Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

Prediction of Type 2 Diabetes through Risk Factors using Binary Logistic Regression

Imad Yagoub Hamid

Shaqra University, Faculty of Science and Humanities Studies, Department of Mathematics, Dawadmi-Saudi Arabia, E-mail: emad@su.edu.sa

ARTICLE INFO

Article history:
Received: 01/10/2020
Rrevised form: //
Accepted : 03 /11/2020
Available online: 16/11/2020

Keywords:

Type-2 Diabetes, Risk Factors, Logistic Regression, Prediction.

ABSTRACT

The main objective of this study is to arrive at a highly efficient prediction model for early prediction of diabetes by relying on risk factors for diabetes as predictor variables.
Using a binary logistic regression model, a model was built for the data of study which taken from a sample of diabetics and non-diabetics persons.
The results have shown the high ability of the binary logistic regression model in predicting the diabetes-infected persons. All indicators confirm the validity and quality of the model. The results of Chi-square test (sig=0.945>.05) indicated that the model is significant. Likewise, the results of Hosmer&Lemeshow test (sig=0.945>.05) confirm that the model represents the data very well. Classification table findings were also high, as the overall percentage for the correct classification was 91%.
The significant risk factors that influential in predicting diabetes can be arranged as follows: (High blood pressure, Diabetes in the family first degree, High cholesterol, Smoking, Age above 35, Overweight and Gender).
All these results confirm the quality and accuracy of this model in predicting the disease, the thing which may indicate that the model can be used as a primary tool for predicting type-2 diabetes through risk factors.

MSC : 30C45 , 30C50

DOI : https://doi.org/10.29304/jqcm.2020.12.3.709

1. Introduction

Diabetes is considered as one of the chronic diseases that have rapidly proliferated with high ratio in recent years, and become one of the most challenging health issues of 21st century. In 2019, it is estimated that 463 million people have diabetes, and the number is expected to double in the next twenty years (Cataloguing, 2016; Atlas, 2019). Diabetes is one of the main causes of death in the working age population, the number of deaths annually caused by diabetes is around 4.2 million. Furthermore, diabetes is the fourth leading cause of death in most

∗Corresponding author :***Imad Yagoub Hamid***

Email addresses: *emad@su.edu.sa*

Communicated by : ***Alaa Hussein Hamadi***

developed countries (Cataloguing, 2016; Atlas, 2019). Among the factors that increase the risk of this disease is that a third of infected people are not diagnosed early, because the symptoms may not be so clear for a long time. Many people may discover the disease when doing routine medical checkup, and it may often be discovered after the emergence of one its complications, such as cardiovascular disease, kidney failure, nerve damage, blindness, strokes and lower limb amputations (Cataloguing, 2016). However, whenever diabetes is diagnosed early enough, many of its complications can be avoided.

In order to predict this disease early and know the effect of risk factors on its prediction, the logistic regression model was drawn on to build up a reliable model for early disease prediction. Medical diagnosis is one of the fields in which logistic regression models have been successfully applied, as the results obtained from these studies showed high accuracy compared to other statistical prediction methods. Due to their flexibility in dealing with most types of data, logistic regression models are frequently used in the medical field, as the nature of the data in this sector ranges from qualitative and quantitative.

Logistic regression models are a special case of general regression models, and prediction is the main objective of most regression models. The logistic model is used in studies that aim to build predictive models when the dependent variable in study is a qualitative one (Sperandei, 2014). In medical studies, it is used to predict a specific disease through symptoms and risk factors, define and arrange the impact of risk factors for disease, classify a group of people as infected or uninfected, or to categorize a number of diseases based on a number of influencing factors. Logistic regression models may play an active role in knowing the effect of risk factors on predicting diabetes and determining the degree of influence of these factors, and this may help in early diagnosis of this disease and reduce the risk of its complications.

Most cases of type-2 diabetes are late for several years before being diagnosed, this may be because the symptoms are usually mild at first. This delay in diagnosis may result in serious complications, the negative effects of which extend to the kidneys, eyes, heart and other vital body systems. In this study, we try to initiate a statistical model for early prediction of type-2 diabetes via risk factors, by using binary logistic regression, to help early diagnosis of the disease. Therefore, the research is intended to give answers to these questions: To what extent can type-2 diabetes be predicted by relying only on risk factors?, What is the efficiency and accuracy of binary logistic regression model in building a prediction model for diabetes?, What

are the most influential risk factors in predicting type-2 diabetes? And What is the degree of influence of any risk factor in predicting diabetes?.

