020, not including reports from private health facilities. If all health facilities reported their cases, the number of TB cases in Aceh Tamiang could be higher. In recent years, health services in Aceh Province have focused on data analysis to improve medical care and quality of life. Data mining and forecasting are essential in modern social and medical fields. However, data processing at the Aceh Tamiang District Health Office is still done manually using Excel, so it has yet to be able to categorize the spread of TB disease effectively[2].

The advancement of technological knowledge and its applications in all fields cannot be separated from computer devices. The need for information services is significant, especially in the health sector[3]. One is an application that uses data mining to facilitate the analysis process in determining the pattern of Tuberculosis disease spread. This grouping aims to determine which areas have low, medium and highest levels of Tuberculosis disease. The results of this study are expected to be a recommendation for the Aceh Tamiang Regency government to provide drugs against Tuberculosis disease more effectively. The Government can later focus more on areas that have high levels of Tuberculosis disease cases so that Tuberculosis cases in Aceh Tamiang Regency will decrease[4].

The purpose of this research is to build a web-based Tuberculosis disease distribution grouping system that uses the K- Medoids method to group the distribution of Tuberculosis (TB) disease types in Aceh Tamiang Regency. It is hoped that this system can be useful for the Aceh Tamiang Regency government to see the distribution of community data affected by TB disease and help the government in prevention and awareness efforts in dealing with the symptoms of TB disease. Clustering is a technique used to group a set of objects or data that have similar characteristics into several groups or partitions based on specific criteria[5]. The clustering process is done by observing and analyzing the characteristics of each data object so that it can be grouped into different classes according to their similarities. One method that is often used in clustering is the calculation of the distance between data using Euclidean Distance. This method involves calculating the distance between data points in a multidimensional space, which is then used to determine the closeness or similarity between objects in a clustering algorithm[6].

Literature Review

1. Tuberculosis

Tuberculosis, or TB, is an infectious disease that affects the lungs and is caused by the bacterium Mycobacterium tuberculosis. The disease is characterized by the formation of necrotic granulation tissue (greening) in response to the bacterial infection. The lungs are a very important organ of the body as they play a major role in supporting the respiratory system.

2. Causes of Tuberculosis (TB) Disease

To date, the most common bacteria found is Mycobacterium tuberculosis which is associated with the following five bacteria: Mycobacterium tuberculosis, Mycobacterium bovis, Mycobacterium africanum, Mycobacterium microti, and Mycobacterium cannettii. These five types of bacteria are known as the Mycobacterium tuberculosis complex and have the ability to cause disease in both humans and animals.

3. Symptoms of Tuberculosis (TB)

Symptoms of tuberculosis may vary depending on the location of the infection in the body. Latent tuberculosis infection causes no symptoms, while active tuberculosis causes symptoms. Symptoms are usually related to the respiratory system which can affect other parts of the body, depending on where the TB bacteria grow. Common symptoms for people with TB disease include:

1. Bloody cough
2. Pain in the chest
3. Shortness of breath
4. Cough that lasts for more than 3 weeks
5. Drastic weight loss
6. Loss of appetite
7. Fever and chills
8. Night sweats

4. Types of Tuberculosis (TB) Disease

1. Pulmonary TB
2. Extra Pulmonary TB
3. Latent TB
4. Drug Resistant TB
5. Miliary tuberculosis (TB)

5. Data Mining

Data mining is often used to uncover information hidden in large databases. Data mining is also known as an activity that involves collecting and utilizing historical data to identify patterns, rules, and relationships in large datasets. Data mining is also called an activity that involves the collection and utilization of historical data to identify patterns, rules, and relationships in large datasets.

6. Geographic Information System

The term "spatial" means relating to space or place involving geographical coordinates. Spatial information can be known from the Geographic Information System (GIS) has been referred to by this context. In this case, information related to location or space on the earth, we can gain insight into the coordinates of an object on its surface as well as information related to the attributes and characteristics of the area. Geographic Information Systems, abbreviated as GIS, are computer-based systems that convey information using data with a spatial information component.

Materials & Methods

1. Clustering

Clustering is a technique used to group a set of objects or data that have similar characteristics into several groups or partitions based on specific criteria. [8]. The clustering process is done by observing and analyzing the characteristics of each data object so that it can be grouped into different classes according to their similarities. One method that is often used in clustering is the calculation of the distance between data using Euclidean Distance. This method involves calculating the distance between data points in a multidimensional space, which is then used to determine the closeness or similarity between objects in a clustering algorithm [9].

