JURNAL SISFOTENIKA
Vol. 15 No. 2, Juli 2025
p-ISSN: 2087-7897 ; e-ISSN : 2460-5344
DOI : 10.30700/sisfotenika.v15i1.569

Model Prediksi Kanker Paru-Paru dengan Random Forest

Lung Cancer Prediction Model with Random Forest

Adha Maulana*[1], Aditya Pratama[2], Denny Primanda[3], Nurman Hariyanto[4]
[1,2,3,4]Universitas Nahdlatul Ulama Kalimantan Barat
[3]Jurusan Sistem Informasi, Fakultas Teknik UNU Kalbar, Kubu Raya
e-mail: ***[1]adhamaulana@unukalbar.ac.id**, [2]adityapratamabadra@unukalbar.ac.id,
[3]dennyprimanda@unukalbar.ac.id, [4]nurmanhariyanto@unukalbar.ac.id

Abstrak

Kanker paru-paru merupakan penyakit serius dengan morbiditas dan mortalitas tinggi, seringkali terdiagnosis pada stadium lanjut karena gejala awal yang tidak spesifik, sehingga deteksi dini menjadi krusial. Penelitian ini bertujuan untuk membangun model prediksi kanker paru-paru menggunakan algoritma Random Forest guna mendukung identifikasi dini individu berisiko tinggi. Data yang digunakan diperoleh dari dataset 'survey lung cancer' yang tersedia di Kaggle,. Dataset awal terdiri dari 309 entri, yang mencakup berbagai faktor risiko dan gejala. Proses diawali dengan pemuatan dan pra-pemrosesan data menggunakan bahasa pemrograman Python dan pustaka seperti Pandas dan Scikit-learn,, termasuk penanganan nilai yang hilang (tidak ditemukan) dan penghapusan 33 entri duplikat, menghasilkan 276 entri unik, serta encoding variabel kategorikal. Data kemudian dibagi menjadi set pelatihan dan pengujian. Model Random Forest dilatih pada set pelatihan dan dievaluasi menggunakan metrik akurasi dan classification report. Hasil evaluasi menunjukkan akurasi model sebesar 90.36%, menunjukkan bahwa model Random Forest mampu memprediksi kanker paru-paru dengan akurasi yang tinggi. Analisis feature importance juga dilakukan untuk mengidentifikasi faktor-faktor paling berpengaruh dalam prediksi kanker paru-paru. Penelitian ini menunjukkan potensi penggunaan Random Forest dalam pengembangan sistem prediksi kanker paru-paru.

Kata kunci— *prediksi, model, kanker paru-paru, random forest*

Abstract

Lung cancer is a serious disease with high morbidity and mortality rates, often diagnosed at advanced stages due to nonspecific early symptoms, making early detection crucial. This study aims to build a lung cancer prediction model using the Random Forest algorithm to support the early identification of high-risk individuals. The data used was obtained from the 'survey lung cancer' dataset available on Kaggle. The initial dataset consisted of 309 entries, which included various risk factors and symptoms. The process began with loading and preprocessing the data using the Python programming language and libraries such as Pandas and Scikit-learn, including handling missing values and removing 33 duplicate entries, resulting in 276 unique entries, as well as encoding categorical variables. The data was then divided into training and testing sets. The Random Forest model was trained on the training set and evaluated using accuracy metrics and a classification report. The evaluation results showed a model accuracy of 90.36%,

indicating that the Random Forest model is capable of predicting lung cancer with high accuracy. Feature importance analysis was also conducted to identify the most influential factors in lung cancer prediction. This research demonstrates the potential use of Random Forest in the development of lung cancer prediction systems.

Keywords— *prediction, model, lung cancer, random forest*

1. PENDAHULUAN

Kanker paru-paru merupakan penyakit serius yang mengancam kesehatan global, ditandai dengan morbiditas dan mortalitas yang tinggi. Pada tahun 2022, kanker paru-paru menyumbang 12,5% dari total kasus kanker di dunia, dengan sekitar 2.480.675 kasus tercatat dalam lima tahun (2018-2022) secara global . Di Indonesia, kanker paru-paru menjadi kanker kedua terbanyak setelah kanker payudara, dengan sekitar 38.904 kasus tercatat dalam lima tahun (2018-2022) berdasarkan data Globocan . Kanker paru-paru didefinisikan sebagai kondisi di mana zat karsinogen memicu pertumbuhan dan pembelahan sel yang tidak terkendali di dalam paru-paru. Strategi esensial untuk meminimalisir angka kematian akibat kanker ini adalah dengan berfokus pada pencegahan dan deteksi dini [1].

Gejala awal kanker paru-paru seringkali tidak spesifik, menyebabkan sebagian besar penderita mengabaikannya sebagai gangguan pernapasan umum, yang berujung pada penundaan diagnosis dan pengobatan yang tepat [2]. Kurangnya kesadaran di kalangan profesional kesehatan untuk melakukan pemeriksaan lanjutan juga memperburuk situasi ini, meningkatkan tingkat keparahan dan mortalitas penyakit [3]. Kanker paru-paru umumnya terkait dengan kebiasaan merokok dan gaya hidup tidak sehat, menjadikannya jenis kanker ketiga tersering di Indonesia. Mayoritas kematian akibat kanker paru-paru berhubungan dengan merokok dan paparan asap rokok, di mana merokok secara signifikan merusak paru-paru. Merokok juga menjadi salah satu kebiasaan yang sulit dihentikan. Gangguan pada paru-paru ini dapat mengurangi efisiensi organ dalam menyerap oksigen dari udara. Oleh karena itu, deteksi dini dan intervensi cepat sangat krusial dalam mengurangi angka kematian akibat kanker paru-paru.

