Application of Machine Learning K-Means Clustering and Linear Regression in Determining the Risk Level of Pulmonary Tuberculosis

Ziaul Islam Bablu
Department of Computer Science and Engineering
BGC Trust University Bangladesh
Chattogram, Bangladesh
ziaulislambablu2@gmail.com

Towhidul Haque Limon
Department of Computer Science and Engineering
BGC Trust University Bangladesh
Chattogram, Bangladesh
towhidulhaque4455@gmail.com

Sowmik Barua
Department of Computer Science and Engineering
BGC Trust University Bangladesh
Chattogram, Bangladesh
sowmikbarua7878@gmail.com

Piyal Dey
Department of Computer Science and Engineering
BGC Trust University Bangladesh
Chattogram, Bangladesh
piyaldey6@gmail.com

Mowmita Tajnin Jiba
Department of Computer Science and Engineering
BGC Trust University Bangladesh
Chattogram, Bangladesh
mowmitatajninj@gmail.com

Abhijit Pathak
Department of Computer Science and Engineering
BGC Trust University Bangladesh
Chattogram, Bangladesh
abhijitpathak@bgctub.ac.bd

ABSTRACT

Pulmonary tuberculosis (TB) remains a pressing public health concern in densely populated regions, particularly in Bireuen, Bangladesh. Despite efforts to combat the disease, Bireuen reported a substantial burden of approximately 755 cases of pulmonary TB in 2019, amidst a population of approximately 400,000. This study leveraged data from Bangabandhu Sheikh Mujib Medical University Hospital and the Health Department across 17 districts to not only identify high-risk areas but also predict disease incidence. Through the application of advanced analytical methodologies such as K-Means clustering and Clusterwise Regression, the analysis delineated two high-risk areas within Cluster 1, six areas within Cluster 2, and nine areas within Cluster 3. The regression analysis demonstrated a coefficient of determination (R-squared) of 0.5740, indicating a moderate predictive capacity. These findings offer critical insights for public health authorities, empowering them to devise targeted interventions and allocate resources effectively to combat the spread of pulmonary tuberculosis. By implementing tailored strategies in identified high-risk areas, such as targeted screening programs and enhanced access to diagnostic and treatment facilities, authorities can mitigate the disease's impact and improve health outcomes in affected communities. Additionally, these findings underscore the importance of continued surveillance and monitoring efforts, alongside collaborative initiatives between government agencies, healthcare providers, researchers, and community stakeholders, to achieve the ultimate goal of tuberculosis elimination in Bangladesh.

General Terms

Epidemiology, K-means clustering, Public health, K-means clustering, Linear Regression, Predictive analytics, Intervention strategies.

Keywords

Pulmonary Tuberculosis, Linear Regression, K-Means Clustering, Predictive modeling, Disease mapping.

1 INTRODUCTION

Pulmonary tuberculosis is a slowly progressive disease caused by Mycobacterium tuberculosis. It is now a major global health challenge causing death Worldwide. According to the World Health Organization, Tuberculosis is one of the top 10 diseases causing death globally [1]. The risk level for TB is increasing globally, but India has more Tuberculosis cases than any other country. According to WHO, the 30 high TB burden countries: Cambodia, the Russian Federation, and Zimbabwe have transitioned out of the list; Gabon, Mongolia, and Uganda have joined the list. The 30 high TB/HIV burden countries: Angola, Chad, Ghana, and Papua New Guinea have transitioned out of the list; Gabon, Guinea, Philippines, and the Russian Federation have joined the list. The 30 high MDR/RR-TB burden countries. Ethiopia, Kenya, and Thailand have transitioned out of the list; Mongolia, Nepal, and Zambia have joined the list. In 2000-2001 the annual risk of tuberculosis infection was 1.5% which decreased to 1.1% in 2009-2010 according to the surveys of the National annual risk of TB infection (ARTI). There is a program named The Revised National Tuberculosis Control Programme (RNTCP) which provides free treatment to TB patients and over 15 million patients treated successfully [2]. Despite this, the risk level of PTB is still a global concern for the people. This study aimed to determine the risk level of pulmonary tuberculosis. As pulmonary tuberculosis is a bigger global health concern, it a need for innovative approaches to determine the risk level accurately using machine learning k-means clustering and linear regression.

In Bangladesh, efforts to combat tuberculosis are underway through programs like the National Tuberculosis Control Programme (NTP). However, despite these efforts, the risk level of PTB remains a concern due to undetected cases and challenges in providing effective treatment and prevention measures. This study aims to determine the risk level of pulmonary tuberculosis in Bangladesh using innovative approaches such as machine learning techniques like k-means clustering and linear regression. By analyzing data sets containing information on population density, geographical distribution, and TB patient numbers, the study seeks to identify clusters of areas that are more susceptible to PTB and to forecast disease trends based on various factors. The utilization of machine learning algorithms like k-means clustering enables the grouping of districts based on disease distribution patterns, allowing for the identification of areas with higher PTB risk. Additionally, the application of clusterwise regression with linear regression algorithms enhances the accuracy of forecasting models by considering the interplay between different variables. With the large number of sufferers who have not been detected, the possibility of spreading pulmonary TB disease is increasing. The data entered is in the form of a data set to be processed in the k-means algorithm model which is grouped based on the objects/districts that are seen regarding the distribution of the disease. Then we can see the level of clusters that are and are not susceptible to pulmonary TB disease in each area. Next, a forecasting model using clusterwise regression with a linear regression algorithm is used to see the forecasting results of variables that have a mutual influence on other variables. The prediction results are more accurate because they are seen from the relationship between variables.

