Proceedings of the 2nd International Conference on Mathematical Physics and Computational Simulation
DOI: 10.54254/2753-8818/42/2024CH0212

Analysis of influencing factors of diabetes based on logistic regression

Jiayi Li

School of Statistics and Data Science, Beijing Wuzi University, Beijing, 101125, China

2121330025@bwu.edu.cn

Abstract. As a global chronic disease, diabetes has a serious impact on human health and imposes a significant economic burden. Facing this challenge, researchers are actively developing and optimizing predictive models to improve early diagnosis and management of diabetes. This paper analyzed the influence of independent variables on the risk of diabetes by logistic regression from 8 aspects including body mass index (BMI), glycosylated hemoglobin (HbA1c) and heart disease. The logistic regression model, an effective binary classification method, optimizes parameters using maximum likelihood estimation to predict diabetes probability. The model will be evaluated by ROC curve, cross-validation, standardized residual analysis and confusion matrix to comprehensively test its predictive power, stability and classification performance. The results showed that hemoglobin A1c level (HbA1c.level) had the most significant effect on diabetes risk. Other relevant variables, including blood glucose level and body mass index (BMI), demonstrated significant positive correlations, particularly with hypertension and heart disease. The findings will enhance early diabetes identification and provide data to support the development of targeted prevention and intervention measures to reduce the burden of diabetes on individuals and society.

Keywords: Diabetes, logistic regression, ROC curve, cross-validation.

1. Introduction

Diabetes, as an incurable chronic endocrine disease, seriously affects the health of a large number of people around the world [1]. According to the International Diabetes Federation (IDF) Diabetes Atlas (2021), 10.5% of the adult population (ages 20-79) have diabetes, and nearly half do not know they have it. IDF forecasts show that the total number of people with diabetes is expected to rise to 643 million by 2030. By 2045, 1 in 8 adults (about 783 million) will have diabetes, representing an increase of 46% [2].

In order to effectively predict and manage diabetes, many researchers and experts are working to develop and optimize various predictive models and treatment strategies [3]. Zhang et al. used K-nearest neighbor (KNN), support vector machine, logistic regression, and other single algorithms, as well as complex analysis models such as random forest and voting methods to predict diabetes data [4]. Wen et al. established the prediction equation through multi-factor Logistic regression analysis, drew the ROC curve, and evaluated the clinical effectiveness of the prediction equation for diabetic kidney disease by external diagnostic verification [5]. Liu et al. evaluated the effectiveness of various ensemble learning algorithms and decision trees (DT) for developing a risk assessment model for type 2 diabetes in

DOI: 10.54254/2753-8818/42/2024CH0212

individuals aged 45 and older in China. Their findings offer a theoretical foundation for using ensemble learning techniques in the prevention and management of diabetes among this age group [6].

Diabetes impacts both the physical well-being of individuals and imposes significant economic costs on their families and society. As a common chronic disease, diabetes often leads to multiple serious complications [7]. Diabetic retinopathy (DR) consists of a series of fundus lesions resulting from retinal microangiopathy and nerve damage caused by diabetes mellitus (DM). It is the most common ocular complication in DM patients and also one of the fundus diseases with a high rate of blindness [8]. In addition, diabetes can involve the heart and cause changes in the structure and function of cardiac microvessels and myocardium, which affects the normal physiological function of the heart and can ultimately cause sudden cardiac death in patients [9]. Among them, aortic calcification will increase the risk of cardiovascular disease in diabetic patients, and aortic calcification is one of the common complications of diabetes, which is also an important reason for the increased mortality of diabetic patients [10]. In the face of this global health challenge, accurate diabetes prediction has become a crucial focus of current research. By leveraging big data analytics and machine learning algorithms, researchers hope to be able to identify at-risk populations ahead of time and take effective prevention and intervention measures to reduce the onset and progression of diabetes [11, 12]. Developing and promoting public health policies is also vital, including health education, healthy lifestyle promotion, and efficient use of medical resources [13].

In conclusion, diabetes is a major global health and economic challenge [14, 15]. Addressing it effectively requires scientific research, interdisciplinary cooperation, and collective societal effort to improve health outcomes and quality of life for patients.

2. Methodology

2.1. Data source

The data set used in this article is sourced from the Kaggle website (Diabetes Prediction Dataset). Formed from electronic health records (EHRs) provided by multiple healthcare providers, the data set contained a total of 100,001 data entries. After removing incomplete records, 48,451 samples were selected. The original data set is kept in CSV format.

