Journal of Computer Science, Information Technology and Telecommunication Engineering (JCoSITTE)
Vol. 6, No. 1, March 2025, pp. 730~736
ISSN: 2721-3838, DOI: 10.30596/jcositte.v6i1.21669

Optimization of K-Means Clusteting with Elbow Method for Identification of TB Prone in Central Java

Liem, Angelita Rafika Hanjaya Putri[1], Nur Wakhidah[2]
[1,2]Study Program of Informatics Engineering, Semarang University, Indonesia

ABSTRACT

Tuberculosis (TB) is an infectious disease caused by the bacillus Mycobacterium tuberculosis. This infectious disease needs more attention because the transmission rate is still high, especially in areas with high case finding rates and low treatment success rates. This study was conducted to monitor areas prone to TB case transmission, especially areas in Central Java Province. The K-Means method was applied to determine areas prone to TB transmission by considering the success rate of TB treatment, through the Elbow approach to determine the optimal number of clusters. The results of clustering show areas that are classified as vulnerable areas, which are areas with high transmission cases but low treatment success. Visualization of clustering results in scatter plots that clarify the division of areas into clusters and help in identifying treatments and areas that need further intervention.

Keyword : Tuberculosis; K-Means; Elbow Method.

Corresponding Author:
Nur Wakhidah,
Universitas Semarang
Jl. Soekarno Hatta, RT.7/RW.7, Tlogosari Kulon, Kec. Pedurungan, Kota Semarang, Jawa Tengah 50196.
Email : ida@usm.ac.id

Article history:
Received Nov 04, 2024
Revised Nov 06, 2024
Accepted Mar 10, 2025

1. INTRODUCTION

Tuberculosis (TB) is a contagious infectious disease caused by the mycobacterium tuberculosis bacillus (Adhanty and Syarif 2023). This infectious disease remains a public health problem worldwide, attacking the lungs and other parts of the body (Desmiany Duri, Afriansya, and Rizal Maulana 2023). TB has long been a challenge in infectious disease control efforts because of the high level of transmission, especially in densely populated areas, such as Central Java. If TB treatment is not completed, it can lead to failure from serious, potentially fatal complications to death (Pardosi et al. 2024).

Significant TB case finding rates are still found in various regions in Indonesia, including Central Java. In addition to the high case finding rate in each region, the treatment success rate is still relatively low. It is necessary to identify areas that are classified as vulnerable, with high case finding rates and low treatment success rates. In an effort to control and manage TB cases, in order to focus health interventions on the right areas.

There have been many studies using clustering to determine areas prone to disease cases. In previous research (Nur Amalia, Umaidah, and Mayasari 2024), researchers used the K-Means algorithm to determine areas prone to infectious diseases in Karawang Regency using the Knowledge Discovery in Database method in the 2020 to 2023 study. In a previous study (Asmiatun 2019), researchers used the K-Medoids algorithm to classify road conditions in Semarang City using the Manhattan Distance approach and evaluate the results with the Silhouette Coefficient. Other previous research (Stepanus Ginting et al. 2022) focuses on calculating the distance to determine the level of domestic violence with various distance measurement methods and K-Means.

In this study, after knowing the TB cases that need attention, researchers conducted clustering using K-Means as a solution. K-Means will classify the cities/districts in Central Java. From two indicators, namely "TB Treatment Success Rate (%)" and "TB Discovery Rate per 100,000 population" in 2023. The Elbow Method approach was used in this study to determine the optimal number of clusters (Safira and Agus Sugianto 2024). So that it can clarify the areas that are classified as prone to the spread of tuberculosis.

Overall, this research aims to identify areas that need special attention to deal with TB in Central Java Province. The K-Means algorithm with the Elbow method approach is a practical solution to map vulnerable areas, so that health interventions can be carried out on target. The results of this study are

expected to help the health sector in Central Java Province to pay attention to TB cases. So that there is an effective policy in handling TB cases.

