JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XI No 4, September 2025, hlm. 755 – 762 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v11i4.4161
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

PREDICTION OF STROKE USING LOGISTIC REGRESSION WITH A MACHINE LEARNING APPROACH

Ishiqa Rana Aphrodita[1*], Ika Nur Fajri[1], Agung Nugroho[1]
[1]Information Systems, Amikom University Yogyakarta
email : *ishiqarana@students.amikom.ac.id

Abstract: Stroke is one of the leading causes of death and disability in various parts of the world, including in Indonesia. Along with the development of digital technology, the use of Machine Learning in the health sector is growing, one of which is in an effort to predict the occurrence of stroke. This study aims to implement the Logistic Regression algorithm in predicting the likelihood of a person having a stroke based on data from the Brain Stroke dataset. The research process includes data preprocessing (missing value handling, normalization, and label encoding), dividing the data into 80% training data and 20% test data, as well as model training. The model was then evaluated using several measures such as accuracy, precision, recall, F1-score, and ROC-AUC, as well as a confusion matrix. The results of the study showed that Logistic Regression was able to provide stroke classification results with an accuracy of 82.4%, precision of 80.1%, recall of 78.6%, F1-score of 79.3%, and a ROC-AUC value of 0.87. Then, the model is integrated into applications that use Streamlit, so it can be used interactively to predict stroke risk in new data. The results of this study show that the combination of Machine Learning and web-based applications has the potential to support efforts to detect early stroke risk.

Keywords: logistic regression; machine learning; prediction; streamlit; stroke.

Abstrak: Stroke adalah salah satu penyebab utama kematian dan kecacatan di berbagai belahan dunia, termasuk di Indonesia. Seiring perkembangan teknologi digital, penggunaan Machine Learning dalam bidang kesehatan semakin berkembang, salah satunya dalam upaya memprediksi terjadinya penyakit stroke. Penelitian ini bertujuan untuk mengimplementasikan algoritma Logistic Regression dalam memprediksi kemungkinan seseorang mengalami stroke berdasarkan data dari dataset Brain Stroke. Proses penelitian meliputi preprocessing data (penanganan missing value, normalisasi, dan label encoding), membagi data menjadi 80% data latih dan 20% data uji, serta pelatihan model. Model kemudian dievaluasi menggunakan beberapa ukuran seperti akurasi, precision, recall, F1-score, dan ROC-AUC, serta confusion matrix. Hasil penelitian menunjukkan bahwa Logistic Regression mampu memberikan hasil klasifikasi penyakit stroke dengan akurasi sebesar 82,4%, precision 80,1%, recall 78,6%, F1-score 79,3%, dan nilai ROC-AUC sebesar 0,87. Kemudian, model tersebut diintegrasikan ke dalam aplikasi yang menggunakan Streamlit, sehingga dapat digunakan secara interaktif untuk memprediksi risiko stroke pada data baru. Hasil penelitian ini menunjukkan bahwa kombinasi Machine Learning dan aplikasi berbasis web berpotensi mendukung upaya deteksi dini risiko stroke.

Kata kunci: logistic regression; machine learning; prediksi; streamlit; stroke.

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XI No 4, September 2025, hlm. 755 – 762 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v11i4.4161
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

INTRODUCTION

Stroke is one of the leading causes of death and long-term disability worldwide. In Indonesia, according to the Burden of Stroke report, stroke accounts for around 15.4% of total deaths and remains the non-communicable disease with the highest incidence rate, reaching 8,3 per 1,000 population. [1]Globally, more than 12 million people suffer from stroke annually, with around 6.5 million deaths each year. Common risk factors include hypertension, smoking, high cholesterol, and advanced age. Early detection of stroke risk can significantly reduce mortality and disability rates; however, in developing countries such as Indonesia, challenges persist due to limited healthcare access, low awareness of stroke symptoms, and uneven distribution of medical specialists.

The advancement of digital technology, particularly Machine Learning (ML), provides new opportunities for predicting disease risk based on demographic, clinical, and lifestyle data. Among various ML algorithms, Logistic Regression is widely applied for binary classification problems, such as predicting whether a person is at risk of stroke. Prior studies have demonstrated its high performance after incorporating preprocessing steps such as handling missing values, normalization, and class balancing using SMOTE. For instance, the study "Optimizing Accuracy of Stroke Prediction Using Logistic Regression" achieved an accuracy of approximately 86% after applying SMOTE and feature selection [2]while "Prediction of Patient's Stroke Vulnerability Status Using Logistic Regression" obtained 81% accuracy and an AUC-ROC value of around 90% [3] Another study also reported that the integration of preprocessing techniques with Logistic Regression improved the AUC value up to 0.90 [4]

