DOI: 10.54254/2753-8818/52/2024CH0149

The Effect of Dataset Limitation on the Diabetes Risk Factors

Yueyang Ma

School of Mathmatics, Harbin Institute of Technology, Harbin, 150000, China

2022110634@stu.hit.edu.cn

Abstract. Diabetes research is an important topic in the medical domain. However, the research on diabetes risk factors always restricts objects to a certain category. This research wants to figure out the effect of such limitations. The result of this study could give guidance in this field. A dataset built by the Center of Disease Control and Prevention, which has been cleaned is used in this study. The binomial logistic regression analysis is applied to this research and the decision tree model is used for testing. The results suggest that many environmental factors are associated with diabetes. However, taking these risk factors together did not accurately predict diabetes. By analyzing and comparing, limitations on the dataset could increase the accuracy of prediction. The result states that limitations during the study on diabetes risk factors could lead to high-accuracy correlation results. This study provides a good guide for other research on diabetes risk factors to investigate people with more common points in their daily lifestyle.

Keywords: Diabetes, Limitation, Logistic Regression, Decision Tree model.

1. Introduction

In recent years, people with diabetes are increasing rapidly, and the patients are of all ages. Diabetes can be classified into many types, the most common 3 groups are Type 1 diabetes, Type 2 diabetes, and Gestational diabetes [1]. Diabetes not only changes the lifestyle of patients but also leads to mental disease [2]. Although Diabetes is a chronic disease, it has already become one of the ten main causes of death [3]. Risks that accompany diabetes leads to public attention to this disease.

Diabetes is mainly caused by two types of factors, the first is genetic factors and the second is environmental factors [4, 5]. Research in the diabetes field is rich and diverse, which covers therapy methods, risk factors, life influence, and so on. Many research studies focus on the risk factors of diabetes, which makes this field a good academic development prospect. A study in 1997 found that men have an increased chance of developing clinical diabetes when they become weight, increase fat distribution, and become obese [6]. Some researchers studied the relationship between diet, lifestyle, and other physiological factors and diabetes in women [7]. Jing made an effort to verify which factor has a relationship with diabetes by Logistic Regression method [8]. Oppermann and Spritzer conducted a study on Brazilian women and focused on some women-only factors, which gain early age at menarche, and systolic BP were highly and independently related to diabetes [9]. A published paper focused on factors affecting diabetes in the liver transplant population [10]. There was also a study that tried to find the corresponding factors of Type 1 diabetes among children and adolescents [11].

Among all reference studies, there are some common points among them. First of all, every study limits the object of study to a class of people, such as women, men, or the place objects live. Moreover,

DOI: 10.54254/2753-8818/52/2024CH0149

some studies even focus on typical types of diabetes. All such restrictions will lead to more accurate results, but such restrictions also cause the limitations of research conclusions. The study that used the method of Logistic Regression only gained a fuzzy relationship between diabetes and the risk factors.

The reason why a large number of studies abandoned the general applicability of the findings, but focused on small and specific samples is worth studying. Therefore, this article will investigate whether additional restrictions are necessary for diabetes risk factors. Datasets that are not restricted in all aspects will be used, and the logistic regression method will be used for analysis to determine whether a conclusion with high precision and wide application can be obtained. This paper will help other researchers who want to study diabetes to conduct better sample collection.

2. Methods

2.1. Data Source

The data to be used in this paper is from the Kaggle website, and the dataset contains 17 risk factors and detailed information was collected from 70693 people. The strength of this dataset is it contains a large amount of data, and the author who built this dataset balanced all its classes. This dataset is actually from The Behavioral Risk Factor Surveillance System (BRFSS), which is a health-related telephone survey made by the Center of Disease Control and Prevention (CDC) annually. Each year, the survey collects responses from over 400,000 Americans on health-related risk behaviors, chronic health conditions, and the use of preventative services. It has been conducted every year since 1984. This original dataset contains responses from 441,455 individuals and has 330 features. These features are either questions directly asked of participants, or calculated variables based on individual participant responses.

2.2. Variable Introduction

The variables used in this paper are as table 1 shows. The specific situation of the various variables is shown in Table 1 below, and the introduction of some of the abbreviations and criteria are described as follows. High cholesterol (High Chol), Cholesterol check (Chol Check), Body Mass Index (BMI), Coronary heart disease or myocardial infarction (Heart Diseaseor Attack), Physical activity (Phys Activity), Drinks per week (Hvy Alcohol Consump), General health (Gen Hlth), Poor mental health (Ment Hlth), Physical illness or injury (Phys Hlth), Difficulty walking (Diff Walk).

