Penerapan Metode Klasifikasi Decision Tree dalam Prediksi Kanker Payudara Menggunakan Algoritma C4.5

Nurjoko[1], Hendra Kurniawan*[2], Cici Cahyati[3], Elvira Uthia Rustanti[4], Hiya Cahya Mujahidah[5], Amanda Putri Maharani[6], Rosita[7], Agus Rahardi[8]
[1,2,3,4,5,6,7]Institut Informatika dan Bisnis Darmajaya; Jl. Z.A Pagar Alam No. 93 Bandar Lampung 35142 Indonesia, Telp. (0721) 787214/Fax. (0721) 700261

e-mail: [1]nurjoko@darmajaya.ac.id, *[2]hendra.kurniawan@darmajaya.ac.id, [3]agusrahardi@darmajaya.ac.id

Abstrak

Kanker payudara merupakan penyakit mematikan yang memerlukan deteksi dini dan prediksi akurat untuk meningkatkan peluang kesembuhan. Penelitian ini bertujuan memprediksi kanker payudara menggunakan teknik Data Mining dengan algoritma Decision Tree C4.5. Dataset ini mencakup atribut seperti ukuran tumor, status estrogen, status progesteron, Status Progesteron, Bulan Kelangsungan Hidup, dan status. Atribut-atribut ini dipilih berdasarkan relevansi klinis dan potensi prediktifnya dalam konteks kanker payudara. Hasil klasifikasi menunjukkan tingkat akurasi tinggi dengan riwayat prediksi jiwa penderita kanker payudara yang masih hidup sebanyak 658 jiwa dan precision class sebesar 91,90%. Penelitian ini memiliki tingkat keakuratan sebesar 89,81%. Temuan ini berpotensi dikembangkan menjadi sistem pendukung keputusan medis untuk membantuk diagnosis penyakit kanker payudata yang lebih objektif dan efisien.

Kata kunci*—Kanker Payudara, Data Mining, Decision Tree, C4.5, Prediksi, Akurasi*

Abstract

Breast cancer is a deadly disease that require early detection and accurate prediction to improve recovery chance. This research aims to predict breast cancer using Data Mining technique with Decision Tree C4.5 algorithm. The dataset includes attributes such as tumor size, estrogen status, progesterone status, Progesterone Status, Survival Month, and status. These attributes were selected based on their clinical relevance and predictive potential in the context of breast cancer. The classification results showed a high level of accuracy with a prediction history of 658 surviving breast cancer patients and a precision class of 91.90%. This study has an accuracy rate of 89,81%. These findings have the potential to be developed int a medical decision support system to assist in more objective and efficient.

Keywords—*Breast Cancer, Data Mining, Decision Tree, C4.5, Prediction, Accuracy*

1. PENDAHULUAN

Kanker merupakan penyakit penyebab kematian kedua di dunia dengan kisaran angka kematian mencapai 9.6 juta kasus pada tahun 2018 [1]. Kanker payudara merupakan salah satu jenis kanker yang paling sering terjadi pada wanita di seluruh dunia. Penyebab umum kanker payudara biasanya adalah kombinasi faktor genetik dan lingkungan, termasuk mutasi gen, paparan hormonal yang berkepanjangan, gaya hidup tidak sehat, dan faktor risiko lainnya yang

dapat menyebabkan pertumbuhan sel abnormal di jaringan payudara. Kanker payudara dapat menyerang wanita di berbagai kelompok umur, meskipun risikonya meningkat seiring bertambahnya usia, dengan mayoritas kasus didiagnosis pada wanita di atas 50 tahun. Penting untuk dicatat bahwa meskipun pengobatan kanker payudara telah mengalami kemajuan signifikan, penyakit ini seringkali masih sulit disembuhkan terutama jika terdeteksi pada stadium lanjut, menekankan pentingnya deteksi dini dan perawatan yang tepat. Kanker payudara merupakan salah satu penyakit tidak menular yang paling banyak menyerang perempuan, lebih dari 400.000 wanita meninggal akibat kanker payudara. Penderita kanker payudara akan merasa malu karena perubahan fisik yang terjadi pada dirinya, dan akan berpengaruh pada kualitas hidupnya [2]. Angka kejadian dan kematian pada penderita kanker payudara terus mengalami peningkatan di setiap tahunnya. Prevalensi kejadian 8-9% wanita mengalami kanker payudara. Menurut data *World Health Organization* (WHO) tahun 2018 menyebutkan angka kejadian kanker payudara di Indonesia sebesar 42,1 per 100.000 penduduk dengan rata-rata angka kematian 17 per 100.000 penduduk [3].

Kanker merupakan penyakit tidak menular dimana terjadi pertumbuhan dan perkembangan yang sangat cepat, tanpa terkendali dari sel maupun jaringan.Pertumbuhan ini dapat menggangu proses metabolisme tubuh dan menyebar antarsel dan jaringan tubuh (Hero, 2021; Susmini & Supriayadi, 2020) [4]. Kanker payudara (*Carcinoma mammae*) merupakan suatu neoplasma ganas yang timbul dari parenkim. Penyakit ini paling umum diderita oleh Perempuan. Penyakit ini terdaftar dalam *International Classification of Diseases* (ICD) Organisasi Kesehatan Dunia (WHO) dengan nomor kode 17 [5]. Di Indonesia, angka kejadian penyakit tidak menular, termasuk kanker, semakin meningkat (RISKESDAS) tahun 2018. Berdasarkan hasil Riset Kesehatan Dasar (RISKESDAS) 2018, di mana kanker payudara menempati peringkat ketujuh dalam daftar penyakit kanker di semua penyakit kanker, dan menduduki peringkat pertama baik di rawat inap maupun rawat jalan di seluruh RS Indonesia (28,7%). danProvinsi Daerah Istimewa Yogyakarta (DIY) mempunyai angka kejadian kanker payudara tertinggi yaitu sebesar 2,4%, disusul Provinsi Kalimantan Timur sebesar 1,0% dan Provinsi Sumatera Barat sebesar 0,9% (Kemenkes RI. 2020) [6].

