Institut Riset dan Publikasi Indonesia (IRPI)
MALCOM: Indonesian Journal of Machine Learning and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/malcom
Vol. 4 Iss. 4 October 2024, pp: 1485-1495
ISSN(P): 2797-2313 | ISSN(E): 2775-8575

The Application of C4.5 Decision Tree Algorithm for Predicting the Survival Rate of Thyroid Cancer Patients

Penerapan Algoritma Decesion Tree C4.5 untuk Memprediksi Tingkat Kelangsungan Hidup Pasien Kanker Tiroid

Adinda Dwi Putri[1], Fitriana Sholekhah[2], Eric Dadynata[3], Lusiana Efrizoni[4], Rahmaddeni[5*], Nur Sapina[6]

[1,2,3,4,5]Program Studi Teknik Informatika, Universitas Sains dan Teknologi Indonesia, Indonesia

E-Mail: [1]2110031802112@sar.ac.id, [2]2110031802099@sar.ac.id, [3]2110031802021@sar.ac.id, [4]lusiana@sar.ac.id, [5]rahmaddeni@sar.ac.id, [6]2310031802147@sar.ac.id

Received Jul 12th 2024; Revised Aug 55th 2024; Accepted Sept 10th 2024
Corresponding Author: Rahmaddeni

Abstract

One of the diseases with varying survival rates among patients is thyroid cancer. To predict patient survival rates based on clinical characteristics, this study employs the Decision Tree C4.5 algorithm. This method utilizes Natural Language Processing (NLP) with Count Vectorizer to convert text into numerical data. In evaluating prediction accuracy, the model's performance is assessed using a confusion matrix to analyze classification performance. Additionally, Area Under Curve (AUC) is calculated to evaluate the model's performance. Experimental results demonstrate that this method accurately predicts survival rates of thyroid cancer patients, achieving 97% accuracy and an AUC of 0.95, indicating excellent performance. This research contributes to deepening the understanding of Decision Tree algorithm applications in medical contexts and its potential to support clinical decision-making in the future.

Keyword: Area Under Curve (AUC), Confusion Matrix, Decision Tree C4.5, Survival Prediction, Thyroid Cancer

Abstrak

Salah satu penyakit yang memiliki tingkat kelangsungan hidup yang bervariasi di antara pasien adalah kanker tiroid. Untuk memprediksi tingkat kelangsungan hidup pasien berdasarkan karakteristik klinis, penelitian ini menggunakan algoritma Decision Tree C4.5. Metode ini memanfaatkan pengolahan bahasa alami (NLP) dengan Count Vectorizer untuk mengubah teks menjadi data numerik. Dalam penilaian keakuratan prediksi, evaluasi dilakukan dengan matriks kebingungan (confusion matrix) untuk mengukur kinerja model dalam klasifikasi. Selain itu, Area Under Curve (AUC) juga dihitung untuk mengevaluasi performa model. Hasil eksperimen menunjukkan bahwa metode ini memberikan prediksi yang akurat tentang tingkat kelangsungan hidup pasien dengan kanker tiroid, mencapai akurasi sebesar 97% dan AUC sebesar 0.95, menunjukkan kinerja yang sangat baik. Penelitian ini penting untuk memperdalam pemahaman tentang penerapan Decision Tree dalam konteks medis dan potensi algoritma ini dalam mendukung pengambilan keputusan klinis di masa depan.

Kata Kunci: Area Under Curve (AUC), Confusion Matrix, Decision Tree C4.5, Kanker Tiroid, Kelangsungan Hidup

1. PENDAHULUAN

Decision Tree C4.5 merupakan salah satu algoritma yang paling populer dan sederhana untuk dipahami. Hasilnya menyerupai cara kerja otak manusia, sehingga aturan-aturan yang dihasilkannya juga mudah dimengerti. Karena sifat pohonnya, maknanya, dan kesimpulan yang dapat dipahami oleh orang banyak, algoritmaini bahkan berguna di bidang medis, membantu dokter membuat keputusan penting tentang laporan patologi tertentu [1]. Algoritma C4.5 adalah evolusi dari algoritma ID3, dan Ross Quinlan adalah orang yang mengembangkannya. Tujuan pengembangan algoritma C4.5 adalah untuk meningkatkan akurasi pengklasifikasian data dan memperbaiki atribut yang tidak lengkap dari ID3 [2].

Sekitar 7% insidensi kanker tiroid meningkat setiap tahun, menunjukkan peningkatan yang lebih cepat dibandingkan dengan jenis kanker solid lainnya[3]. Kelenjar endokrin yang dikenal sebagai tiroid terletak di bagian depan leher, tepatnya di belakang otot *sternothyroideus* dan *sternohyoideus*, sejajar dengan *vertebra cervicalis V* hingga *vertebra thoracica* [4]. Kelenjar tiroid merupakan organ yang memproduksi dua hormon

utama: tiroksin dan kalsitonin. Tiroksin berperan dalam mengatur metabolisme sel tubuh, sedangkan kalsitonin bertanggung jawab atas pengaturan *metabolisme* kalsium. Oleh karena itu, kelenjar tiroid memainkan peran penting dalam menjaga keseimbangan metabolisme kalsium dalam tubuh manusia.

Kanker merupakan penyakit yang muncul akibat pertumbuhan sel-sel jaringan tubuh yang abnormal. Sel-sel kanker berkembang dengan cepat dan dapat menyebar ke area lain dalam tubuh, yang akhirnya bisa berujung pada kematian [5]. Kanker yang berasal dari kelenjar tiroid organ kecil berbentuk kupu-kupu di bagian depan leher terjadi ketika sel-sel tiroid tumbuh secara tidak terkendali dan membentuk tumor ganas. Karsinoma folikuler dan papiler adalah dua jenis kanker tiroid yang paling umum. Meskipun gejala kanker tiroid dapat bervariasi, beberapa yang paling umum termasuk pembengkakan atau benjolan di leher, perubahan suara, masalah menelan, dan kadang-kadang nyeri. Jenis dan stadium kanker tiroid menentukan pengobatannya, yang dapat mencakup operasi, terapi radiasi, atau terapi hormon.

