Jurnal Penelitian Rumpun Ilmu Teknik (JUPRIT)
Vol.2, No.4 November 2023
e-ISSN: 2963-7813; p-ISSN: 2963-8178, Hal 155-164
DOI: https://doi.org/10.55606/juprit.v2i4.3039

Implementasi Algoritma Random Forest Dalam Klasifikasi Diagnosis Penyakit Stroke

Ary Prandika Siregar, Dwi Priyadi Purba , Jojor Putri Pasaribu , Khairul Reza Bakara
Universitas Negeri Medan

Jl. Willem Iskandar Pasar V Medan Estate
Email : [1] aryprandika0902@gmail.com, [2] purbaa666@gmail.com, [3] jojorputripasaribu@gmail.com
[4] khairulrezabakara@gmail.com

Abstract. *The most common disease in Indonesia is stroke, this disease occurs when blood flow to the brain is disrupted, either due to rupture of blood vessels or due to blockage of blood vessels. The data mining process can be a solution in identifying early symptoms of stroke. By using the Random Forest Method, it is hoped that it can be the right choice for preprocessing data in identifying early symptoms. The model results produce an adjustment of 96% of the training score and from the results table of precision, recall, F1-score, and accuracy which results in an accuracy of 0.95 or 95%, as well as the final result of AUC of 0.80 which shows that the model results are included in the good classification*

Keywords: *Classification, Random Forest, Stroke.*

Abstrak. *Penyakit yang paling umum terjadi di indonesia adalah stroke, penyakit ini terjadi ketika aliran darah ke otak terganggu, baik karena pecahnya pembulu darah atau karena sumbatan pembulu darah. Proses Data mining bisa menjadi solusi dalam mengidentifikasi gejala dini stroke. Dengan menggunakan Metode Random Forest di harapkan bisa menjadi pilihan tepat dalam melakukan preprosessing data dalam mengidentifikasi gejala awal. Hasil model penyesuaian menghasilkan 96% skor pelatihan dan dari tabel hasil precision, recall, F1-score, dan accuracy Yang mendapatkan hasil akurasi sebesar 0.95 atau 95%, serta hasil akhir dari AUC sebesar 0.80 yang menunjukan hasil model tersebut termasuk ke dalam klasifikasi baik*.

Kata kunci: Klasifikasi, Random Forest, Stroke.

LATAR BELAKANG

Salah satu penyakit yang paling umum di Indonesia adalah stroke, yang menjadi penyebab kematian tertinggi kedua setelah diabetes dan hipertensi. Stroke adalah salah satu kondisi yang memiliki banyak konsekuensi kesehatan yang serius dan berpotensi fatal. Penyakit ini terjadi ketika aliran darah ke otak terganggu, baik karena pecahnya pembuluh darah atau karena sumbatan pembuluh darah. Stroke dapat menyebabkan kerusakan permanen pada otak, jadi sangat penting untuk mendeteksi dan mengklasifikasikan stroke dengan benar.

Metode Random Forest adalah salah satu algoritma machine learning yang termasuk dalam kategori ensemble learning. Ensemble learning melibatkan penggabungan hasil dari beberapa model untuk meningkatkan kinerja dan ketepatan prediksi dibandingkan dengan penggunaan satu model tunggal. Dalam konteks Random Forest, model yang digunakan adalah pohon keputusan (decision trees). Saat ini, teknologi sedang berkembang pesat. Perkembangan teknologi sangat membantu komunitas medis. Salah satunya adalah program yang menggunakan AI untuk mendeteksi stroke. Machine learning adalah salah satu bidang kecerdasan yang dapat digunakan. Untuk memungkinkan mesin untuk belajar secara otomatis,

Received September 07, 2023; Revised Oktober 22, 2023; Accepted November 04, 2023
* Ary Prandika Siregar, aryprandika0902@gmail.com

cabang kecerdasan buatan yang disebut machine learning dibuat. Machine learning berkonsentrasi pada analisis data untuk menentukan hubungan yang diinginkan antara input dan output. Random forest adalah salah satu algoritma pengajaran mesin yang paling terkenal. Ini adalah kombinasi pohon klasifikasi yang bekerja secara independen dan berasal dari distribusi yang sama, yang kemudian menghasilkan hasil akhir melalui proses pemungutan suara.

