Dean&Francis
ISSN 2959-6157

A Machine Learning-Based Prediction Study for Type 2 Diabetic Mellitus

Yuhao Xia

School of Statistics and Mathematics, Shanghai Lixin University of Accounting and Finance, Shanghai, 20120, China

221330134@stu.lixin.edu.cn

Abstract:

Diabetes mellitus is one of the major diseases in the world. The pathogenesis is complex and difficult to cure, and accurate prediction of the risk of diabetes can help improve the treatment rate. Machine learning, as an important branch of artificial intelligence, can discover potential relationships between data, to efficiently carry out disease prediction and risk assessment. In this paper, a logistic regression prediction model was constructed, and an authoritative dataset was used as the research object, including 8 medical characteristics as the variables for predicting diabetes. First, some data pretreatment steps were taken to eliminate any invalid or missing data, and then the model was trained the prediction test set was carried out, and finally, the five main evaluation indicators of the model were obtained. The results showed that the logistic regression model had an accuracy rate of 0.95, a precision rate of 0.85, a recall rate of 0.63, an F1 Score of 0.73, and an AUC Value of 0.96. The results show that the prediction model has good stability and accuracy, can effectively predict diabetes, and has certain potential clinical application value.

Keywords: Diabetes prediction, machine learning, logistic regression.

1 Introduction

Globally, diabetes is becoming a significant public health issue due to its rising incidence and prevalence. The number of individuals with diabetes is predicted to rise from 537 million in 2021 to 643 million by 2030, according to the International Diabetes Federation and others [1]. This trend highlights the urgency and importance of diabetes prevention and control. Diabetes, a generally chronic condition, is brought on by either insufficient insulin production by the pancreas or ineffective insulin utilization by the body. Individuals with diabetes are at least twice as likely to die as their counterparts without the disease [2].

Type 2 diabetes mellitus (T2DM) is a chronic disease affected by a combination of host genetic and environmental factors, and its pathogenesis is complex and diverse, some patients with T2DM are often undetectable due to lack of relevant knowledge or asymptomatic [3]. Traditional medical methods are difficult to predict and accurately predict and assess the risk of diabetes in patients. Therefore, it is neces-

ISSN 2959-6157

sary to construct an appropriate method to predict T2DM in a timely and effective manner, to achieve early detection and early treatment to avoid further deterioration of T2DM.

The methodical process of learning and training from data and making precise predictions is known as machine learning. It is a sort of artificial intelligence used in bioinformatics data processing and is a multidisciplinary field of knowledge. Machine learning methods are better than most traditional statistical methods by processing and analyzing complex data sets, discovering potential relationships and patterns between data, digging deep into latent patterns in medical data, and constructing high-precision prediction models [4]. Machine learning prediction research can also provide a scientific basis for the research and formulation of prevention and control strategies for diabetes, and its powerful data processing and pattern recognition capabilities provide new ideas and methods for the prediction and intervention of diabetes, which is helpful for detection and intervention in the early stage of the disease, and reduces the incidence of complications and mortality.

Big data technology is developing quickly, and machine learning algorithms are always being improved. A large amount of patient data has been accumulated in the medical field, including multi-dimensional data such as biochemical indicators, genetic information, and lifestyle habits. These data provide a rich data source and powerful technical support for the use of machine learning methods for diabetes prediction. Previous studies have shown that the main influencing factors of type 2 diabetes are blood pressure, triglycerides, total cholesterol, age, and many other factors [5, 6]. The T2DM prediction model based on machine learning methods can conduct a personalized risk assessment based on the specific situation of the patient (the multiple influencing factors mentioned above), which helps provide more accurate health guidance and intervention for patients. At the same time, it can also deeply understand the pathogenesis and influencing factors of diabetes, and provide support for formulating more effective prevention and control strategies.

The logistic regression model has a high ability in the prediction of many diseases, such as cardiovascular disease, heart disease, and other major diseases [7, 8]. For example, according to the research results of Zhao, the accuracy of the logistic regression model in predicting heart disease reaches 0.85, indicating that the logistic regression prediction model has high accuracy [8]. And because the mathematical principle and calculation process behind it are relatively intuitive and simple, easy to understand and implement, the logistic regression model is the first choice in this paper to achieve the prediction of diabetes.

This paper takes the authoritative data set in Kaggle as the research object, selects the logistic regression prediction model, and takes the accuracy and stability of the results as the evaluation criteria for its effectiveness. The importance of influencing factors of different dimensions of type 2 diabetes was studied to improve the accuracy and stability of type 2 diabetes. This paper aims to provide timely and effective preventive measures for potential patients and assist doctors in early detection and treatment to reduce the risk of disease onset and deterioration.