MERS-COV Virus Clustering Analysis Research provides that MERS-CoV DNA sequences can be clustered using the k-means algorithm. The stages carried out in this research are converting data to numeric, then normalizing the data, grouping DNA sequences using the spectral clustering method with the stages of building a similarity graph, managing the Laplacian normalization matrix, calculating k-eigen values, and calculating k-eigen values. The results of the normalized data are then clustered using the k-means partition algorithm. The research results show that the clustering results using the k-means algorithm have three clusters and are more

homogeneous compared to clustering which only uses k-means. Machine Learning algorithms can easily analyze trends and identify viral diseases that spread easily. This research consists of 5 main scopes and processes for predicting viruses and can group each region or determine the type. The hybrid K-means model offers high evaluation metric results in the development of hybrid models such as CNN-Bi-LSTM and its high accuracy and quality performance values exceed 90% for classifying viruses in various studies and almost 100% at its peak.

Pulmonary tuberculosis (TB) remains a significant global health concern, with early detection and intervention crucial for effective management. Traditional methods for assessing TB risk often lack precision and scalability. Therefore, there is a need for innovative approaches leveraging machine learning techniques to accurately determine the risk level of pulmonary tuberculosis. How can machine learning algorithms such as K-Means clustering and linear regression be applied to effectively determine the risk level of pulmonary tuberculosis, enhancing early detection and intervention strategies?

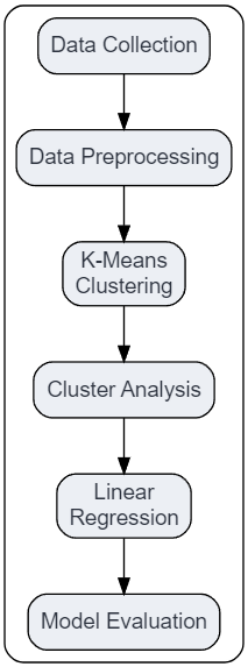

Fig 1. Stages to determine the Risk Level of Pulmonary Tuberculosis

In this study, the authors aim to employ advanced machine learning techniques to enhance the understanding of pulmonary tuberculosis (TB) epidemiology and inform targeted prevention and control strategies. The objectives of this research endeavor are as follows:

- Implement a K-Means clustering algorithm to identify distinct patterns and clusters within datasets containing variables relevant to pulmonary tuberculosis risk factors.
- Utilize the identified clusters to stratify individuals into different risk groups based on their demographic, clinical, and environmental characteristics.
- Apply linear regression modeling to assess the relationship between TB risk factors and the likelihood of developing pulmonary tuberculosis within each identified cluster.
- Evaluate the performance of the developed models in predicting TB risk levels and compare it with traditional risk assessment methods.
- Provide insights into the most influential risk factors contributing to the development of pulmonary tuberculosis, based on the results obtained from the machine learning models.
- Suggest potential interventions or targeted strategies for TB prevention and control based on the identified risk factors and clusters.

2 RELATED WORKS

The reviewed paper provides an in-depth examination of pulmonary tuberculosis, starting with an introductory exploration characterizing the lung as the primary gateway for tuberculosis infection in the majority of cases. The symptoms and signs of both primary and reactivation tuberculosis were described. Laboratory Examination of pulmonary tuberculosis was also described. Chest radiography was also described. In this paper, primary tuberculosis and reactivation tuberculosis were also described. Computed tomography (CT) in pulmonary tuberculosis was described. The occurrence of tuberculosis in the elderly and those patients on anti-

tumor necrosis factor alpha inhibitors was described. Pleural tuberculosis and its diagnosis were described. The pulmonary findings of tuberculosis in HIV infection and complications of pulmonary tuberculosis were described.

The paper undertakes a comprehensive review of childhood pulmonary tuberculosis, offering insights into established concepts and principles. It elucidates how this wealth of knowledge, often regarded as "old wisdom," remains pertinent in addressing contemporary and forthcoming challenges within the realm of childhood tuberculosis. The chemotherapy literature that described the natural history of disease in children identified three central concepts: (1) the need for accurate case definitions, (2) the importance of risk stratification, and (3) the diverse spectrum of disease pathology. These three concepts are linked with the diagnosis of childhood tuberculosis. The diagnosis of childhood tuberculosis was also discussed in this paper. The concepts are also linked with the principles of antituberculosis treatment and treatment of childhood tuberculosis.

The paper critically examines the risk factors associated with extra-pulmonary tuberculosis in comparison to pulmonary tuberculosis. Despite advancements in healthcare, tuberculosis (TB) persists as a significant global health concern, contributing to substantial disability and mortality rates worldwide. One-third of the world's population is estimated to be currently infected with Mycobacterium tuberculosis. The diagnosis of TB was based on: 1) sputum, pleural effusion, pericardial effusion, ascites, urine, cerebral spinal fluid, synovial fluid, and abscess or tissue culture that yielded M. tuberculosis. 2) histological findings of granulomatous inflammation combined with positive acid-fast stain in the pathology specimens and favorable clinical response to anti-tuberculosis chemotherapy. The definition of EPTB was based on the guidelines of the American Thoracic Society and the US Centers for Disease Control and Prevention. EPTB was defined as extra-pulmonary involvement with or without concomitant pulmonary involvement. Smokers had a higher risk for PTB than non-smokers [3].