Dalam konteks ini, teknologi *machine learning* menawarkan solusi yang menjanjikan untuk memodelkan dan memprediksi pola penyakit [4][5]. Perkembangan teknologi yang pesat telah mendorong peningkatan volume pertukaran data, memicu kebutuhan akan *data mining* sebagai teknik untuk mengekstrak pola dan informasi berharga dari dataset besar [6]. *Data mining* semakin banyak diterapkan dalam sektor perawatan kesehatan, khususnya untuk penelitian dan penanganan kanker[7]. Klasifikasi, sebagai bentuk analisis data, menggunakan model untuk memprediksi label kelas, menjadikannya sangat relevan untuk menganalisis data gejala awal guna mendukung pengambilan keputusan medis [8][9] .

Meskipun demikian, sebagian besar penelitian yang ada cenderung berfokus pada analisis citra medis seperti CT scan yang memerlukan infrastruktur dan keahlian khusus, serta seringkali kurang dapat diakses untuk skrining awal berbasis gejala umum. Di sisi lain, studi yang menggunakan data survei gejala dan faktor risiko, meskipun lebih mudah dikumpulkan, seringkali menghadapi tantangan seperti ukuran dataset yang terbatas yang membatasi generalisasi, masalah ketidakseimbangan kelas, atau kurangnya interpretasi model yang jelas, yang dapat mengurangi kepercayaan di kalangan profesional kesehatan dan pasien.

Oleh karena itu, terdapat celah penelitian yang signifikan dalam pengembangan model prediksi kanker paru-paru yang tidak hanya akurat tetapi juga mudah diinterpretasikan dan dapat diimplementasikan menggunakan data gejala dan faktor risiko yang lebih mudah diakses. Urgensi penelitian ini muncul dari kebutuhan mendesak akan alat skrining awal yang efektif dan dapat dipercaya, yang dapat membantu mengidentifikasi individu berisiko tinggi sebelum penyakit berkembang ke stadium lanjut, sehingga memungkinkan intervensi cepat dan berpotensi mengurangi angka kematian. Dalam studi ini, kami mengusulkan penggunaan algoritma Random

Forest untuk memprediksi kanker paru-paru. Random Forest dikenal sebagai metode yang efektif karena kemampuannya dalam menangani data non-linear, memiliki kapabilitas generalisasi yang baik, serta menawarkan kinerja tinggi dan kemudahan interpretasi dalam konteks medis [10]. Penelitian ini bertujuan untuk mengembangkan model prediksi kanker paru-paru menggunakan metode Random Forest, dengan pemilihan fitur berdasarkan gejala-gejala dan faktor risiko yang umum terjadi di Indonesia.

Melalui penelitian ini, kami berharap dapat memberikan kontribusi signifikan terhadap manajemen kanker paru-paru di Indonesia dan menyediakan alat yang berguna bagi pembuat kebijakan serta masyarakat umum dalam menghadapi isu kesehatan yang semakin menantang ini. Kami mengantisipasi bahwa temuan studi ini akan membantu dalam perumusan kebijakan pengendalian kanker paru-paru yang lebih efektif dan meningkatkan kesadaran publik akan pentingnya deteksi dini. Selain itu, penelitian ini juga bertujuan untuk memperkaya pengetahuan tentang penerapan teknik *machine learning* dalam pemantauan kesehatan dan mengeksplorasi potensi Random Forest untuk meningkatkan prediksi penyakit. Dengan model yang akurat dan andal, kami berharap hasil ini dapat mendorong implementasi sistem pemantauan kesehatan yang lebih canggih dan responsif, yang tidak hanya memberikan informasi kepada publik tetapi juga mendukung upaya pencegahan dan penanganan penyakit secara lebih efektif. Pada akhirnya, kami berharap penelitian ini akan menjadi fondasi yang kuat untuk studi di masa mendatang dalam bidang prediksi penyakit dan *machine learning*, serta membuka pintu bagi studi interdisipliner yang dapat memanfaatkan data dan teknologi untuk meningkatkan kesehatan masyarakat.

2. METODE PENELITIAN

Kecerdasan Buatan (AI) secara cepat membentuk kembali penelitian kanker dan perawatan klinis yang dipersonalisasi. Ketersediaan dataset berdimensi tinggi, ditambah dengan kemajuan dalam komputasi kinerja tinggi dan arsitektur *deep learning* yang inovatif, telah menyebabkan peningkatan pesat penggunaan AI dalam berbagai aspek penelitian onkologi [10]. Aplikasi ini mencakup deteksi dan klasifikasi kanker, karakterisasi molekuler tumor dan lingkungan mikro mereka, penemuan dan penggunaan kembali obat, hingga prediksi hasil pengobatan untuk pasien. Seiring dengan penetrasi kemajuan ini ke dalam klinik, diperkirakan akan terjadi pergeseran paradigma dalam perawatan kanker yang akan sangat didorong oleh AI. Kanker adalah penyakit yang disebabkan oleh perubahan sel yang menyebabkan pertumbuhan dan pembelahan sel tidak terkendali. Dalam konteks Indonesia, kasus kanker paru mencapai 8,6% atau 30.023 kasus, dengan angka kematian 12,6% atau 26.095 kematian akibat kanker paru-paru.

