 ISSN:2302-8149
e-ISSN:2540-9719

Klasifikasi Kanker Paru-Paru menggunakan Metode *Naïve Bayes* dengan SMOTE

Lung Cancer Classification using the Naïve Bayes Method with SMOTE

[1]Ananda Ikhwana Khairur Akbar, [2]Yani Parti Astuti*
[1,2]Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro
[1,2]Jl. Imam Bonjol 207 Semarang Jawa Tengah Indonesia
*e-mail: 111202113476@mhs.dinus.ac.id, yanipartiastuti@dsn.dinus.ac.id

(***received***: 24 July 2025, ***revised***: 1 September 2025, ***accepted***: 3 September 2025)

Abstrak

Permasalahan utama dalam penelitian ini mencakup keterlambatan deteksi dini kanker paru-paru akibat gejala awal yang tidak spesifik, keterbatasan algoritma Naive Bayes dalam mengolah data kategorikal seperti gejala, jenis kelamin, dan kebiasaan merokok, serta tantangan ketidakseimbangan kelas pada dataset yang dapat memengaruhi akurasi model. Untuk itu, metode SMOTE digunakan sebagai pendekatan untuk meningkatkan performa klasifikasi. Penelitian ini bertujuan untuk mengimplementasikan algoritma Naive Bayes pada proses klasifikasi kanker paru-paru dan membandingkan performa antara data tidak seimbang dengan data yang telah diseimbangkan menggunakan SMOTE. Pendekatan yang digunakan mencakup tahapan preprocessing data, proses encoding, penerapan SMOTE untuk balancing data, dan klasifikasi menggunakan Naive Bayes. Evaluasi dilakukan pada tiga rasio pembagian data, yaitu 80:20, 70:30, dan 60:40. Hasil penelitian menunjukkan bahwa Penerapan teknik SMOTE menghasilkan peningkatan akurasi yang bervariasi di setiap rasio pembagian data. Peningkatan paling signifikan terlihat pada rasio 60:40, di mana akurasi model naik dari akurasi 88.29% dengan kelas 'Yes', model memiliki precision 0.96, recall 0.91, dan F1-score 0.93, sementara untuk kelas 'No', precision adalah 0.40, recall 0.60, dan F1-score 0.48 menjadi 93.19% dengan kelas 'Yes', model memiliki precision 0.96, recall 0.91, dan F1-score 0.93, sementara untuk kelas 'No', precision adalah 0.90, recall 0.96, dan F1-score 0.93. Sebaliknya, pada rasio 80:20 dan 70:30 justru terjadi sedikit penurunan setelah diterapkan SMOTE. Penelitian ini menunjukkan bahwa SMOTE memberikan peningkatan signifikan pada rasio 60:40, tidak hanya dari sisi akurasi, tetapi juga pada metrik Recall dan F1-score yang berperan penting dalam meminimalkan False Negatives pada kelas minoritas bagian kelas yes. Hal ini sangat krusial dalam konteks deteksi dini, karena keberhasilan mendeteksi kasus kanker yang sebenarnya lebih bermakna dibanding hanya mempertahankan akurasi keseluruhan. Meskipun pada rasio lain akurasi tidak selalu meningkat, SMOTE tetap berkontribusi pada perbaikan deteksi kasus kanker, sehingga penerapannya perlu dipertimbangkan dengan memperhatikan keseimbangan antara akurasi dan metrik yang lebih bermakna secara medis.

Kata kunci: machine learning, naive bayes, kanker paru, klasifikasi, evaluasi model

Abstract

The primary challenges addressed in this study include delays in the early detection of lung cancer due to non-specific initial symptoms, the limitations of the Naïve Bayes algorithm in processing categorical data such as symptoms, gender, and smoking habits, as well as class imbalance issues in the dataset that can affect model accuracy. To overcome these challenges, the SMOTE (Synthetic Minority Over-sampling Technique) method was applied to improve classification performance. This study aims to implement the Naïve Bayes algorithm for lung cancer classification and compare its performance on imbalanced data versus data balanced using SMOTE. The methodology consists of data preprocessing, encoding, applying SMOTE for balancing, and classification using Naïve Bayes. Evaluation was performed using three data split ratios: 80:20, 70:30, and 60:40. The results show that applying SMOTE led to performance improvements, with the most significant gains observed at the 60:40 split ratio. In this case, model accuracy improved from 88.29% to 93.19%. For the "Yes" (positive) class, precision remained at 0.96, recall at 0.91, and F1-score at 0.93. However, for the

 ISSN:2302-8149
e-ISSN:2540-9719

"No" (negative) class, precision improved from 0.40 to 0.90, recall from 0.60 to 0.96, and F1-score from 0.48 to 0.93. Conversely, slight decreases in accuracy were observed for the 80:20 and 70:30 ratios after SMOTE application. These findings demonstrate that SMOTE significantly enhances model performance at the 60:40 ratio, not only in terms of accuracy but also in recall and F1-score, which are crucial for reducing false negatives in the minority ("Yes") class. This is especially critical in early detection, as correctly identifying actual cancer cases is more important than merely maintaining overall accuracy. Although SMOTE did not always improve accuracy at other ratios, it still contributed to better cancer case detection. Therefore, its application should be considered carefully, balancing overall accuracy with clinically meaningful metrics.

Keywords: *machine learning, naive bayes, lung cancer, classification, model evaluation*

1 Pendahuluan

Kanker paru-paru merupakan salah satu jenis kanker paling mematikan di Indonesia [1]. Data dari Globocan menunjukkan bahwa kanker paru-paru menyumbang sekitar 34.783 kasus baru dengan 30.843 kematian per tahun di Indonesia, menjadikannya sebagai penyebab kematian akibat kanker tertinggi di tanah air [2]. Tingginya angka kematian disebabkan oleh deteksi yang sering terlambat, karena gejala awal kanker paru-paru sering kali tidak spesifik. Untuk mengatasi permasalahan tersebut, pendekatan berbasis data seperti machine learning mulai banyak digunakan dalam bidang medis, khususnya untuk membantu proses klasifikasi dan prediksi penyakit [3].

Namun, tantangan besar muncul pada data medis yang umumnya tidak seimbang, di mana jumlah pasien dengan hasil negatif kanker jauh lebih sedikit dibandingkan dengan pasien positif. Kondisi ini dapat membuat model klasifikasi menjadi bias terhadap kelas mayoritas. Untuk itu, digunakan metode *Synthetic Minority Over-sampling Technique* (SMOTE) yang dapat menyeimbangkan distribusi kelas dengan membangkitkan data sintetis pada kelas minoritas [4]. Pada penelitian ini dipilih algoritma Naive Bayes, yaitu metode klasifikasi probabilistik berbasis Teorema Bayes yang mengasumsikan independensi antar fitur [5]. Algoritma ini dikenal sederhana namun efektif, terutama untuk dataset kategorikal seperti gejala atau riwayat penyakit [6][7], serta memiliki keunggulan dalam hal komputasi yang ringan dan kemudahan implementasi.