The paper offers a comprehensive analysis of the chemotherapy protocols employed in the treatment of pulmonary tuberculosis. The right use of modern methods of chemotherapy now makes it possible to aim at 100% success in the treatment of pulmonary tuberculosis. At the beginning of 1954, an integrated service for tuberculosis was introduced in Edinburgh, with the result that proper methods of chemotherapy were used for all patients requiring treatment. There are three groups in which the question of prophylactic chemotherapy might be considered. 1. Children with a positive tuberculin reaction as the only evidence -of infection. 2. Adolescents with strongly positive tuberculin reactions but no detectable disease only evidence -of infection. 3. Lung lesions of doubtful activity. The greatest disaster that can happen to a patient with tuberculosis is that his organisms become resistant to two or more of the standard drugs. It is the authors' view that all patients with tuberculosis are required to continue chemotherapy for at least one year. Patients with drug-resistant bacilli have to take various forms of "salvage" chemotherapy [4].

The paper critically evaluates the efficacy of rifapentine and isoniazid in the continuation phase of pulmonary tuberculosis treatment. It highlights the promising potential of rifapentine, a rifamycin exhibiting complete cross-resistance with rifampin, especially when administered once weekly, demonstrating comparable effectiveness to daily rifampin in experimental murine tuberculosis models. This suggests a feasible option for once-weekly dosage regimens in human tuberculosis treatment. Rifapentine was first developed in 1965 by Lepetit. A once-weekly regimen including rifapentine starting early in treatment would have the greatest operational advantages, the first ethically justifiable step was to explore once-weekly rifapentine and isoniazid in the continuation phase in regimens that began with a conventional 2 mo of intensive 4-drug chemotherapy. Between December 1991 and July 1995, 672 patients were admitted to the study, 633 from 10 outpatient clinics and 33 from hospitals. Of these, 75 did not fulfill the admission criteria; 53 had negative pretreatment cultures, eight were infected with nontuberculous mycobacteria, eight had other serious diseases (cancer; active hepatitis B), three had extrapulmonary tuberculosis, and three were excluded for miscellaneous reasons [5].

When a case-control study in the city of Samara, 700 miles southeast of Moscow was undertaken, it came with the two most important risk factors for the development of pulmonary tuberculosis in Russia were exposure to raw milk and unemployment. The paper recruited 334 cases between 1 January 2003 and 31 December 2003. Poverty, unemployment, drinking unpasteurized milk, diabetes, living with a relative with tuberculosis, living in overcrowded conditions, a prison or detention history, and low accumulated wealth were associated with an increased risk of tuberculosis. Ensuring a safe milk supply and employment for everyone would be a public health priority. The potential rule of HIV is not included in the paper because of ethical [6].

The objective of this paper was to estimate the incidence of pulmonary tuberculosis (PTB) and identify potential risk factors associated with its occurrence. It defined high-risk areas in Portugal between 2004 to 2006. The paper used data from the National Tuberculosis Control Programme at the General Directorate of Health, the National Statistics Institute, and many more for the study. The spatial analysis of PTB risk factors is characterized by a high prevalence of HIV/ AIDS. There were also included the prison population, unemployment, and overcrowded housing. Living in an area of high TV Incidence is also a risk factor for the development of PTB. Knowledge about the risk factors and proper sanitation will help to decrease the development of PTB [7].

The paper focuses on smear-positive tuberculosis cases and their household contacts in Vitoria, Brazil. The paper included 894 household contacts where 464 had TB infection and 23 developed TB disease. A biomarker to target preventive therapy is urgently needed in the intensity of exposure to increased risk of TB infection and disease among household contacts. The paper reported on the rates of TB infection and secondary TB disease among the people who lived near the patients of TB. Those people are at higher risk for M. tuberculosis infection and disease according to this paper. The study was conducted between February 2008 and October 2013 in Vitoria, Brazil. The main fact that came from the paper is close contact with an infectious TB case is known to be the strongest predictor for infection [8].

3 METHODOLOGY

The stages of research methodology in applying clustering k-means and linear regression for determining the level of risk of pulmonary tuberculosis are as follows:

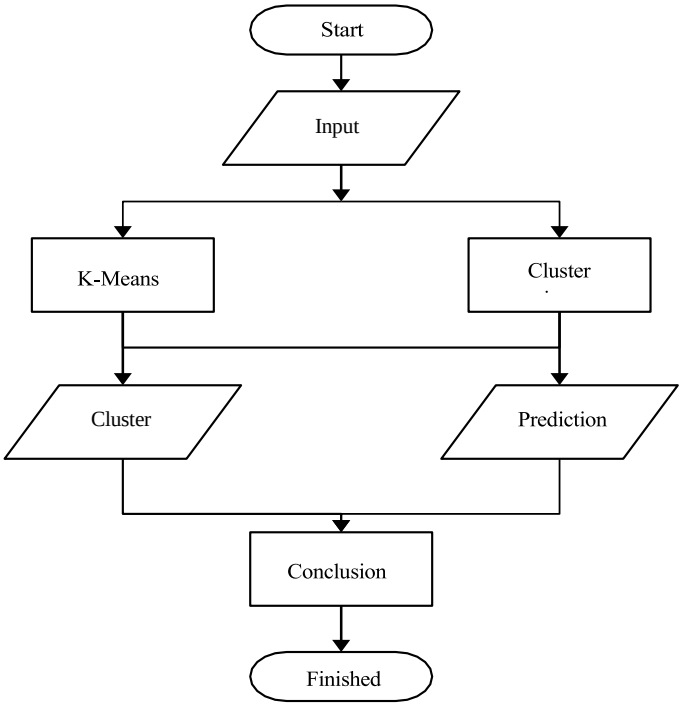

Fig 2. Stages of Research Methodology for Determining Pulmonary TB

Figure 1- The collected tuberculosis dataset is divided into two groups using k-means clustering and linear regression. The value of k will determine the number of clusters. The datasets in k-means show the cluster results. Then clusterwise Regression is the prediction results. The combination of cluster results and prediction results is the final result. Analyzing the data in this process is more effective in determining high-risk areas for pulmonary tuberculosis.