2. Method K-Medoids

K-Medoids is a clustering method that divides a dataset into groups based on the proximity between data, using the actual data object as the cluster centre, called the "medoid." Unlike K-Means, which uses the average of the data as the centroid, K- medoids are more resistant to outliers and noise, as they are not affected by extreme values. The process starts with randomly selecting several objects as medoids and then calculating the distance of each object in the dataset to the nearest medoid to determine the initial clustering [11].

After clustering, the K-Medoids algorithm iteratively updates the medoids by searching for new candidate medoids

that minimize the total distance within the cluster. If the new medoid reduces the total overall distance, it replaces the old medoid [12]. This process is repeated until the medoids no longer change or until a specific stopping criterion is reached, resulting in a more stable and accurate clustering. This method is particularly effective on datasets with outliers, as the chosen medoid provides a more robust representation than the centroid in K-Means.

3. Stages of the K-Medoids Method

The completion stage in the K-Medoids algorithm involves several main steps, starting from the initialization of cluster centres to the final determination of the medoids in each cluster. Here are the complete steps:

1. Initialize k cluster centres (number of clusters). Randomly select k objects from the dataset as initial medoids. These
   selected objects will be the initial centres for each cluster.
2. Count each object to the nearest cluster using the Euclidian Distance equation. Group each object to the cluster whose medoid has the closest distance to the object, which is the following formula:

$$d(x_{i,y_i}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

   Where d = document point
   x = criteria data
   y = point medoids

3. Randomly select a new medoid; for each cluster, randomly select an object as a new medoid candidate.
4. Calculate the distance of objects in each cluster to the new medoid candidate.
5. Calculate the total deviation (S) by calculating the new total distance - old total distance value. If S<0, swap objects with cluster data to form a new set of k objects as a medoid.
6. Iterate the previous step until the medoid does not change so that the cluster and its members are obtained. Once the medoid is stable, the algorithm is complete, and you get k clusters, each with a medoid, and cluster members are determined based on the closest distance to the medoid.

4. K-Medoids Scheme

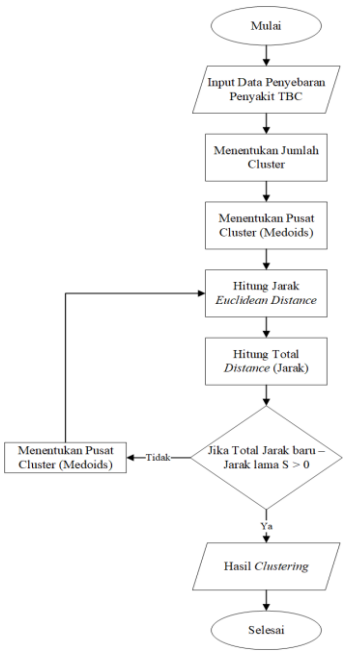

Figure 1. Scheme K-Medoids

Figure 1 is a process or system design scheme in this research, from start to finish:

1. The schematic process of the K-Medoids method begins.
2. The system starts from inputting data on the spread of Tuberculosis disease in 2019, 2020, 2021, 2022 and 2023 for each sub-district in Aceh Tamiang Regency where the data is obtained from the Aceh Tamiang GeneralHospital recap.
3. After inputting, the next step is to determine the value of the number k, which is to determine 3 initial cluster centers randomly.
4. Next calculate the distance between the object and the temporary medoid. After that, mark the object's smallest distance value to the medoid and calculate the sum using the Euclidean Distance formula.
5. After obtaining the distance results from the Euclidean Distance formula, the next step is calculating the total distance.
6. Next, calculate the total distance by calculating the difference between the new total cost and the old total cost.

7. If the total deviation is less than 0, then the calculation will continue until the total deviation is more than 0.
8. The next step is determining the new medoid as the cluster centre at the calculated iteration.
9. Recalculate the distance of each data to each medoid using the Euclidean Distance to obtain the results of measuring the data distance for each cluster until the total deviation is reached. S> 0.
10. After the calculations on all data are complete, the next step is to group the data into clusters. The cluster selected for each data is determined based on the smallest value of each cluster calculation, and this process ends when the final cluster result is obtained.
11. Then, the system displays the output of the clustering results of Tuberculosis disease, which is grouped according to each cluster.
12. The last stage is completion, which marks the end of the system.

