SJCS Scientific Journal of Computer Science

futuristech PT. Teknologi Futuristik Indonesia

Date of publication August 10, 2025, date of current version August 10, 2025.
Digital Object Identifier ***10.64539/sjcs.v1i1.2025.32***

Article

Analysis of Suspected Factors in Tuberculosis Cases in Semarang City Using a Logistic Regression Model

Ihsan Fathoni Amri[1,*], Febrian Hikmah Nur Rohim[1], Muhammad Ivan Ardiansyah[1], Farid Sam Saputra[1], Supriyanto[2,3], Ariska Fitriyana Ningrum[1], Arman Mohammad Nakib[4]

[1] Department of Data Science, Faculty of Science and Agriculture Technology, Universitas Muhammadiyah Semarang, Semarang 50273, Indonesia; ihsanfathoni@unimus.ac.id, ariskafitriyana@unimus.ac.id
[2] Department of Chemistry Education, Faculty of Educational Sciences and Humanities, Universitas Muhammadiyah Semarang, Semarang 50273, Indonesia
[3] Yayasan Mentari Sehat Indonesia, Bangetayu Wetan 50194, Indonesia
[4] Artificial Intelligence, Nanjing University of Information Science &Technology, Nanjing, Jiangsu, China
* Correspondence

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Abstract: Tuberculosis (TB) is one of the world's deadliest infectious diseases, with Indonesia being among the countries with the highest TB burden. Semarang City, as an urban area with a dense population, faces significant challenges in controlling TB, particularly among vulnerable populations. This study identifies significant risk factors influencing TB incidence in Semarang City using a binary logistic regression model. Descriptive analysis reveals an imbalance in the data, with the majority of patients categorized as "not indicated for TB." Chi-Square tests show that variables such as shortness of breath, persistent fever for more than one month, diabetes mellitus, and household contact are significantly associated with TB incidence. The logistic regression model demonstrates overall significance (G statistic = 275.13; p-value = 1.23×10^{-55}), with shortness of breath and diabetes mellitus emerging as major risk factors based on odds ratio interpretation. However, the model's performance in detecting the "indicated for TB" category is very low (Precision 36.36%; Recall 2.05%; F1-Score 3.88%), despite an overall accuracy of 87.25%. The poor performance in the "1" category and the Pseudo R^2 value of 7% are likely related to data imbalance, where the number of cases in the "1" category is much smaller than in the "0" category, leading to bias toward the majority class. Additionally, the distribution of predictor variables that do not provide sufficient information to distinguish the "1" category from the "0" category further contributes to the model's limited ability to explain data variability overall.

Keywords: Tuberculosis; Logistic Regression; Risk Factors; Classification

1. Introduction

Tuberculosis (TB) is one of the deadliest infectious diseases in the world and has long been a global concern, including in Indonesia [1, 2]. This disease is caused by the bacterium Mycobacterium tuberculosis, which primarily attacks the lungs [3, 4]. Data from the World Health Organization (WHO) indicate that tuberculosis is one of the top ten causes of death worldwide [5, 6]. In Indonesia, tuberculosis is classified as a national priority due to its high incidence and mortality rates [7, 8].

Factors such as population density, lack of access to healthcare services, and low public awareness regarding tuberculosis prevention and treatment further exacerbate its spread [9]. Additionally, resistance to anti-tuberculosis drugs (MDR-TB) poses a major challenge in controlling this disease [10, 11]. Therefore, integrated and sustainable efforts are needed to address this issue, including public education, improvement of healthcare facilities, and further research on effective diagnostic and treatment methods [12, 13].

Semarang City, as one of the urban areas with a dense population, is indirectly facing significant challenges in the control and diagnosis of tuberculosis (TB) [14, 15]. These challenges become even more complex among vulnerable populations, such as low-income communities, individuals with HIV/AIDS, and groups with limited access to healthcare services [16].

The urban environment, characterized by overcrowding, slum settlements, inadequate sanitation, and high population mobility, contributes to the spread of tuberculosis in this region [17, 18]. Furthermore, social stigma against tuberculosis patients often prevents individuals from seeking timely diagnosis and treatment [19]. On the other hand, the capacity of healthcare services in detecting tuberculosis cases, particularly multidrug-resistant tuberculosis (MDR-TB), still needs to be improved to ensure effective control.

A comprehensive approach involving local governments, healthcare workers, and the community is necessary to address the tuberculosis challenge. Measures such as increasing access to healthcare facilities, educational programs, and the implementation of more advanced diagnostic technologies can serve as solutions to reduce the TB burden in Semarang City [19].

Data and studies on specific risk factors in Semarang City are essential to strengthen tuberculosis control efforts effectively [20]. Analyzing risk factors that contribute to the increasing prevalence of tuberculosis is a key step in early diagnosis and appropriate intervention, such as household contact tracing, physical symptoms (cough, shortness of breath), and comorbid conditions (diabetes mellitus and smoking habits), which are often associated with TB risk [21-24].

Individuals living in the same household as tuberculosis patients are at high risk of infection, especially if their living environment has poor ventilation or high occupancy density. Tuberculosis transmission occurs through airborne droplets released by infected individuals when they cough, sneeze, or speak, as these droplets contain Mycobacterium tuberculosis bacteria [25, 26]. The main symptoms of tuberculosis include a persistent cough lasting more than two weeks, often accompanied by sputum or blood [27, 28]. Additionally, other commonly observed symptoms in individuals with active tuberculosis include shortness of breath, fever, night sweats, and unexplained weight loss. Factors such as Diabetes Mellitus also increase the risk of tuberculosis by weakening the immune system, making the body more susceptible to infections [29, 30]. Furthermore, smoking habits contribute to the risk of tuberculosis, as smoking can damage the respiratory tract and reduce the lungs' ability to clear pathogens, including tuberculosis-causing bacteria [31, 32].

Logistic regression is a widely used statistical method for analyzing the relationship between independent variables, such as risk factors, and a categorical dependent variable, such as tuberculosis indication [33-35]. This method is highly effective in medical research due to its ability to capture nonlinear relationships among the studied variables, especially when categorical data [36-38], such as tuberculosis-positive or negative status, is the primary focus. Through a logistic regression approach, significant risk factors such as age, medical history, and environmental conditions can be identified. Additionally, this method allows for estimating the probability of a patient being indicated for tuberculosis based on specific characteristics [39, 40].

