Prediction of the Probability of Having Type 2 Diabetes by Using Logistic Regression

Yinzhu Chen *

Department of Mathematics, University College London, London, UK

* Corresponding Author Email: yinzhu.chen.20@ucl.ac.uk

Abstract. Diabetes has been the primary cause of disability and mortality on a global scale. It is crucial to take action to reduce the rising incidence of diabetes. The treatment and control of diabetes are aided by early diagnosis of diabetics and pre-diabetics. A deeper comprehension of the diabetes risk factors is necessary to perform the early screening of diabetics more effectively. In order to determine the type 2 diabetes risk variables and forecast the likelihood that Pima Indian women will develop the disease, logistic regression is utilized in this essay. Pregnancy, glucose, diabetes pedigree function and body mass index(BMI) are four variables in the model that have been identified to be significant predictors. All relevant predictors had positive relationships with the result, according to the analysis. The cross-validated test error rate for the model is 15.74%, and its accuracy is 76.28%. In conclusion, it is suggested that it is conceivable to forecast type 2 diabetes using the data currently available, and the model found suggests that lowering the rate of pregnancies, blood glucose levels, and body mass index may be a way to minimize the prevalence of diabetes.

Keywords: Type 2 Diabetes, Logistic Regression.

1. Introduction

A chronic metabolic condition called type 2 diabetes is characterized by elevated blood sugar levels in the case of insulin intolerance and relative insulin deficiency [1]. Ninety percent of all cases of diabetes are caused by type 2 diabetes, a widespread epidemic disease with a high prevalence. Globally, 450 million persons were predicted to have diabetes in 2017; by 2045, that figure was projected to increase to 693 million [2]. Diabetes is a serious danger to death and disability globally, ranking as the ninth and sixth leading causes of death and disability, respectively [3, 4]. Additionally, diabetics are constantly at risk for developing diabetes complications. Some of the complications can be life-threatening, most notably cardiovascular disease, such as heart failure and stroke [5]. Furthermore, diabetics are at risk for kidney failure, foot ulcers, and issues with the eyes [6]. As a consequence, in response to the increased incidence of type 2 diabetes worldwide, the financial burden is rising, not only on individuals but also on the healthcare system. The budget of the healthcare system on diabetes in 2015 was estimated at 825 billion dollars [7].

It is worth noting that the estimated undiagnosed diabetes rate for adults is close to 50% [8]. While undiagnosed patients are reported to be more likely than diagnosed patients to experience diabetic complications [3]. This means a huge number of diabetics are unaware of their diseases so as to miss out on early treatment, which can result in a more serious condition. While early identification of type 2 diabetes and patient-centered management have been reported to be useful to lessen symptoms and reduce complications [4]. As a consequence, determining the risk factors of diabetes and strengthening the early detection of high-risk populations are crucial. Therefore, targeted treatment can be carried out for those patients with early diagnosis or prediabetics so as to avoid diabetes and lessen complications. For instance, among high-risk individuals who are obese or overweight (BMI > 25), the likelihood of being diabetics can be lowered by 29% to 58% through losing merely 5% to 7% of body weight [1]. Many researchers have investigated the causes of diabetes and reached their own conclusions. One previous study suggested that the ethnicities, obesity, history of gestational diabetes mellitus and cardiovascular disease are main risk factors. People with these risk factors are more likely to be diagnosed with diabetes [9]. In addition, it is reported that the gender and pregnancy are also considered to be causes of diabetes [10].

Many people have conducted the analyses of the triggers for type 2 diabetes. A variety of different algorithms have been employed to improve the predictive accuracy in the investigations of diabetes triggers. For example, Zou *et al.* [11] predicted the incidence of diabetes based on the data set from Luzhou, China and Pima India by different machine learning approaches and compare the performance of the different approaches; Nguyen *et al.* [12] employed deep learning methods to enhance the efficiency of the diabetes prediction analysis. Logistic regression model was used in this essay to analyze the triggers of diabetes and to predict the occurrence of diabetes mellitus. Best subset selection as well as the forward and backward stepwise selection along with selection criteria —C_p, AIC, BIC and adjusted R^2 were employed to select statistically significant predictors in the model. Moreover, the quality of the model was assessed by cross-validated error test. Extend earlier research on diabetes prediction by investigating basic machine learning methods – logistic regression. The prediction model gives four key predictors of diabetes: pregnancy, glucose, BMI and diabetes pedigree function. It aids in the classification of high-risk populations, as well as the prevention, diagnosis, and management of diabetes. The regression model's prediction accuracy with the four predictors mentioned above was 76.28 percent, while the cross-validation error rate was 15.74 percent. This research can assist policymakers in developing health strategies to combat the diabetes pandemic. Because diverse data sets need different methods, future research should be wary about relying on a single model or approach to predict diabetes risk.