a. Problem Identification and Data Processing

Data analysis in machine learning involves two stages: clustering model with the K-Means algorithm and Cluster Regression method to determine clusters of high-risk pulmonary TB areas and forecasting models to examine the impact of population density on the number of pulmonary tuberculosis patients and to find solutions to the problems generated based on the results of the data sets entered in the analysis.

b. Research Data Analysis

Data analysis in classifying high-risk areas for pulmonary TB. The criterion data are variables used to determine high-risk areas for pulmonary tuberculosis. The research data are as follows:

Table 1. Research Data

No	District	No	District
1	Dhaka	10	Chittagong
2	Barisal	11	Sylhet
3	Khulna	12	Rajshahi
4	Rangpur	13	Mymensingh
5	Comilla	14	Jessore
6	Narayanganj	15	Bogra
7	Gazipur	16	Tangail

8	Jamalpur	17	Dinajpur
9	Faridpur		

c. K-Means Clustering Algorithm

Data clustering is performed for each area by processing it into a cluster, and the k-means algorithm can determine cluster levels in each area based on the values of objects with different values in each group. The cluster model can also identify complex clustering values in determining high-risk areas for diseases. Furthermore, the k-means model can be divided into one or more clusters/groups. This method divides data into clusters or groups, grouping data with similar characteristics into the same cluster, and grouping data with different characteristics into other groups.

The analysis of the Clustering model with the K-Means algorithm in determining risk levels in clustering can divide data into clusters into several groups. Data mining models can group data with similar characteristics into the same cluster and group data with different characteristics into another group.

1. Determine the number of clusters to be formed.
2. Decide on random centroids and initialize clusters according to the number of clusters.
3. Calculate the distance to the centroid using the Euclidean Distance formula, as follows:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

This formula can be generalized to higher dimensions as well. In three-dimensional space, for example, with points (x_1, y_1, z_1) and (x_2, y_2, z_2), the Euclidean distance *d* is:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$$

And in *n* dimensions:

$$\mathrm{d} = \sqrt{\sum_{i=1}^{n} (x_{2i} - x_{1i})^2}$$

where (x_{1i}, x_{2i}) are the coordinates of the points in each dimension.

4. Observe the clustering data with the closest distance value to the centroid.
5. Determining the data center or new centroid.

Next, Data analysis of the Linear regression model used in notation X and one response variable that can be represented by Y. Linear regression is used to assess the extent of influence between one variable and another.

d. Cluster wise Regression

There are three methods for clusterwise regression: linear regression, Finite Mixture Method (FMM), and cluster-weighted method (CWM) [19]. The linear regression method consists of one or more independent variables commonly denoted as X and one response variable represented by Y [20]. Linear regression is used to obtain the forecast value with a variable related to another variable and can make better predictions. Therefore, it is preferable to implement a cluster initialization process using Clusterwise Regression modeling in the next stage. One effective technique for cluster initialization is the Clustering method [21], [22].

e. Research Data

The dataset for the study on the Application of K-Means Clustering and Linear Regression in Determining the Risk Level of Pulmonary Tuberculosis is as follows:

Table 2. Research Data Set

Number	Area	Population (People)	Area (km²)	Pulmonary TB Cases
1	Dhaka	23,936,000	369	1
2	Barisal	549,000	13.23	1
3	Khulna	965,483	59.57	0
4	Rangpur	445,677	2308	6
5	Comilla	670,775	153	3

6	Narayanganj	286330	33.57	5
7	Gazipur	213 061	49.32	3
8	Jamalpur	150 172	2031.98	11
9	Faridpur	122 425	66.24	19
10	Chittagong	5,513,609	5,282.98	13
11	Sylhet	999,374	26.5	0
12	Rajshahi	983,707	34,513	18
13	Mymensingh	497,562	91.32	2
14	Jessore	110 541	2610	0
15	Bogra	944,877	72.5	5
16	Tangail	180,144	29.04	7
17	Dinajpur	206,234	20.7	1

4 RESULTS AND DISCUSSION

Research utilizing the K-Means algorithm for determining the risk levels of pulmonary tuberculosis (TB) has effectively divided data into distinct clusters. This analysis categorizes areas into different risk groups: two areas fall into the first cluster (low risk), several areas into the second cluster (moderate risk), and the remaining areas into the third cluster (high risk).

Following the clustering analysis, the study employs the Clusterwise Regression method to predict the impact of population density on the number of pulmonary TB cases. This approach assesses how population density influences the incidence of pulmonary TB across different areas, providing insights into the correlation between these variables and aiding in targeted public health interventions.

4.1 Application of the K-Means Algorithm Method

- **Data on the population distribution of each sub-district**

The population distribution in the application of k-means clustering and linear regression in determining the risk level of pulmonary tuberculosis is as follows:

Table 3. Data on the distribution of the number of pulmonary TB patients in each district

No	District	Population (People)	Area (Km²)	Number of TB Patients
1	Dhaka	23,936,000	369	1
2	Barisal	549,000	13.23	1
3	Khulna	965,483	59.57	0
4	Rangpur	445,677	2308	6
5	Comilla	670,775	153	3
...	...	...	...	...
...	...	...	...	...
15	Bogra	944,877	72.5	5
16	Tangail	180,144	29.04	7
17	Dinajpur	206,234	20.7	1

- **Initialize the Cluster Center**

The following table shows the population, area, and number of TB patients in each district, along with their assigned clusters after iteration:

Table 4. Iteration Details

Iteration	Information	Population	Area (km²)	Number of Pulmonary TB

Dhaka	Cluster1= C1	23,936,000	369	1
Barisal	Cluster 2 = C2	549,000	13.23	1
Khulna	Cluster 3 = C3	965,483	59.57	0

- ## Nearest Cost Values Using Euclidean Distance

Euclidean Distance is a method used to determine the shortest path or minimum distance between two points in a multidimensional space. In the context of clustering and the given table, the Euclidean Distance helps determine the nearest cluster for each data point based on their calculated cost. The nearest cost values are determined by calculating the Euclidean Distance.