The most important objectives of the study are: Building initiating a statistical model that can predict early incidence of diabetes through risk factors, knowing the efficiency of the binary logistic regression model in predicting the probability of diabetes, determining the most important risk factors that influence the prediction of type-2 diabetes.

To achieve the objectives of this study, an analytical statistical approach will be followed. A model of prediction is constructed by using binary logistic regression for the data investigated in this study, which are samples of people with and without diabetes. The variable to be predicted in the study is the infection of diabetes and it is a qualitative variable that has two categories (1: infected, 0: uninfected). As for the predictor variables, they were the risk factors for diabetes. The quality and validity of the model are confirmed by relying on some statistical tests such as Chi square test, Hosmer-Lemeshow test, and other criteria.

Data for this study were collected using a questionnaire for a sample of 189 case, (97 male, 92 female). (140 diabetes, 49 non-diabetes). The questionnaire included a number of questions concerning risk factors and some personal information. Data of the diabetic infected people were collected from patients attending medical follow-up in the primary health care centers in the city of Dawadmi, Saudi Arabia.

The model included 10 independent variables that represent the most common risk factors for diabetes, as follows: Gender, Age above 35, High blood Pressure, High cholesterol, Diabetes in family 1st degree, Physical activity-sport, Physical activity-work, Food habits (vegetables, fruits), Smoking, overweight. And the dependent variable is diabetes prediction (diabetes, non-diabetes).

2. Literature Review

Many studies have been done through various statistical methods to find a proper model for predicting diabetes. Some of these studies were summarized as follows:

(Rahman et al., 2013), Using neural network and Logistic regression to build models to for classifying diabetes and non-diabetes subjects, they investigated demographic, enthropometric and clinical data. They found that logistic regression correctly classified 70.4% of all cases. The neural network sensitivities were 84.33%. (Maulana, Badriyah and Syarif, 2018), also examined

the Influence of logistic regression models for prediction and analysis of diabetes risk factors. The data used in this study were taken from the dataset of Soewandhie Hospital. The accuracy level obtained is 94.77%. (Rastogi and Singh, 2019), Logistic regression was used to assess the effect of risk factors on the prevalence of diabetes in urban and rural areas of India. They found that age, BMI and intake of fast food were associated with diabetes in rural. While age and BMI were associated with diabetes in urban. (Rahimloo and Jafarian, 2016), Combining neural network and logistic regression to build a model used to predict diabetes. The data set was taken from the Association of diabetic's city of Urmia. The error of the combined model is equal to 0.0002. (Senthilvel, Radhakrishnan and Sathiyamoorthi, 2011), the logistic model used to examine the influence of various factors in the prediction of diabetic retinopathy among diabetics. The result was that the probability to develop the diabetic retinopathy in a patient was 0.98. (Islam and Rahman, 2012), used Chi-square test and logistic regression to find out the risk factors of Type 2 diabetes in Bangladesh. They used clinical data collected from the diabetic patients of Rajshahi Diabetes Association, Bangladesh. (Saied and Abdallah, 2019), used Logistic regression to find out factors affecting diabetes in the red sea state. The data were collected from a diabetes treatment center in Port Sudan city, Sudan. He found that the factors affecting diabetes are age, gender, and kinship. (Niyikora, 2015), used Logistic regression modeling to evaluate the effect of risk factors on diabetes. The data were collected from Gitwe Hospital. The results showed that age, alcohol consumption, cholesterol level, occupation status and hypertension were associated with the outcome of having diabetes. (Zehra et al., 2018), also applied logistic regression on the data of National Health and Morbidity Survey Malaysia 2011, to predict diabetes in Malaysians. The results showed that the prediction accuracy was 93%.

3. Basic Concepts

3.1. Diabetes:

Diabetes Mellitus (DM) is a disease that occurs because of insulin production, either when the pancreas does not produce enough insulin or when the body cells can not properly use the insulin it produces (Cataloguing, 2016).

There are various types of diabetes, but the following are the most common:

1. Type-1 Diabetes: Is a form of diabetes that occurs when the pancreas unable to produce insulin. Type-1 diabetes can develop at any age but usually appears before the age of 40, and

the majority occurs in children and adolescents. 10% of all diabetes cases are nearly found in type1. The exact causes of type-1 diabetes are unknown (Cataloguing, 2016; Atlas, 2019).

2. Type-2 Diabetes: Is a form of diabetes that occurs when the pancreas does not produce enough insulin, or when the body cannot properly use the insulin it produces. It usually appears in people over the age of 40, but it can develop at any age. 90% of all worldwide diabetes cases are classified as type 2. Symptoms of type-2 are often less marked or absent. As a result, the disease may go undiagnosed for several years, until complications emerge (Cataloguing, 2016; Risk and Collaboration, 2016).

