Dean&Francis

ISSN 2959-6157

Enhancing Early Diabetes Risk Prediction: Optimization and Application of Logistic Regression Models

Manwen Luo[1, *]

Department of Statistics, Chengdu University of Information Technology, Chengdu, 610000, China
*Corresponding author: 2023201061@stu.cuit.edu.cn

Abstract:

Diabetes has become a significant challenge in global public health, with its incidence steadily rising and increasingly affecting younger populations. Statistics indicate that over 422 million people worldwide are currently living with diabetes. This chronic metabolic disease severely impacts patients' quality of life and imposes substantial economic pressure on healthcare systems. Therefore, timely identification of high-risk groups and the implementation of effective preventive strategies are crucial. However, the multifactorial nature of diabetes presents limitations in existing prediction methods. This study utilizes logistic regression analysis to highlight the key roles of factors such as gender, age, frequent urination, excessive thirst, fatigue, localized paralysis, genital fungal infections, emotional instability, and increased appetite in assessing diabetes risk. The model achieved prediction accuracies close to 93% in both the training and test sets. Additionally, the p-values for frequent urination, localized movement disorders, excessive drinking, mood fluctuations, weakness, genital fungal infections, and increased appetite were significantly below 0.01, strongly indicating a close association between these characteristics and the risk of developing diabetes. The findings of this study lay a critical foundation for the early warning system of diabetes and emphasize the need for future research to incorporate richer clinical data and innovative technologies to optimize prediction models.

Keywords: Diabetes; logistic regression; prediction model.

1. Introduction

Diabetes has become a significant challenge in global public health, with its prevalence rising steadily over the past few decades, and the age of onset showing a downward trend. Currently, the global number of di-

abetes patients has exceeded 422 million [1]. As a chronic metabolic disorder, it significantly reduces patients' quality of life and imposes substantial economic pressure on healthcare systems. Therefore, it is crucial to promptly identify individuals at high risk of developing diabetes and implement effective prevention strategies to mitigate the associated risks. However, given the complex and multifaceted etiology of diabetes, which includes genetic, environmental, and lifestyle factors, current predictive methods still have significant shortcomings and limitations [2].

In the field of predicting potential disease risks, statistical modeling methods, particularly logistic regression models, have gained favor among researchers and clinical experts, becoming mainstream tools due to their intuitive understanding and broad application range [3]. These models excel at handling multiple independent variables, allowing them to quantify the contribution of each variable to disease risk and provide refined risk assessment results. However, existing research trends have predominantly focused on traditional biomarkers and demographic data, such as blood glucose levels, body mass index (BMI), and genetic background, while relatively overlooking the importance of symptom information in predicting early-stage diabetes risk [4]. Although blood glucose is directly linked to diabetes diagnosis, its levels may not fully reveal an individual's actual risk in the early stages of the disease [5].

Against this research background, this study specifically focuses on a commonly overlooked symptom-oral thrush, a fungal infection of the oral mucosa closely related to immune system dysfunction and widely recognized as a key factor in the pathogenesis of diabetes [6]. Compared to the general population, individuals who have experienced oral thrush are significantly more likely to develop diabetes. However, current diabetes prediction models have not fully considered this important symptom in their construction [7].

Inspired by previous research, this study aims to enhance prediction accuracy by incorporating symptom variables such as oral thrush into the diabetes risk assessment model, thereby expanding the classic logistic regression framework [8]. Guan et al. have preliminarily tested the potential of integrating symptom information into disease risk assessment models, demonstrating that this approach can effectively improve model recognition rates and discriminatory power [9, 10]. Nevertheless, these preliminary efforts still face limitations in fine-tuning the model and validating its clinical applicability. Therefore, this study not only integrates symptom data into the model construction but also deepens the structural optimization of the logistic regression model by adding interaction terms and nonlinear features to more effectively reveal the complex interaction patterns between symptoms.

Moreover, to ensure the stability of the model, this study utilized a large dataset of patient records and implemented cross-validation methods to evaluate and fine-tune the model. This methodology aims to strengthen the model's broad applicability across various population groups and present clinicians with an effective tool for identifying high-risk groups for diabetes. Specific measures include thoroughly exploring the interaction between oral thrush and other indicators of diabetes risk, aiming to uncover possible pathological mechanisms and lay a theoretical foundation for future research activities.

