ISSN (Online) 2747-0563
Volume 4 Tahun 2024

Perbandingan Algoritma Decision Tree dan Random Forest dengan Hyperparameter Tuning dalam Mendeteksi Penyakit Stroke

Alvin Ryan Dana[1], Raja Valentino Kristananda[2], Muhammad Bagas Satrio Wibowo[3],Dwi Arman Prasetya[4]

[1,2,3] (Sains Data, Universitas Pembangunan Nasional "Veteran" Jawa Timur)
[1]21083010035@student.upnjatim.ac.id
[2]21083010068@student.upnjatim.ac.id
[3]21083010071@student.upnjatim.ac.id

[4] (Sains Data, Universitas Pembangunan Nasional "Veteran" Jawa Timur)
**Corresponding author email*: arman.prasetya.sada@upnjatim.ac.id

***Abstrak*— Penelitian ini bertujuan untuk membandingkan model prediksi dalam menentukan risiko seseorang mengalami stroke berdasarkan berbagai variabel pasien yang relevan, termasuk jenis kelamin, usia, riwayat hipertensi, penyakit jantung, status pernikahan, jenis pekerjaan, tingkat glukosa darah rata-rata, indeks massa tubuh (BMI), dan status merokok. Stroke merupakan penyebab kematian kedua terbanyak di dunia menurut Organisasi Kesehatan Dunia (WHO), oleh karena itu deteksi dini dan prediksi risiko stroke sangat penting untuk pencegahan dampak buruk yang mungkin terjadi. Hasil dari penelitian ini menunjukkan bahwa model Random Forest memberikan akurasi yang lebih tinggi (0.97) dan akurasi *hyperparameter* (0.98) dibandingkan dengan model Decision Tree, sehingga dipilih sebagai model prediksi yang lebih cocok dalam konteks dataset ini. Akurasi dipilih sebagai ukuran evaluasi karena penting untuk mengklasifikasikan dengan benar antara yang berisiko dan tidak berisiko mengalami stroke, yang memiliki implikasi yang sangat penting dalam aplikasi medis dan pencegahan stroke.**

Kata Kunci— **Decision Tree, Random Forest, Hypertuning, Stroke**

I. PENDAHULUAN

Stroke, sebuah kondisi yang mengancam jiwa, telah menjadi fokus utama dalam bidang kesehatan masyarakat dan kedokteran karena dampaknya yang signifikan terhadap kesehatan global. Stroke adalah penyebab kecacatan nomor satu di dunia dan penyebab kematian nomor tiga di dunia[12]. Dengan menyadari dampak serius ini, para peneliti, ilmuwan, dan praktisi medis terus berupaya keras untuk mengembangkan strategi baru dalam mendeteksi dini dan mencegah kejadian stroke. Secara menyeluruh 15 juta orang terserang stroke setiap tahunnya, satu pertiga diantaranya meninggal dunia dan sisanya mengalami kecacatan permanen[14].Word Health Organization menyebutkan bahwa data kematian akibat stroke sekitar 12,8% dari jumlah total seluruhnya[8] Pentingnya deteksi dini dan pengelolaan risiko stroke tidak bisa diabaikan. Sebagian besar stroke dapat dicegah melalui pengelolaan faktor risiko tertentu, seperti tekanan darah tinggi, kolesterol tinggi, diabetes, dan gaya hidup tidak sehat seperti merokok dan kurangnya aktivitas fisik. Oleh karena itu, ada kebutuhan yang mendesak untuk memperbaiki alat prediksi yang ada dan mengembangkan model prediktif yang lebih akurat untuk mengidentifikasi individu yang berisiko tinggi.

Dalam hal ini, data menjadi kunci. Penggunaan teknologi dan metode analisis data canggih telah memungkinkan para peneliti untuk mengeksplorasi hubungan yang lebih kompleks antara faktor risiko stroke dan kemungkinan kejadian stroke. Namun, meskipun ada upaya yang signifikan dalam penelitian ini, masih ada ruang untuk peningkatan dalam akurasi prediksi. Di samping itu, penting juga untuk mempertimbangkan kebutuhan akan model yang dapat digunakan dalam praktik klinis sehari-hari. Analisis prediktif adalah penggunaan analisis data, pemodelan statistik, dan teknologi pembelajaran mesin untuk memprediksi kemungkinan hasil [5] Model prediktif yang akurat namun mudah digunakan oleh praktisi medis dapat membantu dalam identifikasi dini pasien yang berisiko tinggi, memungkinkan intervensi yang tepat waktu dan pengelolaan risiko yang lebih efektif.

Penelitian ini bertujuan untuk mengembangkan model prediksi yang canggih dan valid untuk menilai risiko stroke pada individu berdasarkan data klinis mereka. Penelitian ini tidak hanya diharapkan dapat meningkatkan pemahaman kita tentang faktor risiko stroke, tetapi juga memberikan kontribusi signifikan dalam upaya pencegahan dan pengelolaan stroke secara global.

II. METODOLOGI PENELITIAN

Dalam penelitian ini, metode klasifikasi dengan algoritma Decision Tree dan Random Forest dengan penerapan metode hyperparameter tuning digunakan untuk mengklasifikasikan sebuah penyakit stroke. Metode *hyperparameter* tuning dilakukan oleh [10]. Penelitian tersebut bertujuan untuk meningkatkan peforma dari model *stance detection*, dan dalam penelitian tersebut dapat disimpulkan bahwa model memiliki peforma yang lebih baik dengan nilai micro f1-score yang lebih besar 6,6%. Metode *Hyperparameter* tuning juga dilakukan oleh [6] penelitian tersebut akan membandingkan konfigurasi *default* dengan *hyperparameter* tuning. Representasi dari

ISSN (Online) 2747-0563
Volume 4 Tahun 2024

penggunaan *hyperparameter* tuning dalam penelitian ini ditunjukkan dalam bentuk diagram alur sebagai berikut:

Gambar. 1 Block Diagram.