Penelitian ini juga menyoroti celah penelitian terkait pemanfaatan kombinasi metode. Sebagai contoh, studi Almeyda et al. menggunakan algoritma k-Nearest Neighbor (k-NN) dalam klasifikasi penyebab kanker paru-paru dengan hasil akurasi yang baik [8]. Berbeda dengan penelitian tersebut, studi ini berfokus pada integrasi Naive Bayes dengan teknik SMOTE untuk mengeksplorasi bagaimana resampling data dapat meningkatkan efektivitas model klasifikasi. Tujuan penelitian ini adalah untuk menerapkan algoritma Naive Bayes dalam klasifikasi kanker paru-paru berdasarkan data medis kategorikal, memanfaatkan metode SMOTE guna mengatasi ketidakseimbangan kelas pada dataset, mengevaluasi efektivitas kombinasi kedua metode tersebut dalam meningkatkan akurasi, precision, recall, dan F1-score.

2 Tinjauan Literatur

Machine Learning adalah teknik berbasis komputasi yang digunakan untuk membangun model prediktif atau pengambilan keputusan berdasarkan data, dan merupakan bagian integral dari bidang Artificial Intelligence (AI) [9]. Machine Learning memiliki banyak algoritma seperti Suport Vector Machine,Random Forest,KNN dan Naive Bayes. Beberapa penelitian telah memanfaatkan algoritma machine learning untuk melakukan klasifikasi dan prediksi dalam bidang kesehatan, terutama dalam mendeteksi penyakit seperti kanker paru-paru

Berdasarkan Penelitian oleh Almeyda et al. [8] mengkaji penerapan algoritma K-Nearest Neighbor (K-NN) dalam klasifikasi penyebab kanker paru-paru. Hasil pengujian menunjukkan bahwa K-NN mampu menghasilkan akurasi sebesar 90,32%, dengan precision sebesar 96% pada kelas "YES" dan recall sebesar 92%, yang mencerminkan kinerja yang baik dalam klasifikasi penyakit. Penelitian ini juga menegaskan bahwa K-NN cocok digunakan dalam pengolahan data medis karena kesederhanaannya serta kemampuannya menangani data numerik maupun kategorikal secara efektif.

Penelitian yang dilakukan oleh Juliani dan Soleh [10] mengeksplorasi penerapan algoritma Naïve Bayes dalam klasifikasi kasus kanker paru-paru yang diintegrasikan dengan sistem chatbot berbasis

 ISSN:2302-8149
e-ISSN:2540-9719

Natural Language Processing (NLP) menggunakan data pasien. Temuan penelitian ini menunjukkan bahwa metode tersebut berhasil mencapai akurasi hingga 81%. Dengan menggabungkan algoritma klasifikasi dan interaksi chatbot, sistem ini dirancang untuk memberikan informasi pendahuluan secara efisien dalam mendukung deteksi awal kanker paru-paru.

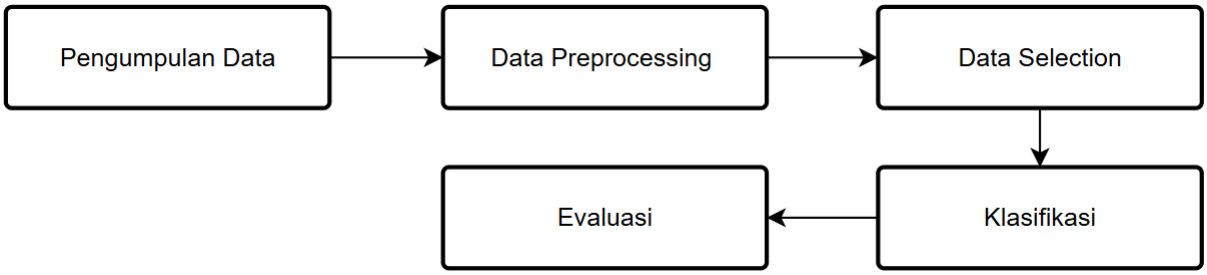
Sementara itu, penelitian oleh Dhini Septhya dan tim [11] berfokus pada penerapan algoritma Decision Tree dan Support Vector Machine (SVM) untuk klasifikasi penyakit kanker paru-paru. Penelitian ini menggunakan dataset kanker paru-paru yang diperoleh dari sumber terbuka dan melakukan proses pra-pemrosesan data, seleksi fitur dengan metode forward selection, serta pembagian data menggunakan rasio 60:40, 70:30, dan 80:20 untuk pelatihan dan pengujian. Tujuan utamanya adalah untuk membandingkan performa dari kedua algoritma tersebut dalam hal akurasi. Hasil dari penelitian tersebut menunjukkan bahwa algoritma SVM dengan forward selection menghasilkan tingkat akurasi terbaik sebesar 62,3% pada rasio 80:20

Perkembangan penelitian terkait penerapan algoritma machine learning dalam klasifikasi kanker paru-paru menunjukkan pendekatan yang beragam. Almeyda et al. [8] mengkaji algoritma K-Nearest Neighbor (K-NN) dan memperoleh hasil akurasi tinggi sebesar 90,32% dengan nilai precision dan recall yang baik. Penelitian ini menegaskan bahwa K-NN cukup andal dalam pengolahan data medis, meskipun fokus utamanya terbatas pada evaluasi satu algoritma tanpa eksplorasi strategi penanganan ketidakseimbangan data. Di sisi lain, Juliani dan Soleh [10] mengeksplorasi penggunaan algoritma Naïve Bayes yang diintegrasikan dengan sistem chatbot berbasis NLP, menghasilkan akurasi sebesar 81%. Kontribusi utama penelitian ini terletak pada inovasi interaktif melalui chatbot, namun belum menitikberatkan pada optimasi performa model klasifikasi itu sendiri. Sementara itu, penelitian oleh Dhini Septhya et al. [11] menggunakan Decision Tree dan Support Vector Machine (SVM) dengan teknik *forward selection* untuk seleksi fitur. Hasilnya menunjukkan bahwa SVM hanya mencapai akurasi 62,3%, menandakan keterbatasan algoritma tersebut dalam memberikan prediksi yang andal pada kasus medis.

Berdasarkan perkembangan tersebut, terlihat bahwa meskipun beberapa algoritma machine learning telah diuji untuk klasifikasi kanker paru-paru, tantangan utama seperti ketidakseimbangan data dan pemilihan metode klasifikasi yang tepat masih menjadi masalah. Oleh karena itu, penelitian ini menawarkan pendekatan berbeda dengan mengintegrasikan algoritma Naïve Bayes dan teknik SMOTE untuk mengatasi isu *class imbalance* sekaligus mengevaluasi sejauh mana kombinasi ini dapat meningkatkan efektivitas klasifikasi kanker paru-paru

3 Metode Penelitian

Untuk Tahapan-tahapan yang akan dilakukan dalam penelitian ini disajikan pada Gambar 1.



