Transactions on Materials, Biotechnology and Life Sciences
ISSN: 2960-1932, eISSN: 2960-2319 | Volume 5
MMTCS 2024

Research on Diabetes Risk Prediction Using Multiple Machine Learning Models

Mingxuan Zhang *

Department of Wenli, Northeast Agricultural University, Harbin, China, 150030

* Corresponding Author Email: 19513205115@163.com

Abstract. Diabetes is a chronic hyperglycemic disease caused by insufficient insulin secretion. Traditional diagnostic methods cannot provide early prevention, while data-driven machine learning methods can analyze medical data for early prediction. This study employs four machine learning algorithms—Logistic Regression, Support Vector Machine, Decision Tree, and Random Forest—to analyze and model relevant data. By comparing model performance indicators such as accuracy and recall, it was found that the Random Forest model performs better overall. Using SHAP waterfall and beeswarm plots further to explain the prediction results of the Random Forest model and comparing these results with the variable importance ranking within the Random Forest, it was discovered that Serum Creatinine, Blood Urea, Triglycerides, and Hemoglobin are the most significant factors contributing to diabetes. This finding suggests that these factors should be the focus in predicting the risk of diabetes onset.

Keywords: Disease Prediction, Machine Learning, Predictive Model, SHAP.

1. Introduction

Diabetes is a chronic disease characterized by high blood sugar levels, caused by absolute or relative insulin deficiency and utilization impairment. The disease is mainly classified into three types: Type 1, Type 2, and gestational diabetes. With the development of the socio-economic environment and lifestyle changes, the incidence of diabetes is increasing year by year. According to the International Diabetes Federation, the number of adult diabetes patients worldwide reached 537 million in 2021, affecting approximately 1 in 10 adults globally. By 2025, the prevalence of diabetes is expected to rise to 12.2%, with the number of patients increasing to 783 million. Research shows that more than 50% of diabetes patients are not diagnosed and treated promptly, leading to significant increases in individual healthcare costs and potentially severe complications, thereby increasing the risk of mortality[1-2]. Therefore, it is crucial to implement measures for predicting the risk of diabetes onset, identifying high-risk individuals at an early stage, and promoting personalized intervention plans. This is significant for improving patients' quality of life, reducing healthcare burdens, and mitigating socio-economic losses.

The pathogenesis of diabetes is complex and influenced by genetic, environmental, and lifestyle factors. Traditional methods of diagnosing diabetes primarily rely on blood sugar level tests, such as fasting blood glucose and oral glucose tolerance tests. However, these methods can only detect diabetes once it has developed, making early prevention and intervention impossible. In contrast, data-driven machine learning methods can utilize vast amounts of medical data for mining and analysis, establishing predictive models to screen and intervene in high-risk populations [3]. This is of great significance for preventing and controlling the occurrence of the disease. These methods help identify risk factors and patterns associated with diabetes onset, providing strong support for early diagnosis and intervention of diabetes.

This study predicts diabetes onset risk using multiple machines learning models, including Logistic Regression, Support Vector Machine, Decision Tree, and Random Forest. Initially, the acquired data is cleaned, handling outliers and missing values, followed by validation of the processed dataset. In the model construction phase, this study first builds a Logistic Regression model, calculate the model's loss function value, and perform the Hosmer-Lemeshow test. Next, this study constructs an

SVM model, using grid search to determine the optimal combination of parameters. Then, this study builds a Decision Tree model, optimizing algorithm performance by plotting the AUC values on the training and test sets as model parameters change and adjusting the loss matrix parameters. Finally, this study constructs a Random Forest model, identifying the optimal parameters by comparing various indicators on the test set and measuring variable importance using the average decrease in Gini impurity. Ultimately, this study compares the performance of the four models using metrics such as accuracy and interpret the results using SHAP waterfall and beeswarm plots.

2. Data Source and Preprocessing

2.1. Data Source

The data for this study is sourced from www.ncmi.cn. The original dataset includes a total of 88 items, such as age, gender, fatty liver, and renal failure. For our analysis, this study selected 11 items: age (AGE), gender (SEX), body mass index (BMI), triglycerides (0.4-1.7 mmol/L) (TG), total cholesterol (3.1-5.7 mmol/L) (TC), fibrinogen (FBG), blood urea (BU), serum creatinine (SCR), hemoglobin (HB), platelets (PLT), and lactate dehydrogenase (40-250 U/L) (LDH_L). These selected items were used to predict the risk of diabetes onset.

