JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 4, August 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i4.8914
Hal 521-530
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Classification of Lung TB Levels by Region in Medan City Using Logistic Regression Algorithm

Sri Purnamawati*, Ilka Zufria

Fakultas Sains dan Teknologi, Ilmu Komputer, Universitas Islam Negeri Sumatera Utara, Medan, Indonesia
Email: [1*]sripurnamawati535@gmail.com, [2]ilkazufria@uinsu.ac.id
Correspondence Author Email: sripurnamawati535@gmail.com
Submitted **21-07-2025;** Accepted **11-08-2025;** Published **14-08-2025**

Abstract

Tuberculosis (Pulmonary TB) is an infectious disease that is still a serious problem in Medan City, especially due to the high population density and the increasing number of cases. This condition requires an analytical approach that is able to accurately map areas with high risk levels. This study aims to develop a risk classification model for Lung TB in Medan City based on three main variables: number of patients, population density, and area. The model was built by combining K-Means Clustering algorithm to form risk category labels and Logistic Regression to predict binary classification (High Risk or Low Risk). The dataset used is secondary data for 2024 with coverage of 21 sub-districts, organized in monthly time series format. The data is divided by time into training data (January-October) and test data (November-December). The model evaluation results show that Logistic Regression successfully classifies regions with an accuracy of 92.86%, recall 100%, precision 90.32%, and F2-score of 97.90%. Visualization of the prediction results per sub-district also shows the consistency of classification over time and space. These findings indicate that the model performs well and can be utilized as a decision support system to map high-risk areas of Pulmonary TB in a more focused manner in Medan City.

Keywords: Classification; Pulmonary TB; Medan City; Logistic Regression; K-Means

1. INTRODUCTION

Tuberculosis (Lung TB) is an infectious disease caused by the bacterium *Mycobacterium tuberculosis*. This bacteria usually attacks the lungs, but can also spread to other parts of the body such as the kidneys, spine, and brain [1]. Transmission occurs through droplets of saliva or sputum that are inhaled into the air, especially when the patient coughs or sneezes, which allows the bacteria to spread into the surrounding environment and be inhaled by others [2]. Infections that do not immediately progress to disease often cause no symptoms, however, a small percentage of cases may show symptoms after the bacteria have multiplied and invaded certain organs. Symptoms include coughing for more than two weeks, coughing up phlegm that may be mixed with blood, chest pain, shortness of breath, malaise, weight loss, loss of appetite, chills, fever, and night sweats. Pulmonary TB not only affects the lungs, but can reach other organs such as the skin, brain, and spine [3]. The bacteria that cause pulmonary TB consist of five types that are closely related to the infection, namely *Mycobacterium tuberculosis*, *Mycobacterium bovis*, *Mycobacterium africanum*, *Mycobacterium microti*, and *Mycobacterium cannettii*, with *M. tuberculosis* being the most common type and is airborne between humans [4].

The spread of Pulmonary TB in Indonesia continues to show an increasing trend every year and is not only a public health problem in general, but also a challenge at the provincial level. North Sumatra Province is one of the regions that recorded a high number of pulmonary TB cases, with a report from the Central Bureau of Statistics mentioning 51,827 cases in 2024 [5]. This condition shows that the number of Lung TB cases in the province occupies a sizable number on a national scale. One factor that is thought to influence this increase is the high population density in some areas of North Sumatra, which provides greater opportunities for the transmission of Pulmonary TB between individuals. High social interaction in dense neighborhoods and limited access to health facilities are the main triggers in expanding the chain of transmission of this disease [6].

Medan City, as one of the largest cities in Indonesia with a high population density, shows an increasing trend of Pulmonary TB cases every year which is influenced by factors of population density, living environment, and access to health services [7]. Based on data from the North Sumatra Central Bureau of Statistics, Medan City has a population density of 8,902 people per square kilometer and reported 17,161 TB cases in 2024 [8]. This figure places Medan City as an area with a high number of TB cases among other regions in North Sumatra . High population density in urban areas, such as Medan City, is often associated with suboptimal environmental conditions that support health. Densely populated areas have intense levels of social interaction that can accelerate the transmission of Pulmonary TB. In this case, it is important to predict the risk of the spread of Pulmonary TB, considering that population density is one of the main indicators that need to be analyzed to understand the pattern of spread of this disease [9].

This condition shows that Medan City is an area with social and demographic characteristics that can accelerate the spread of Lung TB, so a prediction system is needed that is able to identify areas with high risk levels so that countermeasures can be carried out in a focused manner. Risk modeling that considers quantitative indicators such as population density, number of patients, and area is considered important to support the decision-making process in efforts to control Lung TB. Previous research conducted by [10] has utilized *machine learning* methods by combining K-Means Clustering and linear regression to group Lung TB cases based on certain characteristics and predict disease risk based on linear relationships between variables. Patient clustering is done based on the characteristics identified in the *dataset*,

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 4, August 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i4.8914
Hal 521-530
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

while risk prediction is based on the linear relationship between risk factors and the severity of Pulmonary TB. The prediction results through *Clusterwise Regression* showed a 57% contribution to the variation in the number of cases, while the rest was influenced by other factors not explained in the model. However, this approach has limitations as it only produces quantitative predictions of the number of cases on a continuous basis, rather than the classification of risk categories that are more needed in mapping priority areas [11]. Another weakness is that there is no time weighting and limited spatial coverage, so the modeling does not reflect the dynamic fluctuations of disease spread at the administrative level [12]. To overcome these shortcomings, the prediction model in this study was designed by combining K-Means Clustering as a classifier and Logistic Regression to classify the risk of Lung TB into High Risk and Low Risk categories based on population density, number of patients, and area. The modeling was conducted temporally every month throughout 2024 and focused on Medan City areas characterized by high population density, so as to capture the pattern of the spread of Lung TB in a more contextual and applicable manner [13].