The advantage of logistic regression lies in its flexibility in processing various types of data, both numerical and categorical, as well as its ability to present results in an easily interpretable form, such as the odds ratio [41, 42]. The information generated from this model can serve as a reference for Mentari Sehat Indonesia in determining intervention priorities based on individual risk levels. Furthermore, logistic regression enables more accurate data-driven decision-making [43, 44], supports early diagnosis, and assists the government or health institutions in formulating evidence-based policies to control tuberculosis transmission [45, 46]. With this approach, health data management can be conducted more effectively, thereby making a tangible contribution to improving the quality of Mentari Sehat Indonesia's services.

This study aims to analyze the key risk factors influencing tuberculosis cases in Semarang City using logistic regression. It is expected to identify critical variables such as household contact with tuberculosis patients, symptoms of cough and shortness of breath, night sweats without strenuous activity, fever lasting more than one month, diabetes mellitus (DM), and smoking habits. By identifying these factors, this study seeks to determine the most significant risks affecting tuberculosis incidence in the Semarang City community. Furthermore, this research is designed to develop a predictive model capable of estimating the likelihood of an individual being indicated for tuberculosis based on specific characteristics, making it a practical and efficient screening tool.

Through this predictive model, the study results are expected to provide a meaningful contribution to healthcare services, particularly in improving diagnostic accuracy and intervention effectiveness. The findings from this research are also anticipated to offer new insights into tuberculosis transmission patterns at the local level and serve as a foundation for more targeted and evidence-based tuberculosis prevention and management efforts. Moreover, this study can serve as a reference for similar studies in other regions with different demographic and epidemiological conditions, potentially offering broader benefits at both regional and national levels. This research also supports the national agenda for tuberculosis control as part of the global commitment to the Sustainable

Development Goals (SDGs), which aim to end the global TB epidemic.

2. Research Methods

2.1. Data Source

This study utilizes data obtained from Yayasan Mentari Sehat Indonesia. The data originate from screenings conducted by healthcare cadres on suspected Tuberculosis (TB) patients and have been recorded over the past six months of 2024. In total, this study uses 5,180 samples, reflecting the current epidemiological condition of TB in Semarang City. With a substantial amount of data, this research provides a more comprehensive empirical overview of the risk factors contributing to suspected TB cases. Furthermore, the use of this data has received official authorization, ensuring compliance with research ethics standards and applicable regulations.

2.2 Research Variables

The analysis conducted in this study uses several variables that reflect risk factors related to suspected Tuberculosis cases. These variables include patient characteristics, clinical symptoms, and health conditions. To provide a clearer overview, the following summarizes the symbols, variable definitions, and data categories used, as presented in Table 1.

Table 1. Research Variables.

Symbol	Variable	Category
Y_1	Patient Type	1 = Indicated Positive for Tuberculosis 0 = Not Indicated Positive for Tuberculosis
X_1	Cough	1 = Yes 2 = No
X_2	Shortness of Breath	1 = Yes 2 = No
X_3	Night Sweats Without Activity During the Day	1 = Yes 2 = No
X_4	Fever and Chills for More Than One Month	1 = Yes 2 = No
X_5	Diabetes Mellitus	1 = Yes 2 = No
X_6	Smoker	1 = Yes 2 = No
X_7	Household Contact with Tuberculosis Patient	1 = Yes 2 = No

2.3. Binary Logistic Regression

Binary logistic regression is a statistical technique used to explore the relationship between a dependent variable with two possible outcomes (e.g., yes or no, success or failure) and one or more independent variables. This method is often applied to predict the probability of an event occurring based on influencing factors. The results obtained from binary logistic regression are typically probability values ranging from 0 to 1 [47, 48]. Additionally, a more complex version, known as polychotomous or multinomial logistic regression, is used when the dependent variable has more than two categories. In this study, the dependent variable is dichotomous (with two possible values: positive (1) or negative (0)), making binary logistic regression the chosen analytical method. The binary logistic regression model, where the response variable takes values 0 and 1 and follows a Bernoulli distribution, is defined as follows: let y_k represent the value of the response variable y for the k-th observation, then the probability function of y is given by [49]:

$$f(y_k) = \pi(x_k)^{y_k}(1 - \pi(x_k))^{1-y_k} \quad (1)$$

Where y_t= 0; 1, $\pi(x_t)$ represents the probability of success, and $1 - \pi(x_t)$ represents the probability of failure. The index t = 1,2,3 … , n denotes the observation index, with n being the total number of observations. The binary logistic regression model used is as follows [49]:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}} \quad (2)$$

Where k is the number of independent variables. The logistic regression model in the second equation is transformed using the logit function, resulting in Equation 3 [49].

$$g(x) = \log\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \quad (3)$$

2.4. G Likelihood Ratio Test for Regression Model

The G Likelihood Ratio Test is used to measure the extent to which a more complex model (full model) provides a significant improvement in goodness-of-fit compared to a simpler model (reduced model). This test is calculated based on the difference in log-likelihood values between the two models. In this study, a simultaneous test is conducted using the G-test statistic or the Likelihood Ratio Test [50, 51]. The simultaneous test is performed to determine the significance of parameters on the response variable as a whole. The G-test statistic follows a Chi-Square distribution. Mathematically, the G-test can be expressed using Equation 4 [52]:

$$G = 2\,(ln\, L_1 - ln\, L_0) \quad (4)$$

Explanation:

L_0	:	The maximum likelihood value of the function without predictor variables (reduced model).
L_1	:	The maximum likelihood value of the function with all predictor variables (full model).

Hypotheses Used:

Null Hypothesis H_0	:	$\beta 1 = \beta 2 = \cdots = \beta_k = 0$ (all predictor variables have no effect on the dependent variable).
Alternative Hypothesis H_1	:	$\beta_k \neq 0$; $k = 1, 2, \ldots, k$ (at least one predictor variable influences the dependent variable).

The G statistic follows a Chi-Square distribution with degrees of freedom p, where p is the number of predictor variables (excluding the intercept) in the model. The null hypothesis H_0 is rejected if $G > \chi^2_{kritis}$ or if $p - value < \alpha$. A larger G value indicates that a more complex model provides a better fit to the data. Conversely, a smaller G value suggests that the simpler model cannot be rejected, meaning that it is sufficient to describe the data.

2.5. Wald Test

The Wald test is a statistical method used to assess the significance of coefficients in regression models, including logistic regression. This test aims to measure the contribution of each independent variable to the model by comparing the estimated coefficient value with its standard error. The Wald test statistic is obtained by dividing the estimated coefficient (β_k) by its standard error ($SE(\hat{\beta}_k)$), resulting in a test statistic that follows a normal distribution [53, 54]. A large Wald statistic indicates that the independent variable significantly contributes to the model. In this study, the Wald test is used to evaluate the significance of each risk variable included in the logistic regression model, helping to determine which factors have the most influence on the likelihood of tuberculosis occurrence.

