Research on Diabetes based on the Logistic Regression Model

Zhixiang Zhou *

School of Business Administration, Southwestern University of Finance and Economics, Henan, China

* Corresponding Author Email: 42412020@smail.swufe.edu.cn

Abstract. Diabetes, as a globally prevalent metabolic disease, poses a serious threat to public health. Due to the subtle early symptoms of diabetes, most patients have already developed complications by the time they are diagnosed. Therefore, risk prediction becomes crucial for disease prevention and control. This study employs a logistic regression model to conduct predictive analysis on a dataset from Kaggle and utilizes a confusion matrix to determine the accuracy of the model's predictions, while also performing comparative auxiliary analysis. The results showed that the model had a prediction accuracy of 74.500%, indicating good predictive performance. This study holds dual value. At the clinical practice level, the use of logistic regression models can assist in identifying high-risk populations and implementing early intervention to reduce the risk of disease, providing significant guidance for primary care. In terms of diabetes prevention, people can start by focusing on factors that affect diabetes, identifying which factors have a more significant impact on the development of diabetes, and thereby reducing the risk of developing diabetes.

Keywords: Diabetes, confusion matrix, logistic regression model.

1. Introduction

Diabetes and its complications have become one of the major health burdens and causes of death globally. Diabetes facts and figures indicate that the global diabetes burden on individuals, families, and countries is increasing. In 2020, the estimated number of people aged 65 and above with diabetes worldwide was 727 million, and this figure is expected to more than double to 1.5 billion by 2050 [1]. In addition, the International Diabetes Federation (IDF) Diabetes Atlas 2025 report states that 11.1% (or 1 in 9) of people have diabetes, and 4 in 10 people are unaware that they have the disease [2]. Its high rates of disability and premature mortality constitute a global public health crisis that urgently needs to be addressed. As a preventable disease, early screening and intervention for diabetes are crucial, so the accuracy and reliability of model prediction results deserve special attention from researchers [3].

Regarding the prediction of diabetes, scholars have utilized many different methods for predictive analysis. Zaman et al. proposed a diabetes classification model using support vector machine (SVM), naive Bayes, random forest (RF), and decision tree (DT). They utilized the PIMA diabetes dataset and achieved an accuracy rate of 81% through a Naive Bayes classifier [4]. Amilo has adopted seven machine learning algorithms, including XGBoost and SVM, to identify key risk factors for diabetes, with XGBoost demonstrating superior performance [5]. Xiang Junjie, a scholar, compared and analyzed the accuracy of predicting whether people have diabetes using the k-Nearest Neighbor (KNN) algorithm, logistic regression prediction, and decision tree prediction methods. The results showed that logistic regression excels in the quantitative metric of recall rate and performs well for linear data, but may not be able to handle complex nonlinear data in some cases [6]. Currently, the use of clinical indicators for diabetes classification lacks stability and accuracy, and further exploration and verification will still require large-scale clinical data in the future [1]. In addition, machine learning exhibits unique advantages in data integration, pattern recognition, and prediction accuracy, supporting precise predictions and personalized treatment strategies [7]. Researchers can still further predict and discuss diabetes predictions by using machine learning.

This study will utilize a logistic regression model to conduct predictive analysis on whether individuals suffer from diabetes, while verifying the applicability and accuracy of the logistic

regression model in predicting diabetes, and identifying the risk factors that affect individuals' susceptibility to diabetes.

2. Methods

2.1. Data Source and Description

This study selected the synthetic diabetes dataset from the Kaggle platform, which is owned by Mr Simple and has an availability score of 10.0 [8]. In addition, the dataset consists of 1000 sample data points and 9 variables. The dataset contains a series of features related to diabetes risk factors. These characteristics encompass variables such as blood glucose levels, body mass index (BMI), age, family history, and other pertinent health indicators. Each set of feature values is accompanied by a diagnostic label indicating whether the individual has diabetes. This dataset is highly valuable for training machine learning models to predict the likelihood of developing diabetes based on the provided risk factors. It can be used for the development of research, analysis, and prediction models aimed at improving the diagnosis and management of diabetes.