2.2. Data Preprocessing

First, to identify outliers in different indicators, this study attempts to use box plots. Here, this study takes the first four items in the dataset—AGE, SEX, BMI, and TG—as examples, with the same method applied to other items. As shown in Figure 1:

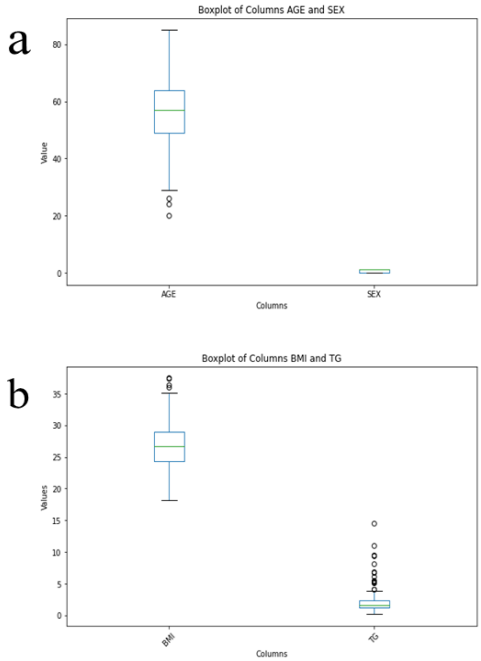

Figure 1. (a) Box Plot of AGE and SEX;(b) Box Plot of BMI and TG

Identify the missing values in the data and visualize them. The results are shown in Figure 2, Figure 3:

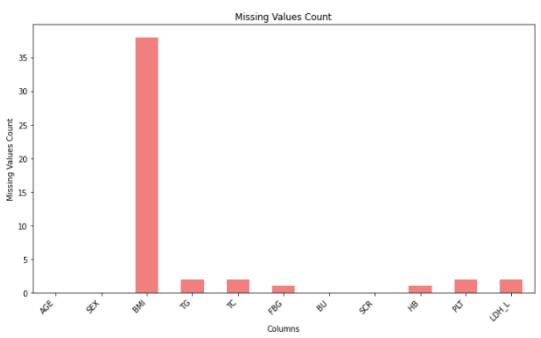

Figure 2. Missing Value Statistics

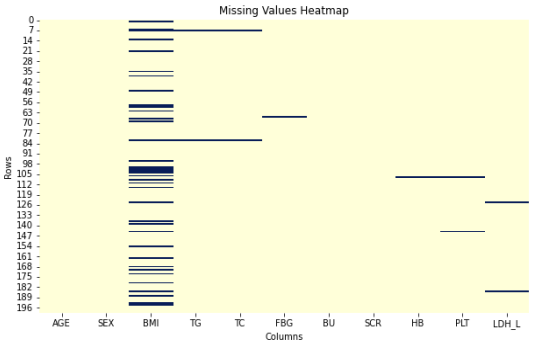

Figure 3. Visualization of Missing Values

Calculate the Spearman correlation coefficient between each data item and create a heatmap of the correlation coefficients. By analyzing the magnitude of the correlation coefficients, determine the specific type of missing values and thereby identify the appropriate interpolation method. The heatmap of the correlation coefficients is shown in Figure 4:

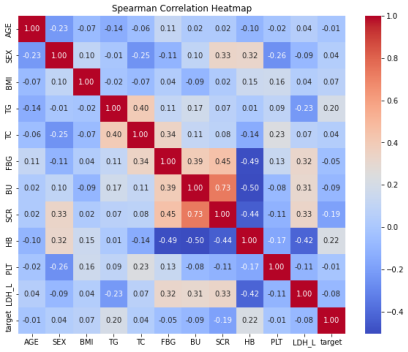

Figure 4. Heatmap of Correlation Coefficients

From the Figure 4, except for the strong correlation between SCR and BU, the relationships between the other data items are weak or very weak. Therefore, this study consider the missing values in the original dataset to be of the Missing At Random (MAR) type. This study chose to use Random Forest to handle the missing values and conducted an F-test on the imputed dataset. The results are shown in Table 1:

