Prediction of Diabetes Risk in Young and Middle-aged Adults: Machine Learning Analysis based on Health Behavior and Physiological Indicators

Ruohan Zhang*

Shanghai University of Medicine & Health Sciences, Shanghai, 201318, China

*Corresponding author: b22020102007@stu.sumhs.edu.cn

Abstract. One of the most common chronic illnesses in the world is diabetes. In recent years, the prevalence of diabetes is increasing among young and middle-aged people. This article aimed to making a predict model about diabetes used a balanced data, focusing on health behavior and physiological indicators mainly, from Behavioral Risk Factor Surveillance System (BRFSS) 2015. The data was analyzed by machine learning method and two models have been constructed in this article, which are logistic regression model and random forest model, in order to choose the model with higher accuracy, the accuracy rate, f1-score and a confusion matrix of that two models have been compared. The findings of the study indicated that logistic regression model is better for using in this dataset with higher accuracy. However, its accuracy is 84.42%, which is not high enough for actual use. There are 82 false positives (FPs) and 228 false negatives (FNs) as the prediction outcome of logistic regression model. Based on these findings, it is suggested that more updated variables, different parameters' selection and other predict models (such as k nearest-neighbor, decision tree etc.) should be considered in model construction.

Keywords: Diabetes; logistic regression; random forest.

1. Introduction

Diabetes mellitus (DU), a chronic noninfectious disease, is caused by insufficient insulin secretion or defective insulin action, or both of them [1]. According to the etiological classification system published by World Health Organization (WHO) in 1999, diabetes mellitus is mainly categorized as four categories: type 1 diabetes (T1D), type 2 diabetes (T2D), gestational diabetes and other special forms of diabetes. Diabetes mellitus is currently one of the fastest growing chronic diseases worldwide. In 2021, the prevalence of diabetes mellitus among adults (aged from 20 to 79) worldwide is estimated at 537 million [2]. By 2045, this prevalence is expected to reach 783 million, an increase of 46 percent, which indicates that diabetes mellitus presents an increasing trend in its incidence in the global significantly [2]. Moreover, there has shown a significantly growing tendency in young-onset T2DM among younger people (aged less than 40) in both developed and developing countries, which indicated that diabetes mellitus is occurring at a younger age [3]. At present, the main complications of diabetes mellitus include diabetic kidney disease, diabetic retinopathy, diabetic neuropathy and cardiovascular disease [4]. All of these complications pose a serious threat towards individual health chronically. Among them, the most common potential cause of death is cardiovascular diseases [5]. Therefore, the degree of the danger of diabetes cannot be ignored, and more measures should be considered. This situation gives rise to the objective of this article, which is to create diabetes prediction models focusing on younger individuals.

The typical symptoms of diabetes mellitus include increased water intake, increased urine output, increased food intake and unexplained weight loss [6]. According to the American Diabetes Association's (ADA) 2024 diabetes care standards, it is recommended to meet any of A1C≥48mmol/mol, FPG≥7.0mmol/L, 2-h PG≥11.1 mmol/L during Oral Glucose Tolerance Test (OGTT), a random plasma glucose≥11.1mmol/L to diagnose diabetes mellitus (non-pregnant status) [7]. Due to the high prevalence and the younger trend of diabetes mellitus as mentioned above, the most important thing for early screening among symptomless adults for prediabetes is to make early detection, early diagnosis and early treatment [8]. Early screening and daily monitoring enable early

intervention in the lifestyle and diet of people at risk, delaying or preventing the morbidity, which will have great significance to diabetes mellitus.

Meanwhile, the methods of Artificial Intelligence-based algorithms and machine learning (such as Regression model, Decision Tree model, Random Forest and so on) in regard with a great many data samples (like laboratory data, examination data, survey data and so on) contribute to a more comprehensive prediction and personalized screening towards the management of chronic diseases [9]. There are many prediction models now for prediabetes and diabetes have been proven with a high accuracy of prediction outcome [10]. A 10-year prospective cohort study based on population genomes in Korea used a machine learning model combined logistic regression and random forest to predict type 2 diabetes risk with AUC up to 0.883 in succeed [11]. The availability of such methods presents the risk factors that lead to diabetes mellitus sufficiently and provides guidance for people at risk, which are good for them. This article will use the data of Behavioral Risk Factor Surveillance System (BRFSS), which is a survey collected by America CDC annually, to create predict models for predicting diabetes among young and middle-aged adults (aged from 18 to 44), based on health behavior and physiological indicators mainly. Through logistic regression model and random forest model, this article is trying to figure out whether the survey questions provided in BRFSS are able to predict diabetes with high accuracy.