2.2. Indicator Selection

The dataset consists of 1000 sample data points and 9 variables. There are no missing values in the sample data. The diagnosis of diabetes is selected as the dependent variable, with positive=1 and negative=0. The remaining variables are independent, and $x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8$ and y are used to correspond to each variable for simplification of variable representation. The names and explanations of each variable are shown in Table 1.

Table 1. Introduction to each variable

Explanatory variable	Name of Variables	Definition of the Variables
y	Diagnosis	Binary label indicating whether the individual has diabetes (1) or not (0).
x_1	Pregnancies	The number of pregnancies the individual has had.
x_2	Glucose	Plasma glucose concentration (mg/dL) measured during an oral glucose tolerance test.
x_3	Blood Pressure	Diastolic blood pressure (mm Hg).
x_4	Skin Thickness	Thickness of skin fold (mm) at the triceps.
x_5	Insulin	2-Hour serum insulin(mu U/ml).
x_6	BMI	BMI (body mass index) is calculated by weight (kg) divided by height (m)² to assess body fat and health risks
x_7	Diabetes Pedigree	Diabetes pedigree function, which represents the likelihood of diabetes based on family history
x_8	Age	Age of the individual (years)

2.3. Method Introduction

Firstly, the dataset is divided into a training set and a test set. The training set contains 80% of the cases in the dataset, while the test set accounts for 20% of the total sample size. Secondly, the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (lbfgs) optimization algorithm is used for modeling, combined with L2 regularization, to prevent overfitting of the model and improve its generalization ability. In this study, the LBFGS algorithm was employed for model parameter optimization. As an improved form of the quasi-Newton method, lbfgs stores historical iteration information through

limited memory to approximate the inverse Hessian matrix, reducing computational complexity while maintaining high optimization efficiency, making it suitable for high-dimensional parameter scenarios. In addition, to alleviate the overfitting problem, the model introduces L2 regularization, which constrains the scale of parameters by adding a parameter L2 norm penalty term to the loss function to enhance generalization ability. The ultimate optimization objective is a loss function with regularization, where parameters are iteratively updated using LBFGS until convergence. Finally, this study will utilize a confusion matrix to analyze the prediction results of the logistic regression model, to evaluate its classification performance for positive and negative diabetes samples. Based on the accuracy of the model and comparative analysis of the data, reasonable prevention suggestions can be provided.

Logistic regression is a widely used statistical learning model for binary classification tasks. It constructs a probability model to predict the probability of a sample belonging to a certain category, and then determines the category based on a probability threshold. The calculation formula used in this study is as follows:

$$g(x) = \frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2+\cdots+\theta_n x_n)}} \quad (1)$$

Among them, g(x) represents the prediction probability within the range of (0, 1); $\theta_0, \theta_1, \theta_2 \cdots \theta_n$ are model parameters; $x_1, x_2 \cdots x_n$ are input features.

3. Results and Discussion

3.1. Spearman Correlation Analysis

Before conducting logistic regression analysis, this article first performed a Spearman correlation analysis to quantify the degree and direction of linear correlation between two continuous variables. From the Spearman correlation analysis (triangle format) in Table 2, it can be seen that the absolute values of the correlation coefficients between the dependent variable y and each independent variable ($x_1 \sim x_8$) are all ≤0.061, and there are no significant markers, indicating that there is neither a significant monotonic association nor a strong correlation between y and all independent variables. Only three pairs of independent variables exhibit statistically significant weak correlations: x_4 and x_6 (r=0.070*), x_4 and x_7 (r=-0.073*), x_1 and x_8 (r=-0.070*). There are no statistically significant correlations between the remaining independent variables, and the absolute values of the correlation coefficients are generally small. The dataset can be analyzed and predicted using a logistic regression model.

