Vol. 5, No. 2, Agustus 2024, hal. 367-372

Jurnal Computer Science and Information Technology (CoSciTech)

p-ISSN: 2723-567X
e-ISSN: 2723-5661

http://ejurnal.umri.ac.id/index.php/coscitech/index

Komparasi Algoritma Menggunakan Teknik SMOTE Dalam Melakukan Klasifikasi Penyakit Stroke Otak

Fitri Handayani*[1], Reny Medikawati Taufiq[2]
Email: [1]fitrihandayani@umri.ac.id, [2]renymedikawati@umri.ac.id

[1,2]Teknik Informatika, Ilmu Komputer, Universitas Muhammadiyah Riau

Diterima: 12 Juli 2024 | Direvisi: 10 Agustus 2024 | Disetujui: 15 Agustus 2024
©2020 Program Studi Teknik Informatika Fakultas Ilmu Komputer,
Universitas Muhammadiyah Riau, Indonesia

Abstrak

Stroke merupakan salah satu penyakit yang mematikan. Hal ini dapat terjadi karena gangguan fungsi otak yang terjadi secara mendadak progresif dan cepat. Namun untuk mengetahui gejala dini penyakit stroke sulit untuk diketahui. Penerapan pengetahuan data mining dapat digunakan untuk mendiagnosa penyakit. Penelitian ini dilakukan untuk mengimplementasikan data mining dalam melakukan klasifikasi penyakit stroke otak. Dataset yang digunakan didapat dari Kaggle berjumlah 4891 data. Namun dataset tersebut tidak seimbang jumlah data setiap kelas. Untuk melakukan penyeimbangan data digunakan teknik SMOTE bertujuan untuk dapat meningkatkan akurasi. Penerapan algoritma klasifikasi yang digunakan yaitu algoritma *Logistic Regression* (LR), *Random Forest* (RF), *Support Vector Machine* (SVM) dan *K-Nearest Neighbors* (KNN) bertujuan untuk mengetahui performa algortima terbaik. Penelitian ini menghasilkan bahwa perbandingan dari keempat algoritma didapatkan bahwa algoritma LR, RF dan SVM menghasilkan nilai akurasi, precision, recall dan f1-score tertinggi yaitu akurasi 95%, precision 95%, recall 100% dan f1-score 97%.. Sedangkan algoritma KNN dihasilkan nilai akurasi, precision, recall dan f1-score lebih rendah yaitu akurasi 90%, precision 95%, recall 85% dan f1-score 90%.

Kata kunci: *Confusion Matrix, Logistic Regression, K-Nearest Neighbors, Random Forest, Support Vector Machine, Stroke Otak*

Comparison Of Algorithms Using The SMOTE Technique In Classification Brain Stroke Diseases

Abstract

Stroke is a deadly disease. This can occur due to disturbances in brain function that occur suddenly, progressively and quickly. However, it is difficult to know the early symptoms of stroke. The application of data mining knowledge can be used to diagnose disease. This research was conducted to implement data mining in classifying brain stroke. The dataset used was obtained from Kaggle, totaling 4891 data. However, the dataset does not have a balanced amount of data for each class. To balance the data, the SMOTE technique is used which aims to increase accuracy. The application of the classification algorithms used, namely the Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) algorithms aims to determine the best algorithm performance. This research resulted in a comparison of the four algorithms which showed that the LR, RF and SVM algorithms produced the highest accuracy, precision, recall and f1-score values, namely 95% accuracy, 95% precision, 100% recall and 97% f1-score. The KNN algorithm produces lower accuracy, precision, recall and f1-score values, namely 90% accuracy, 95% precision, 85% recall and 90% f1-score.

Keywords: *Confusion Matrix, Logistic Regression, K-Nearest Neighbors, Random Forest, Support Vector Machine, Brain Stroke*

1. PENDAHULUAN

Di dunia stroke merupakan tingkat kematian tertinggi kedua dengan 70% penderita penyakit stroke terdapat pada Negara yang berpenghasilan rendah hingga menengah. Adapun kematian dan kecacatan didapati 87% disebabkan penyakit stroke. Hal ini menunjukkan penyakit stroke merupakan penyakit dalam kasus emergensi [1].

