Bol. Soc. Paran. Mat.
©SPM – E-ISSN-2175-1188
SPM: www.spm.uem.br/bspm

(3s.) **v. 2026 (44) 2** : 1–7.
ISSN-0037-8712
doi:10.5269/bspm.80611

Analyzing Factors Influencing Diabetes and Predicting its Occurrence Using Logistic Regression

Manohar Dingari, S. Hariprasd, N. Subadra and V. Sumalatha*

ABSTRACT: The current work seeks to examine the determinants affecting diabetes and to create a predictive model grounded in those determinants utilizing Logistic Regression. The dataset utilized for this study is obtained from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) and comprises data on 768 female patients. The main goal is to find the most important factors that lead to diabetes and to see how effectively the model can tell the difference between those who are diabetic and those who are not. Logistic regression, a prevalent statistical technique for binary classification, is utilized to estimate the chance of diabetes occurrence based on many independent variables, including glucose level, blood pressure, body mass index (BMI), insulin level, skin thickness, and age. The Hosmer and Lemeshow goodness-of-fit test is used to check how well the model works and how accurate it is. The results show that the logistic regression model fits the data well and may be used to make predictions. The investigation revealed that skin thickness, insulin levels, and age are negligible predictors, whereas glucose levels, body mass index (BMI), and blood pressure significantly influence diabetes prediction. This discovery is consistent with medical knowledge that obesity and high glucose levels are significant risk factors for diabetes. While other studies have investigated machine learning techniques for diabetes prediction, this work underscores a statistical modeling approach utilizing logistic regression. The model not only predicts diabetes risk but also helps us understand the main elements that affect it in the examined population.

Keywords: Diabetes, logistic regression, classification table, percentage accuracy, specificity, sensitivity.

Contents

1. Introduction

High blood glucose levels (hyperglycemia) resulting from insulin secretion or action abnormalities, or both, characterize a chronic metabolic condition known as diabetes. It is probable that the pancreas to not create enough insulin, or for the insulin that is produced to not be used efficiently by the physical form. By promoting glucose uptake into cells for energy synthesis, the hormone insulin is crucial in regulating blood glucose levels. A disruption in this regulation mechanism causes blood sugar levels to rise, which, if left unchecked, can harm many bodily systems and organs, such as the nervous system and blood vessels. Kidney failure, cardiovascular disease, nerve damage, and visual issues are all more likely in those with hyperglycemia who experience it for a long amount of duration.

The World Health Organization discovered that 8.5% of adults aged 18 and over had diabetes in 2014. In 2019, 48% of deaths before 70 were directly related to diabetes. Diabetes also contributed to nearly 20% of heart failures and an estimated 0.46 million deaths from kidney disease. Diabetes has emerged as one of the leading global health challenges of the 21st century due to causes that include urbanization,

* Corresponding author.
2020 *Mathematics Subject Classification:* 62J12.
Submitted December 05, 2025. Published March 14, 2026

sedentary lifestyles, and changing dietary patterns. As a consequence, early prediction and diagnosis of diabetes are crucial for preve4ntion and effective management.

The application of both classis statistical models and cutting-edge machine learning algorithms to forecast the onset of diabetes has been the focus of multiple experiments. Logistic Regression (LR) is a common and easily understandable statistical method for binary classification issues. It is especially useful for diseases like diabetes, where the response variable can only be one of the two potential values: whether someone has diabetes or not. Employing independent factors including glucose level, BMI, blood pressure, age, and family history, logistic regression evaluates the likelihood of diabetes. Logistic regression allows researchers to comprehend the influence of each predictor on the risk of acquiring diabetes by modeling the dependent variable's log-odds, as compared to linear regression's analysis of continuous outcomes.

