JISS Journal of Indonesian Social Science

Vol. 4, No.09, September 2023
E-ISSN: 2723-6692
P-ISSN: 2723-6595
http://jiss.publikasiindonesia.id/

Analysis of Tuberculosis Disease Case Growth From Medical Record Data, Viewed Through Clustering Algorithms (Case Study: Islamic Hospital Bogor)

La Dodo[1], Nenden Siti Fatonah[2], Gerry Firmansyah[3], Habibullah Akbar[4]
[1,2,3,4] Universitas Esa Unggul, Indonesia
Email : ladodo881@student.esaunggul.ac.id, nenden.siti@esaunggul.ac.id, gerry@esaunggul.ac.id, habibullah.akbar@esaunggul.ac.id

KEYWORDS
Tuberculosis; Clustering; K-means; Fuzzy C-Means; Gaussian Mixture;

ABSTRACT
Tuberculosis is a chronic infectious disease caused by Mycobacterium tuberculosis infection. Tuberculosis can spread from one person to another through airborne transmission. This disease is most commonly found in the Asian region. Currently, Indonesia ranks second after India in terms of tuberculosis cases. The discovery of tuberculosis cases by province in Indonesia reveals that West Java Province is one of the contributors to the highest tuberculosis cases. It is known that the tuberculosis case rate in Bogor Regency is one of the highest in West Java. This serves as the foundation for the focus of this research, which will be conducted at Islamic Hospital Bogor, to determine the average age and gender of patients who are more susceptible to tuberculosis. One way to understand the growth of tuberculosis cases is through clustering using Data Mining Techniques, specifically several clustering algorithms such as k-means clustering, fuzzy c-means, and Gaussian mixture. These techniques aim to identify the growth of tuberculosis cases based on age range and gender. Therefore, the research results are expected to provide new insights, which could be valuable for decision-makers in various capacities, such as preventive measures, healthcare facility provision, and medication considerations.

1. Introduction

Tuberculosis is a chronic infectious disease caused by infection with Mycobacterium tuberculosis. This tuberculosis is most commonly found in Southeast Asia (44%) and Africa (24%) (Istiharoh, Djannah, &; Ajmala, 2022). Currently , Indonesia ranks second after India related to tuberculosis. Currently, it is known that Indonesia ranks second after India related to tuberculosis (TB), with 969 thousand cases and 93 thousand deaths per year or equivalent to 11 deaths per hour (WHO, 2022). The coverage of TB case discovery according to provinces in Indonesia in 2017, the highest cases are in West Java Province with a population of 48,037,827 people with case findings of 31,598 cases, East Java with a population of 39,292,972 people with 22,585 case findings, Central Java with a population of 34,257,865 people (Yani, Pebrianti, &; Purnama, 2022).

Tuberculosis has a relationship between humans and their environment, especially in urban areas that have the highest population and density, so accurate information about the urban environment of tuberculosis areas is important. According to data obtained from West Java Open Data, namely Tuberculosis Data in West Java Province displays data that in 2020 all cities and regencies in the West Java region had a number of Tuberculosis cases starting from 320 cases in Banjar Regency which is the lowest case, and 10,248 cases in Bogor Regency which is the highest case in West Java (Fadhlan Sulistiyo Hidayat1, Rizma Berliana Putri Affandi2, Virgaria Zuliana3, 2022).

The basis of research that will be the focus of this study will be carried out at Bogor Islamic Hospital, to find out patients with what average age and gender are more susceptible to tuberculosis. One way to determine the growth of tuberculosis cases is to cluster with Data Mining Techniques, namely in several clustering algorithms, namely cluster k-means, fuzzy c-means and gaussian mixture to determine the growth of tuberculosis cases based on age range and gender. Thus, the results of the research are expected to become new information, which can later be one of the considerations for related parties in decision making, such as preventive measures, provision of health facilities, and medicines.

2. Materials and Methods

This section will explain the Literature Review, and the methods used in building basic knowledge in the context of the topic under study.

