Social Medicine and Health Management (2025)
Clausius Scientific Press, Canada

DOI: 10.23977/socmhm.2025.060115
ISSN 2616-2210 Vol. 6 Num. 1

Study on Influencing Factors of Tuberculosis Based on Logistic Regression and Decision Tree Model

Kexin Guo[1], Xiaoran Xu[1], Qinge Zhan[1], Li Guo[2], Feng Feng[1,*]

[1]School of Medicine, Shihezi University, Shihezi, 832003, Xinjiang, China
[2]Xinjiang Production and Construction Corps Second Division 30th Regiment Hospital, Tiemenguan, 841000, Xinjiang, China
[*]Corresponding author: j467720872@163.com

Keywords: Tuberculosis, Influencing Factors, Logistic Regression, Decision Tree Model

Abstract: Tuberculosis (TB), long established as a key factor in morbidity and mortality throughout the world. TB not only jeopardizes the health of individuals, but also imposes a heavy burden on society and the economy. Therefore, there is an urgent need for prevention and treatment studies to address this health problem. The aim of this study was to evaluate the predisposing factors of tuberculosis and develop predictive models to identify high-risk groups. The incidence of tuberculosis in 2022 was 133/100,000, which is an increase of 3.9% over the period 2020-2022, against the target of "ending the tuberculosis epidemic". The study collected data from 2032 patients and analyzed key factors such as age, history of tobacco use, gender, alcohol consumption, malnutrition and diabetes through logistic regression, decision tree and random forest models. The results showed that history of tobacco use, history of alcohol consumption, malnutrition and diabetes mellitus were the main causative factors, while age had a weaker relationship. Among the models, logistic regression (91.97% correct for logit and 88.68% for probit), decision tree (89.77% correct), and random forest (97.87% correct) predicted well, with random forest being the best. This research contributes to optimizing the detection and management processes for high-risk populations through enhanced preventive strategies.

1. Introduction

Tuberculosis (TB) was the leading cause of infectious death worldwide until the COVID-19 pandemic, which reduced case reporting and disrupted TB diagnosis and services[1], long established as a key factor in morbidity and mortality throughout the world [2]. Tuberculosis is also one of the most significant public health problems worldwide. In line with the World Health Organization (WHO), the incidence of tuberculosis is high globally, accounting for about one-third of the population, with the majority of patients concentrated in developing countries. It has been seen that apart from physical symptoms, The expenses of treating tuberculosis has become a barrier of their access and adherence to treatment, affect health outcomes, increase the risk of disease transmission, and add to the household`s economic burden [3]. This disease not only seriously jeopardizes the health of patients, but also imposes a heavy burden on society and economy [4]. Therefore, research on the prevention and treatment of TB is not only an urgent need to solve the

current public health problems, but also an important task to ensure the long-term stable development of the society. Since the initial lung signs are unknown, understanding the factors affecting the development of tuberculosis has a positive significance in reducing the incidence of tuberculosis.

Tuberculosis is one of the key concerns in China. For this reason, Diagnosis and Treatment Policy for Tuberculosis Patients In China, After diagnosis, non-rifampicin-resistant TB patients can receive a single free dose of anti-tuberculosis medication uniformly provided by the State [5]. Many scholars have also invested in the study of the disease, some of them studying the factors of the disease. Risk-factors for PTLD include multiple episodes of tuberculosis, drug-resistant tuberculosis, delays in diagnosis, and possibly smoking [6]. Anete Trajman confirm along with well-established risk factors (such as human immunodeficiency virus (HIV), malnutrition, and young age), emerging variables such as diabetes, indoor air pollution, alcohol, use of immunosuppressive drugs, and tobacco smoke play a significant role at both the individual and population level [7]. Li Baiyuan (2021) explored the influence of various factors including poor lifestyle behaviors on the severity of disease in PTB patients, and studied the correlation between poor lifestyle factors such as smoking, drinking, malnutrition, and staying up all night and the development of tuberculosis, and the results side-by-side confirmed that poor habits are one of the predisposing factors that cause tuberculosis disease [8]. In many studies, it has been confirmed that age, gender, and bad habits increase the prevalence of the disease, but they have not been statistically analyzed or applied in real life. There are also some scholars who have conducted relevant studies about the prediction methods of tuberculosis. Mingtao Zeng review summarizes the commonly employed animal models, and their characteristics as used in TB vaccine research, and provides a basis for selecting appropriate animal models according to specific research needs [9].

