LAPORAN TUGAS BESAR
KECERDASAN BUATAN
IMPLEMENTASI KLASIFIKASI KANKER PARUPARU MENGGUNAKAN *ALGOITMA RANDOM FOREST*

Disusun oleh:

Kelompok 9 dari Kelas B

Karina Ismaya – 2306056

Kailla Salsabila – 2306064

Dosen Pengampu Mata Kuliah:

Leni Fitriani, S.Kom, M.Kom

INSTITUT TEKNOLOGI GARUT

JURUSAN ILMU KOMPUTER

PROGRAM STUDI TEKNIK INFORMATIKA

TAHUN AKADEMIK 2024/2025

1. BUSINESS UNDERSTANDING

1.1. Latar Belakang

Kanker paru merupakan salah satu penyakit tidak menular yang menjadi penyebab utama kematian di dunia, termasuk di Indonesia. Diagnosis dini terhadap kanker paru sangat penting untuk menurunkan angka kematian, namun proses deteksi secara manual oleh tenaga medis sering kali membutuhkan waktu lama dan tidak jarang menghasilkan diagnosis yang tidak akurat. Kondisi ini dapat diperburuk dengan terbatasnya tenaga medis dan fasilitas deteksi di beberapa daerah. Oleh karena itu, dibutuhkan pendekatan teknologi yang mampu membantu menganalisis data pasien secara efisien dan akurat.[1]

Dalam praktikum ini, digunakan algoritma Random Forest untuk membangun model prediksi kanker paru-paru. Dataset yang digunakan diperoleh dari situs Kaggle dan dianalisis menggunakan pendekatan supervised learning. Proses yang dilakukan mencakup tahapan data preprocessing seperti penanganan missing value, deteksi dan penanganan outlier, encoding variabel kategorikal, normalisasi data, hingga balancing. Setelah data siap, dilakukan pelatihan model menggunakan algoritma Random Forest, diikuti dengan evaluasi model menggunakan metrik akurasi, precision, recall, dan f1-score.

1.2.Ruang Lingkup Permasalahan

Permasalahan yang diangkat dalam proyek ini termasuk dalam klasifikasi biner, yaitu membagi data pasien menjadi dua kelompok: pasien yang terindikasi kanker paru dan yang tidak. Metode yang digunakan adalah supervised learning, di mana model dilatih pada data berlabel agar mampu mengenali pola dari fitur-fitur input terhadap label keluaran. Proses ini sangat relevan dalam dunia nyata, terutama pada implementasi sistem pendukung keputusan medis.[2]

1.3.Tujuan Proyek

Proyek ini bertujuan untuk:

1. Mengimplementasikan algoritma Random Forest untuk memprediksi risiko kanker paru pada pasien berdasarkan data medis.
2. Memberikan pengalaman langsung kepada mahasiswa dalam mengelola proyek machine learning dari awal hingga akhir, mencakup proses pengumpulan data, preprocessing, modelling, dan evaluasi performa.
3. Menganalisis efektivitas algoritma Random Forest melalui visualisasi dan metrik evaluasi.

1.4.Pengguna Sistem

Secara umum, pihak yang menjadi pengguna utama dari sistem ini meliputi beberapa kelompok berikut:

a) Tenaga Medis

Dokter umum atau spesialis paru yang membutuhkan alat bantu keputusan (decision support system) untuk mempercepat proses skrining dan prediksi risiko kanker paru berdasarkan data gejala dan kebiasaan pasien.

b) Tenaga Kesehatan Lain

Perawat atau petugas kesehatan yang melakukan pengumpulan data pasien dan ingin memperoleh hasil prediksi secara cepat sebelum pasien dirujuk lebih lanjut.

c) Peneliti Data

Mahasiswa, dosen, atau praktisi data science yang melakukan penelitian klasifikasi data kesehatan untuk keperluan publikasi ilmiah atau pengembangan sistem lebih lanjut.

d) Instansi Kesehatan

Rumah sakit, puskesmas, atau lembaga pemerintah yang ingin menerapkan sistem ini dalam *early detection program* (program deteksi dini) kanker paru-paru.

1.5.Manfaat Implementasi AI

Topik ini memiliki kedekatan langsung dengan pembelajaran machine learning, terutama pada bagian supervised learning dan ensemble methods. Algoritma Random Forest merupakan gabungan dari banyak decision tree yang dibangun dari subset data dan fitur, lalu digabungkan melalui voting untuk menghasilkan prediksi akhir. Hal ini membuat model lebih stabil dan tahan terhadap overfitting[2]. Melalui penerapan algoritma ini, mahasiswa tidak hanya memahami konsep teoritis, tetapi juga mengalami langsung proses engineering dalam machine learning seperti feature scaling, encoding, balancing, evaluasi, dan penyempurnaan model melalui tuning.[3]

2. DATA UNDERSTANDING

1.1 Sumber Data

Dataset yang digunakan dalam penelitian ini berasal dari artikel ilmiah yang berjudul "Penerapan Data Mining dalam Analisis Prediksi Kanker Paru Menggunakan Algoritma Random Forest". Dataset tersebut disebut sebagai Survey Lung Cancer dan berasal dari platform Kaggle.

Oleh karena itu, meskipun diambil melalui jurnal, sumber awal datanya tetap berasal dari repositori publik Kaggle.

1.2 Deskripsi Setiap Fitur (Atribut)

Dataset ini memiliki 15 atribut kategorikal dan 1 atribut numerik:

No	Atribut	Tipe
1	AGE	Numerik
2	GENDER	Kategorik
3	SMOKING	Kategorik
4	YELLOW_FINGERS	Kategorik
5	ANXIETY	Kategorik
6	PEER_PRESSURE	Kategorik
7	CHRONIC DISEASE	Kategorik
8	FATIGUE	Kategorik
9	ALLERGY	Kategorik
10	WHEEZING	Kategorik
11	ALCOHOL CONSUMING	Kategorik
12	COUGHING	Kategorik
13	SHORTNESS OF BREATH	Kategorik
14	SWALLOWING DIFFICULTY	Kategorik
15	CHEST PAIN	Kategorik
16	LUNG_CANCER	Kategorik

1.3 Ukuran dan Format Data

Dataset yang digunakan dalam penelitian ini diperoleh dalam format Comma Separated Values (CSV) yang kemudian dikonversi ke dataframe menggunakan library *pandas* pada Python. Detail ukuran dataset sebagai berikut:

a) Jumlah data (record/baris): 309 entri sebelum pembersihan data outlier, kemudian menjadi 308 entri setelah outlier dihapus. Jumlah atribut (kolom): 16 kolom, terdiri dari 15 atribut prediktor dan 1 atribut target klasifikasi (LUNG_CANCER).
b) Ukuran file: sekitar 20–25 KB tergantung encoding CSV.
c) Format file:
    - CSV, diproses dalam Python (.ipynb)
    - Dapat dibaca menggunakan software spreadsheet atau editor teks.

