Prediksi Penyakit Stroke Menggunakan Machine Learning Dengan Algoritma Random Forest

Mellya Putri[1], Erlin[2]

[1,2] *Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Insititut Bisnis dan Teknologi Pelita Indonesia*

[1*]mellya.putri@student.pelitaindonesia.ac.id
[2]erlin@lecturer.pelitaindonesia.ac.id

Abstrak— Sistem rekomendasi Penyakit stroke merupakan salah satu penyakit yang sering menimbulkan dampak yang serius bagi penderitanya. Oleh karena itu, prediksi penyakit stroke menjadi penting untuk dapat melakukan tindakan pencegahan yang tepat. Penelitian ini bertujuan untuk melakukan prediksi penyakit stroke menggunakan teknik machine learning dengan algoritma Random Forest. Data yang digunakan dalam penelitian ini berasal dari kumpulan data pasien yang terdiri dari berbagai atribut seperti usia, jenis kelamin, tekanan darah, dan lainnya. Proses prediksi dilakukan dengan memanfaatkan algoritma Random Forest untuk mengidentifikasi faktor-faktor risiko yang berkontribusi terhadap penyakit stroke. Hasil evaluasi menggunakan confusion matrix menunjukkan tingkat akurasi model yang baik dalam memprediksi kemungkinan seseorang terkena penyakit stroke. Berdasarkan pengujian yang telah dilakukan menghasilkan 933 stroke dengan benar (TP), 11 stroke yang sebenarnya bukan stroke (FP), 0 stroke yang sebenarnya stroke (FN), dan 988 stroke yang sebenarnya bukan stroke (TN).
Kata kunci: Penyakit Stroke, Prediksi, Machine Learning, Random Forest

Abstract— Stroke disease is one of the diseases that often causes serious impacts on sufferers. Therefore, prediction of stroke disease is important to be able to take appropriate preventive measures. This research aims to predict stroke disease using machine learning techniques with the Random Forest algorithm. The data used in this study comes from a collection of patient data consisting of various attributes such as age, gender, blood pressure, and others. The prediction process is carried out by utilizing the Random Forest algorithm to identify risk factors that contribute to stroke disease. Evaluation results using confusion matrix show a good level of model accuracy in predicting the likelihood of a person having a stroke. Based on the testing that has been done, it produces 933 strokes correctly (TP), 11 strokes that are not actually strokes (FP), 0 strokes that are actually strokes (FN), and 988 strokes that are not actually strokes (TN).
Keywords: Stroke Disease, Prediction, Machine Learning, Random Forest

I. PENDAHULUAN

Pada saat ini, penyakit stroke merupakan masalah kesehatan serius yang dapat menyebabkan kerusakan otak yang parah dan bahkan berujung pada kematian [1]. Di seluruh dunia, jumlah kasus stroke terus meningkat, dan memahami faktor-faktor yang mempengaruhi resiko kematian pada pasien stroke adalah langkah penting dalam penanganan dan perawatan yang lebih baik. Stroke adalah penyakit dengan ditandai oleh terganggunya fungsi otak disebabkan kurangnya pasokan oksigen dan aliran darah ke otak sehingga mempengaruhi beberapa fungsi otak yang membuat penyintas mengalami kesulitan dalam melakukan aktifitas.

Menurut data World Health Organization (WHO) diperkirakan 17,5 juta orang meninggal dunia akibat penyakit kardiovaskular dengan 6,7 juta orang meninggal akibat stroke, yaitu urutan kedua tertinggi mengakibatkan kematian setelah penyakit jantung koroner. Penyakit stroke merupakan salah satu masalah kesehatan yang mendesak, dengan dampak yang signifikan pada tingkat global. Pada tingkat individu, faktor-faktor yang mempengaruhi resiko kematian akibat stroke sangat beragam, seperti riwayat kesehatan, gaya hidup, dan karakteristik biologis.

Stroke merupakan salah satu penyakit yang paling banyak diderita oleh masyarakat Indonesia dan menjadi urutan pertama penyebab kematian tertinggi disusul oleh diabetes dan hipertensi. Diperkirakan ada 4,5 juta kematian per tahun akibat stroke di dunia dan lebih dari 9 juta penderita stroke. Risiko kekambuhan selama 5 tahun adalah 15-40%. Diperkirakan pada tahun 2023 akan ada absolut peningkatan jumlah pasien yang mengalami pertama kali stroke meningkat sekitar 30% dibandingkan dengan 1983. Stoke adalah yang terdepan penyebab kecacatan pada orang dewasa [2].

