Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th February 2025. Vol. 4. No. 2; e-ISSN: 2808-4519

Improving Regional Clustering Based on Tuberculosis Cases using the K-Means Algorithm of the Cirebon City Health Office

Wilda Rusmiati Rahayu[1*], Ade Irma Purnamasari[2], Agus Bahtiar[3], Kaslani[4]

[1,2]*Informatics Engineering,* STMIK IKMI Cirebon, Indonesia
[3]*Information System,* STMIK IKMI Cirebon, Indonesia
[4]*Computerized Accounting,* STMIK IKMI Cirebon, Indonesia
Wildarusmiati573@gmail.com[1], irma2974@yahool.com[2], agusbahtir038@gmail.com[3] , kaslani@ikmi.ac.id[4]*

Abstract

Tuberculosis (TB) is a highly infectious disease prevalent in Indonesia, including Cirebon City. This study utilizes the K-Means algorithm to optimize the clustering of areas based on TB case data from the Cirebon Health Office. By analyzing the number of cases, population density, and other factors, the study aims to identify regional clusters with similar TB case characteristics. The research employed Rapid Miner software and the Knowledge Discovery Database (KDD) methodology. The K-Means analysis categorized the study area into two clusters. **Cluster_0**, representing 20 areas, had lower TB risk, characterized by higher population density, smaller geographic size, and fewer TB cases. **Cluster_1**, representing two areas, exhibited higher TB risk, marked by lower population density, larger area, and more TB cases. The clustering quality was evaluated using the Davies-Bouldin Index (DBI), which yielded an optimal value of 0.189 at K=2K = 2. Additionally, the Avg within Centroid Performance Vector Analysis supported the clustering validity the clusters with value of 19851032.925.The results demonstrate that this clustering approach effectively identifies TB risk areas, aiding targeted interventions. The findings provide the Cirebon Health Office with a framework for better resource allocation, focusing intensive programs in high-risk regions and preventive measures in low-risk areas.

Keywords: *Tuberculosis, K-means, Clustering, Health Office, Davies-Bouldin Index.*

1. Introduction

One of the infectious diseases that still poses public health problems in Indonesia and around the world is tuberculosis (TB). TB requires urgent attention due to its rapid transmission and high number of cases in different locations, including within the jurisdiction of the Cirebon City Health Office. TB patients are still distributed unequally in some areas, but various efforts are being made to prevent TB. In such a situation, it is important to analyze the spread pattern of TB to determine the appropriate intervention methods. Clustering, a technique for analyzing spread patterns, can be used to group areas based on the frequency of TB. K-Means can cluster data quickly and accurately, making it one of the most effective clustering algorithms.

However, the use of the K-Means algorithm for TB community participation still faces several challenges. First, the data used is often heterogeneous, so the clustering process needs to be adjusted. Second, selecting the appropriate number of clusters or groups is often difficult, as it requires considering the characteristics of the area and the number of cases. Third, the K-Means algorithm cannot handle outliers due to data values that are significantly different from others, which may affect the clustering results. In addition, the lack of a suitable predictive model is also an obstacle to optimizing the clustering results. For the Cirebon City Health Office, the distribution of TB cases varies from region to region, so a thorough analysis is needed to determine which regions need additional measures. This study is necessary to develop a clustering model that can improve the accuracy of region clustering based on TB cases. The ultimate goal of this study is to better classify regions so that TB control programs can be implemented better and more effectively. It is hoped that this study will help the Cirebon City Health Office utilize data to reduce the number of TB cases in the region.

To date, there have been many studies on the K-Means algorithm for analyzing the spread of diseases. One particular study [1]**,**addressed the use of K-Means to analyze and cluster areas in Karawang district based on the distribution patterns of TB cases. Another study [2]used a combination of K-Means clustering and machine learning from linear regression to assess the risk level of pulmonary tuberculosis (TB). The results of this study were classified into groups based on low, medium, and high risk of developing pulmonary TB. Recent studies [3]

focus on using data mining techniques, particularly K-Means algorithm clustering, to analyze health data and find patterns in the spread of diarrheal diseases.

This study uses a quantitative approach with clustering techniques using the K-means algorithm. The study used secondary data from the Cirebon City Health Office, which includes the number of tuberculosis cases per district over a period of time. Clustering was used to group the districts based on their tuberculosis case rates to identify high, medium, and low risk areas. The K-Means algorithm was chosen in part for its ability to quickly and efficiently cluster complex data. Additionally, evaluation techniques such as the elbow method were used to analyze the cluster results and determine the ideal number of clusters.

The results of this study are expected to provide a clearer picture of the prevalence of TB in the Cirebon City Health Office's area of operation. By dividing areas according to the level of TB prevalence, the Health Office can develop more targeted prevention strategies. The study can also serve as a basis for more effective resource allocation; an example of this would be to place education programs and health workers in areas where there is a greater need. This study will also serve as a reference for the introduction of clustering methods to analyze the spread of infectious diseases in other medical settings. Furthermore, the results of this study are expected to improve the performance of the Cirebon City Health Office's tuberculosis control program.