2. Methodology

2.1. Statistics description

A brief introduction to the basic information about the dataset used is as follow. In this essay, the data utilized is the dataset of Pima Indian women. It has eight variables which are pregnancy, glucose, skin thickness, blood pressure, BMI, insulin, family history and age with a column of result where 0 represents individuals without diabetes and 1 represents diabetics. After the data pre-process which is replacing the null values in variables with their mean value, the statistics description of the data is displayed by Table 1.

Table 1. Statistics description.

Predictor	Explanation	Count	Mean	Std. deviation
Pregnancy	The number of pregnancies	768	3.84	3.36
Glucose	A two-hour oral glucose tolerance test plasma glucose concentration	768	120.89	31.97
Skin Thickness	Skinfold thickness of triceps	768	20.53	15.95
Blood Pressure	Diastolic blood pressure	768	69.10	19.35
BMI	Body mass index	768	31.9	7.88
Insulin	Two-hour test on serum insulin	768	79.79	115.24
Pedigree	Family history of type 2 diabetes	768	0.47	0.33
Age	Age (years)	768	33.24	11.76

Some previous researches have claimed that a few variables are associated to the incidence of type 2 diabetes. Pregnant women are at risk for gestational diabetes mellitus (GDM) [13], and those who have had GDM in the past have a significantly higher chance of acquiring type 2 diabetes [14]. Therefore, the risk of type 2 diabetes will increase as the number of pregnancies rises. Many previous researches reported that high blood glucose is the characteristic of diabetics and this can be explained

by the pathogenesis of type 2 diabetes. Due to insulin's inability to function effectively (insulin resistance and insulin secretion defect), insulin's ability on glucose decomposition is affected, which leads to the rise of blood sugar levels [15]. A higher BMI indicates obesity which is another significant cause of type 2 diabetes since more evidence show that the rise in obesity is highly associated with the growth in the incidence of type 2 diabetes [16]. Defronzo [15] explained that obesity can lead to insulin resistance and the rise of blood glucose level. Another important risk factor of type 2 diabetes that reflects the genetic possibility of having diabetes is genetic predisposition of diabetes [17].

2.2. Logistic regression

Logistic regression is a common and useful classifier to forecast a qualitative response in the statistics. Logistic regression models the probability that the response Y falls into a specific category rather than explicitly modelling the response itself. The probability is called success probability, denoted by P(X) where X are independent variables. For instance, in this essay the probability specifically means the probability of an individual having type 2 diabetes. In order to make sure the probability will strictly range between 0 and 1, the logistic function is used in the logistic regression. The *logistic function* is given by,

$$P(X) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k}} \quad (1)$$

Where $X = (X_1, X_2, \ldots X_k)$ are k predictors.

Clearly, the logistic function will always give a sensible probability irrespective of$X$'s value. By transforming the logistic function,

$$\log\left(\frac{P(X)}{1+P(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \quad (2)$$

There is a quantity $\log\left(\frac{P(X)}{1+P(X)}\right)$ which is linear in X in the logistic regression model. The *odds* is denoted by the expression$\left(\frac{P(X)}{1+P(X)}\right)$, while the *log odds* or *logit* is denoted by$\log\left(\frac{P(X)}{1+P(X)}\right)$.

2.3. Maximum likelihood

To estimate the coefficients β in the logistic regression model, maximum likelihood is the preferred approach as a result of its better statistical properties. Looking at the logistic function, the principle behind the method of maximum likelihood is to find the estimates for coefficients such that the probability P(X) is close to what an individual is observed in the default status which means the model for P(X) will give a number close to 1 when all individuals have type 2 diabetes and close to 0 when no one has type 2 diabetes. Expressing the principle in the mathematical term, an equation called likelihood function is formed:

$$L(\beta_0, \beta_1) = \prod_{i:\, y_i=1} p(x_i) \prod_{i':\, y_{i'}=0} (1 - p(x_{i'})) \quad (3)$$

The value of estimates β which maximizes the likelihood function L is denoted by $\hat{\beta}$ and it is called the maximum likelihood estimator (MLE) ofβ.