2. Methods

2.1. Data Source

The dataset used in this article is available on the Kaggle website organized by 70692 survey responses from cleaned BRFSS 2015. The original dataset contains 70692 observations and 18 variables, and after selecting the aim population (aged from 18 to 44), the dataset used in this article contains 10737 observations and 18 variables in the .XLS format with no missing value. Furthermore, the cleaned dataset is balanced, as the ratio of people with diabetes to those without diabetes is 20.57% (higher than 20%).

2.2. Sample Selection

Table 1 shows all the variable names, types and explanations used in this article.

Table 1. Variable explanation

Variable	Type	Description and values
Age	X1	1 = Age 18 to 24; 2 = Age 25 to 29; 3 = Age 30 to 34; 4 = Age 35 to 39; 5 = Age 40 to 44
Sex	X2	0 = female; 1 = male
HighChol	X3	0 = no high cholesterol; 1 = high cholesterol
CholCheck	X4	Cholesterol check in 5 years (0 = no; 1 = yes)
BMI	X5	Body Mass Index
Smoker	X6	Smoked at least 100 cigarettes in your entire life? (0 = no; 1 = yes)
HeartDiseaseAttack	X7	CHD or MI (0 = no; 1 = yes)
PhysActivity	X8	Physical activity in past 30 days (0 = no; 1 = yes)
Fruits	X9	Consume fruit 1 or more times per day (0 = no; 1 = yes)
Veggies	X10	Consume vegetables 1 or more times per day (0 = no; 1 = yes)
HvyAlcoholCons	X11	Men >= 14 drinks/week; Women >= 7 drinks/week(0 = no; 1 = yes)
GenHlth	X12	health (1 = excellent; 2 = very good; 3 = good; 4 = fair; 5 = poor)
MentHlth	X13	Days of poor mental health scale 1-30 days
PhysHlth	X14	Physical illness or injury days in past 30 days scale 1-30
DiffWalk	X15	Had serious difficulty walking or climbing stairs? (0 = no; 1 = yes)
Stroke	X16	Have you ever had a stroke? (0 = no; 1 = yes)
HighBP	X17	0 = no high BP; 1 = high BP
Diabetes	Y	0 = no diabetes; 1 = diabetes

This dataset mainly contains categorical variables based on health behavior and physiological indicators. Moreover, this dataset comprises 15 categorical variables which had been meticulously encoded and 3 numerical variables (Table 1).

2.3. Method Introduction

This article will choose logistic regression and random forest model to predict diabetes among young and middle-aged adults. The dataset that contains 10737 observations and 18 variables mentioned above is divided into training set and testing set in a ratio of 8:2. For accuracy evaluation, accuracy rate, precision rate, recall rate and f1-score (the harmonic mean of precision rate and recall rate). Meanwhile, the confusion matrix will be used to visualize the false positives (FPs) and the false negatives (FNs).

In logistic regression, stepwise regression analysis is used to construct a model that reduces variables but fits the dependent variables well to decrease the complexity of the model and maintain the prediction accuracy of diabetes. The formula of logistic regression that will be used in this article is:

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

The parameters of random forest model are selected based on cross-validation, in order to obtain the best random forest model with high accuracy of predicting diabetes.

3. Results and Discussion

3.1. Variable Distribution

BMI (X5) and GenHlth (X12), having a higher correlation with diabetes, are visualized in Figure 1. From the boxplot, the distribution of BMI and GenHlth in non-diabetes (0) or diabetes (1) has been displayed clearly that the people with diabetes tend to have a higher X5 and X12 than those without diabetes. Additionally, the median of people with diabetes is higher than that of people without diabetes, which demonstrates a more concentrated trend. Besides, the dashed line of X5 extends from the minimum value to the maximum value, and no outliers are shown.