Studi oleh [11] mengeksplorasi pemanfaatan teknologi Kecerdasan Buatan (AI) dalam meningkatkan efisiensi proses diagnostik medis. Penelitian ini menunjukkan bahwa AI mampu memproses volume besar data pasien dengan cepat dan akurat, mendukung profesional medis dalam pengambilan keputusan klinis, serta mempercepat diagnosis, mengurangi beban kerja dokter, dan meningkatkan akurasi serta kualitas layanan kesehatan. Namun, tantangan seperti integrasi sistem, privasi data, dan pertimbangan etis juga disoroti sebagai aspek yang harus diatasi untuk implementasi AI yang luas dalam bidang medis.

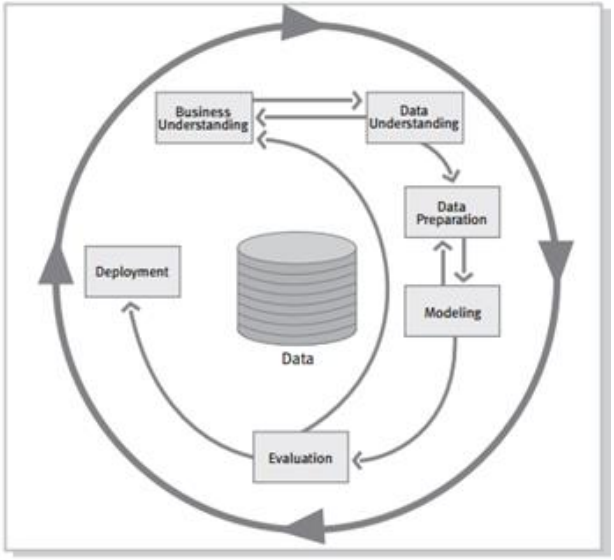
Dalam konteks deteksi dini kanker, penelitian oleh [12] berjudul "Early Diagnosis of Breast Cancer Prediction using Random Forest Classifier" menunjukkan potensi sistem deteksi penyakit otomatis. Studi ini menjelaskan bagaimana kanker payudara dapat diprediksi menggunakan teknik *machine learning* Random Forest Classifier, yang menyusun data menjadi banyak pohon dan menghasilkan keputusan akhir apakah seseorang berisiko menderita kanker payudara atau tidak. Model ini dilaporkan memiliki akurasi 98%, menawarkan respons cepat, keandalan, dan efektivitas dalam mengurangi risiko kematian.

Menurut penelitian [13] dengan judul "Penerapan Machine Learning dengan Algoritma Logistik Regresi untuk Memprediksi Diabetes" menunjukkan potensi *machine learning* dalam memprediksi penyakit kronis seperti diabetes. Studi ini menggarisbawahi bagaimana algoritma regresi logistik dapat digunakan untuk mengidentifikasi kemungkinan adanya diabetes berdasarkan data pasien, yang relevan dengan upaya prediksi penyakit di bidang kesehatan.

Metodologi CRISP-DM (Cross-Industry Standard Process for Data Mining) adalah kerangka kerja yang terstruktur dan diterima secara luas untuk proyek data mining, yang memandu proses analisis dan prediksi melalui enam fase utama: Pemahaman Bisnis, Pemahaman Data, Persiapan Data, Pemodelan, Evaluasi, dan Penyebaran. Pendekatan ini memastikan pemahaman menyeluruh terhadap tujuan bisnis, kualitas data, pemilihan dan pembangunan model yang tepat, serta evaluasi kinerja yang komprehensif, sebelum akhirnya model diintegrasikan ke dalam sistem yang dapat digunakan oleh pemangku kepentingan[14].

3. HASIL DAN PEMBAHASAN

Dalam penelitian ini, metodologi CRISP-DM (Cross-Industry Standard Process for Data Mining) digunakan sebagai pendekatan terstruktur untuk memandu proses analisis dan prediksi kanker paru-paru menggunakan algoritma Random Forest . CRISP-DM adalah kerangka kerja yang diterima secara luas yang menyediakan model proses komprehensif untuk proyek data mining, terdiri dari enam fase utama: Pemahaman Bisnis (Business Understanding), Pemahaman Data (Data Understanding), Persiapan Data (Data Preparation), Pemodelan (Modeling), Evaluasi (Evaluation), dan Penyebaran (Deployment) [14][15].



Gambar 1. Tahapan CRISP-DM

Pemahaman Bisnis (Business Understanding)

Fase pertama CRISP-DM melibatkan pemahaman tujuan proyek dan persyaratan dari perspektif bisnis. Dalam penelitian ini, tujuan utamanya adalah memprediksi kemungkinan adanya kanker paru-paru berdasarkan gejala awal dan faktor risiko, untuk mendukung kebijakan kesehatan masyarakat dan pengambilan keputusan klinis. Fase ini mencakup identifikasi pemangku kepentingan utama, seperti lembaga kesehatan dan profesional medis, serta pemahaman kebutuhan mereka akan prediksi yang akurat untuk meningkatkan proses pengambilan keputusan terkait manajemen penyakit.

Pemahaman Data (Data Understanding)

Setelah fase pemahaman bisnis, langkah selanjutnya adalah mengumpulkan data yang relevan untuk mendukung analisis. Fase ini melibatkan pengumpulan data historis yang mencakup berbagai parameter seperti gejala awal (misalnya, kesulitan bernapas, batuk berdarah, batuk tak kunjung sembuh, nyeri dada/perut, penurunan berat badan, suara serak, kesulitan menelan, nyeri bahu/dada/lengan, bronkitis, pneumonia) dan faktor risiko (misalnya, merokok,

polusi udara, makanan dan minuman, zat kimia, pekerjaan, riwayat penyakit paru-paru/kanker turunan, kurang berolahraga, konsumsi alkohol, riwayat penyakit kanker paru-paru). Sumber data mencakup kuesioner yang disebarkan kepada responden yang telah didiagnosis kanker paru-paru dalam tiga tahun terakhir, serta kelompok kontrol non-pasien. Selama fase ini, analisis awal dilakukan untuk menilai kualitas data, mengidentifikasi nilai yang hilang, dan memahami distribusi fitur-fitur yang berbeda.