The development of this classification model is aimed at predicting the risk of Lung TB based on data on population density, number of patients, and area in Medan City using the Logistic Regression algorithm. In addition to building a classification model, analysis was conducted to identify and test the relationship between the independent variables and the risk level of Lung TB. The resulting model was then evaluated by measuring accuracy and prediction performance to determine the extent to which the Logistic Regression algorithm was able to map areas with High Risk and Low Risk classifications appropriately [14].

The use of the Logistic Regression algorithm is done because this method can model the relationship between independent variables and binary dependent variables, and can be used for both numerical and categorical variables [15]. The use of *machine learning* allows machines to recognize patterns from data and generate predictions without explicit programming [16], so that data on population density, number of patients, and area can be processed to project the risk of spreading Pulmonary TB automatically. By utilizing the characteristics of Medan City, which has a high density and a large number of Lung TB cases, the classification results of this model are expected to support the risk mapping process and become a reference in decision making that focuses on preventing the transmission of Lung TB in areas that are classified as vulnerable [17].

2. RESEARCH METHODOLOGY

2.1 Data Collection

This research dataset is the result of integrating data on Pulmonary TB cases and demographic data from the Medan City Health Office and the North Sumatra Central Bureau of Statistics (BPS), which is compiled in a monthly time-series format from January to December 2024. The dataset includes 251 rows of data representing 21 sub-districts in Medan City for 12 months, with a total of 11,285 Pulmonary TB cases. The data structure consists of three independent variables, namely population density (people/km²), number of Lung TB cases per month, and area (km²), and one dependent variable in the form of a Lung TB risk label generated through the K-Means clustering technique, with a value of 1 for High Risk and 0 for Low Risk. This label was used as a target in the classification process using the Logistic Regression model [18]. Data access was done through formal requests to both agencies, then curated and combined into one structured dataset. Data collection was conducted through: observation of data at puskesmas, hospitals, and other health service centers; interviews with Mrs. Sumartini, SKM, M.Kes, TB Supervisor at the P2P Division of the Medan City Health Office; and literature review to understand the relationship between population density and the spread of Lung TB, as well as the application of the K-Means algorithm and Logistic Regression in infectious disease risk prediction [19].

2.2 Data Preprocessing

Data preprocessing consisted of three main stages: (1) data cleaning, including removal of missing values, removal of data with invalid date formats, and elimination of duplicates based on Year, Month, and Subdistrict combinations; this stage ensures clean, complete, and non-redundant data; (2) feature extraction and aggregation, including extraction of Month and Year variables from the date column, selection of relevant columns (Subdistrict, Density, Area, Date), aggregation of the number of patients per subdistrict per month, and generation of one row of data for each combination of time and area without repetition; The result is a data structure ready for spatial-temporal classification; (3) feature normalization, which is the scaling of the numerical features Number of Patients, Population Density, and Area using the Z-score standardization method with StandardScaler from the scikit-learn library; this normalization equalizes the scale between features so that no variable dominates the model learning process [16].

2.3 Clustering

The K-Means Clustering algorithm was used as an *unsupervised learning* approach to form classification target variables or generate Lung TB risk labels. This algorithm is applied to normalized numerical features, namely Number of Patients, Population Density, and Area [8]. The number of clusters was determined as two (k = 2), assuming that administrative areas can be categorized into two main groups based on risk: *High Risk* and *Low Risk*. After the clustering process was completed, each data entry was labeled 0 for Low Risk and 1 for High Risk. The selection of K-Means as the clustering method is based on its ability to group data based on spatial proximity in the feature space, without requiring a prior label.

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 4, August 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i4.8914
Hal 521-530
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

The result of this clustering is used as the target variable (dependent variable) in the *Logistic Regression* classification model [20].

2.4 Data Splitting

Dataset splitting was performed using the *time-based split* method to distinguish training and test data based on time sequence. The variable 'Month' was used as the basis for splitting, with data from January to October used as training data, and data from November to December used as test data. The separation process was done manually using logical conditions on the time column to ensure no leakage of information from the future into the model training process. The training dataset consists of 209 rows of data, while the test dataset consists of 42 rows of data. Based on the clustering result labels, the distribution in the training data includes 142 data labeled High Risk and 67 Low Risk data, while in the test data there are 29 High Risk data and 13 Low Risk data.

2.5 Modeling

A classification model was developed using the Logistic Regression algorithm to predict the probability of an area falling into the High Risk (label 1) or Low Risk (label 0) category, based on three features: number of Lung TB patients, population density, and area. The model was developed using the scikit-learn library with the LogisticRegression class, where the regularization parameter was set with a value of C = 0.1 (L2 penalty) and a maximum iteration limit of 1000. The training process was conducted on training data covering the months of January to October. The model output in the form of a probability value is calculated through a sigmoid activation function, then converted into a binary class prediction with a threshold of 0.5. Logistic Regression was chosen because it is a suitable method for binary classification and can provide interpretative probability estimates. The use of L2 regularization aims to prevent overfitting and improve model generalization to unseen data.