2.2. Variable selection

The diabetes prediction dataset comprises patients' medical and demographic information, including their diabetes status (positive or negative). It features variables such as age, sex, body mass index (BMI), blood pressure, heart disease, smoking history, glycosylated hemoglobin, and blood sugar levels. Table 1 offers a comprehensive overview of this dataset.

Table 1. List of Variables

Variable	Logogram	Meaning
Gender	X_1	The biological sex of the individual. Female(1), male(2)
Age	X_2	The biological age of the individual
Hypertension	X_3	A disease in which the arterial blood pressure continues to rise. Not sick (0), sick (1)
Heart disease	X_4	Not sick (0), sick (1)
Smoking history	X_5	Ever(1), former(2), never(3)
BMI	X_6	A measure of body fat based on weight and height.
HbA1c level	X_7	Indicates average blood sugar levels over the past 2-3 months.
Blood glucose level	X_8	Indicates the glucose amount in the blood at a specific time.
Diabetes	Y	Not sick (0), sick (1)

DOI: 10.54254/2753-8818/42/2024CH0212

2.3. Variable selection

Machine learning classification algorithms are widely used in diabetes risk prediction [16]. A logistic regression model was employed to assess how independent variables affect the likelihood of diabetes. A logistic regression model predicts diabetes risk by mapping a linear combination of variables to a probability between 0 and 1. The basic principle of the model is to optimize parameters using maximum likelihood estimation to maximize the likelihood function between predicted probabilities and actual observed results. The basic construction of the model includes the linear combination, the logistic function (Sigmoid function), and the log-likelihood function. The logistic regression model first calculates the linear combination of the independent variables:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{1}$$

In the formula above: z is the result of a linear combination, β_0 is the intercept, $\beta_1, \beta_2 \ldots \beta_n$ is the regression coefficient, and $x_1, x_2 \ldots x_n$ is the independent variable. The linear combination z is mapped to a probability value between 0 and 1 using the logical function (Sigmoid function):

$$\hat{p} = \frac{1}{1+e^{-z}} \tag{2}$$

In the formula above: $\hat{p}$ is the predicted probability of the event occurring, and e is the base of the natural logarithm. During training, logistic regression estimates the model parameters by maximizing the log-likelihood function:

$$Log - Likelihood = \sum_{i=1}^{m} [y_i log(\hat{p}_i) + (1 - y_i) log(1 - \hat{p}_i)] \tag{3}$$

In the formula above: m represents the number of samples, y_i represents the actual class of the i-th observation (0 or 1), and $\hat{p}_i$ represents the predicted probability of the i-th observation.

Using the above formula, the logistic regression model can effectively address classification problems and optimize model parameters by maximizing the log-likelihood function, thus achieving accurate predictions of event probabilities [17]. After establishing the model, its predictive ability, stability, and classification performance are comprehensively evaluated using the ROC curve, cross-validation, standardized residual analysis, and confusion matrix.

3. Results and discussion

3.1. Data feature analysis

Feature selection analysis can reveal the correlation and interaction between features. Studying the selected feature set helps infer relationships and interactions, providing insights into the data's structure and patterns. This analysis provides insights into feature importance, helping to optimize models, enhance feature engineering, improve model accuracy and explanatory power, and simplify model complexity.

Feature selection analysis can also effectively examine the relationships between diabetes-related features. In this paper, the ggplot() function in the ggplot2 package, combined with the geom_tile() and geom_text() functions, was used to generate heat maps in R Studio to observe correlations among diabetes influencing factors. Figure 1 illustrates:

DOI: 10.54254/2753-8818/42/2024CH0212

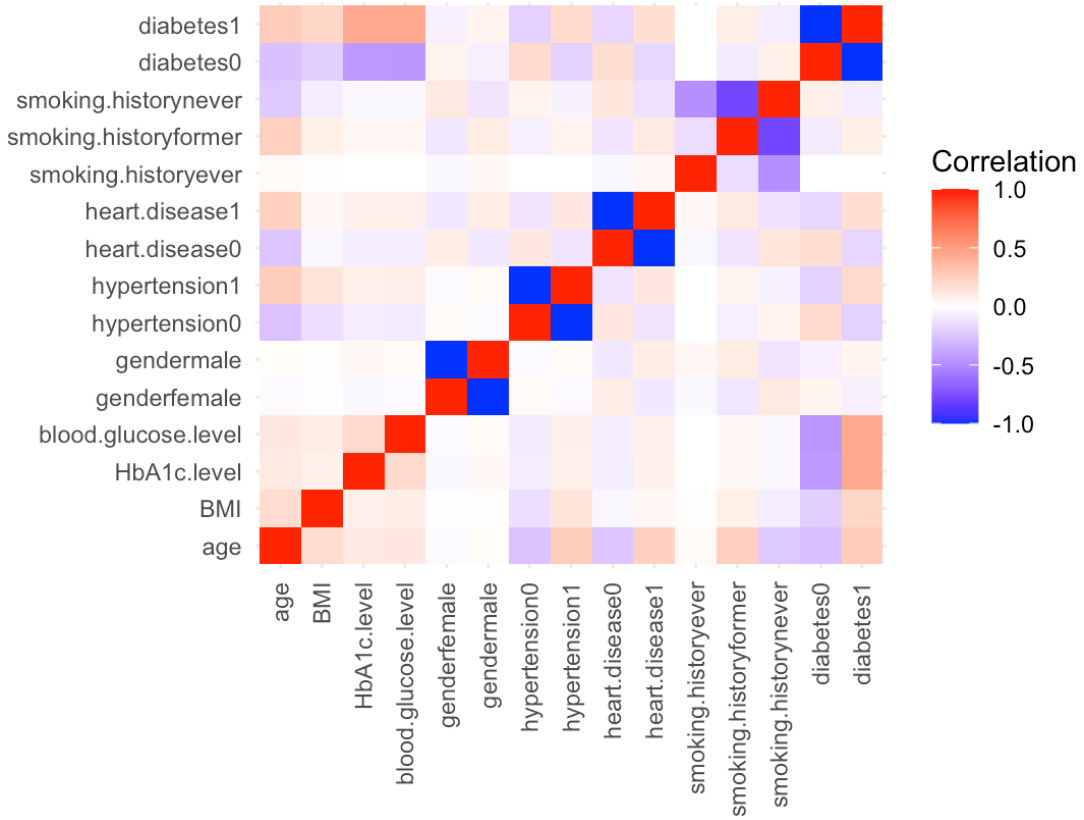

Figure 1. Correlation Heatmap

Figure 1 reveals that diabetes incidence is most strongly associated with blood glucose and HbA1c levels. Heart disease and hypertension are the next most associated. Additionally, age and BMI are also important characteristics to monitor.

3.2. Model building and analysis

The logistic regression algorithm is applicable to binary and multi-class classification problems. It is characterized by a simple model, easy implementation and interpretation, and the ability to handle large-scale data sets [18, 19]. This study analyzed the impact of various factors, including demographic and health-related variables, on the occurrence of diabetes. The model results revealed the degree and significance of each predictor's impact on diabetes risk.

Table 2. Logistic regression model coefficient

Predictor	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-26.800	0.377	-71.111	< 2e-16 ***
Gender male	0.330	0.046	7.119	1.09e-12 ***
Age	0.049	0.002	32.485	< 2e-16 ***
Hypertension1	0.678	0.057	11.944	< 2e-16 ***
Heart disease1	0.654	0.076	8.587	< 2e-16 ***
Smoking history former	-0.077	0.087	-0.885	0.376
Smoking history never	-0.101	0.081	-1.254	0.210
BMI	0.089	0.003	27.598	< 2e-16 ***
HbA1c.level	2.286	0.045	50.566	< 2e-16 ***
Blood glucose level	0.032	0.001	52.383	< 2e-16 ***

DOI: 10.54254/2753-8818/42/2024CH0212

First, HbA1c level had the most significant effect on diabetes risk, with a coefficient of 2.286 ($p < 0.001$), indicating that an increase in HbA1c level significantly increased the risk of diabetes. The coefficient for blood glucose level was 0.032 ($p < 0.001$), also showing a significant positive association. This suggests that higher blood glucose levels elevate diabetes risk. Hypertension and heart disease were both strongly associated with diabetes, showing coefficients of 0.678 ($p < 0.001$) and 0.654 ($p < 0.001$), respectively, which highlights their significant role in increasing diabetes risk.

