Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK)
Volume 9, Nomor 1, Februari 2024, pp 94-106
Terakreditasi Nomor 204/E/KPT/2022 | ISSN: 2527-5771/EISSN: 2549-7839
https://tunasbangsa.ac.id/ejurnal/index.php/jurasik

Implementasi *Decision tree* Untuk Prediksi Kanker Paru-Paru

Faurika[1], Ahsanun Naseh Khudori[2*], M Syauqi Haris[3]
[1,2,3]Informatika, Institut Teknologi, Sains, dan Kesehatan RS.DR. Soepraoen Kesdam V/BRW, Indonesia
Email: ahsanunnaseh@itsk-soepraoen.ac.id

Abstract

Lung cancer is a disorder of the lungs due to changes in respiratory tract epithelial cells which cause uncontrolled cell division and growth. Lung cancer is caused by several factors such as radiation exposure, smoking, heredity, gender, air pollution, and unhealthy lifestyles. Lung cancer can be detected when the cancer has entered an advanced stage. The large amount of lung cancer diagnosis data currently available can be used to predict lung cancer based on patterns in the data. One of the results of technological advances that can learn patterns in data is machine learning, which has currently made many positive contributions in the health sector. This research aims to predict lung cancer using a decision tree algorithm. This research produces rules based on decision trees which are built and then evaluated to produce the same accuracy, precision, recall, and F1-Score of 100%.

Keywords*: lung cancer, machine learning, decision tree*

Abstrak

Kanker paru-paru adalah gangguan pada paru-paru karena perubahan sel epitel saluran pernapasan yang menyebabkan pembelahan dan pertumbuhan sel tak terkendali. Kanker paru-paru disebabkan oleh beberapa faktor seperti paparan radiasi, perokok, keturunan, jenis kelamin, polusi udara dan pola hidup tidak sehat. Kanker paru-paru dapat dideteksi ketika kanker telah memasuki tahap stadium lanjut. Banyaknya data diagnosis kanker paru-paru saat ini, dapat digunakan untuk memprediksi kanker paru-paru berdasarkan pola pada data. Salah satu hasil kemajuan teknologi yang mampu mempelajari pola pada data yakni machine learning yang saat ini telah banyak memberikan kontribusi positif dibidang kesehatan. Penelitian ini bertujuan untuk prediksi kanker paru-paru menggunakan algoritma decision tree. Penelitian ini menghasilkan aturan (rules) berdasarkan pohon keputusan decision tree yang dibangun kemudian dievaluasi sehingga menghasilkan akurasi, presisi, recall, dan F1-Score yang sama yakni sebesar 100%.

Kata kunci*: lung cancer, machine learning, decision tree*

1. PENDAHULUAN

Kanker paru-paru menjadi salah satu penyakit mematikan di seluruh dunia dengan jumlah kasus dan kematian yang terus meningkat setiap tahunnya [1]. Menurut WHO (2022) pada tahun 2020, terdapat 2.2 juta penderita kanker paru-paru di seluruh dunia dan 1.8 juta mengalami kematian. Di Amerika diperkirakan terdapat 236.740 kasus kanker paru-paru dan 130.180 kasus mengalami kematian [3]. Sementara itu, di Indonesia kasus kanker paru-paru mencapai 34.78 di tahun 2020 [4]. Terlambatnya deteksi dini kanker paru-paru menyebabkan tingginya angka kematian,keterlambatan deteksi ini karena biaya yang mahal, sehingga penderita kanker paru-paru enggan melakukan pemeriksaan ke dokter pada tahap stadium awal, kebanyakan dari mereka pergi ke dokter setelah memasuki stadium lanjut saat terjadi komplikasi [5]–[7].

Deteksi kanker paru-paru biasanya dilakukan dengan cara melakukan tes darah untuk menilai fungsi organ lain ketika sel kanker telah menyebar dan meminimalisir infeksi. Pemeriksaan dahak juga dilakukan untuk deteksi sel

kanker dari dahak, pemeriksaan spirometri untuk menilai fungsi paru dengan mengukur jumlah udara yang keluar masuk saat bernapas, thoracentesis juga dilakukan dengan mengambil cairan yang menumpuk di rongga dada untuk sampel uji laboratorium, pemeriksaan radiologi dengan X-Ray untuk menunjukkan adanya massa berwarna putih keabuan, tetapi itu juga dapat disebabkan oleh abses paru [2], [8]. Banyaknya kemungkinan tersebut membuat diagnosa medis lebih lama dan membutuhkan biaya yang mahal sehingga masyarakat enggan untuk berobat ke dokter [6].

Seiringnya berkembangnya teknologi, deteksi dini kanker paru-paru bisa dibantu oleh teknologi. Salah satunya menggunakan teknik komputasi *Machine learning* (ML). ML adalah bagian dari kecerdasan buatan yang dapat diterapkan untuk melakukan prediksi, klasifikasi, dan mengenal pola [9]. Beberapa algoritma ML yang banyak digunakan untuk melakukan prediksi adalah support vector machine (SVM), *decision tree* (DT), naive bayes, *k-nearest neighbor* (KNN), *random forest*, *artificial neural network* dan sebagainya [10]. Pada penelitian ini, Peneliti memilih menggunakan algoritma *decision tree* dikarenakan algoritma ini mudah untuk divisualisasikan serta memiliki kinerja yang bagus untuk melakukan prediksi [11], [12].

