منشورات جامعة وادي النيل
مجلة النيل للآداب والعلوم الانسانية
(ISSN: 1858 – 7054)
المجلدالسادس، العدد الأول2025 م
nilevalley.edu.sd

Estimation and Analysis of Factors Affecting Diabetes Using Full and Stepwise Logistic Regression Models: A Comparative Study

Elnazeer Mohamed Elnoor
IBN SINA University
Corresponding Author :ac.aff.isu@gmail.com

Abstract:

This study aims to evaluate the performance of a predictive model of diabetes using a classification table; a complete logistic regression model (8 variables) and a graduated logistic regression model (5 variables) were studied. The performance of the two models was analyzed using various metrics including weighting logarithm, determination coefficients (Nagelkerke and Cox & Snell), Hosmer and Lemeshow test, Omnibus test, and rating accuracy. The results showed a slight superiority of the complete model in the overall classification accuracy (78.3% versus 77.5% for the graduated model). The full model also outperformed the following: Data relevance: Lower weighting logarithm (723.445 vs. 728.560). Explanatory power: higher determination coefficients (0.408 and 0.296 vs. 0.402 and 0.292). Statistical significance: Chi-square value is higher in the Omnibus test (270.039 vs. 264.924). While the graduated model showed slight superiority in the Hosmer and Lemeshow test (higher p-value: 0.421 vs. 0.403), this did not compensate for the full model's models, but lower sensitivity, which means it is difficult to identify the actual infected. The study concluded that the complete model is generally better, emphasizing the importance of including all possible variables. However, both models still need to be improved, especially in sensitivity, by collecting additional data or using sophisticated modeling techniques. The study also recommends the use of additional assessment measures such as precision and the F1 coefficient for a more comprehensive assessment. In short, the complete model offers better performance, but it needs improvements, especially in identifying actual casualties to reduce negative errors and ensure accurate diagnosis.

Keywords*: diabetes, prediction, classification model, classification table, specific sensitivity, negative error.*

تقدير العوامل المؤثرة في مرض السكري
باستخدام النماذج اللوجستية الكاملة والتدريجية: دراسة مقارنة

المستخلص :

تهدف هذه الدراسة إلى تقييم أداء نموذج تنبؤي للإصابة بمرض السكري باستخدام جدول التصنيف، تمت دراسة نموذج الانحدار اللوجستي الكامل (8 متغيرات) ونموذج الانحدار اللوجستي المتدرج (5 متغيرات). تم تحليل أداء النموذجين باستخدام مقاييس متنوعة تشمل لوغاريثم الترجيح، معاملات التحديد (Nagelkerke و Snell&Cox)، اختبار Hosmer and Lemeshow، اختبار Omnibus، ودقة التصنيف. أظهرت النتائج تفوقًا طفيفًا للنموذج الكامل في دقة التصنيف العامة (78.3% مقابل 77.5% للنموذج المتدرج). كما تفوق النموذج الكامل في: ملاءمة البيانات: لوغاريثم ترجيح أقل (723.445 مقابل 728.560). القوة التفسيرية: معاملات تحديد أعلى (0.408 و 0.296 مقابل 0.402 و 0.292). الدلالة الإحصائية: قيمة Chi-square أعلى في اختبار Omnibus (270.039 مقابل 264.924). بينما أظهر النموذج المتدرج تفوقًا طفيفًا في اختبار Hosmer and Lemeshow(قيمة p أعلى: 0.421 مقابل 0.403)، إلا أن هذا لم يعوض عن تفوق النموذج الكامل في المقاييس الأخرى. كشف تحليل جدول التصنيف عن نوعية عالية في كلا النموذجين، لكن حساسية أقل، ما يعني صعوبة في تحديد المصابين الفعليين. خلصت الدراسة إلى أن النموذج الكامل أفضل بشكل عام، مؤكدة أهمية تضمين جميع المتغيرات المحتملة. مع ذلك، لا يزال كلا النموذجين بحاجة لتحسين، خاصةً في الحساسية، عبر جمع بيانات إضافية أو استخدام تقنيات نمذجة متطورة. توصي الدراسة أيضًا باستخدام مقاييس تقييم إضافية كالدقة (Precision) ومعامل F1 لتقييم أشمل. باختصار، يُقدم النموذج الكامل أداءً أفضل، لكنه يحتاج لتحسينات، خاصةً في تحديد المصابين الفعليين لتقليل الأخطاء السلبية وضمان دقة التشخيص.

كلمات مفتاحية : *مرض السكري التنبؤ , نموذج التصنيف, جدول التصنيف,الحساسية النوعية,الخطأ السلبي.*

Introduction:

Diabetes mellitus (DM) is one of the major global health challenges facing different age groups, representing a growing health and economic burden on individuals and societies. The disease is caused by an imbalance in the production or utilization of the hormone insulin, resulting in elevated blood glucose levels, with serious complications affecting the health of patients. Although diabetes encompasses Type 1, caused by an interruption in insulin production, and Type 2, linked to lifestyle and genetic factors, Type 2 is the most prevalent and presents a major challenge due to its association with behavioral, economic and social changes.

Over recent decades, diabetes rates have risen significantly, especially in developing countries such as Sudan, where factors such as obesity, unhealthy dietary patterns, and lack of health awareness contribute to the increasing prevalence of the disease. Despite efforts to combat diabetes, there are still significant gaps in understanding the relationship between factors influencing the likelihood of developing diabetes and the effectiveness of the predictive models used.

Statement of the Study:

The main issue revolves around the growing challenges posed by diabetes globally and is summarized in the following points:

- Rising incidence: Diabetes, especially type II diabetes, is on the rise due to multiple factors such as obesity, unhealthy diets, and lack of health awareness.
- Gaps in understanding: There is still a lack of understanding of the exact relationship between the factors that influence the likelihood of developing diabetes and the effectiveness of the predictive models used.
- Need for accurate models: There is a need to develop accurate statistical models, such as logistic regression, to analyze these factors systematically and effectively.
- The issue is the lack of ability to identify the factors most associated with diabetes, and the unclear impact of age, gender, BMI, and glucose levels on the likelihood of developing diabetes. In addition, there is a need to develop accurate statistical models, such as logistic regression, to systematically analyze these factors, allowing a deeper understanding of the factors influencing the disease.

