

OPTIMALISASI FITUR DENGAN FORWARD SELECTION PADA ESTIMASI TINGKAT PENYAKIT PARU-PARU MENGGUNAKAN ALGORITMA KLASIFIKASI RANDOM FOREST

Lukman Rangga Aditya Tarigan, Dahlan

Teknik Informatika, Politeknik TEDC Bandung

Jl. Politeknik-Pesantren KM2 Cimahi Utara – Cimahi – Jawa Barat - Indonesia

lukmanradityaa@gmail.com

ABSTRAK

Penyakit paru-paru merupakan masalah kesehatan serius di Indonesia, dengan peningkatan kasus dan kematian yang signifikan. Penelitian ini bertujuan untuk meningkatkan akurasi klasifikasi penyakit paru-paru menggunakan algoritma *Random Forest* dengan metode optimasi fitur *Forward Selection*. Dataset yang digunakan terdiri dari 1.000 sampel dengan 10 atribut. Penelitian dilakukan dalam dua tahap: pertama, pengujian awal algoritma *Random Forest* tanpa optimasi fitur, dan kedua, pengujian dengan optimasi fitur menggunakan metode *Forward Selection*. Hasil pengujian awal menunjukkan akurasi sebesar 89,45%, Presisi 83,20%, Recall 100,00%. Setelah menerapkan optimasi fitur, akurasi meningkat menjadi 92,46%, presisi 95,88%, recall 89,42%. Peningkatan akurasi sebesar 3,01% ini menegaskan pentingnya optimasi fitur dalam meningkatkan performa model klasifikasi. Analisis atribut penting mengidentifikasi bahwa aktivitas olahraga, kebiasaan merokok, dan usia merupakan faktor yang paling signifikan dalam memprediksi penyakit paru-paru. Penelitian ini menyoroti bahwa penggunaan metode optimasi fitur seperti *Forward Selection* dapat secara signifikan meningkatkan akurasi prediksi model klasifikasi penyakit paru-paru.

Kata kunci : Penyakit Paru-Paru, Random Forest, Optimasi Fitur, Forward Selection, Klasifikasi.

1. PENDAHULUAN

Kesehatan adalah aset berharga yang tidak bisa dibeli oleh siapa pun, menjadikannya hal yang sangat penting dan tak ternilai bagi setiap orang. Salah satu organ vital dalam tubuh manusia yang sangat mempengaruhi kesehatan adalah paru-paru. Paru-paru memiliki peran penting dalam sistem pernapasan manusia [1].

Penyakit paru-paru dapat mengakibatkan kerusakan pada sistem pernapasan baik dalam jangka pendek maupun jangka panjang. Zat kimia beracun seperti yang terdapat dalam rokok bisa menjadi penyebab penyakit paru-paru. Berdasarkan informasi dari *Lung Foundation Australia*, penyakit paru-paru dapat menyerang pria, wanita, anak-anak, perokok, mantan perokok, dan bahkan orang yang tidak merokok [2].

Menurut Global Cancer Statistics jumlah kasus kanker paru-paru di Indonesia bertambah 34.783 dengan angka kematian 30.843 [3]. Berdasarkan data dari TBC Pada tahun 2020, angka kematian akibat tuberkulosis di Indonesia meningkat menjadi 98.000 dengan kasus baru mencapai 845.000. Berdasarkan temuan tersebut, penyakit paru-paru di Indonesia merupakan masalah serius yang memerlukan pengobatan yang memadai dan tepat waktu [3].

Data Mining adalah serangkaian proses yang melibatkan eksplorasi atau penambangan data untuk mengungkap informasi yang sebelumnya tersembunyi atau tidak diketahui [4].

Data mining merupakan disiplin ilmu yang menggabungkan teknik pembelajaran mesin, pengenalan pola, statistik, basis data, dan visualisasi

untuk menangani masalah penggalian informasi dari basis data yang sangat besar [5].

Meskipun teknik *Data Mining* sering kali berfokus pada kumpulan data yang besar, beberapa algoritma juga dapat diterapkan pada kumpulan data yang lebih kecil. *Data Mining* menggunakan kumpulan data sederhana yang terdiri dari baris yang mewakili kejadian tertentu (juga dikenal sebagai observasi atau contoh) dan kolom yang menentukan atribut atau variabel kasus. Algoritma yang digunakan bergantung pada jenis data (*nominal*, *ordinal*, *rasio*, atau *interval*). Proses *Data Mining* melibatkan pemilihan data dari sumber tertentu, pra-pemrosesan untuk memastikan kualitasnya, transformasi data, penambangan, serta interpretasi dan evaluasi hasil. Hasil akhir dari proses ini adalah data baru yang dapat digunakan untuk memperbaiki artikel [6].

Data mining juga dapat diartikan sebagai proses mengungkap informasi dari sejumlah data sehingga informasi yang dibutuhkan lebih mudah ditemukan dengan menggunakan pola yang dibangun berdasarkan kedekatan titik-titik data. Pendekatan atau algoritma yang digunakan dalam *Data Mining* haruslah tepat karena berkaitan langsung dengan tujuan dan keseluruhan proses *Knowledge Discovery in Database* (KDD) [7].