2.4. Stepwise selection and best subset selection

The (forward) stepwise selection and best subset selection are each illustrated in Tables 3 and Table 2 [13] below. Only the forward stepwise selection is shown in Table 3 since the backward stepwise selection is pretty identical to the forward one only with the opposite orientation. The process of each selection can be divided into three steps.

Table 2. Best subset selection.

Definition
1. Denote N_0 as null model with no variables. For each observation, N_0 represents the predicted mean of the sample. 2. For i = 1,2, ... p: (i) Fit all $\binom{p}{i}$ models with precisely i variables. (ii) In the $\binom{p}{i}$ models, select the best one and label it N_i. The best one here is defined as the model with the biggest R^2 or smallest RSS. 3. From $N_0, \ldots, N_p$ based on the test error by cross-validation, C_p, adjusted R^2 or BIC, choose an optimal model.

Table 3. Forward stepwise selection.

Definition
1. Denote N_0 as null model with no variables. 2. For i = 0,1,2, ... p-1: (i) Fit all p – i models adding one variable to the predictions in N_i. (ii) In the p – i models, select the best one and label it N_{i+1}. (same as the definition above) 3. From $N_0, \ldots, N_p$ based on the test error by cross-validation, C_p, adjusted R^2or BIC, choose an optimal model.

It seems two methods are somewhat similar. However, in step 2, the best subset method will select the best model from all possible variable combinations, while stepwise selection only chooses variables based on the previous step which means the variables used in the previous step will always contained. The calculation of best subset selection method is larger than that of stepwise selection method, but the selected variables are more accurate. In general, however, the two methods will give relatively similar results.

2.5. C_p, BIC and adjusted R^2

C_p, adjusted R^2 and BIC are used to choose the optimal model. They all estimate test error through adjusting to training error while using different ways of adjustment. The best model is selected by the lowest C_p and BIC statistics and the largest adjusted R^2 statistic.

C_p Estimates the test error using the equation:

$$C_p = \frac{1}{n}\ (RSS + 2d\hat{\sigma}^2) \tag{4}$$

Where RSS is residual sum of squares and $\hat{\sigma}^2$ is the variance of error with each response.

BIC is calculated as,

$$BIC = \frac{1}{n}\ (RSS + \log(n) d\hat{\sigma}^2) \tag{5}$$

With total sum of squares (TSS) $=\sum(y_i - \bar{y})^2$, the adjusted R^2 is given by,

$$Adjusted\ \ R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)} \tag{6}$$

2.6. Validation and cross-validation

Validation approach works by extracting a portion of the training set as a test set and using the test set to estimate test errors. The cross-validation method is a method to improve the selection of test sets of validation method. The data is divided into k pieces of similar size, the random one of them is regarded to be the test set, with the remaining k-1 pieces in training set are fitted to obtain the mean Square Error (MSE1), and then repeated k times. The K-fold CV estimate is obtained,

$$CV_k = \frac{1}{k}\sum_{i=1}^{k} MSE_i \quad (7)$$

Cross-validation is a very useful tool which can be used both to evaluate the quality of the model by estimating the value of test error and to figure out the quantity of variables [18].

3. Results and Discussion

3.1. Analysis of the correlation matrix

The correlation matrix which measures the strength of linear relationship between two variables is displayed in Figure 1. It shows all the correlation coefficients for nine variables in pairs. Among the single predictor of outcome, glucose is most correlated with outcome with correlation coefficient 0.5. For correlations between predictors, there are two relatively large correlation coefficients, which are between age and pregnancies (0.57), and between skin thickness and BMI (0.55) respectively. The large coefficients between two predictors indicates that they are highly correlated with each other.