Ada berbagai cara untuk mendiagnosis kanker payudara. Sampai saat ini, pemeriksaan histopatologi masih menjadi standar emas diagnosis kanker payudara. Pemeriksaan *histopatologi* dapat mengetahui jenis kanker payudara [7]. Kanker payudara secara umum terbagi menjadi dua jenis yaitu jinak (*benign*) dan ganas (*malignant*). Kanker payudara jinak biasanya ditandai dengan benjolan kecil, bulat, dan lunak [8]. Kanker payudara jinak biasanya pertumbuhan nya tidak bersifat kanker. Meskipun kanker ini dapat dideteksi, namun tidak menyebar atau merusak jaringan di sekitarnya. Pada kasus, kanker payudara ganas ditandai dengan bentuk yang asimetris, kasar, dan nyeri. Pada kasus lainnya, kanker payudara biasanya bermetastasis dan merusak jaringan dan organ lain di sekitarnya. Oleh karena itu, diperlukan model untuk memprediksi penyakit kanker payudara sejak dini, yang merupakan langkah yang sangat penting untuk mencegah dan mengobati kanker payudara secara lebih efektif [9]. Keterlambatan penderita kanker payudara melakukan pemeriksaan awal ke pelayanan kesehatan di Indonesia mencapai lebih dari 80%. Hal ini mengakibatkan kejadian kanker payudara banyak yang ditemukan pada stadium lanjut[10].

Penderita penyakit kanker payudara dapat diprediksi dengan cara *data mining* [11]. Model prediktif ini dapat membantu mengidentifikasi individu yang berisiko tinggi terkena penyakit ini sehingga mereka dapat dengan cepat menerima perawatan dan intervensi yang diperlukan. Menurut *Daryl Pregibon*, *data mining* merupakan kombinasi statistik, kecerdasan buatan, dan penelitian basis data dan masih dalam pengembangan [12]. Tujuan utama dari data mining adalah untuk mengekstraksi wawasan berharga dari sejumlah besar data. Berbagai teknik data mining telah dikembangkan, antara lain teknik asosiasi*, clustering, sequence modeling,* dan klasifikasi [9]. Dengan menggunakan teknik pengumpulan data yang tepat, informasi penting dapat diidentifikasi dan pola yang terkait dengan perkembangan penyakit dapat ditemukan, yang dapat membantu diagnosis dini dan pendukung keputusan dalam pengelolaan penyakit.

Kanker payudara sebagai penyebab utama kematian pada wanita global [13] (WHO, 2021) memerlukan pendekatan diagnostik yang lebih canggih. Studi terbaru menunjukkan perkembangan signifikan dalam penerapan *machine learning* untuk prediksi kanker payudara, dimana algoritma XGBoost telah terbukti efektif dengan akurasi mencapai 99,27% [14]. Namun, penelitian sebelumnya masih memiliki keterbatasan dalam menangani data tidak seimbang [15] Penelitian lainnya, dimana 6 (enam) algoritma *machine learning,* seperti Random Forest, Decision-Tree, K-Nearest Neigbhbors, Logistic Regression, Support Vector Classifier, Linier Support Vector Classifier digunakan untuk mendeteksi dan pencegahan kanker payudara [16] dari 3002 gambar yang didapatkan dari 1501 individu. Penelitian ini menunjukkan bahwa algoritma *machine learning* memberikan nilai efisiensi dan akurasi yang tinggi untuk mendeteksi kanker payudara. Penelitian lainnya adalah kombinasi algoritma *machine learning* Support Vector Machine dan Principal Component Analysis mendapatkan akurasi 89% dalam deteksi dini kanker payudara dalam klasifikasi gambar [17]. Dilanjutkan dengan kombinasi 5 (lima) algoritma *machine learning,* seperti Random Forest, Decision-Tree, K-Nearest Neigbhbors, Logistic Regression, Support Vector Machine menunjukkan akurasi 97,2% dalam mendeteksi kanker payudara [18].

Oleh karena itu, penggunaan *data mining* atau *machine learning* dalam analisis data medis memiliki potensi besar untuk meningkatkan deteksi dini dan pengobatan kanker payudara [9]. Pada penelitian ini, peneliti melakukan klasifikasi penyakit kanker payudara dengan menggunakan metode *Decision Tree* dengan algoritma C4.5. *Decision Tree* (Pohon Keputusan) adalah struktur diagram alur berbasis pohon di mana setiap simpul internal mewakili pengujian suatu atribut, setiap cabang mewakili hasil pengujian, dan simpul daun mewakili kelas atau distribusi kelas [18]. Algoritma C4.5 mempunyai masukan berupa sampel pelatihan dan sampel, dan sampel adalah *field data* yang nantinya akan digunakan sebagai parameter untuk klasifikasi data[19]. Hasil klasifikasi Algoritma C4.5 divalidasi menggunakan *cross validation* untuk mengetahui tingkat akurasinya.

2. METODE PENELITIAN

Metode penelitian yang dilakukan untuk prediksi kanker payudari dengan klasifikasi Decision Tree menggunakan algoritma C4.5 dapat dilihat pada Gambar 1. Adapun penjelasan metode sebagai berikut:



Gambar 1. Metode Penelitian

2.1 Pengumpulan Data

Sumber Data: Dataset ini diambil dari sumber yang terpecaya dan terdiri dari 4025 dataser yang relevan dengan kanker payudara. Atribut yang Tersedi: Dataset ini memiliki 15 atribut yang berbeda, yaitu:

1. Umur: Informasi usia pasien dalam *dataset*.
2. Ras: Latar belakang ras atau etnis pasien, (misalnya hitam, putih dan lainnya).
3. Status Perkawinan: Status perkawinan pasien, misalnya (lanjang, menikah, dan cerai)
4. Stadium T: Tingkat perkembangan tumor primer.
5. Stadium N: Status kelenjar getah bening regional.
6. Stadium 6: Tahap penyakit berdasarkan sistem penilaian tertentu.
7. Derajat Pembeda: Tingkat diferensiasi sel kanker.
8. Stadium A: Klasifikasi stadium kanker payudara berdasarkan sistem tertentu.
9. Ukuran Tumor: Dimensi atau besaran tumor yang terdeteksi.
10. Status Estrogen: Kondisi reseptor estrogen pada sel kanker.
11. Status Progesteron: Kondisi reseptor progesteron pada sel kanker.
12. Pemeriksaan Nodus Regional: Jumlah kelenjar getah bening yang diperiksa.
13. Nodus Regional Positif: Jumlah kelenjar getah bening yang positif terkena kanker.
14. Bulan Bertahan Hidup: Durasi kelangsungan hidup pasien dalam bulan.
15. Status: Status akhir pasien (misalnya, hidup atau meninggal).