Table 2. Spearman correlation - triangular line format

	y	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
y	1								
x_1	0.004	1							
x_2	0.001	0.009	1						
x_3	0.017	0.052	0.015	1					
x_4	-0.008	-0.043	0.006	-0.048	1				
x_5	0.032	-0.027	-0.019	0.039	0.037	1			
x_6	-0.025	-0.017	0.010	0.005	0.070*	0.052	1		
x_7	0.061	-0.003	0.007	0.003	-0.073*	-0.022	-0.019	1	
x_8	-0.037	0.070*	-0.017	0.050	-0.051	0.044	-0.054	0.027	1
* p<0.05 ** p<0.01									

3.2. Logistic Regression Model

The logistic regression model for predicting diabetes was constructed based on 1000 patient samples, with 80% (n=800) used for training and 20% (n=200) for testing. Key physiological indicators include pregnancy history, blood glucose levels, blood pressure values, skin thickness, insulin usage, body mass index (BMI), and family history of diabetes. The accuracy of this model in predicting whether a person has diabetes is 74.500%. As shown in Fig. 1, the vast majority of patients predicted as negative by the logistic regression model were indeed negative. Among the 200 test set samples, 149 were correctly predicted, indicating a certain predictive performance. Based on the model evaluation results of the training and test sets in Table 3, the training set has 800 samples, while the test set has 200 samples. Among the two types of data, the precision, recall, and F1-score for item 1.0 are all 0, indicating that the model has deficiencies in recognizing this category. The training set has a precision of 0.68 and an F1-score of 0.81 for item 0.0, while the test set has a precision of 0.74 and an F1-score of 0.85 for item 0.0. The better performance of the test set for item 0.0 indicates that the model has improved its recognition accuracy for this category during the testing phase. In terms of accuracy, the training set has an accuracy of 0.68, while the test set has an accuracy of 0.74, indicating that the overall recognition accuracy of the test set is higher. The precision, recall, and f1-score of the average (comprehensive) values on the test set (0.56, 0.74, 0.64) are all superior to those on the training set (0.46, 0.68, 0.55), indicating that the model has better comprehensive performance on the test set. However, due to the failure of class 1.0 recognition, there is still considerable room for optimization overall. In the future, emphasis should be placed on improving the recognition ability of class 1.0 samples and enhancing the comprehensiveness and effectiveness of model classification.

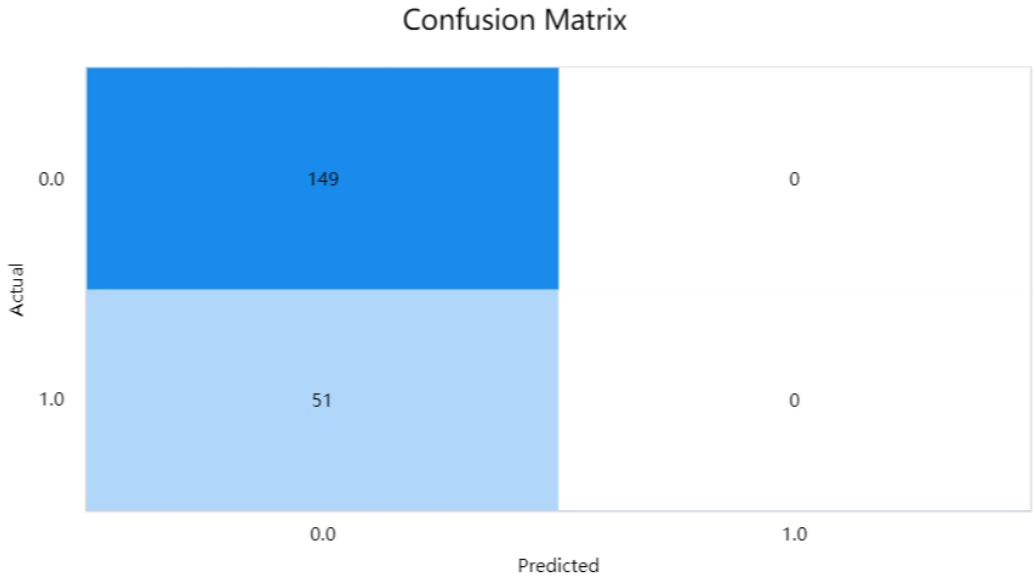

Fig. 1 Confusion matrix (Photo/Picture credit: Original).