Seiring dengan perkembangan teknologi informasi dan komputasi, penggunaan kecerdasan buatan, khususnya Machine Learning, telah menjadi fokus dalam berbagai bidang, termasuk bidang medis. Salah satu aplikasi penting dari Machine Learning adalah dalam menganalisis data kesehatan dan membantu dalam pengambilan keputusan klinis yang lebih cerdas. Machine learning merupakan bidang lain dari ilmu komputer yang merancang sebuah algoritma agar memungkinkan sebuah komputer untuk belajar melalui data sehingga sering dikatakan sebagai learn from data. Jadi machine learning adalah pemrograman komputer yang menggunakan data masa lalu yang digunakan untuk pembelajaran model sehingga mendapatkan performa yang optimal dalam menggali informasi dari suatu kumpulan data. Inti machine learning adalah untuk membuat model yang merefleksikan pola-pola data [3].

Pada aplikasi machine learning ini, algoritma atau urutan proses statistik dilatih untuk menemukan pola tertentu dalam jumlah data yang besar. Salah satu algoritma Machine Learning adalah Random Forest, telah terbukti sangat efektif dalam mengatasi masalah klasifikasi dan prediksi. Dalam konteks penyakit stroke, penggunaan Random Forest dapat membantu dalam mengidentifikasi faktor-faktor yang dapat meningkatkan resiko kematian pada pasien stroke. Ini termasuk variabel seperti usia pasien, jenis kelamin, riwayat medis, gejala awal, dan hasil pemeriksaan klinis lainnya. Dalam penelitian ini, kita akan menjelajahi penerapan Machine Learning, khususnya algoritma Random Forest, untuk mengklasifikasikan resiko kematian pada pasien stroke. Pendekatan ini memungkinkan kita untuk memanfaatkan data

yang tersedia, baik data pasien maupun data medis, untuk mengembangkan model yang dapat memprediksi dengan lebih akurat kemungkinan resiko kematian pasien. Menurut [4]), Random Forest adalah teknik bagging yang memiliki karakteristik signifikan yang berjalan efisien pada dataset besar. Random Forest dapat menangani ribuan variabel masukan tanpa penghapusan variabel dan memperkirakan fitur penting untuk klasifikasi. Beberapa penelitian terdahulu menggunakan algoritma ini dengan algoritma lainnya untuk membandingkan seberapa akurat dari penggunaan algoritma ini.

Berdasarkan peneliti terdahulu sebelumnya oleh [5], dengan melakukan perbandingan algoritma regresi linier dan regresi random forest dalam memprediksi kasus positif covid-19 dapat disimpulkan bahwa model random forest regression lebih baik dalam memprediksi kasus positif Covid-19 dibandingkan model regresi linier. Hal ini ditunjukkan dari hasil akurasi, RMSE, dan MAPE yang lebih tinggi pada model random forest regression. Model random forest regression memiliki akurasi sebesar 97,7%, sedangkan model regresi linier memiliki akurasi sebesar 94%. Hal ini menunjukkan bahwa model random forest regression mampu memprediksi kasus positif Covid-19 dengan lebih akurat dibandingkan model regresi linier.

Penelitian yang dilakukan oleh [6], yaitu perbandingan model decision tree, naive bayes dan random forest untuk prediksi klasifikasi penyakit jantung menggunakan Data yang digunakan dalam penelitian ini adalah data pasien penyakit jantung dari Rumah Sakit Cipto Mangunkusumo. Data tersebut terdiri dari 200 data pasien, dengan 150 data untuk pelatihan dan 50 data untuk pengujian. Hasil penelitian menunjukkan bahwa model random forest memiliki akurasi tertinggi sebesar 97,6%, diikuti oleh model naive bayes sebesar 96,1%, dan model decision tree sebesar 95,2%. Berdasarkan hasil penelitian tersebut, dapat disimpulkan bahwa model random forest merupakan model yang paling baik untuk prediksi klasifikasi penyakit jantung. Hal ini karena model random forest dapat menggabungkan prediksi dari beberapa decision tree, sehingga dapat menghasilkan prediksi yang lebih akurat.

Penelitian yang dilakukan oleh [7], Klasifikasi Penyakit Liver dengan Algoritma Machine Learning, pada penelitian membahas tentang klasifikasi penyakit liver menggunakan algoritma Random Forest. Data yang digunakan adalah dataset ILPD dari UCI Machine Learning Repository. Hasil penelitian menunjukkan bahwa algoritma Random Forest dapat menghasilkan akurasi yang tinggi dalam memprediksi penyakit liver, yaitu sebesar 78,63%. Akurasi sebesar 78,63% menunjukkan bahwa algoritma Random Forest dapat memprediksi dengan baik apakah seseorang menderita penyakit liver atau tidak. Nilai akurasi ini lebih unggul dari algoritma lainnya yang diuji dalam penelitian ini, yaitu algoritma K-Nearest Neighbor (KNN), Naive Bayes, dan Support Vector Machine (SVM). Berdasarkan hasil penelitian ini, dapat disimpulkan bahwa algoritma Random Forest dapat menjadi pilihan yang tepat untuk prediksi penyakit liver. Namun, perlu diperhatikan bahwa akurasi prediksi dapat bervariasi tergantung pada kualitas data yang digunakan.