Table 1. F-test Results of the Imputed Dataset

Category	AGE	SEX	BMI	TG	TC	FBG	BU	SCR	HB	PLT	LDH_L
P	0.010	0.000	0.000	0.000	0.010	0.000	0.000	0.000	0.000	0.000	0.000

The imputed data items are significant at the 0.01 level, providing sufficient reason to consider the imputed results as reasonable. Descriptive statistics for the imputed data are shown in Table 2:

Table 2. Descriptive Statistics of the Imputed Dataset

	Number of Observations	Standard Error	Minimum Value	Maximum Value
AGE	200	0.74	29.0	85.0
SEX	200	0.03	0	1.0
BMI	200	0.21	18.1	35.1
TG	200	0.05	0.2	3.9
TC	200	0.08	2.1	7.8
FBG	200	0.08	2.1	7.3
BU	200	0.16	2.4	13.2
SCR	200	2.53	30.1	182.9
HB	200	1.65	68.0	175.0
PLT	200	3.55	90.0	352.0
LDH_L	200	2.35	90.6	255.1

3. Algorithm Selection and Interpretability Analysis

3.1. Logistic Regression

Logistic Regression models the linear combination of input features and uses the sigmoid function to map the linear output to a probability value between 0 and 1, thereby predicting the probability of an event occurring. The sigmoid function is as follows:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\alpha+\beta X)}} \quad (1)$$

The value of the model's loss function is 0.5914. The results of the Hosmer-Lemeshow test are shown in Table 3:

Table 3. Hosmer-Lemeshow Test

Chi-square	df	Significance
9.383	8	0.311

The P-value is greater than 0.05, thus this study rejects the null hypothesis and consider the model fit to be good.

Chen et al pointed out that using the Logistic Regression model can accurately predict whether patients have heart disease, guiding clinical prediction [4]. Chen et al used the example of general surgery to establish a Logistic prediction model, providing a reference for preventing surgical complications [5].

3.2. Support Vector Machine

Support Vector Machine works by finding support vectors that have good discriminative power for individual classifications, and then constructing a separating hyperplane. In high-dimensional space, it completely separates the two classes defined by the binary outcome variable, ultimately achieving the goal of building the optimal classifier to maximize the margin between classes. The optimal separating hyperplane is determined by solving the following equation:

$$\begin{cases} \min_{(w,b)} \frac{\|w\|^2}{2} + C\sum_{i=1}^{n} \varphi_i \\ s.t. y_i(w^T x_i + b) \geq 1 - \varphi_i . i = 1,2,\cdots,n \end{cases} \quad (2)$$

This study use grid search to determine the optimal combination of parameters for the Support Vector Machine. The range for C is set to [0.01, 0.1, 1, 10, 100], the range for kernel types is ['linear', 'rbf'], and the range for gamma is [0.01, 0.1, 1, 10, 100]. This study choose the negative mean squared error as the evaluation metric and use 10-fold cross-validation. The optimal parameters for the SVM are found to be C=0.1, kernel='linear', and gamma=0.1. The evaluation metric corresponding to the optimal parameters, the lowest cross-validation score (negative mean squared error), is -0.0706.

Ding et al integrated the Artificial Bee Colony algorithm into a single SVM algorithm, resulting in higher prediction accuracy and greater stability [6]. Chen et al developed a visualization system based on SVM that can identify and analyze common respiratory diseases, filling the gap in the visualization of current prediction models [7].

3.3. Decision Tree

The Decision Tree algorithm recursively splits the dataset into smaller subsets, representing the decision process in a tree structure. Each node represents a feature or attribute, each branch represents a split based on a feature value or attribute value, and each leaf node represents a class or prediction value.

By observing the AUC values on the training and test sets as the model parameters change, this study determines the optimal parameters to be: max_depth=7, min_samples_split=40, min_samples_leaf=4, and the loss matrix= (0, 1, 1, 0).

Wang et al and others constructed a prediction model for the risk of hospital-acquired infections in cardiovascular inpatients based on the Decision Tree algorithm, providing a basis for improving patient survival rates [8]. Zhang et al significantly improved the accuracy of predicting asthma occurrences by integrating the Decision Tree algorithm with other algorithms [9].

3.4. Random Forest

Random Forest builds a forest randomly composed of many decision trees, with each decision tree being parallel to the others. Whenever a new sample is input, each decision tree is used to determine its category, and the final predicted category for the sample is given by combining all the results.

