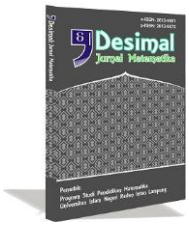
Desimal: Jurnal Matematika Vol 8 No 1 (2025) 41-50

Contents lists available at DJM

DESIMAL: JURNAL MATEMATIKA

p-ISSN: 2613-9073 (print), e-ISSN: 2613-9081 (online), DOI 10.24042/djm

http://ejournal.radenintan.ac.id/index.php/desimal/index



Diabetes risk prediction using logistic regression model

Linda Rassiyanti[1,*], Fajri Farid[1], Rizka Pitri[2]

[1] Institut Teknologi Sumatera, Indonesia
[2] UIN Raden Intan Lampung, Indonesia

ARTICLE INFO

Article History
Received : 04-03-2025
Revised : 08-03-2025
Accepted : 09-04-2025
Published : 30-04-2025

Keywords:
Diabetes; Cook's Distance; Regresi Logistik; ROC-AUC.

*Correspondence: E-mail:
linda.rassiyanti@sd.itera.ac.id

Doi:
10.24042/djm.v8i1.26493

ABSTRACT

This study aims to analyze the factors that contribute to diabetes using the logistic regression method. The data used in this study include variables of number of pregnancies, glucose levels, blood pressure, skin thickness, insulin levels, body mass index (BMI), family history of diabetes, and age. The logistic regression model was applied to determine the effect of each variable on the likelihood of a person having diabetes. Evaluation of model performance was carried out using the ROC (Receiver Operating Characteristic) curve, and the results obtained showed an AUC value of 0.8391, which indicated a very good classification ability of the model. The results of the analysis showed that the number of pregnancies, glucose levels, blood pressure, BMI, and family history of diabetes had a significant effect on the risk of diabetes.

http://ejournal.radenintan.ac.id/index.php/desimal/index

INTRODUCTION

According to the World Health Organization (2023), diabetes is one of the non-communicable diseases (NCD) that is the leading cause of death in the world. This disease can cause various serious complications, such as cardiovascular disease, nephropathy (kidney failure), neuropathy (nerve disorders), and retinopathy, which can lead to blindness. Currently, nearly half a billion people in the world are living with diabetes, and this number is projected to increase by 25% by 2030 and 51% by 2045 (Cho et al., 2018). The NCD Risk Factor Collaboration (2022) reports that the prevalence of diabetes is increasing globally, with major risk factors including unhealthy diet, lack of physical activity, and increasing obesity rates in various countries. Various studies have shown that the main risk factors for type 2 diabetes mellitus include obesity, hypertension, family history, food consumption patterns, and physical activity (Putri & Kismiantini, 2024).

Early identification of diabetes risk factors is very important so that preventive and treatment measures can be carried out more effectively. Research on diabetes risk factors has developed rapidly, especially using statistical

methods to analyze the relationship between variables that affect diabetes. One method that is often used in epidemiological research is logistic regression, which is able to model the relationship between risk factors and the likelihood of a person experiencing diabetes (Yasmin, 2023). This method is very useful in estimating the probability of a person suffering from diabetes based on various factors, such as the number of pregnancies, glucose levels, blood pressure, skin thickness, insulin levels, body mass index (BMI), diabetes pedigree function, and age. Several comparative studies have shown that logistic regression has competitive accuracy compared to machine learning-based methods in early diabetes prediction (Anthony & Eloy, 2023).

Previous research conducted by Smith et al. (2020) revealed that blood glucose levels and body mass index have a significant effect on the risk of diabetes, with a high odds ratio value. In addition, other studies have also shown that blood glucose levels and BMI play an important role in determining the risk of diabetes (Sisodia & Sisodia, 2018). Meanwhile, a study by Johnson & Lee, 2021) compared several machine learning methods and found that logistic regression remains one of the effective methods in predicting diabetes, especially because of its ease of interpretation of the results.

This study used a diabetes dataset containing patients' medical and lifestyle information to analyze factors that significantly contribute to the likelihood of a person being diagnosed with diabetes. The logistic regression model was evaluated using various metrics, such as Pseudo R-Square to measure the model's fit, Odds Ratio to assess the impact of each variable on the risk of diabetes, and ROC-AUC curve to assess the model's ability to distinguish healthy and at-risk individuals.

