**Journal of Applied Informatics and Computing (JAIC)**
Vol.9, No.6, December 2025, pp. 3363~3369
e-ISSN: 2548-6861


# Analysis of Naive Bayes Algorithm for Lung Cancer Risk Prediction Based on Lifestyle Factors

**Sheila Anggun Vabilla [1]*, Majid Rahardi [2]***
* Informatika, Fakultas Ilmu Komputer, Universitas Amikom Yogyakarta
sheila@students.amikom.ac.id [1], majid@amikom.ac.id [2]

**Article Info**

***Article history:***

Received 2025-10-08
Revised 2025-11-01
Accepted 2025-11-12

***Keyword:***

*Lung Cancer,*
*Lifestyle,*
*Gaussian Naive Bayes,*
*SMOTE,*
*Model Mutual Information.*

**ABSTRACT**

Lung cancer is one of the types of cancer with the highest mortality rate in the world, which is often difficult to detect in the early stages due to minimal symptoms. This study aims to build a lung cancer risk prediction model based on lifestyle factors using the Gaussian Naive Bayes algorithm. Data fit is addressed using the Synthetic Minority Over-sampling Technique (SMOTE), and feature selection is carried out using the Mutual Information. The dataset used consists of 1000 patient data with 24 features related to lifestyle and environmental factors. Model validation is carried out using 5-fold Stratified Cross Validation, and evaluated based on accuracy, precision, recall, and confusion matrices. The results show that the application of SMOTE successfully increases the model accuracy to 91.00% with high precision and recall values in all risk classes (Low, Medium, High). The features "Passive Smoker" and "Coughing up Blood" are identified as the most influential factors in the prediction. The results of this study indicate that the combination of Gaussian Naive Bayes with SMOTE and Mutual Information is able to produce an accurate prediction model.



## I. PENDAHULUAN

Kanker merupakan salah satu penyebab kematian diseluruh dunia. Diantara banyaknya jenis kanker, kanker paru-paru menjadi salah satu jenis kanker yang paling banyak dialami dan memiliki tingkat kematian yang tinggi [1]. Kanker paru-paru ialah jenis kanker yang terjadi ketika sel di paru-paru berkembang secara tidak normal dan tidak terkontrol [2]. Menurut *American Cancer Society* pada tahun 2022 terdapat pasien sebanyak 236.740 mengidap kanker paru-paru dan setengahnya dari mereka meninggal sebanyak 130.180 [3].

[4] di Indonesia, kanker paru-paru terdapat pada urutan tiga besar jenis kanker terbanyak, dengan kasus yang terus meningkat setiap tahun. Pemeriksaan awal penyakit ini masih menjadi hambatan karena gejala awal kanker paru-paru seperti batuk yang lama sembuh, sesak napas dan nyeri dada [5]. Seringkali penderita mengira gejala yang muncul merupakan tanda penyakit biasa sehingga dianggap tidak perlu melakukan pemeriksaan lebih lanjut [6]. Beberapa penelitian sebelumnya mengungkapkan bahwa pola gaya hidup memiliki peran penting terhadap risiko terjadinya kanker paru-paru. Kebiasaan merokok, terkena paparan asap rokok, konsumsi alkohol, kurangnya aktivitas fisik dan pola makan tidak sehat termasuk faktor utama yang dapat meningkatnya terkena kanker paru-paru [7]. Penelitian oleh Laily dkk. menemukan bahwa usia, jenis kelamin, serta kebiasaan merokok memiliki hubungan yang kuat terhadap risiko terjadinya adenokarsinoma paru-paru [8]. Selain itu, penelitian oleh Suratman menunjukkan bahwa rata-rata pasien wanita penderita kanker paru-paru non- sel kecil (NSCLC) di Indonesia ternyata tidak merokok. Temuan ini menujukkan selain rokok, faktor lingkungan juga sangat berpengaruh [9].

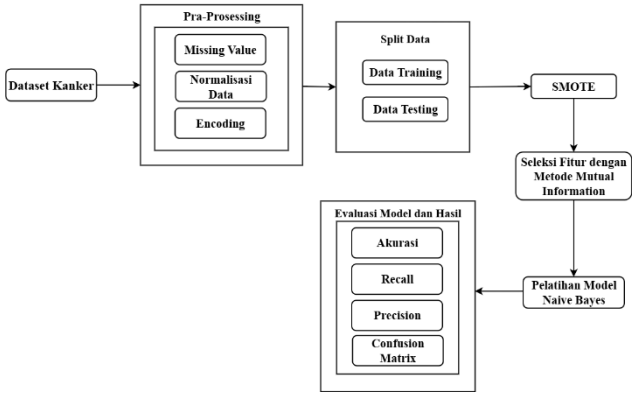
Kanker paru-paru terbagi menjadi dua jenis yaitu, Kanker paru-paru sel kecil (Small Cell Lung Cancer/SCLC) jenis ini lebih jarang terkena kanker paru-paru, sekitar 10-15% dari kasus, tetapi lebih agresif dan cepat menyebar ke bagian tubuh lain. SCLC berkaitan dengan kebiasaan merokok. Serta Kanker paru-paru non sel kecil (Non-Small Cell Lung Cancer/NSCLC) jenis ini merupakan yang paling umum terkena, sekitar 80-87% dari kasus kanker paru-paru. NSCLC terdiri dari beberapa subtipe utama, yaitu adenokarsinoma

(biasanya di bagian luar paru-paru dan sel penghasil lendir), karsinoma sel skuamosa (biasanya di saluran pernapasan dalam paru), dan karsinoma sel besar. NSCLC cenderung tumbuh lebih lambat dan jika terdeteksi dini [10].