The parameters that need to be determined for Random Forest include the maximum number of decision trees and the number of randomly selected features. By constructing models with different parameter values and comparing their prediction performance on the test set, this study ultimately selects 250 as the maximum number of decision trees and 6 as the number of randomly selected features.

Li et al proposed a hybrid model combining Random Forest and Artificial Neural Network. Comparative experiments demonstrated that this model effectively improves prediction accuracy and recall rate, achieving an accuracy of 96% [10]. Xie et al explored the influencing factors of related diarrhea and constructed a Random Forest model, achieving an accuracy of 76.27%, which can provide a reference for healthcare personnel to develop personalized preventive measures [11].

3.5. Algorithm Selection

This study calculates the accuracy, recall, and other metrics for the predictions of the four algorithms on the training and test sets, respectively, to compare the models. The results are shown in Table 4 and 5:

Table 4. Performance Comparison of the Four Models on the Training Set

Evaluation Metrics	LR	SVM	DT	RF
Accuracy	0.6563	0.6875	0.7813	0.7563
Recall	0.8387	0.8710	0.7634	0.75
F1 score	0.7393	0.7642	0.8023	0.7797
Positive Likelihood Ratio	1.4048	1.5357	3.9347	3.1875
Negative Likelihood Ratio	0.4002	0.2981	0.2935	0.3269
Positive Predictive Value	0.6610	0.6807	0.8452	0.8118
Negative Predictive Value	0.6429	0.7073	0.7105	0.6933

Table 5. Performance Comparison of the Four Models on the Test Set

Evaluation Metrics	LR	SVM	DT	RF
Accuracy	0.575	0.55	0.6	0.625
Recall	0.7619	0.8095	0.5238	0.6667
F1 score	0.6531	0.6538	0.5789	0.6154
Positive Likelihood Ratio	1.2063	1.0986	1.6587	1.6296
Negative Likelihood Ratio	0.6463	0.7238	0.6960	0.5641
Positive Predictive Value	0.5714	0.5484	0.6471	0.5714
Negative Predictive Value	0.5833	0.5556	0.5652	0.6842

Considering the results from both the training set and the test set, the Random Forest shows the best predictive performance. When comparing the Random Forest with the other three models, the P-values are all less than 0.01, which this study consider to be statistically significant ($P<0.05$).

The ROC curves of the four models on the training set and the test set are plotted as shown in Figure 5:

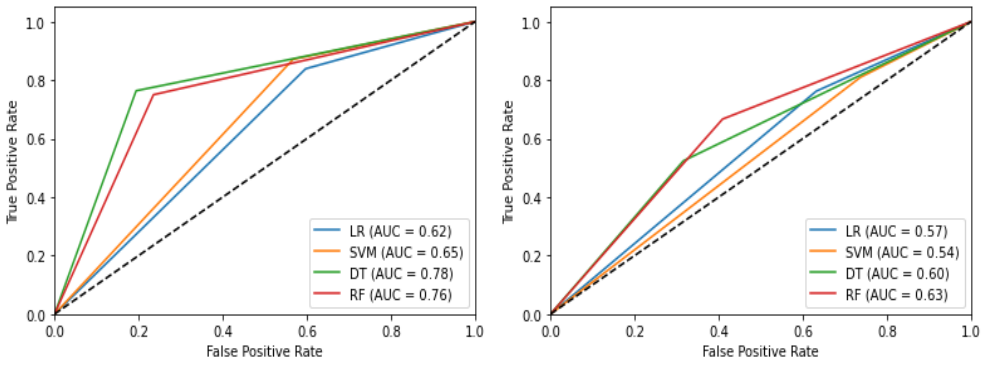

Figure 5. ROC Curves of the Four Models on the Training Set (left) and Test Set (right)

On the training set, all four models exhibit good predictive performance, with the Decision Tree model performing the best. On the test set, the Decision Tree and Random Forest models show good predictive performance,while the Logistic Regression and Support Vector Machine models perform moderately.Overall, the Random Forest model performs better.

3.6. SHAP

SHAP is an algorithm used to interpret the prediction results of machine learning models. It is based on the concept of Shapley values, calculating the contribution of each feature to the prediction result, thus providing consistent and locally accurate explanations. This method decomposes the model output into a baseline value and feature contribution values, making the prediction process for each data point transparent, thereby enhancing the interpretability and reliability of the model.