Considering the use of 768 female patients' wealth of data from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), the current investigation intends to use logistic regression for predicting the eventual development of diabetes. The primary objective is to establish a model that is both statistically solid and easy to understand in order to determine the variables that significantly influence the occurrence of diabetes. To assess if the logistic regression model satisfies the criteria for a good fit, the Hosmer-Lemeshow test is employed. Some variables, such as age, skin thickness, and insulin levels, may not add much to the model, judging by preliminary results, but other predictors have considerable effects on the risk of diabetes.

2. Literature Review

Millions of people all around the entire world endure the effects by diabetes mellitus, an occasionally deadly metabolic disorder that can persist over time. It is associated with major health issues and mortality [16]. In order to lower the burden of illness, early detection and prevention are crucial, which explains precisely researchers have constructed predictive models using ML approaches. Datasets related to diabetes were initially examined using conventional classifiers such as logistic regression, decision trees, and support vector machines (SVM) [9,15,17]. For more accurate predictions, Battineni et a1. [1] used cross validation methods; for type 2 diabetes, Birjais et a1. [2] and Lai et a1. [10] validated logistic regression models. Corresponding with this, Sisodia and Sisodia [15] reviewed several classification algorithms and concluded that SVM provided the best results. Although these early models were straight forward, they often experienced from difficulties stemming from their reliance on limited feature sets and tiny sample sizes.

Later studies used superior algorithms and hybrid approaches that incorporate feature optimization and ensemble learning to make predictions more accurate. Hasan et al. [5,6] inspected the impact of health-related parameters, as well as glucose concentration, BMI, and insulin levels, on the onset of diabetes. Jagannathan et al. [7] stressed the prerequisite of clinically approved diagnostic tests, like the Oral Glucose Tolerance Test (OGTT), to appear how medical and computational models are associated.Ahmad and Khan [18] generated a hybrid model that combined logistic regression and random forest. This model did better than standard methods at finding diabetes early. Rahman et al. [13] and Rawat et al. [14] suggested ensemble techniques that enhanced classification accuracy and mitigated bias. Kaur and Kumari [23] have proposed improved feature selection to improve computational efficiency. Ghosh et al. [4] and Mujumdar and Vaidehi [11] also talked about comparing algorithms to get the best prediction tool.

Recent progress has concentrated on deep learning (DL) and explainable artificial intelligence (XAI) to enhance model interpretability and foster therapeutic trust. Deepa and Priya [21] showed that DL models are better than logistic regression at finding complicated nonlinear correlations in health data. Jian et al. [8] and Zhang and Zhou [24] employed XAI frameworks to forecast diabetes complications, enhancing transparency in diagnostic choices. Sun et al. [25] utilized deep learning for real-time risk analysis of electronic health records, whereas Wang and Chen [26] examined novel AI applications in diabetes treatment. Research conducted by Aljame et al. [22] and Li et al. [19] substantiates that the amalgamation of feature selection, neural networks, and hybrid modeling improves both predictive accuracy and interpretability. Abdar et al.'s [20] thorough reviews and The World Health Organization's [16] and the National Institute of Diabetes and Digestive and Kidney Diseases' [12] global reports show how AI is changing healthcare. These works collectively illustrate a paradigm change from traditional statis-

tical models to intelligent, explicable, and clinically integrated AI systems that enhance early diagnosis and individualized diabetes management.

3. Methodology

Logistic Regression is a statistical regression analysis which is used when there is a single dichotomous outcome and one or more independent variables which may be either continuous or categorical. Thus the logit model represents the log odds of an as a linear combination of predictor variables. Since the response variable is binary, it takes two values which are generally coded as 1 and 0. Code '1' indicates the success of the event and '0' failure of the event. The probability of occurrence of the code 1 lies between 0 and 1. In any regression the conditional means can be expressed as a linear regression model

$$E(\frac{Y}{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k \tag{3.1}$$

Where Y is the response variable which is binary, $X_1, X_2 \ldots . X_k$ are independent variables andβ_0, β_1, $\beta_2 ....\beta_k$ are the model parameters.