Results and Discussion

1. Culcuation Analysis of K-Method

The dataset that has been collected will be analyzed using the K-Medoids method. The analysis is carried out using 4 criteria attributes, namely disease spread data in 2019, 2020, 2021, 2022 and 2023. The spread of TB disease is divided into 5 types of TB disease that will be grouped, namely Pulmonary TB, Extra Pulmonary TB, Latent TB, Billiary TB, and Drug.

Table 1. Cluster Labels

Cluster	Label
C1	High
C2	Medium
C3	Low

The following pulmonary TB data is organized per sub-district each year to form a dataset that is grouped in Table 2.

Table 2. Pulmonary TB Data

No	Kecamatan	2019	2020	2021	2022	2023
1	Banda Mulia	4	3	2	1	0
2	Bandar Pusaka	1	1	1	2	2
3	Bendahara	1	4	2	3	1
4	Karang Baru	3	4	6	5	2
5	Kejuruan Muda	7	7	2	5	1
6	Kota Kuala Simpang	3	6	2	6	1
7	Manyak Payed	1	1	1	2	0
8	Rantau	3	5	2	7	6
9	Sekerak	1	2	0	0	0
10	Seruway	3	8	4	1	1
11	Tamiang Hulu	4	2	3	1	0
12	Tenggulun	0	3	2	1	1

The steps taken in completing the calculation of the K-Meodids Method are as follows:

1. Determining the Number of Clusters; The number of clusters used in grouping the spread of TB disease is 3 clusters. The clusters include High (C1), Moderate (C2) and Low (C3) distribution based on data in 2019-2023 with a total of 12 sub- districts in Aceh Tamiang.
2. Data Normalization: The data normalization function in the clustering process aims to adjust the values of the features in the data so that they are on the same scale. This is important because if the data is not normalized, large-scale features can dominate small-scale features, affecting the clustering results. The normalization function used is min-max scaling for the following results:

   $x_1y_1 = (3 - 0)/(7 - 0) = 0{,}571$

   $x_2y_1 = (3 - 1)/(8 - 1) = 0{,}285$

   $x_3y_1 = (2 - 0)/(6 - 0) = 0{,}333$

   Continue until all rows and columns of data are last. So, the final normalization result is as follows.

Table 3. Lung TB Normalization Data

No	Kecamatan	2019	2020	2021	2022	2023
1	Banda Mulia	0,5714	0,2857	0,3333	0,1429	0
2	Bandar Pusaka	0,1429	0	0,1667	0,2857	0,3333
3	Treasurer	0,1429	0,4286	0,3333	0,4286	0,1667
4	New Coral	0,4286	0,4286	1	0,7143	0,3333
5	Junior Vocational	1	0,8571	0,3333	0,7143	0,1667
6	Kuala Simpang City	0,4286	0,7143	0,3333	0,8571	0,1667
7	Manyak Payed	0,1429	0	0,1667	0,2857	0

8	Rantau	0,4286	0,5714	0,3333	1	1
9	Sekerak	0,1429	0,1429	0	0	0
10	Seruway	0,4286	1	0,6667	0,1429	0,1667
11	Tamiang Hulu	0,5714	0,1429	0,5000	0,1429	0
12	Tenggulun	0	0,2857	0,3333	0,1429	0,1667

3. Determining Initial Medoids: Determining the initial centre of the cluster (Medoids) is done randomly or randomly, for the initial Medoids can be seen in Table 3.

Table 4. Initial Lung TB Medoids

Label	Kecamatan	2019	2020	2021	2022	2023
C1	Banda Mulia	0,5714	0,2857	0,3333	0,1429	0
C2	Bandar Pusaka	0,1429	0	0,1667	0,2857	0,3333
C3	Treasurer	0,1429	0,4286	0,3333	0,4286	0,1667

4. Calculate the *Euclidean* Distance with the initial medoids that have been determined;

$d(x_1y_1) = \sqrt{(0{,}5714 - 0{,}5714)^2 + (0{,}2857 - 0{,}2857)^2 + (0{,}3333 - 0{,}3333)^2 + (0{,}1429 - 0{,}1429)^2 + (0 - 0)^2} = 0$

$d(x_2y_1) = \sqrt{(0{,}1429 - 0{,}5714)^2 + (0 - 0{,}2857)^2 + (0{,}167 - 0{,}333)^2 + (0{,}2857 - 0{,}1429)^2 + (0{,}3333 - 0)^2} = 0{,}6516$

$d(x_3y_1) = \sqrt{(0{,}1429 - 0{,}5714)^2 + (0{,}4286 - 0{,}285)^2 + (0{,}333 - 0{,}333)^2 + (0{,}4286 - 0{,}1429)^2 + (0{,}167 - 0)^2} = 0{,}5599$

Continue until the last row of data by calculating the same *Euclidean* distance as before. After getting all the *Euclidean* distances, choose the smallest value and the total distance from all the cluster data. The following is described in the form of table 5.