2 Research Object and Method

In reality, logistic regression is a popular machine-learning technique for categorization issues. The core idea is that the logic function maps the output of linear regression into the interval (0, 1), which converts the probability value into a specific class label (such as 0 or 1) according to a set threshold value (usually 0.5).

Logistic regression is widely used in many fields because of its simplicity, efficiency, and interpretability, especially those scenarios that require binary prediction. Simultaneously, the logistic regression model may provide the weight and degree of contribution of each characteristic to the prediction result, which aids in determining which factors significantly affect the outcome. Secondly, logistic regression is relatively simple to compute, and when dealing with some large-scale data sets, its training speed is usually faster and more efficient than some complex deep learning models. In addition, the logistic regression model can also be extended to multi-classification problems through some strategies, such as one-to-many, many-to-many, etc., which makes the logistic regression model more flexible.

The data selected for this study came from the diabetes prediction dataset on the Kaggle website. Its data sets are obtained from a variety of sources, such as research in some professional fields, healthcare institutions, or other data providers, so the number is relatively large [9]. This dataset is commonly used in the field of diabetes prediction and has been extensively studied and analyzed by researchers and data scientists with a high degree of confidence in its data. The diabetes prediction dataset has a total of 100,000 original sample data, including 8 medical predictors and 1 outcome variable, as shown in Table 1.

Table 1. Analysis of medical characteristics of diabetes data set

Feature name	Analysis of medical characteristics
Gender	Patients' gender
Age	Patients' age
Hypertension	1: having hypertension, 0: not having hypertension
Heart disease	1: having heart disease, 0: not having heart disease

Continue Table 1.

Smoking history	1: smoking currently, 0.5: smoking formerly, 0: never smoking
BMI	Patients' Body Mass Index
HbA1c level	Patients' Hemoglobin A1c level
Blood glucose level	Patients' Blood glucose level
Diabetes	1: having diabetes, 0: not having diabetes

In the original diabetes prediction dataset, 91,500 individuals did not have diabetes, while 8,500 patients did.

To increase the precision and consistency of machine learning model predictions, this research must extract more representative and predictive characteristics from the original data set. To accomplish this, the original data must be pre-processed, the significantly overlapping cases of patients with or without diabetes must be removed, the cases without smoking history information, or the data that is not suitable for the model or inaccurate, and then defined the three characteristics of hypertension, heart disease, and smoking history as numerical variables. To better predict the model, then lastly ensure that the pre-processed data completely satisfies every prediction model criterion.

In the revised data set, there were 64,184 valid cases, of which 7046 were diabetic and 57,138 were non-diabetic.

3. Experimental Results

3.1 Experimental Design

First, import the data set file. Then the feature variable and the target variable are separated, and its function is to facilitate the prediction of the target variable based on the feature variable. Define numerical features and categorical features, which play different roles in machine learning, to help describe discrete data and type features of the data. Then different feature types are processed, including unique thermal coding of classification features and data preprocessing, which can effectively improve the accuracy of model prediction. The training and test sets, training model, and prediction test sets are then separated from the data set. Lastly, the model is assessed, and the primary evaluation metrics are produced.

The five main evaluation indicators of this experiment are as follows:

Accuracy: The most logical performance metric is accuracy, which shows the percentage of samples that the model accurately predicted out of all the samples.

Precision: Out of all the samples that the model predicted, precision is the percentage of samples that are truly positive examples.

Recall: The percentage of samples that the model accurately identified as positive examples out of all samples that are in fact positive examples is known as recall. A high recall rate indicates that the majority of real positive cases can be found by the model.

F1 Score: The F1 Score, which is used to thoroughly assess the model's performance, is the harmonic average of the accuracy rate and recall rate. An improved balance between accuracy and recall is indicated by a higher F1 score. A more thorough evaluation index is the F1 score, which can provide a compromise assessment, particularly when the accuracy rate and recall rate are at odds.

AUC Value: AUC Value mainly measures the ranking ability of the model, that is, whether the model can make the prediction probability of the positive sample higher than that of the negative sample.

ISSN 2959-6157

3.2 Experimental Results

Table 2 shows the performance evaluation results of the logistic regression model.

Table 2. Logistic regression evaluation index table

Data Set	Accuracy	Precision	Recall	F1 Score	AUC Value
Test Set	0.95	0.85	0.63	0.73	0.96

As can be seen from Table 2, the logistic regression model is very accurate in predicting on the whole. The accuracy rate of 0.85 indicates that the model has high accuracy in predicting positive samples. Additionally, the experimental results' F1 score is impacted by the poor recall rate. The model is also quite good at differentiating between positive and negative samples. Overall, the performance of the logistic regression model is good, it can effectively predict most samples and has good classification effect and generalization ability, but there is a certain imbalance between the accuracy rate and the recall rate.

