Proceedings of the 5th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/112/2024.17911

Research of Some Factors Leading to Diabetes among Women by Logistic Regression Model

Qihan Wu[1,a,*]

[1]School of Mathematical Sciences, Zhejiang University, Zhejiang, 310058, China
a. 3220103534@zju.edu.cn
**corresponding author*

Abstract: With the continuous development of the economic level, diabetes has become an increasingly concerning problem, and the incidence of diabetes is also rising, which has caused huge losses to human health and economic development. It is very important to build a self-improving prediction model with considerable accuracy. As the cause of diabetes is complex and there are differences between people, this paper selects a widely recognized medical diabetes dataset for research, which contains 768 samples and 9 variables. After grouping, the data set is simulated and predicted by the logistic regression model, and it is found that the accuracy reaches 77.9% without grouping. After grouping, the accuracy reached 79.5% and 86.9%, respectively. It was found that the logistic model could effectively predict diabetes. At the same time, reasonable classification of the diabetes data set could effectively improve the accuracy of the prediction model. The main risk factors were Diabetes Pedigree Function, BMI, pregnancy, and Glucose. This provides some new perspectives and ideas for future research.

Keywords: Diabetes, Logistic regression model, Characteristic coefficient.

1. Introduction

Diabetes is a globally prevalent chronic disease that can lead to serious complications such as diabetic feet, vision loss, and kidney failure. It is also one of the fastest-growing non-communicable diseases in the world. It is predicted that the number of people with diabetes worldwide will reach 643 million by 2030, and this number will increase to 783 million by 2045 [1]. The rate of death and disability caused by diabetes is continuing to rise, and its economic losses are incalculable [2]. Type 2 diabetes is the most typical and most common form of diabetes, accounting for 90 percent of all cases. Unlike the cause of type 1 diabetes, type 2 diabetes is often caused by a combination of environmental factors and gene expression. Examples include obesity, poor eating habits during pregnancy, and family history of the disease. Therefore, most people with type 2 diabetes can fight diabetes by eating a healthy diet, exercising properly, and intervening early to delay its onset or even avoid it. However, awareness of diabetes is not high, as shown in the survey, half of adults with diabetes worldwide do not know they have it. Therefore, establishing an effective prediction model, through data collection and intelligent management, to remind vulnerable or suspected patients to take active and effective measures is one of the important ways to fight type 2 diabetes [3].

The causes of type 2 diabetes are very complex, it also varies among different populations in different regions. Its occurrence is a complex expression of muti-factors and muti-steps. Therefore, in the research, it is necessary to select aspects of a specific gender in specific areas with comparative research value to explore. For women, obesity, pregnancy, and a family history of diabetes are all factors contributing to diabetes [4,5]. Many women are at high risk of developing gestational diabetes during pregnancy or after pregnancy, and newborns may also be born with diabetes. Even the same woman has different rates of gestational diabetes at different ages and with different medical conditions. The prediction of gestational diabetes risk and the avoidance of maternal diabetes risk after pregnancy are very important for the well-being of newborns and pregnant women [6]. Existing studies usually focus on the impact of a single indicator on the development of diabetes in women, and there is a lack of models that comprehensively consider the differences in multiple indicators among individuals. This makes it difficult to provide accurate diabetes risk prediction for individuals in practical applications. For example, in women of normal weight who have a family history of diabetes, the risk of diabetes at different stages of pregnancy is difficult to assess by a single indicator. Therefore, it is important to build a prediction model that integrates multiple individual differences. At present, the existing machine learning methods for some diabetes data sets, such as random forest, decision tree, Deep Convolutional Neural Network Based-Bayesian Optimization, etc. have achieved considerable accuracy [7-9]. There are also analyses through Cite Space[10]. But they all do not solve the problem above.

Based on the particularity of the first pregnancy, this paper intends to first group women who have had a history of pregnancy and women who have not had a history of pregnancy, determine the main factors and linear equations affecting the disease of the two groups of women through principal component analysis and linear regression, and then test the accuracy through mechanical learning. Analyzing 768 data items on the Kaggle website, including Pregnancy, Glucose, Blood Pressure, Skin Thickness, insulin, BMI, Diabetes Pedigree Function, Age, and outcome to look for relationships between variables is the main purpose of this paper. However, on account of a small amount of data, and the data set containing a relatively limited population, the model and conclusion obtained in this paper have limited application. There are also a lot of variables that are not taken into account. Therefore, another aim of this paper is to verify the validity and testability of the logistic regression model and propose a more accurate diabetes prediction and management path based on artificial intelligence management.