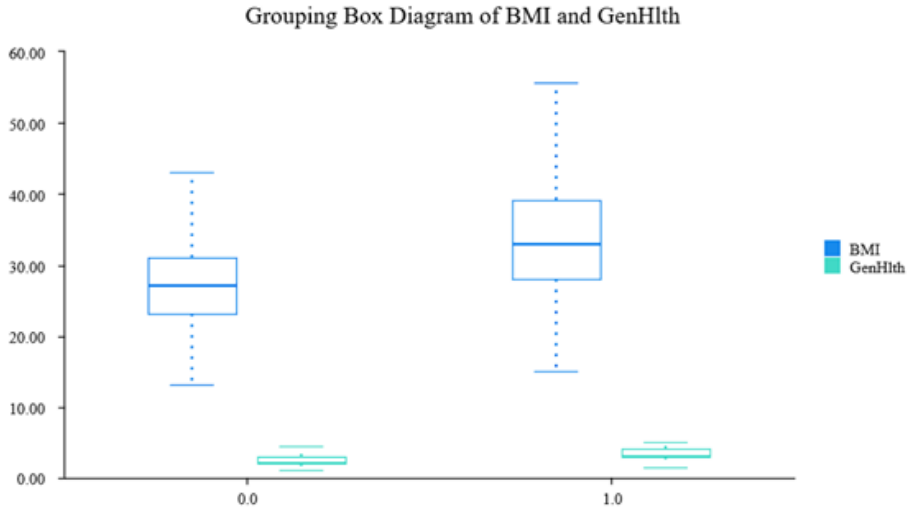

Fig. 1 Grouping Box Diagram of BMI and GenHlth

3.2. Correlation Results

Table 2 shows the result of spearman correlation analysis among all these variables. According to spearman correlation table, all variables show significant correlations with the y outcome (Diabetes), with p values less than 0.01. Among them, there is a significant negative correlation between

"Diabetes (X18)" and "Sex (X2)", "Fruits (X9)", "PhysActivity (X8)", "HvyAlcoholConsump (X11)", "Veggies (X10)". The remaining variables show a significant positive correlation with X18, and there is no collinearity between all the variables. Therefore, all 18 variables will be included.

Table 2. Spearman correlation

	Diabetes
Stroke	0.095**
Age	0.197**
Sex	-0.027**
HighChol	0.326**
CholCheck	0.095**
Smoker	0.076**
BMI	0.335**
HeartDiseaseorAttack	0.146**
Fruits	-0.065**
PhysActivity	-0.116**
GenHlth	0.406**
PhysHlth	0.228**
HighBP	0.342**
DiffWalk	0.251**
HvyAlcoholConsump	-0.028**
MentHlth	0.104**
Veggies	-0.065**

* $p<0.05$ ** $p<0.01$

3.3. Logistic Regression Results

The result of logistic regression is all presented in Table 3. The standardized coefficients of X3, X5, X12 and X17 are the four highest-ranking ones, which means that for each of these variables, if one unit is added, the log odds will increase by 0.163, 0.153, 0.250 and 0.136 respectively. Besides, the coefficients of X2 and X13 are negative values, meaning that an increase in X2 and X13 will reduce the probability of diabetes occurring.

The value of R^2 in Table 3 has shown that X1, X2, X3, X4, X5, X7, X12, X15 and X17 can explain 29.2% (R^2= 0.292) of the changes in diabetes. Moreover, the F test (F=441.371, p=0.000<0.05) indicates that the model is valid, and the formula of the model is:

$$y(diabetes) = -0.495 + 0.022x_1 - 0.020x_2 + 0.152x_3 + 0.099x_4 + 0.008x_5 + 0.109x_6 + 0.097x_7 - 0.001x_8 + 0.073x_9 + 0.132x_{10} \quad (2)$$

Therefore, only these nine variables are used in logistic regression model.

Table 3. The result of logistic regression (n=10737)

	Unstandardized Coefficients		Standardized Coefficients	t	p	Collinearity Diagnosis	
	B	Standard Error	Beta			VIF	Tolerance
Constant	-0.495	0.021	-	-23.573	0.000**	-	-
Age	0.022	0.003	0.070	8.272	0.000**	1.081	0.925
Sex	-0.020	0.007	-0.024	-2.943	0.003**	1.032	0.969
HighChol	0.152	0.008	0.163	18.475	0.000**	1.178	0.849
CholCheck	0.099	0.014	0.057	6.966	0.000**	1.008	0.992
BMI	0.008	0.000	0.153	17.131	0.000**	1.215	0.823
HeartDiseaseorAttack	0.109	0.023	0.040	4.814	0.000**	1.060	0.943
GenHlth	0.097	0.004	0.250	25.598	0.000**	1.439	0.695
MentHlth	-0.001	0.000	-0.019	-2.205	0.027*	1.180	0.847
DiffWalk	0.073	0.013	0.050	5.458	0.000**	1.260	0.793
HighBP	0.132	0.009	0.136	14.858	0.000**	1.276	0.784
R^2				0.292			
Adjusted R^2				0.291			
F				F (10,10726)=441.371, p=0.000			
D-W value				0.579			

Comment: dependent variable = Diabetes

* p<0.05 ** p<0.01

The confusion matrix displays the logistic regression model's accuracy and performance that its accuracy and f1-score are 84.42% and 0.84, respectively.

