ENHANCED KNOWLEDGE IN SCIENCES AND TECHNOLOGY
e-ISSN: 2773-6385

EKST

Vol. 4 No. 2 (2024) 191-200
https://publisher.uthm.edu.my/periodicals/index.php/ekst

Prediction of Tuberculosis Prevalence in Kelantan, Malaysia: Comparison of Model Prediction based on Lasso and Elastic Net Regression

Christy Vun Wan Sze[1], Khuneswari Gopal Pillay[1]*

[1] *Department of Mathematics and Statistics, Faculty of Applied Sciences and Technology, UTHM Kampus Cawangan Pagoh, Hab Pendidikan Tinggi Pagoh, KM 1, Jalan Panchor, 84600 Pagoh, Muar, Johor, MALAYSIA*

*Corresponding Author: khuneswari@uthm.edu.my
DOI: https://doi.org/10.30880/ekst.2024.04.02.023

Article Info

Received: 27 December 2023
Accepted: 11 January 2024
Available online: 12 December 2024

Keywords

Tuberculosis (TB), LASSO Regression, Elastic Net Regression, *K-F*old Cross-Validation

Abstract

Tuberculosis (TB), caused by the *bacillus Mycobacterium tuberculosis*, is a leading cause of death, with an untreated mortality rate of approximately 50%. This study aims to identify factors influencing the modelling prediction of TB prevalence in Kelantan, Malaysia, using LASSO and Elastic Net regression. The objectives included selecting the best model through model selection criteria and subsequently applying the chosen model to predict TB prevalence in Kelantan and utilising cross-sectional data from TB patients in Kelantan, Malaysia, collected from individuals who underwent TB screening between 2019 and 2020. The results indicate that variables including low TB incidence areas (-0.846), primary education (0.848), ex-smokers (0.482), passive smokers (-0.560), BCG vaccination (1.649), night sweats (3.375), weight loss (1.859), and loss of appetite (0.142) significantly influence the TB prevalence in logistic LASSO. Comparison of AICc and BIC indicate the logistic LASSO regression (-76.665 and -54.037 respectively) outperformed than logistic Elastic Net regression (-72.319 and -44.357 respectively) for TB prevalence prediction. The goodness of fit test further proved the model's effectiveness of logistic LASSO regression in predicting TB prevalence. The model predicts varying probabilities for TB prevalence across different scenarios and conditions. In future research, it is essential to collaborate with healthcare institutions for comprehensive medical data and explore alternative methodologies to contribute more impactful insights into predicting TB prevalence.

1. Introduction

Tuberculosis (TB), caused by the *bacillus Mycobacterium tuberculosis*, primarily affects the lungs before potentially spreading to other organs [1]. Common symptoms include coughing, chest pain, weight loss, fever, and night sweats. Globally, TB has a substantial impact, resulting in millions of deaths annually. According to WHO's Global Tuberculosis Report 2022, the death rate of Tuberculosis is about 50% without any treatment, which represents a remarkably high number. In 2021, the estimated global death toll from TB reached 1.6 million, with significant prevalence in regions like Africa and Southeast Asia [2].

The COVID-19 pandemic has further complicated TB management and monitoring, with similar symptoms leading to challenges in distinguishing between the two diseases. As stated by WHO's Global Tuberculosis Report 2022, the number of reported cases of Tuberculosis has dramatically dropped from 7.1 million in year 2019 to 5.8

million in year 2020 [2]. However, the decrease in reported Tuberculosis cases does not mean the effort to end TB has improved, but it suggests that the number of undiagnosed and untreated Tuberculosis patients has increased. Due to the lockdown during the COVID-19 pandemic, it is hard for people to get treatment and diagnosis.

While Malaysia isn't among the top countries for TB cases, it reported 21,727 cases in 2021, showing an 8.1% decline from the previous year [3]. Despite this decline, Malaysia is committed to becoming TB-free by 2035. The migrant workforce, constituting 13% of reported cases, poses an unpredictable challenge. The origin countries of these migrants often have high TB incidence rates, contributing to the disease's spread in Malaysia [4]. Efforts are being made to understand and control TB clusters and risk groups, as evidenced by a 2020 study in Sabah, Malaysia, which highlighted the importance of early diagnosis and control measures.

Regression analysis, particularly LASSO (Least Absolute Shrinkage and Selection Operators), plays a crucial role in understanding the relationships between variables. LASSO addresses multicollinearity by introducing a penalty term that can shrink coefficients to zero, aiding in variable selection [5]. It has found success in diverse applications, from public health datasets to predicting the existence of breast cancer [6]. However, biases may emerge, and there is a risk of information loss in specific conditions.

Elastic net regression, introduced to overcome limitations in Ridge and LASSO methods, combines L1 and L2 penalties, providing both shrinkage and variable selection. Acting like a "stretchable fishing net," it sets some coefficients to zero while reducing others [7]. The versatility of the elastic net is evident in various fields, from dealing with multicollinearity in data to predicting vitamin D deficiency [8].

This study aimed to identify factors influencing the modelling prediction of TB prevalence in Kelantan, Malaysia, using logistic LASSO and Elastic Net regression. The objectives included selecting the best model through model selection criteria and subsequently applying the chosen model to predict TB prevalence in Kelantan. This study is limited to the TB patients in Kelantan, Malaysia, who had a TB screening between 2019 and 2020.

2. Materials and Methods

Table 1 The description of the variables of Tuberculosis patients

Variable	Description	Variable	Description
Dependent, y		Smoking history	
Diagnosed	0 = No, 1 = Yes	Smoking status	0 = Active smoker,
Sociodemographic			1 = Ex-smoker,
Age	Year		2 = Passive smoker
Gender	0 = Male, 1 = Female	Tuberculosis History	
Marital Status	0 = Never married,	BCG vaccination	0 = No, 1 = Yes
	1 = Married,	Exposed to the TB index case	0 = No, 1 = Yes
	2 = Divorced/widowed		
Crowded home	0 = No, 1 = Yes	History of TB disease	0 = No, 1 = Yes
Resided in high TB incidence area	0 = No, 1 = Yes		
		Tuberculosis symptom	
Work in a non-medical field	0 = No, 1 = Yes	Cough	0 = No, 1 = Yes
		any sputum	0 = No, 1 = Yes
Education status	0 = Primary school level,	weight loss	0 = No, 1 = Yes
	1 = Secondary school level,	night sweat	0 = No, 1 = Yes
	2 = Tertiary level/College/University	chest pain	0 = No, 1 = Yes
Medical history		loss of appetite	0 = No, 1 = Yes
Diabetic patient	0 = No, 1 = Yes	Fever	0 = No, 1 = Yes
HIV patient	0 = No, 1 = Yes	Haemoptysis	0 = No, 1 = Yes
Immunosuppressed	0 = No, 1 = Yes		

The dataset that was used in this study is obtained from the Zenodo website at the link given https://zenodo.org/record/6349779. The dataset that is obtained consists of the patients' details who came for Tuberculosis screening between 2019 and 2020 [9]. Information such as sociodemographic, medical history,

Enhanced Knowledge in Sciences and Technology Vol. 4 No. 2 (2024) p. 191-200

smoking history, Tuberculosis history, and symptom history was gathered for the research used. A total of 159 patients' data were selected for the analysis process. Table 1 indicates the details and the description of the variables of the information of Tuberculosis patients.

2.1 Data Pre-processing

To enhance the accuracy of predictions and address issues like noise, missing, or inconsistent data, thorough data pre-processing is crucial. This involves eliminating irrelevant variability and preparing the data for effective modelling. Various methods, such as complete case analysis, missing indicators, and multiple imputations, are employed to manage missing data. In the case of the Tuberculosis prevalence prediction study, certain variables are unrelated to the prediction and will be excluded. Additionally, due to substantial missing values, considerations will be made to either eliminate the variable or the corresponding column. Since the dataset comprises binary and categorical variables, label encoding will be applied to convert them into numerical values, streamlining the analysis and ensuring compatibility with the chosen method.

2.2 Cross Tabulation

Cross-tabulation, or a contingency table, is a method displaying the frequencies of multiple categorical variables, facilitating the exploration of relationships between them [10]. This technique proves valuable in uncovering trends and dependencies among variables, such as gender and smoking status. The constructed table summarizes categorical data, offering insights for various statistical tests, often utilizing the chi-squared distribution. In this study, cross-tabulation will be employed to unveil correlations between sociodemographic, medical history, smoking history, Tuberculosis history, and symptom history variables, providing a comprehensive understanding of their relationships.

2.3 Chi-square Test and Fisher's Exact Test

The Chi-square test is a statistical method utilized to assess the association between two categorical variables [11]. Before the test, cross-tabulation displays observation frequencies for variable combinations, facilitating subsequent chi-square statistic computation using the formula in equation (1) [12].

$$\chi^2 = \frac{\sum (O-E)^2}{E} \tag{1}$$

Where χ^2 represents the test statistic, O is the observed value and E is the expected value.

The Chi-square test suggests the null hypothesis that there is no association between the categorical variables, with the alternative hypothesis suggesting the presence of an association. H_0 is rejected if the chi-square statistic exceeds the critical value or if the *p*-value is less than the significance value, indicating a statistically significant association between variables [11].

Notably, meticulous examination of the contingency table is essential. If any cell has a frequency of five or less, an adjustment for the *p*-value is necessary, achievable through Fisher's exact test as shown in equation (2) [13]. Both tests share a similar interpretation of *p*-values, where variables with values less than 0.05 as statistically significant.

$$p = \frac{(a+b)!(c+d)!(a+c)!(b+d)}{(a!b!c!d!n!)} \tag{2}$$

Where *a, b, c, d* are observations in respective cells, *n* is the total number of observations and '!' denotes factorial.

2.4 *K*-fold Cross-Validation

K-fold cross-validation emerges as a widely adopted method for model evaluation, particularly in classification and regression tasks, effectively mitigating overfitting concerns [14]. By partitioning the dataset into *K* groups and designating each as a validation set, *K*-fold cross-validation facilitates robust model assessment. In the context of LASSO and Elastic Net Regression, where the regularization parameter (lambda) and mixing parameter (alpha) play pivotal roles, *K*-fold cross-validation becomes instrumental. It systematically explores different parameter combinations, aiding in the optimization of hyperparameters. The process involves dividing the data into equal-sized subsets, with *K*=10 chosen for this study. Subsequently, the model is trained on *K*-1 folds, and the remaining fold serves as the validation set, allowing the determination of optimal lambda by minimizing mean squared error or maximizing *R*-squared value.

2.5 Logistic Regression

Logistic regression is a statistical modelling technique extensively used in medical research and healthcare datasets, especially for exploring binary outcomes like disease presence or absence [15]. In this study, the binary outcome variable "diagnosed" (0 for "No" and 1 for "Yes") is examined about independent variables. The logistic regression model estimates coefficients using the log odds of probability, then further clarifies the transformation of the binary outcome into probabilities, providing a robust framework for assessing the associations between variables.

2.6 LASSO Regression

LASSO is a regression method crucial in statistical analysis to address overfitting. It performs shrinkage and variable selection by constraining model parameters, shrinking regression coefficients toward zero or ensuring their total absolute value is less than a predefined λ value [5]. Cross-validation, involving dataset division into k sub-samples or "folds," is employed to determine optimal λ values. The LASSO model is fitted iteratively during cross-validation, and its performance is assessed using metrics like mean squared error. A regularization path visualizes coefficient changes as λ varies, aiding in selecting the optimal λ associated with the best performance metric. The logistic LASSO regression, utilized in this research, is implemented through the "*glmnet*" package in *R*. The estimation of the logistic LASSO regression model is defined by equation (3) [16].

$$\hat{\beta} = \arg\min\left\{\frac{1}{2n}\|y - x\beta\|_2^2 + \lambda\|\beta\|_1\right\} \tag{3}$$

Where $\hat{\beta}$ represents the estimated coefficients, n denotes sample sizes, y represents the dependent variable, X represents the independent variables and λ becomes the tuning parameter.

2.7 Elastic Net Regression

Elastic Net regression combines both L1 and L2 penalties in the regression model, integrating aspects of both LASSO and Ridge regression techniques. This approach results in a reduction of coefficients like Ridge regression and the selection of variables like LASSO. The logistic Elastic Net model is implemented using the "*glmnet*" package in *R*, like the logistic LASSO model. The *K*-fold cross-validation is also applied, and a grid for both λ and α is defined to find optimal parameters, ensuring effective model performance. The regularization path aids in selecting the optimal parameters, optimizing the trade-off between regularization strength and variable selection. The model estimation is defined by equation (4) [16].

$$\hat{\beta} = \left(1 + \frac{\lambda_2}{n}\right)\left\{\arg\min_{\beta}\|y - X\beta\|_2^2 + \lambda_2\|\beta\|_2^2 + \lambda_1\|\beta\|_1\right\} \tag{4}$$

Which $\hat{\beta}$ represents the estimated coefficients, n is sample sizes, y is the dependent variable, X is the independent variables, and λ is the tuning parameter.

2.8 Model selection criteria

After constructing logistic LASSO and Elastic Net regression models, the choice between them is determined using model selection criteria. Two criteria, Akaike Information Criterion corrected (AICc) and Bayesian Information Criterion (BIC), are employed for this purpose, considering both sample size and the number of parameters in the model. AICc, an extension of the maximum likelihood principle, addresses estimation bias for small sample sizes, making it suitable for this study. The AICc formula equation (5) incorporates a penalty for parameters and an additional term for model complexity [17].

$$AICc = -2\log\ell\left(\hat{\theta}\right) + 2p + \frac{2p(p+1)}{n-p-1} \tag{5}$$

Where $\log\ell\left(\hat{\theta}\right)$ represents log-likelihood, p denotes the number of parameters and n is the sample size.

Bayesian Information Criterion (BIC) (equation (6)), introduced by Schwarz in 1978, competes with AIC and AICc for model selection [18]. The model with the lowest AICc and BIC is chosen, where BIC introduces a penalty based on parameters and sample size. The comparison of AICc and BIC aids in selecting the optimal model between logistic LASSO and Elastic Net regression [17].

Enhanced Knowledge in Sciences and Technology Vol. 4 No. 2 (2024) p. 191-200

$$BIC = -2\log \ell\left(\hat{\theta}\right) + p\ln\left(n\right) \tag{6}$$

Where $\log \ell\left(\hat{\theta}\right)$ represents log-likelihood, p is the number of parameters and n is the sample size.

2.9 Goodness-of-fit test

The model's goodness of fit is assessed through a deviance goodness-of-fit test. This test compares the deviance of the fitted logistic regression model to a saturated model perfectly fitting the data. The lower the deviance, the better the model's fit. The deviance is calculated using the formula shown in equation (7) [19].

$$2*\left[LL_{\delta} - LL_{f}\right] \tag{7}$$

Where LL_{δ} represents the log-likelihood of the fitted model and LL_f becomes the log-likelihood of the saturated model.

A chi-square test is then conducted on the difference in deviance, with a p-value > 0.05 indicating an acceptable and well-fitted model. If not, the model iteration process continues until an optimal model is obtained. Additionally, examining the deviance residual plot provides insights into the model's fit, with patterns indicating consistency in predictive performance, such as the presence of parallel lines.

2.10 Model Prediction

In public health, predictive models guide preventive interventions for individuals at elevated disease risk. These models also estimate the likelihood of disease presence, signaling treatment necessity for high probabilities and suggesting further diagnostic testing for low probabilities [20]. After passing the deviance goodness-of-fit test, the chosen model proceeds to prediction using new data. The model assigns probabilities to individuals or scenarios, crucial for decision-making. These probabilities, ranging from zero to one, provide actionable insights, directing appropriate actions based on perceived outcome likelihoods.

3. Results and Discussion

3.1 Data pre-processing

Data pre-processing aimed to enhance dataset quality. Firstly, the binning process categorized the continuous "Age" variable into six age groups which are <30, <40, <50, <60, <70, and 70+, simplifying data complexity. Secondly, addressing missing data, the variable "Name of immunosuppressed" with significant missing values was removed to improve data accuracy. Additionally, non-relevant variables, such as Index No, New Education status, Monthly Income, Income group, Alcohol, Smoke Exposure, and Duration symptoms were deleted to reduce bias and focus on key research objectives.

3.2 Cross tabulation, Chi-square Test and Fisher's Exact Test

To illustrate the distribution of frequencies for the dependent and significant independent variables, cross-tabulation was conducted on categorical variables. The resulting tables, generated in *R* software, reveal the relationships between the "Diagnosed" variable and the relevant independent variables. These tables have provided insights into the associations. Subsequently, the Chi-square test or Fisher's Exact test will be employed to assess the statistical significance of these associations. The *sjPlot* package in *R* was instrumental in this process, allowing for the visualization of cross-tabulated frequencies and simultaneous application of the Chi-square test or Fisher's exact test.

In Table 2, the chi-square test was prioritized over Fisher's exact test due to larger sample sizes or adherence to chi-square assumptions. Variables with p-values less than 0.05 were considered statistically significant, indicating an association. The results revealed variables like Smoking status, Weight loss, and Loss of appetite to be statistically significant with tuberculosis diagnosis. While non-significant variables, including Gender, Crowded Home, DM, History of TB disease, Sputum, and Fever, were excluded from further analysis.

In addition to the Chi-square test, Fisher's exact test was employed to evaluate associations between categorical variables and tuberculosis diagnosis, particularly when dealing with small sample sizes or non-compliance with chi-square assumptions. Table 3 presents the p-values obtained from Fisher's exact test for various variables. Those with p-values less than 0.05, including TB incidence area, Education status, BCG vaccination, exposure to the TB index case, Night sweat, Chest pain, and Haemoptysis, indicate a statistically significant relationship with TB diagnosis. Conversely, variables like Age Group, Marital status, Health Care

Worker, HIV, Immunosuppressed, and Cough displayed *p*-values above 0.05, signifying no significant association. Based on the combined results from both tests, 13 variables will be excluded due to non-significance, while the remaining variables will proceed to *K*-fold cross-validation.

Table 2 The p-value obtained by Chi-square test

Variables	*p*-value
Gender	0.080
Crowded Home	0.160
DM	0.680
Smoking status	0.004
History of TB disease	0.350
Sputum	0.109
Weight loss	0.000
Loss appetite	0.000
Fever	0.336

Table 3 The p-value obtained by Fisher's Exact test

Variables	*p*-value
Age Group	0.536
Marital status	0.847
TB incidence area	0.032
Health Care Worker	0.323
Education status	0.009
HIV	0.487
Immunosuppressed	0.202
BCG scar	0.035
Exposed to the TB index case	0.031
Cough	0.202
Night sweat	0.000
Chest pain	0.000
Hemoptysis	0.124

3.3 *K*-fold cross-validation

K-fold cross-validation with $K = 10$ was implemented to assess the model's predictive performance. This technique involves dividing the dataset into ten folds for training and testing purposes, with Mean Squared Error (MSE) used as the evaluation metric. The MSE measures the accuracy of the model's predictions compared to the actual data in each fold, and the best fold with the lowest MSE is considered indicative of high predictive performance. The fold with an MSE of 0, signifying perfect alignment between predicted and actual values, was identified as the best fold. This optimal fold is subsequently utilized for further regression analysis. The process ensures a robust evaluation of the model's performance, and a training set and testing set are established through this methodology.

3.4 Regression Model

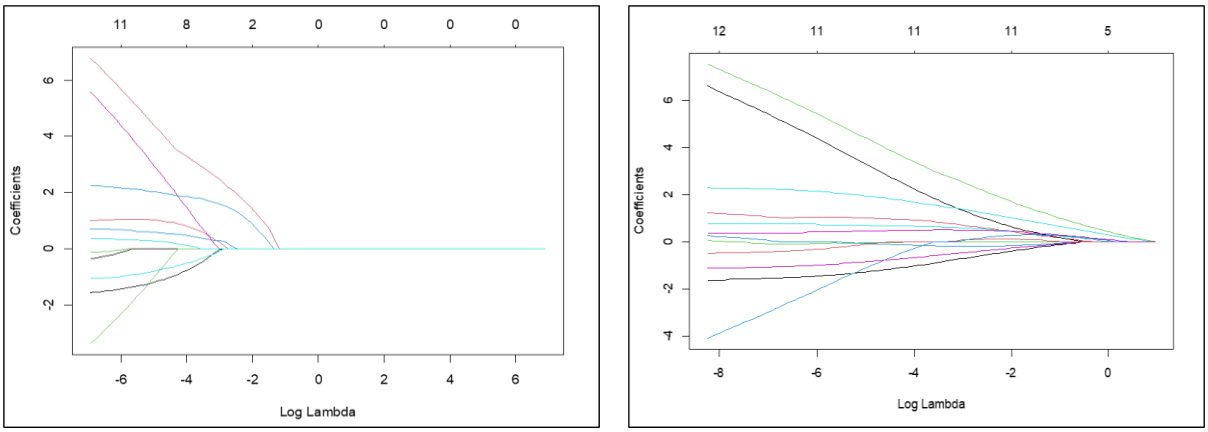
Logistic LASSO and Logistic Elastic Net regression were employed for Tuberculosis prevalence prediction. The logistic LASSO model underwent K-fold cross-validation to determine the optimal lambda, with a value of 0.0163 identified as the best fit. The resulting coefficients (Table 4) revealed variable selection, with certain predictors eliminated (e.g., Secondary Education, Tertiary above Education, History of TB, and Chest Pain). The regularization path in Fig. 1(a) depicted how coefficients changed with varying lambda values, highlighting the impact of regularization. AICc and BIC were then utilized, with AICc at -76.6650 and BIC at -54.0369.

Enhanced Knowledge in Sciences and Technology Vol. 4 No. 2 (2024) p. 191-200

Logistic Elastic Net regression also underwent K-fold cross-validation, resulting in optimal lambda (0.0248) and alpha (0.112) values. Coefficients (Table 4) indicated variable importance, with some eliminated (e.g., Secondary Education, Tertiary above Education, and History of TB). The regularization path in Fig. 1(b) illustrates the impact of regularization strength on coefficients. AICc for Elastic Net was -72.3190, and BIC was -44.3572. The subsequent sections will compare the AICc and BIC values to determine the superior model.

Table 4 The regression coefficients by LASSO and Elastic Net regression

Predictor Variable	LASSO	Elastic Net
Intercept	-3.2496	-3.4279
TB Incidence in Low-Incidence Areas	-0.8460	-0.9483
Primary Education	0.8482	0.8605
Secondary Education	-	-
Tertiary above Education	-	-0.1829
Ex-Smokers	0.4818	0.6390
Passive Smoker	-0.5598	-0.6208
BCG Vaccination	1.6485	1.9317
History of TB	-	-
Night Sweat	3.3750	3.0841
Chest Pain	-	-0.0764
Weight Loss	1.8593	1.5705
Loss Appetite	0.1421	0.4825



(a) **(b)**

Fig. 1 Regularization path for (a) logistic LASSO; (b) logistic Elastic Net regression

The model equation based on the provided coefficients of the logistic LASSO regression and logistic Elastic net regression models are shown in equation (8) and equation (9) respectively [6]. Where *P* indicates the probability, *Y* represents the Diagnosed variable (*0=No, 1=Yes*) and *x* represents the independent variable.

$$P_{LASSO}(Y=1|x_i)=\frac{exp\begin{pmatrix}-3.250-0.846*lowTBarea+0.848*primaryEducation+0.482*exSmoker\\-0.560*passiveSmoker+1.649*BCGvaccination+3.375*nightSweat\\+1.859*weightLoss+0.142*lossAppetite\end{pmatrix}}{1+exp\begin{pmatrix}-3.250-0.846*lowTBarea+0.848*primaryEducation+0.482*exSmoker\\-0.560*passiveSmoker+1.649*BCGvaccination+3.375*nightSweat\\+1.859*weightLoss+0.142*lossAppetite\end{pmatrix}} \quad (8)$$

$$P_{EN}(Y=1|x_i)=\frac{exp\begin{pmatrix}-3.428-0.948*lowTBarea+0.861*primaryEducation-0.183*TertiaryEducation\\+0.639*exSmoker-0.621*passiveSmoker+1.932*BCGvaccination\\+3.084*nightSweat-0.076*chestPain+1.571*weightLoss+0.483*lossAppetite\end{pmatrix}}{1+exp\begin{pmatrix}-3.428-0.948*lowTBarea+0.861*primaryEducation-0.183*TertiaryEducation\\+0.639*exSmoker-0.621*passiveSmoker+1.932*BCGvaccination\\+3.084*nightSweat-0.076*chestPain+1.571*weightLoss+0.483*lossAppetite\end{pmatrix}} \quad (9)$$

Table 5 Comparison of AICc and BIC for logistic Lasso and Elastic Net regression

Methods	AICc	BIC
LASSO	-76.6650	-54.0369
Elastic Net	-72.3190	-44.3572

A comparison of AICc and BIC values for logistic LASSO and Elastic Net regression shown in Table 5 revealed lower values for logistic LASSO (-76.6650 and -54.0369, respectively). This indicates that logistic LASSO regression outperforms Elastic Net and is deemed the optimal model for predicting Tuberculosis prevalence. The next step involves conducting a goodness-of-fit test to assess the selected model's performance and determine its suitability for predictions.

3.5 The goodness of fit Test

The goodness-of-fit test, specifically the deviance residual plot, was employed to assess the logistic LASSO regression model's efficacy in predicting Tuberculosis prevalence. The deviance residual plot for both the training and testing sets revealed consistent parallel lines, indicating the model's ability to maintain reliable predictions without systematic underestimation or overestimation. This alignment suggests a good fit between observed and predicted values. Additionally, the deviance goodness-of-fit test further supported the model's performance, with a calculated *p*-value of 0.367, exceeding the significance level of 0.05. This result indicates that the logistic LASSO model adequately captures the data distribution, reinforcing its reliability in explaining Tuberculosis prevalence patterns in the studied context.

3.6 Prediction of Tuberculosis Prevalence

For predictive purposes, a logistic LASSO regression model was applied to fifteen randomly generated tuberculosis patient cases, utilizing categorical variables. The model yielded probabilities ranging from low to high, reflecting the varying risk levels based on different combinations of predictor variables. For instance, the first case, characterized by a high TB incidence area, no formal education, and a history of ex-smoking, exhibits a low probability (0.0675). In contrast, the twelfth case, marked by a high TB incidence area, primary education, ex-smoker, BCG vaccination, and symptoms like Night Sweats and Weight Loss, presents a high predicted probability (0.9931), indicating an elevated risk of tuberculosis. Conversely, the sixth case, featuring a low TB incidence area, secondary education, ex-smoking, and an absence of significant symptoms, is associated with a moderate probability (0.1229).

4. Discussion

The logistics LASSO regression model proved superior to Elastic Net in predicting tuberculosis prevalence in Kelantan, Malaysia. A comparison using AICc, and BIC favored LASSO, indicating its better fit. The deviance goodness-of-fit test confirmed LASSO's strong fitting capability. Significant predictors such as TB Incidence in Low-Incidence Areas, Primary Education, Ex-Smokers, BCG Vaccination, Night Sweats, Weight Loss, and Loss of Appetite were identified, each with noticeable coefficients. Lower education levels and a history of smoking showed a positive relationship with TB prevalence. While BCG Vaccination had a strong negative influence, highlighting its effectiveness. Symptoms like Night Sweat, Weight Loss, and Loss of Appetite were proven associated with higher TB risk. Null coefficients for variables like Secondary Education, Tertiary above Education, History of TB, and Chest Pain indicated their lack of influence on TB prevalence. The findings enable more targeted approaches for tuberculosis management and control in Kelantan, Malaysia.

Conclusion

The research study on Tuberculosis prevalence in Kelantan, Malaysia, using logistic LASSO and Elastic Net regression identified key predictors such as low TB incidence areas, primary education, ex-smokers, passive smokers, BCG vaccination, night sweats, weight loss, and loss of appetite. The comparison of AICc and BIC revealed

that logistic LASSO regression (-76.665 and -54.037 respectively) outperformed logistic Elastic Net regression (-72.319 and -44.357 respectively), making it the optimal model for Tuberculosis prevalence prediction. The model demonstrated robust predictive performance, aiding public health planning and resource allocation. However, limitations include obtaining comprehensive medical data and the study's localised focus on Kelantan, limiting generalizability. To address these limitations, future research should prioritize collaborative efforts with healthcare institutions, consider an expanded scope to cover multiple regions in Malaysia and explore alternative methodologies such as Artificial Neural Networks. These actions can contribute to more impactful and universally applicable research on Tuberculosis prevalence in Malaysia.

Acknowledgement

The authors would thank the Faculty of Applied Sciences and Technology, Universiti Tun Hussein Onn Malaysia for its support.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of the paper.

Author Contribution

The authors confirm their contribution to the paper as follows: ***study conception and design****: Christy Vun Wan Sze, Khuneswari Gopal Pillay;* ***data collection****: Christy Vun Wan Sze;* ***analysis and interpretation of results****: Christy Vun Wan Sze, Khuneswari Gopal Pillay;* ***draft manuscript preparation****: Christy Vun Wan Sze, Khuneswari Gopal Pillay. All authors reviewed the results and approved the final version of the manuscript.*

References

[1] Koch, A., & Mizrahi, V. (2018). Mycobacterium tuberculosis. *Trends in Microbiology*, *26*(6), 555–556. https://doi.org/10.1016/j.tim.2018.02.012

[2] World Health Organization: WHO. (2022). Tuberculosis. *WHO*. https://www.who.int/news-room/fact-sheets/detail/tuberculosis

[3] The Star. (2022, March 24). *TB cases in Malaysia dropped in 2021, says KJ*. https://www.thestar.com.my/news/nation/2022/03/24/tb-cases-in-malaysia-dropped-in-2021-says-kj

[4] Bernama. (2019, December 16). TB deaths high due to late treatment. *Malaysiakini*. https://www.malaysiakini.com/news/503769

[5] Emmert-Streib, F., & Dehmer, M. (2019). High-Dimensional LASSO-Based Computational Regression Models: Regularization, Shrinkage, and Selection. *Machine Learning and Knowledge Extraction*, *1*(1), 359–383. https://doi.org/10.3390/make1010021

[6] Kim, S. Y., Kim, Y., Jeong, K., Jeong, H., & Kim, J. (2018). Logistic LASSO regression for the diagnosis of breast cancer using clinical demographic data and the BI-RADS lexicon for ultrasonography. *Ultrasonography*, *37*(1), 36–42. https://doi.org/10.14366/usg.16045

[7] Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B-statistical Methodology*, *67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

[8] Garcia-Carretero, R., Vigil-Medina, L., Barquero-Perez, O., Mora-Jiménez, I., Soguero-Ruiz, C., Goya-Esteban, R., & Ramos-Lopez, J. (2020). Logistic LASSO and Elastic net to Characterize Vitamin D Deficiency in a Hypertensive Obese Population. Metabolic Syndrome and Related Disorders, 18(2), 79–85. https://doi.org/10.1089/met.2019.0104

[9] Maharana, K., Mondal, S., & Nemade, B. P. (2022). A review: Data pre-processing and data augmentation techniques. Global Transitions Proceedings, 3(1), 91–99. https://doi.org/10.1016/j.gltp.2022.04.020

[10] Momeni, A., Pincus, M. R., & Libien, J. (2018). Introduction to Statistical Methods in Pathology. In Springer eBooks. https://doi.org/10.1007/978-3-319-60543-2

[11] Scott, M., Flaherty, D. L., & Currall, J. (2012). Statistics: dealing with categorical data. Journal of Small Animal Practice, n/a. https://doi.org/10.1111/j.1748-5827.2012.01298.x

[12] Franke, T., Ho, T. S., & Christie, C. A. (2011). The Chi-Square Test. *American Journal of Evaluation*, *33*(3), 448–458. https://doi.org/10.1177/1098214011426594

[13] Schober, P., & Vetter, T. R. (2019). Chi-square tests in medical research. Anesthesia & Analgesia, 129(5), 1193.

https://doi.org/10.1213/ane.0000000000004410
[14] Tamilarasi, P., & Rani, R. U. (2020). *Diagnosis of Crime Rate against Women using k-fold Cross Validation through Machine Learning*. https://doi.org/10.1109/iccmc48092.2020.iccmc-000193
[15] Wang, Q., Yu, S., Qi, X. S., Hu, Y., Zheng, W. J., Shi, J., & Yao, H. (2019). [Overview of logistic regression model analysis and application]. PubMed, 53(9), 955–960. https://doi.org/10.3760/cma.j.issn.0253-9624.2019.09.018
[16] Ogutu, J. O., Schulz-Streeck, T., & Piepho, H. (2012). Genomic selection using regularized linear regression models: ridge regression, LASSO, elastic net and their extensions. *BMC Proceedings*, *6*(S2). https://doi.org/10.1186/1753-6561-6-s2-s10
[17] Steyerberg, E. W. (2008). Applications of prediction models. In Statistics in the health sciences (pp. 11–31). https://doi.org/10.1007/978-0-387-77244-8_2
[18] Pho, K., Ly, S., Ly, S., & Lukusa, T. M. (2019). Comparison among Akaike Information Criterion, Bayesian Information Criterion and Vuong's test in Model Selection: A Case Study of Violated Speed Regulation in Taiwan. *Journal of Advanced Engineering and Computation*, *3*(1), 293. https://doi.org/10.25073/jaec.201931.220
[19] Xie, X. J., Pendergast, J. F., & Clarke, W. (2008). Increasing the power: A practical approach to goodness-of-fit test for logistic regression models with continuous predictors. *Computational Statistics & Data Analysis*, *52*(5), 2703–2713. https://doi.org/10.1016/j.csda.2007.09.027
[20] Shipe, M. E., Deppen, S. A., Farjah, F., & Grogan, E. L. (2019). Developing prediction models for clinical use using logistic regression: an overview. Journal of Thoracic Disease, 11(S4), S574–S584. https://doi.org/10.21037/jtd.2019.01.25