The partial test using the Wald statistic is applied to examine the individual effect of each parameter coefficient (β_k) on the obtained model [55, 56]. The results of the partial or individual test can be used to assess whether a predictor variable should be included in the model. Mathematically, the Wald statistic can be expressed by the Equation 5.

$$W = \frac{\hat{\beta}_k}{SE(\hat{\beta}_k)} \ \& \ SE(\hat{\beta}_k) = \sqrt{(\sigma^2(\hat{\beta}_k)} \tag{5}$$

Explanation:

$\hat{\beta}_k$	:	Estimated parameter value of the k-th predictor variable.
$SE(\hat{\beta}_k)$	:	Standard error of the estimate of the k-th predictor variable.

Hypotheses Used:

H_0	:	$\beta_k = 0$, $k = 1,2, \ldots, k$ (the k-th predictor variable has no significant effect on the response variable).
H_1	:	$\beta_k \neq 0$; $k = 1,2, \ldots, k$ (the k-th predictor variable has a significant effect on the response variable).

H_0 will be rejected if the test statistic $W > Z_{kritis}$ or if the p-value < α, indicating that the k-th predictor variable significantly affects the response variable.

2.6. Odds Ratio Analysis

The odds ratio (OR) represents the comparison between the probability of an event occurring and the probability of it not occurring. In logistic regression, OR is used to interpret the model results by measuring how much the odds of an event change based on a one-unit increase in the independent variable. Mathematically, the odds ratio can be expressed by the following equation [34]:

$$OR = \frac{(P \leq j \mid 1\,)/(P > j|1)}{(P \leq j \mid 0\,)/(P > j|0)} = \frac{\exp\,(\alpha_j + \beta_k)}{\exp\,(\alpha_j)} = \exp\,(\beta_k) \tag{6}$$

An OR value greater than 1 indicates that the tested factor increases the likelihood of the event occurring, whereas an OR value less than 1 suggests a decreasing effect on the event's probability. In this study, odds ratio analysis is used to identify significant risk factors and evaluate the extent of their influence on the likelihood of developing tuberculosis, aiding in prioritizing public health interventions.

2.7. Research Steps

The analytical steps used in this study are shown in Figure 1. According to the flowchart in Figure 1, the detailed explanation of the research steps is as follows:

- Conduct a literature review to find relevant references and sources related to the study.
- Collect data related to the research variables for further analysis.
- Perform data cleaning to ensure that the data used is free from duplication, errors, or invalid values.
- Conduct descriptive statistical analysis on pulmonary tuberculosis (TB) patient data.
- Perform independence testing between variables using the Pearson Chi-Square test.
- Use simultaneous testing with the G-test to conduct multivariate analysis on the available data.
- Apply the Wald test to partially test the hypothesis.
- Analyze the influence of research variables by calculating the odds ratio.
- Develop a predictive model using logistic regression to classify suspected tuberculosis cases.

- Draw conclusions based on the analysis results obtained throughout the study.

3. Result and Discussion

3.1. Descriptive Analysis

The first step in this study is conducting a descriptive analysis to provide a general overview of the characteristics of the data used. Bar charts facilitate the visualization of patterns for each variable in the dataset, such as the proportion of patients with certain symptoms or habits. The analysis results are presented in the form of bar charts representing the frequency distribution of each variable, including Cough, Shortness of Breath, Night Sweats Without Activity, Fever for More Than One Month, Diabetes Mellitus, Smoking, and Household Contact, as shown in the following figure.

Based on Figure 2, the descriptive analysis results for each variable can be explained as follows:

- The majority of patients in category "0" experienced a cough (positive), with 4,398 patients, while in category "1," there were 622 patients with a positive cough. Conversely, the number of patients without a cough was 134 in category "0" and only 24 in category "1."
- Most patients in category "0" did not experience shortness of breath, with a total of 4,157 patients, while in category "1," 460 patients did not have shortness of breath. Patients who experienced shortness of breath were recorded as 375 in category "0" and 186 in category "1."
- In category "0," 4,056 patients did not experience night sweats, whereas in category "1," the number was 547 patients. Meanwhile, patients who experienced night sweats were recorded as 467 in category "0" and 99 in category "1."
- The majority of patients in category "0" did not experience fever, totaling 3,737 patients, while in category "1," there were 490 patients. Patients who experienced fever were recorded as 795 in category "0" and 156 in category "1."
- Patients without diabetes mellitus dominated category "0" with 4,455 patients, while in category "1," there were 594 patients. Patients who tested positive for diabetes mellitus totaled 77 in category "0" and 52 in category "1."
- Most patients in category "0" were non-smokers, with 4,122 patients, while in category "1," there were 593 non-smokers. The number of smoking patients was 410 in category "0" and 53 in category "1."
- In category "0," patients with household contact numbered 2,272, while in category "1," there were 230 patients. For patients without household contact, the total was 2,260 in category "0" and 416 in category "1."

3.2. Independent Chin-Square Test

After conducting the descriptive analysis, the next step is to perform the Chi-Square independence test to analyze the relationship between independent variables and determine whether there is a statistically significant association between them. Table 4 presents the results of the predictor variable influence test on the response variable.

Based on the independence test results presented in Table 2, the variable Shortness of Breath has a p-value of $4.59{\times}10^{-55}$, which is smaller than the significance level α = 0.05. This indicates that Shortness of Breath has a statistically significant effect on tuberculosis occurrence and will be included in the logistic regression model analysis.

The same applies to other variables, namely Night Sweats Without Activity, Fever for More Than One Month, Diabetes Mellitus, and Household Contact, all of which have p-values < 0.05. Therefore, these variables will also be included in the logistic regression model analysis to evaluate their influence on tuberculosis occurrence.

On the other hand, the variable Cough has a p-value of 0.35430, which is greater than α = 0.05. This suggests that Cough does not have a statistically significant effect on tuberculosis occurrence and, therefore, will not be included in the logistic regression model analysis. The same applies to the Smoking variable, which also does not meet the statistical significance criteria.

3.2. Simultaneous Logistic Regression Model Test

After conducting the Chi-Square independence test to examine the relationships between independent variables, the next step is to perform a simultaneous logistic

Table 2. Independence test.

