

Optimasi Algoritma C4.5 Untuk Mendeteksi Kanker Paru-Paru Menggunakan Metode Penyeimbangan Data

Optimization of C4.5 Algorithm to Detect Lung Cancer Using Balanced Data Method

**Cut Alna Fadhilla^{a,1}, Muhammad Arif Sanjaya^{b,2}, Cahya Amanda Sahfitri^{d,3}, Yessy Agustin^{e,4},
Agellia ramadani^{e,5}**

^aProgram Studi Informatika, Langsa, 24416

^bProgram Studi Bahasa Inggris, Langsa, 24416

^cProgram Studi Informatika, Langsa, 24416

^dProgram Studi Informatika, Langsa, 24416

^eProgram Studi Informatika, Langsa, 24416

cutalnafadhilla@unsam.ac.id¹, ariev2209@unsam.ac.id², cahhyaamanda@gmail.com³, yessya@gmail.com⁴,
agelliar@gmail.com⁵

ABSTRAK

Kanker paru-paru merupakan salah satu jenis kanker yang umum terjadi di Indonesia, dengan estimasi sekitar 34 ribu kasus baru setiap tahunnya. Tingginya angka ini menunjukkan pentingnya deteksi dini, namun proses deteksi masih terkendala oleh keterbatasan akurasi metode konvensional. Untuk mengatasi permasalahan tersebut, penelitian ini menggunakan algoritma C4.5 dalam membangun model prediksi yang efektif untuk mendeteksi kanker paru-paru berdasarkan data medis pasien. Dataset yang digunakan berasal dari Kaggle dan mencakup atribut seperti jenis kelamin, usia, kebiasaan merokok, penyakit kronis, kelelahan, alergi, konsumsi alkohol, serta faktor lainnya yang diduga berperan dalam risiko terkena kanker paru-paru. Mengingat adanya ketidakseimbangan kelas dalam data, penelitian ini menerapkan Synthetic Minority Oversampling Technique (SMOTE) untuk memperbaiki distribusi data sehingga algoritma dapat mempelajari pola secara lebih optimal. Dengan pendekatan ini, penelitian bertujuan menghasilkan prediksi yang lebih akurat dalam mendeteksi kanker paru-paru. Hasil penelitian menunjukkan bahwa algoritma C4.5 mampu mencapai tingkat akurasi sebesar 87,65% dalam mengklasifikasikan kondisi pasien, sehingga dapat diandalkan sebagai pendekatan yang efektif untuk mendukung diagnosis medis kanker paru-paru.

Kata Kunci : C.45, Kanker Paru-paru, Klasifikais Data Medis

ABSTRACT

Lung cancer is one of the most common types of cancer in Indonesia, with an estimated 34 thousand new cases each year. This high number shows the importance of early detection, but the detection process is still constrained by the limited accuracy of conventional methods. To overcome this problem, this study uses the C4.5 algorithm to build an effective prediction model to detect lung cancer based on patient medical data. The dataset used comes from Kaggle and includes attributes such as gender, age, smoking habits, chronic diseases, fatigue, allergies, alcohol consumption, and other factors that are thought to play a role in the risk of developing lung cancer. Given the class imbalance in the data, this study applies the Synthetic Minority Oversampling Technique (SMOTE) to improve data distribution so that the algorithm can learn patterns more optimally. With this approach, the study aims to produce more accurate predictions in detecting lung cancer. The results of the study show that the C4.5 algorithm is able to achieve an accuracy level of 87.65% in classifying patient conditions, so it can be relied on as an effective approach to support the medical diagnosis of lung cancer.

Keywords : C.45, Kanker Paru-paru, Klasifikais Data Medis

Disubmit: 12 Mei 2024

Info Artikel :
Direview: 10 Juni 2024

Diterima : 11 Juli 2024

Copyright © 2024 - Journal UPU. All rights reserved.

1. PENDAHULUAN

Kanker paru-paru adalah salah satu jenis kanker yang memiliki angka kematian tertinggi di seluruh dunia. Berdasarkan data Global Cancer Observatory Organisasi Kesehatan Dunia, kanker paru menduduki posisi pertama dengan 1.796.144 kematian. Di Indonesia kanker paru menjadi jenis kanker dengan angka kejadian tertinggi sekitar 34.783 kasus baru dan 30.843 kematian selama tahun 2020 [1]. Dengan mendeteksi penyakit paru-paru lebih awal, penanganan dapat dimulai lebih cepat, dan ini dapat meningkatkan peluang kesembuhan atau pengendalian penyakit [3]. Hal ini menyebabkan perlunya metode yang lebih efektif untuk mendeteksi kanker paru-paru sejak dini.

