*"BIMASAKTI"*
*Jurnal Riset Mahasiswa Bidang Teknologi Informasi*

***Volume 8 Nomor 2***
***P ISSN : 2355-4401 – EISSN 3109-1652***
*Available online at https://ejournal.unikama.ac.id/index.php/JFTI*

# PENERAPAN ALGORITMA LOGISTIC REGRESSION UNTUK KLASIFIKASI PENYAKIT STROKE

**Rachel Trivica Amelia**[1)] **, Danang Aditya Nugraha**[2)]**, Moh. Ahsan**[3)]

*Universitas PGRI Kanjuruhan Malang*[1,2,3)]
*email : racheltrivica88@gmail.com*

## *Abstrak*

*Stroke merupakan salah satu penyakit dengan tingkat kematian tertinggi di dunia setelah penyakit jantung dan kanker. Pentingnya deteksi dini terhadap potensi stroke adalah untuk memungkinkan penanganan lebih cepat dan tepat. Tujuan dari penelitian ini adalah guna menerapkan algoritma Logistic Regression dalam mengklasifikasikan penyakit stroke berdasarkan faktor-faktor risiko seperti Jenis kelamin, Umur, Hipertensi, Penyakit jantung, status pernikahan, jenis pekerjaan, Jenis tempat tinggal, avg glucose, BMI, status Merokok, Status stroke. Dataset yang digunakan bersumber dari Kaggle dengan jumlah 5110 data pasien. Proses penelitian meliputi tahapan data cleaning, transformasi data, normalisasi memanfaatkan Min-Max Scaler, serta pembagian data menjadi training dan testing menggunakan pendekatan beberapa variasi proporsi (90%-10%, 85%-15%, 80%-20%, 70%-30%, dan 65%-35%). Evaluasi dijalankan dengan memanfaatkan Confusion Matrix menggunakan pendekatan metrik akurasi, presisi, recall, dan F1-score. Berdasarkan hasil analisis didapatkan bahwa proporsi pembagian data training dan testing sebesar 90%-10% memberikan akurasi tertinggi yaitu 76,17%, dengan nilai precision dan recall yang menunjukkan kemampuan model dalam mengenali data non-stroke secara baik. Namun, performa pada kelas minoritas (stroke) masih rendah sehingga perlu dikerjakan perbaikan, seperti penanganan ketidakseimbangan data. Secara keseluruhan, penerapan algoritma Logistic Regression terbukti cukup efektif dalam klasifikasi awal penyakit stroke, namun akurasinya dapat ditingkatkan menggunakan pendekatan pengembangan metode resampling atau optimasi model lanjutan.*

***Kata Kunci :*** *Stroke; Logistic Regression; Klasifikasi; Data Mining; Confusion Matrix*

## *Abstract*

*Stroke is one of the leading causes of death worldwide, ranking after heart disease and cancer. Early detection of stroke risk is essential to enable faster and more accurate treatment. The purpose of this study is to apply the Logistic Regression algorithm to classify stroke cases based on several risk factors, including gender, age, hypertension, heart disease, marital status, occupation, residence type, average glucose level, body mass index (BMI), smoking status, and stroke status. The dataset used in this research was obtained from Kaggle and consists of 5,110 patient records. The research process involves several stages, including data cleaning, data transformation, and normalization using the Min-Max Scaler method, followed by splitting the data into training and testing sets with various proportions (90%-10%, 85%-15%, 80%-20%, 70%-30%, and 65%-35%). The evaluation was conducted using a Confusion Matrix with performance metrics such as accuracy, precision, recall, and F1-score. The analysis results show that the 90%-10% data split achieved the highest accuracy of 76.17%, with precision and recall values indicating that the model performs well in identifying non-stroke cases. However, performance on the minority class (stroke) remains relatively low, suggesting the need for improvement through data imbalance handling. Overall, the application of the Logistic Regression algorithm proved to be effective for initial stroke classification, although accuracy can still be improved through resampling techniques or advanced model optimization.*

***Keywords*** **:** *Stroke; Logistic Regression; Classification; Data Mining; Confusion Matrix*

## 1. PENDAHULUAN

Stroke atau *Cerebrovascular Accident* (CVA) merupakan gangguan fungsi saraf yang terjadi akibat terganggunya aliran darah ke otak karena penyumbatan pembuluh darah arteri