   The most common risk factors of type 2 diabetes in addition to genetic factors are: overweight and obesity, unhealthy diet, physical inactivity, family history of diabetes, age, high blood pressure, increased body fat, alcohol or smoking addiction.

3. Gestational diabetes (GD): is a temporary condition that affects females during pregnancy. It occurs because the body cannot produce enough insulin to meet the extra needs of pregnancy (Cataloguing, 2016).

3.2. Logistic Regression (LR)

Logistic regression models are a special case of general regression models. It is used to analyzes the relationship between multiple independent variables (continuous or categorical) and a categorical dependent variable, and to predict the probability of an event occurring based on the values of explained variables that can be related to that event. There are three types of logistic regression models, which are classified according to the type of the dependent variable: Binary logistic regression is used when the dependent variable is dichotomous. Multinomial logistic regression is used when the dependent variable is has more than two categories. Ordinal logistic regression and is used when the dependent variable is ordinal. (Kleinbaum and Klein, 2002; Hosmer and Lemeshow, 2005).

The Logistic distribution is a continuous probability density function that is symmetric and uni-modal, and receives its name from its cumulative distribution function, which is an instance of the family of logistic functions:

$$f(x;\alpha,\beta)=\frac{1}{1+e^{(x-\alpha)/\beta}}=\frac{1}{2}+\frac{1}{2}\tanh(\frac{x-\alpha}{2\beta}) \qquad (1)$$

The probability density function of the logistic distribution is given by:

$$f(x;\alpha,\beta)=\frac{e^{(x-\alpha)/\beta}}{(1+e^{(x-\alpha)/\beta})^2}=\frac{1}{4\beta}\operatorname{sech}^2(\frac{x-\alpha}{2\beta}) \qquad (2)$$

For a binary dependent random variable Y, and k number of independent random variables $(X_1, X_2, ...X_k)$, the logistic regression equation can be written as follows (Hosmer and Lemeshow, 2005):

$$p_i = E(Y_i/X_i) = \frac{e^{(\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_kX_k)}}{1+e^{(\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_kX_k)}} \quad ; \quad 0 < P < 1 \qquad (3)$$

This equation can be converted to linear formula by using a logit transformation as follows:

$$\log it = \log(\frac{P}{1-P}) = \beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_kX_k \qquad (4)$$

Where:

- $P/(1-P)$: Odds Ratio, it indicates the probability that something will happen (Y=1) or will not happen (Y=0). It takes $(0 < odds < \infty)$. Where P: represents the probability of occurring event , and $(1-P)$: the probability of event not occurring.
- $Logit$: logarithm of the odds (log odds), refers to regression models that includes the logit as a dependent variable in the equation.
- β_j: Regression coefficients, refers to the effect of X_j on the log odds that $Y=1$.
- β_0: the intercept (the constant)

4. Results and Discussion

The model of logistic regression was created using spss program and the analysis was based on enter method to determine the ability of the model to accurately predict diabetes. The size of the study sample included in analysis 189 cases, all cases were selected and no missing cases. The analysis results and discussion are as follows:

Table 1: Omnibus Tests of Model Coefficients

		Chi-square	Df	Sig.
Step 1	Step	140.096	10	.000
	Block	140.096	10	.000
	Model	140.096	10	.000

Table 1 shows the Chi-square test. It is used to test the null hypothesis "the model is not significant". The model p-value (sig=.000<.05), which is significant at statistical value (.05), for that the null hypothesis is rejected, which means the model is significant.

Table 2: Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	76.226[a]	.523	.768

Table 2 shows the Cox-Snell-R^2 and the adjusted Nagelkerke-R^2. From the table, between 52% and 77% of the variation in predicting diabetes can be explained by the model.

Table 3: Hosmer-Lemeshow Test

Step	Chi-square	df	Sig.
1	2.240	7	.945

Table 3 shows the Hosmer-Lemeshow test. It is used to test the null hypothesis "the model is represent the data well". The significance value of Chi-square (sig.=.945>.05), for that the null hypothesis is accepted, which mean the model is represents the data well.