Seiring berkembangnya tehnologi, penggunaan machine learning dibidang kesehatan berkembang sangat cepat. Salah satu algoritma yang sering digunakan ialah Naive Bayes, algoritma ini bekerja berdasarkan prinsip probabilitas dan menganggap setiap fitur data saling independen. Kelebihan dari algoritma ini adalah memiliki proses klasifikasi yang cepat dan dapat mengelola data besar secara efektif [11]. [12] menujukkan bahwa prediksi kanker paru-paru menggunakan algoritma Naive Bayes dapat memberikan hasil yang akurat. Hasil penelitian menunjukkan bahwa Naive Bayes merupakan algoritma yang efesien dalam mnegklasifikasi risiko kanker paru-paru. Penelitian oleh Widyawati dkk. menghasilkan akurasi Naive Bayes sebesar 0,89% meskipun hasilnya sudah baik, penelitian ini tidak melakukan teknik penyeimbangan data yang dapat mempengaruhi akurasi pada kelas minoritas dan tidak melakukan validasi silang [13]. Sementara itu, penelitian oleh Shafa dkk. menunjukkan bahwa walaupun adanya penerapan normalisasi dan pemilihan fitur pada tahapan pre-processing, algoritma Naive Bayes mudah mengalami overfitting dan belum membandingkan secara mendalam dengan metode boosting [14]. Selain itu, sebagian besar penelitian terdahulu belum banyak yang menggabungkan teknik balancing seperti SMOTE, serta pemilihan fitur dengan Mutual Information. Oleh karena itu, masih terdapat celah untuk mengembangkan penelitian untuk mengembangkan model prediktif yang memadukan algoritma Naive Bayes dengan cara penyeimbangan data serta interpretasi hasil prediksi secara mendalam. Seperti pada penelitian yang dilakukan oleh Karunia dkk. Yang juga menggunakan metode Mutual Information [15].

Oleh karena itu, penelitian ini bertujuan untuk membangun model prediksi risiko kanker paru-paru berbasis algoritma Naive Bayes dengan menggunakan data faktor pola gaya hidup seperti kebiasaan merokok, paparan lingkungan dan aktivitas fisik. Harapannya, model ini tidak hanya bermanfaat dalam konteks akademik, tetapi juga dapat digunakan untuk mendukung kebijakan kesehatan dan meningkatkan kesadaran masyrakat terhadap pentingnya pola hidup sehat. Serta bertujuan juga untuk mengisi celah tersebut dengan membangun model prediksi resiko kanker paru-paru berbasis Naive Bayes yang menggabungkan dengan metode SMOTE untuk mengatasi ketidakseimbangan data dan Mutual Information untuk memilih fitur yang paling berpengaruh. Dengan menggabungkan berbagai metode, penelitian ini diharapkan dapat menghasilkan model prediksi yang mencapai akurasi tinggi dan performa lebih baik dibandingkan metode-metode yang telah digunakan pada penelitian sebelumnya.

## II. METODE




Gambar 1. Alur Penelitian

### A. *Dataset*

Metode penelitian ini menggunakan Dataset yang berasal dari platform Kaggle (Lung Cancer Dataset) dirilis tahun 2023 dan bersifat open-source, dengan nama Lung Cancer Prediction yang berisi 1000 data pasien dengan 26 fitur. Fitur dalam Dataset ini terdiri dengan berbagai aspek termasuk karakteristik gaya hidup seperti (kebiasaan merokok, obesitas, konsumsi alkohol dan pola makan), gejala klinis yang relevan seperti (sesak napas, nyeri dada, dan batuk berdarah), serta paparan lingkungan seperti (polusi udara dan bahaya pekerjaan). Fitur dalam dataset ini berkaitan dengan kasus kanker paru-paru, sehingga target penelitian ini adalah memprediksi Tingkat Risiko Kanker Paru-Paru yang dikelompokkan ke dalam tiga kelas yaitu Low(rendah), Medium(sedang), High(tinggi).

Dataset yang digunakan terdiri dari 26 fitur yang mencakup faktor gaya hidup, lingkungan, serta gejala klinis pasien. Fitur gaya hidup seperti Merokok dan Perokok pasifr menggambarkakn kebiasaan merokok aktif maupun paparan asap rokok orang lain, yang keduanya mempunyai kaitan kuat dengan peningkatan risiko kanker paru-paru. Sementara fitur Batuk berdarah merupakan gejala klinis yang berpengaruh terhadap nilai Mutual Information dalam prediksi. Selain itu, fitur seperti Usia dan Polusi Udara merupakan faktor lingkungan serta kondisi fisiologis pasien yang juga berkontribusi paa tingkat nrisiko kanker paru-paru.

### B. *Pra-Prosessing*

Setelah dataset kanker dikumpulkan, langkah selanjutnya ialah melakukan pra-prosessing data. Pra-prosessing yaitu membersihkan data dan menyiapkan data sebelum digunakan untuk melatih model. Proses ini bertujuan untuk memastikan apakah data yang digunakan dalam pelatihan model memiliki kualiatas yang baik, konsisten, dan sesuai dengan kebutuhan klasifikasi algoritma. Tahap pra-pemrosesan dilakukan untuk memastikan kualitas data sebelum proses pelatihan model. Proses ini mencangkup pemeriksaan nilai kosong (missing value) yang hasilnya menunjuukkan tidak adanya nilai yang hilang, penghapusan pada kolom yang tidak relevan seperti

"Patient ID" dan "Index", dihapus karena tidak ada kontribusi terhadap prediksi serta seluruh fitur kategorikal yang masih dalam bentuk teks atau label diubah ke bentuk numerik menggunakan teknik encoding. Selain itu, normalisasi seluruh fitur numerik menggunakan metode *StandardsScaler* agar setiap fitur memiliki skala yang sama.