Pada gambar 1 yaitu gambar Block Diagram dimana diagram tersebut menggambarkan alur kerja dari awal yaitu tahap pengumpulan data hingga hasil untuk mengklasifikasikan sebuah penyakit stroke atau tidak dari sebuah pasien. Tahap awal berfokus pada pemrosesan data untuk mempersiapkan dataset yang akurat dan lengkap. Selanjutnya, dilakukan pemrosesan data agar data yang digunakan lebih akurat dalam langkah selanjutnya. Tahap implementasi algoritma hingga evaluasi dan hasil menggunakan Python untuk mengklasifikasikan pasien ke dalam kelas mengalami penyakit stroke atau tidak. Evaluasi hasil klasifikasi dilakukan untuk menilai sejauh mana klasifikasi dapat mencerminkan karakteristik pasien, menggunakan confusion matriks untuk melihat hasil prediksi dari data dan classification report dari sebuah library scikit learn digunakan untuk melihat bagaimana hasil kinerja model yang telah dibangun.

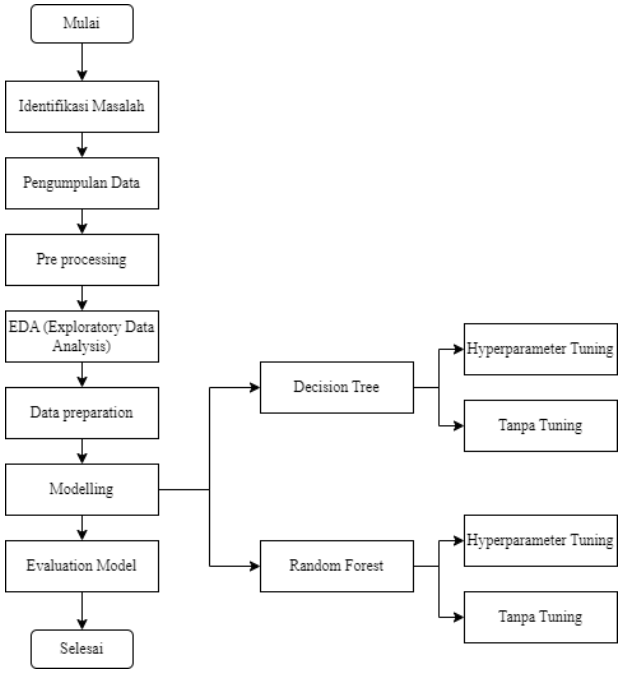
Pada pengumpulan data untuk penelitian ini diambil dari Kaggle, sebuah platform yang menyediakan berbagai dataset publik untuk digunakan dalam penelitian dan analisis. Dataset yang digunakan adalah "Stroke Prediction Dataset" yang berisi informasi yang mencakup informasi seperti Jenis Kelamin, Status Perkawinan, Usia, Hipertensi, Penyakit Jantung, Status Pernikahan, Jenis Pekerjaan, Tipe Tempat Tinggal, Rata-rata Kadar Glukosa Darah, Indeks Massa Tubuh (BMI), Status Merokok, dan keberadaan stroke.

TABEL I TABEL KOLOM DESKRIPSI

***Feature* / Kolom**	**Deskripsi**
"Id"	Identifikasi unik
"Gender"	Jenis kelamin (Laki-laki / Perempuan")
"Age"	Usia
"Hypertension"	Status Hipertensi (Ya / Tidak)
"Heart Disease"	Kelainan jantung dari lahir (Ya / Tidak)
"Ever Married"	Status Menikah (Ya / Tidak)
"Work Type"	Status pekerjaan (Anak-anak, Pekerjaan Pemerintah, Tidak Bekerja, Swasta, atau Pekerja Lepas)
"Residence Type"	Status wilayah (Pedesaan / Perkotaan)
"Avg Glucose"	Kadar gula darah (mg/dL)
"Body Mass Index"	Indeks massa tubuh
"Smoking Status"	Status merokok (Dulu Merokok, Tidak Pernah Merokok, Merokok, atau Tidak Diketahui)
"Stroke"	Penyakit pembuluh darah otak (Ya / Tidak)

Data di atas terdapat beberapa variabel yakni variabel independen dan variabel dependen. Variabel independen meliputi Id (identifikasi unik), Gender (jenis kelamin), Age (usia), Hypertension (status hipertensi), Heart Disease (kelainan jantung), Ever Married (status menikah), Work Type (status pekerjaan), Residence Type (status wilayah), Avg Glucose (rata-rata kadar gula darah), Body Mass Index (BMI), dan Smoking Status (status merokok). Variabel dependen adalah Stroke (keberadaan penyakit pembuluh darah otak).. Dari sumber data tersebut tentu akan dilakukan pengolahan data agar data yang digunakan lebih akurat dalam langkah selanjutnya.

Tahapan-tahapan pelaksanaan penelitian klasifikasi penyakit stroke sebagai berikut :



Gambar. 2 Flowchart Penelitian.

A) Identifikasi Masalah

Tahap ini mengidentifikasi dan menjelaskan masalah yang dibahas. Masalah utama yang diidentifikasi adalah tingginya risiko terjadinya stroke pada individu dan perlunya memahami faktor-faktor yang berkontribusi terhadap risiko tersebut. Stroke merupakan salah satu penyakit tidak menular tetapi merupakan kegawatdaruratan neurologi yang dapat menimbulkan disabilitas bahkan kematian[16]. Jika faktor-faktor risiko ini tidak dipahami dengan baik, dampaknya bisa sangat

ISSN (Online) 2747-0563
Volume 4 Tahun 2024

serius, termasuk meningkatnya angka kematian, kualitas hidup yang menurun, dan biaya kesehatan yang tinggi. Serta evaluasi tingkat akurasi algoritma *Decision Tree* dan *Random Forest* dengan penerapan metode *hyperparameter tuning* maupun *non tuning* menjadi fokus utama dalam penelitian ini. Algoritma Decision Tree merupakan salah satu algoritma yang populer dan mudah dipahami karena hasilnya seperti otak manusia sehingga mudah untuk memahami aturan yang didapat dari hasil tersebut[1]. Sedangkan Random Forest adalah algoritma yang digunakan untuk mengklasifikasikan data dalam volume besar, melibatkan penggabungan beberapa pohon keputusan dan melatihnya menggunakan sampel data yang ada[13] Penelitian ini akan menguji algoritma tersebut dan membandingkan kinerja mereka untuk menemukan metode yang paling optimal dalam memprediksi risiko stroke.