2.2 Proses Pengumpulan Data

a. Kegiatan Pengumpulan: Informasi tentang pokok bahasan materi dikumpulkan dari sumber yang dapat dipercaya. Pengumpulan data ini dapat mencakup survei, catatan pasien, atau sumber informasi kesehatan lainnya.
b. Keakuratan data: Proses pengumpulan data dilakukan untuk memastikan konsistensi, keakuratan, dan integritas data. Setiap atribut yang dikumpulkan harus mempunyai dokumentasi dan spesifikasi yang jelas.
c. Privasi dan Etika: Penting untuk memastikan bahwa data yang dikumpulkan konsisten dengan peraturan privasi dan etika penelitian yang berlaku, seperti memastikan anonimitas subjek atau mendapatkan persetujuan sebelum mengumpulkan data sensitif.
d. Pengolahan Awal: Setelah pengumpulan, data dapat diproses terlebih dahulu untuk membersihkan data dari duplikasi, kesalahan, atau nilai yang hilang untuk digunakan di masa mendatang.

Proses pengumpulan data ini merupakan langkah penting dalam memastikan bahwa dataset yang diperoleh valid, lengkap, dan sesuai dengan kebutuhan analisis dan pengujian model.

Tabel 1. *Dataset* Prediksi Penyakit Kanker Payudara

Age	Race	Marital Status	T Stage	N Stage	Tumor Size	Estrogen Status	Progesterone Status	Survival Months	Status
68	White	Married	T1	N1	4	Positive	Positive	60	Alive
50	White	Married	T2	N2	35	Positive	Positive	62	Alive
58	White	Divorced	T3	N3	63	Positive	Positive	75	Alive
58	White	Married	T1	N1	18	Positive	Positive	84	Alive
47	White	Married	T2	N1	41	Positive	Positive	50	Alive
51	White	Single	T1	N1	20	Positive	Positive	89	Alive
51	White	Married	T1	N1	8	Positive	Positive	54	Alive
40	White	Married	T2	N1	30	Positive	Positive	14	Dead
40	White	Divorced	T4	N3	103	Positive	Positive	70	Alive
69	White	Married	T4	N3	32	Positive	Positive	92	Alive
68	White	Widowed	T1	N1	13	Positive	Positive	64	Dead
46	White	Married	T3	N1	59	Negative	Negative	92	Alive
65	White	Married	T2	N1	35	Positive	Positive	56	Alive

48	White	Married	T1	N2	15	Positive	Positive	38	Alive
62	White	Divorced	T2	N3	35	Positive	Positive	64	Alive
61	White	Married	T1	N1	19	Positive	Positive	49	Alive
56	White	Single	T2	N1	46	Positive	Positive	105	Alive
43	White	Married	T2	N1	24	Positive	Positive	62	Alive
48	Black	Divorced	T2	N1	25	Positive	Positive	107	Alive
60	White	Divorced	T2	N1	29	Negative	Negative	77	Alive
48	Other	Married	T2	N1	30	Positive	Positive	81	Alive
57	White	Married	T2	N2	40	Positive	Positive	50	Alive
55	White	Single	T2	N1	29	Positive	Positive	78	Alive
48	White	Married	T3	N1	70	Positive	Negative	102	Alive

2.3. Algoritma C4.5

Algoritma C4.5 yaitu sebuah algoritma yang digunakan untuk membangun decision tree (pengambilan keputusan). Algoritma C.45 adalah salah satu algoritma induksi pohon keputusan yaitu ID3 (Iterative Dichotomiser 3). ID3 dikembangkan oleh J. Ross Quinlan. Dalam prosedur algoritma ID3, input berupa sampel training, label training dan atribut. Algoritma C4.5 merupakan pengembangan dari ID3. Beberapa pengembangan yang dilakukan pada C4.5 adalah sebagai antara lain bisa mengatasi missing value, bisa mengatasi continu data, dan pruning (Faradillah, 2013)[21].

Dalam penelitian yang menggunakan algoritma C4.5 untuk memprediksi kanker payudara, nilai gain dihitung untuk mengukur kemampuan setiap atribut dalam memisahkan data berdasarkan label kelas (misalnya, kanker payudara ganas atau jinak). Atribut dengan nilai gain tertinggi dianggap paling informatif dan paling kuat dalam memprediksi apakah seseorang memiliki risiko tinggi atau rendah terkena kanker payudara. Secara matematis, rumus untuk menghitung nilai gain pada C4.5 adalah sebagai berikut:

$$Entropy(S) = \sum_{i=1}^{m} pi \log 2(pi)$$

Keterangan:

S : Himpunan Kasus
m : Jumlah partisi S
pi : Proporsi dari Si terhadap S

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

Keterangan:

S : Himpunan kasus
A : Atribut
n : Jumlah partisi atribut A
|Si| : Jumlah kasus pada partisi ke i
|S| : Jumlah kasus dalam S

Dalam algoritma C4.5, nilai gain dihitung untuk menentukan atribut mana yang paling efektif dalam mengelompokkan data menjadi subset yang lebih homogen berdasarkan risiko kanker payudara. Atribut dengan nilai gain tertinggi dipilih untuk membuat keputusan di setiap node pada saat membangun pohon keputusan. Hal ini membantu mengidentifikasi atribut yang paling penting dalam memprediksi apakah seseorang berisiko tinggi atau rendah terkena kanker payudara.

2.4 RapidMiner

RapidMiner adalah platform perangkat lunak yang kuat untuk ilmu data dan pembelajaran mesin. Ini menyediakan beragam alat untuk persiapan data, pemodelan, evaluasi, dan implementasi. RapidMiner dirancang untuk mudah digunakan dan memungkinkan pengguna untuk dengan mudah membangun dan menguji berbagai model, bahkan tanpa pengalaman pemrograman [22]. RapidMiner merupakan perangkat lunak yang bersifat terbuka (*open source*) yang diciptakan dengan menggunakan bahasa pemrograman java sehingga bisa diakses oleh semua sistem operasi. RapidMiner dapat dijadikan solusi dalam melakukan analisis terhadap datamining, dengan menggunakan Teknik deskriptif dan prediksi yang diberikan kepada pengguna sehingga dapat mengambil keputusan yang paling baik [23]. RapidMiner tersedia dalam versi *open source* yang memungkinkan akses gratis untuk sebagian besar fungsionalitas dasar, dan memiliki komunitas pengguna yang aktif. Hal ini memungkinkan pengguna untuk berbagi pengetahuan, pengalaman, dan solusi terkait analisis data, serta memfasilitasi pertukaran informasi yang berguna [24].