Gambar 1 Metodologi penelitian

3.1 Pengumpulan Data

Data yang digunakan dalam penelitian ini merupakan dataset mengenai kanker paru-paru yang diperoleh dari situs Kaggle. Dataset tersebut berformat CSV dan memuat 16 fitur serta 309 data. Setiap kolom merepresentasikan variabel atau atribut yang diamati, sedangkan setiap baris menggambarkan data dari masing-masing responden,dataset tersebut ditunjukan pada Tabel 1.

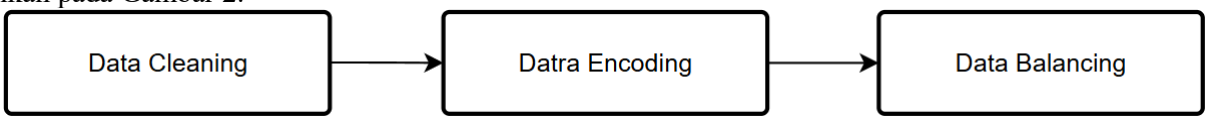
 ISSN:2302-8149
e-ISSN:2540-9719

Tabel 1 Dataset lung cancer

GENDER	AGE	SMOKING	YELLOW FINGERS	…	SHORTNESS OF BREATH	SWALLOWING DIFFICULTY	CHEST PAIN	LUNG CANCER
M	69	1	2	...	2	2	2	YES
M	74	2	1	…	2	2	2	YES
F	59	1	1	…	2	1	2	NO
M	63	2	2	…	1	2	2	NO
F	63	1	2	…	2	1	1	NO
...	...	...	...	…	...	...	...	...
F	56	1	1	…	2	2	1	YES
M	70	2	1	…	2	1	2	YES
M	58	2	1	…	1	1	2	YES
M	67	2	1	…	2	1	2	YES
M	62	1	1	…	1	2	1	YES

3.2 Data Preprocessing

Data preprocessing adalah tahapan awal dalam pengolahan data yang memiliki tujuan untuk menyiapkan data yang belum di olah agar siap digunakan dalam pengujian machine learning. Data yang diperoleh dari sumber eksternal sering kali tidak langsung dapat digunakan karena adanya data yang tidak konsisten, tidak lengkap, atau tidak terstruktur. Oleh karena itu, dilakukan beberapa tahapan pembersihan dan transformasi data agar model yang digunakan dapat menghasilkan hasil klasifikasi yang optimal [12]. Dalam penelitian ini, proses ini dibagi menjadi tiga tahapan, tiga tahapan tersebut adalah Data Cleaning, Data Encoding, dan Data Balancing,Tahapan-tahapan tersebut disajikan pada Gambar 2.



Gambar 2 Data preprocessing

3.2.1 Data Cleaning

Tahapan ini bertujuan untuk menghapus atau memperbaiki data yang tidak valid, duplikat, dan missing value. Pada dataset kanker paru-paru, proses cleaning mencakup pengecekan nilai kosong atau tidak logis. Data yang tidak sesuai akan dibuang atau diperbaiki sesuai aturan yang ditentukan, sehingga dataset menjadi bersih dan konsisten [13].

3.2.2 Data Encoding

Proses encoding dilakukan dengan mengonversi nilai kategori menjadi format numerik yang bersifat kategori dan mengantikan nilai yang dari kolom-kolom agar mudah dipahami dan bisa diproses oleh algoritma machine learning. Sebagai contoh, label target LUNG CANCER diubah menjadi nilai 1 untuk “YES” dan 0 untuk “NO”. Dan contoh target lainnya yaitu SMOKING dengan nilai 2 untuk “YES” dan 1 untuk “NO” diganti menjadi 1 untuk “YES” dan 0 untuk “NO”. Demikian juga pada fitur lain seperti gender atau gejala lainnya, dilakukan encoding untuk merepresentasikan nilai-nilai tersebut dalam bentuk angka. Transformasi ini penting agar model dapat memahami dan memproses fitur-fitur tersebut[14].

3.2.3 Data Balancing

Dataset kanker paru-paru yang digunakan memiliki ketidakseimbangan antara jumlah data positif dan negatif. Ketidakseimbangan ini dapat menyebabkan model bias terhadap kelas mayoritas. Untuk mengatasi hal tersebut, dilakukan teknik data balancing menggunakan SMOTE, yaitu metode yang menghasilkan data sintetis untuk kelas minoritas berdasarkan kemiripan fitur dengan data aslinya. Dengan teknik ini, distribusi antar kelas menjadi lebih seimbang sehingga model dapat belajar secara adil dan memberikan hasil prediksi yang lebih akurat

 ISSN:2302-8149
e-ISSN:2540-9719

3.3 Data Selection

Tahap data selection dilakukan dengan memilih variabel yang relevan dan membagi data menjadi data latih dan data uji. Dalam proses ini, fitur target atau label dipisahkan sebagai variabel Y, sedangkan fitur-fitur lain seperti usia, jenis kelamin, dan gejala-gejala lainnya dijadikan variabel X. Selanjutnya, data dibagi dengan rasio umum seperti *70:30, 80:20, dan 60:40* di mana sebagian besar data digunakan untuk dilatih dan diuji. Pembagian ini penting untuk memastikan model dapat belajar dari data yang cukup dan tetap dievaluasi pada data yang tidak dikenalnya.

3.4 Klasifikasi

Setelah variabel target ditentukan dan data dibagi sesuai rasio tertentu untuk keperluan pelatihan dan pengujian, langkah berikutnya adalah mengimplementasikan algoritma klasifikasi menggunakan algoritma CategoricalNB dari pustaka scikit-learn. Algoritma Categorical Naive Bayes dipilih karena dirancang khusus untuk menangani fitur-fitur kategorikal, sehingga sangat sesuai digunakan pada dataset seperti gejala atau riwayat medis yang memiliki tipe data diskrit. Proses ini mencakup pelatihan model pada data training serta evaluasi kinerjanya menggunakan data testing dengan imbalance data dan SMOTE. Implementasi ini merupakan bagian utama dari eksperimen untuk menilai kemampuan model dalam melakukan prediksi akurat terhadap data kanker paru-paru yang telah melalui tahap preprocessing.Secara matematis, rumus dasar Teorema Bayes dituliskan pada persamaan (1).

$$P(C \mid X) = \frac{P(X|C) \cdot P(C)}{P(X)} \quad (1)$$

Dimana:

P(C| X) :Probabilitas hipotesis C (kelas) diberikan data X (fitur),
P(X|C) :Probabilitas fitur X muncul jika diketahui kelas C,
P(C) :Probabilitas awal dari kelas C,
P(X) :Probabilitas dari data X secara umum.

3.5 Evaluasi

Evaluasi model dilakukan setelah data preprocessing dan data selection untuk data pelatihan dan data pengujian selesai untuk mengetahui sejauh mana model mampu melakukan klasifikasi dengan benar. Salah satu metode evaluasi yang banyak digunakan yaitu confusion matrix, yang memberikan informasi mengenai jumlah prediksi yang benar maupun salah berdasarkan kategori tertentu [15]. Berikut adalah susunan Tabel Confusion Matrix yang dapat dilihat pada Tabel 2

Tabel 2 Confusion matrix

Aktual	Prediksi	
	Positif	Negatif
Positif	*TP (True Positive)*	*FP (False Positive)*
Negatif	*FN (False Negative)*	*TN (True Negative)*

Dari matriks *Confusion Matrix* ini, diperoleh metrik seperti akurasi, presisi, recall, dan F1-score, yang sangat berguna dalam menilai kinerja dan keandalan model Berikut ini adalah Persamaan akurasi, presisi, recall, dan F1-score dapat dilihat pada persamaan (2),(3),(4),dan (5).

$$Accuracy = \frac{TP+TN}{TP+FN+FP+FN} \cdot 100\% \quad (2)$$

$$Precission = \frac{TP}{TP+FP} \cdot 100\% \quad (3)$$

 ISSN:2302-8149
e-ISSN:2540-9719

$$Recall = \frac{TP}{TP+FN} \cdot 100\% \quad (4)$$

$$F1 - Score = 2 \cdot \frac{(Precission \cdot Recall)}{(Precission \cdot Recall)} \cdot 100\% \quad (5)$$

4 Hasil dan Pembahasan

Hasil dari metode penelitian yang telah melalui tahapan pengumpulan data,preprocessing,data selection,klasifikasi,dan evaluasi confusion matrix, disajikan pada bagian ini sebagai bentuk keluaran akhir dari proses analisis. Seluruh langkah tersebut dilakukan secara sistematis untuk memastikan bahwa data yang digunakan telah siap dan layak untuk dievaluasi lebih lanjut menggunakan algoritma Naive Bayes yang akan diterapkan dalam penelitian ini.

4.1 Data Preprocessing

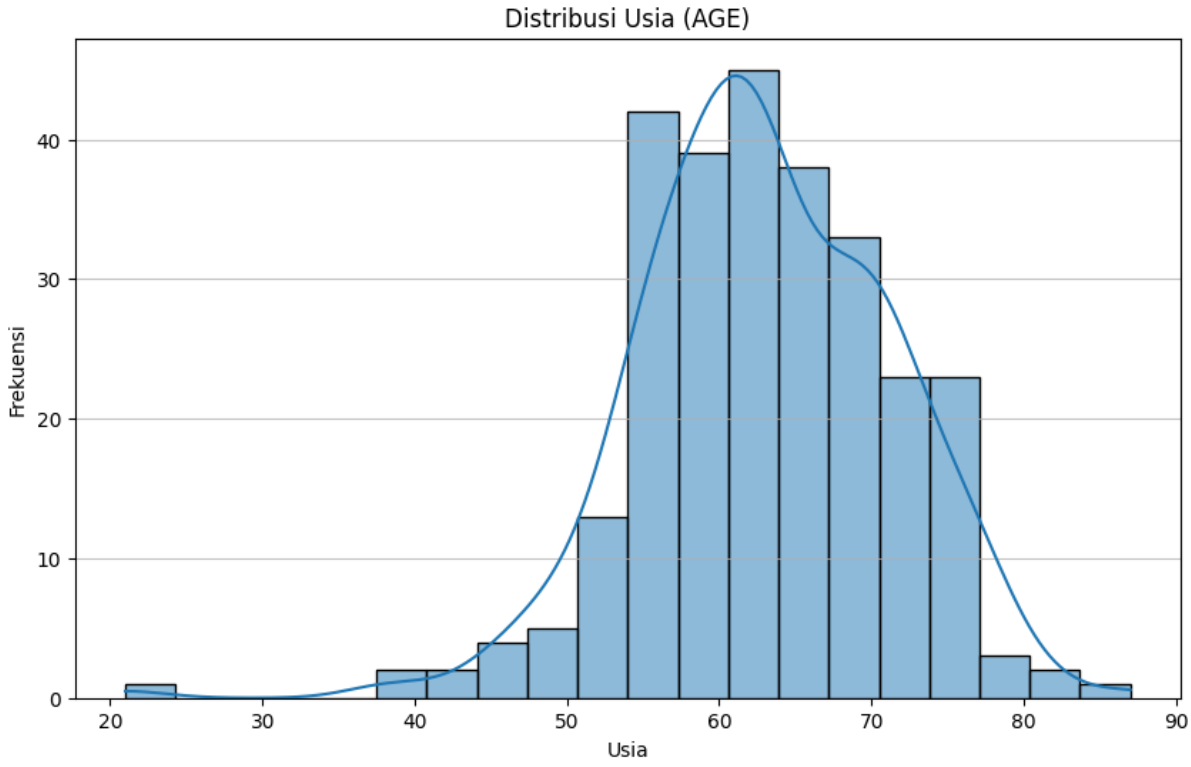
Setelah melakukan pengumpulan data di Kaggle selesai , tahapan selanjutnya adalah preprocessing data, yang berperan penting dalam menghasilkan performa klasifikasi yang optimal. Pada tahap ini, dilakukan serangkaian proses persiapan data sebelum diterapkan ke dalam model machine learning. Secara umum, data preprocessing terdiri dari tiga langkah utama berikut:

4.1.1 Data Cleaning

Pada tahap awal Preprocessing ini, dataset dimuat dan dilakukan pemeriksaan untuk missing values serta duplicate rows. Hasil menunjukkan bahwa tidak ada missing values dalam dataset. Namun, terdeteksi sebanyak 33 baris duplikat yang dapat mempengaruhi analisis. Baris-baris duplikat ini kemudian berhasil dihapus, menghasilkan data baru dengan 276 baris data unik yang siap untuk tahap pemrosesan selanjutnya.

4.1.2 Data Encoding

Setelah melakukan data cleaning, langkah awal dalam proses ini dimulai dengan pemeriksaan fitur umur, yang kemudian digunakan untuk mengklasifikasikan data usia ke dalam kategori tertentu agar bisa di proses oleh model klasifikasinya.



Gambar 3 Distribusi fitur AGE

Gambar 3 menunjukkan bahwa usia termuda pada fitur tersebut berada di atas 20 tahun. Oleh karena itu, data numerik pada fitur AGE dikonversi menjadi data kategorikal berdasarkan rentang

 ISSN:2302-8149
e-ISSN:2540-9719

usia, yaitu kategori dewasa (20–59 tahun) dan lansia (60 tahun ke atas).Hasil dari normalisasi fitur age dapat dilihat pada Tabel 3.