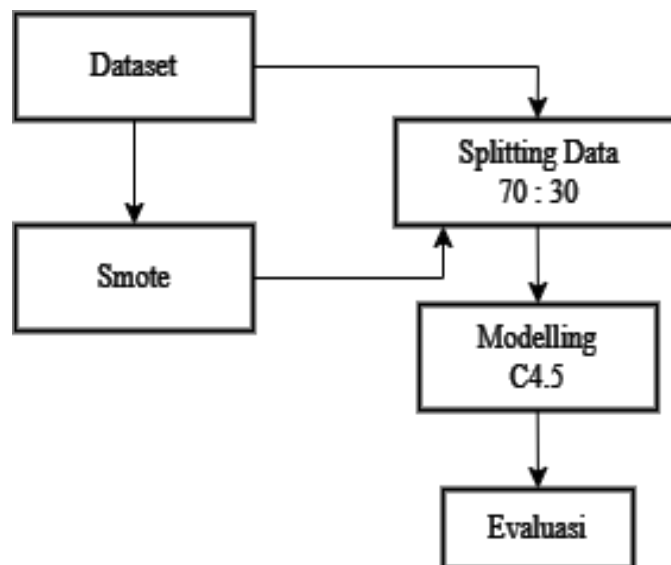
Salah satu pendekatan yang menjanjikan adalah penerapan algoritma pembelajaran mesin. Algoritma C4.5 merupakan suatu algoritma yang dapat digunakan untuk melakukan klasifikasi atau pengelompokan pada dataset [4]. Algoritma ini dapat menganalisis data medis dengan efisien dan menghasilkan model klasifikasi yang akurat. Algoritma ini telah terbukti efektif dalam berbagai penelitian untuk memprediksi dan mengklasifikasikan penyakit berdasarkan data klinis. Di dalam [2], para penulis menggunakan algoritma Random Forest dengan capaian akurasi sebesar 98% . Di dalam [5], para penulis menggunakan algoritma regresi linear dengan capaian akurasi sebesar 90%. Di dalam [6], para penulis menggunakan algoritma K-Nearest Neighbor dengan capaian akurasi sebesar 80,40%. Namun demikian, penelitian ini menunjukkan bahwa algoritma C4.5 juga memiliki performa yang baik dengan tingkat akurasi sebesar 87,65%, menjadikannya alternatif yang layak untuk digunakan dalam mendeteksi kanker paru-paru, terutama dengan kompleksitas dan interpretasi model yang lebih sederhana.

Kanker paru-paru merupakan salah satu penyebab utama kematian di seluruh dunia, dengan angka kejadian yang terus meningkat setiap tahunnya. Deteksi dini kanker paru-paru sangat penting untuk meningkatkan peluang kesembuhan pasien. Dalam konteks ini, algoritma pembelajaran mesin, khususnya algoritma C4.5, telah menunjukkan potensi yang signifikan dalam menganalisis data medis dan membantu dalam pengambilan keputusan klinis [7]. Algoritma C4.5, yang merupakan pengembangan dari algoritma ID3, digunakan untuk membangun pohon keputusan yang dapat mengklasifikasikan data berdasarkan atribut yang relevan. Namun, salah satu tantangan utama dalam penerapan algoritma ini adalah ketidakseimbangan data, di mana jumlah contoh dari kelas positif (misalnya, pasien dengan kanker paru-paru) jauh lebih sedikit dibandingkan dengan kelas negatif. Ketidakseimbangan ini dapat menyebabkan model yang dihasilkan cenderung bias terhadap kelas mayoritas, sehingga mengurangi akurasi dan efektivitas deteksi kanker. Oleh karena itu, penting untuk menerapkan metode penyeimbangan data yang dapat meningkatkan representasi kelas minoritas dan, pada gilirannya, meningkatkan kinerja algoritma C4.5 [8].

Selain itu, penelitian ini juga akan membahas pentingnya pemilihan fitur yang tepat dalam proses klasifikasi. Fitur-fitur yang relevan dan informatif dapat meningkatkan kinerja algoritma C4.5 secara signifikan. Dengan menggunakan teknik pemilihan fitur, kami bertujuan untuk mengurangi kompleksitas model dan meningkatkan interpretabilitas hasil [9]. Penelitian ini diharapkan tidak hanya memberikan kontribusi terhadap pengembangan metode deteksi kanker paru-paru yang lebih efektif, tetapi juga memberikan wawasan yang lebih dalam mengenai penerapan algoritma pembelajaran mesin dalam bidang kesehatan, khususnya dalam konteks diagnosis penyakit yang kompleks. Dalam publikasi ini, kami akan membahas optimasi algoritma C4.5 untuk mendeteksi kanker paru-paru dengan menggunakan metode penyeimbangan data. Kami akan mengeksplorasi berbagai teknik penyeimbangan, seperti oversampling, undersampling, dan kombinasi keduanya, serta menganalisis dampaknya terhadap akurasi

dan sensitivitas model [10]. Dengan pendekatan ini, diharapkan dapat diperoleh model yang lebih akurat dan dapat diandalkan dalam mendeteksi kanker paru-paru, sehingga berkontribusi pada upaya peningkatan diagnosis dan perawatan pasien. Penelitian ini bertujuan untuk mengeksplorasi penerapan algoritma C4.5 dalam mendiagnosis kanker paru-paru menggunakan dataset yang berisi informasi klinis pasien. Dengan meningkatnya akurasi dalam prediksi, model ini memiliki potensi untuk digunakan dalam praktik medis dan dunia kesehatan secara luas (Rofiani et al., n.d.). Dengan harapan bahwa penelitian ini dapat memberikan kontribusi dalam mengembangkan alat bantu diagnosis yang lebih cepat dan akurat, serta memberikan wawasan lebih lanjut mengenai potensi penggunaan pembelajaran mesin dalam bidang kedokteran.

