ISSN 2959-409X

The Research on Factors Influencing Diabetes

Benhao Wang[1]

[1]School of Statistics, Shandong University, Weihai, China
Corresponding author: 202000820204@mail.sdu.edu.cn

Abstract:

Diabetes is a serious disease for people, causing massive damage to body. However, people learn a little about the diabetes, showing that it is very important to analyze the relationships between diabetes and several factors. We incorporate multiple factors such as age and gender into a unified framework to jointly test the impact of multiple factors on the dependent variable. Whether a person owning the diabetes is been seen as the result. This paper established logistic regression model and using a comprehensive dataset sourced from Kaggle containing these factors to estimate paraments and using F1-score and Receiver Operating Characteristic (ROC) curve to evaluate the whole model. The results show that the estimation of parameter HbA1c is 2.53, playing the most important role in the diabetes. Blood glucose, Age, and BMI account, either. Apart from that, it could be learned from the Stand Error (SE) and P-value about the accuracy. The characters of F1-score and ROC curve show that this model performs exceptionally well in identifying diabetes cases. A conclusion could be drawn that a person could try to exercise or improve diet to release the symptom of diabetes.

Keywords: Diabetes; Logistic regression model; F1-score; ROC curve; SE; P-value

1. Introduction

Diabetes mellitus is a kind of disease that impacts the body's insulin which is a hormone to convert glucose (sugar) into energy. The people with diabetes always don't have enough insulin supply to decrease the sugar.

Diabetes has been a global concern, according to the research from Michael and Elizabeth about the spread of diabetes in the world [1]. More specifically, we could learn from the data about more than 1,486 nonpregnant adults that the population of diabetes has increased a lot compared with the old ones [1]. The tendency about the growth of the patients is difficult to prevent, and the number has been doubled from approximate 11,000,000 cases to 22,000,000 cases during 27 years [2]. These diseases can considerably worsen the quality of life and the probability of death is rising,either. Undoubtedly, the situation will be more difficult in the foresee future [3].

This disease might cause several serious results. Research on diabetes has been ongoing for decades, with significant progress made in understanding the disease. Above all, as research going, it might be easier for these people several genetic complications [4]. Moreover, the happening of End stage renal disease (ESRD) might be related to the worldwide disease [5]. Last but not the least, the disorder even has an association with heart diseases [6]. It should come as no surprise that it could a severe result if we didn't treat it carefully.

Diabetes could result in serious consequences and it is an extremely urgent to gain a deeper understanding of diabetes. To achieve the object, the predecessors has studied a lot about the essential aspects about the causes of the diabetes from the macroscopic factors like age, smoking and gender [7 8, 9]. However, though many things have been found there is still much to be learned about diabetes, and people learn a little about it. The comprehensive managements containing more efficient and cost-effective ways to diagnose, manage, and treat diabetes is rare, which can help reduce the economic burden of diabetes on healthcare systems, making care more accessible and sustainable for patients and even gain insights into preventing and managing those complications.

Because of the terrible consequences, the purpose of the paper is to help people have a better understanding of diabetes and it is important to have an analysis of the factors of the causing of diabetes.

An essential step is acquiring an appropriate dataset. It includes data from patients about medical and demographic prospects, also containing their diabetes situations. By obtaining this dataset, researchers can analyze multiple features such as age and body mass index (BMI). The next step involves creating a statistical model to examine the correlation between diabetes and these factors, which helps in understanding the disease's mechanisms.

Compared to the traditional research which just discuss one factor, this paper takes 7 variates into consideration. Furthermore, the article can analyze the relation between diabetes and every factor and even can compare the estimation of each factor to get the rough ranking of influence. In this way, people can find the major factors of diabetes and do some targeted measures to solve it. The result shows that the most effective index is HbA1c and the estimation is 2.53, followed by blood glucose and age whose estimation is 1.370 and 1.027. Others can't cause decisive effect.

