ALGORITMA RANDOM FORETS UNTUK PENINGKATAN MODEL KLASIFIKASI PADA DATA DIAGNOSA PASIEN PUSKESMAS PEKALANGAN KOTA CIREBON

Arisa, Rudi Kurniawan, Umi Hayati
Teknik Informatika, STMIK IKMI Cirebon
Jl. Perjuangan No.10B, Karyamulya, Kec. Kesambi, Kota Cirebon, Jawa Barat, Indonesia
arisar375@gmail.com

ABSTRAK

Puskesmas Pekalangan Kota Cirebon menghadapi tantangan dalam memanfaatkan data diagnosis pasien secara efektif untuk mendukung keputusan medis. Pengelolaan data yang tidak seimbang dan kompleks sering kali menghambat akurasi klasifikasi diagnosis. Penelitian ini bertujuan untuk meningkatkan akurasi model klasifikasi diagnosis pasien menggunakan algoritma Random Forest dengan optimasi parameter "number of trees" dan "max depth". Metode penelitian ini mengadopsi pendekatan Knowledge Discovery in Databases (KDD) yang mencakup prapemrosesan data, seleksi fitur, dan evaluasi model. Dataset terdiri dari 3.769 data rekam medis pasien Puskesmas Pekalangan periode Januari hingga Juni 2024. Hasil penelitian menunjukkan bahwa parameter optimal "number of trees" adalah 28 dan "max depth" adalah 10, keduanya menghasilkan akurasi model sebesar 76,39%. Selain itu, atribut "keluhan utama" terbukti menjadi faktor yang paling berpengaruh terhadap prediksi, dengan akurasi mencapai 53,58%. Temuan ini menegaskan pentingnya pemilihan parameter yang tepat dan seleksi fitur dalam meningkatkan efisiensi serta keandalan model klasifikasi. Implementasi model ini diharapkan mampu mendukung pengambilan keputusan medis yang lebih akurat dan cepat di tingkat puskesmas.

Kata kunci : *Random Forest, Akurasi, Parameter Optimal, KDD Process, Data Klinis*

1. PENDAHULUAN

Perkembangan pesat di bidang informatika telah memberikan dampak yang signifikan pada berbagai aspek kehidupan, termasuk teknologi, bisnis, pendidikan, dan layanan kesehatan. Dalam layanan kesehatan, penerapan teknologi berbasis kecerdasan buatan dan machine learning telah membantu mengoptimalkan pengolahan data yang kompleks dan besar, terutama pada bidang diagnosis penyakit. Algoritma Random Forest, sebagai salah satu metode machine learning yang populer, telah menunjukkan kinerjanya dalam menghasilkan prediksi yang akurat pada berbagai aplikasi data medis [14].

Misalnya, penelitian sebelumnya menunjukkan bahwa algoritma ini mampu meningkatkan efisiensi dan akurasi pada prediksi stunting, penyakit jantung, serta diabetes [13]

Potensi besar algoritma ini juga terlihat pada berbagai studi yang memanfaatkannya untuk klasifikasi data diagnosis dan pengambilan keputusan yang lebih efektif di bidang Kesehatan [4].

Dalam dunia kesehatan, pengelolaan data medis menjadi tantangan yang semakin kompleks seiring meningkatnya jumlah data yang dihasilkan setiap hari. Puskesmas, sebagai fasilitas kesehatan tingkat pertama, sering menghadapi tantangan dalam memanfaatkan data diagnosis pasien secara efektif untuk mendukung pengambilan keputusan medis. Salah satu masalah utama adalah bagaimana mengklasifikasikan data dengan akurasi tinggi untuk menghasilkan diagnosa yang lebih tepat [12].

Puskesmas Pekalangan Kota Cirebon, sebagai salah satu fasilitas kesehatan utama di tingkat kecamatan, memainkan peran penting dalam memberikan layanan kesehatan dasar kepada masyarakat. Dengan jumlah pasien yang terus meningkat, Puskesmas ini menghadapi tantangan besar dalam pengelolaan dan analisis data medis pasien. Data yang terus berkembang ini sering kali tidak dimanfaatkan secara optimal untuk mendukung pengambilan keputusan medis yang cepat dan akurat.

Penelitian ini bertujuan untuk meningkatkan akurasi diagnosa penyakit melalui penerapan algoritma Random Forest pada data diagnosa pasien Puskesmas Pekalangan Kota Cirebon. Algoritma ini dipilih karena kemampuannya dalam menangani data kompleks dan besar, serta meningkatkan akurasi klasifikasi dengan mengurangi overfitting [18].

Selain itu, penelitian ini juga berupaya mengatasi permasalahan ketidakseimbangan data yang sering ditemukan dalam data kesehatan.

Jika penelitian ini berhasil mencapai tujuannya, implementasi algoritma Random Forest diharapkan dapat memberikan kontribusi signifikan terhadap sistem informasi kesehatan. Hasil penelitian ini tidak hanya meningkatkan akurasi diagnosa penyakit, tetapi juga berpotensi mengurangi kesalahan medis dan meningkatkan kualitas pelayanan kesehatan di tingkat puskesmas.