Penelitian ini menggunakan machine learning dengan implementasi algoritma Random Forest untuk memprediksi penyakit jantung dengan tujuan memberikan wawasan baru dalam pengembangan metode prediksi yang lebih akurat untuk mendukung diagnosis dan pencegahan stroke menggunakan pendekatan *machine learning*.

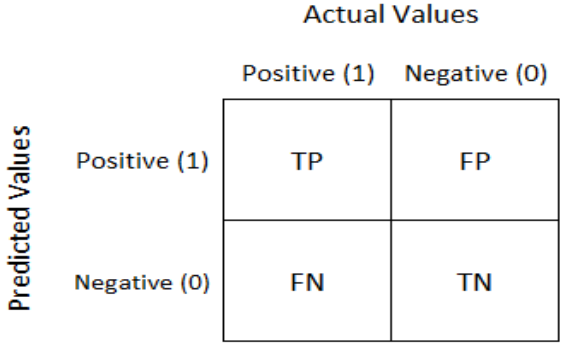
A. Tinjauan Pustaka

Penyakit stroke adalah penyakit yang terjadi ketika aliran darah ke otak terhambat atau terputus, sehingga otak tidak mendapatkan oksigen dan nutrisi yang dibutuhkan [8]. Hal ini dapat menyebabkan kerusakan otak dan kematian sel-sel otak. Stroke dapat diklasifikasikan menjadi dua jenis, yaitu stroke iskemik dan stroke hemoragik. Stroke iskemik terjadi ketika aliran darah ke otak terhambat oleh gumpalan darah, sedangkan stroke hemoragik terjadi ketika pembuluh darah di otak pecah. Penderita hipertensi akan mengalami aneurisma yang disertai disfungsi endoteliel pada jaringan pembuluh darahnya. Apabila gangguan yang terjadi pada pembuluh darah ini berlangsung terus dalam waktu yang lama akan dapat menyebabkan terjadinya stroke.

Organisasi Kesehatan Dunia melaporkan bahwa kelemahan wajah yang tiba-tiba dan bahkan mati rasa di lengan, kaki, atau di satu sisi tubuh adalah gejala yang paling sering dialami. Selain itu, fungsi sensorik tubuh hilang, dan sakit kepala yang sangat parah dirasakan, yang dapat menyebabkan pingsan atau tidak sadarkan diri [9].

Confusion Matrix merupakan metode pengukuran untuk mencari masalah klasifikasi yang dilakukan oleh machine learning dengan keluaran berupa dua kelas atau lebih, confusion Matrix adalah tabel dengan 4 kombinasi berbeda dari nilai prediksi dan nilai aktual [10]. Confusion matrix adalah tabel yang digunakan untuk mengukur kinerja model klasifikasi. Confusion matrix terdiri dari empat kuadran terlihat pada **Gambar 2.1**, yaitu:

1. True Positive (TP): Jumlah data yang diprediksi positif dan benar-benar positif.
2. False Positive (FP): Jumlah data yang diprediksi positif tetapi sebenarnya negatif.
3. True Negative (TN): Jumlah data yang diprediksi negatif dan benar-benar negatif.
4. False Negative (FN): Jumlah data yang diprediksi negatif tetapi sebenarnya positif.



Gambar 2. 1 Cunfuxion Matrix Recall

Ada beberapa metrik yang dapat digunakan untuk mengukur kinerja model klasifikasi berdasarkan confusion

matrix seperti diperlihatkan pada persamaan 1 sampai 4, antara lain:

1. Akurasi: Persentase data yang diprediksi dengan benar, baik positif maupun negatif.
2. Presisi: Persentase data positif yang diprediksi dengan benar.
3. Recall: Persentase data positif yang benar-benar diprediksi positif.
4. F1 score: Rata-rata presisi dan recall.

