

Analisa Perbandingan Algoritma Classification Berdasarkan Komposisi Label

Diterima:

10 Juni 2024

Revisi:

10 Juli 2024

Terbit:

1 Agustus 2024

^{1*}Theo Krisna Amarya, ²Asye Candra Andy G, ³Ridho Achmad,
⁴Erna Daniati, ⁵Aidina Ristyawan

¹⁻⁵Universitas Nusantara PGRI Kediri Fakultas Teknik dan Ilmu
Komputer

¹theokrisnaamarya.9i.23@gmail.com, ²man94837@gmail.com,

³andygaluh935@gmail.com, ⁴ernadaniati@unpkediri.ac.id,

⁵aidinaristi@unpkediri.ac.id

Abstrak—Di Indonesia, salah satu penyebab kematian dan gangguan neurologis paling umum adalah stroke. Tujuan penelitian ini adalah untuk membandingkan kinerja algoritma klasifikasi sebelum dan sesudah balance serta cross validation dalam mengklasifikasi penyakit stroke menggunakan dataset predict-stroke dari Kaggle. *Random Forests, KNN, Naive Bayes, Decision Trees, SVM, Neural Networks, dan Logistic Regression* adalah algoritma yang diuji. Pada penelitian ini, metode percobaan (eksperimen) digunakan. Langkah-langkah pemrosesan termasuk preprocessing, pembagian data untuk membedakan data pelatihan dan pengujian, dan evaluasi dengan cross-validation. nilai yang akan dinilai adalah akurasi, presisi, recall, dan F1. Hasilnya menunjukkan bahwa algoritma Random Forest menunjukkan tingkat akurasi tertinggi 94% setelah cross-validation dan penyeimbangan data. Meskipun, karena proses penyeimbangan, akurasi berkurang, Random Forest tetap memiliki performa yang baik dalam klasifikasi.

Kata Kunci—Data Mining; Algoritma Klasifikasi; Cross Validation

Abstract— Stroke is a prevalent cause of deaths and neurological problems in Indonesia. The goal of this study is to evaluate the efficiency of classification algorithms in accurately classifying stroke disorders. This will be done by comparing their performance before and after implementing balance and cross validation techniques. The predict-stroke dataset from kaggle will be utilized for this analysis. The algorithms studied include *Random Forests, KNN, Naive Bayes, Decision Trees, SVM, Neural Networks, and Logistic Regression*. The study employed the experimental approach. The processing phases consist of preprocessing, dividing the data into training and testing sets, and evaluating it by cross-validation. The metrics to be evaluated include accuracy, precision, recall, and F1 score. The results indicate that the Random Forest algorithm obtains the maximum accuracy rate of 94% following cross-validation and data balancing. Despite the reduction in accuracy caused by the balancing process, Random Forest still demonstrates strong performance in categorization.

Keywords—Data Mining; Classification Algorithm; Cross Validation

This is an open access article under the CC BY-SA License.



Penulis Korespondensi:

Theo Krisna Amarya,

Sistem Informasi,

Universitas Nusantara PGRI Kediri,

Email: theokrisnaamarya.9i.23@gmail.com

ID Orcid: [<https://orcid.org/0009-0005-1260-8210>]