From this research, people can have a deeper understanding of the correlation between different factors, so they can detect relevant index in daily life to judge the condition roughly. Furthermore, they can take more precautions in advance to deal with the disease, like changing diets and doing exercise.

2.Method

2.1 Data Source

The date of the research is from a website named Kaggle, including several factors relating to the disease. It contains 49,989 samples, which could make sure that the result would be concise enough.

2.2 Variable Selection

Large number of factors are relevant to diabetes, as illustrated in Table.1.

Table 1. Data of diabetes

Factors	Number	Diabetes	Range	Variable type
Gender	49,989	4,320	1 or -1	Typed
Age	49,989	4,320	0.08 ~ 80	Continuous
Hypertension	3,760	1,067	0 or 1	Typed
Heart disease	1,946	637	0 or 1	Typed
BMI	36,924	4,000	10.01 ~ 88.72	Continuous
HbA1c	10311	2650	3.5 ~ 9	Continuous
Blood glucose	5,450	1,983	80 ~ 300	Continuous

HbA1c level and blood glucose level play an important role in judging the diabetes. On the one hand, HbA1c refers to hemoglobin bound with glucose, and it can express the glucose level during 2-3 months. Because of the short life of red blood cells, HbA1c can be considered to be a symbol of blood glucose level. On the other hand, blood glucose level means the density of glucose in the blood. Fasting Blood Glucose (FBG) and Oral Glucose Tolerance Test (OGTT) are common measurement methods. Typically, if the HbA1c level of ≥6.5%, a 2-hour blood glucose level of ≥200 mg/dL during OGTT or the BMI is higher than 23.9, the patient can be diagnosed diabetes. Apart from that, gender, age, hypertension, heart disease and BMI are selected as the effecting factors either.

2.3 Method Introduction

It could be easily found that the result of data is 0 or 1, which is correlated to binary classification problem and logistic model might be the essential model about this kind of problem.

By learning from the dataset, the paper might find there are some potential relationships between these factors like if a person following with high level of HbA1c or high level of blood glucose is used to have a diabetes. The paper designs a logistic regression model from which people can get a better understanding and achieving of the research and simplify the situation to a easier one. The formula is expressed as followed.

$$P(Y=1|X)=\frac{1}{1+e^{-\sum_{i=1}^{7}a_i x_i+b}} \tag{1}$$

All the samples will be cited into the model and soon get a estimated values of the paraments. After training the model through the large samples and accurately estimating the coefficients, much information could be concluded. Each factor was contained and we could judge the importance according the regression coefficient. After that, the prediction whether a person owns the disease as long as he could provide enough information into the model will come true. If the value of the estimated model is greater than 0.5, the person may be diagnosed as a patient.

3.Results and Discussions

3.1 Data Processing

Before the analysis, the research needs to handle de-

fault-value. Usually, the solution is using median or mean value to replace the default-value and in this paper, the median is adopted. The next step is ensuring training set and test set, and they account for 80% and 20% of the total, respectively. Lastly, it is necessary to use function StandardScaler from sklearn to make sure the data follow the standard normal distribution. By the approaches above, the information would have a better usage and the result would be more accurate.

3.2 Model Estimation

Much information can be learned from the Table 2. The parameter of gender Hypertension and Heart disease are very small, which means that they almost do not have influence in diabetes. Meanwhile, the outcome of the parameters of Age Blood glucose and BMI show that the three factors are effective. A person with older age, higher blood glucose or greater BMI is more likely to have this disease. At last, HbA1c is determining. In most situations, a person with high HbA1c could be diagnosed owning the diabetes.

Table 2. The estimation

Factors	Estimate of parameters	Stand Error	P-value
Gender	0.140	0.028	4.944×10^{-07}
Age	1.027	0.039	1.453×10^{-153}
Hypertension	0.205	0.020	1.646×10^{-25}
Heart disease	0.137	0.019	1.003×10^{-13}
BMI	0.667	0.027	7.666×10^{-136}
HbA1c	2.530	0.061	$0.000\times10^{+00}$
Blood glucose	1.370	0.031	$0.000\times10^{+00}$
Constant	-5.266		

3.3 Model Evaluation and Discussion

Having estimated the model, it is necessary to guarantee the accuracy, which is related to the F1-score.