2.1 Tuberculosis

Tuberculosis (TB) is a chronic infectious disease caused by Mycobacterium tuberculosis infection and can be cured. Tuberculosis can spread from one person to another through airborne transmission (phlegm droplets of tuberculosis patients). Patients infected with tuberculosis will produce droplets containing a number of TB germ bacilli when they cough, sneeze, or talk. People who inhale these TB germ bacilli can become infected with Tuberculosis (Oktaviani, Sumarni, & Supriyanto, 2023).

2.2 Data Mining

Data mining integrates data modeling and analytics. Although based on several disciplines, data mining differs from them in its orientation towards the end rather than the means to achieve it, utilizing all these disciplines to extract patterns, describe trends, and predict behavior, utilizing information. Data mining is only one stage, but the most important, in the process of knowledge discovery in databases (KDD). Note that KDD is defined as a non-trivial process for identifying valid, new, potentially useful, and ultimately understandable patterns in often large data sets, and for extracting relevant information from available databases. The KDD methodology includes an iterative and interactive process in which the subject's experience is combined with various analytical techniques including ML algorithms for pattern recognition and modeling development (Palacios, Reyes-Suárez, Bearzotti, Leiva, & Marchant, 2021).

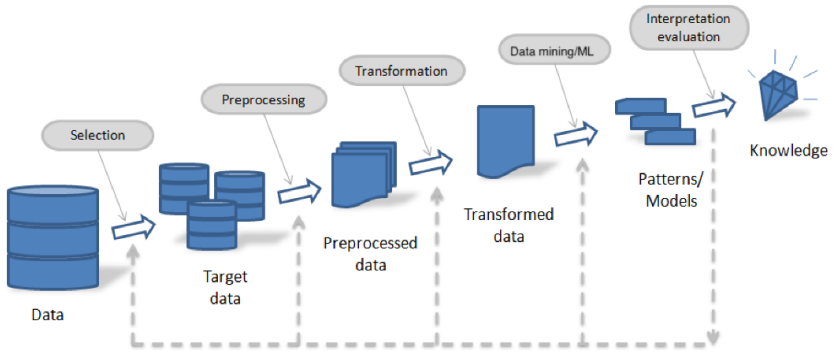

Figure 1. KDD Methodology Scheme.

2.3 Clustering

Clustering is an unattended procedure for organizing data into groups of similar pattern items typical for each group. Grouping procedures can be classified as hierarchical or non-hierarchical. Group objects group hierarchical into clusters and define relationships between items in the cluster. In contrast, non-hierarchical methods group items into clusters without establishing relationships between objects in the same cluster (Agapito, Milano, & Cannataro, 2022).

2.4 K-means

The K-Means algorithm is one of the clustering algorithms in data mining to group data. The K-Means algorithm can partition data into two or more groups based on nonhierarchical data groupings. This method will group data into groups that have the same data characteristics while data with different characteristics will be added to other groups (H. Syukron et al. 2022). The K-means algorithm method starts with a specific value for K (number of categories) and tries to categorize a specific set of samples in group K so that the hypothesis is expressed in the Equation (Shirazy, Hezarkhani, Shirazi, Khakmardan, & Rooki, 2022).

2.5 Fuzzy c-means

The Fuzzy C-Means method groups data by degree of membership. Data can be grouped by degree of membership, which ranges from 0 to 1, and there are some data types that only display partial membership. Fuzzy Clustering is used by fuzzy C-Means to assign data ownership to each cluster that each has a different membership. The degree of membership controls the range between 0 and 1 of data presence in the cluster. The Fuzzy C-Means algorithm has an excellent advantage in detecting high-level clusters and revealing relationships between various cluster models (Syukron, Fayyad, Fauzan, Ikhsani, & Gurning, 2022).

2.6 Gaussian mixture model

Gaussian Mixture Model is a method that models or clusters the data of a dataset into several groups of data that have a Gaussian or Normal probabilistic distribution. This method assumes that all individuals are a mixture of Gaussian probability distributions, representing Gaussian distributions where each distribution typically has distribution parameters (Joko Riyono et al. 2022).