Beberapa penelitian yang telah mengimplementasikan algoritma *decision tree* untuk prediksi kanker paru-paru diantaranya dilakukan oleh Saeed (2019) yang menggunakan dataset dari salah satu rumah sakit di Karachi, Pakistan yang memuat informasi medis pasien kanker paru-paru yakni usia, jenis kelamin, kadar kolesterol, berat badan, golongan darah, dan riwayat rokok. Penerapan algoritma *decision tree* pada penelitian ini menghasilkan akurasi sebesar 60%. Penelitian lainnya dilakukan oleh Kumar Mohan dan Bhraguram Thayyil (2023) yang menggunakan database kesehatan yang berisi informasi kesehatan pasien kanker paru-paru yang meliputi jenis kelamin, usia, riwayat rokok, riwayat konsumsi alkohol, tekanan darah, dan indeks massa tubuh. Pada penelitian ini, algoritma *decision tree* menghasilkan akurasi sebesar 88%. Penelitian lainnya dilakukan Gupta (2023) yang menggunakan dataset dari cbioportal yang memuat informasi medis pasien kanker paru-paru NSCLSC yakni id pasien, usia, riwayat rokok, persentase pasien yang menerima manfaat perawatan ICI. Penelitian tersebut, algoritma *decision tree* menghasilkan akurasi sebesar 85.7%.

Berdasarkan penelitian-penelitian tersebut, dapat disimpulkan *decision tree* dapat diterapkan untuk prediksi kanker paru-paru dengan hasil akurasi cukup bagus, namun masih memiliki celah perbaikan. Berdasarkan penelitian tersebut, peneliti bermaksud memodifikasi penelitian tentang *decision tree* untuk prediksi kanker paru-paru diantaranya menggunakan dataset yang lebih beragam dan menerapkan *decision tree* ID3. Penelitian ini bertujuan untuk menganalisis kinerja algoritma *decision tree* untuk prediksi kanker paru-paru. Sehingga dapat membantu tenaga medis dalam mendiagnosa kanker paru-paru sejak dini yang dapat menekan angka kematian yang disebabkan kanker paru-paru.

2. TINJAUAN PUSTAKA

2.1. Kanker paru-paru

Kanker paru-paru adalah salah satu jenis kanker yang disebabkan oleh gangguan paru-paru karena adanya perubahan pada sel epitel saluran pernapasan yang menyebabkan pembelahan dan pertumbuhan sel yang tidak terkendali [16]. Kanker paru-paru dapat menyebabkan kerusakan sistem kekebalan tubuh, tumor dan gangguan lain sehingga tubuh tidak mampu melakukan fungsinya dengan baik [17]. Neoplasma pada paru-paru menjadi penyebab utama kanker dan kematian di seluruh dunia [18]. Berdasarkan asal selnya, kanker paru-paru dibedakan menjadi kanker paru-paru sel kecil atau *small-cell lung cancers* (SCLC) dan kanker paru-paru non sel kecil atau *non-small cell lung cancers* (NSCLC) [18]. NSCLC merupakan tipe kanker paru-paru yang paling dominan dengan tingkat keganasan tinggi [19].

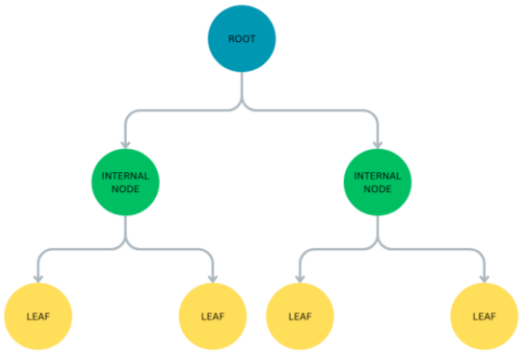
2.2. *Machine learning*

ML merupakan salah satu bagian kecerdasan buatan yang mampu membuat model dari pengalaman dan mampu membuat prediksi di masa depan dengan akurat [20]–[22]. Dalam ML terdapat dua model pembelajaran yaitu *supervised* dan *unsupervised learning*, dimana *supervised learning* ini banyak diterapkan untuk klasifikasi sedangkan *unsupervised learning* banyak digunakan untuk permasalahan pengelompokan [23], [24]. Terdapat banyak algoritma dalam kedua model tersebut. Beberapa algoritma dalam ML yang termasuk pada kategori *supervised learning* diantaranya *k-nearest neighbor* (KNN), *neural network, linear regression*, *support vector machine* (SVM), *random forest*, *decision tree*, dan sebagainya. Dalam penelitian ini, algoritma yang akan digunakan yaitu *decision tree* ID3.

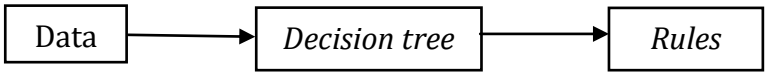
2.3. *Decision tree ID3*

DT merupakan salah satu algoritma supervised learning yang populer dan memiliki performa yang bagus untuk klasifikasi [6], [25]. Dalam klasifikasi, DT akan membentuk pohon keputusan yang terdiri dari beberapa *node* membentuk *root* dan *node* lain bukan *root* tetapi memiliki input disebut node internal dan node lain yang mewakili nilai salah satu kelas disebut daun atau *leaf* [9]. *Root* merepresentasikan pilihan yang akan menghasilkan *subtree* semua data kedalam dua atau lebih subset [26]. *Node* internal merepresentasikan kemungkinan pilihan pada titik tertentu dalam struktur *tree*, bagian atas node terhubung dengan *node* induk dan tepi bawah terhubung dengan simpul daun atau turunannya [27]. *Leaf* merepresentasikan hasil akhir dari kombinasi DT [28]. Input yang digunakan pada DT berdasar pada kriteria tertentu dan outputnya dipetakan dalam *true* dan *false* [29]. Nilai dalam node diperoleh dengan membandingkan atribut yang digunakan berdasarkan *information gain* yang outputnya ditampilkan pada *leaf* [29]. Struktur *decision tree* sebagaimana ditunjukkan pada Gambar 1 dan konsep dari *decision tree* ditunjukkan pada Gambar 2.

Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK)
Volume 9, Nomor 1, Februari 2024, pp 94-106
Terakreditasi Nomor 204/E/KPT/2022 | ISSN: 2527-5771/EISSN: 2549-7839
https://tunasbangsa.ac.id/ejurnal/index.php/jurasik



Gambar 1. Struktur *decision tree*



Gambar 2. Konsep *decision tree*

Hasil dari *decision tree* berupa pohon keputusan yang dibuat dengan alur perhitungan sebagai berikut.

a) Menghitung *entropi* dan *information gain*

Rumus yang digunakan untuk menghitung *entropi* sebagai berikut.

$$Entropy\ (S) = -p_+\ log_2\ p_+ - p_-\ log_2\ p_- \quad (1)$$

Dimana S merupakan Himpunan semesta, p_+ merupakan proporsi jumlah atribut terhadap semesta pertama, dan p_- merupakan proporsi jumlah atribut terhadap semesta kedua. Setelah menghitung *entropi*, selanjutnya menghitung *information gain* dengan menggunakan rumus berikut.

$$Gain\ (S,A) = Entropy\ (S) - \sum_{v\ \epsilon nilai\ (A)} \frac{|S_v|}{|S|}\ Entropy\ (S_v) \quad (2)$$

Dimana S merupakan himpunan semesta, A menyatakan atribut, $Entropy\ (S)$ merupakan *entropi* dari semesta, S_v merupakan proporsi atribut, S merupakan proporsi semesta, dan $Entropy\ (S_v)$ merupakan *entropi* atribut.

b) Membentuk *Root*

Root ditentukan berdasarkan *information gain* yang telah dihitung sebelumnya. Atribut yang memiliki *information gain* terbesar akan ditetapkan sebagai *root*.

c) Membuat *Subtree*

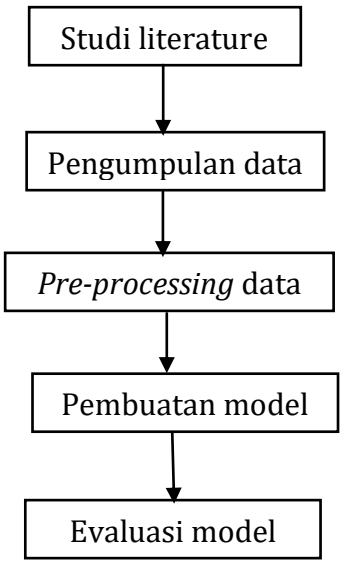
Subtree dibentuk berdasarkan percabangan pada *root*. Jika *entropi* pada cabang *root* bernilai 0 berarti tidak ada percabangan lagi, namun jika entropi-nya bukan 0 maka dilakukan perhitungan ulang information gain atribut berdasarkan cabang dari *root* tersebut. Langkah ini dilakukan terus menerus sampai diperoleh hasil akhir *node* tidak memiliki cabang lagi.

d) Membuat *leaf* atau daun

Leaf atau daun menjadi hasil akhir dari DT yang merepresentasikan output apakah data itu bernilai *true* atau *false*.

3. METODOLOGI PENELITIAN

Metode penelitian yang digunakan dalam penelitian ini sebagaimana ditunjukkan pada Gambar 3 berikut.



Gambar 3. Metode penelitian

Dari Gambar 3 dapat disimpulkan bahwa langkah-langkan pada penelitian ini terdiri dari:

1) Studi Literatur, pada langkah ini dilakukan pencarian penelitian terdahulu yang membahas implementasi algoritma *decision tree* untuk memprediksi kanker paru-paru.
2) Pengumpulan Data, data yang digunakan dalam penelitian ini adalah dataset dari Data *World*. Dataset ini berisi informasi tentang pasien kanker paru-paru dari berbagai negara.
3) *Pre-processing* data, Pada tahap ini data yang telah diperoleh dilakukan pengecekan dan pembersihan. Tujuannya untuk memastikan data yang akan digunakan telah sesuai dengan kriteria kualitas data yang ditentukan. Adapun data yang digunakan pada penelitian ini memiliki kriteria atribut datanya lengkap (tidak ada *missing values*), memiliki tipe data yang seragam dan sesuai, dan data tidak *redundant*.
4) Pembuatan Model, pada tahap ini dilakukan pembangunan pembelajaran mesin yang dapat digunakan untuk membuat prediksi berdasarkan data. Model algoritma *decision tree* dibangun dengan algoritma *decision tree* ID3

Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK)
Volume 9, Nomor 1, Februari 2024, pp 94-106
Terakreditasi Nomor 204/E/KPT/2022 | ISSN: 2527-5771/EISSN: 2549-7839
https://tunasbangsa.ac.id/ejurnal/index.php/jurasik

5) Evaluasi Model, pada tahap ini akan dilakukan evaluasi model menggunakan data uji. Pada penelitian ini model *decision tree* dievaluasi dengan beberapa teknik, yaitu:
   a. Akurasi, yakni rasio jumlah data prediksi yang benar terhadap jumlah total data, dihitung dengan rumus:

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

   b. Presisi, Presisi adalah rasio jumlah data prediksi yang benar terhadap jumlah data yang diprediksi positif, dihitung dengan rumus:

$$Presisi = \frac{TP}{TP + FP} \quad (4)$$

   c. *Recall,* yakni rasio jumlah data prediksi yang benar terhadap jumlah data yang sebenarnya positif, dihitung dengan rumus:

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

   d. *F1-Score* adalah kombinasi dari presisi dan *recall*, dihitung dengan rumus

$$F1 = 2 * \frac{Presisi * Recall}{Presisi + Recall} \quad (6)$$

Dimana TP merupakan *True Positive* yang berarti nilai prediksi dan nilai aktualnya bernilai positif, TN merupakan *True Negative* yang berarti nilai prediksi dan nilai aktualnya bernilai negatif, FP merupakan *False Positive* yang berarti nilai prediksi bernilai positif sedangkan nilai aktualnya bernilai negatif, FN merupakan *False Negative* yang berarti nilai prediksi bernilai false sedangkan nilai aktualnya bernilai positif.

