Jurnal Sisfotek Global
ISSN (Online): 2721 - 3161, ISSN (Print): 2088 – 1762
DOI: http://dx.doi.org/10.38101/sisfotek.v14i2.15658
Vol. 14, No. 2, September 2024, pp. 86-92

Implementation Machine Learning of K-Means Clustering Method and Linear Regression for Detecting the Risk of Tuberculosis Spread in Bangka Regency

Nurhaeka Tou[1*], Putri Mentari Endraswari[2], Syafiranur Iftizam[3], Itlahtul Mu'Anah[4]

[1,2,3,4] Universitas Bangka Belitung, Bangka, Indonesia, 33172
E-mail: [1]nurhaeka@ubb.ac.id, [2]putrimentari@ubb.ac.id, [3]syafiranur.sn@gmail.com, [4]madkholil1203@gmail.com

ARTICLE HISTORY

Received : July 4, 2024
Revised : July 23, 2024
Accepted : August 15, 2024

KEYWORDS

Clustering
K-Means
Linear Regression
Prediction
Tuberculosis

ABSTRACT

The number of tuberculosis (TB) sufferers in Bangka Regency tends to increase and is becoming a very serious public health problem. In 2023, there will be around 331,581 residents of Bangka Regency, with around 1,489 of them suffering from TB. The high number of TB cases in Bangka Regency requires comprehensive monitoring. Surveillance can be carried out by grouping TB cases by region. One clustering method is K-Means which groups data based on similar criteria. Researchers also used the Linear Regression method to predict the effect of population density on increasing TB prevalence. This research was conducted to cluster the distribution pattern of TB by region and identify the influence of population density variables on the increase in TB. Based on the results of the K-Means method analysis by looking at the Performance Davies Bouldin-Index (Dbi) test, researchers obtained a pattern of TB distribution which was categorized into: Areas with High Prevalence (Sungailiat), Medium Prevalence (Belinyu and West Mendo) and Low Prevalence (Bakam, Merawang, Pemali, Pudding Besar, and Riau Silip). Meanwhile, the prediction results show that there is a positive influence of 81% of the population density variable on increasing the number of TB spreads and 19% is influenced by other variables. The results of this research can be used as a reference for the Bangka District Health Service to assist decision-making in preventing an increase in TB cases.

1. Introduction

Tuberculosis (TB) is a type of disease caused by infection with the microscopic bacteria Mycobacterium Tuberculosis which attacks the lungs of humans [1]. This disease is one of the 10 causes of death in the world and is a public health challenge for sustainable health development [2][3]. Globally there are 10.6 million people who suffer from TB disease, there are 1.4 million deaths caused by TB disease including HIV-negative, and 187,000 deaths including HIV-Positive [4].

Indonesia is the second highest country contributing to TB cases after India with a total of 969 thousand cases in 2022 and death data reaching 93 thousand every year or the equivalent of 11 deaths per hour [5][6]. Based on WHO World Health data published in the 2022 Global TB Report, TB disease in the world mostly attacks those aged 45-54 years. Transmission of this disease is through the saliva or phlegm of sufferers containing pulmonary tuberculosis bacilli [4][7].

In Indonesia, TB disease spreads throughout the region, one of which is the Bangka Belitung Islands Province. According to the Head of the P2M Section of the Bangka Belitung Provincial Health Service, TB cases will experience an increasing trend of 26% from 2021 to 2022. In 2022 there will be 2,500 confirmed cases of TB patients and it is estimated that there will still be many TB cases that have not been confirmed. Currently, the Health Service continues to strive to detect as many TB cases as possible in the community with a target of 6,200 cases. One of the districts in Bangka Belitung Province that was most affected was Bangka Regency [8].

In 2023, Bangka Regency recorded 662 TB cases, or around 44.46 percent of the total 1489 cases. So, it is estimated that around 55.44% of TB cases in Bangka Regency have not been detected. With so many cases of TB that have not been detected, the possibility of spreading TB disease is increasing [9]. To quickly detect TB cases that have not been reported, the important thing that can be done is to recognize the pattern of TB disease spread based on sub-regions in Bangka Regency. So, it can be determined which regional sub-local cluster quality is close to high, medium, or low TB cases. In this way, the Health Service can focus its outreach goals and prevention efforts on areas that have high TB case clusters. To

determine clusters in an area, an accurate method is needed. One method that can be used is the K-Means Clustering method. This method is considered effective in determining clusters in piles of data [10].