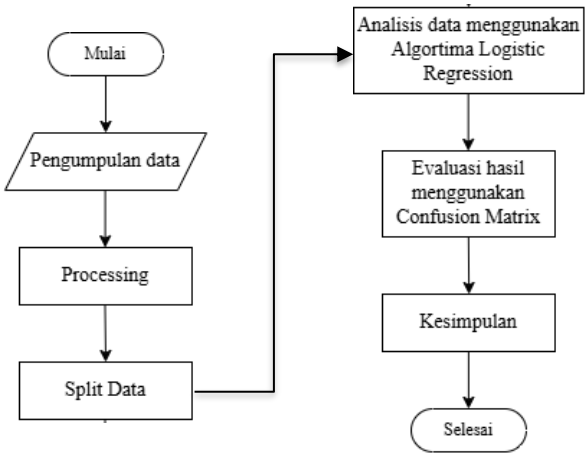
dari penumpukan darah pada pembuluh darah, pecahnya pembuluh darah akibat melemahnya dinding pembuluh darah atau kelainan pada kondisi darah itu sendiri. Hal ini mengakibatkan kurangnya cadangan oksigen dan nutrisi ke otak yang menimbulkan kerusakan di jaringan otak. Stroke merupakan penyebab kematian ketiga tersering di negara maju, setelah penyakit jantung dan kanker.

Berdasarkan data yang di peroleh pada World Stroke Organization (2022) terdapat 12.224.551 kasus stroke baru setiap tahunnya dan kondisi saat ini sebesar 101.474.558 individu pernah mengalami stroke [1]. Hampir setiap hari seorang warga negara dari berbagai usia, meninggal dunia karena stroke. Seorang penderita penyakit stroke tidak menyadari atau merasakan adanya gejala dini penyakit stroke, sehingga Ketika penderita mengalami gejala yang serius barulah si penderita di evaluasi ke rumah sakit guna mendapatkan penanganan yang lebih baik. Berdasarkan analisis faktor risiko stroke meliputi hipertensi, kebiasaan merokok, dislipidemia, diabetes melitus, obesitas, serta konsumsi alkohol. Faktor risiko yang tidak dapat diubah meliputi usia, jenis kelamin, dan riwayat keluarga.

Berdasarkan penelitian sebelumnya yang dikerjakan oleh [2] membandingkan model Decision Tree dan Logistic Regression guna menentukan klasifikasi penyakit jantung. Informasi yang digunakan dalam penelitian ini berdasarkan data tentang pasien yang menderita penyakit jantung. Data tersebut terdiri 297 pasien, 61 untuk pengujian dan 236 untuk pelatihan. Berdasarkan hasil analisis didapatkan bahwa bahwa model Logistic Regression memiliki akurasi yang tinggi sebesar 87% diikuti oleh model Decision Tree sebesar 75%. Penelitian selanjutnya dikerjakan oleh [3] yang membandingkan tiga model yaitu Naïve Bayes, Logistic Regression dan K-Nearest Neighbor guna menentukan klasifikasi peminatan masyarakat terhadap kandidat bakal calon presiden 2024. Data tersebut diambil dari data twitter sebanyak 2.201 data. Hasil dari penelitian tersebut menunjukkan bahwa algoritma Logistic Regression mempunyai akurasi 0.85% diikuti oleh model Naïve Bayes sebesar 0.77%, dan K-Nearest Neighbor sebesar 0.71%. Berdasarkan temuan ini, dapat disimpulkan bahwa bahwa algoritma Logistic Regression berpotensi mampu memberikan hasil yang akurat dalam proses klasifikasi.

## 2. METODE / ALGORITMA

Metode yang digunakan dalam penelitian ini adalah Algoritma Logistic Regression untuk mengklasifikasikan penyakit stroke. Langkah-langkah yang digunakan dalam penelitian ini dapat dilihat pada gambar dibawah ini.




**Gambar 1. Alur Penelitian**

### 2.1 Pengumpulan Data

Metode pengumpulan data dalam penelitian ini memanfaatkan studi literatur, di mana data yang akan digunakan adalah data siap pakai mengenai penyakit stroke yang di ambil dari situs

web https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset. Data yang akan dianalisis terdiri dari 5110 records terkait penyakit stroke. Berdasarkan data tersebut, klasifikasi akan dikerjakan menggunakan pendekatan mempertimbangkan indikator yang meliputi data fisik seperti Jenis kelamin, Umur, Hipertensi, Penyakit jantung, status pernikahan, jenis pekerjaan, Jenis tempat tinggal, avg glucose, BMI, status Merokok, Status stroke dan atribut lainnya guna klasifikasi penyakit stroke memanfaatkan algoritma Logistic Regression.