Research Objectives:

The study aims to achieve the following:

- Emphasize the importance of statistics in analyzing diabetes data, using a logistic regression model to identify factors influencing the likelihood of developing diabetes, such as age, gender, body mass index, and glucose levels
- Improve the accuracy of predictive models: To develop and improve the accuracy of predictive models used in predicting disease incidence.
- Provide recommendations based on statistical analysis to support strategies for prevention and early diagnosis of diabetes, especially in developing countries.
-. In addition, the study seeks to improve the accuracy of predictive models and provide recommendations based on statistical analysis to support prevention strategies and early diagnosis of diabetes, especially in developing countries where the disease represents a major challenge

Significance/Importance of Research:

The importance of the study represents in the following:

- Highlighting the role of statistics in public health: The study highlights the importance of using advanced statistical methods, such as logistic regression, to analyze public health data and make evidence-based health decisions.
- Addressing a global health challenge: The study contributes to addressing one of the most prominent global health challenges, diabetes, which represents a significant health and economic burden.
- Deeper understanding of influencing factors: The study provides a deeper understanding of the factors that influence the likelihood of developing diabetes, helping to develop more effective prevention strategies.
- Improved prediction accuracy: Improving the accuracy of predictive models contributes to earlier diagnosis of the disease, enabling early intervention and minimizing complications.
- Support prevention and diagnostic strategies: The study provides recommendations based on statistical evidence to support prevention and early diagnosis strategies, especially in developing countries that face greater challenges in combating the disease.

Research Methodology:

The study adopts a quantitative approach, both descriptive and analytical

- Statistical Analysis: Statistical methods were used to analyze diabetes data.
- Logistic Regression Model: The logistic regression model was used to identify the factors affecting the likelihood of developing the disease.
- Factor analysis: Analyzing the impact of specific factors, such as age, gender, BMI, and glucose levels, on the likelihood of developing the disease.
- Develop and optimize predictive models: Developing and improving the accuracy of predictive models used to predict diabetes.
- Provide evidence-based recommendations: Provide recommendations based on the results of statistical analyses to support prevention and early diagnosis strategies.

Previous Studies:

- The study of Ram D. Joshi and Chandra K. Dhakal (2001) analyzed the factors influencing the incidence of type 2 diabetes using logistic regression model and decision tree algorithm. The study focused on data from Indian Pima women and identified glucose, pregnancy, body mass index (BMI), diabetes incidence, and age as the most important influencing factors. The study had a prediction accuracy of 78.26 per cent and an error rate of 21.74 per cent. The researchers emphasized that the model can support preventive measures to reduce the prevalence of the disease and the costs associated with the study of Zainab Ahmed Al-Birmani and Aisha Abdul Khaliq Ismail (2019) aimed to study the factors influencing diabetes using logistic regression. The study included a random sample of 150 elderly people in the city of Hilla, and 14 independent variables were analyzed. The study concluded that smoking, exercise, vitamin D, and blood pressure were the most influential factors, while the rest of the variables did not show statistical significance. The correct classification rate of the model was 92.7%, highlighting the efficiency of the model in predicting diabetes.
- The study of Olufemi, C. Obunadeke et al (2023) analyzed the impact of different factors on early prediction of diabetes in the United States using logistic regression. The study showed that 'frequent glucose' and 'excessive dehydration' were the most important factors in predicting the disease. The study used a Lasso regularization method to optimize the model, achieving 95% prediction accuracy based on the area under the ROC curve. The results confirmed the effectiveness of the model in predicting diabetes and its use as a reliable tool in early diagnosis.
- The study of Rajendra and Shahram Latifi (2021) indicated the importance of logistic regression as an effective tool in clinical analysis and prediction of diseases, including diabetes. The study

emphasized that early diagnosis using predictive models can reduce disease progression and associated complications. The study used machine learning techniques to improve the accuracy of diabetes prediction, enhancing the model's potential for use in the medical field.

Comparison between Current Study and the Previous Studies:

The current study characterizes the risk of diabetes by using an advanced logistic regression model based on comprehensive data. Unlike previous studies that focused on limited techniques or datasets, this study seeks to provide accurate predictions based on a variety of factors and analyzed using state-of-the-art statistical methods. The study aims to improve predictive models and develop more effective prevention and treatment strategies to combat diabetes.

Theoretical Frameworks:

Diabetes:

Diabetes is a global health challenge that affects millions of people around the world. This chronic disease causes serious health complications such as blindness, kidney failure, heart attacks, strokes, and lower limb amputations. Understanding the disease, including its causes, symptoms, and treatments, is critical for prevention, early diagnosis, and effective treatment of diabetes [1][2]:

Type 1 diabetes:

Develops when the immune system attacks the pancreatic cells responsible for producing insulin, requiring daily insulin administration. This type is usually diagnosed in childhood or young adulthood.(International Journal of Environmental Research and Public Health, 2021, p. 7346).

Type II diabetes:

This is the most common type, caused by the body's resistance to insulin or lack of insulin production. Often associated with obesity and an unhealthy lifestyle, it mainly affects adults, but is increasingly appearing in children due to the increase in obesity rates.

3. Diabetes that appears during pregnancy in some women, and often disappears after birth. However, it increases the risk of developing type 2 diabetes later on.

4. Other rare types include those caused by genetic mutations, issues with the pancreas due to infections or surgery, or as a result of the use of certain medications. (International Journal of Environmental Research and Public Health, 2021, p. 7346).