2. RESEARCH METHODS

2.1 Research Stages

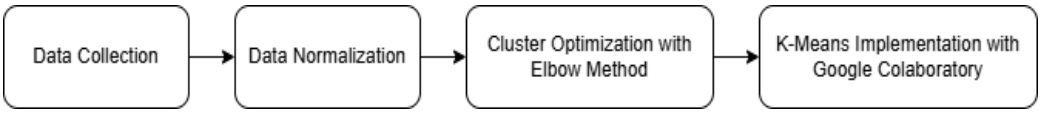

Fig 1. Research Stages

This study begins with the collection of data obtained from Statistics Center Agency Central Java Province. The raw data in excel form will be normalized to neutralize the attributes. After normalization, cluster search is conducted by applying Elbow Method and clustered using K-Means algorithm with Google Collaboratory tools.

2.2 Data Collection

The data used in this study is a secondary source taken from the Central Java Provincial Statistics Agency website related to Tuberculosis in Central Java Province for the year 2023. The data can be found in table 1 below:

Table 1. TB data from Statistics Center Agency Cental Java Province

District/City	TB discovery rate /100,000 population	TB Treatment Success Rate (%)
	2023	2023
CENTRAL JAVA PROVINCE	226	87
Cilacap Regency	218	83
Banyumas Regency	377	90
Purbalingga Regency	241	90
Banjarnegara Regency	137	90
Kebumen Regency	208	82
Purworejo Regency	123	81
Wonosobo Regency	201	89
Magelang Regency	75	87
Boyolali Regency	125	82
Klaten Regency	140	88
Sukoharjo Regency	197	91
Wonogiri Regency	151	90
Karanganyar Regency	107	93
Sragen Regency	124	85
Grobogan Regency	154	92
Blora Regency	156	88
Rembang Regency	216	88
Pati Regency	189	85
Kudus Regency	304	88
Jepara Regency	144	87
Demak Regency	163	90
Semarang Regency	86	90
Temanggung Regency	104	80
Kendal Regency	239	81
Batang Regency	176	85
Pekalongan Regency	209	94

Pemalang Regency	222	91
Tegal Regency	315	89
Brebes Regency	295	83
Magelang Town	1122	79
Surakarta Town	503	86
Salatiga Town	610	90
Semarang Town	443	84
Pekalongan Town	460	95
Tegal Town	1226	87

2.3 Data Normalization

Data normalization aims to change the data scale so that each variable has the same range of values (Kusnaidi, Gulo, and Aripin 2022). In the K-Means method, data normalization is necessary, because the distance between data greatly affects the clustering results. The formula for normalization used in this study is (AKILLI and ATIL 2020):

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

Description:
Xmax = Maximum value in the data
Xmin = Minimum value in the data
X = Original value in the data
X' = Normalized value

2.4 Cluster Optimization with Elbow Method

K-Means is an algorithm that requires an input number of clusters, choosing the right number of clusters is important to obtain accurate clustering results (Ahmed, Seraj, and Islam 2020). The Elbow method is used to identify the point where increasing the number of clusters does not procide a significant improvement in the reduction of within-cluster sum of squares (WCSS) (Putu et al. 2024). Within-cluster sum of squares is the total distance between the data and the cluster center (Sari et al. 2022). The Elbow method helps to find the point where the decline in WCSS begins to slow down, known as the "elbow point". WCSS is calculated by the formula (Cui 2020):

$$WCSS = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \quad (2)$$

Description:
C_i = i[th] cluster
X = data in the cluster C_i
μ_i = cluster center C_i
$\|x - \mu_i\|$ = distance between data x and the cluster center μ_i

3. RESULTS AND DISCUSSION

3.1Implementation of Data Normalization

Explanation sub section 1 Normalization is done by giving the command MinMaxScaler() and fit_transform (X). MinMaxscaler is one of the classes from sklearn.preprocessing in the scikit learn library (sklearn) which is used to perform normalization with the Min-Max Scaling method (Waldia et al. 2021). The fit_transform(X) command is a method that performs two steps at once, namely fit(x) and transform(x). In fit(x), python learns the minimum and maximum values of each feature in the data x. While in transform(x), python will use the minimum and maximum values to scale the data into a new range between 0 and 1. Here are the results of applying data normalization:

Table 2. Data Normalization Results

District/City	TB discovery rate /100,000 population 2023	TB Treatment Success Rate (%) 2023
CENTRAL JAVA PROVINCE	0.131190	0.5000
Cilacap Regency	0.124240	0.2500
Banyumas Regency	0.262381	0.6875
Purbalingga Regency	0.144222	0.6875
Banjarnegara Regency	0.053866	0.6875
Kebumen Regency	0.115552	0.1875
Purworejo Regency	0.041703	0.1250
Wonosobo Regency	0.109470	0.6250
Magelang Regency	0.000000	0.5000
Boyolali Regency	0.043440	0.1875
Klaten Regency	0.056473	0.5625
Sukoharjo Regency	0.105995	0.7500
Wonogiri Regency	0.066030	0.6875
Karanganyar Regency	0.027802	0.8750
Sragen Regency	0.042572	0.3750
Grobogan Regency	0.068636	0.8125
Blora Regency	0.070374	0.5625
Rembang Regency	0.122502	0.5625
Pati Regency	0.099044	0.3750
Kudus Regency	0.198957	0.5625
Jepara Regency	0.059948	0.5000
Demak Regency	0.076455	0.6875
Semarang Regency	0.009557	0.6875
Temanggung Regency	0.025195	0.0625
Kendal Regency	0.142485	0.1250
Batang Regency	0.087750	0.3750
Pekalongan Regency	0.116421	0.9375
Pemalang Regency	0.127715	0.7500
Tegal Regency	0.208514	0.6250
Brebes Regency	0.191138	0.2500
Magelang Town	0.909644	0.0000
Surakarta Town	0.371851	0.4375
Salatiga Town	0.464813	0.6875
Semarang Town	0.319722	0.3125
Pekalongan Town	0.334492	1.0000
Tegal Town	1.000000	0.5000

3.2 Implementation of Elbow Method

The Elbow Method is user to determine the optimal total clusters in the K-Means algorithm. In Google Colaboratory, the elbow method is applied with the following command:

```
inertia = []
K = range(1, 11)
for k in K:
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(X)
    inertia.append(kmeans.inertia_)
```

Fig 2. Elbow Method Implementation

The inertia variable is used to store the WCSS value (Maulindar 2023), for each value of k (number of clusters). The range of the number of clusters is tested from 1 to 10 through a for loop on the variable K. In each loop, the K-Means algorithm is initialized with k number of clusters. Then the data X is trained using the fit(X) function. In inertia.append(kmeans.inertia_), the inertia value is stored, which will represent the total distance between the data and its cluster center for each k. The graph of the inertia values will be used to determine the elbow point, where the decline starts to slow down indicating the optimal number of clusters (Rykov et al. 2024).

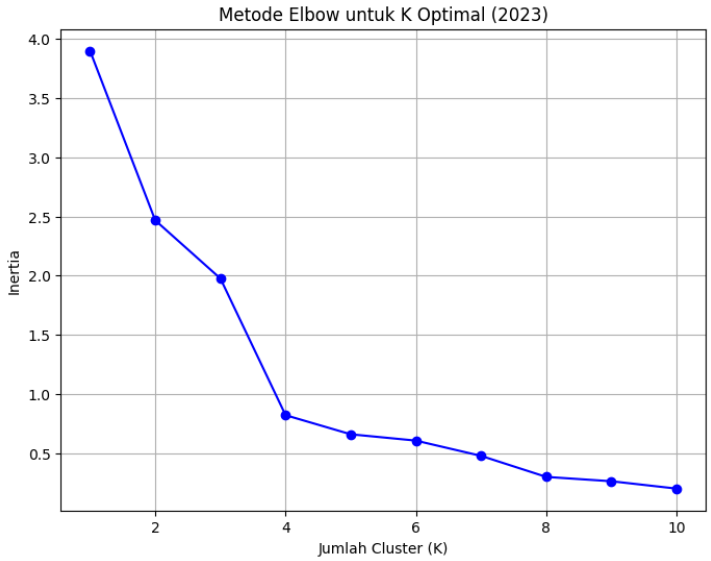

Fig 3. Visualization of Elbow Method Results

In Figure 3, the optimal cluster result of 4 is shown. At the point with avalue of 4, there is the end of the significant decline for the Elbow method. After that point, the decline that occurs is slowing down. So it can be seen that the optimal cluster of this research is 4.