4. Discussions

The logistic regression model was used to predict diabetes risk in the test set, and the results are shown in Table 2. As can be seen from Table 2, the analysis and comparison of the prediction results on the test set found that the logistic regression model performed well in the accuracy, F1 score, and AUC Value of the logistic regression model, indicating that the logistic regression model has high stability and accuracy and good classification ability in the binary classification of diabetes prediction.

However, the risk of diabetes is often related to a non-linear combination of multiple factors, and there are often some complex relationships among its features, such as hierarchical or interdependent features [10, 11]. The predictive ability of logistic regression is limited when dealing with complex and nonlinear data. In addition, logistic regression is usually used to solve the binary classification problem, and the prediction of diabetes may sometimes involve multiple categories, such as pre-diabetes, post-diabetes, etc., at which time logistic regression cannot give an accurate prediction. However, this problem can be solved by using multiple logistic regression prediction models, and a more accurate result can be obtained after many experiments. In short, logistic regression has a wide range of applications. Although the prediction may be less accurate when dealing with the data of complex relationships, this model can still be chosen by most prediction research institutes due to its simplicity and efficiency.

In the future, the application of the logistic regression model in the prediction of diabetes can be combined with genomics, metabolomics, and other disciplines to explore the relationship between genetic and environmental factors and the incidence of diabetes [12, 13].

5. Conclusions

Based on the patient's age, hypertension, heart disease, smoking history, BMI, hemoglobin level, blood sugar level, and other characteristics, this paper constructed a T2DM prediction model based on the logistic regression machine learning method, which can be effectively applied to the physiological characteristics of different patients to predict disease risk. The obtained data set was preprocessed, and 64,184 valid cases were obtained. After training the model and testing the model, the main evaluation indexes of the five models were obtained, which were accuracy rate, precision rate, recall rate, F1 Score, and AUC Value, respectively. Analysis of the logistic regression assessment indices revealed that the model had good classification capacity, high stability and accuracy, and a good prediction effect. However, there are some defects in the ability to accurately predict positive samples and further analysis is needed to effectively solve this problem.

In conclusion, the logistic regression model has a strong overall prediction ability, which can effectively and accurately achieve T2DM prediction and early screening, so that the risk population can receive early prevention and treatment, so as to reduce the incidence of diabetes. In addition, more targeted personalized treatment programs can be developed to reduce unnecessary medical costs and ultimately achieve accurate treatment of T2DM.

References

[1] International Diabetes Federation. IDF diabetes atlas. Brussels: International Diabetes Federation, 2021.

[2] Ebrahim O, Derbew G. Application of supervised machine learning algorithms for classification and prediction of type-2 diabetes disease status in Afar regional state, Northeastern Ethiopia 2021. Scientific Reports, 2023, 13(1): 7779.

[3] Li L. Machine learning for predicting diabetes risk in western China adults. Diabetology & Metabolic Syndrome, 2023, 15(1): 165.

[4] Gu J. A personalized mRNA signature for predicting hypertrophic cardiomyopathy applying machine learning

methods. Scientific Reports, 2024, 14(1): 17023.
[5] Mansoori A. Prediction of type 2 diabetes mellitus using hematological factors based on machine learning approaches: a cohort study analysis. Scientific Reports, 2023, 13(1): 663.
[6] Ling X, Wang J. Prediction of diabetes based on machine learning algorithms. Modern Information Technology, 2024, 8(14): 59-63+68.
[7] Fang Y, Zhang J. The cumulative and single effect of 12 aldehydes concentrations on cardiovascular diseases: An analysis based on Bayesian kernel machine regression and weighted logistic regression. Reviews in Cardiovascular Medicine, 2024, 25(6): 206.
[8] Zhao H. Visualization analysis and logistic regression-based heart disease risk prediction. Proceedings of the 5th International Conference on Computing and Data Science (part2), Faculty of Medical and Health Sciences and Bioengineering Institute, University of Auckland, ITM Department, Illinois Institute of Technology, USA, Department of Computer Science and Technology, Shandong University of Finance and Economics, 2023: 8.
[9] Mustafa M. Diabetes prediction dataset, 2023.9.27, 2024.9.27. https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset
[10] Ashwini T, Devi S P. Machine learning for diabetes clinical decision support: A review. Advances in Computational Intelligence, 2022, 2(2): 22-22.
[11] Guttikonda R, Pothu UK. Role of biochemical parameters in prediction of diabetic peripheral neuropathy. Journal of Research in Applied and Basic Medical Sciences, 2024, 10(2): 169-177.
[12] Guo J, Gao Y, Gao H, et al. Comparative study of type 2 diabetes risk prediction models. Chinese Journal of Bioengineering, 2023, 43(11): 35-42.
[13] Carrasco-Zanini J. Multi-omic prediction of incident type 2 diabetes. Diabetologia, 2024, 67(1): 102-112.