This study aims to identify the main risk factors that contribute to the possibility of someone suffering from diabetes and assess the effectiveness of the logistic regression model in predicting this disease. The results of the study are expected to provide deeper insight for medical personnel in developing strategies for preventing and managing diabetes more optimally.

METHOD

This study uses secondary data in the form of a 2024 diabetes dataset obtained from the National Institute of Diabetes and Digestive and Kidney Diseases link. The diabetes dataset used is data on the number of pregnancies (X1), glucose levels (X2), blood pressure (X3), skin thickness (X4), insulin levels (X5), BMI (X6), age (X7), history of diabetes (X8), and diabetes diagnosis (Y). The diabetes diagnosis variable is in the form of binary data in the form of yes and no. This study was conducted using a quantitative research approach with logistic regression analysis. The stages used in this study are:

a. Downloading Data

The authors searched for and downloaded data on the number of pregnancies (X1), glucose levels (X2), blood pressure (X3), skin thickness (X4), insulin levels (X5), BMI (X6), age (X7), history of diabetes (X8), and diagnosis of diabetes (Y) from the National Institute of Diabetes and Digestive and Kidney Diseases website.

b. Data Preprocessing

After all the data is downloaded, preprocessing is performed. Data preprocessing is a crucial stage in data analysis that includes data cleaning, outlier detection, and filling in missing values using various techniques, including mean imputation (García, Luengo, & Herrera, 2015).

c. Data Exploration

After the data obtained is clean from empty data and outliers, data exploration

is carried out so that the data analysis that will be presented will be clearer and can be easily understood by the reader (Hidayat, Tolago, Dako, & Ilham, 2023).

d. Performing Logistic Regression Analysis

According to Asyiah (2008), logistic regression is often used in health research because of its ability to model the relationship between binary dependent variables and several independent variables. Several previous studies have shown that logistic regression is an effective method in analyzing diabetes risk factors because of its ability to interpret the relationship between variables (Li, Wang, & Zhao, 2023). In addition, this method is often compared to machine learning algorithms such as Random Forest and Deep Learning, which, although more complex, still face challenges in the interpretability of the results (Alghamdi, Alzahrani, & Alharthi, 2023).

The logistic regression model is based on the logit function, which relates the probability p (of an event) to the independent variable X, which is expressed as follows (Kupper, Hosmer, & Lemeshow, 1990):

$$logit(p) = log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Thus, the probability of an event can be calculated using the sigmoid function (Kleinbaum & Klein, 2010):

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

Agresti (2018) also discusses the maximum likelihood method as the main approach to estimating the parameters of the logistic regression model. This approach aims to find the parameter value that maximizes the likelihood function, thus producing the most likely estimate describing the relationship between the dependent and independent variables. The likelihood function is defined as (Menard, 2002):

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{(1-y_i)}$$

Where y_i is the value of the dependent variable and p_i is the probability of the event based on the logistic regression model..

Odds Ratio (OR) is a measure used to assess the strength of the relationship between independent variables and binary dependent variables. OR shows the ratio of the odds of an event occurring in the exposed group compared to the unexposed group (Agresti, 2018). Odds Ratio is calculated as:

$$OR = e^{\beta}$$

Interpretation of OR is very important in understanding the impact of each independent variable in the logistic regression model. Several main assumptions in logistic regression need to be considered to ensure the validity and reliability of the results (Pampel, 2000):

1) The dependent variable is binary.
2) The relationship between the independent variable and the logarithm of the odds is linear.
3) Observations must be independent of each other.
4) There is no multicollinearity.
5) There are no extreme outliers.
6) Sufficiently large sample size.

Pseudo R^2 (Nagelkerke R^2 and Cox & Snell R^2) is used to measure the extent to which the independent variables in the model are able to explain the variability in the data (Pampel, 2000).

e. Conduct Parameter Tests Simultaneously

This stage is carried out to determine the overall influence of

independent variables, namely data on the number of pregnancies (X1), glucose levels (X2), blood pressure (X3), skin thickness (X4), insulin levels (X5), BMI (X6), age (X7), and history of diabetes (X8), on the diagnosis of diabetes (Y).

f. Perform Partial Parameter Testing

This stage is carried out if the test results show real significance, which indicates the influence of the independent variables as a whole on the diagnosis of diabetes (Y). Furthermore, partial parameter testing is applied to identify independent variables that have a significant influence on the diagnosis of diabetes (Y) (Susanti, Anugrawati, Fitrah, Usman, & Yusrianto, 2023).

g. Conduct a Logistic Regression Model Evaluation

This stage aims to evaluate the performance of the logistic regression model and whether there is a difference between the prediction and observation values (Parsaulian, Tarno, & Ispriyanti, 2021).