2. METODE



Gambar 1 Metode Penelitian

a) Dataset

Data yang digunakan dalam penelitian ini merupakan dataset Lung Cancer berbentuk csv yang diperoleh melalui web www.kaggle.com. Data terdiri dari 16 kolom dan 309 baris. Dimana kolom mewakili variabel dan baris mewakili banyak responden. Tabel 1 menunjukkan sebuah dataset yang berisi informasi terkait pasien yang berpotensi mengalami kanker paru-paru. Berikut adalah penjelasan mengenai kolom-kolom yang terdapat dalam dataset tersebut:

- 1) Gender: Jenis kelamin pasien, yang dapat berupa 'M' untuk laki-laki dan 'F' untuk perempuan.
- 2) Age: Usia pasien dalam tahun, yang merupakan faktor penting dalam risiko terkena kanker paru-paru.
- 3) Smoking: Status merokok pasien, yang biasanya dinyatakan dalam kategori (misalnya, 1 untuk tidak merokok, 2 untuk perokok aktif, dan seterusnya). Merokok adalah salah satu faktor risiko utama untuk kanker paru-paru.
- 4) Chest Pain: Tingkat atau frekuensi nyeri dada yang dialami pasien, yang dapat menjadi gejala awal dari masalah paru-paru.
- 5) Lung Cancer: Status diagnosis kanker paru-paru, yang dinyatakan sebagai 'YES' (ya) atau 'NO' (tidak). Ini adalah variabel target yang ingin diprediksi oleh model.

Dataset ini dapat digunakan untuk analisis lebih lanjut, seperti penerapan algoritma pembelajaran mesin (misalnya, C4.5) untuk mengklasifikasikan pasien berdasarkan risiko kanker paru-paru. Dengan memanfaatkan informasi yang ada, model dapat dilatih untuk mengenali pola dan faktor yang berkontribusi terhadap diagnosis kanker, sehingga dapat membantu dalam deteksi dini dan pengambilan keputusan klinis.

Tabel 1 Lung Cancer Dataset

NO	Gender	Age	Smoking	...	Chest Pain	Lung Cancer
0	M	69	1	...	2	YES
1	M	74	2	...	2	YES
2	F	59	1	...	2	NO
3	M	63	2	...	2	NO
4	F	63	1	...	1	NO
...	...	...	...	...	...	...
307	M	67	2	...	2	YES
308	M	62	1	...	1	YES

b) Dataset Synthetic Minority Oversampling Technique (SMOTE)

SMOTE merupakan salah satu turunan dari oversampling. Teknik ini mensintesis sampel baru dari kelas minoritas untuk menyeimbangkan dataset dengan cara sampling ulang sampel kelas minoritas [6]. Synthetic Minority Oversampling Technique (SMOTE) adalah metode yang digunakan untuk mengatasi masalah ketidakseimbangan kelas dalam dataset, terutama dalam konteks pembelajaran mesin. Ketidakseimbangan kelas terjadi ketika jumlah contoh dari kelas minoritas (misalnya, pasien dengan kanker paru-paru) jauh lebih sedikit dibandingkan dengan kelas mayoritas (misalnya, pasien tanpa kanker). Hal ini dapat menyebabkan model pembelajaran mesin cenderung bias terhadap kelas mayoritas, sehingga mengurangi akurasi dan efektivitas dalam mendeteksi kelas minoritas.

SMOTE bekerja dengan cara menciptakan contoh sintesis baru untuk kelas minoritas. Alih-alih hanya menyalin contoh yang ada, SMOTE menghasilkan data baru dengan cara mengambil contoh dari kelas minoritas dan menciptakan titik data baru di antara mereka. Proses ini melibatkan pemilihan satu atau lebih contoh dari kelas minoritas dan kemudian menghasilkan contoh baru dengan interpolasi antara contoh yang dipilih. Dengan cara ini, SMOTE tidak hanya meningkatkan jumlah contoh dalam kelas minoritas, tetapi juga memperkaya keragaman data, yang dapat membantu model belajar lebih baik. Penerapan SMOTE dapat meningkatkan kinerja model dalam berbagai tugas klasifikasi, termasuk deteksi penyakit. Dalam konteks deteksi kanker paru-paru, penggunaan SMOTE dapat membantu dalam meningkatkan sensitivitas model, sehingga lebih mampu mendeteksi pasien yang benar-benar memiliki kanker. Selain itu, dengan mengurangi bias terhadap kelas mayoritas, SMOTE dapat menghasilkan model yang lebih seimbang dan akurat, yang pada gilirannya dapat berkontribusi pada diagnosis yang lebih baik dan pengambilan keputusan klinis yang lebih tepat.