Tabel 3 Normalisasi fitur AGE

GENDER	AGE	SMOKING	YELLOW FINGERS	…	SHORTNESS OF BREATH	SWALLOWING DIFFICULTY	CHEST PAIN	LUNG_CANCER
M	Lansia	1	2	…	2	2	2	YES
M	Lansia	2	1	…	2	2	2	YES
F	Dewasa	1	1	…	2	1	2	NO
M	Lansia	2	2	…	1	2	2	NO
F	Lansia	1	2	…	2	1	1	NO
…	...	...	...	…	...	...	...	...
F	Dewasa	1	2	…	1	2	1	YES
F	Dewasa	2	1	…	2	1	1	NO
M	Dewasa	2	1	…	2	1	2	NO
M	Dewasa	1	2	…	1	2	2	NO
M	Lansia	1	2	…	2	2	2	YES

Setelah Mengubah data pada fitur AGE,langkah ini mengubah data yang ada di semua fitur menjadi format numerik yang dapat diproses oleh algoritma machine learning. Menggunakan LabelEncoder dari pustaka scikit-learn, nilai-nilai unik dalam fitur 'GENDER' (M & F) dan 'LUNG_CANCER' (YES & NO) diubah menjadi representasi numerik (0 dan 1), dan untuk age akan diubah secara manual Dewasa = 0 , Lansia = 1.Hasil Encoding pada dataset dapat dilihat pada Tabel 4.

Tabel 4 Dataset setelah encoding

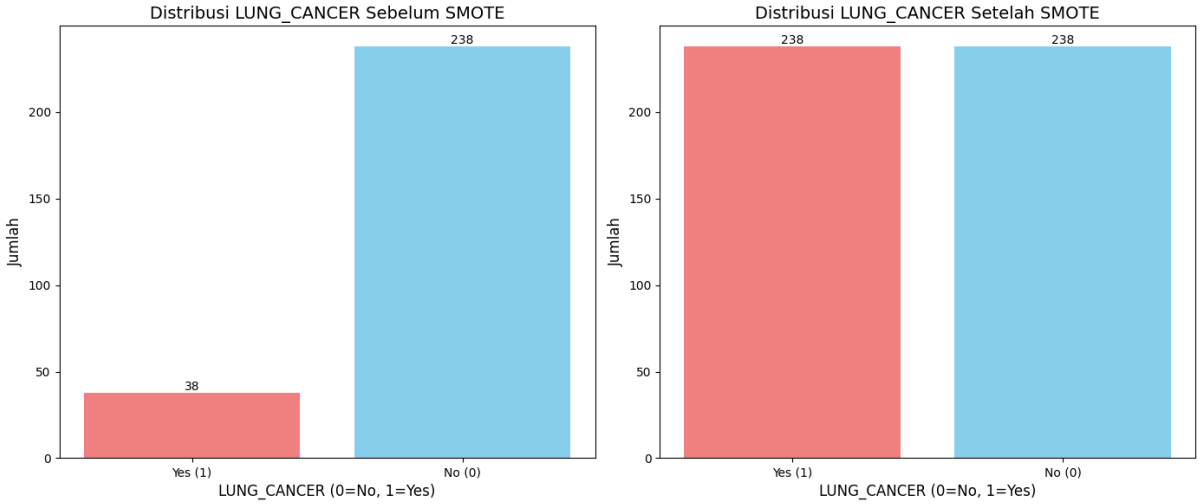
GENDER	AGE	SMOKING	YELLOW FINGERS	…	SHORTNESS OF BREATH	SWALLOWING DIFFICULTY	CHEST PAIN	LUNG_CANCER
1	1	0	1	…	1	1	1	1
1	1	1	0	…	1	1	1	1
0	0	0	0	…	1	0	1	0
1	1	1	1	…	0	1	1	0
0	1	0	1	…	1	0	0	0
...	...	...	...	…	...	...	...	...
0	0	0	1	…	0	1	0	1
0	0	1	0	…	1	0	0	0
1	0	1	0	…	1	0	1	0
1	0	0	1	…	0	1	1	0
1	1	0	1	…	1	1	1	1

4.1.3 Data Balancing

Setelah missing values dan duplicate rows dan encoding telah ditangani, distribusi kelas pada kolom target 'LUNG_CANCER' diperiksa. Terlihat bahwa jumlah sampel untuk kelas 'YES' (238) jauh lebih banyak dibandingkan dengan kelas 'NO' (38), menunjukkan adanya ketidakseimbangan kelas. Untuk mengatasi ini, teknik oversampling SMOTE diterapkan pada data training. SMOTE menghasilkan sampel sintetis untuk kelas 'NO', sehingga jumlah sampel untuk kedua kelas menjadi seimbang, masing-masing 238. Hasil penyeimbangan ini divisualisasikan dengan grafik batang yang menunjukkan jumlah sampel yang merata untuk kedua kelas.Pada Gambar 4 menunjukan hasil grafik

 ISSN:2302-8149
e-ISSN:2540-9719

batang Perbandingan sebelum balancing dan sesudah balancing dan Tabel 5 menunjukan hasil dataset setelah balancing



Gambar 4 Perbandingan sebelum data balancing dan sesudah data balancing

Tabel 5 Dataset setelah balancing

GENDER	AGE	SMOKING	YELLOW FINGERS	…	SHORTNESS OF BREATH	SWALLOWING DIFFICULTY	CHEST PAIN	LUNG_CANCER
1	1	0	1	…	1	1	1	1
1	1	1	0	…	1	1	1	1
0	0	0	0	…	1	0	1	0
1	1	1	1	…	0	1	1	0
0	1	0	1	…	1	0	0	0
…	…	…	…	…	…	…	…	…
0	0	1	0	…	1	0	0	0
1	1	0	0	…	0	0	0	0
1	0	0	1	…	0	1	0	0
1	0	0	1	…	0	1	1	0
0	0	1	1	...	0	0	0	0

4.2 Data Selection

Tahap ini memisahkan dataset menjadi dua bagian utama,yaitu fitur dan target. Kolom 'LUNG_CANCER' yang merupakan variabel target dipisahkan menjadi variabel Y, sedangkan semua kolom lainnya yang merupakan fitur-fitur yang akan digunakan untuk memprediksi target disimpan dalam variabel X. Setelah memisahkan variabel X dan Y, data selanjutnya dibagi menjadi dua subset dengan rasio 80:20 untuk proses pelatihan dan pengujian model.grafik perbandingan kelas yes dan no dengan data ratio 80:20 dapat dilihat pada Gambar 5.

 ISSN:2302-8149
e-ISSN:2540-9719

Gambar 5 Perbandingan jumlah kelas yes dan no pada data rasio 80:20

4.3 Klasifikasi Naive Bayes

Bagian ini membahas penerapan proses klasifikasi terhadap dua kondisi data, yaitu data asli yang memiliki distribusi kelas tidak seimbang dan data yang telah melalui proses penyeimbangan menggunakan metode SMOTE, guna mengevaluasi perbedaan kinerja model pada kedua skenario tersebut.