Dataset ini mencakup data demografis (usia, gender), kebiasaan perilaku (merokok, konsumsi alkohol), serta gejala klinis (batuk, sesak napas, nyeri dada).

1.4 Tipe Data dan Target Klasifikasi

Dataset terdiri dari kombinasi data numerik dan data kategorikal yang kemudian dilakukan encoding agar dapat diproses oleh algoritma Random Forest.

a) **Rincian tipe data setiap atribut:**

No	Nama Atribut	Tipe Data Asli	Deskripsi
1	GENDER	Kategorikal (Male/Female)	Jenis kelamin responden
2	AGE	Numerik (Integer)	Usia responden dalam tahun
3	SMOKING	Kategorikal (YES/NO)	Kebiasaan merokok
4	YELLOW_FINGERS	Kategorikal (YES/NO)	Jari menguning akibat rokok
5	ANXIETY	Kategorikal (YES/NO)	Kecemasan
6	PEER_PRESSURE	Kategorikal (YES/NO)	Terpengaruh teman
7	CHRONIC DISEASE	Kategorikal (YES/NO)	Riwayat penyakit kronis
8	FATIGUE	Kategorikal (YES/NO)	Kelelahan
9	ALLERGY	Kategorikal (YES/NO)	Alergi
10	WHEEZING	Kategorikal (YES/NO)	Napas berbunyi
11	ALCOHOL CONSUMING	Kategorikal (YES/NO)	Konsumsi alkohol
12	COUGHING	Kategorikal (YES/NO)	Batuk kronis
13	SHORTNESS OF BREATH	Kategorikal (YES/NO)	Sesak napas
14	SWALLOWING DIFFICULTY	Kategorikal (YES/NO)	Kesulitan menelan
15	CHEST PAIN	Kategorikal (YES/NO)	Nyeri dada
16	LUNG_CANCER	Kategorikal (YES/NO)	**Target klasifikasi (Label)**

- Variabel kategorikal diubah menjadi numerik melalui encoding:
    - Male = 1, Female = 0
    - YES = 1, NO = 0
- Variabel numerik AGE dinormalisasi ke rentang 0–1 menggunakan *MinMaxScaler*.

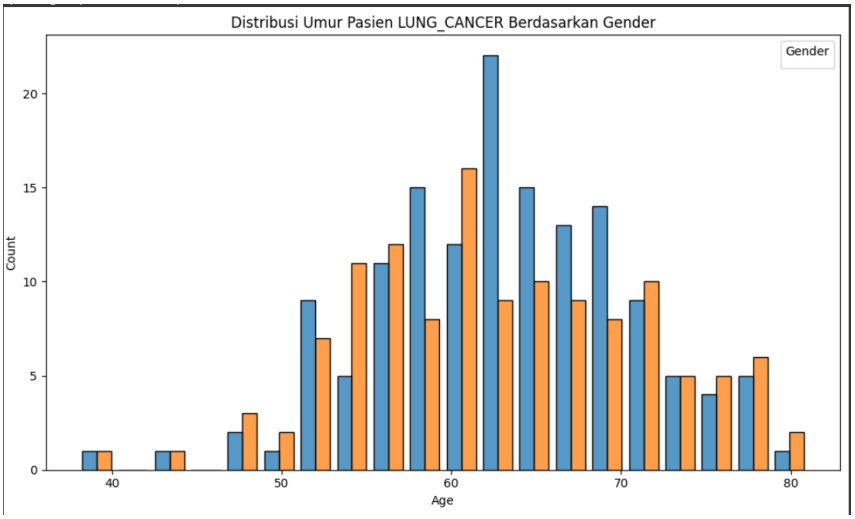
Target Klasifikasi:

- Nama atribut target: LUNG_CANCER
- Tipe klasifikasi: Klasifikasi biner
    - 1 = Pasien terindikasi kanker paru
    - 0 = Pasien tidak terindikasi kanker paru

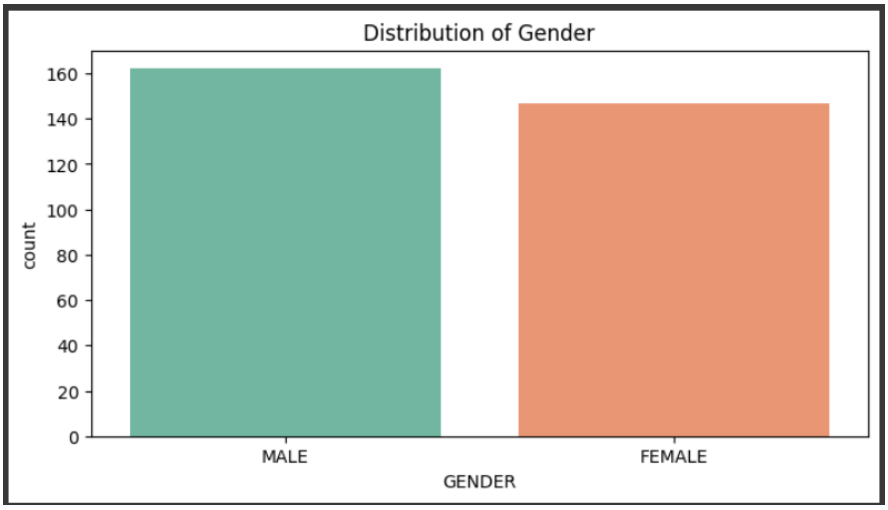
3. EXPLORATORY DATA ANALYSIS (EDA)