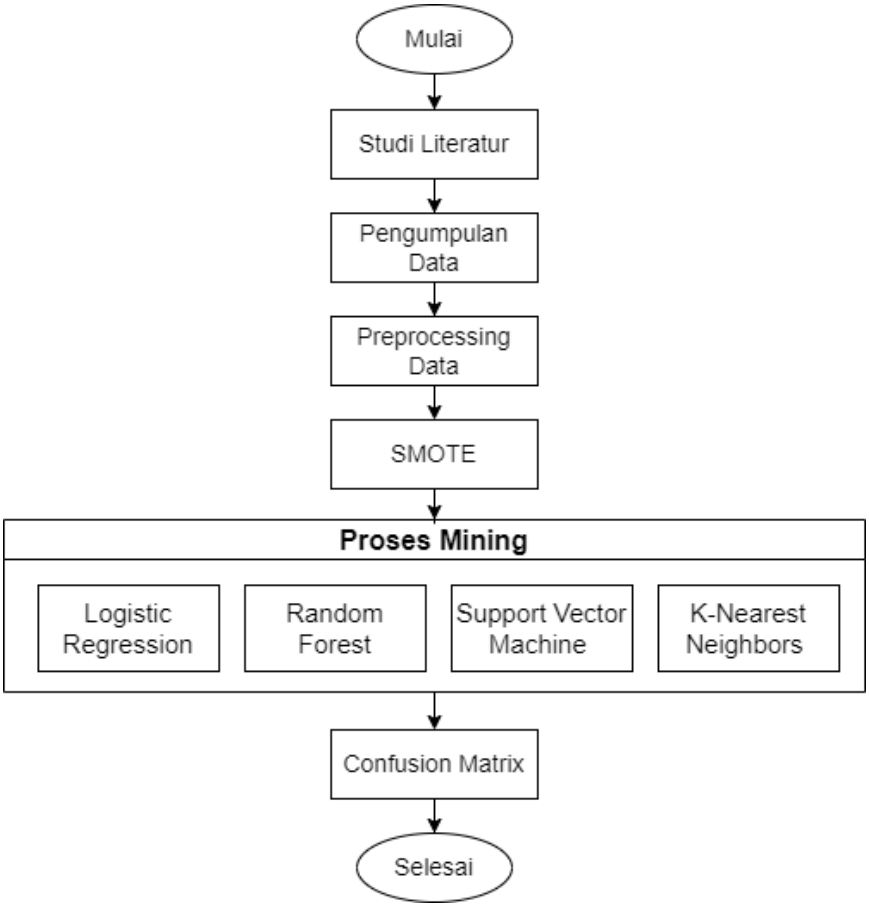
Penyakit stroke di Indonesia merupakan penyakit yang paling mematikan. Penyakit ini terjadi dikarenakan gangguan fungsi otak yang terjadi mendadak progresif dan cepat. Hal ini mengakibatkan gangguan peredaran darah otak non traumatik bekerja lebih dari 24 jam ataupun dapat berakhir dengan kematian [2]. Namun untuk mengetahui gejala dini penyakit stroke sulit untuk diketahui. Penerapan pengetahuan data mining dapat digunakan untuk mendiagnosa penyakit [1].

Data mining dapat dipahami bahwa merupakan proses dalam pengumpulan serta pengolahan data sehingga didapat informasi penting. Klasifikasi merupakan salah satu metode pada data mining untuk memprediksi kelas pada suatu objek [3]. Penelitian [4] melakukan penerapan klasifikasi penyakit stroke menggunakan metode *Support Vector Machine* (SVM) dan Logistic Regression (LR) menggunakan data 4981 baris dan 11 kolom menghasilkan akurasi terbaik pada SVM + SMOTE bernilai 95.3%. Penelitian [5] melakukan klasifikasi stroke menggunakan algoritma Logistic Regression dengan 4981 data didapat akurasi bernilai 94%. Penelitian [6] penerapan algoritma C4.5 dan teknik SMOTE (*Synthetic Minority Oversampling Technique*) untuk melakukan imbalance data dihasilkan nilai akurasi 95%. Penelitian [7] melakukan perbandingan klasifikasi beberapa algoritma pada percobaan data tidak seimbang algoritma *Logistic Regression*, *Random Forest*, *Support Vector Machine*, dan *K-Nearest Neighbors* dihasilkan akurasi 98.63% sedangkan pada data seimbang algoritma Random Forest bernilai 72.17%, SVM bernilai 76.52%, KNN bernilai 72.61% dan *Logistic Regression* 75.65%. Penelitian [8] dihasilkan bahwa algoritma Random Forest memproses jumlah data yang besar efektif mendeteksi penyakit stroke [8].

Berdasarkan beberapa penelitian yang telah dilakukan dapat dilakukan penelitian kembali menggunakan teknik SMOTE untuk mengolah data yang tidak seimbang. Penerapan teknik ini dapat berhasil mengatasi imbalance class pada dataset[8] dan meningkatkan akurasi agar dihasilkan nilai lebih baik [9]. Oleh karena itu, penelitian ini bertujuan untuk meningkatkan akurasi penelitian [7] dan untuk mengetahui performa algoritma terbaik.