2. TINJAUAN PUSTAKA

Tinjauan pustaka membahas teori dan penelitian terdahulu yang relevan sebagai dasar dalam mendukung penelitian ini. Landasan teori ini digunakan untuk memahami konsep dan metode yang diterapkan dalam penelitian ini [2].

2.1. Random Forest

Perkembangan pesat di bidang informatika telah memberikan dampak yang signifikan pada berbagai aspek kehidupan, termasuk teknologi, bisnis, pendidikan, dan layanan kesehatan. Dalam layanan kesehatan, penerapan teknologi berbasis kecerdasan buatan dan machine learning telah membantu mengoptimalkan pengolahan data yang kompleks dan besar, terutama pada bidang diagnosis penyakit. Algoritma Random Forest, sebagai salah satu metode machine learning yang populer, telah menunjukkan kinerjanya dalam menghasilkan prediksi yang akurat pada berbagai aplikasi data medis [14].

Misalnya, penelitian sebelumnya menunjukkan bahwa algoritma ini mampu meningkatkan efisiensi dan akurasi pada prediksi stunting, penyakit jantung, serta diabetes [13].

Potensi besar algoritma ini juga terlihat pada berbagai studi yang memanfaatkannya untuk klasifikasi data diagnosis dan pengambilan keputusan yang lebih efektif di bidang kesehatan [4].

2.2. Support Vector Machine (SVM)

Support Vector Machine (SVM) adalah algoritma yang digunakan untuk klasifikasi dan regresi dengan mencari hyperplane yang dapat memisahkan kelas-kelas dalam ruang fitur. SVM sangat efektif dalam masalah klasifikasi biner, seperti dalam aplikasi diagnostik medis, di mana membedakan antara dua kondisi atau penyakit sangat penting. SVM bekerja dengan memilih hyperplane yang memaksimalkan margin antara dua kelas untuk meningkatkan akurasi klasifikasi [4].

2.3. K-Nearest Neighbor (KNN)

K-Nearest Neighbor (KNN) adalah metode pembelajaran berbasis instance yang mengklasifikasikan data berdasarkan kedekatannya dengan data pelatihan lainnya. Algoritma ini bekerja dengan mencari "tetangga terdekat" dari titik data yang belum diketahui labelnya dan kemudian memberikan label berdasarkan mayoritas kelas dari tetangga tersebut. KNN cukup efektif dalam kasus dengan data kecil dan sederhana, tetapi bisa menjadi lambat dengan dataset besar karena harus menghitung jarak untuk setiap titik data [5].

2.4. Optimasi Hyperparameter

Optimasi hyperparameter adalah proses pencarian nilai terbaik untuk parameter dalam model pembelajaran mesin yang tidak dipelajari langsung dari data, seperti kedalaman pohon dalam Random Forest atau parameter kernel dalam SVM. Teknik yang sering digunakan untuk optimasi ini adalah Grid Search, yang mencoba kombinasi parameter yang telah ditentukan, dan Random Search, yang secara acak memilih kombinasi parameter. Optimasi ini bertujuan untuk meningkatkan akurasi model dan meminimalkan error prediksi dengan memilih pengaturan yang optimal [6].

2.5. Synthetic Minority Oversampling Technique (SMOTE)

SMOTE adalah metode oversampling yang digunakan untuk mengatasi masalah ketidakseimbangan kelas, di mana satu kelas memiliki lebih sedikit data dibandingkan kelas lainnya. SMOTE bekerja dengan mensintesis data baru untuk kelas minoritas dengan cara mengambil dua atau lebih titik data yang ada dan menghasilkan data sintetis berdasarkan perbedaan antar titik tersebut. Teknik ini efektif untuk meningkatkan kinerja model pada data yang tidak seimbang, yang sering terjadi dalam kasus medis di mana kondisi tertentu jarang terjadi [7].

2.6. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) adalah teknik reduksi dimensi yang digunakan untuk mengurangi jumlah fitur dalam dataset sambil mempertahankan sebagian besar informasi penting. PCA mengubah data ke dalam ruang baru yang disebut komponen utama, yang merupakan kombinasi linier dari fitur asli yang menjelaskan variasi terbesar dalam data. Dalam konteks algoritma klasifikasi, PCA digunakan untuk mengurangi kompleksitas model dan meningkatkan efisiensi komputasi, terutama ketika data memiliki banyak fitur dan berisiko menyebabkan overfitting [8].