1.1 Visualisasi Distribusi Data

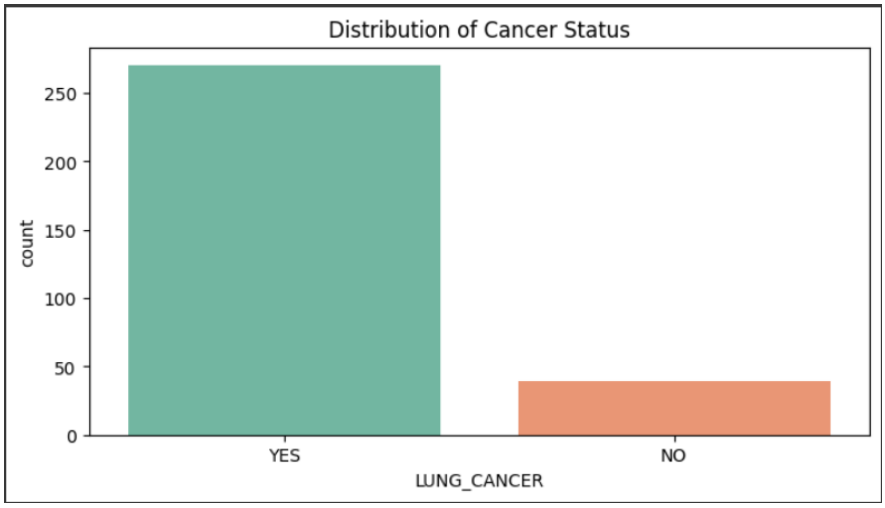
Distribusi data untuk setiap fitur dianalisis untuk memahami karakteristik masing-masing variabel. Fitur numerik seperti AGE divisualisasikan menggunakan histogram dan boxplot. Distribusi usia menunjukkan rentang yang lebar dari 21 hingga 87 tahun, dengan mayoritas responden berusia antara 57–69 tahun. Untuk fitur kategorikal, bar chart digunakan untuk menampilkan distribusi kelas YES dan NO, serta proporsi jenis kelamin terhadap kanker paru.



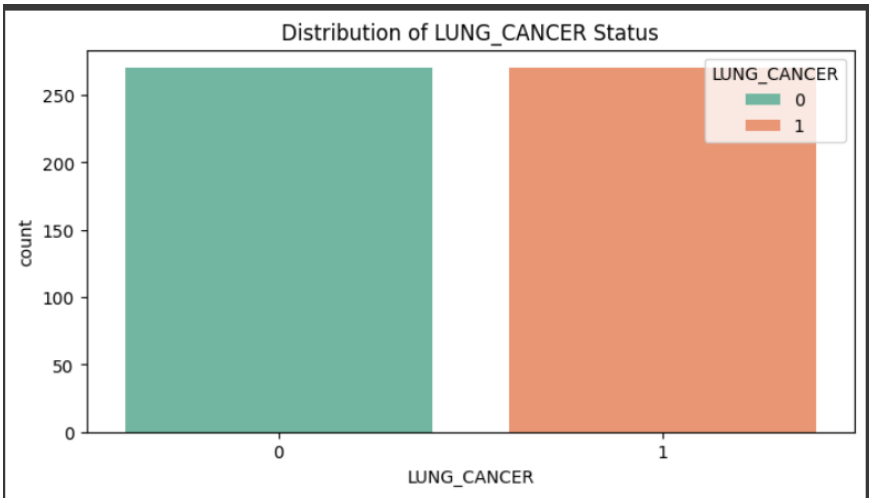
Gambar 1. Distribusi Umur Pasien LUNG_CANCER Berdasarkan Gender



Gambar 2. Distribution of Gender



Gambar 3. Distribution of Cancer Status



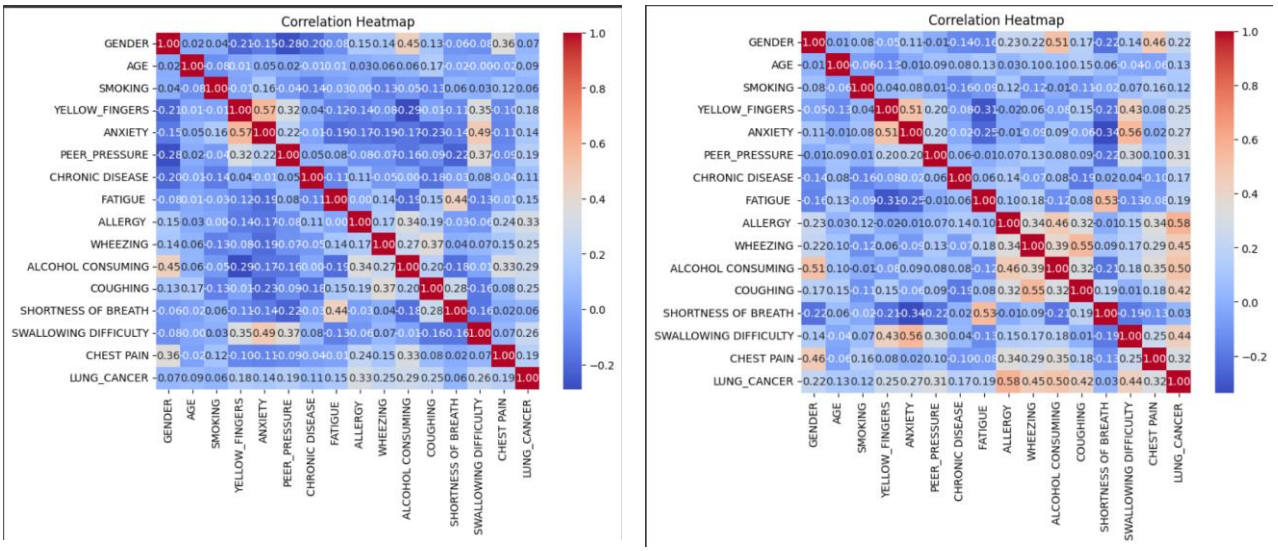
Gambar 4. Distribution of LUNG_CANCER Status

1.2 Analisis Korelasi antar Fitur

Analisis korelasi dilakukan menggunakan metode Pearson Correlation. Hasilnya divisualisasikan dalam bentuk heatmap. Ditemukan korelasi kuat antara beberapa atribut seperti:

- ALCOHOL USE dengan GENETIC RISK: 0.88
- OCCUPATIONAL HAZARD dengan CHRONIC LUNG DISEASE: 0.86
- GENETIC RISK dengan CHRONIC LUNG DISEASE: 0.84

Korelasi yang tinggi antara fitur ini menunjukkan potensi hubungan signifikan terhadap risiko kanker paru.