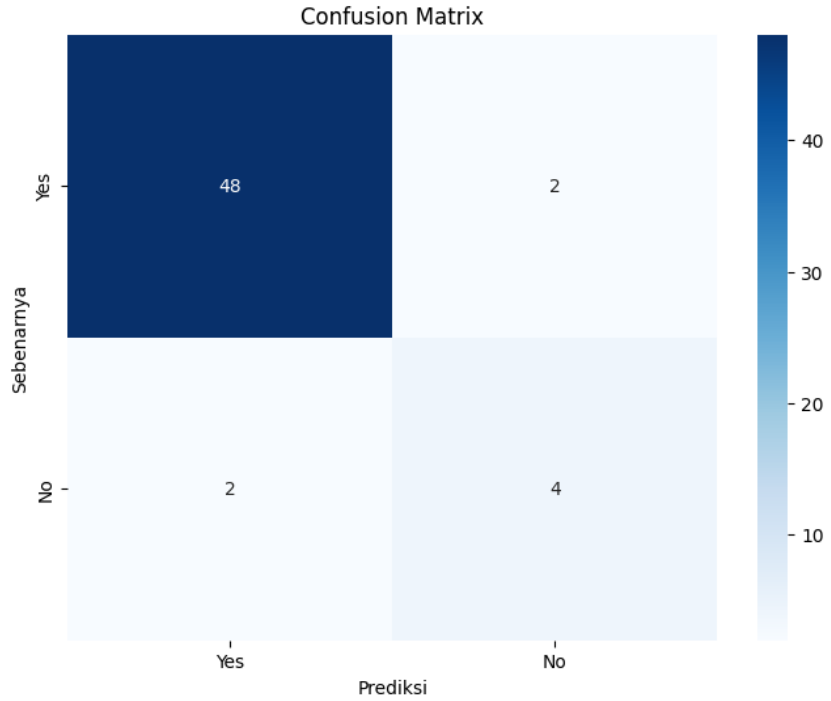
4.3.1 Imbalance Data

Pada bagian ini, dilakukan proses klasifikasi pada Imbalance Data dengan rasio 80:20 untuk mengidentifikasi serta prediksi mengelompokkan data ke dalam dua kategori: Lung Cancer dan Non-Lung Cancer. Tabel 6 dan Gambar 6 di bawah ini menyajikan hasil dari proses klasifikasi tersebut, yang menunjukkan tingkat akurasi model dan confusion matrix.

Tabel 6 Hasil klasifikasi dengan imbalance data pada rasio 80:20

	Precision	Recall	F1-Score	Support
1	0.96	0.96	0.96	50
0	0.67	0.67	0.67	6
Accuracy			0.93	56
Macro Avg	0.81	0.81	0.81	56
Weighted Avg	0.93	0.93	0.93	56

 ISSN:2302-8149
e-ISSN:2540-9719



Gambar 6 Confusion matrix imbalance data pada data rasio 80:20

Hasil evaluasi model Naive Bayes pada data *testing* dengan rasio 80:20 menunjukkan akurasi keseluruhan sebesar 0.9286 yang dibulatkan menjadi dua angka desimal menjadi 0.93. Analisis *confusion matrix* merinci kinerja ini dengan 48 True Positives (TP), 2 False Positives (FP), 2 False Negatives (FN), dan 4 True Negatives (TN). *Classification report* lebih lanjut menguraikan bahwa untuk kelas 'Yes' (LUNG_CANCER=1), model mencapai precision 0.96, recall 0.96, dan F1-score 0.96. Sementara itu, untuk kelas 'No' (LUNG_CANCER=0), Menunjukan bahwa precision, recall, dan F1-score menghasilkan nilai yang sama yaitu 0.67. Hasil ini mengindikasikan model CategoricalNB mempunyai kemampuan yang baik dalam memprediksi kasus positif (kanker paru-paru), namun kinerjanya pada kelas minoritas negatif masih bisa ditingkatkan.

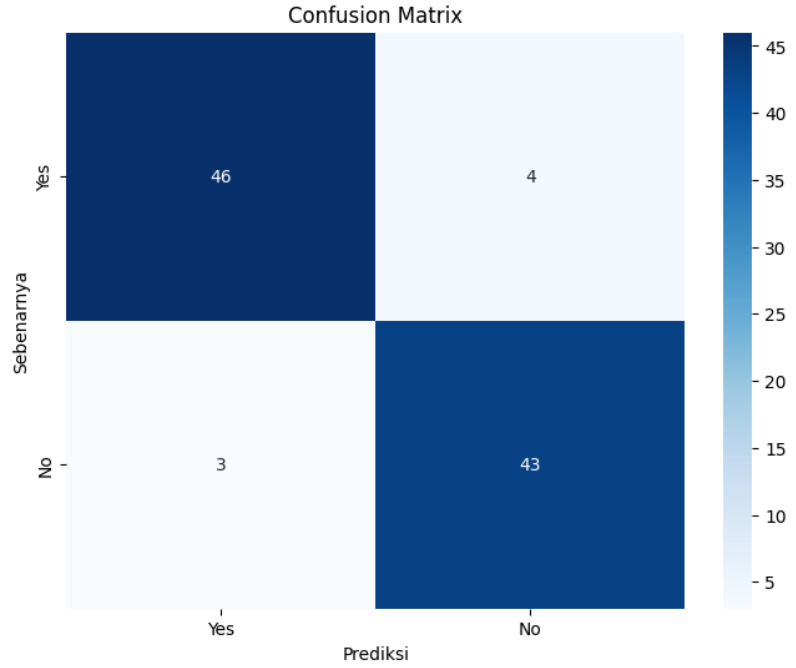
4.3.2 Klasifikasi SMOTE

Pada bagian ini, juga dilakukan proses klasifikasi pada data yang telah di balancing menggunakan metode SMOTE dengan rasio 80:20 untuk mengidentifikasi serta prediksi mengelompokkan data ke dalam dua kategori: Lung Cancer dan Non-Lung Cancer. Tabel 7 dan Gambar 7 di bawah ini menyajikan hasil dari proses klasifikasi tersebut, yang menunjukkan tingkat akurasi model dan confusion matrix.