4. HASIL DAN PEMBAHASAN

Dari hasil pengumpulan data dari data *world* (https://data.world/cancerdatahp/lung-cancer-data) diperoleh data sejumlah 1000. Data tersebut terdiri 25 atribut yang akan digunakan sebagai atribut input pada penelitian ini. Masing-masing atribut tersebut berisikan tingkat keparahan yang dijabarkan dalam skala 1 sampai 9. Penjelasan dari setiap tingkat keparahan tersebut sebagaimana ditunjukkan pada Tabel 1.

Tabel 1. Tingkat keparahan

No	Atribut	Nilai	Kategori
1	*age*	14-73 tahun	5-14 tahun = remaja 15-39 tahun = dewasa awal 40-59 tahun = dewasa akhir 60-69 tahun = lansia muda 70-79 tahun = lansia madya
2	*gender*	1, 2	1 = laki-laki 2 = perempuan
3	*air pollution*	1-8	1-3 = memiliki risiko kesehatan rendah. 4-6 = memiliki risiko kesehatan sedang. 7-10 = memiliki risiko kesehatan tinggi.
4	*alcohol use*	1-8	1-4 = memiliki risiko sedang dan tidak terlalu menyebabkan kecanduan berlebih.

No	Atribut	Nilai	Kategori
			5-8 = memiliki risiko tinggi terhadap kesehatan menyebabkan kecanduan berlebih.
5	*dust allergy*	1-8	1-3 = allergy ringan dan tidak mengganggu aktifitas sehari-hari. 4-6 = gejala alergi lebih terlihat dan mengganggu aktifitas sehari-hari. 7-10 = alergi parah yang dapat mengganggu aktifitas sehari-hari secara signifikan.
6	*occupational hazards*	1-8	1-3 = pekerjaan berisiko rendah seperti dalam lingkungan perkantoran 4-6 = pekerjaan berisiko sedang seperti dalam lingkungan industri 7-10 = pekerjaan berisiko tinggi seperti dalam lingkungan sektor konstruksi
7	*genetic risk*	1-7	1-3 = faktor risiko genetik rendah 4-6 = faktor risiko genetik sedang 7-10 = faktor risiko genetik tingi
8	*chronic lung disease*	1-7	1-3 = gejala penyakit ringan 4-6 = gejala mulai memburuk 7-10 = keparahannya terus meningkat
9	*balanced diet*	1-7	1-3 = diet yang sangat tidak seimbang 4-6 = fase diet yang cukup seimbang 7-10 = fase diet yang sangat seimbang
10	*obesity*	1-7	1-3 = kadar BMI berada dikisaran normal 4-6 = kadar BMI 30-34.9 7-10 = kadar BMI 35-39.9
11	*smoking*	1-8	1-3 = konsumsi rokok dalam skala rendah 4-6 = konsumsi rokok secara rutin 7-10 = konsumsi rokok terus menerus hingga menyebabkan kecanduan.
12	*passive smokers*	1-8	1-3 = jumlah paparan rokok sangat minim 4-6 = durasi paparan cukup lama dan sering 7-10 = paparan asap rokok terus menerus
13	*chest pain*	1-9	1-3 = nyeri yang dirasakan pada waktu tertentu 4-6 = nyeri yang mulai mengganggu aktifitas, namun masih dapat diatasi 7-10 = nyeri yang sering terjadi dan mungkin diikuti oleh gelaja lain.
14	*coughing of blood*	1-9	1-3 = jumlah darah yang keluar sedikit dan jarang terjadi 4-6 = jumlah kejadian lebih sering dan darah yang keluar lebih banyak 7-10 = jumlah kejadian yang meningkat signifikan dan darah yang keluar dalam jumlah banyak.
15	*fatigue*	1-9	1-3 = belum merasakan lelah berlebih 4-6 = mulai terjadi penurunan tenaga dan daya tahan fisik 7-10 = terjadi penurunan tenaga yang signifikan
16	*weight loss*	1-8	1-3 = penurunan berat badan normal 4-6 = penurunan berat badan sekitar 5-10% 7-10 = secara signifikan melebihi 10%
17	*shortness of breath*	1-9	1-3 = masih jarang terjadi 4-6 = lebih sering terjadi dan mengganggu aktifitas