*1. Missing Value*

Penanganan nilai kosong merupakan tahap penting dalam melakukan pemrosesan data. Pada dataset yang diperoleh, peneliti melakukan pemeriksaan secara menyeluruh untuk mendeteksi keberadaaan nilai kosong disetiap kolom. Hasil dari pengecekan menunjukkan bahwa seluruh data bersih dan tidak adanya misiing value. Oleh karena itu, tidak perlu memerlukan proses pengisian nilai mneggunakan rata-rata, modus, atau median. Kondisi ini menjadi kelebihan karena model dapat dilatih dengan data yang utuh dari setiap sampel, sehingga mengurangi kemungkinanna terjadinya risiko bias ataupun kesalahan prediksi akibat data yang hilang.

*2. Normalisasi Data*

Normalisasi data dilakukan untuk meningkatkan kestabilan model dan memastikan bahwa semua fitur numerik memiliki kontribusi yang seimbang selama proses pelatihan. Proses ini menggunakan metode StandardScaler, yaitu tehnik yang dapat mengubah data sehingga memiliki distribusi dengsn rata-rata nol dan standar devisi satu. Langkah ini penting, khususnya untuk algoritma Gaussian Naive Bayes yang menganggap bahwa fitur mengikuti distribusi Gaussian (normal) dan sangat sensitif terhadap distribusi data. Tanpa proses normalisasi, fitur dengan nilai yang lebih besar dapat mengontrol hasil prediksi sehingga dapat mengurangi performa model secara keseluruhan. Eksperimen di jalankan menggunakan *Python 3.10.12* di Google Colab, dengan menggunakan pustaka utama *scikit-learn (1.4.2), imbalanced-learn (0.11.0), matplotib (3.7.1), dan numpy (1.26.4).* Semua dilakukan pada CPU runtime standar Colab tanpa menggunakan akselerasi GPU.

*3. Encoding Variabel Kategorikal*

Pada penelitian yang dilakukan, proses encoding difokuskan pada variabel target yaitu "Tingkat Risiko Kanker", yang awalnya berupa nilai kategori seperti "Low", "Medium", dan "High". Variabel ini perlu diubah menjadi format numerik agar dapat diproses oleh algoritma klasifikasi. Proses dilakukan menggunakan LabelEncoder dari ppustaka sklearn.preprocessing, yang secara otomatis mengonversi setiap label menjadi angka, misalnya "Low" menjadi 0. "Medium" menjadi 1 dan "High" menjadi 2. Langkah ini sangat penting agar model Gaussian Naive Bayes dapat memahami label target dalam bentuk matematis. Penting untuk diperhatikan bahwa tidak ada fitur input lain yang bersifat kategorikal, sehingga encoding hanya diterapkan pada kolom target saja.

*C. Split Data*

Split data merupakan proses machine learning untuk memastikan apakah model dapat dilatih dan diuji secara adil. Dalam penelitian ini, dataset dibagi menjadi dua bagian yaitu data training dan data testing menggunakan train_test_split dari pustaka scikit-learn. Data latih dalam penelitian ini digunakan untuk membangun serta melatih model, sedangkan data testing digunakan untuk menilai sejauh mana model mampu melakukan penyamarataan pada model yang belum pernah ditemui sebelumnya. Pembagian data dilakukan dengan proporsi 80% untuk data training dan 20% untuk data testing. Dengan menggunakan metode stratified split agar distribusi kelas target seimbang dikedua bagian.

*D. Penyeimbangan Data Menggunakan SMOTE*

Dataset yang digunakan dalam penelitian ini memiliki kelas yang tidak seimbang, dengan jumlah data pada setiap kelas (Low, Medium, High) tidak merata. Ketidakseimbangan ini dapat menyebabkan model lebih cenderung ke kelas mayoritas dan mengabaikan kelas minoritas. Untuk mengatasi masalah tersebut, digunakannya metode SMOTE (Synthetic Minority Over-sampling Technique) pada data pelatihan. SMOTE bekerja dengan membuat data sintesis pada kelas minoritas melalui interpolasi antar sampel yang sudah ada, sehingga menghasilkan dataset baru yang lebih seimbang antar kelas. Sebelum dilakukan *oversampling*, distribusi jumlah data pada tiap kelas tidak seimbang, dimana kelas High memiliki 365 data, kelas Medium sebanyak 332, dan kelas Low sebanyanyak 303 data. Setelah diterapkan SMOTE pada data latih, distrubusi data menjadi seimbang untuk semua kelas. Total pada data latih setelah dilakukan SMOTE berjumlah 876 sampel dengan 23 fitur, sebagaimana terlihat pada Gambar 7. Penerapan SMOTE dilakukan setelah pembagian data yaitu menjadi data latih dan data uji, dimana hanya data latih yang dilakukan oversampling. Proses SMOTE menggunakan paramenter *k-neighboars* = 5 dan *sampling strategy* = 'auto', sehingga jumlah sampel pada setiap kelas menjadi seimbang. Pemilihan k=5 didasarkan pada rekomendasi umum untuk dataset yang berukuran sedang agar sintesis data tetap beragam dan menghindari *overfitting*. Proses ini menghasilkan distribusi kelas menjadi seimbang dan memungkinkan model untuk belajar dengan seimbang dari semua kelas dan meningkatkan kemampuan generalisasi pada data uji.

*E. Seleksi Fitur dengan Metode Mutual Information*

Untuk mengetahui fitur-fitur mana yang paling berpengaruh dalam memprediksi risiko kanker paru-paru, dilakukan seleksi fitur menggunakan metode Mutual Information (MI). MI ini berfungsi untuk mengukur tingkat ketergantungan antara setiap fitur independen dengan variabel target, di mana semakin tinggi nilainya, semakin besar kontribusi fitur tersebut terhadap prediksi. Hasil analisis MI menunjukkan bahwa fitur "Perokok Pasif" dan "Batuk Berdarah" memiliki pengaruh yang sangat dominan dalam

model prediksi. MI menampilkan bentuk grafik batang horizontal yang menampilkan 10 fitur yang paling berpengaruh terhadap kanker. Proses ini bertujuan untuk memperkuat pemahaman bahwa fitue-fitur yang dipilih memang memilikin hubungan kuat dengan risiko kanker paru-paru.

