Jurnal Teknik Mesin, Industri, Elektro Dan Informatika (JTMEI)
Vol. 2 No. 4 Desember 2023
E-ISSN : 2963-7805 dan P-ISSN : 2963-8208, Hal 156-170
DOI: https://doi.org/10.55606/jtmei.v2i4.3004

Analisis Terjadinya Kanker Paru-Paru Pada Pasien Menggunakan Decision Tree: Penerapan Algoritma C4.5 Dan RapidMiner Untuk Menentukan Risiko Kanker Pada Gejala Pasien

Deigo Anugrah Pratama
Program Studi Sistem Informasi FTI, Universitas Bina Sarana Informatika

Ibnu Rizal Mutaqin
Program Studi Sistem Informasi FTI Universitas Bina Sarana Informatika

Kevin Rafael Manuela
Program Studi Sistem Informasi FTI Universitas Bina Sarana Informatika

Jl. Gatot Subroto No.8, Cimone, Kec. Karawaci, Kota Tangerang, Banten 15114, Indonesia.
Korespondensi Penulis: deigoanugrahp@gmail.com

Abstract. *The innovative approach to cancer patient data modeling has been employed in this research. We utilized the "Decision Tree" concept as a machine learning algorithm to analyze a dataset containing detailed information about patients, including age, gender, family history, and other medical test results. Through meticulous data study steps, we compiled a relevant dataset and then performed data classification to determine the target variable, whether a patient can be categorized as likely to have lung cancer or not. Input variables were carefully grouped to ensure the accuracy of the analysis. Data analysis using the Decision Tree algorithm provided profound insights into the significant factors in predicting cancer symptoms in patients. The results of this analysis were interpreted carefully, and performance model evaluation metrics, such as accuracy and precision, were provided to offer a comprehensive understanding of the reliability of the generated model. The findings of this research have important implications for the understanding and management of cancer in patients. The application of this method can enhance accuracy in predicting cancer status, assist in clinical decision-making, and ultimately improve the quality of patient care.*

Keywords*: Data analysis, Data Mining, RapidMiner, Decision Tree, C4.5 Algorithm*

Abstrak. Pendekatan inovatif dalam pemodelan data pasien kanker telah digunakan dalam penelitian ini. Kami memanfaatkan konsep "Pohon Keputusan" sebagai algoritma pembelajaran mesin untuk menganalisis dataset yang mencakup informasi rinci tentang pasien, termasuk usia, jenis kelamin, riwayat keluarga, dan hasil tes medis lainnya, Melalui langkah-langkah studi data yang cermat, kami mengumpulkan dataset yang relevan dan kemudian melakukan klasifikasi data untuk menentukan variabel target, apakah pasien dapat dikategorikan berpeluang terkena (kanker paru-paru atau tidak). Variabel input dikelompokkan dengan seksama untuk memastikan keakuratan analisis. Analisis data menggunakan algoritma Pohon Keputusan memberikan wawasan mendalam tentang faktor-faktor yang signifikan dalam memprediksi gejala kanker pada pasien. Hasil analisis ini diinterpretasikan dengan seksama, dan metrik evaluasi performa model, seperti akurasi dan presisi, diberikan untuk memberikan pemahaman yang komprehensif terhadap kehandalan model yang dihasilkan. Temuan penelitian ini memiliki implikasi penting dalam pemahaman dan penanganan kanker pada pasien. Penerapan metode ini dapat meningkatkan ketepatan dalam memprediksi status kanker, membantu dalam pengambilan keputusan klinis, dan pada gilirannya, meningkatkan kualitas perawatan pasien.

Kata kunci : Data analisis, Data Mining, RapidMiner, Pohon Keputusan, Algoritma C4.5

Received September 27, 2023; Revised Oktober 22, 2023; Accepted November 30, 2023
* Deigo Anugrah Pratama, *deigoanugrahp@gmail.com*

PENDAHULUAN

Kemajuan pesat dalam bidang data mining tak lepas dari perkembangan ilmu pengetahuan dan teknologi yang terus dikembangkan. Data mining merupakan suatu proses yang bertujuan untuk menemukan hubungan, pola, dan kecenderungan yang signifikan dalam sekumpulan data. Proses ini melibatkan penggunaan teknik pengenalan pola, seperti teknik statistik dan matematika (Larose, 2014).

Salah satu teknik yang umum digunakan dalam pengolahan data mining adalah klasifikasi. Klasifikasi merupakan proses identifikasi pola atau fungsi yang mampu mendeskripsikan dan memisahkan kelas-kelas data. Teknik ini juga digunakan untuk melakukan prediksi terhadap data yang belum memiliki kelas tertentu (Han & Kamber, 2006). Salah satu metode klasifikasi yang efektif adalah menggunakan pohon keputusan.

Pohon keputusan menjadi salah satu teknik yang dapat diandalkan untuk mengklasifikasikan sekumpulan objek. Pembuatan pohon keputusan dapat dilakukan secara manual dengan cermat atau dibuat secara otomatis melalui penerapan algoritma pohon keputusan. Algoritma ini digunakan untuk memodelkan himpunan data yang belum terklasifikasi.

Beberapa algoritma klasifikasi yang umum digunakan untuk konstruksi pohon keputusan antara lain adalah algoritma *Classification and Regression Trees* (CART), algoritma *Iterative Dichotomiser* (ID3), dan algoritma C4.5. Dalam artikel ini, penulis melakukan analisis terhadap beberapa algoritma klasifikasi pohon keputusan yang saat ini umum digunakan, khususnya Decision Tree dan algoritma C4.5. Analisis ini bertujuan untuk memberikan pemahaman lebih mendalam mengenai keefektifan algoritma decision tree untuk menganalisis kemungkinan terjadinya kanker paru-paru pada gejala pasien yang telah diberikan.

METODOLOGI PENELITIAN

Data Mining

Data mining adalah suatu proses analisis besar set data untuk menemukan pola yang bermanfaat, informasi tersembunyi, atau pengetahuan yang dapat digunakan untuk membuat keputusan lebih baik. Dalam data mining, teknik statistik, matematika, dan kecerdasan buatan digunakan untuk menjelajahi dan menganalisis dataset besar dengan tujuan menemukan pola yang tidak terlihat secara langsung. (Larose, 2014). Proses data mining menjadi fokus utama untuk menemukan pola dalam sejumlah data besar. Data mining yang digunakan dalam jurnal penelitian ini, di dapat melalui situs Kaggle. Proses ini merupakan bagian integral dari

Knowledge Discovery in Database (KDD), yang bertujuan untuk mengekstraksi informasi berharga dari basis data berukuran besar (Han & Kamber, 2006), Menurut (Larose, 2014) data mining dapat diklasifikasikan ke dalam beberapa kelompok berdasarkan tugas yang dapat dilakukannya. Pertama, klasifikasi melibatkan target variabel kategori, seperti penggolongan pendapatan ke dalam kategori tertentu. Kedua, estimasi mirip dengan klasifikasi, namun variabel target berfokus pada nilai numerik. Model dibangun menggunakan record lengkap yang memberikan nilai variabel target sebagai prediksi. Ketiga, prediksi, meskipun serupa dengan klasifikasi dan estimasi, difokuskan pada nilai hasil di masa mendatang.

- Klasifikasi

Klasifikasi melibatkan pemisahan data ke dalam kategori atau kelas tertentu. Sebagai contoh, penggolongan orang yang beresiko tinggi terkena kanker paru paru dapat dipisahkan dalam tiga kategori yaitu : perokok, usia, dan riwayat penyakit kronis.

- Estimasi

Walau memiliki kesamaan dengan klasifikasi, terdapat perbedaan pada variabel target estimasi yang lebih ke arah numerik daripada ke arah kategori. Ketika kita berbicara tentang estimasi, fokus utamanya adalah pada prediksi nilai numerik yang mungkin dihasilkan oleh suatu instansi data. Dalam hal ini, data mining berfungsi sebagai alat untuk memahami pola atau tren yang ada dalam dataset dan membangun model yang dapat memberikan perkiraan yang sesuai untuk variabel target yang bersifat kontinu. Misalnya, dalam dunia kesehatan, estimasi dapat diterapkan untuk memprediksi penyakit pada suatu pasien berdasarkan berbagai faktor seperti faktor umur, faktor tingkat aktivitas fisik, dan faktor genetik.

- Prediksi

Prediksi dalam data mining merujuk pada kemampuan untuk mengidentifikasi atau memproyeksikan nilai atau kejadian di masa mendatang berdasarkan pola atau tren yang ditemukan dalam dataset historis. Dalam konteks analisis data, prediksi melibatkan penggunaan model matematis atau statistik untuk memahami dan menganalisis hubungan antara variabel-variabel dalam data. Tujuannya adalah untuk memperoleh pemahaman yang lebih mendalam tentang bagaimana variabel-variabel tersebut berinteraksi dan bagaimana perubahan pada satu variabel dapat memengaruhi variabel lainnya. Contoh praktis dari prediksi dalam data mining dapat mencakup berbagai bidang, seperti pola prediksi risiko kesehatan berdasarkan riwayat medis.

- Pengklasteran

Pengklasteran dalam data mining merupakan suatu teknik yang digunakan untuk

mengelompokkan data atau objek-objek dalam dataset ke dalam kelompok-kelompok yang memiliki kemiripan tertentu. Tujuan utama dari pengklasteran adalah untuk mengidentifikasi struktur atau pola dalam data yang mungkin tidak terlihat secara langsung. Dalam proses ini, algoritma pengklasteran akan memisahkan data menjadi kelompok-kelompok atau klaster berdasarkan kesamaan karakteristik tertentu, sehingga objek-objek dalam satu klaster memiliki kemiripan yang lebih tinggi dibandingkan dengan objek-objek di klaster lainnya.

- Asosiasi

Asosiasi merujuk pada suatu tugas atau teknik analisis yang bertujuan untuk menemukan keterkaitan dan pola kejadian bersama antara atribut atau item dalam suatu dataset. Tugas asosiasi umumnya digunakan untuk mengidentifikasi hubungan yang muncul bersama-sama atau seringkali terjadi bersama-sama dalam suatu kumpulan data. Proses asosiasi melibatkan pencarian aturan asosiasi, yang menggambarkan hubungan antara berbagai atribut atau item. Aturan ini sering diungkapkan dalam bentuk "jika A maka B," yang berarti jika suatu kondisi atau atribut tertentu (A) terpenuhi, maka ada kemungkinan atribut atau item lain (B) juga muncul.

Analisis

Analisis adalah pemeriksaan terperinci terhadap segala sesuatu yang kompleks untuk memahami sifat atau menentukan ciri-ciri esensialnya. Ini adalah proses memecah topik atau substansi yang kompleks menjadi bagian-bagian yang lebih kecil untuk mendapatkan pemahaman yang lebih baik. Dalam konteks ilmiah atau penelitian, analisis sering kali melibatkan pemecahan masalah atau pertanyaan penelitian menjadi elemen-elemen yang dapat dianalisis secara terpisah. Proses analisis ini dapat melibatkan penggunaan metode atau teknik tertentu, seperti metode statistika, analisis data kualitatif, atau metode matematis, tergantung pada bidang atau masalah yang sedang dipelajari, Secara umum, analisis dapat mencakup beberapa tahapan seperti:

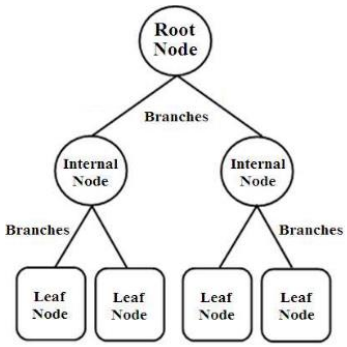
- Pengidetifikasian Komponen: Menentukan elemen-elemen atau komponen-komponen yang relevan dalam suatu konteks atau dataset.
- Penguraian: Membongkar atau memecah informasi menjadi bagian-bagian yang lebih kecil atau detail untuk mempermudah pemahaman.
- Pemahaman: Memeriksa dan memahami hubungan antara komponen-komponen tersebut.
- Interpretasi: Menafsirkan hasil analisis dan mengambil kesimpulan atau implikasi dari temuan tersebut.

Pohon Keputusan (Decision Tree)

Pohon keputusan merupakan struktur pohon yang setiap simpul internalnya mewakili pengujian terhadap suatu atribut, cabang mewakili hasil pengujian, dan simpul daun mewakili label kelas atau keputusan. Dalam konteks algoritma C4.5, pohon keputusan digunakan untuk mengklasifikasikan *new instance* berdasarkan data pelatihan yang disediakan. Algoritma membangun pohon keputusan menggunakan konsep *entropy* informasi dan memilih atribut data yang paling efektif membagi kumpulan sampelnya menjadi himpunan bagian yang diperkaya dalam satu kelas atau kelas lainnya. Komponen kunci dari pohon keputusan meliputi:

- *Root Node:* Node paling atas dari pohon, yang mewakili keseluruhan masalah atau keputusan yang harus diambil.
- *Note Internal:* Node yang mewakili pengujian pada atribut. Setiap node internal memiliki sejumlah pointer ke node anak, yang mewakili kemungkinan hasil pengujian.
- *Leaf Nodes:* Node yang mewakili label atau keputusan kelas. Node daun tidak memiliki pointer ke *child node*, dan node tersebut memberikan *output* dari pohon keputusan.
- *Branches:* Node yang mewakili label atau keputusan kelas. Node daun tidak memiliki pointer ke node anak, dan node tersebut memberikan keluaran akhir dari pohon keputusan.

Contoh Pohon Keputusan (*Decision Tree*) ditunjukkan pada Gambar 1.



Gambar 1. *Contoh Pohon Keputusan.*

Dahan-dahan dalam struktur pohon keputusan mengemban peran sebagai pertanyaan klasifikasi, sementara pada ujung-ujungnya terletak kelas-kelas atau kelompok data. Fungsi utama dari algoritma C4.5 adalah menjalankan tugas klasifikasi dengan mengelompokkan data ke dalam kategori-kategori yang spesifik melalui pengolahan dataset. Pohon keputusan memiliki kegunaan yang luar biasa dalam mengeksplorasi dataset dan mengungkapkan keterkaitan tersembunyi antara sejumlah variabel input potensial dengan variabel target. Dengan demikian, keunggulan utama penggunaan pohon keputusan terletak pada kapasitasnya dalam menyederhanakan proses pengambilan keputusan yang rumit, memberikan kemudahan

bagi pengambil keputusan untuk mengartikan solusi dari permasalahan yang dihadapi.

Berikut adalah keuntungan metode pohon keputusan.

- Pohon keputusan dapat diinterpretasikan dengan mudah oleh manusia. Struktur pohon yang berhierarki memungkinkan pengambil keputusan untuk memahami alur logika dan faktor-faktor yang memengaruhi keputusan.
- Pohon keputusan mampu melakukan pemrosesan data secara otomatis. Setelah pohon keputusan dibuat, pengguna dapat dengan cepat dan mudah memprediksi hasil untuk *instance data* yang baru.
- Pohon keputusan dapat menangani dengan baik data kategorikal maupun numerik tanpa memerlukan transformasi khusus. Hal ini menjadikanya fleksibel dalam menangani berbagai jenis data.
- Pohon keputusan merupakan metode *nonparametric*, yang berarti tidak memerlukan asumsi tertentu tentang distribusi data.
- Pohon keputusan selama pembuatannya memiliki algoritma yang secara otomatis memilih fitur-fitur yang paling informatif untuk membuat keputusan, mengurangi kebutuhan intervensi manusia dalam proses ini.
- Dalam analisis *multivariat*, dengan kriteria dan kelas yang jumlahnya cukup banyak, seorang penguji biasanya perlu untuk mengestimasikan distribusi dimensi tinggi ataupun parameter tertentu dari distribusi kelas tersebut. Metode pohon keputusan menghindari munculnya permasalahan dengan menggunakan kriteria yang jumlahnya sedikit pada setiap node internal.

Berikut adalah kekurangan pada pohon keputusan.

- Pohon keputusan rentan terhadap *overfitting*, terutama jika pohon keputusan dibuat dengan terlalu detail dan tidak diberlakukan teknik pemotongan yang memadai.
- Pohon keputusan cenderung menjadi tidak stabil, terutama pada kecilnya perubahan dalam data yang dapat menghasilkan struktur pohon yang berbeda. Pada akhrinya dapat mengakibatkan keputusan yang kurang konsisten.
- Pohon keputusan memiliki keterbatasan dalam menangani masalah *XOR* dan masalah logika yang lebih kompleks, karena memerlukan beberapa level pemisahan untuk mengatasi masalah tersebut.
- Pohon keputusan dapat sangat sensitif terhadap perubahan kecil dalam data, yang dapat menyebabkan varian yang tinggi dalam struktur pohon keputusan.
- Pohon keputusan memiliki keterbatasan dalam merepresentasikan hubungan linear

kompleks dalam data, sehingga dapat kalah efektif jika data memiliki hubungan yang lebih kompleks secara linear.

Algoritma C4.5

Algoritma C4.5, adalah algoritma pembelajaran mesin yang digunakan untuk membuat pohon keputusan dalam konteks analisis data dan klasifikasi yang dibuat oleh Ross Quinlan. Algoritma ini merupakan pengembangan dari pendahulunya, *ID3 (Iterative Dichotomiser 3),* dan dirancang untuk menangani berbagai jenis data, baik data kategorikal maupun numerik. Tujuan utama dari C4.5 adalah membangun pohon keputusan yang efisien untuk mengklasifikasikan instance data ke dalam kategori atau kelas tertentu berdasarkan serangkaian aturan keputusan. Algoritma C4.5, terdapat beberapa *input* atau *field* yang menjadi bagian integral dalam proses pembentukan pohon keputusan. Input utama dari algoritma ini adalah dataset atau sampel pelatihan (*training samples*). Dataset ini terdiri dari sejumlah *instance data* yang memiliki atribut atau fitur yang dapat diukur. Setiap *instance data* dalam dataset membawa informasi tentang variabel input dan variabel target yang ingin diprediksi atau diklasifikasikan.

Secara umum Algorima C4.5 untuk membangun pohon keputusan adalah sebagai berikut:

1. Pilih atribut sebagai akar
2. Buat cabang untuk tiap-tiap nilai
3. Bagi kasus dalam cabang
4. Ulangi proses untuk tiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.

Untuk memilih atribut sebagai node akar, didasarkan pada nilai Gain tertinggi dari atribut-atribut yang ada. Untuk menghitung Gain digunakan rumus seperti tertera dalam Gambar 2 berikut:

$$Gain\ (S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy\ (Si)\ (1)$$

Gambar 2. *Rumus Gain*

Keterangan:

S : Kelompok kasus

A : Atribut

n : Total partisi atribut

|Si| : Total kasus pada partisi ke-i

|S| : Total S dalam S

Setelah memperoleh nilai Gain, terdapat langkah tambahan yang harus di lakukan dalam perhitungan, yakni menentukan nilai *Entropy*. *Entropy* digunakan untuk mengevaluasi seberapa informatif sebuah atribut masukan dalam menghasilkan atribut keluaran. Rumus dasar *Entropy* tertera dalam Gambar 3 berikut:

$$Entropy(S) = \sum_{i=1}^{n} -pi \ . \ log2 \ pi$$

Gambar 3. *Rumus dasar Entropy*

Keterangan:

S : Kelompok kasus

n : Total partisi atribut

pi : Proporsi dari Si terhadap S

RapidMiner

RapidMiner merupakan *software* platform analitik data yang digunakan untuk mengelola, menganalisis, dan memodelkan data. Platform ini menyediakan lingkungan pengembangan visual yang memungkinkan pengguna, sistem nya di buat lebih mudah untuk pengguna yang pemula dalam bidang data visual dan analisis agar dapat melakukan proses visual dan analisis data tanpa harus menulis kode secara manual. RapidMiner menawarkan berbagai alat dan fitur, termasuk penyajian data, pemrosesan data, pemodelan prediktif, dan evaluasi model. Dengan antarmuka pengguna yang intuitif, pengguna dapat membuat alur kerja analisis data dengan menyusun dan menghubungkan berbagai operasi analisis. RapidMiner juga mendukung integrasi dengan berbagai sumber data, termasuk database, file teks, dan sumber data online. Dengan fitur-fitur ini, RapidMiner memudahkan para profesional dan analis data untuk menjelajahi, memahami, dan memanfaatkan potensi informasi yang terkandung dalam data mereka.

Pemahaman Kanker Paru-paru

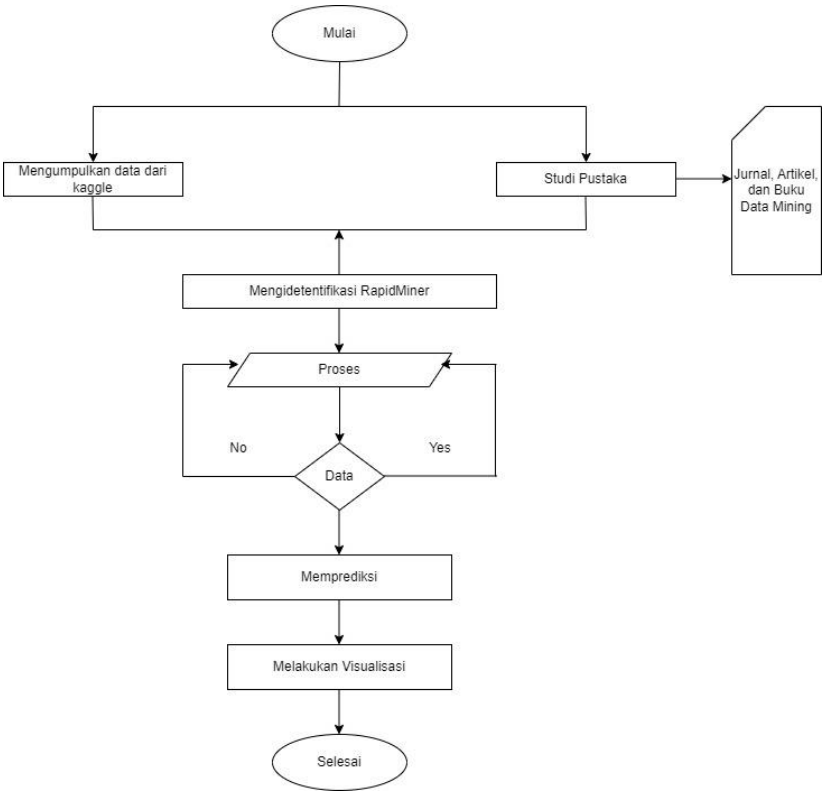
Kanker paru-paru merupakan jenis kanker yang bermula ketika sel-sel di dalam paru-paru mengalami pertumbuhan tidak terkendali. Paru-paru adalah organ vital dalam sistem pernapasan yang berfungsi untuk mengambil oksigen dari udara dan membuang karbon dioksida. Kanker paru-paru dapat dibagi menjadi dua kategori utama: *Non-Small Cell Lung Cancer (NSCLC)* dan *(Small Cell Lung Cancer): (SCLC)*. Pengobatan kanker paru-paru tergantung pada stadium penyakit, namun melibatkan berbagai metode seperti pembedahan, kemoterapi, dan radioterapi.

Penyebab utama kanker paru-paru umumnya terkait dengan paparan jangka panjang terhadap zat-zat berbahaya, terutama asap rokok. Meskipun demikian, tidak semua kasus kanker paru-paru terkait dengan merokok, karena faktor-faktor genetik dan lingkungan juga

dapat memainkan peran. Gejala kanker paru-paru meliputi batuk persisten, sesak napas, nyeri dada, kehilangan berat badan, dan batuk berdarah. Dikutip dari jurnal *Predicting lung cancer in patients with cough and risk factors*, Perkiraan kemungkinan terjadinya kanker paru-paru pada pasien yang mengalami gejala kanker paru-paru dan memiliki faktor risiko kanker paru-paru adalah sekitar 50% (Christopher A. Miller, 2022). Dan dikutip dari jurnal *Age-specific prognosis of non-small cell lung cancer in patients with cough*, Perkiraan kemungkinan terjadinya kanker paru-paru pada pasien yang berusia di atas 50 tahun dan mengalami gejala kanker paru-paru adalah sekitar 70% (Lee, J. H, 2021). Walau begitu dokter tempat anda dirawat atau berobat memiliki faktor persentasenya secara individual sebelum dilakukannya test-test terkait, deteksi dini melalui pemeriksaan rutin dan gaya hidup sehat dapat membantu mengurangi risiko terkena kanker paru-paru.

Rancangan Penelitian

Perancangan penelitian ini digunakan untuk menganalisis dan memberikan rekomendasi keputusan tentang penyebab terbesar pada pasien kanker paru paru yang dapat dilihat dalam rancangan penelitian pada Gambar 4 berikut:



***Gambar 4**. Rancangan penelitian*

1. Mengumpulkan Data

   Pada tahapan ini data yang di analisis, dalam jurnal ini diperoleh dari situs :

   https://www.kaggle.com/code/sandragracenelson/lung-cancer-prediction/input

2. Studi Pustaka

   Pada tahapan ini kami mengumpulkan dan melengkapi pengetahuan dasar, bahan bahan penelitian dan teori teori untuk melengkapi penelitian analisis ini.

3. Mengidentifikasi Masalah Dengan RapidMiner

   Pada tahapan ini memproses tahap konservasi data dan mengidentifikasi masalah yang akan di selesaikan dengan analisis dari data yang telah di tentukan.

4. Proses

   Pada tahapan ini dilakukanya proses analisis data dengan melakukan preview untuk memastikan data dapat digunakan kedalam aplikasi RapidMiner.

5. Menguji Hasil Pengolahan data

   Pada tahapan ini kami melakukan uji coba terhadap hasil pengolahan data dengan menggunakan aplikasi RapidMiner.

6. Memprediksi

   Pada tahapan ini Prediksi dilakukan untuk melihat seberapa akurat data pasien kanker paru-paru yang dimiliki dengan metode Algortima C4.5

7. Melakukan Visualisasi

   Pada tahapan ini kami melakukan visualiasasi dengan pohon keputusan (*Decision Tree*) untuk memprediksi alasan terbanyak pasien mengalami penyakit kanker paru-paru pada visual yang dihasilkan.

PEMBAHASAN

Dalam melakukan analisis terhadap pasien kanker paru-paru serta faktor-faktor apa saja yang dapat meningkatkan terjadinya kanker paru-paru pada pasien, dapat menggunakan beberapa tahapan antara lain sebagai berikut:

- Menyiapkan data yang telah di olah
- Menentukan atribut yang akan di gunakan untuk melakukan metode *Decision Tree*
- Melakukan visualisasi dengan metode Algortima C4.5, dan *Decision Tree* menggunakan
- aplikasi RapidMiner

Pengolahan Data

Pengolahan data digunakan untuk melakuakan pembersihan dan *preview* pada data yang akan di analisis, adapun sebagai berikut :

- Kriteria Data

Kriteria Data yang digunakan dapat dilihat pada Tabel menggunakan gambar berikut ini:

	GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC DISEASE	FATIGUE	ALLERGY	WHEEZING	ALCOHOL CONSUMING	COUGHING	SHORTNESS OF BREATH	SWALLOWING DIFFICULTY	CHEST PAIN	LUNG_CANCER
0	M	69	1	2	2	1	1	2	1	2	2	2	2	2	2	YES
1	M	74	2	1	1	1	2	2	2	1	1	1	2	2	2	YES
2	F	59	1	1	1	2	1	2	1	2	1	2	2	1	2	NO
3	M	63	2	2	2	1	1	1	1	1	2	1	1	2	2	NO
4	F	63	1	2	1	1	1	1	1	2	1	2	2	1	1	NO
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
304	F	56	1	1	1	2	2	2	1	1	2	2	2	2	1	YES
305	M	70	2	1	1	1	1	2	2	2	2	2	2	1	2	YES
306	M	58	2	1	1	1	1	1	2	2	2	2	1	1	2	YES
307	M	67	2	1	2	1	1	2	2	1	2	2	2	1	2	YES
308	M	62	1	1	1	2	1	2	2	2	2	1	1	2	1	YES

Gambar 5. *Tabel Kriteria Data*

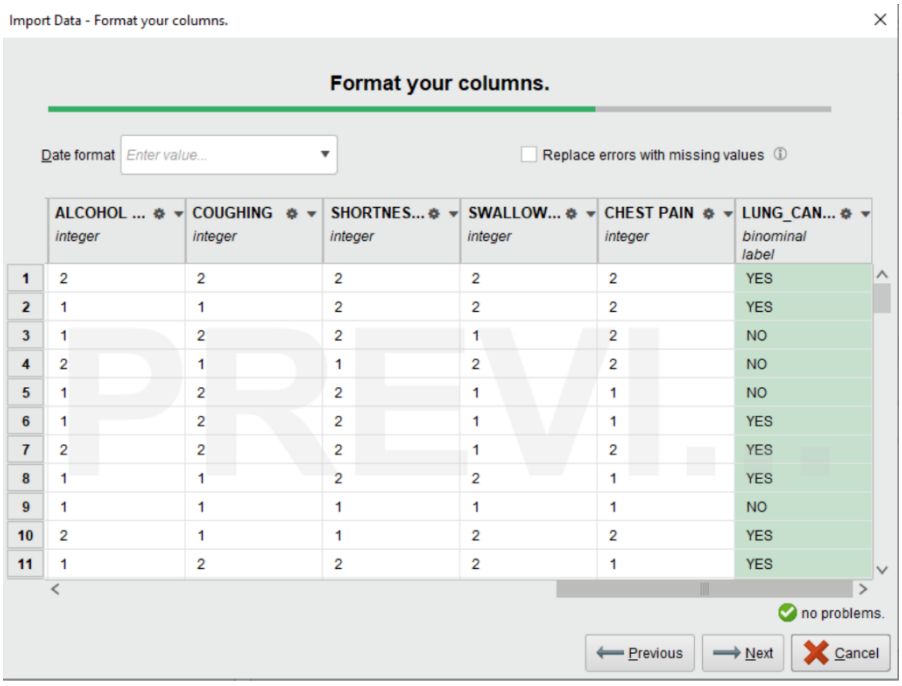
Hasil Pengujian Data

Data Uji merujuk pada dataset yang telah dipersiapkan sebelumnya untuk menjalani proses pengujian. Setelah melibatkan data uji, selanjutnya adalah dilakukanya pengkategorian menggunakan variabel dan atribut yang akan dijadikan sebagai data training, data training terdiri dari 309 record data, dengan 279 pasien dalam kategori *"YES"* atau positif kanker paru-paru dan 30 pasien dalam kategori *"NO"* atau tidak positif kanker paru-paru. Dari langkah-langkah tersebut, dilakukan perhitungan menggunakan Algoritma C4.5 guna memprediksi kemungkinan terjadinya penyakit kanker paru-paru pada pasien.

Dalam konteks penelitian ini, Setelah penjabaran kriteria data, langkah berikutnya adalah menentukan *"label"* berdasarkan "*coloumn"* yang akan digunakan untuk menjadi *leaf nodes* yang terbentuk dari proses *Decision Tree* pada visual RapidMiner. Setelah ini akan diulas langkah-langkah pembuatan *Decision Tree* menggunakan alat bantu RapidMiner.

Proses Transformasi Data Menggunakan Rapid Miner

Setelah melalui proses percobaan dan klasifikasi data, langkah berikutnya adalah melibatkan RapidMiner untuk pembuatan *Decision Tree* melalui data yang telah disiapkan. Berikut adalah bentuk tranformasi data pada coloumn *"LUNG_CANCER"* yang digunakan untuk pengujian klasifikasi data pasien kanker paru-paru ini adalah RapidMiner, seperti yang terlihat pada Gambar 6 di bawah:



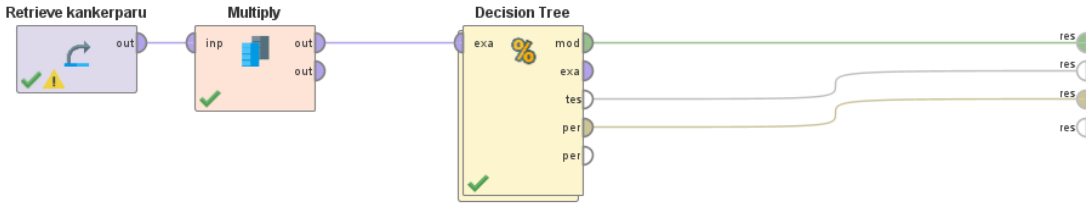
	ALCOHOL ... integer	COUGHING integer	SHORTNES... integer	SWALLOW... integer	CHEST PAIN integer	LUNG_CAN... binominal label
1	2	2	2	2	2	YES
2	1	1	2	2	2	YES
3	1	2	2	1	2	NO
4	2	1	1	2	2	NO
5	1	2	2	1	1	NO
6	1	2	2	1	1	YES
7	2	2	2	1	2	YES
8	1	1	2	2	1	YES
9	1	1	1	1	1	NO
10	2	1	1	2	2	YES
11	1	2	2	2	1	YES

***Gambar 6**.Tranformasi Data LUNG_CANCER Menjadi Binominal Label*

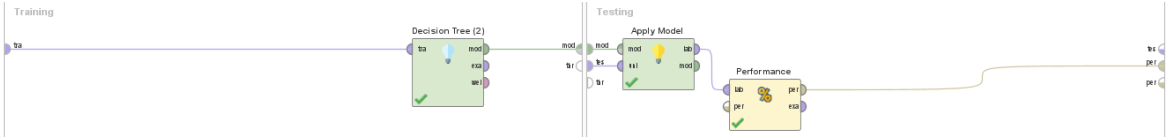
Langkah awal dalam proses ini adalah import data dari database ke dalam *Local Repository* RapidMiner, selanjutnya melakukan pengecekan pada data, jika sudah selanjutnya adalah menentukan coloumn yang akan di rubah menjadi label pada RapidMiner.

Penghubungan Port Operator Algoritma C4.5 Menggunakan RapidMiner

Setelah melalui proses transformasi data selanjutnya adalah penghubungan port pada setiap operator yang digunakan, operator yang digunakan disini telah menggunakan algoritma C4.5, operator yang digunakan antara lain : *Retrieve Data, Multiply, Decision Tree, Apply Model, dan Performance.* Perhubungan port algoritma C4.5 pada RapidMiner dapat dilihat pada gambar 7 dan 8 berikut:



***Gambar 7**.Penghubungan port Retrieve kankerparu, Multiply, dan Decision Tree*



***Gambar 8**.Penghubungan port Decision Tree ke mod apply model, dan performace*

Perhitungan Tingkat Akurasi Pada Decision Tree (C4.5)

Setelah melalui proses penghubungan port pada aplikasi RapidMiner, selanjutnya adalah pembuatan tampilan tingkat akurasi pada *Decision Tree (C4.5)* yang dapat dilihat pada gambar 9 berikut:

accuracy: 90.29% +/- 3.40% (micro average: 90.29%)

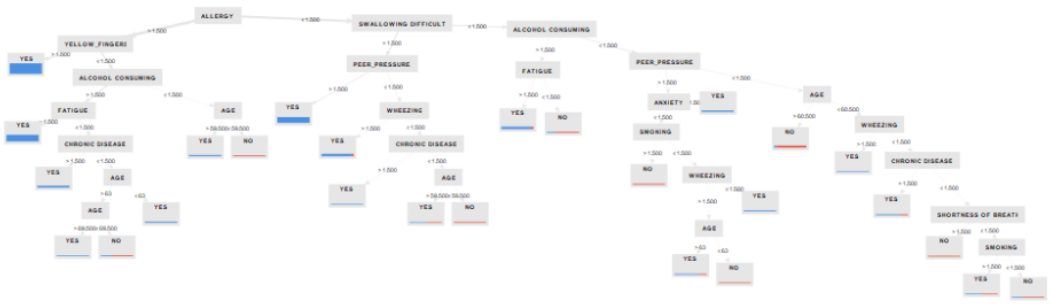
	true YES	true NO	class precision
pred. YES	257	17	93.80%
pred. NO	13	22	62.86%
class recall	95.19%	56.41%	

***Gambar 9**. Tampilan Akurasi Decision Tree (C4.5)*

Dengan pengolahan data menggunakan aplikasi RapidMiner didapat nilai akurasi sistem sebesar 93.80%. Dari gambar di jelaskan bahwa prediksi YES adalah 257 dan prediksi NO adalah adalah 22 dengan nilai precision sebesar 93.80% dan 62.86%, dan nilai recall sebesar 95.19% dan 56,41%.

Hasil Graph Visualiasasi Data Dengan Operator Decision Tree (C4.5)

Setelah melalui proses pembuatan tampilan tingkat akurasi pada *Decision Tree (C4.5)* selanjutnya adalah menganalisis dan menampilkan hasil visualisasi data menggunakan operator *Decision Tree (C4.5)*. Algoritma C4.5, yang digunakan dalam *Decision Tree*, memungkinkan penguraian data yang lebih baik untuk memahami faktor-faktor yang signifikan dalam memprediksi persentase terjadinya kanker pada suatu pasien menggunakan data pasien kanker paru-paru yang sudah ada. Hasil visualisasi data dapat dilihat pada gambar 10 berikut:



***Gambar 10**. Visualiasasi Data Dengan Operator Decision Tree (C4.5)*

Dengan melakukan visualisasi dapat dilihat bahwa persentase untuk pasien yang memiliki umur diatas 65 tahun serta pencandu alkohol dan rokok memiliki peluang 75% terjadinya kanker paru-paru, dan untuk pasien yang dibawah 60 tahun apabila memiliki penyakit kronis, sesak nafas, maka memiliki peluang 50 % terkena kanker paru-paru untuk perokok aktif, dan 33% untuk yang bukan perokok aktif.

KESIMPULAN

Dari hasil penelitian yang dilakukan pada proses pengujian dengan 309 rekaman data testing, ditemukan bahwa penggunaan algoritma C4.5 menghasilkan tingkat akurasi sebesar 93.80%. Tingkat akurasi tersebut mencerminkan kemampuan algoritma dalam menganalisis dengan baik, terutama dalam konteks mempercepat pengambilan keputusan terkait kemungkinan terjadinya kanker pada pasien, seperti melakukan test X-Ray, CT Scan Thorax, PET-CT Scan, DLL.

Penerapan data mining melalui metode Algoritma C4.5 menunjukkan bahwa algoritma ini efektif dalam mengevaluasi gejala yang telah di-visualkan pada pasien. Hal ini memungkinkan dokter dengan cepat memperkirakan apakah pasien perlu menjalani tes kanker paru-paru atau tidak. Keberhasilan algoritma C4.5 dalam memberikan tingkat akurasi yang tinggi dapat menjadi landasan penting untuk mendukung keputusan klinis yang lebih cepat dan tepat.

Sebagai langkah pengembangan lebih lanjut, perlu dilakukan penelitian tambahan dengan mempertimbangkan penambahan data pasien dan variasi gejala. Hal ini dapat meningkatkan ketepatan algoritma dalam menganalisis dan memprediksi kemungkinan terjadinya kanker paru-paru. Selain itu, mempertimbangkan integrasi dengan metode lain atau teknologi di bidang kedokteran juga dapat menjadi arah penelitian yang menarik untuk meningkatkan kehandalan sistem ini dalam mendukung praktik klinis sehari-hari.

REFERENCES

Böhlen, M., Gamper, J., & Polasek, W. (Eds.). (2013). Data Analysis, Machine Learning, and Applications. Springer.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (Eds.). (1996). Advances in Knowledge Discovery and Data Mining. AAAI Press.

Han, J. & M. Kamber. (2006). Data Mining Concept and Techniques, Morgan Kaufmann Publishers, San Francisco.

Larose, D. T. (2014). Discovering Knowledge in Data: An Introduction to Data Mining (2nd ed.). Wiley.

Lee, J. H., et al. (2021). Age-specific prognosis of non-small cell lung cancer in patients with cough. Lung Cancer, 208, 1-9.

Miller, C. A., et al. (2022). Predicting lung cancer in patients with cough and risk factors. Cancer Epidemiology, Biomarkers & Prevention, 31(1), 1-10

Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann.

Travis, W. D., Brambilla, E., Noguchi, M., Nicholson, A. G., Geisinger, K., Yatabe, Y., & Flieder, D. B. (2011). International association for the study of lung cancer/American thoracic society/European respiratory society international multidisciplinary classification of lung adenocarcinoma. Journal of Thoracic Oncology, 6(2), 244-285.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.