**Jurnal Bumigora Information Technology (BITe)**
Vol. 7, No. 1, Juni 2025, Hal. 51∼62
E-ISSN: 2685-4066, DOI: 10.30812/bite.v7i1.4909 

# Perbandingan Metode Berbasis *Decision Tree* untuk Mendeteksi Penyakit Paru

# *Comparison of Decision Tree-Based Methods in Lung Disease Detection*

**Lely Kurniawati***, **Dadang Priyanto**, **Neny Sulistianingsih**, **Moch. Syahrir**, **Ria Rismayati**
Universitas Bumigora, Mataram, Indonesia

**Informasi Artikel:**
Diterima: 22 Januari 2025, Direvisi: 5 Februari 2025, Disetujui: 27 Juni 2025

**Abstrak-**
**Latar Belakang:** Penyakit paru-paru menjadi penyebab utama kematian global dengan lebih dari 4 juta kasus setiap tahun, termasuk 500.000 kasus baru di Indonesia yang mayoritas terdeteksi pada stadium lanjut.
**Tujuan:** Penelitian ini bertujuan membandingkan performa tiga algoritma *Decision Tree* XGBoost, C4.5, dan Random Forest dalam deteksi penyakit paru-paru, serta menentukan metode terbaik berdasarkan metrik evaluasi.
**Metode:** Sebanyak 30.000 sampel data dari Kaggle diproses melalui tahap pembersihan menggunakan metode IQR, pengkodean atribut kategorikal, serta pembagian data menjadi 80% untuk pelatihan dan 20% untuk pengujian. Model klasifikasi yang digunakan meliputi XGBoost, C4.5, dan Random Forest. Evaluasi kinerja model dilakukan menggunakan *confusion matrix*, akurasi, presisi, *recall*, dan *F1-score*.
**Hasil:** Hasil menunjukkan algoritma C4.5 memiliki performa terbaik dengan akurasi 94,33% dan *zero false negative*. XGBoost menyusul dengan akurasi 93,18%, sedangkan Random Forest terendah (90,07%).
**Kesimpulan:** Hasil menunjukkan algoritma C4.5 memiliki performa terbaik dengan akurasi 94,33% dan *zero false negative*. XGBoost menyusul dengan akurasi 93,18%, sedangkan Random Forest terendah (90,07%).

**Kata Kunci**: C4.5, *Decision Tree*, *Machine Learning*, Penyakit Paru-paru, *Random Forest*, XGBoost.

***Abstract-***
***Background:*** *Lung disease is a leading cause of death globally, with more than 4 million cases each year, including 500,000 new cases in Indonesia, most of which are detected at an advanced stage.*
***Objective:*** *This study aims to compare the performance of three decision tree algorithms, XGBoost, C4.5, and Random Forest, in detecting lung disease and to determine the best method based on evaluation metrics.*
***Methods:*** *: A total of 30,000 data samples from Kaggle were processed through a cleaning stage using the IQR method, categorical attribute coding, and data division into 80% for training and 20% for testing. The classification models used include XGBoost, C4.5, and Random Forest. Model performance evaluation used a confusion matrix, accuracy, precision, recall, and F1-score.*
***Result:*** *The results showed that the C4.5 algorithm had the best performance with an accuracy of 94.33% and zero false negatives. XGBoost followed with an accuracy of 93.18%, while Random Forest was the lowest (90.07%).*
***Conclusion:*** *These findings indicate that C4.5 has great potential in an accurate early detection system, helping to reduce the risk of misdiagnosis, especially in false negative cases, and supporting clinical decision making in health facilities.*

***Keywords****: C4.5, Decision Tree, Lung Disease, Machine Learning, Random Forest, XGBoost.*

***Penulis Korespondensi:***

Lely Kurniawati,
Program Studi Ilmu Komputer, Universitas Bumigora, Mataram, Indonesia,
Email: lelyykurniawatii88@gmail.com

***How to Cite***: L. Kurniawati, D. Priyanto, N. Sulistianingsih, M. Syahrir, dan R. Rismayati, "Perbandingan Metode Berbasis *Decision Tree* untuk Mendeteksi Penyakit Paru," *Jurnal Bumigora Information Technology (BITe)*, vol. 7, no. 1, pp. 51-62, Jun. 2025. doi: 10.30812/bite.v7i1.4909.

## 1. PENDAHULUAN

Penyakit paru-paru merupakan ancaman serius bagi kesehatan global, dengan angka kematian mencapai 4 juta jiwa setiap tahun [1]. Di Indonesia, lebih dari 500.000 kasus baru tercatat setiap tahun, dan mayoritas dari kasus tersebut baru terdeteksi pada stadium lanjut [2]. Keterlambatan diagnosis ini menekankan urgensi deteksi dini melalui pendekatan teknologi *machine learning*, yang memungkinkan sistem untuk mempelajari data dan mengenali pola penyakit secara otomatis tanpa perlu diprogram secara eksplisit. Berbagai studi menunjukkan bahwa algoritma *machine learning* mampu meningkatkan akurasi diagnosis, mempercepat proses deteksi, serta mengurangi tingkat kesalahan prediksi.

Di antara algoritma *machine learning*, pendekatan berbasis *decision tree* seperti XGBoost, C4.5, dan *Random Forest* banyak digunakan karena kemampuannya dalam menangani data medis yang kompleks dan menghasilkan model yang dapat diinterpretasikan. XGBoost dikenal efisien dalam menangani dataset besar dengan teknik *boosting* yang memperbaiki kelemahan model sebelumnya [3]. C4.5 menawarkan transparansi keputusan yang baik melalui perhitungan *gain ratio* [4]. Sementara *Random Forest* menggunakan pendekatan *ensemble* untuk meningkatkan akurasi dan mengurangi *overfitting* [5].