Handphone: 081230924799

I. PENDAHULUAN

Stroke adalah suatu kondisi yang secara tiba-tiba mengganggu fungsi otak akut dalam waktu 24 jam[1]. Stroke merupakan pemicu kematian serta keadaan neurologis yang amat universal di Indonesia. Stroke dapat menyerang individu dari segala usia. Penyakit stroke lebih banyak terjadi pada individu yang memiliki Perilaku buruk termasuk merokok, minum alkohol secara berlebihan, dan kurang berolahraga, kurangnya asupan serat, dan tekanan emosional[2]. Prediksi penyakit stroke dapat difasilitasi melalui pemanfaatan teknologi, yang melibatkan pengumpulan dan pemrosesan data masa lalu untuk menghasilkan wawasan yang informatif. Namun, pengelolaan sistem atau teknologi tersebut memerlukan metode selama pemrosesan. Metode yang dimaksud mirip dengan algoritma yang cocok untuk memastikan adanya stroke pada seseorang. Dengan mengevaluasi hasil dari strategi yang berbeda, berbagai metode telah digunakan untuk meningkatkan probabilitas terjadinya stroke. Beberapa metode klasifikasi digunakan dalam bidang penggalian data. Salah satu contoh studi tersebut yakni studi yang dicoba oleh Dikan Ismafillah, Tatang Rohana, dan Yana Cahyana dari Jurnal Riset Komputer (JURIKOM) pada tahun 2023. Penelitian tersebut berjudul “Implementasi Model *Support Vector Machine* dan *Logistik Regresi* untuk Memprediksi Penyakit Stroke”. Tujuan dari penelitian tersebut adalah untuk menentukan model dengan tingkat akurasi tertinggi. Hasil penelitian menunjukkan bahwa model *Support Vector Machine* + *SMOTE* mencapai akurasi sebesar 95,3%, sedangkan model Logistik Regression + *SMOTE* mencapai akurasi sebesar 93,1%[3].

Algoritma *classification* data mining yang dimaksud adalah KNN (*K-Nearest Neighbors*), *Naive Bayes*, *Neural Network* (Jaringan Saraf Tiruan), *Decision Tree*, *Support Vector Machine* (SVM), serta *Random Forest*. Penulis menggunakan dataset *stroke-predict*, yang dapat diunduh dari <https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>[4], untuk penelitian ini. Dataset ini mencakup parameter seperti jenis kelamin, usia, tekanan darah, riwayat jantung, jenis posisi, kadar glukosa rata-rata, berat badan, status merokok, dan status pernikahan. Algoritma klasifikasi yang disebutkan di atas akan diimplementasikan dengan menggunakan perangkat Jupyter Notebook versi 7.0.8. Penulis mengantisipasi bahwa penelitian ini akan menghasilkan wawasan baru dalam penerapan algoritma klasifikasi dalam memprediksi penyakit stroke, sehingga dapat memberikan kontribusi terhadap kemajuan dalam klasifikasi mesin dan perawatan kesehatan

II. METODE

Dalam penelitian ini, metode eksperimen digunakan untuk menguji keakuratan setiap model algoritma klasifikasi untuk memprediksi stroke[5]. Dengan Langkah – langkah sebagai berikut.

- A. Pengumpulan Data: Untuk penelitian ini, data digunakan dari dataset *stroke-predict*, yang dapat diunduh dari <https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>[4]. Pada langkah pertama, data ini dianalisis dan disiapkan untuk pemodelan.
- B. Pre-processing: Mengecek dan memperbaiki kolom dengan nilai yang terduplikat dan nilai data yang nihil (tidak terisi). Untuk melakukan perataan atau perimbangan data, menggunakan modul Imblearn.
- C. Pembagian Data: Setelah proses pre-processing selesai, data akan dibagi menjadi dua bagian: data pelatihan dan data uji. Sebagian besar dataset akan digunakan untuk pelatihan model, dan sebagian kecil lagi akan digunakan untuk pengujian model dengan ukuran $train\ size = 0.7$ dan $test\ size = 0.3$.
- D. Evaluasi Model: Setelah pelatihan, model yang dihasilkan dari evaluasi diuji dengan berbagai data pengujian. Menilai performa model masing-masing algoritma klasifikasi melibatkan penggunaan metrik evaluasi seperti akurasi, presisi, recall, dan skor f1. Selanjutnya, hasil performa akan ditampilkan dalam tabel.

2.1 K-Nearest Neighbors

Algoritma KNN adalah Metode klasifikasi objek didasarkan pada informasi pendidikan yang setidaknya mirip dengan objek yang tengah diuji. Hasil dari komputasi hendak digabungkan dengan informasi pelatihan yang menampilkan jumlah saudara paling tinggi yang terletak dalam rentang nilai yang sudah diresmikan lebih dahulu. Persamaan Eclidean menghitung jarak antara informasi uji serta informasi pelatihan[6]. Pada penelitian ini penulis menggunakan nilai $n_neighbors = 5$.