2.6 Evaluation

The classification model was evaluated using a confusion matrix, namely Accuracy, Precision, Recall, and F1-Score, to map the model's predicted results to High Risk and Low Risk classes, so that classification errors could be identified. These metrics were used to provide an assessment of the model's performance, where F2-score was prioritized as the main focus of the study was to ensure High Risk cases were not missed by the model.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$\text{F1} - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

3. RESULT AND DISCUSSION

3.1 Clustering

The results of clustering using the K-Means algorithm show that the data can be divided into two clusters representing two levels of risk of Lung TB, namely Low Risk (label 0) and High Risk (label 1). The clustering process was performed on normalized numerical features: number of Lung TB patients, population density, and area. The number of clusters was set to two (k = 2), in accordance with the binary classification objective. To determine the classification target labels, the risk labels were generated using the K-Means algorithm applied to the normalized features (Number of Patients, Density, Area), with the clustering results assigned as: 0 = Low Risk; 1 = High Risk.

To gain a deeper understanding of how the K-Means algorithm operates in practice, a manual simulation of the clustering process was carried out using 10 sample data points extracted from the dataset. The features selected for this simulation were the three normalized variables. The sample data used is shown below:

Table 1. Sample Data for Manual K-Means Calculation

Kecamatan	X1	X2	X3
Medan Barat	0,3194	-0,7279	-0,7415
Medan Labuhan	-1,2874	1,6994	2,2522
Medan Barat	0,3194	-1,3899	-0,7415
Medan Perjuangan	1,7644	0,5960	-0,9289
Medan Perjuangan	1,7644	2,3614	-0,9289
Medan Timur	0,1710	1,2580	-0,4759
Medan Sunggal	-0,3015	-0,2866	-0,0209
Medan Deli	-0,2909	2,3614	0,5590
Medan Selayang	-0,8997	-0,5072	0,3112
Medan Marelan	-0,8938	2,1407	1,7253

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 4, August 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i4.8914
Hal 521-530
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

To illustrate the clustering mechanism in greater detail, manual Euclidean distance calculations were performed iteratively for one representative sub-district, Medan Marelan, across three clustering iterations, until the centroids stabilized.

a. Iteration 1

The initial centroid is randomly selected as follows:

Table 2. Initial Centroid

	C0_X1	C0_X2	C0_X3	C1_X1	C1_X2	C1_X3
Value	0,31	-0,939	1,2421	-0,1477	0,4474	-0,5918

The value of Medan Marelan is: $x = (X^1, X^2, X^3) = (-0.8939, 2.1407, 1.7254$. Calculation of Euclidean distance to the two centroids:

1. Distance to C0

$$D_{c0} = \sqrt{(-0.8939 - 0.3100)^2 + (2.1407 + 0.9390)^2 + (1.7254 - 1.2421)^2} = 3.3418$$

2. Distance to C1

$$D_{c0} = \sqrt{((-0.8939 + 0.1477)^2 + (2.1407 - 0.4474)^2 + (1.7254 + 0.5918)^2)} = 2.9654$$

Since $D_{c1} < D_{c0}$, then Medan Marelan is classified into Cluster 1 (High Risk).

Table 3. Iteration 1

Kecamatan	X1	X2	X3	C0	C1	
Medan Barat	0,3194	-0,7279	-0,7415	1,9948	1,2736	C1
Medan Labuhan	-1,2874	1,6994	2,2522	3,2455	3,3098	C0
Medan Barat	0,3194	-1,3899	-0,7415	2,0342	1,9017	C1
Medan Perjuangan	1,7644	0,5960	-0,9289	3,0307	1,9473	C1
Medan Perjuangan	1,7644	2,3614	-0,9289	4,2097	2,7264	C1
Medan Timur	0,1710	1,2580	-0,4759	2,7925	0,8787	C1
Medan Sunggal	-0,3015	-0,2866	-0,0209	1,5475	0,9425	C1
Medan Deli	-0,2909	2,3614	0,5590	3,4235	2,2379	C1
Medan Selayang	-0,8997	-0,5072	0,3112	1,5862	1,5140	C1
Medan Marelan	-0,8938	2,1407	1,7253	3,3418	2,9653	C1

b. Iteration 2

The centroid is updated based on the average value of each cluster from the first iteration:

Table 4. Updated Centroid Values

	C0_X1	C0_X2	C0_X3	C1_X1	C1_X2	C1_X3
Value	-1,10592	-0,09622	1,608563	0,28201	0,024537	-0,41018

1. Distance to C0

$$D_{c0} = \sqrt{\left(-0.8939 - (-1.10592)\right)^2 + (2.1407 - (-0.09622)^2 + (1.7254 - 1.6086)^2} = 2.2500$$

2. Distance to C1

$$D_{c1} = \sqrt{(-0.8939 - 0.2820)^2 + (2.1407 - 0.0245)^2 + (1.7254 - (-0.4102))^2} = 3.2283$$

Because $D_{c0} < D_{c1}$, Medan Marelan is now included in Cluster 0 (Low Risk).

Table 5. Iteration 2

Kecamatan	X1	X2	X3	C0	C1	
Medan Barat	0,3194	-0,7279	-0,7415	2,8202	0,8230	C1
Medan Labuhan	-1,2874	1,6994	2,2522	1,9161	3,5152	C0
Medan Barat	0,3194	-1,3899	-0,7415	3,0378	1,4532	C1
Medan Perjuangan	1,7644	0,5960	-0,9289	3,8932	1,6713	C1
Medan Perjuangan	1,7644	2,3614	-0,9289	4,5517	2,8156	C1
Medan Timur	0,1710	1,2580	-0,4759	2,7946	1,2402	C1
Medan Sunggal	-0,3015	-0,2866	-0,0209	1,8271	0,7673	C1
Medan Deli	-0,2909	2,3614	0,5590	2,7938	2,5939	C1
Medan Selayang	-0,8997	-0,5072	0,3112	1,3764	1,4831	C0
Medan Marelan	-0,8938	2,1407	1,7253	2,2500	3,2282	C0

c. Iteration 3

Performed with the following new centroid:

Table 6. Updated Centroid Values

	C0_X1	C0_X2	C0_X3	C1_X1	C1_X2	C1_X3

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 4, August 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i4.8914
Hal 521-530
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Value	-1,14857	0,11359	1,534371	0,352947	-0,03491	-0,4715

1. Distance to C0

$$D_{c0} = \sqrt{(-0.8939 - (-1.1486))^2 + (2.1407 - 0.1136)^2 + (1.7254 - 1.5344)^2} = 2.0520$$

2. Distance to C1

$$D_{c0} = \sqrt{(-0.8939 - 0.3529)^2 + (2.1407 - (-0.0349))^2 + (1.7254 - (-0.4715))^2} = 3.3338$$

Because $D_{c0} < D_{c1}$, Medan Marelan remains in Cluster 0.

Table 7. Iteration 3

Kecamatan	X1	X2	X3	C0	C1	
Medan Barat	0,3194	-0,7279	-0,7415	2,8360	0,7445	C1
Medan Labuhan	-1,2874	1,6994	2,2522	1,7462	3,6218	C0
Medan Barat	0,3194	-1,3899	-0,7415	3,0976	1,3820	C1
Medan Perjuangan	1,7644	0,5960	-0,9289	3,8453	1,6124	C1
Medan Perjuangan	1,7644	2,3614	-0,9289	4,4279	2,8185	C1
Medan Timur	0,1710	1,2580	-0,4759	2,6632	1,3057	C1
Medan Sunggal	-0,3015	-0,2866	-0,0209	1,8156	0,8334	C1
Medan Deli	-0,2909	2,3614	0,5590	2,5960	2,6868	C0
Medan Selayang	-0,8997	-0,5072	0,3112	1,3940	1,5507	C0
Medan Marelan	-0,8938	2,1407	1,7253	2,0519	3,3338	C0

The manual calculation of Medan Marelan sub-district shows the dynamics of the cluster position changing during the initial iterations before finally settling. In the first iteration, Medan Marelan was initially classified as High Risk because the Euclidean distance to the centroid of cluster 1 was smaller than that of cluster 0. However, after updating the centroid based on the initial clustering results, the centroid position moved and caused Medan Marelan to enter the Low Risk cluster in the second iteration, and remained in that cluster in the third iteration and so on.

This change in cluster position reflects how K-Means adapts the group center representation based on the actual data structure. It also shows that even if a data is initially close to one centroid, data redistribution can change the center of gravity of the group and affect the final classification result. In this case, Medan Marelan has a very high population density value and a large area, which initially makes it closer to the High Risk centroid. However, after a centroid recalibration that considers all cluster members, the sub-district is closer to the Low Risk cluster.

This manual calculation simulation proves that the clustering result does not depend on a single calculation of distance, but rather through an iterative process that continuously improves the position of the centroid until the data distribution stabilizes. Thus, the clustering result label does not only reflect the initial numerical proximity, but the result of the process of forming a group structure that represents the real distribution of the data.

This process is continued automatically by the system until it converges at the 7th iteration. Based on the final results, all data were labeled with a Lung TB risk classification according to the cluster results: 0 for Low Risk and 1 for High Risk. This label becomes the target variable in the Logistic Regression classification model.

3.2 Modeling

The classification model that has been built using the Logistic Regression algorithm produces parameters in the form of one intercept value (bias) and three coefficients for each feature: number of Lung TB patients, population density, and area, which have previously been normalized. The values of model parameters obtained after the training process are as follows: Intercept (bias) = 1.1779; Coefficient of Number of Patients = -0.3540; Coefficient of Population Density = 0.9883; Coefficient of Area = -1.4857. These coefficients are used to form a linear function $F_1(x)$ and calculate the classification probability using a sigmoid function. The probability value is then used to determine the binary prediction with a threshold of 0.5.

To ensure that the interpretation of the model results could be thoroughly understood, manual calculations were performed on three samples of sub-district data. Each calculation includes the linear function value $F_1(x)$, the prediction probability, and the final classification based on the threshold.

$$F_1(x) = Bias + w_1x_1 + w_2x_2 + w_3x_3$$

$$Predict\ Probability = \frac{1}{1 + e^{-F_1(x)}}$$

a. Medan Deli

X_1 (*Scaled* Number of Patients) = 2.361414
X_2 (*Scaled* Density) = -0.290986
X_3 (Area *Scaled*) = 0.559086

1. Calculate $F_1(x)$

$$F_1(x) = 1.1779 + (-0.3540 \times 2.361414) + (0.9883 \times -0.290986) + (-1.4857 \times 0.559086)$$

$$F_1(x) = -0.7763$$

2. *Predict Probability*

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 4, August 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i4.8914
Hal 521-530
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

$$Predict\ Probability = \frac{1}{1 + e^{0.7762}} = \frac{1}{1 + 2.173} = 0.3151$$