Persiapan Data (Data Preparation)

Persiapan data adalah langkah krusial dalam kerangka CRISP-DM yang melibatkan pembersihan dan transformasi data ke dalam format yang sesuai untuk analisis. Dalam penelitian ini, data yang terkumpul menjalani beberapa langkah pra-pemrosesan, termasuk penanganan nilai yang hilang, normalisasi data, dan pengodean variabel kategorikal. Selain itu, analisis data eksplorasi (EDA) dilakukan untuk memvisualisasikan hubungan antara fitur-fitur yang berbeda dan untuk mengidentifikasi tren atau pola yang dapat menginformasikan proses pemodelan. Fase ini memastikan bahwa data siap untuk fase pemodelan, meningkatkan akurasi dan efisiensi prediksi.

Pemodelan (Modeling)

Fase pemodelan melibatkan pemilihan teknik pemodelan yang sesuai dan pembangunan model prediktif menggunakan Random Forest. Berbagai konfigurasi Random Forest diuji, termasuk penyetelan hyperparameter untuk mengoptimalkan kinerja model. Model dilatih menggunakan sebagian dari dataset, sementara sisanya dicadangkan untuk validasi. Teknik cross-validation digunakan untuk memastikan bahwa model kuat dan dapat digeneralisasi dengan baik ke data yang belum terlihat. Fase ini berpuncak pada pemilihan model Random Forest berkinerja terbaik berdasarkan metrik evaluasi seperti akurasi, presisi, recall, dan f1-score.

Evaluasi (Evaluation)

Setelah model dibangun, fase evaluasi menilai kinerjanya dalam memprediksi kanker paru-paru. Model Random Forest yang dipilih dievaluasi terhadap dataset uji terpisah untuk menentukan akurasi dan keandalannya. Fase ini melibatkan analisis prediksi model terhadap diagnosis aktual dan penghitungan metrik kinerja. Evaluasi tidak hanya membantu dalam memahami efektivitas model tetapi juga memberikan wawasan tentang area potensial untuk peningkatan, seperti pemilihan fitur atau penggabungan sumber data tambahan.

Penerapan (Deployment)

Akhirnya, fase penerapan berfokus pada integrasi model prediktif ke dalam sistem yang mudah digunakan untuk pemangku kepentingan. Ini melibatkan pembuatan dasbor atau aplikasi yang menampilkan prediksi, peringatan, dan tren historis terkait risiko kanker paru-paru secara real-time. Fase penerapan juga mencakup dokumentasi model dan prosesnya untuk memastikan bahwa pemangku kepentingan dapat secara efektif memanfaatkan sistem untuk pengambilan keputusan yang terinformasi mengenai manajemen kesehatan.

Kerangka CRISP-DM menyediakan pendekatan terstruktur untuk melakukan penelitian ini tentang prediksi kanker paru-paru menggunakan Random Forest. Dengan mengikuti fase-fase CRISP-DM, penelitian ini memastikan analisis yang komprehensif dan mengembangkan model prediktif yang andal yang dapat mendukung manajemen kesehatan yang efektif dan berkontribusi pada inisiatif kesehatan masyarakat. Metodologi ini tidak hanya memfasilitasi aspek teknis penelitian tetapi juga menyelaraskan proyek dengan kebutuhan pemangku kepentingan, menjadikannya alat yang berharga untuk mengatasi tantangan kesehatan.

Dalam penelitian ini, peneliti menggunakan metodologi CRISP-DM untuk mengimplementasikan algoritma random forest dalam memprediksi kanker paru-paru. Proses penelitian ini terbagi dalam enam fase utama, yaitu Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, dan Deployment

Pemahaman Bisnis (Business Understanding)

Pada tahap pemahaman bisnis, peneliti mengidentifikasi masalah, menentukan tujuan penelitian, dan memahami konteks terkait kanker paru-paru. Peneliti menganalisis data survei yang mencakup gejala dan faktor risiko, mempelajari literatur dan temuan medis terkait kanker

paru-paru, serta mengidentifikasi pemangku kepentingan yang terlibat, seperti pasien, profesional medis, dan lembaga kesehatan. Tujuan dari tahap ini adalah untuk mendapatkan pemahaman komprehensif mengenai isu kanker paru-paru dan bagaimana solusi analitis dapat memberikan nilai tambah. Tujuan dari penelitian ini adalah membangun model prediksi kanker paru-paru, memanfaatkan algoritma Random Forest, untuk membantu identifikasi individu berisiko tinggi, sehingga mendukung upaya deteksi dini dan kebijakan kesehatan masyarakat yang lebih efektif. Dengan kata lain, peneliti mengumpulkan informasi dan mendefinisikan masalah secara jelas sebelum melanjutkan ke tahap berikutnya.