Sudah banyak penelitian yang dilakukan diantaranya: Menurut jurnal Prediksi Penyakit Stroke Menggunakan Metode Random Forest Penelitian ini berhasil melalui semua tahapan, mulai dari pra-pemrosesan hingga pemrosesan serta evaluasi. Hasil yang diperoleh dari penelitian ini menunjukkan tingkat akurasi sebesar 99%. Penelitian lainnya pada jurnal Klasifikasi Diagnosis Penyakit Stroke Dengan Menggunakan Metode Random Forest Metode Random Forest dapat digunakan untuk mendiagnosis penyakit stroke, pada penelitian ini dihasilkan nilai accuracy dengan selisih yang sedikit pada setiap penggunaan jumlah pohon yang berbeda, serta menghasilkan nilai precision, sensitivity dan f-measure yang cukup berbeda. Penggunaan jumlah pohon yang banyak pada penelitian ini tidak membuat nilai accuracy meningkat cukup banyak dan mengakibatkan proses komputasi menjadi lebih lama. Model dengan penggunaan jumlah pohon 90 menghasilkan nilai yang optimal, dengan memperoleh nilai accuracy 95.2%, nilai sensitivity 4.1%, nilai specificity 99.8%, nilai precision 66.7%, dan nilai F-measure 7.6%. Serta nilai ROC Curve 0.8048 yang menunjukkan bahwa model masuk ke dalam Good Classification.

Berdasarkan penjelasan yang telah disampaikan, akan dilakukan penelitian tentang Implementasi Algoritma Random Forest dalam Klasifikasi Diagnosis Penyakit stroke. Data yang didapatkan berasal dari website Kaggle.com. harapan dari penelitian ini adalah dapat membantu kalangan medis untuk dengan mudah mendiagnosa seseorang terkena penyakit stroke. Penanganannya lebih cepat jika penyakit terdeteksi lebih awal.

KAJIAN TEORITIS

Stroke merupakan penyakit yang berbahaya karena dapat menyebabkan kematian. Penyakit ini terjadi karena aliran darah ke otak terhambat atau terganggu sehingga menyebabkan kerusakan sel-sel otak. Oleh karena itu, harus diatasi secepat mungkin dengan pengobatan yang efektif dan efisien. Menurut Organisasi Kesehatan Dunia (WHO) yang dikutip oleh Junaidi (2011), Stroke adalah suatu sindrom klinis, khususnya gejala disfungsi otak fokal atau global, yang menyebabkan kecacatan atau berujung pada kematian, kematian yang berlangsung selama 24 jam tanpa sebab lain, kecuali gangguan pembuluh darah. Faktor

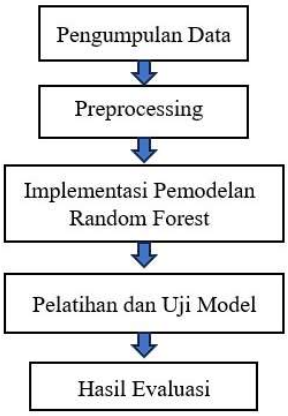
penyebab stroke antara lain jenis kelamin, usia, genetika, ras, hipertensi, diabetes, merokok, dan penyakit jantung. Dan gaya hidup yang buruk dan tidak sehat secara fisik dapat menyebabkan diagnosis stroke.

Random forest adalah algoritma machine learning yang menggunakan kombinasi pohon keputusan untuk membuat prediksi yang akurat guna menentukan cara yang lebih tepat dalam memproses data. Kelebihan Random Forest adalah dapat menangani kumpulan data yang besar dengan banyak fitur yang beragam untuk dapat mengolah data tersebut dengan baik serta mengatasi masalah overfitting yang dapat terjadi pada pohon hutan keputusan tunggal. Dan dapat menjaga stabilitas kinerja yang tinggi dan baik.

METODE PENELITIAN

Secara garis besar penelitian ini mencakup rangkaian langkah yang diambil selama pelaksanaan penelitian ini. Proses penelitian dapat kita lihat dalam gambar di bawah ini.