Prediction: Low Risk (0)

b. Medan Tembung
   $X_1 = 1.699409$
   $X_2 = 1.108465$
   $X_3 = -0.584294$
   1. Calculate $F_1(x)$
   $$F_1(x) = 1.1779 + (-0.3540 \times 1.699409) + (0.9883 \times 1.108465) + (-1.4857 \times -0.584294)$$
   $$F_1(x) = 2.5401$$
   2. *Predict Probability*
   $$Predict\ Probability = \frac{1}{1 + e^{2.5401}} = \frac{1}{1 + 2.173} = 0.9269$$
   Prediction: High Risk (1)

c. Medan Marelan
   $X_1 = -0.507273$
   $X_2 = -0.893894$
   $X_3 = 1.725376$
   1. Calculate $F_1(x)$
   $$F_1(x) = 1.1779 + (-0.3540 \times -0.507273) + (0.9883 \times -0.893894) + (-1.4857 \times 1.725376)$$
   $$F_1(x) = -2.0890$$
   2. *Predict Probability*
   $$Predict\ Probability = \frac{1}{1 + e^{2.0890}} = \frac{1}{1 + 2.173} = 0.1101$$
   Prediction: Low Risk (0)

Based on the results of manual calculations on the three sample sub-districts, the logistic regression model shows a consistent classification pattern and conforms to the logic of the relationship between features and the risk of Pulmonary TB. The linear function values and prediction probabilities show the sensitivity of the model to the combination of normalized values of the number of patients, population density, and area. The model not only performs a binary prediction function, but also conveys a classification logic that can be quantitatively traced and explained. The pattern of the relationship between the direction of the coefficients and the classification results can also be intuitively understood: an increase in population density increases the risk, while an increase in the number of patients and area tends to decrease the probability of High Risk classification, in accordance with the negative sign of the coefficients.

In the case of Medan Deli, despite having a very high number of patients (2.36), the low population density (-0.29) and medium area (0.56) resulted in a negative linear value and probability of only 0.3151, thus being classified as Low Risk. This shows that a high number of patients does not automatically lead to a High Risk prediction if the population density is also low and the area is relatively large. The negative coefficients on the number of patients and area indicate that increasing these two features actually decreases the risk probability value, as long as the density is not high. While in Medan Tembung, the high number of patients (1.69), high population density (1.11), and small area (-0.58) together produce a large positive F1(x)F_1(x)F1(x) value (2.5401), which results in a probability of 0.9269. This indicates that when two supporting factors of high risk (number of patients and density) increase, and one protective factor (area) decreases, the probability of predicting High Risk is very high. This also confirms that the model effectively captures the directional interaction of each feature on the classification results. In Medan Marelan, all features showed the opposite direction towards High Risk: low patient count (-0.51), very low density (-0.89), and large area (1.72). This combination results in a very low F1(x)F_1(x)F1(x) value (-2.0890), with a probability of only 0.1101, thus predicting Low Risk. This is in line with the interpretation that areas with loose population, few cases, and wide geographical coverage are less likely to be classified as vulnerable.

3.3 Evaluation

Evaluation of the classification model's performance was conducted using test data consisting of 42 rows of data. The purpose of the evaluation is to determine how well the model is able to classify the risk of Lung TB into two classes, namely High Risk and Low Risk. Based on the prediction results of the model on the test data, the values of the confusion matrix components were obtained. : True Positive (TP) = 28 (High Risk prediction, and correct); False Positive (FP) = 3 (High Risk prediction, but actually Low Risk); True Negative (TN) = 11 (Low Risk prediction, and correct); False Negative (FN) = 0 (Low Risk prediction, but actually High Risk).

With these components, the evaluation metrics are calculated using the following formula:

$$Precision = \frac{TP}{TP + FP} = \frac{28}{28 + 3} = 0.9032$$

$$Recall = \frac{TP}{TP + FN} = \frac{28}{28 + 0} = 1.0000$$

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 4, August 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i4.8914
Hal 521-530
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} = \frac{28 + 11}{28 + 0 + 11 + 3} = 0.9286$$

$$F2\ Score = \frac{5 \times Precision\ \times\ \text{Recall}}{(4 \times Precision) +\ Recall} = \frac{5 \times 0{,}9032 \times 1{,}0000}{4 \times 0.9032 + 1.0000} = 0.9790$$

Based on these calculations, an accuracy of 92.86% was obtained, which means that out of all test data, the model managed to make correct predictions in 39 out of 42 cases. The precision rate of 90.32% indicates that the majority of High Risk predictions by the model are indeed data with the High Risk label, although there are still 3 cases that are misclassified to that class. Meanwhile, recall reached 100%, indicating that all High Risk data was successfully recognized without missing any. This is very important in the context of classifying areas prone to Lung TB, as missing High Risk areas can be potentially fatal to disease management.

The F2-score value of 97.90% supports the conclusion that the model has excellent sensitivity to the High Risk class. Since the F2-score gives greater weight to recall, this high score confirms that the model successfully fulfills the main priority in this study, which is to optimally detect all high potential areas.