Zeng, Guansheng (2024) collected patient information and data statistics, used logistic regression to analyze the clinical characteristics of smear-negative pulmonary tuberculosis, and constructed a column-line graph prediction model based on the regression results [10]. Kiarash Ghazvini analyzed using SPSS software version 22 through Logistic regression model and Chi-square test. The results showed that the variables of vitamin D3, hemoglobin and body mass index (BMI) have a better prediction of TB in the logistic regression model [11]. Until now, domestic and foreign scholars have confirmed the correlation between multiple factors and the development of tuberculosis, but the traditional prediction is to use a single model for prediction analysis, such as Logistic regression prediction model, risk prediction model, logistic regression model and so on. However, this study utilizes multiple models to predict and analyze the related indexes to enhance the accuracy of prediction and thus reduce the incidence rate of tuberculosis.

2. The establishment of predictive model

2.1 Structure of Logistic Regression Model

Logistic regression is an extended linear regression model, which is widely used in many fields such as data mining, automatic disease diagnosis and economic forecasting. The model predicts the likelihood of an event by analyzing data sets of independent variables, and its output is a probability value between 0 and 1.

Logit model and Probit model belong to probabilistic nonlinear multiple regression model. The former is based on cumulative logical distribution function, and the disturbance term follows logical distribution. Its expression is:

$$P_i = \frac{1}{1+e^{-Y_i}} \tag{1}$$

Among them, P_i represents the probability of tuberculosis, Y_i is the factor, and the corresponding dummy variable Y_i is taken as 1 when suffering from tuberculosis, otherwise it is taken as 0. Suffering from tuberculosis is influenced by several factors such as age, sex, history of tobacco use, history of alcohol consumption, diabetes and malnutrition.

Take the inverse function to get the linear model:

$$Y_i = \ln \frac{P_i}{1-P_i} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \mu_i \tag{2}$$

Probit model is on the foundation of cumulative normal probability distribution function, and the disturbance term $\mu_i$follows logical distribution, and the expression is:

$$P_i = F(Y_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Y_i} e^{-\frac{t^2}{2}} dt \tag{3}$$

Among them, P_i represents the probability of tuberculosis, and Y_i is a variable factor to measure whether it is sick or not, which is influenced by several factors. As mentioned above about Logit model, Probit model uses its inverse function to convert the probability value into the value of Y_i .

$$Y_i = F^{-1}(P_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \mu_i \tag{4}$$

The binary selection model usually adopts maximum likelihood estimation, and the process is as follows.

According to the symmetry of standard normal or logical distribution.

$$F(-t) = 1 - F(t) \tag{5}$$

$$P(Y_i = 1) = P(Y_i^* > 0) = P(\mu_i^* > -X_i B) = 1 - P(\mu_i^* > -X_i B) = 1 - F(-X_i B) = F(X_i B) \tag{6}$$

$$P(Y_1, Y_2, \cdots, Y_n) = \Pi_{Y_{i=0}} \left(1 - F(X_i B)\right) \Pi_{Y_{i=1}} F(X_i B) \tag{7}$$

$$L = \Pi_{i=1}^{n} \left(F(X_i B)\right)^{Y_i} \left(1 - F(X_i B)\right)^{1-Y_i} \tag{8}$$

After logarithmic transformation on both sides of the function, the derivative of parameter B is obtained to obtain the first-order conditions of maximum likelihood estimation. The parameters are further inferred by the relation between the probability distribution function and the density function.

$$\frac{\partial \ln L}{\partial B} = \sum\nolimits_{n-1}^{n} \left[\frac{Y f_1}{F_i} + (1 - Y_i) \frac{-f_i}{(1 - F_i)} \right] X_i = 0 \tag{9}$$

The methods to test the validity of Probit model and Logit model mainly include goodness-of-fit test and prediction effect test.