Namun, meskipun SMOTE memiliki banyak keuntungan, ada beberapa hal yang perlu diperhatikan. Salah satunya adalah risiko overfitting, di mana model mungkin belajar terlalu banyak dari contoh

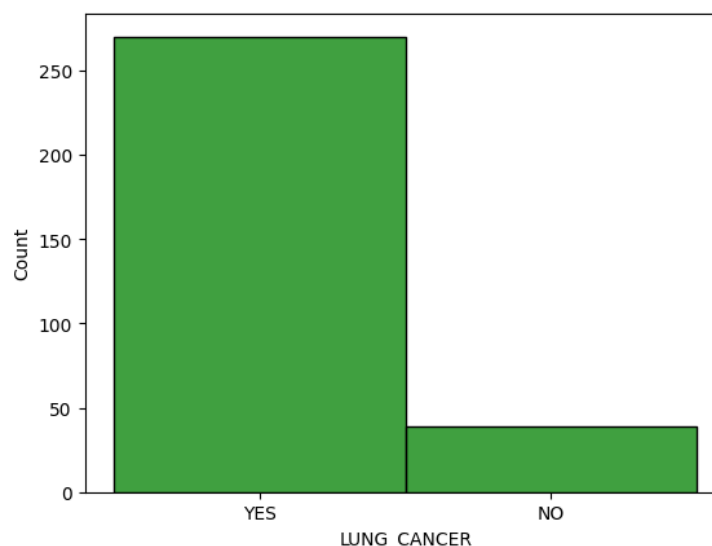
sintetis yang dihasilkan, sehingga mengurangi kemampuannya untuk generalisasi pada data baru. Oleh karena itu, penting untuk menggabungkan SMOTE dengan teknik lain, seperti validasi silang dan pemilihan fitur, untuk memastikan bahwa model yang dihasilkan tetap robust dan dapat diandalkan dalam praktik klinis.

c) Algoritma C4.5

Algoritma C4.5 merupakan salah satu teknik dalam pengolahan Data Mining [12]. Algoritma ini menjadi salah satu algoritma untuk mengubah fakta yang besar menjadi pohon keputusan yang merepresentasikan aturan. Beberapa pengembangan yang dilakukan pada C4.5 antara lain dapat mengatasi *missing value*, *data kontinu*, dan *pruning*. Algoritma C4.5 adalah salah satu metode pembelajaran mesin yang digunakan untuk membangun pohon keputusan, yang merupakan alat yang efektif untuk klasifikasi data. Dikembangkan oleh Ross Quinlan, C4.5 merupakan pengembangan dari algoritma sebelumnya, ID3, dan dirancang untuk mengatasi beberapa keterbatasan ID3, seperti penanganan data yang hilang dan kemampuan untuk bekerja dengan atribut kontinu. C4.5 menggunakan pendekatan berbasis informasi untuk memilih atribut yang paling relevan dalam membagi dataset, dengan tujuan memaksimalkan informasi yang diperoleh dari setiap pembagian. Proses ini melibatkan perhitungan rasio informasi, yang membantu dalam menentukan atribut mana yang paling efektif dalam memisahkan kelas-kelas yang berbeda [13].

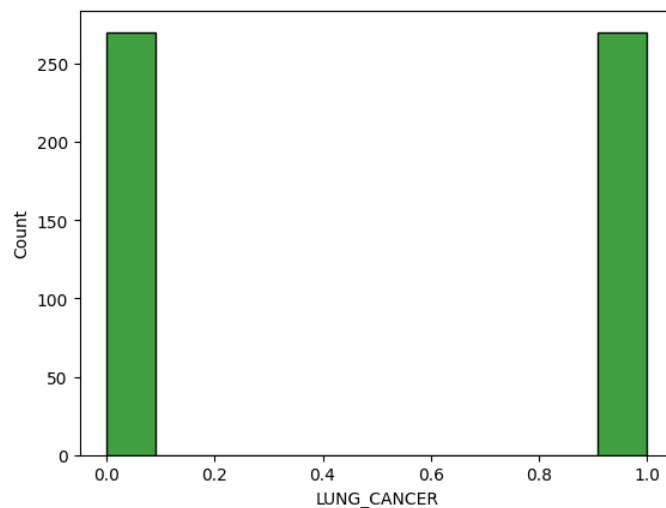
Salah satu keunggulan C4.5 adalah kemampuannya untuk menghasilkan pohon keputusan yang mudah dipahami dan diinterpretasikan, sehingga sangat berguna dalam konteks di mana transparansi dan interpretabilitas model sangat penting, seperti dalam diagnosis medis. Selain itu, C4.5 juga dapat menangani data yang tidak seimbang dengan lebih baik dibandingkan dengan algoritma lain, meskipun tetap memerlukan perhatian khusus terhadap masalah ketidakseimbangan kelas. Dengan kemampuannya untuk menangani berbagai jenis data dan menghasilkan model yang efisien, C4.5 telah menjadi salah satu algoritma yang populer dalam berbagai aplikasi, termasuk analisis data, pengenalan pola, dan pengambilan keputusan berbasis data.