The special standards are following. Cholesterol check is whether cholesterol check in 5 years. 1 of smoker means smoked at least 100 cigarettes in their entire life.1 for physical activity means having physical activity in the past 30 days not including job. The value of Mental health relates to days of poor mental health scale of 1-30 days. Physical illness or injury value indicates physical illness or injury days in the past 30 days scale 1-30.

Table 1. Variable introduction.

Variable	Types	Range/Meaning
Age(x1)	Categorical	13-level age category
Sex(x2)	Categorical	0=female,1=male
High Chol(x3)	Categorical	0 = no,1 =yes
Chol Check(x4)	Categorical	0 = no,1 =yes
BMI(x5)	Numeric	12 to 98
Smoker(x6)	Categorical	0 = no,1 =yes
Heart Diseaseor Attack(x7)	Categorical	0 = no,1 =yes
Phys Activity(x8)	Categorical	0 = no,1 =yes
Fruits(x9)	Categorical	0 = no,1 =yes
Veggies(x10)	Categorical	0 = no,1 =yes
Hvy Alcohol Consump(x11)	Categorical	0 = no,1 =yes

DOI: 10.54254/2753-8818/52/2024CH0149

Table 1. (continued).

Gen Hlth(x12)	Categorical	1 = excellent 2 = very good 3 = good 4 = fair 5 = poor
Ment Hlth(x13)	Numeric	1 to 30
Phys Hlth(x14)	Numeric	1 to 30
Diff Walk(x15)	Categorical	0 = no,1 =yes
Stroke(x16)	Categorical	0 = no,1 =yes
Diabetes(x17)	Categorical	0 = no,1 =yes

2.3. Method Introduction

2.3.1. Logistic Regression

The Logistic function is also called the Sigmoid function. This function has the following form: Logistic Regression is a classical classification method, which is especially good for binary classification problems. It is a classification algorithm based on a generalized linear model. Logistic regression can easily solve the classification problem only by using a logistic function based on the general linear regression model. The Logistic function is also called the Sigmoid function. This function has the following form:

$$y = \frac{1}{1+e^{-z}} \tag{1}$$

By expressing z as $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots$ (2), substituting z back to function (1), then having logarithmic operation on both sides simultaneously. This paper can get the logistic regression mode:

$$\ln\left(\frac{y}{1-y}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p \tag{3}$$

Based on the data type of Diabetes, is classified into whether have diabetes or not. Binomial logistic regression analysis is applied. The paper uses SPSS to receive the previous information and gain the result of the regression function and the evaluation of fitting results.

2.3.2. Decision Tree Model

Decision Tree model is a model that displays decision rules and classification results with a tree-like structure. As a type of inductive learning algorithm, the core of the decision tree model lies in transforming seemingly disordered and chaotic known data into a model resembling a tree structure and then using certain technical means to predict unknown data.

3. Results and Discussion

3.1. Descriptive Analysis

In the process of dealing with practical problems, this paper randomly extracts the data, retain its original distribution characteristics, and uses the random 10,246 data for analysis. According to statistics and analysis, age and whether have diabetes were taken as examples. As shown in Figure 1 and Figure 2, the sample size this paper extracted also maintained the same balance as the overall sample and could be analyzed. After a similar analysis, the other variables maintained the same balance. This result shows the randomly selected sample could be used for analysis.