Berikut adalah deskripsi atribut yang digunakan dalam penelitian ini dapat dilihat pada tabel 2.1

**Tabel 1. Deskripsi Atribut**

| No | Atribut | Keterangan |
|---|---|---|
| 1. | Jenis Kelamin | Pada dataset tersebut mencakup laki-laki dan perempuan (0 untuk laki-laki dan 1 untuk perempuan) |
| 2. | Umur | Untuk usia responden bervariasi, meliputi anak-anak, remaja, dewasa muda, dewasa dan lansia. Rentang usia berkisar dari 1 sampai 82 tahun |
| 3. | Hipertensi | Kriteria Pasien yang hipertensi dikategorikan menjadi 2 meliputi 0 untuk non-hipertensi dan 1 yang hipertensi |
| 4. | Penyakit Jantung | Pasien tidak memiliki riwayat penyakit jantung ditulis sebagai 0 dan Pasien yang memiliki riwayat penyakit jantung di tulis sebagai 1 |
| 5. | Pernah Menikah | Untuk pasien yang sudah menikah di catat sebagai Yes (1) dan No (0) untuk pasien yang belum menikah |
| 6. | Jenis Pekerjaan | Untuk jenis pekerjaan di kategorikan menjadi pribadi (0), wiraswasta (1), pekerjaan pemerintah (2), anak-anak (3), dan tidak pernah bekerja (4) |
| 7. | Jenis Tempat Tinggal | Untuk jenis tempat tinggal pasien di catat sebagai 1 (perkotaan) dan 0 (untuk pedesaaan) |
| 8. | Tingkat Glukosa rata rata | Untuk tingkat glukosa rata-rata dalam dataset ini berkisaran 55.1 sampai 272 |
| 9. | Indeks Massa tubuh | Indeks yang dibutuhkan dalam dataset ini berkisaran 10.3 sampai 97.6 |
| 10. | Merokok | Untuk status pasien merokok dicatat sebagai No (0) tidak merokok dan Yes (1) yang merokok |
| 11. | Status Stroke | Status penyakit Stroke di kelompokkan sebagai 0 (tidak terdiagnosa penyakit stroke) dan 1 (terdiagnosa penyakit stroke) |

## 2.2 Preprosesing Data

Hal ini dijalankan guna memastikan bahwa data yang digunakan dalam penelitian ini akurat. Dalam penelitian ini, dataset penyakit stroke yang di peroleh dari situs web akan melalui tahap prosesing sebelum dikerjakan pemodelan. Oleh karena itu pada pemrosesan ini akan dikerjakan transformasi dan normalisasi data.

### 2.2.1 Data Cleaning

Data Cleaning adalah tahapan yang dilakukan untuk menghilangkan missing value apabila atribut pada data tidak tersedia atau kosong. Hal tersebut dilakukan agar dapat memastikan dataset yang digunakan siap untuk dianalisis lebih lanjut. Data yang akan dihapus pada proses ini adalah data pada atribut BMI Sebanyak 201 data. Terdapat outlier pada atribut tingkat glukosa rata-rata dan BMI sebanyak 627 data untuk tingkat glukosa rata-rata dan 110 data untuk BMI. Untuk mengatasi hal tersebut dengan memasukkan data median dari class.

### 2.2.2 Transformasi Data

Transformasi data adalah proses perubahan struktur atau format data mentah menjadi bentuk yang sesuai dalam analisis data. Proses ini melibatkan berbagai teknik seperti mengubah tipe data, menggabungkan data, atau membersihkan data yang tidak konsisten.

Pada proses ini merubah data kategorikal menjadi data numerik contohnya positif = 1, dan negatif = 0.

### 2.2.3 Normalisasi Data

Normalisasi berfungsi untuk mengubah rentang nilai data sehingga semua data memiliki skala yang sama. Biasanya, data dinormalisasi ke dalam rentang 0 hingga 1 memanfaatkan metode Min-Max. Normalisasi juga membantu algoritma pembelajaran mesin guna belajar menggunakan pendekatan lebih efektif. Pada penelitian ini akan dikerjakan normalisasi data pada atribut tingkat glukosa rata-rata dan indeks massa tubuh.

### 2.3 Split Data

Data pelatihan digunakan untuk proses pembelajaran algoritma, sedangkan data pengujian berfungsi untuk mengevaluasi kinerja model yang telah dihasilkan. Sehingga sebelum data dianalisis oleh algoritma Logistic Regression data akan dibagi terlebih dahulu menjadi data latih dan data uji. Penentuan data latih dan data uji pada data penyakit stroke di lakukan random sebanyak lima kali percobaan menggunakan pendekatan presentase yang berbeda yaitu 90% dan 10%, 80% dan 20%, 85% dan 15%, 70% dan 30%, 65% dan 35%.