Pemahaman Data (Data Understanding)

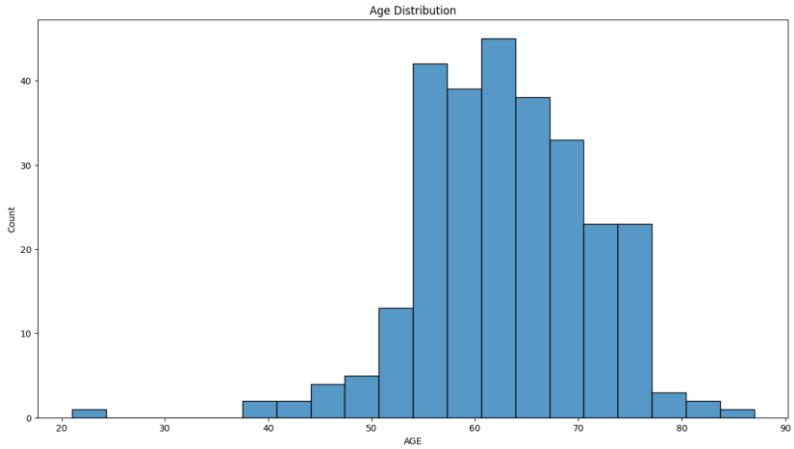
Pada tahap ini, data survei terkait faktor risiko dan gejala kanker paru-paru dikumpulkan. Dataset awal terdiri dari 309 entri dengan 16 kolom yang mencakup informasi demografis (seperti GENDER dan AGE) serta berbagai faktor risiko dan gejala (seperti SMOKING, YELLOW_FINGERS, ANXIETY, PEER_PRESSURE, CHRONIC DISEASE, FATIGUE, ALLERGY, WHEEZING, ALCOHOL CONSUMING, COUGHING, SHORTNESS OF BREATH, SWALLOWING DIFFICULTY, CHEST PAIN) dan variabel target (LUNG_CANCER).

```
data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 309 entries, 0 to 308
Data columns (total 16 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   GENDER                 309 non-null    int64
 1   AGE                    309 non-null    int64
 2   SMOKING                309 non-null    int64
 3   YELLOW_FINGERS         309 non-null    int64
 4   ANXIETY                309 non-null    int64
 5   PEER_PRESSURE          309 non-null    int64
 6   CHRONIC DISEASE        309 non-null    int64
 7   FATIGUE                309 non-null    int64
 8   ALLERGY                309 non-null    int64
 9   WHEEZING               309 non-null    int64
 10  ALCOHOL CONSUMING      309 non-null    int64
 11  COUGHING               309 non-null    int64
 12  SHORTNESS OF BREATH    309 non-null    int64
 13  SWALLOWING DIFFICULTY  309 non-null    int64
 14  CHEST PAIN             309 non-null    int64
 15  LUNG_CANCER            309 non-null    int64
dtypes: int64(16)
memory usage: 38.8 KB
```

Gambar 2. Info Dataset

Analisis awal dilakukan untuk memahami struktur dan karakteristik data. Pemeriksaan informasi dataset (data.info()) menunjukkan bahwa semua kolom memiliki tipe data integer (int64) dan tidak ada nilai yang hilang pada setiap kolom terlihat pada gambar 2. Hal ini mengindikasikan kelengkapan data pada setiap fitur.



Gambar 3. Distribusi Umur

Eksplorasi data visual juga dilakukan untuk mendapatkan wawasan awal mengenai distribusi data dan hubungan antar variabel. Distribusi usia pada gambar 3 responden divisualisasikan menggunakan histogram (sns.histplot). Selain itu, korelasi antar variabel ditampilkan menggunakan heatmap (sns.heatmap(data.corr(), annot=True)) untuk mengidentifikasi potensi hubungan linear antar fitur dan antara fitur dengan variabel target (LUNG_CANCER). Fase pemahaman data ini memberikan landasan yang kuat mengenai kondisi data dan isu-isu yang perlu ditangani pada tahapan selanjutnya.

Persiapan Data (Data Preparation)

Fase persiapan data merupakan tahap krusial untuk memastikan kualitas dan format data sesuai untuk pemodelan. Berdasarkan analisis awal, dataset yang digunakan tidak memiliki nilai yang hilang (missing values), sebagaimana diverifikasi melalui pemeriksaan. Namun, ditemukan adanya data duplikat. Sebanyak 33 entri duplikat teridentifikasi. Untuk menjaga integritas dataset dan menghindari bias dalam pelatihan model, entri-entri duplikat ini dihapus.

```
[ ]  from sklearn.preprocessing import LabelEncoder

     encoder = LabelEncoder()
     data['GENDER'] = encoder.fit_transform(data['GENDER'])

     data['LUNG_CANCER'] = encoder.fit_transform(data['LUNG_CANCER'])
```

Gambar 4. Encoding

Selanjutnya, dataset mengandung variabel kategorikal ('GENDER' dan 'LUNG_CANCER') yang perlu diubah menjadi format numerik agar dapat diproses oleh algoritma machine learning. Transformasi ini dilakukan menggunakan Label Encoding. Variabel 'GENDER' diubah menjadi representasi numerik (0 dan 1). Variabel target 'LUNG_CANCER' juga di-encode menjadi nilai numerik (0 dan 1). Setelah tahapan persiapan data ini, dataset siap untuk digunakan dalam fase pemodelan.