2.2 Naïve Bayes

Metode klasifikasi didasarkan pada teori Bayes. Teori ini berasal dari metode probabilitas yang diusulkan oleh ilmuwan Inggris Thomas Bayes, yang memungkinkan prediksi masa depan berdasarkan pengalaman sebelumnya. Ciri metode Naive Bayes adalah asumsi yang kuat (Naive) terhadap independensi parameter. Salah satu kelebihan metode ini adalah bahwa itu tidak membutuhkan data set yang besar[7].

2.3 Decision Tree

Decision Tree adalah Struktur *flowchart* yang mempunyai pohon, atau pohon, disebut divisi pohon. Setiap simpul internal menunjukkan tes atribut, setiap cabang menunjukkan hasil, dan simpul daun menunjukkan distribusi kelas[8].

2.4 Support Vector Machine

Untuk pertama kalinya, *Support Vector Machine* (SVM) dibuat oleh Vapnik pada tahun 1992 sebagai set konsep harmonis, yang merupakan konsep unggulan dalam pengidentifikasi pola. Dengan menggunakan prinsip *Strukyural Risk Minimazation* (SRM), SVM berusaha menemukan hyperlane terbaik yang dapat memisahkan dua kelas dan ruang input[9].

2.5 Neural Network

Jaringan syaraf telah digunakan untuk aktivitas seperti klasifikasi, pengelompokan, prediksi, penambangan pola, dan mengenali pola. Jaringan syaraf bekerja sebagai analogi untuk otak dalam hal pemrosesan informasi[10].

2.6 Logistic Regression

Logistic Regression adalah teknik statistik yang merupakan bagian dari analisis regresi yang umumnya digunakan untuk memprediksi kelas biner[11], Selain itu, digunakan untuk menentukan kemungkinan terjadinya kejadian biner dan menyelesaikan masalah klasifikasi[12].

2.7 Random Forest

Algoritma pembelajaran ansambel *Random Forest* menggunakan kumpulan pohon keputusan untuk mengurangi variasi dalam prediksi, sehingga meningkatkan ketepatan prediksinya. Algoritma pohon keputusan memilih kumpulan variabel acak. Hutan tanpa aturan ini disebut hutan tanpa aturan[13]. *Random Forest* telah diakui sebagai algoritma klasifikasi terbaik karena kemampuannya menciptakan nilai yang sangat akurat. *Random Forest* juga dapat mengidentifikasi karakteristik penting dalam data dan menangani data yang tidak seimbang secara signifikan jika dibandingkan dengan model lain[14].

2.8 Cross Validation

Cross-Validation (CV) adalah metode untuk menilai dan menguji kinerja (akurasinya) model Pengajaran Mesin. Dalam proses CV, sampel data asli dibagi secara acak menjadi beberapa subset, dan model Pengajaran Mesin dilatih pada setiap subset kecuali satu subset yang akan digunakan untuk menguji model dalam membuat prediksi [15].

III. HASIL DAN PEMBAHASAN

Langkah pertama adalah membaca dataset menggunakan library *pandas*, kemudian dilakukan pengisian nilai yang hilang, serta penghapusan nilai yang sama pada kolom "bmi", label pada dataset ini adalah kolom "stroke".

Keterangan Tabel:

A = id E = penyakit jantung I = rata – rata kadar gula
 B = jenis kelamin F = status pernikahan J = bmi
 C = usia G = pekerjaan K = status merokok
 D = tekanan darah H = jenis tempat tinggal L = stroke

Tabel 1. Data Stroke

A	B	C	D	E	F	G	H	I	J	K	L
9046	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
51676	Female	61	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
31112	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
60182	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
1665	Female	79	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1

Pada Tabel 1, terdapat 4 *record* data dari 5110 *record* data dengan jumlah 12 kolom, terdapat sebanyak 201 data pada kolom "bmi" kosong yang akan penulis isi dengan nilai rata – rata dari kolom "bmi". Langkah berikutnya adalah melakukan penghapusan pada kolom "id", seperti yang ditampilkan pada Tabel 2