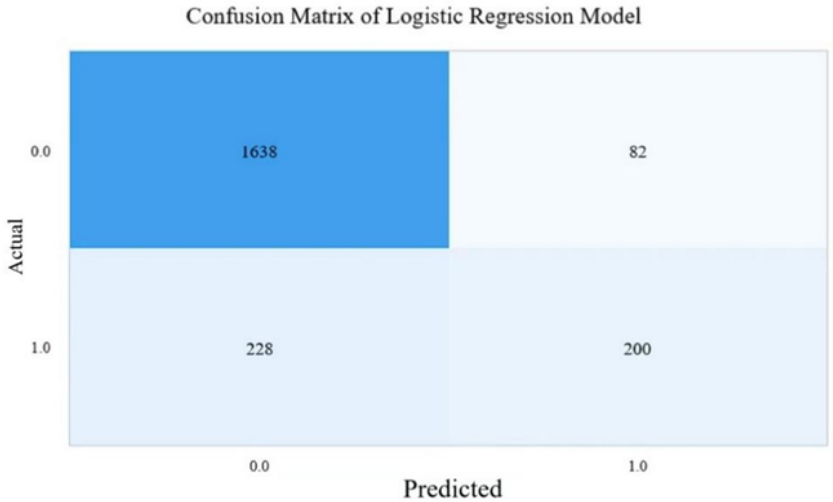

Fig. 2 Confusion Matrix of Logistic Regression Model

Figure 2 demonstrates that a subset of those who did not have diabetes were predicted to have diabetes, while a subset of those who did were predicted not to have diabetes. What is more, the

logistic regression model has a prediction with 82 false positives (FPs) and 228 false negatives (FNs). Ideally, the smaller FPs and FNs are, the better the model will be in actual use. However, there is actually a trade-off relationship between them, so that it will be balanced with f1-score.

3.4. Random Forest Model

In random forest model, methods of parameter net and grid search training are carried out to explore the optimum parameter. Ultimately, this model has 200 decision trees, with a maximum tree depth of 10 and a minimum node sample size of 10. Figure 3 represents the feature importance in the random forest model, which stands for the degree of importance of each variable in contributing to the model. The proportion of X12 is 24.09%, which holds the highest weight and plays a crucial role in the model construction. The proportion of X5 is 20.30%, which is of the second highest importance and also plays an important role in the model construction. The combined proportion of the first five variables accounted for 74.74% in this random forest model, which is of significant important. Finally, the accuracy of the model on the test set is 83.95%, the value of f1-score is 0.84, and the result of confusion matrix is presented in Figure 4.