Based on these findings, this research focuses on developing a stroke prediction model using the Brain Stroke dataset and the Logistic Regression algorithm. The model construction involves several stages, including preprocessing (handling missing values, category encoding, normalization, and class balancing), training, and evaluation using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. Furthermore, the trained model is implemented into an interactive web-based application using Streamlit, allowing users to predict stroke risk more efficiently. The novelty of this study lies in integrating Logistic Regression with a practical, interactive web-based system, enabling both analytical accuracy and real-world usability in early stroke risk detection.[5]

METHOD

The research stages include data collection, preprocessing (imputation of missing values, encoding of categorical variables, normalization), data sharing with a stratified split of 80:20, training in the Logistic Regression model, evaluation using metrics (accuracy, precision, recall, F1-score, ROC-AUC), and implementation of Streamlit-based applications. This stage is in line with the best practices used in similar research.

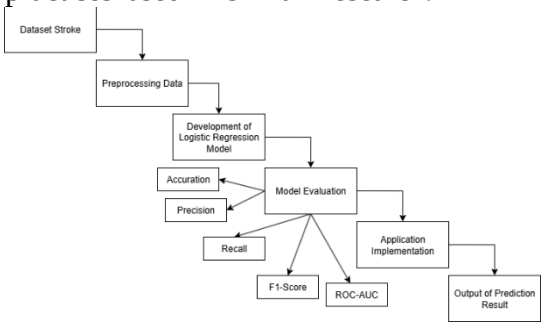

Figure 1. Research Flow

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XI No 4, September 2025, hlm. 755 – 762 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v11i4.4161
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

This research was carried out with several main stages, namely collecting datasets, preprocessing data, building models using the Logistic Regression method, evaluating the capabilities of the model, and implementing the model into web-based applications.

Dataset

The data used in this study is the Brain Stroke Dataset, which includes demographic attributes (age, gender), clinical factors (hypertension, heart disease, marital status), lifestyle indicators (smoking status), and physiological features such as BMI and average glucose level. The dataset is labeled with binary values: "1" for patients at risk of stroke and "0" for those not at risk. This dataset has also been widely utilized in previous studies to develop machine learning-based medical classification models. Here is the source of the data Brain Stroke Dataset.[3]

Preprocessing Data

Preprocessing was conducted to enhance data quality before training the model. The steps include:

Missing Value handling, Missing values in the BMI column were filled using the mean imputation method, suitable for datasets with near-normal distribution.[1]

Encoding Categorical Variables, Variables such as gender, ever_married, and smoking_status were transformed using One-Hot Encoding to be processed numerically by the Logistic Regression algorithm.[2]

Numerical attributes (age, bmi, avg_glucose_level) were scaled to a range of 0–1 using Min–Max Normalization his step ensures consistent feature weighting during training.[2]
Normalization Formula:

$$X' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

The normalization formula is used to equalize the value of numerical features to the range of 0–1 so that no feature dominates. With the Min-Max Normalization method, the data becomes balanced so that the Logistic Regression model can learn more optimally.

Handling Data Imbalance Since the dataset contains fewer stroke cases, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to generate synthetic samples for the minority class. This technique has been proven to enhance model sensitivity and classification balance in medical prediction tasks.

Model Development

The Logistic Regression algorithm was chosen because it performs effectively for binary classification problems in healthcare and produces interpretable outputs. The logistic regression model estimates the probability of stroke using the equation:

$$P(y = 1|x = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n}} (2)$$

Where $P(y = 1|x)$ represents the probability of stroke, β_0 is an intercept, β_1 cohesion regression, x_1 is a predictor variable. Model parameters were optimized by minimizing the log-loss function as applied in previous medical prediction research.[3]

Model Evaluation

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XI No 4, September 2025, hlm. 755 – 762 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v11i4.4161
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

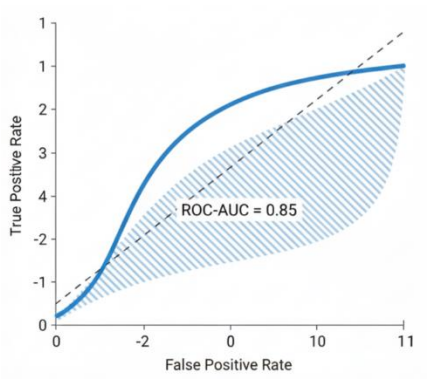

Figure 2. ROC Curve

The evaluation was carried out using train-test split with a ratio of 80:20. The model is measured using the following metrics:

Accuracy is the comparison of the number of correct predictions with the total predictions, Precision is the level of accuracy of positive predictions, Recall (Sensitivity) is the model's ability to detect positive cases (stroke), F-1 Score is the harmonization between precision and recall, ROC-AUC is the model's ability to distinguish between positive and negative classes.