Table 3. Evaluation Results

	precision	recall	f1-score	sample size	accuracy
Training set average (comprehensive)	0.34	0.68	0.55	800	0.68
0.0 (training set)	0.68	1.00	0.81	545	
1.0 (training set)	0.00	0.00	0.00	255	
Test set average (comprehensive)	0.56	0.74	0.64	200	0.74
0.0 (test set)	0.74	1.00	0.85	149	
1.0 (test set)	0.00	0.00	0.00	51	

3.3. Comparative Data Analysis

This study divided all data samples into two categories: positive (red) and negative (blue), and compared the mean values of the two categories. As shown in Fig. 2, there were no significant differences in variables such as age, BMI, skin thickness, blood pressure, blood glucose, and number of pregnancies between the two groups. The average genetic function value of diabetes in the positive group was 0.47, slightly higher than that in the negative group. However, after the overall sample was injected with 2 hours of serum insulin (mu U/ml), the insulin content in the positive group was 5.75mu U/ml higher than that in the negative group. Therefore, insulin levels may be an important indicator for distinguishing diabetes diagnosis results.

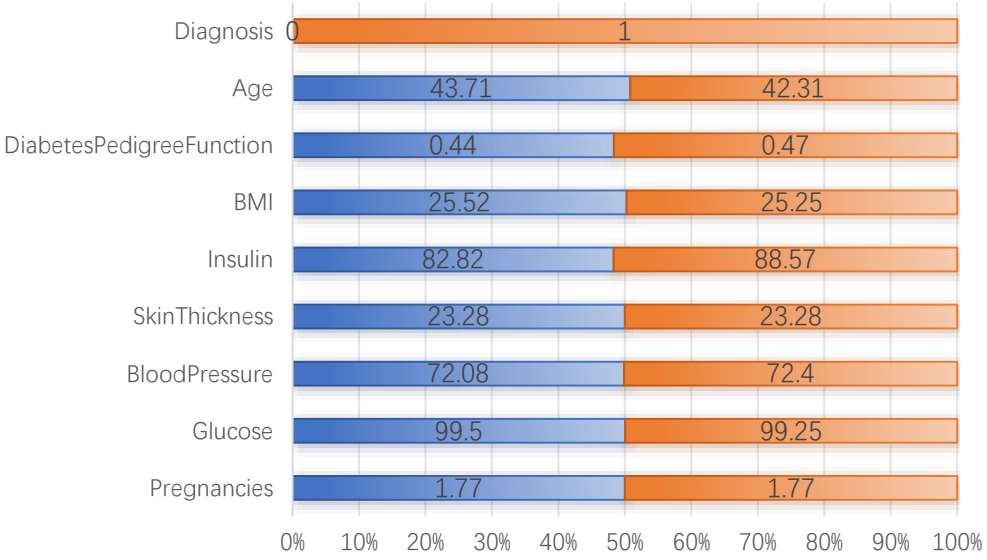

Fig. 2 Bar percentage stacked chart of the diabetes positive group and the negative group (Photo/Picture credit: Original).

3.4. Discussion

In this study, the logistic regression model exhibits certain variability in predicting diabetes. Specifically, the model performs well in identifying negative cases of diabetes (class 0.0), achieving an accuracy rate and F1-score of 0.74 and 0.85, respectively, in the test set, with an overall accuracy of 74.500%. This indicates that the model has certain application value in negative prediction. However, the model completely fails to identify positive cases (class 1.0). This phenomenon may be related to data imbalance - the proportion of positive samples in the dataset is relatively low, with 255 positive samples in the training set and 51 positive samples in the test set. This is consistent with the conclusion proposed by Gupta et al. that dataset imbalance can affect model performance [3].

Further Spearman analysis results showed that the correlation between the dependent variable and the independent variable was extremely weak, and only insulin levels exhibited significant differences between the positive and negative groups. This suggests that a single logistic regression model may struggle to capture the complex influence relationships in diabetes prediction, and also confirms Xiang Junjie's viewpoint on the limitations of logistic regression in handling nonlinear data [5]. Nan Rui and other scholars have concluded through research that compared to other machine learning models, XGBoost's accuracy, recall rate, and AUC value have all increased by about 2% [9]. This indicates that other machine learning algorithms may be more suitable for diabetes prediction.