Let L0 be the likelihood function value of all explanatory variables in the model with coefficients of 0, then there are:

$$\ln L_0 = n\left(P \ln P + (1 - P) \ln(1 - p)\right) \tag{10}$$

Where P represents the proportion of the disease, n represents the number of samples, and L represents the likelihood function value estimated by the model to construct a statistic.

$$R^2 = 1 - \frac{\ln L}{\ln L_0} \tag{11}$$

When the model does not fit the observed values of samples at all, L=L0, then R2=0; When the model completely fits the sample observation, L=1, then R2=1. Therefore, R2 is used as a statistic to test the goodness of fit of the model, and the larger R2 is, the more effective the model is.

When Logit model and Probit model are estimated, the observed values of explanatory variables of all test samples are substituted into the models, and the probability of choosing 1 as the explained variable of each sample is calculated, then compared with the actual observed values of the explained variable of each test sample, and finally the prediction accuracy of the two models is calculated to compare the effectiveness of the two models.

2.2 The Determination of the Number of Network Layers

Decision tree modeling is a classification and regression method based on tree structure. In this study, we used categorical decision tree model to classify and predict tuberculosis predisposing factors. First, a decision tree is generated by recursively dividing the dataset, with each node representing a feature attribute, branches representing different attribute values, and leaf nodes representing categories or predicted values. This inductive learning framework demonstrates classification efficacy when applied to previously unobserved observational data. Quantitative assessment of classification fidelity and algorithmic consistency provides empirical evidence for the clinical utility of decision tree architecture in tuberculosis susceptibility determinant analysis.

A decision tree model is constructed using selected features as inputs. Model performance was optimized by continuously adjusting parameters such as the depth of the tree and the minimum number of samples for leaf nodes. The model was evaluated using metrics such as accuracy, recall, and F1 score to ensure that the model could accurately identify and predict the predisposing factors of tuberculosis.

Influence variable X: {Age, gender, history of tobacco use, history of alcohol consumption, diabetes mellitus, malnutrition };

Outcome variable Y: {Tuberculosis }

Decision Tree Modeling Analysis Steps The basic steps are as follows:

1) Build decision data classification model by training set data to get decision tree structure.

2) Calculate the feature importance through the established decision tree.

3) Apply the established decision tree classification model to the training and testing data to get the classification evaluation results of the model.

2.3 The Determination of the Number of Network Layers

Random forests (RF) are a class of combinatorial methods designed specifically for decision tree classifiers. Random forests use a fixed probability distribution to generate random vectors. Decision tree bagging is a special case of random forests. By randomly selecting N samples from the original training set there and back, randomness is added to the process of constructing the model, and the bagging uses a uniform distribution to generate self-help samples.

The margin of the classifier M(Xy)=P(Y0=Y)-maxP(Y0=Z)

Where: Y0 is the prediction class for X made by a classifier constructed from some random variable. The larger the margin, the greater the likelihood that the classifier will correctly predict a given sample Randomization helps to reduce the correlation between decision trees thereby improving the generalization error of the combined classifier.

The typical steps for analyzing a decision tree model include:

1) Building a classification model from the training data to determine the decision tree's structure.

2) Assessing the significance of features using the constructed decision tree.

3) Evaluating the model's classification performance by applying it to both training and test datasets.

If put-back sampling is used during tree construction, roughly one-third of the records remain unselected. These unselected records naturally form a control dataset for model validation. As a result, Random Forest does not require a separate dataset for cross-validation, as its inherent algorithm mimics cross-validation, with out-of-bag error serving as an unbiased estimate of prediction error. The sampling process is repeated 10 times with replacement, generating training datasets labeled D1, D2, ..., D10. Decision trees are then built on each dataset, denoted as T1, T2, ...,T10, collectively forming the random forest T*.

The Random Forest Modeling Steps is shown in figure 1.