3. HASIL DAN PEMBAHASAN

Penelitian ini menggunakan metode C4.5 untuk menganalisis data dan memprediksi gejala kanker payudara. Data diolah melalui tahap seleksi dan transformasi untuk memastikan kualitas dan kesesuaiannya dengan analisis. Hasil analisis akan digunakan sebagai panduan dalam memprediksi gejala kanker payudara dengan lebih akurat.

3.1 Proses Pengumpulan Data

Penelitian ini menggunakan data sekunder yang diperoleh dari rumah sakit. Data ini telah tersedia sehingga tidak memerlukan pengumpulan data melalui observasi, wawancara, atau survei. Data yang digunakan terdiri dari 4025 sampel, yang kemudian dibagi menjadi 80% untuk pelatihan model dan 20% untuk pengujian. Pohon keputusan dibangun menggunakan RapidMiner dengan data pelatihan yang berfokus pada prediksi "Alive" atau "Dead" untuk kemungkinan terkena kanker payudara. Data pelatihan ini terdiri dari 120 sampel, dengan 658 sampel dikategorikan sebagai "Alive" dan 65 sampel dikategorikan sebagai "Dead".

Tabel 2. Daftar atribut data keseluruhan

Atribut	Tipe Data	Keterangan
Tumor Size	Integer	1-140 (mm)
Estrogen Status	Binominal	Positive/Negative
Progesterone Status	Binominal	Positive/Negative
Survival Months	Integer	1-107 (month)
Status	Label	Alive/Dead

Berikut adalah penjelsan mengenai Tabel 2:

a. Tumor Size: Data ukuran tumor dalam dataset ini diidentifikasi dengan atribut "Tumor Size" dan diukur dalam mm dengan rentang 1-140; digunakan untuk menilai keparahan kanker dan merencanakan pengobatan.

b. Estrogen Status: Atribut "Estrogen Status" menggambarkan status reseptor estrogen pada sel kanker dengan opsi nilai "Positive" dan "Negative"; digunakan untuk menentukan jenis pengobatan hormonal.
c. Progesterone Status: Atribut "Progesterone Status" menunjukkan status reseptor progesteron pada sel kanker dengan opsi nilai "Positive" dan "Negative"; digunakan untuk menentukan jenis pengobatan hormonal.
d. Survival Months: Data lama hidup pasien dalam dataset ini diidentifikasi dengan atribut "Survival Months" dan diukur dalam bulan dengan rentang 1-107; digunakan untuk mengukur efektivitas pengobatan dan kelangsungan hidup pasien.
e. Status: Atribut "Status" menunjukkan apakah pasien masih hidup atau sudah meninggal dengan opsi nilai "Alive" dan "Dead"; digunakan untuk analisis outcome dan risiko.

3.2 Data Selection

Pemilihan data dilakukan karena beberapa atribut dianggap kurang relevan dalam memprediksi gejala kanker payudara. Tujuannya adalah untuk memilih atribut yang paling tepat untuk membangun model prediksi kanker payudara yang akurat. Tabel 3 di bawah ini menunjukkan daftar atribut yang akan digunakan dan diolah dalam perangkat lunak RapidMiner.

Tabel 3. Atribut Terpilih

No	Atribut	Keterangan
1	Tumor Size	Atribut Fitur
2	Estrogen Status	Atribut Fitur
3	Progesterone Status	Atribut Fitur
4	Survival Months	Atribut Fitur
5	Status	Label

Setelah melalui proses analisis data yang menyeluruh, beberapa atribut terpilih untuk membangun model prediksi kanker payudara dengan lebih akurat Atribut yang telah terpilih adalah tumor size, estrogen status, progesterone status, Progesterone Status, Survival Months, and status. Pemilihan atribut-atribut ini didasarkan pada analisis statistik dan korelasi mereka dengan risiko terkena kanker payudara. Atribut-atribut yang telah dipilih akan diolah menggunakan algoritma C4.5 dalam perangkat lunak RapidMiner. Algoritma ini akan membangun pohon keputusan dan mengevaluasi akurasinya dalam memprediksi kemungkinan terkena kanker payudara. Pemrosesan ini dilakukan karena beberapa atribut dirasa kurang relevan dalam memprediksi penyakit kanker payudara. Tujuannya adalah untuk memilih atribut yang paling tepat dan efektif dalam membangun model prediksi yang akurat. Gambar 2 di bawah ini menunjukkan daftar atribut yang akan digunakan dan diolah dalam RapidMiner.

3.3 Data Transformation

Tahap transformasi data ini penting untuk mempersiapkan data agar sesuai dengan format yang dibutuhkan dalam proses data mining. Transformasi dilakukan menggunakan *software* RapidMiner dengan mengubah tipe data atribut yang digunakan dari polinomial menjadi binomial. Hal ini dapat dilihat pada Gambar 2 yang menunjukkan keseluruhan data setelah transformasi.

T Stage *polynominal*	N Stage *polynominal*	Tumor Size *integer*	Estrogen St... *binominal*	Progestero... *binominal*	Survival M... *integer*	Status *binominal label*
T1	N1	4	Positive	Positive	60	Alive
T2	N2	35	Positive	Positive	62	Alive
T3	N3	63	Positive	Positive	75	Alive
T1	N1	18	Positive	Positive	84	Alive
T2	N1	41	Positive	Positive	50	Alive
T1	N1	20	Positive	Positive	89	Alive
T1	N1	8	Positive	Positive	54	Alive
T2	N1	30	Positive	Positive	14	Dead
T4	N3	103	Positive	Positive	70	Alive
T4	N3	32	Positive	Positive	92	Alive
T1	N1	13	Positive	Positive	64	Dead
T3	N1	59	Negative	Negative	92	Alive
T2	N1	35	Positive	Positive	56	Alive
T1	N2	15	Positive	Positive	38	Alive
T2	N3	35	Positive	Positive	64	Alive
T1	N1	19	Positive	Positive	49	Alive
T2	N1	46	Positive	Positive	105	Alive
T2	N1	24	Positive	Positive	62	Alive