DOI: 10.54254/2753-8818/52/2024CH0149

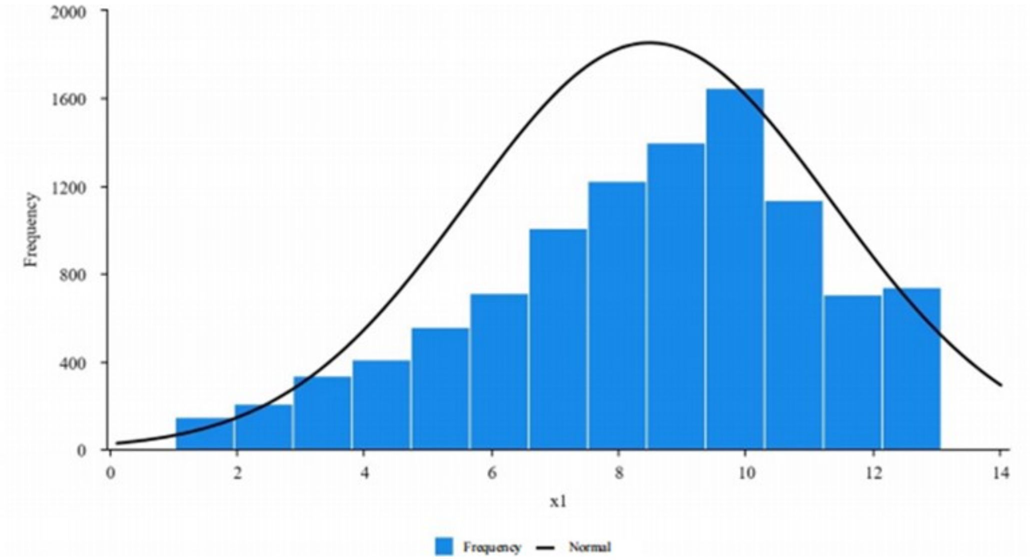

Figure 1. Age distribution in sample.

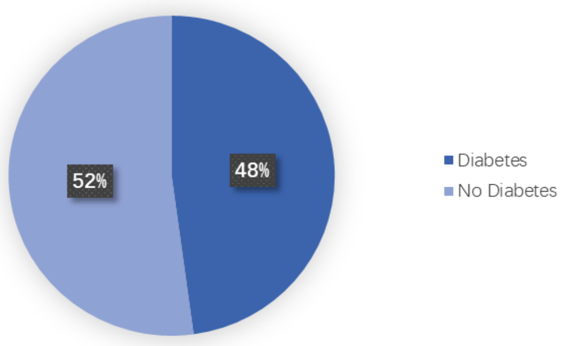

Figure 2. The proportion of diabetes in sample.

3.2. Correlation Analysis

By using the tool SPSS, the correlation between all risk factors and diabetes is shown in Figure 3. By analyzing the result, even though the correlation exists and is remarkable, the responding degree is low. There are 4 negative correlation factors and 13 positive correlation factors. The selection of relevant factors is confirmed by the analysis result. All risk factors could be chosen and put into binomial logistic regression analysis. The next step is to check the colinearity relation between risk factors.

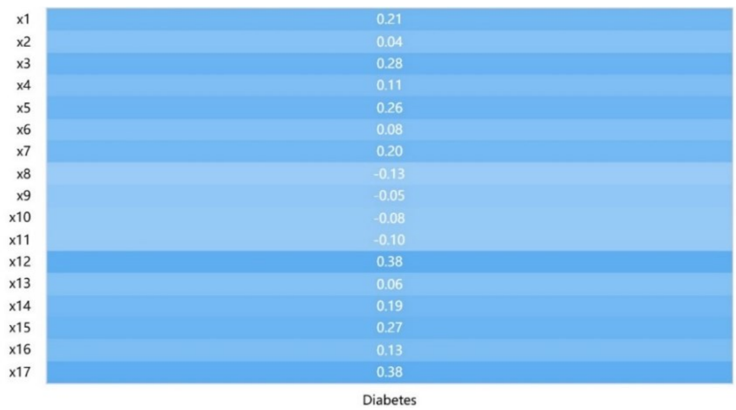

Figure 3. The correlation between dependent and independent variables.

Based on the Kendall corresponding analysis the relationship between the two factors is diverse, including positive relation, negative relation, and irrelevancy (Figure 4). However, for each risk factor, there are no colinearity or multicollinearity risk factors. This implies that the binomial logistic regression hypothesis is satisfied. The binomial logistic regression analysis could be used now.

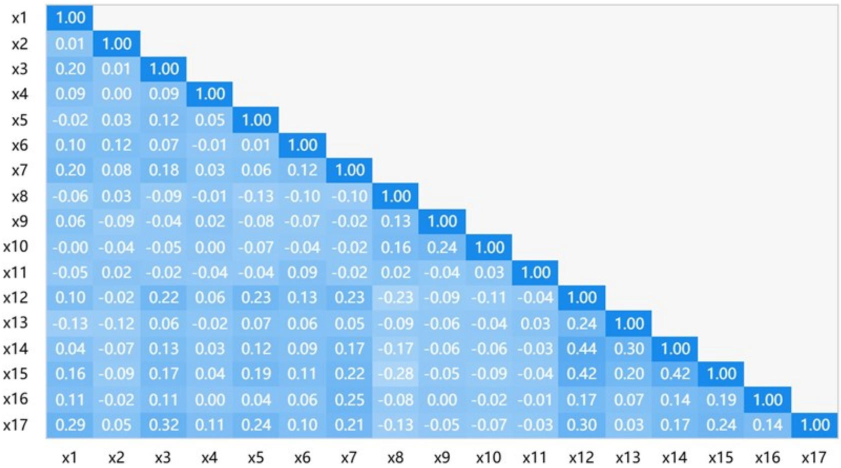

Figure 4. The correlation between independent variables.