2. Methods

2.1. Data Source and Data Processing

The dataset used in this article, sourced from the Kaggle website and originally from the National Institute of Diabetes, Digestive and Kidney Diseases, is a typical example of predictive analytics used in the medical field. It is important to help predict and understand the development of diabetes and to provide personalized treatment options for patients.

This dataset contains 768 cases and 9 variables, while this paper processed the data and added a variable(possibility) to indicate the possibility of a person developing diabetes, a total of 10 variables. The missing data of some variables of some samples in the data set will be replaced by linear interpolation so that the data in the whole data set is within a reasonable interval. The results are shown in Table 1.

In addition, the sample was grouped according to whether the sample was pregnant (whether X_{j1} was equal to 0). After grouping, the sample size of the pregnant group was 657, and the sample size of the non-pregnant group was 111

Table 1: Different types of variables

Elements	Type	Range	Logogram
Pregnancies	Numeric	0 to 17 times	X1
Glucose	Numeric	44 to 199 milligrams per deciliter	X2
Blood Pressure	Numeric	24 to 122 MMHG	X3
Skin Thickness	Numeric	7 to 99 millimeters	X4
Insulin	Numeric	15 to 846 microunits per milliliter	X5
BMI	Numeric	18.2 to 57.3 kilogram per square meter	X6
Diabetes Pedigree Function	Numeric	0.084 to 2.349	X7
Age	Numeric	21 to 81 years old	X8
Outcome	Categorical	0-Without diabetes 1-Have diabetes	Y
Possibility	Numeric	0 to 1	f(x_j)

2.2. Method Introduction

The method used in this study is logistic regression, which is a classification algorithm widely used in statistics and machine learning, especially in binary classification problems. Its goal is to predict the probability that a data point belongs to a certain category.

The core of Logistic regression is the use of the Sigmoid function shown in Figure 1. This function maps any real value to 0 or 1, converting the output of a linear combination into a distribution over 0 and 1, yielding a probability value.

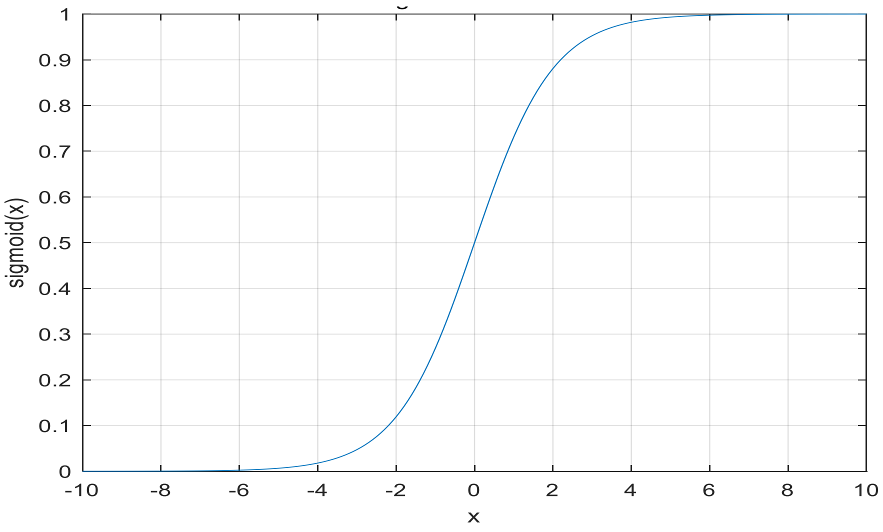

Figure 1: Sigmoid function

X_j=(X_{j1}, X_{j2}...X_{j8}) is the vector consisting of nine variables in the data set, since there are 768 samples, j goes from 1 to 768.