2. METODE PENELITIAN

Metode penelitian merupakan tahapan yang digunakan untuk menghasilkan suatu tujuan. Adapun metode penelitian yang dilakukan terdapat pada Gambar 1:



Gambar 1 Metodologi Penelitian

2.1 Studi Literatur

Studi literatur merupakan tahapan penting dalam melakukan penelitian dimana merupakan sebagai referensi terdahulu sebagai landasan ilmiah serta sumber data [10]. Referensi tersebut bersumber dari internet terdiri dari artikel, jurnal buku ataupun e-book [1].

2.2 Pengumpulan Data

Penelitian ini menggunakan dataset yang di dapat dari Kaggle berjumlah 4891 data, 11 variabel terdiri dari jenis kelamin, usia, hipertensi, penyakit jantung, status menikah, jenis pekerjaan, jenis tempat tinggal, kadar glukosa rata-rata, bmi, status merokok, dan stroke. Jumlah kelas terdiri dari iya dan tidak. Masing-masing kelas dengan jumlah data yang tidak seimbang. Kelas iya atau terkena stroke berjumlah 248 data dan yang tidak terkena stroke berjumlah 4733 data. Berikut Tabel Dataset Brain Stroke.

Tabel 1Variabel Dataset Brain Stroke

Atribut Label		
Variabel	**Type Data**	**Nilai**
Stroke	Binominal	Yes
		No

Atribut Variabel		
Variabel	**Type Data**	**Nilai**
Gender	Binominal	Male
		Female
Age	Integer	1-82
Hypertension	Binominal	1
		0
Heart_disease	Binominal	1
		0
Ever_married	Binominal	Yes
		No
Work_type	Polynominal	Private
		Self employed
		Govt_job
		Children
Residence_type	Binominal	Urban
		Rural
Avg_glucose_level	Real	59-160
BMI	Real	12-78
Smoking_status	Polynominal	Formerly smoked
		Never smoked
		Smoked

2.3 *PreProcessing* Data

Tahapan preprocessing data bertujuan untuk mendapatkan data yang lebih efektif. Beberapa tahapan yang dilakukan yaitu:

1. Menghapus Variabel *Residence Type* dan *Work Type*

   Tahap ini dilakukan karena variabel pada dataset tidak mempengaruhi hasil penelitian klasifikasi penyakit stroke otak

2. Mengkonversi Variabel Biner ke Numerik

   Tahap ini dilakukan mengubah type data biner ke dalam bentuk numerik bertujuan agar dapat diproses dataset tersebut. Variabel tersebut terdiri dari *gender* dan *ever married*.

3. Mengkonversi Variabel Multiclass

   Tahap ini dilakukan konversi variabel yang memiliki multiclass dan mengubah bentuk type data menjadi numerik. Variabel tersebut yaitu *smoking status*.

4. Mengisi Nilai Hilang Variabel BMI

   Tahap ini dilakukan dengan menambahkan data yang hilang pada kolom variabel BMI dengan cara dilakukan perhitungan nilai rata-rata.

5. Normalisasi

   Tahap ini dilakukan dengan cara mengubah type data numerik dengan rentang sama untuk menghindari perbedaan skala sehingga setiap variabel mempunyai pengaruh sebanding pada model.

6. Pembagian Dataset

   Tahap ini dilakukan bertujuan untuk membagi data menjadi data latih dan data uji. Pembagian data latih yang digunakan dengan perbandingan 80:20.

7. Data Imbalance

   Tahap ini dilakukan proses untuk mengatasi data yang tidak seimbang jumlah kelas pada dataset yang digunakan. Tahapan ini bertujuan untuk meningkatkan akurasi. Adapun teknik yang digunakan yaitu SMOTE (*Synthetic Minority Over-sampling Technique*). SMOTE merupakan teknik pendekatan untuk melakukan penyeimbangan data sampel untuk kelas tidak seimbang (mayoritas) dengan berfokus pada kelas yang minoritas bertujuan agar efisiensi metode klasifikasi meningkat [11]. Penerapan teknik ini dilakukan oversampling dengan cara data pada kelas minoritas direplikasi menggunakan data sintetik yang didapat dari reproduksi data pada kelas minoritas. Sehingga diharapkan kelas mayoritas dapat seimbang [12]. Menurut Chawla, et al (2002) SMOTE memiliki prinsip yaitu menyetarakan kelas minor dan kelas mayor dengan cara menambah jumlah data dan membangkitkan synthesis data atau data buatan [13].