Berikut adalah rumus untuk menghitung metrik-metrik tersebut:

$$\text{Akurasi} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (1)$$

$$\text{Presisi} = \frac{TP}{TP+FP} \quad (2)$$

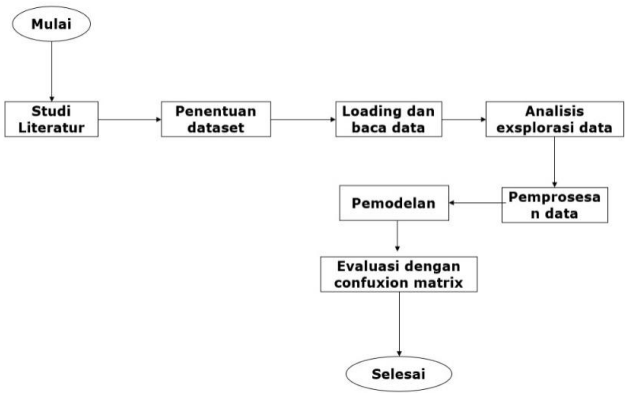
$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

$$\text{F1 score} = \frac{2*presisi*recall}{presisi+recall} \quad (4)$$

Confusion matrix merupakan alat yang berguna untuk memahami kinerja model klasifikasi. Dengan memahami confusion matrix, kita dapat mengetahui kelebihan dan kekurangan model klasifikasi, dan kemudian melakukan penyesuaian untuk meningkatkan kinerjanya.

II. METODOLOGI PENELITIAN

Sistem penyelesaian penelitian ini, dilakukan tahap – tahap penelitian yang merupakan proses dari penyelesaian penelitian ini. Adapun kerangka penelitian yang dilakukan dapat dilihat pada **Gambar 3.1** dibawah ini.



Gambar 3. 1 Kerangka Penelitian

a. Studi literatur
Studi literatur adalah kegiatan membaca dan menganalisis sumber-sumber tertulis yang relevan dengan topik penelitian. Tujuan dari studi literatur adalah untuk mendapatkan landasan teori, kerangka berpikir, dan hipotesis penelitian.

b. Penentuan dataset
Penentuan dataset dilakukan untuk dapat memperoleh informasi yang dibutuhkan selama melakukan penelitian dalam rangka untuk mencapai tujuan yang diinginkan. Dataset yang digunakan dalam penelitian ini berasal dari dataset UCI Machine Learning Repository.

c. Loading dan baca data
Loading dan membaca data adalah proses memasukkan data dari sumber eksternal ke dalam program komputer. Sumber eksternal ini dapat berupa file, database, atau bahkan sumber data streaming. Data yang dimuat biasanya akan disimpan dalam memori komputer sehingga dapat diakses dan diproses oleh program komputer.

d. Analisis eksplorasi data
Analisis eksplorasi data bertujuan menganalisis dataset yang digunakan untuk meringkas karakteristik utama dataset tersebut menggunakan bantuan statistika dan mempresentasikannya melalui teknik visual. Pada tahap ini, data diperiksa sebelum dibangunnya model, sehingga didapatkan wawasan maksimal dari dataset yang dimiliki.

e. Pemprosesan data
Pemprosesan data dapat membantu peneliti untuk menyampaikan hasil penelitian, menampilkan hasil dari pengolahan data. Data yang sudah melalui proses data cleaning dan preprocessing akan disajikan sehingga menghasilkan informasi yang berguna dengan menggunakan model algoritma Random Forest. Pada tahap ini, pengecekan dilakukan terhadap nilai data yang hilang karena dataset bisa saja memuat data yang tidak lengkap. Nilai data yang hilang digantikan dengan nilai median dari setiap variabel, sehingga setiap data pada variabel dataset memiliki nilai yang lengkap.

f. Pemodelan
Pemodelan ini menggunakan algoritma random forest. Random forest merupakan algoritma pembelajaran mesin yang berbasis pohon keputusan. Algoritma ini bekerja dengan membangun banyak pohon keputusan dari data yang sama, dan kemudian menggabungkan prediksi dari setiap pohon keputusan untuk menghasilkan prediksi akhir. Algoritma ini dapat menghasilkan akurasi yang tinggi pada berbagai jenis data.

g. Evaluasi Dengan Confuxion Matrix
Evaluasi yang digunakan untuk mengukur kinerja algoritma/model adalah akurasi, presisi, recall, dan F1-score dalam bentuk confusion matrix yang sudah banyak digunakan oleh peneliti lainnya. Confusion matrix merupakan tabel yang bekerja dengan cara membandingkan jumlah prediksi yang benar dan yang salah yang terdapat pada masing-masing kelas, sehingga memberikan wawasan mengenai kesalahan model yang dibangun.