Table 5. Euclidean Distance Iteration-1 Lung TB

No	C1	C2	C3	Smallest Distance	Label
1	0	0,651615818	0,55990362	0	C1
2	0,651615818	0	0,509546061	0	C2
3	0,55990362	0,509546061	0	0	C3
4	0,960678143	1,069310075	0,797174716	0,797174716	C3
5	0,929791628	1,307140689	1	0,929791628	C1
6	0,861431072	0,986875323	0,589015089	0,589015089	C3
7	0,55990362	0,333333333	0,509546061	0,333333333	C2
8	1,355261854	1,179232619	1,059724433	1,059724433	C3
9	0,579310717	0,490845908	0,63576333	0,490845908	C2
10	0,818230489	1,174656798	0,775181933	0,775181933	C3
11	0,219512963	0,668365183	0,634424409	0,219512963	C1
12	0,595238095	0,421905837	0,349927106	0,349927106	C3
Total Distance				**5,54450711**	

5. Randomly Select New Medoids: New medoids are selected randomly with the condition that they cannot take medoids that have been used before, for new medoids can be seen in Table 6.

Table 6. New Medoids of Pulmonary Tuberculosis

Label	Kecamatan	2019	2020	2021	2022	2023
C1	New Coral	0,4286	0,4286	1	0,7143	0,3333
C2	Junior Vocational	1	0,8571	0,3333	0,7143	0,1667
C3	Kuala Simpang City	0,4286	0,7143	0,3333	0,8571	0,1667

6. Recalculating *Euclidean* distance: After getting new medoids, the data is recalculated with *Euclidean* using the new centre point (medoids). The calculation is the same as before in the first iteration. The following are the results of the 2nd iteration calculation in Table 7.

Table 7. Euclidean Distance Iteration-2 Lung TB

No	C1	C2	C3	Smallest Distance	Label
1	0,960678143	0,929791628	0,861431072	0,861431072	C3
2	1,069310075	1,307140689	0,986875323	0,986875323	C3
3	0,797174716	1	0,589015089	0,589015089	C3
4	0	0,991174205	0,757801451	0	C1
5	0,991174205	0	0,606091527	0	C2
6	0,757801451	0,606091527	0	0	C3
7	1,120060332	1,307140689	0,986875323	0,986875323	C3

8	0,995454522	1,088228084	0,857473481	0,857473481	C3
9	1,335881918	1,37622343	1,132142231	1,132142231	C3
10	0,88991579	0,885765488	0,838419851	0,838419851	C3
11	0,888640838	1,03728671	0,955352507	0,888640838	C1
12	1,001416231	1,285714286	0,936776932	0,936776932	C3
Total Distance				**8,07765014**	

7. After getting the 2nd iteration distance, then calculate the total deviation. When the total deviation S>0 is reached, the calculation will be stopped. If S<0, then iterate again as before. The total deviation is as follows:
Deviation (S) = Total New Distance - Total Old Distance S = 8,07765014 - 5,54450711 = 2,53314303
Since the total deviation result is more than 0, the object does not need to be swapped, and this test stops at the 2nd iteration. The last iteration result will be the clustering parameter, so the final cluster result is as follows:

Table 8. Final Result of PulmonaryTB K-Medoids Calculation

No	Kecamatan	2019	2020	2021	2022	2023	Cluster/Label
1	Banda Mulia	4	3	2	1	0	C3 (Low)
2	Bandar Pusaka	1	1	1	2	2	C3 (Low)
3	Treasurer	1	4	2	3	1	C3 (Low)
4	New Coral	3	4	6	5	2	C1 (High)
5	Junior Vocational	7	7	2	5	1	C2 (Medium)
6	Kuala Simpang City	3	6	2	6	1	C3 (Low)
7	Manyak Payed	1	1	1	2	0	C3 (Low)
8	Rantau	3	5	2	7	6	C3 (Low)
9	Sekerak	1	2	0	0	0	C3 (Low)
10	Seruway	3	8	4	1	1	C3 (Low)
11	Tamiang Hulu	4	2	3	1	0	C1 (High)
12	Tenggulun	0	3	2	1	1	C3 (Low)

8. Final Result of TB Type Clustering
Based on the calculation of K-Medoids at the previous point, it is implemented to the entire dataset, namely, 5 types of TB disease. The results of all clusters can be seen in the following table.