2.7. Evaluasi Model

Evaluasi model Evaluasi model adalah langkah penting dalam proses pembelajaran mesin, di mana kinerja model diukur menggunakan berbagai metrik. Metrik yang umum digunakan antara lain akurasi, precision, recall, F1-score, dan ROC-AUC. Akurasi mengukur persentase prediksi yang benar, sementara precision dan recall lebih fokus pada keseimbangan antara deteksi positif dan negatif. F1-score adalah rata-rata harmonis dari precision dan recall, dan ROC-AUC mengukur kemampuan model dalam membedakan antara kelas yang berbeda. Persamaan yang digunakan untuk menghitung metrik evaluasi antara lain Akurasi, Precision, Recall, F1-score. Pemilihan metrik yang tepat bergantung pada jenis masalah dan tujuan penelitian [9].

2.8. Penerapan Algoritma pada Diagnosis Penyakit

Berbagai algoritma pembelajaran mesin, seperti Random Forest dan SVM, telah diterapkan dalam diagnosis penyakit untuk meningkatkan akurasi diagnosis dan mendukung keputusan medis. Algoritma ini digunakan untuk menganalisis data medis, seperti hasil tes laboratorium, riwayat medis, dan citra medis, untuk memprediksi kemungkinan penyakit. Misalnya, algoritma ini dapat digunakan untuk mendiagnosis penyakit jantung, diabetes, atau kanker dengan tingkat akurasi yang tinggi, membantu dokter dalam membuat keputusan yang lebih cepat dan tepat [10].

2.9. Kesehatan Mental dan Algoritma Klasifikasi

Di bidang kesehatan mental, algoritma klasifikasi seperti Random Forest dan SVM telah digunakan untuk mendiagnosis gangguan mental, termasuk depresi dan kecemasan. Penggunaan teknik klasifikasi ini melibatkan analisis data survei atau tes psikologis untuk mendeteksi pola yang mengindikasikan kondisi mental tertentu. Dengan menggunakan data historis yang dikumpulkan dari individu, algoritma ini dapat memberikan diagnosis yang akurat dan membantu dalam merancang rencana perawatan yang lebih personalisasi [11].

2.10. Pengaruh Reduksi Dimensi pada Kinerja Model

Reduksi dimensi adalah langkah penting dalam mengurangi kompleksitas data dan mempercepat waktu komputasi. Teknik seperti PCA membantu mengurangi jumlah fitur dalam dataset dengan mempertahankan informasi yang paling relevan. Dalam banyak kasus, pengurangan dimensi dapat meningkatkan kinerja model, terutama ketika data memiliki banyak fitur yang tidak terlalu relevan. Dengan mengurangi dimensi, model menjadi lebih efisien dan kurang rentan terhadap overfitting, yang sangat penting dalam aplikasi medis di mana akurasi dan kecepatan sangat dibutuhkan [12].

3. METODE PENELITIAN

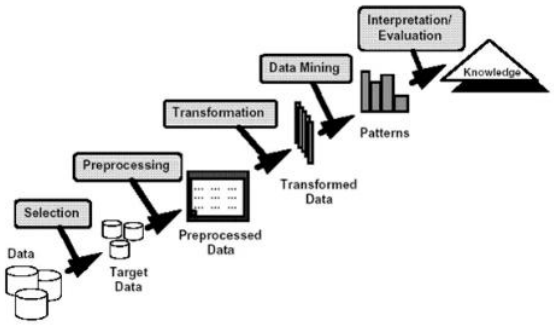
Penelitian ini menggunakan pendekatan Knowledge Discovery in Database (KDD) untuk meningkatkan akurasi algoritma Random Forest dalam klasifikasi data diagnosis pasien di Puskesmas Pekalangan, Kota Cirebon. Tahapan penelitian meliputi analisis karakteristik dataset untuk memahami pola data, seleksi atribut relevan seperti nama pasien, NIK, RT/RW, kelurahan, umur, jenis kunjungan, poli, asuransi, keluhan utama, suhu, dan diagnosa, serta pra-pemrosesan data untuk menangani data hilang, inkonsistensi, dan normalisasi. Selanjutnya, data ditransformasi ke format yang sesuai sebelum diterapkan pada algoritma Random Forest, yang dioptimalkan untuk meningkatkan akurasi model prediktif. Evaluasi dilakukan menggunakan metrik akurasi, presisi, recall, dan F1-score untuk memastikan performa model. Pendekatan ini dirancang untuk menghasilkan model prediksi diagnosis yang akurat dan mendukung pengambilan keputusan medis berbasis data klinis.

3.1. Sumber Data

Penelitian ini menggunakan data sekunder dari rekam medis elektronik (EMR) Puskesmas Pekalangan, Kota Cirebon, periode Januari hingga Juni 2024. Data mencakup atribut seperti nama pasien, nomor eRM, NIK, RT/RW, kelurahan, umur, jenis kunjungan, poli/ruangan, jenis asuransi, dokter/tenaga medis, keluhan utama, suhu tubuh, dan diagnosis. Pengumpulan data dilakukan melalui observasi terhadap data historis pasien, baik manual maupun terkomputerisasi, dengan format seperti CSV atau spreadsheet, sambil menjaga privasi pasien sesuai regulasi. Validasi dilakukan untuk memastikan keandalan dan akurasi data, yang kemudian digunakan untuk analisis model prediktif dengan algoritma Random Forest. Proses ini dilakukan secara sistematis dengan izin resmi dan koordinasi pihak Puskesmas guna menjamin integritas data yang relevan untuk penelitian.