Random Forest adalah metode klasifikasi yang terdiri dari sekumpulan pohon keputusan yang dibangun berdasarkan data pelatihan, dengan pemilihan fitur secara acak selama pembuatan setiap pohon. Setelah sejumlah pohon keputusan dibuat, setiap pohon memberikan suaranya untuk menentukan kelas yang paling populer. *Random Forest* memiliki

beberapa kelebihan, termasuk kemampuannya meningkatkan akurasi meskipun ada data yang hilang, tahan terhadap outliers, dan efisien dalam penyimpanan data. Selain itu, Random Forest memiliki proses seleksi fitur yang dapat memilih fitur terbaik, sehingga dapat meningkatkan performa model klasifikasi [8] [9].

Proses pengambilan suara kolektif dari sejumlah pohon keputusan dikenal sebagai Random Forest (RF). Algoritma klasifikasi *Random Forest* (RF) memiliki dua parameter utama: jumlah pohon dan jumlah fitur yang digunakan. Banyaknya pohon yang ditanam membentuk sebuah hutan yang kemudian dianalisis. Prediksi respons dilakukan dengan menggabungkan prediksi dari setiap pohon, dan masalah klasifikasi diselesaikan dengan memilih suara terbanyak. Algoritma *Random Forest* (RF) sangat terkait dengan pohon keputusan, karena algoritma ini merupakan gabungan suara mayoritas dari banyak pohon keputusan. Pohon keputusan adalah teknik pengumpulan kondisional yang mencari nilai pemisahan yang optimal berdasarkan nilai entropi (ukuran informasi dalam suatu keadaan) untuk memaksimalkan perolehan informasi [10]. *Random Forest* (RF) adalah pengklasifikasi ansambel yang membangun hutan dari sejumlah pohon keputusan. Kelas data masukan ditentukan oleh suara terbanyak (*voting*) dari pohon keputusan. Metode ini mengatasi masalah pengklasifikasian yang sering muncul ketika hanya menggunakan satu pohon keputusan, yang mungkin tidak ideal. Dengan menggunakan beberapa pohon keputusan, *Random Forest* dapat mencapai tingkat akurasi yang optimal secara keseluruhan [11].

2. TINJAUAN PUSTAKA

Tinjauan pustaka mencakup definisi, konsep, dan pengertian dari berbagai topik berikut: Paru-Paru, Optimasi Fitur, Klasifikasi, *Random Forest*, *Forward Selection*, dan *Confusion Matrix*.

2.1. Paru-Paru

Paru-paru adalah organ penting untuk pernapasan, di mana respirasi mengubah bahan organik menjadi CO₂, H₂O, dan energi menggunakan O₂. Penyakit paru-paru umum terjadi karena instalasi udara terkontaminasi oleh debu, asap, virus, atau bakteri, dan dapat menyerang sejak bayi hingga dewasa dengan pengobatan yang kompleks. Di Indonesia, penyakit paru-paru meliputi PPOK, yang menurut WHO akan menjadi penyebab kematian ketiga pada 2019-2030, serta bronkitis, asma, kanker paru-paru, pneumonia, pneumotoraks, emfisema, efusi pleura, dan TB. Gejala dan penanganan tiap penyakit berbeda, sehingga penting bagi penderita memahami kondisi mereka. Penyakit ini dapat menyebabkan hipoksia dan kesulitan bernapas, serta berpotensi fatal [12].

2.2. Optimasi Fitur

Optimasi fitur bertujuan mengurangi atribut data yang tidak relevan untuk mempercepat algoritma dan meningkatkan akurasi model klasifikasi. Dengan mengurangi atribut ini, algoritma menjadi lebih efisien dan model lebih akurat. Selain itu, optimasi fitur membantu mencegah *overfitting*, di mana model terlalu kompleks dan tidak dapat digeneralisasi untuk data baru. Optimasi ini juga menyederhanakan model, memudahkan interpretasi, dan mendukung pengambilan keputusan yang lebih baik [13].

2.3. Klasifikasi

Klasifikasi adalah teknik *machine learning* untuk memprediksi kelas atau label data berdasarkan fitur-fiturnya. Tujuannya adalah membangun model yang dapat mempelajari pola dari data latih dan memprediksi kelas dari data uji yang baru. Metode klasifikasi meliputi *Naive Bayes*, *Decision Tree*, *Random Forest*, *Support Vector Machine*, dan *Deep Neural Network*, masing-masing dengan kelebihan dan kekurangannya. Pemilihan metode yang tepat sangat penting untuk mencapai akurasi tinggi. Klasifikasi digunakan di berbagai bidang seperti pengenalan wajah, deteksi spam, dan klasifikasi citra [14].

2.4. Random Forest

Random Forest adalah algoritma *Machine Learning* untuk klasifikasi dan regresi yang menggabungkan metode *Bagging* dan *Random Subspaces*. Algoritma ini membangun banyak *Decision Tree* secara acak dan menggabungkan prediksi setiap pohon untuk hasil akhir yang lebih akurat dan stabil. Salah satu keunggulan *Random Forest* adalah kemampuannya mengatasi *overfitting*, sehingga model lebih generalis dan efektif untuk data baru. Algoritma ini sukses digunakan di berbagai bidang seperti kesehatan, keuangan, dan industri [15].

Random Forest adalah algoritma *Machine Learning* populer yang menggabungkan beberapa *Decision Tree* untuk prediksi akhir. Setiap pohon dibangun menggunakan subset acak dari fitur dan data latih, membuat model lebih robust dan mampu mengatasi *overfitting*. Keunggulan *Random Forest* adalah kemampuannya dalam pemilihan fitur acak dan pemilihan simpul, yang menghasilkan tingkat kesalahan rendah [16].