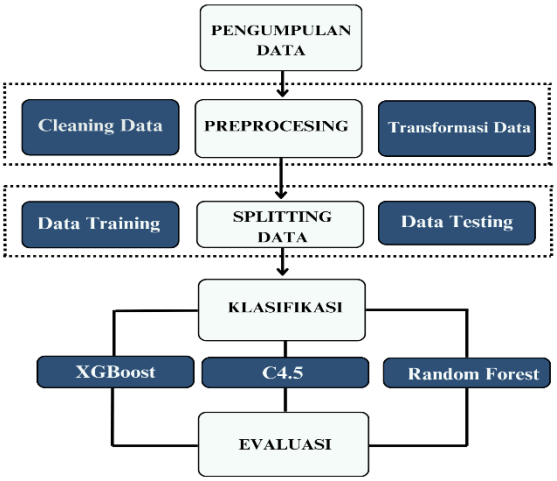
Namun demikian, meskipun ketiga algoritma tersebut telah banyak diterapkan di berbagai domain kesehatan, studi yang secara langsung membandingkan kinerja XGBoost, C4.5, dan *Random Forest* dalam konteks deteksi penyakit paru-paru masih sangat terbatas.

Beberapa penelitian terdahulu menunjukkan berbagai keterbatasan dalam pendekatan yang digunakan. Seperti penelitian [4] hanya membandingkan dua algoritma dengan keterbatasan data yang mencakup atribut demografi saja, sehingga model belum sepenuhnya akurat untuk diagnosis medis sebenarnya. Studi [6] juga terbatas pada perbandingan dua algoritma tanpa mengeksplorasi algoritma lain. Demikian pula dengan penelitian [7] yang menghadapi keterbatasan dalam generalisasi model, ketergantungan pada kualitas data, serta tantangan dalam *tuning* parameter yang optimal. Penelitian [5] menunjukkan kelemahan dalam evaluasi yang hanya berfokus pada akurasi tanpa mempertimbangkan metrik penting lainnya seperti sensitivitas, spesifisitas, dan *F1-score*. Studi [8] mengalami kendala berupa performa LSTM yang relatif rendah dengan kompleksitas tinggi dan risiko *overfitting*. Penelitian [9] memiliki keterbatasan generalisasi data terhadap kondisi masa depan yang lebih beragam. Sementara itu, studi [10] dinilai kurang lengkap dalam penggunaan metrik evaluasi, terutama pada dataset yang tidak seimbang. Penelitian [11] mengalami kendala data yang terbatas, kurangnya eksplorasi parameter, dan analisis variabel yang belum menyeluruh. Studi [12] menunjukkan berbagai kekurangan seperti akurasi rendah, evaluasi terbatas, *preprocessing* data yang minim, tanpa *tuning parameter*, dan analisis yang dangkal. Terakhir, penelitian [13] memiliki kelemahan dalam metodologi yang kurang optimal, analisis data yang dangkal, eksperimen yang tidak transparan, dan evaluasi yang tidak komprehensif sehingga mengurangi keandalan hasil . Terdapat kesenjangan atau gap penelitian ini dengan sebelumnya adalah belum adanya studi yang secara komprehensif membandingkan kinerja tiga algoritma *decision tree*: XGBoost, C4.5, dan *Random Forest* dalam mendeteksi penyakit paru-paru berbasis data non-citra secara sistematis dan terukur. Kebanyakan penelitian sebelumnya hanya membandingkan sebagian dari algoritma tersebut atau berfokus pada penyakit lain, serta belum menerapkan evaluasi performa yang menyeluruh menggunakan metrik seperti akurasi, presisi, *recall*, dan *F1-score*.

Penelitian ini berkontribusi dalam memberikan dasar empiris yang kuat tentang kelebihan dan kelemahan masing-masing algoritma dalam klasifikasi penyakit paru-paru, serta memberikan rekomendasi terhadap metode paling optimal yang dapat diimplementasikan dalam sistem deteksi dini di fasilitas pelayanan kesehatan. Dengan demikian, hasil penelitian ini diharapkan dapat menjadi acuan dalam pengembangan sistem pendukung keputusan medis berbasis *machine learning* yang lebih akurat, cepat, dan dapat diandalkan dalam mengurangi angka kematian akibat keterlambatan diagnosis. Berdasarkan hal tersebut, peneliti memutuskan untuk melakukan penelitian dengan judul "Perbandingan Metode Berbasis *Decision Tree* dalam Deteksi Penyakit Paru-Paru"

## 2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatatif komparatif untuk mengevaluasi dan membandingkan performa tiga algoritma *machine learning* berbasis *decision tree*, yaitu XGBoost, C4.5, dan *Random Forest*, dalam mendeteksi penyakit paru-paru. Pendekatan ini dipilih karena sesuai dengan tujuan penelitian, yaitu membandingkan performa tiga algoritma *machine learning* dalam klasifikasi penyakit paru-paru berdasarkan data non-citra [14]. Proses penelitian ini dilakukan secara sistematis dengan beberapa tahapan sebagai berikut:




Gambar 1. Tahap metodologi penelitian

### 2.1. Pengumpulan Data

Tahap awal penelitian ini dimulai dengan proses pengumpulan data dari sumber atau instansi tepercaya yang berkaitan dengan pasien paru-paru. Data yang digunakan pada penelitian ini diambil melalui *website Kaggle* di-*upload* oleh Bsrc Andot03, Dataset ini terdiri dari 30.000 data pasien dengan 9 atribut dan 1 kolom target yang merepresentasikan status kesehatan paru-paru (sehat atau sakit). Dataset ini bersifat non-citra dan dipilih berdasarkan pertimbangan ukuran data yang besar, keberagaman atribut (numerik dan kategorikal), serta kesesuaiannya dengan permasalahan klasifikasi medis. Penggunaan data non-citra ini memungkinkan pendekatan kuantitatif komparatif dalam menguji dan membandingkan performa berbagai algoritma klasifikasi untuk deteksi dini penyakit paru-paru secara sistematik dan terukur.

