Ambidextrous: Journal of Innovation, Efficiency and Technology in Organizations

https://journal.takaza.id/index.php/ambidextrous

Vol. 4, No. 2, 2026, pp. 204-211
E-ISSN: 3031-7002

E-mail:ambidex@takaza.id

Application of Logistic Regression Method for Predicting Diabetes Mellitus

Dendi Pratama Riawan[1], Dede Brahma Arianto[2]
Faletehan University Informatics Study Program
Corresponding Author e-mail:dendipratamar@gmail.com,dedebrahma@uf.ac.id

Article History:
Received: 04-07-2026
Revised: 07-07-2026
Accepted: 07-07-2026

Keywords:*Diabetes Mellitus, Logistic Regression, Data Mining, Confusion Matrix, ROC Curve*

Abstract:*Diabetes is a chronic disease that requires early detection to prevent complications. This study refers to the analysis of diabetes prediction using the Logistic Regression algorithm. The data used comes from the open dataset platform, namely Kaggle, including health attributes such as Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Age, Outcome. The process in this study includes data cleaning, model development, and prediction. Model assessment was carried out using Confusion Matrix to calculate accuracy, Precision, Recall, and F1-Score, which is supported by ROC Curve analysis. The findings in this study show that the Logistic Regression model achieved an accuracy level of 75.32% and an AUC of 0.8232, indicating that the classification performance is quite good in predicting diabetes conditions.*

How to Cite: Dendi Pratama Riawan, Dede Brahma Arianto. (2026). Application of the Logistic Regression Method to Predict Diabetes Mellitus.*04* (02). pp https://doi.org/10.61536/ambidextrous.539

https://doi.org/10.61536/ambidextrous.539

Introduction

Diabetes mellitus is a chronic disease that poses a serious threat to public health worldwide. Diabetes is a disorder of carbohydrate metabolism in the body characterized by high blood sugar levels (hyperglycemia) due to problems with insulin secretion, insulin function, or both. Insulin is a hormone that plays a role in regulating blood glucose levels (Oktaviana et al. 2022). When the body does not produce enough insulin or if the response to insulin is suboptimal, glucose cannot enter cells and accumulates in the bloodstream, resulting in excessively high blood sugar levels.

The causes of diabetes are complex and involve an interaction between genetic and non-genetic factors. Type 1 diabetes is caused by damage to pancreatic cells due to autoimmune disorders or unknown causes, resulting in decreased or cessation of insulin production. This

type of diabetes generally appears before a person reaches the age of 20 (Guarango, 2022). Type II diabetes mellitus is most common in older adults, but cases are increasing among children, adolescents, and young adults. The causative factors for type II diabetes are closely related to being overweight (obesity), age, family health history, and high consumption of sugary drinks (Suyani, 2022).

The body's inability to control blood glucose levels over the long term can lead to serious complications affecting other organs. These complications include atherosclerosis, nerve damage, kidney dysfunction, and retinopathy.(Bina et al. 2025)Early symptoms of diabetes are often nonspecific and go unnoticed by sufferers, so many cases are diagnosed only after complications have developed or the disease is already quite advanced. This further highlights the importance of early detection and risk-based prevention efforts.

Diabetes is also a significant global public health problem. According to the 2023 Indonesian Health Survey (SKI), the prevalence of diabetes in Indonesia is 11.7%. This represents an increase from the 2018 Basic Health Research (Riskesdas) of 10.9% (Ministry of Health 2023). Projections indicate that the number of sufferers will continue to rise substantially in the coming decades. This increasing prevalence is influenced by modern lifestyle changes such as high-calorie food consumption, increased obesity, and low physical activity, which are the main risk factors for type 2 diabetes.

Advances in computing and statistical technology are now being utilized to manage health data and identify disease risks. This approach enables predictions of health conditions based on clinical and demographic data patterns. One statistical method frequently used to predict binary conditions, such as determining whether a person is at risk of diabetes, is Logistic Regression. This method estimates the probability of an event, such as a diabetes diagnosis, based on a combination of several predictor variables, including age, body mass index, blood glucose levels, blood pressure, and other relevant risk factors (Rahma et al. 2025).

Logistic Regression has advantages in the health context because, in addition to its classification capabilities, it produces interpretive models; the model coefficients can provide information on the influence of each variable on a person's likelihood of developing diabetes. Several studies have applied Logistic Regression to predict diabetes risk, with good classification performance, making this method worthy of consideration as an initial analytical tool in data-driven disease prediction studies (Hamid, 2020).

