Prediction of Diabetes Mellitus based on Logistic Regression and Random Forest

Dingkang Qian*

School of Public Health, Hebei Medical University, Shijiazhuang, 050017, China

* Corresponding Author Email: yiwan@ldy.edu.rs

Abstract. Diabetes mellitus is a metabolic disease characterized by hyperglycemia caused by impaired insulin secretion and various degrees of peripheral insulin resistance. However, the primary mechanism is still unclear. The study analyzed medical data and used machine learning to build two predictive models to predict a patient's risk of heart disease. The dataset consisted of 3,000 clinical cases and 9 independent variables, such as high blood pressure, heart disease, and glycosylated hemoglobin. The target variable was whether the patient was diagnosed with diabetes mellitus. During this study, the logistic regression model initially had an accuracy of about 96% for the training set and then the fitted model was tested and the accuracy remained good at 94%. The prediction results of this logistic regression model were superior. In order to better predict the occurrence of heart disease, a second set of prediction models was developed in this study using Random Forest, with an accuracy of 100% in the training set, and then the fitted model was tested and the accuracy remained excellent at 96%. The two models used in this study provide a foundation for future testing of diabetes mellitus predictor variables and the discovery of even better predictive models.

Keywords: Diabetes mellitus, logistic regression, random forest, machine learning.

1. Introduction

Diabetes mellitus is a metabolic disease. More than 90% of patients have type 2 diabetes mellitus(T2DM) [1]. The most important factor that leads to diabetes mellitus is problems with insulin production and uptake in the body [2]. If diabetes is not controlled, it will lead to a series of complications [3]. Diabetes mellitus, caused by a combination of genetic and environmental factors, is a clinical syndrome of unknown pathogenesis. Diabetes has been linked to a decline in memory, language, and executive skills [4].

The pathogenesis of diabetes mellitus has not been fully elucidated, but the quest for it never stops. Previous studies have found diabetes mellitus to be a risk factor for cognitive dysfunction [5]. Type 1 diabetes mellitus(T1DM) accounts for more than 90% of childhood diabetes mellitus [6]. In recent years, the incidence of gestational diabetes mellitus has increased [7]. At the same time, during the recent years of new coronavirus disease 2019 (COVID-19) epidemics, studies have added a deeper understanding of diabetes mellitus. Studies suggested that COVID-19 may increase the risk of developing diabetes [8]. One of the common complications is diabetic nephropathy [9].

In addition, in terms of the method of prediction model, Liu used a logistic regression model to build the risk prediction model of T2DM and cognitive dysfunction like most experiments [10]. In order to better reflect the prediction data, this paper also added other machine learning model methods to build the prediction model.

In conclusion, diabetes mellitus prediction research has an important background and significance. By understanding the pathogenesis of diabetes mellitus, the influence of genetic and environmental factors, and by using new predictive models and biomarkers, people can improve the prevention and treatment of diabetes mellitus, reduce the occurrence of complications, and improve the quality of life of patients. These findings will help the medical community better understand diabetes mellitus and provide valuable guidance for public health.

2. Methods

2.1. Data Source

The dataset is a collection of medical and demographic data on people with diabetes mellitus and their diabetes status, organized by individuals and posted on the Kaggle website. The raw data is saved in CSV format. The dataset provides various risk factors that may be associated with diabetes mellitus and is important for research in the field of diabetes mellitus related research. The dataset consists of 3000 case studies that categorize individuals based on the presence or absence of diabetes mellitus. The diagnostic criteria for diabetes mellitus are based on the diagnostic criteria proposed by the WHO Expert Committee on Diabetes Mellitus (1999), which is the current internationally accepted diagnostic criteria for diabetes mellitus.

2.2. Data Introduction

The data used in this paper contains 3000 instances and 9 variables without any missing values. An algorithm was used to categorize diabetes mellitus patients from normal people using all available clinical characteristics including age, gender, presence of high blood pressure, presence of heart disease, BMI, and smoking history, among other factors. Table 1 lists all 9 variables.