Faktor genetik dan riwayat keluarga adalah salah satu dari banyak penyebab kanker tiroid. Memiliki riwayat keluarga yang didiagnosis dengan kanker tiroid atau kondisi genetik tertentu, seperti sindrom neoplasia endokrin multipel atau poliposis adenomatosa keluarga, memiliki potensi untuk lebih rentan. Selain itu, paparan radiasi pada kepala, leher, atau dada, terutama pada anak-anak, dapat meningkatkan risiko terkena kanker tiroid ini dapat terjadi karena paparan radiasi lingkungan atau terapi radiasi untuk kanker lain. Merokok juga dapat menyebabkan otak kekurangan oksigen, dan nikotin yang terkandung dalam rokok dapat menyebabkan reaksi inflamasi menjadi lebih intens. Selain itu, stres berkorelasi dengan antibodi yang menentang antibodi TSH-reseptor [6].

Wanita lebih sering terkena dibandingkan pria dan mayoritas kasus terjadi pada orang berusia tiga puluh hingga lima puluh tahun. Risiko terkena kanker tiroid juga dapat meningkat jika memiliki riwayat penyakit tiroid, seperti tiroiditis Hashimoto atau goiter, yang merupakan pembesaran kelenjar tiroid. Selain itu, diet yang sangat rendah yodium juga dapat meningkatkan risiko terkena kanker tiroid, meskipun ini lebih sering terjadi di daerah dengan diet yang lebih sederhana. Beberapa penulis menyebutkan faktor-faktor kondisi penting, seperti Disabilitas intelektual, pasien dengan kondisi komorbid seperti gangguan mental atau fisik, penyakit kronis yang berdampak pada kualitas hidup, termasuk diabetes melitus, penyakit arteri koroner, gagal ginjal kronis, kejang, *multiple sclerosis*, penyakit rematik, penyakit vaskular serebral, serta riwayat operasi dalam kurun waktu kurang dari 6 bulan.

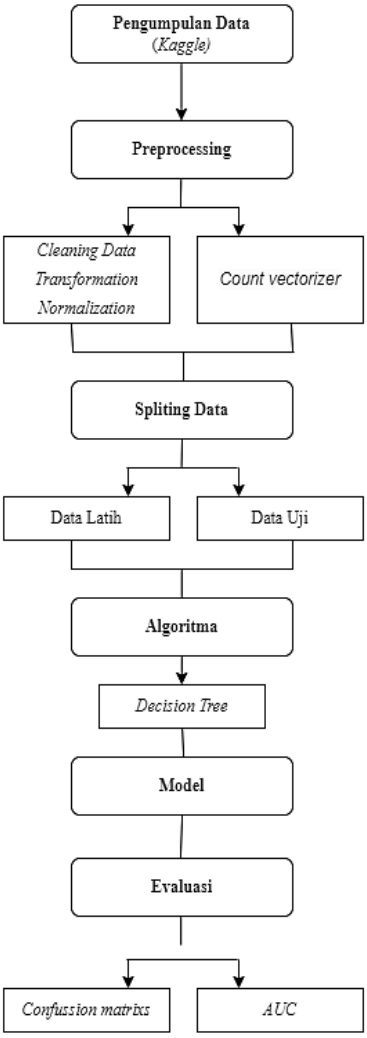
Salah satu metode data mining yang dapat diterapkan untuk memprediksi penyakit adalah teknik klasifikasi. Data mining melibatkan serangkaian proses yang memanfaatkan satu atau lebih teknik pembelajaran mesin untuk secara otomatis menganalisis dan mengekstrak informasi dari kumpulan data, sehingga menghasilkan wawasan tambahan yang tidak dapat diperoleh secara manual [7]. Penelitian ini menggunakan teknik klasifikasi berbasis algoritma Decision Tree C4.5. Decision Tree, atau Pohon Keputusan, adalah metode yang memanfaatkan struktur pohon untuk membuat keputusan dengan mengikuti serangkaian aturan sederhana yang diterapkan pada data input. Pohon keputusan membagi data berdasarkan karakteristik tertentu sehingga pada setiap langkahnya memaksimalkan kehomogenan grup yang dihasilkan, yang memungkinkan untuk menghasilkan prediksi atau keputusan dari data yang diberikan. Serta penggunaan *CountVectorizer* untuk fitur kelas perhitungan numerik yang menggunakan metode ekstraksi fitur teks. *CountVectorizer* ini mengubah teks kata menjadi matriks frekuensi kata; kemudian, fungsi matriks *fit_transform* digunakan untuk menghitung berapa kali setiap kata muncul[8].

Peneliti sebelumnya yang menggunakan Decision Tree mencapai akurasi sebesar 95,80%, dengan standar deviasi sebesar 2,74%. Nilai presisi yang diperoleh adalah 94,29%, nilai recall mencapai 95,90%, dan nilai kelas yang tidak berprestasi adalah 95,65%[9].. Peneliti juga akan melakukan evalusi *Matrix Confusion* dan *Area Under the Curve (AUC)*. Mengevaluasi kedua metrik ini digunakan untuk memastikan bahwa model yang dikembangkan berfungsi dengan baik, membedakan antara berbagai kelas dalam dataset dan dapat diandalkan untuk memprediksi atau mengklasifikasikan data baru dengan benar. Namun, AUC (Area Under Curve) adalah ukuran kemampuan model untuk membedakan kelas negatif dan positif akurasi memberikan gambaran umum tentang kebenarannya dalam klasifikasinya secara keseluruhan.

Model *Decision Tree* C4.5 kami menunjukkan kinerja yang luar biasa dalam memprediksi kelangsungan hidup pasien kanker tiroid, dengan akurasi 97% dan AUC 0,95. Sebagai perbandingan[10], penelitian ini menggunakan Random Fores*t* menunjukkan akurasi 87,75% dan F1-Score 0,922. Terlepas dari kenyataan bahwa Random Forest menawarkan keseimbangan yang sangat baik antara ketepatan dan ingatan, model C4.5 kami menunjukkan tingkat akurasi yang lebih tinggi, yang menunjukkan bahwa ia lebih efektif dalam menemukan kasus positif. Hasil ini menunjukkan bahwa model kami dapat diterapkan secara klinis untuk memprediksi kelangsungan hidup pasien. Hasil evaluasi ini juga sangat penting untuk menilai keakuratan dan efektivitas model Decision Tree C4.5 dalam konteks spesifik prediksi kelangsungan hidup pasien kanker tiroid. Evaluasi ini membantu memahami seberapa baik model bekerja dan memungkinkan perbaikan atau penyesuaian yang diperlukan selama pengembangan model.