$\Pi(x) = \frac{\beta_0+\beta_1 X_1+\beta_2 X_2+.....+\beta_k X_k}{1+\beta_0+\beta_1 X_1+\beta_2 X_2+.....+\beta_k X_k}$ is called logistic function and it is a monotonic function tends to 0 as x tends to infinity.

Non linearity in the logistic regression can be converted to linearity with help of odds ratio. The odds is the probability that a particular outcome is a case divided by the probability that is a non-case i.e., $\frac{\Pi(x)}{1-\Pi(x)} = \frac{P(Y=1)}{P(Y=0)}$.

The parameters of the logistic regression are estimated by Maximum likelihood method. In this study, Hosmer - Lemeshow test is used to check the adequacy of the fitted Logistic regression.

Data: The dataset utilized in this study was sourced from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), a reputable provider of diabetes research data. There are 768 female patients in the dataset, all of them are Pima Indians and are at least 21 years old. Due to its thorough inclusion of clinically relevant health factors, this dataset has been utilized in several research that tried to predict diabetes. The primary aim of this study is to examine the impact of various physiological and demographic factors on the probability of developing diabetes and to develop a predictive model utilizing Logistic Regression, a reliable statistical method appropriates for binary classification scenarios—where the dependent variable has two potential outcomes: diabetic or non-diabetic.

This analysis involves a look at eight different characteristics to see how each of them promote diabetes, both individually and together. Age, since the likelihood of developing diabetes rises with each passing year; Body Mass Index (BMI), an essential indicator of obesity and insulin resistance; and skin thickness, an indirect reflection of body fat distribution, are all among these factors. One of the most striking symptoms that a person has diabetes is a high glucose level, which indicates the amount of sugar in the blood. You can find out your risk of developing diabetes by looking at your family genealogy with the diabetes pedigree characteristic. Your pancreas's effectiveness of manufacturing insulin can be determined by your insulin level. Last but not least, blood pressure is brought up because it is frequently associated with metabolic issues and cardiovascular risks that accompany diabetes.

Predictions in the fields of medicine and social science are frequently generated using SPSS, a program for statistical analysis. It was with this application that the Logistic Regression model was developed and tested. By incorporating each of these variables, the model determines the likelihood that a person has hyperglycemia. Using this method, we can better understand which variables significantly impact diabetes risk and which ones have a more minimal impact. In addition to providing us with understandable coefficients, the logistic regression model is a helpful tool to assess the degree of significance and direction of the association between diabetes status factors. Early diagnosis, clinical decision-making, and the implementation of preventive healthcare programs are all made easier with the help of this study's results, which contribute to the development of a statistically robust and understandable framework for diabetes risk forecasting.

4. Results

Present study used Hosmer Lemeshow test to check the best fit of the logisyic regression model.

Table 1: Hosmer and Lemeshow test

Step	Chi-square	Df	Sig.
1	8.323	8	.403

Table1 demonstrates the Hosmer and Lemeshow test which is used to diagnose the adequacy of the logistic regression. Since $p > 0.05$, it is clear that the fitted logistic regression is best fit for the data.

Table 2: Classification Table

Observed		**Predicted**		**Percentage Correct**
	Outcome	**0**	**1**	
Step 1	0	445	55	89.0
	1	112	156	58.2
Overall Percentage				78.3

5. Discussion

Essential insights regarding the model's effectiveness in predicting diabetes among those who participated may be found in the classification table generated from the logistic regression study. The tool contrasts the expected categories with the actual results, allowing one to see how accurate and dependable the predictions model is. You can see how well the model performs and how well it differentiates between people with diabetes and non- diabetics through examining each of the statistical metrics in this table, such as percentage accuracy, specificity, sensitivity, positive value and negative predictive value.