3.2. Teknik Analisis Data

Penelitian ini menggunakan teknik Knowledge Discovery in Databases (KDD) untuk menganalisis data diagnosis pasien di Puskesmas Pekalangan, Kota Cirebon dengan tujuan mengevaluasi akurasi model prediktif berdasarkan pola antar atribut data. Algoritma Random Forest diterapkan untuk memprediksi diagnosis menggunakan atribut seperti nama pasien, no.eRM, NIK, RT/RW, kelurahan, umur, jenis kunjungan, poli/ruangan, asuransi, dokter/tenaga medis, keluhan utama, suhu, dan diagnosa. Analisis dilakukan menggunakan RapidMiner , mencakup seleksi data, pra-pemrosesan, transformasi data, pembangunan model, dan evaluasi akurasi. Penelitian ini bertujuan memahami pola diagnosis pasien, memberikan rekomendasi atribut paling berpengaruh, dan menjelaskan batasan generalisasi hasil. Hasil diharapkan dapat meningkatkan akurasi diagnosis dan mendukung pengambilan keputusan medis di Puskesmas Pekalangan. Proses KDD dirangkum seperti pada gambar 1.



Gambar 1. Alur tahapan Knowledge Discovery in Databases (KDD)
Sumber : Jurnal Swabumi [15].

4. HASIL DAN PEMBAHASAN

Penelitian ini menggunakan metode Knowledge Discovery in Databases (KDD) pada dataset Electronic Medical Records (EMR) dari Puskesmas Pekalangan, Kota Cirebon periode Januari–Juni 2024, untuk menganalisis tingkat akurasi model Random Forest, mengidentifikasi pola data, dan mengevaluasi kontribusi atribut rekam medis terhadap hasil prediksi diagnosis pasien.

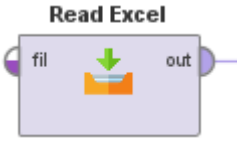
4.1. Data understanding

Data Electronic Medical Records atau EMR diperoleh dari database Puskesmas Pekalangan, Kota

Cirebon, yang mencakup informasi rekam medis pasien selama periode Januari hingga Juni 2024. Data yang dikumpulkan merupakan data sekunder yang relevan dengan penelitian.

4.2. Seleksi Data

Penelitian ini menggunakan dataset "Rekam Medis Elektronik" berisi 3.769 baris data dengan 16 atribut terkait diagnosis pasien, diekspor dari database Puskesmas Pekalangan, Kota Cirebon untuk periode Januari–Juni 2024 dan diimpor ke RapidMiner melalui file Dataset Rekamedis.xlsx seperti pada Gambar 2.



Gambar 2. Operator read excel

Gambar 2. menunjukkan representasi visual dari operator Read Excel, yang digunakan untuk membaca data dari file Excel. Tabel 1. memberikan rincian mengenai parameter-parameter yang digunakan dalam operator tersebut.

Tabel 1. Isi operator read excel

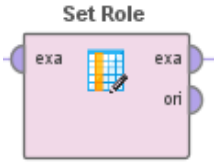
Parameter	**Isi**	
Import Configuration wizard		
Read Exel	exel file	DatasetRekamedis.xlsx
	sheet number	1
	date format	

Tabel 1. menampilkan data dari hasil penggunaan Operator Read excel, yang telah di dapatkan dari hasil observasi di Puskesmas Pekalangan Kota Cirebon.

Tabel 2. Hasil operator read exel

No	**Nama Pasien**	**No.eRM**	**…**	**Diagnosa 1**
1	NOVA SELASARI	986920	…	Acute amoebic dysentery
2	PRIYA PRAYOGA	584010	…	Acute amoebic dysentery
3	VONI LEGISTA	579036	…	Acute amoebic dysentery
4	SOFIA DEVIANI	1036350	…	Acute amoebic dysentery
…	…	…	…	…
3768	RASNATA	1091726	…	Other specified headache syndromes
3769	SITI SADIAH	586993	…	Other specified headache syndromes

Pada Tabel 2 setelah data diimpor, atribut Diagnosa 1 ditetapkan sebagai label untuk memprediksi diagnosis, sementara atribut NO sebagai ID sebagai pengenal unik. Langkah ini memastikan struktur dataset siap untuk analisis, seperti ditunjukkan pada Gambar 3.



Gambar 3. Operator set role

Tabel 3 menunjukkan bahwa atribut "Diagnosa 1" ditetapkan sebagai "Label" untuk variabel target prediksi, sementara atribut "No" berfungsi sebagai "ID" data. Pengaturan ini penting karena memengaruhi cara algoritma machine learning memproses data dan membangun model.