2.5. Forward Selection

Forward Selection adalah metode fitur seleksi dalam *machine learning* untuk memilih subset fitur yang paling relevan dalam dataset. Proses ini dimulai dengan satu fitur dan secara iteratif menambahkan fitur yang paling meningkatkan kinerja model pada setiap langkah. Proses ini berlanjut hingga tidak ada peningkatan kinerja signifikan atau hingga mencapai jumlah fitur yang diinginkan. *Forward Selection* menggunakan F-statistik untuk memilih variabel

dengan peningkatan kinerja tertinggi secara berurutan [17].

2.6. Confusion Matrix

Confusion Matrix adalah metode untuk mengukur kinerja sistem klasifikasi dengan membandingkan hasil klasifikasi sistem dengan hasil klasifikasi yang sebenarnya. *Confusion Matrix* terdiri dari empat komponen: *True Positive* (TP), *False Positive* (FP), *True Negative* (TN), dan *False Negative* (FN) [18]. *Confusion Matrix* memungkinkan perhitungan berbagai metrik evaluasi seperti akurasi, presisi, *recall*, dan *F1-score*, yang membantu dalam menilai kinerja sistem klasifikasi. Berikut 4 bagian *Confusion Matrix* :

a. True Positives (TP)

Merupakan sejumlah besar data nilai sebenarnya adalah positif dan kelas prediksi sebagai nilai positif;

b. False Positives (FP)

Merupakan sejumlah besar data nilai sebenarnya adalah negatif dan kelas prediksi sebagai nilai positif;

c. False negatives (FN)

Merupakan sejumlah besar data nilai sebenarnya adalah positif dan kelas prediksi sebagai nilai negatif;

d. True negatives (TN)

Merupakan kumpulan data negatif dengan klasifikasi sebagai nilai negatif.

Tabel 1. *Confusion Matrix*

		Nilai Sebenarnya	
		TRUE	FALSE
Nilai Prediksi	TRUE	TP	FP
	FALSE	FN	TN

Pada tabel 1 *Confusion Matrix* umumnya ditampilkan 3 metode *Testing* yaitu:

a. Accuracy

Accuracy atau akurasi merupakan salah satu metode *Testing* suatu algoritma berdasarkan tingkat kedekatan antara nilai prediksi dengan nilai sebenarnya. Rumus menghitung akurasi yaitu:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100$$

b. Precision

Precision atau presisi merupakan salah satu metode *Testing* suatu algoritma dengan melakukan perbandingan data benar yang diperoleh sistem dengan jumlah seluruh data yang diambil oleh sistem yang benar maupun salah. Rumus menghitung presisi yaitu:

$$Precision = \frac{TP}{TP + FP} \times 100$$

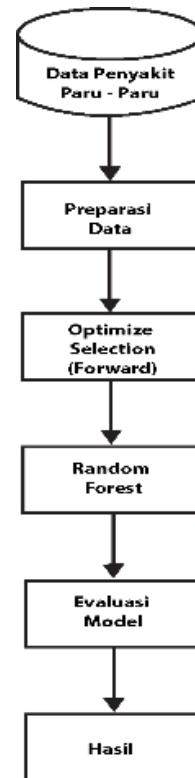
c. Recall

Recall adalah merupakan salah satu metode *Testing* suatu algoritma dengan melakukan perbandingan jumlah data support benar yang diambil atau tidak diambil oleh sistem. Rumus menghitung *recall* yaitu:

$$Recall = \frac{TP}{TP + FN} \times 100$$

3. METODE PENELITIAN

Dalam penelitian ini, metode *Random Forest* diterapkan untuk mengklasifikasikan tingkat kesehatan paru-paru pada remaja, sementara metode *Optimize Forward Selection* digunakan untuk memilih karakteristik yang memberikan akurasi prediksi terbaik. Proses penelitian ini dilakukan secara bertahap, sebagaimana yang ditunjukkan pada Gambar 1.



Gambar 1. Metode Yang Digunakan

3.1. Pengumpulan Data

Tahap pertama penelitian ini mengumpulkan data, peneliti memanfaatkan *dataset* public yang tersedia di *kaggle.com*. *dataset* tersebut berisi data prediksi terkena penyakit paru-paru, *Dataset* ini terdiri dari 10 Atribut dan salah satunya adalah atribut class, data terdiri dari 30.001 data. Data yang penulis gunakan hanya 1000 data. Berikut *dataset* penyakit paru-paru pada gambar 2.

No	Usia	Jenis Kelamin	Merokok	Bekerja	Rumah tangga	Aktivitas Begadang	Aktivitas Olahraga	Asuransi	Penyakit Bawaan	Hasil
1.	Tua	Pria	Pasif	Tidak	Ya	Ya	Sering	Ada	Tidak	Ya
2.	Tua	Pria	Aktif	Tidak	Ya	Ya	Jarang	Ada	Ada	Tidak
3.	Muda	Pria	Aktif	Tidak	Ya	Ya	Jarang	Ada	Tidak	Tidak
.	.	.	.	.	.	.	.	.	.	.
1000	Tua	Wanita	Pasif	Tidak	Ya	Tidak	Sering	Tidak	Tidak	Ya

Gambar2 . Data Penyakit Paru-Paru

3.2. Preparasi Data

Pada gambar 3, salah satu atribut yang digunakan untuk mendiagnosa ada atau tidaknya penyakit paru-paru pada seseorang adalah atribut hasil. Setelah itu,

data tersebut dipreparasi secara manual, seperti pengambilan atribut yang memiliki pengaruh paing besar terhadap penyakit paru-paru seperti kebiasaan merokok dan aktivitas olahraga memiliki pengaruh langsung terhadap kesehatan paru-paru. Jadi atribut data yang penulis hilangkan seperti atribut asuransi dan rumah tangga. Dapat dilihat pada gambar 3 hasil preparasi data.