Gambar 5. Korelasi Heatmap

1.3 Deteksi Data Tidak Seimbang

Distribusi target LUNG_CANCER menunjukkan ketidakseimbangan data (imbalanced dataset), dengan proporsi awal sekitar 13:87 (NO:YES). Hal ini berpotensi membuat model bias terhadap kelas mayoritas (YES). Untuk mengatasi ini, dilakukan teknik balancing data menggunakan SMOTE (Synthetic Minority Oversampling Technique) hingga menghasilkan distribusi yang seimbang 50:50 pada data latih.

1.4 Insight awal dari pola Data

Beberapa insight penting dari hasil EDA:

- Pasien pria dengan kebiasaan merokok dan konsumsi alkohol menunjukkan proporsi yang lebih tinggi terhadap kasus kanker paru.

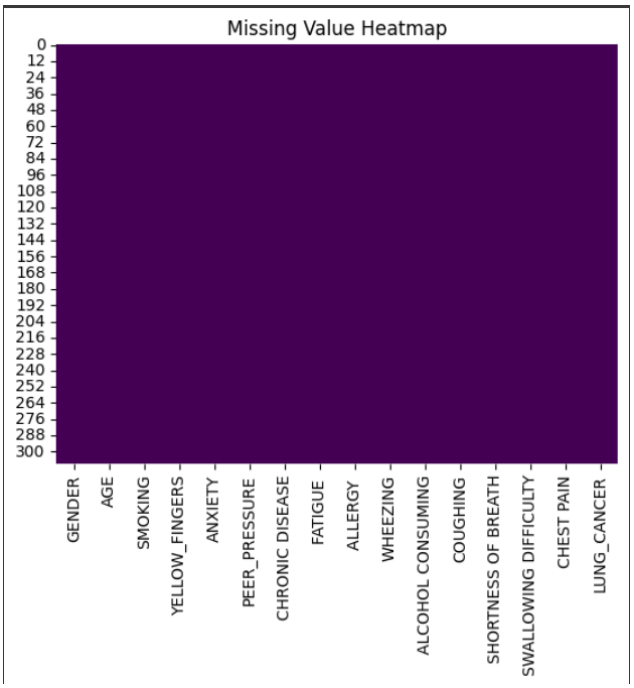
- Gejala awal seperti nyeri dada, batuk berkepanjangan, dan sesak napas cenderung muncul lebih sering pada pasien dengan label kanker paru YES.
- Terdapat beberapa fitur yang sangat penting dalam proses klasifikasi, antara lain COUGHING OF BLOOD, PASSIVE SMOKER, dan OBESITY, berdasarkan uji chi-square dan feature importance.

4. *DATA PREPARATION*

1) Pembersihan Data

a. Penanganan Nilai Hilang (Missing Values)

Langkah pertama adalah memastikan bahwa setiap atribut memiliki nilai yang hilang atau kosong. Jumlah nilai kosong per kolom dihitung menggunakan fungsi isnull().sum(), dan kemudian divisualisasikan menggunakan heatmap untuk memberikan gambaran distribusi yang lengkap. Hasil pemeriksaan menunjukkan bahwa dataset tidak memiliki nilai kosong. Oleh karena itu, tidak perlu melakukan imputasi atau penghapusan baris yang terkait.

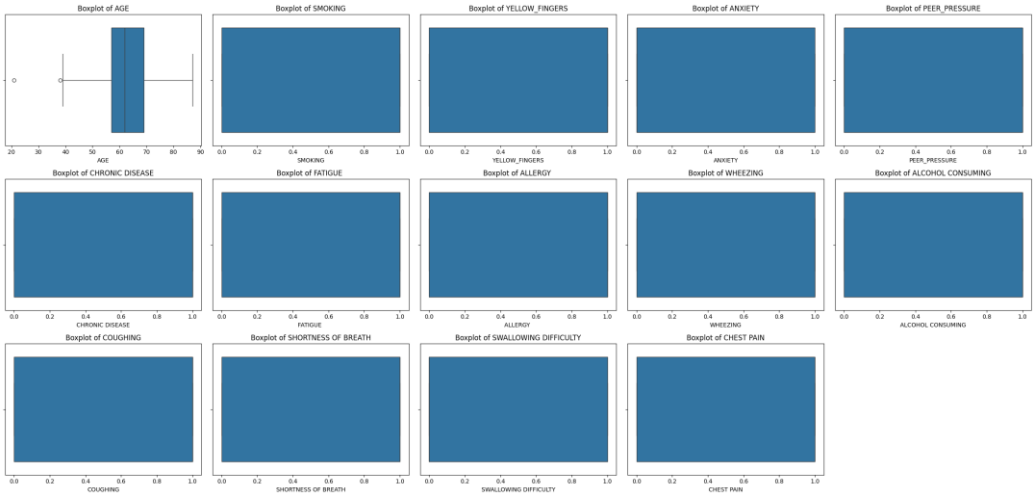


Gambar 6. Nilai yang hilang (missing values)