Based on the problems above, research is needed with the title "Detection of the Spread of Tuberculosis in Bangka Regency Using the K-Means Clustering Machine Learning Method and Linear Regression". In this research, the input data is in the form of data sets which are processed using the K-Means Clustering method to be grouped based on objects/districts as seen from the distribution of TB disease. Determining the best cluster is based on the cluster that has the optimum Dbi (Davies Boulding Index) value. From this process, it will then be seen which sub-districts are prone to TB disease. Furthermore, this research will also carry out a prediction model using Linear Regression to see the cause-and-effect relationship from one variable to another variable. The results of this research can be used as a reference by the health service in detecting TB disease more quickly and providing treatment for TB patients. It is hoped that the Health Service and the relevant Government can focus more on identifying and treating TB in areas that have a higher risk of spreading TB. So, the number of TB cases in Bangka Regency can continue to decline.

This research refers to several previous studies that also used the K-Means Clustering method to carry out clustering, namely: the first study with the title "Detection of the Spread of Tuberculosis with the K-Means Clustering Algorithm Using Rapid Miner". This research clustered the spread of TB disease in the city of Bandung in 2020. The clustering results showed that the largest cluster was in cluster 0 with 10 members, the medium cluster with 8 members, and the small cluster with 1 member [5].

The next research is entitled "Clustering of Tuberculosis and Normal Lungs Based on Image Segmentation Results of Chan-Vese and Canny with K-Means". This research clustered normal lung conditions and lung conditions suffering from tuberculosis based on the segmentation results of chest X-ray images, producing a cluster accuracy of 73% [11].

The latest research is entitled "Implementation of the K-Means Algorithm in Clustering Stunting Cases among Toddlers in Randudongkal Village". This research used 200 datasets which were processed using the K-Means method with results of 190 stunted toddlers and 10 normal toddlers based on a DBI value of -0.673. This shows that the cluster results are good because the DBI value is close to 0[12].

2. Research Methodology

To ensure the achievement of the objectives of this research, the researcher created the research stages which are presented in Figure 1.

2.1 Identification of Problems

The first stage of this research was identifying and formulating problems at the Bangka District Health Service. The problem experienced is that the number of TB cases in Bangka Regency tends to increase. Apart from that, monitoring or prevention patterns are not yet on target in areas where the spread of TB is high.

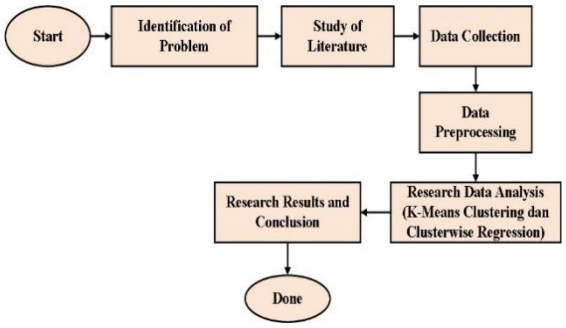

Figure 1. Research Diagram

2.2 Study of Literature

The literature study stage was carried out to find references that were used as references in this research. Researchers conduct studies through books, journals, and other publications that are appropriate to the research topic to obtain basic information and knowledge that can support the implementation of this research.

2.3 Data Collection

The data collection technique used in this research is an observation technique, where the research team visited the Bangka District Health Office directly to collect data on TB cases. Apart from that, the research team also conducted interviews with the TBC Technical Office of the Bangka District Health Service to find out about the development of TB disease, especially in Bangka Regency.

2.4 Data Preprocessing

Data preprocessing is an important stage in data analysis in machine learning. The data preprocessing stage is carried out to clean noise data, duplicate data, or missing data from the obtained data set. This stage is also carried out by researchers to minimize errors when modeling data. In this study, duplicate or missing data was treated in the form of data deletion. So, the data that goes into the model-building stage is truly valid and accurate.