Symptoms of diabetes:

- Excessive thirst. - Feeling tired. - Frequent need to urinate. - Blurred vision. - Feeling hungry. - Weight loss. - Slow wound healing. - Frequent fungal infections or urinary tract infections. (World Health Organization [WHO], n.d.).

Causes of diabetes:

- Increased obesity. - Older age. - Family history of diabetes. - Lack of physical activity.
- Previous gestational diabetes. - High blood pressure or triglycerides.

Diagnosis of diabetes:

Diagnosis is based on tests, the most important of which are:

1. Fasting blood sugar test: Measurement of glucose level after an 8-hour fast. If the result is ≥126 mg/dl in two separate tests, it indicates diabetes.

2. Random glucose test: Glucose measurement at any time. A result of ≥200 mg/dl with symptoms of diabetes confirms the diagnosis.

3. Oral glucose tolerance test: Drink a sugary solution and measure the sugar level two hours later. A result of ≥200 mg/dl confirms the diagnosis.

4. Hemoglobin A1c test: Measures the average blood sugar over the past three months. A result of ≥6.5% indicates diabetes [1][2]

The relationship between hypertension and diabetes:

70% per cent of people with diabetes also have high blood pressure. This association increases the risk of complications such as heart disease, stroke, retinopathy, kidney disease, and peripheral vascular disorders.

Factors affecting diabetes: - Genetic factors and being overweight. - Lack of physical activity. - Unhealthy diet. - Age. - High blood pressure. - Pregnancy. (World Health Organization [WHO], n.d.).

Logistic Regression:

Logistic regression is a common statistical method used to analyze binary (categorical) data and estimate the likelihood of an event occurring based on a set of independent variables. This model is characterized by its flexibility and ability to deal with non-linear data by using a logistic function to convert predicted values into probabilities between 0 and 1, making facilitate analyzing and classifying different phenomena.(Al-Ali, 2020, p. 25)

Logistic regression model formulation:

The model is represented as:

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

where:

- (P): The probability of an event occurring.
- Log $\left(\frac{\mathrm{p}}{1-\mathrm{p}}\right)$The natural logarithm of the likelihood ratio ($logit$).
- (β_0): Constant.
- (β_i): Estimated coefficients of the independent variables.
- (X_i): The independent variables.

Applications of logistic regression:

-Medicine: Estimating the likelihood of certain diseases based on health factors.

- Economics: Analyze consumer behavior or predict financial crises.

-Marketing: Categories customers and determine their likelihood of responding to marketing offers.(Al-Ali, 2020, p. 25)

Advantages of logistic regression:

1. Simplicity of the model: Easy to interpret.
2. Handles binary data: Suitable for analyzing binary variables.
3. Accurate prediction: Allows accurate estimation of the probability of events.

Limitations of logistic regression:

1. Specific assumptions: Relies on the assumption of a linear relationship between the independent variables and the natural logarithm of the proportional number.
2. Handling interactions: May require modification to include interaction effects between variables.

Rationale for using logistic regression:

When using linear regression analysis with binary variables, two main issues arise:

1. Conceptual issues:
   - Probabilities must be between 0 and 1, making linear regression inappropriate for these data.
   - The data is fitted using an ‘S’ shaped logistic curve instead of a straight line.
2. Statistical issues:
   - Violation of assumptions of linear regression analysis, such as normal distribution and homogeneity of variance.
   - The binary nature of the dependent variable causes variances that affect the accuracy of the results.(Al-Ali, 2020, p. 25)

Steps to build a logistic regression model:

1. Analyze the primary relationship:
   - Test the correlation between the binary nominal dependent variable and the independent variables using tests such as chi-square.
2. Examine the linear relationship:
   - Check for a linear relationship between the natural logarithm of the dependent variable and the independent variables.
3. Check for overlap:
   - Evaluate the linear relationship between the independent variables.
4. Adding and Deleting Variables:
   - Optimize the model by adding or deleting variables with limited impact.

Assessing the quality of the model:

The R^2 coefficient of determination is used to assess the quality of linear models, while statistics such as (R^2_ Cox & Snell) (R^2 Nagelkerke) are used in logistic regression to achieve the same goal(Al-Ali, 2020, p. 25)

The most important statistical tests in logistic regression:

1. Wald Test:
   - It is used to assess the significance of the model coefficients.
   - The probability value of the Wald statistic is compared to a predefined level of statistical significance.
2. Hosmer and Lemeshow Test:
   - Evaluates how well the model fits the data.
   - It relies on the chi-square test to compare predicted and observed values.

Conditions for using logistic regression:

1. Large sample size: Ensures accuracy of results.
2. Correlation of variables: There should be a relationship between the independent variables and the dependent variable.
3. Lack of high correlation between independent variables: Avoid multicollinearity.

Importance of Logistic Regression:

Logistic regression is a powerful tool for analyzing binary data due to its flexibility and ability to make accurate predictions. It is widely used in medical, social and marketing fields to analyze the relationships between independent variables and predict the probability of events.

The model not only provides accurate results, but also helps to identify the most influential variables, which enhances the understanding of the relationships between different factors and contributes to make decisions based on strong statistical evidence.

Stepwise Logistic Regression:

Stepwise logistic regression is a statistical method used in logistic regression analysis to select a subset of the most important predictor (independent) variables for predicting a binary outcome (a variable with only two values, such as yes/no, affected/unaffected). This method aims to build a simple and interpretable model while maintaining predictive power.

This model works through an iterative process that involves adding or removing predictor variables from the model based on specific statistical criteria. There are three main types of stepwise selection methods:

- Forward Selection: The model starts with no predictor variables. In each step, the variable that improves the model the most (usually measured by a statistical test such as the Wald test or the likelihood ratio test) is added. This process continues until no other variable significantly improves the model.