### 2.2. *Pre-Processing*

Data yang telah dikumpulkan selanjutnya dapat melalui proses *pre-processing* data yang bertujuan untuk meningkatkan kualitas data sebelum digunakan dalam proses pelatihan dan evaluasi model. Tiga tahap utama dilakukan, yaitu: pembersihan data (*data cleaning*), transformasi data, dan analisis data eksploratif (E*xploratory Data Analysis/EDA*). Tahapan *Cleaning* data meliputi pembersihan data untuk menangani *missing values*, mendeteksi dan menghapus *outliers*, dan penghapusan data duplikat. Juga terdapat transformasi data pada penelitian ini digunakan untuk mengkonversi atribut kategorikal menjadi format numerik *(encoding*), memastikan bahwa semua algoritma dapat memproses atribut fitur dengan benar. Hal ini memungkinkan beberapa algoritma *machine learning* memerlukan data dalam format numerik. Kemudian proses menampilkan *Exploratory Data Analysis* (EDA) untuk mempermudah memahami karakteristik, struktur, dan komponen penting dari dataset sebelum melanjutkan ke analisis pemodelan.

#### 2.2.1. *Cleaning Data*

*Cleaning Data* merupakan proses penting dalam *pre-processing* untuk memastikan kualitas dataset yang digunakan dalam analisis atau pelatihan model *machine learning*. Tahapan ini bertujuan untuk mengidentifikasi

dan menangani nilai hilang (*missing values*), *outlier*, dan data duplikat. *Missing values* ditangani dengan metode imputasi berbasis nilai modus atau median tergantung pada tipe data. *Outlier* diidentifikasi menggunakan metode *Interquartile Range* (IQR), dan duplikasi data dihapus untuk menghindari bias model. Pembersihan data merupakan tahap krusial agar model dilatih menggunakan data yang bersih, konsisten, dan representatif [15].

### 2.2.2. Transformasi Data

Transformasi data adalah langkah penting dalam *preprocessing* untuk memastikan data berada dalam format yang sesuai bagi algoritma *machine learning*. Pada penelitian ini, atribut kategorikal diubah menjadi bentuk numerik menggunakan *one-hot encoding*, sedangkan atribut numerik dinormalisasi menggunakan metode Standard *Scaler* untuk menyeragamkan skala antar fitur. Transformasi ini diperlukan karena algoritma seperti *Random Forest*, C4.5, dan XGBoost bekerja lebih optimal dengan input numerik. Proses ini juga memastikan model dapat mengenali pola secara konsisten tanpa bias akibat skala fitur yang berbeda.

### 2.2.3. Exploratory Data Analysis (EDA)

EDA dalam penelitian ini bertujuan untuk memahami struktur dan karakteristik data, termasuk distribusi kelas, korelasi antar fitur, dan identifikasi ketidakseimbangan data. Visualisasi seperti histogram, *boxplot*, dan *heatmap* korelasi digunakan untuk mengidentifikasi pola penting dan potensi anomali dalam *dataset*. Tahapan ini membantu menyusun strategi pemodelan dan evaluasi yang lebih efektif serta mendukung interpretasi hasil analisis model [16].

## 2.3. Splitting Data

Setelah proses *Pre-Processing data* selesai, akan dilakukan tahap *splitting data*, yaitu pembagian dataset menjadi dua subset utama yaitu *data training* dan *data testing*. Pada tahap ini, dataset dibagi secara acak menjadi 80% *data training* dan 20% *data testing*. Pemilihan skema 80:20 dilakukan karena dataset yang digunakan cukup besar dan representatif, serta untuk memastikan efisiensi waktu komputasi dalam pelatihan dan evaluasi model [17]. *Data training* digunakan untuk melatih model sehingga dapat mempelajari pola dari dataset, sedangkan data testing digunakan untuk mengevaluasi performa model terhadap data baru yang belum pernah dilihat sebelumnya. Pembagian secara acak memastikan bahwa data yang digunakan tidak bias atau terlalu terfokus pada satu pola tertentu. Dengan cara ini, performa model dapat diukur secara objektif, dan hasilnya merefleksikan kemampuan model dalam memprediksi data yang sebenarnya. *Splitting* data adalah langkah penting untuk menghindari *overfitting*, yaitu situasi di mana model terlalu cocok dengan *data training* dan kehilangan kemampuan untuk melakukan generalisasi pada data baru.

## 2.4. Klasifikasi

Tahapan klasifikasi pada penelitian ini direncanakan untuk melatih model pembelajaran mesin agar mampu mengklasifikasikan data ke dalam kategori tertentu berdasarkan pola yang ditemukan dalam data pelatihan. Proses klasifikasi dirancang untuk mengidentifikasi apakah seseorang berisiko terkena penyakit paru-paru atau tidak, berdasarkan informasi dalam dataset. Untuk mencapai tujuan tersebut, tiga algoritma utama akan digunakan, yaitu XGBoost, C4.5, dan *Random Forest*.

### A. XGBoost

XGBoost merupakan algoritma berbasis *gradient boosting* yang membangun model secara bertahap untuk mengurangi kesalahan prediksi sebelumnya. XGBoost dikenal karena efisiensinya dalam menangani data besar serta kemampuannya menghasilkan akurasi tinggi dengan *overfitting* yang rendah [3].

### B. C4.5

Algoritma C4.5 membangun model klasifikasi dengan membentuk pohon keputusan berdasarkan kriteria *Gain Ratio*. Pohon keputusan ini bekerja dengan memilih fitur terbaik pada setiap *node* untuk membagi

data hingga mencapai hasil yang paling "murni" (homogen) [18].

**C. *Random Forest***

*Random Forest* merupakan metode *ensemble learning* yang membentuk sejumlah pohon keputusan dari *subset* data yang diambil secara acak dan menggabungkan hasilnya menggunakan *voting* mayoritas. Teknik ini memperkuat stabilitas prediksi dan mengurangi risiko *overfitting* [19].