Several previous studies have shown that Logistic Regression performs well in predicting diabetes risk. According to research conducted by Fitri and Dede, the Logistic Regression model applied to clinical data yielded an accuracy of 79.1%, with a precision of 77% and a recall of 74%, making it considered quite stable in differentiating between diabetic and non-diabetic groups.(Kurniawati & Arianto 2023). These results are in line with the findings of Gunawan et al which showed that Logistic Regression was able to provide more than 80% accuracy and a high AUC value on type 2 Diabetes Mellitus patient data, so the model was considered effective for use as an early detection tool in health services (Gunawan et al, 2025) Similar findings were also reinforced by Rassiyanti et al who reported an ROC-AUC value of 0.83, with blood glucose and body mass index (BMI) variables as the most influential predictors of diabetes risk (Rassiyanti et al, 2025). Overall, these results show that Logistic Regression is not only accurate, but also able to provide a clear interpretation of risk factors, so

https://doi.org/10.61536/ambidextrous539

it is relevant to use in this study as the main method for predicting the possibility of diabetes based on clinical data.

Overall, a thorough understanding of diabetes and statistical approaches to predicting its risk is crucial in the context of public health and medical data analysis. Discussing diabetes prediction analysis using the Logistic Regression method in the context of a journal assignment is relevant as part of understanding how statistics can help systematically and data-drivenly identify individuals at higher risk for this disease.

Research Methods

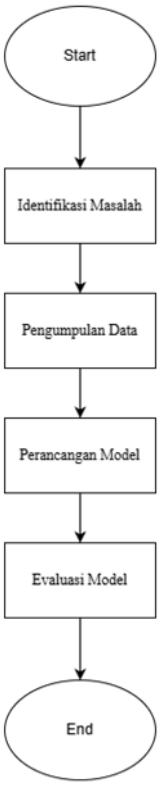

Figure 1. Research Method

This research flow is designed to create and evaluate a diabetes prediction model using the Logistic Regression algorithm. The research process is described in a flowchart that includes several steps: Problem Identification, Data Collection, Model Design, and Model Evaluation.

Problem Identification

Diabetes mellitus is a chronic disease with a growing prevalence and requires early detection to avoid more complex complications. Traditional diabetes diagnosis methods still rely on manual examinations, which are time-consuming and expensive. Furthermore, the shortage of medical personnel and the increasing number of patients present challenges to the diagnostic process.

The main problem in this research is how to create a diabetes prediction model that can effectively classify patients into diabetic or non-diabetic categories by utilizing their medical data. Therefore, an approach using data mining and machine learning is needed to accelerate and simplify the disease prediction process. The Logistic Regression algorithm was chosen because it is considered suitable for solving binary classification problems.

https://doi.org/10.61536/ambidextrous.539

Data collection

Data collection was conducted using a diabetes dataset obtained from an open data source, a public dataset commonly used in data mining and machine learning research. The dataset is available in CSV format and contains 768 rows of patient medical data for analysis.

This dataset includes various attributes such as number of pregnancies (Pregnancies), glucose level (Glucose), blood pressure (Blood Pressure), skin thickness (Skin Thickness), insulin level (Insulin), body mass index (BMI), family history of diabetes (Diabetes Pedigree Function), and patient age (Age) and outcome. This dataset was selected by considering the completeness of the attributes, data quality, and relevance to the research problem.

The attributes used in this study can be seen in the following table:

Table 1. Attributes and Categories of Diabetes Disease

ATTRIBUTE	CATEGORY
Pregnancies	0 – 5
	6 – 10
	11 – 17
Glucose	0 – 79
	80 – 89
	90 – 99
	100 – 145
	≥ 146
Blood Pressure	0 – 80
	82 – 98
	100 - 184
Skin Thickness	0 -20
	21 – 30
	31 - 45
	46 – 99
Insulin	0 – 59
	60 – 100
	101 - 200
	≥ 201
BMI	< 18.5
	18.5 – 24.9
	25 – 29.9
	≥ 30
Age	21 – 30
	31 – 40
	41 – 50
	> 50
Outcome	0 = No Diabetes
	1 = Diabetes

Model Design

The model design stage is the process of designing a system to predict diabetes using the Logistic Regression algorithm. This step includes initial data processing, dataset division, model development, and model storage for analysis.

1. Dataset Sharing

   The pre-processed dataset was then divided into two parts: training data and testing data. The data separation process was carried out with a proportion of 80% for training data and 20% for

https://doi.org/10.61536/ambidextrous539

testing data.

Training data is used to develop the Logistic Regression model, while testing data is used to assess how well the model predicts previously analyzed data.

2. Logistic Regression Algorithm

*Logistic Regression*is a classification algorithm that functions to estimate the possibility of an event occurring with the result being a value of two choices. In this study, the Logistic Regression algorithm was applied to predict the patient's diabetes condition based on existing medical data.