Zeng et al and others constructed a prediction model for death or readmission due to acute heart failure and then used the SHAP algorithm to identify important clinical features influencing the model's output [12]. Li et al and others utilized the SHAP algorithm to provide detailed explanations for influenza-like illness predictions, which helped in formulating effective measures [13].

Since the Random Forest model has the best predictive performance, this study use the SHAP algorithm to interpret the results of the Random Forest model. This study create SHAP waterfall plots based on the Random Forest model, along with the corresponding beeswarm plots, as shown in Figure 6:

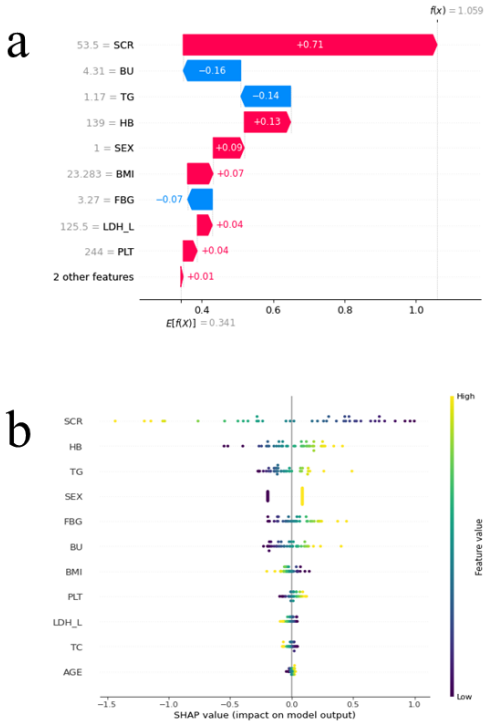

Figure 6. (a)SHAP Waterfall Plot;(b) Beeswarm Plot

The Random Forest model's predicted result value based on this dataset is 1.059. The four features contributing the most to this result are SCR, BU, TG, and HB. SCR and HB increase the model's predicted value by 0.71 and 0.13, respectively, while BU and TG decrease the model's predicted value by 0.16 and 0.14, respectively. The length of the lines in the beeswarm plot also shows that SCR, HB, and TG have the greatest contribution to the model. SCR and HB have a positive impact on the model prediction (lines pointing upward), while TG hurts the model prediction (lines pointing downward).

The measurement of variable importance in the Random Forest model is achieved by comparing the reduction in Gini impurity. This study calculate this for the 11 variables, and the ranking of variable importance is visualized as shown in Figure 7:

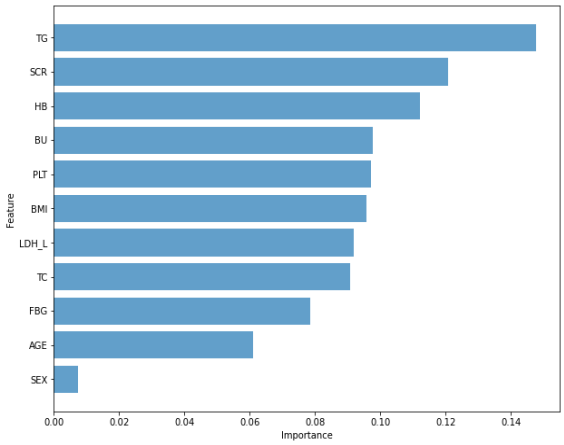

Figure 7. Variable Importance Ranking Results

The results obtained from the SHAP waterfall plot are essentially the same as those obtained from the reduction in Gini impurity. SCR, BU, TG, and HB have the greatest contribution, indicating that this study should focus on these variables when predicting the risk of diabetes onset.

4. Conclusions:

This study aims to use machine learning techniques to establish a diabetes risk prediction model for early diagnosis and intervention. After acquiring the data, this study handle the missing and outlier values in the dataset. The imputed data items are significant at the 0.01 level, indicating that the imputation results are reasonable. This study establish a binary Logistic Regression model and find that TG, SCR, and HB levels significantly impact diabetes incidence. This study calculate the model's loss function value and conduct the Hosmer-Lemeshow test, with a P-value of 0.311 > 0.05. Therefore, this study reject the null hypothesis and consider the model fit to be good. This study then construct a Support Vector Machine, Decision Tree, and Random Forest models. For the SVM model, this study use grid search and 10-fold cross-validation to find the optimal parameters, with the lowest cross-validation score being -0.0706. In the Decision Tree model, this study determine the model parameters by observing changes in AUC values on the training and test sets. In the Random Forest model, this study determine the optimal parameters by comparing different parameters on the test set. Through model evaluation and comparison, this study find that the Random Forest model performs the best in predicting diabetes risk. By plotting SHAP waterfall and beeswarm plots based on the Random Forest model, this study discover that SCR, BU, TG, and HB have the greatest contributions.