## 2.5. Evaluasi

Evaluasi model dilakukan untuk menilai kinerja algoritma dalam mendeteksi penyakit paru-paru berdasarkan data uji. Proses evaluasi bertujuan untuk mengukur seberapa baik model dapat mengklasifikasikan data dengan benar serta mengetahui tingkat kesalahan prediksi. Pengujian dilakukan menggunakan berbagai metrik evaluasi yang mencakup akurasi, *precision*, *recall*, dan *F1-score*. Akurasi mengukur persentase prediksi yang benar dari keseluruhan data uji. *Precision* digunakan untuk menilai tingkat ketepatan prediksi positif dari semua prediksi positif yang dihasilkan model. *Recall* mengukur sensitivitas model dalam mendeteksi kasus positif yang sebenarnya ada dalam dataset. Sementara itu, *F1-score* adalah rata-rata harmonik antara *precision* dan *recall* yang memberikan gambaran keseimbangan antara kedua metrik tersebut.

Selain metrik evaluasi, *confusion matrix* digunakan untuk mengidentifikasi distribusi kesalahan prediksi yang dilakukan oleh model. *Confusion matrix* mengevaluasi prediksi model dengan membandingkan hasil prediksi terhadap label sebenarnya dengan mengklasifikasikan prediksi ke dalam kategori *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). Dengan menggunakan *confusion matrix*, dapat diketahui seberapa baik model dapat menghindari kesalahan tipe I (*False Positive*) dan kesalahan tipe II (*False Negative*). Hasil evaluasi dari masing-masing algoritma akan dianalisis secara komparatif untuk mengetahui metode mana yang memberikan hasil terbaik dalam mendeteksi penyakit paru-paru, baik dari segi akurasi maupun keseimbangan antara *precision* dan *recall*.

## 2.6. Persamaan

Dalam penelitian ini, beberapa persamaan digunakan untuk menghitung metrik evaluasi model yang digunakan. Persamaan metrik Evaluasi dihitung sebagai berikut dengan Persamaan (1):

$$Accuracy = \frac{TP + TN}{TP + TN + FT + FN} \tag{1}$$

Akurasi (*Accuracy*) digunakan untuk menilai kinerja suatu metode klasifikasi berdasarkan ketepatan hasil klasifikasinya. Semakin tinggi nilai akurasi yang diperoleh, maka metode tersebut dinilai semakin efektif.

$$Precision = \frac{TP}{TP + FT} \tag{2}$$

Presisi mencakup tingkat sensitivitas atau akurasi sistem dalam menilai informasi yang diberikan, sehingga dapat mengidentifikasi data positif atau negatif secara tepat. Berikut Persamaan (2) untuk menghitung presisi.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

*Recall* merupakan proporsi data positif yang terdeteksi secara akurat dibandingkan dengan seluruh data positif, baik yang benar maupun yang keliru dikategorikan sebagai negatif. Nilai *recall* dapat dihitung menggunakan Persamaan (3).

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

*F1-Score* adalah metrik evaluasi yang digunakan untuk mengukur kinerja model klasifikasi, terutama data memiliki distribusi kelas yang tidak seimbang. Perhitungannya dapat dilihat pada Persamaan (4). Di bawah ini

merupakan hasil dari metrik evaluasi tiga algoritma yang digunakan.

## 3. HASIL DAN PEMBAHASAN

Hasil penelitian ini diperoleh melalui penerapan algoritma XGBoost, C4.5, dan *Random Forest* pada dataset kesehatan paru-paru yang terdiri dari 30.000 data dengan 9 atribut dan 1 kolom yang mewakili *class*, yaitu kolom hasil. Dataset tersebut telah melalui tahap *preprocessing* yang mencakup pembersihan data dengan penanganan nilai Syang hilang (*missing values*), deteksi dan penghapusan outlier, penghapusan data duplikat. Selain itu terdapat *Exploratory Data Analysis* (EDA) pada tahap *preprocesing* untuk memahami karakteristik data, struktur data, serta menemukan pola dan informasi yang mungkin tersembunyi dalam data, berikut hasil dalam setiap tahapan penelitiannya.

### 3.1. Pengumpulan Data

Pada tahap pengumpulan data yang digunakan dalam dataset *predict* terkena penyakit paru-paru yang di upload oleh Andot Bsrc memalui *website kaggle*. Data-data tersebut dapat dilihat pada Tabel 1.

Tabel 1. Dataset penyakit paru-paru

| No | Usia | Jenis _Kelamin | Merokok | Bekerja | Rumah _Tangga | Aktivitas _Begadang | Aktivitas _Olahraga | Asuransi | Penyakit _Bawaan | Hasil |
|---|---|---|---|---|---|---|---|---|---|---|
| 1. | Tua | Pria | Pasif | Tidak | Ya | Ya | Sering | Ada | Tidak | Ya |
| 2. | Tua | Pria | Aktif | Tidak | Ya | Ya | Jarang | Ada | Ada | Tidak |
| 3. | Muda | Pria | Aktif | Tidak | Ya | Ya | Jarang | Ada | Tidak | Tidak |
| 4. | Tua | Pria | Aktif | Ya | Tidak | Tidak | Jarang | Ada | Ada | Tidak |
| 5. | Muda | Wanita | Pasif | Ya | Tidak | Tidak | Sering | Tidak | Ada | Ya |

### 3.2. *Pre-processing*

Proses dilanjutkan ke tahap *preprocessing*, yang merupakan langkah penting untuk meningkatkan kualitas data sebelum digunakan dalam model. Pada tahap *preprocesing* ini terdapat beberapa tahapan yaitu *cleaning data*, trasformasi data, dan juga menampilkan *Exploratory Data Analysis* (EDA).

