Proceedings of the 2023 International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/8/20230158

The incidence of diabetes prediction based on the logistic regression model

Fengyuan Yang

Wharton International senior school, Liaoning, 110016, China

15060340328@xs.hnit.edu.cn

Abstract. Diabetes is a rapidly growing global public health challenge that demands effective prevention and treatment strategies. This study aims to explore the relationship between diabetes incidence and relevant indicators such as BMI, blood pressure, and skin thickness by utilizing data from the Kaggle dataset. In this study, the logistic regression model was employed to identify risk factors associated with the incidence of diabetes. The logistic regression model allows this study to test the effect of correlation between predictor variables and the outcome variable, and develop a model to predict the likelihood of diabetes incidence. Using a Python implementation of the model, nearly 77% precision in predicting the incidence of diabetes was obtained. The analysis revealed a strong correlation between age and BMI and the incidence of diabetes, aligning with previous findings in the domain knowledge of diabetes. The results of the study may help individuals and healthcare providers to identify and manage the risk factors associated with diabetes, ultimately reducing the incidence of this disease. The proposed approach of utilizing the logistic regression model provides a valuable tool for predicting the onset of diabetes, contributing to the ongoing effort to combat this global public health issue.

Keywords: the incidence of diabetes prediction, machine learning, logistic regression.

1. Introduction

Diabetes is a complex metabolic disorder characterized by elevated blood glucose levels resulting from defective insulin secretion or impaired biological effects, or both. The chronic hyperglycaemia associated with diabetes can lead to extensive tissue damage and dysfunction in various organs such as the kidneys, heart, blood vessels, eyes, and nerves. The damage inflicted by diabetes can be classified into two broad categories: acute complications and chronic complications [1, 2].

In terms of acute complications, there may be acute complications such as diabetic ketoacidosis, hypertonic coma, acute infection, lactic acidosis etc. For the Chronic complications: first, there may be macrovascular complications, including diabetic cardiovascular disease and diabetic cerebrovascular disease, which can lead to myocardial infarction, cerebral infarction, disability and death; second, there may be small vascular complications, such as kidney damage, retinopathy, fundus haemorrhage and even blindness. Third, there may be neurological complications, diabetic peripheral neuropathy, which will cause abnormal sensation in the limbs of patients, and may also lead to diabetic feet, such as foot ulcers, infection, gangrene and so on. The detection rate of diabetes is low, about 50% of patients with diabetes are missed, and even some patients have serious cardiovascular and cerebrovascular diseases before they are found to have diabetes. In addition, the speed of diagnosing diabetes is very slow, and

© 2023 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/8/20230158

the labor cost is very high, such as the cost of the doctor and the machine. Therefore, it is necessary to find an alternative way to alleviate the problem more quickly, and it is generally believed that the available method is the prediction based on the artificial intelligence algorithms.

Artificial Intelligence (AI) is a nascent technical discipline that is concerned with the study and development of theories, methods, technologies, and application systems designed to simulate, extend, and amplify human intelligence [3, 4]. As a subfield of computer science, AI strives to comprehend the fundamental nature of intelligence and create intelligent machines that can approximate human cognition. Research endeavors in this domain encompass a wide range of topics, including robotics, speech recognition, image processing, natural language understanding, and expert systems [5-7]. Over the years, the theory and technology of AI have become increasingly sophisticated, and its applications have proliferated exponentially. The future scientific and technological advancements facilitated by AI are likely to embody the essence of human wisdom. By replicating the information processing mechanisms of human consciousness and thinking, AI is capable of emulating human-like intelligence, and in some cases, even surpassing human cognition. Although there are many related applications, the application of artificial intelligence in medicine is still less. In the past, when diagnosing diabetes, people mostly relied on doctors' domain knowledge, urine tests, and other complex medical tests, which were time-consuming and laborious and insensitive. combined with artificial intelligence to predict diabetes, the labor cost can be saved, and the misdiagnosis rate can be greatly reduced.

Based on the fact mentioned above, this paper focuses on the use of logistic regression model and utilize the large amount of data to quickly get the prediction results. The data in this paper is collected by kaggle which is a data collection website. the final experimental result based on the linear regression model is that the accuracy of this algorithm is about 91.42%.