This study provides a research approach and framework applied to the field of diabetes, employing methods such as Logistic Regression, Support Vector Machine, Decision Tree, and Random Forest to predict diabetes risk. By integrating these machine learning algorithms, this study can not only improve the accuracy of diabetes risk prediction but also provide reliable support for clinical decision-making. This, in turn, offers robust technical support for public health management and the formulation of personalized prevention strategies.

References

[1] CHARBONNEL B, SIMON D, DALLONGEVILLE J, et al. Direct medical costs of type 2 diabetes in France: an insurance claims database analysis [J]. Pharmacoecon Open, 2018, 2(2):209-219.

[2] He Xiaoning, Zhang Yawen, Ruan Zhen, et al. Study on the Prevalence of Chronic Complications and Per Capita Medical Expenses in Chinese Patients with Type 2 Diabetes [J]. Chinese Journal of Endocrinology and Metabolism, 2019, 35(3): 6.

[3] Ma Yujia, Che Qianzi, Zheng Qiwen, et al. Common Evaluation Methods for Risk Prediction Models of Type 2 Diabetes [J]. Chinese Journal of Prevention and Control of Chronic Diseases, 2020, 28(02): 94-100.

[4] Chen Mengmeng, Fang Zhenhong, Tu Wenyi, et al. Construction and Effect Analysis of a Heart Disease Prediction Model Based on Logistic Regression [J]. Hospital Management Forum, 2022, 39(02): 32-35.

[5] Chen Wangyue, Xue Fang, Han Wei, et al. Prediction of Five Types of General Surgery Complications Based on Logistic Regression Model [J]. Basic & Clinical Medicine, 2023, 43(06): 974-980.

[6] Ding Weijing, Zhang Shaomeng, Pei Yuntao. Research on a Prediction Model for Children's Influenza Based on an Artificial Bee Colony Optimized Support Vector Machine [J]. Pharmacy Today, 2024, 34(01): 74-80.

[7] Chen Jingwen, Zhang Pengpeng, Xu Siyu, et al. Visualization System for Respiratory Disease Prediction Based on Machine Learning [J]. Internet of Things Technology, 2023, 13(02): 68-70.

[8] Wang Silu, Liu Lei, Wu Wei, et al. Construction of a Prediction Model for Hospital-acquired Infection Risk in Cardiovascular Inpatients Based on Decision Tree Algorithm [J]. Journal of Shenyang Medical College, 2023, 25(03): 312-317.

[9] Zhang Qing, Duan Liyao, Liu Yanxiang, et al. Research on a Risk Prediction Model for Asthma Incidence by Integrating Multiple Machine Learning Algorithms [J]. Journal of Environmental Hygiene, 2024, 14(02): 113-120.

[10] Li Dan, Lu Yan, Wu Peishan, et al. Disease Risk Prediction Based on an Improved Random Forest Ensemble Model [J]. Laboratory Research and Exploration, 2023, 42(09): 95-99+109.

[11] Xie Wenliang, Wang Shufang, Li Xuguang, et al. Construction and Validation of a Risk Prediction Model for Enteral Nutrition-Related Diarrhea in ICU Patients [J]. Chinese Journal of Nursing, 2022, 57(19): 2324-2332.

[12] Zeng Jing, He Xiaolong, Hu Huajuan, et al. Construction of a Risk Prediction Model for Death or Readmission During Vulnerable Periods in Patients with Acute Heart Failure Based on Machine Learning [J]. Journal of Army Medical University, 2024, 46(07): 738-745.

[13] Li Jin, Wei Yanlong, Xue Hongxin. Analysis of Influencing Factors for Influenza-like Cases Based on Random Forest Model and SHAP Algorithm [J]. Information Technology and Informatization, 2024(02): 3-6.