2.5 Research Data Analysis

The next stage is to carry out an analysis of the data that has been collected using two methods, namely K-means clustering and Linear Regression.

2.5.1 K-Means Clustering

The K-Means Clustering method is a method that groups data based on similar characteristics. Data that

has the same characteristics will be grouped into the same cluster and data that has different characteristics will be grouped into another cluster[13]. Parameter K is the number of clusters that must be determined first before carrying out the clustering process. In general, the K-Means method has four main stages as follows:

1. Determine the number of K values or clusters to be created. Next, the randomly determined K value will be used as the initial center of mass.
2. Calculate the distance values to all centroid centers using the Euclidean Distance equation. The Euclidean equation is presented in equation (1).

$$D_e = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2} \quad (1)$$

Information:
$x_i - s_i$ = First coordinate point
$y_i - t_i$ = Second coordinate point

3. Move each object in the cluster based on the closest centroid distance value.
4. If a change occurs, the next process is to calculate the centroid value. The calculation of the new centroid value uses the average value of the objects in each cluster. The next process will continue to repeat from step 2, if there are no further changes then the k-means algorithm process will stop.

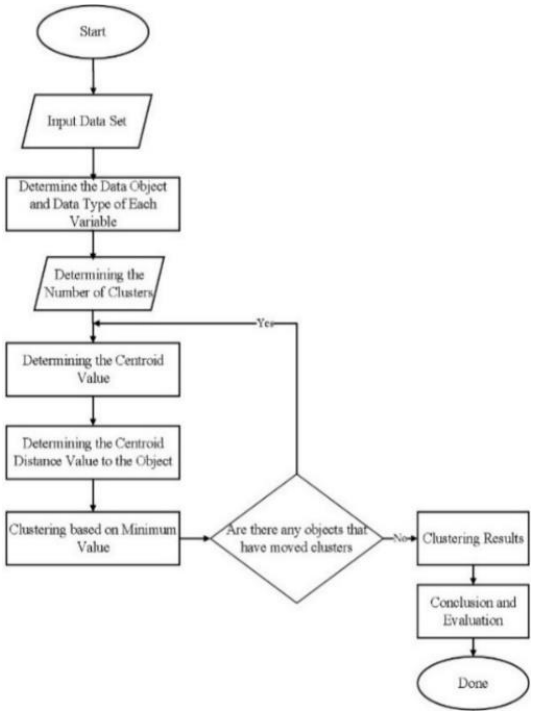

Figure 2. Flowchart K-Means Clustering

The K-Means method is generally presented in Figure 2. Data that is clean and complete indicates that the data is ready for modeling using the K-Means method.

2.5.2 Linear Regression

Linear regression is a method that consists of one or more independent variables symbolized by (X) and one dependent or response variable symbolized by (Y) [14]. In this research, the linear regression method will be used to see the causal relationship from one variable to another using the following equation:

$$\gamma = \alpha + \beta x \quad (2)$$

$$\beta = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2} \quad (3)$$

$$a = \frac{\sum Y \sum X^2 - \sum X \sum XY}{n\sum X^2 - (\sum X)^2} \quad (4)$$

Information:
γ = Criterion Variable
x = Predictor Variable
a = Constant value
β = Linear Regression Direction Coefficient Value
n = Number of data

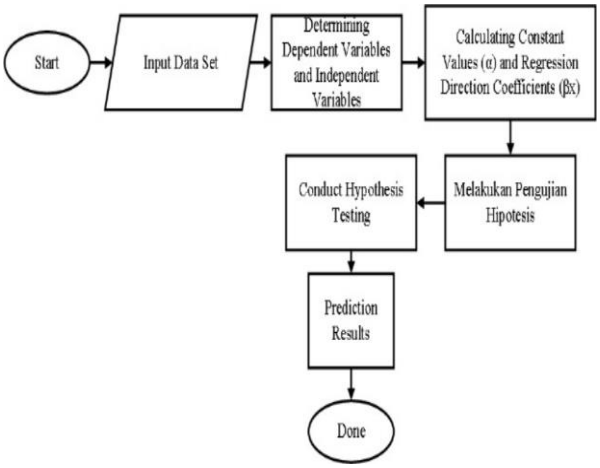

Figure 3. Flowchart Linear Regression

The general linear regression method is presented in Figure *3*. Furthermore, to find out the variables that influence the increase in the number of TB, additional data is needed, namely the population density variable in each sub-district.