Pemodelan (Modeling)

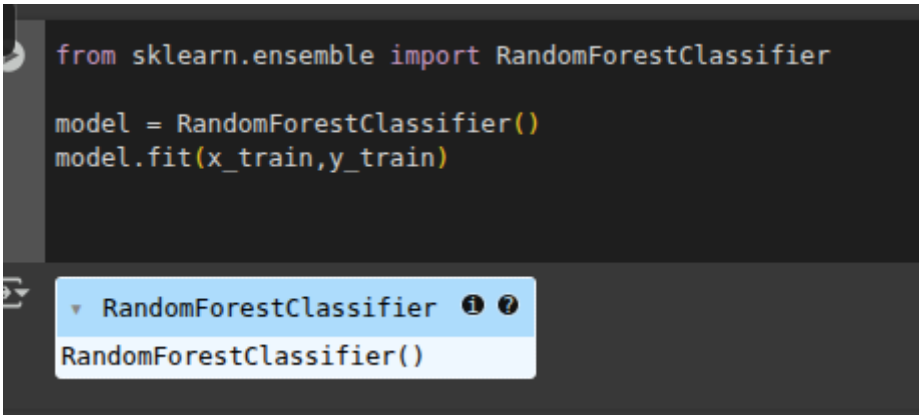
Fase Pemodelan dalam CRISP-DM melibatkan pemilihan teknik pemodelan, pembangunan model, dan kalibrasi parameternya. Dalam penelitian ini, algoritma Random Forest dipilih untuk membangun model prediksi kanker paru-paru. Pendekatan ini membantu mengurangi overfitting dan meningkatkan robustnes model.

```
[ ]  from sklearn.model_selection import train_test_split

     x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.3, random_state=42)
```

Gambar 5. Data Train

Sebelum melatih model, dataset yang telah dipersiapkan dibagi menjadi dua bagian: set pelatihan (training set) dan set pengujian (testing set) tampak pada gambar 5. Pembagian ini dilakukan menggunakan fungsi train_test_split dari library sklearn.model_selection. Proporsi pembagian ditetapkan sebesar 70% untuk data pelatihan dan 30% untuk data pengujian (test_size=0.3). Pembagian ini dilakukan secara acak, namun dengan random_state yang ditetapkan (misalnya, 42) untuk memastikan bahwa pembagian data konsisten dan dapat direproduksi. Variabel independen (fitur) disimpan dalam variabel x, yang merupakan data asli tanpa kolom target 'LUNG_CANCER', sedangkan variabel dependen (target) disimpan dalam variabel y, yang hanya berisi kolom 'LUNG_CANCER'.



Gambar 6. Data Train

Model Random Forest diinisialisasi (RandomForestClassifier()) dengan menggunakan pengaturan default tampak pada gambar 6. Model ini kemudian dilatih menggunakan data pelatihan (x_train dan y_train) melalui metode model.fit(x_train, y_train). Proses pelatihan ini melibatkan pembangunan ensemble decision tree berdasarkan pola dan hubungan dalam data pelatihan untuk mempelajari cara memprediksi variabel target. Setelah proses pelatihan selesai, model siap untuk dievaluasi pada data yang belum pernah dilihat sebelumnya, yaitu set pengujian.

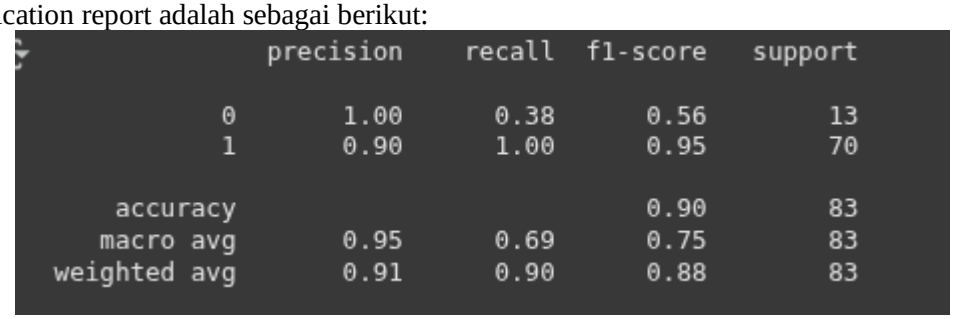
Evaluasi (Evaluation)

Fase Evaluasi dalam CRISP-DM bertujuan untuk menilai kualitas dan performa model yang telah dibangun serta menentukan apakah model tersebut memenuhi tujuan bisnis/riset yang ditetapkan pada fase awal. Pada tahap ini, model Random Forest yang telah dilatih digunakan untuk membuat prediksi pada set pengujian (x_test), yang merupakan data yang belum pernah dilihat oleh model selama proses pelatihan. Hasil prediksi ini disimpan dalam variabel y_pred.

Evaluasi performa model dilakukan menggunakan beberapa metrik standar untuk klasifikasi biner:

Akurasi (Accuracy): Mengukur proporsi total prediksi yang benar (baik True Positives maupun True Negatives) dari seluruh jumlah kasus. Akurasi dihitung menggunakan accuracy_score dari library sklearn.metrics. Berdasarkan hasil eksekusi notebook, akurasi model adalah 90.36%. Ini menunjukkan bahwa model secara keseluruhan mampu memprediksi status kanker paru-paru dengan benar pada sekitar 90.36% kasus dalam set pengujian.