2. Literature Review

2.1. Data Mining

In this phase, patterns are found within the selected data using specific methods or algorithms based on the Knowledge Discovery in Database (KDD) process. The technique used in this final project is K-Means clustering, a data analysis technique that groups data based on the similarity of features. In this final project, the K-Means clustering algorithm is implemented using Rapid Miner as a data management tool [4]K-Means clustering has been widely used to analyze disease distribution patterns in various studies. [5] compared K-means and fuzzy C-means in medical data analysis and showed that K-means is more efficient for homogenous data such as hospital patient data. Furthermore, [6] used K-means to cluster areas based on the prevalence of HIV cases, which helped to develop more targeted public health policies. In another study, [7] used K-means to analyze the prevalence of ARI diseases in Karawang Regency and demonstrated the potential of K-means in identifying high-risk areas.

2.2. Clustering

One of the known techniques in data mining is clustering. The definition of clustering in data mining is grouping a number of data or objects into clusters (groups) such that all data in a cluster contain data that is as similar and different as possible from the objects in other clusters. The most commonly used clustering method is the K-means clustering method. The main drawback of this method is that the results are sensitive to the selection of the initial cluster centers and the calculation of local solutions to achieve optimal conditions. Cluster analysis is a multivariate method that aims to group objects based on their properties. Cluster analysis classifies objects such that all objects that are most similar to other objects fall within the same cluster. K-means has been used in various studies to analyze disease distribution patterns. [8] applied K-means to cluster disease data and found two major clusters that separate patient groups by diagnosis and gender. Furthermore, [9]used K-means to analyze the HIV epidemic in Karawang and grouped areas into several clusters based on their prevalence and susceptibility to the disease. Both studies show how effective K-means is at recognizing patterns.

2.3. K-Means

K-Means is a clustering algorithm that uses a partitioning method. The algorithm divides each data item into clusters based on its proximity to the cluster center or centroid. Below are the steps of the K-Means algorithm used in this study:

1. Determine the number of clusters (k) in the dataset.
2. Determine the centroids. ** The determination of the centroid value is done randomly at the initial stage, but at the iterative stage the following formula is used:

$$Vk=Nk\sum i=1NkXi$$

Where:
Vk = centroid of the k-th cluster
Xi = i-th data
Nk = number of objects that are members of the k-th cluster

3. Calculate the shortest distance to the centroid for each dataset. The centroid distance used is the Euclidean distance with the formula:

$$D_E = \sqrt{(x-s)^2 + (y-t)^2}$$

Where:
DE= Euclidean distance
i= number of objects
(x,y) = object coordinates
(s,t) = centroid coordinates

4. Group objects based on distance to nearest centroid.
5. Repeat step 2 until the centroid reaches the optimal value.

K-means has been widely used to analyze data distribution in various studies. For example, Kurniawan et al. (2023) used K-means to cluster the pharmaceutical data of Puskesmas, and the clustering results helped optimize pharmaceutical inventory management. Furthermore,

3. Research Methods

This studies technique makes use of K-Means Clustering to cluster regions primarily based totally on Tuberculosis case information within side the operating region of the Cirebon City Health Office. This quantitative technique makes use of the K-Means clustering algorithm, information collection, information pre-processing, utility of the K-Means algorithm, assessment of clustering outcomes and evaluation of outcomes represent the study's methodology.

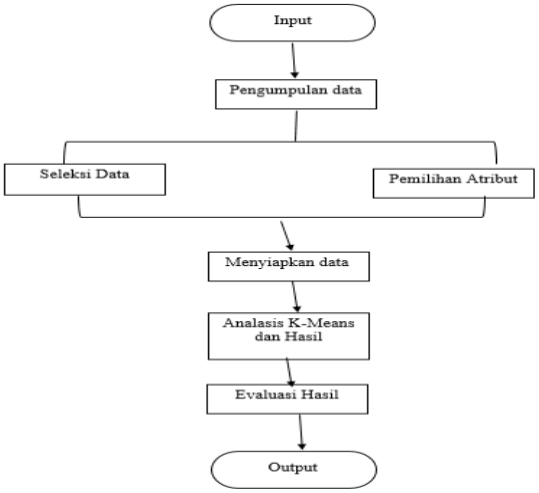

Fig. 1: Research Methods

Table 1: Activity Description Research Methods

Stages	Activity	Activity Description
Selection	Data Collection	The data used in this research is data on Tuberculosis cases such as the name of the village, population (people), area (km2) and the number of Pulmonary TB patients.
Pre-Procession	Data Selection	Selecting data is used to analyze attributes in data that are suitable and relevant for data mining methods so that the data will be selected first.
	Attribute Selection	Attribute Selection In the data mining thesis, it is necessary to select attributes to find attributes that appear at one time.
Data Transformation	Preparing Data for the analysis process	After the data is processed, the next step is to change the data format according to the requirements of the K-Means algorithm. This includes converting category values into numbers and selecting the appropriate four for clustering.
Data Mining	K-Means Clustering Method.	In this research dataset, the data collection is basic calculation, random and rotating so it has the risk of inaccurate data collection. The K-Means method can cluster areas based on high TB case rates and is aided by Rapid Miner Application.
Evaluation	Evaluation of results.	Data mining is presented in a form that must be understood. Davies Bouldin Index (DBI) method to test the quality of the cluster by measuring the extent to which the cluster is relevant and accurate.
Knowledge	Results	Data Mining process decisions and actions in research from Rapid miner results and DBI value results.