B) Mengumpulkan Data

Mengumpulkan data merupakan tahapan krusial dalam penelitian ini karena data yang akurat dan representatif sangat penting untuk mendapatkan hasil yang valid dan dapat diandalkan. Berikut merupakan pemaparan lebih lengkap mengenai data yang digunakan pada tahap penelitian ini:

1. Informasi yang Dikumpulkan
   Informasi yang dikumpulkan dalam penelitian ini mencakup data terkait faktor-faktor risiko yang berhubungan dengan penyakit stroke. Variabel yang diamati meliputi informasi seperti usia, jenis kelamin, status hipertensi, penyakit jantung, status pernikahan, jenis pekerjaan, tipe tempat tinggal, rata-rata kadar glukosa darah, indeks massa tubuh (BMI), dan status merokok.
2. Sumber Data
   Pada pengumpulan data untuk penelitian ini diambil dari Kaggle, sebuah platform yang menyediakan berbagai dataset publik untuk digunakan dalam penelitian dan analisis.
3. Subjek Penelitian
   Melibatkan individu yang telah mengalami stroke atau yang berisiko tinggi terkena stroke. Memilih individu yang telah mengalami stroke atau berisiko tinggi terkena stroke sebagai subjek penelitian sesuai dengan konteks penelitian ini yang berkaitan dengan identifikasi faktor-faktor risiko dan pencegahan stroke.
4. Validasi Data
   Setelah data dikumpulkan, langkah berikutnya adalah memvalidasi keakuratan data. Proses ini melibatkan pemeriksaan kembali terhadap data untuk memastikan bahwa tidak ada kesalahan entri atau ketidaksesuaian yang dapat mempengaruhi hasil analisis.

C) Data *preprocessing*

Data preprocessing adalah langkah kritis dalam pipeline data science yang dapat mempengaruhi hasil akhir analisis secara signifikan. Dengan pemahaman yang mendalam tentang langkah-langkah preprocessing, data akan lebih siap menghadapi proyek yang dibangun [4] . Selain itu Data Preprocessing bertujuan untuk mengubah data mentah menjadi data yang berkualitas sehingga data layak untuk diolah pada tahapan selanjutnya [2].

1. Memeriksa *missing valueS*
   Memeriksa *missing value* dalam data penting untuk memastikan integritas dan validitas hasil penelitian. *Missing value* dapat mengganggu analisis data dan menghasilkan kesimpulan yang tidak akurat. Oleh karena itu, dengan memeriksa missing value, dapat mengidentifikasi data yang hilang atau tidak lengkap.
2. Memastikan tidak ada data yang terduplikat
   Dalam tahap ini kita menggunakan metode `duplicated()` pada data untuk mengidentifikasi baris-baris yang memiliki data yang sama dan memastikan ketiadaan duplikasi data adalah faktor penting dalam analisis data karena dapat mempengaruhi validitas dan hasil dari pemodelan atau analisis yang dilakukan.
3. Mengganti tipe data
   Mengganti tipe data adalah langkah di mana tipe data dari suatu variabel atau kolom dalam dataset diubah menjadi tipe data yang lebih sesuai untuk analisis atau pemodelan yang akan dilakukan.

D) EDA *(Exploratory Data Analysis)*

EDA atau *Exploratory Data Analysis* langkah ini dilakukan proses eksplorasi dan analisis data yang dilakukan untuk memahami karakteristik, pola, dan hubungan antara variabel dalam dataset. Tujuan dari EDA adalah untuk mendapatkan wawasan yang lebih dalam tentang data, mengidentifikasi outlier atau data yang tidak biasa, serta menemukan pola atau tren yang mungkin terdapat dalam data.

1. Statistika deskriptif
   Statistika Deskriptif tahap ini berkaitan dengan pengumpulan, penafsiran, dan penyajian data secara numerik atau grafis untuk memberikan gambaran yang ringkas tentang karakteristik data. Tujuan dari statistika deskriptif adalah untuk menggambarkan dan meringkas data secara statistik agar mudah dipahami.
2. Korelasi
   Korelasi digunakan untuk mengukur sejauh mana hubungan antara dua variabel. Korelasi menggambarkan arah (positif atau negatif) dan kekuatan hubungan linier antara variabel-variabel tersebut
3. Visualisasi
   Visualisasi data berisi proses menggunakan grafik atau representasi visual lainnya untuk

ISSN (Online) 2747-0563
Volume 4 Tahun 2024

menyajikan informasi atau pola yang terkandung dalam data. Visualisasi data memiliki peran penting dalam eksplorasi data, pemahaman karakteristik data, dan komunikasi hasil analisis kepada pemangku kepentingan.

E) Data *preparation*

Data preparation adalah tahap dalam analisis data di mana data mentah yang dikumpulkan dari berbagai sumber diperbaiki, disusun, dan ditransformasikan untuk mempersiapkannya menjadi bentuk yang sesuai untuk analisis lebih lanjut.

1. Label encoding

   Encoding adalah salah satu tahap preprocessing data, di mana tipe data kategori diubah menjadi tipe data numerik sebelum diproses oleh algoritma machine learning[9].

   - *Gender = 0 / 1 = e*
   - *Ever married = 0 / 1*
   - *Work type = 0 / 1 / 2 / 3*
   - *Residence type = 0 / 1*
   - *Smoking status = 0 / 1*

   Hal ini diperlukan karena sebagian besar algoritma machine learning membutuhkan input berupa bilangan numerik, sedangkan data kategorikal biasanya direpresentasikan dalam bentuk string.

2. *Splitting* data

   Pembagian data pada langkah ini pemodelan yang melibatkan memisahkan dataset menjadi subset yang berbeda menjadi data *training* 80 % dan data *testing* 20 %. Tujuan utama dari pembagian data adalah untuk mengukur kinerja model secara objektif.

3. Balancing data

   Balancing data menggunakan Metode Synthetic Minority Over-sampling Technique(SMOTE) merupakan metode yang populer diterapkan dalam rangka menangani ketidak seimbangan kelas. Teknik ini mensintesis sampel baru dari kelas minoritas untuk menyeimbangkan dataset dengan cara sampling ulang sampel kelas minoritas[3]. SMOTE membantu mengatasi masalah ini dengan mensintesis contoh baru dari kelas minoritas. Teknik ini bekerja dengan memilih sampel dari kelas minoritas dan kemudian menciptakan sampel baru yang terletak di antara sampel yang dipilih dan tetangganya terdekat dalam ruang fitur.