Gambar 1. Tahapan Penelitian



Pengumpulan Data

Pada tahap pengumpulan data dilakukan yaitu menyiapkan dataset, data yang digunakan bersumber dari Kaggle Link: https://www.kaggle.com/code/mechatronixs/machine-learning-and-modeling-stroke-data/input. Dataset yang digunakan adalah Stroke Prediction Dataset dengan nama file tersebut healthcare-dataset-stroke-data.csv. Data yang digunakan terdiri dari 40910 record. Dataset ini terdiri dari 11 indeks stroke. Di bawah ini adalah detail indikator atau atribut kumpulan data yang digunakan.

Tabel 1. Atribut dari Dataset

Atribut	Keterangan

Gender	"Pria", "Wanita" atau "Lainnya"
Age	Umur/ Usia
Hypertension	Pernah menderita hipertensi (1) atau tidak (0)
Heart_disease	Memiliki penyakit jantung apa pun (1) atau tidak memiliki penyakit jantung (0)
Ever_married	Menikah (1) atau tidak menikah (0)
Work_type	Tipe pekerjaan
Residence_type	Area tempat tinggal pasien, perkotaan (1) atau pedesaan (0)
Avg_glucose_level	Rata-rata kadar glukosa dalam darah
BM	Indeks Massa Tubuh
Smoking_status	Merokok (1) atau tidak pernah merokok (0)
Stroke	Apakah pasien mengalami stroke (1) atau tidak (0)

Preprocessing

Data asli harus melalui tahap preprocessing sebelum dapat digunakan oleh sistem. Oleh karena itu, beberapa langkah Preprocessing harus diterapkan untuk mengubah beberapa data guna meningkatkan kualitasnya. Preprocessing dilakukan untuk membersihkan data, menghilangkan noise dan nilai yang hilang sebelum melakukan langkah pemodelan.

Di antara 12 atribut tersebut terdapat satu atribut yang tidak digunakan pada saat pengolahan yaitu atribut ID, sehingga pada penelitian ini hanya menggunakan 11 atribut. Selama tahap prapemrosesan data, pembersihan dan transformasi data dilakukan. Pembersihan data menghapus data dari nilai yang hilang dan outlier. Pemisahan data train dan data test menggunakan pemisahan data, yaitu 80% untuk data train saat melatih model dan 20% untuk data uji saat menguji model.

a) Handling Missing Value

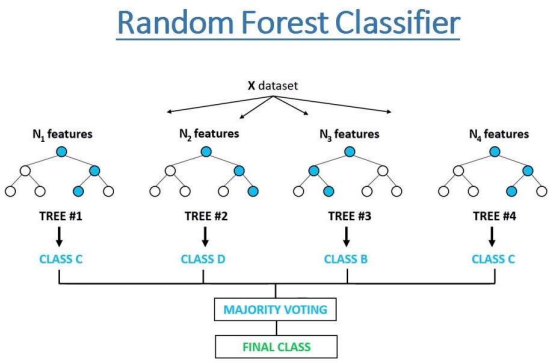
Pada tahap preprocessing awal, data masukan diproses terlebih dahulu dengan nilai yang hilang (missing value). Missing value dapat terjadi karena kesalahan input data atau data yang hilang. Karena algoritma machine learning tidak dapat menangani data dengan nilai yang hilang, sebelum melakukan pemodelan, nilai yang hilang harus ditangani terlebih dahulu. Peneliti menggunakan rata-rata 4.444 teknik amputasi. Dengan demikian, data yang nilainya hilang akan diisi dengan nilai rata-rata pada kolom tersebut.

b) Normalisasi Data

Normalisasi data merupakan salah satu teknik penting yang dilakukan pada tahap preprocessing. Hal ini karena data seringkali memiliki rentang nilai antar variabel yang sangat luas. Pada penelitian ini digunakan metode minmax scaling untuk melakukan normalisasi data. MinMax Scaler adalah metode preprocessing data dalam analisis data yang digunakan untuk mengubah nilai fitur dalam suatu kumpulan data agar terdistribusi dalam rentang 0 hingga 1.