3.3 Implementation of K-Means Algorithm

```
kmeans_optimal = KMeans(n_clusters=4, random_state=42)
data_cleaned['Cluster'] = kmeans_optimal.fit_predict(X)
```

Fig 4. K-Means Algorithm Implementation Command

The command kmeans_optimal = Kmeans(n_cluster=4, random_state=42) is used to initialize the K-Means algorithm with the number of clusters 4. The random_state=42 parameter is used to ensure consistent results every time the algorithm is rum, by setting the same initialization starting point for the centroid. Then, the fit_predict(x) command performs two operations at once, namely fit which groups the X data in 4 clusters and predict which returns the cluster label of each data. The results of the K-Means algorithm are presented in the following scatter-plot visualization:

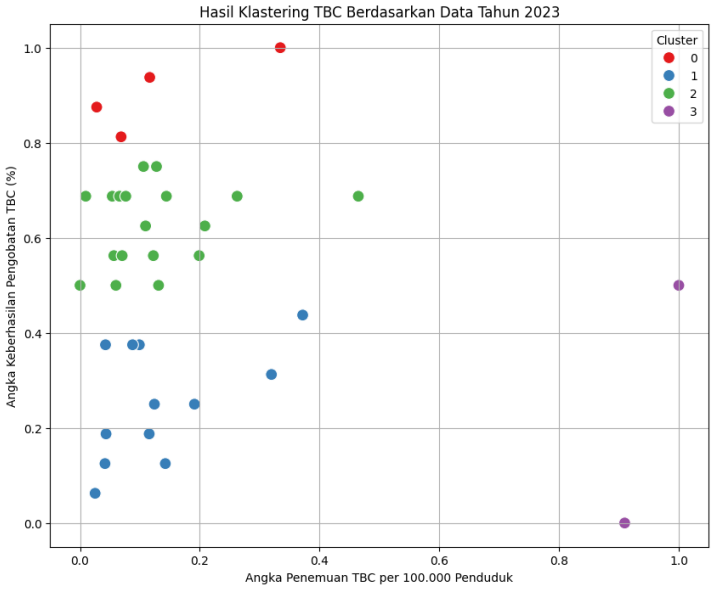

Fig 5. Visualization of K-Means Algorithm Results

4. CONCLUSION

Based on the research conducted, the Elbow method is effective in determining the optimal number of clusters. The analysis showed that areas with high TB case finding rates and low treatment success were located in cluster 3, which included Magelang Town and Tegal Town. The use of K-Means is proven to help map vulnerable areas, which can be used to support more effective health intervention strategies in various regions, especially Central java Province. In further development, the research can be expanded by examining other cases of infectious disease spread or predicting the number of infectious disease case findings in arious provinces in Indonesia.

REFERENCES

Adhanty, Shania, and Syahrizal Syarif. 2023. "Kepatuhan Pengobatan Pada Pasien Tuberkulosis Dan Faktor-Faktor Yang Mempengaruhinya: Tinjauan Sistematis." *Jurnal Epidemiologi Kesehatan Indonesia* 7 (1): 7. https://doi.org/10.7454/epidkes.v7i1.6571.

Ahmed, Mohiuddin, Raihan Seraj, and Syed Mohammed Shamsul Islam. 2020. "The K-Means Algorithm: A Comprehensive Survey and Performance Evaluation." *Electronics* 9 (8): 1295.

AKILLI, Asli, and Hulya ATIL. 2020. "Evaluation of Normalization Techniques on Neural Networks for the Prediction of 305-Day Milk Yield." *Turkish Journal of Agricultural Engineering Research* 1:354–67. https://doi.org/10.46592/turkager.2020.v01i02.011.

Asmiatun, Siti. 2019. "Penerapan Metode K-Medoids Untuk Pengelompokkan Kondisi Jalan Di Kota Semarang." *JATISI (Jurnal Teknik Informatika Dan Sistem Informasi)* 6 (2): 171–80. https://doi.org/10.35957/jatisi.v6i2.193.