F) Modelling

1. Tanpa Tuning

   Pemodelan adalah proses membangun model prediktif atau model analitis yang digunakan untuk mengambil informasi, membuat prediksi, atau melakukan analisis pada data. Dalam konteks analisis data, pemodelan bertujuan untuk menggambarkan hubungan antara variabel-variabel dalam dataset dan memprediksi nilai atau perilaku yang diinginkan. Pada proses ini pemodelan dibangun dua model yaitu model *Decision Tree* dan *Random Fores*t tanpa *tuning*.

2. *Hyperparameter Tuning*

   Pengaturan parameter merupakan langkah dalam mengoptimalkan model machine learning, yang bertujuan untuk meningkatkan akurasi prediksi. Dalam penelitian "Prediksi Penyakit Stroke menggunakan Algoritma *Decision Tree* dan *Random Forest* dengan Metode *Hyperparameter Tuning*", digunakan metode *grid search* untuk mencari kombinasi parameter terbaik dari model. Grid search merupakan salah satu metode yang paling umum digunakan dalam eksplorasi hyperparameter tuning. Metode ini bekerja dengan melakukan pencarian secara menyeluruh terhadap semua kombinasi hyperparameter yang telah ditentukan pada grid konfigurasi. Kelebihan grid search adalah kemampuannya untuk dijalankan secara paralel, di mana setiap percobaan berjalan secara mandiri [7].

G) Evaluasi performa

1. Classfication report

   Dalam penelitian ini, evaluasi performa klasifikasi metode Decision Tree dan Random Forest dilakukan untuk mengetahui seberapa baik kedua model tersebut dalam memprediksi kategori yang sesuai untuk data yang diberikan. Metode yang digunakan untuk melakukan evaluasi performa adalah menggunakan sklearn classification report memberikan ringkasan yang komprehensif tentang metrik evaluasi kinerja model klasifikasi, termasuk akurasi, presisi, recall, dan F1-score.

2. Confusion Matriks

   Matrik konfusi adalah tabel yang digunakan untuk menggambarkan kinerja model klasifikasi atau prediksi pada suatu set data testing yang nilai nilai sebenarnya sudah diketahui[15]. Confusion Matrix mengevaluasi kinerja klasifikasi. False Positive (FP) dan False Negative (FN) mengacu pada contoh di mana kelas positif dan negatif salah diklasifikasikan. Sebaliknya, True Negative (TN) mengacu pada contoh di mana kelas positif dan negatif diklasifikasikan dengan benar[11].

III. HASIL DAN PEMBAHASAN

Pada bab ini menjelaskan tentang hasil implementasi klasifikasi untuk prediksi penyakit stroke menggunakan algoritma Decision Tree dan Random Forest yang sudah dituliskan secara runtut pada Bab III Metodologi Penelitian. Bab ini menganalisis dan mengidentifikasi data dari hasil pengolahan dengan metode hyperparameter tuning dan hasil

ISSN (Online) 2747-0563
Volume 4 Tahun 2024

klasifikasi untuk menentukan model yang paling akurat dalam memprediksi risiko stroke pada pasien. Hasil implementasi ini akan mencakup evaluasi kinerja masing-masing model sebelum dan setelah penerapan hyperparameter tuning, serta interpretasi dari metrik evaluasi yang digunakan untuk menilai efektivitas model.

A. *Pengumpulan data*

Mengimpor library yang dibutuhkan yaitu numpy dan pandas sebagai np dan pd. Numpy adalah pustaka yang digunakan untuk melakukan operasi matematika dan manipulasi array multidimensi, sedangkan pandas adalah pustaka yang digunakan untuk analisis data dan manipulasi dataset tabular. Selanjutnya menggunakan fungsi pd.read_csv() untuk membaca file "healthcare-dataset-stroke-data.csv" yang telah kita dapatkan dari Kaggle dan menyimpannya ke dalam objek DataFrame yang disebut 'df'. Fungsi ini membaca file CSV dan mengonversinya menjadi struktur data tabular yang disebut DataFrame, yang memungkinkan untuk melakukan analisis dan manipulasi data lebih lanjut.

[17]:

	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	NaN	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	NaN	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	Unknown	0

5110 rows × 12 columns

Gambar. 3 Stroke Dataset.

B. *Preprocessing Data*

1. *Memeriksa Duplikat Data*

   Pada tahap ini tujuan utamanya adalah untuk menghitung dan menampilkan jumlah kemunculan data redundan pada setiap baris dalam DataFrame 'df' , baris-baris duplikat dalam DataFrame 'df' akan dihapus namun memang tidak ditemukan data redundan pada data.

2. *Mengganti Tipe Data*

   Melakukan penyesuaian data untuk pemodelan prediksi stroke. Pertama, kolom "id" yang tidak relevan dihapus dari dataset. Selanjutnya, variabel kategorik seperti gender, ever_married, work_type, Residence_type, dan smoking_status dimasukkan kedalam variabel numerik ataupun kategorik.