2.7 Literature review

Systematic Literature Review (SLR) was chosen with the aim of justifying based on previous research related to mathematical proof ability. This research stage includes data collection, data analysis, and conclusion drawing (Niken Shofiana Dewi et al. 2023). At the data collection stage, researchers traced and collected data in the form of primary research conducted at Bogor Islamic Hospital, where the data taken was 2022 data related to tuberculosis, where the data was sourced directly from patient medical record data.

3. Results and Discussions

This study aims to analyze the growth of tuberculosis (TB) cases through the application of *clustering algorithms*. *Clustering* is a data analysis technique used to group entities that have similar characteristics into larger groups. The dataset used in this study is a collection of medical record data from 2217 patients diagnosed with tuberculosis. The dataset used in this study consists of three main attributes that carry critical information about patients suffering from Tuberculosis (TB). These attributes are Control Month, Gender and Age or Age.

3.1 *Output*

Silhouette Score and Davies-Bouldin Score are two evaluation metrics used to measure clustering quality in data analysis. These two metrics provide guidance on how well data has been grouped into meaningful clusters. The Silhouette Score measures how close each sample is to the cluster in which it is located compared to its nearest neighboring cluster. This metric provides information about how well objects are in their own cluster and how separated that cluster is from other clusters. In cluster analysis, the higher the Silhouette Score, the better the clustering results. But it should be remembered that Silhouette Score values need to be analyzed along with visual interpretation of clustering results. The Davies-Bouldin Score measures how well segregated the clusters are. This metric measures the average distance between each cluster and the other clusters that are most similar to that cluster. The lower the Davies-Bouldin Score, the better the separation between clusters.
Interpretasi Skor Davis Bouldin dan Silhouette Score

Table 4. Clustering Algorithm Score

Fuzzy C Means		
Number of Clusters	**Silhouette Score**	**Davies-Bouldin Score**
2	0.7131426951986491	0.473580151568806
3	0.6682199080834739	0.44616784733542697

4	0.6672766680235653	0.46084480850830717
5	0.6410211226039462	0.49929350991129373
6	0.6110511492572396	0.5043799808225232
7	0.6371414429917497	0.4691595839153332
8	0.647705746685325	0.46136161836197553
9	0.6460257591825395	0.46685695249570475
10	0.6488268141245808	0.4554664738833975
11	0.6116747817753289	0.4848766791787212
12	0.6381535519277991	0.46490983263892033
13	0.6367596197792532	0.46344160654927224
14	0.6012438348431733	0.5156456343120637
15	0.620962198747955	0.48205268446337557
16	0.6225560879216999	0.472789780864476
17	0.617671528579297	0.485212024963306
18	0.5931398622921469	0.5054598618227805
19	0.5995111030478238	0.48754718778101
Gaussian Mixture		
Number of Clusters	**Silhouette Score**	**Davies-Bouldin Score**
2	0.7131426951986491	0.473580151568806
3	0.6026823034662715	0.43008430558318445
4	0.5601863236966083	0.47450630140941785
5	0.6001630026569527	0.5029887001405555
6	0.5504530089149884	0.4877933110982933
7	0.5526637284270811	0.4971928519875118
8	0.5658261969577871	0.47962907060109805
9	0.628364525158927	0.4474447809829992
10	0.6296352062103897	0.4384929752507222
11	0.6106823932381024	0.44794456843755764
12	0.6264848888694038	0.470053175691856
13	0.6212655701926757	0.4797683103655699
14	0.6267127341368781	0.47167731733492885
15	0.6190359926890503	0.4737341744724074
16	0.6005795982805149	0.48416480536128736
17	0.6045909304733575	0.4792843955920918
18	0.5922989833543447	0.4801290343991462
19	0.6011277277858451	0.4749104386077319
K-Means		
Number of Clusters	**Silhouette Score**	**Davies-Bouldin Score**
2	0.7131426951986491	0.473580151568806
3	0.6967634535817717	0.3963345463696839