PREVIEW

no problem

Previous Next Canc

Gambar 2. Data yang terdapat dari Rumah Sakit

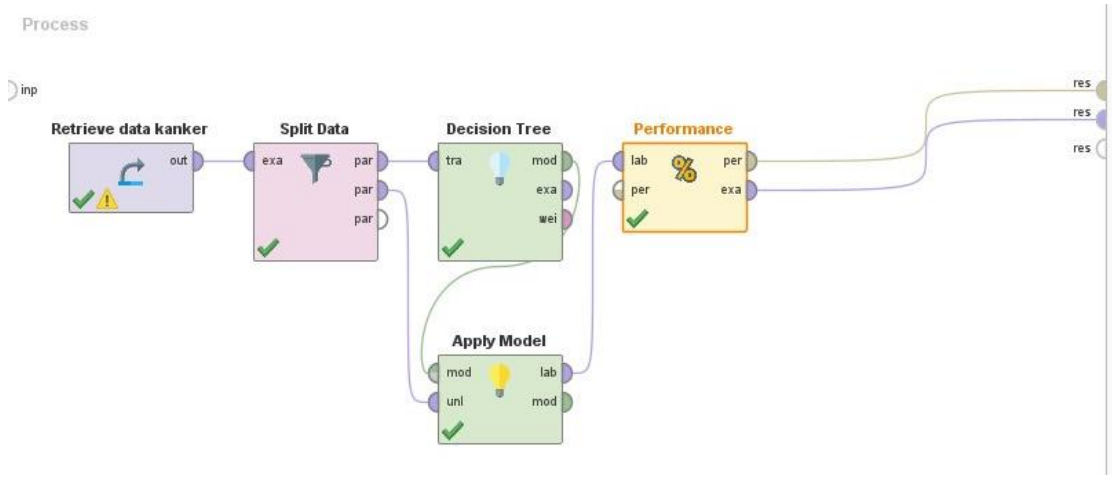
3.4 Data Mining

Data mining adalah proses yang menggunakan statistik, matematika, kecerdasan buatan dan machine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar [25]. Algoritma C4.5 merupakan salah satu algoritma data mining yang populer dan banyak digunakan untuk membangun model prediksi berdasarkan data yang tersedia. Algoritma ini bekerja dengan menganalisis data yang telah melalui proses seleksi dan transformasi untuk memprediksi kemungkinan pasien terkena kanker payudara. Dalam penelitian ini, proses data mining diawali dengan pengumpulan data dari berbagai sumber seperti rekam medis, seperti tumor size, estrogen status, progesterone status, Progesterone Status, Survival Months, and status. Data-data ini kemudian melalui beberapa tahap pengolahan, termasuk pembersihan data (data cleansing) untuk menyingkirkan data yang tidak lengkap, tidak akurat, atau tidak relevan. Selanjutnya, data disiapkan melalui tahap transformasi untuk mengubah data mentah menjadi format yang sesuai untuk analisis.

3.5 Implementasi menggunakan RapidMiner

RapidMiner merupakan salah satu perangkat lunak populer yang digunakan dalam proses analisis data, termasuk dalam konteks data mining. Dalam konteks spesifik Anda, penggunaan RapidMiner bertujuan untuk memproses data menggunakan algoritma C4.5. Setelah proses ini selesai, diharapkan akn dihasilkan pohon keputusan dan aturan - aturan terkait dari analisis tersebut. Proses yang melibatkan algoritma C4.5 dalam RapidMiner dimulai dengan tahap pengolahan data. Data yang telah disiapkan sebelumnya, yang mungkin telah melalui tahap pembersihan, transformasi, dan pemilihan atribut, akan dimuat ke dalam perangkat lunak. RapidMiner akan membantu Anda dalam menyiapkan data tersebut, memungkinkan visualisasi yang lebih baik, serta memfasilitasi penggunaan algoritma C4.5 untuk membangun model pohon keputusan [24].

Dalam penelitian ini, algoritma C4.5 akan dijalankan dalam software RapidMiner untuk menganalisis data yang telah diolah sebelumnya. Algoritma ini akan mengidentifikasi pola-pola yang ada dalam data dan membangun model pohon keputusan berdasarkan atribut-atribut yang relevan dengan kanker payudara. Proses ini melibatkan penentuan atribut-atribut terpenting yang memengaruhi hasil prediksi terkait kemungkinan pasien terkena kanker payudara. Setelah algoritma C4.5 selesai dijalankan, RapidMiner akan menghasilkan visualisasi pohon keputusan yang menunjukkan cabang-cabang keputusan berdasarkan atribut-atribut yang paling signifikan. Saat algoritma C4.5 selesai dieksekusi, RapidMiner akan menghasilkan pohon keputusan yang dapat divisualisasikan. Pohon keputusan ini akan memperlihatkan cabang- cabang keputusan berdasarkan atribut-atribut yang paling signifikan. Selain pohon keputusan, RapidMiner juga dapat menyajikan aturan-aturan yang terkait dengan model yang telah dibangun. Aturan-aturan ini merupakan representasi dari kondisi-kondisi yang mempengaruhi keputusan prediksi terkait kanker paru-paru berdasarkan data yang diolah [26].



Gambar 3. Penggunaan RapidMiner

Pada Gambar di atas merupakan proses perhitungan menggunakan RapidMiner.

a. Pohon Keputusan

Dari hasil pengujian pada RapidMiner, diperoleh pohon keputusan seperti yang terlihat pada Gambar 4.



Gambar 4. Pohon Keputusan

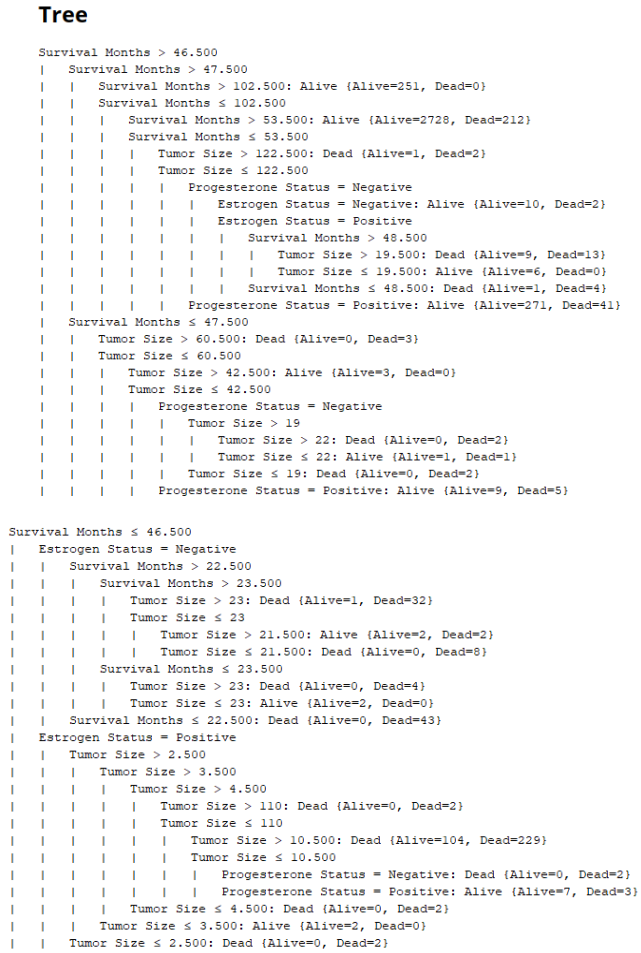
Hasil dari pohon Keputusan diatas yaitu:

a. Jika waktu kelangsungan hidup > 60 bulan, maka pasien masih hidup.
b. Jika waktu kelangsungan hidup ≤ 60 bulan dan ukuran tumor > 122.500, maka pasien meninggal.
c. Jika waktu kelangsungan hidup ≤ 60 bulan, ukuran tumor ≤ 122.500, dan status progesteron negatif, maka pasien hidup.
d. Jika waktu kelangsungan hidup ≤ 60 bulan, ukuran tumor ≤ 122.500, status progesteron positif, dan status estrogen negatif, maka pasien hidup.
e. Jika waktu kelangsungan hidup ≤ 60 bulan, ukuran tumor ≤ 122.500, status progesteron positif, status estrogen positif, dan waktu kelangsungan hidup > 47.500 bulan, maka pasien hidup.
f. Jika waktu kelangsungan hidup ≤ 47.500 bulan, ukuran tumor > 80.500, maka pasien meninggal.
g. Jika waktu kelangsungan hidup ≤ 47.500 bulan, ukuran tumor ≤ 80.500, dan waktu kelangsungan hidup > 22.500 bulan, maka pasien hidup.
h. Jika waktu kelangsungan hidup ≤ 22.500 bulan, ukuran tumor > 23.500, maka pasien meninggal
i. Jika waktu kelangsungan hidup ≤ 22.500 bulan, ukuran tumor ≤ 23.500, dan ukuran tumor > 21.500, maka pasien hidup.
j. Jika waktu kelangsungan hidup ≤ 22.500 bulan, ukuran tumor ≤ 21.500, dan ukuran tumor > 21, maka pasien meninggal.
k. Jika waktu kelangsungan hidup ≤ 22.500 bulan, ukuran tumor ≤ 21, dan ukuran tumor > 2.300, maka pasien hidup.
l. Jika waktu kelangsungan hidup ≤ 22.500 bulan, ukuran tumor ≤ 2.300, maka pasien meninggal.

Keterangan di atas hanya mencakup beberapa jalur utama. Untuk analisis yang lebih menyeluruh, diperlukan evaluasi lebih lanjut terhadap setiap kombinasi faktor risiko yang ditunjukkan dalam diagram.

b. Deskripsi Pohon Keputusan

Berdasarkan analisis pohon keputusan yang disajikan mengungkapkan serangkaian pola dan ketentuan yang kompleks. Visualisasi ini, yang dihasilkan menggunakan perangkat Rapidminer, memetakan berbagai jalur keputusan yang mungkin terjadi, dapat Anda lihat pada Gambar 5 dibawa ini.



Tree

```
Survival Months > 46.500
|   Survival Months > 47.500
|   |   Survival Months > 102.500: Alive {Alive=251, Dead=0}
|   |   Survival Months ≤ 102.500
|   |   |   Survival Months > 53.500: Alive {Alive=2728, Dead=212}
|   |   |   Survival Months ≤ 53.500
|   |   |   |   Tumor Size > 122.500: Dead {Alive=1, Dead=2}
|   |   |   |   Tumor Size ≤ 122.500
|   |   |   |   |   Progesterone Status = Negative
|   |   |   |   |   |   Estrogen Status = Negative: Alive {Alive=10, Dead=2}
|   |   |   |   |   |   Estrogen Status = Positive
|   |   |   |   |   |   |   Survival Months > 48.500
|   |   |   |   |   |   |   |   Tumor Size > 19.500: Dead {Alive=9, Dead=13}
|   |   |   |   |   |   |   |   Tumor Size ≤ 19.500: Alive {Alive=6, Dead=0}
|   |   |   |   |   |   |   Survival Months ≤ 48.500: Dead {Alive=1, Dead=4}
|   |   |   |   |   Progesterone Status = Positive: Alive {Alive=271, Dead=41}
|   Survival Months ≤ 47.500
|   |   Tumor Size > 60.500: Dead {Alive=0, Dead=3}
|   |   Tumor Size ≤ 60.500
|   |   |   Tumor Size > 42.500: Alive {Alive=3, Dead=0}
|   |   |   Tumor Size ≤ 42.500
|   |   |   |   Progesterone Status = Negative
|   |   |   |   |   Tumor Size > 19
|   |   |   |   |   |   Tumor Size > 22: Dead {Alive=0, Dead=2}
|   |   |   |   |   |   Tumor Size ≤ 22: Alive {Alive=1, Dead=1}
|   |   |   |   |   Tumor Size ≤ 19: Dead {Alive=0, Dead=2}
|   |   |   |   Progesterone Status = Positive: Alive {Alive=9, Dead=5}

Survival Months ≤ 46.500
|   Estrogen Status = Negative
|   |   Survival Months > 22.500
|   |   |   Survival Months > 23.500
|   |   |   |   Tumor Size > 23: Dead {Alive=1, Dead=32}
|   |   |   |   Tumor Size ≤ 23
|   |   |   |   |   Tumor Size > 21.500: Alive {Alive=2, Dead=2}
|   |   |   |   |   Tumor Size ≤ 21.500: Dead {Alive=0, Dead=8}
|   |   |   Survival Months ≤ 23.500
|   |   |   |   Tumor Size > 23: Dead {Alive=0, Dead=4}
|   |   |   |   Tumor Size ≤ 23: Alive {Alive=2, Dead=0}
|   |   Survival Months ≤ 22.500: Dead {Alive=0, Dead=43}
|   Estrogen Status = Positive
|   |   Tumor Size > 2.500
|   |   |   Tumor Size > 3.500
|   |   |   |   Tumor Size > 4.500
|   |   |   |   |   Tumor Size > 110: Dead {Alive=0, Dead=2}
|   |   |   |   |   Tumor Size ≤ 110
|   |   |   |   |   |   Tumor Size > 10.500: Dead {Alive=104, Dead=229}
|   |   |   |   |   |   Tumor Size ≤ 10.500
|   |   |   |   |   |   |   Progesterone Status = Negative: Dead {Alive=0, Dead=2}
|   |   |   |   |   |   |   Progesterone Status = Positive: Alive {Alive=7, Dead=3}
|   |   |   |   Tumor Size ≤ 4.500: Dead {Alive=0, Dead=2}
|   |   |   Tumor Size ≤ 3.500: Alive {Alive=2, Dead=0}
|   |   Tumor Size ≤ 2.500: Dead {Alive=0, Dead=2}
```