As a proportion of the total instances investigated, the classification accuracy percentage reveals the proportion of correct model predictions. Overall, the classification accuracy of the fitted logistic regression model in this study is 78.3%. This indicates that in 78.3% of the cases, a determination of having diabetes or non- diabetes was correctly made. This illustrates that, depending to the specified criteria, the model could differentiate between people with diabetes and non- diabetics quite effectively.

The specificity of the model is defined as the proportion of correctly predicted true negatives. This indicator evaluates the model's ability in recognizing people who are fit. The model's ability to prevent false positives, in which persons who do not have diabetes are mistakenly believed to have it, is shown by a high specificity number. The specificity of the model used in the present research was 89%. Which indicates that it accurately forecast that 89% of the people involved who did not have diabetes actually did not have the disease. Because of its high specificity, the logistic regression model consistently classifies healthy individuals accurately.

You can tell how many actual cases of diabetes the model accurately recognized by looking at its sensitivity, which is also known as its true positive rate. This demonstrates how accurate the model is at identifying diabetics. In this study, the model's sensitivity was 58.2%, meaning it correctly identified 58.2% of the diabetic cases. While this score falls short in specificity, it demonstrates that the model is capable of reasonably predicting diabetes patients. To increase the model's sensitivity, you may alternatively incorporate more predictive factors or adjust the model parameters so that they better reflect diabetes-related features.

The PPV measures the accuracy of the diagnosis of diabetes relative to the number of individuals who were predicted to have the disease. It sheds light on the reliability of optimistic predictions. The above scenario has a positive prediction value of 73.9% as estimated by the model. So, around 73.9% of those who were predicted to have diabetes ended up having the disease. With a high PPV, the model is accurate in predating the general incidence of diabetes among the general public across the globe.

The negative predictive value (NPV) tells you how many of the people who were projected to be non-diabetic were indeed non-diabetic. It shows how well the model can guess that someone doesn't have diabetes. The NPV for this study is 79.9%, which means that almost 79.9% of the patients that were predicted to be non-diabetic were indeed non-diabetic. This is because $100 \times (445 \,/\, (445 + 112)) = 79.9\%$. This illustrates that the approach works well to leave out people who don't have diabetes.

In conclusion, the classification table gives a full picture of how well the model can diagnose. The logistic regression model has a high level of specificity and a fair level of sensitivity, which means that it can accurately find non-diabetic people while still being able to find diabetes people with a fair level of accuracy. The model is statistically sound because it has an overall predictive accuracy of 78.3%. This means that it can be a useful tool for predicting diabetes based on the independent variables that were chosen.

Table 3: Variables in the Equation

Step	**Variable**	**B**	**S.E.**	**Wald**	**df**	**Sig.**	**Exp(B)**
Step 1â	Pregnancies	0.123	0.032	14.747	1	0.000	1.131
	Glucose	0.035	0.004	89.897	1	0.000	1.036
	BloodPressure	-0.013	0.005	6.454	1	0.011	0.987
	SkinThickness	0.001	0.007	0.008	1	0.929	1.001
	Insulin	-0.001	0.001	1.749	1	0.186	0.999
	BMI	0.090	0.015	35.347	1	0.000	1.094
	DiabetesPedigreeFunction	0.945	0.299	9.983	1	0.002	2.573
	Age	0.015	0.009	2.537	1	0.111	1.015
	Constant	-8.405	0.717	137.546	1	0.000	0.000

a. Variable(s) entered on step 1: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age.

Statistical significance of the independent variables is determined by Wald test, from table 3 it can be observed that number of pregnancies (X_1), glucose (X_2), blood pressure (X_3), BMI(X_6), diabetes pedigree function (X_7) are significantly contributing the model as the significance value corresponding to these variables is > 0.05, on the other hand the significance values of skin thickness (X_4), insulin (X_5) and age (X_8) are >¿ 0.05 which indicates that these 3 variables do not contribute significantly to the model. The logistic model for predicting the diabetes is

$P(Y = 1) = \Pi(0.123X_1 + 0.035X_2 - 0.013X_3 + 0.001X_4 - 0.001X_5 + 0.090X_6 + 0.945X_7 + 0.015X_8 - 80405)$

Where $\Pi(x) = \frac{e^x}{(1+e)^x}$

If this probability is > 0.5 the subject is positive with diabetes otherwise negative.