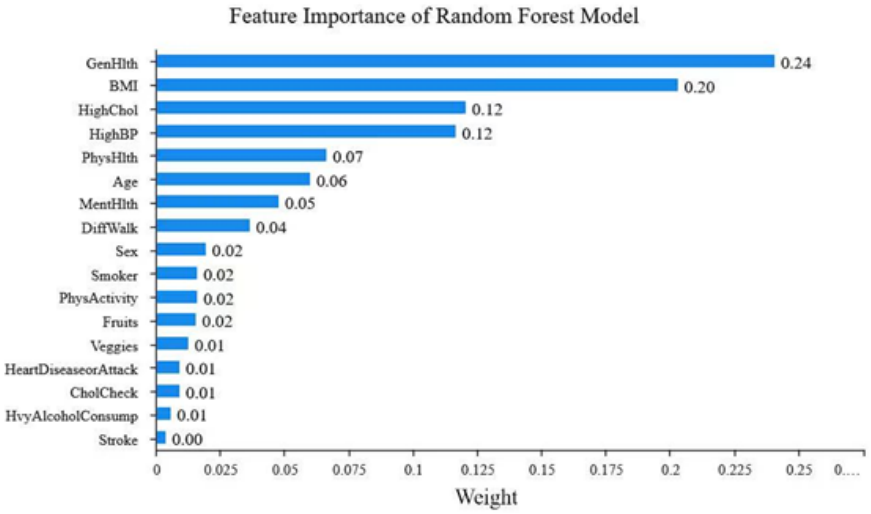

Fig. 3 Feature Importance of Random Forest Model

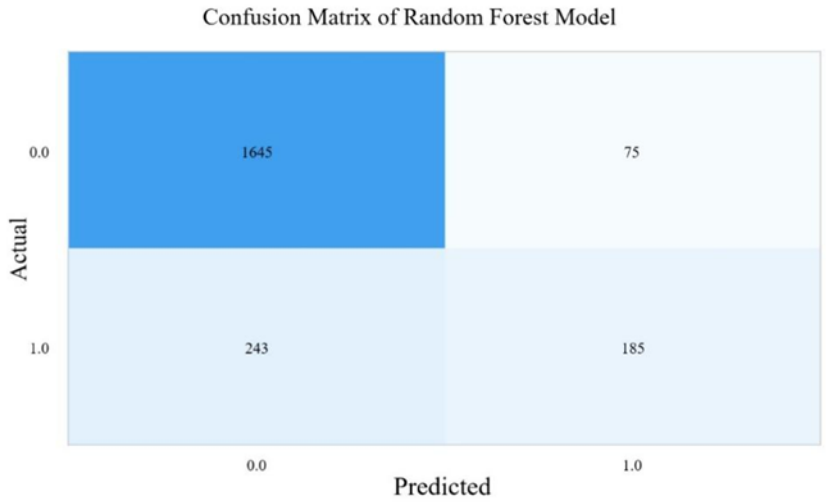

Fig. 4 Confusion Matrix of Random Forest Model

It can be seen from Figure 4 that 75 people who did not have diabetes were mistakenly predicted to have diabetes, and 243 people who did have diabetes were mistakenly predicted not to have diabetes.

3.5. Model Comparison Results

The accuracy of logistic regression model is 84.42% and its f1-score is 0.84. The accuracy of random forest model is 83.95% and its f1-score is 0.84 that is equal to logistic regression model. Consequently, by comparing two prediction model, logistic regression model and random forest model, about diabetes concentrated on health behavior and physiological indicators data from BRFSS, logistic regression model is finally been selected in this article due to its higher accuracy.

4. Conclusion

In conclusion, the results showed in logistic regression model indicate that all the variables included in the model have significant influence on BRFSS survey about diabetes prediction among the young and middle-aged adults. Additionally, X3, X5, X12 and X17 occupy an important role in the logistic regression model, which have shown that the four variables are of great significance for diabetes prediction. However, the accuracy of the model is not very high enough to be used in actual prediction. The higher of accuracy of the model is, the more precise the prediction results obtained will be, perhaps more variables should be taken into account to improve the model accuracy. It also offers a train of thought for BRFSS to include or explore more survey questions about health behavior and physiological indicators in order to predict diabetes or prediabetes without large-scale physical examination or invasive examination. Besides, there are other models that can be constructed (such as k nearest-neighbor or decision tree, etc.) and different parameters should be used in searching the model with the highest accuracy.

References

[1] Antar S A, et al. Diabetes mellitus: Classification, mediators, and complications; A gate to identify potential targets for the development of new effective treatments. Biomed Pharmacother, 2023, 168: 115734.

[2] Harreiter J, Roden M. Diabetes mellitus: definition, classification, diagnosis, screening and prevention (Update 2023). Wien Klin Wochenschr, 2023, 135(Suppl 1): 7-17.

[3] Magliano D J, et al. Young-onset type 2 diabetes mellitus - implications for morbidity and mortality. Nat Rev Endocrinol, 2020, 16(6): 321-331.

[4] Cole J B, Florez J C. Genetics of diabetes mellitus and diabetes complications. Nat Rev Nephrol, 2020, 16(7): 377-390.

[5] Morrish N J, et al. Mortality and causes of death in the WHO Multinational Study of Vascular Disease in Diabetes. Diabetologia, 2001, 44(2): S14-21.

[6] Diabetes Care Center. Diagnosis and Classification of Diabetes: Standards of Care in Diabetes-2025. Diabetes Care, 2025, 48(Supplement_1): S27-S49.

[7] Diabetes Care Center. Diagnosis and Classification of Diabetes: Standards of Care in Diabetes-2024. Diabetes Care, 2024, 47(Suppl 1): S20-S42.

[8] Davidson K W, et al. Screening for Prediabetes and Type 2 Diabetes: US Preventive Services Task Force Recommendation Statement. JAMA, 2021, 326(8): 736-743.

[9] Subramanian M, et al. Precision medicine in the era of artificial intelligence: implications in chronic disease management. J Transl Med, 2020, 18(1): 472.

[10] Oikonomou E K, Khera R. Machine learning in precision diabetes care and cardiovascular risk prediction. Cardiovasc Diabetol, 2023, 22(1): 259.

[11] Hahn S J, et al. Prediction of type 2 diabetes using genome-wide polygenic risk score and metabolic profiles: A machine learning analysis of population-based 10-year prospective cohort study. Ebio Medicine, 2022, 86: 104383.