3.1. Data Source

In this study using secondary data whose research data sources were identified directly from secondary data obtained from the Cirebon City Health Office in this study. This data includes tuberculosis cases from January 01 to November 01, 2024. The data includes various important elements relevant for epidemiological analysis and mapping of disease patterns, such as the geographical distribution of tuberculosis cases across sub-districts of Cirebon City, as well as the demographic characteristics of patients, including age, gender, and

socio-economic status. This demographic information makes it possible to identify at-risk groups and patterns of disease spread that may differ among different segments of the population, so these data can be used to make inferences about how tuberculosis.

3.2. Population

Population is a generational field consisting of subjects or objects with certain numbers and characteristics that have been grouped determined by researchers to study and draw conclusions, this study comes from Tuberculosis (TB) cases in the Cirebon City Health Office work area. The population includes all TB patients recorded in this area during the time period specified in the study, namely Tuberculosis cases in 2024 which were recapitulated from January 01 to November 01, 2024.

3.3. Data Collation Techniques

The data used in this study were obtained from reports of Tuberculosis cases recorded at the Cirebon City Health Office. Data collection can be done through medical record reports. The records include variables such as the name of the village, the number of TB cases per area, the area, and the population density included in the record. In this study, the data collection method used is observation, which means a data collection method that involves direct observation of the object of research in the field. To obtain data for clustering analysis, the author made direct observations at the Cirebon City Health Office. After the data is collected, the next stage is pre-processing or data cleaning. This is done to ensure that the information to be fed into the algorithm has no errors or inconsistencies that could affect the results. This step involves reviewing the data to identify and process data containing outliers or extreme values, which may need to be removed or sorted to avoid distorting the clustering results. Before being used in the analysis, data such as the number of Tuberculosis cases and population density were normalized to ensure that each variable had the same scale. The normalization process is important because scale differences between variables can affect the performance of the K-Means algorithm, which is sensitive to scale differences.

3.4. Data Analysis Techniques

The data analysis technique uses the K-Means method to group data into clusters that have the same characteristics into one clusters, K-Means is a process to achieve an accurate result in research, and can ensure that the data has been grouped accurately. By using the K-Means Algorithm to classify the results of the spread of TB disease in the Cirebon City Health Office Region, the data is processed using the KDD (knowledge Discovery Database) stage assisted by the Rapid miner application. The following is an explanation of the KDD process in outline as shown in Figure 2 below.

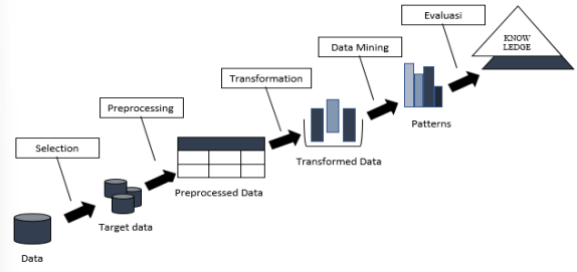

Figure 2 Process

3.5.1. Data Selection

This stage is the preparatory stage of the data selection process. After the attribute selection procedure is completed, the data from the observations will be used for preliminary work [3].

3.5.2. Preprocessing/Cleaning Data

This stage includes a number of tasks, including removing duplicates and correcting data inaccuracies, correcting data errors, and so on. This stage will produce quality that is good enough and clean enough to enable the next stage of transformation [10].

3.5.3. Transformation

Transformation is a set of instructions to convert inputs into usable outputs, following an input-process-output pattern. The processes involved include data processing to achieve the desired shape, customized according to the applied Clustering algorithm [4].

3.5.4. Data Mining

This stage is the process of finding patterns in the selected data using certain methods or algorithms based on the overall KDD process, the method used in this final project is K-Means Clustering. K-Means Clustering is a data analysis method to group data based on the similarity of its characteristics. In this final project, the K-Means Clustering algorithm will be implemented into rapid miner as a data processing method [1]**.**

3.5.5. Data Interpretation (Evaluation)

It requires presenting the information patterns generated from data mining in a format that can be easily understood by interested parties. Evaluation is a crucial step in the KDD process that involves examining how the patterns or information found fit with pre-existing facts or hypotheses. [2].

4. Results and Discussion

4.1. Research Result

4.1.1 Data Selection

The data obtained in the form of the number of TB patients in the Cirebon City Health Office area and selection is carried out to obtain attributes, for this research selection, namely the attributes of Kelurahan Name, Population, Area, and Number of TB Patients. After selecting attributes, Village Name, Population, Area, and Number of TB Patients, the data is grouped based on TB Patient Data which is carried out data selection there are 4,282 total Tuberculosis patient data from 22 villages so that the dataset can be like table 2.