### 2.4 Klasifikasi Algoritma Logistic Regression

Klasifikasi algoritma Logistic Regression melibatkan penggunaan algoritma guna memprediksi nilai biner dari suatu variabel terikat berdasarkan nilai beberapa variabel bebas. Variabel dependen dalam regresi logistik biasanya berupa kategori menggunakan pendekatan dua nilai, yaitu 1 (sesuai) dan 0 (tidak sesuai). Proses klasifikasi dalam studi ini memanfaatkan perangkat lunak Jupyter Notebook. Pada tahap ini dikerjakan perhitungan proses klasifikasi.

### 2.5 Evaluasi Hasil menggunakan Confusion Matrix

Evaluasi model dilakukan untuk menghitung akurasi dari algoritma Logistic Regression dalam pengklasifikasian penyakit stroke. Untuk menghitung matrix evaluasi digunakan metode validasi Confusion Matrix seperti akurasi, presisi, recall dan F1-score.

## 3. HASIL DAN PEMBAHASAN

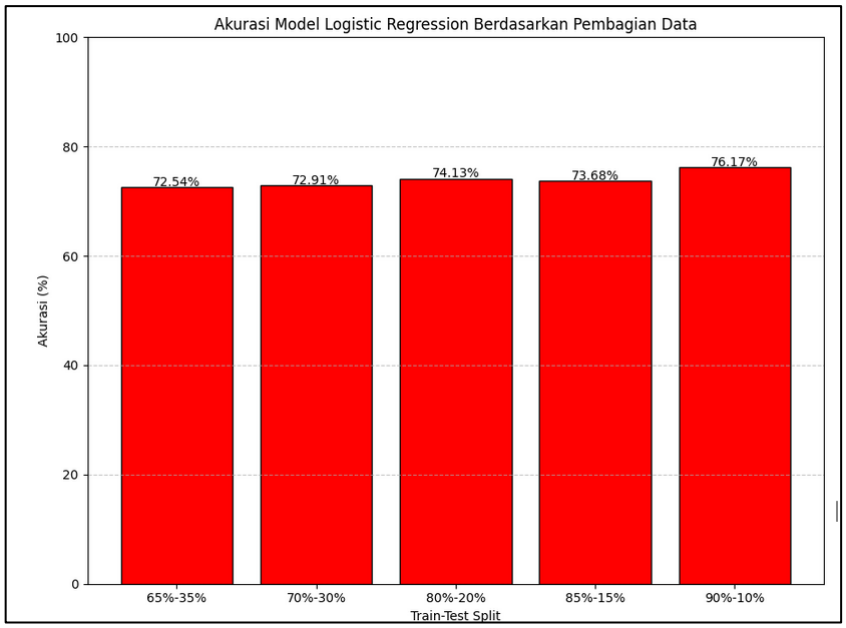
Pengujian algoritma Logistic Regression memanfaatkan data yang sudah dibagi menjadi dua bagian, pelatihan dan pengujian menggunakan pendekatan ukuran yang berbeda, dapat dilihat pada tabel berikut ini.

**Tabel 2. Hasil Pengujian Logistic Regression**

| Data Training | Data testing | Accuracy |
|---|---|---|
| 90% | 10% | 76.17% |
| 85% | 15% | 73.68% |
| 80% | 20% | 74.13% |
| 70% | 30% | 72.91% |
| 65% | 35% | 72.54% |

### 3.1 Tampilan dan Fungsi Utama Sistem

Penerapan metode Logistic Regression digunakan dalam penelitian ini untuk mengklasifikasikan data penyakit stroke. Pada klasifikasi Logistic Regression data dibagi menjadi dua yaitu, data latih dan data uji dengan presentase 90% dan 10%, 80% dan 20%, 85% dan 15%, 70% dan 30%, 65% dan 35%. Dalam pengujian algoritma Logistic Regression dengan melakukan percobaan sebanyak lima kali pada setiap presentase, sehingga hanya diambil hasil akurasi yang tertinggi pada setiap percobaan. Hal ini juga diterapkan pada metrik yang lain seperti precision, recall dan F1-score untuk masing masing presentase:




**Gambar 1. Grafik hasil pengujian Logistic Regression**

## 4. KESIMPULAN

Studi ini berhasil menerapkan algoritma Logistic Regression untuk klasifikasi penyakit stroke menggunakan bahasa pemrogaman python dengan dataset yang diperoleh dari Kaggle. Setelah dilakukan analisis dan pengujian hasil evaluasi menunjukkan bahwa model Logistic Regression mampu menghasilkan tingkat akurasi yang tinggi, dengan akurasi tertinggi mencapai 76.17% pada pembagian data 90% training dan 10% testing. Hasil evaluasi mengggunakan confusion matrix cukup baik dalam mengenali kelas non-stroke, namun kesulitan dalam mengenali kelas stroke, ditunjukkan dengan nilai Precision dan F1-score yang rendah pada kelas 1. Hal ini disebabkan oleh ketidakseimbangan data atar kelas. Saran untuk saran penelitian selanjutnya guna mengembangkan algoritma Logistic Regression dalam memprediksi: 1). Untuk meningkatkan nilai akurasi dapat menggunakan model lain seperti Random Forest, Gradient Boosting, atau XGBoots. 2). Untuk meningkatkan kinerja model dalam mengklasifikasikan kasus stroke, dapat menggunakan teknik penyeimbangan data seperti SMOTE.

## REFERENSI

[1] V. L. Feigin *et al.*, “World Stroke Organization (WSO): Global Stroke Fact Sheet 2022,” *Int. J. Stroke*, vol. 17, no. 1, pp. 18–29, 2022, doi: 10.1177/17474930211065917.

[2] F. Bukhari, S.- Nurdiati, M. K. Najib, and R. N. Amalia, “Deteksi Penyakit Jantung Menggunakan Metode Klasifikasi Decision Tree dan Regresi Logistik,” *Sains, Apl. Komputasi dan Teknol. Inf.*, vol. 5, no. 1, p. 41, 2024, doi: 10.30872/jsakti.v5i1.10780.

[3] Z. Febrian, “Perbandingan Algoritma Klasifikasi Naïve Bayes, Logistic Regression, Dan KNN Untuk Analisis Sentimen Peminatan Masyarakat Terhadap Kandidat Bakal Calon Presiden 2024,” pp. 1–78, 2023, [Online].

[4] W. C. F. Mariel, S. Mariyah, and S. Pramana, “Sentiment analysis: A comparison of deep learning neural network algorithm with SVM and naïve Bayes for Indonesian text,” *J. Phys. Conf. Ser.*, vol. 971, no. 1, 2018, doi: 10.1088/1742-6596/971/1/012049.

[5] Annas, S., Aswi, Abdy, M., & Poerwanto, B. (2021). Stroke Classification Model Using Logistic Regression. Journal of Physics: Conference Series, 2123(1), 012016. https://doi.org/10.1088/1742-6596/2123/1/012016

[6] Wang, L. (2023). Logistic Regression for Stroke Prediction: An Evaluation of its

Accuracy and Validity. Highlights in Science, Engineering and Technology, 39.

[7] Guhdar, M., Melhum, A. I., & Ibrahim, A. L. (2023). Optimizing Accuracy of Stroke Prediction Using Logistic Regression. Journal of Technology and Informatics, 4(2).

[8] Annas, S., Poerwanto, B., Aswi, Abdy, M., & Fa'rifah, R. Y. (2022). Classification Model for Type of Stroke Using Kernel Logistic Regression. Communications in Mathematical Biology and Neuroscience.

[9] Annas, S., Aswi, Abdy, M., & Poerwanto, B. (2021). *Stroke Classification Model Using Logistic Regression*. Journal of Physics: Conference Series, 2123(1), 012016. https://doi.org/10.1088/1742-6596/2123/1/012016. Penelitian ini membangun model klasifikasi stroke menggunakan Logistic Regression dan menganalisis faktor-faktor yang berpengaruh terhadap jenis stroke.

[10] Okoye, G. C., & Umeh, E. U. (2024). *Predicting Functional Outcome After Ischemic Stroke Using Logistic Regression and Machine Learning Models*. Earthline Journal of Mathematical Sciences.

[11] Guhdar, M., Melhum, A. I., & Ibrahim, A. L. (2023). *Optimizing Accuracy of Stroke Prediction Using Logistic Regression*. Journal of Technology and Informatics, 4(2). https://doi.org/10.37802/joti.v4i2.278.

[12] World Health Organization. (2024). *Stroke*.

[13] Kementerian Kesehatan Republik Indonesia. (2023). *Profil Kesehatan Indonesia*.