Tabel 2. Penghapusan Kolom Id

A	B	C	D	E	F	G	H	I	J	K
Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
Female	61	0	0	Yes	Self-employed	Rural	202.21	28.8	never smoked	1
Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
Female	79	1	0	Yes	Self-employed	Rural	174.12	24	never smoked	1

Pada Tabel 2, terdapat 5 *record* data dari 5110 *record* data dan *missing value* (data kosong) pada kolom "bmi" sudah terisi, dengan 11 kolom. Selanjutnya mengubah nilai pada kolom *gender* (jenis kelamin), *ever_married* (status pernikahan), *work_type* (jenis pekerjaan), *Residence_type* (tipe tempat tinggal), dan *smoking status* (status merokok) diubah menjadi bentuk numerik menggunakan teknik *Labelencoder*. Hal ini perlu sebab beberapa algoritma classification hanya dapat mengelola data berupa numerik, data dapat dilihat pada Tabel 3.

Tabel 3. Transformasi Data

A	B	C	D	E	F	G	H	I	K	L
1	67	0	1	1	2	1	228.69	36.6	1	1
0	61	0	0	1	3	0	202.21	28.8	2	1
1	80	0	1	1	2	0	105.92	32.5	2	1
0	49	0	0	1	2	0	171.23	34.4	3	1
0	79	1	0	1	3	1	174.12	24	2	1

Pada table 3, nilai dari kolom “jenis kelamin (B)” sudah menjadi numerik dimana “1” untuk pria dan “0” untuk Wanita. Berikutnya melakukan pelatihan data (*data training*) dengan perbandingan *train_size = 0.7* dan *test_size* 0.3, pada data yang sudah melalui proses *Replace Missing Value*, namun belum dilakukan proses *balance data*. Hasil akurasi dari pelatihan data dapat dilihat pada tabel 4.

Tabel 4. Hasil Akurasi Sebelum Balance Data dan *Cross Validation*

No	Algoritma	Sebelum Balancing Data dan Cross Validation				
		Akurasi	Presisi	Recall	F1 – Score	Average
1	K-Nearest Neighbors	94%	91%	94%	92%	93%
2	Naïve Bayes	88%	88%	88%	90%	89%
3	Decission Tree	92%	91%	92%	91%	92%
4	Support Vector Machine	95%	90%	95%	92%	93%
5	Neural Network	95%	90%	95%	92%	93%
6	Logistic Regression	95%	90%	95%	90%	93%
7	Random Forest	95%	90%	95%	92%	93%

Pada Tabel 4, terdapat 4 algortima dengan akurasi tertinggi yaitu *Support Vector Machine*, *Neural Network*, *Logistic Regression*, *Random Forest* dengan nilai akurasi sebesar 95%. Meskipun akurasinya menunjukkan nilai 95%, hasil traning data ini hanya dapat menilai data mayor (mayoritas) pada kolom ”*stroke*” dengan nilai “No” atau angka “0” dengan jumlah data sebanyak 4861. Sedangkan bayak data orang yang mengidap penyakit stroke sebesar 249, jika dilakukan perbandingan maka dihasilkan sebesar “No” sebanyak 95.13% sedangkan “Yes” sebanyak 4.87%. Untuk menghindari terjadi kesalahan klasifikasi jika seseorang mengidap penyakit stroke namun sistem mengklasifikasi bahwa orang tersebut tidak mengidap penyakit stroke, maka penulis melakukan proses *balancing data* dengan metode *hybrid*. *Hybrid* adalah sebuah metode *balance data* gabungan dari *oversampling* dan *under sampling*, serta penulis melakukan *cross validation*

untuk menentukan algoritma klasifikasi terbaik. Hasil perbandingan akurasi sebelum dan setelah dilakukan balance data serta *cross validation* dapat dilihat pada Tabel 5.