#### 3.2.1. Cleaning Data

Pada tahap ini dilakukan proses pembersihan data dengan mengidentifikasi outlier dan menghapusnya dengan menggunakan metode *Interquartile Range* (IQR), di mana data ekstrem yang dapat memengaruhi performa model secara negatif diidentifikasi dan dikeluarkan. Selain itu, nilai hilang ditangani dengan mengisi data numerik menggunakan rata-rata dan data kategorikal menggunakan modus. Data duplikat juga dihapus untuk mengurangi bias yang mungkin timbul dari data berulang.

#### 3.2.2. Transformasi Data

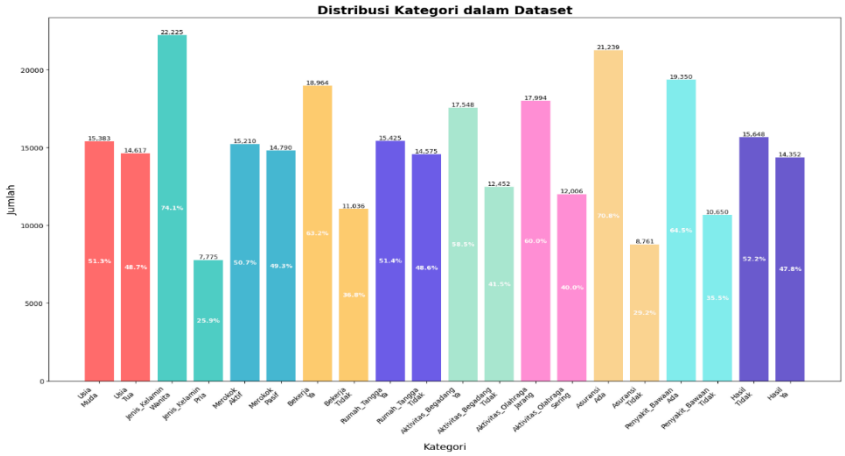
Setelah melakukan pembersihan data, selanjutnya data di proses ke tahapan transformasi data yaitu mengubah data kategorikal menjadi numerik memastikan data berada dalam format yang sesuai bagi algoritma *machine learning* (lihat Tabel 2).

Tabel 2. Dataset penyakit paru-paru

| No | Usia | Jenis_ Kelamin | Merokok | Bekerja | Rumah_ Tangga | Aktivitas Begadang | Aktivitas_ Olahraga | Asuransi | Penyakit_ Bawaan | Hasil |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 |
| 2 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| 4 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 |

### 3.2.3. Menampilkan *Exploratory Data Analysis* (EDA)

Gambar 2 menampilkan *Exploratory Data Analysis* (EDA) pada dataset yang telah dibersihkan guna untuk memahami pola data secara lebih mendalam dan mengetahui jumlah total data setelah melalui *cleaning data* dan transformasi data. Distribusi setiap variabel dianalisis menggunakan grafik seperti histogram, bar chart, dan *pie chart.*




Gambar 2. EDA dengan *bar chart*

Pada output EDA dengan bar chart ini menampilkan distribusi data pada semua fitur-fitur yang ada pada dataset sehingga dapat dilihat jumlah label pada setiap fiturnya beserta persentase yang memudahkan untuk memahami jumlah dataset setelah melewati tahapan pembersihan dan trasformasi data.

## 3.3. Splitting Data

Pada tahapan ini juga menampilkan jumlah data pada masing-masing set, sehingga kita tahu berapa banyak data yang akan digunakan untuk pelatihan dan pengujian, Data *training* digunakan untuk melatih model sehingga dapat mempelajari pola dari dataset, sedangkan data *testing* digunakan untuk mengevaluasi performa model terhadap data baru yang belum pernah dilihat sebelumnya [17]. Pembagian dataset ditampilkan dalam Tabel 3.

Tabel 3. Pembagian dataset

| Keterangan | Data Training | Data Testing | Total |
|---|---|---|---|
| Proporsi | 80% | 20% | 100% |
| Jumlah | 24.000 | 6.000 | 30.000 |

## 3.4. Klasifikasi

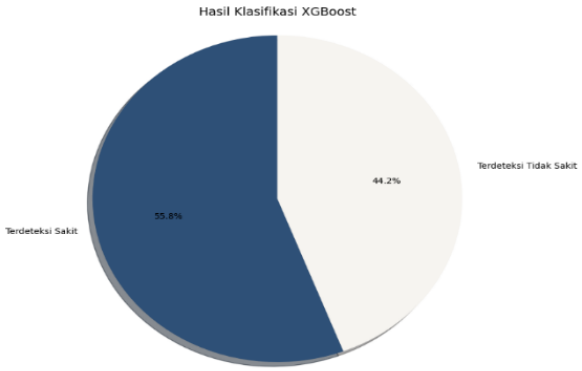
Setelah melalui proses Preprocesing dan Splitting data maka proses selanjutnya adalah Proses klasifikasi mencakup penerapan tiga Model algoritma, yaitu XGBoost, C4.5, dan Random Forest Proses klasifikasi terdiri dari dua tahap utama yaitu pelatihan model dan prediksi. Pada tahap pelatihan, data training (x_train dan y_train) digunakan untuk mengajari model mengenali pola antara fitur dan label target.

Setelah itu, pada tahap prediksi, model menggunakan data testing (x_test dan y_test) untuk memprediksi label target berdasarkan pola yang telah dipelajari. Hasil prediksi kemudian dibandingkan dengan label target sebenarnya untuk mengevaluasi kinerja model. Proses ini memastikan model dapat memahami pola data dengan benar.

### A. XGBoost

Pada tahap ini, dilakukan proses klasifikasi menggunakan dataset yang telah melalui tahap *splitting data* menjadi data pelatihan dan data pengujian. Model berhasil mendeteksi kondisi "sakit" dengan persentase

55,8% dan "tidak sakit" dengan persentase 44,2%. Hasil klasifikasi ini ditampilkan dalam *pie chart* pada Gambar 3.