Tabel 3. Isi dari parameter operator set role

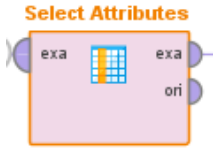
Parameter	**Isi**	
Set Role	Attributes name	Target role
	Diagnosa 1	Label
	No	Id

Berdasarkan hasil operator Set Role pada Tabel 3, atribut 'No' telah ditetapkan sebagai (ID) data, sedangkan atribut 'Diagnosa 1' telah ditetapkan sebagai label. Seperti ditampikan pada Tabel 4.

Tabel 4. Hasil data dari operator set role

No	**Diagnosa 1**	**Nama Pasien**	**No.eRm**	**…**	**Suhu**
1	Acute amoebic dysentery	NOVA SELASARI	986920	…	36.6 C
2	Acute amoebic dysentery	PRIYA PRAYOGA	584010	…	36 C
…	…	…	…	…	…
3768	Other specified headache syndromes	RASNATA	1091726	…	36.9 C
3769	Other specified headache syndromes	SITI SADIAH	586993	…	36.9 C

Selanjutnya, kita memilih atribut yang relevan untuk proses analisis menggunakan Operator Select Attribute. Pada dataset rekam medis semua atribut di pilih kecuali, atribut diagnosa 1 serta No, karena sudah di jadikan label dan Id. Tampak seperi Gambar 4.



Gambar 4. Select Attributes

Tabel 5 menjelaskan parameter Select Attributes dalam pemilihan atribut data rekam medis, dengan

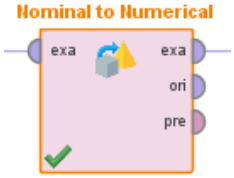
Type disetel ke "Include attributes" dan Attributes filter type ke "a subset". Atribut yang dipilih meliputi "Diagnosa 1", "Asuransi", "Jenis Kelamin", "Dokter/Tenaga Medis", dan lainnya.

Tabel 5. Isi parameter Select Attributes

Parameter	Isi
Type	Include attributes
Attributes filter type	a subset
Select subset	Select Atribut
Attributes	Diagnosa 1, Asuransi, No., Dokter/Tenaga Medis, Jenis Kelamin, Jenis Kunjungan, Keluhan Utama, Kelurahan, Nama Pasien, NIK, No. eRM, Poli/Ruangan, RT, RW, Suhu, Umur Tahun

4.3. Transformasi Data

Transformasi data menggunakan operator "Nominal to Numerical" mengubah format non-numerik menjadi numerik agar kompatibel dengan algoritma pembelajaran mesin, seperti terlihat pada Gambar 5.



Gambar 5. Operator Nominal to Numerical

Gambar 5 menunjukkan konfigurasi parameter operator "Nominal to Numerical" di RapidMiner, yang digunakan untuk mengubah atribut data tipe nominal menjadi numerik agar kompatibel dengan algoritma machine learning.

Tabel 6. Isi dari parameter Nominal to numerical

Parameter	Isi	
Nominal To Numerical	Aattribute filter type	All
	Invert sellection	False
	Include special attributes	FALSE
	Coding type	Unique integers

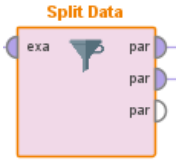
Tabel 6 adalah hasil konfigurasi dari operator Nominal to Numerical. Konfigurasi ini bertujuan untuk mengkonversi semua atribut nominal yang ada dalam dataset secara otomatis.

Tabel 7. Hasil dari data dari operator Nominal to numerical

No.	Diagnosa	Nama Pasien	…	RW
1	Acute amoebic dysentery	0	…	2
2	Acute amoebic dysentery	1	…	5
3	Acute amoebic dysentery	2	…	4
…	…	…	…	…
3768	Other specified headache syndromes	1681	…	5
3769	Other specified headache syndromes	1482	…	4

4.4. Proses Data Mining

Tahapan ini menghasilkan model klasifikasi diagnosis pasien dengan Random Forest, memisahkan dataset 80% untuk pelatihan dan 20% untuk pengujian, seperti ditunjukkan pada Gambar 6.



Gambar 6. Operator Split Data

Untuk membagi dataset, RapidMiner menyediakan operator Split Data. Detail konfigurasi parameter yang dapat diatur pada operator ini, termasuk cara pembagian dan tipe sampling, disajikan dalam Tabel 8.

Tabel 8. Isi dari parameter Split data

Parameter	Isi	
Split data	Partitions	Edit Enumber
		ratio
		80%
		20%
	sampling type	automatic

Konfigurasi parameter untuk operator "Split Data" dapat dilihat pada Tabel 8. Hasil pembagian data dengan rasio 80% untuk pelatihan dan 20% untuk pengujian menghasilkan 754 data rekord untuk pengujian model, yang ditampilkan pada Tabel 9.