This study has certain limitations. Firstly, the sample may be limited to the female population, which to some extent restricts the universality of the research conclusions. Secondly, the model fails to fully explore the correlations between features, affecting the identification of positive samples.

Based on the aforementioned analysis, future research can be improved in the following aspects. Firstly, the sample scope should be expanded to break gender constraints, thereby enhancing the

applicability of the model. Secondly, measures should be taken to balance data distribution and address the issue of a low proportion of positive samples. Finally, one can attempt to utilize other machine learning algorithms such as Random Forest, XGBoost, Catboost, and k-Nearest Neighbors. By comparing the accuracy of different models, a more optimal predictive model can be selected, thereby enhancing the identification capability for positive samples and the overall performance of the model. In addition, effective control of blood glucose can help improve patients' quality of life [10]. Studies have found that there is a significant difference in insulin levels between the positive and negative groups 2 hours after insulin injection, suggesting that this indicator may be a key factor in predicting diabetes. Future research could pay more attention to this.

4. Conclusion

In summary, by analyzing and predicting the logistic regression model using a confusion matrix, it is found that the accuracy of the logistic regression model in predicting whether people have diabetes is 74.500%. Although it can predict whether people have diabetes to a certain extent, for real applications in the medical field, the prediction accuracy of the model generally needs to reach over 90%. Clearly, the logistic regression model may not be the most effective and accurate model for predicting diabetes. If a model for predicting diabetes is to be selected in the future, other machine learning algorithms such as random forest, k-nearest neighbors, or Catboost can be considered. The optimal predictive model can be determined by comparing the accuracy of each model. Meanwhile, this study has learned through comparing the diabetes-positive group and the negative group that the insulin content 2 hours after insulin injection may serve as a key factor in predicting whether a person has diabetes. It should be noted that the low accuracy of the logistic regression model may stem from the limited selection of sample subjects, which only focuses on whether female groups have diabetes. Future researchers can start by expanding the sample subjects to make the predictions of the logistic regression model more accurate.

References

[1] Weihao W, Qi P. The Current Application of Artificial Intelligence in the New Classification of Diabetes. China Medical Frontiers Journal (Electronic Edition), 2023, 15 (12): 91-92.

[2] IDF. Diabetes Atlas. April.7, 2025.July.19, 2025.https://idf.org/about-diabetes/diabetes-facts-figures/

[3] Lina W. Do You Know About the Early Diagnosis of Diabetes?. Gansu Science and Technology Daily, 2024-08-29 (007).

[4] Gupta S, Goel N. Performance Evaluation of Diabetes Prediction Model Based on Imbalanced Dataset Using Feature Selection and Hyperparameter Tuning of Classifiers. Cureus Journal of Computer Science, 2025, 2 (1):4842-4842.

[5] Amilo D, Sadri K, Hincal E, et al. An integrated machine learning and fractional calculus approach to predicting diabetes risk in women. Healthcare Analytics, 2025, 8100402-100402.

[6] Junjie X. Application of Machine Learning in Predicting Diabetes. China Science & Technology Information, 2025, (14): 77-80.

[7] Zhenkun F, Lisha M. Precision Diabetes Treatment: The Potential of Multi-omics and Machine Learning in Islet Transplantation. Organ Transplantation, 2025, 16 (04): 626-631.

[8] MrSimple. Diabetes Prediction. 1/1/2024. 7/21/2025. https://www.kaggle.com/datasets/mrsimple07/diabetes-prediction.

[9] Rui N, Yanli L, Qiyang W, et al. Diabetes Prediction and Factor Analysis Based on Machine Learning. Intelligent Computer and Applications, 2025, 15 (06): 140-145.

[10] Qian Z, Shuxian Q, Xi J, et al. The impact of multimedia-based continuous care on blood glucose levels and cardiovascular disease risk factors in elderly patients with diabetes. Nursing Practice and Research, 2025, 22 (08):1187-1192.