2.4 Proses Data Mining

Proses data mining merupakan tahapan melakukan klasifikasi penyakit stroke otak menggunakan beberapa algoritma yaitu:

1. *Logistic Regression* (LR)

   Logistic Regression merupakan model statistik untuk melakukan klasifikasi biner. LR menggunakan fungsi logistik untuk menghasilkan model keluaran biner [14]. Hasil keluaran akan berupa probabilitas $(0 \leq x \leq 1)$ dan digunakan untuk mempresiksi biner 0 atau 1 sebagai keluaran (jika x<0.5, keluaran = 0, jika tidak keluaran = 1) [15].

2. *Random Forest* (RF)

   Random Forest merupakan metode pembelajaran esemble dengan menggabungkan beberapa pohon keputusan sehingga didaptkan hasil akurasi yang akurat[14]. Kemampuan RF yaitu menangani data berdimensi tinggi dan mengurangi overfitting[15].

2. *Support Vector Machine* (SVM)

   Support Vector Machine merupakan metode klasifikasi melakukan proses dengan menemukan hyperplane optimal yang memisahkan atau menyesuaikan titik data dengan margin maksimum sehingga dalam menangani data berdimensi tinggi lebih efektif dan hubungan non—linier melalui fungsi kernel [14]. SVM mencakup instance klasifikasi yang dipisahkan dengan cara linier [7]. Regresi metode atau ketepatan klasifikasi pada metode SVM ditentukan berdasarkan kumpulan parameter yang digunakan optimal [16].

3. *K-Nearest Neighbors* (KNN)

   K-Nearest Neighbors merupakan algoritma klasifikasi dan regresi yang efektif. KNN menetapkan titik data baru ke kelas atau memprediksi nilai berdasarkan jumlah terbanyak atau rata-rata k tetangga terdekat [14]. Algoritma ini melakukan pengelompokkan dilihat dari jarak antar data. Untuk menghitung jarak digunakan persamaan Euclidean Distance [17]. Logika dasar KNN yaitu menjelajahi lingkungan sekitar, menganggap titik data pengujian serupa dengan titik data tersebut dan didapatkan keluaran [15]. Menurut Abdudrrahman et.al (2019) bahwa nilai k terbaik pada algoritma KNN dipengaruhi oleh data. Nilai k tertinggi akan mengurangi efek noise pada klasifikasi, akan tetapi membuat batas antar kelas semakin menjauh [18].

2.5 Confusion Matrix

Confusion matrix merupakan proses memvalidasi pengujian kinerja algoritma *machine learning* berdasarkan data uji. Indikator penilaian terdiri dari akurasi dan presisi [4].

	Positive	Negative	
Positive	True Positive (TP)	False Negative (FN)	Recall $\frac{TP}{(TP+FN)}$
Negative	False Positive (TP)	True Negative (TN)	accuracy $\frac{TP+TN}{(TP+TN+FP+FN)}$
	Precision $\frac{TP}{(TP+FP)}$	F1 score $2x\frac{precission\ x\ recall}{(precission+recall)}$	

Gambar 2 Perhitungan Confusion Matrix

3. HASIL DAN PEMBAHASAN

Hasil dan pembahasan penelitian yang telah dilakukan dilakukan beberapa percobaan algoritma diantanya:

1. Ekperimen pertama, dilakukan uji coba data seimbang menggunakan teknik SMOTE dan algoritma LR didapatkan nilai akurasi 95%, precision 95%, recall 100% dan f1-score 97%.
2. Ekperimen kedua, dilakukan uji coba data seimbang menggunakan teknik SMOTE dan algoritma RF didapatkan nilai akurasi 95%, precision 95%, recall 100% dan f1-score 97%.
3. Ekperimen ketiga, dilakukan uji coba data seimbang menggunakan teknik SMOTE dan algoritma SVM didapatkan nilai akurasi 95%, precision 95%, recall 100% dan f1-score 97%.
4. Ekperimen keempat, dilakukan uji coba data seimbang menggunakan teknik SMOTE dan algoritma KNN didapatkan nilai akurasi 90%, precision 95%, recall 85% dan f1-score 90%.

4. KESIMPULAN

Berdasarkan hasil dari penelitian yang telah dilakukan, dapat disimpulkan bahwa teknik SMOTE dapat meningkatkan nilai akurasi penelitian terdahulu pada algoritma LR, RF, SVM, dan KNN. Dimana proses pengujian menggunakan pembagian data 80:20 dan jumlah data seimbang 4733 data. Perbandingan dari keempat algoritma didapatkan bahwa algoritma LR, RF dan SVM menghasilkan nilai akurasi, precision, recall dan f1-score tertinggi yaitu akurasi 95%, precision 95%, recall 100% dan f1-score 97%.. Sedangkan algoritma KNN dihasilkan nilai akurasi, precision, recall dan f1-score lebih rendah yaitu akurasi 90%, precision 95%, recall 85% dan f1-score 90%.