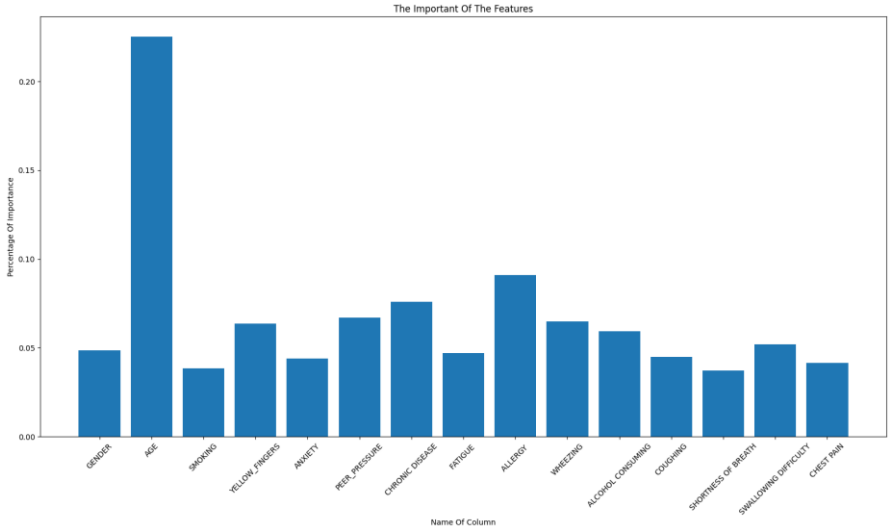
Classification Report: Menyediakan metrik evaluasi yang lebih rinci untuk setiap kelas (dalam hal ini, kelas 0 untuk 'Tidak Kanker' dan kelas 1 untuk 'Kanker Paru-paru'), termasuk Presisi (Precision), Recall (Sensitivity), F1-Score, dan Support (jumlah kasus aktual per kelas). Classification report dihitung menggunakan classification_report dari sklearn.metrics. Hasil classification report adalah sebagai berikut:



	precision	recall	f1-score	support
0	1.00	0.38	0.56	13
1	0.90	1.00	0.95	70
accuracy			0.90	83
macro avg	0.95	0.69	0.75	83
weighted avg	0.91	0.90	0.88	83

Gambar 7. Confusion Matrix

Dari classification report, dapat diamati bahwa model memiliki presisi 1.00 dan recall 0.38 untuk kelas 0 (Tidak Kanker), serta presisi 0.90 dan recall 1.00 untuk kelas 1 (Kanker Paru-paru). Recall 1.00 untuk kelas 1 (Kanker Paru-paru) sangat penting dalam konteks medis, karena ini berarti model berhasil mengidentifikasi semua kasus positif kanker paru-paru dalam set pengujian (tidak ada False Negatives). Presisi 0.90 untuk kelas 1 menunjukkan bahwa 90% dari kasus yang diprediksi sebagai kanker paru-paru memang benar positif. Sedangkan untuk kelas 0, presisi 1.00 berarti semua yang diprediksi negatif memang benar negatif, namun recall 0.38 menunjukkan model hanya mengidentifikasi 38% dari total kasus negatif yang sebenarnya. Fokus pada recall yang tinggi untuk kelas positif (kanker) seringkali menjadi prioritas dalam deteksi penyakit untuk meminimalkan risiko False Negatives.



Gambar 8. Feature Importances

Selain evaluasi performa model secara keseluruhan, analisis pentingnya fitur (feature importance) juga dilakukan pada tahap ini. Hasil analisis model.feature_importances_ menunjukkan kontribusi relatif setiap fitur dalam proses prediksi model Random Forest. Fitur-fitur dengan nilai feature importance tertinggi dianggap paling berpengaruh dalam menentukan prediksi kanker paru-paru oleh model. Visualisasi feature importance (plt.bar(column, important)) memberikan gambaran jelas mengenai fitur mana yang paling dominan. Berdasarkan visualisasi, Age, Allergy, Chronic Disease tampak menjadi fitur yang paling penting dalam model ini.

Secara keseluruhan, hasil evaluasi menunjukkan bahwa model Random Forest memiliki performa yang baik dalam memprediksi kanker paru-paru berdasarkan dataset yang digunakan, dengan kekuatan utama pada kemampuannya mengidentifikasi kasus positif (recall tinggi untuk kelas 1).

4. KESIMPULAN

Penelitian ini telah berhasil mengembangkan model prediksi kanker paru-paru menggunakan algoritma Random Forest berdasarkan data survei faktor risiko dan gejala. Mengikuti metodologi CRISP-DM, penelitian melalui tahapan pemahaman bisnis, pemahaman data, persiapan data, pemodelan, dan evaluasi. Data survei yang telah dipersiapkan, termasuk penanganan duplikat dan encoding variabel kategorikal, digunakan untuk melatih dan menguji model. Evaluasi model Random Forest pada set pengujian menunjukkan performa yang menjanjikan, dengan akurasi keseluruhan sebesar 90.36%. Secara spesifik, model menunjukkan kemampuan yang sangat baik dalam mengidentifikasi kasus positif kanker paru-paru, sebagaimana tercermin dari nilai recall yang tinggi untuk kelas positif dalam classification report.

Hasil penelitian ini menunjukkan potensi penerapan algoritma Random Forest sebagai alat bantu dalam identifikasi dini individu yang berisiko tinggi terkena kanker paru-paru. Analisis feature importance juga memberikan wawasan mengenai faktor-faktor risiko dan gejala yang paling berpengaruh dalam prediksi model, yang dapat menjadi informasi berharga untuk upaya pencegahan dan edukasi kesehatan masyarakat. Meskipun model menunjukkan performa yang baik pada dataset yang digunakan, penelitian lanjutan dengan dataset yang lebih besar dan beragam, serta eksplorasi algoritma lain dan teknik optimasi model, dapat dilakukan untuk meningkatkan generalisasi dan robustnes model prediksi kanker paru-paru di masa mendatang. Penelitian ini diharapkan dapat menjadi landasan bagi pengembangan sistem pendukung keputusan klinis yang lebih canggih.