### *F. Pelatihan Model Naive Bayes*

Setelah data latih diseimbangkan menggunakan SMOTE, model dikembangkan dengan menggunakan algoritma Gaussian Naive Bayes, yaitu salah satu metode klasifikasi berbasis probabilitas yang mengansumsikan fitur-fitur saling saling independen dan mengikuti distribusi Gaussian (normal). Peneliti memilih menggunakan Algoritma Gaussian Naive Bayes karena efisiensinya yang tinggi serta performa yang baik pada data numerik yang sudah dinormalisasi. Model dilatih dengan menggunakan data hasil SMOTE , dan kemudian divalidasi dengan metode Stratifield 5-Fold Cross Validation untuk memastikan setiap kelas terwakili secara seimbang disetiap fold. Hasil validasi silang 5-Fold menunjukkan rata-rata akurasi sebesar 88.58% dengan *standard deviation* sebesar 1.92%, yang dapat menandakan model mampu mengenali pola dari semua kelas risiko secara konsisten dan stabil.

Alasan pemilihan varian Gaussian Naive Bayes karena seluruh fitur pada dataset bersifat numerik dan telah melalui proses normalisasi menggunakan *StandardScaler.* Varian Gaussian Naive Bayes mengamsumsikan setiap fitur memiliki ditrisbusi normal, sehingga sangat cocok untuk data dengan nilai kontinu. Sementara itu, varian *Multional* dan *Bernoulli* lebih cocok digunakan pada data diskrit atau biner seperti teks atau kategori. Oleh karena itu, pemilihan Gaussian Naive Bayes paling relevan untuk karakteristik data dalam penelitian ini.

### *G. Evaluasi Model dan Hasil*

Evaluasi model ini bertujuan untuk mengukur seberapa efektif algoritma dalam memprediksi tingkat risiko kanker paru-paru berdasarkan data uji dan hasil validasi silang. Model Gaussian Naive Bayes diuji dalam dua tahap, yaitu pada data latih yang sudah diseimbangkan menggunakan SMOTE dan pada data uji. Penilaian dilakukan dengan menggunakan metrik akurasi, precision, recall, f1-score dan confusion matrix. Proses evaluasi ini penting untuk mengevaluasi tidak hanya performa keseluruhan model. Tetapi juga kemampuannya dalam mengklasifikasikan setiap kategori risiko (Low, Medium, High) secara detail. Dengan adanya evaluasi yang menyeluruh, dapat dipastikan bahwa model tidak bias terhadap kelas tertentu dan mampu memberikan hasil yang dapat diandalkan untuk kebutuhan medis.

#### *1. Akurasi*

Akurasi merupakan metrix yang menunjukkan persentase prediksi yang benar dari seluruh data yang diuji. Dalam penelitian ini, model mencapai akurasi sebesar 88.58% pada data latih menggunakan proses Stratified 5-Fold Cross Validation, yang menunjukkan kestabilan dan kemampuan model dalam belajar dari data yang sudah seimbang. Kemudian saat diuji pada data uji, model memperoleh akurasi tinggi sebesar 91.00% yang berarti dari 200 data uji, sebanyak 182 prediksi berhasil dilakukan dengan benar. Akurasi ini menunjukkan kinerja umum secara keseluruhan dan menunjukkan efektivitas kombinasi antara SMOTE dan algoritma Gaussian Naive Bayes.

#### *2. Precision*

Precision ini mengukur ketepatan model dalam memprediksi kelas tertentu, yaitu berapa banyak prediksi positif yang benar-benar akurat. Pada data uji, nilai precision mencapai 1.00 untuk kelas Low, 0.93 untuk kelas Medium, dan 0.84 untuk kelas High. Nilai precision yang tinggi menunjukkan bahwa model memiliki prediksi yang tepat, terutama pada kelas Low dan Medium.

#### *3. Recall*

Recall menunjukkan kemampuan dalam mendeteksi seluruh kasus positif sebenarnya dalam suatu kelas. Pada data uji, recall memperoleh nilai 0.93 untuk kelas Low, 0.82 untuk kelas Medium, dan 0.97 untuk kelas High. Nilai recall yang tinggi pada kelas High menunjukkan bahwa model sangat efektif dalam mengenali pasien dengan risiko tinggi kanker paru-paru, sebingga kecil kemungkinan terjadinya false negative (kasus berisiko tinggi yang tidak terdeteksi) sangat kecil.

#### *4. Confusion Matrix*

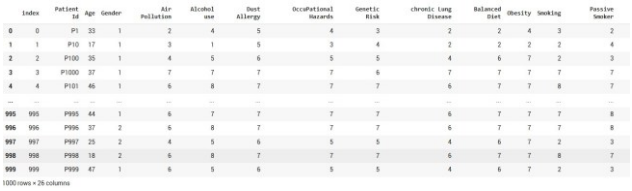
Confusion matrix menampilkan perbandingan antara prediksi model dan label asli dalam bentuk matriks, yang memperlihatkan distribusi prediksi yang benar dan salah unutk setiap kelas. Pada data uji, confusion matrix menunjukkan sebagian besar prediksi berada pada diagonal utama, yang berarti model berhasil mengklasifikasi data ke kelas yang sesuai. Sebagai contoh, dari 73 data yang memiliki label aktual High, sebanyak 71 data berhasil diklasifikasikan dengan tepat oleh model. Penggunaan heatmap untuk visualisasi confusion matrix memudahkan dalam menganalisis kesalahan klasifikasi dan menunjukkan bahwa model menunjukkan performa yang seimbang di seluruh kelas.

## III. HASIL DAN PEMBAHASAN

### *A. Dataset*

Dataset yang digunakan dalam penelitian ini berasal dari platform Kaggle *Lung Cancer Prediction* dan berisikan 1000 data pasien dengan 26 fitur, yang terdiri dari usia, jenis kelamin, polusi udara, konsumsi alkohol, alergi debu, bahaya pekerjaan, risiko genetik, penyakit paru kronis, pola makan

seimbang, obesitas, merokok, perokok pasif, nyeri dada, batuk berdarah, kelelahan, penurunan berat badan, sesak napas, mengi, sulit menelan, pembesaran ujung jari, sering flu, batuk kering, mendekur dan tingkat risiko kanker. Dalam penelitian ini terdapat kolom target "Level" yang menunjukkan tingkatan level risiko terkena kanker. Gambar 2. Tampilan dataset



| | Index | Patient Id | Age | Gender | Air Pollution | Alcohol use | Dust Allergy | OccuPational Hazards | Genetic Risk | chronic Lung Disease | Balanced Diet | Obesity | Smoking | Passive Smoker |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | P1 | 33 | 1 | 2 | 4 | 5 | 4 | 3 | 2 | 2 | 4 | 3 | 2 |
| 1 | 1 | P10 | 17 | 1 | 3 | 1 | 5 | 3 | 4 | 2 | 2 | 2 | 2 | 4 |
| 2 | 2 | P100 | 35 | 1 | 4 | 5 | 6 | 5 | 5 | 4 | 6 | 7 | 2 | 3 |
| 3 | 3 | P1000 | 37 | 1 | 7 | 7 | 7 | 7 | 6 | 7 | 7 | 7 | 7 | 7 |
| 4 | 4 | P101 | 46 | 1 | 6 | 8 | 7 | 7 | 7 | 6 | 7 | 7 | 8 | 7 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 995 | 995 | P995 | 44 | 1 | 6 | 7 | 7 | 7 | 7 | 6 | 7 | 7 | 7 | 8 |
| 996 | 996 | P996 | 37 | 2 | 6 | 8 | 7 | 7 | 7 | 6 | 7 | 7 | 7 | 8 |
| 997 | 997 | P997 | 25 | 2 | 4 | 5 | 6 | 5 | 5 | 4 | 6 | 7 | 2 | 3 |
| 998 | 998 | P998 | 18 | 2 | 6 | 8 | 7 | 7 | 7 | 6 | 7 | 7 | 8 | 7 |
| 999 | 999 | P999 | 47 | 1 | 6 | 5 | 6 | 5 | 5 | 4 | 6 | 7 | 2 | 3 |

1000 rows × 26 columns

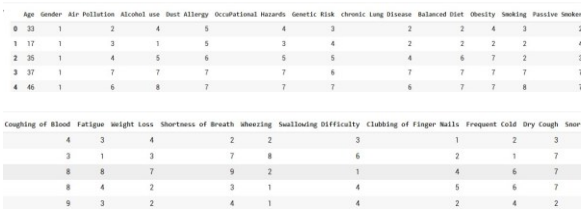
Gambar 2. Tampilan Dataset Sebelum Pra-Prosessing

## B. *Pra-Prosessing*

Pada tahap ini, langkah pertama yaitu memeriksa nilai null, kemudian melakukan penghapusan pada kolom yang tidak relevan seperti "Patient ID" dan "Index". Serta dilakukan juga normalisasi data menggunakan metode StandardScaler. Gambar 4 Tampilan untuk tahap pra-prosessing, sedangkan Gambar 5 Tampilan Dataset menggunakan metode StandardScaler.

```
=== Pemeriksaan Nilai Null ===
index                     0
Patient Id                0
Usia                      0
Jenis Kelamin             0
Polusi Udara              0
Konsumsi Alkohol          0
Alergi Debu               0
Bahaya Pekerjaan          0
Risiko Genetik            0
Penyakit Paru Kronis      0
Pola Makan Seimbang       0
Obesitas                  0
Merokok                   0
Perokok Pasif             0
Nyeri Dada                0
Batuk Berdarah            0
Kelelahan                 0
Penurunan Berat Badan     0
Sesak Napas               0
Mengi                     0
Sulit Menelan             0
Pembesaran Ujung Jari     0
Sering Flu                0
Batuk Kering              0
Mendengkur                0
Tingkat Risiko Kanker     0
dtype: int64
```

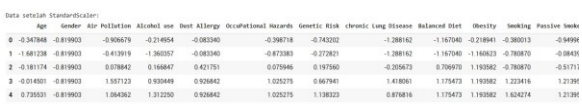
Gambar 3. Tampilan Missing Value



| | Age | Gender | Air Pollution | Alcohol use | Dust Allergy | OccuPational Hazards | Genetic Risk | chronic Lung Disease | Balanced Diet | Obesity | Smoking | Passive Smoke |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 33 | 1 | 2 | 4 | 5 | 4 | 3 | 2 | 2 | 4 | 3 | 2 |
| 1 | 17 | 1 | 3 | 1 | 5 | 3 | 4 | 2 | 2 | 2 | 2 | 4 |
| 2 | 35 | 1 | 4 | 5 | 6 | 5 | 5 | 4 | 6 | 7 | 2 | 3 |
| 3 | 37 | 1 | 7 | 7 | 7 | 7 | 6 | 7 | 7 | 7 | 7 | 7 |
| 4 | 46 | 1 | 6 | 8 | 7 | 7 | 7 | 6 | 7 | 7 | 8 | 7 |

| Coughing of Blood | Fatigue | Weight Loss | Shortness of Breath | Wheezing | Swallowing Difficulty | Clubbing of Finger Nails | Frequent Cold | Dry Cough | Snor |
|---|---|---|---|---|---|---|---|---|---|
| 4 | 3 | 4 | 2 | 2 | 3 | 1 | 2 | 3 | |
| 3 | 1 | 3 | 7 | 8 | 6 | 2 | 1 | 7 | |
| 8 | 8 | 7 | 9 | 2 | 1 | 4 | 6 | 7 | |
| 8 | 4 | 2 | 3 | 1 | 4 | 5 | 6 | 7 | |
| 9 | 3 | 2 | 4 | 1 | 4 | 2 | 4 | 2 | |

Gambar 4. Tampilan Dataset Setelah Pra-Prosessing



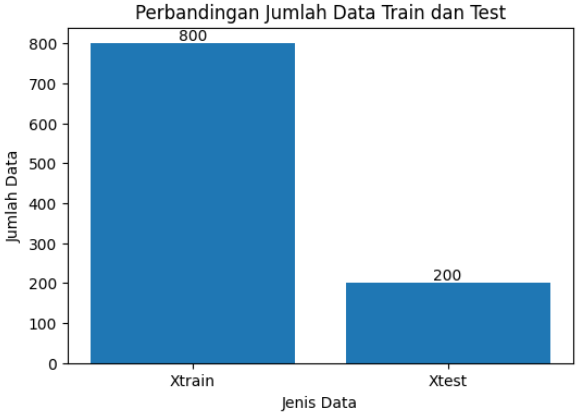
Data setelah StandardScaler:

| | Age | Gender | Air Pollution | Alcohol use | Dust Allergy | OccuPational Hazards | Genetic Risk | chronic Lung Disease | Balanced Diet | Obesity | Smoking | Passive Smoke |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | -0.347848 | -0.819903 | -0.906679 | -0.214954 | -0.083340 | -0.398718 | -0.743202 | -1.288162 | -1.167040 | -0.218941 | -0.380013 | -0.94998 |
| 1 | -1.681238 | -0.819903 | -0.413919 | -1.360357 | -0.083340 | -0.873383 | -0.272821 | -1.288162 | -1.167040 | -1.160623 | -0.780870 | -0.08435 |
| 2 | -0.181174 | -0.819903 | 0.078842 | 0.166847 | 0.421751 | 0.075946 | 0.197560 | -0.205673 | 0.706970 | 1.193582 | -0.780870 | -0.51711 |
| 3 | -0.014501 | -0.819903 | 1.557123 | 0.930449 | 0.926842 | 1.025275 | 0.667941 | 1.418061 | 1.175473 | 1.193582 | 1.223416 | 1.21399 |
| 4 | 0.735531 | -0.819903 | 1.064362 | 1.312250 | 0.926842 | 1.025275 | 1.138323 | 0.876816 | 1.175473 | 1.193582 | 1.624274 | 1.21399 |

Gambar 5. Tampilan Normalisasi menggunakan metode StandardScaler

## C. *Split Data*

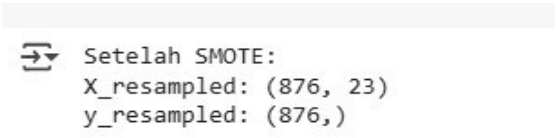
Sebelum dilakukan SMOTE dataset yang terdiri dari 1000 data akan dibagi dengan rasio pembagian sebesar 80:20. Dimana 80% ( 800 data) digunakan sebagai data data latih dan 20% (200 data) sebagai data uji yang digunakan untuk mengevaluasi performa model. Gambar 6 merupakan hasil distribusi.




Gambar 6. Distribusi Data Train dan Test

## D. *Penyeimbangan Data Menggunakan SMOTE*

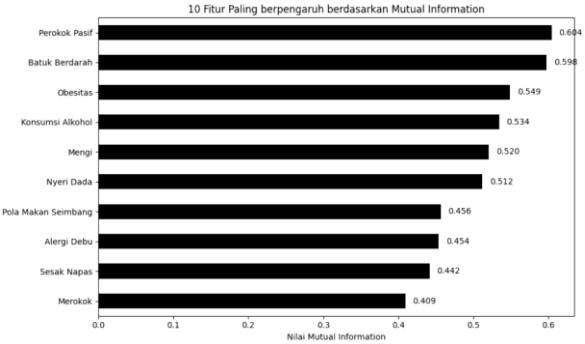
Sebelum melakukan SMOTE, data latih (X_train) menunjukkan distribusi kelas yang tidak seimbang, di mana kelas "High" memiliki jumlah data yang jauh lebih sedikit dibandingkan kelas "Low". Untuk menangani masalah ini pada penelitian ini menerapkan SMOTE. Gambar 7 merupakan Hasil setelah dilakukan SMOTE.



```
Setelah SMOTE:
X_resampled: (876, 23)
y_resampled: (876,)
```

Gambar 7. Tampilan Hasil Setelah SMOTE

*E. Seleksi Fitur dengan Metode Mutual Information*




Gambar 8. Tampilan Grafik Mutual Information

Berdasarkan hasil analisis Mutual Information, fitur "Perokok Pasif" dan "Batuk Berdarah" menepati posisi tertinggi sebagai faktor yang paling berpengaruh terhadap prediksi risiko kanker paru-paru. Secara klinis, kedua faktor ini sangat terkait dengan kanker paru-paru. Menurut WHO (2023) menunjukkan bahwa paparan asap rokok pada perokok pasif dapat meningkatkan risiko terkena kanker paru-paru hingga 30%, sementara batuk berdarah merupakan gejala umum pada pasien kanker paru-paru stadium lanjut [16]. Sedangkan fitur Merokok ada pada urutan paling bawah, karena Perokok Pasif memberikan pengaruh yang lebih konsisten dibanding Merokok aktif.

Berdasarkan perhitungan Mutual Information, sepuluh fitur teratas dipilih sebagai fitur utama untuk pelatihan model, dengan kriteria nilai Mutual Information ≥ 0.05, yang menunjukkan tingkat ketergantungan signifikan tetrhadap kelas target. Visualisasi hasil seleksi menampilkan fitur perokok pasif dan batuk berdarah memberikan kontribusi terbesar dalam memprediksi risiko kanker paru-paru.