2. METODOLOGI PENELITIAN

Penelitian ini bertujuan untuk mengembangkan model klasifikasi untuk kanker tiroid. Proses penelitian dimulai dengan [1] pengumpulan data; [2] *preprocessing*; [3] pembagian data; [4] penerapan algoritma; dan [5] model, seperti yang ditunjukkan pada gambar 1.



Gambar 1. Alur Penelitian

2.1 Pengumpulan Data

Data utama untuk penelitian ini berasal dari kumpulan data publik tentang kanker tiroid yang tersedia melalui platform Kaggle. Kumpulan data ini terdiri dari 384 titik data dengan 17 fitur yang menggambarkan berbagai aspek yang terkait dengan kanker tiroid, termasuk kekambuhan pasien. Secara keseluruhan, data ini memberikan landasan yang kuat untuk melakukan analisis menyeluruh tentang kanker tiroid, yang merupakan fokus penelitian ini.

2.2 *Preprocessing*

Preprocessing data adalah tahap analisis data yang sangat penting di mana data disiapkan dan dibersihkan agar siap untuk proses analisis berikutnya. Tujuan dari pengolahan data pada tahap ini adalah untuk menghindari data yang mengganggu (*noise*) atau tidak konsisten [11].

1. *Data Cleaning*

   Data cleansing adalah proses yang melibatkan pemeriksaan untuk memastikan bahwa tidak ada nilai data yang kosong (*null*) dan untuk menemukan data duplikat. Memeriksa dan menghapus nilai kosong sangat penting untuk menjaga kualitas data, sementara menemukan dan menghapus data duplikat sangat penting untuk menghindari bias dan memastikan akurasi analisis. Kami dapat menjamin bahwa dataset siap untuk analisis lebih lanjut dengan memastikan bahwa tidak ada masalah dengannya [12]. Atau, Mengidentifikasi, memperbaiki, atau menghapus data yang tidak akurat, tidak lengkap, tidak relevan, atau tidak terbaru dari sebuah kumpulan data dikenal sebagai proses pembersihan data. Proses

penyiapan data terdiri dari dari menghilangkan atau menambah nilai kosong di seluruh kumpulan data menggunakan nilai rata-rata dari masing-masing kolom nilai kosong [13].

2. *Data Tranformation*
   Transformasi data berarti mengolah data agar dapat menghasilkan data tambahan yang lebih berkualitas. Menggabungkan data, generalisasi, normalisasi, meng-agregasi, dan menciptakan fitur dan atribut adalah beberapa proses yang akan dilakukan dengan data [14]. Menurut [15] transformasi data dikenal sebagai mengubah data ke dalam format yang dapat digunakan untuk proses data mining.

3. *Normalization*
   Proses mengubah nilai-nilai dari berbagai variabel yang termasuk dalam kumpulan data sehingga rentang nilai masing-masing variabel menjadi seragam atau sebanding dengan satu sama lain dikenal sebagai normalisasi data. Ini adalah proses mengubah Normalisasi data adalah pengurangan atribut numerik menjadi skala yang lebih sederhana seperti 0 hingga 1 [16].

4. *Count vectorizer*
   Metode untuk membuat vector kalimat adalah *Count Vectorizer*. Tujuannya adalah untuk mengumpulkan semua kata yang ada dalam setiap kalimat dan membuat kosa kata yang terdiri dari semua kata yang berbeda yang ada dalam setiap kalimat; kemudian, kosa kata ini dapat digunakan untuk membuat vector fitur dari jumlah kata yang ada [17].

2.3 *Splitting Data*

Langkah berikutnya adalah pembagian data, juga dikenal sebagai pembagian data. Data dibagi menjadi dua bagian oleh proses ini: data latihan dan data uji. Ini dilakukan untuk memastikan model dapat digunakan pada data baru, sehingga dibagi tiga kali untuk penelitian ini: 60:40, 70:30, dan 80:20. Algoritma Decision Tree akan digunakan untuk menguji data yang telah dibagi. Data uji digunakan untuk mengevaluasi keakuratan melatih model, sedangkan data latih adalah informasi yang akan digunakan untuk melatih model [18].

1. Data Latih
   Data Training juga dikenal sebagai data pelatihan, adalah sekumpulan data yang diberi label atau kelas yang digunakan mesin untuk mengidentifikasi karakteristiknya., menghasilkan pola atau model data[19]. Training data adalah data di mana data digunakan untuk membangun sebuah model[20]. Tujuannya adalah agar model dapat mengidentifikasi pola dan hubungan dalam data tersebut.

2. Data Uji
   Data testing digunakan untuk mengevaluasi performa algoritma yang sudah dilatih sebelumnya ketika menemukan data baru[21]. Dalam proses pengembangan model pembelajaran mesin, pengujian data memainkan peran penting. Hal ini diperlukan untuk memastikan bahwa model yang telah dilatih tidak hanya menghafal data pelatihan yang telah diberikan kepadanya. Dengan pengujian data, kemampuan model untuk menggeneralisasi dan dapat mengevaluasi dan membuat prediksi yang akurat berdasarkan data baru yang belum pernah dilihat sebelumnya.

2.4 Algoritma *Decision Tree C4.5*

Algoritma adalah instruksi langkah demi langkah yang diberikan kepada komputer untuk menyelesaikan tugas atau mencapai tujuan. Data mining melibatkan serangkaian langkah yang dirancang untuk mengungkap nilai tambahan dari data yang sebelumnya tidak terdeteksi secara manual dari sebuah database, melalui proses penggalian pola. Tujuan dari proses ini adalah mengubah data menjadi informasi dengan mengekstraksi dan mengidentifikasi pola-pola yang signifikan atau menarik dari data yang ada di basis data [22]. Algoritma klasifikasi yang dikenal sebagai Decision Tree digambarkan sebagai pembagian rekursif dari ruang contoh. Decision Tree terdiri dari node yang membentuk pohon berakar, yaitu pohon yang diarahkan dengan sebuah simpul yang disebut akar; node dengan cabang keluar disebut node internal atau tes; dan node lainnya disebut daun. Setiap node internal dalam pohon keputusan, misalnya, membagi ruang menjadi dua atau lebih subruang sesuai dengan fungsi diskrit tertentu dari atribut nilai [23].