Implementasi Pemodelan Random Forest

Langkah utama penelitian ini adalah mengimplementasikan model klasifikasi menggunakan Random Forest. Metode Random Forest merupakan metode yang berbasis pada pohon keputusan, pada saat pelatihan Random Forest akan dibuat banyak pohon keputusan sehingga dari sampel yang ada pada set pelatihan akan menghasilkan beberapa pohon [10].



Gambar 3.1. Random Forest

Random Forest memerlukan kombinasi beberapa pohon keputusan untuk memprediksi hasil secara akurat. Saat menggunakan random forest sebagai pengklasifikasi, setiap pohon keputusan dapat menghasilkan jawaban yang sama atau berbeda. Misalnya pohon keputusan A, B, E dan F memprediksi hasil 1. Sedangkan pohon keputusan C dan D memprediksi hasil 0. Karena banyak alternatif jawaban di pohon keputusan dan probabilitasnya tinggi maka random forest mengambil hasil prediksi tersebut. Hasil dari beberapa pohon keputusan berdasarkan suara mayoritas dan prediksi hasil yang lebih akurat.

Pelatihan dan Uji Model

Selama tahap pengujian model menggunakan data uji (data test). Tahap ini memprediksi nilai target untuk setiap baris data pengujian. Klasifikasi data uji dilakukan dengan memasukkan dan membandingkan nilai kolom dengan pohon terlatih untuk mengambil keputusan akhir berdasarkan nilai yang paling sering muncul (suara terbanyak).

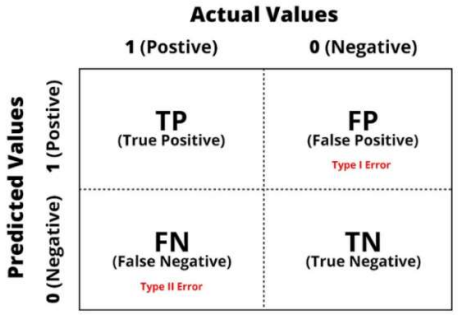
Setelah tahap pemisahan data X dan y, langkah selanjutnya adalah membagi data X menjadi data latih (train) dan data uji (test) dengan menggunakan modul 'train_test_split' dari perpustakaan scikit-learn. Pemisahan dilakukan dengan perbandingan 80% untuk data pelatihan dan 20% untuk data pengujian. Untuk mendapatkan parameter yang optimal, peneliti menerapkan teknik pencarian stokastik dengan validasi silang. Metode ini digunakan untuk mencari parameter terbaik untuk algoritma yang digunakan dalam analisis kasus. Pengacakan dengan validasi silang adalah bagian dari perpustakaan scikit-learn yang bertujuan untuk memvalidasi beberapa model dan hyperparameter individual secara otomatis dan sistematis.

Setelah melakukan pencarian acak dengan validasi silang, akan dibuat model dengan hasil pelatihan dan pengujian.

Tahap Evaluasi Hasil

Pada tahap ini, Confusion matrix yang digunakan untuk mengukur performa model machine learning akan dibuat. Confusion matrix adalah tabel yang digunakan untuk mengevaluasi performa model klasifikasi machine learning. Matriks ini menyajikan ringkasan hasil prediksi model pada sebuah dataset, memungkinkan kita mengevaluasi seberapa akurat atau salah model dalam mengklasifikasikan kumpulan data.

Gambar 3.2 Confusion Matrix



Berdasarkan Confusion Matrix, Berikut Rumus untuk menghitung matrik evaluasi klasifikasi:

- **Accuracy (Akurasi):** $\frac{TP+T}{TP+TN+FP+FN}$
- **Precision (Presisi):** $\frac{TP}{TP+F}$
- **Recall (Sensitivitas atau True Positive Rate):** $\frac{TP}{TP+}$
- **F1-Score:** $2 \times \frac{Precision \times Recall}{Precision+Recall}$

Terdapat 3 istilah sebagai representasi hasil proses klasifikasi pada confusion matrix. Dengan adanya perhitungan dari confusion matrix maka dapat diperoleh Accuracy, Precision, Recall dan F1 Score.