According to Agresti (2018), there are several methods used to evaluate logistic regression models as follows:

1. The Hosmer-Lemeshow Test is used to test the model's fit to the data and determine whether the model is able to describe the relationship between independent and dependent variables well (Kupper et al., 1990).
2. Area Under the Curve (AUC-ROC Curve), used to assess the model's predictive ability to distinguish between categories of the dependent variable (Menard, 2002).

The research flow can be seen in Figure 1.

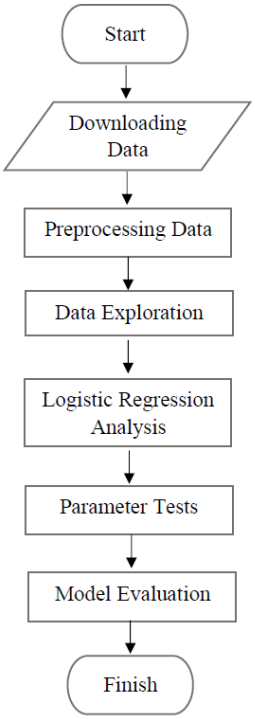


Figure 1. Research Flow

RESULTS AND DISCUSSION

Descriptive analysis in this study aims to see the characteristics and distribution of data that will be processed using logistic regression analysis.

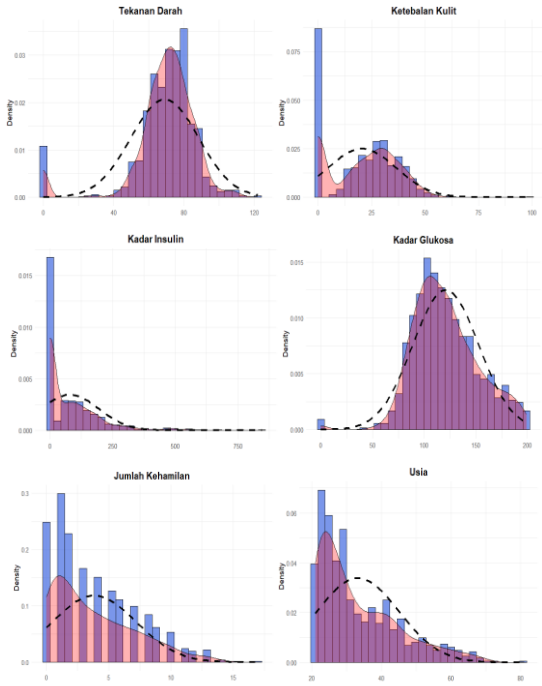


Figure 2. Distribution of Each Variable

Based on Figure 2, it can be seen that the distribution of glucose levels and blood pressure is close to normal but slightly skewed to the right. This indicates that there are individuals with higher glucose and blood pressure levels than the population average. The distribution of skin thickness, insulin levels, and number of pregnancies is skewed to the right, indicating that most individuals have low values for these variables, but there are a few individuals with much higher values. The distribution of age is also skewed to the right, indicating that the majority of individuals in this dataset are young, with fewer individuals in older ages.