3. *Cek Missing Value*

   Dari hasil pemeriksaan nilai yang hilang, hanya kolom bmi yang memiliki nilai yang kosong sebanyak 201 entri. Untuk menangani nilai yang hilang pada kolom BMI, kita dapat menggunakan SimpleImputer dengan strategi 'mean' untuk mengisi nilai kosong dengan rata-rata dari kolom tersebut. Dengan melakukan hal ini, kita akan mempertahankan integritas data dan memastikan konsistensi dalam analisis dan pemodelan selanjutnya.

```
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 Missing Value.

C. *EDA (Exploratory Data Analysis)*

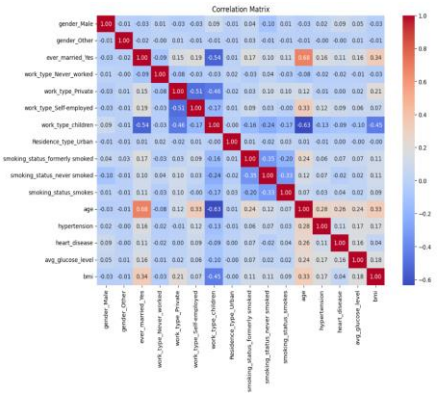
1. *Statistika Deskriptif*

	age	hypertension	heart_disease	avg_glucose_level	bmi	stroke
count	5110.000000	5110.000000	5110.000000	5110.000000	4909.000000	5110.000000
mean	43.226614	0.097456	0.054012	106.147677	28.893237	0.048728
std	22.612647	0.296607	0.226063	45.283560	7.854067	0.215320
min	0.080000	0.000000	0.000000	55.120000	10.300000	0.000000
25%	25.000000	0.000000	0.000000	77.245000	23.500000	0.000000
50%	45.000000	0.000000	0.000000	91.885000	28.100000	0.000000
75%	61.000000	0.000000	0.000000	114.090000	33.100000	0.000000
max	82.000000	1.000000	1.000000	271.740000	97.600000	1.000000

Gambar. 5 Missing Value.

Statistika Deskriptif tahap ini berkaitan dengan pengumpulan, penafsiran, dan penyajian data secara numerik atau grafis untuk memberikan gambaran yang ringkas tentang karakteristik data. Melalui statistika deskriptif, kita dapat memahami distribusi, tendensi pusat, variabilitas yang mungkin ada dalam dataset, sehingga memungkinkan untuk membuat kesimpulan awal dan mengidentifikasi tren atau anomali yang mungkin perlu diteliti lebih lanjut.

2. *Korelasi*

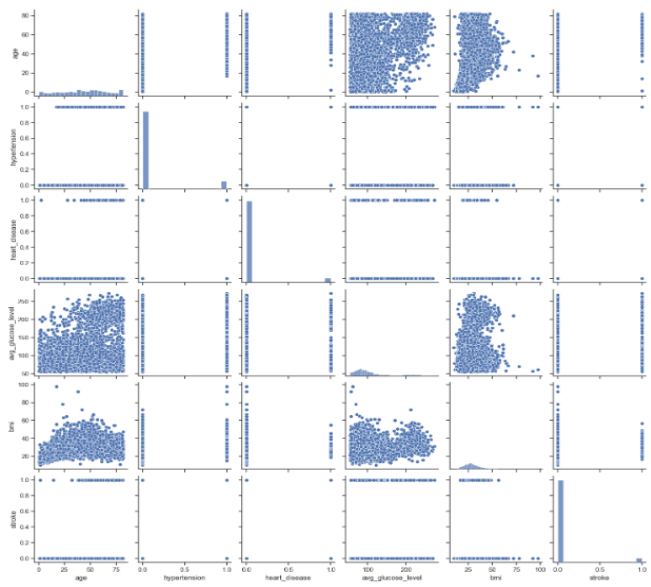


Gambar. 6 Stroke Dataset Correlation.

Setiap sel pada heatmap menunjukkan tingkat korelasi antara dua kolom numerik yang bersesuaian. Angka-angka di dalam setiap sel mewakili nilai korelasi antara pasangan kolom tersebut. Skala warna pada heatmap juga membantu dalam memvisualisasikan tingkat korelasi, di mana warna-warna yang lebih pekat (biru / merah) menunjukkan korelasi yang lebih kuat.

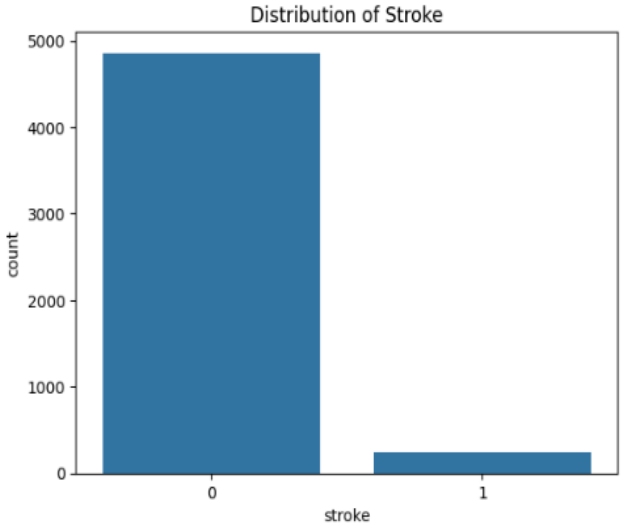
ISSN (Online) 2747-0563
Volume 4 Tahun 2024

3. Visualisasi



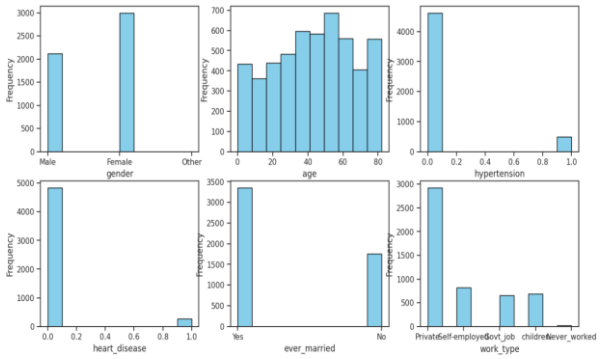
Gambar. 7 Visualisasi Scatter Plot.

Melalui scatter plot, kita dapat dengan cepat melihat apakah ada pola, tren, atau korelasi antara variabel-variabel yang diamati. Ini membantu kita untuk membuat asumsi awal tentang hubungan antara variabel tersebut dan memandu langkah-langkah analisis selanjutnya. Dengan melihat visualisasi ini, kita dapat membuat keputusan yang lebih tepat dalam menjalankan analisis data lebih lanjut atau merancang penelitian lebih lanjut.



Gambar. 8 Distribusi Stroke.

Gambar dsitribusi kelas penyakit stroke diatas memberikan visualisasi yang jelas tentang seberapa banyak data yang termasuk dalam setiap kategori stroke, pada data terlihat bahwa distribusi pada data menampilkan data orang yang tidak terkena stroke (0) lebih banyak daripada orang yang terkena stroke (1).



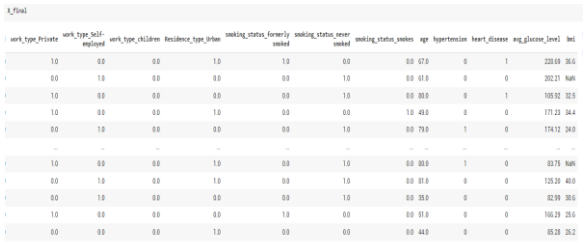
Gambar. 9 Visualisasi Histogram.

Plot histogram yang dihasilkan memberikan gambaran visual tentang distribusi frekuensi dari setiap variabel yang diamati dalam data. Ini memberikan pemahaman tentang sebaran nilai-nilai dalam setiap variabel dan membantu mengidentifikasi pola atau karakteristik penting dalam data tersebut.

D. *Data Preparation*

1. Label Encoding

Melakukan label encoding terhadap variabel kategorik seperti gender, ever_married, work_type, residence_type, dan smoking_status diubah menjadi bentuk numerik dengan one-hot encoding menggunakan OneHotEncoder. Terakhir, kolom-kolom yang telah dienkoding digabungkan kembali dengan kolom-kolom numerik untuk membentuk dataset final yang siap untuk analisis dan pemodelan.



Gambar. 10 Output *Label Encoding*.

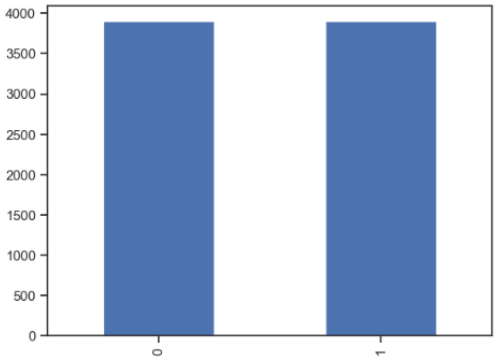
2. Splitting Data

Pada tahap split data, kita menggunakan fungsi train_test_split dari scikit-learn untuk membagi dataset menjadi data latih (X_train, y_train) dan data uji (X_test, y_test). Dengan parameter test_size=0.2, kita memilih untuk mengalokasikan 20% dari data sebagai data uji, sementara 80% sisanya akan digunakan sebagai data latih. Penggunaan random_state=42 memastikan bahwa pembagian data dilakukan secara konsisten setiap kali proses ini dijalankan, sehingga memudahkan reproducibility dalam analisis dan pemodelan. Lalu dihapuskanlah atribut target dari DataFrame serta dibuat DataFrame baru yaitu ‘x’. Lalu untuk atribut target dibuatkan