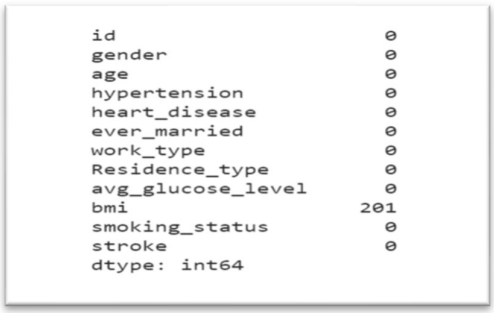
HASIL DAN PEMBAHASAN

Preprocessing

Langkah ini dilaksanakan untuk memproses data sehingga siap untuk tahap pemodelan. Hasil dari proses preprocessing adalah sebagai berikut.

a) Handling Missing Value

Berikut adalah penjelasan tentang data mentah yang masih mengandung nilai yang hilang di dalamnya.

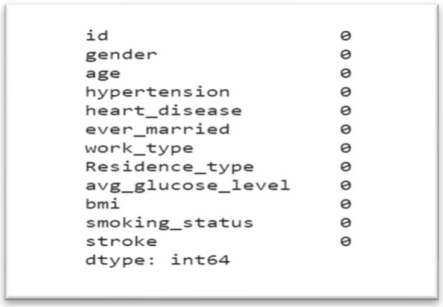


```
id                   0
gender               0
age                  0
hypertension         0
heart_disease        0
ever_married         0
work_type            0
Residence_type       0
avg_glucose_level    0
bmi                201
smoking_status       0
stroke               0
dtype: int64
```

Gambar 4.1. Deskripsi data

Dari gambar 4.1 dapat kita ketahui bahwa kolom “bmi” terdapat data yang hilang. Deskripsi tersebut menunukkan bahwa 201 baris pada data tidak terisi ataupun missing. Oleh karena itu dilakukan handling missing value agar baris pada kolom “bmi” terisi seluruhnya.

Gambar 4.2. Deskripsi data setelah handling missing value

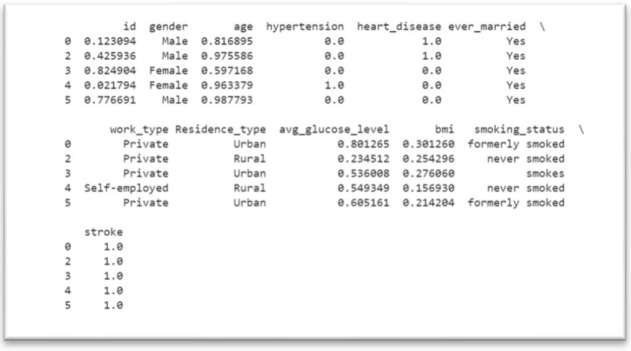


```
id                   0
gender               0
age                  0
hypertension         0
heart_disease        0
ever_married         0
work_type            0
Residence_type       0
avg_glucose_level    0
bmi                  0
smoking_status       0
stroke               0
dtype: int64
```

Berikut merupakan data setelah handling missing value. Dapat dilihat pada gambar 4.2 dimana kolom “bmi” tidak terdapat lagi missing value.

b) Normalisasi Data

Berikut adalah hasil dari normalisasi data menggunakan Min-Max Scaling.



```
         id  gender       age  hypertension  heart_disease ever_married  \
0  0.123094    Male  0.816895           0.0            1.0          Yes
2  0.425936    Male  0.975586           0.0            1.0          Yes
3  0.824904  Female  0.597168           0.0            0.0          Yes
4  0.021794  Female  0.963379           1.0            0.0          Yes
5  0.776691    Male  0.987793           0.0            0.0          Yes

      work_type Residence_type  avg_glucose_level       bmi   smoking_status  \
0       Private          Urban           0.801265  0.301260  formerly smoked
2       Private          Rural           0.234512  0.254296     never smoked
3       Private          Urban           0.536008  0.276060           smokes
4  Self-employed          Rural           0.549349  0.156930     never smoked
5       Private          Urban           0.605161  0.214204  formerly smoked

   stroke
0     1.0
2     1.0
3     1.0
4     1.0
5     1.0
```

Gambar 4.3. Hasil dari normalisasi data

Dalam gambar 4.3, Min-Max Scaling mengubah setiap nilai dalam kolom numerik sehingga mereka berada dalam rentang 0 hingga 1.