4	0.6676012761223927	0.45998825008499394
5	0.6486474763057268	0.49453608817460903
6	0.626405781078509	0.4897458822191436
7	0.6439078252490292	0.4663792522278512
8	0.6495560288433111	0.4607587770754166
9	0.6522026739203765	0.44744170857819615
10	0.6449907868081267	0.46627219634848743
11	0.6370873630338417	0.4703648772380841
12	0.62762713248547	0.47707993812004873
13	0.6346295398784288	0.4887019547928556
14	0.6361333990019239	0.48370525379532614
15	0.6136429520920142	0.4856185767067282
16	0.610142121843062	0.49251461441289013
17	0.6134654911393829	0.4828465633858001
18	0.6190055616032395	0.4600192000140335
19	0.6154924658652092	0.47448213190246197

Basically, choosing the right number of clusters *involves a trade-off* between *Silhouette Score* and *Davies-Bouldin Score*. The main goal is to find the number of clusters that have good separation and cohesion. Based on the results of evaluating clustering metrics, especially *Silhouette Score and* Davies-Bouldin Score*, and* considering the overall performance of the algorithm, it can be concluded that the number of 9 clusters is the most optimal choice used for complex clustering. Overall, the total of 9 clusters provides a good balance between cluster separation and cohesion, which is reflected in the relatively high and low *Silhouette Score* and *Davies-Bouldin Score*, respectively.

3.2 Average Age by Gender Cluster

Table 5. Average Age by Gender Cluster

AVERAGE TUBERCULOSIS AGE BY GENDER CLUSTER									
FUZZY C MEANS									
SEX	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
MALE			52,90		38,42		21,94	68,12	3,70
FEMALE	23,20	4,36		63,78		43,87			
GAUSSIAN MIXTURE									
SEX	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
MALE		21,46		51,93		3,68		66,80	38,01
FEMALE	3,34		43,36		62,74		21,94		
K-MEANS									
SEX	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**

MALE		3,70		52,43		67,47	21,74	38,25
FEMALE	4,36		43,60		23,20		63,48	

From the results of clustering analysis of cases of tuberculosis patients based on age and sex, we can identify several meanings that can be taken:

3.3 Age Difference between Men and Women with Tuberculosis

There is significant variation in the age distribution of men and women with tuberculosis. The age of male sufferers is divided into several groups that cover a wider age range, including younger and older age groups. On the other hand, the age of female sufferers tends to be more focused on certain age groups, namely young age groups and older age groups. The presence of a very young group of men (less than 5 years) may indicate the risk of mother-to-child transmission of tuberculosis or early exposure to the disease at an early age.

3.4 Young Age Group in Women

There is a group of women with tuberculosis with a very young age (less than 5 years). This could indicate a case of mother-to-child transmission or early exposure to tuberculosis at an early age. Special care and attention is needed to prevent transmission and ensure proper care of this group.

3.5 Elderly Age Group in Men and Women

There are older groups of men and women with tuberculosis, especially in some clustering methods. This could indicate a higher risk in the elderly group. Prevention, early detection, and appropriate treatment efforts are needed to overcome these cases.

3.6 Differences in clustering methods

Different clustering methods can result in different age groups. For example, the Fuzzy C Means method tends to produce more groups with greater age variation, while the Gaussian Mixture and K-Means methods are more likely to produce more focused age groups.

These meanings provide a view of the age profile of men and women with tuberculosis in relation to grouping based on clustering methods. This information can provide healthcare professionals with insights into designing more effective prevention, detection, and treatment strategies for different age groups and genders.

3.7 Average Age by Cluster

Table 6. Average Age by Cluster

AVERAGE AGE BY CLUSTER		
FUZZY C	**GAUSSIAN**	**K-MEANS**
AGE	**AGE**	**AGE**
3,71	3,34	3,71
4,37	3,68	4,37

21,94	21,46	21,75
23,21	21,95	23,21
38,43	38,02	38,25
43,88	43,36	43,61
52,90	51,93	52,43
63,78	62,74	63,49
68,13	66,80	67,48

Table 5 illustrates the average age in clusters generated by three clustering algorithms, namely Fuzzy C-Means, Gaussian Mixture, and K-Means. Each row in the table represents a different age group and each column represents a different clustering algorithm.