Gambar 5. Deskripsi Pohon Keputusan

Berdasarkan pohon keputusan yang ditampilkan pada Gambar 5, dapat dijelaskan bahwa dari beberapa atribut penilaian terhadap prediksi kanker payudara, atribut "Survival Months" (Bulan Kelangsungan Hidup) memiliki pengaruh paling signifikan terhadap prediksi kanker payudara. Pohon keputusan ini menunjukkan bagaimana berbagai faktor seperti waktu kelangsungan hidup, ukuran tumor, status estrogen, dan status progesteron berinteraksi untuk memprediksi hasil (hidup atau meninggal) pada pasien kanker payudara.

Dalam konteks ini, pohon keputusan yang dihasilkan menggambarkan penerapan metode klasifikasi (mungkin menggunakan algoritma seperti C4.5 atau yang serupa) untuk menilai prediksi hasil kanker payudara. Visualisasi ini kemungkinan dihasilkan menggunakan perangkat lunak analisis data seperti RapidMiner, yang memungkinkan pengujian keakuratan dan kebenaran model prediktif yang dibangun berdasarkan data yang digunakan. Pohon keputusan ini memberikan gambaran rinci tentang bagaimana berbagai faktor klinik berperan dalam menentukan prognosis pasien kanker payudara, memungkinkan para profesional medis untuk membuat keputusan yang lebih terinformasi dalam penanganan pasien.

4. KESIMPULAN

Melalui pengumpulan data yang teliti dan proses analisis yang cermat, peneliti berhasil mengembangkan model klasifikasi yang mampu memprediksi dengan akurasi yang memuaskan. Penelitian ini menunjukkan bahwa faktor-faktor tertentu seperti ukuran tumor, bentuk sel, dan atribut lainnya dapat menjadi prediktor yang kuat dalam diagnosis kanker payudara. Hasil evaluasi performa model juga menunjukkan tingkat sensitivitas dan spesifisitas yang memadai, mendukung potensi aplikasi praktis dalam klinis. Namun demikian, penelitian ini tidak terlepas dari beberapa batasan, termasuk keterbatasan pada jumlah dan variasi data yang tersedia, serta pengaturan parameter yang optimal pada model *Decision Tree*. Dalam masa depan, penelitian lebih lanjut dapat fokus pada peningkatan akurasi prediksi dengan mempertimbangkan lebih banyak fitur atau menggunakan metode klasifikasi yang lebih kompleks. Secara keseluruhan, penelitian ini memberikan kontribusi dalam bidang deteksi dini kanker payudara melalui penerapan teknik klasifikasi yang efektif.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Tim Redaksi Jurnal Jupiter Politeknik Negeri Sriwijaya yang telah memberi memberi kesempatan, sehingga artikel ilmiah ini dapat diterbitkan.

DAFTAR PUSTAKA

[1] B. Wulandari, “Studi Literatur: Kualitas Hidup Pasien Kanker Kolorektal Pasca Kolostomi,” vol. 2, 2020.

[2] W. R. Hastutiningtyas, S. Andarini, dan L. Supriati, “Pengaruh Logoterapi terhadap Kesejahteraan Psikologis pada Pasien Kanker Payudara di Rumah Sakit TK II dr. Soepraoen’ Malang,” *Care J. Ilm. Ilmu Kesehat.*, vol. 5, no. 3, hlm. 329–344, Nov 2017, doi: 10.33366/jc.v5i3.696.

[3] A. Fristiohady dan L. A. Haruna, “Review Jurnal: Potensi Spons Laut Sebagai Anti Kanker Payudara,” *J. Mandala Pharmacon Indones.*, vol. 6, no. 01, Art. no. 01, Jul 2020, doi: 10.35311/jmpi.v6i01.56.

[4] K. Suparna dan L. M. K. K. S. Sari, “Kanker Payudara: Diagnostik, Faktor Risiko, Dan Stadium,” *Ganesha Med.*, vol. 2, no. 1, hlm. 42–48, Jun 2022, doi: 10.23887/gm.v2i1.47032.

[5] E. S. Wahyuni, “Penerapan Metode Seleksi Fitur Untuk Meningkatkan Hasil Diagnosis Kanker Payudara,” *Simetris J. Tek. Mesin Elektro Dan Ilmu Komput.*, vol. 7, no. 1, hlm. 283, Apr 2016, doi: 10.24176/simet.v7i1.516.

[6] 1 A. Gani M.Kes S. Pd ,. SKM ,. S. Kep ,. M. Kes 2. Yeni Elviani ,. AM Keb ,. SKM ,. S. Kep ,. M. Kes 3. Ns Andre Utama Saputra, S. Kep ,. M. Kep 4. Ns Dedi Fatrida, S. Kep ,. M. Kep 5. Mustakim, Amd Kep ,. S. Kep ,. S. Sos, *Pendidikan Kesehatan Program Pencegahan Kanker Payudara (terhadap pengetahuan, sikap, dan tindakan remaja)*. Penerbit Adab.

[7] Gambaran stadium dan jenis histopatologi kanker payudara di Subbagian Bedah Onkologi RSUP Sanglah Denpasar tahun 2015-2016 | Intisari Sains Medis.” Diakses: 9 Juli 2024. [Daring]. Tersedia pada: https://imap2.isainsmedis.id/index.php/ism/article/view/165

[8] “Analisis performa metode Knn pada Dataset pasien pengidap Kanker Payudara | Indonesian Journal of Data and Science.” Diakses: 9 Juli 2024. [Daring]. Tersedia pada: https://www.jurnal.yoctobrain.org/index.php/ijodas/article/view/13

[9] R. Rofiani, L. Oktaviani, D. Vernanda, dan T. Hendriawan, "Penerapan Metode Klasifikasi Decision Tree dalam Prediksi Kanker Paru-Paru Menggunakan Algoritma C4.5," *J. Tekno Kompak*, vol. 18, no. 1, Art. no. 1, Feb 2024, doi: 10.33365/jtk.v18i1.3525.