*F. Evaluasi Model Dan Hasil*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| High | 0.84 | 0.97 | 0.90 | 73 |
| Low | 1.00 | 0.93 | 0.97 | 61 |
| Medium | 0.93 | 0.82 | 0.87 | 66 |
| accuracy | | | 0.91 | 200 |
| macro avg | 0.92 | 0.91 | 0.91 | 200 |
| weighted avg | 0.92 | 0.91 | 0.91 | 200 |

Gambar 9. Classification Report

Berdasarkan hasil Classification Report pada Gambar 9, dengan menggunakan model Naive Bayes menunjukkan hasil prediksi yang sangat baik dengan akurasi 91.00%. Model mampu untuk mengklasifikasikan kelas High, Low, Medium dengan cukup seimbang. Untuk kelas High, recall mencapai nilai yang tinggi yaitu 0.97, menunjukkan sebagian data asli kelas ini terdeteksi dengan baik, meskipun precisionnya hanya 0.84, yang memiliki prediksi salah (false positive) pada kelas ini. Untuk kelas Low, model memiliki precision sempurna 1.00 dan recall 0.93, yang menunjukkan bahwa prediksi pada kelas Low benar dan hanya sedikit data actual yang tidak dikenali. Sedangkan pada kelas Medium, model ini memiliki precesion yanhg cukup baik 0.93, namun recallnya paling rendah diantara tiga kelas, yaitu 0.82 yang berarti sejumlah data pada kelas Medium tidak berhasil dikenali oleh model. Rata-rata nilai macro average dan weighted average untuk precision, reacll, dan f1-score berada diangka 0.91 dan 0.92, menunjukkan model cukup konsisten pada semua kelas walaupun recall pada kelas Medium masih perlu diperbaiki.

Meskipun model Gaussian Naive Bayes yang digunakan dalam penelitian ini menunjukkan hasil yang cukup baik dengan akurasi mencapai 91.00%, namun masih terdapat hal yang menjadi keterbatasan. Pertama, algoritma Naive Bayes berasumsi bahwa setiap fitur bersifat independen satu sama lain, padahal kenyataannya beberapa faktor gaya hidup seperti kebiasaan merokok dan paparan asap rokok (perokok pasif) saling berkaitan, yang dapat mempengaruhi perhitungan probabilitas dalam model. Kedua, data yang digunakan hanya berasal dari satu sumber yaitu Kaggle, sehingga belum mewakili populasi secara menyeluruh. Oleh karena itu, di masa mendatang, pengujian model menggunakan dataset lain di perlukan untuk mengetahui sejauh mana kemampuan model dalam memprediksi data baru di luar data pelatihan.

Untuk mendapatkan pemahaman yang lebih detail mengenai distribusi hasil prediksi di setiap kelas, dilakukan analisis menggunakan confusion matrix, yang menunjukkan kemampuan model dalam membedakan masing-masing kelas risiko kanker paru-paru. Model ini mengklasifikasikan data terhadap tiga kelas yaitu High, Low, Medium. Pada gambar 9.1 Confusion Matrix Hasil Prediksi Serta dapat diuraikan sebagai berikut.

*1. Kelas High*

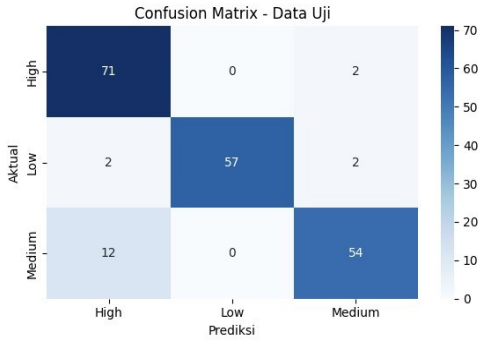
Sebanyak 71 data berhasil diindentifikasi dengan benar sebagai High. Namun, terdapat 2 data yang seharusnya masuk kedalam kategori High, tetapi diprediksi sebagai Medium. Tidak ditemukan kesalahan prediksi kedalam kelas Low.

*2. Kelas Low*

Sebanyak 57 data terklasifikasi dengan bernar sebagai Low. Namun terdapat 2 kesalahan prediksi, 1 ke kelas High dan 1 ke kelas Medium

*3. Kelas Medium*

Sebanyak 54 data berhasil diprediksi dengan benar oleh Medium. Namun, terdapat 12 data yang seharusnya Medium, tetapi diprediksi sebagai High. Tidak ada data Medium yang salah diprediksi ke kelas Low.




Gambar 10. Confusion Matrix Hasil Prediksi

Visualisasi confusion matrix yang menunjukkakn hasil distribusi prediksi benar dan salah antar kelas pada Gambar 9.1, dihasilkan menggunakan fungsi *seaborn.*heatmap*()* yang menampilkan distribusi prediksi antar kelas dalam bentuk warna. Heatmap confucion matrix menggambarkan sitribusi prediksi terhadap tiga kelas risiko: High, Low, dan Medium. Semakin gelap warna kotak menunjukkan jumlah prediksi yang benar semakin tinggi.

## IV. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan, penelitian ini berhasil mengembangkan model prediksi risiko kanker paru-paru berdasarkakn faktor gaya hidup dengan menggunakan algoritma Naive Bayes yang digabungkan dengan metode teknik SMOTE. Hasil akhir menunjukkan bahwa penerapan SMOTE untuk menyeimbangkan kelas dapat meningkatkan akurasi model mencapai 91.00%, dengan nilai precision dan recall yang tinggi diseluruh kelas (Low, Medium, High). Dengan melakukan seleksi fitur menggunakan Mutual Information ditemukan bahwa fitur "Perokok Pasif" dan "Batuk Berdarah" memiliki kontribusi terbesar dalam prediksi. Metode yang digunakan dalam penelitian ini tidak hanya meningkatkan performa model, tetapi juga membuat hasil klasifikasi menjadi mudah dimengerti. Selain itu, hasil penelitian ini juga dapat mengisi celah pada penelitian sebelumnya yang belum secara maksimal mengatasi masalah data yang tidak seimbang dan belum menjelaskan hasil prediksi dengan baik.