W=(w_1,w_2...,w_8) is the coefficient to be solved

The sigmoid function is also introduced.

$$f(x) = \frac{1}{1+e^{-x}} \tag{1}$$

Then the formula for each donor is as follows:

$$f(x_j) = \frac{1}{1+e^{-w^T\mathrm{x}+b}} \tag{2}$$

For each f(x_j), This article specifies a formula as:

$$f(x_j) = \begin{cases} 1 & if\ f(x_j) > 0.5 \\ 0 & if\ f(x_j) \leq 0.5 \end{cases} \tag{3}$$

Divide the data set into a training set and test set, set the ratio of the training set to 0.8, and solve w for all y in the training set, when $min \sum_{j=1}^{n} \quad |f(x_j) - y_j|$ is taken as the coefficient of the model.

Finally, it is brought into the test set to check the accuracy of the model.

3. Results and Discussion

Table 2: Model prediction efficiency without grouping

name	Parameter name	Parameter value
Model evaluation effect	Accuracy rate	77.922%
	Accuracy rate (comprehensive)	77.636%
	Recall rate (comprehensive)	77.922%
	f1-score	0.770

According to the Table 2, the comprehensive accuracy and recall rate of the ungrouped data set by logistic regression can reach 77.636% and 77.922%, The F1 score used to evaluate the performance of the binary model reached 0.77, indicating that logistic regression can better predict the samples.

Table 3 and Table 4 represent the model prediction rates for the pregnant and non-pregnant groups after grouping, respectively.

Table 3: Prediction efficiency of the pregnant group model

name	Parameter name	Parameter value
Model evaluation effect	Accuracy rate	79.545%
	Accuracy rate (comprehensive)	79.243%
	Recall rate (comprehensive)	79.545%
	f1-score	0.792

Table 4: Prediction efficiency of the Non-pregnant group model

name	Parameter name	Parameter value
Model evaluation effect	Accuracy rate	86.957%
	Accuracy rate (comprehensive)	86.743%
	Recall rate (comprehensive)	86.957%
	f1-score	0.867

After grouping, the prediction accuracy of the pregnant group reached 79.243%, the recall rate reached 79.545%, the prediction accuracy of the non-pregnant group was 86.743%, the recall rate was 86.957%, which was significantly higher than the accuracy and recall rate of the non-grouping group.

The above results show that the logistic regression model has a relatively perfect classification ability and good prediction accuracy for female diabetes, and can effectively classify and predict the population prone to diabetes and the population not prone to diabetes. At the same time, the experimental data showed that when the logistic regression method was used, the accuracy of the prediction after grouping pregnant and non-pregnant women was higher than that of the non-grouping, and the accuracy of the non-pregnant group was significantly improved.

Table 5: Characteristic coefficients of each group of equations

name	Ungrouped feature coefficient	Pregnancy group characteristic coefficient	Non-pregnancy group characteristic coefficient
Pregnancies	0.1367	0.1736	0
Glucose	0.0368	0.0383	0.0472
Blood Pressure	-0.0085	-0.0030	-0.0273
Skin Thickness	-0.0030	-0.0028	0.0159
Insulin	-0.0003	-0.0011	-0.0032
BMI	0.0790	0.0882	0.1027
Diabetes Pedigree Function	1.0926	1.0337	0.3339
Age	0.0136	0.0060	-0.0335
intercept	-8.5902	-9.3579	-7.6473

Table 5 is a table of characteristic coefficients corresponding to the respective variables obtained according to the regression model. Characteristic coefficients are used to judge the degree of influence of independent variables on the results. The larger the eigencoefficient is, the greater the influence of the independent variable on the result. By observing the three groups of characteristic coefficients, it can be found that the characteristic coefficients of Diabetes Pedigree Function are 1.0926, 1.0337, and 0.3339, and the characteristic coefficients of BMI are 0.0790, 0.0882, and 0.1027, respectively. The characteristic coefficient of pregnancy was 0.1367 and 0.1736 in the original sample and the pregnant group, and the characteristic coefficient of Glucose was 0.0368, 0.0383, and 0.0472. The characteristic coefficient of the above four variables was significantly higher than that of other variables. These results indicate that Diabetes Pedigree Function, BMI, pregnancy, and glucose have significant effects on female diabetes.

