Diabet Prediction with Logistic Regression

Zijia Li

Abstract:

Diabetes is a common chronic disease that seriously challenges countries worldwide. However, there are problems such as low detection rate, low awareness rate, low acceptance rate, and low treatment compliance rate in the diagnosis and treatment process. This paper aims to construct a Logistic Regression Model to predict whether a patient has diabetes or not and to investigate the key factors that can be used to diagnose whether a patient has diabetes or not to promote the development of diabetes treatment. The research results show that the patient's characteristics, such as the number of times prior, body mass index, and diagnostic measurements such as plasma glucose concentration 2 hours in an oral glucose tolerance test, trips skin fold thickness, 2-hour serum insulin, and diabetes pedigree function will have a significant impact on whether the patient has diabetes. Therefore, the above six variables should be focused on when diagnosing diabetes mellitus.

Key Words: Diabet Prediction; Logistic Regression Model

1. Introduction

Diabetes is a multiple chronic disease that causes many complications, such as vision loss, stroke, and heart attack, making patients suffer physically and mentally. It has been widely a concern in the society. According to the 2017 statistics of the International Diabetes Federation (IDF), there are about 425 million adult patients with diabetes worldwide, the global prevalence of diabetes is about 8.4% for women and 9.1% for men aged 20-79, and it is expected that the number of people with diabetes may reach 629 million by 2045. However, the pathogenesis of diabetes is not clear yet. Moreover, there are problems such as a low detection rate, low awareness rate, low acceptance rate, and low treatment compliance rate in the diagnosis and treatment process. Therefore, it is important to further study the influencing factors of whether patients have diabetes and construct a suitable diabetes risk prediction model for early detection and treatment of diabetes.

In recent years, more scholars have used machine learning algorithms to predict whether a patient has diabetes. Smith et al. (1988) forecast the onset of diabetes mellitus in a high risk population of Pima Indians using an early neural network model, ADAP [1]. Santhanam et al. (2015) used SVM as a classifier to classify whether patients had diabetes or not[2]. Hasan et al. (2020) proposed the weighted ensembling of different ML models to improve the prediction of diabetes [3].

This paper aims to use the Pima Indians Diabetes Database in Kaggle to construct a logistic regression model to predict whether a patient has diabetes or not and to investigate the key factors that can be used to diagnose whether a patient has diabetes or not to assist doctors in diagnosis and treatment, and to promote the development of diabetes treatment.

2. Method

This paper aims to use patient personal characteristics and diagnostic measurements to predict whether a patient has diabetes. It is a typical binary decision problem. The logistic regression function can effectively limit the domain of variable values between [0, 1] and is widely used in regression models where the dependent variable is dichotomous. Therefore, this paper selects the logistic regression model to investigate the key diagnostic variables that can be used to diagnose the existence of diabetes. The diagnostic result of having diabetes is denoted by$y = 1$, and the diagnostic result of not having diabetes is denoted by$y = 0$. The logistic regression model is as follows:

$$P_i = F\left(y_i\right) = \frac{\exp\left(\beta_0 + \sum_{j=1}^{n} \beta_j X_{ij}\right)}{1 + \exp\left(\beta_0 + \sum_{j=1}^{n} \beta_j X_{ij}\right)}$$

whereP_i is the probability that the patient has diabetes, $F\left(y_i\right)$ is the probability distribution function, β_0 is the intercept term, $\beta_j$is the regression coefficient of the j independent variable, n is the number of independent variables, andX_{ij} is the value of the $i$patient on the$j$ variable.

By taking the logarithm of both sides of equation (1), a simplified equation can be obtained.

$$y_i = \ln\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \sum_{j=1}^{n} \beta_j X_{ij}$$

3. Data Analysis and Result Discussion

3.1. Data Introduction and Data Preprocessing

In this paper, we used the Pima Indians Diabetes Database[1] from Kaggle. This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes based on certain diagnostic measurements included in the dataset. All patients here are females at least 21 years old of Pima Indian heritage. The datasets consist of several medical predictors (independent) variables and one target (dependent) variable, Outcome. Independent variables include the number of pregnancies the patient has had, their BMI, insulin level, age, etc. The variable description table is shown in Table 1.

Table1 Variable description table

Variables	Meaning
Pregnancies	Number of times pregnant
Glucose	Plasma glucose concentration 2 hours in an oral glucose tolerance test
Blood Pressure	Diastolic blood pressure (mm Hg)
SkinThickness	Triceps skin fold thickness (mm)
Insulin	2-Hour serum insulin (mu U/ml)
BMI	Body mass index (weight in kg/(height in m)^2)
DiabetesPedigreeFunction	Diabetes pedigree function
Age	Age (years)
Outcome	Class variable (0 or 1) 268 of 768 are 1; the others are 0

The descriptive statistics for each variable are shown in Table 2.

