Jurnal Teknologi dan Manajemen Industri Terapan (JTMIT) Vol. 4, No. 4, Desember 2025 pp. 1164 - 1171
P-ISSN: 2829-0232 E-ISSN: 2829-0038

Implementasi Algoritma Klasifikasi C4.5 Untuk Memprediksi Resiko Terkena Penyakit Paru-Paru

Muhamad Ilhan Mansiz, Ahmad Homaidi, Jarot Dwi Prasetyo
[1,2)] Teknologi Informasi, Fakultas Sains & Teknologi, Universitas Ibrahimy
[3)] Ilmu Komputer, Fakultas Sains & Teknologi, Universitas Ibrahimy
Jl. KHR. Syamsul Arifin. Dusun. Sukorejo, Des. Sumberejo, Kec. Banyuputih, Situbondo
Email: ilhanmansiz2022@gmail.com, ahmadhomaidi@ibrahimy.ac.id, jarot_dwi_prasetyo@yahoo.com

ABSTRAK

Penyakit paru-paru merupakan salah satu masalah kesehatan utama di Indonesia dengan prevalensi yang terus meningkat akibat faktor gaya hidup seperti merokok, paparan polusi, dan pola hidup tidak sehat. Penelitian ini bertujuan mengembangkan model prediksi risiko penyakit paru-paru berbasis algoritma C4.5 dengan menganalisis atribut pola hidup masyarakat, meliputi kebiasaan merokok, paparan polusi, jenis pekerjaan, pola makan, olahraga, dan lingkungan tempat tinggal. Algoritma C4.5 dipilih karena kemampuannya menangani data kategorikal dan nilai hilang (missing values) serta menghasilkan model yang mudah dipahami. Dataset yang digunakan berjumlah 30.000 baris data sekunder yang diperoleh dari platform Kaggle. Hasil pengujian menunjukkan bahwa model C4.5 yang dibangun memiliki akurasi sebesar 94%, precision 95%, recall 94%, dan F1-score 94%. Temuan ini diharapkan dapat membantu tenaga medis dan masyarakat dalam melakukan deteksi dini serta upaya preventif terhadap penyakit paru-paru, sekaligus mengidentifikasi faktor gaya hidup yang paling berkontribusi terhadap risiko penyakit ini, yaitu kebiasaan merokok.

Kata kunci: implementasi, algoritma, klafikasic4.5, prediksi, penyakit paru-paru.

ABSTRACT

Lung disease is a significant health problem in Indonesia, with its prevalence continuing to increase due to lifestyle factors such as smoking, pollution exposure, and unhealthy lifestyles. This study aims to develop a lung disease risk prediction model based on the C4.5 algorithm by analyzing community lifestyle attributes, including smoking habits, pollution exposure, type of work, diet, exercise, and residential environment. It also eliminates missing values and produces an easily understood model. The dataset consists of 30,000 rows of secondary data obtained from the Kaggle platform. The test results show that the C4.5 model has an accuracy of 94%, a precision of 95%, a recall of 94%, and an F1-score of 94%..

Keywords: *implementation, algorithm, c4.5 classification, prediction, lung disease.*

Pendahuluan

Penyakit paru-paru merupakan salah satu gangguan kesehatan yang memiliki dampak serius terhadap kualitas hidup manusia, terutama karena menyerang sistem pernapasan yang berperan vital dalam pertukaran oksigen dan karbon dioksida dalam tubuh[1]. Faktor gaya hidup seperti kebiasaan merokok, baik aktif maupun pasif, menjadi penyebab utama munculnya penyakit ini, bahkan menurut data dari World Health Organization (WHO), lebih dari 40% kematian akibat penggunaan tembakau berkaitan langsung dengan kerusakan paru-paru[2].

Di Indonesia sendiri, penyakit paru-paru menjadi salah satu penyumbang utama beban pembiayaan dalam sektor kesehatan. Berdasarkan Riset Kesehatan Dasar (Riskesdas) tahun 2018 yang dilakukan oleh Kementerian Kesehatan Republik Indonesia, prevalensi penyakit paru kronik mencapai angka 3,7% dan diperkirakan terus meningkat dari tahun ke tahun, terutama di daerah perkotaan dan kawasan industri yang memiliki tingkat polusi udara tinggi[3].

Banyak faktor yang berkontribusi terhadap timbulnya penyakit paru-paru, salah satunya adalah pola hidup yang tidak sehat. Kebiasaan merokok, minum alkohol, kurang olahraga, paparan polusi udara, serta kebiasaan tidur yang buruk merupakan faktor risiko utama yang dapat memperburuk kondisi kesehatan paru-paru[4]. Pola hidup tersebut seringkali diabaikan oleh masyarakat, padahal dampaknya sangat signifikan terhadap kualitas hidup dan harapan hidup seseorang. Oleh karena itu, perlu adanya upaya prediksi dan pencegahan dini untuk menekan angka kejadian penyakit paru-paru[5].

Perkembangan teknologi informasi dan kecerdasan buatan membuka peluang besar dalam penerapan metode analisis data untuk keperluan prediksi kesehatan[6]. Salah satu pendekatan yang dapat digunakan adalah data mining, yaitu ekstraksi informasi atau pola yang penting atau menarik dari data yang ada di database yang besar. Dalam konteks prediksi penyakit,

data mining dapat membantu menemukan pola atau hubungan tersembunyi antara atribut-atribut data pasien, seperti gaya hidup dan riwayat kesehatan, dengan kemungkinan munculnya suatu penyakit tertentu[7].

Metode klasifikasi dalam data mining banyak digunakan untuk membangun sistem prediksi. Salah satu metode klasifikasi yang populer dan efektif adalah algoritma C4.5, yang merupakan pengembangan dari metode ID3 [8]–[12]. Algoritma C4.5 mampu membentuk pohon keputusan berdasarkan atribut yang paling informatif menggunakan perhitungan gain ratio, sehingga sangat cocok digunakan untuk prediksi berbasis data kategorikal seperti kebiasaan merokok, aktivitas fisik, konsumsi makanan, dan sebagainya[13]–[16]. Keunggulan metode ini adalah kemampuannya dalam menangani data dengan nilai yang hilang (missing values), serta menghasilkan model yang mudah dipahami oleh manusia karena berupa struktur pohon keputusan[17].

Beberapa penelitian terdahulu menunjukkan bahwa algoritma C4.5 dapat digunakan secara efektif dalam sistem prediksi penyakit. Misalnya, dalam studi yang dilakukan oleh wina yusnaeni dan widiarina, metode C4.5 digunakan untuk memprediksi penyakit diabetes tahap awal berdasarkan data rekam medis pasien dan menghasilkan tingkat akurasi yang cukup tinggi, yaitu mencapai 88,46%[18].

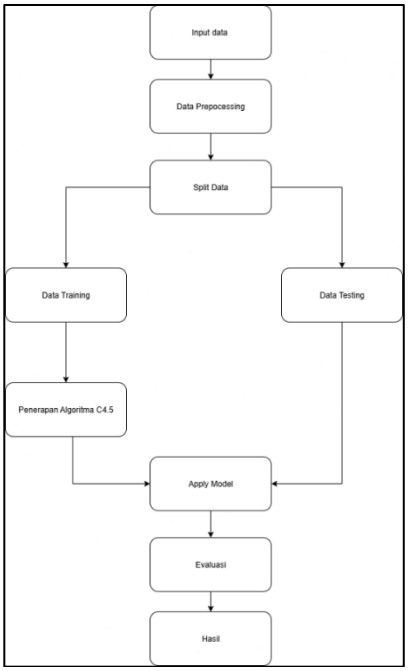
Meskipun prediksi penyakit menggunakan data mining telah banyak dilakukan, sebagian besar penelitian fokus pada penyakit jantung, diabetes, atau kanker, sementara penelitian prediksi penyakit paru-paru masih relatif sedikit, terutama yang menggunakan data pola hidup masyarakat dalam jumlah besar. Hal ini menjadi peluang penelitian untuk mengembangkan model prediksi penyakit paru-paru yang relevan dan mudah diinterpretasikan.

Dengan melihat permasalahan yang ada, maka diperlukan suatu pendekatan analisis yang dapat memprediksi risiko penyakit paru-paru secara dini berdasarkan pola hidup individu. Analisis ini diharapkan dapat menjadi alat bantu bagi tenaga kesehatan atau masyarakat umum dalam mengambil langkah preventif terhadap penyakit paru-paru. Penelitian ini bertujuan untuk mengembangkan model prediksi penyakit paru-paru menggunakan algoritma C4.5 berdasarkan data pola hidup masyarakat. Adapun atribut-atribut pola hidup yang akan dianalisis meliputi kebiasaan merokok, paparan polusi, jenis pekerjaan, pola makan, olahraga, serta kondisi lingkungan tempat tinggal.

Penelitian ini bertujuan untuk mengidentifikasi faktor-faktor utama pada gaya hidup yang dapat mempengaruhi resiko penyakit paru-paru serta untuk melakukan evaluasi terhadap efektivitas algoritma Decision Tree dalam analisis data resiko terkena penyakit paru-paru. Dengan memahami faktor-faktor tersebut, pihak medis dapat mengembangkan langkah yang lebih efektif dalam mencegah penyakit paru-paru berdasarkan faktor gaya hidup yang paling berperan terhadap munculnya penyakit tersebut. Oleh karena itu, penelitian ini memiliki nilai penting dalam pencegahan terhadap gaya hidup yang dapat menyebabkan penyakit paru-paru.

Metode Penelitian

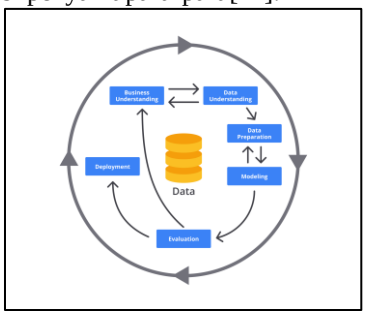
Tahap penelitian adalah kerangka kerja penelitian yang dijelaskan dalam tahapan-tahapan penelitian, tahapan ini dilakukan untuk menyelesaikan masalah yang akan diteliti dan untuk mengumpulkan informasi tentang masalah yang diteliti Untuk menyelesaikan penelitian ini maka digambarkan tahapan-tahapan penelitian yang dilakukan seperti pada berikut yaitu:



Gambar 1. Block diagram

Penelitian ini menggunakan pendekatan metodologi Knowledge Discovery in Database (KDD). KDD merupakan proses untuk menemukan pengetahuan yang bermanfaat dari kumpulan data. Proses ini berlangsung secara interaktif dan berulang, mencakup beberapa tahapan yang memungkinkan keterlibatan pengguna dalam pengambilan keputusan serta memungkinkan untuk kembali ke langkah sebelumnya jika diperlukan. Tahapan-tahapan dalam KDD meliputi: pembersihan data (data cleaning), integrasi data (data integration), pemilihan data (data selection), transformasi data (data transformation), penggalian data (data mining), interpretasi dan evaluasi hasil, hingga akhirnya menghasilkan pengetahuan yang dapat dimanfaatkan[19]. KDD merupakan suatu proses sistematis yang bertujuan untuk mengekstraksi pola atau pengetahuan baru dari sekumpulan data, yang dilakukan melalui serangkaian tahapan, antara lain :

1. Pemahaman Masalah dan Tujuan
   Langkah pertama dimulai dengan mempelajari pentingnya mendeteksi penyakit paru-paru secara dini, sekaligus mencari kemungkinan menggunakan algoritma machine learning sebagai solusi alternatif untuk mengatasi masalah tersebut.
2. Seleksi dan Pengumpulan Data
   penelitian ini menggunakan data yang diambil dari situs Kaggle.com dengan memanfaatkan penyakit paru-paru Dataset, yang memuat sebanyak 30000 entri data serta berbagai atribut yang merepresentasikan kondisi masing-masing individu.
3. Preprocessing Data
   Pada tahap ini, dilakukan proses pembersihan data untuk mengatasi nilai yang hilang maupun data yang tidak konsisten. Selanjutnya, dilakukan transformasi terhadap nilai-nilai numerik dan proses normalisasi agar dataset siap digunakan dalam pelatihan model.
4. Transformasi dan Seleksi Fitur
   Sejumlah atribut diprioritaskan dalam analisis karena dinilai memiliki korelasi yang signifikan terhadap potensi terjadinya penyakit paru-paru, di antaranya adalah Merokok dan Usia.
5. Pemodelan (Data Mining)
   Algoritma Decision Tree C4.5 digunakan sebagai metode utama dalam pengembangan model klasifikasi ini. Metode ini bekerja dengan membangun pohon keputusan berdasarkan perhitungan gain ratio untuk menentukan atribut yang paling berpengaruh dalam proses pemisahan data pada setiap node. Proses pelatihan model dilakukan dengan memanfaatkan struktur pohon yang dihasilkan, di mana setiap cabang merepresentasikan hasil pengujian atribut dan setiap daun merepresentasikan kelas prediksi. Pemilihan atribut dilakukan secara berulang hingga data pada node telah terklasifikasi secara murni atau tidak dapat dibagi lagi, sehingga dihasilkan model yang mampu melakukan klasifikasi secara efektif terhadap data baru.
6. Evaluasi Model
   Model dievaluasi menggunakan sejumlah metrik kinerja, yakni akurasi, presisi, dan recall, guna mengukur efektivitas model dalam memprediksi penyakit paru-paru[14].



Gambar 2. Alur Proses Data

Penelitian ini menggunakan bahasa pemrograman Pyhton. Python merupakan bahasa pemrograman tingkat tinggi yang dikembangkan oleh Guido van Rossum pada tahun 1989 dan dirilis pertama kali pada tahun 1991. Python diciptakan sebagai jawaban atas kebutuhan untuk memudahkan programmer dalam menyelesaikan tugas dengan cepat. Python dirancang untuk memberikan banyak kemudahan bagi pemrogram, baik dalam bidang manajemen waktu yang efisien atau pengembangan program dan kompatibilitas sistem. Python dapat digunakan untuk membuat program mandiri dan pemrograman coretan (Scriting Programming)[20].

Langkah selanjutnya yaitu melakukan klasifikasi menggunakan algoritma decision tree (C4.5). Algoritma C4.5 merupakan versi perbaikan dari ID3, yang awalnya fiturnya bertipe kategorikal (nominal) dan tipe numerik (rasio) tidak dapat digunakan.[4] Sedangkan C4.5 sudah menangani fitur dengan tipe numerik, melakukan pemotongan decision tree, dan penurunan rule set. C4.5 adalah algoritma berbasis gain ratio maka dalam membentuk pohon keputusan diperlukan tahapan perhitungan sebagai berikut:

1. Mempersiapkan data training, dapat diambil dari data history yang pernah terjadi sebelumnya dan sudah dikelompokkan dalam kelas-kelas tertentu.
2. Menentukan akar dari pohon dengan menghitung nilai gain yang tertinggi dari masing-masing atribut atau berdasarkan nilai index entropy terendah. Sebelumnya dihitung terlebih dahulu nilai index entropy, dengan rumus:

$$Entropy(S) = -\sum_{i=1}^{N} \text{pi}\ log2\ (pi) \quad (1)$$

Keterangan:
I = himpunan kasus
m = jumlah partisi i
f(i,j) = proposi j terhadap i

3. Hitung nilai gain dengan rumus:

$$Gain = -\sum_{i=1}^{p} {}^{ni}_{n} IE(i) \quad (2)$$

Keterangan:
P = jumlah partisi atribut
ni = proporsi ni terhadap i
n = jumlah kasus dalam n

4. Untuk menghitung gain ratio perlu diketahui suatu term baru yang disebut Split Inormation dengan rumus:

$$Split\ Information = -\sum_{t=1}^{c} \frac{s1}{s} log_2 \frac{S1}{S} \quad (3)$$

Keterangan:
sampai Se = c subset yang dihasilkan dari pemecahan S dengan menggunakan atribut A yang mempunyai sebanyak c nilai

5. Selanjutnya menghitung rasio

$$Gainratio(S.A) = \frac{Gain(S.A)}{Splitinformation(S.A)} \quad (4)$$

6. Mengulangi langkah ke-2 hingga semua record terpartisi Proses partisi pohon keputusan akan berhenti di saat:
   a. Semua tupel dalam record dalam simpul m mendapatkan kelas yang sama.
   b. Tidak ada atribut dalam record yang dipartisi lagi.
   c. Tidak ada record di dalam cabang yang kosong.

Hasil Dan Pembahasan

Penelitian ini mengimplementasikan algoritma klasifikasi C4.5 sebagai metode untuk memprediksi risiko terkena penyakit paru-paru. Tahapan penelitian mengikuti alur proses data mining, yaitu: data selection, data preprocessing, data transformation, dan evaluation. Alur ini digunakan sebagai acuan dalam membentuk model klasifikasi pohon keputusan berbasis gejala dan karakteristik penderita.

Preprocessing

a. Dataset Sampel

Penelitian ini menggunakan teknik pengumpulan data sekunder yang diperoleh dari platform penyedia data, Kaggle. Pemilihan data sekunder dilakukan karena sumber data telah tersedia, sehingga tidak diperlukan proses observasi, wawancara, atau survei untuk mengumpulkan data.

Tabel 1. Sampel dataset

Usia	Jenis kelamin	merokok	bekerja	Rumah tangga	begadang	olahraga	asuransi	Penyakit bawaan	Hasil
Tua	pria	pasif	tidak	ya	ya	sering	ada	tidak	Ya
Tua	pria	aktif	Tidak	ya	ya	jarang	ada	ada	Tidak
Muda	pria	aktif	tidak	ya	ya	jarang	ada	tidak	Tidak
Tua	pria	aktif	ya	tidak	tidak	jarang	ada	ada	Tidak
muda	wanita	pasif	ya	tidak	tidak	sering	tidak	ada	tidak

b. Analisis Dataset

Analisis data dalam penelitian ini terpusat pada penggunaan dataset yang diunduh dari situs Kaggle.com, yang merupakan salah satu platform terkemuka di bidang data science dan machine learning. Dataset tersebut mencakup 30000 entri, terdiri dari 14352 data orang yang terkena penyakit paru-paru dan 15648 data orang yang tidak terkena penyakit paru-paru. Data ini cocok untuk analisis prediktif penyakit paru-paru. untuk Fitur yang paling berpengaruh adalah merokok dan usia yang menjadikannya tolak ukur dalam data tersebut berikut tabel dibawah ini fitur dari data penyakit paru-paru tersebut. Atribut Usia dibagi menjadi dua kategori, yaitu Muda sebanyak 15.383 data dan Tua sebanyak 14.617 data. Atribut Jenis Kelamin terdiri dari Pria sebanyak 7.775 data dan Wanita sebanyak 22.225 data. Untuk atribut Merokok, kategori Aktif tercatat sebanyak 15.210

data, sedangkan Pasif sebanyak 14.790 data. Atribut Bekerja menunjukkan bahwa responden yang bekerja berjumlah 18.964 data, sedangkan yang tidak bekerja 11.036 data. Pada atribut Rumah Tangga, responden yang menikah atau berumah tangga tercatat sebanyak 15.425 data dan yang belum sebanyak 14.575. Atribut Aktivitas Begadang menunjukkan bahwa 17.548 responden memiliki kebiasaan begadang, sementara 12.452 tidak. Untuk Aktivitas Olahraga, responden yang jarang berolahraga berjumlah 17.994 data, sedangkan yang sering berolahraga sebanyak 12.006 data. Atribut Asuransi mengindikasikan bahwa 21.239 responden memiliki asuransi dan 8.761 responden tidak. Sementara itu, atribut Penyakit Bawaan menunjukkan 19.350 responden memiliki penyakit bawaan dan 10.650 tidak. Terakhir, atribut Hasil mencatat bahwa 14352 responden terdiagnosis penyakit paru-paru dan 15648 data lainnya tidak.

c. Encoding

Encoding di Python adalah proses mengubah teks menjadi bytes agar bisa diproses oleh komputer. Proses sebaliknya disebut decoding. Encoding penting saat menyimpan, membaca, atau mentransfer data teks. Format yang sering digunakan adalah UTF-8 karena mendukung banyak karakter dari berbagai bahasa. Python menggunakan .encode() untuk mengubah string ke bytes dan .decode() untuk mengubah bytes ke string.

```
[ ]  le = LabelEncoder()
     for col in df.columns:
         df[col] = le.fit_transform(df[col])
```

Gambar 3. Encoding

d. Drop Data

Tahap awal pengolahan data dengan menghapus beberapa data merupakan langkah penting sebelum proses pemodelan dimulai. Dalam analisis untuk mendeteksi penyakit paru-paru, dataset yang digunakan memuat atribut ‘Hasil’ yang berfungsi sebagai variabel target atau label, yaitu yang menunjukkan apakah seseorang terdiagnosis penyakit paru-paru atau tidak. Pada langkah ini, atribut ‘Hasil’ dipisahkan dari dataset menggunakan fungsi drop yang ada dalam pustaka pandas di Python. Pemisahan ini bertujuan untuk membedakan antara atribut prediktor (variabel independen) dengan variabel target (variabel dependen) yang nantinya akan digunakan dalam pelatihan
dan penilaian model. Setelah atribut ‘Hasil’ dihapus dari dataset, fitur-fitur prediktor akan disimpan dalam variabel X, sementara nilai dari atribut ‘Hasil’ akan disimpan dalam variabel Y.

```
X = data.drop(['Hasil','No'], axis=1)
y = data['Hasil']
```

Gambar 4. Drop Data

e. Split Data

dataset dibagi menjadi dua bagian, yaitu data pelatihan dan data pengujian. Data input atau fitur disimpan dalam variabel X, sedangkan label atau target disimpan dalam variabel y. Fungsi train_test_split akan membagi X dan y menjadi X_train, X_test, y_train, dan y_test. Sebanyak 80 persen data akan digunakan untuk pelatihan, dan 20 persen sisanya untuk pengujian, sesuai dengan parameter test_size=0.2. Parameter random_state=42 digunakan agar hasil pembagian data tetap sama setiap kali kode dijalankan. Pembagian ini penting untuk melatih model pada satu bagian data dan menguji performanya pada bagian data yang lain.

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Gambar 5. Split data

Implementasi Algoritma C4.5

Pada tahap ini dilakukan proses implementasi program menggunakan bahasa pemrograman Python dengan menerapkan metode klasifikasi C4.5 yang telah dirancang pada tahap sebelumnya. Tujuan dari implementasi ini adalah membangun sebuah model prediksi berbasis data mining yang mampu mengklasifikasikan tingkat risiko seseorang terkena penyakit paru-paru. Proses klasifikasi dilakukan dengan menganalisis atribut-atribut yang relevan, seperti gejala fisik dan riwayat kesehatan, sehingga dapat digunakan sebagai dasar pengambilan keputusan dalam upaya deteksi dini penyakit paru-paru.

a. Klasifikasi

Pada bagian ini dilakukan proses penentuan kriteria atau atribut yang akan digunakan dalam proses klasifikasi untuk memprediksi risiko terkena penyakit paru-paru. Penentuan kriteria ini sangat penting karena akan mempengaruhi tingkat akurasi yang dihasilkan oleh model klasifikasi. Setiap atribut dipilih berdasarkan relevansinya terhadap kemungkinan seseorang mengalami gangguan paru-paru, seperti usia, jenis kelamin,

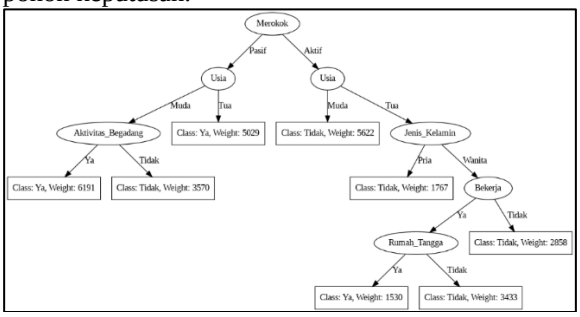
kebiasaan merokok, riwayat kesehatan, dan lingkungan tempat tinggal. Setelah kriteria ditentukan, dilakukan proses entry data atau input data ke dalam sistem, di mana data tersebut telah dipersiapkan dan divalidasi sebelumnya. Data ini kemudian digunakan sebagai dasar untuk melatih model dan menguji akurasinya. Hasil dari proses ini akan ditampilkan dalam bentuk evaluasi performa model, seperti nilai akurasi dan confusion matrix, yang menggambarkan sejauh mana model mampu melakukan klasifikasi secara tepat.

```
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report

# Bagi data menjadi data latih dan data uji
X = data.drop(['Hasil', 'No'], axis=1)
y = data['Hasil']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42

# Melatih model dengan data latih
model = C45Classifier()
model.fit(X_train, y_train)

# Melakukan prediksi terhadap data uji
y_pred = model.predict(X_test)

# Evaluasi performa model
print("Akurasi:", accuracy_score(y_test, y_pred))
print("\nConfusion Matrix:\n", confusion_matrix(y_test, y_pred))
print("\nClassification Report:\n", classification_report(y_test, y_pred))
```

Gambar 6. Klasifikasi

b. Hasil C4.5 Pohon Keputusan

salah satu metode klasifikasi dan prediksi yang digunakan dalam pembelajaran mesin (machine learning). Dalam algoritma seperti C4.5, pohon keputusan dibangun dengan cara membagi dataset ke dalam subset berdasarkan atribut yang memberikan gain informasi tertinggi, sampai mencapai kondisi klasifikasi tertentu. Berikut merupakan hasil dari pohon keputusan.



Gambar 7. Pohon Keputusan

Berdasarkan pohon keputusan yang telah ditampilkan, dapat disimpulkan bahwa atribut yang paling berpengaruh terhadap hasil klasifikasi adalah atribut "Merokok". Atribut ini muncul di posisi paling atas pohon keputusan (sebagai root node), yang berarti model menganggapnya sebagai variabel paling menentukan dalam membedakan antara kelas target "Ya" dan "Tidak". Setelah atribut Merokok, model selanjutnya membagi data berdasarkan atribut "Usia", yang juga muncul pada kedua sisi cabang utama, baik untuk yang merokok pasif maupun aktif. Hal ini menunjukkan bahwa usia juga merupakan variabel yang cukup penting setelah merokok.

Pada bagian cabang untuk perokok pasif, klasifikasi dilanjutkan dengan mempertimbangkan aktivitas begadang. Sementara itu, untuk perokok aktif, setelah mempertimbangkan usia, model memperhitungkan atribut jenis kelamin, kemudian pekerjaan, dan pada bagian akhir, atribut rumah tangga. Ini berarti bahwa pengaruh atribut dalam proses klasifikasi bersifat bertingkat, dimulai dari yang paling kuat di awal hingga yang lebih spesifik pada cabang-cabang akhir.

Secara keseluruhan, model memberikan penekanan utama pada status merokok dan usia sebagai penentu utama, sedangkan atribut-atribut lain seperti aktivitas begadang, jenis kelamin, pekerjaan, dan rumah tangga berperan dalam konteks tertentu yang lebih sempit pada cabang pohon keputusan. Dengan demikian, dapat dikatakan bahwa klasifikasi berjalan sesuai dengan struktur dan logika pohon yang dibangun, serta menunjukkan bahwa model telah sukses dalam mengidentifikasi atribut-atribut yang paling berpengaruh berdasarkan data yang diberikan.

Evaluasi

a. Akurasi

Pada bagian ini, ditampilkan hasil evaluasi akurasi dari proses klasifikasi yang dilakukan menggunakan algoritma Decision Tree (pohon keputusan). Evaluasi ini dilakukan dengan membandingkan hasil prediksi sistem terhadap data aktual, yang kemudian divisualisasikan dalam bentuk confusion matrix. Confusion matrix ini menunjukkan seberapa baik model dalam mengklasifikasikan data ke dalam kategori risiko penyakit paru-paru, seperti benar positif (TP), benar negatif (TN), salah positif (FP), dan salah negatif (FN). Melalui hasil

tersebut, dapat dianalisis tingkat akurasi, presisi, recall, dan f1-score dari model klasifikasi, sehingga diperoleh gambaran sejauh mana sistem yang dibangun mampu memprediksi risiko penyakit paru-paru secara tepat dan andal berdasarkan data yang tersedia.

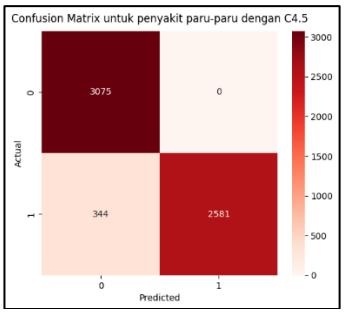
Berdasarkan hasil evaluasi model klasifikasi C4.5, diperoleh nilai akurasi sebesar 94% yang menunjukkan bahwa model mampu memprediksi kategori risiko penyakit paru-paru dengan tingkat ketepatan yang sangat tinggi. Hasil Confusion Matrix memperlihatkan bahwa dari total 3.075 data kelas "Tidak" (tidak berisiko), seluruhnya berhasil diprediksi dengan benar (True Negative = 3.075, False Positive = 0). Sementara itu, dari 2.925 data kelas "Ya" (berisiko), sebanyak 2.581 data diprediksi dengan benar (True Positive) dan 344 data salah diklasifikasikan sebagai "Tidak" (False Negative). Pada hasil Classification Report, kelas "Tidak" memiliki precision sebesar 0,90, recall 1,00, dan F1-score 0,95, yang menunjukkan bahwa seluruh kasus "Tidak" berhasil dikenali model tanpa ada yang terlewat. Sementara itu, kelas "Ya" memiliki precision sempurna sebesar 1,00 dan recall 0,88, dengan F1-score 0,94, yang berarti semua prediksi "Ya" yang dihasilkan model benar, namun masih terdapat sekitar 12% kasus "Ya" yang tidak terdeteksi. Secara keseluruhan, nilai macro average dan weighted average untuk precision, recall, dan F1-score berada di kisaran 0,94–0,95, yang menegaskan konsistensi performa model pada kedua kelas.

```
Akurasi: 0.9426666666666667

Confusion Matrix:
 [[3075    0]
 [ 344 2581]]

Classification Report:
              precision    recall  f1-score   support

       Tidak       0.90      1.00      0.95      3075
          Ya       1.00      0.88      0.94      2925

    accuracy                           0.94      6000
   macro avg       0.95      0.94      0.94      6000
weighted avg       0.95      0.94      0.94      6000
```

Gambar 8.Hasil Akurasi

b. Visualisai Confusion Matrix

Visualisasi confusion matrix adalah bentuk penyajian hasil evaluasi model klasifikasi dalam bentuk tabel matriks yang memperlihatkan perbandingan antara prediksi model dan nilai sebenarnya dari data uji. Matriks ini biasanya berbentuk kotak dengan sumbu horizontal mewakili kelas prediksi dan sumbu vertikal mewakili kelas aktual. Setiap sel di dalam matriks menunjukkan jumlah data yang masuk pada kombinasi tertentu antara prediksi dan kenyataan.



Gambar 9. Visualisai Confusion Matrix

Simpulan

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa algoritma klasifikasi C4.5 mampu digunakan secara efektif dalam memprediksi risiko terkena penyakit paru-paru berdasarkan atribut-atribut gejala dan faktor risiko yang dimiliki oleh pasien. Melalui proses data mining yang mencakup tahap seleksi data, preprocessing, pembentukan pohon keputusan, dan evaluasi model, sistem mampu menghasilkan model klasifikasi yang akurat, logis, dan dapat diinterpretasikan dengan baik.

C4.5 membentuk pohon keputusan berdasarkan nilai gain ratio tertinggi dari setiap atribut, sehingga mampu memisahkan data ke dalam kelas risiko secara efisien. Hasil pengujian menunjukkan bahwa model ini memiliki tingkat akurasi yang cukup tinggi dalam mengklasifikasikan data ke dalam kategori risiko (tinggi, sedang, atau rendah). Selain itu, struktur pohon yang terbentuk juga memberikan kemudahan bagi tenaga medis atau pengguna sistem untuk memahami pola dan hubungan antar gejala yang menjadi penentu risiko. Dengan demikian, implementasi algoritma C4.5 tidak hanya berguna sebagai alat bantu prediksi, tetapi juga sebagai media edukasi dan pendukung pengambilan keputusan dalam bidang kesehatan, khususnya dalam mendeteksi potensi penyakit paru-paru secara lebih dini dan tepat sasaran.

Penelitian ini memiliki keterbatasan karena data yang digunakan bersifat sekunder dan tidak mencakup data klinis langsung maupun data geospasial, serta tidak mempertimbangkan faktor genetik atau riwayat penyakit keluarga secara

mendalam. Untuk penelitian selanjutnya, disarankan agar data klinis real-time, data dari sensor kualitas udara, dan faktor genetik dapat diintegrasikan guna meningkatkan akurasi dan cakupan prediksi.

Daftar Pustaka

[1] L.Sari, A.Romadloni, Andr.Listyaningrum, "Penerapan Data Mining Dalam Analisis Prediksi Kanker Paru Menggunakan Algoritma Random Forest," *Infotekmesin*, Vol. 14, No. 1, Pp. 155–162, 2023, Doi: 10.35970/Infotekmesin.V14i1.1751.

[2] W. H.Organization, "Smoking Is The Leading Cause Of Chronic Obstructive Pulmonary Disease," 2023.

[3] "Riset Kesehatan Dasar (Riskesdas), 'Laporan Riskesdas 2018 Nasional.Pdf,' 2018.".

[4] R.Rofiani, L.Oktaviani, D.Vernanda, Andt.Hendriawan, "Penerapan Metode Klasifikasi Decision Tree Dalam Prediksi Kanker Paru-Paru Menggunakan Algoritma C4.5," *J. Tekno Kompak*, Vol. 18, No. 1, P. 126, 2024, Doi: 10.33365/Jtk.V18i1.3525.

[5] "S. Machfud And Y. Cahyono, 'Klasifikasi Penyakit Kanker Paru-Paru Menggunakan Metode C4 . 5,' Vol. 5, No. 2, Pp. 109–117, 2024, Doi: 10.31284/J.Kernel.2024.V5i2.7315.".

[6] F.Meila Azzahra Sofyan, A.Voutama, Andy.Umaidah, "Penerapan Algoritma C4.5 Untuk Prediksi Penyakit Paru-Paru Menggunakan Rapidminer," *Jati (Jurnal Mhs. Tek. Inform.*, Vol. 7, No. 2, Pp. 1409–1415, 2023, Doi: 10.36040/Jati.V7i2.6810.

[7] M. K.Amril Mutoi Siregar, S.Kom., M.Kom. Dan Adam Puspabhuana, S.Kom., *Data Mining: Pengolahan Data Menjadi Informasi Dengan Rapidminer*, Amril Muto. Cv Kekata Group.

[8] A.Naseh Khudori Andm.Syauqi Haris, "Implementasi Decision Tree Untuk Prediksi Kanker Paru-Paru," *J. Ris. Sist. Inf. Dan Tek. Inform. (Jurasik*, Vol. 9, No. 1, Pp. 94–106, 2024.

[9] Y.Son, "Development Of Methodology For Classification Of User Experience (Ux) In Online Customer Review," *J. Retail. Consum. Serv.*, Vol. 71, 2023, Doi: 10.1016/J.Jretconser.2022.103210.

[10] H. K.Dreiner, "The Abc Of Rpv: Classification Of R-Parity Violating Signatures At The Lhc For Small Couplings," *J. High Energy Phys.*, Vol. 2023, No. 7, 2023, Doi: 10.1007/Jhep07(2023)215.

[11] R.Krittayaphong, "Clinical Phenotype Classification To Predict Risk And Optimize The Management Of Patients With Atrial Fibrillation Using The Atrial Fibrillation Better Care (Abc) Pathway: A Report From The Cool-Af Registry," *Qjm An Int. J. Med.*, Vol. 117, No. 1, Pp. 16–23, 2024, Doi: 10.1093/Qjmed/Hcad219.

[12] G.Houge, "Stepwise Abc System For Classification Of Any Type Of Genetic Variant," *Eur. J. Hum. Genet.*, Vol. 30, No. 2, Pp. 150–159, 2022, Doi: 10.1038/S41431-021-00903-Z.

[13] D.Anugrah Pratama, I.Rizal Mutaqin, Andk.Rafael Manuela, "Analisis Terjadinya Kanker Paru-Paru Pada Pasien Menggunakan Decision Tree: Penerapan Algoritma C4.5 Dan Rapidminer Untuk Menentukan Risiko Kanker Pada Gejala Pasien," *Jtmei*, Vol. 2, No. 4, Pp. 156–170, 2023.

[14] D. T.Dinh Andv. N.Huynh, "*K*-Pbc: An Improved Cluster Center Initialization For Categorical Data Clustering," *Appl. Intell.*, Vol. 50, No. 8, Pp. 2610–2632, 2020, Doi: 10.1007/S10489-020-01677-5.

[15] S.Bensalem, S.Naouali, Andz.Chtourou, "Rough Mode: A Generalized Centroid Proposal For Clustering Categorical Data Using The Rough Set Theory," In *Big Data And Smart Digital Environment*, Y.Farhaoui Andl.Moussaid, Eds., Polytech Sch Tunisia, Bp 743,Rue El Khawarizmi, Tunis 2078, Tunisia Mil Acad Fondouk Jedid, Virtual Real & Informat Technol, Tunis, Tunisia Digital Res Ctr Sfax, Bp 275, Sfax 3021, Tunisia, 2019, Pp. 225–236. Doi: 10.1007/978-3-030-12048-1_24 10.1007/978-3-030-12048-1.

[16] I.Saha, J. P.Sarkar, Andu.Maulik, "Integrated Rough Fuzzy Clustering For Categorical Data Analysis," *Fuzzy Sets Syst.*, Vol. 361, Pp. 1–32, 2019, Doi: 10.1016/J.Fss.2018.02.007.

[17] M. T.Muhammad Arhami, S.Si., M.Kom., Muhammad Nasir, S.T., *Data Mining - Algoritma Dan Implementasi*. Andi Offset, 2020.

[18] "W. Yusnaeni And Widiarina, 'Penerapan Algoritma C4.5 Dalam Prediksi Resiko Diabetes Tahap Awal (Early Stage Diabetes),' J. Tek. Komput. Amik Bsi, Vol. 8, No. 2, Pp. 56–60, 2022, Doi: 10.31294/Jtk.V4i2.".

[19] "Mulaab, Data Mining : Konsep Dan Aplikasi. Media Nusa Creative (Mnc Publishing), 2021. [Online]. Available: Https://Books.Google.Co.Id/Books?Id=X1fkeaaaqbaj".

[20] "M. S. Ariantini Et Al., Sistem Pendukung Keputusan : Konsep,Metode, Dan Implementasi. Pt. Sonpedia Publishing Indonesia, 2023. [Online]. Available: Https://Books.Google.Co.Id/Books?Id=2e_Jeaaaqbaj".