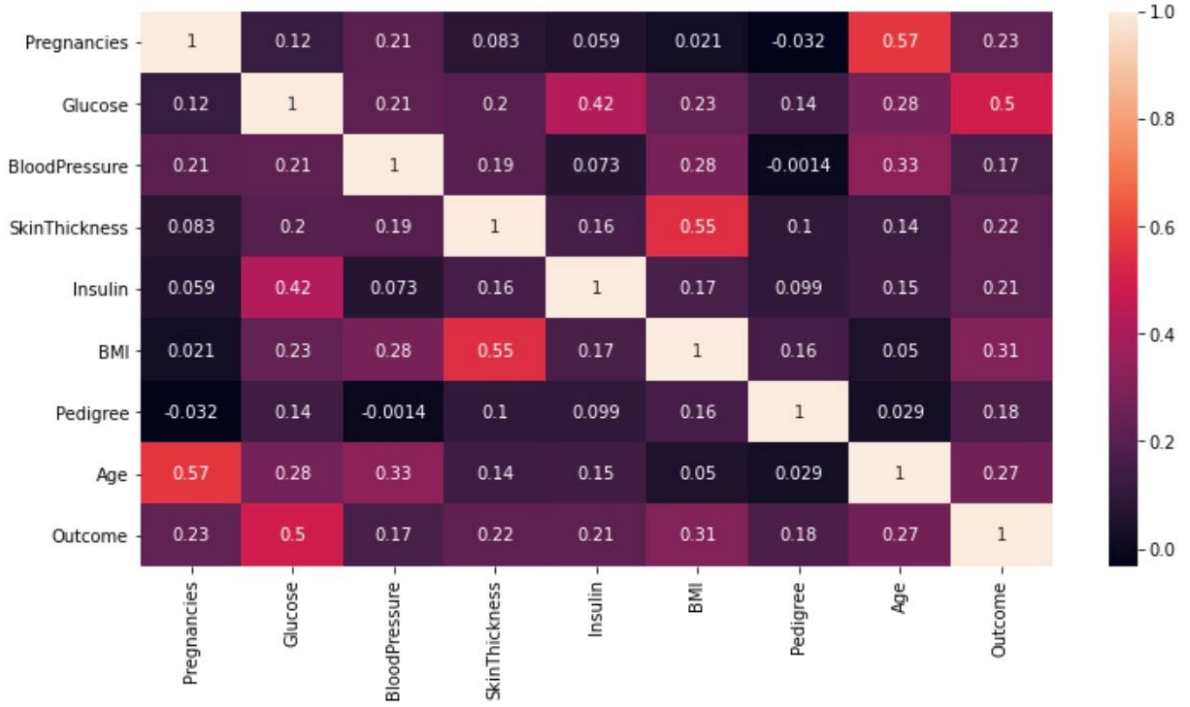


Figure 1. Correlation matrix.

3.2. Estimates of logistic regression model

The logistic regression model is applied to predict whether an individual has type 2 diabetes or not. It is of vital importance to choose a model containing all the relevant variables and excluding irrelevant ones, since the unnecessary variables not only add complexity to model also reduce the interpretability of the model. In this essay, the variable selection methods including both best subset selection and stepwise selection are used to find the appropriate variables. First, looking at the best subset selection.

Table 4 shows the best model for every given number of variables which is generated by the best subset method. The set of models are selected based on the R^2 value. However, the R^2 value is calculated only by the training data, so the maximum R^2 value can only guarantee a lower training error which does not necessarily ensure a lower test error. Once the variables in the model increase, the value of R^2 will inevitably increases accordingly. As shown in Table 4, the R^2 value is only 24.28% with only one variable included and increase monotonically to 32.12% when eight predictors are added in the model. As a result, it will eventually be the model containing all variables which is

meaningless for choosing the optimal model. Therefore, it is reasonable to choose the optimal model based on the other criteria by which the test error can be obtained or estimated. This means that after getting the set of the best models for each subset size, it is appropriate to useC_p, BIC, adjusted R^2 or Cross-validated error criterion to further select the optimal one from the set. In other words, the exact number of variables in the best model is required to determine.

Table 4. Best model for each subset size.