Table2 Descriptive statistics of variables

Variables	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Pregnancies	0.0000	1.0000	3.0000	3.8450	6.0000	17.0000
Glucose	0.0000	99.0000	117.0000	120.9000	140.2000	199.0000
Blood Pressure	0.0000	62.0000	72.0000	69.1100	80.0000	122.0000
SkinThickness	0.0000	0.0000	23.0000	20.5400	32.0000	99.0000
Insulin	0.0000	0.0000	30.5000	79.8000	127.2000	846.0000
BMI	0.0000	27.3000	32.0000	31.9900	36.6000	67.1000
DiabetesPedigreeFunction	0.0780	0.2437	0.3725	0.4719	0.6262	2.4200
Age	21.0000	24.0000	29.0000	33.2400	41.0000	81.0000
Outcome	0.0000	0.0000	0.0000	0.3490	1.0000	1.0000

It can be found that the five variables Glucose, Blood Pressure, SkinThickness, Insulin, and BMI have unrealistic data with the value of 0. Therefore, in this paper, the data with a value of 0 in these five variables are regarded as missing data, and the mean value is used to fill in the data.

3.2. EDA

The pie chart of if they had diabetes or not is shown in Figure 1. From Figure 1, 65.1% of the patients in this dataset did not have diabetes, and 34.9% had diabetes.

[1] https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database?datasetId=228

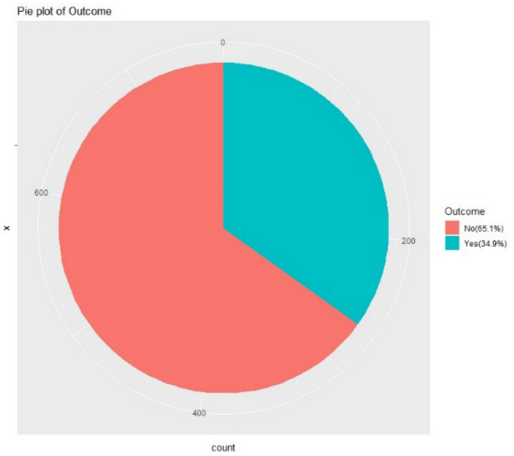

Figure1 Pie Plot of Outcome

The density curves of each independent variable were plotted as shown in Figure 2. From Figure 2, the greater the values of Pregnancies, Glucose, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, and Age, the more likely the patient had diabetes. However, for Blood Pressure, the difference in the value of this variable was smaller whether the person had diabetes or not.