The percentage of true positive samples among the predicted positive results is measured by precision, while recall is the proportion of true positive predictions among all actual positive samples. Higher values for both metrics indicate better performance. The F1-score is a comprehensive index, averaging these two indicators . Although both precision and recall should ideally be high, they often conflict, making the F1-score a useful metric for overall classifier performance evaluation. Its value is positive but less than 1. The closer the value to 1, the better the model is. However, specific evaluations should consider the actual context. The F1-score is frequently used to gauge model improvements.

Table 3. Evaluation results of training set

Diabetes	Precision	Recall rate	F1-score	Samples
Yes	0.97	0.99	0.98	9131
No	0.89	0.64	0.75	867
Mean	0.93	0.82	0.86	9998
Mean(comprehensive)	0.96	0.96	0.96	9998

Table 4. Evaluation results of texting set

Diabetes	Precision	Recall rate	F1-score	Samples
Yes	0.97	0.99	0.98	36538

No	0.87	0.63	0.73	3453
Mean	0.92	0.81	0.85	39991
Mean(comprehensive)	0.96	0.96	0.96	39991

According to the Table 3 and Table 4, it could be easily found that the model has a high F1-score of 0.98 for positive prediction both in training set and text set and the comprehensive mean F1-score is 0.96, meaning that the model performs exceptionally well in identifying diabetes cases.

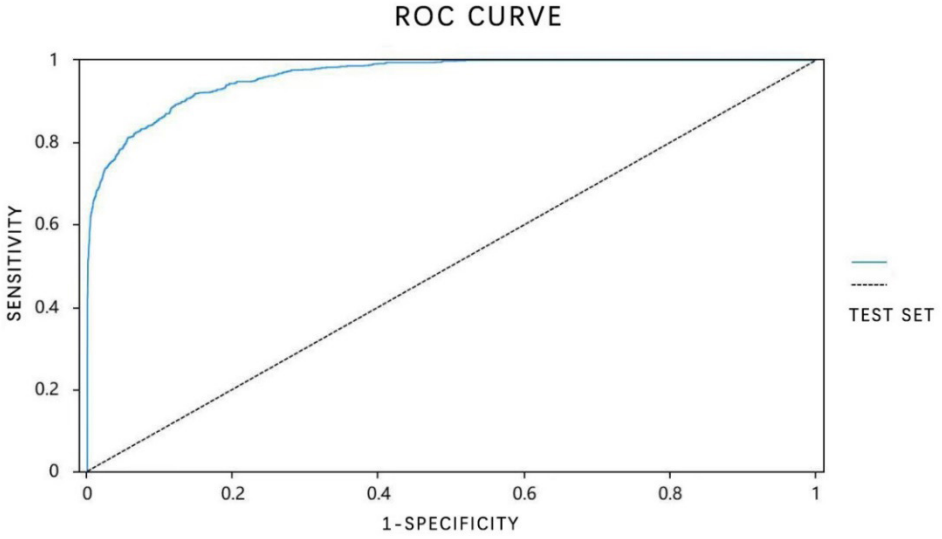

Fig 1. ROC curve

Another important thing is ROC curve which is specifically used for binary classification problems. It plots the true positive rate (sensitivity) against the false positive rate (1-specificity) across various threshold settings, providing an overall view of the model's performance. The curve can make an assessment about the function of classification according to the model's result. A curve closer to the top-left corner signifies better model performance. As shown in Figure 1, the ROC curve for the model is really close to the top-left corner. It is a symbol of high true positive rates and low false positive rates, and thus demonstrating excellent classification performance on the test set.