No	Atribut	Nilai	Kategori
			7-10 = terjadi meningkat secara signifikan
18	*wheezing*	1-8	1-3 = masih jarang terjadi 4-6 = lebih sering terjadi 7-10 = terjadi meningkat signifikan dan mengganggu pernapasan
19	*swallowing difficulty*	1-8	1-3 = kesulitan menelan dalam waktu tertentu 4-6 = lebih membutuhkan perhatian seperti saat mengonsumsi makanan kering 7-10 = memerlukan upaya ekstra yang sangat mengganggu
20	*clubbing of finger nail*	1-9	1-3 = masih kemungkinan terjadi 4-6 = perubahan pada jari dan kuku yang memerlukan perhatian medis 7-10 = perubahan mungkin berpengaruh terhadap fungsi tangan normal
21	*frequent cold*	1-7	1-3 = paparan ringan terhadap dingin 4-6 = paparan suhu dingin dalam waktu lebih lama yang dapat menyebabkan kedinginan 7-10 = paparan suhu dingin yang parah yang dapat berdampak serius terhadap kesehatan.
22	*dry coughing*	1-7	1-3 = terjadi sesekali 4-6 = lebih sering terjadi 7-10 = terjadi terus menerus dan sangat mengganggu aktiftas
23	*snoring*	1-7	1-3 = tidak mengganggu kualitas tidur. 4-6 = terjadi lebih sering dan mempengaruhi kualitas tidur 7-10 = terjadi terus menerus dan sangat mengganggu
24	*level*	*low, medium, high*	*low* = riwayat pasien menderita kanker paru-paru berstatus rendah *medium* = riwayat pasien menderita kanker paru-paru berstatus sedang *high* = riwayat pasien menderita kanker paru-paru berstatus tinggi.

Pada tahap pre-processing data yang dikumpulkan telah memenuhi kriteria. Namun, terdapat 1 kolom *patient id* yang tidak akan digunakan dalam penelitian ini karena tidak berpengaruh terhadap kanker paru-paru. sehingga dilakukan pembersihan (data *cleansing*) untuk memastikan kualitas data yang digunakan, sehingga kolom yang tersisa sebanyak 24 dan jumlah data tetap 1000. Data tersebut dibagi menjadi 2 bagian, yakni sebanyak 800 (80%) data digunakan sebagai data latih dan sebanyak 200 data (20%) dijadikan data uji.

4.1. Pembuatan model *decision tree*

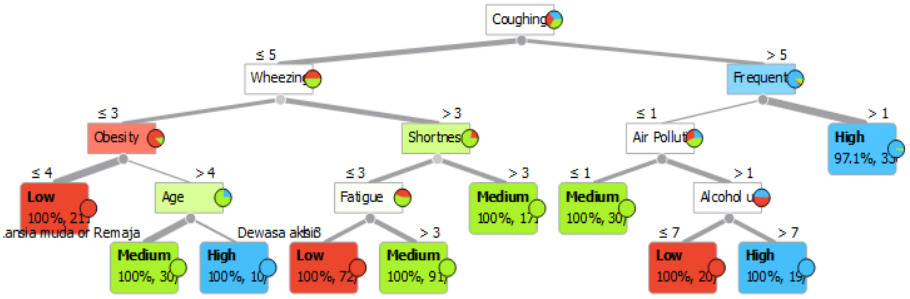
Penerapan *decision tree* pada penelitian ini menghasilkan informasi atau pola dalam bentuk *rules* berupa pohon keputusan *decision tree*. Terdapat beberapa parameter yang digunakan untuk membangun model *decision tree* diantaranya

1) *Induce binary tree*, parameter ini untuk membangun pohon biner dengan 2 node child.

Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK)
Volume 9, Nomor 1, Februari 2024, pp 94-106
Terakreditasi Nomor 204/E/KPT/2022 | ISSN: 2527-5771/EISSN: 2549-7839
https://tunasbangsa.ac.id/ejurnal/index.php/jurasik

2) *Min numbers of instances in leaves*, parameter ini untuk menentukan jumlah minim cabang di leaf atau daun. pada penelitian ini jumlah minimal cabang ditentukan sebanyak 2.
3) *Limit the maximal tree depth to*, parameter ini untuk menentukan jumlah kedalaman atau level pohon yang akan dibangun. pada penelitian ini jumlah kedalaman atau level pohon ditentukan sebanyak 100.
4) *stop when majority reaches* [%], parameter ini untuk menentukan proses pembuatan model berhenti ketika mencapai ambang batas yang ditentukan. pada penelitian ini ambang batas pembuatan model *decision tree* ditentukan pemodelan akan berhenti ketika mencapai ambang batas 100%.

Dari parameter-parameter tersebut, kemudian dibuat model *decision tree* hingga menghasilkan suatu pohon keputusan. Pohon keputusan yang dihasilkan sebagaimana ditunjukkan pada Gambar 4 berikut.



Gambar 4. Pohon keputusan

Berdasarkan pohon keputusan pada Gambar 3 diatas kemudian dibuat aturan (rules) *decision tree* yang ditunjukkan pada Tabel 2.

Tabel 2. Aturan (*rules*) *decision tree*

No	Aturan
1	*IF Coughing of blood* >5 ^ *Frequenty cold* > 1, *THEN High*
2	*IF Coughing of blood* >5 ^ *Frequenty cold* ≤ 1 ^ *Air pollution* > 1 ^ *Alcohol use* > 7, *THEN High*
3	*IF Coughing of blood* >5 ^ *Frequenty cold* ≤ 1 ^ *Air pollution* > 1 ^ *Alcohol use* 7, *THEN Low*
4	*IF Coughing of blood* >5 ^ *Frequenty cold* ≤ 1 ^ *Air pollution* ≤ 1, *THEN Medium*
5	*IF Coughing of blood* ≤ 5 ^ *Wheezing* > 3 ^ *Shortness* > 3, *THEN Medium*
6	*IF Coughing of blood* ≤ 5 ^ *Wheezing* > 3 ^ *Shortness* ≤ 3 ^ *Fatigue* > 3, *THEN Medium*
7	*IF Coughing of blood* ≤ 5 ^ *Wheezing* > 3 ^ *Shortness* 3 ^ *Fatigue* ≤ 3, *THEN Low*
8	*IF Coughing of blood* ≤ 5 ^ *Wheezing* ≤ 3 ^ *Obesity* > 4 ^ *Age* = "Dewasa akhir", *THEN High*
9	*IF Coughing of blood* ≤ 5 ^ *Wheezing* ≤ 3 ^ *Obesity* > 4 ^ *Age* = "Dewasa awal" *Age*= "Lansia madya" *Age*="Lansia Muda" *Age*="Remaja", *THEN Medium*
10	*IF Coughing of blood* ≤ 5 ^ *Wheezing* ≤ 3 ^ *Obesity* ≤ 4, *THEN Low*

Dari model *decision tree* tersebut makan didapatkan sebuah aturan sebagai berikut:

Aturan 1 : Jika seseorang mengalami batuk berdarah atau *coughing of blood* dengan tingkat keparahan diatas 5 dan mengalami demam atau *frequenty cold* dengan tingkat keparahan diatas 1, maka orang itu memiliki risiko menderita kanker paru-paru yang tinggi atau *high*.

Aturan 2 : Jika seseorang mengalami batuk berdarah atau *coughing of blood* dengan tingkat keparahan diatas 5 dan mengalami demam atau *frequency cold* dengan tingkat keparahan diatas 1 dan polusi udara dengan tingkat keparahan diatas 1, serta mengalami kecanduan terhadap alkohol dengan tingkat keparahan diatas 7, maka orang itu memiliki risiko menderita kanker paru-paru di tingkat yang tinggi atau *high*.

Aturan 3: Jika seseorang mengalami batuk berdarah atau *coughing of blood* dengan tingkat keparahan diatas 5 dan mengalami demam atau *frequency cold* dengan tingkat keparahan diatas 1 dan polusi udara dengan tingkat keparahan diatas 1, serta mengalami kecanduan terhadap alkohol dengan tingkat keparahan kurang dari atau sama dengan 7, maka orang itu memiliki risiko menderita kanker paru-paru di tingkat rendah atau *low*.

Aturan 4: Jika seseorang mengalami batuk berdarah atau *coughing of blood* dengan tingkat keparahan diatas 5 dan mengalami demam atau *frequency cold* dengan tingkat keparahan diatas 1 dan polusi udara dengan tingkat keparahan kurang dari atau sama dengan 1, maka orang itu memiliki risiko menderita kanker paru-paru di tingkat sedang atau *medium*.

Aturan 5: Jika seseorang mengalami batuk berdarah atau *coughing of blood* dengan tingkat keparahan kurang dari atau sama dengan 5 dan tingkat keparahan mengi atau *wheezing* diatas 3 dan mengalami sesak napas atau *shortness* dengan tingkat keparahan diatas 3, maka orang itu memiliki risiko menderita kanker paru-paru di tingkat sedang atau *medium*.

Aturan 6: Jika seseorang mengalami batuk berdarah atau *coughing of blood* dengan tingkat keparahan kurang dari atau sama dengan 5 dan tingkat keparahan atau *wheezing* dengan tingkat keparahan diatas 3 dan mengalami sesak napas atau *shortness* dengan tingkat keparahan diatas 3 dan mengalami kelelahan berlebih atau *fatigue* dengan tingkat keparahan diatas 3, maka orang itu memiliki risiko menderita kanker paru-paru di tingkat sedang atau *medium*.

Aturan 7: Jika seseorang mengalami batuk berdarah atau *coughing of blood* dengan tingkat keparahan kurang dari atau sama dengan 5 dan tingkat keparahan mengi atau *wheezing* diatas 3 dan mengalami sesak napas atau *shortness* dengan tingkat keparahan kurang dari atau sama dengan 3 dan mengalami kelelahan berlebih atau *fatigue* dengan tingkat keparahan kurang dari atau sama dengan 3, maka orang itu memiliki risiko menderita kanker paru-paru di tingkat rendah atau *low*.

Aturan 8: Jika seseorang mengalami batuk berdarah atau *coughing of blood* dengan tingkat keparahan kurang dari atau sama dengan 5 dan tingkat keparahan mengi atau *wheezing* dengan tingkat keparahan kurang dari atau sama dengan 3 dan mengalami obesitas atau *obesity* dengan tingkat keparahan diatas 4 dan tergolong kategori usia atau age dewasa akhir , maka orang itu memiliki risiko menderita kanker paru-paru tingkat tinggi atau *high*.

Aturan 9: Jika seseorang mengalami batuk berdarah atau *coughing of blood* dengan tingkat keparahan kurang dari atau sama dengan 5 dan tingkat keparahan mengi atau *wheezing* kurang dari atau sama dengan 3 dan mengalami obesitas atau *obesity* dengan tingkat keparahan diatas 4 dan tergolong kategori usia atau *age* dewasa awal atau lansia madya atau lansia muda atau remaja, maka orang itu memiliki risiko menderita kanker paru-paru di tingkat sedang atau *medium*.

Aturan 10: Jika seseorang mengalami batuk berdarah atau *coughing of blood* dengan tingkat keparahan kurang dari atau sama dengan 5 dan tingkat keparahan mengi atau *wheezing* dengan tingkat keparahan kurang dari atau sama dengan 3 dan mengalami obesitas atau *obesity* dengan tingkat keparahan kurang dari atau sama dengan 4 dan tergolong kategori usia atau *age* dewasa awal atau lansia madya atau lansia muda atau remaja, maka orang itu memiliki risiko menderita kanker paru-paru di tingkat rendah atau *low*.

Berdasar dari aturan tersebut, seseorang dapat diprediksi apakah menderita kanker paru-paru atau tidak. Namun, untuk memastikan kebenaran dan keakuratan dari hasil yang diperoleh, maka diperlukan evaluasi untuk memeriksa keakuratan dan kebenaran yang ditemukan telah sesuai atau belum.

4.2. Evaluasi model

Proses evaluasi dalam penelitian ini menggunakan *tools orange* dengan menggunakan teknik evaluasi untuk mengukur akurasi, presisi, *recall*, dan *F1-Score*. Evaluasi ini menggunakan teknik *random sampling* yang membagi data secara acak untuk evaluasi model *decision tree*. Proses evaluasi dilakukan dengan 100 iterasi dan training set sebesar 80%.

Dari hasil evaluasi didapatkan nilai akurasi *level low, medium*, dan *high* sebesar 100%. Hal ini berarti bahwa model *decision tree* yang dibangun jika digunakan untuk memprediksi data baru, maka kemungkinan besar prediksinya akan benar sebesar 100%. Angka tersebut menunjukkan bahwa model *decision tree* memiliki akurasi yang sangat tinggi. Namun perlu diingat bahwa akurasi model *decision tree* dapat dipengaruhi oleh beberapa faktor yakni pertama, kualitas data, semakin berkualitas data baru yang digunakan, maka kemungkinan akan semakin tinggi tingkat akurasinya. kedua, jumlah fitur yang digunakan,semakin banyak fitur yang digunakan, akan semakin tinggi akurasi model *decision tree*.

Sedangkan hasil evaluasi presisi untuk level *low, medium*, dan *high* sebesar 100% dan recall sebesar 100%. Hal ini menunjukkan bahwa model *decision tree* yang dibangun jika digunakan untuk memprediksi data baru, maka kemungkinan besar ketepatan prediksi sebesar 100%. Hasil tersebut menunjukkan bahwa model *decision tree* memiliki presisi dan *recall* yang sangat tinggi.

Evaluasi yang terakhir menggunakan teknik *F1-Score*, dari hasil evaluasi *F1-Score* juga didapatkan nilai *level low, medium*, dan *high* sebesar 100%. Hal ini menunjukkan bahwa model *decision tree* yang dibangun memiliki kinerja yang sangat bagus dengan hasil *F1-Score* yang tinggi menunjukkan bahwa presisi dan recall yang dihasilkan sangat bagus.

4. SIMPULAN

Implementasi algoritma *decision tree* untuk prediksi kanker paru-paru menghasilkan rules dan model untuk prediksi kanker paru-paru. Setelah diperoleh pohon keputusan dan *rules* prediksi kanker paru-paru, kemudian dilakukan evaluasi menggunakan Orange dengan teknik akurasi, presisi, *recal*l, dan *F1-Score*. Hasil evaluasi kinerja *decision tree* dengan rasio pembagian data sebanyak 800 (80%) sebagai data latih dan sisanya sebanyak 200 (20%) data sebagai data uji diperoleh hasil yang sama untuk akurasi, presisi, *recall*, dan *F1-Score* yakni sebesar 100%. berdasarkan hasil yang diperoleh, dapat disimpulkan bahwa algoritma *decision tree* memiliki performa yang sangat bagus untuk memprediksi kanker paru-paru.

DAFTAR PUSTAKA

[1] M. Abdul, R. Wahid, A. Nugroho, and A. H. Anshor, "Prediksi Penyakit Kanker Paru-Paru Dengan Algoritma Regresi Linier," *Bull. Inf. Technol.*, vol. 4, no. 1, pp. 63–74, 2023.

[2] WHO, "Lung cancer," 2022, [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/lung-cancer

[3] American Lung Association, "State of Lung Cancer: Texas," 2021, [Online]. Available: https://www.lung.org/research/state-of-lung-cancer/states/texas

[4] Globocan, "Cancer in Indonesia," *JAMA J. Am. Med. Assoc.*, vol. 247, no. 22, pp. 3087–3088, 2020, doi: 10.1001/jama.247.22.3087.

[5] I. Buana and D. A. Harahap, "Asbestos, Radon Dan Polusi Udara Sebagai Faktor Resiko Kanker Paru Pada Perempuan Bukan Perokok," *AVERROUS J. Kedokt. dan Kesehat. Malikussaleh*, vol. 8, no. 1, p. 1, 2022, doi: 10.29103/averrous.v8i1.7088.

[6] D. Septhya *et al.*, "Implementasi Algoritma Decision Tree dan Support Vector Machine untuk Klasifikasi Penyakit Kanker Paru," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 3, no. 1, pp. 15–19, 2023.

[7] J. Sinurat, "Jaringan Saraf Tiruan Diagnosa Penyakit Kanker Paru-Paru Menggunakan Metode Hebb Rule," *Bull. Inf. Technol.*, vol. 2, no. 1, pp. 20–27, 2021.

[8] Rokom, "One Stop Service, Deteksi Dini Kanker Paru di RSUP Persahabatan," 2023, [Online]. Available: https://sehatnegeriku.kemkes.go.id/baca/rilis-media/20230405/3242727/one-stop-service-deteksi-dini-kanker-paru-di-rsup-persahabatan/

[9] N. R. Muntiari and K. H. Hanif, "Klasifikasi Penyakit Kanker Payudara Menggunakan Perbandingan Algoritma Machine Learning," *J. Ilmu Komput. dan Teknol.*, vol. 3, no. 1, pp. 1–6, 2022, doi: 10.35960/ikomti.v3i1.766.

[10] R. A. Sowah, A. A. Bampoe-Addo, S. K. Armoo, F. K. Saalia, F. Gatsi, and B. Sarkodie-Mensah, "Design and Development of Diabetes Management System Using Machine Learning," *Int. J. Telemed. Appl.*, vol. 2020, 2020, doi: 10.1155/2020/8870141.

[11] S. Ghosh, A. Dasgupta, and A. Swetapadma, "A study on support vector machine based linear and non-linear pattern classification," *Proc. Int. Conf. Intell. Sustain. Syst. ICISS 2019*, no. Iciss, pp. 24–28, 2019, doi: 10.1109/ISS1.2019.8908018.

[12] A. Septhiani, "Analisis Perbandingan Algoritma Supervised Learning untuk Prediksi Kasus Covid-19 di Jakarta," vol. 7, no. September, pp. 583–594, 2023.

[13] S. Saeed, A. Abdullah, N. Jhanjhi, and T. Malaysia, "Analysis of the Lung Cancer patient's for Data Mining Tool," *IJCSNS Int. J. Comput. Sci. Netw. Secur.*, vol. 19, no. 7, p. 90, 2019.

[14] Kumar Mohan and Bhraguram Thayyil, "Machine Learning Techniques for Lung Cancer Risk Prediction using Text Dataset," *Int. J. Data Informatics Intell. Comput.*, vol. 2, no. 3, pp. 47–56, 2023, doi: 10.59461/ijdiic.v2i3.73.

[15] T. Gupta, T. Qawasmeh, and S. McCalla, "Predictions of Programmed Cell Death Ligand 1 Blockade Therapy Success in Patients with Non-Small-Cell Lung Cancer," *BioMedInformatics*, vol. 3, no. 4, pp. 1060–1070, 2023, doi: 10.3390/biomedinformatics3040063.

[16] Juwita, N. Amalita, and M. D. Parma, "Faktor-Faktor Risiko yang Mempengaruhi Kanker Paru-Paru dengan Menggunakan Analisis Regresi Logistik," *UNPjoMath*, vol. 4, no. 1, pp. 38–42, 2021, [Online]. Available: https://ejournal.unp.ac.id/students/index.php/mat/article/download/11550/4620

[17] S. S. A.-N. Ibrahim M. Nasser, "Lung Cancer Detection Using Artificial Neural Network," vol. 3, no. 3, pp. 17–23, 2019.

[18] L. Wheless, J. Brashears, and A. J. Alberg, "Epidemiology of lung cancer," *Lung Cancer Imaging*, pp. 1–15, 2021, doi: 10.1007/978-1-60761-620-7_1.

[19] Z. Bing, Z. Zheng, and J. Zhang, "Risk factors influencing chemotherapy compliance and survival of elderly patients with non-small cell lung cancer," *Afr. Health Sci.*, vol. 23, no. 3, pp. 291–300, 2023, doi: 10.4314/ahs.v23i3.35.

[20] Q. Aini, N. Lutfiani, H. Kusumah, and M. S. Zahran, "Deteksi dan Pengenalan Objek Dengan Model Machine Learning: Model Yolo," *CESS (Journal Comput. Eng. Syst. Sci.*, vol. 6, no. 2, p. 192, 2021, doi: 10.24114/cess.v6i2.25840.

[21] M. Idris, R. I. Adam, Y. Brianorman, R. Munir, and D. Mahayana, "Kebenaran dalam Perspektif Filsafat Ilmu Pengetahuan dan Implementasi dalam Data Science dan Machine Leaning," *J. Filsafat Indones.*, vol. 5, no. 2, pp. 173–181, 2022, doi: 10.23887/jfi.v5i2.42207.

[22] N. Wiranda, H. S. Purba, and R. A. Sukmawati, "Survei Penggunaan Tensorflow pada Machine Learning untuk Identifikasi Ikan Kawasan Lahan Basah," *IJEIS (Indonesian J. Electron. Instrum. Syst.*, vol. 10, no. 2, p. 179, 2020, doi: 10.22146/ijeis.58315.

[23] R. R. Pratama, "Analisis Model Machine Learning Terhadap Pengenalan Aktifitas Manusia," vol. 19, no. 2, pp. 302–311, 2020.

[24] A. Saputra, U. Nahdlatul, U. Sidoarjo, and K. Sidoarjo, "KLASIFIKASI PENGENALAN BUAH MENGGUNAKAN ALGORITMA NAIVE," vol. 2, no. 2, pp. 83–88, 2019.

[25] M. A. Rosid, A. S. Fitrani, Y. Findawati, S. Winata, and V. A. Firmansyah, "Classification of Dengue Hemorrhagic Disease Using Decision Tree with Id3 Algorithm," *J. Phys. Conf. Ser.*, vol. 1381, no. 1, 2019, doi: 10.1088/1742-6596/1381/1/012039.

[26] S. B. Begenova and T. V. Avdeenko, "Building of fuzzy decision trees using ID3 algorithm," *J. Phys. Conf. Ser.*, vol. 1015, no. 2, 2018, doi: 10.1088/1742-6596/1015/2/022002.

[27] A. Rajeshkanna and K. Arunesh, "ID3 Decision Tree Classification: An Algorithmic Perspective based on Error rate," *Proc. Int. Conf. Electron. Sustain. Commun. Syst. ICESC 2020*, no. Icesc, pp. 787–790, 2020, doi: 10.1109/ICESC48915.2020.9155578.

[28] E. E. Ogheneovo and P. A. Nlerum, "Iterative Dichotomizer 3 (ID3) Decision Tree: A Machine Learning Algorithm for Data Classification and Predictive Analysis," *Int. J. Adv. Eng. Res. Sci.*, vol. 7, no. 4, pp. 514–521, 2020, doi: 10.22161/ijaers.74.60.

[29] P. Sathiyanarayanan, S. Pavithra, M. Sai Saranya, and M. Makeswari, "Identification of breast cancer using the decision tree algorithm," *2019 IEEE Int. Conf. Syst. Comput. Autom. Networking, ICSCAN 2019*, pp. 1–6, 2019, doi: 10.1109/ICSCAN.2019.8878757.