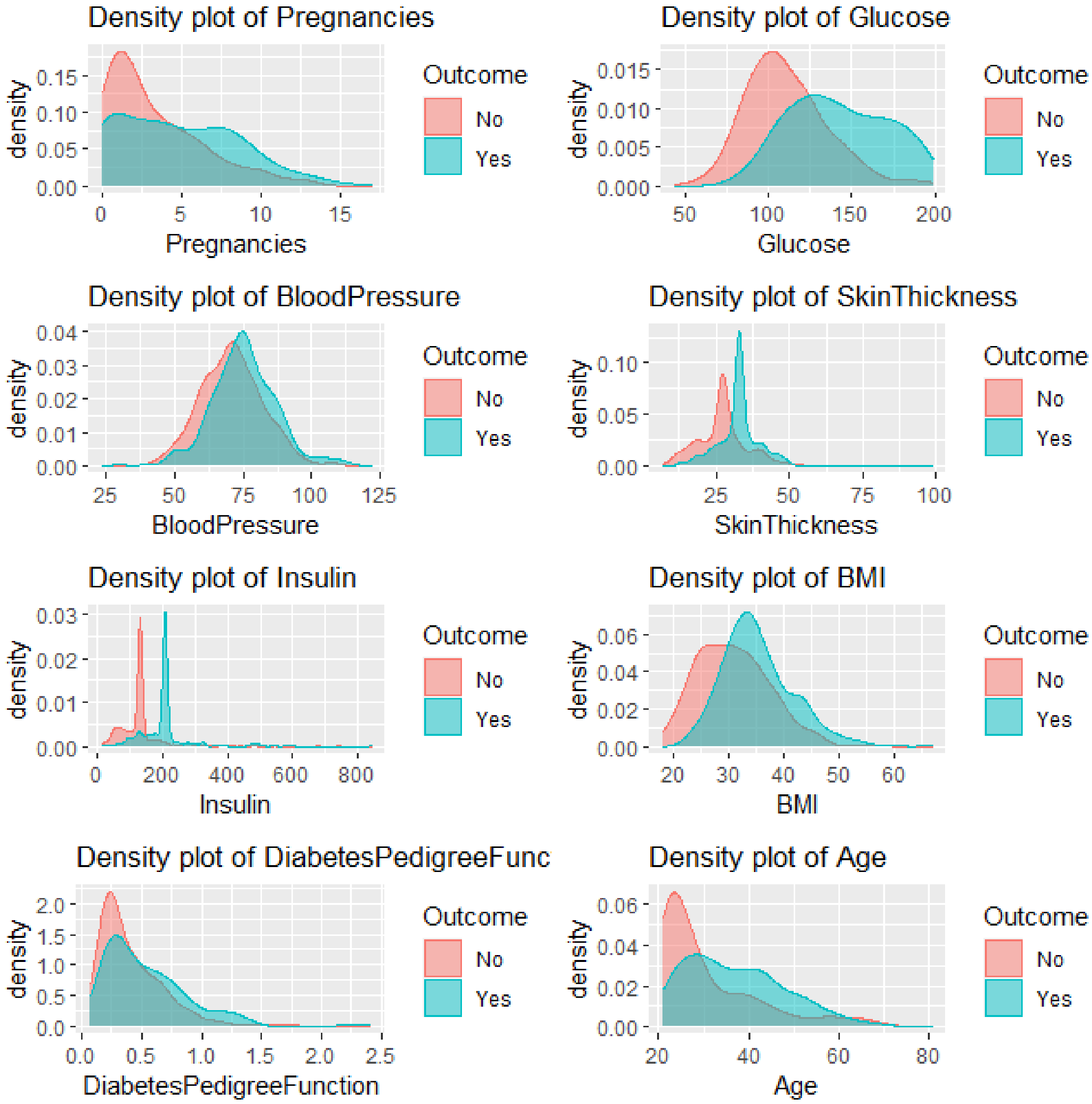

Figure2 Density Plot

3.3. Build a Logistic Regression Model

In this paper, we used Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction and Age as the independent variables and Outcome as the dependent variable to construct a Logistic Regression Model to explore the key diagnostic variables that can be used to diagnose the presence or absence of diabetes. The model regression results are shown in Table 3.

Table3 Results of Logistic Regression Model[2]

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-9.4707	0.8458	-11.197	< 2e-16 ***
Pregnancies	0.1131	0.0322	3.513	0.0004 ***
Glucose	0.0287	0.0039	7.376	1.63e-13 ***
BloodPressure	-0.0047	0.0089	-0.528	0.5976
SkinThickness	0.0445	0.0140	3.167	0.0015 **
Insulin	0.0075	0.0016	4.724	2.31e-06 ***
BMI	0.0530	0.0181	2.930	0.0034 **
DiabetesPedigreeFunction	0.8187	0.3034	2.698	0.0070 **
Age	0.0120	0.0097	1.239	0.2153

As can be seen from Table 3, among the eight independent variables, six variables were significant influencing factors, namely Pregnancies, Glucose, SkinThickness, Insulin, BMI, and DiabetesPedigreeFunction. And all of them had a significant positive effect on Outcome. The higher the values of Pregnancies, Glucose, SkinThickness, Insulin, BMI, and DiabetesPedigreeFunction, the higher the probability of diabetes. It follows that patient personal characteristics such as the number of times pregnant, body mass index, and plasma glucose concentration 2 hours in an oral glucose tolerance test, triceps skin fold thickness, 2-hour serum insulin, and diabetes pedigree function have a significant impact on whether a patient has diabetes. Therefore the focus should be on these six variables when making a diagnosis of diabetes. When the diagnostic measurements of these six variables are large, they should be taken seriously, and the patient has a higher probability of having diabetes mellitus.

Reference

[1] Smith J W, Everhart J E, Dickson W C, et al. Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus[J]//Proceedings of the annual symposium on computer application in medical care. American Medical Informatics Association, 1988: 261-265.

[2] Santhanam T, Padmavathi M S. Application of K-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis[J]. Procedia Computer Science, 2015, 47: 76-83.

[3] Hasan M K, Alam M A, Das D, et al. Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers[J]. IEEE Access, 2020, 8: 76516-76531.

Appendix1 Code

```
# include all the required packages
library(VIM)
library(Hmisc)
library(tidyverse)
library(skimr)
library(cowplot)

# read the data
diabetes <- read.csv(“C:/Users/86156/Desktop/diabetes.csv”)
head(diabetes)
summary(diabetes)

# convert the data
diabetes$Outcome[diabetes$Outcome==0] <- ‘No’
diabetes$Outcome[diabetes$Outcome==1] <- ‘Yes’
diabetes$Outcome <- as.factor(diabetes$Outcome)

# process the missing value
diabetes$Glucose[diabetes$Glucose==0] <- NA
diabetes$BloodPressure[diabetes$BloodPressure==0] <- NA
diabetes$SkinThickness[diabetes$SkinThickness==0] <- NA
diabetes$Insulin[diabetes$Insulin==0] <- NA
diabetes$BMI[diabetes$BMI==0] <- NA
aggr(diabetes, prop=T, numbers=T)
pMiss <- function(x){sum(is.na(x)) / length(x) * 100}
apply(diabetes, 2, pMiss)
```