Klasifikasi Random Forest

Dalam klasifikasi menggunakan algoritma Random Forest, data awal akan dibagi menjadi dua bagian, yaitu data latih dan data uji, dengan rasio 80% untuk data latih dan 20% untuk data uji. Setelah pembagian data, kemudian melakukan validasi untuk memperoleh estimasi kinerja yang lebih konsisten. Untuk mengatur hiperparameter, digunakan metode cross-validation, yang memungkinkan penyesuaian parameter dilakukan dengan efisiensi waktu. Hasil dari proses penyesuaian tersebut menunjukkan skor pelatihan sekitar 96%.

```
Cross-Validation Scores: [0.95723014 0.95723014 0.95621181 0.95723014 0.95718654]
Average Accuracy: 0.96
```

Gambar 4.4. Hasil dari cross validation

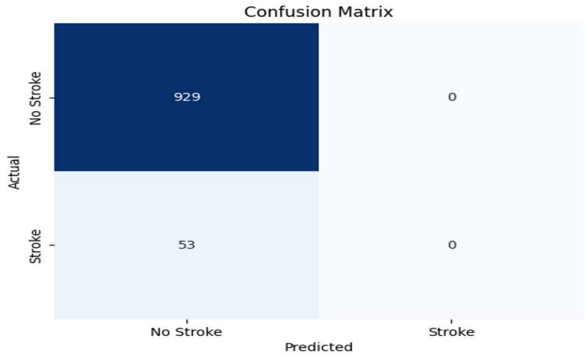
Evaluasi Hasil

Evaluasi model merupakan suatu proses untuk mengukur seberapa baik kinerja model yang telah dibangun. Hal ini penting untuk memahami sejauh mana model dapat menggeneralisasi pada data yang belum pernah dilihat sebelumnya. Dapat dilihat pada Tabel 2 menampilkan hasil precision, recall, f1-score, dan accuracy. Pada penelitian ini didapatkan akurasi sebesar 0.95 atau 95%

Tabel 2. Classification Report

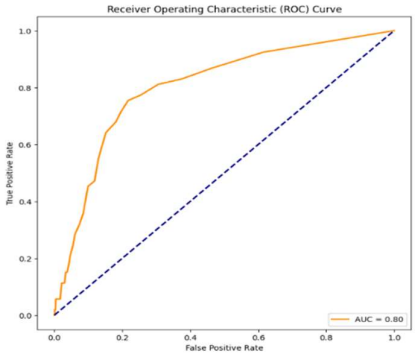
	Precision	Recall	f1-score	support
0	0.95	1.00	0.97	929
1	0.00	0.00	0.00	53
accuracy			**0.95**	**982**
macro avg	**0.47**	**0.50**	**0.49**	**982**
weighted avg	**0.89**	**0.**	**0.92**	**982**

Berikut ini adalah Confusion Matrix dari model Random Forest



Gambar 4.5 Confusion Matrix

Untuk melihat kurva ROC (Receiver Operating Characteristic), digunakan matplotlib dan sklearn guna untuk mengevaluasi kinerja model. Berikut adalah tampilan ROC pada penelitian ini.



Gambar 4.6 Kurva ROC

Dapat dilihat dalam gambar 3.6 menunjukkan bahwa pada model ini, AUC (Area under the ROC Curve) menghasilkan nilai 0.80, yang mengindikasikan bahwa model tersebut termasuk dalam klasifikasi yang baik