2. Methodology

2.1. Dataset preparation

The present dataset has been sourced from the National Institute of Diabetes and Digestive and Kidney Diseases [8], and serves the purpose of predicting the likelihood of a patient being diagnosed with diabetes, based on various diagnostic measurements contained within the dataset. The selection of instances within the dataset was subject to a number of specific constraints, such that all patients included are females aged at least 21 years and of Pima Indian descent. The dataset, which is provided in a comma-separated value (.csv) format, comprises several independent medical predictor variables, alongside a single target dependent variable (Outcome). Here is a piece of data that included blood pressure, skin thickness and other seven factors which can influence diabetes. This data set is classified data. It is divided into nine categories. The statistical information of a part of the collected dataset is shown in Figure 1.

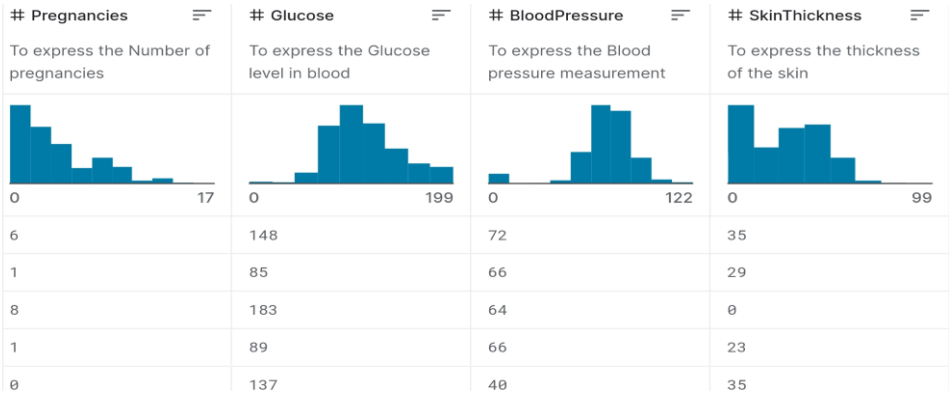

Figure 1. The statistical information for the collected dataset.

DOI: 10.54254/2755-2721/8/20230158

2.2. Logistic regression

Logistic Regression (LR) is a kind of special classification algorithm with "regression" in its name, which is actually used to solve the problem of classification, and the idea of solving the problem still refers to the idea of regression [9]. The core idea of the logistic regression algorithm is to find a line (surface) in the space and classify it according to the relative position of the undetermined point and the dividing line (interface). The Sigmoid function is generally used to deal with the dichotomy problem. The linear regression coefficient can be employed to calculate the importance of features. It can be easy for the study to find which features are more important and which features are less important in comparison.

Scikit-learn, also known as sklearn, is an open-source machine learning library for the Python programming language. It emphasizes on simplicity and ease of use, with a consistent API and clear documentation. This makes it accessible to both novice and experienced machine learning practitioners, and allows for rapid experimentation and prototyping, which is considered in this study. In the first step, sklearn was employed to read the data, and visualize the distribution of it and preprocess the data. In the next step, the data was divided into the training set and the test set, and this study uses the training set to train the logistic regression model and employs the test set to test model.

3. Result and discussion

As can be seen from the below Figure 2, which shows the Pearson coefficient of the factors of Diabetes. Age accounts for the largest percentage, which is most likely to lead to Diabetes (0.54). It was followed by Glucose and BMI, accounting for 0.49 and 0.31, respectively. The factor that has little effect on diabetes is Diabetes, which only accounts for 0.03. The experimental results show that age is closely related to the causes of diabetes, which is also extremely mild with the relevant knowledge of the real life. In life, people with high incidence of diabetes are also concentrated in the middle-aged and elderly, teenagers or adults rarely have diabetes. So, this is reasonable. There is also a strong correlation between BMI. If the BMI is too high, it means that the person is obese, and obese people are more likely to develop diabetes than ordinary people.