6. Conclusions

Diabetes mellitus is one of the most common non-communicable diseases in the world, impacting millions of people. It is a long-term metabolic condition that causes high blood sugar levels because the pancreas doesn't make enough insulin or the body can't use the insulin it does make. The rising number of people with diabetes is a big health problem around the world since it leads to cardiovascular disorders, renal failure, nerve damage, and other problems with organs. Early diagnosis and appropriate care of diabetes are essential in mitigating these risks and enhancing the quality of life for patients.

This study aims to create a predictive model for diabetes utilizing Logistic Regression, a robust and comprehensible statistical method for binary classification challenges. The research employs an open dataset sourced from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), comprising data on 768 female participants aged 21 years and older. Eight independent variables—number of births, glucose level, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, and age—were examined to evaluate their impact on the incidence of diabetes. The dependent variable was binary, indicating the presence or absence of diabetes in an individual.

We used SPSS statistical software to fit the logistic regression model and the Hosmer and Lemeshow test to see how well it worked. The test findings indicated that the model fit was statistically sound, showing that logistic regression adequately captures the link between the independent variables and the probability of having diabetes. The study found that glucose level, BMI, number of pregnancies, and diabetes pedigree function were strong predictors of diabetes. Other characteristics, like age, insulin, and skin thickness, were less strongly linked to diabetes. The model has an overall classification accuracy of

78.3%, which shows that it is a good predictor. The model's specificity (89%) and sensitivity (58.2%) also showed that it was good at telling the difference between diabetes and non-diabetic cases.

The results of this study highlight the efficacy of logistic regression as a viable and comprehensible statistical approach for medical diagnosis and illness forecasting. Logistic regression is easier to understand and interpret than complicated machine learning models since it shows how each variable affects the outcome. This ability to understand is very important in healthcare settings, where judgments need to be clear and fair. The study also gives clinicians and public health officials a way to find those who are at high risk and take steps to stop them from getting sick early on.