Tabel 7 Hasil klasifikasi dengan smote pada rasio 80:20

	Precision	Recall	F1-Score	Support
1	0.94	0.92	0.93	50
0	0.91	0.93	0.92	46
Accuracy			0.93	96
Macro Avg	0.93	0.93	0.93	96
Weighted Avg	0.93	0.93	0.93	96

 ISSN:2302-8149
e-ISSN:2540-9719



Gambar 7 Confusion matrix smote pada data rasio 80:20

Berdasarkan Hasil evaluasi model Naive Bayes pada data testing dengan rasio 80:20, didapatkan akurasi sebesar 0.9271 dan dibulatkan menjadi dua angka desimal menjadi 0.93. Confusion matrix menunjukkan 46 True Positives (TP), 3 False Positives (FP), 4 False Negatives (FN), dan 43 True Negatives (TN). Dari Classification Report, terlihat bahwa untuk kelas 'Yes' (1), model memiliki precision 0.94, recall 0.92, dan F1-score 0.93, sementara untuk kelas 'No' (0), precision adalah 0.91, recall 0.93, dan F1-score 0.92. meskipun akurasi sedikit menurun menjadi 0.9271,tetapi performa model terhadap kedua kelas menjadi jauh lebih seimbang, ditunjukkan oleh F1-Score yang hampir setara antara kelas 1 (0,93) dan kelas 0 (0,92). Peningkatan nilai rata-rata makro dan berbobot (masing-masing menjadi 0,93) mengindikasikan bahwa SMOTE berhasil meningkatkan keadilan klasifikasi antar kelas tanpa mengorbankan akurasi secara signifikan.

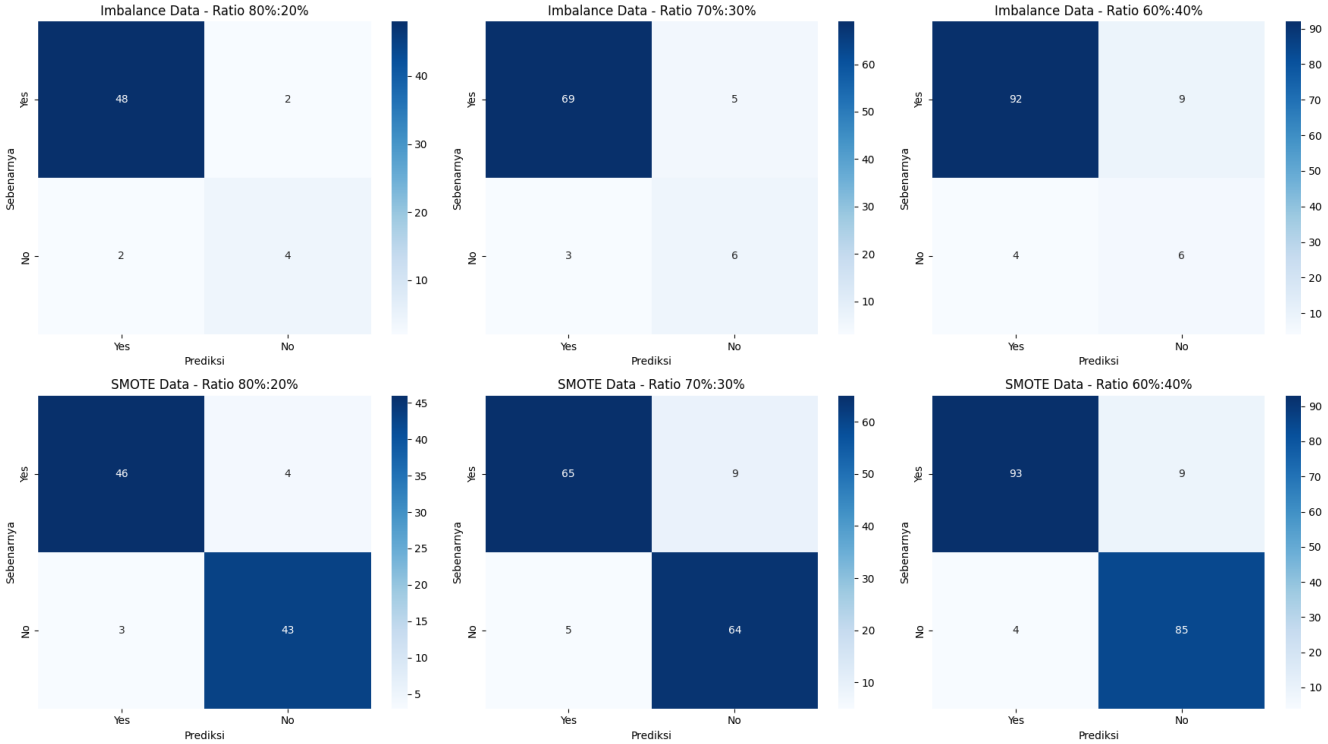
4.4 Evaluasi dan perbandingan

Tahap ini mengevaluasi hasil kinerja klasifikasi algoritma Naive Bayes dengan membandingkan hasil pada data yang tidak seimbang dan data yang telah diseimbangkan menggunakan metode SMOTE, berdasarkan variasi rasio pembagian data latih dan uji, di antaranya 80:20, 70:30, dan 60:40. Hasil akurasi dan Confusion Matrix dari masing-masing skenario disajikan dalam Tabel 8 dan Gambar 8 berikut.

Tabel 8 Hasil perbandingan akurasi

Rasio Data Training : Testing	Akurasi Imbalance Data	Akurasi SMOTE
80:20	92.86%	92.71%
70:30	90.36%	90.21%
60:40	88.29%	93.19%

 ISSN:2302-8149
e-ISSN:2540-9719



Gambar 8 Confusion matrix imbalance data dan smote pada data rasio 80:20, 70:30, 60:40

Berdasarkan Tabel 8, dapat dilihat perbandingan akurasi model klasifikasi Naive Bayes pada berbagai rasio data pelatihan dan pengujian. Secara umum, akurasi model cenderung stabil baik sebelum maupun sesudah penerapan SMOTE, meskipun peningkatan tidak selalu signifikan. Pada rasio 80:20 dan 70:30, akurasi data sebelum dilakukan balancing sedikit lebih tinggi dibandingkan setelah diseimbangkan. Sebaliknya, pada rasio 60:40 justru terjadi peningkatan akurasi cukup mencolok setelah menggunakan SMOTE, dari 88.29% dengan kelas 'Yes' (1), model memiliki precision 0.96, recall 0.91, dan F1-score 0.93, sementara untuk kelas 'No' (0), precision adalah 0.40, recall 0.60, dan F1-score 0.48 menjadi 93.19% dengan kelas 'Yes' (1), model memiliki precision 0.96, recall 0.91, dan F1-score 0.93, sementara untuk kelas 'No' (0), precision adalah 0.90, recall 0.96, dan F1-score 0.93.