Table 4: Contingency Table for Hosmer and Lemeshow Test

		Prediction = non-diabetes		Prediction = diabetes		
		Observed	Expected	Observed	Expected	Total
Step 1	1	19	18.788	0	.212	19
	2	16	16.365	3	2.635	19
	3	7	7.759	13	12.241	20
	4	5	4.677	15	15.323	20
	5	2	.923	17	18.077	19
	6	0	.346	20	19.654	20
	7	0	.114	21	20.886	21
	8	0	.022	19	18.978	19
	9	0	.006	32	31.994	32

Table 4 shows frequencies observed and expected for diabetic people and non-diabetic. From the table it is clear that the differences between the observed and expected values are not significant, which indicates the ability of the model to represent the data.

Table 5: Classification Table[a]

	Observed		Predicted		
			Prediction		Percentage Correct
			non-diabetes	diabetes	
Step 1	Prediction	non-diabetes	36	13	73.5
		diabetes	4	136	97.1
	Overall Percentage				91.0

a. The cu t value is .500

Table 5 shows the number of cases that were correctly predicted. 97.1% of people who diabetes are correctly predicted. 73.5% of people who non-diabetes are correctly predicted. Overall percentage 91% of the cases are correctly predicted, this percentage is considered very good.

Table 6: Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
							Lower	Upper
Gender	-1.813	.729	6.178	1	.013	.163	.039	.682
Age above 35	-2.544	.952	7.142	1	.008	.079	.012	.507
High Blood Pressure	-3.461	1.482	5.456	1	.020	.031	.002	.573
High Cholesterol	-2.749	.804	11.687	1	.001	.064	.013	.309
Diabetes in family 1st degree	-2.750	.717	14.698	1	.000	.064	.016	.261
Physical Activity: sports	.186	.743	.063	1	.802	1.204	.281	5.171
Physical Activity: works	-.154	.668	.053	1	.818	.857	.232	3.174
Food Habits (vegetables, fruits)	.564	.668	.713	1	.398	1.758	.475	6.513
Smoking	-2.647	1.153	5.267	1	.022	.071	.007	.680
Overweight	-2.405	.711	11.446	1	.001	.090	.022	.364
Constant	23.926	5.369	19.855	1	.000	2.459e+10		

Table 6 shows the regression coefficients (B), standard error, Wald statistics and Exp(B) for the variables included in the models. From this table Consider Wald and Sig. column. The significant variables are: (Gender, Age above 35, High Blood Pressure, High Cholesterol, Diabetes in family 1st degree, Smoking and Overweight). (sig.<0.05) for all these variables, which refer to significant effect of these variables in predicting diabetes. On the other hand, the not-significant variables are: (Physical Activity-ports, Physical Activity-works and Food Habits). (sig >0.05).

Column (B) and Exp(B), contained the model coefficients by logit and the odds ratio, which shows the effect of each predictor in predicting diabetes. For example, the variable Gender which coded (female=0, male=1), (B=-1.813 & Exp.B=.163), that means when gender goes from female to male lead to decrease the probability of diabetes by (1.813) in comparison to female, Exp(B)=.163 means that females are 0.163 times as likely to diabetes than males. The variable Age above 35 which coded (yes=1, no=2), yes mean age above 35 and no refer to less than 35, (B=-2.544& exp.B=0.079). That means increase by one value from 1 to 2 which transfers the variable from above 35 to less than 35, that lead to decrease the probability of diabetes by (2.544) in comparison by Age above 35. And exp(B)=0.079 means that the person whose age is above 35 is 0.079 times as likely to diabetes than less than 35. For all other significant variables, any increase by one unit from 1 to 2 decrease the probability of diabetes by the value of B.

From table 6, using the parameters estimates of significant variables, can get the following model:
Logit(diabetes predication)= 23.926 - 1.813(Gender) - 2.544(Age above 35) - 3.461(High Blood Pressure) -2.749(High Cholesterol) -2.750(Diabetes in family 1st degree) -2.647 (Smoking) - 2.405 (Overweight)

Table 7: Order and importance of the variables in the model

Predictor Variable	Exp(B)	Sig.	(1-expB)*100
High Blood Pressure	.031	.020	96.9
Diabetes in family 1st degree	.064	.000	93.6
High Cholesterol	.064	.001	93.6
Smoking	.071	.022	92.9
Age above 35	.079	.008	92.1
Overweight	.090	.001	91.0
Gender	.163	.013	83.7

Table 7 shows the order of the significant risk factors depends on their importance in the model, through their statistical significance and the ratio of predicting diabetes.

5. Conclusions

In order to obtain a highly efficient statistical model that can predict type-2 diabetes , a binary logistic regression was used to build a model to predict the probability of developing diabetes depending on risk factors for diabetes as predictor variables.

According to the results of the model quality measures, the model adopted in this study was highly efficient in predicting diabetes, as the results of the Chai square test based on the significant value of the test (sig=.000<.05) showed that the model was significant. The result of Hosmer-Lemeshow test based on the significance value (sig=0.945>.05) also confirmed that the model represents the data well. Likewise, Nagelkerke-R2=0.768 indicated the ability of independent variables in interpreting the change in the dependent variable.

What ensures the efficiency of the model and its high accuracy rate in predict, is the high values in classification tables, as the correct classification ratio for the infected reached 97.1%, and the overall percentage for the correct classification was 91%.

Through the sig-values of the Wald test (sig.<0.05) related to the significance of the independent variables in the model, the most significant and influential factors in predicting diabetes were arranged according to their degree of influence as follows: (High blood pressure, Diabetes in family first degree, High cholesterol, Overweight, Age above 35, Smoking, and Gender). As for

non-significant factors, they were (Physical Activity-sports, Physical Activity-works and Food Habits).

We conclude, on the base of the high efficiency of the constructed model, it is possible to rely on logistic regression in predicting type-2 diabetes through risk factors.

References

[1] Atlas, I. D. F. D. (2019) 463 PEOPLE LIVING WITH DIABETES million.

[2] Cataloguing, W. L. (2016) 'Global Report on Diabetes', Isbn, 978, pp. 6–86. Available at: http://www.who.int/about/licensing/copyright_form/index.html%0Ahttp://www.who.int/about/licensing/copyright_form/index.html%0Ahttps://apps.who.int/iris/handle/10665/204871%0Ahttp://www.who.int/about/licensing/.

[3] Hosmer, D. W. and Lemeshow, S. (2005) Applied Logistic Regression, Applied Logistic Regression. doi: 10.1002/0471722146.

[4] Islam, R. and Rahman, O. (2012) 'The Risk Factors of Type 2 Diabetic Patients Attending Rajshahi Diabetes Association, Rajshahi, Bangladesh and Its Primary Prevention', Food and Public Health, 2(2), pp. 5–11. doi: 10.5923/j.fph.20120202.02.

[5] Kleinbaum, D. G. and Klein, M. (2002) Logistic Regression A Self-Learning Text Second Edition, Survival.

[6] Maulana, Y. I. R., Badriyah, T. and Syarif, I. (2018) 'Influence of Logistic Regression Models For Prediction and Analysis of Diabetes Risk Factors', EMITTER International Journal of Engineering Technology, 6(1), pp. 151–167. doi: 10.24003/emitter.v6i1.258.

[7] Niyikora, S. (2015) 'Multiple logistic regression modeling on risk factors of diabetes. Case study of Gitwe Hospital (2011-2013)'. Available at : http: // www.jkuat.ac.ke/ campuses/kigali/wp-content/uploads/2014/04/NiyikoraSylivere2015-Multiple-logistic-regression-modeling-on-risk-factors-of-diabetescase-study-of-Gitwe-hospital-2011-2013.pdf.

[8] Rahimloo, P. and Jafarian, A. (2016) 'Prediction of Diabetes by Using Artificial Neural Network, Logistic Regression Statistical Model and Combination of Them', Bulletin de la Société Royale des Sciences de Liège, 85, pp. 1148–1164.

[9] Rahman, A. (2013) 'Application of Artificial Neural Network and Binary Logistic Regression in Detection of Diabetes Status', Science Journal of Public Health, 1(1), p. 39. doi: 10.11648/j.sjph.20130101.16.

[10] Rastogi, P. and Singh, B. K. (2019) 'A multivariate binary logistic regression modeling for a ssessing various risk factors that affect diabetes', International Journal of Scientific and Technology Research, 8(8), pp. 589–599.

[11] Risk, N. C. D. and Collaboration, F. (2016) 'Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4.4 million participants', Lancet (London, England), 387(10027), pp. 1513–1530. doi: 10.1016/S0140-6736(16)00618-8.

[12] Saied, A. and Abdallah, R. (2019) 'Using logistic regression models to determine factors affecting diabetes in the red sea state', 4(4), pp. 12–17.

[13] Senthilvel, V., Radhakrishnan, R. and Sathiyamoorthi, R. (2011) 'Prediction of diabetic retinopathy among diabetics using binary logistic regression approach', Indian Journal of Medical Specialities, 3(1). doi: 10.7713/ijms.2012.0005.
[14] Sperandei, S. (2014) 'Understanding logistic regression analysis', Biochemia Medica, 24(1), pp. 12–18. doi: 10.11613/BM.2014.003.
[15] Zehra, A. et al. (2018) 'Statistical modeling for prediction of diabetes in Malaysians', Life Science Journal, 15(6).