- Backward Elimination: The model starts with all available predictor variables. In each step, the least statistically significant variable is removed. This process continues until only the statistically significant variables remain in the model.
- Stepwise Selection: This is a combination of forward and backward selection. It usually starts with forward selection, but after adding each variable, all variables already in the model are checked, and any variable that is no longer statistically significant is removed. This method allows the model adapt to changes in the significance of variables as other variables are added or removed.

Selection Criteria:

These methods use statistical criteria to determine whether a variable should be added or removed. These criteria include:

- P-value: Used to determine if a variable is statistically significant. A p-value less than 0.05 is usually used as a cutoff.
- Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC): These criteria are used to compare different models and determine the model that balances goodness of fit and model complexity. Lower AIC and BIC values indicate a better model.

Advantages of Stepwise Logistic Regression:

- Model Simplification: Helps build a simpler model with fewer variables, making it easier to interpret and understand.
- Avoiding Multicollinearity: Can help avoid the problem of multicollinearity between predictor variables.
- Improving Predictive Power (in some cases): In some cases, it can improve predictive power by removing variables that add noise to the model.

Disadvantages of Stepwise Logistic Regression:

- Data Dependence: The model's results are highly dependent on the data used, and the results may vary with different samples.
- Potential Loss of Important Information: May lead to the removal of important variables, especially if they are correlated with other variables in the model.
- Statistical Issues: There are some statistical issues associated with stepwise selection, such as artificially inflating the significance of some variables.

This method is used for a large number of predictor variables to identify the most important ones. It is commonly used in fields such as medicine, marketing, and social sciences (Al-Ali, 2020, p. 25)

Methodological Frameworks:

This research is based on two main approaches: descriptive and analytical:

The descriptive approach:

It aims to describe the factors affecting the incidence of diabetes based on qualitative and quantitative data. This approach helps in collecting information about various aspects related to the study, which contributes to recognizing the current reality and developing it in the future. Data from statistical reports and information about diabetes were used to identify the influencing factors. This approach relies on analyzing the likelihood ratio based on these factors while correctly classifying the cases.

Analytical approach:

The analytical approach was used to analyze the study data using a logistic regression model. This approach helps in measuring the impact of different factors on the incidence of diabetes and exploring effective statistical methods to develop a statistical model that fits the available data. The model aims to estimate the likelihood of incidence and select the best statistical model by

comparing different models. The analysis was performed using the Statistical Package for the Social Sciences (SPSS) version 27, which is a popular and powerful tool for analyzing statistical data in social and medical fields.

Data collection and analysis

- Data collection: Data were collected from a group of patients with or at risk of diabetes mellitus. The data included various demographic and health information such as age, gender, glucose and insulin levels, skin thickness, family history of diabetes, and previous pregnancies.

-Data analysis: A logistic regression model was used to analyze the effect of independent variables on the likelihood of diabetes. The independent variables include demographic and health factors, while the dependent variable represents the final status of diabetes (with or without diabetes).

Study Population

The study population consists of individuals with or at risk of diabetes and includes all age groups and both sexes. The population includes patients with specific risk factors such as abnormal blood glucose and insulin levels or a family history of diabetes.

Study Sample

The data represents a comprehensive sample of patients whose data was collected from multiple hospitals and clinics spread across different regions and geographical areas, ensuring the comprehensiveness and diversity of the results.

instrument:

- SPSS software: SPSS version 27 was used to analyze the data and apply the logistic regression model. SPSS has a user-friendly interface and advanced analysis tools that allow for statistical analyses and accurate interpretation of results.

Temporal and Spatial Limitations:

- Temporal boundary: The study period extends from the start of data collection to the present, providing a comprehensive view of the changes in diabetes rates during this period.
- Spatial boundaries: The study includes patient data from different geographical regions, ensuring a broad and comprehensive representation of the study population.

Data source: The data used in the study were obtained from Kaggle (https://www.kaggle.com), a reliable source that provides various databases for scientific research purposes.

Studied Variables:

The study included a set of variables:

- Independent variables: Age, gender, number of previous pregnancies, skin thickness, glucose levels, and insulin levels.

-Dependent variable: The final status of diabetes mellitus (diabetic/non-diabetic).

Objective of the study: The aim of the study is to analyze the factors influencing the incidence of diabetes using a logistic regression model and make recommendations based on the results to improve prevention and early diagnosis strategies.

The Applied Study and Discussions of the Results:

The applied aspect forms the basis for analyzing the factors influencing the likelihood of developing diabetes using a logistic regression model. This aspect relies on accurate data collected from reliable sources, including demographic and health variables such as age, gender, and glucose and insulin levels.

The application aims to identify the most influential factors and estimate their strength using statistical tools such as SPSS. The fit of the model is assessed through statistical tests such as the Wald test and the Hosmer-Lemeshow test.

This aspect contributes to providing a practical model for predicting the incidence of diabetes, which enhances prevention and early intervention strategies in the medical field.

This data contains the following variables:

1. Dependent Variable:

-Status of diabetes:(0) Not having diabetes (1) Infection.

2. Independent Variables:

A. Demographic variables:

- Age. - Gender. - Number of previous pregnancies (women)

B. Health variables:

- Blood glucose levels.- Insulin levels- Skin thickness- Body mass index (BMI) - Diabetes Pedigree Function.