3. HASIL DAN PEMBAHASAN

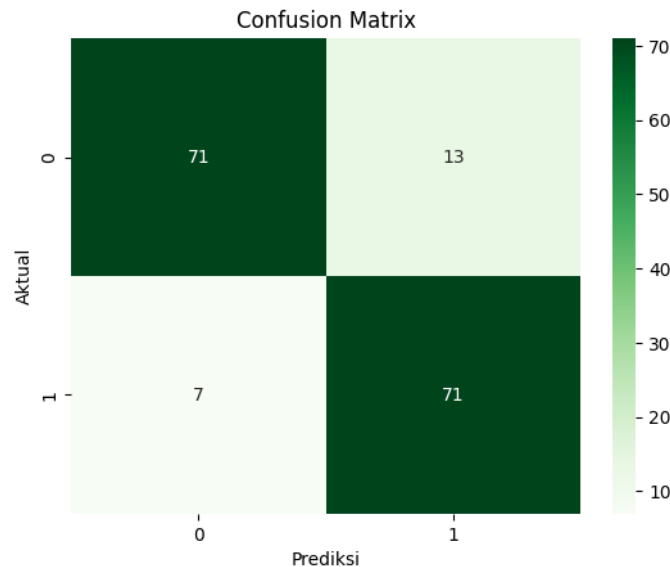


Gambar 2 Data Sebelum Proses SMOTE

Gambar di atas menunjukkan distribusi data awal pada variabel LUNG_CANCER yang divisualisasikan menggunakan histogram untuk mengidentifikasi potensi ketidakseimbangan data. Gambar di atas adalah histogram yang menunjukkan distribusi jumlah kasus berdasarkan status diagnosis kanker paru-paru, dibagi menjadi dua kategori: "YES" (ada kanker) dan "NO" (tidak ada kanker). Pada sumbu X, kategori "LUNG_CANCER" menampilkan dua status tersebut, sementara sumbu Y menunjukkan jumlah kasus untuk masing-masing kategori. Dari histogram ini, terlihat bahwa jumlah pasien yang didiagnosis dengan kanker paru-paru ("YES") jauh lebih sedikit dibandingkan dengan jumlah pasien yang tidak memiliki kanker ("NO"). Hal ini mencerminkan adanya ketidakseimbangan kelas yang signifikan dalam dataset, di mana kelas mayoritas adalah pasien tanpa kanker. Ketidakseimbangan ini dapat menjadi tantangan dalam analisis data dan penerapan algoritma pembelajaran mesin, karena model mungkin cenderung bias terhadap kelas mayoritas. Oleh karena itu, teknik seperti SMOTE (Synthetic Minority Oversampling Technique) dapat digunakan untuk meningkatkan representasi kelas minoritas dan meningkatkan kinerja model dalam mendeteksi kanker paru-paru. Histogram ini memberikan gambaran yang jelas tentang distribusi data dan pentingnya penanganan ketidakseimbangan kelas dalam analisis lebih lanjut. Ketidakseimbangan ini kemudian diatasi dengan teknik oversampling menggunakan SMOTE (Synthetic Minority Oversampling Technique), yang menghasilkan sampel sintesis pada kategori minoritas agar distribusi menjadi lebih seimbang. Hasil dari penggunaan SMOTE untuk menyeimbangkan data adalah sebagaimana Gambar 3.



Gambar 3 Data Setelah Proses SMOTE

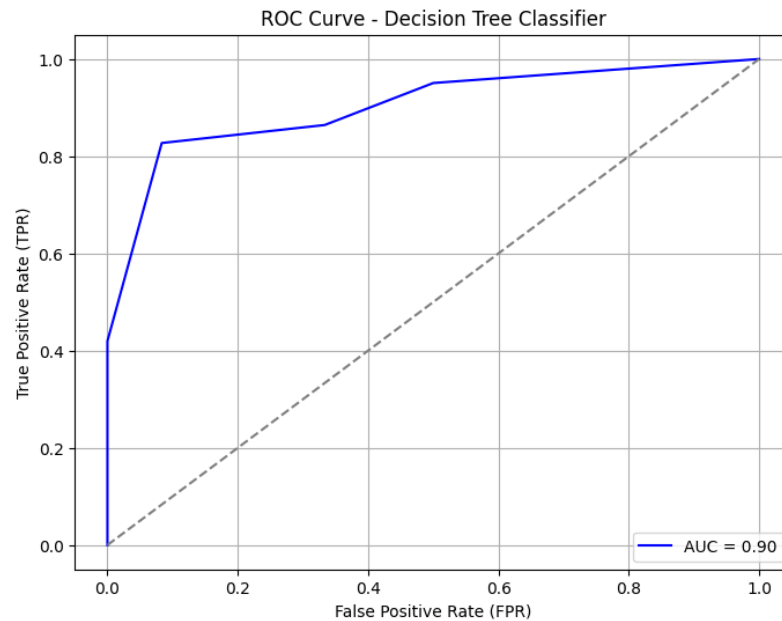


Gambar 4 Confusion Matrix

Gambar 4 adalah matriks kebingungan (confusion matrix) yang digunakan untuk mengevaluasi kinerja model klasifikasi dalam mendeteksi kanker paru-paru. Matriks ini memberikan gambaran yang jelas tentang bagaimana model melakukan prediksi dibandingkan dengan label aktual. Pada sumbu X, kategori prediksi dibagi menjadi dua, yaitu 0 (tidak ada kanker) dan 1 (ada kanker), sementara sumbu Y menunjukkan label aktual yang juga terbagi dalam dua kategori yang sama. Dengan cara ini, matriks kebingungan memungkinkan analisis yang lebih mendalam terhadap hasil prediksi model.