Apart from that, according to the research of Samuel Klein [10], fatness plays an important role in diabetes for the reason that it causes both insulin resistance and β cell dysfunction, resulting in a contradiction with this paper. However, he is just sharing some introduction of theory instead of doing detailed experiments, which causing the contradiction. Moreover, the estimation of hypertension is considerable low and it is just 0.205, expressing that it almost has no effect to the diabetes. This result is opposite to the findings of Murray and James [11] whose finding shows that diabetes mellitus and hypertension frequently coexist, occurring together more often than would be expected by chance. The causing may be their ignorance for the physical influence or the effect of diets.

However, there are still several problems in this model. The lower recall for the Non-Diabetes indicates a high false negative rate, where actual non-diabetes cases are incorrectly classified as diabetes The F1-score of Non-Diabetes is just 0.73, showing there is a significant imbalance in the number of samples between the classes. There might has multicollinearity between different variables and the affection of obesity is dispersed to other factors. Therefore, there are still several drawbacks.

4.conclusion

Diabetes is a serious disease causing badly damage to people. For helping people learn more about it, this paper has research on the possible factors of diabetes. To solve this problem, a logistic model is established and after estimation and a conclusion could be drew that HbA1c and blood glucose are determining factors. The ROC curve and F1-score shows that the model is very accurate for positive prediction. However, a significant imbalance in the number of samples results in low recall rate and F1-score.

According to the paper, people could learn more about diabetes. For preventing diabetes or releasing symptoms, some suggestions could be provided to people. Because

of the importance of HbA1c and blood glucose, a person could try to exercise or improve diet. From national aspect, it is useful to build more sports facilities to encourage fitness.

References

[1] Michael Fang, Elizabeth Selvin; Thirty-Year Trends in Complications in U.S. Adults With Newly Diagnosed Type 2 Diabetes. Diabetes Care 1 March 2021; 44 (3): 699–706.

[2] Liu, Jinli, et al. "Trends in the incidence of diabetes mellitus: results from the Global Burden of Disease Study 2017 and implications for diabetes mellitus prevention." BMC public health 20 (2020): 1-12.

[3] Chollette C. Olisah, Lyndon Smith, Melvyn Smith, Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective, Computer Methods and Programs in Biomedicine, Volume 220,2022

[4] Cole, J.B., Florez, J.C. Genetics of diabetes mellitus and diabetes complications. Nat Rev Nephrol 16, 377–390 (2020)

[5] Hui-Teng Cheng, Xiaoqi Xu, Paik Seong Lim, Kuan-Yu Hung; Worldwide Epidemiology of Diabetes-Related End-Stage Renal Disease, 2000–2015. Diabetes Care 1 January 2021; 44 (1): 89–97

[6] Triposkiadis, F.; Xanthopoulos, A.; Bargiota, A.; Kitai, T.; Katsiki, N.; Farmakis, D.; Skoularigis, J.; Starling, R.C.; Iliodromitis, E. Diabetes Mellitus and Heart Failure. J. Clin. Med. 2021, 10, 3682.

[7] Bahour, N., Cortez, B., Pan, H. et al. Diabetes mellitus correlates with increased biological age as indicated by clinical biomarkers. GeroScience 44, 415–427 (2022).

[8] Bar-Zeev, Yael PhD, MD; Haile, Zelalem T. PhD, MPH; Chertok, Ilana Azulay PhD, MSN. Association Between Prenatal Smoking and Gestational Diabetes Mellitus. Obstetrics & Gynecology 135(1):p 91-99, January 2020.

[9] Ciarambino, T.; Crispino, P.; Leto, G.; Mastrolorenzo, E.; Para, O.; Giordano, M. Influence of Gender in Diabetes Mellitus and Its Complication. Int. J. Mol. Sci. 2022, 23, 8850.

[10] Klein, Samuel, Amalia Gastaldelli, Hannele Yki-Järvinen, and Philipp E. Scherer. "Why does obesity cause diabetes?" *Cell metabolism* 34, no. 1 (2022): 11-20.

[11] Epstein, Murray, and James R. Sowers. "Diabetes mellitus and hypertension." Hypertension 19.5 (1992): 403-418.