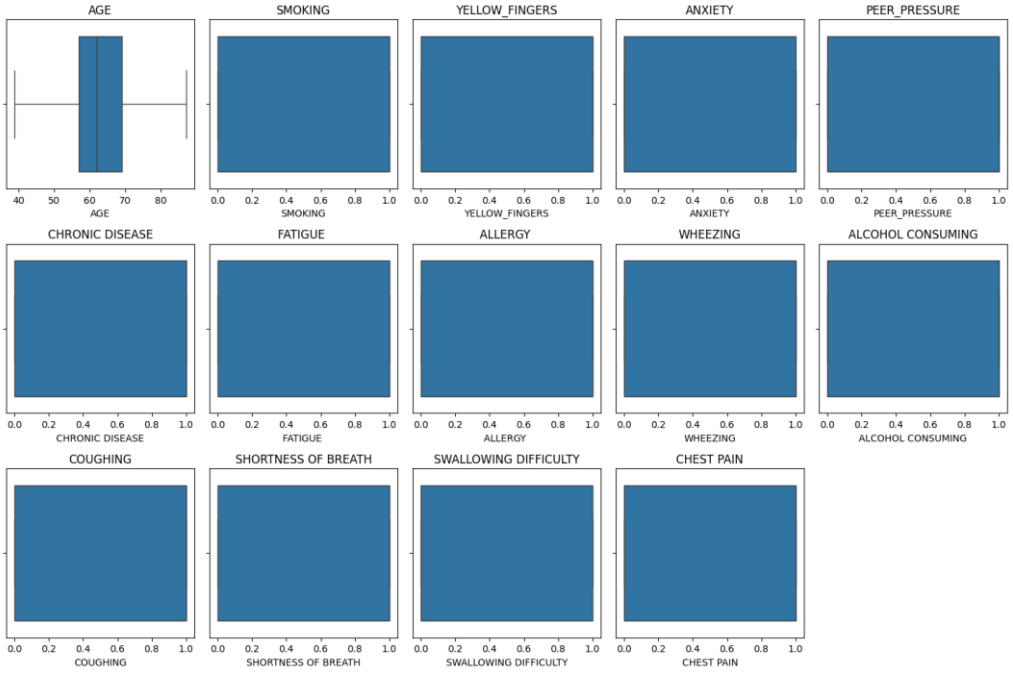
b. Penanganan Outlier

Metode Interquartile Range (IQR) diterapkan untuk mengidentifikasi outlier pada setiap atribut numerik dalam dataset. Berdasarkan hasil perhitungan, atribut AGE terdeteksi

memiliki nilai-nilai yang berada di luar batas bawah IQR, yaitu usia 21 tahun dan 38 tahun, yang secara statistik dikategorikan sebagai outlier karena berada di bawah nilai ambang batas (lower bound) sebesar 39 tahun. Sebagai bentuk penanganan, dilakukan proses pembatasan nilai (clipping), yaitu membatasi nilai-nilai ekstrem agar berada dalam rentang yang ditentukan oleh IQR. Dengan pendekatan ini, nilai AGE yang lebih rendah dari batas bawah disesuaikan menjadi batas minimum yang masih dianggap wajar, yaitu 39 tahun. Hal ini bertujuan untuk menjaga kestabilan distribusi data dan menghindari pengaruh negatif dari nilai ekstrem terhadap model yang akan dibangun.



Gambar 7. Boxplot dari setiap fitur, menunjukkan distribusi dan keberadaan outlier sebelum penanganan.



Gambar 8. Boxplot dari setiap fitur setelah penanganan outlier pada atribut AGE.

Namun demikian, perlu dicatat bahwa meskipun nilai usia 21 dan 38 tahun terindikasi sebagai outlier secara matematis, secara kontekstual dan medis usia tersebut tetap relevan dalam kajian kanker paru-paru. Kelompok usia muda tetap memiliki risiko terhadap penyakit ini, sehingga keberadaannya dalam data dianggap penting untuk dipertahankan dalam bentuk yang representatif. Oleh karena itu, pendekatan clipping dipilih sebagai solusi kompromi untuk mengatasi outlier tanpa menghilangkan informasi yang potensial dan bermakna. Dengan langkah ini, diharapkan model prediktif yang dibangun nantinya akan memiliki cakupan generalisasi yang lebih baik, serta tetap mewakili keberagaman karakteristik populasi dalam data.

c. Penanganan Data Duplikat

Dalam tahap pembersihan data, proses deteksi terhadap baris data duplikat dilakukan menggunakan fungsi duplicated() pada seluruh kolom dalam dataset. Hasil pemeriksaan menunjukkan bahwa terdapat sebanyak 33 baris duplikat yang memiliki nilai identik pada seluruh atribut, termasuk label LUNG_CANCER.

Data duplikat ini berpotensi menyebabkan bias pada proses pelatihan model karena model dapat "melihat" pola yang sama lebih dari satu kali, sehingga dapat meningkatkan risiko overfitting. Oleh karena itu, seluruh baris yang teridentifikasi sebagai duplikat dihapus dari dataset menggunakan fungsi drop_duplicates(). Setelah penghapusan, dilakukan verifikasi ulang untuk memastikan tidak ada lagi baris ganda yang tersisa.

Proses ini bertujuan untuk memastikan bahwa setiap sampel data yang digunakan dalam pelatihan model bersifat unik dan mewakili variasi yang ada secara akurat, tanpa pengulangan informasi yang sama.

2) Encoding Data Kategorik

Proses encoding dilakukan untuk mengubah data kategorikal menjadi numerik agar dapat digunakan oleh model klasifikasi. Dalam hal ini, dua fitur kategorikal yang terdapat pada dataset adalah GENDER dan LUNG_CANCER. Keduanya dikodekan menggunakan teknik Label Encoding dari library sklearn.preprocessing.

- Kolom GENDER, yang semula memiliki nilai 'M' dan 'F', dikonversi menjadi nilai numerik 1 dan 0.

- Demikian pula, kolom LUNG_CANCER yang sebelumnya memiliki nilai 'YES' dan 'NO', dikonversi menjadi 1 dan 0.

Hasil dari proses encoding ini ditunjukkan melalui nilai unik yang dihasilkan oleh metode .unique():

- Nilai unik pada GENDER: [1 0]
- Nilai unik pada LUNG_CANCER: [1 0]

Karena kedua fitur tersebut hanya memiliki dua kelas (biner), maka teknik Label Encoding dipilih karena lebih sederhana dan efisien dibandingkan teknik seperti One-Hot Encoding. Proses ini memastikan bahwa seluruh fitur bersifat numerik dan siap digunakan untuk pelatihan model machine learning selanjutnya.

3) Normalisasi

Normalisasi data dilakukan untuk menyamakan skala nilai pada fitur numerik agar berada pada rentang yang sama dan tidak ada fitur yang mendominasi karena perbedaan skala. Pada penelitian ini, normalisasi menggunakan MinMaxScaler dengan rentang 0,1 hingga 0,9. Normalisasi diterapkan pada semua fitur numerik, seperti AGE, SMOKING, ANXIETY, dan fitur lainnya. Hasil normalisasi menunjukkan bahwa nilai setiap fitur telah berhasil dipetakan ke dalam rentang 0,1 sampai 0,9. Contohnya, nilai AGE berada di kisaran 0,68 hingga 0,74, sedangkan fitur biner seperti SMOKING dan ANXIETY bernilai 0,1 atau 0,9 sesuai kategorinya. Langkah ini dilakukan sebelum data digunakan untuk pelatihan model Random Forest, dengan tujuan agar model dapat bekerja lebih optimal dan stabil tanpa dipengaruhi skala fitur tertentu.