DAFTAR PUSTAKA

[1] Iskandar N A, Ernawati I and Widiastiwi Y 2022 Klasifikasi Diagnosis Penyakit Stroke Dengan Menggunakan Metode Random Forest *Semin. Nas. Mhs. Ilmu Komput. dan Apl.* 676–85
[2] Riany A F and Testiana G 2023 Penerapan Data Mining untuk Klasifikasi Penyakit Stroke Menggunakan Algoritma Naïve Bayes *J. Saintekom* **9** 42–54
[3] Setiawan R 2021 Apa itu Data Mining dan Bagaimana Metodenya?
[4] Ismafillah D, Rohana T and Cahyana Y 2023 Implementasi Model Support Vector Machine dan Logistic Regression Untuk Memprediksi Penyakit Stroke *JURIKOM (Jurnal Ris. Komputer)* **10** 248–56
[5] Suhliyyah, Handayani H H and Baihaqi K A 2023 Implementasi Algoritma Logistic Regression Untuk Klasifikasi Penyakit Stroke *Syntax J. Inform.* **12** 15–23
[6] Nabila F, Afrianty I, Sanjaya S and Syafria F 2023 Implementasi Algoritma C4.5 dalam Melakukan Klasifikasi Penyakit Stroke Otak *J. Inform. Univ. Pamulang* **8** 229–35
[7] Azhar Y, Firdausy A K and Amelia P J 2022 Perbandingan Algoritma Klasifikasi Data Mining Untuk Prediksi Penyakit Stroke *Sintech J.* **5** 191–7
[8] Mualfah D, Fadila W and Firdaus R 2022 Teknik SMOTE untuk Mengatasi Imbalance Data Pada Deteksi Penyakit Stroke Menggunakan Algoritma Random Forest *J. Comput. Sci. Inf. Technol. (CoSciTech)* **3** 107–13
[9] Ayuningtyas Y and Suartana I M 2023 Klasifikasi Penyakit Stroke Menggunakan Support Vector Machine (SVM) dan Particle Swarm Optimization (PSO) *JINACS (Journal Informatics Comput. Sci.* **4** 452–7
[10] Akmal K, Faqih A and Dikananda F 2023 Perbandingan Metode Algoritma Naive Bayes dan K-Nearest Neighbors Untuk Klasifikasi Penyakit Stroke *JATI (Jurnal Mhs. Tek. Inform.* **7** 470–7
[11] Siringoringo R 2018 Klasifikasi Data Tidak Seimbang Menggunakan Algoritma SMOTE Dan K-Nearest Neighbor *J. Inf. Syst. Dev.* **3**
[12] Firdaus R, Mualfah D and Hasanah J S 2023 Klasifikasi Multi-Class Penyakit Jantung Dengan SMOTE dan Pearson's Correlation Menggunakan MLP *J. Comput. Sci. Inf. Technol. (CoSciTech)* **4** 262–71
[13] Sofyan S and Prasetyo A 2021 Penerapan Synthetic Minority Oversampling Technique (SMOTE) Terhadap Data Tidak Seimbang Pada Tingkat Pendapatan Pekerja Informal Di Provinsi D.I. Yogyakarta Tahun 2019 *Semin. Nas. Off. Stat.* **2019** 868–77
[14] Mahesh 2023 Exploring Decision Trees, Random Forest, Logistic Regression, KNN, Linear Regression, SVM, RNN, LSTM, and LightGBM for Effective Data Analysis
[15] Varghese D 2018 Comparative Study on Classic Machine learning Algorithms

[16] Maulaya A K and Junadhi 2022 Analisis sentimen menggunakan support vector machine masyarakat indonesia di twitter terkait bjorka *J. Comput. Sci. Inf. Technol.* **3** 495–500
[17] Insany G P, Yustiana I and Rahmawati S 2023 Penerapan KNN dan ANN pada Klasifikasi Status Gizi Balita Berdasarkan Indeks Antropometri *J. Comput. Sci. Inf. Technol. (CoSciTech)* **4** 385–93
[18] Oktavyani A R, Wicaksono A, Seanne A F, Nofana A D K, Putra R S and Kurniawan M 2023 Perbandingan Metode Naive Bayes, K-NN dan Decision Tree *Semin. Nas. Tek. Elektro, Sist. Informasi, dan Tek. Inform.* 276–81