The 'male' category in the gender variable had a significant positive association (coefficient of 0.329, $p < 0.001$), suggesting that men are at a greater risk of diabetes than women. Age was a significant positive factor (coefficient of 0.049, $p < 0.001$), with the logarithmic odds of diabetes increasing by 0.049 for each additional year of age. This indicates a gradual increase in diabetes risk with age. The coefficient for body mass index (BMI) was 0.089 ($p < 0.001$), indicating that diabetes risk increased by approximately 0.089 for each additional unit of BMI. In contrast, the effect of smoking history on diabetes was not significant in this model. Specifically, the 'former smoker' category had a coefficient of -0.077 ($p = 0.376$), and the 'never smoker' category had a coefficient of -0.101 ($p = 0.210$), neither of which reached statistical significance.

The model demonstrated a high goodness of fit, with a residual deviation of 13,692, significantly lower than the 33,904 of the model without predictive variables. This suggests that the model effectively interprets the data. The AIC value of 13,712 suggests that the model fits the data well.

3.3. Model performance evaluation

The accuracy rate, recall rate, F1-score, and AUC values of the logistic regression model were calculated and evaluated, and the ROC curve was plotted, as shown in Figure 2. Table 3 shows the evaluation indicators for the logistic regression model.

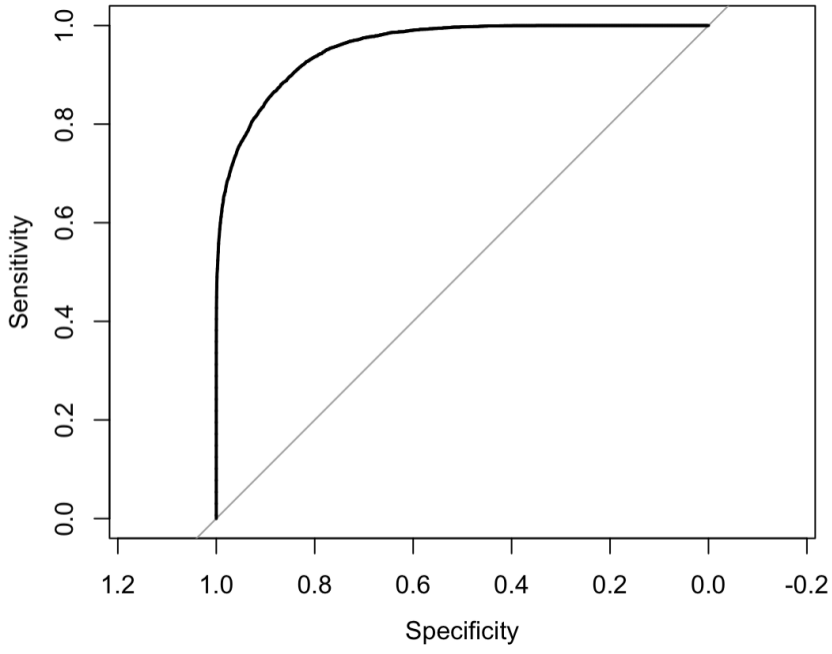

Figure 2. Logistic regression ROC curve

Table 3. Logistic regression evaluation index table

Accuracy	Precision	Recall	F1 Score	AUC
0.9478649	0.8584577	0.6381287	0.7320747	0.9569

DOI: 10.54254/2753-8818/42/2024CH0212

Based on the confusion matrix, the model's performance on the test set includes 3451 true positives, 569 false positives, 42470 true negatives, and 1957 false negatives. The accuracy is 94.79%, precision is 85.85%, recall is 63.81%, and the F1 score is 73.21%. An AUC of 0.9569 demonstrates the model's strong classification ability.

These statistical indicators show that although the model performs well in terms of overall accuracy and precision, its recall for minority classes is relatively low, indicating that further optimization is needed to improve the recognition of minority classes.

Additionally, this paper applied 10-fold cross-validation to assess the model [20]. Initially, all 48,451 samples were split into 10 equal-sized subsets. Each subset was used as the test set once, while the remaining nine subsets served as the training set. This process was repeated for all 10 subsets, allowing a comprehensive evaluation of the model's performance across different data subsets. The results of cross-validation show that the accuracy of the model is 94.78%, indicating that in the classification task, the model has a very high proportion of correct classification on the test data. Kappa value is 0.703, indicating that the classification result of the model is much higher than the random guess level, indicating the consistency and reliability of the model classification. These indicators show that the model is stable on different data subsets, with strong predictive power and high accuracy.

4. Conclusion

This paper employed a logistic regression model to thoroughly analyze diabetes risk factors using 100,001 electronic health record (EHR) datasets and validated the model's performance from multiple perspectives. The results showed that hemoglobin A1c level (HbA1c.level) had the most significant effect on diabetes risk. Other relevant variables, including blood glucose level and body mass index (BMI), demonstrated significant positive correlations, particularly with hypertension and heart disease. This suggests that managing these chronic conditions is crucial for reducing diabetes risk.