Table 9. Final Result of TB Type Clustering

Types Of Tuberculosis Disease	Kecamatan		
	C1 (High)	C2 (Medium)	C3 (Low)
Tuberculosis Of The Lungs	2	1	9
Extra-Pulmonary Tuberculosis	3	1	8
Latent Tb	1	7	4
Billionaire Tb	1	7	4
Drug-Resistant Tuberculosis	2	2	8

The results are visualized with a bar graph in the following figure:

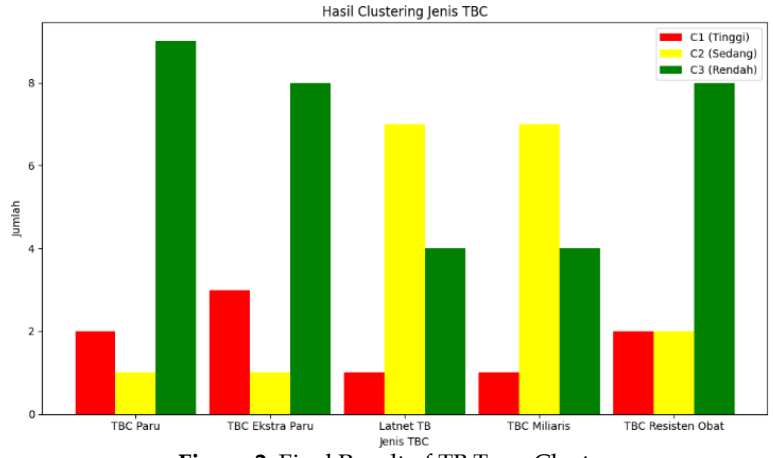

Figure 2. Final Result of TB Type Cluste

The conclusion from the clustering results of the spread of various types of TB disease in Aceh Tamiang District

shows that most sub-districts are in clusters with low spread. Although most areas have a controlled level of spread, there are several sub-districts with high TB spread, especially in Extra Pulmonary TB and Drug Resistant TB, which require more intensive intervention and more aggressive control strategies. Meanwhile, sub-districts with low transmission rates should be closely monitored to ensure the situation remains under control and prevent future increases in cases. A more focused and specific approach to the conditions of each cluster is essential for the effectiveness of TB prevention and management programs in this region.

9. System Implementation

The following are the results of the system implementation using the PHP and Java Script programming languages according to the previously designed system scheme.

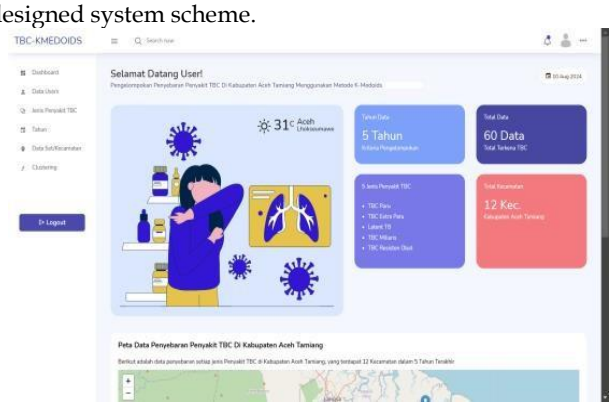

Figure 3. System Home Page

In Figure 3, the main page of the system displays dataset information, namely year, type of disease, and total sub-districts. For the sidebar, there is a CRUD (Create, Read, Update and Delete) menu for sub-district data, user data, year data, TB disease type data and the Clustering menu. Then, the *clustering* process can be seen in Figure 4.