Results and Discussions:

Table (1): Some descriptive statistical measures of the patients participating in the study

Study Variables	Mean	Standard Deviation	Minimum Value	Maximum Value
Glucose	120.89	31.973	0	199
Blood Pressure	69.11	19.356	0	122
Skin Thickness	20.54	15.952	0	99
Insulin	79.80	155.244	0	846
Body Mass Index (BMI)	31.993	7.8842	0.01	67.1
Diabetes Pedigree Function	0.47188	0.331329	0.078	2.420
Age	33.24	11.760	21	81

Source: Researcher preparation using spss27, 2025

Tables (1) present the descriptive statistics of the study sample, showing the basic characteristics of the studied variables, as the results showed the following:

- Glucose: The mean blood glucose was 120.89, with a range between 0 and 199. This data suggests that high glucose levels are associated with an increased risk of diabetes.
- Blood pressure: The average blood pressure was 69.11 mmHg, with a range of 0 to 122. Blood pressure is an important indicator because high blood pressure is associated with an increased risk of diabetes.
-Skin thickness: The average skin thickness was 20.45, with a range of 0 to 99. Studies show that increased skin thickness may be associated with an increased risk of diabetes.
- Insulin: The average insulin level was 79.80, with a range of 0 to 846. Results suggest that low levels of insulin in the body increase the likelihood of developing diabetes.
- Body Mass: The average body mass were 31.993, with a range of 0 to 67.1. These values indicate that increased body mass is associated with a higher risk of diabetes.
- Family Ancestry Function for Diabetes: The average family pedigree function was 0.47188, with a range of 0.078 to 2.420. The results suggest that higher this index increases the risk of developing diabetes.
- Age: The mean age of the participants was 33.24 years, with a range of 21 to 81 years. The sample includes a wide range of ages, with a focus on the middle and older age groups, which are at higher risk of developing diabetes.

Table (2) shows the frequencies and percentage of patients participating in the study

Variable	Measurement/Value	(%) Frequency
Pregnancies	0	(14.5) 111
	1	(17.6) 135
	2	(13.4) 103
	3	(9.8) 75
	4	(8.9) 68
	5	(7.4) 57
	6	(6.5) 50
	7	(5.9) 45
	8	(4.9) 38
	9	(3.6) 28
	10	(3.1) 24
	11	(1.4) 11
	12	(1.2) 9
	13	(1.3) 10
	14	(0.3) 2
	15	(0.1) 1
	17	(0.1) 1
	Total	(100) 768
Test Outcome	Yes	(34.9) 268
	No	(65.1) 500
	Total	(100) 768

Source: Researcher preparation using spss27, 2025

Table (2) presents the frequency and percentage distribution of patients in the study based on the number of pregnancies and test outcomes for diabetes.

- Pregnancies Distribution:
  - The majority of participants had 0 to 3 pregnancies, with the highest percentage in the 1-pregnancy category (17.6%), followed by 0 pregnancies (14.5%) and 2 pregnancies (13.4%).
  - A smaller proportion of participants had 4 to 9 pregnancies, with percentages decreasing gradually.
  - Only a few participants had 10 or more pregnancies, with the lowest frequency observed for 15 and 17 pregnancies (0.1%) each.
- Diabetes Test Outcome:
  - 34.9% (268 participants) tested positive for diabetes.
  - 65.1% (500 participants) tested negative.
  - The total sample size was 768 participants.

Table (3) shows omnibus tests of model coefficients

		Chi-square	Df	Sig.
Step 1	Step	270.039	8	.000
	Block	270.039	8	.000
	Model	270.039	8	.000

Source: Researcher preparation using spss27, 2025

Table (3) shows the results of the Chi-square statistical test used to assess the fit of the model as a whole. The results showed that the Chi-square statistic value was 270.039 with 8 degrees of freedom, and the probability value ($p - value$) was less than 0.001, which is less than the significance level of 0.05.

This result reflects the high statistical significance of the general model, which means that the independent variables have a significant effect on the dependent variable (diabetes). In other words, the developed model relies well on the independent variables to improve the predictive power of patients' health status.

Table (4) injects the summary of the model using the full model method

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	723.445[a]	.296	.408

Source: Researcher preparation using spss27, 2025

Table (4) indicates a set of metrics and tests that evaluate the quality of the model used to explain the relationship between the variables. Among the most notable results, we note that the value of -2 Log Likelihood decreased to 723.445 compared to the initial model value of 993.550. The programme also provides two values for the coefficient of determination R square (the square of the correlation coefficient) to estimate the amount of change in the dependent variable explained by the model. The Nagelkerke R square value is about 0.408, indicating that the model explains about 40.8% of the variance in the dependent variable. This result indicates that the model has a moderate level of explanation for the relationship between the independent variables (e.g. glucose, insulin, and age) and the incidence of diabetes. This reflects the model's ability to provide an acceptable explanation of the relationships between these important health factors.

Table (5) shows the Hosmer and lemeshow test using the full model method

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

Source: Researcher preparation using spss27, 2025

Table (5) indicates the results of the Hosmer & Lemeshow test for goodness of fit. The key point of this test is the value of the test statistic itself, which are 8.323 with a statistical significance value of 0.403.

This test aims to check whether the observed data is significantly different from the predicted values generated by the model. For a well-fitting model, the non-significant statistical significance value must be greater than 0.05.

In this case, the result indicates that the significance value is greater than 0.05, which means that the differences between the observed and predicted values are not significant. Therefore, the model can be considered suitable to represent the data well.

Table (6) shows the classification of cases using the full model method:

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

Source: Researcher preparation using spss27, 2025

Table (6) Testing the classification table where the classification table indicates the extent to which the model predicts the cluster items, and gives a classification of the expected values calculated on the basis of the final model estimated for the data

The model correctly classifies 78.3% of cases, which means that the model is relatively accurate in predicting the incidence of diabetes. The correct classification ratio for unaffected people is 89.0, while the correct classification ratio for affected people is 58.2, which means that the model is able to predict with high accuracy for unaffected people compared to affected people.

Table (7) shows the variables in the replay using the full form method

		B	S.E.	Wald	Df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1	Pregnancies	.123	.032	14.747	1	.000	1.131	1.062	1.204
	Glucose	.035	.004	89.897	1	.000	1.036	1.028	1.043
	Blood Pressure	-.013	.005	6.454	1	.011	.987	.977	.997
	Skin Thickness	.001	.007	.008	1	.929	1.001	.987	1.014
	Insulin	-.001	.001	1.749	1	.186	.999	.997	1.001
	BMI	.090	.015	35.347	1	.000	1.094	1.062	1.127
	Diabetes Pedigree Function	.945	.299	9.983	1	.002	2.573	1.432	4.625
	Age	.015	.009	2.537	1	.111	1.015	.997	1.034
	Constant	-8.405	.717	137.546	1	.000	.000		

Source: Researcher preparation using spss27, 2025

Table (7) presents the logistic regression results using the full model method. It provides estimates for the impact of different independent variables on the likelihood of developing diabetes.