From the results of this table, it can be seen that the average age in clusters generated by Fuzzy C-Means and K-Means algorithms tends to be relatively similar. For example, in the first cluster, the average age for Fuzzy C-Means was 3.71 while K-Means was 3.71 as well. Similarly, for the latter cluster, the average age of Fuzzy C-Means was 68.13 while K-Means was 67.48. This suggests that both algorithms tend to generate age groups that have similar age characteristics. On the other hand, the Gaussian Mixture algorithm has more significant differences in some cases. For example, in the first cluster, the mean age for the Gaussian Mixture was 3.34, which is lower compared to Fuzzy C-Means and K-Means. This suggests that the Gaussian Mixture algorithm can generate different age groups than other algorithms in some situations. In conclusion, this table provides an overview of how the three clustering algorithms behave in age grouping. There are notable differences in some cases, especially in Gaussian Mixture algorithms, while Fuzzy C-Means and K-Means algorithms tend to produce more similar results in terms of average age in clusters.

3.8 Cluster Growth

Table 8. Cluster Growth Fuzzy c-means Male

Fuzzy C-Means					
Month	**MALE (Year)**				
	0-13	**14-30**	**31-45**	**46-60**	**61-85**
JANUARY	40	27	20	17	12
FEBRUARY	38	18	13	21	8
MARCH	56	12	25	20	10
APRIL	58	16	22	24	8
MAY	29	16	19	21	11
JUNE	6	6	3	2	3
JULY	42	17	12	19	16
AUGUST	25	9	15	8	11
SEPTEMBER	0	4	8	22	16
OCTOBER	31	13	14	10	9
NOVEMBER	37	16	24	13	14
DESEMBER	16	16	12	34	31
TOTAL	378	170	187	211	149
AVG	34,52%	15,53%	17,08%	19,27%	13,61%

Overall, the average percentage of TB sufferers in the age group tends to be high in the age group 0-13 years, with a value of around 34.52%, and in the age group of 46-60 years with a value of 19.27%. While the age groups of 14-30 years, 31-45 years, and 61-85 years have lower percentages, respectively are 15.53%, 17.08%, and 13.61%. These results suggest that in the male population, the age group of children and the elderly group tend to be more susceptible to TB infection, while the age group of young adults has a lower risk.

Table 9. Cluster Growth FCM Female

FUZZY C-Means				
Month	FEMALE (Years)			
	0-13	14-33	34-54	55-81
JANUARY	31	32	24	16
FEBRUARY	30	28	20	12
MARCH	35	31	25	19
APRIL	40	29	23	20
MAY	28	37	26	14
JUNE	7	4	6	4
JULY	40	29	17	20
AUGUST	18	25	14	14
SEPTEMBER	4	22	38	30
OCTOBER	37	22	14	12
NOVEMBER	31	27	22	14
DESEMBER	14	28	39	50
TOTAL	315	314	268	225
AVG	28,07%	27,99%	23,89%	20,05%

This information indicates that in the female population, the age groups of young adolescents (14-33 years) and children (0-13 years) are the groups that tend to be susceptible to TB infection. However, the middle adult age group (34-54 years) and the elderly age group (55-81 years) are also inseparable from the risk.

Table 10. Cluster Growth GM Male

Gaussian Mixture					
Month	MALE (Years)				
	0-11	12-28	29-45	46-58	59-86
JANUARY	40	26	21	16	13
FEBRUARY	38	17	14	20	9
MARCH	56	11	26	19	11
APRIL	57	15	24	22	10
MAY	29	15	20	20	12
JUNE	6	6	3	2	3
JULY	42	16	13	15	20
AUGUST	25	9	15	6	13
SEPTEMBER	0	4	8	17	21
OCTOBER	31	12	15	8	11

NOVEMBER	37	15	25	12	15
DESEMBER	16	16	12	27	38
TOTAL	377	162	196	184	176
AVG	34,43%	14,79%	17,90%	16,80%	16,07%

These results suggest that TB sufferers in the male population spread evenly across age groups, but young children and adolescents (1-28 years) appear to be more susceptible to infection. In addition, older age groups, especially 59-86 years, also have a significant risk of the disease.