Tabel 9. Hasil Operator split data 80% melatih model

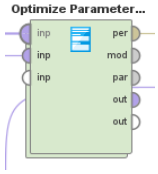
No.	Diagnosa	Nama Pasien	…	RW
1	Acute amoebic dysentery	0	…	2
2	Acute amoebic dysentery	1	…	5
3	Acute amoebic dysentery	2	…	4
…	…	…	…	…
3768	Other specified headache syndromes	1681	…	5
3769	Other specified headache syndromes	1482	…	4

Tabel 9. menunjukkan data 80% untuk pelatihan model, sementara Tabel 10 menunjukkan data 20% untuk pengujian model.

Tabel 10. Hasil Operator split data 20% menguji model

No.	Diagnosa	Nama Pasien	Jenis Kelamin		RW
8	Acute amoebic dysentery	7	1	…	7
10	Acute amoebic dysentery	9	1	…	7
11	Acute amoebic dysentery	10	1	…	8
…	…	…	…	…	…
3762	Other specified headache syndromes	1679	0	…	4
3764	Other specified headache syndromes	1680	1	…	1

Langkah terakhir adalah optimasi parameter menggunakan operator "Optimize Parameters (Grid)" dengan metode grid search. Proses ini mencari kombinasi parameter terbaik, termasuk jumlah pohon (number of trees) dari 10 hingga 100 dengan langkah 10, untuk menghasilkan model optimal Gambar 7.



Gambar 7. Operator Optimize Parameters (Grid)

Gambar 7 menunjukkan tampilan operator Optimize Parameters (Grid), sedangkan Tabel 11 merinci parameter-parameter yang dapat dioptimasi beserta konfigurasinya.

Tabel 11. Isi Operator Optimize Parameters (Grid)

Parameter	Isi
Optimize Parameters	Edit Parameter Settings
Operators & Parameters	Edit settings, Random Forest (Criterion: Number of Trees), Apply Model (Pruning: Max Depth), Performance (Confidence, Prepruning Options)
Grid/Range	Min: 10, Max: 100, Steps: 10, Scale: Linear
Error Handling	Fail on Error
Log Performance	Log Criteria

Setelah menentukan parameter pada Tabel 11. eksperimen dilakukan untuk mencari akurasi terbaik pada jumlah pohon (number of trees) yang ditemukan pada angka 28 dengan akurasi 76,39%, seperti yang ditunjukkan pada Tabel 12.

Tabel 12. Hasil ekperimen number of tree

number of tree	maksimal depth	Akurasi
10.0	1.0	14.72%
19.0	1.0	14.72%
28.0	1.0	14.72%
…	…	…
19.0	10.0	71.62%
28.0	10.0	76.39%
37.0	10.0	69.89%
…	…	…
82.0	10.0	72.81%
91.0	10.0	73.21%
100.0	10.0	72.15%

Berikutnya, parameter maximal depth dievaluasi dalam rentang 1 hingga 10 untuk menemukan akurasi terbaik, seperti yang ditunjukkan pada Tabel 13.

Tabel 13. Hasil Random Forest Maximal depth

Parameter	Isi
Optimize Parameters	Edit Parameter Settings
Operators & Parameters	Random Forest (Criterion: Number of Trees), Apply Model (Pruning: Max Depth), Performance (Confidence, Prepruning Options: Gain, Leaf Size, Split Size, Alternatives)
Grid/Range	Min: 1, Max: 10, Steps: 1, Scale: Linear
Error Handling	Fail on Error
Log Performance	Log Criteria

Setelah menentukan parameter optimasi Tabel 13, eksperimen menunjukkan maximal depth terbaik adalah 10 dengan akurasi 76,39%, seperti disajikan pada Tabel 14.

Tabel 14. Hasil ekperimen maksimal depth

number of tree	maksimal depth	Akurasi
10.0	1.0	14.72%
19.0	1.0	14.72%
28.0	1.0	14.72%
…	…	…
19.0	10.0	71.62%
28.0	10.0	76.39%
37.0	10.0	69.89%
…	…	…
82.0	10.0	72.81%
91.0	10.0	73.21%
100.0	10.0	72.15%

Setelah menemukan parameter optimal, analisis dilakukan dengan menghapus atribut satu per satu untuk mengidentifikasi yang paling berpengaruh terhadap akurasi. Hasilnya ditunjukkan pada Tabel 15.

Tabel 15. hasil ekperimen atribut yang berpengaruh

Atribut	Nilai akurasi
Dokter / Tenaga Medis	78,38%
Jenis Kelamin	78,91%
Jenis Kunjungan	79,44%

Atribut	Nilai akurasi
Keluhan Utama	53,58%
Kelurahan	79,58%
NIK	76,26%
Asuransi	78,91%
No. eRM	75,99%
Poli/Ruangan	77,72%
RT	75,86%
RW	75,73%
Suhu	77,98%
Umur Tahun	77,98%
Nama Pasien	71,88%

Tabel 15 menyajikan hasil eksperimen tentang pengaruh atribut terhadap akurasi model, yang divisualisasikan dalam diagram batang pada Gambar 8. Gambar tersebut, dapat disimpulkan bahwa atribut "keluhan utama" memiliki pengaruh besar terhadap akurasi, terbukti dengan nilai akurasi terendah yang mencapai 53,58%.