3.3. Model Results

The results of binomial logistic regression are shown in Table 2. According to the analysis of p-values less than or equal to 0.05, Sex, High Chol, Chol Check, BMI, Heart Diseaseor Attack, Gen Hlth, Diff Walk, Stroke, Veggies, Hvy Alcohol Consump, PhysHlth produce significant positive effects on Diabetes. Veggies, Hvy alcohol consump, Phys hlth produce obvious negative effects on Diabetes. However, Smoker, Phys Activity, Fruits, and Ment Hlth do not have any effect on Diabetes.

Table 2. Binomial logistic regression analysis result summary.

Term	β	SD	z	Wald χ2	p	OR	OR 95% CI
x1	0.138	0.010	13.916	193.643	0.000	1.148	1.126 ~ 1.170
x2	0.227	0.049	4.647	21.590	0.000	1.255	1.140 ~ 1.381
x3	0.565	0.049	11.509	132.467	0.000	1.760	1.598 ~ 1.937
x4	1.207	0.200	6.044	36.531	0.000	3.344	2.261 ~ 4.947
x5	0.066	0.004	16.228	263.350	0.000	1.068	1.060 ~ 1.077
x6	-0.011	0.049	-0.215	0.046	0.830	0.990	0.899 ~ 1.089
x7	0.209	0.076	2.746	7.542	0.006	1.233	1.062 ~ 1.432
x8	0.054	0.056	0.971	0.943	0.332	1.055	0.947 ~ 1.177
x9	-0.020	0.051	-0.392	0.153	0.695	0.980	0.887 ~ 1.083
x10	-0.147	0.061	-2.427	5.889	0.015	0.863	0.766 ~ 0.972
x11	-1.030	0.133	-7.748	60.024	0.000	0.357	0.275 ~ 0.463
x12	0.619	0.029	21.537	463.863	0.000	1.857	1.755 ~ 1.965
x13	-0.000	0.003	-0.139	0.019	0.889	1.000	0.993 ~ 1.006
x14	-0.008	0.003	-2.723	7.412	0.006	0.992	0.986 ~ 0.998
x15	0.170	0.067	2.547	6.488	0.011	1.185	1.040 ~ 1.350
x16	0.179	0.108	1.653	2.731	0.098	1.196	0.967 ~ 1.479
x17	0.734	0.052	14.220	202.208	0.000	2.084	1.883 ~ 2.306
Intercept	-6.895	0.273	-25.218	635.937	0.000	0.001	0.001 ~ 0.002

DOI: 10.54254/2753-8818/52/2024CH0149

The gained formula of binomial logistic regression analysis result is:

$$\ln\left(\frac{y}{1-y}\right) = -6.895 + 0.138x_1 + 0.227x_2 + \cdots + 0.179x_{16} + 0.734x_{17} \tag{4}$$

Table 3. Binomial logistic regression analysis accuracy.

		Prediction 0	Prediction 1	Accuracy	Error
True value	0	2516	1782	58.54%	41.46%
	1	1070	5488	83.68%	16.32%
Collect				73.73%	26.27%

Although the results show that about 12 factors relate to diabetes. The accuracy of the model is only about 70%. As shown in Table 4, the prediction accuracy obtained by binary logistic regression is relatively low, to make sure that the output of this model is correct, the decision tree model is used to test twice. The outcome is shown in Table 4, which shows that the accuracy is similar to the result by using binary logistic regression. The accuracy of the decision tree model is 70%, which is close to the accuracy of the binomial logistic regression method.

Table 4. Decision tree model accuracy.

Term	Accuracy	Recall rate	f1-score	N of samples
0	0.74	0.70	0.72	1100
1	0.67	0.71	0.69	951
Accuracy			0.70	2051
Mean value	0.70	0.70	0.70	2051
Mean value(synthesis)	0.70	0.70	0.70	2051

To find the effect of the limitation, the data who maintain the same age stage and gender are chosen randomly. As shown in Table 5, by again using binomial logistic regression, the result this paper gain has a higher accuracy, which improves to 81.10%. Meanwhile, the number of risk factors that may have a significant influence on diabetes is decreased.

Table 5. Binomial logistic regression analysis accuracy with limitation.

		Prediction 0	Prediction 1	Accuracy	Error
True value	0	1262	138	90.10%	9.90%
	1	235	341	59.20%	40.80%
Collect				81.10%	18.90%

3.4. Discussion

In this study, the binomial logistic regression model was used to analyze the effects of different risk factors on diabetes. The decision tree model is used for testing. The accuracy of the binomial logistic regression model is about 70% even changing the method to the decision tree model, the problem of low accuracy cannot be solved. By adding the limitation, the model accuracy under data with limitation is higher than data without it, which shows limitation is a point to the study on risk factors.

Even with the limitation, the accuracy is still only about 80%. The reason why accuracy is low could be analyzed from three aspects. First, the absence of restrictions leads to more and larger differences between individuals, which causes many unrelated factors to be taken into account with large volume. These unrelated, contingent factors showed an association with diabetes and affected the final regression results. Second, the data has no classification on the type of diabetes. Since the causes of different types of diabetes are different, now the data analysis may be influenced. The type and proportion of diabetes patients currently under consideration are unknown. Both type and proportion affect the accuracy of regression results. Third, the risk factors this paper considered only contain the environmental factors. There is a possibility that the genetic factor has a higher relation with diabetes.

4. Conclusion

This paper shows that with the limitation on data, the binomial logistic regression could have higher accuracy and the correlated risk factors of diabetes will be less. This indicates that data with more common points, such as age stage, and gender, even with smaller amounts of data, the result could be simpler and more accurate. Research that based on limitation, could gain better results with a smaller workload.

The results of the study show that in the process of research, it is really necessary to limit the study object so that the study of diabetes risk factors can get better accuracy in correlation analysis. It is possible to include geographical restrictions, gender restrictions, diabetes type restrictions, and restrictions in different diabetes periods to better obtain fitting results. The researcher recommends that freshmen should limit their subjects to risk factors if they want to minimize the influence of unrelated factors and get better correlation results.

Using the results of this paper as a starting point, how much different restrictions can lead to improvements in the accuracy of regression could be investigated. The results trend under the change of number of limitations could be studied to find the most suitable number of limitations. All the branches of this study will make guidance on other studies to focus on the analysis of how environmental factors influence diabetes. The thought of this study could also be applied to the study of other types of disease, which is useful to research pathogenic factors in the medical field.

References

[1] Ondiagnosis, T.E.C. (1997) The Report of the Expert Committee on the Diagnosis and Classification of Diabetes Mellitus. Diabetes Care, 22.
[2] Richardson, L.K., Egede, L.E. and Mueller, M. (2008) Effect of race/ethnicity and persistent recognition of depression on mortality in elderly men with type 2 diabetes and depression. Diabetes Care, 31(5), 880.
[3] D'Souza, M.J. , Li, R.C., Gannon, M.L., et al. (2019) Leading Causes of Death Information Due to Diabetes, Neoplasms, and Diseases of the Circulatory System, Issues Cautionary Weight-Related Lesson to the US Population at Large. IEEE network.
[4] Green, A. (1990) The role of genetic factors in the development of insulin-dependent diabetes mellitus. Current Topics in Microbiology & Immunology, 164, 3-16.
[5] Longnecker, M.P., et al. (2001). Environmental contaminants as etiologic factors for diabetes. Environmental Health Perspectives Supplements.
[6] Chan, J.M., Rimm, E.B., Colditz, G.A., et al. (1994) Obesity, Fat Distribution, and Weight Gain as Risk Factors for Clinical Diabetes in Men. Diabetes Care, 17(9), 961-969.
[7] Lako, J.V. and Nguyen, V.C. (2015) Dietary patterns and risk factors of diabetes mellitus among urban indigenous women in Fiji.. Asia Pacific Journal of Clinical Nutrition, 10(3), 188-193.
[8] Jing, Z. (2006) Logistic Regression Analysis of Diabetes Pathogenic Factors. Medicine and Society.
[9] Oppermann, K. and Spritzer, P.M. (2024) Prevalence and risk factors associated with diabetes mellitus among middle-aged women in southern Brazil: a population-based study. Menopause, 31(3), 225-230.
[10] Campos, et al. (2022) Risk factors associated with diabetes after liver transplant. Archives of endocrinology and metabolism 66, 2, 182-190.
[11] Andrea, L., Varga, B., et al. (2014) Factors influencing the diabetes-specific health-related quality of life in children and adolescents with type 1 diabetes mellitus. Journal of Child Health Care for Professionals Working with Children in the Hospital & Community, 18(3), 253-60.