This algorithm model uses the sigmoid function which is formulated as follows:

$$P(Y\ 1|X) = \frac{1}{1 + e-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}$$

Information:

P(Y 1|X) = is the probability that the patient has diabetes

$x, x_2 \cdots x_n$= is the input variable

β_0= is a constant (Intercept)

$\beta_{1\ \beta_2} \cdots \beta_n$= is the regression coefficient

Model Evaluation

Model evaluation was conducted to assess the performance of the Logistic Regression algorithm in predicting diabetes based on test data. The purpose of this evaluation process was to assess the model's usability and ensure effective classification before further analysis of the results.

The evaluation process is carried out by comparing the model prediction results to the original data using various evaluation matrices that are commonly used for binary classification research, such as*precision, recall. F1-score, and accuracy*obtained from the results of the confusion matrix.

*Confusion matrix*comparing the model prediction results with the actual data, resulting in the main components, namely true positive (TP), true negative (TN), false positive (FP), false negative (FN).

*Precision*used to assess the model's accuracy in predicting data categorized as positive. This matrix shows the proportion of positive predictions that actually match the actual situation.

*Recall*Used to measure how effectively the model can identify all data. This matrix shows how well the model can identify diabetes cases from all data.

*F1-Score*is the harmonic mean between precision and recall. This matrix serves to balance the accuracy and completeness of predictions, especially in data with an unbalanced class distribution.

*Accuracy*used to assess the overall accuracy of the model in classifying data. This matrix provides an overview of the model's performance on all test data.

Results and Discussion

Preprocessing Data

At this stage it is done*Preemption*to ensure the quality of data information before it is used in the model building process. The process carried out in this study includes checking for empty data with invalid values, especially for data attributes that function as target variables. The goal of this stage is to improve the accuracy and stability of the model in predicting diabetes.

This preprocessing stage plays an important role in ensuring that the data used is ready to be processed by the algorithm and can produce more accurate predictions.

Here is the script for the Logistic Regression model:

```
def evaluate_model(model, X_test, y_test):
    y_pred = model.predict(X_test)
    y_prob = model.predict_proba(X_test)[:,1]
```


https://doi.org/10.61536/ambidextrous.539

```
    accuracy = accuracy_score(y_test, y_pred)
    # Confusion Matrix
    cm = confusion_matrix(y_test, y_pred)
      plt.figure()
      sns.heatmap(cm, annot=True, fmt='d', cmap='Blues')
      plt.xlabel("Predicted")
plt.ylabel("Actual")
    plt.savefig("confusion_matrix.png")
plt.close()
    # ROC Curve
    fpr, tpr, _ = roc_curve(y_test, y_prob)
    roc_auc = auc(fpr, tpr)
    plt.figure()
    plt.plot(fpr, tpr, label=f"AUC = {roc_auc:.2f}")
    plt.plot([0,1], [0,1], linestyle='--')
    plt.legend()
      plt.savefig("roc_curve.png")
    plt.close()
    report = classification_report(y_test, y_pred)
    with open("evaluation_result.txt", "w") as f:
        f.write(f"Model Accuracy: {accuracy}\n\n")
    f.write(report)
    return accuracy, roc_auc
```

Logistic Regression Evaluation Results

This stage presents the performance prediction results of the model.*Logistic Regression*The assessment was conducted using test data and presented in the form of a confusion matrix and ROC curve to provide a visual overview of the model's performance. The test results yielded an accuracy value of 0.7532, or 75.32%, indicating that the model was able to classify data on both diabetic and non-diabetic patients with a fairly good level of accuracy.

```
Akurasi Model: 0.7532467532467533

              precision    recall  f1-score   support

           0       0.80      0.83      0.81        99
           1       0.67      0.62      0.64        55

    accuracy                           0.75       154
   macro avg       0.73      0.72      0.73       154
weighted avg       0.75      0.75      0.75       154
```

Figure 2. Logistic Regression Evaluation Results

Confusion Matrix

Based on*confusion matrix*This model is able to correctly classify most of the data. The distribution of true positive, true negative, false positive, and false negative values provides an overview of the model's error rate and success in predicting diabetes status in patients. Further analysis of the confusion matrix components forms the basis for calculating evaluation metrics such as precision, recall, F1-score, and accuracy.

https://doi.org/10.61536/ambidextrous539

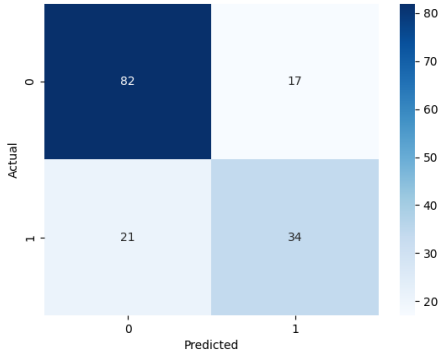

Figure 3. Confusion Matrix of Logistic Regression Prediction Results

ROC Curve

The evaluation results on the model are also displayed in the form*ROC Curve*, this form is used to describe the model's ability to distinguish between diabetic and non-diabetic classes at various threshold values.