ISSN (Online) 2747-0563
Volume 4 Tahun 2024

DataFrame baru yaitu ‘y’ yang berisikan atribut target saja.

3. Balancing Data



Gambar. 11 Output *Balancing Data*.

Setelah dilakukan data balancing pada data terlihat bahwa distribusi pada data menampilkan data orang yang tidak terkena stroke (0) telah seimbang dengan orang yang terkena stroke (1).

E. *Modeling*

1. *Decision Tree*

Dengan menggunakan fungsi Decision Tree Classifier yang diimpor dari library sklearn dilakukan prediksi pada ‘x_train’ dan ‘y_train’. Setelah dibuat model random forestnya, model tersebut digunakan untuk memprediksi atribut target dari ‘x_test’.

2. *Random Forest*

Dengan menggunakan fungsi RandomForestClassifier yang diimpor dari library sklearn dilakukan prediksi pada ‘x_train’ dan ‘y_train’. Setelah dibuat model random forestnya, model tersebut digunakan untuk memprediksi atribut target dari ‘x_test’.

F. *Evaluasi Performa*

1. *Decision Tree Tanpa Tuning*

TABEL II PERFORMA *DECISION TREE* TANPA TUNING

	PRECISION	**RECALL**	**F1-SCORE**	**SUPPORT**
0	0.95	0.95	0.95	781
1	0.95	0.96	0.95	780
ACCURACY			0.95	1561
MACRO AVG	0.95	0.95	0.95	1561
WEIGHTED AVG	0.95	0.95	0.95	1561

Model decision tree tanpa tuning menunjukkan performa yang sangat baik dengan precision, recall, dan F1-score semuanya berada di 0.95 untuk kedua kelas. Akurasi keseluruhan model adalah 95%. Ini menunjukkan bahwa model mampu membuat prediksi yang sangat akurat dan seimbang antara kedua kelas.

TABEL III CONFUSION MATRIX PERFORMA DECISION TREE TANPA TUNING

Actual/Predicted	Peredicted 0	Peredicted 1
Acctual 0	744	37
Actual 1	35	745

Confusion matrix ini menunjukkan bahwa model decision tree memiliki performa yang seimbang dalam mendeteksi kedua kelas. Dengan 744 true negatives dan 745 true positives, model menunjukkan kemampuan yang kuat dalam memprediksi kedua label. Jumlah false positives (37) dan false negatives (35) relatif kecil dibandingkan dengan total prediksi, menunjukkan tingkat kesalahan yang rendah. Ini mendukung interpretasi awal bahwa model bekerja dengan sangat baik tanpa tuning tambahan.

2. *Random Forest Tanpa Tuning*

TABEL IV PERFORMA *RANDOM FOREST* TANPA TUNING

	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.96	0.99	0.97	781
1	0.99	0.96	0.97	780
ACCURACY			0.97	1561
MACRO AVG	0.97	0.97	0.97	1561
WEIGHTED AVG	0.97	0.97	0.97	1561

Model random forest tanpa tuning menunjukkan performa yang sangat baik dengan akurasi keseluruhan 97%. Untuk kelas 0, precision mencapai 0.96 dan recall 0.99, sedangkan untuk kelas 1, precision mencapai 0.99 dan recall 0.96. F1-score untuk kedua kelas adalah 0.97, menunjukkan keseimbangan yang baik dalam mendeteksi kejadian dari kedua kelas. Rata-rata makro dan berbobot untuk precision, recall, dan F1-score semuanya berada di 0.97, mengindikasikan konsistensi model dalam memprediksi kedua kelas.

TABEL V PERFORMA *RANDOM FOREST* TANPA TUNING

Actual/Predicted	Peredicted 0	Peredicted 1
Acctual 0	773	8
Actual 1	35	745

ISSN (Online) 2747-0563
Volume 4 Tahun 2024

Confusion matrix ini menunjukkan bahwa model random forest tanpa tuning memiliki performa yang sangat baik dan seimbang dalam mendeteksi kedua kelas. Dengan 773 true negatives dan 745 true positives, model menunjukkan kemampuan yang kuat dalam memprediksi kedua label. Jumlah false positives (8) sangat kecil, menunjukkan model memiliki tingkat kesalahan yang sangat rendah dalam memprediksi kejadian kelas 0 sebagai kelas 1. Meskipun ada 35 false negatives, jumlah ini masih kecil dibandingkan dengan total prediksi.

3. *Decision Tree Dengan Hyperparameter Tuning*

TABEL VI PERFORMA DECISION TREE DENGAN HYPERPARAMETER TUNING

	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.95	0.96	0.96	781
1	0.96	0.95	0.96	780
ACCURACY			0.96	1561
MACRO AVG	0.96	0.96	0.96	1561
WEIGHTED AVG	0.96	0.96	0.96	1561

Model decision tree dengan hyperparameter tuning menunjukkan performa yang sangat baik dengan akurasi keseluruhan 96%. Untuk kelas 0, precision mencapai 0.95 dan recall 0.96, sedangkan untuk kelas 1, precision mencapai 0.96 dan recall 0.95. F1-score untuk kedua kelas adalah 0.96, menunjukkan keseimbangan yang baik dalam mendeteksi kejadian dari kedua kelas. Rata-rata makro dan berbobot untuk precision, recall, dan F1-score semuanya berada di 0.96, mengindikasikan konsistensi model dalam memprediksi kedua kelas. Hasil ini menunjukkan bahwa tuning hyperparameter berhasil meningkatkan kemampuan model untuk membuat prediksi yang akurat dan seimbang.