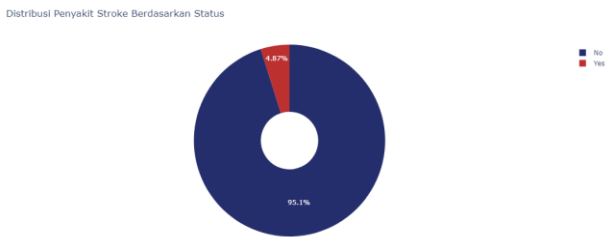
III. HASIL DAN PEMBAHASAN

A. Explorasi Data Analisis

1) Distribusi penyakit stroke

Berdasarkan Data healtcare Stroke yang menderita penyakit stroke yaitu 4.87% dan yang tidak memiliki penyakit stroke sebanyak 95.1% terdapat pada **Gambar 4.1**. Hal ini dengan jelas menggambarkan ketidakseimbangan kelas kumpulan data, dengan jumlah kasus "Tanpa Stroke" yang jauh lebih besar dibandingkan

dengan kasus "Stroke". Namun, laporan ini hanya menyampaikan distribusi kasus stroke dan mungkin tidak memberikan wawasan yang lebih mendalam mengenai faktor-faktor lain.

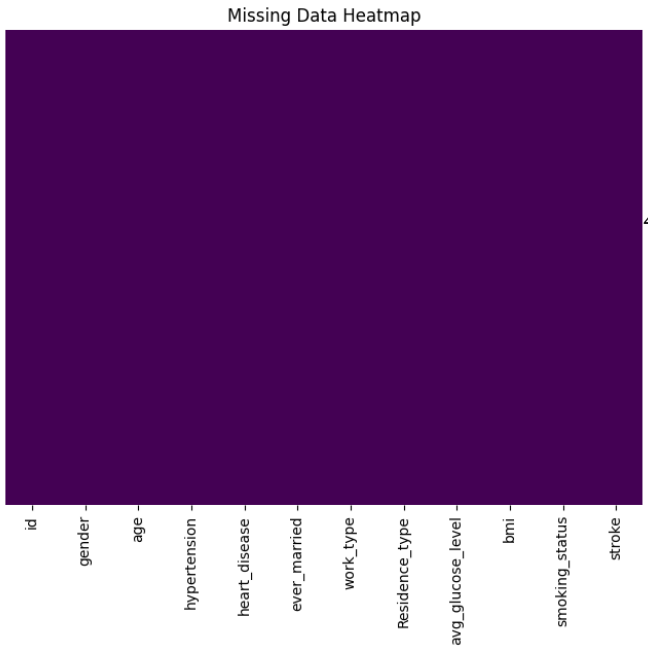


Gambar 4. 1 Distribusi Penyakit Stroke Berdasarkan Status

B. Pemprosesan Data

1) Mengatasi nilai data yang kosong pada data set

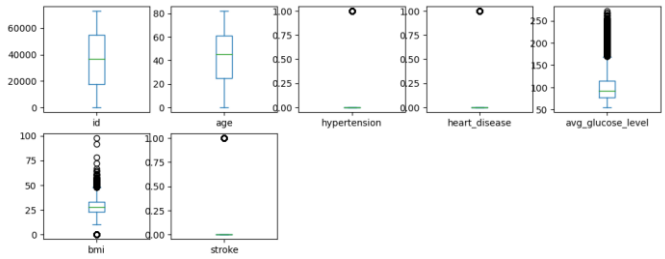
Terlihat bahwa pada **Gambar 4.15** Mengatasi Data Kosong sudah tidak ada lagi data yang kosong pada gambar tersebut.



Gambar 4. 2 Mengatasi Data Kosong

2) Mengatasi outliers

Pada gambar **Gambar 4.16** Mengatasi outliers, nilai maksimum yang terdapat dalam kolom 'bmi' adalah 97.6. Jumlah data dalam kolom 'bmi' yang memiliki nilai lebih besar atau sama dengan 65, dan data-data ini dianggap sebagai pencilan (outliers) berdasarkan kondisi yang ditetapkan.



Gambar 4. 3 Mengatasi Outliers

3) Mengatasi duplikat data

Berdasarkan Data healtcare Stroke, duplikasi dalam dataset dapat dapat dilihat pada **Gambar 4.18** mempengaruhi analisis. Kami mengidentifikasi dan menghapus baris duplikat untuk memastikan integritas data. Berikut merupakan data yang duplikat:

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	0.0	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1
...	...	...	...	...	...	...	...	...	...	...	...	...
5105	18234	Female	80.0	1	0	Yes	Private	Urban	83.75	0.0	never smoked	0
5106	44873	Female	81.0	0	0	Yes	Self-employed	Urban	125.20	40.0	never smoked	0
5107	19723	Female	35.0	0	0	Yes	Self-employed	Rural	82.99	30.6	never smoked	0
5108	37544	Male	51.0	0	0	Yes	Private	Rural	166.29	25.6	formerly smoked	0
5109	44679	Female	44.0	0	0	Yes	Govt_job	Urban	85.28	26.2	never smoked	0

5077 rows × 12 columns

Gambar 4. 4 Data duplikat

4) Pengkodean kategori

Untuk menyiapkan data kategori analisis ini menerapkan teknik pengkodean *one hot encoding*. Dengan mengubah data-data yang kategori menjadi numerik. menggunakan pengkodean label untuk 'jenis_kelamin', 'pernah_menikah', dan 'jenis_tempat_tinggal'. Selain itu, kami menggunakan pengkodean satu titik untuk 'jenis_pekerjaan' dan 'status_merokok' untuk mengubahnya menjadi variabel biner yang sesuai untuk analisis dan pemodelan. Teknik-teknik pengkodean ini membantu mengubah data kategori menjadi format yang dapat digunakan oleh algoritma machine learning secara efektif.