Table 1. Different types of variables

Term	Types	Range
Age	Numeric	0 to 80
Gender	Categorical	0-Female, 1-male
Hypertension	Categorical	Have high blood pressure 0-no, 1-yes
Heart disease	Categorical	Have heart disease 0-no, 1-yes
Smoking history	Categorical	1-Not current, 2-no information, 3-never, 4-former, 5-ever, 6-current
BMI	Numeric	10.30 to 69.37
HbA1c level	Numeric	3.50 to 9.00
Blood glucose level	Numeric	80 to 300
Diabetes	Categorical	0-no, 1-yes

2.3. Method Introduction

2.3.1 Logistic regression

The method used in this study was logit regression analysis by SPSS. On the basis of clinical variables, a logistic regression model for predicting diagnosis was constructed. A logistic regression model can be expressed as:

$$P = \frac{exp(\beta_0+\beta_1x_1+\beta_2x_2+\cdots\beta_mx_m)}{1+exp(\beta_0+\beta_1x_1+\beta_2x_2+\cdots\beta_mx_m)} \quad (1)$$

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

$$ln\frac{p}{1-p} = \beta_0 + \beta_1x_1 + \beta_2x_2 \ldots + \beta_mx_m \quad (2)$$

Logistic regression analysis belongs to non-linear regression, which is a method of multiple regression analysis to study the relationship between the dependent variable.

2.3.2 Random forest

Random forest is an algorithm that integrates multiple trees through the Bagging idea of ensemble learning, its basic unit is the decision tree.

3. Results and Discussion

3.1. Descriptive Analysis

Figure 1 shows a bar graph of the age of diabetes mellitus patients. The age of patients is more common in the range of 50 to 80 years. Trends in age and prevalence can also be seen in Figure 1. The prevalence is higher in the elderly.

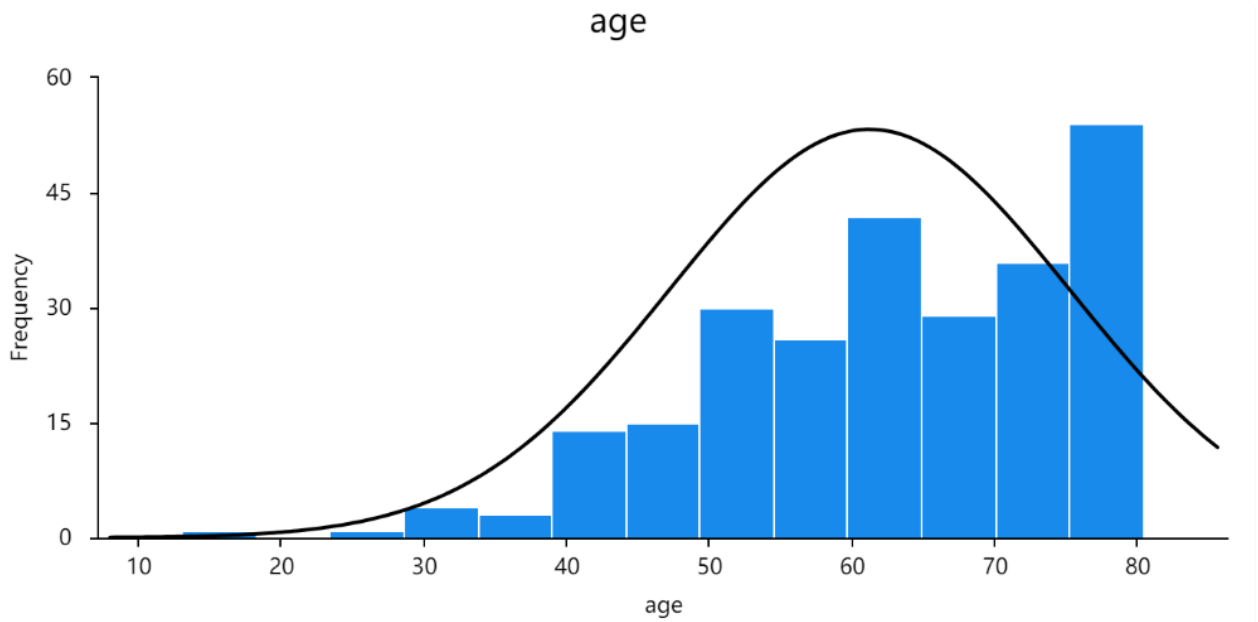

Fig. 1 Histogram of patients' age

Figure 2 shows the patient's hemoglobin A1c level, which is a measure of a person's average blood glucose level over the past 2-3 months. The data shows that most of the patients had hemoglobin A1c levels greater than 7, with the highest values reaching 9 or more.