In this research, the number of clusters will be determined using the Devices Boulding-Index (Dbi) value. The cluster that has the minimum Dbi value is the best. The first research will create a clustering model for the spread of TB using the K-means method using RapidMiner tools. Future research will look for the relationship between population density variables and the increase in the number of TB.

2.6 Research Results and Conclusion

In the final stage of this research, researchers presented the results of the clustering of TB distribution patterns and looked at the characteristics of each cluster. Furthermore, from the results of the analysis and evaluation, conclusions will be drawn from this research

3. Results and Discussion

3.1 Research Data

The data used in this research were 773 TB cases spread across 8 sub-districts in Bangka Regency. This data is data for 2023 and 2024 obtained from the Bangka District Health Service. The details of the data set that will be used in this research can be seen in Table 1.

Table 1. Distribution of Data on the Number of TB Patients in Each District

Sub-district	*An Area (km^2)*	*Total Population*	*Number of TBC Patients*
Bakam	593.52	19219	34
Belinyu	746.5	50977	93
Mendo Barat	614.37	51486	75
Merawang	207.27	31225	63
Pemali	127.87	35435	93
Puding Besar	383.29	20177	43
Riau Silip	523.68	29018	40
Sungailiat	147.99	94044	296

3.2 Analysis of K-Means Clustering and Linear Regression

3.2.1 Data Modelling using the K-Means Method

The data model created is a machine learning model using the K-means method. Researchers use the RapidMiner tool to process the cleaned data. The clustering process is carried out based on the steps previously explained.

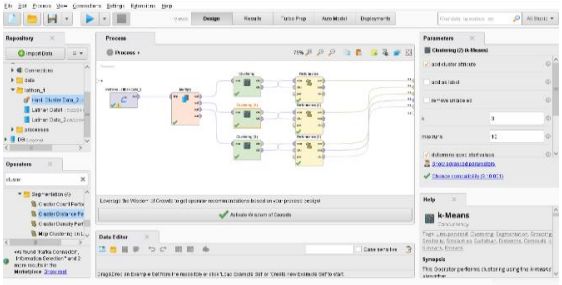

Figure 4. Clustering Process using the K-Means Method

Figure 4 shows a general overview of the clustering process in Rapid Miner. In the Rapidminer tool, 2 supporting operators can be used to support the clustering process, namely the multiply operator and the performance operator. The multiply operator is used to create more than one copy of an object. Meanwhile, the performance operator evaluates the performance of the resulting model.

The number of clusters is determined by looking at the clusters that have the optimum Dbi (Davies Bouldin-Index) value. The smaller or closer to 0 the Dbi value, the better the resulting cluster.Table 2 presents the Dbi values for each cluster.

Table 2. Cluster Calculation Results with K-Means

Cluster (k)	*Dbi Value*	*Cluster Member*
2	0.167	Cluster_0: 1 items, Cluster_1: 7 items
3	0.106	Cluster_0: 5 items, Cluster_1: 1 items, Cluster_2: 2 items
4	0.186	Cluster_0: 1 items, Cluster_1: 3 items, Cluster_2: 2 items, Cluster_3: 2 items

In Table 2 , the Clustering process with values K= 2 to K= 4 with 3 iterations, obtained the smallest Dbi value in cluster 3, namely 0.106. Thus, the optimal or best clustering to describe the spread of TB disease in Bangka Regency is in cluster 3. The initial stage of the clustering process using the K-means method is to initialize the initial centroid in the first iteration. An example of a cluster set in the first iteration is presented in Table 3.