Size	R^2	**Features**
1	0.24283	Glucose
2	0.28432	Glucose BMI
3	0.31048	Pregnancies Glucose BMI
4	0.31868	Pregnancies Glucose BMI Pedigree
5	0.31993	Pregnancies Glucose BMI Pedigree Age
6	0.32059	Pregnancies Glucose BloodPressure BMI Pedigree Age
7	0.32123	Pregnancies Glucose BloodPressure BMI Insulin Pedigree Age
8	0.32124	Pregnancies Glucose BloodPressure SkinThickness BMI Insulin Pedigree Age

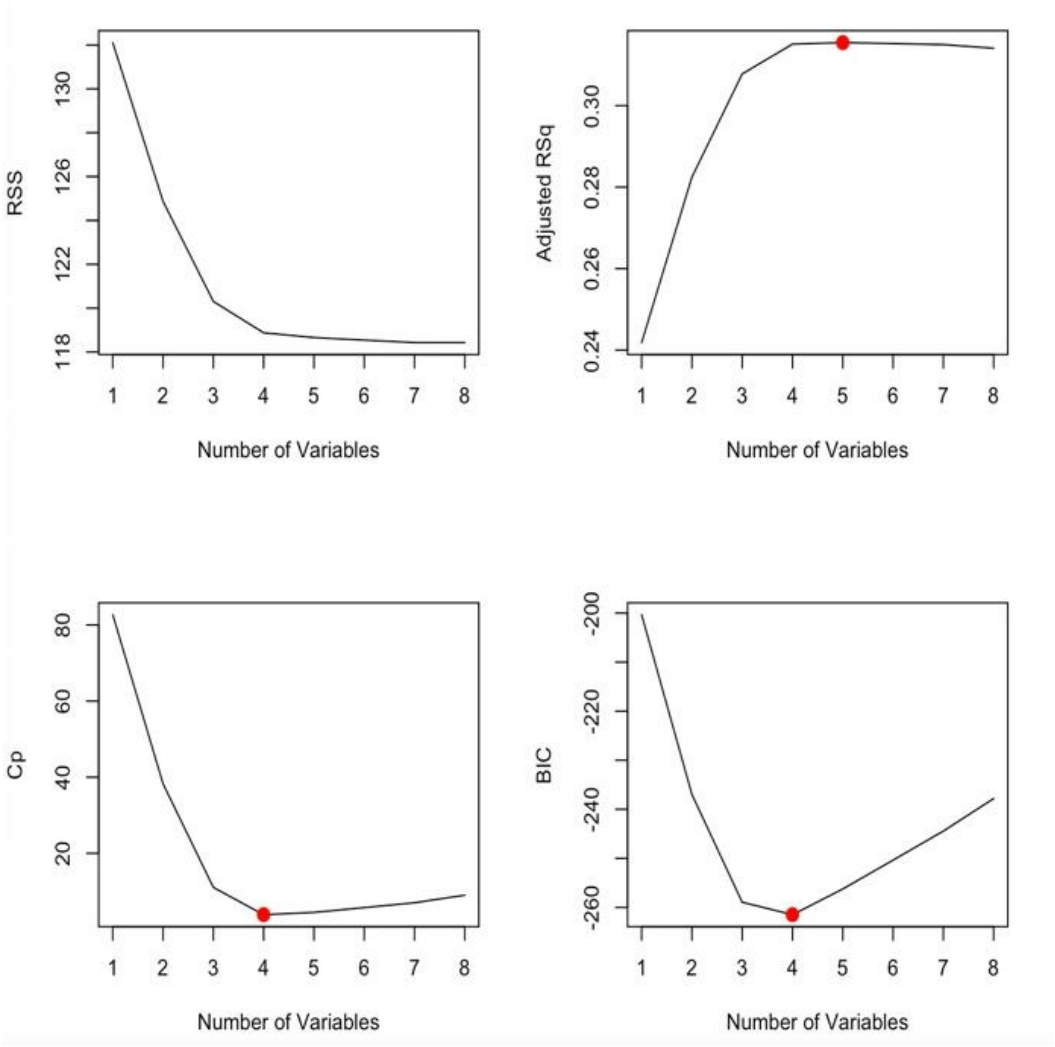


Figure 2. The number of predictors by different criteria.