[10] G. A. R. Dyanti dan N. L. P. Suariyani, "Faktor-Faktor Keterlambatan Penderita Kanker Payudara Dalam Melakukan Pemeriksaan Awal Ke Pelayanan Kesehatan," *J. Kesehat. Masy.*, vol. 11, no. 2, Art. no. 2, Jan 2016, doi: 10.15294/kemas.v11i2.3742.

[11] "Penerapan Algoritma Decision Tree Dan Regresi Linear Untuk Klasifikasi Kanker Payudara | SENTIMAS: Seminar Nasional Penelitian dan Pengabdian Masyarakat." Diakses: 9 Juli 2024. [Daring]. Tersedia pada: https://journal.irpi.or.id/index.php/sentimas/article/view/571

[12] "Data Mining: The Textbook | SpringerLink." Diakses: 9 Juli 2024. [Daring]. Tersedia pada: https://link.springer.com/book/10.1007/978-3-319-14142-8?ref=https%3A%2F%2Fgithubhelp.com

[13] T. I. A. for R. on Cancer (IARC), "Global Cancer Observatory." Accessed: Jun. 19, 2025. [Online]. Available: https://gco.iarc.fr/

[14] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Systems with Applications*, vol. 73, pp. 220–239, May 2017, doi: 10.1016/j.eswa.2016.12.035.

[15] T. E. Mathew, "Breast Cancer Classification Using an Extreme Gradient Boosting Model with F-Score Feature Selection Technique," *JAIT*, vol. 14, no. 2, pp. 363–372, 2023, doi: 10.12720/jait.14.2.363-372. [16] T. I. A. for R. on Cancer (IARC), "Global Cancer Observatory." Accessed: Jun. 19, 2025. [Online]. Available: https://gco.iarc.fr/

[16] A. Khalid *et al.*, "Breast Cancer Detection and Prevention Using Machine Learning," *Diagnostics (Basel)*, vol. 13, no. 19, p. 3113, Oct. 2023, doi: 10.3390/diagnostics13193113.

[17] N. Manjunathan, N. Gomathi, and S. Muthulingam, "Early Detection of Breast Cancer using Machine Learning," in *2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS)*, Jun. 2023, pp. 165–169. doi: 10.1109/ICSCSS57650.2023.10169777.

[18] M. A. Naji, S. E. Filali, K. Aarika, E. H. Benlahmar, R. A. Abdelouhahid, and O. Debauche, "Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis," *Procedia Computer Science*, vol. 191, pp. 487–492, Jan. 2021, doi: 10.1016/j.procs.2021.07.062.

[19] http://download.garuda.kemdikbud.go.id/article.php?article=990863&val=15134&title=PENENTUAN%20JADWAL%20KERJA%20BERDASARKAN%20KLASIFIKASI%20DATA%20KARYAWAN%20MENGGUNAKAN%20METODE%20DECISION%20TREE%20C45%20Studi%20Kasus%20Universitas%20Muhammadiyah%20Surabaya." Diakses: 9 Juli 2024. [Daring]. Tersedia pada: http://download.garuda.kemdikbud.go.id/article.php?article=990863&val=15134&title=PENENTUAN%20JADWAL%20KERJA%20BERDASARKAN%20KLASIFIKASI%20DATA%20KARYAWAN%20MENGGUNAKAN%20METODE%20DECISION%20TREE%20C45%20Studi%20Kasus%20Universitas%20Muhammadiyah%20Surabaya

[20] "http://download.garuda.kemdikbud.go.id/article.php?article=2410730&val=13467&title=Simulasi%20Penerapan%20Metode%20Decision%20Tree%20C45%20Pada%20Penentuan%20Status%20Gizi%20Balita." Diakses: 9 Juli 2024. [Daring]. Tersedia pada: http://download.garuda.kemdikbud.go.id/article.php?article=2410730&val=13467&title=Simulasi%20Penerapan%20Metode%20Decision%20Tree%20C45%20Pada%20Penentuan%20Status%20Gizi%20Balita

[21] L. Bachtiar dan M. Mahradianur, "Analisis Data Mining Menggunakan Metode Algoritma C4.5 Menentukan Penerima Bantuan Langsung Tunai," *J. Inform.*, vol. 10, no. 1, hlm. 28–36, Mar 2023, doi: 10.31294/inf.v10i1.15115.

[22] M. Rafi Nahjan, Nono Heryana, dan Apriade Voutama, "Implementasi Rapidminer Dengan Metode Clustering K-Means Untuk Analisa Penjualan Pada Toko Oj Cell," *JATI J. Mhs. Tek. Inform.*, vol. 7, no. 1, hlm. 101–104, Jan 2023, doi: 10.36040/jati.v7i1.6094.

[23] A. Apriyadi, M. R. Lubis, dan B. E. Damanik, "Penerapan Algoritma C5.0 Dalam Menentukan Tingkat Pemahaman Mahasiswa Terhadap Pembelajaran Daring," *Komputa J. Ilm. Komput. Dan Inform.*, vol. 11, no. 1, hlm. 11–20, Apr 2022, doi: 10.34010/komputa.v11i1.7386.

[24] R. Rofiani, L. Oktaviani, D. Vernanda, dan T. Hendriawan, "Penerapan Metode Klasifikasi Decision Tree dalam Prediksi Kanker Paru-Paru Menggunakan Algoritma C4.5," *J. Tekno Kompak*, vol. 18, no. 1, Art. no. 1, Feb 2024, doi: 10.33365/jtk.v18i1.3525.

[25] N. Manullang, R. W. Sembiring, I. Gunawan, I. Parlina, dan I. Irawan, "Implementasi Teknik Data Mining untuk Prediksi Peminatan Jurusan Siswa Menggunakan Algoritma C4.5," *J. Ilmu Komput. Dan Teknol.*, vol. 2, no. 2, hlm. 1–5, Nov 2021, doi: 10.35960/ikomti.v2i2.700.

[26] I. M. A. O. Gunawan, I. D. A. I. Saraswati, I. D. G. R. Agung, dan I. P. E. Putra, "Klasifikasi Penyakit Jantung Menggunakan Algoritma Decision Tree Series C4.5 Dengan Rapidminer," *J. Teknol. Dan Sist. Inf. Bisnis*, vol. 5, no. 2, Art. no. 2, Apr 2023, doi: 10.47233/jteksis.v5i2.775.