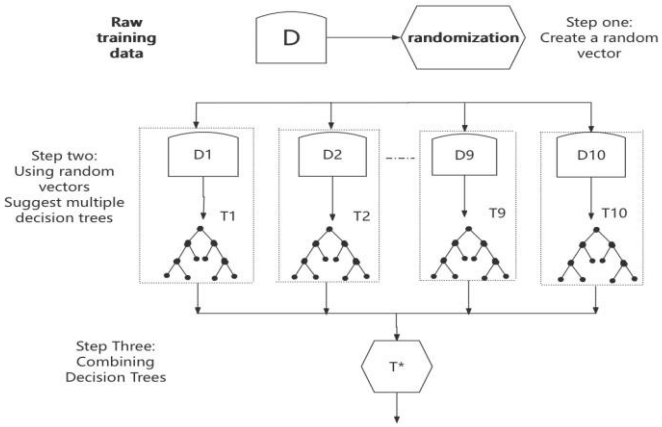

Figure 1 Random forest modeling steps

3. Results

3.1 Analysis of Decision Tree Results

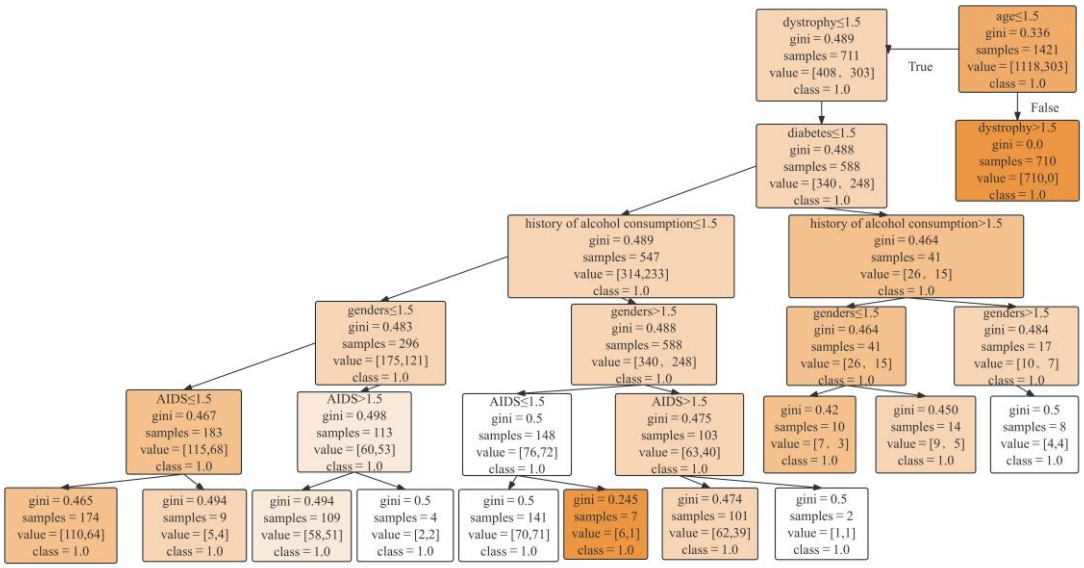

Figure 2 Decision tree output results

The Decision tree output results is shown in Figure 2 and Figure 3.

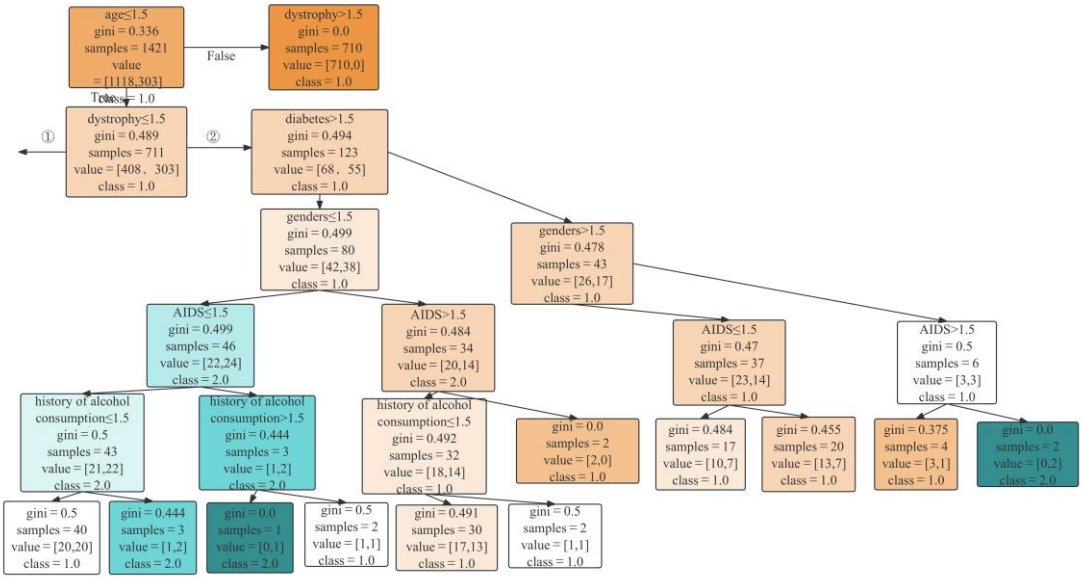

Figure 3 Decision tree output results

The Application of Decision Tree Models for Effectiveness Measurement is shown in table 1.