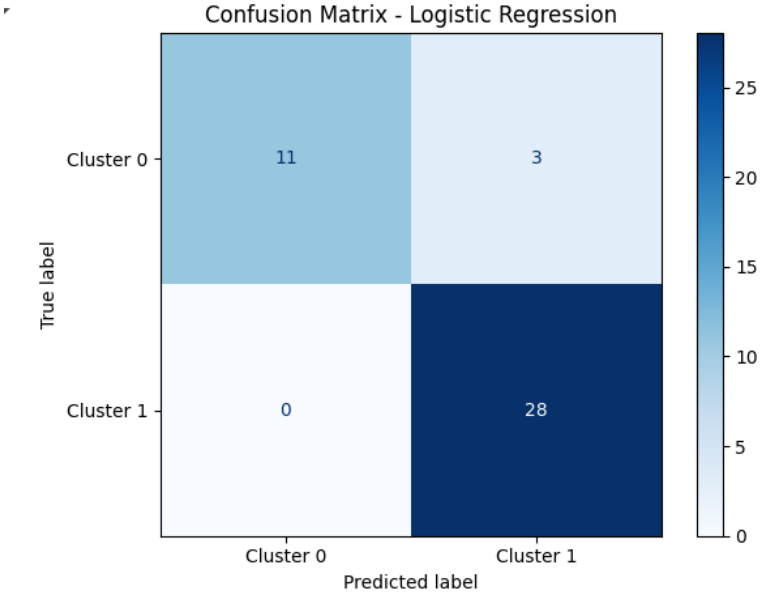
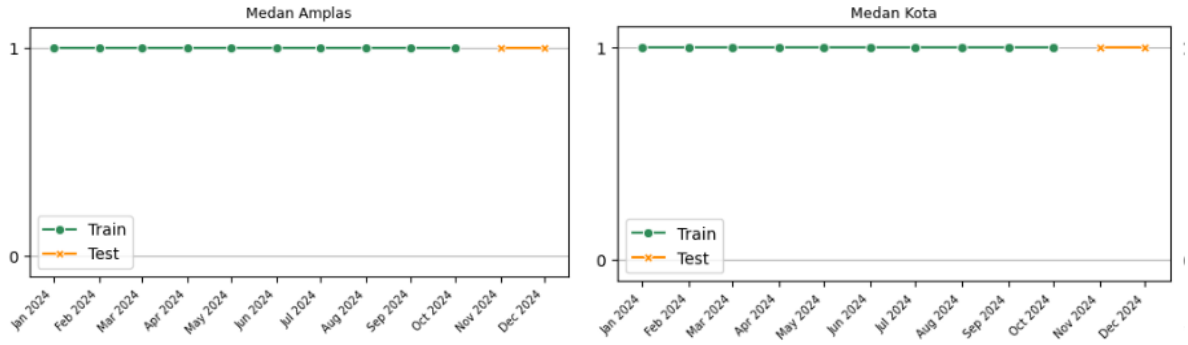


Figure 1. Confusion Matrix

The classification model using Logistic Regression showed consistent and stable performance on the test data, with high detection ability of the High Risk category and relatively small prediction errors. The error composition is more likely to be *false positive* than *false negative*, which in the context of controlling infectious diseases such as Lung TB, is much more acceptable than failing to recognize high risk areas.

To evaluate the prediction results, a visualization of the Lung TB risk classification by month and sub-district was conducted. This visualization not only serves as a complement to the numerical evaluation, but also provides a dynamic view of the spatial and temporal distribution of the prediction results. Each sub-graph displays the change in risk status for one sub-district from month to month during 2024, based on the training data (January-October) and test data (November-December).

From the overall visualization, it can be observed that most sub-districts are stably classified in the High Risk category throughout the year. This can be seen in Medan Amplas, Medan Area, Medan Kota, Medan Baru, and Medan Helvetia, where the prediction status does not change significantly between the training and testing months.



JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 4, August 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i4.8914
Hal 521-530
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

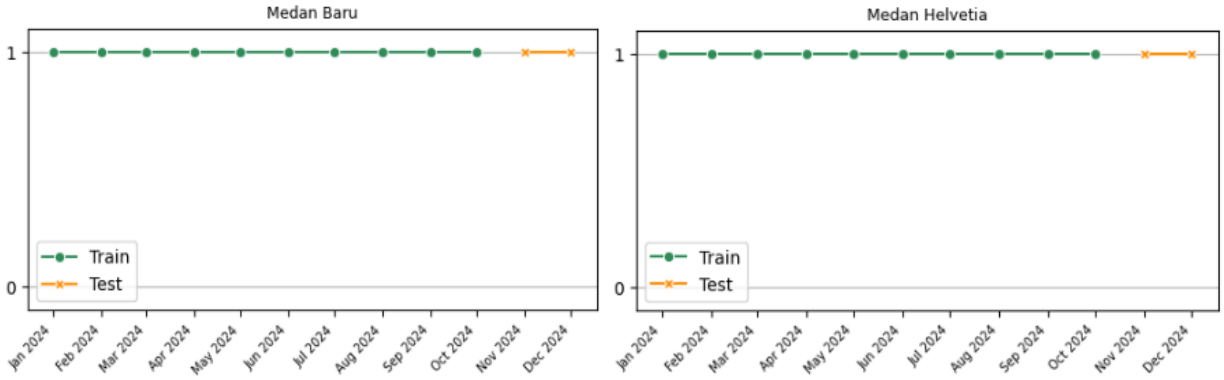


Figure 2. Predicted Medan Amplas, Medan Area, Medan Kota, Medan Baru, and Medan Helvetia,

Some other sub-districts, such as Medan Marelan, Medan Labuhan, and Medan Tuntungan, were consistently in the Low Risk category during the same period. Meanwhile, sub-districts such as Medan Deli showed a change in classification from Low Risk to High Risk in the middle of the year and decreased again in the following month.