Table 3. Centroid Value for the First Iteration

Cluster	**An Area**	**Total Population**	**Number of TBC Patients**
Cluster_0	593.52	19219	34
Cluster_1	746.5	50977	93
Cluster_2	614.37	51486	75

Calculate the Euclidean distance value of all objects to the centroid center of each cluster. Next, compare the three Euclidean distance values for each row of data. This stage will produce the closest distance value.

Calculates a new centroid using the average value of objects in each cluster member. This stage continues to repeat from step two if there is a change in cluster position. The iteration process will continue until convergence is achieved or there is no change in position from one cluster to another. The final results of the clustering of TB cases in Bangka Regency are presented in Table 4.

Table 4. Grouping TB Cases using the K-means Method

C0	**C1**	**C2**	**Minimum Distance**	**Cluster**
5776.87	71811.86	29951.16	5776.87	C0
25220.48	40824.892	1041.13	1041.13	C2
23140.92	42901.104	1041.13	1041.13	C2
4819.259	61221.472	19367.79	4819.25	C0

4970.51	61073.34	19221.94	4970.51	C0
6266.431	72304.825	30446.47	6266.43	C0
2263.068	63781.62	21921.60	2263.06	C0
66038.40	0	41862.92	0	C1

Table 4 is the process of completing the last iteration of centroid 3. In this last iteration, the position of each cluster does not change, so the iteration process is stopped. Therefore, it can be concluded that cluster_0 has five members, cluster_1 has one member, and cluster_2 has two members. The cluster results for the spread of TB across each sub-district using the K-Means Clustering method are visualization in Figure 5.

Figure 5. Visualization of the Distribution of TBC Data in Each Cluster

Figure 5 shows data on the spread of TB disease in eight sub-districts which are divided into 3 categories, namely: the area with the high spread category (orange) is Sungailiat, the area with the medium spread category (Gray) is Belinyu and West Mendo, and the area with the low spread category (Blue) are Bakam, Merawang, Pemali, Pudding Besar, Riau Silip.

The results of this research show that a high distribution of TB cases has been identified in Sungailiat District, which is a densely populated area in Bangka Regency. Then, the West Mendo and Belinyu regions were identified as areas with moderate prevalence of TB cases. These two areas are also densely populated residential areas. Meanwhile, the sub-districts of Bakam, Merawang, Pemali, Puding Besar and Riau Silip are in the category of low spread of TB cases because they have a large area and are not densely populated. The findings of this research are supported by previous research which states that conditions in dense areas can increase exposure to people suffering from TB, making it easier for TB bacilli to spread to other people [15].

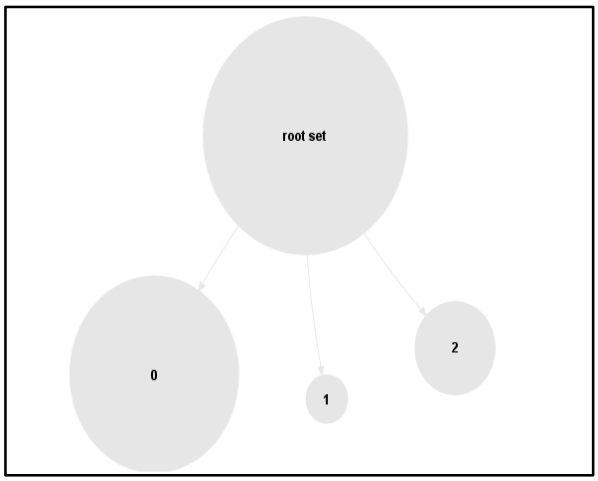

Figure 6. Data Comparison with Model Tree

Figure 6 shows the results of clustering mapping in Tree form which describes the size of the comparison between clusters and all data. With this tree visualization, we can see a breakdown of the root set or main data into several derivative data called clusters.