Table 2: Data Selection

No	Name Kelurahan	Population	Area	Number of TB Patients
1	Kejaksan	10.188	0,663	479
2	Sukapura	17.400	1,571	59
3	Kesenden	14.360	1,456	48
4	Kebonbaru	9.565	0,742	17
5	Pegambiran	24.509	4,447	2
6	Lemahwungkuk	8.903	0,649	16
7	Kesepuhan	17.294	0,767	170
8	Panjunan	10.640	1,322	127
9	Kalijaga	38.397	4,225	46
10	Hajarmukti	22.943	2,344	176
11	Kecapi	24.448	2,294	281
12	Larangan	16.813	1,903	133
13	Argasunya	27.066	6,835	66
14	Jagasatru	10.768	0,353	30
15	Pekalipan	6.571	0,426	18
16	Pulasaren	8.001	0,313	19
17	Pekalangan	6.136	0,493	34
18	Pekiringan	13.175	1,263	151
19	Sunyaragi	13.105	2,265	188
20	Kesambi	9.296	1,006	1.203
21	Drajat	15.989	0,934	654
		Total		**4.282**

4.1.2. Preprocessing/ Cleaning

With the potential for incomplete and irrelevant information, data pre-processing is essential. To effectively cluster Tuberculosis cases, it is necessary to identify and select the variables that are most significant in clustering the data. The operator used in this context is attribute selection, which aims to select relevant attributes, can be seen data processing results in Figure 3 below.

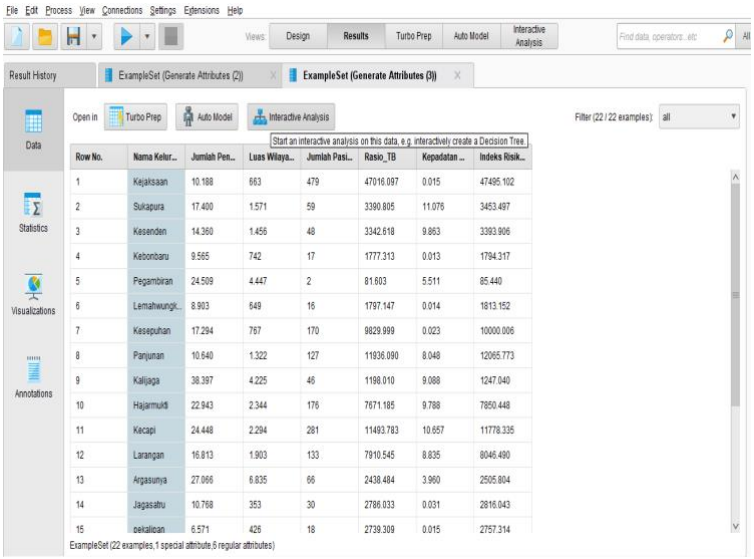

Fig. 3: Data processing results

4.1.3. Transfomation

This process is done to ensure that each attribute has equal influence in the clustering analysis and that attributes with larger values do not dominate the clustering results. The Normalize operator is used after pre-processing is complete, converting attribute values to a range of 0 to 1 or Z-Transformation with methods such as Min-Max. Clustering algorithms such as K-Means can perform better and find more representative cluster patterns with normalized data. Normalization is performed on the relevant columns of the TB Risk Index attribute so that all values are in the same range. Before running the process, make sure to select the filter type parameter as "Single". After the Normalize operator is applied, the TB Risk Index attribute values are in the same range. The results can be found in figure 4 below.

File Edit Process View Connections Settings Extensions Help

Views: Design | Results | Turbo Prep | Auto Model | Interactive Analysis

Find data, operators...etc

Result History | Z-Transformation (Normalize) | ExampleSet (Generate Attributes (3)) | ExampleSet (Normalize)

Data | Statistics | Visualizations | Annotations

Open in Turbo Prep | Auto Model | Interactive Analysis

Filter (22 / 22 examples): all

Row No.	Nama Kelur...	Indeks Risik...	Jumlah Pen...	Luas Wilaya...	Jumlah Pasi...	Rasio_TB	Kepadatan ...
1	Kejaksaan	3.104	10.188	663	479	47016.097	0.015
2	Sukapura	-0.480	17.400	1.571	59	3390.805	11.076
3	Kesenden	-0.485	14.360	1.456	48	3342.618	9.863
4	Kebonbaru	-0.615	9.565	742	17	1777.313	0.013
5	Pegambiran	-0.754	24.509	4.447	2	81.603	5.511
6	Lemahwungk...	-0.614	8.903	649	16	1797.147	0.014
7	Kesepuhan	0.052	17.294	767	170	9829.999	0.023
8	Panjunan	0.221	10.640	1.322	127	11936.090	8.048
9	Kalijaga	-0.660	38.397	4.225	46	1198.010	9.088
10	Hajarmukti	-0.122	22.943	2.344	176	7671.185	9.788
11	Kecapi	0.197	24.448	2.294	281	11493.783	10.657
12	Larangan	-0.107	16.813	1.903	133	7910.545	8.835
13	Argasunya	-0.557	27.066	6.835	66	2438.484	3.960
14	Jagasatru	-0.532	10.768	353	30	2786.033	0.031
15	pekalipan	-0.537	6.571	426	18	2739.309	0.015

ExampleSet (22 examples,1 special attribute,6 regular attributes)

Fig. 4: Transformation Result

4.1.4. Data Mining

In the data mining process, the first step is to use the K-Means algorithm to group regions based on the spread of Tuberculosis cases using Rapidminer. Algortima is implemented in the process, with the determination of parameters that can be seen in the figure, namely K = 2 which is done a maximum of 10 times the clustering. An image of the K-Means clustering process can be seen in Figure 5 below.