The results are as follows:

Table 5. Distance (Cost) Values in the First Iteration

C1	C2	C3	Cluster Proximity	Cluster Assignment
0	23908700	238670517	0.0000	C1
239087000	0	416483.83	0.0000	C2
238670517	416483.0026	0	0.0000	C3
5952.07231	2645.354067	17979.00106	2645.3541	C2
4164.530169	868.3768975	19771.00014	868.3769	C2
1799.241467	5104.561437	25728.26698	1799.2415	C1
...	...	...	...	...
...	...	...	...	...
24150.14635	27457.17456	48081.03475	24150.1464	C1
16433.14822	13126.26785	7498.120815	7498.1208	C3
8848.596239	5542.23089	15083.18804	5542.2309	C2
Total Proximity			**98672.3064**	

Information

Distance of Dhaka to Clusters

$$D_1(c_1) = \sqrt{(239636000 - 239636000)^2 + (369 - 369)^2 + (1 - 1)^2} = 0$$
$$D_1(c_2) = \sqrt{(239636000 - 549{,}000)^2 + (369 - 13.23)^2 + (1 - 1)^2} = 239087000$$
$$D_1(c_3) = \sqrt{(239636000 - 965{,}483)^2 + (369 - 59.57)^2 + (1 - 0)^2} = 238670517$$

Distance of Barisal to Clusters

$$D_2(c_1) = \sqrt{(549{,}000 - 239636000)^2 + (13.23 - 369)^2 + (1 - 1)^2} = 23908700$$
$$D_2(c_2) = \sqrt{(549{,}000 - 549{,}000)^2 + (13.23 - 13.23)^2 + (1 - 1)^2} = 0$$
$$D_2(c_3) = \sqrt{(549{,}000 - 965{,}483)^2 + (13.23 - 59.57)^2 + (1 - 0)^2} = 416483.0026$$

Distance of Khulna to Clusters

$$D_3(c_1) = \sqrt{(965483 - 239636000)^2 + (59.57 - 369)^2 + (0 - 1)^2} = 238670517$$
$$D_3(c_2) = \sqrt{(965483 - 549000)^2 + (59.57 - 13.23)^2 + (0 - 1)^2} = 416483.83$$
$$D_3(c_3) = \sqrt{(965483 - 965483)^2 + (59.57 - 59.57)^2 + (0 - 1)^2} = 0$$

Each row in the table shows the distance of a point from each cluster center (C1, C2, C3). The "Cluster Proximity" column shows the smallest distance, and the "Nearest Cluster" column indicates the corresponding cluster assignment based on this minimum distance.

- ## Selection of the Fourth Centroid

- **Results of the Fourth Iteration Process**

The cluster centers (centroids) remain unchanged in this iteration. The new centroids obtained from the previous iteration are used to calculate the distances using Euclidean distance. The results are as follows:

Table 6. Final Centroid Centers for the Fourth Iteration

Cluster	Population (people)	Area (km²)	Number of TB Patients
Cluster 1 (C1)	55617,5	37,995	18,5
Cluster 2 (C2)	19646,33333	105,1	3,444444444
Cluster 3 (C3)	30597,16667	129,3933333	5,333333333

These centroids represent the centers of the clusters after the fourth iteration, calculated based on the population, area, and number of TB patients. The Euclidean distance is used to determine the proximity of each district to these centroids.

- **Results of the Fourth Iteration**

The calculations for the fourth iteration of applying machine learning clustering using k-means and linear regression for determining the risk level are as follows:

Table 7. Results of the Fourth Iteration

No	Population	Area (km²)	Number of TB Patients	Cluster 1	Cluster 2	Cluster 3	Cluster Group
12	983707	34513	18	1704.63	37675.69	42895.36	C1
8	150172	203198	11	16410.66	19560.67	24780.32	C1
9	122425	6624	19	1704.63	34266.78	23316.17	C2
3	965483	5957	0	46376.57	10405.34	5185.74	C2
5	670775	153	3	43113.60	7142.37	1923.08	C2
13	497562	91.32	2	39703.54	3732.35	1487.30	C2
14	110541	2610	0	42924.57	6953.34	1733.85	C2
..	...	...	...	...	...	...	...
..	...	...	...	...	...	...	...
1	23,936,000	369	1	22445.75	13525.72	18745.46	C3
2	549,000	13.23	1	25752.77	10218.79	15438.41	C3
4	445,677	2308	6	28397.60	7573.67	12793.34	C3
6	286330	33.57	5	26607.92	9366.97	14586.03	C3
16	180,144	29.04	7	30465.50	5505.97	10725.79	C3
17	206,234	20.7	1	31293.50	4678.14	9897.47	C3

The results in Table 7 represent the completion of the iteration process for the third centroid. Then, the fourth iteration was carried out using the new centroids. The result of this process shows that there are differences in the cluster numbers for each area based on the new centroid values. In the fourth iteration, the cluster results did not change, meaning that the third and fourth clusters remained the same, and the iteration process was stopped. Therefore, it can be concluded that:

- 2 regions are included in Cluster 1,
- 6 regions are included in Cluster 2,
- 9 regions are included in Cluster 3.

- **Results of the K-Means Clustering Algorithm**

a. Cluster Results with K-Means Using Python

The cluster results for each sub-district, distributed across each region using the k-means algorithm, are as follows:

Table 8. K-Means Cluster Results

Sub-District	Cluster_km
Dhaka	2
Barisal	2
Khulna	1
Rangpur	2
Comilla	1
Narayanganj	2
Gazipur	2
Jamalpur	2
Faridpur	0
Chittagong	1
Sylhet	2
Rajshahi	0
Mymensingh	1
Jessore	1
Bogra	1
Tangail	2
Dinajpur	2

b. K-Means Cluster Graph Results

The graphical representation of the implementation of K-Means Clustering and Linear Regression in determining the risk level of pulmonary tuberculosis is as follows:

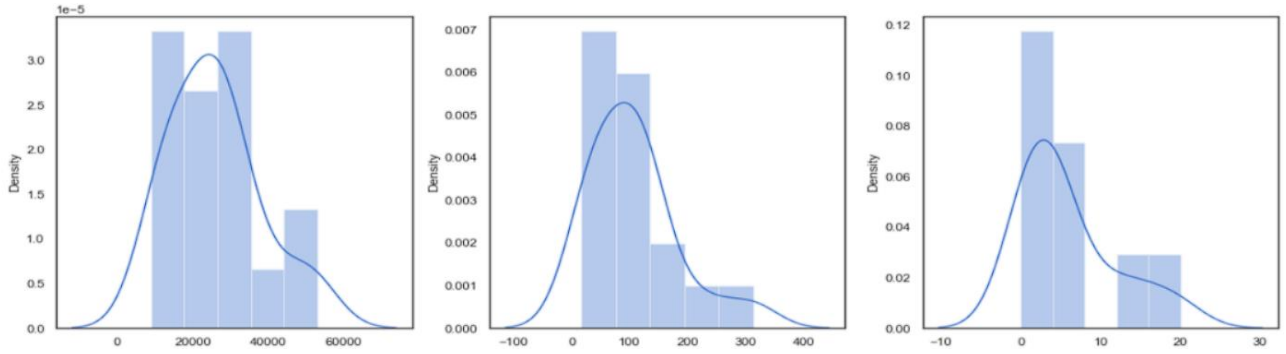

Figure 3. K-Means Cluster Graph

- **Linear Regression Algorithm**

a. Results of X^2, Y^2 and XY

The results for the values of X^2, Y^2, and XY to find the total are as follows:

Table 9. Results for $X2X2$, $Y2Y2$, and $XYXY$

No	X	Y	X^2	Y^2	XY
1	235	1	55225	1	235
2	192	1	36864	1	192
3	81	0	6561	0	0
4	242	6	58564	36	1452
5	98	3	9604	9	294
6	93	5	8649	25	465
7	151	3	22801	9	453
8	359	11	128881	121	3949
..	..	..	..	..	..
..	..	..	..	..	..
15	233	5	54289	25	1165
16	536	7	287296	49	3752
17	629	11	395641	121	6919
Total	**9251**	**100**	**14182549**	**1160**	**109865**

This table shows the calculated values for X^2, Y^2, and XY, which are used in the linear regression analysis to determine the relationship between the variables X and Y. The totals at the bottom are the sums of each column.

b. Calculation of the Model Coefficients a and b

To calculate the coefficients a (the intercept) and b (the slope) in the linear regression model $Y=a+bX$, we use the following formulas:

Slope (b):

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

Intercept (a):

$$\mathbf{a} = \frac{(\sum \mathbf{Y})(\sum \mathbf{X^2}) - (\sum \mathbf{X})(\sum \mathbf{XY})}{\mathbf{n}(\sum \mathbf{X^2}) - (\sum \mathbf{X})^2} = 0{,}00606$$

$$\mathbf{r^2} = \frac{\mathbf{b}(\sum \mathbf{XY})}{\sum \mathbf{Y^2}} = 0{,}57403$$

Y= 2,58415+ 0,00606X

Given the totals from Table 9:

- $\sum X$=9251
- $\sum Y$=100

- $\sum X^2$=14182549
- $\sum Y^2$=1160
- $\sum XY$=109865
- n=17 (number of observations)

This means that approximately 57% of the variation in the dependent variable (x), which is population density, can explain the variation in the number of pulmonary tuberculosis patients. In other words, the variable (x) has an influence of 57% on the variable (y).

- **Calculating the Linear Regression Model Equation**

The simple linear regression equation is given by: $Y=a+bX$

Using the calculated coefficients a=2.584154827 and b=0.006060898, we can predict the values of Y for given values of X.

Calculation of Predicted Values

For X=235:
Y=2.584154827+0.006060898×235
=2.584154827+1.424311044
=4.008465871

For X=192:
Y=2.584154827+0.006060898×192
=2.584154827+1.163692428
=3.747847255

For X=242:
Y=2.584154827+0.006060898×242
=2.584154827+1.46673733
=4.050892157

For X=98:
Y=2.584154827+0.006060898×98
=2.584154827+0.594268004
=3.178422831

For X=93:
Y=2.584154827+0.006060898×93
=2.584154827+0.563664614
=3.147819441

For X=151:
Y=2.584154827+0.006060898×151
=2.584154827+0.915195898
=3.499350725

For X=359:
Y=2.584154827+0.006060898×359
=2.584154827+2.176055794
=4.760210621

For X=1138:

Y=2.584154827+0.006060898×1138

=2.584154827+6.89730199

=9.481456817

By applying the regression formula, we can predict the number of tuberculosis patients based on the given population densities.

- **Graph of the Influence of Population Density on the Number of Pulmonary TB Patients**

The graph depicting the influence of population density on the number of pulmonary tuberculosis patients is as follows:

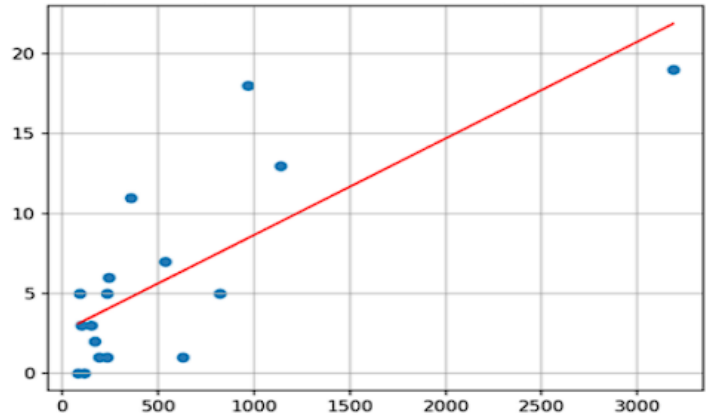

Figure 4. Influence of Population Density on the Number of Pulmonary TB Patients

5 CONCLUSION

Based on the analysis conducted using the K-means algorithm, we have identified specific areas that are particularly susceptible to pulmonary tuberculosis. These findings are instrumental in directing targeted intervention efforts and resource allocation to mitigate the spread of the disease effectively. The clustering results reveal that Cluster 1 encompasses two regions, cluster 2 comprises six areas, and cluster 3 includes nine locations. Such clustering allows for a more nuanced understanding of geographical patterns and enables tailored strategies to address the unique needs of each cluster. Moreover, employing Clusterwise Regression analysis sheds light on the relationship between population density and the incidence of pulmonary tuberculosis. The model indicates that approximately 57% of the variation in the number of tuberculosis patients can be explained by population density. This finding underscores the significance of demographic factors in shaping disease prevalence within communities. However, it's important to acknowledge that while population density plays a substantial role, other variables also contribute to the remaining 43% of variation in tuberculosis cases. Exploring these additional factors could provide further insights into the complex dynamics of disease transmission and inform comprehensive public health interventions. In summary, the combined use of clustering techniques and regression analysis offers a powerful framework for understanding disease patterns and informing targeted interventions. By leveraging these analytical approaches, policymakers and healthcare professionals can devise more effective strategies to combat pulmonary tuberculosis and improve public health outcomes. While this analysis sheds light on the relationship between population density and pulmonary tuberculosis, it has limitations. Data quality issues, the narrow scope of variables, and assumptions of linearity may impact the findings. Future research should integrate diverse data sources, employ non-linear modeling approaches, and conduct spatial-temporal analyses for a more comprehensive understanding. Validating findings and exploring sensitivity will enhance the reliability of results, guiding more effective tuberculosis control strategies.

REFERENCES

1. B. Ula Mutammimul, Bakhtiar, Desvina Yulisda, Badriana, “Application Of The Fuzzy Time Series Model In Clothing Material Stock Forecasting,” J. Sist. Inf. Dan Ilmu Komput. Prima (JUSIKOM PRIMA), vol. 6, no. 1, pp. 56–61, 2022, doi: https://doi.org/10.34012/jurnalsisteminformasidanilmukomputer.v6i1.286 2.
2. CLUSTERWISE REGRESSION PADA STATISTICAL DOWNSCALING UNTUK PENDUGAAN CURAH HUJAN BULANAN,” Indones. J. Stat. Its Appl., vol. 3, no. 3, pp. 236–246, Oct. 2019, doi: 10.29244/IJSA.V3I3.310.
3. F. Hardiyanti, H. S. Tambunan, and I. S. Saragih, “PENERAPAN METODE K- MEDOIDS CLUSTERING PADA PENANGANAN KASUS DIARE DI INDONESIA,” KOMIK (Konferensi Nas. Teknol. Inf. dan Komputer), vol. 3, no. 1, Dec. 2019, doi: 10.30865/komik.v3i1.1666.
4. G. Gustientiedina, M. H. Adiya, and Y. Desnelita, “Penerapan Algoritma K- Means Untuk Clustering Data Obat-Obatan,” J. Nas. Teknol. dan Sist. Inf., vol. 5, no. 1, pp. 17–24, Apr. 2019, doi: 10.25077/TEKNOSI.V5I1.2019.17-24.
5. Gustientiedina Gustientiedina, M. Hasmil Adiya, and Yenny Desnelita, “Penerapan Algoritma K-Means Untuk Clustering Data Obat-Obatan,” J. Nas. Teknol. dan Sist. Inf., vol. 5, no. 1, pp. 17–24, Apr. 2019, doi: 10.25077/TEKNOSI.V5I1.2019.17-24.
6. Iwan Stia Budi, Yustini Ardillah, Indah Purnama Sari, and Dwi Septiawati, “Analisis Faktor Risiko Kejadian penyakit Tuberculosis Bagi Masyarakat Daerah Kumuh Kota Palembang,” J. Kesehat. Lingkung. Indones., vol. 17, no. 2, pp. 87–94, Oct. 2018, doi: 10.14710/JKLI.17.2.87-94.
7. J. Khatib Sulaiman, M. Fatkuroji, Taslim, E. Sabna, and K. Warti Ningsih, “Optimasi Nilai K Pada Algoritma k-Means untuk Klasterisasi Data Pasien Covid-19,” Indones. J. Comput. Sci., vol. 11, no. 2, Sep. 2022, Accessed: Feb. 20, 2023. [Online]. Available: http://ijcs.stmikindonesia.ac.id/ijcs/index.php/ijcs/article/view/3088
8. K. Moulaei, M. Shanbehzadeh, Z. Mohammadi-Taghiabad, and H. Kazemi- Arpanahi, “Comparing machine learning algorithms for predicting COVID-19 mortality,” BMC Med. Inform. Decis. Mak., vol. 22, no. 1, pp. 1–12, Dec. 2022, doi: 10.1186/S12911-021-01742-0/FIGURES/3.

9. M. U. Fitria, Rahma, Desvina Yulisda, “Data Mining Classification Algorithms For Diabetes Dataset Using Weka Tool,” J. Sist. Inf., vol. 2, no. 1, 2021.

10. M. Ula, A. F. Ulva, Mauliza, M. A. Ali, and Y. R. Said, “Application Of Machine Learning In Predicting Children’s Nutritional Status With Multiple Linear Regression Models,” MULTICA Sci. Technol. J., vol. 2, no. 2, pp. 124–130, Feb. 2022, doi: 10.47002/MST.V2I2.363.

11. N. Puspitasari, N. Puspitasari, and F. Urmila Jannah Helmi Puadi, “Klasterisasi Wilayah Penghasil Tanaman Lada Menggunakan Algoritma K-Means,” Indones. J. Comput. Sci., vol. 11, no. 3, Dec. 2022, Accessed: Feb. 20, 2023. [Online]. Available: http://ijcs.stmikindonesia.ac.id/ijcs/index.php/ijcs/article/view/3104

12. R. A. Rizal, N. O. Purba, L. A. Siregar, K. Sinaga, and N. Azizah, “Analysis of Tuberculosis (TB) on X-ray Image Using SURF Feature Extraction and the K- Nearest Neighbor (KNN) Classification Method,” JAICT, vol. 5, no. 2, pp. 9–12, Oct. 2020, doi: 10.32497/JAICT.V5I2.1979.

13. R. Ula, M., Ulva, A. F., Mauliza, M., Sahputra, I., Ridwan, “Implementation of Machine Learning in Determining Nutritional Status using the Complete Linkage Agglomerative Hierarchical Clustering Method,” J. Mantik, vol. 5, no. 3, pp. 1910–1914, 2021.

14. Rahmana Dwi Shaputra and Syarif Hidayat, “Implementasi regresi linear untuk prediksi penjualan pada aplikasi point of sales restoran,” AUTOMATA, vol. 2, no. 1, Jan. 2021, Accessed: Feb. 15, 2023. [Online]. Available: https://journal.uii.ac.id/AUTOMATA/article/view/17355

15. Raisuli Ramadhan, Eka Fitria, and Rosdiana Rosdiana, “DETEKSI Mycobacterium Tuberculosis Dengan Pemeriksaan Mikroskopis Dan Teknik Pcr Pada Penderita Tuberkulosis Paru Di Puskesmas Darul Imarah,” J. Penelit. Kesehat., vol. 4, no. 2, pp. 73–80, Nov. 2017, doi: 10.22435/SEL.V4I2.1463.

16. Ratih Sari Wardani, Purwanto, Sayono, and Aditya Paramananda, “Clustering tuberculosis in children using K-Means based on geographic information system,” AIP Conf. Proc., vol. 2114, no. 1, p. 060012, Jun. 2019, doi: 10.1063/1.5112483.

17. Septian Wulandari and Dian Novita, “Analisis Clustering Virus MERS-CoV Menggunakan Metode Spectral Clustering Dan Algoritma K-Means,” STRING (Satuan Tulisan Ris. dan Inov. Teknol., vol. 5, no. 3, pp. 315–323, Apr. 2021, doi: 10.30998/STRING.V5I3.7942.

18. Suhandio Handoko, F. Fauziah, and Endah Tri Esti Handayani, “Implementasi Data Mining Untuk Menentukan Tingkat Penjualan Paket Data Telkomsel Menggunakan Metode K-Means Clustering,” J. Ilm. Teknol. dan Rekayasa, vol. 25, no. 1, pp. 76–88, May 2020, doi: 10.35760/TR 2020.V25I1.2677.

19. V. P. Butar-butar, A. M. Soleh, and A. H. Wigena, “PEMODELAN

20. W. M. Baihaqi, M. Dianingrum, K. Aswin, and N. Ramadhan, “Regresi Linier Sederhana Untuk Memprediksi Kunjungan Pasien Di Rumah Sakit Berdasarkan Jenis Layanan Dan Umur Pasien,” Simetris J. Tek. Mesin, Elektro dan Ilmu Komput., vol. 10, no. 2, pp. 671–680, Nov. 2019, doi: 10.24176/SIMET.V10I2.3484.

21. W. Santoso, K. Hulliyah, W. Nurjannah, and A. H. Setianingrum, “Systematic Literature Review: Virus Prediction Based on DNA Sequences using Machine Learning and Deep Learning method,” in 2022 10th International Conference on Cyber and IT Service Management (CITSM), Sep. 2022, pp. 1–7. doi: 10.1109/CITSM56380.2022.9935921.

22. Yogi Arvendo Pratama, “Karakteristik Klinis Penyakit Tuberkulosis Paru pada Anak,” J. Penelit. Perawat Prof., vol. 3, no. 2, pp. 237–242, Mar. 2021, doi: 10.37287/JPPP.V3I2.403.

23. Yusmandin Idris and Mursal Ismail, “Masih Banyak Penderita TBC di Bireuen, Berikut Sebab dan Cara Pencegahannya - Serambinews.com,” Serambinews, Aug. 23, 2019.https://aceh.tribunnews.com/2019/08/23/masih-banyak-penderita-tbc-di- bireuen-berikut-sebab-dan-cara-pencegahannya (accessed Feb. 15, 2023).