3.2.2 Linear Regression Calculations

After clustering TB disease which is spread across 8 sub-districts, this research then wants to look at the causal relationship between the influence of population density variables on the increase in the number of TB diseases in Bangka Regency using the linear regression method with the following stages:

1. Calculating Values XY, X^2, Y^2

Table 5.Value Calculation Results XY, X2, Y2

ID	X	Y	XY	X^2	Y^2
1	32.38	34	1100.92	1048.46	1156
2	68.29	93	6350.97	4663.52	8649
3	83..8	75	6285	7022.44	5625
4	150.65	63	9490.95	22695.42	3969
5	277.12	93	25772.16	76795.49	8649
6	52.64	43	2263.52	2770.96	1849
7	55.41	40	2216.4	3070.26	1600
8	644.82	296	190866.72	415792.83	87616
	1365.11	**737**	**244346.6**	**533859.4**	**119113**

Table 5 is the result of calculating the variables XY, X^2, Y^2 to find the total value.

2. Calculating Coefficient Values α and β

The next stage is to calculate the coefficient values α dan βusing the following equation:

$$a = \frac{\sum Y \sum X^2 - \sum X \sum XY}{n\sum X^2 - (\sum X)^2} = 24.880$$

$$\beta = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2} = 0.394$$

Table 6. Calculation of Coefficient Values with SPSS

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	24.880	12.875		1.932	.101			
	kepadatan_penduduk	.394	.050	.955	7.907	.000	.955	.955	.955

In Table *6* there is information on the significance value and t-count value as follows:

1. Signifance Value 0.000 < 0.05
2. T-Count > t-table = 9.248 > 1.943

Based on the two bases for decision-making above, it can be concluded that population density has a positive and very significant effect on increasing the number of TB diseases in each sub-district in Bangka Regency. This means that the denser the population in a sub-district, the quantity of TB disease spreading in Bangka Regency will increase.

After getting the coefficient values α and β, and determining their significance status, then look for the value of the coefficient of determination (r2) with the following equation:

$$r^2 = \frac{b(\sum XY)}{\sum Y^2} = 0.808$$

$$kd = r^2 * 100 = 81\%$$

Based on the results of calculating the coefficient of determination above, the coefficient of determination value is: 0.808 or a percentage of 81%. This means that the population density variable (X) influences the increase in the number of TB diseases (Y) by 81% while the remaining 19% is influenced by other variables.

Figure 7. The Effect of Population Density on the Increase in the Number of TB Diseases

Figure 7 shows the effect of population density on the increase in the number of TB diseases in Bangka Regency.

The results of research regarding the prediction of a causal relationship between population density variables to the increase in the spread of TB disease are supported by previous research [14]. This research also explains that population density has a positive influence on increasing the spread of tuberculosis. Congested areas can increase exposure to people with TB, making it easier for germs to spread to other people [15].

4. Conclusion

This research conducted effective clustering using the K-Means algorithm which resulted in a TB distribution pattern in Bangka Regency into three categories, namely areas with high prevalence (Sungailiat sub-district), areas with moderate prevalence (Belinyu and Mendo Barat sub-districts), and areas with low prevalence (Bakam, Merawang, Pemali, Puding Besar, and Riau Silip sub-districts). The researcher also found that population density affected the increase in TB cases by 81% and 19% was influenced by other variables. The results of the analysis of the two methods identified that population density was one of the factors affecting the spread of tuberculosis. The results of this study can then be used as a reference by the Bangka Regency Health Office to conduct surveillance, prevention, and treatment of TB in sub-districts included in the high-spread cluster. Further research is recommended to explore the parameters used in the K-means method and linear regression, supporting algorithms for the K-means method, and exploration tools to improve the results and performance of the clustering model.

Acknowledgement

Thank you to LPPM Bangka Belitung University for supporting this research program through the 2024 Young Researcher (PM) Funding Grant with contract number: 593/UN50/L/PP/2024. Thank you also to the Bangka District Health Service who have helped a lot in the smooth running of this research.

References

[1] T. K. A. Teibo, R. L. de P. Andrade, R. J. Rosa, R. B. V. Tavares, T. Z. Berra, and R. A. Arcêncio, "Geo Spatial High Risk Clusters of Tuberculosis in the Global General Population: a Systematic Review," *BMC Public Health*, vol. 23, no. 1, pp. 1–10, 2023.