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	1	67.0	0	1	1	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	0	61.0	0	0	1	Self-employed	Rural	202.21	0.0	never smoked	1
2	31112	1	80.0	0	1	1	Private	Rural	105.92	32.5	never smoked	1
3	60182	0	49.0	0	0	1	Private	Urban	171.23	34.4	smokes	1
4	1665	0	79.0	1	0	1	Self-employed	Rural	174.12	24.0	never smoked	1
...	...	...	...	...	...	...	...	...	...	...	...	...
5105	18234	0	80.0	1	0	1	Private	Urban	83.75	0.0	never smoked	0
5106	44873	0	81.0	0	0	1	Self-employed	Urban	125.20	40.0	never smoked	0
5107	19723	0	35.0	0	0	1	Self-employed	Rural	82.99	30.6	never smoked	0
5108	37544	1	51.0	0	0	1	Private	Rural	166.29	25.6	formerly smoked	0
5109	44679	0	44.0	0	0	1	Govt_job	Urban	85.28	26.2	never smoked	0

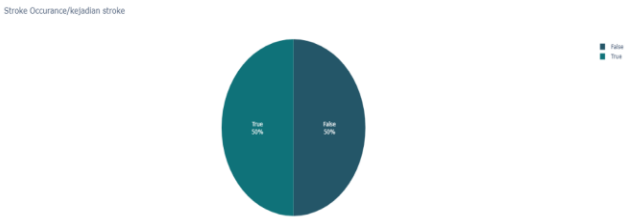
5076 rows × 12 columns

Gambar 4. 5 Pengkodean Data

5) Menyeimbangkan data

Dataset adalah tidak seimbang sesuai dengan gambar sebelumnya, apabila dataset yang tidak seimbang ini tetap diproses akan menghasilkan model yang tidak atau

kurang akurasinya, mengahsilkan akurasi model yang tidak baik, oleh sebab itu diperlukan untuk menyeimbangkan atau membalance kan data. Data dibalance kan atau di seimbangkan dengan menggunakan teknik upsampling, proses meningkatkan resolusi atau ukuran data. Teknik upsampling digunakan untuk membuat data memiliki dimensi yang lebih tinggi dari pada data asli, dimana data yang jumlah awal nya sedikit akan di buat data imitasi (data tambahan). Terlihat pada **Gambar 4.20**



Gambar 4. 6 Kejadian Stroke

C. Pemodelan

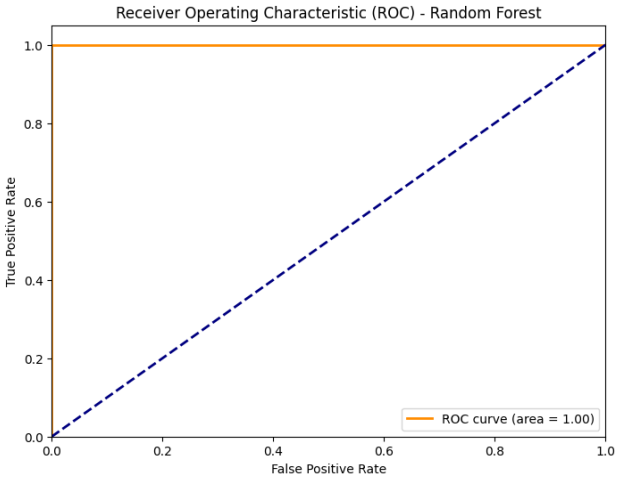
1) Pembagian data

Pembagian data 80% untuk training dan 20% untuk testing adalah pembagian data yang umum digunakan dalam machine learning, termasuk dalam algoritma random forest. Pembagian data ini dimaksudkan untuk memisahkan data menjadi dua set, yaitu set data training dan set data testing. Set data training digunakan untuk melatih model random forest. Model random forest akan mempelajari hubungan antara fitur-fitur data dan label kelas dari set data training. Set data testing digunakan untuk mengevaluasi kinerja model random forest. Model random forest akan digunakan untuk memprediksi label kelas dari set data testing. Hasil prediksi dari set data testing akan dibandingkan dengan label kelas sebenarnya untuk mengetahui akurasi dan kinerja model random forest.