[1] Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

```
diabetes$Glucose[diabetes$Outcome==’No’] <- impute(diabetes$Glucose[diabetes$Outcome==’No’], mean)
diabetes$Glucose[diabetes$Outcome==’Yes’] <- impute(diabetes$Glucose[diabetes$Outcome==’Yes’], mean)
diabetes$BloodPressure[diabetes$Outcome==’No’] <- impute(diabetes$BloodPressure[diabetes$Outcome==’No’], mean)
diabetes$BloodPressure[diabetes$Outcome==’Yes’] <- impute(diabetes$BloodPressure[diabetes$Outcome==’Yes’], mean)
diabetes$SkinThickness[diabetes$Outcome==’No’] <- impute(diabetes$SkinThickness[diabetes$Outcome==’No’], mean)
diabetes$SkinThickness[diabetes$Outcome==’Yes’] <- impute(diabetes$SkinThickness[diabetes$Outcome==’Yes’], mean)
diabetes$Insulin[diabetes$Outcome==’No’] <- impute(diabetes$Insulin[diabetes$Outcome==’No’], mean)
diabetes$Insulin[diabetes$Outcome==’Yes’] <- impute(diabetes$Insulin[diabetes$Outcome==’Yes’], mean)
diabetes$BMI[diabetes$Outcome==’No’] <- impute(diabetes$BMI[diabetes$Outcome==’No’], mean)
diabetes$BMI[diabetes$Outcome==’Yes’] <- impute(diabetes$BMI[diabetes$Outcome==’Yes’], mean)

# EDA
ggplot(data=diabetes) + geom_bar(aes(x=””, fill=Outcome), width = 1) + coord_polar(theta=”y”) + scale_fill_discrete(labels=c(paste(‘No’, paste(‘(‘, round(sum(diabetes$Outcome==’No’) / nrow(diabetes) * 100, 1), ‘%)’, sep = ‘’), sep = ‘’), paste(‘Yes’, paste(‘(‘, round(sum(diabetes$Outcome==’Yes’) / nrow(diabetes) * 100, 1), ‘%)’, sep = ‘’), sep = ‘’))) + ggtitle(“Pie plot of Outcome”)

p1 <- ggplot(data=diabetes, mapping=aes(x=Pregnancies, fill=Outcome, color=Outcome)) + geom_density(alpha=0.5) + ggtitle(“Density plot of Pregnancies”)
p2 <- ggplot(data=diabetes, mapping=aes(x=Glucose, fill=Outcome, color=Outcome)) + geom_density(alpha=0.5) + ggtitle(“Density plot of Glucose”)
p3 <- ggplot(data=diabetes, mapping=aes(x=BloodPressure, fill=Outcome, color=Outcome)) + geom_density(alpha=0.5) + ggtitle(“Density plot of BloodPressure”)
p4 <- ggplot(data=diabetes, mapping=aes(x=SkinThickness, fill=Outcome, color=Outcome)) + geom_density(alpha=0.5) + ggtitle(“Density plot of SkinThickness”)
p5 <- ggplot(data=diabetes, mapping=aes(x=Insulin, fill=Outcome, color=Outcome)) + geom_density(alpha=0.5) + ggtitle(“Density plot of Insulin”)
p6 <- ggplot(data=diabetes, mapping=aes(x=BMI, fill=Outcome, color=Outcome)) + geom_density(alpha=0.5) + ggtitle(“Density plot of BMI”)
p7 <- ggplot(data=diabetes, mapping=aes(x=DiabetesPedigreeFunction, fill=Outcome, color=Outcome)) + geom_density(alpha=0.5) + ggtitle(“Density plot of DiabetesPedigreeFunction”)
p8 <- ggplot(data=diabetes, mapping=aes(x=Age, fill=Outcome, color=Outcome)) + geom_density(alpha=0.5) + ggtitle(“Density plot of Age”)
cowplot::plot_grid(p1, p2, p3, p4, p5, p6, p7, p8, ncol = 2)

# build the logistic regression model
model <- glm(Outcome~., family=binomial(), data=diabetes)
summary(model)
predict_result <- predict(model, type = “response”)
summary(predict_result)
threshold_0.5 <- table(diabetes$Outcome, predict_result > 0.5)
threshold_0.5
accuracy_0.5 <- round(sum(diag(threshold_0.5))/sum(threshold_0.5), 3)
sprintf(“Accuracy is %s”, accuracy_0.5)
```