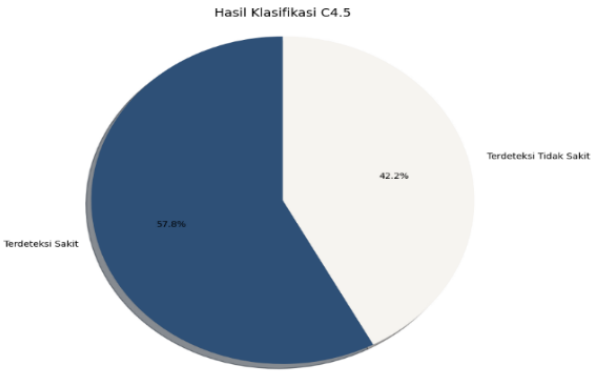



Gambar 3. Hasil klasifikasi XGBoost

Dengan proses membangun dan melatih model klasifikasi mencakup proses pembuatan model, pelatihan dengan data training (x_train dan y_train), prediksi kelas pada data testing (x_test dan y_test), serta memastikan format data yang konsisten untuk evaluasi. Akhirnya, visualisasi hasil prediksi melalui pie chart membantu dalam memahami performa model secara intuitif.

## B. C4.5

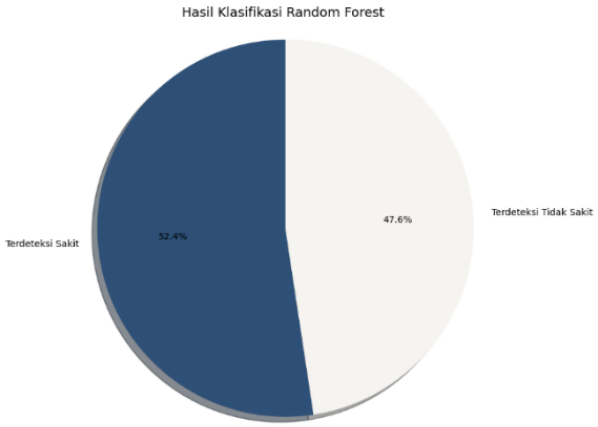
Model diinisialisasi dengan parameter yang tepat untuk mengoptimalkan pembagian *node* dalam pohon keputusan, kemudian dilatih dengan data pelatihan untuk mengenali pola yang ada. Setelah model dilatih, ia mampu memprediksi kelas pada data pengujian, dan hasil prediksi tersebut divisualisasikan dalam bentuk *pie chart*. Berikut *output* yang ditampilkan dengan *pie chart* (Gambar 4). Hasil pada proses ini yang terdeteksi sakit dengan jumlah 57,8% dan yang terdeteksi tidak sakit berjumlah 42,4%.




Gambar 4. Hasil klasifikasi C4.5

## C. *Random Forest*

Model diinisialisasi dengan parameter yang tepat untuk mengoptimalkan pembangunan hutan keputusan, lalu dilatih dengan data pelatihan untuk mengenali pola yang ada. Setelah model dilatih, ia mampu memprediksi kelas pada data pengujian, dan hasil prediksi tersebut divisualisasikan dalam bentuk *pie chart* untuk memudahkan pemahaman tentang distribusi prediksi dibandingkan dengan label sebenarnya. Gambar 4 menunjukkan *output* yang ditampilkan dengan *pie chart*. Hasil pada proses ini yang terdeteksi sakit dengan jumlah 52,4% dan yang terdeteksi tidak sakit berjumlah 47,6%.



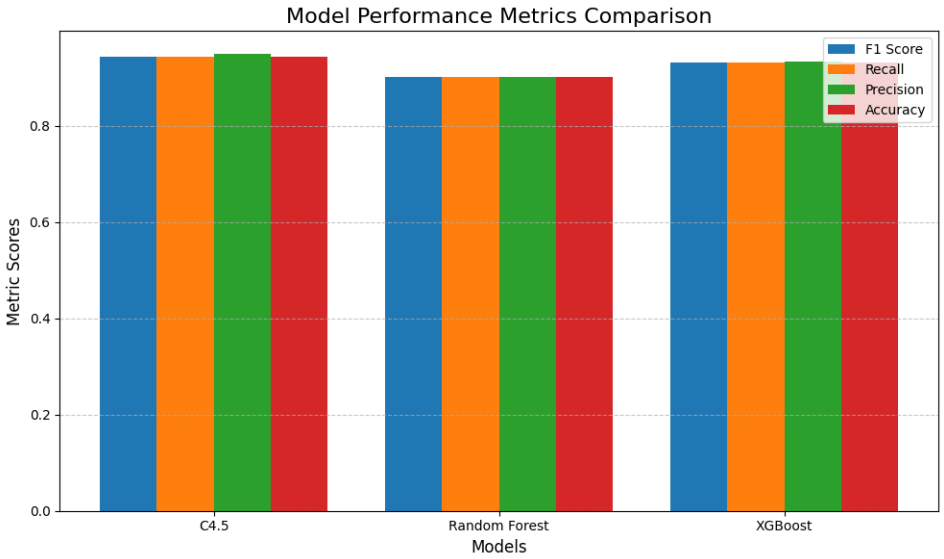

Gambar 5. Hasil klasifikasi *random forest*

### 3.5. Evaluasi

Tahapan evaluasi model melibatkan pengukuran performa setiap algoritma menggunakan metrik seperti *F1-Score*, *Recall*, *Precision*, dan *Accuracy*. *Confusion matrix* digunakan untuk menganalisis hasil prediksi secara lebih rinci, seperti memisahkan *True Positive*, *False Positive*, *True Negative*, dan *False Negative*. Evaluasi ini bertujuan untuk membandingkan kekuatan dan kelemahan setiap algoritma, sehingga dapat menentukan metode terbaik untuk digunakan. Berikut ini merupakan tabel perbandingan performa dari tiga algoritma yang digunakan yaitu pada Tabel 4.

Tabel 4. Perbandingan hasil metrik evaluasi

| | **XGBoost** | **C4.5** | **Random forest** |
|---|---|---|---|
| *Accuracy* | 0,9318 | 0,9433 | 0,9007 |
| *Precision* | 0,9339 | 0,9489 | 0,9007 |
| *Recall* | 0,9318 | 0,9433 | 0,9007 |
| *F1-score* | 0,9316 | 0,9490 | 0,9007 |




Gambar 6. Perbandingan metrik evaluasi

Hasil penelitian menunjukkan bahwa algoritma berbasis pohon keputusan, yaitu XGBoost, C4.5, dan *Random Forest*, memiliki keunggulan masing-masing dalam mendeteksi penyakit paru-paru. Berdasarkan Gambar 6, evaluasi menggunakan metrik seperti akurasi, *precision*, recall, dan *F1-Score*, algoritma C4.5 memberikan

hasil terbaik dibandingkan dua algoritma lainnya [20]. Algoritma ini mencapai akurasi sebesar 94,33%, *precision* 94,89%, *recall* 94,33%, dan *F1-Score* 94,90%. Posisi kedua ditempati oleh XGBoost, dengan akurasi 93,18%, *precision* 93,39%, *recall* 93,18%, dan *F1-Score* 93,16% (lihat Tabel 5). Algoritma *Random Forest* memiliki performa paling rendah di antara ketiganya, dengan *accuracy*, *precision, recall,* dan *F1-Score* masing-masing 90,07%.

Tabel 5. Perbandingan hasil *confusion matrix*

| | **True Positive (TP)** | **True Negative (TN)** | **False Positive (FP)** | **False Negative (FN)** |
|---|---|---|---|---|
| XGBoost | 3035 | 2556 | 314 | 90 |
| C4.5 | 3130 | 2530 | 340 | 0 |
| Random Forest | 3839 | 2565 | 305 | 291 |

### 3.6. Interpretasi dan Pembahasan

XGBoost, dan Random Forest memiliki kemampuan yang baik dalam mendeteksi penyakit paru-paru. Namun, algoritma C4.5 memberikan performa terbaik dibandingkan dua algoritma lainnya.

Keunggulan C4.5 dapat dijelaskan secara ilmiah karena algoritma ini mampu menangani atribut kategorikal secara langsung dan memiliki proses *tuning* yang baik, sehingga mengurangi *overfitting*. XGBoost juga menunjukkan performa kompetitif berkat teknik *boosting* yang kuat dalam mengurangi bias dan varians.

Hasil ini sejalan dengan penelitian [6] yang berjudul "*Comparison of the Performance Results of C4.5 and Random Forest Algorithm in Data Mining to Predict Childbirth Process*" memiliki hasil yang berbeda juga yaitu algoritma C4.5 lebih unggul dari *Random forest* dan *Catboost*. Sebaliknya, hasil ini berbeda dari penelitian [4] yang berjudul "Perbandingan Algoritma *Random Forest* Dan XGBoost Untuk Klasifikasi Penyakit Paru-Paru Berdasarkan Data Demografi Pasien" meliliki hasil *accuracy* 94% XGBoost dan 91% *Random Forest*. Perbedaan ini kemungkinan besar disebabkan oleh variasi karakteristik dataset, seperti jumlah fitur, jenis atribut, dan distribusi kelas. Lebih detailnya, perbandingan hasil penelitian dapat dilihat pada Tabel 6.

Tabel 6. Perbandingan dengan Penelitian Sebelumnya

| **Studi** | **Algoritma Terbaik** | **Accuracy** | **Keterangan** |
|---|---|---|---|
| Penelitian ini | C4.5 | 94,33% | Mengungguli XGBoost dan Random Forest |
| "Perbandingan Algoritma Random Forest Dan Xgboost Untuk Klasifikasi Penyakit Paru-Paru Berdasarkan Data Demografi Pasien" [4]. | XGBoost | 94.00% | Random Forest 91% |
| "Comparison of the Performance Results of C4.5 and Random Forest Algorithm in Data Mining to Predict Childbirth Process" [6]. | C4.5 | 96.00% | Random forest 95.00% |

## 4. KESIMPULAN

Penelitian ini menunjukkan bahwa algoritma C4.5 memiliki kinerja terbaik dalam mendeteksi penyakit paru-paru dibandingkan XGBoost dan *Random Forest*, berdasarkan metrik akurasi, presisi, *recall*, dan *F1-score*. Hasil ini menegaskan bahwa algoritma klasik masih mampu bersaing secara efektif dalam konteks data tertentu. Kontribusi utama studi ini adalah memberikan analisis komparatif terhadap tiga metode klasifikasi berbasis pohon keputusan untuk deteksi penyakit berbasis data non-medis. Penelitian ini memiliki keterbatasan pada jenis dan sumber data, sehingga disarankan untuk menguji pendekatan serupa pada data klinis yang lebih kompleks di masa depan. Studi ini dapat menjadi referensi awal bagi penelitian lanjutan dalam pengembangan sistem deteksi penyakit berbasis *machine learning*, khususnya yang memanfaatkan data demografi dan perilaku.

## UCAPAN TERIMA KASIH

Terima kasih atas dukungan dan kesempatan yang telah diberikan oleh Universitas Bumigora dan kepada semua pihak yang telah membantu dan membimbing dalam penelitian ini.

## DAFTAR PUSTAKA

[1] S. A. Naufal, A. Adiwijaya, dan W. Astuti, "Analisis Perbandingan Klasifikasi Support Vector Machine (SVM) dan K-Nearest Neighbors (KNN) untuk Deteksi Kanker dengan Data Microarray," *JURIKOM (Jurnal Riset Komputer)*, vol. 7, no. 1, p. 162, Feb. 2020. DOI: 10.30865/jurikom.v7i1.2014.

[2] M. Y. Haffandi et al., "Klasifikasi Penyakit Paru-Paru dengan Menggunakan Metode Naïve Bayes Classifier," *Jurnal Teknik Informasi dan Komputer (Tekinkom)*, vol. 5, no. 2, p. 176, Dec. 2022. DOI: 10.37600/tekinkom.v5i2.649.

[3] D. Adimanggala, *Algoritme XGBoost dengan Contohnya*, https://dindaadi.medium.com/algoritme-xgboost-dengan-contohnya-28e958a3e2f6, Mar. 2023.

[4] R. Harahap et al., "Perbandingan Algoritma Random Forest dan XGBoost untuk Klasifikasi Penyakit Paru-Paru Berdasarkan Data Demografi Pasien," *Jurnal Ilmiah Betrik*, vol. 15, no. 2, pp. 130–141, 2024.

[5] Y. Amelia, "Perbandingan Metode Machine Learning untuk Mendeteksi Penyakit Jantung," *IDEALIS : InDonEsiA journaL Information System*, vol. 6, no. 2, pp. 220–225, Jul. 2023. DOI: 10.36080/idealis.v6i2.3043.

[6] M. Muhasshanah et al., "Comparison of the Performance Results of C4.5 and Random Forest Algorithm in Data Mining to Predict Childbirth Process," *CommIT (Communication and Information Technology) Journal*, vol. 17, no. 1, pp. 51–59, Mar. 2023. DOI: 10.21512/commit.v17i1.8236.

[7] J. M. A. S. Dachi dan P. Sitompul, "Analisis Perbandingan Algoritma XGBoost dan Algoritma Random Forest Ensemble Learning pada Klasifikasi Keputusan Kredit," *Jurnal Riset Rumpun Matematika dan Ilmu Pengetahuan Alam*, vol. 2, no. 2, pp. 87–103, Jul. 2023. DOI: 10.55606/jurrimipa.v2i2.1470.

[8] F. T. Kristanti et al., "Advancing financial analytics: Integrating XGBoost, LSTM, and Random Forest Algorithms for precision forecasting of corporate financial distress," *Journal of Infrastructure, Policy and Development*, vol. 8, no. 8, p. 4972, Aug. 2024. DOI: 10.24294/jipd.v8i8.4972.

[9] A. S. Sunge et al., "Performance Comparison of Decision Tree, Random Forest, and XGBoost Models; And Its Interpretability Using Shap for Recognizing the Necessity of Caesareans Section of Childbirth," *Journal of Theoretical and Applied Information Technology*, vol. 101, no. 9, p. 3297, 2023.

[10] Y. Yennimar et al., "Comparison of data mining algorithms (random forest, C4.5, catboost) based on adaptive boosting in predicting diabetes mellitus," *Jurnal Teknik Informatika C.I.T Medicom*, vol. 16, no. 1, pp. 1–12, Mar. 2024. DOI: 10.35335/cit.Vol16.2024.730.pp1-12.

[11] E. Ismanto dan M. Novalia, "Komparasi Kinerja Algoritma C4.5, Random Forest, dan Gradient Boosting untuk Klasifikasi Komoditas," *Techno.Com*, vol. 20, no. 3, pp. 400–410, Aug. 2021. DOI: 10.33633/tc.v20i3.4576.

[12] A. Wahid, "Komparasi Algoritma C4.5 dengan Random Forest untuk Rekomendasi Penjualan Gaun aliexpress.com," Skripsi, Universitas Muhammadiyah Jember, Jan. 2020. DOI: 10/ARTIKEL%20.pdf.

[13] T. R. Karin et al., "Enhancing Bank Customer Protection Against Phishing Attacks Through XGBoost-Based Feature Analysis," *Transmisi: Jurnal Ilmiah Teknik Elektro*, vol. 26, no. 3, pp. 114–121, Nov. 2024. DOI: 10.14710/transmisi.26.3.114-121.

[14] H. H. Sinaga dan S. Agustian, "Pebandingan Metode Decision Tree dan XGBoost untuk Klasifikasi Sentimen Vaksin Covid-19 di Twitter," *Jurnal Nasional Teknologi dan Sistem Informasi*, vol. 8, no. 3, pp. 107–114, Dec. 2022. DOI: 10.25077/TEKNOSI.v8i3.2022.107-114.

[15] K. L. Kohsasih dan Z. Situmorang, "Analisis Perbandingan Algoritma C4.5 dan Naïve Bayes dalam Memprediksi Penyakit Cerebrovascular," *Jurnal Informatika*, vol. 9, no. 1, pp. 13–17, Apr. 2022. DOI: 10.31294/inf.v9i1.11931.

[16] E. D. Wahyuni, A. A. Arifiyanti, dan M. Kustyani, "Exploratory Data Analysis dalam Konteks Klasifikasi Data Mining," *Prosiding Seminar Nasional ReTII Ke-14 2019*, pp. 263–269, Nov. 2019.

[17] R. Adinugroho, "Perbandingan Rasio Split data Training dan data Testing Menggunakan Metode LSTM dalam Memprediksi Harga Indeks Saham Asia," Skripsi, Fakultas Sains dan Teknologi UIN Syarif Hidayatullah Jakarta, Feb. 2023.

[18] T. Tukino, "Penerapan Algoritma C4.5 untuk Memprediksi Keuntungan pada PT SMOE Indonesia," *Jurnal Sistem Informasi Bisnis*, vol. 9, no. 1, p. 39, May 2019. DOI: 10.21456/vol9iss1pp39-46.

[19] B. S. C. Putra et al., "Efektivitas Algoritma Random Forest, XGBoost, dan Logistic Regression dalam Prediksi Penyakit Paru-paru," *Techno.Com*, vol. 23, no. 4, pp. 909–922, Nov. 2024. DOI: 10.62411/tc.v23i4.11705.

[20] A. Sugarda et al., "Penerapan Metode Data Mining C4.5 dalam Penentuan Kelayakan Rehabilitas Rumah Warga," *Journal of Computing and Informatics Research*, vol. 1, no. 3, pp. 56–64, Jul. 2022. DOI: 10.47065/comforch.v1i3.321.