Table 1 Application of decision tree models for effectiveness measurement

	Accuracy	Recall Rate	Precision rate	F1
Training set	0.8895	0.8895	0.8766	0.8770
Test Set	0.8977	0.8977	0.8736	0.8740

The experimental results demonstrate classification accuracy of 89.77%, reflecting strong predictive consistency across the sample population. Notably, the precision rate achieved equivalent values (89.77%), confirming metric alignment that validates the model's structural integrity through concordant performance measures; The proportion of results predicted to be positive samples that are actually positive is 87.4%, which also indicates the high precision of the model; both the precision law and the recall law are high, and F1 as the reconciled average of the two is also high, which also indicates the high accuracy of the model; All of these data amount to more than 85%, which indicates that the success of our model construction is high as well as the correctness.

3.2 Analysis of Experimental Results

The Disaggregated assessment indicator results for Random Forest Training and Test Sets is shown in table 2.

Table 2 Disaggregated assessment indicator results for random forest training and test sets

	Accuracy	Recall Rate	Precision rate	F1
Training set	0.9751	0.9751	0.9526	0.9529
Test Set	0.9824	0.9824	0.9621	0.9636

The accuracy rate of this experiment reaches 98.24%, indicating that the predicted correct samples are close to the total samples; in the results of the actual positive samples, both the proportion of the predicted positive samples and the proportion of the predicted out positive samples are similar to the prediction, which indicate that the correctness of the model construction is high; both the precision law and the recall law are high, and F1 as the reconciled average of the two is also high, which also

indicates the high accuracy of the model; all of these data amount to 90% or more, indicating high success as well as high correctness of our model construction.

The Predictive consequences of decision tree modeling and random forest modeling for tuberculosis prevalence is shown in table 3.

Table 3 Disaggregated assessment indicator results for random

Decision tree modeling					Random Forest				
Original classification		Predictive classification		Total (N=610)	Original classification		Predictive classification		Total (N=610)
		1	0				1	0	
Count	1	440	32	472	Count	1	466	4	470
	0	35	103	138		0	9	131	140
Percentage	1	93.22	6.78	89.77	Percentage	1	99.15	0.85	97.87
	0	25.36	74.64			0	0.06	93.57	

Comparing the prediction results of TB prevalence between the decision tree model and the random forest model, the correct prediction rate of not having TB and the correct prediction rate of having TB of the random forest were higher than the results of the decision tree, which indicated that the random forest model had excellent prediction effect on the prevalence of TB and was much better than the decision tree model, and the relevant conclusions had the validity.

3.3 Model Comparison, Optimization and Selection

Logistic regression model: It has obvious advantages in dealing with linear relationships, and can more accurately reflect the linear relationship between tuberculosis and various factors, but its ability to deal with non-linear data is relatively weak.

Decision tree model: it can deal with the nonlinear relationship in the data well, and does not need to carry out complex pre-processing of the data. However, it may have a large impact on small changes in the data, leading to model instability.

Random Forest Model: It can produce highly accurate predictions by integrating predictions from multiple decision tree models, but at the same time it is computationally difficult.

Logistic regression model, decision tree model and random forest model have their own advantages and disadvantages in predicting TB predisposing factors. However, the accuracy of random forest is significantly higher than the other two models, so it may be more advantageous to select the random forest model in the study of TB prevalence. To enhance the predictive accuracy of the model, medical professionals and researchers may explore the integration of regression analysis and decision tree algorithms. This approach leverages the strengths of both methodologies while compensating for the limitations inherent in random forest techniques. By combining these models, a more robust and comprehensive predictive framework can be developed, potentially leading to improved diagnostic and prognostic outcomes.