Variable	Chin^2	P-Value	Decision
Cough	0.8579	0.3543	Not Significant
Out of Breath	244.2774	$4.59{\times}10^{-55}$	Significant
Sweating at night without activity	14.1270	0.0002	Significant
Feverish fever >1 month	16.0226	$6.26{\times}10^{-5}$	Significant
Feverish fever	91.2695	$1.25{\times}10^{-21}$	Significant
Smoker	0.3948	0.5298	Not Significant
Household contact	47.2128	$6.37{\times}10^{-12}$	Significant

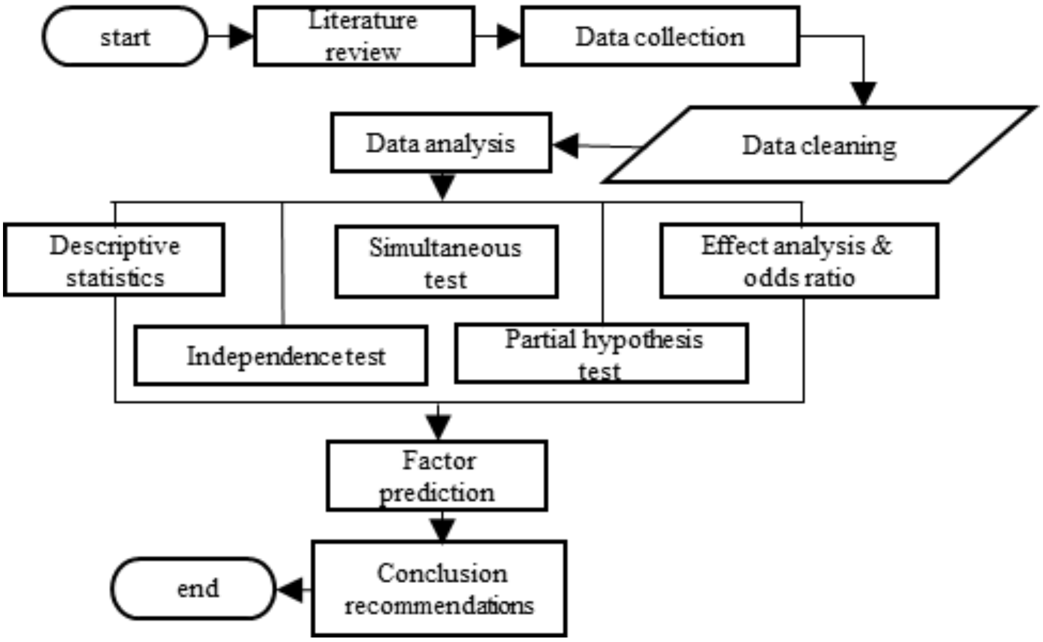

Figure 1. Flowchart of Research Steps.

Table 3. Simultaneous test.

G Statistic	P-value
275.1328	1.23×10^{-55}

regression model test. This test aims to evaluate whether the independent variables collectively have a significant influence on the dependent variable. The results are presented in the Table 3.

Based on the simultaneous test results presented in Table 3, the obtained test statistic is G = 275.1328001 with a p-value of 1.23×10^{-55}. Using a significance level of α = 5% and degrees of freedom (df) = 13, the critical value from the Chi-Square distribution table is χ critical^2 = 22.362. Since G > χ critical^2, the null hypothesis H0 is rejected. This indicates that at least one parameter βi ≠ 0, meaning that one or more predictor variables have a significant influence on the response variable in this study.

3.4. Partial Test of Predictor Variables

After conducting the simultaneous logistic regression model test to evaluate the combined influence of the variables, the next step is to perform a partial test using the Wald statistic for each predictor variable. This test aims to identify the significant contribution of each predictor variable individually to the model. The results of this partial analysis are presented in Table 4.

From the results of the partial Wald test presented in Table 4, four predictor variables significantly influence the occurrence of the disease:

- Shortness of breath with a p-value of $1.54\times10^{-34} < 0.05$
- Fever for more than one month with a p-value of $0.0208 < 0.05$
- Diabetes Mellitus (DM) with a p-value of $9.74\times10^{-11} < 0.05$
- Household contact with a p-value of $6.68\times10^{-10} < 0.05$

The Pseudo R^2 value obtained from this analysis is 0.07060, indicating that the seven predictor variables included in the logistic regression equation explain only 7% of the variability in tuberculosis (TB) incidence.

The low Pseudo R^2 value suggests that the model captures only a small portion of the data variability. This may be due to the model's tendency to focus on the majority category. Additionally, the low value could be attributed to data imbalance and the limited contribution of predictor variables in explaining the response category. Based on this analysis, the logistic regression model for pulmonary tuberculosis incidence in Semarang City is as follows:

$$\pi(x) = \frac{e^{\substack{-1.6434+(-0.4047)x_1+1.3185x_2+0.1225x_3+0.2508x_4+ \\ 1.2830x_5+(-0.1635)x_6+-0.5594x_7}}}{1+e^{\substack{-1.6434+(-0.4047)x_1+1.3185x_2+0.1225x_3+0.2508x_4+ \\ 1.2830x_5+(-0.1635)x_6+-0.5594x_7}}}$$

3.5. Interpretation of Odds Ratio Values

The odds ratio (OR) values in logistic regression analysis are obtained from exp(β) in the output of the partial test. Based on Table 4, the interpretation of the odds ratio for each predictor variable is as follows:

- Cough (OR = 0.667): Individuals experiencing cough have 0.667 times the odds of developing tuberculosis compared to those without cough. Since OR < 1, coughing tends to reduce the likelihood of tuberculosis occurrence.

- Shortness of breath (OR = 3.737): Individuals with shortness of breath have 3.737 times the odds of developing tuberculosis compared to those without shortness of breath. Since OR > 1, shortness of breath is a significant risk factor for tuberculosis.
- Night sweats without activity (OR = 1.1302): Individuals experiencing night sweats without activity have 1.1302 times the odds of developing tuberculosis compared to those without this symptom. Since OR is close to 1, this symptom has a minor effect on increasing tuberculosis risk.
- Fever for more than one week (OR = 1.284): Individuals with a fever lasting more than one week have 1.284 times the odds of developing tuberculosis compared to those without fever. Since OR > 1, prolonged fever slightly increases tuberculosis risk.
- Diabetes Mellitus (OR = 3.607): Individuals with diabetes mellitus have 3.607 times the odds of developing tuberculosis compared to those without diabetes. This indicates that diabetes is a significant risk factor for tuberculosis.
- Smoking (OR = 0.849): Smokers have 0.849 times the odds of developing tuberculosis compared to non-smokers. Since OR < 1, smoking slightly reduces tuberculosis risk in this model.
- Household contact with tuberculosis patients (OR = 0.571): Individuals in close household contact with tuberculosis patients have 0.571 times the odds of developing tuberculosis compared to those without such contact. Since OR < 1, household contact reduces the likelihood of tuberculosis in this model.