Algoritma Decision Tree C4.5 adalah algoritma pembelajaran mesin yang digunakan untuk tugas klasifikasi. Algoritma ini dikembangkan oleh Ross Quinlan sebagai evolusi dari algoritma Iterative Dichotomiser 3, atau ID3. Algoritma C4.5 menggunakan fitur data yang digunakan untuk membuat prediksi. Pohon keputusan untuk data pelatihan dibuat oleh algoritma C4.5, yang terdiri dari kasus-kasus atau catatan (tupel) yang ada dalam basis data [24]. Algoritma C4.5 untuk menciptakan pohon keputusan secara global melakukan hal-hal berikut [28]:

1. Memilih atribut untuk digunakan sebagai akar
2. Membuat cabang untuk setiap nilai
3. Membagi kasus dalam cabang

4. Mengulangi proses sampai setiap cabang agar semua kasus pada cabang memiliki kelas yang sama [25].

2.5 Model

Model ini adalah hasil dari proses pembelajaran yang dilakukan oleh algoritma pembelajaran mesin, yang mengidentifikasi pola, struktur, dan hubungan dalam data melalui data pelatihan. Algoritma adalah representasi matematis atau komputasional yang dapat membuat prediksi atau keputusan berdasarkan input baru melalui serangkaian iterasi dan penyesuaian parameter. Model ini kemudian dapat digunakan untuk berbagai tujuan, seperti klasifikasi, regresi, deteksi anomali, dan prediksi, dengan kinerja yang dievaluasi menggunakan berbagai metrik untuk memastikan bahwa prediksi yang dihasilkan adalah akurat dan dapat diandalkan [27].

2.6 Evaluasi

Dua metrik utama yang digunakan dalam evaluasi pohon keputusan adalah *confusion matrix* dan *Area Under the Curve* (AUC). *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN) adalah empat kategori prediksi yang dikategorikan sebagai benar dan salah dalam tabel *confusion matrix*. Untuk menghitung akurasi, presisi, dan recall, ini digunakan. Kemampuan model untuk membedakan kelas negatif dan positif diukur dengan AUC. Ini diperoleh dari grafik ROC, yang menunjukkan perbedaan antara rasio positif asli (sensitivitas) dan rasio positif palsu (1—spesifisitas). Nilai AUC berkisar antara 0 dan 1; nilai AUC yang lebih tinggi menunjukkan bahwa model bekerja lebih baik. Seberapa baik model pohon keputusan dalam prediksi dapat dilihat dengan menggabungkan kedua metrik ini.

2.6.1 *Confusion Matrix*

Confusion Matrix juga dikenal sebagai matriks kebingungan, adalah alat yang digunakan untuk mengevaluasi bagaimana algoritma klasifikasi bekerja dalam pembelajaran mesin. Matriks ini menunjukkan bagaimana prediksi model dibandingkan dengan label data uji yang sebenarnya. Elemen-elemen dalam *Confusion matrix* untuk masalah klasifikasi biner terdiri dari empat komponen utama:

1. *True Positive (TP):* Frekuensi model memprediksi kelas positif dengan benar.
2. *True Negatives (TN):* Berapa kali model memprediksi kelas negatif dengan benar.
3. *False Positive (FP):* Berapa kali model salah memprediksi kelas positif (juga dikenal sebagai kesalahan Tipe I atau positif palsu).
4. *False Negatives (FN):* Berapa kali model salah memprediksi kelas negatif (juga dikenal sebagai kesalahan atau kesalahan Tipe II).

Untuk menghitung *accuracy, precision,* dan r*ecal* rumus *confusion matrix* berikut digunakan:

$$Accuracy = \frac{TP+TN}{Total} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+Fn} \quad (3)$$

$$F1 - score = 2 \times \frac{Precission \times Recall}{Precission+Recall} \quad (4)$$

2.6.2 *Area Under Curve* (AUC)

Area Under the Curve (AUC) adalah metrik evaluasi yang sangat penting dan sering digunakan dalam pembelajaran mesin untuk mengukur kinerja model klasifikasi. Metrik ini memberikan gambaran tentang kemampuan model dalam membedakan kelas-kelas yang berbeda, khususnya dalam konteks klasifikasi biner. AUC menunjukkan seberapa baik model dapat memprediksi kelas positif dan negatif daripada kurva ROC (*Receiver Operating Characteristic*). AUC adalah ukuran total kemampuan model untuk membedakan antara kelas positif dan negative. Nilai mulai dari 0 hingga 1:

1. Jika AUC = 1, model memiliki kemampuan sempurna untuk membedakan kelas positif dan negatif.
2. Jika AUC = 0,5, model tidak dapat membedakan antara kelas positif dan negatif, sama seperti melakukan prediksi acak.
3. Jika AUC kurang dari 0,5 = model melakukan prediksi acak dengan lebih buruk, yang menunjukkan bahwa model mungkin membalikkan kelas.

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Penelitian ini dimulai dengan mengumpulkan data dari situs *Kaggle*. Analisis yang dilakukan menunjukkan bahwa dataset ini mencakup 384 entri., dengan variabel independen kanker tiroid masing-masing dengan nilai 0 dan 1. Nilai 0 menunjukkan bahwa pasien tidak menderita kanker tiroid. Dengan algoritma pohon keputusan, data ini akan digunakan sebagai dasar untuk memprediksi tingkat kelangsungan hidup pasien. Sebagai sampel data ditunjukkan pada tabel 1.