2. Methods

2.1 Data Source

The data used in this paper for predicting diabetes comes from clinical surveys of patients at Sylhet Diabetes Hospital in Bangladesh. Saved in CSV format, the dataset includes 520 cases that detail patients' age, gender, and symptoms like excessive urination, thirst, sudden weight loss, and weakness. This data helps understand how these factors relate to diabetes risk, with a "class" column shows whether each patient was diagnosed with diabetes.

2.2 Variable Introduction

This dataset contains 520 instances and 16 variables, with 1 missing value (Table 1).

Table 1. Different types of variables

Term	Type	Range	Logogram
Age	Numeric	26-70	x_1
Gender	Categorical	0-Female,1-Male	x_2
Polyuria	Feature	0-false,1-ture	x_3
Polydipsia	Feature	0-false,1-ture	x_4
sudden weight loss	Feature	0-false,1-ture	x_5

ISSN 2959-6157

weakness	Feature	0-false,1-ture	x_6
Polyphagia	Feature	0-false,1-ture	x_7
Genital thrush	Feature	0-false,1-ture	x_8
visual blurring	Feature	0-false,1-ture	x_9
Itching	Feature	0-false,1-ture	x_{10}
Irritability	Feature	0-false,1-ture	x_{11}
delayed healing	Feature	0-false,1-ture	x_{12}
partial paresis	Feature	0-false,1-ture	x_{13}
muscle stiffness	Feature	0-false,1-ture	x_{14}
Alopecia	Feature	0-false,1-ture	x_{15}
Obesity	Feature	0-false,1-ture	x_{16}
class	Feature	0-negatice,1-positive	y

It includes 177 females and 342 males, with patient ages ranging from 26 to 72 years. The data includes 16 variables (Age, Gender, Polyuria, Polydipsia, sudden weight loss, weakness, Polyphagia, Genital thrush, visual blurring, Itching, Irritability, delayed healing, partial paresis, muscle stiffness, Alopecia, Obesity, class).

2.3 Method Introduction

The methodology employed in this study involves conducting binary logistic regression using SPSS. A logistic regression model is fitted to predict the binary outcome variable, diagnosis_bin, based on each predictor variable. The logistic regression model can be expressed as follows:

$$P = \frac{exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_m x_m\right)}{1 + exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_m x_m\right)} \quad (1)$$

Where β_0 is constant term, $\beta_1, \beta_2, \cdots, \beta_m$ are partial regression coefficients. Perform a logit transformation on $f\left(x\right) = \frac{1}{1+e^{-x}}$. So, the logistic regression model can be expressed in the following linear form:

$$ln\frac{p}{1-p} = \beta_0 + \beta_1 x_1, \beta_2 x_2 \ldots, \beta_m x_m \quad (2)$$

3. Results and Discussion

3.1 Descriptive Analysis

In patients with diabetes, the early onset of signs and symptoms was compared with the average, and the average was higher.