[2] L. Dodo, N. S. Fatonah, G. Firmansyah, and H. Akbar, "Analysis of Tuberculosis Disease Case Growth From Medical Record Data, Viewed Through Clustering Algorithms (Case Study: Islamic Hospital Bogor)," *J. Indones. Sos. Sains*, vol. 4, no. 09, pp. 915–927, 2023.

[3] R. S. Wardani, Purwanto, Sayono, and A. Paramananda, "Clustering Tuberculosis in Children Using K-Means Based on Geographic Information System," in *AIP Conference Proceedings*, 2019, vol. 2114.

[4] S. Bagcchi, "WHO's Global Tuberculosis

Report 2022," *The Lancet Microbe*, vol. 4, no. 1, p. e20, 2023.

[5] P. A. Kusuma and A. U. Firmansyah, "Deteksi Penyebaran Penyakit Tuberkulosis dengan Algoritma K-Means Clustering Menggunakan Rapid Miner," *Teknol. Inform. dan Komput.*, vol. 8, no. 2, pp. 41–54, 2022.

[6] Y. Yao *et al.*, "Identification of spinal tuberculosis subphenotypes using routine clinical data: a study based on unsupervised machine learning," *Ann. Med.*, vol. 55, no. 2, p. , 2023.

[7] R. Iman, B. Rahmat, and A. Junaidi, "Implementasi Algoritma K-Means dan Knearest Neighbors (KNN) Untuk Identifikasi Penyakit Tuberkulosis Pada Paru-Paru," *Publ. Tek. Inform. dan Jar.*, vol. 2, no. 3, pp. 12–25, 2024.

[8] R. Afriansyah, D. Lanaya, L. Sari, M. Azrul, and M. Riyadi, "Perancangan Aplikasi Re-Tuberis (Remember Tuberculosis) Dalam Pelayanan Informasi Dan Kepatuhan Penggunaan Obat," *J. Manaj. Inf. Kesehat. Indones.*, vol. 11, no. 2, pp. 157–164, 2023.

[9] Rudini, Helda, and M. Qomariah, "The Effect of Cadres Training on Competence Of Tuberculosis Health Cadres At The Muntok Health Center In West Bangka Regency," *Eduhealth*, vol. 14, no. 02, pp. 1041–1047, 2023.

[10] M. Kossakov, A. Mukasheva, G. Balbayev, S. Seidazimov, D. Mukammejanova, and M. Sydybayeva, "Quantitative Comparison of Machine Learning Clustering Methods for Tuberculosis Data Analysis," *Eng. Proc.*, vol. 60, no. 1, pp. 1–14, 2024.

[11] F. N. R. Putri, N. C. H. Wibowo, and H. Mustofa, "Clustering of Tuberculosis and Normal Lungs Based on Image Segementation Results of Chan-Vese and Canny with K-Means," *Indones. J. Artif. Intell. Data Min.*, vol. 6, no. 1, p. 237, 2011.

[12] R. N. Pratistha and B. Kristianto, "Implementasi Algoritma K-Means dalam Klasterisasi Kasus Stunting pada Balita di Desa Randudongkal," *J. Indones. Manaj. Inform. dan Komun.*, vol. 5, no. 2, pp. 1193–1205, 2024.

[13] P. Apriyani, A. R. Dikananda, and I. Ali, "Penerapan Algoritma K-Means dalam Klasterisasi Kasus Stunting Balita Desa Tegalwangi," *Ilmu Komput.*, vol. 2, no. 1, pp. 20–33, 2023.

[14] M. Ula, A. Zulfikri, A. F. Ulva, and R. R. Achmad, "Penerapan Machine Learning Clustering K-Means dan Linear Regression dalam Penentuan Tingkat Resiko Tuberkulosis Paru," *Indones. J. Comput. Sci.*, vol. 12, no. 1, pp. 336–348, 2023.

[15] A. A. Lestar, M. R. Makful, and C. Okfriani, "Analisis Spasial Kepadatan Penduduk Terhadap Kasus Tuberkulosis di Provinsi Jawa Barat 2019-2021," *J. Cahaya Mandalika*, pp. 577–584, 2021.