Both gender and age showed significant effects in the model, with men having a relatively high risk of diabetes and increasing age also increasing the logarithmic odds of diabetes. Although smoking history did not show a significant effect in this model, this finding needs further validation and exploration.

The model demonstrated a high goodness of fit and excellent data fitting ability, confusion matrix and ROC curve showed strong classification ability, good accuracy and other performance indicators. 10-fold cross-validation confirmed the model's stability, reliability, and consistency across various data subsets. Despite the overall excellent performance, the recall ability of the model in a few classes still needs to be improved. Future work should focus on optimizing this aspect to further improve the forecasting effect.

References

[1] Ling X J and Wang J J 2024 Diabetes forecast based on machine learning algorithm. Journal of modern information technology, 8(14), 59-63 + 68.

[2] International Diabetes Federation. IDF Diabetes Atlas, 10th edition. Brussels, Belgium: 2021. Website: https://www.diabetesatlas.org

[3] Gao Z X and Liu J F 2019 Research progress on the classification and application of adult diabetes mellitus. Clinical Review, 39(07), 650-653.

[4] Zhang Y X, He S and You S M 2019 Application of ensemble learning in diabetes prediction. Intelligent Computers and Applications, 9(5), 176-179.

[5] Wen X C, Ma X Y and Gong C J 2023 Progress in diabetic kidney disease risk factors and prediction model building. Journal of liaoning university of traditional Chinese medicine, 25(01), 161-170.

[6] Liu R Y, Qu Y M, Liu X, et al. 2023 Integrated learning and decision tree in the prospective risk assessment on the application of type 2 diabetes. China's chronic disease prevention and control, 31(4), 278-283+288.

DOI: 10.54254/2753-8818/42/2024CH0212

[7] Zhang Y, et al. 2024 The pathogenesis of diabetes complications and treatment of drug research progress. Chinese pharmacological bulletin, 10, 1808-1813.
[8] Chen X Y, CAI W Q, Wang S Z, et al. 2019 Construction and verification of prediction model for diabetic retinopathy. International Journal of Ophthalmology, 24(08), 1297-1302.
[9] Su L W, Wu Y J, Zhu Y, et al. 2024 Based on the theory of target state syndrome differentiation treatment of diabetic cardiomyopathy. Chinese medicine.
[10] Dai G Y, Zhao H, Tang D Q, et al. 2019 Advances in imaging studies of diabetes-associated aortic calcification plaques. Modern Medicine & Hygiene, 40(17), 2995-3000.
[11] Wang P 2021 Research on pathologic speech detection and classification based on acoustic and kinematic characteristics. Taiyuan University of Technology.
[12] Duan J Y and Wei S H 2019 Effect of therapeutic communication on anxiety and depression in patients with Parkinson's disease. Nursing Research, 37(19), 3592-3596.
[13] Mou X G, Tao J X and Chen L 2023 Parkinson's Disease diagnosis based on speech feature fusion. Digital Manufacturing Science, 21(3), 225-230.
[14] Chen Z Y, Tan W Z, Zheng J H, et al. 2024 Strategy to explore the diagnosis and treatment for diabetes. Lifescience, 1-18.
[15] Wang M, Yang M Z, Ding X, et al. 2019 Visual analysis of community diabetes management research based on Cite Space. Chinese Journal of Primary Health Care, 38(06), 1-5.
[16] Wu Y 2024 Early diagnosis model of diabetes mellitus based on machine learning and interpretable analysis. Chongqing university of science and technology.
[17] Fang Y H, Li Z W, Xu Y Y, et al. 2024 Patients with bladder cancer survival prediction model based on machine learning research. Journal of modern information technology, 8(16), 83-87.
[18] Zhang X X, Zhang Z T, Liu L, et al. 2019 Research and Implementation of Assistive device adaptation Model Based on Decision Tree and logistic Regression Algorithm. Chinese Journal of Rehabilitation Medicine, 38(8), 1108-1113.
[19] Jiang M Y and Zhang H. 2024 Study on the prediction efficiency of machine learning algorithm for heart disease. Chinese Journal of Medical Physics, 41(07), 905-909.
[20] Zhang X Y and Yuan H J 2020 Regularization and cross validation in the application of combination forecast model. Computer system application, 29(4), 18-23.