Cui, Mengyao. 2020. "On the Elbow Method," 5–8. https://doi.org/10.23977/accaf.2020.010102.

Desmiany Duri, Iin, Roni Afriansya, and Mochamad Rizal Maulana. 2023. "Pendampingan Edukasi Penyakit Tuberkulosis, Penggunaan Obat TB, Hand Hygiene Dan Etika Batuk Di Kelurahan Bangetayu Wetan." *Abdi Reksa* 4 (2): 56–61. https://doi.org/10.33369/abdireksa.v4.i2.56-61.

Kusnaidi, Muhammad Rafli, Timotius Gulo, and Soeb Aripin. 2022. "Penerapan Normalisasi Data Dalam Mengelompokkan Data Mahasiswa Dengan Menggunakan Metode K-Means Untuk Menentukan Prioritas Bantuan Uang Kuliah Tunggal." *Journal of Computer System and Informatics (JoSYC)* 3 (4): 330–38. https://doi.org/10.47065/josyc.v3i4.2112.

Maulindar, Joni. 2023. "Pengembangan Klastering Untuk Penanganan Ibu Hamil Menggunakan K-Means." *Prosiding Seminar Nasional Teknologi Informasi Dan Bisnis (SENATIB)*, 703–8.

Nur Amalia, Indira, Yuyun Umaidah, and Rini Mayasari. 2024. "Penerapan Data Mining Untuk Klasterisasi Daerah Rawan Penyakit Menular Di Kabupaten Karawang Dengan Menggunakan Algoritma K-Means." *JATI (Jurnal Mahasiswa Teknik Informatika)* 8 (4): 5582–91. https://doi.org/10.36040/jati.v8i4.9953.

Pardosi, Lenny Christina, Donal Nababan, Nettietalia Br Brahmana, Mindo Tua Siagian, and Rosetty Sipayung. 2024. "Faktor Yang Berhubungan Dengan Keberhasilan Terapi Penderita TB Paru Di Puskesmas Siatas Barita." *PREPOTIF:Jurnal Kesehatan Masyarakat* 8 (2): 3643–52.

Putu, Ni, Esti Utami, I Made Joel, Jaya Dilaga, and Rika Lusiana Simbolon. 2024. "Pengelompokan Toko Pupuk Termurah E-Commerce Shopee Dengan Metode Klasterisasi" 2024 (Senada): 319–31.

Rykov, Andrei, Renato Cordeiro De Amorim, Vladimir Makarenkov, and Boris Mirkin. 2024. "Inertia-Based Indices to Determine the Number of Clusters in K-Means: An Experimental Evaluation." *IEEE Access* 12 (December 2023): 11761–73. https://doi.org/10.1109/ACCESS.2024.3350791.

Safira, Rifa, and Castaka Agus Sugianto. 2024. "Optimalisasi Algoritma K-Means Menggunakan Metode Elbow Dalam Pengelompokan Data Stunting." *Brahmana : Jurnal Penerapan Kecerdasan Buatan* 5 (2): 257–64.

Sari, Wahyuni Eka, Muslimin Muslimin, Annafi Franz, and Putu Sugiartawan. 2022. "Deteksi Tingkat Kematangan Tandan Buah Segar Kelapa Sawit Dengan Algoritme K-Means." *SINTECH (Science and Information Technology) Journal* 5 (2): 154–64. https://doi.org/10.31598/sintechjournal.v5i2.1146.

Stepanus Ginting, Roni, Hamdani Hamdani, Anindita Septiariani, and Faza Alameka. 2022. "The Clustering Tindak Kekerasan Dalam Rumah Tangga Di Kota Samarinda Menggunakan Algoritma K-Means." *Metik Jurnal* 6 (2): 172–77. https://doi.org/10.47002/metik.v6i2.378.

Waldia, Akshita, Pragati Garg, Priyanka Garg, Rachna Tewani, A. Kumar Dubey, and Anurag Agrawal. 2021. "Crop Recommendation Using Machine Learning." *Fusion: Practice and Applications* 6 (2): 57–63. https://doi.org/10.54216/FPA.060203.