Pembagian data 80% untuk training dan 20% untuk testing ini memiliki beberapa kelebihan, yaitu model random forest akan memiliki akurasi yang lebih baik,
model random forest akan lebih robust terhadap overfitting. Overfitting adalah suatu kondisi di mana model machine learning terlalu cocok dengan data training, sehingga kinerja model menjadi buruk ketika diterapkan pada data baru. Pembagian data 80% untuk training dan 20% untuk testing dapat membantu mengurangi overfitting dengan cara memberikan set data testing yang cukup untuk mengevaluasi kinerja model.

2) Model dengan Random Forest

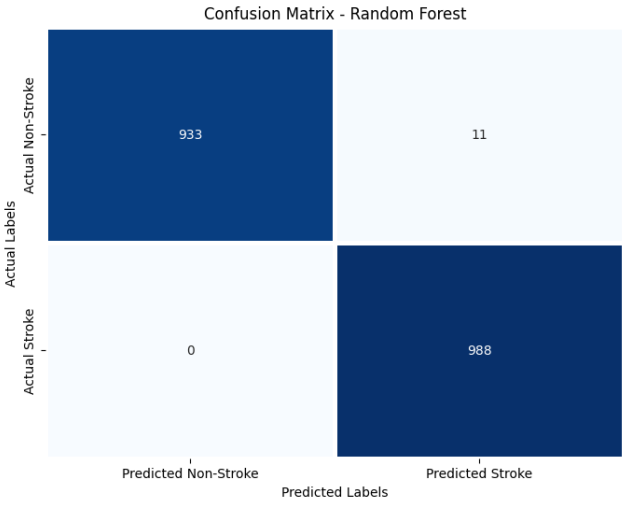
Kurva ROC *(Receiver Operating Characteristic)* untuk model Random Forest. Kurva memiliki nilai AUC *(Area Under the Curve)* 1.00, yang menunjukkan performa model yang sempurna dalam membedakan antara kelas positif dan negatif. Kurva tersebut menggambarkan kemampuan model yang sangat baik dalam membedakan kelas-kelas yang diprediksi. Dapat dilihat Pada **Gambar 4.21** Kurva ROC (Receiver Operating Characteristic)



Gambar 4. 7 Kurva ROC (Receiver Operating Characteristic)

D. Evaluasi dengan confuxion matrix

Model random forest pada **Gambar 4.22**, memprediksi 933 stroke dengan benar (TP), 11 stroke yang sebenarnya bukan stroke (FP), 0 stroke yang sebenarnya stroke (FN), dan 988 stroke yang sebenarnya bukan stroke (TN). Secara keseluruhan, model random forest memiliki kinerja yang sangat baik dalam memprediksi stroke.



Gambar 4. 8 Confuxion Matrix

Model random forest memiliki presisi, recall, dan f1-score yang tinggi untuk kedua kelas, yaitu stroke (1) dan bukan stroke (0). Presisi untuk kelas stroke sebesar 1,00 artinya model tidak pernah salah memprediksi stroke. Recall untuk kelas stroke sebesar 0,99 artinya model memprediksi stroke dengan benar sebesar 99% dari data aktual yang benar-benar stroke. F1-score untuk kelas stroke sebesar 0,99 artinya model memprediksi stroke dengan benar sebesar 99% dari data aktual yang benar-

benar stroke. Presisi untuk kelas bukan stroke sebesar 0,99 artinya model tidak pernah salah memprediksi bukan stroke.

Recall untuk kelas bukan stroke sebesar 1,00 artinya model memprediksi bukan stroke dengan benar sebesar 100% dari data aktual yang benar-benar bukan stroke. F1-score untuk kelas bukan stroke sebesar 0,99 artinya model memprediksi bukan stroke dengan benar sebesar 99% dari data aktual yang benar-benar bukan stroke. Secara keseluruhan, model random forest memiliki kinerja yang sangat baik dalam memprediksi stroke. Model ini dapat memprediksi stroke dengan benar dengan presisi, recall, dan f1-score yang tinggi.

IV. KESIMPULAN

Berdasarkan penelitian ini dijelaskan bahwa untuk memprediksi penyakit stroke menggunakan machine learning dengan algoritma random forest, langkah-langkah yang dilakukan meliputi penentuan dataset, loading data, analisis explorasi data, pemprosesan data, pemodelan, dan evaluasi dengan confixion matrix. Pertama-tama, data yang relevan terkait dengan pasien-pasien yang telah mengalami stroke atau memiliki faktor risiko yang berkaitan dengan stroke dikumpulkan. Data ini mencakup informasi seperti usia, jenis kelamin, riwayat medis, gaya hidup, dan faktor-faktor lain yang dapat berkontribusi terhadap risiko stroke. Selanjutnya, data tersebut diproses untuk membersihkan, menghilangkan nilai-nilai yang hilang, dan menyesuaikan format data agar sesuai untuk digunakan dalam model machine learning.