- Significant Predictors:

- Pregnancies (B = 0.123, p < 0.001): A higher number of pregnancies increase the likelihood of diabetes.

- Glucose (B = 0.035, p < 0.001): Higher glucose levels are strongly associated with increased diabetes risk.
- Blood Pressure (B = -0.013, p = 0.011): Slightly lower blood pressure reduces the risk.
- BMI (B = 0.090, p < 0.001): Increased BMI significantly raises the risk.
- Diabetes Pedigree Function (B = 0.945, p = 0.002): A strong genetic predisposition to diabetes increases the likelihood of developing the disease.

- Non-Significant Predictors:
- Skin Thickness (p = 0.929), Insulin (p = 0.186), and Age (p = 0.111) did not show statistically significant effects.

Model Fit:
- The constant term (B = -8.405, p < 0.001) indicates the baseline probability of diabetes when all predictors are zero.

Table (8) shows the summary of the model using the gradient model method:

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	808.720	.214	.295
2	771.403	.251	.346
3	744.125	.277	.382
4	734.306	.286	.395
5	728.560	.292	.402

Source: Researcher preparation using spss27, 2025

Table (8) presents the model summary using the gradient model method across multiple steps.
- The -2 Log Likelihood value decreases from 808.720 to 728.560, indicating an improved model fit with each step.
- The Cox & Snell R Square increases from 0.214 to 0.292, suggesting a gradual improvement in the model's explanatory power.
- The Nagelkerke R Square also increases from 0.295 to 0.402, showing an enhanced ability of the model to explain the variation in the dependent variable.

Table (9) shows the Hosmer and Lemeshow test using the gradient model method

Step	Chi-square	df	Sig.
1	2.697	8	.952
2	11.871	8	.157
3	11.152	8	.193
4	10.180	8	.253
5	8.128	8	.421

Source: Researcher preparation using spss27, 2025

Table (9) presents the Hosmer and Lemeshow test results for evaluating the goodness-of-fit of the gradient model method across multiple steps.
- The Chi-square values range from 2.697 to 11.871, indicating variations in model fit across different steps.
- The degrees of freedom (df) remain constant at 8 throughout all steps.
- The significance values (Sig.) are all greater than 0.05, ranging from 0.157 to 0.952, indicating that the model does not significantly differ from the observed data, suggesting a good fit.

Table (10) shows the classification of cases using the graded model method

	Observed		Predicted		
			Outcome		Percentage Correct
			0	1	
Step 1	Outcome	0	443	57	88.6
		1	138	130	48.5
	Overall Percentage				74.6
Step 2	Outcome	0	445	55	89.0
		1	126	142	53.0
	Overall Percentage				76.4
Step 3	Outcome	0	437	63	87.4
		1	116	152	56.7
	Overall Percentage				76.7
Step 4	Outcome	0	442	58	88.4
		1	117	151	56.3
	Overall Percentage				77.2
Step 5	Outcome	0	441	59	88.2
		1	114	154	57.5
	Overall Percentage				77.5

Source: Researcher preparation using spss27, 2025

Table (10) presents the classification accuracy of cases using the graded model method across multiple steps. The table shows how well the model predicts the actual outcomes (0 or 1) at each step and the overall percentage of correct classifications.

- Step 1: The model achieved 88.6% accuracy in classifying non-diabetic cases (0) but only 48.5% for diabetic cases (1), resulting in an overall accuracy of 74.6%.
- Step 2: Accuracy improved slightly, with 89.0% for non-diabetic cases and 53.0% for diabetic cases, increasing the overall accuracy to 76.4%.
- Step 3: A minor decline was observed in non-diabetic classification (87.4%), but the diabetic classification improved to 56.7%, leading to an overall accuracy of 76.7%.
- Step 4: The accuracy for non-diabetic cases rose to 88.4%, while diabetic classification slightly decreased to 56.3%, increasing the overall accuracy to 77.2%.
- Step 5: The final model maintained 88.2% accuracy for non-diabetic cases and improved diabetic classification to 57.5%, achieving the highest overall accuracy of 77.5%.

Table (11) shows the variables in the equation using the gradient model method:

		B	S.E.	Wald	Df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1	Glucose	.038	.003	135.649	1	.00	1.039	1.032	1.045
	Constant	-5.35	.421	161.624	1	.00	.005		
Step 2	Glucose	.035	.003	114.341	1	.00	1.036	1.029	1.042
	BMI	.076	.013	32.747	1	.00	1.079	1.051	1.108
	Constant	-7.51	.605	154.167	1	.00	.001		
Step 3	Pregnancies	.137	.027	26.230	1	.00	1.147	1.088	1.209
	Glucose	.034	.003	106.418	1	.00	1.035	1.028	1.041
	BMI	.082	.014	35.249	1	.00	1.085	1.056	1.115
	Constant	-8.12	.638	161.897	1	.00	.00		
Step 4	Pregnancies	.142	.027	27.417	1	.00	1.152	1.093	1.215
	Glucose	.034	.003	102.246	1	.00	1.034	1.028	1.041
	BMI	.078	.014	32.162	1	.00	1.081	1.052	1.111
	Diabetes Pedigree Function	.901	.292	9.547	1	.002	2.463	1.390	4.362
	Constant	-8.41	.657	164.130	1	.00	.00		
Step 5	Pregnancies	.153	.028	30.408	1	.00	1.166	1.104	1.231
	Glucose	.035	.003	104.305	1	.00	1.035	1.028	1.042
	Blood Pressure	-.012	.005	5.697	1	.017	.988	.978	.998
	BMI	.085	.014	36.071	1	.00	1.089	1.059	1.119
	Diabetes Pedigree Function	.911	.294	9.592	1	.002	2.486	1.397	4.423
	Constant	-7.95	.676	138.551	1	.00	.00		

Source: Researcher preparation using spss27, 2025

Table (11) presents the coefficient estimates for the independent variables included in the model. This section of the results displays the values associated with the fixed term and the independent variables, providing information about the effect of each independent variable on the dependent variable.