No	Usia	Jenis Kelamin	Merokok	Bekerja	Aktivitas Begadang	Aktivitas Olahraga	Penyakit Bawaan	Hasil
1.	Tua	Pria	Pasif	Tidak	Ya	Sering	Tidak	Ya
2.	Tua	Pria	Aktif	Tidak	Ya	Jarang	Ada	Tidak
3.	Muda	Pria	Aktif	Tidak	Ya	Jarang	Tidak	Tidak
...								
1000	Tua	Wanita	Pasif	Tidak	Tidak	Sering	Tidak	Ya

Gambar 3. Hasil Preparasi Data

3.3. Optimize Selection (Forward)

Optimize Selection (Forward) merupakan langkah penting dalam proses pemodelan untuk meningkatkan kinerja model. Pada penelitian ini, teknik *Forward Selection* digunakan sebagai pendekatan iteratif yang bertujuan untuk menambahkan fitur yang memberikan peningkatan terbesar pada kinerja model. Proses ini dimulai dengan subset fitur kosong, di mana tidak ada fitur yang dipilih. Selanjutnya, fitur ditambahkan satu per satu berdasarkan kontribusinya terhadap performa model, yang diukur melalui metrik evaluasi seperti akurasi.

Setiap kali sebuah fitur baru ditambahkan, kinerja model dievaluasi menggunakan validasi silang untuk memastikan model tidak overfitting terhadap data pelatihan. Proses ini berlanjut sampai penambahan fitur baru tidak lagi memberikan peningkatan signifikan terhadap kinerja model. Pada tahap akhir, subset fitur yang memberikan kinerja terbaik dipilih untuk digunakan dalam pelatihan model selanjutnya, yaitu *Random Forest*.

3.4. Random Forest

Setelah subset fitur terbaik teridentifikasi melalui proses *Optimize Selection* (Forward), langkah berikutnya adalah membangun model prediksi menggunakan algoritma *Random Forest*. *Random Forest* adalah metode pembelajaran ansambel yang terdiri dari banyak pohon keputusan (*decision trees*) yang bekerja bersama untuk menghasilkan prediksi yang lebih akurat dan stabil.

3.5. Proses Dan Hasil Penelitian

Algoritma *Random Forest* diterapkan pada dataset untuk mengklasifikasikan penyakit paru-paru. Model ini dilatih menggunakan bagian dari dataset dan diuji dengan bagian lainnya untuk mengukur performa awalnya. Performa awal model diukur menggunakan metrik akurasi, untuk memahami sejauh mana model dapat mengklasifikasikan penyakit paru-paru dengan benar. Untuk meningkatkan performa model, dilakukan optimasi fitur seleksi menggunakan metode *Forward Selection*. Proses ini melibatkan penambahan fitur secara iteratif dan evaluasi performa model pada

setiap langkah untuk menemukan kombinasi fitur terbaik yang meningkatkan akurasi.

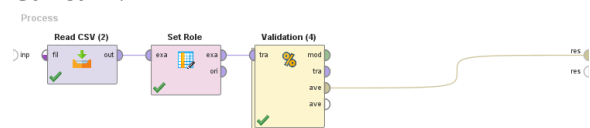
4. HASIL DAN PEMBAHASAN

Pada penelitian ini, dilakukan dua tahap pengujian untuk mengukur performa algoritma *Random Forest* dalam mengklasifikasikan penyakit paru-paru. Pengujian pertama dilakukan tanpa menggunakan teknik optimasi fitur, sementara pengujian kedua dilakukan menggunakan optimasi fitur dengan metode *Forward Selection*. Tujuan dari pengujian kedua ini adalah untuk mengevaluasi apakah optimasi fitur dapat meningkatkan akurasi dan performa keseluruhan model *Random Forest*.

Pada tahap ini, dilakukan penerapan algoritma Genetika untuk optimasi pemilihan fitur dengan metode *Forward Selection*. Algoritma ini berfungsi untuk memilih fitur-fitur yang paling relevan dalam memprediksi penyakit paru-paru. Setelah fitur-fitur optimal terpilih, model *Random Forest* kembali dilatih dan diuji untuk mengukur peningkatan performanya.

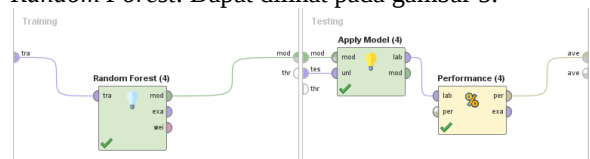
4.1. Proses Pengujian Algoritma Random Forest Tanpa Optimasi Fitur

Pengujian akan dilakukan menggunakan aplikasi *RapidMiner*. Dalam hal ini, kita akan memanfaatkan operator di *RapidMiner*, seperti yang ditunjukkan pada Gambar 4.



Gambar 4. Proses Pengujian Algoritma Random Forest Tanpa Optimasi Fitur

Pada tahap awal, data penyakit paru-paru diimpor ke *RapidMiner* menggunakan operator *Read CSV*. Setelah data dibaca, operator *Set Role* digunakan untuk menentukan peran setiap atribut dalam dataset. Dalam konteks penelitian ini, atribut Hasil ditetapkan sebagai label target yang akan diklasifikasikan oleh model, sedangkan atribut lainnya digunakan sebagai fitur prediksi. Untuk memastikan model yang dibangun memiliki performa yang baik dan tidak overfitting, dilakukan validasi menggunakan operator *Split Validation*. Proses validasi ini melibatkan pemisahan data menjadi data latih (training) dan data uji (testing). Rasio pembagian data ditetapkan pada 0.8, yang berarti 80% dari data digunakan sebagai data uji dan 20% sebagai data latih. Setelah itu pada detail proses *Split Validation* dilakukan pelatihan model *Random Forest*. Dapat dilihat pada gambar 5.



Gambar 5. Pelatihan dan pengujian model random forest tanpa optimasi fitur

Pada tahap pelatihan, algoritma *Random Forest* digunakan untuk membangun model klasifikasi. Data latih yang telah dipisahkan sebelumnya digunakan untuk melatih model ini. Operator *Random Forest* menghasilkan model yang menggabungkan beberapa pohon keputusan untuk meningkatkan akurasi prediksi. Setelah model dilatih, operator *Apply Model* digunakan untuk menerapkan model yang telah dilatih pada data uji. Langkah ini bertujuan untuk memprediksi label dari data uji dan mengevaluasi performa model pada data yang belum pernah dilihat sebelumnya. Operator *Performance* digunakan untuk mengevaluasi performa model yang telah diterapkan pada data uji. Metrik seperti akurasi dihitung untuk memberikan gambaran yang komprehensif mengenai seberapa baik model dapat mengklasifikasikan data penyakit paru-paru. Berikut hasil kinerja algoritma *random forest* tanpa optimasi fitur pada gambar 6.

accuracy: 89.45%

	true Ya	true Tidak	class precision
pred. Ya	74	0	100.00%
pred. Tidak	21	104	83.20%
class recall	77.89%	100.00%	

Gambar 6. Tabel Hasil Akurasi Pengujian *Random Forest* Tanpa Optimasi Fitur

Pada gambar 6 hasil dari kinerja algoritma *random forest* tanpa menggunakan optimasi fitur, Didapatkan hasil akurasi sebesar 89.45%.

precision: 83.20% (positive class: Tidak)

	true Ya	true Tidak	class precision
pred. Ya	74	0	100.00%
pred. Tidak	21	104	83.20%
class recall	77.89%	100.00%	

Gambar 7. Tabel Hasil Presisi Pengujian *Random Forest* Tanpa Optimasi Fitur

Pada gambar 7 hasil dari kinerja algoritma *random forest* tanpa menggunakan optimasi fitur, Didapatkan hasil presisi sebesar 83.20%.

recall: 100.00% (positive class: Tidak)

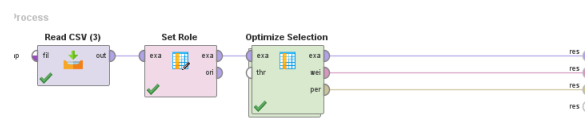
	true Ya	true Tidak	class precision
pred. Ya	74	0	100.00%
pred. Tidak	21	104	83.20%
class recall	77.89%	100.00%	

Gambar 8. Tabel Hasil Recall Pengujian *Random Forest* Tanpa Optimasi Fitur

Pada gambar 8 hasil dari kinerja algoritma *random forest* tanpa menggunakan optimasi fitur, Didapatkan hasil recall sebesar 100.00%.

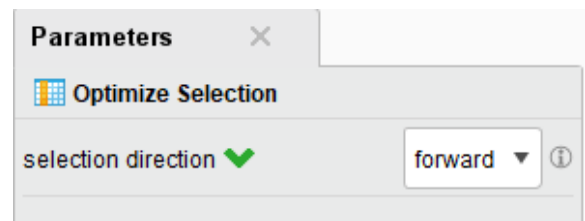
4.2. Proses Pengujian Algoritma *Random Forest* menggunakan Optimasi Fitur

Pada tahap ini dilakukan pengujian Algoritma *Random Forest* menggunakan Optimasi Fitur, pada proses ini ditambahkan operator *Optimize Selection* dengan metode *Forward Selection*. Berikut gambar 9 proses pengujian algoritma *random forest* menggunakan optimasi fitur.



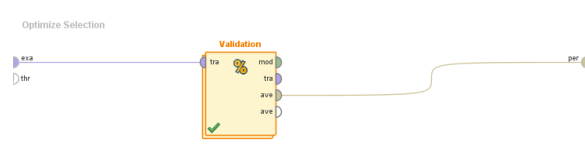
Gambar 9. Proses Pengujian Algoritma *Random Forest* menggunakan Optimasi Fitur

Pada gambar 9 Operator *Optimize Selection* digunakan untuk melakukan optimasi pemilihan fitur. Pada tahap ini, *Optimize Selection* digunakan untuk memilih fitur-fitur yang paling relevan. Parameter *selection direction* diatur ke 'forward' untuk menggunakan metode *Forward Selection*. Berikut gambar 10 detail *Optimize Selection*.



Gambar 10. Detail *Optimize Selection*

Pada gambar 10 metode *forward selection* ini menambahkan fitur satu per satu ke dalam model dan mengevaluasi peningkatan performa setiap kali fitur baru ditambahkan. Setelah fitur-fitur terbaik terpilih, operator *Validation* digunakan untuk melakukan validasi model. Berikut gambar 11 Detail Proses *Optimize Selection*.



Gambar 11. Validasi Model

Pada gambar 11 Setelah proses validasi, model dilatih menggunakan algoritma *Random Forest* pada data latih dengan fitur-fitur terpilih. Operator *Apply Model* digunakan untuk menerapkan model pada data uji, dan operator *Performance* digunakan untuk mengevaluasi performa model berdasarkan metrik seperti akurasi. Berikut gambar 12 metrik akurasi hasil kinerja algoritma *random forest* menggunakan optimasi fitur dengan metode *forward selection*.

accuracy: 92.46%

	true Ya	true Tidak	class precision
pred. Ya	91	11	89.22%
pred. Tidak	4	93	95.88%
class recall	95.79%	89.42%	

Gambar 12. Tabel Hasil Akurasi Pengujian *Random Forest* Menggunakan Optimasi Fitur

Pada gambar 12 hasil dari kinerja algoritma *random forest* menggunakan optimasi fitur, Didapatkan hasil akurasi sebesar 92.46%.

precision: 95.88% (positive class: Tidak)

	true Ya	true Tidak	class precision
pred. Ya	91	11	89.22%
pred. Tidak	4	93	95.88%
class recall	95.79%	89.42%	

Gambar 13. Tabel Hasil Presisi Pengujian *Random Forest* Menggunakan Optimasi Fitur

Pada gambar 13 hasil dari kinerja algoritma *random forest* menggunakan optimasi fitur, Didapatkan hasil presisi sebesar 95.88%.

recall: 89.42% (positive class: Tidak)

	true Ya	true Tidak	class precision
pred. Ya	91	11	89.22%
pred. Tidak	4	93	95.88%
class recall	95.79%	89.42%	

Gambar 14. Tabel Hasil Recall Pengujian *Random Forest* Menggunakan Optimasi Fitur

Pada gambar 14 hasil dari kinerja algoritma *random forest* menggunakan optimasi fitur, Didapatkan hasil *recall* sebesar 89.42%.

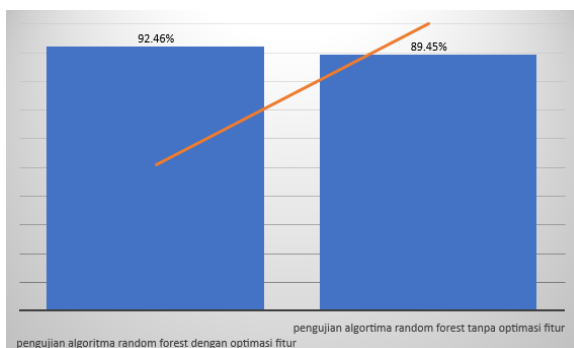
4.3. Kinerja Hasil Pengujian

4.3.1. Pengujian Tanpa Optimasi Fitur

Pada pengujian pertama, algoritma *Random Forest* diterapkan tanpa menggunakan optimasi fitur. Model dilatih dan diuji menggunakan dataset penyakit paru-paru. Hasil evaluasi menunjukkan bahwa model *Random Forest* mencapai akurasi sebesar 89.45%. Meskipun hasil ini menunjukkan kinerja yang cukup baik, masih terdapat ruang untuk perbaikan melalui optimasi fitur.

4.3.2. Pengujian dengan Optimasi Fitur

Pada pengujian kedua, dilakukan optimasi fitur menggunakan metode *Forward Selection*. Metode ini membantu dalam memilih fitur-fitur yang paling relevan, yang kemudian digunakan untuk melatih model *Random Forest*. Setelah proses optimasi fitur dan pelatihan model selesai, hasil evaluasi menunjukkan peningkatan akurasi yang signifikan. Model *Random Forest* dengan optimasi fitur berhasil mencapai akurasi sebesar 92.46%.



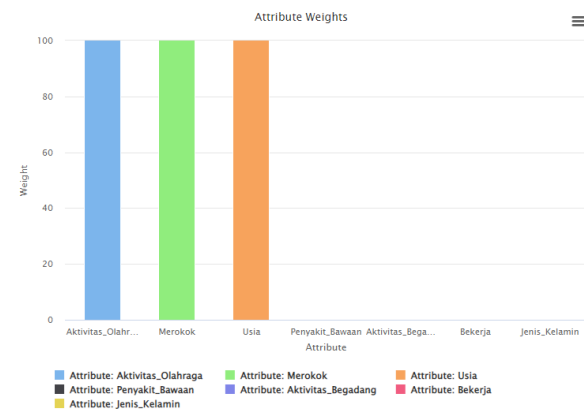
Gambar 15. Grafik Hasil Pengujian Tanpa Optimasi Fitur Dan Dengan Optimasi Fitur

Dari kedua pengujian yang telah dilakukan, terlihat bahwa penggunaan optimasi fitur dapat

meningkatkan performa *model* secara signifikan. Berikut gambar 11 grafik hasil pengujian algoritma *random forest* tanpa menggunakan optimasi fitur dan pengujian algoritma *random forest* dengan optimasi fitur.

Pada gambar 15 grafik hasil akurasi pengujian tanpa optimasi fitur dan dengan optimasi fitur.

4.4. Hasil Atribut Algoritma *Random Forest* menggunakan Optimasi Fitur



Gambar 16. Hasil Atribut Yang Dioptimalkan

Pada gambar 16 Dari 7 atribut dari data penyakit paru-paru, Atribut yang dipilih hanya tiga atribut seperti Aktivitas_olahraga, Merokok, dan Usia. dari keseluruhan atribut hanya 3 ini memiliki bobot tertinggi, masing-masing dengan nilai 100. Hal ini menunjukkan bahwa aktivitas olahraga, kebiasaan merokok, dan usia merupakan faktor yang paling signifikan dalam memprediksi hasil pada dataset penyakit paru-paru.

4.5. Evaluasi Model

Pada penelitian ini, evaluasi dilakukan untuk menilai kinerja algoritma *Random Forest* sebelum dan setelah penerapan metode optimasi fitur *Forward Selection*. Penilaian dilakukan menggunakan beberapa metrik evaluasi yang mencakup akurasi, presisi, recall,. Metrik-metrik ini memberikan gambaran yang komprehensif tentang performa model dalam mengklasifikasikan data penyakit paru-paru.

4.5.1. Evaluasi Awal Tanpa Optimasi Fitur

Pada tahap awal, algoritma *Random Forest* diterapkan tanpa optimasi fitur untuk membangun model klasifikasi. Hasil pengujian awal ini menunjukkan bahwa model mampu mencapai akurasi sebesar 89.45%, Presisi 83.20%, Recall 100.00%. Meskipun akurasi ini cukup tinggi, penting untuk mencermati metrik lain seperti presisi dan recall untuk memahami performa model secara lebih mendalam.

4.5.2. Evaluasi Setelah Optimasi Fitur dengan *Forward Selection*

Setelah menerapkan metode optimasi fitur *Forward Selection*, model *Random Forest* dibangun

kembali menggunakan subset fitur yang telah dioptimasi. Evaluasi terhadap model ini menunjukkan adanya peningkatan signifikan pada berbagai metrik evaluasi. akurasi model meningkat menjadi 92,46%, presisi 95,88%, recall 89.42%.

4.5.3. Hasil Evaluasi

Hasil evaluasi menunjukkan bahwa penerapan metode optimasi fitur *Forward Selection* berhasil meningkatkan kinerja model secara signifikan. Peningkatan akurasi sebesar 3,01% menunjukkan bahwa optimasi fitur memungkinkan model untuk membuat prediksi yang lebih akurat. Selain itu, peningkatan pada metrik presisi, recall, mengindikasikan bahwa model tidak hanya lebih akurat, tetapi juga lebih seimbang dalam mengidentifikasi prediksi positif dan negatif. Secara keseluruhan, penggunaan *Forward Selection* dalam optimasi fitur terbukti efektif dalam meningkatkan performa algoritma *Random Forest* dalam klasifikasi penyakit paru-paru. Hasil ini menunjukkan pentingnya pemilihan fitur yang relevan dalam proses pembelajaran mesin untuk menghasilkan model yang lebih andal dan efisien.

5. KESIMPULAN DAN SARAN

Penelitian ini menunjukkan bahwa algoritma *Random Forest* tanpa optimasi fitur mencapai akurasi sebesar 89.45%, Presisi 83.20%, Recall 100.00%. yang menunjukkan kinerja yang cukup baik, meskipun masih ada ruang untuk perbaikan. Setelah menerapkan optimasi fitur menggunakan metode *Forward Selection*, kinerja model meningkat secara signifikan, dengan akurasi mencapai 92.46%, presisi 95,88%, recall 89.42%. Peningkatan akurasi sebesar 3.01% ini menunjukkan bahwa pemilihan fitur yang tepat memiliki pengaruh besar terhadap performa model. Analisis atribut penting mengidentifikasi bahwa dari tujuh atribut yang digunakan dalam dataset penyakit paru-paru, hanya tiga atribut yang dipilih sebagai yang paling signifikan: aktivitas olahraga, kebiasaan merokok, dan usia. Ketiga atribut ini memiliki bobot tertinggi dan memberikan kontribusi terbesar dalam prediksi penyakit paru-paru. Hasil penelitian ini menekankan pentingnya optimasi fitur dalam meningkatkan performa model klasifikasi. Penggunaan metode seperti *Forward Selection* dapat membantu dalam memilih fitur-fitur yang paling relevan dan meningkatkan akurasi prediksi. Selain *Forward Selection*, ada berbagai metode optimasi fitur lain seperti *Backward Elimination*, *Recursive Feature Elimination*, atau metode berbasis algoritma genetik yang dapat dieksplorasi untuk melihat apakah ada peningkatan performa yang lebih signifikan.

DAFTAR PUSTAKA

[1] A. Wenda, Kraugusteeliana, A. Suryanto, and S. Alam, "Sistem Pakar Untuk Mendiagnosa Penyakit Paru-Paru dengan Menggunakan Metode Teorema Bayes," *J. Media Inform.*

Budidarma, vol. 7, no. 2, pp. 82–88, 2023, doi: 10.30865/mib.v7i1.5394.

- [2] F. Meila Azzahra Sofyan, A. Voutama, and Y. Umaidah, "Penerapan Algoritma C4.5 Untuk Prediksi Penyakit Paru-Paru Menggunakan Rapidminer," *JATI (Jurnal Mhs. Tek. Inform.,* vol. 7, no. 2, pp. 1409–1415, 2023, doi: 10.36040/jati.v7i2.6810.
- [3] A. Buchori, S. Khotijah, and A. S. Ramdan, "Sistem Pakar Diagnosa Penyakit Paru-Paru Menggunakan Metode Naive Bayes Classifier Berbasis Java," *Sist. Pakar Diagnosa Penyakit Paru-Paru Menggunakan Metod. Naive Bayes Classif. Berbas. Java*, pp. 1–7, 2022.
- [4] M. M. A. Andika Dany Prasetya, "Implementasi Data Mining Untuk Strategi Penjualan Pada Menu Warkoppitulur Dengan Menggunakan Metode Association Rule," vol. 6, no. 1, pp. j4-1-j4-16, 2024.
- [5] D. P. Utomo and M. Mesran, "Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung," *J. Media Inform. Budidarma*, vol. 4, no. 2, p. 437, 2020, doi: 10.30865/mib.v4i2.2080.
- [6] A. F. Riany and G. Testiana, "Penerapan Data Mining untuk Klasifikasi Penyakit Stroke Menggunakan Algoritma Naïve Bayes," *J. SAINTEKOM*, vol. 13, no. 1, pp. 42–54, 2023, doi: 10.33020/saintekom.v13i1.352.
- [7] Yuli, "Data Mining : Klasifikasi Menggunakan Algoritma C4 . 5 Data mining merupakan bagian dari tahapan proses Knowledge Discovery in Database (KDD) . Jurnal Edik Informatika," *J. Edik Inform.*, vol. 2, no. 2, pp. 213–219, 2019.
- [8] A. K. Febianto, C. A. Sugianto, P. Studi, T. Informatika, and J. Barat, "Optimalisasi Algoritma Klasifikasi Ensemble Menggunakan Algoritma Genetika Untuk Prediksi Resiko Diabetes," vol. 5, no. 2, pp. 205–213, 2024.
- [9] H. Marlina, A. Zulus, and O. L. W. H. Wijaya, "Penerapan Algoritma Random Forest Dalam Klasifikasi Penjurusan Di SMA Negeri Tugumulyo," *BRAHMANA J. Penerapan Kecerdasan Buatan Terakreditasi*, vol. 4, no. 2, pp. 138–143, 2023, [Online]. Available: https_tunasbangsa.ac.id/pkm/index.php/brahmana/article/view/188
- [10] M. Azhari, Z. Situmorang, and R. Rosnelly, "Perbandingan Akurasi, Recall, dan Presisi Klasifikasi pada Algoritma C4.5, Random Forest, SVM dan Naive Bayes," *J. Media Inform. Budidarma*, vol. 5, no. 2, p. 640, 2021, doi: 10.30865/mib.v5i2.2937.
- [11] A. I. Kusumarini, P. A. Hogantara, M. Fadhlurohman, and S. K. . M. K. Nurul Chamidah, "Perbandingan Algoritma Random Forest, Naive Bayes, Dan Decision Tree Dengan Oversampling Untuk Klasifikasi Bakteri E.Coli," *Pros. Semin. Nas. Mhs. Bid. Ilmu Komput. dan Apl.*, vol. 2, no. 1, pp. 792–799, 2021.

- [12] L. Faradila, A. Ismawardani, and A. N. F. Utami, "Pengaruh Kebiasaan Merokok Terhadap Kapasitas Paru- Paru dan Aktivitas Fisik Masyarakat Indonesia," *Semin. Nas. Potensi dan Kemandirian Drh.*, pp. 1–5, 2022, [Online]. Available: https_seminar.umpo.ac.id/index.php/SEMNASP_SP2KD/article/download/846/616
- [13] S. Rukmana, "Optimalisasi Peranan Financial Technology (Fintech Sharia) Menuju Era Digitalisasi Koperasi," *e-journal Ekon. Bisnis dan Akunt.*, vol. 9, no. 1, pp. 30–35, 2022.
- [14] H. Z. Yahya and Y. Ramdhani, "Klasifikasi Bumbu Dapur Pasar Menggunakan Metode Deep Neural Network Berbasis Android," *J. Nas. Komputasi dan Teknol. Inf.*, vol. 6, no. 1, pp. 19–33, 2023, doi: 10.32672/jnkti.v6i1.5569.
- [15] W. Aprilia, I. Kurniawan, M. Baydhowi, and T. Haryati, "Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest," *J. Sist. Inf.*, vol. 10, no. 1, pp. 163–171, 2021, [Online]. Available: <http://sistemasi.ftik.unisi.ac.id>
- [16] E. Fitri and D. Riana, "Analisa Perbandingan Model Prediction Dalam Prediksi Harga Saham Menggunakan Metode Linear Regression, Random Forest Regression Dan Multilayer Perceptron," *METHOMIKA J. Manaj. Inform. dan Komputerisasi Akunt.*, vol. 6, no. 1, pp. 69–78, 2022, doi: 10.46880/jmika.vol6no1.pp69-78.
- [17] R. Sanjaya and F. Fitriyani, "Prediksi Bedah Toraks Menggunakan Seleksi Fitur Forward Selection dan K-Nearest Neighbor," *J. Edukasi dan Penelit. Inform.*, vol. 5, no. 3, pp. 316–320, 2019, doi: 10.26418/jp.v5i3.35324.
- [18] L. Sari, A. Romadloni, and R. Listyaningrum, "Penerapan Data Mining dalam Analisis Prediksi Kanker Paru Menggunakan Algoritma Random Forest," *Infotekmesin*, vol. 14, no. 1, pp. 155–162, 2023, doi: 10.35970/infotekmesin.v14i1.1751.