Setelah itu, data dibagi menjadi dua bagian, yaitu data latih (training data) dan data uji (testing data). Data latih digunakan untuk melatih model algoritma random forest. Model ini akan belajar dari pola-pola yang terdapat dalam data latih untuk memprediksi kemungkinan terjadinya stroke. Kemudian, data uji digunakan untuk mengevaluasi kinerja model. Lalu di evaluasi menggunakan confusion matrix meliputi akurasi, presisi, recall, dan F1-score.

Berdasarkan pengujian yang telah dilakukan pada bab sebelumnya, untuk mengetahui performa metode *Random forest* pada prediksi penyakit stroke. Diprediksi 933 stroke dengan benar (TP), 11 stroke yang sebenarnya bukan stroke (FP), 0 stroke yang sebenarnya stroke (FN), dan 988 stroke yang sebenarnya bukan stroke (TN). Secara keseluruhan, model random forest memiliki kinerja yang sangat baik dalam memprediksi stroke.

Pada penelitian ini parameter yang digunakan untuk memprediksi penyakit stroke menggunakan algoritma random forest adalah faktor risiko yang telah teridentifikasi dalam literatur medis, seperti umur, jenis kelamin, tekanan darah, kadar gula, riwayat merokok, riwayat pekerjaan, tempat tinggal, status pernikahan, dan faktor-faktor lain yang telah terbukti berkontribusi terhadap risiko stroke.

REFERENSI

[1] Fadli, M., & Saputra, R. A. (2023). *Klasifikasi Dan Evaluasi Performa Model Random Forest Untuk Prediksi Stroke Classification And Evaluation Of Performance Models Random Forest For Stroke Prediction*. *12*(02), 72–80.

[2] Sulaeman, K. R. (2022). Analisis Algoritma Support Vector Machine Dalam Klasifikasi Penyakit Stroke Support Vector Machine Algorithm Analysis In Stroke Disease Classification. *E-Proceeding of Engineering*, *9*(3), 922–928.

[3] Sidik, A. D., & Ansawarman, A. (2022). Prediksi Jumlah Kendaraan Bermotor Menggunakan Machine Learning. *Formosa Journal of Multidisciplinary Research*, *1*(3), 559–568. https://doi.org/10.55927/fjmr.v1i3.745

[4] Rizky, M., & Andarsyah, R. (2023). Klasifikasi MIT-BIH Arrhythmia Database Metode Random Forest dan CNN dengan Model ResNet-50: A Systematic Literature Review. *Jurnal Teknologi Dan Sistem Informasi Bisnis*, *5*(3), 190–196. https://doi.org/10.47233/jteksis.v5i3.825

[5] Fachid, S., & Triayudi, A. (2022). Perbandingan Algoritma Regresi Linier dan Regresi Random Forest Dalam Memprediksi Kasus Positif Covid-19. *Jurnal Media Informatika Budidarma*, *6*(1), 68. https://doi.org/10.30865/mib.v6i1.3492

[6] Depari, D. H., Widiastiwi, Y., & Santoni, M. M. (2022). Perbandingan Model Decision Tree, Naive Bayes dan Random Forest untuk Prediksi Klasifikasi Penyakit Jantung. *Informatik : Jurnal Ilmu Komputer*, *18*(3), 239. https://doi.org/10.52958/iftk.v18i3.4694

[7] Mining Dengan Algoritma Machine Larning Untuk Prediksi Penyakit Liver. *Technologia : Jurnal Ilmiah*, *14*(2), 134. https://doi.org/10.31602/tji.v14i2.10093

[8] Sinaga, S. H., Duha, A. A. M., & Banjarnahor, J. (2023). Analisis Prediksi Deteksi Stroke Dengan Pendekatan Eda Dan Perbandingan Algoritma Machine Learning. *Jurnal Ilmiah Betrik*, *14*(02 AGUSTUS), 355–367.

[9] Azwanti, N., & Elisa, E. (2019). Analisis Pola Penyakit Hipertensi Menggunakan Algoritma C4.5. *InfoTekJar (Jurnal Nasional Informatika Dan Teknologi Jaringan)*, *3*(2), 116–123. https://doi.org/10.30743/infotekjar.v3i2.944

[10] Hovi, H. S. W., Id Hadiana, A., & Rakhmat Umbara, F. (2022). Prediksi Penyakit Diabetes Menggunakan Algoritma Support Vector Machine (SVM). *Informatics and Digital Expert (INDEX)*, *4*(1), 40–45. https://doi.org/10.36423/index.v4i1.895