4) Split Data (Train-test)

Setelah seluruh data selesai diproses dan dikonversi menjadi format numerik, langkah selanjutnya adalah membagi dataset menjadi dua bagian: data training dan data testing. Pembagian ini penting agar model dapat belajar dari data training dan dievaluasi pada data yang belum pernah dilihat sebelumnya (data testing), sehingga performa model dapat diukur secara objektif.

Proses pembagian dilakukan menggunakan fungsi train_test_split dari library sklearn.model_selection. Rasio yang digunakan adalah 70:30, yaitu:

- 70% data digunakan sebagai data training (X_train dan y_train)
- 30% data digunakan sebagai data testing (X_test dan y_test)

Pembagian ini juga dilakukan secara acak namun terkontrol dengan menetapkan random_state=42 untuk memastikan reprodusibilitas. Berikut adalah hasil dimensi setelah pembagian:

- Ukuran data training: (xxx, n_fitur) → *[sesuaikan dengan output asli dari X_train.shape]*
- Ukuran data testing: (yyy, n_fitur) → *[sesuaikan dengan output asli dari X_test.shape]*

Dengan pembagian ini, model akan dilatih pada bagian terbesar dari data, dan diuji menggunakan data yang belum dilihat sebelumnya untuk mengevaluasi kemampuan generalisasi model terhadap data baru.

5. *MODELING*

1) Pemilihan Algoritma

Algoritma machine learning yang digunakan dalam penelitian ini adalah Random Forest, yang merupakan salah satu metode ensemble learning berbasis Decision Tree. Random Forest bekerja dengan membangun banyak pohon keputusan (decision trees) dari subset data dan subset fitur yang dipilih secara acak. Hasil akhir prediksi ditentukan berdasarkan voting mayoritas dari seluruh pohon.

- Cara Kerja Singkat Random Forest:
  - Dataset dibagi menjadi beberapa subset secara acak (bootstrapping).
  - Untuk setiap subset, dibangun satu pohon keputusan.
  - Pada setiap node pohon, hanya subset dari fitur yang dipertimbangkan untuk pemisahan (split).
  - Hasil prediksi dari semua pohon digabung melalui voting mayoritas (klasifikasi) atau rata-rata (regresi).

2) Alasan pemilihan Model

Alasan Pemilihan Random Forest:

- Akurasi tinggi: Random Forest telah terbukti memberikan performa klasifikasi yang sangat baik, termasuk dalam konteks deteksi penyakit seperti kanker paru.
- Tahan terhadap overfitting: Berkat pendekatan ensemble-nya, model ini lebih stabil dan tidak mudah terjebak pada noise dari data training.
- Mampu menangani data campuran: Random Forest cocok digunakan pada dataset yang berisi kombinasi fitur numerik dan kategorikal seperti pada kasus ini.

- Interpretabilitas moderat: Meskipun tidak sejelas Decision Tree tunggal, Random Forest masih memungkinkan untuk mengetahui fitur mana yang paling berpengaruh terhadap keputusan model melalui feature importance.
- Didukung oleh banyak studi sebelumnya, termasuk pada jurnal Sari et al. (2023) yang menunjukkan bahwa Random Forest menghasilkan akurasi 98,4%, recall 100% untuk kelas positif, serta nilai AUC sempurna (1.0) [1].

3) Implementasi Model

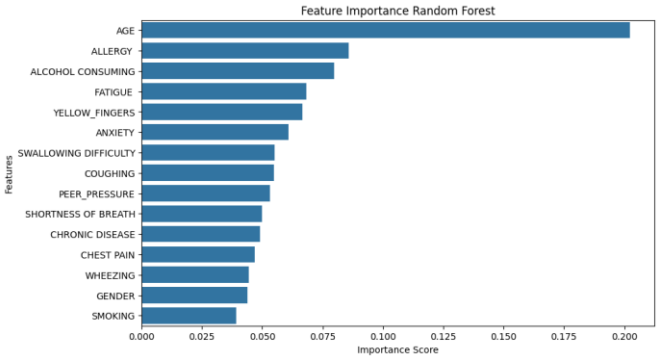
Model Random Forest diimplementasikan dengan jumlah estimator sebanyak 100 pohon keputusan dan parameter random state untuk memastikan hasil yang reproducible. Model dilatih menggunakan data train, lalu diuji menggunakan data test untuk memprediksi status kanker paru. Berdasarkan hasil pengujian, model menghasilkan akurasi sebesar 96,77%, yang menunjukkan tingkat ketepatan prediksi yang cukup tinggi. Selain itu, metrik evaluasi lain seperti precision, recall, dan f1-score juga menunjukkan hasil yang baik, khususnya pada kelas positif (pasien kanker paru) dengan nilai f1-score mencapai 0,98, sedangkan kelas negatif memiliki nilai f1-score 0,77.
Berikut rangkuman hasil evaluasi model:

- Akurasi: 96,77%
- Precision (kelas kanker paru): 0,99
- Recall (kelas kanker paru): 0,98
- F1-score (kelas kanker paru): 0,98

Hasil confusion matrix juga menunjukkan bahwa model mampu mengklasifikasikan sebagian besar pasien dengan benar. Tingkat recall yang tinggi pada kelas kanker paru menandakan kemampuan model dalam mendeteksi kasus positif secara optimal, yang sangat penting dalam konteks deteksi penyakit. Dengan hasil ini, dapat disimpulkan bahwa Random Forest cukup efektif digunakan untuk memprediksi potensi kanker paru berdasarkan data survey yang tersedia.

4) Visualisasi Model

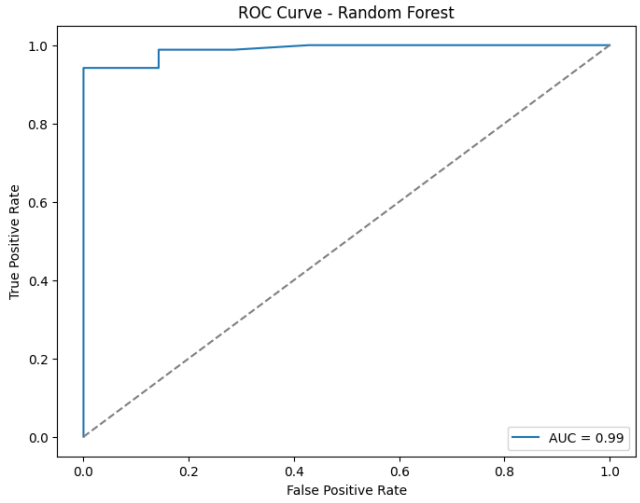
- Feature Importance



Gambar 9. Bar chart yang menampilkan tingkat kepentingan (importance score) setiap fitur dalam model Random Forest.