The above results are in good agreement with the results obtained in previous studies and have certain feasibility, which proves that the logistic regression model can indeed be applied to the management prediction of modern medicine. After establishing a reliable data set, the study can then obtain the sigmoid function used to calculate whether each individual in the data set is susceptible to disease, continuously optimize the accuracy through machine learning, and explore other factors that have a greater impact on female diabetes mellitus based on the results of this experiment.

The innovation of this paper lies in the evaluation of influencing factors, and it is found that grouping by whether or not pregnant can significantly improve the accuracy of model prediction, providing a way of thinking for subsequent research - the objects in the data set can be further classified by different data characteristics for research.

4. Conclusion

In this paper, a logistic regression model was used to predict female diabetes, and the accuracy reached 77.636% when the group was not grouped, 79.545% when the group was pregnant, and 86.743% when the group was not pregnant, indicating that logistic regression model can effectively predict female diabetes. Grouping samples can get better prediction results. According to the characteristic coefficient, it can also be concluded that Diabetes Pedigree Function, BMI, pregnancy, and glucose are the main influencing factors of female diabetes.

The shortcomings of this paper are that there are too few valid data used, the processing of abnormal data may lack certain scientificity, and the classification and discussion of variables are relatively limited, which makes the accuracy of the final results imperfect. However, with the increase of data sets and the exploration of relevant influencing variables, the model will be continuously optimized and improved.

In general, the logistic model has a good predictive and self-improving effect, which can effectively participate in the modern medical risk assessment of diabetes in the population, and help to find some pathogenic factors that are yet to be explored. Continuous optimization combined with other algorithms is of great significance for reducing the incidence of diabetes and delaying the onset of diabetes.

References

[1] The international diabetes federation. The IDF in 2021 is the global map of diabetes (version 10). (2021-12-06) . http://www.diabetesatlas.org.

[2] Zhang, J., Ding, X. L., Long, Y., et al. (2019). The incidence trend of type 2 diabetes mellitus in China from 1990 to 2019 and its prediction from 2020 to 2030. Journal of Huazhong University of Science and Technology (Med Edition), 53(3), 315-320 (in Chinese) DOI:10.3870/j.issn.1672-0741.23.08.017

[3] Huang, R., Feng, W., Lu, S., et al. (2024). An artificial intelligence diabetes management architecture based on 5G. Digital Communication and Network (English Edition), 10(1), 75-82.

[4] Somi, S. P. (2019). High-fat-diet induced obesity and diabetes mellitus in Th1 and Th2 biased mice strains: A brief overview and hypothesis. Chronic Diseases and Translational Medicine, 9(1), 14-19.

[5] Dicky, T. L., Syahidatul, W., Tricaesario, C., et al. (2019). Chronic complications risk among type 2 diabetes patients with a family history of diabetes. Chronic Diseases and Translational Medicine, 9(4), 336-340.

[6] Yumei, W., Juan, J., Rina, S., et al. (2022). Risk of gestational diabetes recurrence and the development of type 2 diabetes among women with a history of gestational diabetes and risk factors: A study among 18 clinical centers in China. Chinese Medical Journal, 135(6), 665-671.

[7] Liu, H. W. (2020). Study on risk prediction and online calculation of gestational diabetes mellitus based on machine learning algorithm. Tianjin Medical University.

[8] Cai, S. S., Zheng, T. Y., Wang, K. Y., Zhu, H. P. (2024). Clinical study of different prediction models in predicting diabetic nephropathy in patients with type 2 diabetes mellitus. World Journal of Diabetes, 15(1), 43-52.

[9] Alqushaibi, A., Hasan, M. H., Abdulkadir, S. J., Muneer, A., Gamal, M., Al-Tashi, Q., Taib, S. M., Alhussian, H. (2023). Type 2 diabetes risk prediction using deep convolutional neural network based-Bayesian optimization. Computers, Materials & Continua, 75(5), 3223-3238.

[10] Lingmei, Y., Miaojing, L., Taolan, S., et al. (2019). Construction of core index set for diabetes risk prediction based on CiteSpace analysis. Chinese Journal of Primary Health Care, 38(3), 25-29 .