5 Kesimpulan

Penelitian ini menerapkan algoritma Naive Bayes untuk klasifikasi kanker paru-paru dengan membandingkan performa model pada data tidak seimbang dan data yang diseimbangkan menggunakan SMOTE pada berbagai rasio data training dan testing. Hasil penelitian menunjukkan bahwa meskipun akurasi model relatif tinggi pada kedua kondisi, SMOTE memberikan peningkatan lebih signifikan pada rasio data yang lebih kecil (60:40) dengan akurasi naik dari 88,29% menjadi 93,19%. Namun, akurasi saja terbukti menyesatkan pada data tidak seimbang. Oleh karena itu, Recall (sensitivitas) dan F1-score pada kelas minoritas, yaitu penderita kanker, menjadi metrik yang lebih penting karena selain mencerminkan sejauh mana model mampu mendeteksi kasus kanker yang sebenarnya, juga berperan dalam meminimalkan False Negatives yang krusial untuk mendukung deteksi dini agar pasien dapat segera memperoleh perawatan yang tepat waktu. Hal ini sangat krusial dalam medis, di mana keberhasilan deteksi dini pasien kanker jauh lebih bermakna dibanding hanya mempertahankan akurasi keseluruhan. Penelitian ini memiliki keterbatasan, yaitu transformasi variabel usia yang berpotensi menghilangkan detail informasi, penggunaan Naive Bayes kategorikal yang mengasumsikan independensi antar fitur, serta belum diterapkannya validasi silang untuk memperkuat estimasi kinerja. Oleh karena itu, penelitian selanjutnya disarankan menekankan evaluasi menggunakan metrik yang lebih relevan bagi kelas minoritas (kanker), menguji variasi rasio data

 ISSN:2302-8149
e-ISSN:2540-9719

untuk menilai stabilitas SMOTE, menerapkan validasi silang, serta membandingkan dengan metode balancing lain agar hasil yang diperoleh lebih representatif dan benar-benar mendukung peningkatan kualitas deteksi kanker paru-paru.

Referensi

[1] S. Andarini, A. A. Santoso, M. A. Arfiansyah, et al., "*Indonesian Society of Respirology Position Paper on Lung Cancer Control in Indonesia*," *J. Respirologi Indones.*, Vol. 44, No. 4, Dec. 2024, doi: 10.36497/jri.v44i4.884.

[2] A. F. Hamdani, W. Purbaningsih, and W. Y. Nalapraya, "Karakteristik Demografi dan Klinikopatologi Pasien Kanker Paru di RSUD Al−Ihsan," *J. Ris. Kedokt.*, pp. 97–102, Dec. 2023, doi: 10.29313/jrk.v3i2.2959.

[3] R. Prakasha, M. Urs, and S. Babu, "*International Journal of Intelligent Systems and Applications in Engineering Machine Learning Approach for Lung Cancer Detection and Classification-A Comparative Analysis*," Mar. 2024. [Online]. Available: www.ijisae.org

[4] B. Dunn, M. Pierobon, and Q. Wei, "*Automated Classification of Lung Cancer Subtypes using Deep Learning and CT-Scan based Radiomic Analysis*," *Bioengineering*, Vol. 10, No. 6, Jun. 2023, doi: 10.3390/bioengineering10060690.

[5] Suprapto, "*Improvement Naive Bayes menggunakan Forward Selection, Information Gain dan Gain Ratio untuk Penanganan Independensi Fitur*," *J. Sos. dan Teknol.*, Vol. 5, No. 4, 2025.doi: 10.59188/jurnalsostech.v5i4.32084

[6] Q. An, S. Rahman, J. Zhou, and J. J. Kang, "*A Comprehensive Review on Machine Learning in Healthcare Industry: Classification, Restrictions, Opportunities and Challenges*," May 01, 2023, *MDPI*. doi: 10.3390/s23094178.

[7] D. Silviana Halawa and R. Mahyuni, "Implementasi *Naive Bayes* pada Sistem Pakar untuk mendiagnosa Penyakit Kelenjar Limfa (Getah Bening)," Nov. 2024, [Online]. Available: https://ojs.trigunadharma.ac.id/index.php/jsi

[8] H. P. Almeyda, Z. F. Khoiri, M. S. Haris, N. H. Alkaff, and S. Sukmadiningtyas, "*Implementation of K-Nearest Neighbor Algorithm for Classification of Lung Cancer Causes,*" *JURTEKSI (Jurnal Teknol. dan Sist. Informasi)*, Vol. 11, No. 1, pp. 37–44, Dec. 2024, doi: 10.33330/jurteksi.v11i1.3305.

[9] R. Alifahasni Zakiah, S. Wahjuni, and W. B. Suwarno, "Pemilihan *Algoritma Machine Learning* untuk Perangkat dengan Komputasi Terbatas pada Deteksi Kematangan Buah Melon Berjala *Selection of Machine Learning Algorithms for Limited Computing Device in Netted Melon Ripeness Detection*," 2023. [Online]. Available: http://journal.ipb.ac.id/index.php/jika

[10] D. Juliani and M. Soleh, "Implementasi *Machine Learning* untuk Klasifikasi Penyakit Kanker Paru menggunakan Metode *Naïve Bayes* dengan Tambahan Fitur *Chatbot (Implementation of Machine Learning for Lung Cancer Classification using Naïve Bayes Method with Additional Chatbot Features)*," Aug. 2024. [Online]. Available: https://www.kaggle.com/datasets/mysarahmadb

[11] D. Septhya, K. Rahayu, S. Rabbani, V. Fitria, Y. Irawan, and R. Hayami, "*MALCOM: Indonesian Journal of Machine Learning and Computer Science Implementation of Decision Tree Algorithm and Support Vector Machine for Lung Cancer Classification Implementasi Algoritma Decision Tree dan Support Vector Machine untuk Klasifikasi Penyakit Kanker Paru*," Vol. 3, pp. 15–19, 2023. doi:10.57152/malcom.v3i1.591

[12] S. S. Berutu, H. Budiati, J. Jatmika, and F. Gulo, "*Data Preprocessing Approach for Machine Learning-based Sentiment Classification*," *J. INFOTEL*, Vol. 15, No. 4, pp. 317–325, Nov. 2023, doi: 10.20895/infotel.v15i4.1030.

[13] R. Taufik, R. Jimah, and A. Solichin, "Implementasi dan Analisis Model *Machine Learning Decision Tree* untuk Deteksi Akun Palsu di Twitter," *J. MEDIA Inform. BUDIDARMA*, Vol. 8, No. 2, p. 797, Apr. 2024, doi: 10.30865/mib.v8i2.7548.

[14] M. Guntara and F. D. Astuti, "Komparasi Kinerja *Label-Encoding* dengan *One-Hot-Encoding* pada *Algoritma K-Nearest Neighbor* menggunakan Himpunan Data Campuran," *JIKO (Jurnal Inform. dan Komputer)*, Vol. 9, No. 2, p. 352, Jun. 2025, doi: 10.26798/jiko.v9i2.1605.

 ISSN:2302-8149
e-ISSN:2540-9719

[15] M. K. Suryadewiansyah, T. Endra, and E. Tju, "Jurnal Nasional Teknologi dan Sistem Informasi *Naïve Bayes* dan *Confusion Matrix* untuk Efisiensi Analisa *Intrusion Detection System Alert*," Aug. 2022, doi: 10.25077/TEKNOSI.v8i2.2022.081-088.