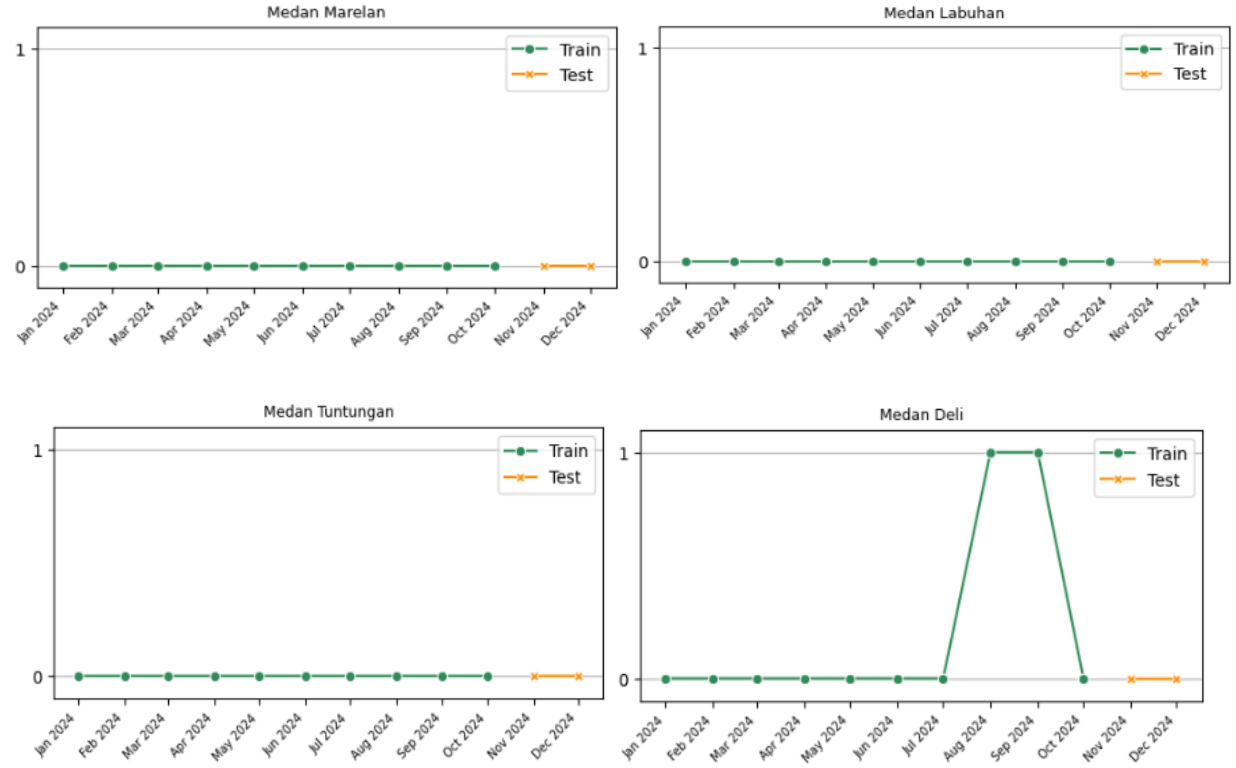


Figure 3. Predictions for Medan Marelan, Medan Labuhan, Medan Tuntungan, and Medan Deli

Similar variations were also observed in Medan Johor and Medan Selayang, which experienced fluctuations between the two risk classes. This finding shows that the model is not only capable of accurate classification, but also sensitive enough to respond to changes in area characteristics over time.

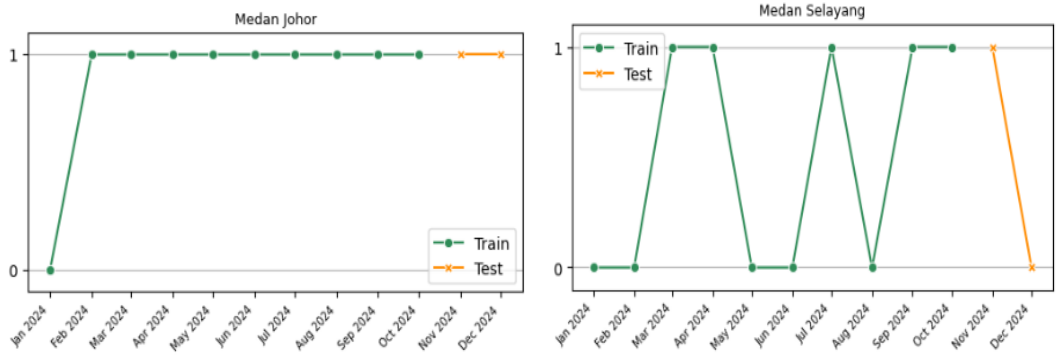


Figure 4. Medan Johor and Medan Selayang

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 4, August 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i4.8914
Hal 521-530
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

In addition to the visualization per sub-district, a recapitulation of the number of sub-districts based on the prediction results for each risk class per month was also conducted. This graph presents an aggregate view of the predicted classifications for 2024 and helps in understanding the trend of risk class dominance at the overall city level.

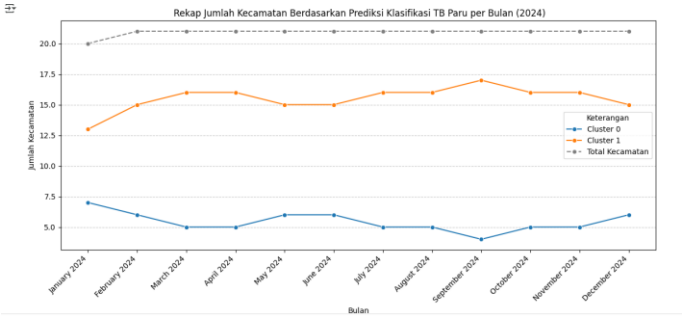


Figure 5. Prediction Recap

Based on the recapitulation graph, the majority of sub-districts each month were classified into the High Risk category (Cluster 1), with the number ranging from 14 to 17 sub-districts out of a total of 21. The highest point of High Risk classification was recorded in September 2024, with 17 sub-districts. In contrast, the number of sub-districts classified as Low Risk (Cluster 0) tends to be low and never exceeds 7 sub-districts in a month. Despite a slight increase in the Low Risk classification in June and December, the overall proportion still shows the dominance of the High Risk classification. This indicates that the model captures the distribution of features (number of patients, density, and area) that generally leads to high risk in most areas of Medan City.