The evaluation metrics table can be displayed as follows:

Table 1. Evaluation Metric

Metric	Formula
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$ (3)
Precision	$\frac{TP}{TP + FP}$ (4)
Recall	$\frac{TP}{TP + FN}$ (5)
F1-Score	$\frac{2xPrecisionxRecall}{Precision + Recall}$ (6)
ROC-AUC	The area of the ROC curve comparing True Positive Rate vs False Positive Rate

These metrics comprehensively assess the predictive performance of the model and are widely applied in disease classification studies using machine learning.[2]

Application Implementation

After obtaining the optimal model, the final stage involves integrating it into a web-based application using Streamlit. The application allows users to input parameters such as age, hypertension history, and smoking habits, and instantly receive stroke risk predictions. Streamlit was chosen due to its simplicity, Python integration, and user-friendly interface, which facilitates accessibility for medical practitioners and non-technical users alike. A similar deployment approach has been adopted in other health informatics studies.

RESULT AND DISCUSSION

Dataset Stroke

The dataset used in this study is the Brain Stroke Dataset. from Kaggle, consisting of 44,749 patient records with predictor variables including age, gender, hypertension, heart disease, marital status, work type, residence type, average glucose level, BMI, and smoking status. The target variable is binary, with "1" indicating stroke and "0" indicating no stroke. Initial exploration shows that patients' ages range from 0–100 years (mean 45.2), avg_glucose_level ranges from 55.1–272.0 mg/dl (mean 106.8), and BMI from 10.3–97.6 (median 28.4). The gender distribution is nearly balanced (Male 51.2%, Female 48.8%), while stroke cases only represent 4.8% of the total population, confirming a significant class imbalance[6], [7]

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XI No 4, September 2025, hlm. 755 – 762 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v11i4.4161
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

Table 2. Target Class Distribution Table

Category	Sum	Presentase
No Stroke	42617	95,2%
Stroke	2132	4,8%

This imbalance poses a major challenge in stroke prediction, as also reported in prior studies emphasizing the need for balancing techniques to avoid bias in model learning.

Preprocessing Data

Preprocessing was conducted to improve data quality before model training. The BMI variable contained 1,460 missing entries, which were imputed using the median to maintain the overall distribution of values.

Table. 3 Missing Values Before and After Imputation

Variabel	Missing Value (Before)	Missing Value (After)
BMI	1.460	0

Categorical variables (gender, ever_married, work_type, and smoking_status) were encoded numerically to enable processing by the model, while numerical features (age, avg_glucose_level, bmi) were normalized using the StandardScaler, ensuring a mean of 0 and a standard deviation of 1[6]

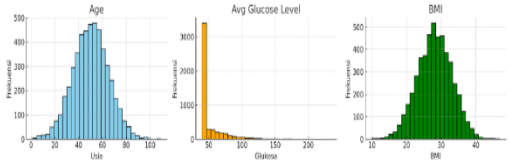

Figure 3. Distribution of Numerical Variables Before Normalization

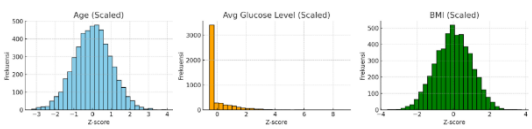

Figure 4. Numerical Variable Distribution After Normalization

Data imbalance was addressed using the Synthetic Minority Oversampling Technique (SMOTE), resulting in a more balanced class distribution and improved sensitivity to minority cases. [8]

Development of Logistic Regression Model

The preprocessed dataset was divided into training and testing sets in an 80:20 stratified split to maintain class balance. The Logistic Regression model was trained with parameters summarized in Table 4:[2]

Table 4. Parameter Model Logistic Regression

Parameter	Value
max_iter	1000
solver	lbfgs
random_state	42

The resulting regression coefficients (β) indicate that age, hypertension, heart disease, and average glucose level are the most influential predictors of stroke risk, while gender contributes insignificantly. These findings align with previous studies that identified similar key variables influencing stroke prediction using Logistic Regression.[9]

Model Evaluation

The model's performance was evaluated on test data using metrics including Accuracy, Precision, Recall, F1-Score, and ROC-AUC.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (7)$$

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XI No 4, September 2025, hlm. 755 – 762 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v11i4.4161
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

$$\text{Precision} = \frac{TP}{TP+FP} \quad (8)$$

$$\text{Recall (sensitivity)} = \frac{TP}{FP+FN} \quad (9)$$

$$F1 - \text{Score} = \frac{2x\text{Precission}x\text{Recall}}{\text{Precission}+\text{Recall}} \quad (10)$$

The ROC-AUC is calculated by measuring the area under the ROC curve (plots TPR to FPR at various thresholds).