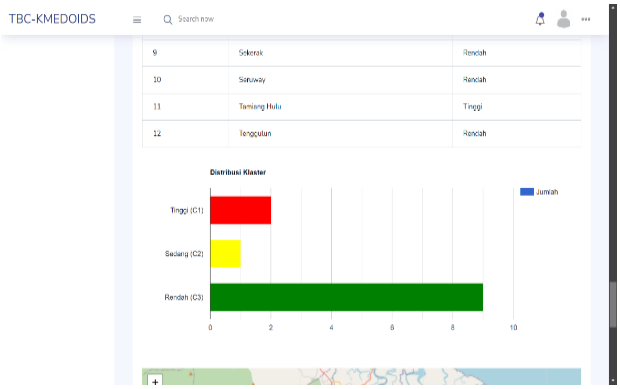

Figure 4. Data Clustering Page

Conclusions

Clustering the spread of TB disease in Aceh Tamiang using the K-Medoids method has been successfully carried out, which is grouped into 3 clusters, namely high, medium, and low, and obtained output in the form of a web-based system built with PHP and JavaScript programming languages. The results of the K-Medoids Implementation obtained 3 clusters in 12 sub-districts in Aceh Tamiang Regency from 5 types of TB disease. In the type of Lung TB disease, C1 of 16.67% or two sub-districts, C2 of 8.33% or one sub-district, and C3 of 75% or nine sub-districts were obtained. In the Extra Pulmonary TB disease type, C1 was obtained at 25%, C2 at 8.33%, and C3 at 66.67%. For the type of Latent TB disease, C1 is 8.33%, C2 is 58.33%, and C3 is 33.33%. For the type of biliary TB disease, C1 is 8.33%, C2 is 58.33%, and C3 is 33.33%. For the type of drug-resistant TB, C1 is as much as 16.67%, C2 is as much as 16.67%, and C3 is as much as 66.67%.

References

[1] A. Sutanti, M. Komaruddin, P. Damayanti, And P. U. Metro Information System Studies, "Designing A Keliling Library Application Using A Structured App," Journal Of Computer And Information Science (Komputa), Vol. 9, No. 1, 2020.

[2] M. Farid, I. Al-Rizki, I. Widaningrum, And G. A. Buntoro, "Prediction Of Tb Disease Spread With The K-Means Clustering Method Using The Rapidminer Application," Journal Of Engineering Technology), Vol. 5, No. 1, 2020, Doi: 10.31544/Jtera.V5.I1.2020.1-10.

[3] D. Marlina, N. Fauzer Putri, A. Fernando, And A. Ramadhan, "Implementation Of K-Medoids And K-Means Algorithms For Grouping Disability Distribution Areas In Children," Coreit Journal, Vol. 4, No. 2, 2018.

[4] T. Juninda And E. Andri, "Application Of K-Medoids Algorithm For Disease Clustering In Pekanbaru Riau," 2019.

[5] A. P. Fialine, D. Alya Alodia, D. Endriani, E. Widodo, And P. Statistika, "Implementation Of The K-Medoids

Clustering Method For Grouping Provinces In Indonesia Based On Education Indicators," 2021..

[6] Y. Puspita Sari Et Al., "Implementation Of K-Means Algorithm For Tuberculosis Distribution Clustering In Karawang Regency," Vol. 5, No. 2, 2020.

[7] M. Harahap, W. Fuadi, L. Rosnita, E. Darnila, And R. Meiyanti, "Featured Vegetable Clustering Using K-Means Algorithm," Journal Of Informatics Engineering And Information Systems, Vol. 8, No. 3, Dec. 2022, Doi: 10.28932/Jutisi.V8i3.5277.

[8] R. A. Permana, D. P. Hapsari, And R. R. Muhima, "Snestik National Seminar On Electrical Engineering, Information Systems, And Informatics Engineering Application Of K-Medoids Clustering Method For Web-Based Abstract Mapping," P. 277, 2022, Doi: 10.31284/P.Snestik.2022.2812.

[9] F. Sulistiyo Hidayat, R. Berliana, P. Affandi, V. Zuliana, And T. N. Padilah, "Application Of K-Means Clustering In Grouping Tuberculosis
Cases In West Java Province," Scientific Journal Of Wahana Pendidikan, Vol. 8, No. 15, Pp. 213-227, 2022, Doi: 10.5281/Zenodo.7049113.

[10] D. Sutris Martua Simanjuntak, I. Gunawan, I. Purnama Sari, T. Informatika, And S. Tunas Bangsa, "Application Of The K-Medoids Algorithm For Grouping Unemployment Ages 25 And Over In North Sumatra." [Online]. Available: Https://Ejournal.Catuspata.Com/Index.Php/Jkdn/Index