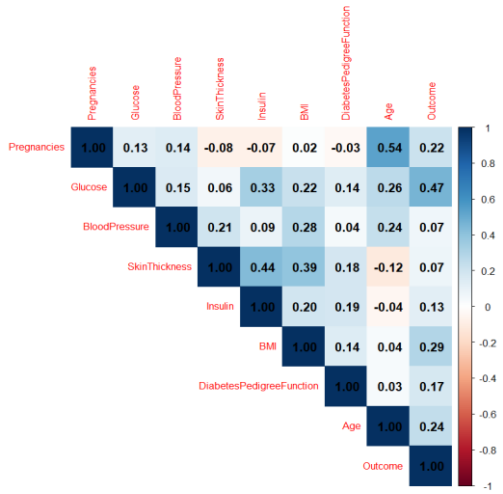


Figure 3. Correlation

Based on Figure 3, glucose levels have the highest correlation with diabetes diagnosis (0.47), indicating that the higher the glucose level, the greater the risk of diabetes. BMI (0.29) and age (0.24) are also positively correlated, indicating that higher BMI and older age increase the risk. Insulin levels and skin thickness are quite strongly correlated (0.44), indicating a relationship between fat storage and insulin production. Blood pressure and other variables have low correlations with diabetes. Overall, glucose levels are the main factor, followed by BMI and age, while other variables contribute less.

Based on the analysis results, glucose levels, BMI, number of pregnancies, and family history have a significant effect on the diagnosis of diabetes. This is in line with research by Garcia & Hernandez (2023), which found that a combination of metabolic and genetic factors can increase the accuracy of diabetes risk prediction. The logistic regression model obtained is as follows:

$$log\left(\frac{P}{1-P}\right) = -8.428 + 0.118X_1 + 0.035X_2 - 0.013X_3 + 0.0001X_4 - 0.001X_5 + 0.090X_6 + 0.016X_7 + 0.932X_8$$

where Y is the probability of a person being diagnosed with diabetes, X_1 is the number of pregnancies, X_2 is glucose level, X_3 is blood pressure, X_4 is skin thickness, X_5 is insulin level, X_6 is body mass index (BMI), X_7 is age, and X_8 is history of diabetes.

Table 1. Logistic Regression Analysis Results

Variable	Odds Value	*P-value*	Significance
Number of Pregnancies	1.125	0.000	Significant
Glucose Level	1.036	0.000	Significant
Blood Pressure	0.987	0.012	Significant
Skin Thickness	1.000	0.987	Not Significant
Insulin Level	0.999	0.186	Not Significant
BMI	1.095	0.000	Significant
History of Diabetes	2.540	0.002	Significant
Age	1.016	0.078	Not Significant

Based on the obtained model, the number of pregnancies, glucose levels, BMI, and family history are the main factors in predicting diabetes, in line with the research of Smith et al. (2020) and Sisodia & Sisodia (2018). Table 1 shows that family history has an OR of 2.540, meaning that individuals with a family history of diabetes have a 2.54 times higher risk of developing diabetes. Harrison et al. (2017) also emphasized

that family history plays a significant role in increasing the risk of this disease.

The number of pregnancies with an OR of 1.125 increased the odds of diabetes by 12.5% for each additional pregnancy. Glucose levels (OR = 1.036) and BMI (OR = 1.094) also significantly increased the risk. In contrast, blood pressure and insulin levels had ORs lower than 1, which indicates that increasing these variables slightly decreased the risk in this model.

Generally, these findings confirm that metabolic and genetic factors, particularly glucose levels, BMI, and family history, play a major role in predicting diabetes, as demonstrated in previous studies.

After obtaining the logistic regression model, the next step is to test the hypothesis of the model. Based on Table 1, it can be seen that the VIF value in all variables is lower than 11, which indicates that there is no multicollinearity.

Table 2. VIF Value

Variable	VIF Value
Number of Pregnancies	1.392
Glucose Level	1.211
Blood Pressure	1.174
Skin Thickness	1.524
Insulin Level	1.470
BMI	1.220
History of Diabetes	1.034
Age	1.490

Cook's Distance is used to identify observations that have a large influence on the model. Based on Figure 4, some points have higher Cook's Distance values than others, such as observations 10, 229, and 707. Points with high Cook's Distance have the potential to be influential outliers, which can significantly change the regression coefficients if removed or modified. However, the Cook's Distance value in this plot is still relatively low (lower than 0.06), indicating that although there are some influential points, they are not too extreme.