The process of entering the K-Means Clustering Operator Next is the stage of adding the Performance Operator involves the process of finding the distance value and DBI value to effectively manage the performance of a communication system or device. The image of the Performance stage can be seen in Figure 6 below.

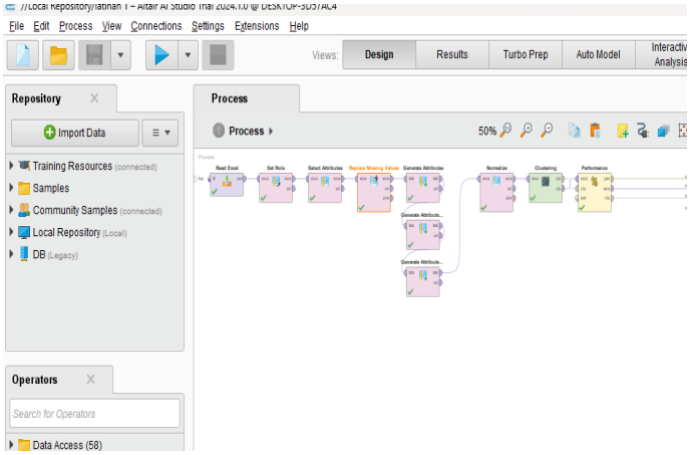

Fig. 5: Clustering K_Means

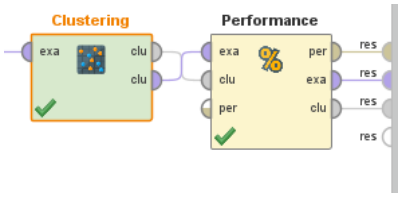

Fig. 6: Performance

After going through a series of careful and thorough performance stages, such as data analysis, performance measurement, and evaluation of various relevant parameters, the average result (Avg) is finally found which reflects the level of efficiency and effectiveness of a system or process. The results of the Avg search can be seen in Figure 7 below.

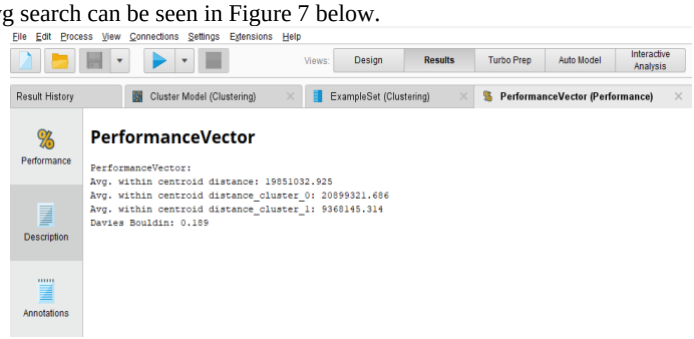

Fig. 7: Avg.Within Centroid Results

4.1.2 Interpretation / Evaluation

In the process of clustering data, model evaluation is performed to determine the best cluster. In this method, the index value in Davies-Bouldin (DBI) is evaluated with the lowest DBI value to show better cluster quality. In the applied K-Means algorithm model, in this modeling, varying k values are tested from K=2 to K=10, and each value is evaluated to determine which DBI value is the smallest and which cluster is formed. Below is Table 3 Experimental results of K=2 to K=10 using DBI evaluation.

Table 3: Evaluation DBI Results

K	DBI Value Result	Number Of Cluster Members	Avg.within centroid distance_cluster	Value Avg.within centroid distance
2	0.189	Cluster 0: 20 items Cluster 1: 2 items	Cluster 0: 20899321.686 Cluster 1: 9368145.314	19851032.925
3	0.281	Cluster 0: 12 items Cluster 1: 2 items Cluster 2: 8 items	Cluster 0: 2123842.140 Cluster 1: 9368145.314 Cluster 2: 4670349.250	3708417.741
4	0.360	Cluster 0: 11 items Cluster 1: 2 items	Cluster 0: 1269074.568 Cluster 1: 9368145.314	2187731.349