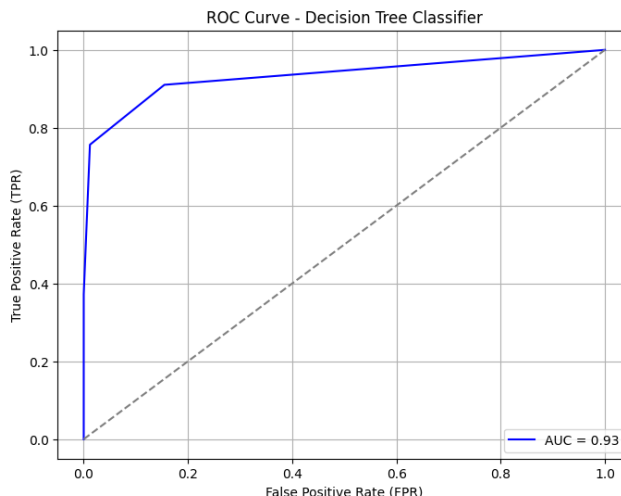
Dalam matriks ini, nilai-nilai yang terdapat di dalamnya memberikan informasi penting mengenai kinerja model. True Negatives (TN) sebanyak 71 menunjukkan jumlah pasien yang benar-benar tidak memiliki kanker dan diprediksi dengan tepat sebagai tidak memiliki kanker. Di sisi lain, False Positives (FP) sebanyak 13 menunjukkan jumlah pasien yang sebenarnya tidak memiliki kanker tetapi diprediksi sebagai memiliki kanker. Sementara itu, False Negatives (FN) sebanyak 7 menunjukkan pasien yang sebenarnya memiliki kanker tetapi tidak terdeteksi oleh model. True Positives (TP) sebanyak 71 menunjukkan pasien yang benar-benar memiliki kanker dan diprediksi dengan tepat sebagai memiliki kanker.

Secara keseluruhan, model ini menunjukkan kinerja yang baik dengan jumlah True Positives dan True Negatives yang tinggi. Namun, adanya 13 False Positives dan 7 False Negatives menunjukkan bahwa masih ada ruang untuk perbaikan. Khususnya, perhatian harus diberikan pada False Negatives, karena ini berarti ada pasien yang sebenarnya memiliki kanker tetapi tidak terdeteksi oleh model, yang dapat berakibat serius bagi kesehatan mereka. Matriks kebingungan ini sangat berguna untuk menghitung metrik evaluasi lainnya, seperti akurasi, presisi, recall, dan F1-score, yang dapat memberikan wawasan lebih dalam tentang kinerja model dalam konteks deteksi kanker paru-paru.



Gambar 5 ROC Curve Sebelum SMOTE

Gambar 5 adalah kurva Receiver Operating Characteristic (ROC) yang digunakan untuk mengevaluasi kinerja model klasifikasi, dalam hal ini, klasifikasi menggunakan algoritma pohon keputusan. Sumbu X menunjukkan tingkat positif palsu (False Positive Rate, FPR), sementara sumbu Y menunjukkan tingkat positif sejati (True Positive Rate, TPR). Kurva ROC menggambarkan trade-off antara sensitivitas (kemampuan model untuk mendeteksi kasus positif) dan spesifisitas (kemampuan model untuk menghindari kesalahan dalam mengklasifikasikan kasus negatif) pada berbagai ambang batas. Dari kurva yang ditampilkan, terlihat bahwa model memiliki kinerja yang baik, dengan TPR yang tinggi pada FPR yang rendah. Area di bawah kurva (AUC) sebesar 0.90 menunjukkan bahwa model memiliki kemampuan yang sangat baik dalam membedakan antara kelas positif dan negatif. Nilai AUC mendekati 1 menunjukkan bahwa model dapat mengklasifikasikan dengan akurasi tinggi, sementara nilai AUC mendekati 0.5 menunjukkan kinerja yang tidak lebih baik dari tebakan acak. Dengan demikian, hasil ini menunjukkan bahwa algoritma pohon keputusan yang digunakan dalam penelitian ini efektif dalam mendeteksi kanker paru-paru, memberikan keyakinan dalam penerapan model untuk diagnosis klinis. Namun, terdapat ketidakseimbangan data yang memungkinkan model cenderung bias terhadap kelas mayoritas.



Gambar 6 ROC Curve Sesudah SMOTE

Gambar 6 adalah kurva Receiver Operating Characteristic (ROC) yang menunjukkan kinerja model klasifikasi menggunakan algoritma pohon keputusan setelah penerapan metode Synthetic Minority Oversampling Technique (SMOTE). Pada sumbu X, terdapat tingkat positif palsu (False Positive Rate, FPR), sedangkan sumbu Y menunjukkan tingkat positif sejati (True Positive Rate, TPR). Kurva ROC ini menggambarkan bagaimana model berperilaku dalam membedakan antara pasien yang memiliki kanker paru-paru dan yang tidak, dengan berbagai ambang batas. Setelah penerapan SMOTE, terlihat bahwa model menunjukkan peningkatan kinerja yang signifikan, dengan AUC (Area Under the Curve) sebesar 0.93. Nilai AUC yang tinggi ini menunjukkan bahwa model memiliki kemampuan yang sangat baik dalam mengidentifikasi kasus positif, dengan TPR yang tinggi pada FPR yang rendah. Penerapan SMOTE membantu mengatasi masalah ketidakseimbangan kelas dalam dataset, yang sebelumnya dapat menyebabkan bias dalam prediksi. Dengan demikian, hasil ini menunjukkan bahwa penggunaan SMOTE tidak hanya meningkatkan jumlah contoh dalam kelas minoritas, tetapi juga meningkatkan akurasi dan keandalan model dalam mendeteksi kanker paru-paru, memberikan keyakinan lebih dalam penerapan model untuk diagnosis klinis.

Tabel 2 Performa Model C4.5

SMOTE		Accuracy	Precision	Recall	f1-Score
Sebelum	0	89,25	0,60	0,50	0,55
	1		0,93	0,95	0,94
Sesudah	0	87,65	0,91	0,85	0,88
	1		0,85	0,91	0,88

Model C4.5 menunjukkan performa yang sangat baik, model ini mampu memprediksi dengan benar sebagian besar data uji dengan akurasi sebesar 87,65%. Hasil ini menunjukkan bahwa model memiliki keseimbangan yang baik antara kemampuan mendeteksi kelas positif dan meminimalkan kesalahan prediksi. Tabel 2 menyajikan perbandingan metrik evaluasi model klasifikasi sebelum dan setelah penerapan metode Synthetic Minority Oversampling Technique (SMOTE). Metrik yang ditampilkan meliputi akurasi, presisi, recall, dan f1-score untuk dua kelas: 0 (tidak ada kanker) dan 1 (ada kanker). Sebelum penerapan SMOTE, model menunjukkan akurasi sebesar 89,25% untuk kelas 0 dan 93% untuk kelas 1, dengan presisi 0,60 dan recall 0,50 untuk kelas 0, serta presisi 0,93 dan recall 0,95 untuk kelas 1. Setelah penerapan SMOTE, akurasi untuk kelas 0 menurun menjadi 87,65%, sementara untuk kelas 1

tetap tinggi di 85%. Meskipun akurasi untuk kelas 0 mengalami penurunan, presisi dan recall untuk kelas 0 meningkat menjadi 0,91 dan 0,85, menunjukkan bahwa model menjadi lebih baik dalam mengidentifikasi pasien yang tidak memiliki kanker. Untuk kelas 1, presisi sedikit menurun menjadi 0,85, tetapi recall tetap tinggi di 0,91. F1-score, yang merupakan rata-rata harmonis dari presisi dan recall, menunjukkan peningkatan untuk kelas 1 menjadi 0,88, sementara kelas 0 tetap di 0,88. Secara keseluruhan, penerapan SMOTE membantu meningkatkan keseimbangan antara presisi dan recall, yang berkontribusi pada kinerja model yang lebih baik dalam mendeteksi kanker paru-paru.

4. KESIMPULAN

Penelitian ini membuktikan bahwa algoritma C4.5 merupakan salah satu metode yang efektif dalam mendeteksi kanker paru-paru berdasarkan data klinis pasien. Dengan menggunakan dataset Lung Cancer dan menerapkan teknik SMOTE untuk mengatasi ketidakseimbangan data, model C4.5 berhasil mencapai akurasi sebesar 87,65%. Penelitian ini berhasil mengevaluasi kinerja model klasifikasi dalam mendeteksi kanker paru-paru menggunakan algoritma pohon keputusan. Melalui penerapan metode Synthetic Minority Oversampling Technique (SMOTE), penelitian ini berhasil mengatasi masalah ketidakseimbangan kelas yang signifikan dalam dataset, yang sebelumnya dapat mempengaruhi akurasi dan keandalan model. Hasil analisis menunjukkan bahwa setelah penerapan SMOTE, model menunjukkan peningkatan dalam metrik evaluasi seperti presisi, recall, dan f1-score, terutama untuk kelas yang lebih jarang, yaitu pasien dengan kanker. Kurva ROC yang dihasilkan menunjukkan AUC sebesar 0,93, menandakan bahwa model memiliki kemampuan yang sangat baik dalam membedakan antara pasien yang memiliki kanker dan yang tidak. Meskipun terdapat sedikit penurunan dalam akurasi untuk kelas 0 setelah penerapan SMOTE, peningkatan dalam presisi dan recall menunjukkan bahwa model menjadi lebih efektif dalam mengidentifikasi kasus positif dan negatif.