Tabel 5. Perbandingan Sebelum dan Setelah Balance data serta *Cross Validation*

No	Algoritma	Sebelum Balancing Data dan Cross Validation					Setelah Balancing Data dan Cross Validation				
		Akurasi	Presi	Recall	F1-Score	Average	Akurasi	Presi	Recall	F1-Score	Average
1	K-Nearest Neighbors	94%	91%	94%	92%	93%	89%	90%	89%	89%	89%
2	Naïve Bayes	88%	88%	88%	90%	89%	79%	79%	79%	79%	79%
3	Decision Tree	92%	91%	92%	91%	92%	91%	91%	91%	91%	91%
4	Support Vector Machine	95%	90%	95%	92%	93%	77%	77%	77%	77%	77%
5	Neural Network	95%	90%	95%	92%	93%	83%	83%	83%	83%	83%
6	Logistic Regression	95%	90%	95%	90%	93%	81%	81%	81%	81%	81%
7	Random Forest	95%	90%	95%	92%	93%	94%	94%	94%	94%	94%

Pada Tabel 5, Setelah dilakukan proses balancing data menggunakan metode *hybrid* dapat diketahui bahwa nilai akurasi, presisi, recall, dan f1 score sudah seimbang. Sehingga didapatkan nilai akurasi algoritma *K-Nearest Neighbors* sebesar 89% dengan simpangan baku sebesar 0.00, *Naive Bayes* sebesar 79% dengan simpangan baku sebesar 0.01, *Decision Tree* sebesar 91% dengan simpangan baku sebesar 0.01, *Support Vector Machine* sebesar 77% dengan simpangan baku sebesar 0.01, *Neural Network* sebesar 83% dengan simpangan baku sebesar 0.03, *Logistic Regression* sebesar 81% dengan simpangan baku sebesar 0.01, dan *Random Forest* sebesar 94%

dengan simpangan baku sebesar 0.02. Dari perbandingan nilai akurasi 7 algoritma yang dibahas diketahui bahwa algoritma terbaik untuk klasifikasi penyakit stroke adalah algoritma *Random Forest* dengan tingkat akurasi sebesar 94%.

IV. KESIMPULAN

Hasil dari pengujian yang dilakukan pada dataset *stroke-predict*, dilakukannya proses balancing data dengan metode *hybrid*, membuat nilai dari *akurasi*, *presisi*, *recall*, dan *f1 score* tiap algoritma sudah seimbang namun akurasi rata-ratanya menurun. Dari hasil akurasi rata-rata sebelum dan setelah proses balance data dan cross validation, ditemukan algoritma terbaik untuk klasifikasi penyakit stroke yaitu algoritma *Random Forest*. Setelah proses balancing data nilai akurasi rata-rata *Random Forest* naik sebesar 1% yang awalnya 93% menjadi 94% dengan simpangan baku (*standart deviasi*) sebesar 0.02 yang artinya algoritma *Random Forest* memiliki tingkat kesalahan sebesar 2%. Rekomendasi untuk program penelitian kedepannya dapat mencoba menerapkan proses *data cleanning*, *preprocessing* yang lebih baik lagi, dan menambahkan beberapa metode serta output seperti waktu eksekusi (*run time*) tercepat.