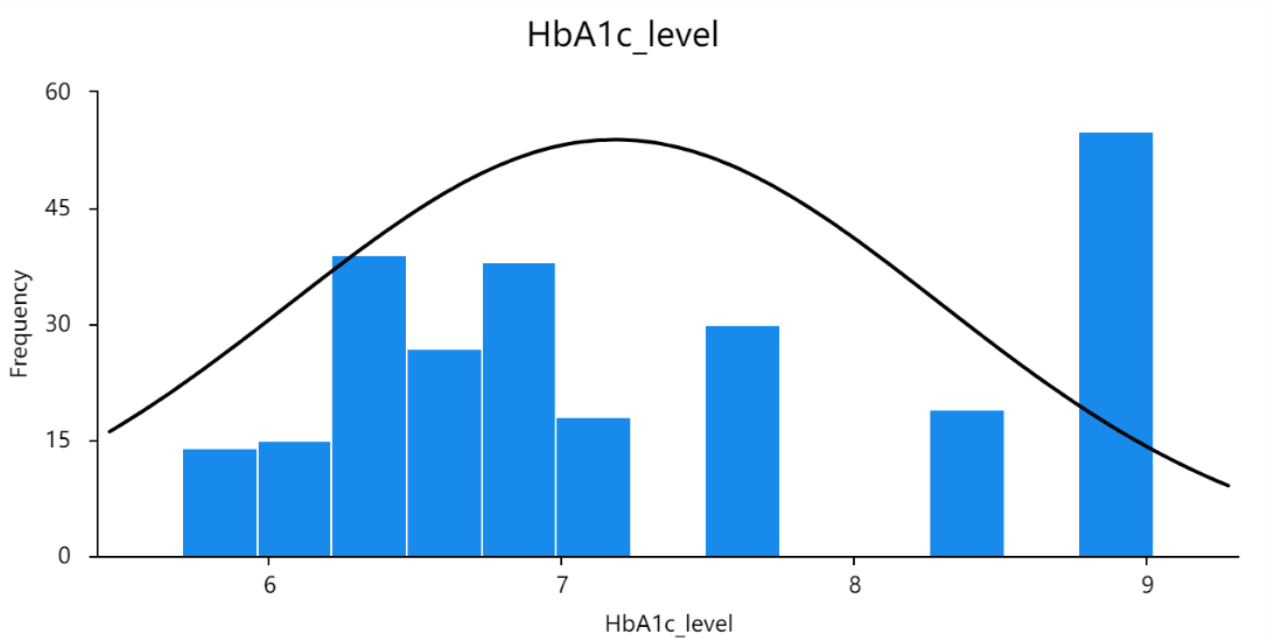

Fig. 2 Histogram of patients' hemoglobin_A1c_level

Body Mass Index (BMI) is a measure of body fat based on weight and height. Normal values for BMI lie between 18.5 and 24.9, and the patient BMI values presented in Figure 3 suggest that the majority of patients are not within the normal range of BMI.

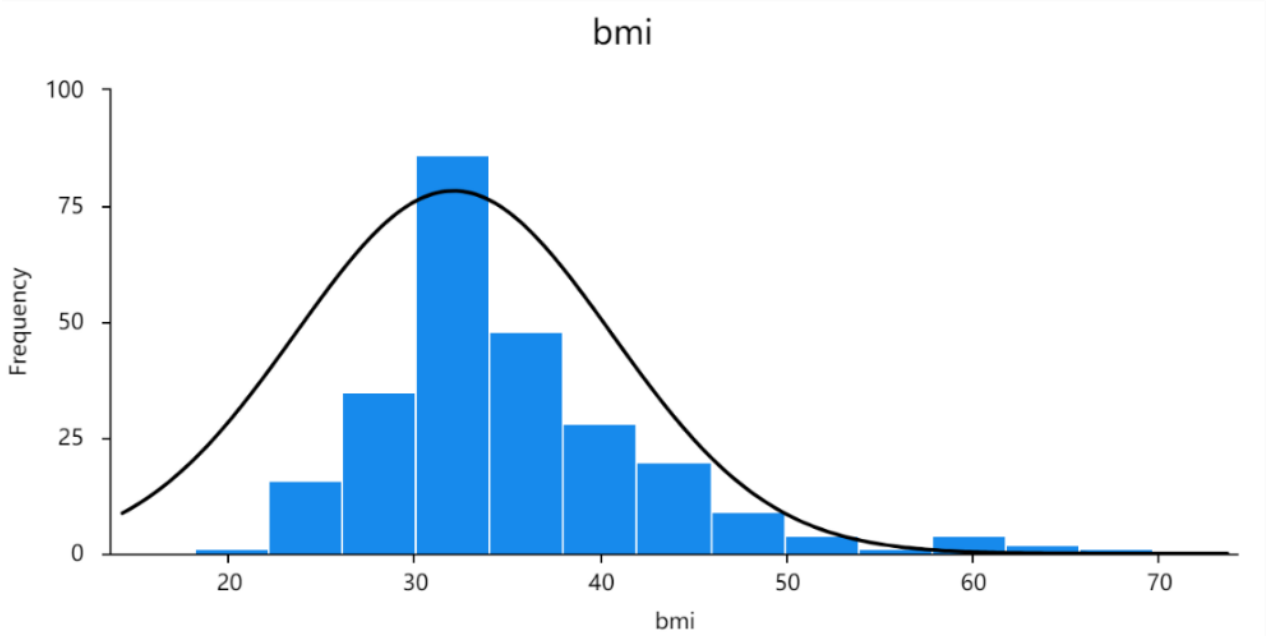

Fig. 3 Histogram of patients' bmi

3.2. Logistic Regression Results

The independent variables comprised gender, age, hypertension, heart disease, smoking history, BMI, HbA1c_level, and blood glucose level, with diabetes as the dependent variable. The training set was scaled to 0.8 and optimized using the lbfgs algorithm, employing logistic regression with L2 regularization and an intercept. Variables were dichotomized into present and absent states. Notably, the training set achieved a 96% accuracy rate, as evidenced in Table 2.

Table 2. Training set model evaluation results

	Precision	Recall	f1-score	Samples
Absence	0.97	0.99	0.98	2195
Presence	0.87	0.68	0.77	205
Accuracy	-	-	0.96	2400
Average	0.92	0.84	0.87	2400
Average(comprehensive)	0.96	0.96	0.96	2400

Upon testing the fitted model, Table 3 confirms sustained accuracy, affirming the robust predictive capacity of the logistic regression model. Notably, the final model attains a 94% accuracy on the test set, accompanied by a 94% recall and an f1 score of 0.93, thus establishing the acceptability of the model results.

Table 3. Testing set model evaluation results

	Precision	Recall	f1-score	Samples
Absence	0.95	0.99	0.97	550
Presence	0.87	0.40	0.53	50
Accuracy	-	-	0.94	600
Average	0.80	0.70	0.75	600
Average(comprehensive)	0.94	0.94	0.93	600

3.3. Random Forest Results

Using gender, age, hypertension, heart disease, smoking history, BMI, HbA1c_level, and blood glucose level as independent variables, with diabetes as the dependent variable, the training set scaled to 0.8. Employing random forest modeling with 100 decision trees, utilizing the gini criterion for node splitting, and no maximum depth restrictions. Table 4 showcases the final model achieving a flawless 100% accuracy on the test set, affirming the model's predictive robustness.

Table 4. Training set model evaluation results

	Precision	Recall	f1-score	Samples
Absence	1.00	1.00	1.00	2195
Presence	1.00	1.00	1.00	205
Accuracy	-	-	1.00	2400
Average	1.00	1.00	1.00	2400
Average(comprehensive)	1.00	1.00	1.00	2400

Following testing of the proposed model, Table 5 demonstrates consistent accuracy, underscoring the robust predictive performance of the random forest model. Notably, the final model achieves a 96% accuracy on the test set, alongside a 96% recall and an f1 score of 0.95, thus indicating the excellence of the model results.