Secara keseluruhan, penelitian ini menegaskan pentingnya penggunaan teknik oversampling seperti SMOTE dalam meningkatkan kinerja model klasifikasi, terutama dalam konteks deteksi dini kanker paru-paru. Hasil ini diharapkan dapat memberikan kontribusi positif bagi pengembangan metode diagnosis yang lebih akurat dan efektif dalam praktik klinis, serta membuka jalan untuk penelitian lebih lanjut dalam bidang ini. Selain itu, peningkatan nilai AUC dari 0,90 menjadi 0,93 setelah penerapan SMOTE menunjukkan bahwa model semakin mampu membedakan antara seseorang yang menderita kanker paru-paru dan yang tidak menderita.

UCAPAN TERIMA KASIH

Terimakasih kepada semua pihak yang telah memberikan dukungan, baik secara langsung maupun tidak langsung, dalam penyelesaian penelitian ini. Ucapan terima kasih juga kepada semua pihak yang telah berkontribusi dalam penyelesaian penelitian ini. Penulis juga ingin menyampaikan terima kasih kepada rekan-rekan sejawat yang telah memberikan masukan dan saran konstruktif, serta berbagi pengetahuan yang sangat berarti. Diskusi dan kolaborasi dengan mereka telah memperkaya pemahaman penulis mengenai topik yang diteliti dan meningkatkan kualitas penelitian ini.

REFERENSI

- [1] J. Hall and K. M. Brown, "Creating feelings of inclusion in adventure tourism: Lessons from the gendered sensory and affective politics of professional mountaineering," *Ann Tour Res*, vol. 97, p. 103505, Nov. 2022, doi: 10.1016/j.annals.2022.103505.
- [2] 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, Jan. 2023, doi: 10.35970/infotekmesin.v14i1.1751.
- [3] R. Rofiani, L. Oktaviani, D. Vernanda, and T. Hendriawan, "Penerapan Metode Klasifikasi Decision Tree dalam Prediksi Kanker Paru-Paru Menggunakan Algoritma C4.5," vol. 18, no. 1.

- [4] H. W. N. S. Putra, V. Atina, and J. Maulindar, "Penerapan Algoritme Decision Tree Pada Klasifikasi Penyakit Kanker Paru-Paru," *Jutisi : Jurnal Ilmiah Teknik Informatika dan Sistem Informasi*, vol. 12, no. 3, p. 967, Dec. 2023, doi: 10.35889/jutisi.v12i3.1323.
- [5] M. Abdul, R. Wahid, A. Nugroho, and A. H. Anshor, "Prediksi Penyakit Kanker Paru-Paru Dengan Algoritma Regresi Linier," *Bulletin of Information Technology (BIT)*, vol. 4, no. 1, pp. 63–74, 2023, doi: 10.47065/bit.v3i1.
- [6] T. Abdi Mangun, O. Nurdiawan, and A. Irma Purnamasari, "LUNG CANCER ANALYSIS USING K-NEAREST NEIGHBOR ALGORITHM," 2023. [Online]. Available: https://ejournal.ubibanyuwangi.ac.id/index.php/jurnal_tinsika
- [7] N. Maleki, Y. Zeinali, and S. T. A. Niaki, "A k-NN method for lung cancer prognosis with the use of a genetic algorithm for feature selection," *Expert Syst Appl*, vol. 164, p. 113981, Feb. 2021, doi: 10.1016/j.eswa.2020.113981.
- [8] Zudyanti Dwi Rahma Sari, Y. Arvita, and Jasmir Jasmir, "Penerapan Data Mining Untuk Prediksi Penyakit Diabetes Menggunakan Algoritma C4.5 Zudyanti Dwi Rahma Sari1, Ja," *Jurnal Informatika Dan Rekayasa Komputer(JAKAKOM)*, vol. 4, no. 1, pp. 827–834, Apr. 2024, doi: 10.33998/jakakom.2024.4.1.1624.
- [9] A. N. Kasanah, M. Muladi, and U. Pujiyanto, "Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 3, no. 2, pp. 196–201, Aug. 2019, doi: 10.29207/resti.v3i2.945.
- [10] M. A. Rahman Wahid, A. Nugroho, and A. Halim Anshor, "Prediksi Penyakit Kanker Paru-Paru Dengan Algoritma Regresi Linier," *Bulletin of Information Technology (BIT)*, vol. 4, no. 1, pp. 63–74, Mar. 2023, doi: 10.47065/bit.v4i1.501.
- [11] R. Siringoringo, "KLASIFIKASI DATA TIDAK SEIMBANG MENGGUNAKAN ALGORITMA SMOTE DAN k-NEAREST NEIGHBOR," 2018.
- [12] R. Tamami, D. Yanti, and R. Anugrah, "ANALISIS KEPUASAN PELANGGAN MENGGUNAKAN ALGORITMA C4.5 PADA APLIKASI BETUKANG.ID," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 8, no. 6, pp. 12507–12511, Nov. 2024, doi: 10.36040/jati.v8i6.11937.
- [13] X. Chen, "Design and research of MOOC teaching system based on TG-C4.5 algorithm," *Systems and Soft Computing*, vol. 5, p. 200064, Dec. 2023, doi: 10.1016/j.sasc.2023.200064.