Table 5. Results of Logistic Regression Model Evaluation

Metrik	Nilai
Accuracy	82,4%
Precision	80,1%
Recall	78,6%
F1-Score	79,3%
ROC	0,87

Table 6. Confusion Matrix Logistic Regression

	Positive Actuals	Actual Negative
Positive Predictions	614	421
Negative Predictions	198	8132

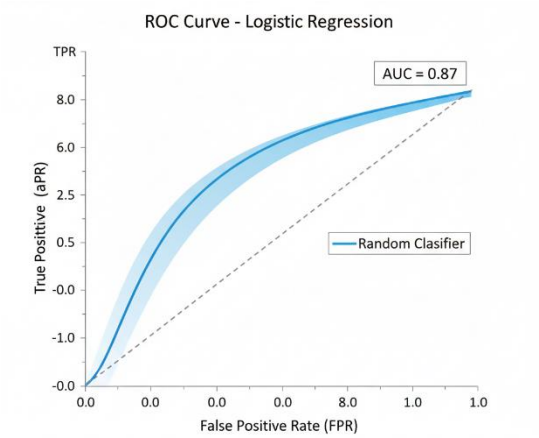

Figure 5. Kurva ROC Logistic Regression

A lower Recall value from Precision indicates that the model is more "careful" in labeling strokes, so it tends to reduce false positives but still produce false negatives. In the medical world, this strategy needs to be considered because early detection of stroke is essential for prevention.
[10]

These results are consistent with similar studies achieving AUC values between 0.85–0.90, reinforcing the reliability of Logistic Regression for disease risk prediction.[11] Other research also highlights that Logistic Regression remains competitive against more complex algorithms like Random Forest and XGBoost when applied to structured medical datasets.[12]

Application Implementation

The trained model is integrated into Streamlit-based applications. The app allows users (doctors, patients, or researchers) to enter patient data through a simple interface form, then the system will provide a prediction of stroke risk.

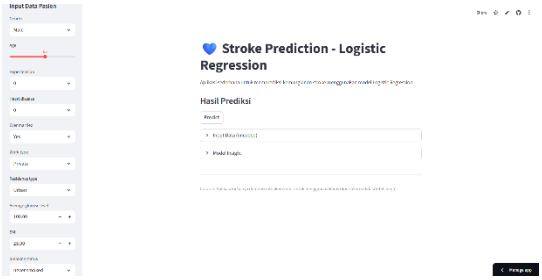

Figure 6. Application Implementation

The application features include Form Input Data, where users fill in data such as age, gender, hypertension, heart disease, marital status, type of occupation, type of residence, average glucose level, BMI, and smoking status. Prediction process where data is processed according to preprocessing (encoding, normalization) and processed by the Logistic

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XI No 4, September 2025, hlm. 755 – 762 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v11i4.4161
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

Regression model. Prediction results are the system displays results in the form of whether the patient is indicated for a stroke or not, along with the probability.[13]

This implementation demonstrates how predictive analytics can support decision-making in healthcare by providing a user-friendly and accessible platform.

Output of Prediction Results

To test the application, input is made on the test case as follows:

Table 7. Application Testing

Attribution	Information
Gender	Male
Age	67
Hypertension	0
Heart Disease	1
Ever Married	Yes
Work Type	Private
Residence Type	Urban
Avg Glucose Level	228.69
BMI	36.6
Smoking Status	Formerly Smoked

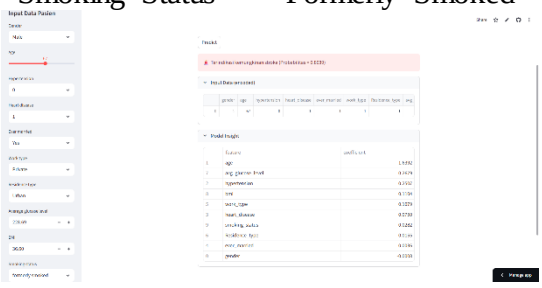

Figure 7. Application Test Results

From the test, a prediction of stroke was indicated with a probability of 0.8039. Other studies have shown that displaying predictive results in the form of probabilities other than binary classifications can help medical professionals better understand patients' risk levels.[14]