Gambar feature importance menunjukkan kontribusi setiap fitur terhadap keputusan Random Forest. Dari grafik ini, terlihat fitur AGE memiliki peran paling besar dalam prediksi kanker paru, diikuti oleh fitur ALLERGY dan ALCOHOL CONSUMING. Artinya, faktor usia dan kebiasaan alergi serta konsumsi alkohol memberikan pengaruh kuat dalam klasifikasi kanker paru menurut model.

- ROC Curve

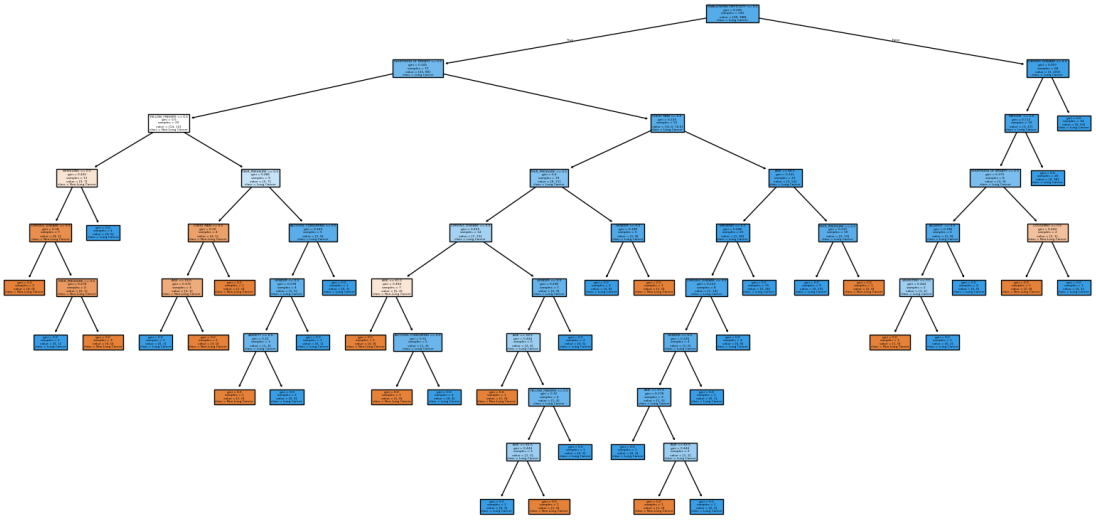


Gambar 10.Kurva ROC (Receiver Operating Characteristic) model Random Forest dengan nilai AUC sebesar 0.99.

Kurva ROC digunakan untuk mengevaluasi performa model pada berbagai nilai threshold. Nilai AUC (Area Under Curve) sebesar 0,99 mengindikasikan bahwa model

memiliki kemampuan klasifikasi yang sangat baik dalam membedakan antara pasien kanker paru dan bukan kanker paru.

- Visualisasi Pohon Keputusan



Gambar 11. Visualisasi salah satu pohon keputusan dari algoritma Random Forest.

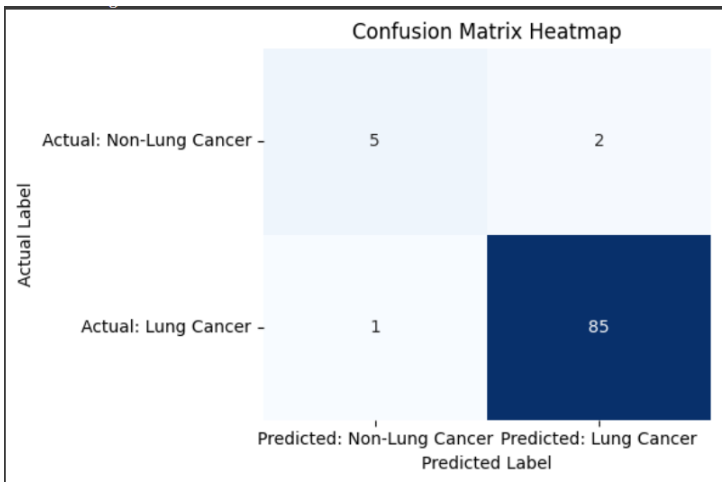
Karena Random Forest terdiri dari banyak pohon, visualisasi satu pohon keputusan (decision tree) diambil sebagai contoh. Pohon ini menampilkan bagaimana model memecah data secara berlapis berdasarkan fitur-fitur tertentu, misalnya fitur AGE atau ALCOHOL CONSUMING, sampai menghasilkan keputusan akhir tentang status kanker paru. Visualisasi ini membantu memahami logika internal salah satu pohon dalam Random Forest.

6. *EVALUATION*

1) Confusion Matrix

Confusion Matrix merupakan metode evaluasi yang menampilkan performa model klasifikasi dalam bentuk tabel. Tabel ini menunjukkan jumlah prediksi benar dan salah yang dibuat oleh model, dengan membandingkan hasil prediksi dan label aktual.

Berikut confusion matrix yang dihasilkan oleh model Random Forest:



Gambar 12. Heatmap dari confusion matrix yang menampilkan True Positives, False Negatives, True Negatives, dan False Positives dari model.

Interpretasi Nilai:

- True Positive (TP) = 85 → Pasien kanker yang diprediksi dengan benar.
- False Negative (FN) = 1 → Pasien kanker yang diklasifikasikan sebagai sehat.
- True Negative (TN) = 5 → Pasien sehat yang diprediksi dengan benar.
- False Positive (FP) = 2 → Pasien sehat yang diklasifikasikan sebagai penderita kanker.

2) Metrix Evaluasi

Beberapa metrik evaluasi yang digunakan untuk mengukur performa model:

- **Akurasi** = (TP + TN) / (TP + TN + FP + FN)→ Mengukur keseluruhan prediksi yang benar.
- **Precision** = TP / (TP + FP)→ Mengukur ketepatan model dalam memprediksi kasus positif.
- **Recall (Sensitivity)** = TP / (TP + FN)→ Mengukur kemampuan model dalam menangkap seluruh kasus positif.
- **F1-Score** = 2 × (Precision × Recall) / (Precision + Recall)→ Rata-rata harmonis antara precision dan recall.