TABEL VII CONFUSION MATRIX PERFORMA DECISION TREE DENGAN HYPER TUNING

Actual/Predicted	Peredicted 0	Peredicted 1
Acctual 0	751	30
Actual 1	37	743

Confusion matrix ini menunjukkan bahwa model decision tree dengan hyperparameter tuning memiliki performa yang sangat baik dan seimbang dalam mendeteksi kedua kelas. Dengan 751 true negatives dan 743 true positives, model menunjukkan kemampuan yang kuat dalam memprediksi kedua label. Jumlah false positives (30) dan false negatives (37) relatif kecil dibandingkan dengan total prediksi, menunjukkan tingkat kesalahan yang rendah.

4. *Random Forest Dengan Hyperparameter Tuning*

TABEL VIII PERFORMA RANDOM FOREST DENGAN HYPERPARAMETER TUNING

	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.96	0.99	0.98	781
1	0.99	0.96	0.97	780
ACCURACY			0.98	1561
MACRO AVG	0.98	0.98	0.98	1561
WEIGHTED AVG	0.98	0.98	0.98	1561

Model random forest dengan hyperparameter tuning menunjukkan performa yang luar biasa dengan akurasi keseluruhan 98%. Untuk kelas 0, precision mencapai 0.96 dan recall 0.99, sedangkan untuk kelas 1, precision mencapai 0.99 dan recall 0.96. F1-score adalah 0.98 untuk kelas 0 dan 0.97 untuk kelas 1, menunjukkan keseimbangan yang sangat baik dalam mendeteksi kejadian dari kedua kelas. Rata-rata makro dan berbobot untuk precision, recall, dan F1-score semuanya berada di 0.98, mengindikasikan bahwa model memiliki konsistensi yang sangat tinggi dalam prediksi. Hasil ini menunjukkan bahwa tuning hyperparameter telah berhasil meningkatkan kemampuan model untuk membuat prediksi yang sangat akurat dan seimbang.

TABEL IX CONFUSION MATRIX PERFORMA RANDOM FOREST DENGAN HYPERPARAMETER TUNING

Actual/Predicted	Peredicted 0	Peredicted 1
Acctual 0	777	4
Actual 1	35	745

Confusion matrix ini menunjukkan bahwa model random forest dengan hyperparameter tuning memiliki performa yang sangat baik dalam mendeteksi kedua kelas. Dengan 777 true negatives dan 745 true positives, model menunjukkan kemampuan yang kuat dalam memprediksi kedua label. Jumlah false positives (4) sangat kecil,

ISSN (Online) 2747-0563
Volume 4 Tahun 2024

menunjukkan bahwa model jarang salah memprediksi kejadian kelas 0 sebagai kelas 1. Meskipun ada 35 false negatives, jumlah ini masih kecil dibandingkan dengan total prediksi. Secara keseluruhan, model bekerja dengan sangat baik setelah tuning hyperparameter, mendukung akurasi keseluruhan 98% dan menunjukkan keseimbangan yang sangat baik dalam mendeteksi kedua kelas.

G. *Hasil*

TABEL X PERFORMA RANDOM FOREST DENGAN HYPERPARAMETER TUNING

	***Accuracy* (Tanpa *Tuning*)**	***Accuracy* (*Hyperparameter Tuning*)**
Decision Tree	95 %	96 %
Random Forest	97 %	98%

Dari hasil evaluasi, model *Decision Tree* menunjukkan peningkatan akurasi dari 95% menjadi 96% setelah proses *hyperparameter tuning*, sementara *Random Forest* juga mengalami peningkatan akurasi, pada 97% menjadi 98%. Hasil yang diperoleh dari evaluasi model mendukung hipotesis pertama dan kedua, di mana model *Decision Tree* dan *Random Forest* memang menunjukkan peningkatan kinerja setelah *hyperparameter tuning*. Namun, memang kedua model ini tetap tidak menunjukkan peningkatan yang signifikan setelah *hyperparameter tuning*.

1. *Best parameters for Decision Tree:*
   {'criterion': 'gini', 'max_depth': 10, 'min_samples_leaf': 4, 'min_samples_split': 10}
   Dengan parameter-parameter ini, model Decision Tree mampu mencapai akurasi sebesar 96%, yang merupakan peningkatan yang signifikan dari akurasi sebelum penyetelan *hyperparameter*.
2. *Best parameters for Random Forest:*
   {'criterion': 'entropy', 'max_depth': 20, 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators': 200}
   Dengan parameter-parameter ini, model *Random Forest* mampu mencapai akurasi sebesar 98%, yang merupakan peningkatan yang signifikan dari akurasi sebelum penyetelan *hyperparameter*.

IV. KESIMPULAN

Dari hasil penelitian ini, dapat disimpulkan bahwa hyperparameter tuning memberikan pengaruh yang kurang signifikan terhadap peningkatan kinerja model *Decision Tree* dan *Random Forest* dalam memprediksi risiko stroke. Meskipun kedua model awalnya telah menunjukkan akurasi yang tinggi, peningkatan kinerja setelah *hyperparameter* tuning tidak signifikan.