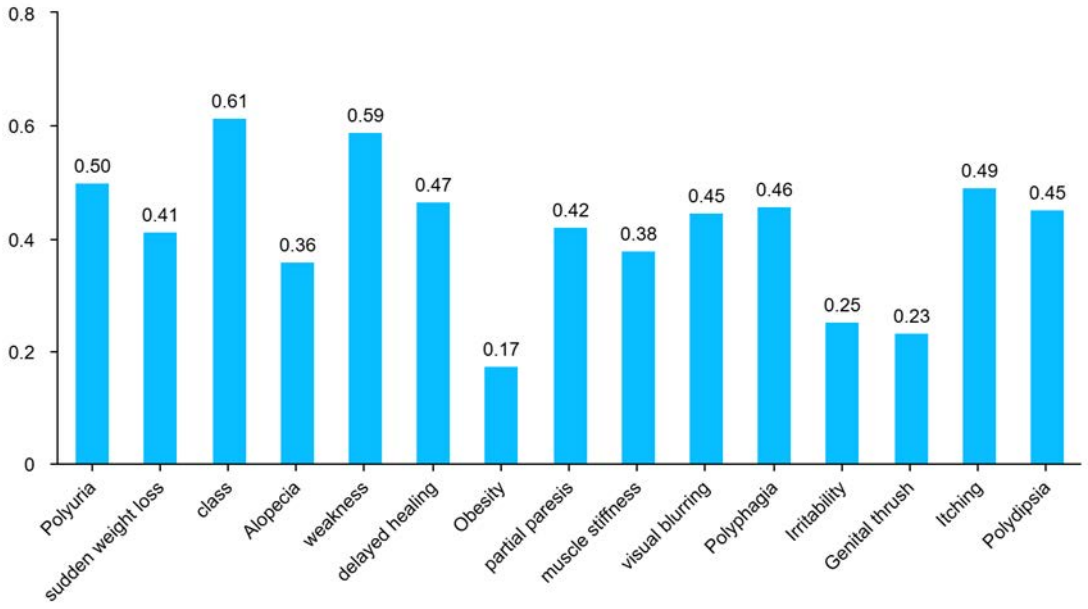

Fig. 1 Comparison chart of averages

Figure 1 shows that Alopecia and weakness are the most important features in early diabetes risk prediction, with scores of 0.61 and 0.59, while Genital thrush and Irritability have the lowest importance, with scores of 0.23 and 0.25. This indicates that different symptoms have varying degrees of influence on early diabetes risk prediction.

3.2 Logistic Regression Results

As shown in Figure 2, this study incorporated various factors that might be associated with diabetes into the model, including Gender, Age, Sudden Weight Loss, Obesity, Alopecia, Polyuria, Partial Paresis, Polydipsia, Irritability, Delayed Healing, Muscle Stiffness, Visual Blurring, Itching, Weakness, Genital Thrush, and Polyphagia.

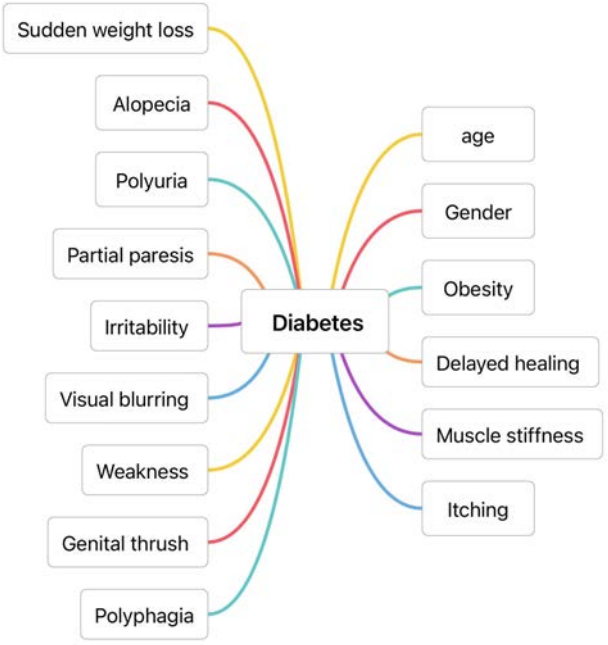

Fig. 2 Variable-related schematic

Through computation, the final logistic regression equation was derived in this study:

$$Probit(p) = 4.685 - 0.056x_1 - 2.009x_2 + ... - 0.219x_{16} \quad (3)$$

Where p represents the probability of diabetes being 1.

Table 2. Model results

Item	coefficient	SE	z-value	p-value	OR-value
Gender	-2.009	0.287	-6.993	0.000	0.134
Age	-0.056	0.025	-2.260	0.024	0.946

ISSN 2959-6157

sudden weight loss	0.136	0.527	0.258	0.797	1.145
Obesity	-0.219	0.537	-0.408	0.683	0.803
Alopecia	0.311	0.597	0.521	0.602	1.365
Polyuria	4.202	0.658	6.389	0.000	66.807
partial paresis	1.230	0.521	2.363	0.018	3.421
Polydipsia	5.188	0.823	6.306	0.000	179.166
Irritability	2.054	0.567	3.626	0.000	7.801
delayed healing	-0.172	0.526	-0.327	0.744	0.842
muscle stiffness	-0.644	0.552	-1.167	0.243	0.525
visual blurring	0.513	0.601	0.854	0.393	1.671
Itching	-2.680	0.638	-4.202	0.000	0.069
weakness	1.060	0.507	2.091	0.037	2.886
Genital thrush	1.685	0.529	3.184	0.001	5.394
Polyphagia	1.455	0.530	2.745	0.006	4.283

Table 2 clearly shows that by examining whether the p-values exceed the 0.05 threshold, the study was able to identify potential factors associated with early-stage diabetes onset. Specifically, sudden weight loss, obesity, hair loss, delayed wound healing, muscle tension, and blurred vision all had p-values exceeding the 0.05 threshold, suggesting that these variables do not significantly affect the likelihood of developing diabetes. Conversely, the p-values for gender, age, and skin itching, while not below 0.01, were less than 0.05, indicating a moderate correlation with diabetes risk. Additionally, polyuria, local paralysis, polydipsia, irritability, weakness, genital fungal infections, and increased appetite all had p-values below 0.01, strongly suggesting a close association with diabetes risk. The combined use of regression coefficients and odds ratios (OR) reveals the detailed mechanisms of these factors: for each unit increase in these factors, the risk of developing diabetes increases by 66.807 times, 3.421 times, 179.166 times, 7.801 times, 2.886 times, 5.394 times, and 4.283 times, respectively. This finding underscores the critical role of these specific factors in predicting early-stage diabetes risk and provides important references for subsequent research and clinical applications.