Table 11. Cluster Growth GM Female

Gaussian Mixture				
Month	**FEMALE (Years)**			
	1-8	**9-34**	**35-51**	**52-81**
JANUARY	27	38	21	17
FEBRUARY	30	30	15	15
MARCH	30	36	24	20
APRIL	36	34	22	20
MAY	23	42	25	15
JUNE	6	5	6	4
JULY	37	33	16	20
AUGUST	13	30	14	14
SEPTEMBER	2	25	29	38
OCTOBER	28	32	13	12
NOVEMBER	27	32	17	18
DESEMBER	11	32	32	56
TOTAL	270	369	234	249
AVG	24,06%	32,89%	20,86%	22,19%

These results indicate that the 9-34 age group, especially adolescents and young adults, has a higher risk of TB infection in the female population. In addition, the age group of 35-51 years also has a significant risk.

Table 12. Cluster Growth K-means Male

K-Means					
Month	**MALE (Years)**				
	0-12	**13-29**	**30-45**	**46-59**	**60-86**
JANUARY	40	27	20	17	12
FEBRUARY	38	18	13	21	8
MARCH	56	12	25	20	10
APRIL	58	14	24	23	9
MAY	29	16	19	21	11
JUNE	6	6	3	2	3
JULY	42	16	13	16	19
AUGUST	25	9	15	6	13
SEPTEMBER	0	4	8	19	19
OCTOBER	31	12	15	8	11

NOVEMBER	37	16	24	12	15
DESEMBER	16	16	12	33	32
TOTAL	378	166	191	198	162
AVG	34,52%	15,16%	17,44%	18,08%	14,79%

In general, the age groups of 1-12 years and 30-45 years have a higher number of cases of TB patients in the male population, with an average percentage of about 34.52% and 17.44% respectively. Followed by the age groups of 13-29 years, 46-59 years, and 60-86 years, with average percentages of around 15.16%, 18.08%, and 14.79%, respectively. These results suggest that the age groups 1-12 years and 30-45 years have a higher risk of TB infection in the male population.

Table 13. Cluster Growth K-means Female

K-Means				
Month	**FEMALE (Years)**			
	0-13	**14-33**	**34-53**	**54-81**
JANUARY	31	32	23	17
FEBRUARY	30	28	20	12
MARCH	35	31	25	19
APRIL	40	29	23	20
MAY	28	37	26	14
JUNE	7	4	6	4
JULY	40	29	17	20
AUGUST	18	25	14	14
SEPTEMBER	4	22	35	33
OCTOBER	37	22	14	12
NOVEMBER	31	27	20	16
DECEMBER	14	28	38	51
TOTAL	315	314	261	232
AVG	28,07%	27,99%	23,26%	20,68%

Overall, the age groups of 1-13 years and 14-33 years have a fairly high number of TB cases in the female population, with an average percentage of about 28.07% and 27.99% respectively. Followed by the age groups of 34-53 years and 54-81 years, with average percentages of around 23.26% and 20.68% respectively. These results suggest that the age group 1-33 years has a higher risk of TB infection in the female population.

The fundamental difference between these three algorithms lies in the mathematical approach and logic behind grouping data. Fuzzy c-means uses fuzzy membership degrees to indicate the extent to which data is involved in each cluster. The Gaussian mixture focuses on modeling the probability distribution of data assuming the data comes from Gaussian distributions that may overlap. K-Means, on the other hand, seeks to group data into cluster centers based on the shortest distance.