B value: The B value in logistic regression plays the same role as it does in linear regression. This value is used in the logistic regression equation to calculate the likelihood of a particular situation occurring within a specific category. In logistic analysis, this coefficient is interpreted as the amount of change in the natural logarithm $logit$ of the dependent variable (outcome), resulting from a one-unit change in the independent variable (predictor). Logit is defined as the natural logarithm of the likelihood (probability) of outcome Y occurring

Critical statistic (Wald): The Wald statistic, which follows a chi-square distribution, is used to test whether the value of the independent variable's coefficient B is significantly different from zero. If this value is significantly different from zero, it indicates that the independent variable makes a significant contribution in predicting the value of the dependent variable Y.

$Exp(B)$test: represents$Exp(B)$ is the change in likelihood resulting from a one-unit change in the independent variable. $Exp(B)$ is easier to interpret than the B coefficient because it does not require a logarithmic transformation.

Confidence Interval for$Exp(B)$: The confidence interval for $Exp(B)$ indicates the range of possible values that $Exp(B)$ in the population based on the sample. For example, if the confidence interval for $Exp(B)$ is between 1.062 and 1.204, this means that in 95% of trials, the calculated confidence intervals will contain the true value of $Exp(B)$ in the population. However, there is a 5% chance that the calculated confidence interval for a particular sample will not include the true value.

This analysis helps in understanding the contribution of each independent variable to the model and how it affects the dependent variable, enabling accurate conclusions to be drawn about the influencing factors. From the data in the final figure, the logistic regression model fitted to the data can be derived as follows:

$$Odds = \log\left(\frac{p}{1-p}\right)$$

$$= -7.955 - .153 * \mathrm{X}_1 + .035 * \mathrm{X}_2 + -.012 * \mathrm{X}_3 + .085 * \mathrm{X}_6 + .911 * \mathrm{X}_7$$

The load: (X_1): 1.166 =$Exp(b)$ indicates that pregnant women are more likely to develop type 2 diabetes later in life, indicating that the effect of pregnancy on diabetes is strong and significant ($P - value = 0$).

Glucose: (X_2) value of 1.035 = $Exp(B)$ indicating that the effect of glucose on diabetesis strong and significant($P - value = 0$).

Blood Pressure: (X_3) $Exp(B) = .988$ which means that for every 98 mmHg increase in blood pressure. A 98 mmHg increase in blood pressure increases the likelihood of developing diabetes by 98%, which is a significant effect($P - value = 0.017$).

Average body mass: (X_6)$Exp(B) = 1.089$, indicating that an increase in body mass increases the risk of diabetes by 8.9 per cent, which is a strong and significant effect ($\mathrm{P} - \mathrm{value} = 0$)

(X_7) Exp(B) $= 2.486$, indicating that the likelihood of having diabetes in a family member increases the incidence of diabetes by about 2.57, which is a significant effect ($\mathrm{P} - \mathrm{value} = 0.002$).

Table (12) Comparison between the full model and the tiered model using the prediction and interpretation criteria

Test/Measure	Full Model	Stepwise Model	Best Model
-2 Log-Likelihood	723.445	728.56	Full Model
Nagelkerke R Square	0.408	0.402	Full Model
Cox & Snell R Square	0.296	0.292	Full Model
Hosmer-Lemeshow Test (p-value)	0.403	0.421	Stepwise Model
Omnibus Test of Model Coefficients (Chi-Square)	270.039	264.92	Full Model
Classification Accuracy (%)	78.3	77.5	Full Model
Number of Independent Variables	8	5	Stepwise Model

Source: Researcher preparation using spss27, 2025

Based on Table (12), the performance and interpretation of both the full model and the tiered model can be analyzed based on a set of statistical tests and measures. Below is the comparative analysis:

Log Likelihood the full model - 723.445 while the scaled model - 728.560 the full model is better here, because a lower log-likelihood value indicates a better fit between the model and the data

Nagelkerke R Square - full model - .408, while the scaled model = .402 The full model explains a greater proportion of the variance in the dependent variable, making it better in terms of explaining the relationship between the independent variables and the dependent variable

Cox &Snell R Square for the full model is .296, while the scaled model - .292 Again, the full model explains a greater proportion of the variance in the dependent variable according to this measure, making it the best in this category.

Hesmer and Lemeshow Test, the full model is $(P - value = 0.403)$. While the scaled model is.421 $\mathrm{P} - \mathrm{value}$. Both models are appropriate, as the probability values are greater than 0.5, but the stepwise model has a higher$\mathrm{P} - \mathrm{value}$, meaning it better represents the data

Omnibus Tests of Model Coefficients Omnibus Tests of Model Coefficients Full Model $\mathrm{Chi} - \mathrm{square} = 270.039$ while the scaled model $\mathrm{Chi} - \mathrm{square} = 264.924$ - the full model shows greater statistical significance, meaning that the independent variables in the full model have a greater influence on the dependent variable.

Classification Accuracy The full model - 78.3%, while the graded model - 77.5% The full model offers better classification accuracy, although the difference between the two models is not significant.

Independent Variables the full model - 7 variables, while the stepwise model - 5 variables the stepwise model is simpler and uses fewer variables, making it easier to interpret and less complex.