Compared to previous studies that often focused on a single factor (such as genetic predisposition or lifestyle choices like diet and exercise), this study, with its broad perspective, offers a more comprehensive understanding of early-stage diabetes risk. By incorporating a wider variety of variables, the study not only alleviates the bias that may arise from isolated examination of individual factors but also paves the way for future diabetes prevention and management research. Such methodologies can enable healthcare professionals to identify high-risk individuals earlier and deploy more precise interventions.

Table 3. Binary probit regression prediction accuracy summary

0		predicted value		forecast accuracy	Prediction error rate
		1			
true value	0	182	18	91.00%	9.00%
	1	19	300	94.04%	5.96%
Summary				92.87%	7.13%

The model's predictive accuracy is used to assess the model's fit (Table 3). As shown in the table above, the overall predictive accuracy of the research model is 92.87%, indicating that the model's fit is acceptable. When the actual value is 0, the predictive accuracy is 91.00%; and when the actual value is 1, the predictive accuracy is 94.04%

4. Conclusion

This study reveals that variables such as gender, age, polyuria, thirst, fatigue, partial paralysis, genital fungal infections, mood instability, and increased appetite show high predictive power in forecasting the onset of diabetes. The accuracy of both the training and validation sets approaches 93%. Through data analysis, this study iden-

tified several key factors that significantly increase the risk of developing diabetes. These findings may provide a basis for developing more targeted diabetes prevention measures, especially for groups at higher risk due to factors such as age or medical history. One focus of future research will be to collect more extensive data. Although this study used mean analysis, extreme values in the data distribution may reveal deeper insights when assessing diabetes risk. Therefore, subsequent research needs to incorporate more diverse clinical parameters, especially focusing on extreme cases, and test the potential advantages of extreme values over conventional averages in predictive performance through model training and outcome comparisons. Furthermore, more sophisticated analytical models should be adopted to systematically explore potential interactions and nonlinear relationships between variables, with the aim of deepening the understanding of the mechanisms underlying the development of diabetes.

References

[1] Liu Jindong. How much do you know about the early symptoms of diabetes. Seeking Medical Advice and Medication, 2013, 5: 4-5.

[2] Bulbul Mehmet. A novel hybrid deep learning model for early stage diabetes risk prediction. Journal of supercomputing, 2024, 80(13): 19462-19484.

[3] Dutta Aishwariya, et al. Early Prediction of Diabetes Using an Ensemble of Machine Learning Models. International journal of environmental research and public health, 2022, 19: 12378.

[4] Wang Xu, et al. Prediction and analysis of risk factors for diabetic retinopathy based on machine learning and interpretable models. Heliyon, 2024, 10(9): 29497.

[5] Tan Yaqi, et al. Early Risk Prediction of Diabetes Based on GA-Stacking. Applied sciences-basel, 2022, 12(2): 632.

[6] Bhat Salliah, et al. Prevalence and Early Prediction of Diabetes Using Machine Learning in North Kashmir: A Case Study of District Bandipora. Computational intelligence and neuroscience, 2022, 1: 2789760.

[7] Chaitanya D, Venkaiah C, Senapati R. Multi Disease Prediction Using Ensembling of Distinct Machine Learning and Deep Learning Classifiers. 5th International Conference on Soft Computing and its Engineering Applications, 2023.

[8] Zhang Zaiheng, et al. A novel evolutionary ensemble prediction model using harmony search and stacking for diabetes diagnosis. Journal of king saud university-computer and information sciences, 2024, 36(1).

[9] Wu Guanrong, et al. Development and validation of a simple and practical model for early detection of diabetic macular edema in patients with type 2 diabetes mellitus using easily accessible systemic variables. Journal of translational medicine, 2024, 22(1): 1-11.

[10] Mundargi Zarinabegam, et al. Diabetes Prediction Using Logistic Regression. Working paper, 2024.