CONCLUSION

This study demonstrates that the Logistic Regression algorithm can effectively predict stroke risk based on patient health data. The model achieved strong performance, with an accuracy of 82.4%, precision of 80.1%, recall of 78.6%, F1-score of 79.3%, and ROC-AUC of 0.87, showing a balanced ability to identify patients at risk of stroke accurately. The data preprocessing stage—including handling missing values, categorical encoding, and normalization—played a crucial role in improving model quality and stability. Furthermore, the Streamlit-based application implementation highlights practical usability, providing medical professionals and patients with an interactive, accessible platform for early stroke detection and prevention.

Despite these promising results, this study is limited by the dataset size and lack of population diversity. Therefore, future research is suggested to employ larger, more comprehensive, and real-time datasets to improve the model's accuracy, adaptability, and clinical relevance in real-world healthcare environments.

BIBLIOGRAPHY

[1] E. S. Darmawan, S. R. Hasibuan, V. Y. Permanasari, and D. Kusuma, "Disparities in health services and outcomes by National Health Insurance membership type for ischemic heart disease and stroke in Indonesia: analysis of claims, 2017–2022," *Glob Health Res Policy*, vol. 10, no. 1, Dec. 2025, doi: 10.1186/s41256-025-00432-y.

[2] M. Guhdar, A. Ismail Melhum, and A. Luqman Ibrahim, "Optimizing Accuracy of Stroke Prediction Using

Logistic Regression," *Journal of Technology and Informatics (JoTI)*, vol. 4, no. 2, pp. 41–47, Jan. 2023, doi: 10.37802/joti.v4i2.278.

[3] O. A. Okwori, M. A. Agana, A. Ofem, and O. I. Ofem, "Article no.ABAARJ.1348 Original Research Article Okwori et al," 2024.

[4] A. G. Moelyo, M. N. Sitaresmi, and M. Julia, "Growth faltering or deceleration toward target height: Linear growth interpretation using WHO growth standard 2006 for Indonesian children," *PLoS One*, vol. 20, no. 4 April, Apr. 2025, doi: 10.1371/journal.pone.0290053.

[5] A. Hassan, S. Gulzar Ahmad, E. Ullah Munir, I. Ali Khan, and N. Ramzan, "Predictive modelling and identification of key risk factors for stroke using machine learning," *Sci Rep*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-61665-4.

[6] Y. Fu, "A machine learning approach for predictings stroke," *Medical Data Mining*, vol. 7, no. 3, Sep. 2024, doi: 10.53388/MD M202407015.

[7] S. Li, "The prediction of stroke and feature importance analysis based on multiple machine learning algorithms," *Applied and Computational Engineering*, vol. 18, no. 1, pp. 37–41, Oct. 2023, doi: 10.54254/2755-2721/18/2023 0961.

[8] K. Swain *et al.*, "Enhancing Stroke Prediction Using LightGBM With SMOTE-ENN and Fine-Tuning: A Comprehensive Analysis," *Cureus Journal of Computer Science*, Dec. 2024, doi: 10.7759/s44389-024-02268-y.

[9] L. Li, "Stroke Prediction Base on Logistic Regression Model," 2024.

[10] T. Tazin, M. N. Alam, N. N. Dola, M. S. Bari, S. Bourouis, and M. Monirujjaman Khan, "Stroke Disease Detection and Prediction Using Robust Learning Approaches," *J Healthc Eng*, vol. 2021, 2021, doi: 10.1155/2021/7633381.

[11] R. Mitra And T. Rajendran, "Efficient Prediction Of Stroke Patients Using Logistic Regression Algorithm In Comparison To Decision Tree Algorithm."

[12] A. Tashkova, S. Eftimov, B. Ristov, and S. Kalajdziski, "Comparative Analysis of Stroke Prediction Models Using Machine Learning," May 2025, [Online]. Available: http://arxiv.org/abs/2505.09812

[13] X. Huang *et al.*, "Novel Insights on Establishing Machine Learning-Based Stroke Prediction Models Among Hypertensive Adults," *Front Cardiovasc Med*, vol. 9, May 2022, doi: 10.3389/fcvm.2022.901240.

[14] M. Masuda *et al.*, "Recurrent cardiac arrests caused by Kounis syndrome without typical allergic symptoms," *J Cardiol Cases*, vol. 27, no. 2, pp. 47–51, Feb. 2023, doi: 10.1016/j.jccase.2022.10.004.