The full model outperforms in most of the measures that relate to the quality of explanation and fit to the data. These indicators show that the full model offers greater explanation and better overall performance in terms of data fit. On the other hand, the stepwise model offers a higher $\mathrm{P} - \mathrm{value}$ (.421) and (.403) and uses fewer variables (5 vs. 8) making it simpler and easier to interpret. If the priority is to fully and comprehensively interpret the data, the full model is preferable, as it provides a more accurate and comprehensive explanation. If the priority is simplicity and predictive efficiency, the stepwise model performs well with fewer variables

Conclusions:

Based on the analysis using the binary logistic regression model, the study came up with a set of important conclusions that explain the factors influencing the likelihood of developing diabetes:

1. **The effect of pregnancy**:

The results showed that pregnant women are more likely to develop type 2 diabetes, as a result of hormonal changes and increased insulin resistance during pregnancy, confirming that pregnancy increases the likelihood of developing the disease.

2. Glucose and obesity:

High glucose levels and average body mass (obesity) are two of the main factors that increase the risk of diabetes. The results suggest that higher glucose levels and higher obesity are strongly associated with an increased likelihood of developing the disease.

3. High blood pressure:

Findings revealed that individuals with high blood pressure are more likely to develop diabetes, emphasizing the influence of this factor on the risk of developing the disease.

4. Family history of diabetes:

Having a family history of diabetes is an influential genetic factor, as individuals who have relatives with diabetes are more likely to develop the disease.

5. Effectiveness of the predictive model:

The logistic regression model proved to be highly effective in predicting the incidence of diabetes, as the prediction accuracy of the full model was 78.3%. This reinforces the use of the model as a reliable analytical tool for diagnosis.

6. Hosmer-Lemeshow test:

Hosmer-Lemeshow tests showed good agreement between the model and the data, increasing confidence in the accuracy of the results from the analysis.

7. The stepwise model:

Use of the stepwise model led to a greater explanation of variance using fewer independent variables, making it simpler and more accurate in prediction, and providing a more effective practical application.

8. Additional effects:

Some other variables, such as skin thickness, insulin levels, and age, have an additional effect on the likelihood of developing diabetes. Insulin, as a key hormone for regulating blood sugar, is an important factor, as any disturbance in its levels can lead to the development of the disease.

These findings demonstrate the importance of using logistic regression to analyze factors influencing diabetes, and highlight the importance of identifying the most influential factors to improve prevention and diagnostic strategies.

Recommendations and Suggestions:

Based on the results of the study, the following recommendations and suggestions can be made to strengthen prevention, diagnosis, and treatment strategies for diabetes:

1. Intensifying examinations during pregnancy:

Regular checks for blood sugar levels for pregnant women, along with follow-up for other health factors such as high blood pressure, are recommended for early detection of diabetes and reduce the risk of infection.

2. Eat a healthy diet:

It is recommended to encourage individuals to adopt a healthy diet and maintain an ideal weight, as obesity is one of the main factors that increase the risk of diabetes.

3. Use the graduated model:

The tiered model is recommended for medical analysis, due to its high performance and accuracy. This model can be an effective tool to support physicians in making decisions based on reliable data to improve the diagnosis of the disease.

4. Take early medical measures:

It is recommended to monitor and treat individuals with high blood pressure, cholesterol or have abnormal blood sugar levels immediately, which contributes to reducing the progression and complications of the disease.

5. Conduct more research on the factors at play:

A deeper study of other genetic and environmental factors such as diet, lack of physical activity, and smoking is recommended to improve the accuracy of predictive models and broaden understanding of the causes of diabetes.

6. Expansion of the scope of the study:

It is proposed to include larger, more diverse population samples covering different age groups, to ensure more comprehensive and accurate results.

7. Use advanced analytical techniques:

Modern technologies such as artificial neural networks and random forests are recommended to analyze data on factors affecting diabetes, improving the accuracy of predictions and enhancing analysis.

8. Conduct long-term follow-up studies:

It is suggested to follow up individuals diagnosed using the current model to assess the accuracy of predictions over a longer period of time, and work to improve the efficiency of the model based on those assessments.

These recommendations contribute to strengthening preventive and curative interventions, and developing more accurate analytical models that help effectively address the challenges of diabetes.

References:

Abdel Moneim, T. M. (2011) Statistical analysis of multiple variables. Anglo-Egyptian Library.

Al-Ali, I. M. (2020) Foundations of multivariate statistical analysis. Faculty of Economics, Tishreen University, Syria.

Al-Bermani, Z. A. A. and Ismail, A. A. K. (2021) ‘Using Logistic Regression to Study the Main Factors Affecting Diabetes in Elderly People in the City of Hilla’, Journal of Physics: Conference Series, 1818(1).

American Diabetes Association (ADA). (n.d.) Causes of Diabetes. Available at: [https://www.diabetes.org].

Azzam, A. H. (1997) Statistical analysis of multiple variables from an applied point of view. (Translated by Richard Johnson). Dar Al-Mareh Publishing.

International Journal of Environmental Research and Public Health. (2021) ‘Predicting type II diabetes using logistic regression’, International Journal of Environmental Research and Public Health, 18(14), p. 7346.

Joshi, R. D. and Dhakal, C. K. (2021) ‘Predicting Type 2 Diabetes Using Logistic Regression and Decision Tree Algorithms’, International Journal of Environmental Research and Public Health, 18(14), p. 7346.

Poussier, R., Zhou, Y., and Standaert, F.-X. (2019) 'Enhancing Logistic Regression Models for Diabetes Prediction Using PCA and K-Means Techniques', Informatics in Medicine Unlocked, 17, p. 100179.

Rajendra, P. and Latifi, S. (2023) 'Application of Logistic Regression in Early Diabetes Prediction across the United States', Operations Density, Science Fiction, and Decision Management Journal, 6(5), pp. 34-48.

World Health Organization (WHO). (n.d.) Diabetes: Causes and Risk Factors. Available at: https://www.who.int](https://www.who.int.