Figure 2 shows the performance of models with different sizes under the selection criteria. Residual sum of square (RSS) displays a monotonic relationship with the number of variables which is similar to R^2 statistic, so it is not used as a reference. The largest adjusted R^2 statistic and the

smallest C_p and BIC statistic are marked in red in Figure 2. It helps identify the best model. Both the lowest C_p statistic and BIC statistic indicate that the model with 4 predictors will perform best. While the largest adjusted R^2 statistic suggests the five-predictor model has the lowest test error. However, on the adjusted R^2 statistic graph, the flat curve starting from the four predictors indicates that four-predictor and five-predictor model show little difference in the performance of prediction accuracy. Usually, among models with equivalent performance, the model with fewer predictors will be preferred. Therefore, combining all three criteria, the model with four predictors is most appropriate.

The selected predictors for each model size according to different statistics are displayed in the order of the size of statistics in Figure 3. The smallest C_p (3.9) and BIC (-260) select four-variable model in which the variables are pregnancies, glucose, pedigree and BMI. While the largest adjusted R^2 suggest five-variable model which contains pregnancies, glucose, pedigree, BMI and age.

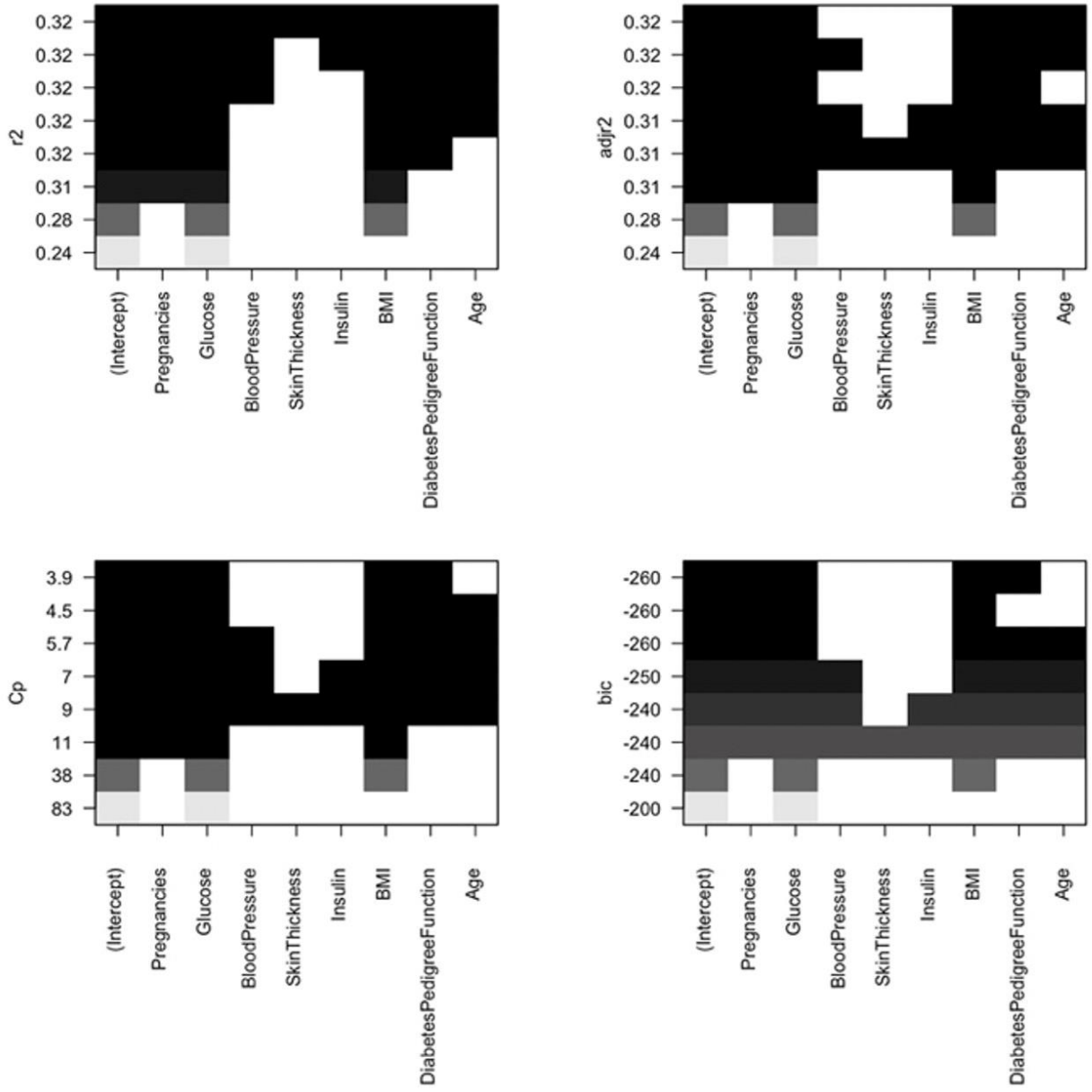


Figure 3. Predictors by different criteria.

Additionally, stepwise selection is also explored for model selection. Since the forward selection and backward selection are almost same except the direction of the process, only the backward selection is displayed in this essay. Six models containing different variables are fitted by the logistic

regression model and their obtained estimates of coefficients and AIC statistic are displayed in the Table 5. Model 1 is a null model only including an intercept which is a start point for the forward selection, while Model 2 is a full model with all variables in the model and it is where the backward selection start. Model 3, 4, 5 and 6 show the process of a backward stepwise selection where the predictor having the highest p-value is deleted in each step. Fit the new model obtained from the last step and stop the process when all the remaining variables have p-values below 5%, then the model with all the variables in it being statistically significant predictors will be obtained. In Table 5, the resulting model is Model 6. Model 6 tells that pregnancies, glucose, pedigree and BMI are four significant predictors. According to the AIC statistic, the four-variable model is the best model. This result is consistent with the conclusion drawn from the best subset selection. Using the *regsubsets( )* function in the R for stepwise selection, the exact same four-variable model is obtained.

Table 5. Estimates for coefficients.

	Model1	**Model2**	**Model3**	**Model4**	**Model5**	**Model6**
Constant	-0.348***	-9.108***	-9.093 ***	-9.052 ***	-9.405 ***	-9.183 ***
Pregnancy		0.124***	0.125 ***	0.125 ***	0.123 ***	0.143 ***
Glucose		0.037***	0.037 ***	0.036 ***	0.036 ***	0.036 ***
BloodP		- 0.009	- 0.009	- 0.008		
SkinTh		0.003				
Insulin		- 0.001	- 0.001			
BMI		0.094***	0.096***	0.094***	0.089***	0.088***
Pedigree		0.875 **	0.876**	0.861 **	0.871 **	0.881 **
Age		0.013	0.013	0.013	0.010	
AIC	1045.3	730.84	728.91	727.96	726.97	726.37

Signif. Codes: 0 < p: '***'< 0.001 <p: '**'< 0.01 <p: '*'<0.05; BloodP represents Blood Pressure, SkinTh represents Skin Thickness, Pedigree represents Diabetes Pedigree Function

3.3. Predictive equation

After selecting the best model, the logistic regression model is fitted to get the estimates of the coefficients. Substituting the estimates into the logistic function, the predictive equation for the type 2 diabetes is obtained. The equation is:

$$Logit = -9.18 + 0.14 * Pregnancies + 0.03 * Glucose + 0.08 * BMI + 0.88 * Pedigree \quad (8)$$

3.4. Prediction accuracy

Cross-validation is a practical tool to assess the model since it can estimate the test error rate. By the 10-fold cross-validation, the test error rate of Model 6 is 15.74%. Logistic regression is the simplest model in machine learning. It can be applied to investigate the illness triggers and forecast the likelihood of the illness incidence based on the triggers. Diabetes has been the major cause of disability and mortality, while many people who are at high risk for being diabetics are unaware of its risk factors. Employing the logistic regression to explore the risk factors and to construct the predictive equation of the probability of being diabetics might help early detection of type 2 diabetes. As a result, measures can be taken to prevent or manage diabetes for the high-risk populations. The dataset explored in this essay is focusing the Pima Indian women. Therefore, the variables in the prediction model contains pregnancy. The probability formula obtained by the logistic regression shows that pregnancy, glucose, BMI and pedigree have the positive relationship with the probability.

The accuracy of the prediction model in this paper reached 76.28%, which is a simple and usable model compared with many other models for Pima Indian database. This demonstrates that the logistic regression models can be effective in disease prediction as well. However, this model also has some shortcomings. For example, the data used in this model focused only on diabetes in women,

and there were only eight predictors to explore. While there are many other causes of diabetes, such as sedentary lifestyle, unhealthy eating habits (excessive intake of sugar), cigarettes smoking, and lack of exercise and so on, which should be included in the future study, so as to better study the risk factors of diabetes, so that the disease can be better prevented and controlled.