Table 5. Testing set model evaluation results

	Precision	Recall	f1-score	Samples
Absence	0.96	1.00	0.98	550
Presence	1.00	0.50	0.67	50
Accuracy	-	-	0.96	600
Average	0.98	0.75	0.82	600
Average(comprehensive)	0.96	0.96	0.95	600

3.4. Comparison Analysis

Based on Table 2.3.4.5, it is easy to see that the prediction of Random Forest is better than logistic regression. This may be related to the type of data and the bias of the method itself. The limitation of logistic regression is its weak ability to model nonlinear relationships. In contrast, Random Forest is an integrated learning method based on decision trees. It performs classification or regression tasks by constructing multiple decision trees and voting or averaging them. This article exemplifies the following advantages of Random Forests. High accuracy. Handling of outliers: Random Forests are able to handle outliers efficiently. Parallelizable processing: since each decision tree in Random Forests is constructed independently, this allows Random Forests to have good scalability on large-scale datasets. It can be seen that Random Forest has obvious advantages in handling complex data, improving accuracy, and dealing with outliers.

4. Conclusion

This study has shown that age, gender, presence of hypertension, presence of heart disease, smoking history, BMI, HbA1c level, and blood glucose level are good predictors of diabetes mellitus. The accuracy of both the test set and the training set was above 95% and Random Forest showed excellent results in this study compared to logistic regression. Based on machine learning models, key factors that contribute to an individual's vulnerability to diabetes mellitus are identified. This will help to target better preventive healthcare measures for diabetes mellitus in general or for those at higher risk (e.g., specific age or having abnormal glycated hemoglobin levels). More relevant and more detailed data should be collected for further research. For predicting the progression of diabetes mellitus, extreme values may lead to more important information that was not addressed in this trial.

In the future, research needs to collect more clinical data, including extreme values, to train the model, and the medical community will be able to fully explore the pathogenesis of diabetes.

References

[1] Huang Tai, Yu Qi, Mu Junfang, et al. Bioinformatics analysis of the association between novel coronavirus infection and type 2 diabetes mellitus. Journal of virus, 2024, 40(01): 65-79.

[2] Syed F Z. Type 1 Diabetes Mellitus. Ann Intern Med, 2022, 175(3): 33-48.

[3] Lian Zheng, Yu Hairong, Liu Peilin, et al. Influence factors and clinical prediction model of positive coronary artery remodeling in patients with coronary heart disease complicated with type 2 diabetes mellitus. Journal of Refractory Diseases, 2018, 23(01): 31-36.

[4] Teixeira M M, et al. Association between diabetes and cognitive function at baseline in the Brazilian Longitudinal Study of Adult Health. Sci Rep, 2020.

[5] Sun Binbin, Zhang Yuanyuan, Jia Jianjun. Research progress of controllable risk factors for cognitive dysfunction in the elderly. Chinese Journal of Multi-organ Diseases of the Elderly, 2020, 21(12): 888-891.

[6] Chen Shun, Yuan Qiaoying, Liu Fang, et al. Analysis and model prediction of type 1 diabetes burden in Chinese children and adolescents from 1990 to 2019. Chinese Journal of Child Health Care, 2024, 1-7.

[7] Ding Xuehua, Zhu Yuhua. Detection of serum Ficolin-3, Omentin-1 and FGF19 levels in patients with gestational diabetes mellitus and their predictive value in the occurrence of GDM. Laboratory Medicine and Clinic, 2024, 21(02): 242-246.

[8] Xiang Pingping, Wei Xiao, Liu Chao. Novel coronavirus pneumonia and diabetes: Where do we go from here. International journal of endocrine metabolism, 2022, 42(1): 4.

[9] Jiang Lan, Liu Leming, Liu Meiling, et al. Diabetic nephropathy merger new coronavirus infection clinical characteristics analysis. Journal of Nanchang university, 2023, 63(06): 76-79.

[10] Liu Mingli, Bian Hongyan. Construction of risk prediction model for cognitive dysfunction in type 2 diabetes patients. Health vocational education, 2024, 42(03): 153-157.