The Area Under Curve (AUC) value generated from the ROC Curve indicates that the model has good classification capabilities. The higher the AUC value, the better the model is at distinguishing between the two classes.

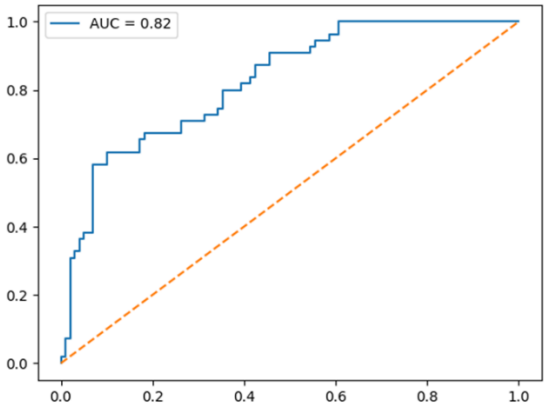

Figure 4. ROC Curve Logistic Regression Model.

Conclusion and Recommendation

Based on the results of this study, it can be concluded that the Logistic Regression algorithm can be used to predict diabetes status using patient medical data. The model developed through data processing, model training, and performance evaluation showed quite satisfactory results.

The evaluation results also achieved an accuracy of 72.325%, indicating that the majority of the test data was correctly categorized. Furthermore, the ROC Curve evaluation yielded an Area Under Curve (AUC) value of 0.8242, indicating that this model effectively differentiates between diabetic and non-diabetic patients.

Based on the results of the research that has been conducted, further research development can be focused on improving data quality, such as handling unbalanced data and cleaning invalid values, so that the model can later detect diabetes cases much better.

https://doi.org/10.61536/ambidextrous.539

In addition, the use of other classification algorithms as a comparison such as Support Vector Machine, Random Forest, or K-Nearest Neighbor, can be done to find out which method can provide more optimal prediction performance.

References

Bina, J. et al. (2025) 'Blood Pressure Profile of Diabetes Mellitus Patients with Uncontrolled Blood Glucose Levels', 21(2), pp. 106–115.

Fitri, K., & Dede, BA (2025). 'Analysis of Feature Selection Implementation in Diabetes Classification Using Correlation Matrix Method and Logistic Regression Algorithm'. Informatik: Journal of Computer Science.https://doi.org/10.52958/iftk.v19i3.6019

Guarango, PM (2022). Analysis of Diet Management That Influences Blood Sugar in DM Patients.

Gunawan, S., Astuti, R., Prihartono, W., & Hamonangan, R. (2025). 'Prediction of Type 2 Diabetes Mellitus with Logistic Regression Algorithm for Early Detection'. Journal of Informatics and Applied Electrical Engineering.https://doi.org/10.23960/jitet.v13i1.5747

Hamid, YI (2020). 'Prediction of Type 2 Diabetes through Risk Factors using Binary Logistic Regression'. Journal of Al-Qadisiyah for Computer Science and Mathematics, 12(3), pp. 1–11.

Ministry of Health of the Republic of Indonesia. (2023). 2023 Indonesian Health Survey (SKI). Jakarta: Ministry of Health of the Republic of Indonesia.

Oktaviana, Elisa et al. (2022). 'Analysis of the Relationship of Blood Glucose Levels with Total Cholesterol and Age of Diabetes Mellitus Patients'. International Journal of Nursing and Health Services (IJNHS), 5(2), pp. 195–202.

Rahma, CF et al. (2025). 'Diabetes Risk Prediction Using Logistic Regression Model', 8(1), pp. 101–114. Available at:https://doi.org/10.24042/djm.

Rassiyanti, L., Farid, F., & Pitri, R. (2025). 'Diabetes Risk Prediction Using Logistic Regression Model'. Desimal: Journal of Mathematics.https://doi.org/10.24042/djm.v8i1.26493

Suyani, S. (2022). 'Factors Associated with the Incidence of Low Birth Weight'. JKM (Journal of Public Health) Cendekia Utama, 10(2), p. 199.https://doi.org/10.31596/jkm.v10i2.1069