Tabel 1. Data Kanker Tiroid

	Age	Gender	Smoking	Hx Smoking	…	Recurred
1	22	F	No	No	…	No
2	24	F	No	Yes	…	No
3	34	F	No	No	…	No
…	…	…	…	…	…	…
384	67	M	Yes	No	…	Yes

Data tentang kanker tiroid ditunjukkan dalam Tabel 1. Ini mencakup berbagai fitur, seperti *Age, Gender, Smoking, Hx Smoking, Hx Radiotherapy, Thyroid Function, Physical Examination, Adenopathy, Pathology, Focality, Risk, T, N, M, Stage, Response, dan Recurred*. bawah ini adalah penjelasan singkat dari setiap fitur:

1. *Age:* Usia pasien saat diagnosis.
2. *Gender:* Jenis kelamin pasien (Laki-laki atau Perempuan).
3. *Smoking:* Status merokok pasien (Ya atau Tidak).
4. *Hx Smoking:* Riwayat merokok pasien (Ya atau Tidak).
5. *Hx Radiotherapy*: Riwayat terapi ramereka
6. si yang diterima pasien (Ya atau Tidak).
7. *Thyroid Function:* Fungsi tiroid pasien berdasarkan hasil tes laboratorium.
8. *Physical Examination:* Hasil pemeriksaan fisik pasien oleh dokter.
9. *Adenopathy:* Adanya pembesaran kelenjar getah bening (Ya atau Tidak).
10. *Pathology*: Hasil patologi dari biopsi tiroid.
11. *Focality:* Jumlah fokus tumor (Unifokal atau Multifokal).
12. *Risk:* Tingkat risiko berdasarkan evaluasi medis.
13. *T:* Ukuran dan extent tumor primer (berdasarkan sistem TNM).
14. *N:* Keterlibatan kelenjar getah bening regional (berdasarkan sistem TNM).
15. *M:* Kehadiran metastasis jauh (berdasarkan sistem TNM).
16. *Stage:* Stadium kanker berdasarkan klasifikasi TNM.
17. *Response:* Respons pasien terhadap pengobatan.
18. *Recurred:* Kekambuhan kanker setelah pengobatan awal (Ya atau Tidak).

3.2 *Preprocessing*

Proses ini termasuk tahap persiapan. Untuk setiap kelas, tangga klasifikasi menyediakan metrik yang lebih rinci ('No' dan 'Yes') atau (‘0’ dan ‘1’). Proses *preprocessing outlier* berhasil dilakukan Tabel 2.

Tabel 2. Hasil *Preprocessing Outliers*

	Age	Gender	Smoking	Hx Smoking	Actual	Recurred
1	22	0	0	0	0	0
2	24	0	0	1	0	0
3	34	0	0	0	0	0
…	…	…	…	…	...	…
384	67	1	1	0	0	1

3.3 *Splitting Data*

Penelitian ini membagi data menjadi beberapa subset (60:40, 70:30, dan 80:20) Pada tahap 60:40: 60% menggunakan menggunakan data pelatihan (training), dan 40% menggunakan data pengujian. Data dibagi dengan rasio 70:30 pada langkah selanjutnya, dengan 70 persen digunakan untuk pelatihan dan 30 persen untuk pengujian. 80:20: 80% menggunakan data pelatihan, dan 20% dipakai dalam pengujian. Hasil *Splitting Data* dapat diliat pada tabel 3.

Tabel 3. Hasil *Spliting Data*

Spliting data	Decision Tree C4.5			
	Akurasi	Presisi	Recall	F1-Score
60.40.00	94%	91%	96%	93%

Spliting data	Decision Tree C4.5			
	Akurasi	Presisi	Recall	F1-Score
70.30.00	97%	97%	96%	97%
80.20.00	96%	94%	96%	96%

3.4 Algoritma Decision Tree C4.5

Untuk tugas klasifikasi, Decision Tree C4.5 adalah algoritma pembelajaran mesin. Penelitian ini menunjukkan bahwa algoritma ini dapat digunakan untuk memprediksi tingkat kelangsungan hidup pasien kanker tiroid dengan tingkat akurasi yang baik. Temuan ini berkontribusi pada upaya peningkatan metode prediksi dalam bidang kesehatan, khususnya dalam penanganan pasien kanker tiroid. *Decision tree*, atau sering disebut sebagai 'pohon keputusan, merupakan model prediktif yang memanfaatkan struktur pohon atau hierarki. Data diolah menjadi pohon keputusan dan serangkaian aturan keputusan. [26].

3.5 Model

Untuk menganalisis data dan membuat pohon keputusan untuk, model ini menggunakan Algoritma Pohon Keputusan C4.5 memproses informasi dan membuat aturan untuk memprediksi tingkat kelangsungan hidup pasien kanker tiroid. Pohon keputusan ini membantu dalam mengklasifikasikan data dengan mengacu pada atribut yang relevan, memungkinkan prediksi yang akurat tentang kelangsungan hidup pasien berdasarkan data yang dianalisis.

3.5.1 Penerapan Algoritma *Decision Tree C4.5*

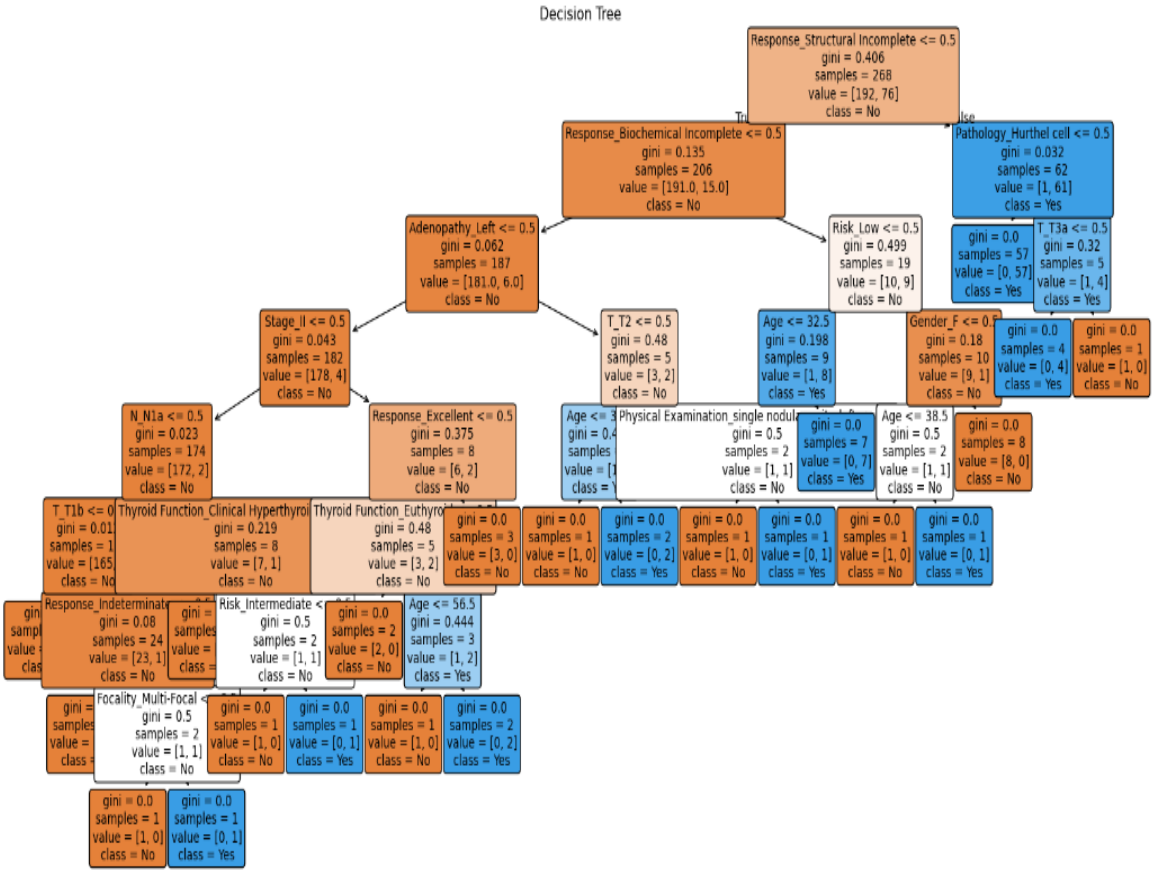
Di bawah ini contoh skripnya disajikan dalam potongan kode ini menunjukkan proses pemrosesan data, mulai dari pemilihan fitur hingga konstruksi pohon keputusan. Berikut ini ditampilkan potongan script dalam mengolah data dengan algoritma *Desicion Tree C4.5*.