Thus, this visualization reinforces the interpretation that the model not only excels in numerical metrics, but is able to provide consistent, structured, and easy-to-analyse predictions based on time and space dimensions. This makes the model potentially applicable in a dynamic monitoring system of high-risk areas of Pulmonary TB on a regular basis.

4. CONCLUSION

This study successfully built a risk classification model for Lung TB in Medan City by combining the K-Means Clustering algorithm and Logistic Regression based on three main variables: the number of Lung TB patients, population density, and area. The clustering results divided the data into two risk categories, which were then used as target labels for model training. The resulting logistic regression model showed logical and interpretable coefficients, and was able to quantitatively calculate classification probabilities. Manual calculations on a sample of sub-districts showed that the model captured the interactions between features well in distinguishing High Risk and Low Risk areas. Evaluation of the test data showed excellent performance, with 92.86% accuracy, 90.32% precision, 100% recall, and 97.90% F2-score. Visualizations of the predictions by month and by sub-district support these findings, showing a consistent classification pattern that matches the characteristics of the region. With a high sensitivity to the High Risk class, this model has the potential to be applied as a tool for mapping the risk of Pulmonary TB on a regular basis, as well as supporting decision-making in disease control efforts in densely populated areas such as Medan City. For future research, it is recommended that the model be developed by considering more supporting variables, such as environmental quality, residential density, population mobility, and access to health services. The addition of a more detailed spatial dimension (e.g. up to the kelurahan level) and the use of panel data or real-time data can also improve the precision of the model in mapping the risk of Pulmonary TB. In addition, exploration of other machine learning methods such as Random Forest, Gradient Boosting, or spatio-temporal deep learning-based models has the potential to provide higher accuracy and more adaptive predictive capabilities to changes in regional dynamics.

REFERENCES

[1] A. Alege *et al.*, “Effectiveness of Using AI-Driven Hotspot Mapping for Active Case Finding of Tuberculosis in Southwestern Nigeria.,” *Trop. Med. Infect. Dis.*, vol. 9, no. 5, 2024, doi: 10.3390/tropicalmed9050099.
[2] A. Amrin and O. Pahlevi, “Implementasi Algoritma Klasifikasi Logistic Regression dan Naïve Bayes untuk Diagnosa Penyakit Hepatitis,” *J. Tek. Komput. AMIK BSI*, vol. Volume 8, 2022.
[3] B. Aribowo and F. Salsa, *Panduan Praktis Machine Learning Klasifikasi Menggunakan Python*. Diandra Kreatif, 2024.
[4] P. Bintoro, R. Ratnasari, E. Wihardjo, I. P. Putri, and A. Asari, *Pengantar Machine Learning*. 2024.
[5] G. Dwi, “Pengaruh Faktor Lingkungan Terhadap Kejadian Tuberkulosis Paru di Wilayah Kerja Puskesmas Tanah Tumbuh,” *J. Pembang. Berkelanjutan*, vol. 7, no. 2, pp. 56–64, 2024.
[6] C. for Disease Control and Prevention, “Tuberculosis Risk Factors,” 2024.
[7] Y. H. G. *et al.*, *METODOLOGI PENELITIAN*. CV. Intelektual Manifes Media, 2023.
[8] S. I. G. I. *et al.*, *METODE PENELITIAN BIDANG ILMU INFORMATIKA (Teori & Referensi Berbasis Studi Kasus)*. PT. Sonpedia Publishing Indonesia, 2023.

[9] A. Indah and B. Hutabarat, *Tuberkulosis Paru : Faktor Penyebab & Penanggulanganya*. Media Nusa Creative (MNC Publishing), 2023.
[10] K. Jaya and Ranatwati, *KEPENDUDUKAN DAN LINGKUNGAN HIDUP*. Feniks Muda Sejahtera, 2022.
[11] N. Khairiah, S. Hajar, D. Amrizal, J. R. Izharsyah, and A. Mahardika, *Prosiding Nasional Perencanaan Pembangunan Daerah dan Kebijakan Daerah 2021*. UMSU Press, 2021.
[12] W. H. Organization, "Tuberculosis," 2024.
[13] A. Pratama, A. C. Nurcahyo, and L. Firgia, "Penerapan Machine Learning dengan Algoritma Logistik Regresi untuk Memprediksi Diabetes," *Pros. CORISINDO 2023*, 2023.
[14] M. R. Pratama and Armansyah, "Forecasting Throughput Capacity on 5Ghz Wireless Radio Network Using Linear Regression Method," *INOVTEK Polbeng - Seri Inform.*, vol. 9, no. 2, pp. 631–643, 2024, doi: 10.35314/zj7aj967.
[15] Prihandoko, R. G. G. Alam, Gunawan, and D. Abdullah, *Memahami Konsep dan Implementasi Machine Learning*. PT. Sonpedia Publishing Indonesia, 2024.
[16] S. R. M. *et al.*, *Klasifikasi Data Mining*. Serasi Media Teknologi, 2024.
[17] K. K. RI, "Pedoman Nasional Pelayanan Kedokteran Tata Laksana Tuberkulosis," 2020.
[18] J. S. *et al.*, *Buku Ajar Machine Learning*. PT. Sonpedia Publishing Indonesia, 2024.
[19] D. R. S. Saputro and Susanto, *WEKA 3.6.9 (Waikato Environment for Knowledge Analysis): Tools untuk Memahami Machine Learning*. Stiletto Book, 2023.
[20] M. Siti *et al.*, *Bahasa Pemrograman Python*. Sada Kurnia Pustaka, 2024.