Gambar 8. Diagram atribut berpengaruh

Pada gambar 8 dapat disimpulkan bahwa Atribut yang paling berpengaruh adalah “keluhan utama” Hal ini terlihat dari rendahnya nilai Akurasi yang paling rendah, yaitu mencapai 53,58%.

4.5. Evaluasi

Pada tahap evaluasi, analisis menunjukkan bahwa atribut "Keluhan Utama" memiliki pengaruh signifikan terhadap akurasi model, dengan penghapusan atribut ini menurunkan akurasi menjadi 53,58%. Atribut lainnya, seperti Jenis Kunjungan dan Kelurahan, juga berkontribusi, tetapi tidak sebesar Keluhan Utama. Evaluasi ini menegaskan pentingnya memilih atribut relevan untuk model prediksi yang optimal.

4.6. Pembahasan

Pada tahap ini membahas hasil penelitian terkait optimasi parameter algoritma Random Forest dalam memprediksi diagnosis pasien, dengan fokus pada pengaruh parameter number of trees, maximum depth, dan atribut yang berpengaruh terhadap performa model. Optimasi parameter ini penting karena mempengaruhi performa algoritma [16].

Optimasi jumlah pohon (number of trees) pada Random Forest mempengaruhi akurasi dan stabilitas model. Eksperimen grid search menunjukkan jumlah pohon optimal adalah 28, dengan akurasi 76,39%. Penelitian sebelumnya menunjukkan bahwa jumlah pohon yang terlalu tinggi tidak selalu meningkatkan akurasi dan dapat meningkatkan kompleksitas komputasi [18], [19].

Kedalaman maksimum optimal pada eksperimen ini adalah 10, dengan akurasi 76,39%. Kedalaman ini cukup menangkap kompleksitas data tanpa menyebabkan overfitting. Penelitian sebelumnya juga menunjukkan bahwa pengaturan kedalaman yang tepat dapat meningkatkan akurasi, terutama pada data tidak seimbang [17].

Analisis menunjukkan bahwa atribut "Keluhan Utama" memiliki kontribusi terbesar terhadap akurasi model, dengan penghapusannya menurunkan akurasi menjadi 53,58%. Temuan ini sejalan dengan penelitian [1] yang menyatakan bahwa atribut klinis yang relevan, seperti keluhan utama, memiliki pengaruh besar terhadap prediksi diagnosis.

5. KESIMPULAN DAN SARAN

Penelitian ini bertujuan meningkatkan akurasi prediksi diagnosis pasien dengan optimasi parameter algoritma Random Forest. Hasil menunjukkan bahwa jumlah pohon optimal adalah 28 dengan akurasi 76,39%, sedangkan kedalaman maksimum optimal adalah 9 dengan akurasi yang sama. Atribut "Keluhan Utama" berpengaruh besar, dengan penghapusan atribut ini menurunkan akurasi menjadi 53,58%. Disarankan untuk menggunakan dataset lebih besar, menerapkan metode optimasi parameter yang lebih kompleks seperti Bayesian Optimization, serta menambah atribut seperti hasil laboratorium dan riwayat penyakit untuk meningkatkan akurasi. Penelitian ini juga menekankan pentingnya menjaga kerahasiaan data pasien dan dapat menjadi dasar pengembangan sistem pendukung keputusan di layanan kesehatan.

DAFTAR PUSTAKA

[1] E. R. B. Sebayang, Y. H. Chrisnanto, and ... (2023), "Klasifikasi Data Kesehatan Mental di Industri Teknologi Menggunakan Algoritma Random Forest," International Journal of .... [Online]. Available: http://ijespgjournal.org/index.php/ijespg/article/view/57

[2] R. Oktafiani, A. Hermawan, and D. Avianto, "Max Depth Impact on Heart Disease Classification: Decision Tree and Random Forest," Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), vol. 8, no. 1, pp. 160–168, 2024. [Online]. Available: https://doi.org/10.29207/resti.v8i1.5574

[3] D. Ismafillah, T. Rohana, and Y. Cahyana, "Analisis algoritma pohon keputusan untuk memprediksi penyakit diabetes menggunakan oversampling smote," INFOTECH: Jurnal Informatika & Teknologi, vol. 4, no. 1, pp. 27–

36, 2023. [Online]. Available: https://doi.org/10.37373/infotech.v4i1.452

[4] A. U. Zailani and N. L. Hanun, "Penerapan Algoritma Klasifikasi Random Forest Untuk Penentuan Kelayakan Pemberian Kredit Di Koperasi Mitra Sejahtera," Infotech: Journal of Technology Information, vol. 6, no. 1, pp. 7–14, 2020. [Online]. Available: https://doi.org/10.37365/jti.v6i1.61