Potongan Kode Dalam Penerapan Algoritma

```
  # Memisahkan data menjadi data latih dan data uji
  X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,
random_state=42)

  # Melatih model Decision Tree
  model = DecisionTreeClassifier()
  model.fit(X_train, y_train)

  # Memastikan model memiliki atribut classes_
  print(f'Model classes: {model.classes_}')

  # Memprediksi data uji
  y_pred = model.predict(X_test)

  # Mengevaluasi performa model
  accuracy = accuracy_score(y_test, y_pred)
      report = classification_report(y_test, y_pred)

  print(f'Accuracy: {accuracy}')
  print('Classification Report:')
  print(report)

  # Visualisasi Confusion Matrix
  conf_matrix = confusion_matrix(y_test, y_pred)
  plt.figure(figsize=(10, 7))
  sns.heatmap(conf_matrix, annot=True, fmt="d", cmap="Blues",
xticklabels=model.classes_, yticklabels=model.classes_)
  plt.xlabel('Predicted')
  plt.ylabel('True')
  plt.title('Confusion Matrix')
plt.show()
```

3.5.2 Hasil Pemodelan Pohon Keputusan

Berikut ini adalah gambar pohon keputusan yang dibuat, yang menunjukkan bagaimana model membuat keputusan. Pohon keputusan ini menunjukkan cara menggunakan setiap fitur untuk membagi data ke dalam kelas yang berbeda. Dalam pohon ini, setiap simpul menunjukkan pilihan yang didasarkan pada nilai fitur tertentu, dan cabang-cabangnya menunjukkan hasil dari pilihan tersebut. Untuk setiap simpul, nilai ini menunjukkan tingkat ketidakmurnian, dengan nilai yang lebih rendah menunjukkan simpul yang lebih murni.



Gambar 2. Hasil Keputusan Dengan Pohon Keputusan

Pohon keputusan ini membuat prediksi dengan menggunakan nilai fitur *Response_Structural Incomplete, Response_Biochemical Incomplete, Adenopathy_Left*, dan *Stage_II*. Nilai fitur ini, serta aturan yang diikuti sepanjang jalur dari simpul akar hingga simpul daun, membentuk keputusan akhir tentang data. Pohon ini menunjukkan bagaimana setiap keputusan dibuat dan bagaimana data dimasukkan ke dalam kelas "No" atau "Yes". Maka, diperoleh rule dari pohon keputusan sebagai berikut:

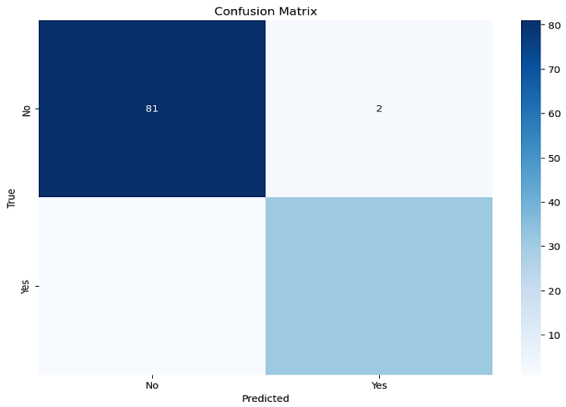
1. Jika "Response_Structural Incomplete" <= 0.5 dan "N_N1a" <= 0.5 dan "T_T1b" <= 0.5 dan "Response_Excellent" <= 0.5 dan "Age" > 39.5, maka class = "No".
2. Jika "Response_Structural Incomplete" > 0.5 dan "Pathology_Hurthel cell" <= 0.5 dan "Response_Biochemical Incomplete" > 0.5 dan "Adenopathy_Left" <= 0.5 dan "Risk_Low" <= 0.5 dan "N_N1a" <= 0.5 dan "Stage_II" > 0.5 dan "Gender_M" > 0.5 dan "Age" <= 30.0, maka class = "Yes".

3.6 Evaluasi

Menggabungkan metrik *confusion matrix* dan *AUC* memberikan gambaran yang lebih baik tentang kemampuan model *decision tree* untuk melakukan prediksi. Dalam penelitian ini, model decision tree menunjukkan nilai *nilai akurasi 97%, presisi 96%, recall 96%, dan skor f1 97%. Nilai AUC yang diperoleh adalah 95%, menunjukkan bahwa model memiliki kemampuan yang baik untuk membedakan kelas negatif dan positif.* Hasil evaluasi ini menunjukkan bahwa model pohon keputusan yang digunakan cukup baik untuk memprediksi tingkat kelangsungan hidup pasien yang didiagnosis dengan kanker tiroid.