Taken together, the Logit model fits better compared to the Probit model, age, history of tobacco use, and history of alcohol consumption are the key factors in the prevalence of tuberculosis, malnutrition has some effect, diabetes has some effect but is unstable, and personality variables are not significant. Thus, history of tobacco use and history of alcohol consumption will somewhat affect the lung organs differently and thus increase the probability of TB prevalence, while age and malnutrition mainly increase the probability of developing TB disease due to lowered immunity, and the gender aspect does not have a significant effect on the prevalence of TB.

In this experiment, both the decision tree model and the random forest model predicted the prevalence of tuberculosis correctly at more than 90%, indicating that both models are effective in prediction, but compared with the decision tree, the correct rate of the random forest is 99.15%, which is much higher than the decision tree model prediction effect. In this experiment, the random forest

model effects the prediction of the prevalence of tuberculosis excellenly, and it is much better than the decision tree model, and the relevant conclusions have validity.

4. Conclusions

The logistic regression model can accurately reflect the linear relationship between tuberculosis and various factors. For decision makers and medical staff, it can more intuitively grasp the inducing factors of tuberculosis, so as to understand and explain the prediction results of the model. The decision tree model does not need to preprocess the data in a complicated way, and can deal with the nonlinear relationship in the data well, which makes the decision tree model have high applicability in dealing with pulmonary tuberculosis inducing factors with complex relationships. Random forest model is able to produce high-accuracy prediction results through integrating multiple decision tree models. This integrated learning method effectively reduces the possible deviation and variance of a single model and improves the overall prediction accuracy. At the same time, the model provides the function of feature importance evaluation, which helps researchers to understand which factors have important influence on the induction of tuberculosis.

Logistic regression model, decision tree model and random forest model have their own advantages and disadvantages in predicting the inducing factors of tuberculosis. However, the accuracy of random forest is obviously higher than the other two models, so in the study of tuberculosis prevalence, it may be more advantageous to choose random forest model.

With the continuous development and wide application of big data technology, Random forest model will play a more important role in future research in public health fields such as tuberculosis. Analysis methods based on regression models and decision tree models will also be further optimized and improved to better adapt to the complex and changing data environment.

References

[1] Hankins E , Khvolis D , Spigos J T ,et al.Acute Presentation of Tuberculosis Empyema in a Healthy Adolescent[J]. American Journal of Case Reports, 2023, 24.

[2] Sudre P, Ten Dam G, Kochi A. Tuberculosis: a global overview of the situation today[J]. Bulletin of the World Health Organization, 1992, 70(2): 149.

[3] Ghazy R M, El Saeh H M, Abdulaziz S, et al. A systematic review and meta-analysis of the catastrophic costs incurred by tuberculosis patients[J]. Scientific Reports, 2022, 12(1): 558.

[4] Li, Surui. Modeling and study of tuberculosis epidemics[D]. Hebei University of Economics and Business, 2024.

[5] Zhou J. Tuberculosis prevention and control policy[J]. Popular Science,2022,(03):36-37.

[6] Allwood B W, Byrne A, Meghji J, et al. Post-tuberculosis lung disease: clinical review of an under-recognised global challenge[J]. Respiration, 2021, 100(8): 751-763.

[7] Narasimhan P, Wood J, MacIntyre C R, et al. Risk factors for tuberculosis[J]. Pulmonary medicine, 2013, 2013(1): 828939.

[8] Li Baiyuan. Investigation of bad life behavior habits of tuberculosis patients and analysis of factors influencing the severity of the disease [D]. Yan'an University, 2021.

[9] Gong W, Liang Y, Wu X. Animal models of tuberculosis vaccine research: an important component in the fight against tuberculosis [J]. BioMed Research International, 2020, 2020(1): 4263079.

[10] Zeng Guansheng, Chen Lixiang, Chen Hui, et al. Clinical characterization and prediction modeling of smear-negative pulmonary tuberculosis [J]. Journal of Tropical Medicine, 2024, 24 (01): 59-64.

[11] Ghazvini K, Yousefi M, Firoozeh F, et al. Predictors of tuberculosis: Application of a logistic regression model[J]. Gene Reports, 2019, 17: 100527.