3.6. Prediction Classification and Model Evaluation

Based on the obtained model equation, the probability (prediction) values can be calculated for each data point. The data is then classified into respective groups based on a cut-off value of 0.5. From this classification, the accuracy of the model is calculated to evaluate how well the model correctly classifies the data. This accuracy is expressed as a percentage and will be presented in Table 5.

Based on Table 5, the model demonstrates strong performance in classifying Category 0, with a Precision of 87.62%, Recall of 99.49%, and an F1-Score of 93.18%. This indicates that almost all actual data belonging to Category 0 were correctly classified, with a very low error rate.

Conversely, the model performs poorly for Category 1, showing a Precision of 36.36%, Recall of only 2.05%, and F1-Score of 3.88%. This poor performance suggests that the model struggles to detect Category 1 data, as most of the Category 1 instances were misclassified as Category 0.

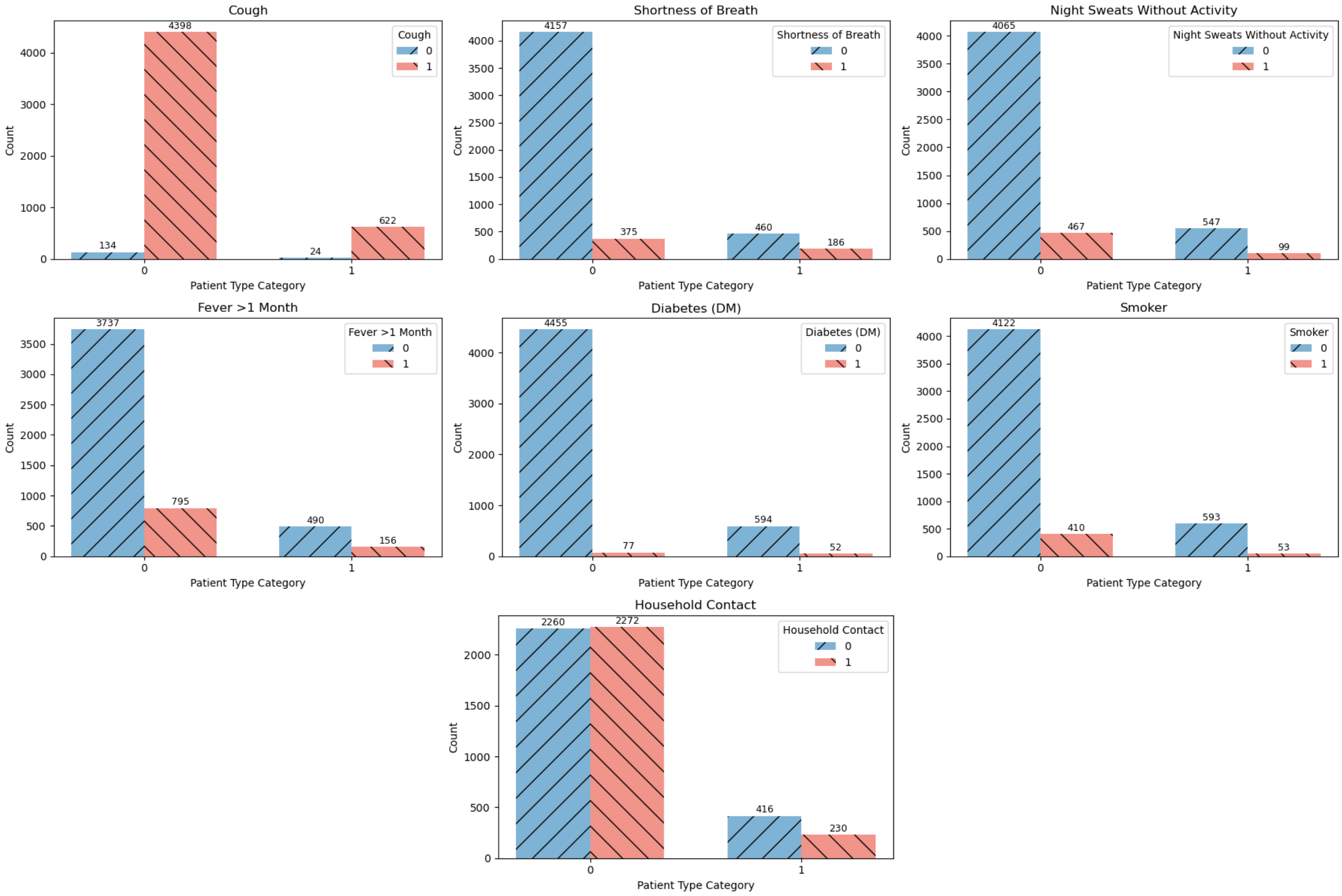

Figure 2. Descriptive Analysis of Predictor Variables.

Table 4. Partial test.

Variable	Coefficient	Standard Error (SE)	Wald Statistic	P-value	EXP(B)	Decision
Const	-1.6434	0.2393	47.1828	6.47×10^{-12}	0.1933	Significant
Cough	-0.4047	0.2342	2.9873	0.0839	0.6672	Not Significant
Out of Breath	1.3185	0.1076	150.2402	1.54×10^{-34}	3.7378	Significant
Sweating at night without activity	0.1225	0.1310	0.8741	0.3498	1.1303	Not Significant
Feverish fever >1 month	0.2508	0.1085	5.3441	0.0208	1.2850	Significant
Feverish fever	1.2830	0.1983	41.8738	9.74×10^{-11}	3.6073	Significant
Smoker	-0.1635	0.1570	1.0839	0.2978	0.8492	Not Significant
Household contact	-0.5594	0.0906	38.1133	6.68×10^{-10}	0.5716	Significant

Table 5. Classification Accuracy.

	Precision	Recall	F1-Score	Support
0	0.8762	0.9948	0.9318	1359
1	0.3636	0.0205	0.0388	195
Accuracy			0.8726	1554
Macro Avg	0.6199	0.5077	0.4853	1554
Weighted Avg	0.8119	0.8726	0.8197	1554

The low performance for Category 1 is likely due to the predictor variables lacking sufficient distinguishing power between Category 0 and Category 1, as seen in Figure 2. The distribution of predictor values in Category 1 appears similar to that of Category 0, making it difficult for the model to accurately recognize Category 1 cases, leading to low prediction accuracy for that category.

4. Conclusions

The study identifies significant risk factors influencing Tuberculosis (TB) incidence in Semarang City using a binary logistic regression model. Descriptive analysis reveals data imbalance, where the majority of patients belong to Category "0" (not indicated for TB), while Category "1" (indicated for TB) is significantly smaller.