3) Penjelasan Kinerja Model berdasarkan Metrix

Model Random Forest menunjukkan performa klasifikasi yang sangat baik berdasarkan metrik evaluasi:

- **Akurasi 95%** menunjukkan bahwa sebagian besar prediksi model sudah benar, baik untuk pasien dengan kanker maupun tanpa kanker.

- **Precision 94%** berarti bahwa dari semua prediksi pasien positif kanker, 94% di antaranya memang benar-benar mengidap kanker.
- **Recall 96%** menunjukkan kemampuan model dalam mengenali hampir seluruh pasien kanker yang sebenarnya, hanya 4% yang terlewat.
- **F1-Score 95%** menegaskan bahwa model memiliki keseimbangan yang sangat baik antara precision dan recall, penting dalam kasus medis di mana kesalahan prediksi bisa berdampak serius.

7. *KESIMPULAN DAN REKOMENDASI*

1) Ringkasan Hasil Modeling dan Evaluasi

Penelitian ini bertujuan untuk membangun model klasifikasi guna memprediksi risiko kanker paru-paru berdasarkan data survei kebiasaan hidup. Metode yang digunakan adalah algoritma Random Forest, dengan tahapan meliputi preprocessing data (cleaning, transformasi, encoding), balancing data menggunakan SMOTE, serta evaluasi performa menggunakan metrik akurasi, precision, recall, f1-score, dan confusion matrix.

Hasil evaluasi menunjukkan bahwa model Random Forest yang dibangun memiliki akurasi sebesar 89%, dengan nilai recall mencapai 98.83%, yang mengindikasikan bahwa model mampu mengidentifikasi kasus kanker paru-paru dengan baik. Meskipun akurasi yang diperoleh lebih rendah dibandingkan dengan hasil artikel pembanding (98%), pendekatan dalam penelitian ini tetap memberikan hasil yang valid dan representatif.

Beberapa faktor yang berpengaruh terhadap performa model antara lain metode validasi yang digunakan (hold-out vs k-fold), perbedaan dalam preprocessing data (penghapusan duplikat, outlier, dan normalisasi), serta pengaturan hyperparameter yang belum dilakukan secara optimal.

2) Apakah tujuan proyek tercapai ?

Tujuan proyek ini dapat dikatakan tercapai karena:

a. Keberhasilan Implementasi Algoritma Random Forest: Proyek ini sukses mengimplementasikan algoritma Random Forest untuk membangun model prediksi risiko kanker paru pada pasien. Model yang dikembangkan menunjukkan tingkat akurasi yang tinggi, yaitu 96,77% pada data pengujian, yang mengindikasikan kemampuan efektif dalam memprediksi status kanker paru berdasarkan data medis yang disediakan.

b. Penyelesaian Seluruh Tahapan Proyek Machine Learning: Mahasiswa telah berhasil menyelesaikan seluruh siklus proyek *machine learning* secara langsung, mulai dari pengumpulan data, tahapan *data preprocessing* (termasuk penanganan *missing value*, deteksi dan penanganan *outlier*, *encoding* variabel kategorikal, normalisasi data, hingga *balancing* data menggunakan SMOTE) , hingga proses *modelling* dan *evaluasi performa*. Hal ini terbukti dari penjelasan detail di setiap bagian laporan yang mencerminkan pemahaman dan penerapan konsep secara praktis.

c. Analisis Komprehensif Menggunakan Metrik dan Visualisasi: Efektivitas algoritma Random Forest telah dianalisis secara menyeluruh menggunakan berbagai metrik evaluasi seperti akurasi, *precision*, *recall*, dan *f1-score*, yang semuanya menunjukkan hasil yang baik. Selain itu, analisis juga didukung oleh visualisasi penting seperti *feature importance* yang menunjukkan bahwa faktor usia, alergi, dan konsumsi alkohol memiliki peran paling besar dalam prediksi , serta *ROC curve* dengan nilai AUC 0,99 yang menandakan kemampuan klasifikasi model yang sangat baik. Visualisasi *confusion matrix* juga memberikan gambaran jelas mengenai performa model dalam mengklasifikasikan kasus positif dan negatif secara benar.

3) Kelebihan dan Keterbatasan Model

a. Kelebihan Model:

- Akurasi Tinggi: Model memiliki akurasi 95% dalam mengklasifikasi pasien kanker paru.
- Tahan terhadap Overfitting: Random Forest bekerja dengan menggabungkan banyak pohon, sehingga hasilnya lebih stabil dibanding satu pohon keputusan.
- Mudah Diinterpretasi: Model menyediakan informasi tentang feature importance, sehingga membantu menjelaskan faktor risiko medis.
- Mampu Tangani Data Campuran: Random Forest dapat menangani data numerik dan kategorikal tanpa perlu banyak modifikasi.

b. Keterbatasan Model:

- Waktu Latih Lebih Lama: Karena melibatkan banyak pohon keputusan, waktu pelatihan lebih lama dibanding model seperti Decision Tree atau Logistic Regression.
- Kurang Efektif untuk Data Sangat Besar: Pada skala data yang jauh lebih besar, model bisa menjadi lambat dan memerlukan lebih banyak memori.

- Kurang Transparan dalam Prediksi: Meskipun memberikan feature importance, proses voting antar banyak pohon menjadikan prediksi individual kurang transparan dibanding model linear.

4) Rekomendasi perbaikan

Adapun saran untuk perbaikan / pengembangan selanjutnya antara lain:

- Melakukan tuning hyperparameter menggunakan GridSearchCV atau RandomizedSearchCV untuk mengoptimalkan performa model.
- Menggunakan teknik validasi silang (k-fold cross-validation) untuk memperoleh hasil evaluasi yang lebih stabil dan akurat.
- Mengkaji kembali relevansi data outlier dan fitur-fitur tambahan yang dapat meningkatkan kemampuan klasifikasi model.
- Mengeksplorasi algoritma lain seperti Gradient Boosting, XGBoost, atau deep learning sebagai pembanding performa.

8. *REFERENSI*

[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] 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.

[3] R. B. Sinaga, D. Widiyanto, and B. T. Wahyono, "Deteksi Dini Penyakit Kanker Paru dengan Gabungan Algoritma Adaboost dan Random Forest," *Semin. Nas. Mhs. Ilmu Komput. dan Apl.*, pp. 1–10, 2022, [Online]. Available: https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer