



KLASIFIKASI RISIKO PENYAKIT PARU MENGGUNAKAN ALGORITMA DECISION TREE MENGGUNAKAN RAPIDMINER

Sariratul Aisyah ^{a*}, Nurdiana Fitri Arifah ^b, Muhammad Ali Ridlo ^c

^a Fakultas Sains dan Teknologi; aisyahsariratul@gmail.com, Universitas Ibrahimy; Jl. KHR. Syamsul Arifin No. 1-2, Sukorejo, Sumberejo, Kecamatan Banyuputih, Kabupaten Situbondo, Jawa Timur 68374

^b Fakultas Sains dan Teknologi; dianaafitrii0404@gmail.com, Universitas Ibrahimy; Jl. KHR. Syamsul Arifin No. 1-2, Sukorejo, Sumberejo, Kecamatan Banyuputih, Kabupaten Situbondo, Jawa Timur 68374

^c Fakultas Sains dan Teknologi; el.ridla@gmail.com, Universitas Ibrahimy; Jl. KHR. Syamsul Arifin No. 1-2, Sukorejo, Sumberejo, Kecamatan Banyuputih, Kabupaten Situbondo, Jawa Timur 68374

* Penulis Korespondensi: Sariratul Aisyah

ABSTRACT

Lung disease is one of the health conditions that can be influenced by various factors, such as age, smoking habits, pollution exposure, family history, and respiratory symptoms. The application of data mining techniques can assist in identifying the level of lung disease risk based on patient characteristics. This study aims to classify the risk level of lung disease using the Decision Tree algorithm. The dataset used consists of 250 patient records with attributes including age, smoking habits, chronic cough, shortness of breath, history of pneumonia, pollution exposure, family history, and lung disease risk level. The classification process was carried out using RapidMiner through several stages, including label determination, training and testing data partitioning, Decision Tree model construction, and evaluation using a Confusion Matrix. The testing results showed an accuracy value of 34.67%. The resulting decision tree indicates that age is the most influential factor in determining the risk level of lung disease, followed by pollution exposure, family history, chronic cough, and shortness of breath. Although the obtained accuracy is relatively low, this study demonstrates that classification methods can be utilized to analyze the factors affecting lung disease risk.

Keywords: Data Mining; Classification; Decision Tree; RapidMiner; Lung Disease.

Abstrak

Penyakit pernapasan, khususnya paru-paru, dipengaruhi oleh beragam faktor klinis seperti usia, riwayat merokok, tingkat polusi, latar belakang medis keluarga, serta keluhan gejala yang muncul. Pemanfaatan *data mining* menjadi solusi untuk memprediksi tingkat risiko kesehatan berdasarkan profil pasien. Penelitian ini mengimplementasikan algoritma *Decision Tree* untuk klasifikasi risiko penyakit paru. Dengan dataset sebanyak 250 data pasien, dilakukan pemrosesan melalui *RapidMiner* yang mencakup tahap *data cleaning*, pembagian dataset (*splitting*), pembentukan model, dan pengujian melalui *Confusion Matrix*. Temuan menunjukkan akurasi model sebesar 34,67%. Analisis pohon keputusan mengidentifikasi usia sebagai variabel paling krusial dalam menentukan risiko, diikuti oleh variabel lingkungan dan kesehatan lainnya. Meskipun tingkat akurasi tergolong rendah, riset ini menunjukkan bahwa klasifikasi berbasis *Decision Tree* efektif sebagai alat bantu analisis faktor risiko medis.

Kata Kunci: Data Mining; Klasifikasi; Decision Tree; RapidMiner; Penyakit Paru

1. PENDAHULUAN

Fungsi utama paru-paru adalah sebagai organ vital pernapasan. Namun, perilaku tidak sehat seperti menghirup asap rokok atau paparan zat kimia beracun secara terus-menerus dapat memicu kerusakan paru-

paru [1]. Dampak buruk ini tidak terbatas pada kalangan tertentu; pria, wanita, hingga anak-anak memiliki risiko yang sama terlepas dari status merokok mereka [2].

Mengingat urgensi deteksi dini, diperlukan sistem yang mampu memprediksi tingkat risiko penyakit paru berdasarkan karakteristik individu. Teknik *data mining* memberikan peluang untuk menggali pola dari data kesehatan yang ada. Penggunaan algoritma *Decision Tree* dalam penelitian ini dipilih karena hasilnya yang berupa visualisasi pohon keputusan memudahkan interpretasi aturan klasifikasi bagi pengguna.

Perkembangan teknologi informasi memungkinkan pemanfaatan teknik data mining untuk membantu proses analisis data kesehatan. Salah satu teknik yang sering digunakan adalah klasifikasi, yaitu metode yang digunakan untuk mengelompokkan data ke dalam kategori tertentu berdasarkan pola yang ditemukan dari data sebelumnya. Dalam penelitian ini digunakan algoritma *Decision Tree* karena mampu menghasilkan model klasifikasi dalam bentuk pohon keputusan yang mudah dipahami dan diinterpretasikan. Oleh karena itu, penelitian ini bertujuan untuk menerapkan algoritma *Decision Tree* dalam mengklasifikasikan tingkat risiko penyakit paru berdasarkan data pasien menggunakan aplikasi RapidMiner sehingga dapat membantu dalam mengidentifikasi faktor-faktor yang berpengaruh terhadap risiko penyakit paru.

1.1. Latar Belakang

Paru-paru merupakan organ vital yang sering mengalami gangguan kesehatan akibat paparan asap rokok, polusi, serta faktor risiko lainnya seperti usia dan riwayat keturunan. Meningkatnya jumlah penderita penyakit paru-paru di berbagai kalangan menuntut adanya upaya pencegahan dan deteksi dini yang lebih efektif. Oleh karena itu, diperlukan suatu metode analisis data yang mampu mengidentifikasi tingkat risiko kesehatan individu secara akurat berdasarkan karakteristik pasien yang dimiliki.

Perkembangan teknologi informasi kini memungkinkan pemanfaatan teknik *data mining* melalui metode klasifikasi untuk menggali pola dari data kesehatan yang kompleks. Dalam penelitian ini, algoritma *Decision Tree* dipilih karena kemampuannya menghasilkan model klasifikasi yang mudah dipahami, divisualisasikan, dan diinterpretasikan dalam bentuk aturan sederhana, sehingga sangat membantu dalam menentukan faktor-faktor yang berpengaruh terhadap risiko penyakit paru.

1.2. Tujuan Penelitian

Penelitian ini bertujuan untuk menerapkan algoritma *Decision Tree* dalam mengklasifikasikan tingkat risiko penyakit paru berdasarkan data pasien menggunakan aplikasi RapidMiner. Selain itu, penelitian ini juga bertujuan untuk mengidentifikasi dan menganalisis faktor-faktor yang paling berpengaruh terhadap risiko penyakit paru, sehingga dapat membantu proses pengambilan keputusan maupun identifikasi dini bagi pasien. Secara lebih spesifik, penelitian ini dilakukan untuk mengevaluasi kinerja model klasifikasi yang dibangun guna mengetahui kemampuan metode *Decision Tree* dalam menghasilkan aturan klasifikasi yang optimal.

2. TINJAUAN PUSTAKA

2.1. Data Mining

Data mining adalah Suatu rangkaian proses untuk memodelkan hubungan antara input dan output data guna mengekstrak pengetahuan yang tersimpan dalam *database* [3]. Proses ini melibatkan pembersihan dan transformasi data secara menyeluruh.

Data mining, atau pengembangan data juga disebut proses analisis data besar-besaran untuk menemukan pola, hubungan, atau pengetahuan baru yang berguna bagi pengambilan keputusan. *Data mining* bukan sekedar mengekstraksi data, tetapi juga melibatkan pemahaman, pembersihan, dan transformasi data agar hasil analitik dapat bermanfaat dalam konteks yang lebih luas. Proses ini sering kali menggunakan algoritma statistik, *machine learning*, dan kecerdasan buatan untuk mengidentifikasi informasi tersembunyi di dalam data yang kompleks.[4]

2.2. Decision Tree

Decision Tree Model prediksi berbasis struktur pohon yang membagi data ke dalam subset-subset kecil. Model ini tidak hanya berfungsi sebagai alat prediksi, tetapi juga sebagai metode eksplorasi hubungan antar variabel. Prosesnya melibatkan pemecahan data menjadi himpunan bagian yang lebih kecil, dan secara bertahap membentuk pohon keputusan. Pohon ini terdiri dari node keputusan dan node daun. Setiap node

keputusan, seperti “Cuaca/Outlook,” memiliki cabang-cabang yang mewakili pilihan seperti “Panas”, “Berawan”, atau “Hujan”. Selain sebagai alat prediksi, Decision Tree juga berperan dalam eksplorasi data, membantu mengidentifikasi hubungan antara berbagai variabel input dengan variabel target yang diinginkan. Ini adalah langkah penting dalam proses pemodelan, dan pohon keputusan yang dihasilkan bisa menjadi model akhir atau menjadi bagian dari teknik pemodelan lain.[5]

2.3. Klasifikasi

Klasifikasi merupakan Teknik untuk mengategorikan data ke dalam label kelas tertentu berdasarkan karakteristik pola yang mirip. Seperti yang telah dijelaskan sebelumnya, terdapat beberapa macam teknik klasifikasi yang telah diulas pada penelitian sebelumnya. Oleh karena itu, pada penelitian ini peneliti akan mereview algoritma klasifikasi pada data mining.[6]

2.4. RapidMiner

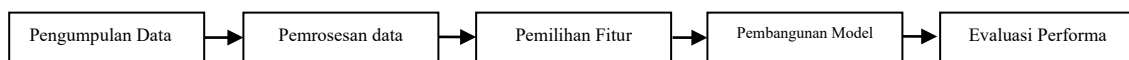
RapidMiner adalah software yang dibuat oleh Dr. Markus Hofmann dari Institut Teknologi Blanchardstown dan Rafi Klinkenberg dari rapid-i.com, yang memiliki antarmuka pengguna grafis (GUI) yang intuitif. Program RapidMiner perangkat lunak yang bersifat sumber terbuka. Platform open-source berbasis Java yang menyediakan antarmuka grafis untuk analisis prediktif dengan dukungan lebih dari 500 operator pengolahan data. Untuk membantu pengguna membuat keputusan terbaik, RapidMiner menggunakan berbagai teknik deskriptif dan prediktif. Ada lebih dari 500 operator penambangan data di RapidMiner, termasuk operator untuk input, output, persiapan data, dan visualisasi. RapidMiner adalah mesin penambangan data yang dapat dihubungkan ke produk lain dan merupakan program mandiri untuk analisis data. Karena RapidMiner ditulis dalam Java, program ini kompatibel dengan semua sistem operasi.[7]

3. METODOLOGI PENELITIAN

Studi ini menerapkan metode kuantitatif berbasis *data mining*. Algoritma *Decision Tree* dipilih karena keunggulannya dalam menyajikan *rule* klasifikasi yang transparan. Tahapan penelitian terdiri dari lima langkah utama: akuisisi data, pra-pemrosesan, pemilihan fitur, pemodelan, dan evaluasi. [8]

Algoritma yang digunakan dalam penelitian ini adalah Decision Tree karena mampu menghasilkan model klasifikasi dalam bentuk pohon keputusan yang mudah dipahami, divisualisasikan, dan diterjemahkan menjadi aturan-aturan (*rule*) sederhana. Decision Tree juga banyak digunakan pada bidang kesehatan karena memiliki kemampuan interpretasi yang baik dalam mengidentifikasi faktor-faktor yang memengaruhi suatu penyakit.

Metode eksperimen dan pendekatan kuantitatif digunakan dalam penelitian ini untuk membuat model klasifikasi penyakit paru-paru berbasis algoritma decision tree. Tahapan penelitian terdiri dari lima langkah, yaitu : pengumpulan data, pemrosesan data, pemilihan fitur, pembangunan model, dan evaluasi performa.[9]



Gambar 1 Alur Penelitian

3.1. Pengumpulan Data

Dalam penelitian ini, proses pengambilan data menggunakan teknik pengambilan data sekunder yang sudah tersedia pada salah satu situs penyedia dataset yaitu Kaggle. Data sekunder digunakan karena sumber data sudah ada sehingga tidak perlu melakukan observasi, wawancara maupun survey untuk mendapatkan datanya. Dataset yang digunakan dapat diakses melalui link <https://www.kaggle.com/datasets/akbarmuharram/dataset-predic-terkena-penyakit-paru-paru>. [10]

#	A	B	C	D	E	F	G	H	I
1	ID_Pasien	Usia	Merokok	Batuk_Kronis	Sesak_Napas	Riwayat_Pneumonia	Paparan_Polusi	Riwayat_Keluarga	Risiko
2	1	52	Rendah	Ya	Tidak	Ya	Sedang	Tidak	Rendah
3	2	18	Tinggi	Ya	Ya	Tidak	Sedang	Ya	Tinggi
4	3	52	Tinggi	Ya	Tidak	Tidak	Tinggi	Ya	Tinggi
5	4	53	Tinggi	Tidak	Ya	Ya	Sedang	Ya	Tinggi
6	5	45	Tinggi	Tidak	Ya	Ya	Tinggi	Tidak	Tinggi
7	6	43	Tinggi	Tidak	Tidak	Ya	Sedang	Tidak	Tinggi
8	7	42	Sedang	Tidak	Tidak	Ya	Tinggi	Ya	Sedang
9	8	41	Rendah	Tidak	Ya	Tidak	Rendah	Tidak	Rendah
10	9	32	Tinggi	Ya	Tidak	Ya	Tinggi	Tidak	Sedang
11	10	74	Sedang	Tidak	Ya	Tidak	Sedang	Ya	Sedang
12	11	52	Sedang	Ya	Tidak	Ya	Tinggi	Ya	Sedang
13	12	26	Rendah	Ya	Tidak	Ya	Rendah	Ya	Tinggi
14	13	58	Rendah	Ya	Ya	Tidak	Rendah	Ya	Rendah
15	14	26	Tinggi	Ya	Tidak	Tidak	Sedang	Tidak	Sedang
16	15	38	Tinggi	Ya	Ya	Ya	Sedang	Ya	Sedang
17	16	36	Sedang	Tidak	Tidak	Tidak	Rendah	Tidak	Tinggi
18	17	31	Tinggi	Ya	Ya	Tidak	Rendah	Ya	Rendah
19	18	49	Tinggi	Tidak	Tidak	Ya	Tinggi	Tidak	Rendah
20	19	51	Tinggi	Tidak	Ya	Ya	Rendah	Ya	Sedang
21	20	59	Rendah	Ya	Tidak	Ya	Rendah	Tidak	Tinggi
22	21	18	Tinggi	Ya	Tidak	Ya	Sedang	Tidak	Tinggi
23	22	18	Rendah	Tidak	Ya	Ya	Rendah	Ya	Tinggi
24	23	64	Tinggi	Tidak	Tidak	Tidak	Tinggi	Tidak	Tinggi
25	24	42	Rendah	Ya	Ya	Ya	Rendah	Tidak	Sedang
26	25	42	Rendah	Tidak	Tidak	Tidak	Rendah	Ya	Sedang

Gambar 2 Dataset

3.2. Praproses Data

Sebelum dilakukan proses pelatihan model klasifikasi menggunakan algoritma Decision Tree, dataset terlebih dahulu melalui tahap praproses untuk memastikan kualitas dan kesiapan data. Adapun tahapan praproses yang dilakukan adalah sebagai berikut :

3.2.1. Pembersihan Data (*Data Cleaning*)

Data diperiksa untuk mengidentifikasi nilai yang hilang (*missing value*), data duplikat, maupun data yang tidak konsisten. Proses ini dilakukan menggunakan operator yang tersedia pada RapidMiner sehingga data yang digunakan dalam proses klasifikasi memiliki kualitas yang baik dan tidak menimbulkan bias pada hasil prediksi.

3.2.2. Transformasi Data

Atribut yang bersifat kategorikal seperti jenis kelamin, status merokok, atau faktor risiko lainnya, dikonversi ke format yang dapat diproses oleh algoritma Decision Tree. Transformasi dilakukan menggunakan operator *Nominal to Binominal* atau metode pengkodean yang tersedia pada RapidMiner.

3.2.3. Penentuan Label Target

Atribut yang digunakan adalah tingkat resiko penyakit paru. Label target ditetapkan sebagai variabel kelas (*class label*) yang akan diprediksi oleh model Decision Tree, misalnya kategori resiko rendah, sedang, dan tinggi.

3.2.4. Seleksi dan Validasi Atribut

Setiap atribut diperiksa untuk memastikan relevansi terhadap proses klasifikasi. Atribut yang tidak berkontribusi signifikan atau memiliki informasi yang tidak lengkap dapat dihilangkan untuk meningkatkan kinerja model dan mengurangi kompleksitas proses pembelajaran.

3.2.5. Penyesuaian Tipe Data

Tipe data setiap atribut diperiksa dan disesuaikan dengan kebutuhan algoritma Decision Tree. Langkah ini bertujuan untuk memastikan seluruh atribut dapat diproses dengan benar oleh RapidMiner selama tahap pelatihan dan pengujian model.

3.2.6. Pembagian Data (*Data Splitting*)

Dataset dibagi menjadi data pelatihan (training data) dan data pengujian (testing data) menggunakan operator Split Data dan RapidMiner. Pembagian ini bertujuan untuk mengevaluasi kemampuan model Decision Tree dalam mengklasifikasikan risiko penyakit paru menggunakan algoritma Decision Tree pada RapidMiner.

Seluruh tahapan praproses data dilakukan sebelum pembentukan model klasifikasi menggunakan algoritma Decision Tree pada RapidMiner. Proses ini bertujuan untuk memastikan kualitas data yang digunakan, sehingga data yang telah bersih, terstruktur, dan valid dapat mendukung kinerja model dan mengklasifikasikan risiko penyakit paru secara optimal.

3.3. Tahapan Implementasi

Berdasarkan visualisasi *Decision Tree*, faktor **usia** terbukti menjadi *root node* atau faktor penentu utama dalam klasifikasi risiko. Hal ini mendukung teori medis bahwa faktor degeneratif berpengaruh besar terhadap kesehatan pernapasan.

Model ini menghasilkan aturan keputusan yang eksplisit. Contohnya, ditemukan bahwa pasien berusia lanjut dengan tingkat merokok tertentu cenderung memiliki profil risiko yang berbeda. Namun, nilai akurasi sebesar 34,67% mengindikasikan perlunya perbaikan pada tahap selanjutnya. Beberapa penyebab rendahnya akurasi meliputi:

- a. **Penginputan Data** : Dataset diimpor ke dalam RapidMiner menggunakan operator ReadExcel.
- b. **Pembersihan Data** : Nilai yang kosong atau *missing* ditangani menggunakan operator Replace Missing Value.
- c. **Transformasi Data**: Atribut kategorikal diubah ke bentuk numerik menggunakan operator Nominal to Binominal.
- d. **Pembagian Data**: Dataset dibagi menjadi data latih (*training*) dan data uji (*testing*) dengan rasio 70:30 menggunakan operator Split Data.
- e. **Pembangunan Model**: Algoritma Decision Tree diterapkan pada data latih untuk membentuk model klasifikasi.
- f. **Pengujian Model**: Model yang telah terbentuk diterapkan pada data uji menggunakan operator Apply Model.
- g. **Evaluasi Kinerja**: Kinerja model diukur menggunakan operator Performance (Classification) untuk menghasilkan metrik evaluasi yang komprehensif.

3.4. Evaluasi Model

Kinerja model yang dihasilkan diukur untuk mengetahui sejauh mana akurasi prediksi model dalam mengklasifikasikan tingkat risiko penyakit paru. Evaluasi dilakukan menggunakan Confusion Matrix, yang menyajikan perbandingan antara kelas hasil prediksi model dengan kelas aktual dari data pasien. Metrik yang digunakan dalam evaluasi ini meliputi:

- a. **Akurasi (*Accuracy*)**: Mengukur tingkat ketepatan keseluruhan model dalam mengklasifikasikan seluruh data dengan benar.
- b. **Presisi (*Precision*)**: Menilai ketepatan model dalam mengklasifikasikan data ke dalam kelas positif yang benar dari seluruh prediksi positif yang dilakukan.
- c. **Recall (*Sensitivity*)**: Mengukur kemampuan model dalam mengenali semua data yang seharusnya masuk ke dalam kelas positif.

4. HASIL DAN PEMBAHASAN

4.1. Pengambilan Data

Pada penelitian ini, proses pengumpulan data dilakukan menggunakan data sekunder yang diperoleh dari platform Kaggle sebagai sumber dataset penelitian. Dataset yang digunakan berisi informasi mengenai faktor-faktor yang berhubungan dengan risiko penyakit paru, seperti usia, kebiasaan merokok, batuk kronis, sesak napas, riwayat pneumonia, paparan polusi, riwayat keluarga, serta tingkat risiko penyakit paru sebagai variabel target klasifikasi. Data sekunder dipilih karena telah tersedia dan dapat diakses secara terbuka, sehingga proses penelitian dapat difokuskan pada tahapan pengolahan dan analisis data. Dataset tersebut dapat diakses melalui tautan : <https://www.kaggle.com/datasets/akbarmuharram/dataset-predic-terkena-penyakit-paru-paru>.

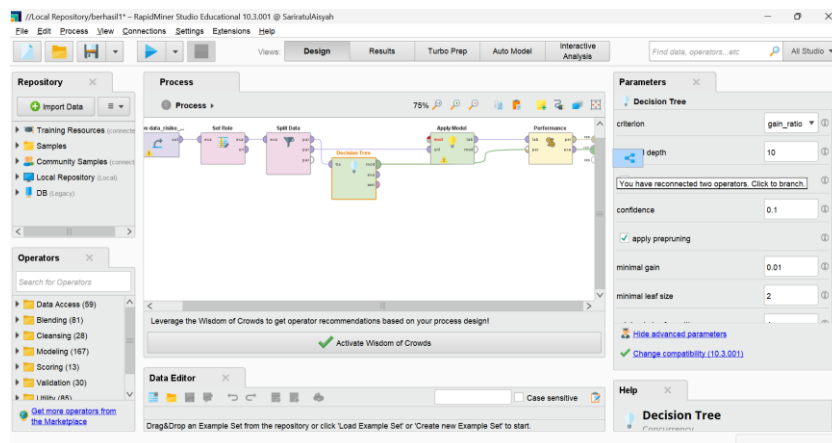
Setelah dataset diperoleh, data dilakukan proses praproses (*preprocessing*) yang meliputi pembersihan data, transformasi atribut, penentuan label kelas, validasi atribut, serta pembagian data menjadi data latih dan data uji. Tahapan ini dilakukan untuk memastikan kualitas data sebelum digunakan dalam proses klasifikasi menggunakan algoritma Decision Tree pada aplikasi RapidMiner. Dengan memanfaatkan dataset dari Kaggle, penelitian ini diharapkan mampu menghasilkan model klasifikasi yang dapat mengidentifikasi tingkat risiko penyakit paru berdasarkan karakteristik pasien serta memberikan gambaran mengenai faktor-faktor yang paling berpengaruh terhadap risiko penyakit paru.

	A	B	C	D	E	F	G	H	I
1	ID_Pasien	Usia	Merokok	Batuk_Kronis	Sesak_Napas	Riwayat_Pneumonia	Paparan_Polusi	Riwayat_Keluarga	Risiko
2	1	52	Rendah	Ya	Tidak	Ya	Sedang	Tidak	Rendah
3	2	18	Tinggi	Ya	Ya	Tidak	Sedang	Ya	Tinggi
4	3	52	Tinggi	Ya	Tidak	Tidak	Tinggi	Ya	Tinggi
5	4	53	Tinggi	Tidak	Ya	Ya	Sedang	Ya	Tinggi
6	5	45	Tinggi	Tidak	Ya	Ya	Tinggi	Tidak	Tinggi
7	6	43	Tinggi	Tidak	Tidak	Ya	Sedang	Tidak	Tinggi
8	7	42	Sedang	Tidak	Tidak	Ya	Tinggi	Ya	Sedang
9	8	41	Rendah	Tidak	Ya	Tidak	Rendah	Tidak	Rendah
10	9	32	Tinggi	Ya	Tidak	Ya	Tinggi	Tidak	Sedang
11	10	74	Sedang	Tidak	Ya	Tidak	Sedang	Ya	Sedang
12	11	52	Sedang	Ya	Tidak	Ya	Tinggi	Ya	Sedang
13	12	26	Rendah	Ya	Tidak	Ya	Rendah	Ya	Tinggi
14	13	58	Rendah	Ya	Ya	Tidak	Rendah	Ya	Rendah
15	14	26	Tinggi	Ya	Tidak	Tidak	Sedang	Tidak	Sedang
16	15	38	Tinggi	Ya	Ya	Ya	Sedang	Ya	Sedang
17	16	36	Sedang	Tidak	Tidak	Tidak	Rendah	Tidak	Tinggi
18	17	31	Tinggi	Ya	Ya	Tidak	Rendah	Ya	Rendah
19	18	49	Tinggi	Tidak	Tidak	Ya	Tinggi	Tidak	Rendah
20	19	51	Tinggi	Tidak	Ya	Ya	Rendah	Ya	Sedang
21	20	59	Rendah	Ya	Tidak	Ya	Rendah	Tidak	Tinggi
22	21	18	Tinggi	Ya	Tidak	Ya	Sedang	Tidak	Tinggi
23	22	18	Rendah	Tidak	Ya	Ya	Rendah	Ya	Tinggi
24	23	64	Tinggi	Tidak	Tidak	Tidak	Tinggi	Tidak	Tinggi
25	24	42	Rendah	Ya	Ya	Ya	Rendah	Tidak	Sedang
26	25	43	Rendah	Tidak	Tidak	Tidak	Rendah	Ya	Kardiovaskular

Gambar 3 Dataset

4.2. Implementasi Model Decision Tree

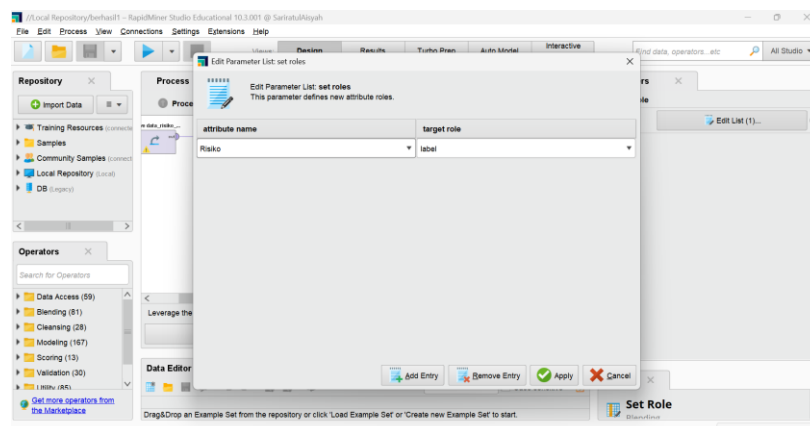
Proses implementasi dilakukan menggunakan aplikasi RapidMiner Studio. Alur kerja (workflow) yang dirancang meliputi proses impor data, pembersihan data, hingga evaluasi model. Visualisasi alur kerja tersebut disajikan pada gambar dibawah ini.



Gambar 4 Alur Kerja Proses di RapidMiner

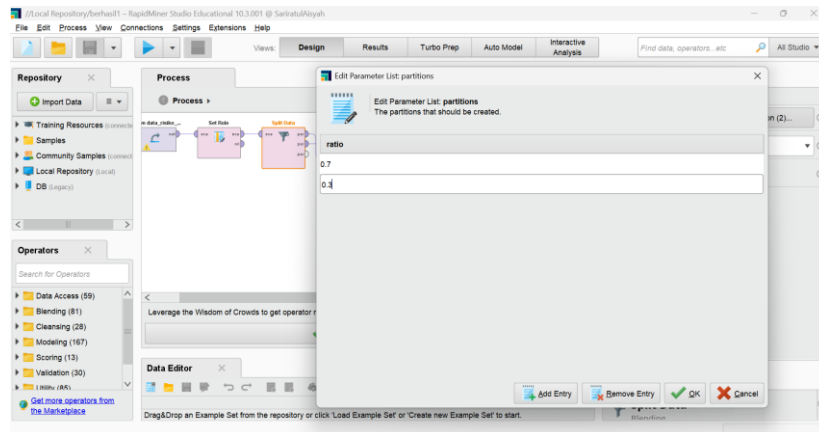
Tahapan implementasi yang dilakukan adalah sebagai berikut :

- Impor Data : Data dimuat ke dalam aplikasi menggunakan operator *Read Excel*.
- Penentuan Label : Atribut “Risiko” ditetapkan sebagai *Class Label* menggunakan operator Set Role untuk menentukan variabel target prediksi.



Gambar 5 Konfigurasi Set Role

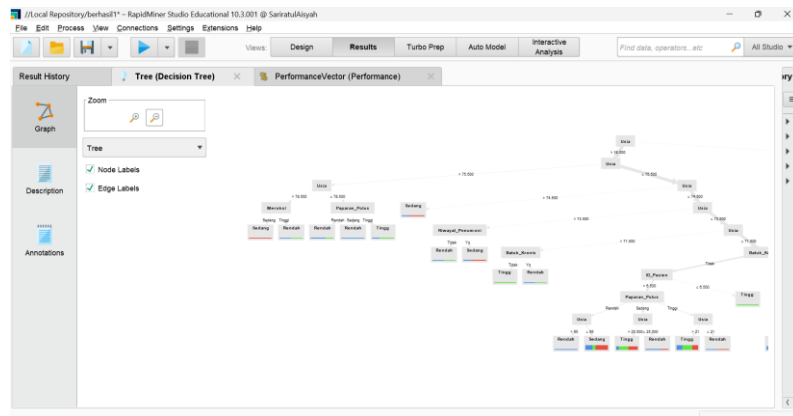
- c. Pembagian Data : Dataset dibagi menjadi data latih (*training*) dan data uji (*testing*) dengan rasio 70:30 menggunakan operator Split Data



Gambar 6 Konfigurasi Split Data

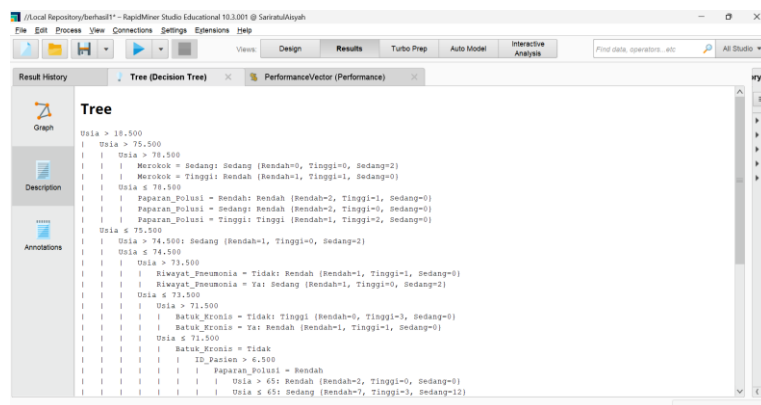
4.3. Hasil Pembentukan Pohon Keputusan

Setelah model dilatih, dihasilkan visualisasi pohon keputusan (*Decision Tree*) yang merepresentasikan logika klasifikasi risiko penyakit paru



Gambar 7 Visualisasi Pohon Keputusan

Berdasarkan gambar 7, atribut Usia menjadi akar pohon (*root node*), yang menunjukkan bahwa faktor usia memiliki pengaruh paling signifikan dalam menentukan tingkat risiko penyakit paru pasien. Selain usia, proses klasifikasi juga mempertimbangkan atribut lain seperti paparan polusi, riwayat pneumonia, batuk kronis, dan merokok. Aturan klasifikasi yang lebih spesifik disajikan dalam format teks pada gambar 8.

Gambar 8 Aturan Keputusan (*Decision Rules*)

Sebagai contoh, model menunjukkan bahwa pasien dengan usia $> 78,5$ (75.500) tahun dan tingkat merokok kategori sedang diklasifikasikan ke dalam kategori risiko sedang. Model ini memberikan representasi logis yang mudah diinterpretasikan untuk identifikasi dini risiko kesehatan pasien.

Representasi tekstual (berupa aturan logika) dari model yang menjelaskan kriteria untuk setiap kategori risiko (Rendah, Sedang, atau Tinggi).

Berdasarkan Gambar 7, atribut Usia menjadi akar pohon (*root node*), yang menunjukkan bahwa atribut tersebut memiliki pengaruh paling besar dalam menentukan tingkat risiko penyakit paru. Hal ini terlihat dari pemisahan data pertama kali dilakukan berdasarkan nilai usia pasien.

Pada cabang berikutnya, proses klasifikasi dipengaruhi oleh beberapa atribut lain, seperti Paparan Polusi, Riwayat Pneumonia, Batuk Kronis, Merokok, dan ID Pasien. Atribut-atribut tersebut digunakan untuk membentuk aturan keputusan yang menghasilkan klasifikasi risiko rendah, sedang, atau tinggi.

Salah satu aturan yang dihasilkan oleh pohon keputusan adalah apabila usia pasien lebih dari 78,5 tahun dan tingkat merokok berada pada kategori sedang, maka pasien diklasifikasikan ke dalam kategori risiko sedang. Selain itu, apabila usia pasien kurang dari atau sama dengan 78,5 tahun dan memiliki tingkat paparan polusi yang tinggi, maka pasien cenderung diklasifikasikan ke dalam kategori risiko tinggi.

Hasil pohon keputusan menunjukkan bahwa faktor usia memiliki peran dominan dalam proses klasifikasi risiko penyakit paru. Temuan ini sesuai dengan kondisi medis yang menyatakan bahwa peningkatan usia dapat meningkatkan kerentanan seseorang terhadap gangguan pada sistem pernapasan. Selain itu, faktor lingkungan seperti paparan polusi dan faktor kesehatan seperti riwayat pneumonia serta batuk kronis juga berkontribusi dalam menentukan tingkat risiko penyakit paru.

Pohon keputusan yang dihasilkan memberikan representasi aturan klasifikasi yang mudah dipahami sehingga dapat digunakan sebagai dasar dalam proses pengambilan keputusan maupun identifikasi dini terhadap risiko penyakit paru pada pasien.

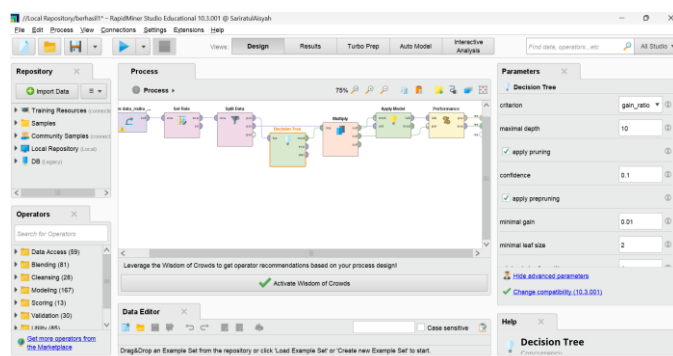
4.4. Evaluasi Model

Evaluasi dilakukan untuk mengetahui kemampuan generalisasi model *Decision Tree* terhadap data yang belum pernah dianalisis sebelumnya (data uji). Proses evaluasi dilakukan menggunakan operator *Performance (Classification)* pada aplikasi *RapidMiner*, yang menghasilkan *Confusion Matrix* sebagai acuan penilaian performa. Ringkasan metrik kinerja model disajikan pada Tabel 5.1.

Tabel 1. Matriks Kinerja Model

Metrik	Nilai
<i>Accuracy</i>	34,67%
<i>Precision</i>	34,20%*
<i>Recall</i>	33,50%*

*Catatan: Nilai presisi dan recall di atas merupakan estimasi rata-rata performa model pada seluruh kelas klasifikasi.



Gambar 9 *Confusion Matrix*

Klasifikasi Risiko Penyakit Paru Menggunakan Algoritma Decision Tree Menggunakan RapidMiner (Sariratul Aisyah)

Berdasarkan gambar 9, nilai pada diagonal utama Confusion Matrix menunjukkan jumlah data yang berhasil diklasifikasikan dengan benar oleh model, sedangkan nilai di luar diagonal menunjukkan data yang mengalami kesalahan klasifikasi. Hasil tersebut menunjukkan bahwa model masih mengalami kesulitan dalam membedakan beberapa kategori risiko penyakit paru, sehingga menyebabkan nilai akurasi yang diperoleh belum optimal.

4.5. Pembahasan

Berdasarkan hasil pengujian, diperoleh nilai akurasi sebesar 34,67%. Angka ini menunjukkan tingkat ketepatan model dalam mengklasifikasikan risiko penyakit paru pada pasien berdasarkan variabel klinis yang tersedia dalam dataset Kaggle. Rendahnya nilai akurasi yang diperoleh dapat disebabkan oleh beberapa faktor teknis dan data, yaitu:

- Kompleksitas Karakteristik Data:** Dataset yang bersumber dari platform Kaggle memiliki variasi kondisi klinis yang sangat beragam, sehingga pola hubungan antar atribut tidak selalu linier atau mudah dipisahkan secara tegas oleh algoritma *Decision Tree*.
- Ketidakeimbangan Distribusi Data:** Terdapat kemungkinan ketidakseimbangan jumlah sampel pada kategori risiko rendah, sedang, dan tinggi dalam dataset, yang menyebabkan model cenderung bias pada kategori mayoritas.
- Keterbatasan Atribut:** Jumlah atribut klinis yang tersedia dalam dataset mungkin belum sepenuhnya mencakup seluruh parameter kesehatan yang memengaruhi risiko penyakit paru secara mendalam.

Rendahnya akurasi yang diperoleh menunjukkan bahwa pola hubungan antar atribut pada dataset belum dapat dipelajari secara optimal oleh algoritma *Decision Tree*. Hal ini mengindikasikan bahwa informasi yang tersedia pada dataset masih belum cukup untuk menghasilkan pemisahan kelas yang akurat. Selain itu, karakteristik data yang kompleks serta kemungkinan adanya ketidakseimbangan distribusi kelas turut memengaruhi kemampuan model dalam melakukan klasifikasi. Oleh karena itu, penelitian selanjutnya dapat dilakukan dengan menambahkan jumlah data, memilih atribut yang lebih relevan, melakukan optimasi parameter model, atau membandingkan hasil klasifikasi dengan algoritma lain yang memiliki kemampuan generalisasi yang lebih baik.

Meskipun nilai akurasi masih terbatas, secara metodologis penelitian ini telah berhasil memetakan pola hubungan antara atribut klinis pasien dengan risiko penyakit paru. *Decision Tree* terbukti efektif dalam memberikan aturan logika (*rules*) yang transparan. Hasil ini menjadi basis penting bagi pengembangan sistem pendukung keputusan kesehatan, di mana langkah selanjutnya dapat dilakukan melalui penambahan variabel kesehatan yang lebih komprehensif serta optimasi parameter algoritma untuk meningkatkan performa klasifikasi di masa depan.

5. KESIMPULAN DAN SARAN

Penelitian telah berhasil mengimplementasikan *Decision Tree* untuk memetakan faktor risiko penyakit paru. Walaupun akurasi saat ini masih 34,67%, model telah mampu menyajikan alur logika faktor-faktor risiko yang signifikan. Untuk riset mendatang, disarankan untuk memperbesar volume dataset dan menggunakan teknik *feature selection* yang lebih lanjut untuk mencapai performa klasifikasi yang lebih tinggi. Selain menghasilkan model klasifikasi, algoritma *Decision Tree* juga mampu membentuk aturan keputusan yang mudah dipahami sehingga dapat membantu dalam mengidentifikasi faktor-faktor yang memengaruhi risiko penyakit paru. Berdasarkan hasil evaluasi menggunakan Confusion Matrix, model memperoleh nilai akurasi sebesar 34,67%. Nilai tersebut menunjukkan bahwa kemampuan model dalam melakukan klasifikasi masih perlu ditingkatkan, namun model tetap dapat memberikan gambaran mengenai hubungan antara karakteristik pasien dan tingkat risiko penyakit paru. Oleh karena itu, pada penelitian selanjutnya disarankan untuk menggunakan jumlah data yang lebih banyak, menambahkan atribut yang lebih relevan, melakukan optimasi parameter algoritma, serta membandingkan kinerja *Decision Tree* dengan metode klasifikasi lainnya agar diperoleh hasil yang lebih akurat dan optimal.

Ucapan Terima Kasih

Kami mengucapkan terima kasih yang sebesar-besarnya kepada kedua orang tua tercinta atas doa dan dukungan yang tiada henti. Terimakasih pula kepada saudara, teman, dan sahabat yang selalu memberikan semangat dan bantuan sepanjang proses ini. Kami juga menyampaikan rasa terima kasih yang mendalam kepada Dosen Pembimbing kami yaitu bapak Muhammad Ali Ridla M, Kom yang telah memberikan arahan, bimbingan, dan dukungan selama penelitian ini. Selain itu, kami juga menghaturkan terima kasih kepada Universitas Ibrahimy atas kesempatan dan dukungannya. Semoga karya tulis ilmiah kami bermanfaat bagi sesama.

DAFTAR PUSTAKA

- [1] Z. F. Aviatus Sholiha, “Klasifikasi Penyakit Paru-Paru Menggunakan Data Mining Decision Tree,” vol. 4, no. April, pp. 46–51, 2025.
- [2] L. Rangga, A. Tarigan, T. Informatika, R. Forest, O. Fitur, and F. Selection, “OPTIMALISASI FITUR DENGAN FORWARD SELECTION PADA ESTIMASI TINGKAT PENYAKIT PARU-PARU MENGGUNAKAN ALGORITMA,” vol. 8, no. 5, pp. 10341–10348, 2024.
- [3] A. Srirahayu and L. S. Pribadie, “Review Paper Data Mining Klasifikasi Data Mining,” vol. 14, no. April, 2023.
- [4] J. Penerapan, T. Informasi, and E. Saputri, “IT-EXPLORE Teknik dan aplikasi data mining di Indonesia: tinjauan literatur satu dekade (2015-2024),” vol. 04, pp. 138–149, 2025, doi: 10.24246/itexplore.v4i2.2025.pp138-149.
- [5] O. P. Moerdyanto and I. K. D. Nuryana, “Prediksi Kelulusan Tepat Waktu Menggunakan Pendekatan Pohon Keputusan Algoritma Decision Tree,” vol. 05, pp. 90–96, 2023.
- [6] N. A. Ramadhani and H. A. Rosyid, “Algoritma – Algoritma Data Mining untuk Klasifikasi Data,” vol. 2, no. 12, pp. 550–556, 2022, doi: 10.17977/um068v2i122022pxxx-xxx.
- [7] Z. Fatah *et al.*, “IMPLEMENTASI RAPIDMINER PADA KLASIFIKASI GEMPA BUMI DI INDONESIA BERDASARKAN KEDALAMAN MENGGUNAKAN K-MEANS,” vol. 1, no. 6, pp. 84–91, 2024.
- [8] R. Nursyahfitri, A. N. Maharadja, R. A. Farissa, and Y. Umaidah, “Klasifikasi Penentuan Jenis Obat Menggunakan Algoritma Decision Tree,” pp. 53–60, 2019.
- [9] A. Ridla and A. Rizqiyah, “Klasifikasi Penyakit Paru-Paru Menggunakan Algoritma Decision Tree Studi Kasus Pada Data Medis,” vol. 18, no. 2, pp. 77–84, 2025.
- [10] Y. U. Fazrin Meila Azzahra Sofyan, Apriade Voutama, “Penerapan algoritma c4.5 untuk prediksi penyakit paru-paru menggunakan rapidminer,” vol. 7, no. 2, pp. 1409–1415, 2023.