The Chi-Square independence test indicates that variables such as Shortness of Breath, Night Sweats Without Activity, Persistent Fever (>1 month), Diabetes Mellitus (DM), and Household Contact are significantly associated with TB incidence, whereas Cough and Smoking do not show statistical significance.

The logistic regression model is statistically significant (G = 275.13, p-value = 1.23×10^{-55}), with four key variables (Shortness of Breath, Persistent Fever (>1 month), Diabetes Mellitus, and Household Contact) contributing significantly based on the Wald test. The odds ratio interpretation confirms that Shortness of Breath (OR = 3.737) and Diabetes Mellitus (OR = 3.607) are the most significant risk factors, increasing TB incidence odds by 3.737 and 3.607 times, respectively.

However, model performance evaluation shows that while Category "0" is classified very well (Precision = 87.62%, Recall = 99.49%, F1-Score = 93.18%), the performance for Category "1" is much lower (Precision = 36.36%, Recall = 2.05%, F1-Score = 3.88%), with an overall accuracy of 87.25%. The poor performance on Category "1" and the Pseudo R^2 value of 7% are likely due to data imbalance, where Category "1" has far fewer cases than Category "0", causing the model to be biased toward the majority class. Additionally, predictor variable distributions may not provide enough information to distinguish Category "1" from Category "0", limiting the model's ability to explain overall data variability. To enhance prediction accuracy and improve TB detection, additional approaches such as data balancing techniques or more complex predictive models are recommended.

5. Conflicts of Interest

The authors declare no conflicts of interest.

6. References

[1] A. Matteelli, S. Lovatti, B. Rossi, and L. Rossi, "Update on multidrug-resistant tuberculosis preventive therapy toward the global tuberculosis elimination," *International Journal of Infectious Diseases*, vol. 155, p. 107849, Jun. 2025, doi: 10.1016/j.ijid.2025.107849.

[2] L. R. Idrus, N. Fitria, F. D. Purba, J.-W. C. Alffenaar, and M. J. Postma, "Analysis of Health-Related Quality of Life and Incurred Costs Among Human Immunodeficiency Virus,

Tuberculosis, and Tuberculosis/HIV Coinfected Outpatients in Indonesia," *Value Health Reg Issues*, vol. 41, pp. 32–40, May 2024, doi: 10.1016/j.vhri.2023.10.010.

[3] S. M. Patil, A. M. Diorio, P. Kommarajula, and N. K. Kunda, "A quality-by-design strategic approach for the development of bedaquiline-pretomanid nanoparticles as inhalable dry powders for TB treatment," *Int J Pharm*, vol. 653, p. 123920, Mar. 2024, doi: 10.1016/j.ijpharm.2024.123920.

[4] S. Mandal, P. Biswas, W. Ansar, P. Mukherjee, and J. J. Jawed, "Tuberculosis of the central nervous system: Pathogenicity and molecular mechanism," in *A Review on Diverse Neurological Disorders*, Elsevier, 2024, pp. 93–102. doi: 10.1016/B978-0-323-95735-9.00050-4.

[5] A. D. Orjuela-Cañón, A. F. Romero-Gómez, A. L. Jutinico, C. E. Awad, E. Vergara, and M. A. Palencia, "Data Fusion of Medical Records and Clinical Data to Enhance Tuberculosis Diagnosis in Resource-Limited Settings," *Applied Sciences*, vol. 15, no. 10, p. 5423, May 2025, doi: 10.3390/app15105423.

[6] K. Bhattacharyya, R. P. Jha, D. Dhamnetiya, P. Patel, N. Shri, and M. Singh, "Exploring secular trends and types of tuberculosis burden in India over past three decades through insights from the Global Burden of Disease Study 2019," *Discover Public Health*, vol. 22, no. 1, p. 439, Jul. 2025, doi: 10.1186/s12982-025-00772-7.

[7] Y. Penyami, M. P. Angkasa, and S. Sumarni, "Using storybooks to enhance health awareness among schoolchildren at risk of tuberculosis," *Malahayati International Journal of Nursing and Health Science*, vol. 7, no. 11, pp. 1338–1343, Feb. 2025, doi: 10.33024/minh.v7i11.567.

[8] R. B. Fanda, A. Probandari, M. O. Kok, and R. A. Bal, "Managing medicines in decentralization: discrepancies between national policies and local practices in primary healthcare settings in Indonesia," *Health Policy Plan*, vol. 40, no. 3, pp. 346–357, Mar. 2025, doi: 10.1093/heapol/czae114.

[9] I. Kumalasari, "Analysis of Risk Factors Associated with Pulmonary Tuberculosis Incidence in Islamic Boarding Schools," *BALABA*, vol. 20, no. 2, pp. 85–95, 2024.

[10] V. Srivastava and A. Verma, "Current Challenges in the Management of Tuberculosis," *Journal of Young Pharmacists*, vol. 16, no. 2, pp. 145–154, Jun. 2024, doi: 10.5530/jyp.2024.16.21.

[11] M. J. Nasiri, K. Lutfy, and V. Venketaraman, "Challenges of Multidrug-Resistant Tuberculosis Meningitis: Current Treatments and the Role of Glutathione as an Adjunct Therapy," *Vaccines (Basel)*, vol. 12, no. 12, p. 1397, Dec. 2024, doi: 10.3390/vaccines12121397.

[12] Md. Faiyazuddin *et al.*, "The Impact of Artificial Intelligence on Healthcare: A Comprehensive Review of Advancements in Diagnostics, Treatment, and Operational Efficiency," *Health Sci Rep*, vol. 8, no. 1, Jan. 2025, doi: 10.1002/hsr2.70312.

[13] Aliu Olalekan Olatunji, Janet Aderonke Olaboye, Chukwudi Cosmos Maha, Tolulope Olagoke Kolawole, and Samira Abdul, "Revolutionizing infectious disease management in low-resource settings: The impact of rapid diagnostic technologies and portable devices," *International Journal of Applied Research in Social Sciences*, vol. 6, no. 7, pp. 1417–1432, Jul. 2024, doi: 10.51594/ijarss.v6i7.1332.

[14] S. Handayani and S. Isworo, "Evaluation of Tuberculosis program implementation in Primary Health Care, Semarang, Indonesia," *International Journal of Public Health Asia Pacific*, pp. 1–11, Jun. 2024, doi: 10.62992/qb8eay62.