		Cluster 2: 5 items Cluster 3: 4 items	Cluster 2: 1150369.725 Cluster 3: 2420532.711	
5	0.272	Cluster 0: 11 items Cluster 1: 1 items Cluster 2: 5 items Cluster 3: 1 items Cluster 4: 4 items	Cluster 0: 1269074.508 Cluster 1: 0.000 Cluster 2: 1150369.724 Cluster 3: 0.000 Cluster 4: 2420532.711	1336081.775
6	0.252	Cluster 0: 11 items Cluster 1: 1 items Cluster 2: 1 items Cluster 3: 3 items Cluster 4: 1 items Cluster 5: 5 items	Cluster 0: 1269074.508 Cluster 1: 0.000 Cluster 2: 0.000 Cluster 3: 1191494.196 Cluster 4: 0.000 Cluster 5: 911023.347	1004064.496
7	0.296	Cluster 0: 1 items Cluster 1: 1 items Cluster 2: 8 items Cluster 3: 3 items Cluster 4: 1 items Cluster 5: 5 items Cluster 6: 3 items	Cluster 0: 0.000 Cluster 1: 0.000 Cluster 2: 403582.448 Cluster 3: 1191494.196 Cluster 4: 0.000 Cluster 5: 911023.347 Cluster 6: 266202.279	552584.806
8	0.446	Cluster 0: 8 items Cluster 1: 1 items Cluster 2: 1 items Cluster 3: 1 items Cluster 4: 4 items Cluster 5: 2 items Cluster 6: 2 items Cluster 7: 3 items	Cluster 0: 403582.448 Cluster 1: 0.000 Cluster 2: 0.000 Cluster 3: 0.000 Cluster 4: 199075.207 Cluster 5: 119969.906 Cluster 6: 1067780.453 Cluster 7: 266202.279	425384.907
9	0.205	Cluster 0: 4 items Cluster 1: 1 items Cluster 2: 1 items Cluster 3: 1 items Cluster 4: 3 items Cluster 5: 8 items Cluster 6: 2 items Cluster 7: 1 items Cluster 8: 1 items	Cluster 0: 199075.207 Cluster 1: 0.000 Cluster 2: 0.000 Cluster 3: 0.000 Cluster 4: 266202.279 Cluster 5: 403582.448 Cluster 6: 14795.258 Cluster 7: 0.000 Cluster 8: 0.000	220598.080
10	0.239	Cluster 0: 2 items Cluster 1: 1 items Cluster 2: 1 items Cluster 3: 2 items Cluster 4: 6 items Cluster 5: 1 items Cluster 6: 4 items Cluster 7: 3 items Cluster 8: 1 items Cluster 9: 1 items	Cluster 0: 635.851 Cluster 1: 0.000 Cluster 2: 0.000 Cluster 3: 14795.258 Cluster 4: 191817.580 Cluster 5: 0.000 Cluster 6: 199075.207 Cluster 7: 185250.459 Cluster 8: 0.000 Cluster 9: 0.000	115173.633

4.1.3 Knowledge

The results of research using the K-Means clustering algorithm using Rapid miner tools and the results of evaluating the Davies-Bouildin Index and using the elbow method can be seen as follows.

a. Davies-Bouldin Index (DBI) value

Clustering using the K-Means clustering algorithm with Rapid miner tools and DBI evaluation obtained the optimal and best cluster value at K-2 with the smallest DBI value of 0.189 with cluster_0 members: 20 items and cluster_1: 2 items. This shows that clustering with two clusters produces optimal clusters and has the best level of separation. Can be seen in Figure 8 below.

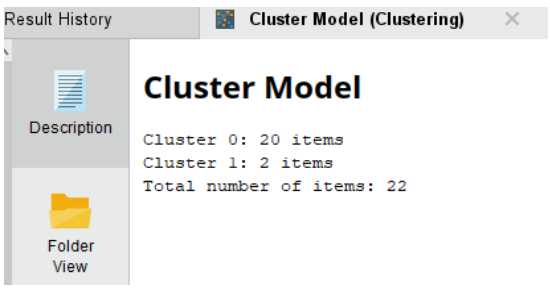

Fig. 8: Model Cluster Results

Attribute	cluster_0	cluster_1
Indeks Risiko TB	-0.286	2.862
Jumlah Penduduk (jiwa)	16.425	13.088
Luas Wilayah (km2)	188.856	798.500
Jumlah Pasien Tuberkulosis (jiwa)	97.360	566.500
Rasio_TB	5739.133	43959.609
Kepadatan Penduduk	5.578	0.016

Fig. 9: Centroid Value

b. Elbow Method

The Elbow method is used to see changes in the average distance within the cluster or called the Avg.within centroid distance value which is used to measure when the number of clusters increases. The elbow graph is obtained by inserting the average within cluster of each K into the graph. The graph starts by looking at the point where the elbow occurs at K=2 which indicates that cluster 2 is the optimal number of clusters. With the evaluation results of the Elbow method which has similarities in the results with the DBI value, it strengthens that cluster 2 is the best choice in this study. Can be seen in Figure 9 below.

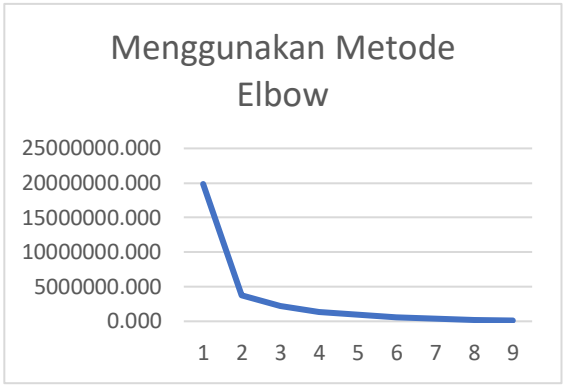

Fig. 10: Method Elbow

4.2 Discussion

a) Apply the use of K-means algorithm to divide the area around the Health Office based on the incidence rate of TB cases.