Model prediksi yang dihasilkan dalam studi ini memiliki potensi besar untuk digunakan dalam sistem deteksi dini kanker paru-paru. Dengan menggunakan data terkait gaya hidup seperti kebiasaan merokok, aktivitas fisik, dan gejala awal seperti batuk berdarah, model tersebut dapat membantu mengidentifikasi individu yang berisiko tinggi secara efektif. Ke depannya, model ini bisa dikembangkan menjadi aplikasi berbasis web atau mobile yang dapat diakses oleh masyarakat umum maupun tenaga medis untuk skrining awal risiko kanker paru. Dengan demikian, penerapan model ini diharapkan dapat mendukung upaya pencegahan dan penanganan kanker paru sejak tahap awal serta dapat meningkatkan kesadaran masyarakat terhadap pentingnya pola hidup sehat.

## DAFTAR PUSTAKA

[1] R. D. Marzuq, S. A. Wicaksono, and N. Y. Setiawan, "Prediksi Kanker Paru-Paru menggunakan Algoritme Random Forest Decision Tree," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 7, pp. 3448–3456, 2023.

[2] D. Anugrah Pratama, I. Rizal Mutaqin, and K. Rafael Manuela, "Analisis Terjadinya Kanker Paru-Paru Pada Pasien Menggunakan Decision Tree: Penerapan Algoritma C4.5 Dan RapidMiner Untuk Menentukan Risiko Kanker Pada Gejala Pasien," *Jtmei*, vol. 2, no. 4, pp. 156–170, 2023, [Online]. Available: https://doi.org/10.55606/jtmei.v2i4.3004

[3] H. Widya, N. Surya Putra, V. Atina, and J. Maulindar, "Penerapan Algoritme Decision Tree Pada Klasifikasi Penyakit Kanker Paru-Paru," *J. Ilm. Tek. Inform. dan Sist. Inf.*, 2023, [Online]. Available: https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer,

[4] J. Ferlay *et al.*, "Cancer statistics for the year 2020: An overview," *Int. J. Cancer*, vol. 149, no. 4, pp. 778–789, 2021, doi: 10.1002/ijc.33588.

[5] S. Zhang *et al.*, "Predicting the risk of lung cancer using machine learning: A large study based on UK Biobank," *Med. (United States)*, vol. 103, no. 16, p. E37879, 2024, doi: 10.1097/MD.0000000000037879.

[6] N. Publikasi, E. Faizal, I. Stimik, and E. Rahma, "Penerapan Sistem Pakar Untuk Mendiagnosa Penyakit Kanker Pada Wanita Dengan Metode Certainty Factor," vol. 8, no. 6, pp. 1–22, 2015.

[7] Y. Sinjanka, V. Kaur, U. I. Musa, and K. Kaur, "ML-based early detection of lung cancer: an integrated and in-depth analytical framework," *Discov. Artif. Intell.*, vol. 4, no. 1, 2024, doi: 10.1007/s44163-024-00204-6.

[8] L. L. Laily, S. Martini, K. D. Artanti, and S. Widati, "Risk factors of lung adenocarcinoma in patients at dr. soetomo district general hospital surabaya in 2018," no. July 2019, pp. 295–303, 2020, doi: 10.20473/ijph.vl15il.2020.295-303.

[9] N. Sutandyo and E. Suratman, "Non-Small Cell Lung Carcinoma in Women: A Retrospective Cohort Study in Indonesia," *Acta Med. Indones.*, vol. 50, no. 4, pp. 291–298, 2018.

[10] M. I. Fajri and L. Anifah, "Deteksi Status Kanker Paru-Paru Pada Citra Ct Scan Menggunakan Metode Fuzzy Logic," *Tek. Elektro*, vol. 7 no. 3, pp. 121–126, 2018.

[11] Q. An, S. Rahman, J. Zhou, and J. J. Kang, "A Comprehensive Review on Machine Learning in Healthcare Industry: Classification, Restrictions, Opportunities and Challenges," *Sensors*, vol. 23, no. 9, 2023, doi: 10.3390/s23094178.

[12] B. Dutta, "Comparative Analysis of Machine Learning and Deep Learning Models for Lung Cancer Prediction Based on Symptomatic and Lifestyle Features," *Appl. Sci.*, vol. 15, no. 8, 2025, doi: 10.3390/app15084507.

[13] Dewi Widyawati and Amaliah Faradibah, "Comparison Analysis of Classification Model Performance in Lung Cancer Prediction Using Decision Tree, Naive Bayes, and Support Vector Machine," *Indones. J. Data Sci.*, vol. 4, no. 2, pp. 80–89, 2023, doi: 10.56705/ijodas.v4i2.76.

[14] B. Shafa, H. H. Handayani, S. Arum, and P. Lestari, "Prediksi Kanker Paru dengan Normalisasi menggunakan Perbandingan Algoritma Random Forest , Decision Tree dan Naïve Bayes," vol. 4, no. 3, pp. 1057–1070, 2024.

[15] S. A. Karunia, R. Saptono, and R. Anggrainingsih, "Online News Classification Using Naive Bayes Classifier with Mutual Information for Feature Selection," *J. Ilm. Teknol. dan Inf.*, vol. 6, no. 1, pp. 11–15, 2017.

[16] Kemenkes RI, "Panduan Penatalaksanaan Kanker Paru," *Kom. Penanggulangan Kanker Nas.*, pp. 1–47, 2015, [Online]. Available: http://kanker.kemkes.go.id/guidelines/PPKProstat.pdf