Namun dari hasil, dapat disimpulkan bahwa kedua model hyper parameter tuning mengalami peningkatan dalam tingkat akurasinya, mendekati tingkat akurasi Random Forest setelah dilakukannya hyperparameter tuning. Dalam kedua kondisi, baik tanpa tuning maupun dengan hyperparameter tuning, Random Forest memiliki tingkat akurasi yang sedikit lebih tinggi daripada Decision Tree. Oleh karena itu, Random Forest dengan hyperparameter tuning dapat dianggap sebagai pilihan yang lebih baik dan dapat diandalkan dalam penelitian ini.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih yang sebanyak-banyaknya kepada Dosen Pembimbing mata kuliah Penelitian Sains Data, Bapak Dr.Eng.Ir.Dwi Arman Prasetya.,ST.,MT.,IPU., Asean. Eng yang telah memberikan berbagai saran dan masukannya dalam penyelesaian artikel ini. Serta Terima kasih disampaikan kepada Tim SANTIKA yang telah meluangkan waktu untuk membuat template ini.

REFERENSI

[1] Akbar, F., Saputra, H. W., Maulaya, A. K., Hidayat, M. F., & Rahmaddeni, R. (2022). Implementasi Algoritma decision Tree C4.5 Dan Support Vector Regression Untuk Prediksi Penyakit stroke. MALCOM: Indonesian Journal of Machine Learning and Computer Science, 2(2), 61–67. https://doi.org/10.57152/malcom.v2i2.426

[2] Alghifari, F., & Juardi, D. (2021). Penerapan Data Mining Pada penjualan Makanan Dan Minuman menggunakan metode algoritma naïve bayes. JURNAL ILMIAH INFORMATIKA, 9(02), 75–81. https://doi.org/10.33884/jif.v9i02.3755

[3] Barus, S. G., Widiyanto, D., & Santoni, M. M. (2022). KLASIFIKASI SENTIMEN DATA TIDAK SEIMBANG MENGGUNAKAN ALGORITMA SMOTE DAN K-NEAREST NEIGHBOR PADA ULASAN PENGGUNA APLIKASI PEDULILINDUNGI. *SENAMIKA*, *3*(2).

[4] Hamdani, I. M., Nurhidayat, N., Karman, A., Fuady Adhalia H, N., & Julyaningsih, A. H. (2024a). Edukasi dan Pelatihan Data Science dan Data Preprocessing. *INTISARI: JURNAL INOVASI PENGABDIAN MASYARAKAT* , *2*(1).

[5] Husein, A. M., Lubis, F. R., & Harahap, M. K. (2021). Analisis Prediktif Untuk Keputusan bisnis : Peramalan Penjualan. *Data Sciences Indonesia (DSI)*, *1*(1), 32–40. https://doi.org/10.47709/dsi.v1i1.1196

[6] Iqbal, M., Hendri Mahmud Nawawi, Ramadhan Saelan, M. R., Sony Maulana, M., Yudhistira, & Mustopa, A. (2023). OPTIMASI hyperparameter multilayer perceptron untuk prediksi Daya Beli mobil. Jurnal Manajemen Informatika Dan Sistem Informasi, 6(1), 73–81. https://doi.org/10.36595/misi.v6i1.739

[7] Joshua Agung Nurcahyo, & Theopilus Bayu Sasongko. (2023). Hyperparameter tuning ALGORITMA supervised learning untuk KLASIFIKASI Keluarga Penerima Bantuan Pangan beras. *Indonesian Journal of Computer Science*, *12*(3). https://doi.org/10.33022/ijcs.v12i3.3254

[8] Kusuma, A. P., Utami, I. T., & Purwono, J. (2022). PENGARUH TERAPI "MENGGENGGAM BOLA KARET BERGERIGI" TERHADAP PERUBAHAN KEKUATAN OTOT PADA PASIEN STROKE DIUKUR MENGGUNAKAN HANGRY DYNAMOMETER RUANG SYARAF RSUD JENDAYANI KOTA METRtO. Jurnal Cendikia Muda, 2(1).

[9] Kristiawan, K., & Widjaja, A. (2021). Perbandingan Algoritma machine learning Dalam Menilai Sebuah Lokasi toko ritel. *Jurnal Teknik Informatika Dan Sistem Informasi*, *7*(1). https://doi.org/10.28932/jutisi.v7i1.3182

ISSN (Online) 2747-0563
Volume 4 Tahun 2024

[10] Nababan, A. H., & Hutagalung, M. Y. (2023). Hyperparameter Tuning Pada Model Stance Detection Menggunakan GridSearchCV. *Jurnal Sains Dan Teknologi* , *5*(1).
[11] Priatna, W. (2024). Dampak Pengambilan Sampel Data untuk Optimalisasi Data tidak seimbang pada Klasifikasi Penipuan Transaksi E-Commerce. *Indonesian Journal of Computer Science*, *13*(2).
[12] Ramadhon, R. N., Ogi, A., & Agung, A. P. (2024). Implementasi Algoritma Decision Tree untuk Klasifikasi Pelanggan Aktif atau Tidak Aktif pada Data Bank . Karimah Tauhid.
[13] Sagita, A., Faqih, A., Dwilestari, G., Siswoyo, B., & Pratama, D. (2024). Penerapan metode random forest Dalam Menganalisis Sentimen pengguna aplikasi capcut di google play store. JATI (Jurnal Mahasiswa Teknik Informatika), 7(6), 3307–3313. https://doi.org/10.36040/jati.v7i6.8205.
[14] Sukmawati, E., Bura Mare, A. C., & Marcello, S. A. (2024). Upaya Pencegahan resiko stroke pada LANSIA melalui pendidikan kesehatan di Panti
[15] Sriyanto, & Supriyatna, A. R. (2023). Prediksi Penyakit Diabetes Menggunakan Algoritma Random Forest . JURNAL TEKNIKA .
[16] Warouw, F., & Wilar, R. (2023). Peningkatan Pengetahuan Tentang Cara Identifikasi Dan Upaya Preventif Faktor-Faktor Resiko Stroke Pada Masyarakat Pesisir Desa Atep Oki. *Jurnal Pengabdian Masyarakat*, *1*(1).
[17] Werdha surya jemursari surabaya. *BERDAYA: Jurnal Pendidikan Dan Pengabdian Kepada Masyarakat*, *6*(1), 111–116. https://doi.org/10.36407/berdaya.v6i1.1162