5. SARAN

Penelitian ini membuktikan bahwa penerapan algoritma Random Forest dalam membangun model prediksi kanker paru-paru memiliki potensi signifikan dalam mendukung deteksi dini penyakit secara efisien di ranah teknologi informasi. Untuk memperkuat keandalan model, disarankan penerapan metode validasi silang (cross-validation) serta perbandingan performa dengan algoritma lain seperti SVM atau XGBoost. Peningkatan kualitas data dapat dilakukan melalui tahapan preprocessing yang tepat dan pemilihan fitur-fitur penting seperti usia pasien dan riwayat merokok. Evaluasi model perlu mencakup metrik kinerja yang lebih komprehensif seperti precision, recall, F1-score, dan AUC-ROC, agar dampak dari kesalahan prediksi (false positive maupun false negative) dapat dianalisis secara mendalam. Pengembangan lanjutan dapat diarahkan pada pembuatan aplikasi berbasis web atau mobile yang mendukung penggunaan prediksi secara langsung dan terintegrasi dengan sistem informasi rumah sakit. Di samping itu, perlindungan data medis serta pertimbangan etis dalam pemanfaatan kecerdasan buatan harus menjadi prioritas utama. Secara umum, hasil penelitian ini berkontribusi terhadap pengembangan sistem pendukung keputusan medis dan mendukung digitalisasi layanan kesehatan secara menyeluruh.

DAFTAR PUSTAKA

[1] Y. You *et al.*, “Artificial intelligence in cancer target identification and drug discovery,” *Signal Transduct. Target. Ther.*, vol. 7, no. 1, pp. 1–24, 2022, doi: 10.1038/s41392-022-00994-0.

[2] B. Bhinder, C. Gilvary, N. S. Madhukar, and O. Elemento, “Artificial Intelligence in Cancer Research and Precision Medicine,” *AACR Journals*, vol. 11, no. 4, pp. 900–915, 2021, doi: https://doi.org/10.1158/2159-8290.CD-21-0090.

[3] H. Shimizu and K. I. Nakayama, “Artificial Intelligence in Oncology,” *Cancer Sci.*, vol. 111, no. 5, pp. 1452–1460, 2000, doi: https://doi.org/10.1111/cas.14377.

[4] W. L. Bi *et al.*, “Artificial intelligence in cancer imaging: Clinical challenges and applications,” *CA. Cancer J. Clin.*, vol. 69, no. 2, pp. 127–157, 2019, doi: 10.3322/caac.21552.

[5] D. Ho, “Artificial Intelligence in Cancer Therapy,” *Science (80-.).*, vol. 367, no. 6481, pp. 982–983, 2020, doi: DOI:10.1126/science.aaz302.

[6] Z. Dlamini, F. Z. Francies, R. Hull, and R. Marima, “Artificial Intelligence (AI) and Big Data in Cancer and Precision Oncology,” *Comput. Struct. Biotechnol. J.*, vol. 18, pp. 2300–2311, 2020, doi: https://doi.org/10.1016/j.csbj.2020.08.019.

[7] M. J. Iqbal *et al.*, “Clinical applications of artificial intelligence and machine learning in cancer diagnosis: looking into the future,” *Cancer Cell Int.*, vol. 21, no. 1, pp. 1–11, 2021, doi: 10.1186/s12935-021-01981-1.

[8] Y. Kumar, S. Gupta, R. Singla, and Y.-C. Hu, “A Systematic Review of Artificial Intelligence Techniques in Cancer Prediction and Diagnosis,” *Arch. Comput. Methods Eng.*, vol. 29, pp. 2043–2070, 2022, doi: doi: 10.1007/S11831-021-09648-W.

[9] R. Perez-Lopez, N. G. Laleh, F. Mahmood, and J. N. Kather, "A guide to artificial intelligence for cancer researchers," *Nat. Rev. Cancer*, vol. 24, pp. 427–441, 2024, [Online]. Available: https://www.nature.com/articles/s41568-024-00694-7.

[10] C. Zhang *et al.*, "Novel research and future prospects of artificial intelligence in cancer diagnosis and treatment," *J. Hematol. Oncol.*, vol. 16, no. 1, pp. 1–29, 2023, doi: 10.1186/s13045-023-01514-5.

[11] P. K. Buatan and N. Husna, "Pemanfaatan Kecerdasan Buatan dalam Meningkatkan Efisiensi Diagnostik Medis," *J. Sains dan Teknol. Indones. STIKES Med. Nurul Islam.*, vol. 1, no. 1, pp. 1–4, 2025, doi: doi: 10.58477/sti.v1i1.282.

[12] P. R. Anisha, C. Kishor Kumar Reddy, K. Apoorva, and C. Meghana Mangipudi, "Early Diagnosis of Breast Cancer Prediction using Random Forest Classifier," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1116, no. 1, p. 012187, 2021, doi: 10.1088/1757-899x/1116/1/012187.

[13] A. Pratama, A. C. Nurcahyo, and L. Firgia, "Penerapan Machine Learning dengan Algoritma Logistik Regresi untuk Memprediksi Diabetes," *Pros. CORISINDO 2023*, pp. 116–121, 2023, [Online]. Available: https://stmikpontianak.org/ojs/index.php/corisindo/article/view/30%0Ahttps://stmikpontianak.org/ojs/index.php/corisindo/article/download/30/22.

[14] A. Pratama, A. Maulana, and R. A. Saputra, "Implementasi Algoritma Linear Regression Untuk Prediksi Harga Rumah di Daerah Tebet," *J. Inf. Technol.*, vol. 5, no. 1, 2025, doi: 10.46229/jifotech.v5i1.986.

[15] A. Pratama, C. Gudiato, A. Maulana, and W. Prayitno, "Prediksi ISPU Berbasis SVM untuk Masa Depan Jakarta yang Berkelanjutan SVM-Driven ISPU Prediction for Jakarta ' s Sustainable Future," vol. 15, no. 1, pp. 66–76, 2025.