4. Conclusion

In this essay, the risk factors for type 2 diabetes are found using logistic regression, which is also used to develop a formula for predicting the likelihood that a person would get diabetes. A rudimentary model is constructed to help understand the main predictors of the disease. The model with cross-validated test error rate (15.74%) and accuracy (76.28%) contains four significant predictors which are pregnancy, glucose, BMI and pedigree respectively and all four predictors are in positive with the outcome. This indicates that in order to reduce the likelihood of acquiring diabetes, the values of the four variables should be lower. As a result, the implication from the prediction model is that interventions to prevent and control diabetes can be structured around four indicators.

References

[1] DeFronzo, R.A., Ferrannini, E., Groop, L., Henry, R.R., Herman, W.H., Holst, J.J., Hu, F.B., Kahn, C.R., Raz, I., Shulman, G.I. and Simonson, D.C. 'Type 2 diabetes mellitus', Nature reviews Disease primers, 2015, 1(1): 1-22.

[2] Cho, N. H., Shaw, J. E., Karuranga, S., Huang, Y., da Rocha Fernandes, J. D., Ohlrogge, A. W., & Malanda, B. IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. Diabetes research and clinical practice, 2018, 138: 271-281.

[3] Zheng, Y., Ley, S. H. and Hu, F. B. 'Global aetiology and epidemiology of type 2 diabetes mellitus and its complications', Nature Reviews Endocrinology, 2018, 14(2): 88.

[4] Chatterjee, S., Khunti, K. and Davies, M. J. 'Type 2 diabetes', The Lancet, 2017, 389(10085): 2239-2251.

[5] Tigga, N.P. and Garg, S. Prediction of type 2 diabetes using machine learning classification methods. Procedia Computer Science, 2020, 167: 706-716.

[6] Schlienger, J.L. Type 2 diabetes complications. Presse medicale (Paris, France: 1983), 2013, 42(5): 839-848.

[7] Seuring, T., Archangelidi, O. and Suhrcke, M. The economic costs of type 2 diabetes: a global systematic review. Pharmacoeconomics, 2015, 33(8): 811-831.

[8] Beagley, J., Guariguata, L., Weil, C. & Motala, A. A. Global estimates of undiagnosed diabetes in adults. Diabetes Res. Clin. Pract, 2014, 103: 150–160.

[9] Vijan, S. Type 2 diabetes. Annals of internal medicine, 2010, 152(5): ITC3-1.

[10] Engelgau, M.M.; Narayan, K.; Herman, W.H. Screening for type 2 diabetes. Diabetes Care 2000, 23, 1563 1580.

[11] Zou, Q.; Qu, K.; Luo, Y.; Yin, D.; Ju, Y.; Tang, H. Predicting diabetes mellitus with machine learning techniques. Front. Genet. 2018, 9: 515.

[12] Nguyen, B.P.; Pham, H.N.; Tran, H.; Nghiem, N.; Nguyen, Q.H.; Do, T.T.; Tran, C.T.; Simpson, C.R. Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records. Comput. Methods Programs Biomed. 2019, 182: 105-155.

[13] Bottalico, J.N., 2007, June. Recurrent gestational diabetes: risk factors, diagnosis, management, and implications. In Seminars in Perinatology, 2007, Vol. 31, No. 3: 176-184. WB Saunders.

[14] Bellamy, L., Casas, J.P., Hingorani, A.D. and Williams, D. Type 2 diabetes mellitus after gestational diabetes: a systematic review and meta-analysis. The Lancet, 2009, 373(9677): 1773-1779.

[15] DeFronzo, R.A. Pathogenesis of type 2 diabetes mellitus. Medical clinics, 2004, 88(4): 787-835.

[16] Maggio, C.A. and Pi-Sunyer, F.X. Obesity and type 2 diabetes. Endocrinology and Metabolism Clinics, 2003, 32(4): 805-822.

[17] Lindstrom, J., & Tuomilehto, J. The diabetes risk score: a practical tool to predict type 2 diabetes risk. Diabetes care, 2003, 26(3): 725-731.

[18] James, G., Witten, D., Hastie, T., & Tibshirani, R. An introduction to statistical learning, 2013, Vol. 112: 18. New York: springer.