3.6.1 *Conffusion Matrix*

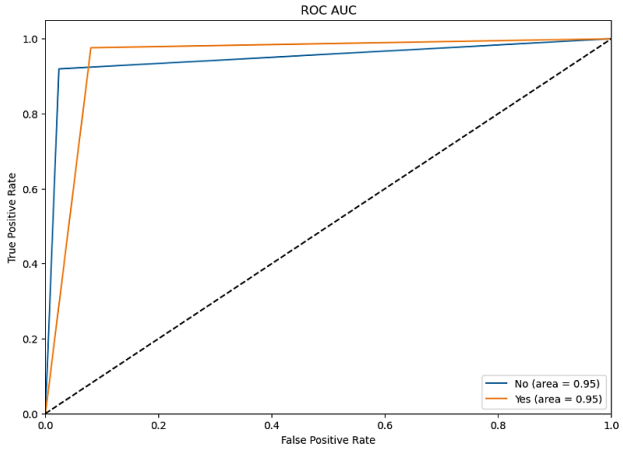
Gambar 3 menunjukkan hasil *Conffusion Matrix True Negative* (TN): `81 (Jumlah prediksi akurat untuk kategori “No”). Model yang menggunakan 'No' dan nilai sebenarnya dalam81 kali. *False Positive* (FP): 2 (Jumlah prediksi salah di mana model memperdiksi 'Yes' tetapi sebenarnya 'No'). Tidak ada skenario yang modelnya memprediksi "Yes", melainkan "No". *False Negative* (FN): 0 (Nilai titik di mana model memprediksi 'No' namun sebenarnya 'Yes'). Model 'No' digunakan, namun hasil sebenarnya kira-kira 19 kali 'Yes. Benar Positif (TP): 17 (Prediksi yang dapat diandalkan untuk kategori "Yes"). 'Yes' tidak ditentukan secara konsisten oleh model.



Gambar 3. Hasil *Conffusion Matrix*

3.6.2 AUC

Gambar 4 menunjukan hasil Nilai area di bawah kurva, atau AUC, memberikan kinerja klasifikasi model dalam membandingkan sampel positif dan negatif. Semakin tinggi nilai tinggi dibandingkan dengan 1, semakin akurat model tersebut dalam menggambarkan kelas yang bersangkutan. Hasil pada AUC yes sebesar 95% dan no sebesar 95%.



Gambar 4. Hasil *AUC*

Nilai AUC sebesar 0.95 untuk kedua kelas menunjukkan bahwa model ini memiliki kemampuan klasifikasi yang sangat baik, termasuk dalam kategori "Excellent Classification" berdasarkan tabel 4.

Tabel 4. *Guide for Classifying the AUC*

Range AUC	Classification Level
0.90 – 1.00	Excellent Classification
0.80 – 0.90	Good Classification
0.70 – 0.80	Fair Classification
0.60 – 0.70	Poor Classification
<0.60	Failure

Dengan nilai AUC 0.95, Model ini menunjukkan bahwa algoritma Decission Tree C4.5 sangat baik dalam memisahkan kelas positif dan negatif. Ini berarti model memiliki tingkat keakuratan yang tinggi dalam memprediksi apakah suatu sampel akan masuk ke dalam kelas "Yes" atau "No".

4. KESIMPULAN

Studi ini menyelidiki penerapan algoritma *Decision Tree C4.5* berdasarkan karakteristik klinis pasien kanker tiroid untuk memprediksi tingkat kelangsungan hidup menggunakan data dari dataset publik di *Kaggle*. Metodologi yang digunakan meliputi pengumpulan data, *preprocessing*, spliting data, penerapan algoritma *Decision Tree C4.5*, serta evaluasi model. Hasil penelitian menunjukkan bahwa model *Decision Tree C4.5* akurat. dalam memprediksi kelangsungan hidup pasien kanker tiroid, dengan akurasi 97% dan nilai AUC 0.95 (kategori *Excellent Classification*). Fitur-fitur seperti *Age*, *Smoking"No"*, *Hx Radiotherapy"No"*, *Thyroid*

Function, dan *Adenopathy"No"* menunjukkan prognosis yang lebih baik. Data klinis juga mendukung kemungkinan bertahan hidup lebih tinggi, seperti hasil patologi yang baik dan stadium kanker yang tidak lanjut, serta respons positif terhadap pengobatan dan tidak adanya kekambuhan kanker. Dengan mempertimbangkan semua faktor ini, model memprediksi bahwa pasien memiliki peluang tinggi untuk bertahan hidup "*Yes*".

REFERENSI

[1] R. A. Saputra *et al.*, "Detecting Alzheimer's Disease by the Decision Tree Methods Based on Particle Swarm Optimization," *J. Phys. Conf. Ser.*, vol. 1641, no. 1, pp. 61–67, 2020, doi: 10.1088/1742-6596/1641/1/012025.

[2] F. M. Hana, "Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma Decision Tree C4.5," *J. SISKOM-KB (Sistem Komput. dan Kecerdasan Buatan)*, vol. 4, no. 1, pp. 32–39, 2020, doi: 10.47970/siskom-kb.v4i1.173.

[3] A. Siswandi, N. Fitriyani, I. Artini, and K. Monitira, "Karakteristik Penderita Kanker Tiroid Di Bagian Bedah Onkologi Di Rumah Sakit Umum Daerah Dr. H. Abdul Moeloek Provinsi Lampung Tahun 2017-2019," *J. Med. Malahayati*, vol. 4, no. 3, pp. 244–248, 2021, doi: 10.33024/jmm.v4i3.2530.

[4] Y. M. Pius Cardia, E. D. Martadiani, and F. P. Sitanggang, "Karakteristik Ultrasonografi Pada Kecurigaan Klinis Kanker Tiroid Di Rsup Sanglah Denpasar Periode Januari 2015-Desember 2015," *E-Jurnal Med. Udayana*, vol. 10, no. 7, p. 45, 2021, doi: 10.24843/mu.2021.v10.i7.p09.

[5] D. Pembimbing, J. Dwi, T. Purnomo, S. Si, and M. Si, "Analisis Regresi Logistik Ordinal pada Faktor-Faktor yang Memperngaruhi Stadium Kanker Tiroid," 2020.

[6] S. Agustiani, A. Mustopa, A. Saryoko, W. Gata, and S. K. Wildah, "Penerapan Algoritma J48 Untuk Deteksi Penyakit Tiroid," *Paradig. - J. Komput. dan Inform.*, vol. 22, no. 2, pp. 153–160, 2020, doi: 10.31294/p.v22i2.8174.