[5] F. Azzahra, N. Suarna, and Y. Arie Wijaya, "Penerapan Algoritma Random Forest Dan Cross Validation Untuk Prediksi Data Stunting," Kopertip: Jurnal Ilmiah Manajemen Informatika dan Komputer, vol. 8, no. 1, pp. 1–6, 2024. [Online]. Available: https://doi.org/10.32485/kopertip.v8i1.238

[6] E. Saputro and D. Rosiyadi, "Penerapan Metode Random Over-Under Sampling Pada Algoritma Klasifikasi Penentuan Penyakit Diabetes," Bianglala Informatika, vol. 10, no. 1, pp. 42–47, 2022. [Online]. Available: https://doi.org/10.31294/bi.v10i1.11739

[7] G. A. B. Suryanegara, A. Adiwijaya, and M. D. Purbolaksono, "Peningkatan Hasil Klasifikasi pada Algoritma Random Forest untuk Deteksi," Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), vol. 1, no. 10, pp. 114–122, 2021.

[8] Y. Azhar, A. K. Firdausy, and P. J. Amelia, "Perbandingan Algoritma Klasifikasi Data Mining Untuk Prediksi Penyakit Stroke," SINTECH (Science and Information Technology) Journal, vol. 5, no. 2, pp. 191–197, 2022. [Online]. Available: https://doi.org/10.31598/sintechjournal.v5i2.1222

[9] U. Sunarya and T. Haryanti, "Perbandingan Kinerja Algoritma Optimasi pada Metode Random Forest untuk Deteksi Kegagalan Jantung," Jurnal Rekayasa Elektrika, vol. 18, no. 4, pp. 241–247, 2022. [Online]. Available: https://doi.org/10.17529/jre.v18i4.26981

[10] D. Derisma, "Perbandingan Kinerja Algoritma untuk Prediksi Penyakit Jantung dengan Teknik Data Mining," Journal of Applied Informatics and Computing, vol. 4, no. 1, pp. 84–88, 2020. [Online]. Available: https://doi.org/10.30871/jaic.v4i1.2152

[11] W. Apriliah, I. Kurniawan, M. Baydhowi, and T. Haryati, "Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest," Sistemasi, vol. 10, no. 1, pp. 163, 2021. [Online]. Available: https://doi.org/10.32520/stmsi.v10i1.1129

[12] M. F. R. Aditya, N. Lutvi, and U. Indahyanti, "Prediksi Penyakit Hipertensi Menggunakan Metode Decision Tree dan Random Forest," Jurnal Ilmiah Komputasi, vol. 23, no. 1, pp. 9–16, 2024. [Online]. Available: https://doi.org/10.32409/jikstik.23.1.3503

[13] A. A. R. Reza and M. Syaifur Rohman, "Prediction Stunting Analysis Using Random Forest Algorithm and Random Search Optimization," Journal of Informatics and Telecommunication Engineering, vol. 7, no. 2, pp. 534–544, 2024. [Online]. Available: https://doi.org/10.31289/jite.v7i2.10628

[14] R. Genuer and J.-M. Poggi, Random Forests with R, R. Genuer and J.-M. Poggi, Eds., Springer International Publishing, pp. 33–55, 2020. [Online]. Available: https://doi.org/10.1007/978-3-030-56485-8_3

[15] E. Mutiara, "Algoritma Klasifikasi Naive Bayes Berbasis Particle Swarm Optimization Untuk Prediksi Penyakit Tuberculosis (Tb)," Swabumi, vol. 8, no. 1, pp. 46–58, 2020. [Online]. Available: https://doi.org/10.31294/swabumi.v8i1.7668

[16] S. Andrian, E. S. Salim, H. Bindan, E. Pranoto, and ..., "Analisa Metode Random Forest Tree dan K-Nearest Neighbor dalam Mendeteksi Kanker Serviks," Jurnal Ilmu Komputer. [Online]. Available: https://www.academia.edu/download/110868411/73-Article_Text-145-1-10-20201010.pdf

[17] E. Erlin, Y. Desnelita, N. Nasution, L. Suryati, and F. Zoromi, "Dampak SMOTE terhadap Kinerja Random Forest Classifier berdasarkan Data Tidak seimbang," MATRIK: Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer, vol. 21, no. 3, pp. 677–690, 2022. [Online]. Available: https://doi.org/10.30812/matrik.v21i3.1726

[18] A. Fauzi, R. Supriyadi, and N. Maulidah, "Deteksi Penyakit Kanker Payudara dengan Seleksi Fitur berbasis Principal Component Analysis dan Random Forest," Jurnal Infortech, vol. 2, no. 1, pp. 96–101, 2020. [Online]. Available: https://doi.org/10.31294/infortech.v2i1.8079

[19] D. Septiani, U. Enri, and N. Sulistiyowati, "Diagnosa Tingkat Depresi Mahasiswa Selama Masa Pandemi Covid-19 Menggunakan Algoritma Random Forest," STRING (Satuan Tulisan Riset Dan Inovasi Teknologi), vol. 6, no. 2, pp. 149, 2021. [Online]. Available: https://doi.org/10.30998/string.v6i2.10361