DAFTAR PUSTAKA

- [1] M. Fadli and R. A. Saputra, "KLASIFIKASI DAN EVALUASI PERFORMA MODEL RANDOM FOREST UNTUK PREDIKSI STROKE Classification And Evaluation Of Performance Models Random Forest For Stroke Prediction," vol. 12, [Online]. Available: <http://jurnal.umt.ac.id/index.php/jt/index>
- [2] W. Riyadina and E. Rahajeng, "Determinan Penyakit Stroke," *Kesmas: National Public Health Journal*, vol. 7, no. 7, p. 324, Feb. 2013, doi: 10.21109/kesmas.v7i7.31.
- [3] D. Ismafillah, T. Rohana, and Y. Cahyana, "Implementasi Model Support Vector Machine dan Logistic Regression Untuk Memprediksi Penyakit Stroke," *Jurnal Riset Komputer*, vol. 10, no. 1, pp. 2407–389, 2023, doi: 10.30865/jurikom.v10i1.5478.
- [4] fedesoriano, "predict-stroke-dataset," <https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>.
- [5] R. W. Putri, A. Ristyawan, and M. N. Muzaki, "Comparison Performance of K-NN and NBC Algorithm for Classification of Heart Disease," *JTECS : Jurnal Sistem Telekomunikasi Elektronika Sistem Kontrol Power Sistem dan Komputer*, vol. 2, no. 2, p. 143, Jul. 2022, doi: 10.32503/jtecs.v2i2.2708.
- [6] F. Sutomo *et al.*, "OPTIMIZATION OF THE K-NEAREST NEIGHBORS ALGORITHM USING THE ELBOW METHOD ON STROKE PREDICTION," *Jurnal Teknik Informatika (JUTIF)*, vol. 4, no. 1, 2023, doi: 10.20884/1.jutif.2023.4.1.839.
- [7] D. Hindarto, R. Eko Indrajit, and E. Dazki, "Perbandingan Kinerja Akurasi Klasifikasi K-NN, NB Dan DT Pada APK Android," vol. 9, no. 1, pp. 486–503, 2022, [Online]. Available: <http://jurnal.mdp.ac.id>
- [8] R. Estian Pambudi, Sriyanto. And Frimasyah, "Klasifikasi Penyakit Stroke Menggunakan Algoritma Decision Tree C.45" *Jurnal Teknika*, vol. 16, No. 02, pp. 221-226, Jul, 2022.
- [9] Fuad Nur Hasan and M. Wahyudi Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri, "ANALISIS SENTIMEN ARTIKEL BERITA TOKOH SEPAK BOLA DUNIA MENGGUNAKAN ALGORITMA SUPPORT VECTOR MACHINE DAN NAIVE BAYES BERBASIS PARTICLE SWARM OPTIMIZATION," *Jurnal AKRAB JUARA*, vol. 3, no. 4, pp.42 – 55, Nov, 2018.

- [10] S. Agustin *et al.*, “Optimasi Feature Selection Menggunakan Algoritma Neural Network Untuk Klasifikasi Brain Stroke,” *Jurnal Penelitian Rumpun Ilmu Teknik (JUPRIT)*, vol. 2, no. 3, pp. 66–74, 2023, doi: 10.55606/juprit.v2i3.2009.
- [11] R. Tyasnurita and A. Y. M. Pamungkas, “Deteksi Diabetik Retinopati menggunakan Regresi Logistik,” *ILKOM Jurnal Ilmiah*, vol. 12, no. 2, pp. 130–135, Aug. 2020, doi: 10.33096/ilkom.v12i2.578.130-135.
- [12] D. Ismafillah, T. Rohana, and Y. Cahyana, “Implementasi Model Support Vector Machine dan Logistic Regression Untuk Memprediksi Penyakit Stroke,” *Jurnal Riset Komputer*, vol. 10, no. 1, pp. 2407–389, 2023, doi: 10.30865/jurikom.v10i1.5478.
- [13] P. Wahyu Setiyo Aji, R. Dijaya, and F. Sains dan Teknologi, “KESATRIA: Jurnal Penerapan Sistem Informasi (Komputer & Manajemen) Prediksi Penyakit Stroke Menggunakan Metode Random Forest.”
- [14] N. Sharfina, and N. Ghaniaviyanto Ramadhan, “Analisis SMOTE Pada Klasifikasi Hepatitis C Berbasis Random Forest dan Naïve Bayes,” *JOINTECS : Jurnal of Information Technology and Computer Science*, vol.8, no.1, pp.33-40, 2023.
- [15] Fadellia Azzahra, N. Suarna, and Y. Arie Wijaya, “Penerapan Algoritma Random Forest Dan Cross Validation Untuk Prediksi Data Stunting,” *Kopertip : Jurnal Ilmiah Manajemen Informatika dan Komputer*, vol. 8, no. 1, pp. 1–6, Feb. 2024, doi: 10.32485/kopertip.v8i1.238.