KESIMPULAN

Random Forest sangat bermanfaat dalam mengklasifikasikan dan mendeteksi gejala penyakit stroke, di dalam penelitian ini menggunakan dua bagian dari data awal yaitu data latih dan data uji dari kedua data tersebut menghasilkan rasio yaitu 80% untuk data latih dan 20% untuk data uji. Setelah itu peneliti melakukan validasi untuk memperoleh estimasi kinerja untuk mengatur hiperparameter, lalu menggunaka metode cross-validation untuk memungkinkan efisiensi waktu, maka hasil penyesuain menghasilkan 96% skor pelatihan. Setelah itu untuk mengukur seberapa baik kinerja model di bangun evaluasi model yaitu dengan menampilkan tabel hasil precision, recall, F1-score, dan accuracy Yang mendapatkan hasil akurasi sebesar 0.95 atau 95% , dari hasil kurva ROC (Receiver Oerating Characteristic) menggunakan matplotlib dan sklearn guna mengevaluasi kinerja model. Hasil akhir AUC (Area under the ROG Curve) menghasilkan nilai sebesar 0.80 dengan mengevaluasi kinerja model yang dapat menunjukkan bahwa model bisa di katakan ke dalam klasifikasi baik.

SARAN

Penelitian ini bisa di kembangkan menjadi lebih baik lagi untuk memberi pengetahuan yang lebih kompleks , beberapa saran-saran yang bisa di terapkan untuk penelitian selanjutnya:

a. menyajikan proses uji coba dengan lebih akurat agar dapat di implementasikan dengan cukup baik.

b. menggunakan model klasifikasi lain agar bervariasi dalam menentukan tingkat akurasi yang lebih sempurna.

UCAPAN TERIMA KASIH

Bagian ini disediakan bagi penulis untuk menyampaikan ucapan terima kasih, baik kepada pihak penyandang dana penelitian, pendukung fasilitas, atau bantuan ulasan naskah. Bagian ini juga dapat digunakan untuk memberikan pernyataan atau penjelasan, apabila artikel ini merupakan bagian dari skripsi/tesis/disertasi/makalah konferensi/hasil penelitian.

DAFTAR REFERENSI

[1] Jawapos.com,"Inilah Penyakit yang Paling Banyak Menyerang Masyarakat Indonesia" *Jawapos.com,* 2017, [Online].Available: https://www.jawapos.com/kesehatan/21/11/2017/inilah-penyakit-yang-paling-banyak-menyerang-masyarakat-indonesia/

[2] M. A. As Sarofi, I. Irhamah, and A. Mukarromah, "Identifikasi Genre Musik dengan Menggunakan Metode Random Forest," Jurnal Sains dan Seni ITS, vol. 9, no. 1, pp. 79–86, 2020, doi: 10.12962/j23373520.v9i1.51311.

[3] Iskandar, N. A., Ernawati, I., & Widiastiwi, Y. (2022, August). Klasifikasi Diagnosis Penyakit Stroke Dengan Menggunakan Metode Random Forest. In Prosiding Seminar Nasional Mahasiswa Bidang Ilmu Komputer dan Aplikasinya (Vol. 3, No. 2, pp. 706-714).

[4] Aji, P. W. S., Suprianto, S., & Dijaya, R. (2023). Prediksi Penyakit Stroke Menggunakan Metode Random Forest. *Kesatria: Jurnal Penerapan Sistem Informasi (Komputer dan Manajemen)*, *4*(4), 916-924.

[5] N. Permatasari, "Perbandingan Stroke Non Hemoragik dengan Gangguan Motorik Pasien Memiliki Faktor Resiko Diabetes Melitus dan Hipertensi," Jurnal Ilmiah Kesehatan Sandi Husada, vol. 11, no. 1, 2020, doi: 10.35816/jiskh.v11i1.273.

[9] A. Byna and M. Basit, "Penerapan Metode Adaboost Untuk Mengoptimasi Prediksi Penyakit Stroke Dengan Algoritma Naïve Bayes," Jurnal Sisfokom (Sistem Informasi dan Komputer), vol. 9, no. 3, pp. 407–411, 2020, doi: 10.32736/sisfokom.v9i3.1023.

[10] D. E. Cahyani, "Penerapan Machine Learning Untuk Prediksi Penyakit Stroke," Jurnal Kajian Matematika dan Aplikasinya, vol. 3, no. January, pp. 8–14, 2022, doi: 10.17977/um055v3i1p15-22

[11] D. Prajarini, S. Tinggi, S. Rupa, D. Desain, and V. Indonesia, "Perbandingan Algoritma Klasifikasi Data Mining Untuk Prediksi Penyakit Kulit," Informatics Journal, vol. 1, no. 3, p. 137, 2016.