The application of the K-Means algorithm in clustering the Cirebon City Health Office Region based on Tuberculosis Cases using Rapidminer by evaluating the DBI value, the smaller the DBI value, the more optimal the cluster formed (Ula et al., 2023). The following is a table of DBI results below.

Table 4: DBI Results

K	DBI
2	0.189
3	0.281
4	0.360
5	0.272
6	0.252
7	0.296
8	0.446
9	0.205
10	0.239

From the table of DBI value results, it can be seen that the optimal cluster is at K = 2 with the smallest DBI value of 0.189, the lower the DBI value, the better the cluster formed. With K = 2 the data is divided into 2 clusters, namely cluster_0 and cluster_1. Members of cluster_0 consist of 20 areas that have a lower risk of TB, high population density, small area and fewer TB patients. While cluster_1 members consist of 2 regions that have a higher risk of TB, lower population density, large area, and more TB patients.

b) Evaluation results of regional clustering based on Tuberculosis cases in Cirebon City.

Based on demographic characteristics and the number of Tuberculosis cases, the study area was grouped into two clusters based on the results of the K-Means method analysis. The results of this clustering provide a clearer picture of the distribution of TB risk in each region, so that it can be a reference for understanding areas with different levels of risk. This can be seen in Figure 4.28 below:

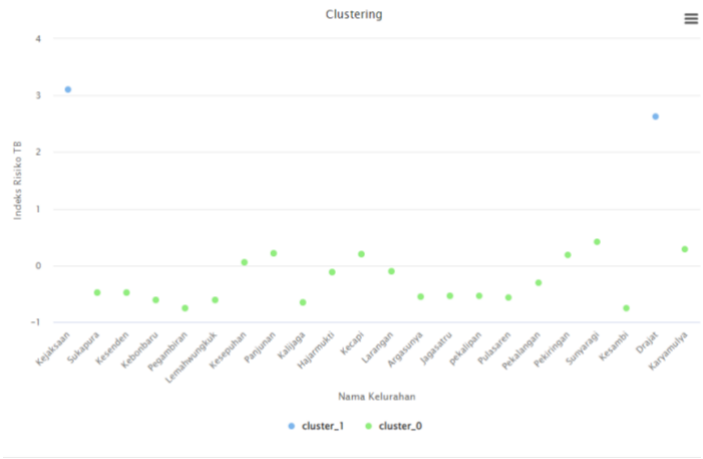

Fig. 11: Graphic Cluster

From the image of the cluster graph results, the following explanation for each cluster: Cluster_0 in green consists of 20 areas, namely Sukapura, Kesenden, Kebonbaru, Pegambiran, Lemahwungkuk, Kesepuhan, Panjunan, Kalijaga, Hajarmukti, Kecapi, Larangan, Argasunya, Jagasatru, Pekalipan, Pulasaren, Pekalangan, Pekiringan, Sunyaragi, and Kesambi. Cluster_0 has common characteristics such as higher population density, smaller area, and a tendency to have fewer TB cases. These areas are considered to have a relatively lower risk of TB compared to areas in other clusters. Whether directly or indirectly, these factors suggest that the spread of TB cases in these areas may be more contained compared to the other clusters. Cluster_1 in blue consists of two areas, Kejaksan and Drajat. The analysis shows that Cluster_1 areas have a higher risk of TB compared to Cluster_0 areas due to their key characteristics of lower population density, larger area, and relatively higher number of TB cases. Thus, Cluster_1 areas of Kejaksan and Drajat may be more vulnerable to the spread of TB than Cluster_0 areas.

The division of areas into two clusters with a DBI value of 0.189 provides a complete picture of the distribution of TB risk in the study area. Areas in Cluster_1, namely Kejaksan and Drajat, are considered to have a higher potential risk than other areas. Meanwhile, areas in Cluster_0 i.e. Sukapura, Kesenden, Kebonbaru, Pegambiran, Lemahwungkuk, Kesepuhan, Panjunan, Kalijaga, Hajarmukti, Kecapi, Larangan, Argasunya, Jagasatru, Pekalipan, Pulasaren, Pekalangan, Pekiringan, Sunyaragi, and Kesambi tend to have more stable characteristics and lower risk. It is possible to use this analysis as a basis for gaining a better understanding of the TB risk conditions in each of these wards to help determine which areas require special handling**.**

5. Conclusion

Based on the results of this research, the discussion that has been carried out, it can be concluded as follows:

1. By applying the K-Means algorithm to group areas based on Tuberculosis cases in the Cirebon City Health Office, it can be seen from the results of the Davies Bouldin Index (DBI) value. The results of clustering areas based on Tuberculosis cases in the Cirebon City Health Office Region with k = 2, obtained a Davies Bouldin Index (DBI) of 0.189.
2. From the results of the evaluation of clustering areas based on Tuberculosis cases having clusters of demographic characteristics and the number of TB cases, the Cluster_0 research area consists of 20 areas namely Sukapura, Kesenden, Kebonbaru, Pegambiran, Lemahwungkuk, Kesepuhan, Panjunan, Kalijaga, Hajarmukti, Kecapi, Larangan, Argasunya, Jagasatru, Pekalipan, Pulasaren, Pekalangan, Pekiringan, Sunyaragi, and Kesambi, including higher population density, smaller area, and fewer TB cases. In contrast, Cluster_1 consists of two areas, Kejaksan and Drajat, and has a higher risk of TB, as indicated by lower population density, larger area, and more TB cases. This division provides a picture of the distribution of TB risk that can be used to understand areas with different levels of risk and make it easier to identify areas that require more specialized TB treatment.

6. Suggestions

Based on the research that has been done, several suggestions can be given for further research:

1. The recommendation from the Cirebon City Health Office is that the regional profile data related to the risk of tuberculosis should be updated and checked regularly. To allocate resources more efficiently, the division of areas into groups can be a reference. This will allow areas with low and high TB risk to receive interventions appropriate to their respective characteristics.
2. The results of the analysis suggest that relevant parties, such as the health office and other health institutions, should use this data to create more efficient health resource management and allocation plans. High-risk areas, such as Kejaksan and Drajat (Cluster_1), may be prioritized for intensive monitoring, regular check-ups, and more in-depth data collection to find out the main causes of the increase in tuberculosis cases in these areas. For now, Cluster_0 areas still need attention to maintain stability and prevent the risk of TB from increasing in the future.

References

[1] K. A. Yatna, N. Rahaningsih, and R. D. Dana, "Penerapan Data Mining untuk Clustering Penyakit Diare Menggunakan algoritma K-Means (Studi Kasus: Puskesmas Beber)," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 3, pp. 3124–3131, 2024, doi: 10.36040/jati.v8i3.9616.

[2] Y. P. Sari, A. Primajaya, and A. S. Y. Irawan, "Implementasi Algoritma K-Means untuk Clustering Penyebaran Tuberkulosis di Kabupaten Karawang," *INOVTEK Polbeng - Seri Inform.*, vol. 5, no. 2, pp. 229–239, 2020, doi: 10.35314/isi.v5i2.1457.

[3] M. Ula, A. Zulfikri, A. F. Ulva, and R. A. Rizal, "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, doi: 10.33022/ijcs.v12i1.3162.

[4] S. Gymnastiar and A. Bahtiar, "PENERAPAN ALGORIMA K-MEANS CLUSTERING UNTUK MENGELOMPOKAN DATA KEJADIAN KEKERINGAN DI KABUPATEN CIREBON," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 2, pp. 2325–2331, 2024, doi: 10.36040/jati.v8i2.8948.

[5] R. A. Farissa, R. Mayasari, and Y. Umaidah, "Perbandingan Algoritma K-Means dan K-Medoids Untuk Pengelompokkan Data Obat dengan Silhouette Coefficient," *J. Appl. Informatics Comput.*, vol. 5, no. 2, pp. 109–116, 2021, doi: 10.30871/jaic.v5i1.3237.

[6] K. Kodratul Munawar and A. Irma Purnamasari, "Implementasi Algoritma K-Means Clustering Pada Klasterisasi Kasus Hiv Di Jawa Barat," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 2, pp. 1092–1099, 2023, doi: 10.36040/jati.v7i2.6372.

[7] I. K. Fauzi, B. A. Dermawan, and T. Padilah, "Penerapan K-Means Clustering pada Penyakit Infeksi Saluran Pernapasan Akut (ISPA) di Kabupaten Karawang," *J. Sist. dan Inform.*, vol. 15, no. 1, pp. 81–87, 2020, doi: 10.30864/jsi.v15i1.350.

[8] R. Anggraini, E. Haerani, J. Jasril, and I. Afrianty, "Pengelompokkan Penyakit Pasien Menggunakan Algoritma K-Means Rahayu," *JURIKOM (Jurnal Ris. Komputer)*, vol. 9, no. 6, pp. 1840–1849, 2022, doi: 10.30865/jurikom.v9i6.5145.

[9] S. D. D. Paramewari, B. Nugraha, and Siska, "ANALISIS CLUSTERING PENYEBARAN HIV DI KARAWANG BERDASARKAN KECAMATAN DENGAN ALGORITMA K-MEANS," *JITET (Jurnal Inform. dan Tek. Elektro Ter.*, vol. 12, no. 3, 2024, doi: 10.23960/jitet.v12i3.4878.

[10] I. J. Putri, F. Riana, and B. Wulandari, "Pengelompokan Kasus Tuberculosis Dengan Algoritma K-Means Berdasarkan Kelurahan di Kota Bogor," *J. Inform.*, vol. 11, no. 1, pp. 42–48, 2024, doi: 10.31294/inf.v11i1.20042.