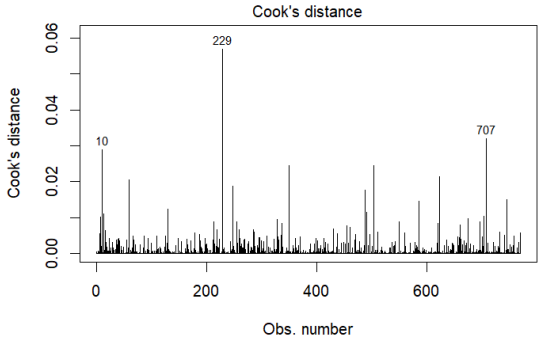


Figure 4. Plot Cook's Distance

The Hosmer-Lemeshow test and the area under the curve (AUC-ROC Curve) value were used in this study to evaluate the obtained model. According to Kupper et al. (1990), the interpretation of the AUC value is 0.5 - 0.6 (bad model), 0.6 - 0.7 (fairly good model), 0.7 - 0.8 (good model), 0.8 - 0.9 (very good model), and greater than 0.9 (very accurate model). Based on the analysis obtained, the area under the curve value is 0.8391, which means that the model obtained in this study is very good. This is also supported by the Hosmer-Lemeshow test value, which has a p-value of 0.1971, which means that the model is considered to fit with the data, namely, the prediction of diabetes diagnosis produced in this study produces predictions that are in accordance with the movement of actual data and a small error rate in predicting future diabetes diagnosis. Based on the results of the model evaluation using the Hosmer-Lemeshow test, in the future, the factors of number of pregnancies, glucose levels, blood pressure, BMI, and history of diabetes will be the main focus in predicting a person's diabetes diagnosis.

Johnson & Lee (2021) compared various machine learning methods, such as Random Forest and Artificial Neural Network (ANN), and found that the ANN model had the highest accuracy (AUC = 0.87). However, they highlighted that although machine learning provides more accurate results, models such as ANN are difficult for medical professionals to interpret. In this context, this study

confirms that logistic regression remains a competitive method, with an AUC of 0.8391, indicating that the model can predict diabetes very well while providing clearer interpretations than machine learning-based models. Wang et al. (2023) also confirmed that logistic regression remains a competitive method in diabetes prediction, especially when combined with other statistical methods.

In addition, the study found that the number of pregnancies and blood pressure were also significant factors in predicting diabetes, which has not always been the focus of previous studies. Most previous studies have focused more on glucose levels and BMI, while this study shows that women with more pregnancies have a higher risk of developing diabetes. This suggests that hormonal and metabolic aspects during pregnancy may play a role in increasing the risk of diabetes later in life.

This study also found that blood pressure has an effect on the risk of diabetes, although the correlation is lower than glucose levels and BMI. These results provide new insights that blood pressure control can also be part of a diabetes prevention strategy, not just focusing on glucose levels alone. Thus, this study contributes by identifying broader risk factors and confirms that logistic regression is still a powerful tool in diabetes risk analysis with a balance between accuracy and clear interpretation.

CONCLUSIONS AND SUGGESTIONS

Based on the results of the analysis in this study, it can be concluded that the variables of the number of pregnancies, glucose levels, blood pressure, body mass index (BMI), and family history of diabetes have a significant effect on a person's diabetes diagnosis. In contrast, the variables of skin thickness, insulin levels, and age did not show a significant effect in determining the diagnosis of diabetes. The logistic regression model developed in this study was proven to have good performance with an area under the curve (AUC) value of 0.8391, so it can be used as a fairly accurate prediction tool in detecting the possibility of someone suffering from diabetes.

Based on the findings of this study, it is recommended that factors that significantly influence diabetes be the main focus to prevent and control this disease, especially glucose levels and body mass index, which can be managed through a healthy diet and active lifestyle. In addition, further research can develop a prediction model by considering other variables such as physical activity levels, diet, and more specific genetic factors. The use of more complex analysis methods, such as machine learning, can also be considered to improve the accuracy of diabetes diagnosis predictions.

REFERENCES

Agresti, A. (2018). *An introduction to categorical data analysis* (3rd ed.). Canada: John Wiley & Sons, Ltd.

Alghamdi, S., Alzahrani, N., & Alharthi, H. (2023). A comparative study of machine learning models for diabetes prediction. *Journal of Biomedical Informatics*, *135*.

Anthony, A.-V. G., & Eloy, C.-V. (2023). *Comparative analysis using classification methods versus early stage diabetes*.

Asyiah, N. (2008). *Regresi logistik dan penerapannya dalam bidang kesehatan*. UIN Sunan Kalijaga, Yogyakarta.