References

1. Battineni G, Sagaro GG, Nalini C, Amenta F, Tayebati SK, *Comparative machine-learning approach: A follow-up study on type 2 diabetes predictions by cross-validation methods* Machines. 2019 Dec 5;7(4):74.
2. Birjais R, Mourya AK, Chauhan R, Kaur H, *Prediction and diagnosis of future diabetes risk: a machine learning approach*, SN Applied Sciences. 2019 Sep;1(9):1112.
3. Deberneh HM, Kim I, *Prediction of type 2 diabetes based on machine learning algorithm. International journal of environmental research and public health*, 2021 Mar 23;18(6):3317.
4. Ghosh P, Azam S, Karim A, Hassan M, Roy K, Jonkman M, *A comparative study of different machine learning tools in detecting diabetes*, Procedia Computer Science. 2021 Jan 1;192:467-77.
5. Ahmad HF, Mukhtar H, Alaqail H, Seliaman M, Alhumam A *Investigating health-related features and their impact on the prediction of diabetes using machine learning*, Applied Sciences. 2021 Jan 27;11(3):1173.
6. Ahmed N, Ahammed R, Islam MM, Uddin MA, Akhter A, Talukder MA, Paul BK, *Machine learning based diabetes prediction and development of smart web application*, International Journal of Cognitive Computing in Engineering. 2021 Jun 1;2:229-41.
7. Jagannathan R, Neves JS, Dorcely B, Chung ST, Tamura K, Rhee M, Bergman M, *The oral glucose tolerance test: 100 years later. Diabetes, metabolic syndrome and obesity*, 2020 Oct 19:3787-805.
8. Jian Y, Pasquier M, Sagahyroon A, Aloul F, *A machine learning approach to predicting diabetes complications*, In Healthcare 2021 Dec 9 (Vol. 9, No. 12, p. 1712). MDPI.
9. Joshi TN, Chawan PM, *Logistic regression and svm based diabetes prediction system*, International Journal For Technological Research In Engineering. 2018 Jul;5:4347-50.
10. Lai H, Huang H, Keshavjee K, Guergachi A, Gao X, *Predictive models for diabetes mellitus using machine learning techniques*, BMC endocrine disorders. 2019 Oct 15;19(1):101.
11. Mujumdar A, Vaidehi V, *Diabetes prediction using machine learning algorithms*, Procedia Computer Science. 2019 Jan 1;165:292-9.
12. *World Health Organization, diabetes fact sheet*, available at: https://www.who.int/news-room/fact-sheets/detail/diabetes (2021).
13. *Rahman, M. A., Rahman, M. H., & Rahman, M. S. (2023). Prediction of diabetes disease using an ensemble of machine learning techniques*, BMC Bioinformatics, 24(1), 212.
14. Rawat, S., et al. (2023), *Optimizing diabetes classification with a machine learning-based approach, BMC Bioinformatics, 24(1), 220.*
15. *Sisodia, D., &Sisodia, D. S. (2018).* Prediction of diabetes using classification algorithms, Procedia Computer Science, 132, 1578–1585.
16. *World Health Organization, (2021). Diabetes fact sheet.*
17. *Yuvaraj, N., &SriPreethaa, K. R. (2017),* Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster, Cluster Computing, 22, 1–9.
18. Ahmad, M. A., & Khan, M. F. (2022), *A hybrid model combining logistic regression and random forest for early prediction of diabetes, Informatics in Medicine Unlocked, 30, 100976.*
19. *Li, Y., Zhao, R., & Xu, X. (2022)* Comparative analysis of logistic regression and neural networks for diabetes prediction, Computers in Biology and Medicine, 150, 106155.
20. Abdar, M., et al. (2021), *A review of advanced machine learning methods for diabetes prediction, Artificial Intelligence in Medicine, 117, 102105.*
21. *Deepa, N., &Priya, R. (2020),* Prediction of type 2 diabetes using logistic regression and deep learning algorithms, Journal of Ambient Intelligence and Humanized Computing, 11, 4433–4445.
22. Aljame, M., et al. (2022), *Machine learning for diabetes prediction and risk factor identification, IEEE Access, 10, 55432–55445.*

23. Kaur, P., &Kumari, V. (2022), Machine learning approach for diabetes prediction using optimized feature selection, Biocybernetics and Biomedical Engineering, 42(3), 870–884.

24. Zhang, J., & Zhou, Y. (2023), *Predicting diabetes complications using explainable artificial intelligence, Expert Systems with Applications, 224, 119907.*

25. Sun, J., et al. (2024), Deep learning–based prediction and interpretation of diabetes risk using health records, Frontiers in Digital Health, 6, 1557467.

26. Wang, H., & Chen, X. (2025), *Advances in artificial intelligence for diabetes prediction, Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 19(1), 102325.*

Manohar Dingari,
Department of Mathematics,
School of Engineering,
ANURAG University,
Hyderabad, India.
E-mail address: manohar.dingari@gmail.com

and

S. Hariprasd,
Department of Mathematics,
School of Engineering,
ANURAG University,
Hyderabad, India.
E-mail address: srinadhunihariprasd@gmail.com

and

N. Subadra,
Department of Mathematics,
Geethanjali College of Engineering and Technology
Cheerryal,
Keesara,
Hyderabad,
India.
E-mail address: nemani.subhadra@gmail.com

and

V. Sumalatha,
Department of Mathematics and Statistics,
Veeranari Chakali IlammaWomen's University,
Hyderabad,
India.
E-mail address: sumanu05@gmail.com