4. Conclusion

The results of this study provide a deep understanding of the growth characteristics of TB disease and its implications for decision making in the health sector. The results of clustering analysis using three different algorithms, namely Fuzzy C-Means (FCM), Gaussian Mixture (GM), and K-Means, have provided mixed views on the growth of TB disease in certain age groups and genders.

Based on these results, several conclusions can be drawn as follows: TB Growth in Age Groups

Clustering results consistently show that the age group of children, especially at the age of 3-4 years, tends to have a higher level of risk of TB infection. This pattern holds true in both male and female populations. Therefore, children were identified as the most vulnerable group to TB disease.

The Effect of Sex on TB Growth

Despite differences in the distribution of the number of TB cases between men and women, clustering results show that the age group of male children is more susceptible to TB disease than the female age group. Nonetheless, these results confirm that young age groups remain the more vulnerable group to infection, regardless of gender.

Health and Decision Implications

The information generated from this clustering analysis has important implications in TB prevention and management. A focus of attention on children's age groups, particularly those aged 3-4 years, is critical in formulating effective prevention strategies. In addition, awareness of higher levels of risk in younger age groups needs to be used as a basis for the allocation of health resources.

Through clustering analysis and interpretation of the results, this study makes an important contribution in understanding the growth characteristics of TB disease in male and female populations at RSI Bogor. This information is expected to provide guidance for efforts to prevent and manage TB disease more effectively and efficiently, especially with a focus on the age group of children.

5. References

Agapito, Giuseppe, Milano, Marianna, & Cannataro, Mario. (2022). A python clustering analysis protocol of genes expression data sets. *Genes*, *13*(10), 1839.

Fadhlan Sulistiyo Hidayat1, Rizma Berliana Putri Affandi2, Virgaria Zuliana3, Tesa Nur Padilah4. (2022). *Penerapan K-Means Clustering dalam Pengelompokan Kasus Tuberkulosis di Provinsi Jawa Barat Fadhlan*. *8*(September), 219–227.

Istiharoh, Tri Riskinie, Djannah, Fathul, & Ajmala, Indana Eva. (2022). Hubungan Antara Gambaran Sitologi Menggunakan Metode Fnab Dengan Respons Terapi Pada Pasien Limfadenitis Tuberkulosis Di Pulau Lombok. *Jurnal Kedokteran Unram*, *11*(4), 1169–1175.

Oktaviani, Siska Dewi, Sumarni, Tri, & Supriyanto, Teguh. (2023). Studi Kasus Implementasi Batuk Efektif pada Pasien dengan Tuberkulosis Paru. *Jurnal Penelitian Perawat Profesional*, *5*(2), 875–880.

Palacios, Carlos A., Reyes-Suárez, José A., Bearzotti, Lorena A., Leiva, Víctor, & Marchant, Carolina. (2021). Knowledge discovery for higher education student retention based on data mining: Machine learning algorithms and case study in Chile. *Entropy*, *23*(4), 485.

Shirazy, Adel, Hezarkhani, Ardeshir, Shirazi, Aref, Khakmardan, Shayan, & Rooki, Reza. (2022). K-means clustering and general regression neural network methods for copper mineralization probability in Chahar-Farsakh, Iran. *Türkiye Jeoloji Bülteni*, *65*(1), 79–92.

Syukron, Hamdi, Fayyad, Muhammad Fauzi, Fauzan, Farin Junita, Ikhsani, Yulia, & Gurning, Umairah Rizkya. (2022). Perbandingan K-Means K-Medoids dan Fuzzy C-Means untuk Pengelompokan Data Pelanggan dengan Model LRFM: Comparison K-Means K-Medoids and Fuzzy C-Means for Clustering Customer Data with LRFM Model. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *2*(2), 76–83.

WHO. (2022). *Global Tuberculosis Report*.

Yani, Desy Indra, Pebrianti, Riani, & Purnama, Dadang. (2022). Gambaran Kesehatan Lingkungan Rumah pada Pasien Tuberkulosis Paru. *Jurnal Keperawatan Silampari*, *5*(2), 1080–1088. https://doi.org/10.31539/jks.v5i2.3548