Cho, N. H., Shaw, J. E., Karuranga, S., Huang, Y., da Rocha Fernandes, J. D., Ohlrogge, A. W., & Malanda, B. (2018). IDF diabetes atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes Research and Clinical Practice*, *138*, 271–281.

https://doi.org/10.1016/j.diabres.2018.02.023

Garcia, M. A., & Hernandez, L. (2023). Using artificial intelligence to enhance diabetes diagnosis: A logistic regression approach. *IEEE Transactions on Healthcare Informatics*, *17*(1), 212–227.

García, S., Luengo, J., & Herrera, F. (2015). *Data preprocessing in data mining* (Vol. 72). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-10247-4

Harrison, T. A., Hindorff, L. A., & Kim, H. (2017). Family history and genetic risk factors for diabetes. *Diabetes Care*, *40*(5), 679–685.

Hidayat, I., Tolago, A. I., Dako, R. D. R., & Ilham, J. (2023). Analisis data eksploratif capaian indikator kinerja utama 3 fakultas teknik. *Jambura Journal of Electrical and Electronics Engineering*, *5*(2), 185–191. https://doi.org/10.37905/jjeee.v5i2.18397

Johnson, R., & Lee, K. (2021). Comparative analysis of machine learning models for diabetes prediction: Logistic regression vs deep learning. *International Journal of Data Science in Health*, *10*(1), 45–60.

Kleinbaum, D. G., & Klein, M. (2010). *Logistic regression*. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4419-1742-3

Kupper, L. L., Hosmer, D. W., & Lemeshow, S. (1990). Applied logistic regression. *Journal of the American Statistical Association*, *85*(411), 901. https://doi.org/10.2307/2290035

Li, J., Wang, Y., & Zhao, Z. (2023). Evaluating the effectiveness of regression models in predicting diabetes risk. *Statistical Methods in Medical Research*, *32*(5), 1014–1030.

Menard, S. (2002). *Applied logistic regression analysis*. 2455 Teller Road, Thousand Oaks California 91320 United States of America : SAGE Publications, Inc. https://doi.org/10.4135/9781412983433

NCD Risk Factor Collaboration. (2022). Worldwide trends in diabetes prevalence and associated risk factors. *The Lancet*, *400*(10362), 881–897.

Pampel, F. (2000). *Logistic regression*. 2455 Teller Road, Thousand Oaks California 91320 United States of America : SAGE Publications, Inc. https://doi.org/10.4135/9781412984805

Parsaulian, A. S., Tarno, T., & Ispriyanti, D. (2021). Analisis faktor-faktor yang mempengaruhi penerima beras raskin menggunakan regresi logistik biner dengan gui r. *Jurnal Gaussian*, *10*(1). https://doi.org/10.14710/j.gauss.v10i1.30934

Putri, E. F., & Kismiantini. (2024). Analisis faktor-faktor yang memengaruhi status diabetes mellitus pada pra lansia dan lansia di indonesia menggunakan model regresi logistik biner. *Statistika*, *24*(1), 54–64. https://doi.org/10.29313/statistika.v24i1.3319

Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. *Procedia Computer Science*, *132*. https://doi.org/10.1016/j.procs.2018.05.122

Smith, J., Doe, A., & Brown, P. (2020). The impact of glucose level and bmiI on diabetes risk: A logistic regression approach. *Journal of Medical Statistics*, *15*(2), 123–135.

Susanti, R. S., Anugrawati, S. D., Fitrah, Usman, J., & Yusrianto. (2023). Analisis faktor risiko penyebab diabetes melitus dengan menggunakan regresi logistik biner. *Jurnal MSA (Matematika Dan*

Statistika Serta Aplikasinya), *11*(2), 37–45. https://doi.org/10.24252/msa.v11i2.41051

Wang, R., Liu, X., & Zhang, Y. (2023). A hybrid approach combining logistic regression and decision trees for diabetes prediction. *Computers in Biology and Medicine*, *153*.

World Health Organization. (2023). Diabetes. Retrieved February 25, 2025, from https://www.who.int/news-room/fact-sheets/detail/diabetes

Yasmin, E. A. (2023). *Analisis faktor-faktor yang mempengaruhi diabetes mellitus dengan menggunakan regresi logistik ordinal di rsi jemursari surabaya*. Institut Teknologi Sepuluh Nopember.