[15] V. R. Aditya, M. Raharjo, and O. Setiani, "Analysis of the Quality of the Physical Environment of the House on the Incidence of Tuberculosis in Tembalang Subdistrict," *Jurnal Penelitian Pendidikan IPA*, vol. 11, no. 5, pp. 677–683, May 2025, doi: 10.29303/jppipa.v11i5.11393.

[16] A. Natalis, "Power, Law, and the Semiotics of Marginalisation: Rethinking Prostitution, Health Risk, and Legal Discourse in Indonesia," *Int J Semiot Law*, Jul. 2025, doi: 10.1007/s11196-025-10310-y.

[17] S. Shafique *et al.*, "Effective community-based interventions to prevent and control infectious diseases in urban informal settlements in low- and middle-income countries: a systematic review," *Syst Rev*, vol. 13, no. 1, p. 253, Oct. 2024, doi: 10.1186/s13643-024-02651-9.

[18] S. N. Ogbonna, C. N. Ochie, and E. C. Aniwada, "Urban slum housing quality, and its public health implications in Nigeria: a case of urban slum residents in Enugu metropolis, South East, Nigeria," *BMC Public Health*, vol. 24, no. 1, p. 3231, Nov. 2024, doi: 10.1186/s12889-024-20764-7.

[19] R. A. Rahmadani, A. A. Sainal, and S. Suprapto, "Community Empowerment to Increase Knowledge About Tuberculosis," *Abdimas Polsaka: Jurnal Pengabdian Masyarakat*, vol. 2, no. 2, pp. 117–123, 2023.

[20] F. Fahdhienie, M. Mudatsir, T. F. Abidin, and N. Nurjannah, "Risk factors of pulmonary tuberculosis in Indonesia: A case-control study in a high disease prevalence region," *Narra J*, vol. 4, no. 2, p. e943, Aug. 2024, doi: 10.52225/narra.v4i2.943.

[21] S. Das *et al.*, "Prevalence, risk factors, and comorbidities of type 2 diabetes among COPD patients at a Bhubaneswar secondary care hospital," *Int J Diabetes Dev Ctries*, Oct. 2024, doi: 10.1007/s13410-024-01404-7.

[22] M. Fayaz, S. A. Zakki, I. U. Haq, M. Afzal, M. Latif, and E. Altaf, "Evaluation of health-related quality of life among patients with chronic obstructive pulmonary disease at District Headquarter Hospital haripur, Pakistan," *Clin Epidemiol Glob Health*, vol. 32, p. 101917, Mar. 2025, doi: 10.1016/j.cegh.2025.101917.

[23] M. S. Bah *et al.*, "Assessment of comorbidities, risk factors, and post tuberculosis lung disease in National Tuberculosis Guidelines: A scoping review," *PLOS Global Public Health*, vol. 5, no. 7, p. e0004935, Jul. 2025, doi: 10.1371/journal.pgph.0004935.

[24] J. Yayan, K.-J. Franke, M. Berger, W. Windisch, and K. Rasche, "Early detection of tuberculosis: a systematic review," *Pneumonia*, vol. 16, no. 1, p. 11, Jul. 2024, doi: 10.1186/s41479-024-00133-z.

[25] M. Coleman, L. Martinez, G. Theron, R. Wood, and B. Marais, "Mycobacterium tuberculosis Transmission in High-Incidence Settings—New Paradigms and Insights," *Pathogens*, vol. 11, no. 11, p. 1228, Oct. 2022, doi: 10.3390/pathogens11111228.

[26] R. Long, M. Divangahi, and K. Schwartzman, "Chapter 2: Transmission and pathogenesis of tuberculosis," *Canadian Journal of Respiratory, Critical Care, and Sleep Medicine*, vol. 6, no. sup1, pp. 22–32, Mar. 2022, doi: 10.1080/24745332.2022.2035540.

[27] E. Garianto *et al.*, "Rifampicin mono resistant tuberculosis (RR-TB): a case report," *Surabaya Medical Journal*, pp. 48–57, May 2024, doi: 10.59747/smjidisurabaya.v2i1.38.

[28] N. Funaguchi *et al.*, "Respiratory/Infection Symptoms," in *Internal Medicine for Dental Treatments*, Singapore: Springer Nature Singapore, 2023, pp. 3–11. doi: 10.1007/978-981-99-3296-2_1.

[29] S. E. Barry, A. Sawka, A. Maldari, J. Inauen, S. LaBroome, and J. B. Geake, "Macrophage Dysfunction in Tuberculosis–Diabetes Mellitus Comorbidity: A Scoping Review of Immune Dysregulation and Disease Progression," *Diabetology*, vol. 6, no. 5, p. 35, May 2025, doi: 10.3390/diabetology6050035.

[30] Z. Ye *et al.*, "Impact of diabetes mellitus on tuberculosis prevention, diagnosis, and treatment from an immunologic perspective," *Exploration*, vol. 4, no. 5, Oct. 2024, doi: 10.1002/EXP.20230138.

[31] Y. Hamada *et al.*, "Tobacco smoking clusters in households affected by tuberculosis in an individual participant data meta-analysis of national tuberculosis prevalence surveys: Time for household-wide interventions?," *PLOS Global Public Health*, vol. 4, no. 2, p. e0002596, Feb. 2024, doi: 10.1371/journal.pgph.0002596.

[32] C. Feldman, A. J. Theron, M. C. Cholo, and R. Anderson, "Cigarette Smoking as a Risk Factor for Tuberculosis in Adults: Epidemiology and Aspects of Disease Pathogenesis," *Pathogens*, vol. 13, no. 2, p. 151, Feb. 2024, doi: 10.3390/pathogens13020151.

[33] M. Abbasian, H. Sadeghi-bazargani, H. Matlabi, N. Havaei, M. Hashemiparast, and H. Allahverdipour, "Factors Affecting Home Injuries in Older Adults: An Analysis Using Binary Logistic Regression," *Health Sci Rep*, vol. 8, no. 7, Jul. 2025, doi: 10.1002/hsr2.71055.

[34] E. O'Shaughnessy, E. Detrinidad, P. Soyer, and A. Lecler, "An introductory guide to statistics for the radiologist," *Diagn Interv Imaging*, vol. 106, no. 2, pp. 49–52, Feb. 2025, doi: 10.1016/j.diii.2024.11.003.

[35] D. Dey *et al.*, "The proper application of logistic regression model in complex survey data: a systematic review," *BMC Med Res Methodol*, vol. 25, no. 1, p. 15, Jan. 2025, doi: 10.1186/s12874-024-02454-5.

[36] Y. Takefuji, "Limitations of logistic regression in analyzing complex ambulatory blood pressure data: a call for non-parametric approaches," *Eur Heart J*, Jul. 2025, doi: 10.1093/eurheartj/ehaf541.

[37] Y. Hua, T. S. Stead, A. George, and L. Ganti, "Clinical Risk Prediction with Logistic Regression: Best Practices, Validation Techniques, and Applications in Medical Research," *Academic Medicine & Surgery*, Mar. 2025, doi: 10.62186/001c.131964.

[38] Q.-Y. Chen, S.-M. Yin, M.-M. Shao, F.-S. Yi, and H.-Z. Shi, "Machine learning-based Diagnostic model for determining the etiology of pleural effusion using Age, ADA and LDH," *Respir Res*, vol. 26, no. 1, p. 170, May 2025, doi: 10.1186/s12931-025-03253-2.

[39] L. M. Faye, C. Magwaza, N. Dlatu, and T. Apalata, "Exploring Determinants and Predictive Models of Latent Tuberculosis Infection Outcomes in Rural Areas of the Eastern Cape: A Pilot Comparative Analysis of Logistic Regression and Machine Learning Approaches," *Information*, vol. 16, no. 3, p. 239, Mar. 2025, doi: 10.3390/info16030239.

[40] A. K. Tiwari and A. Katiyar, "Tuberculosis Disease Detection: Comparative Analysis of Logistic Regression and Decision Tree Models for Predicting TB Positivity Using Demographic and Symptom Data," in *Proceedings of Fourth International Conference on Computing and Communication Networks*, 2025, pp. 359–373. doi: 10.1007/978-981-96-3250-3_29.

[41] S. Rydzi, B. Zahradnikova, Z. Sutova, M. Ravas, D. Hornacek, and P. Tanuska, "A Predictive Quality Inspection Framework for the Manufacturing Process in the Context of Industry 4.0," *Sensors*, vol. 24, no. 17, p. 5644, Aug. 2024, doi: 10.3390/s24175644.

[42] S. Kruschel, N. Hambauer, S. Weinzierl, S. Zilker, M. Kraus, and P. Zschech, "Challenging the Performance-Interpretability Trade-Off: An Evaluation of Interpretable Machine Learning Models," *Business & Information Systems Engineering*, Feb. 2025, doi: 10.1007/s12599-024-00922-2.

[43] Y. Cai, B. de Jonge, and R. H. Teunter, "Data-driven condition-based maintenance optimization given limited data," *Eur J Oper Res*, vol. 324, no. 1, pp. 324–334, Jul. 2025, doi: 10.1016/j.ejor.2025.01.010.

[44] N. Zhang *et al.*, "A data-driven methodology for fragility assessment of hang-off deepwater drilling risers under emergency evacuation conditions," *Ocean Engineering*, vol. 315, p. 119777, Jan. 2025, doi: 10.1016/j.oceaneng.2024.119777.

[45] A. K. Sah *et al.*, "Role of Artificial Intelligence and Personalized Medicine in Enhancing HIV Management and Treatment Outcomes," *Life*, vol. 15, no. 5, p. 745, May 2025, doi: 10.3390/life15050745.

[46] N. Nuha, S. Ali Pitchay, A. H. Ab Halim, M. A. Bin Sahbudin, and I. Sahbudin, "Beyond the outbreak: a review of big data analytics in proactive infectious disease prevention for risk mitigation for COVID-19," *J Big Data*, vol. 12, no. 1, p. 185, Jul. 2025, doi: 10.1186/s40537-025-01245-z.

[47] K. Getu and H. Gangadhara Bhat, "Application of geospatial techniques and binary logistic regression model for analyzing driving factors of urban growth in Bahir Dar city, Ethiopia," *Heliyon*, vol. 10, no. 3, p. e25137, Feb. 2024, doi: 10.1016/j.heliyon.2024.e25137.

[48] J. R. Wilson, K. A. Lorenz, and L. P. Selby, "Standard Binary Logistic Regression Model," in *Modeling Binary Correlated Responses*, 2024, pp. 27–59. doi: 10.1007/978-3-031-62427-8_3.

[49] D. Kartikasari, "Analisis Faktor-Faktor yang Mempengaruhi Level Polusi Udara dengan Metode Regresi Logistik Biner," *Mathunesa: Jurnal Ilmiah Matematika*, vol. 8, no. 1, pp. 55–59, 2020.

[50] G. Sastro, A. Syafiih, and Ilmadi, "Binary Logistic Regression Model of Parental Interest in Islamic Boarding Schools with R Program: A Case Study Islamic Boarding Schools Tahfidz Daarul Qur'an Tangerang," *Ceddi Journal of Education*, vol. 3, no. 1, pp. 8–15, Jun. 2024, doi: 10.56134/cje.v3i1.91.

[51] Ni Made Deviani Prisilia, Adelia Yuniarti, Citra Annisa Rahmania, Made Ayu Asri Oktarini Putri, and Made Susilawati, "Factors That Influence Diabetes Disease," *International Journal of Applied Mathematics and Computing*, vol. 1, no. 3, pp. 31–40, Oct. 2024, doi: 10.62951/ijamc.v1i3.27.

[52] O. Haloho, P. Sembiring, and A. Manurung, "Penerapan Analisis Regresi Logistik Pada Pemakaian Alat Kontrasepsi Wanita (Studi Kasus di desa Dolok Mariah Kabupaten Simalungun)," 2013.

[53] M. P. Woller and C. K. Enders, "Exploration of the MCMC Wald test with linear regression," *Behav Res Methods*, vol. 56, no. 7, pp. 7391–7409, Jun. 2024, doi: 10.3758/s13428-024-02426-z.

[54] F. Sarto, S. Saggese, E. Carbone, and P. Sarnacchiaro, "Integrating SEM, Wald test and ANOM to disentangle the effect of TMT functional background on strategic plans," *Socioecon Plann Sci*, vol. 96, p. 102083, Dec. 2024, doi: 10.1016/j.seps.2024.102083.

[55] H. Hasim *et al.*, "Employing Binary Logistic Regression in Modeling the Effectiveness of Agricultural Extension in Clove Farming: Facts and Findings from Sidrap Regency, Indonesia," *Sustainability*, vol. 17, no. 6, p. 2786, Mar. 2025, doi: 10.3390/su17062786.

[56] E. Kosasih, N. K. W. Asmara Santhi, N. W. A. Febriyanti, E. V. Br Barus, and M. Susilawati, "Identification of Risk Factors for Chronic Kidney Disease Using Binary Logistic Regression," *International Journal of Applied Mathematics and Computing*, vol. 2, no. 3, pp. 09–17, Jul. 2025, doi: 10.62951/ijamc.v2i3.222.