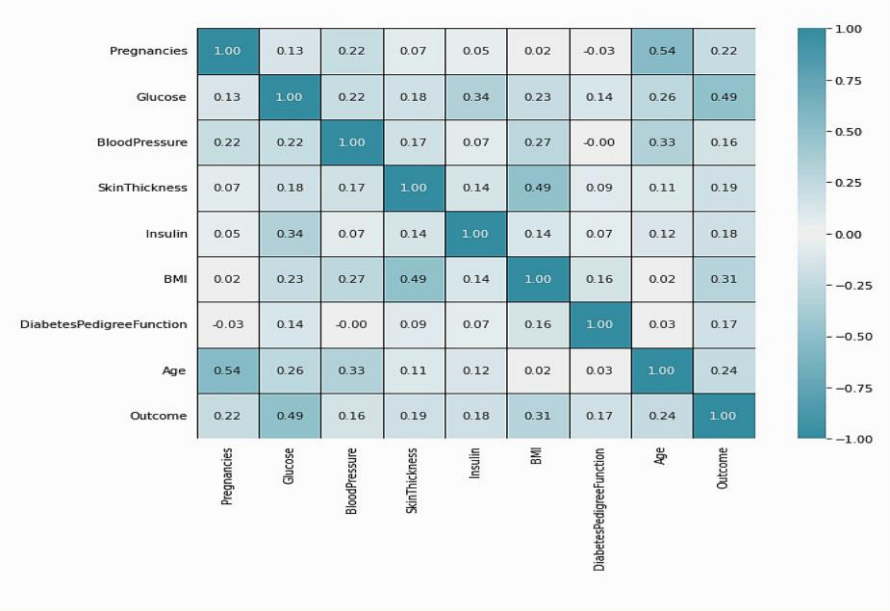

Figure 2. The correlation matrix of different parameters [10].

The following Table 1 shows that the precision of this experiment is as high as 77%, which demonstrates the effectiveness of the proposed method.

DOI: 10.54254/2755-2721/8/20230158

Table 1. The performance of the logistic regression model.

Model name	Precision	Recall	F1-score
Logistic Regression	77%	54%	63%

In addition to the advantages previously mentioned, the logistic regression model possesses several notable benefits when compared to other models. Firstly, logistic regression models are suitable for modeling linear relationships between the predictor variables and the response variable, rendering them useful in scenarios where the nature of the data suggests a linear relationship. Secondly, the computational complexity of logistic regression models is relatively low, enabling fast calculations and making them suitable for handling large datasets. Additionally, the logistic regression model returns classification results in the form of quasi-probability numbers, which can offer useful insights into the likelihood of a given outcome. This is in contrast to other models that provide fixed 0 or 1 classification results, which may not accurately reflect the true probability of the outcome. Finally, the logistic regression model exhibits robustness against noise in the data, providing increased resistance to inaccuracies in the input data that may arise from measurement error or other sources. Taken together, these advantages make the logistic regression model a valuable tool in machine learning and data analysis, with numerous applications across various fields.

4. Conclusion

In conclusion, the present study indicates that the utilization of a logistic regression model in predicting the incidence of diabetes yields a precision exceeding 77%, suggesting that machine learning models possess a promising prospect for reducing the incidence of diabetes by promoting individuals to consciously adopt healthier living habits. However, the model's efficacy is impeded by several limitations, including the suboptimal performance when dealing with a vast feature space, underfitting concerns leading to reduced general accuracy, an inadequate capacity to manage multi-class features or variables, and restriction to binary classifications that must be linearly separable. Moreover, non-linear features require modifications in the logistic regression model to enhance its effectiveness. While acknowledging these limitations, it is anticipated that the advancement of the field will enable the emergence of more sophisticated methods to overcome these challenges.

References

[1] Zimmet P Z Magliano D J Herman W H et al. 2014 Diabetes: a 21st century challenge The lancet Diabetes & endocrinology 2(1) 56-64
[2] Bilous R Donnelly R Idris I 2021 Handbook of diabetes John Wiley & Sons
[3] Russell S J 2010 Artificial intelligence a modern approach Pearson Education Inc.
[4] Zhang B Zhu J Su H 2023 Toward the third-generation artificial intelligence Science China Information Sciences 66(2) 1-19
[5] Yu Q Chen P Lin Z et al. 2020 Clustering Analysis for Silent Telecom Customers Based on K-means++ 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC) IEEE 1: 1023-1027
[6] Yuan F Zhang Z Fang Z 2023 An effective CNN and Transformer complementary network for medical image segmentation Pattern Recognition 136: 109228
[7] Atal D K 2023 Optimal Deep CNN–Based Vectorial Variation Filter for Medical Image Denoising Journal of Digital Imaging1-21
[8] Kaggle Predict diabities 2022 https://www.kaggle.com/datasets/whenamancodes/predict-diabities
[9] Kleinbaum D G Dietz K Gail M et al. 2002 Logistic regression New York: Springer-Verlag
[10] Kaggle Diabetes complete eda and svc tuning 2022 https://www.kaggle.com/code/antoniosabatini/diabetes-complete-eda-and-svc-tuning