[7] I. Khoeri and D. Iskandar Mulyana, "Implementasi Machine Learning dengan Decision Tree Algoritma C4.5 dalam Penerimaan Karyawan Baru pada PT. Gitareksa Dinamika Jakarta," *J. Sos. Teknol.*, vol. 1, no. 7, pp. 615–623, 2021, doi: 10.59188/jurnalsostech.v1i7.126.

[8] K. Rahayu, V. Fitria, D. Septhya, R. Rahmaddeni, and L. Efrizoni, "Klasifikasi Teks untuk Mendeteksi Depresi dan Kecemasan pada Pengguna Twitter Berbasis Machine Learning," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 3, no. 2, pp. 108–114, 2023, doi: 10.57152/malcom.v3i2.780.

[9] A. Supriyadi, "Perbandingan Algoritma Naive Bayes dan Decision Tree(C4.5) dalam Klasifikasi Dosen Berprestasi," *Gener. J.*, vol. 7, no. 1, pp. 39–49, 2023, doi: 10.29407/gj.v7i1.19797.

[10] I. P. Putri, T. Terttiaavini, and N. Arminarahmah, "Analisis Perbandingan Algoritma Machine Learning untuk Prediksi Stunting pada Anak," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 1, pp. 257–265, 2024, doi: 10.57152/malcom.v4i1.1078.

[11] F. Alghifari and D. Juardi, "Penerapan Data Mining Pada Penjualan Makanan Dan Minuman Menggunakan Metode Algoritma Naïve Bayes," *J. Ilm. Inform.*, vol. 9, no. 02, pp. 75–81, 2021, doi: 10.33884/jif.v9i02.3755.

[12] F. Fadli and B. B. Butar, "Penerapan Decision Tree Menggunakan Algoritma C4.5 Untuk Deteksi Demam Berdarah Pada RS. IMC Bintaro," *Indones. J. Softw. Eng.*, vol. 5, no. 1, pp. 75–86, 2019, doi: 10.31294/ijse.v5i1.5866.

[13] Gde Agung Brahmana Suryanegara, Adiwijaya, and Mahendra Dwifebri Purbolaksono, "Peningkatan Hasil Klasifikasi pada Algoritma Random Forest untuk Deteksi Pasien Penderita Diabetes Menggunakan Metode Normalisasi," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 1, pp. 114–122, 2021, doi: 10.29207/resti.v5i1.2880.

[14] B. Prasojo and E. Haryatmi, "Analisa Prediksi Kelayakan Pemberian Kredit Pinjaman dengan Metode Random Forest," *J. Nas. Teknol. dan Sist. Inf.*, vol. 7, no. 2, pp. 79–89, 2021, doi: 10.25077/teknosi.v7i2.2021.79-89.

[15] N. Azwanti and E. Elisa, "Analisis Pola Penyakit Hipertensi Menggunakan Algoritma C4.5," *InfoTekJar (Jurnal Nas. Inform. dan Teknol. Jaringan)*, vol. 3, no. 2, pp. 116–123, 2019, doi: 10.30743/infotekjar.v3i2.944.

[16] M. D. Purbolaksono, M. Irvan Tantowi, A. Imam Hidayat, and A. Adiwijaya, "Perbandingan Support Vector Machine dan Modified Balanced Random Forest dalam Deteksi Pasien Penyakit Diabetes," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 2, pp. 393–399, 2021, doi: 10.29207/resti.v5i2.3008.

[17] F. Pradana Rachman, H. Santoso, and A. History, "Jurnal Teknologi dan Manajemen Informatika Perbandingan Model Deep Learning untuk Klasifikasi Sentiment Analysis dengan Teknik Natural Languange Processing Article Info ABSTRACT," *J. Teknol. dan Manaj. Inform.*, vol. 7, no. 2, pp. 103–112, 2021, [Online]. Available: http://http//jurnal.unmer.ac.id/index.php/jtmi

[18] Adhitya Karel Maulaya and Junadhi, "Analisis Sentimen Menggunakan Support Vector Machine Masyarakat Indonesia Di Twitter Terkait Bjorka," *J. CoSciTech (Computer Sci. Inf. Technol.*, vol. 3, no. 3, pp. 495–500, 2022, doi: 10.37859/coscitech.v3i3.4358.

[19] W. Musu, A. Ibrahim, and Heriadi, "Pengaruh Komposisi Data Training dan Testing terhadap Akurasi Algoritma C4 . 5," *Pros. Semin. Ilm. Sist. Inf. Dan Teknol. Inf.*, vol. X, no. 1, pp. 186–195, 2021.

[20] D. A. Nasution, H. H. Khotimah, and N. Chamidah, "Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma K-NN," *Comput. Eng. Sci. Syst. J.*, vol. 4, no. 1, p. 78, 2019, doi: 10.24114/cess.v4i1.11458.

[21] M. Azhari, Z. Situmorang, and R. Rosnelly, "Perbandingan Akurasi, Recall, dan Presisi Klasifikasi pada Algoritma C4.5, Random Forest, SVM dan Naive Bayes," *J. Media Inform. Budidarma*, vol. 5, no. 2, p. 640, 2021, doi: 10.30865/mib.v5i2.2937.

[22] B. G. Sudarsono, M. I. Leo, A. Santoso, and F. Hendrawan, "Analisis Data Mining Data Netflix Menggunakan Aplikasi Rapid Miner," *JBASE - J. Bus. Audit Inf. Syst.*, vol. 4, no. 1, pp. 13–21, 2021, doi: 10.30813/jbase.v4i1.2729.

[23] A. Riski, "Analisis Komparasi Algoritma Klasifikasi Data Mining Untuk Prediksi Penderita Penyakit Jantung," *J. Tek. Inform. Kaputama*, vol. 3, no. 1, pp. 22–28, 2019, [Online]. Available: https://jurnal.kaputama.ac.id/index.php/JTIK/article/view/141/156

[24] A. H. Nasrullah, "Implementasi Algoritma Decision Tree Untuk Klasifikasi Data Peserta Didik," *J. Pilar Nusa Mandiri*, vol. 7, no. 2, p. 217, 2021.

[25] J. Pangaribuan, C. Tedja, and S. Wibowo, "Perbandingan Metode Algoritma C4.5 dan Extreme Learning Machine untuk Mendiagnosis Penyakit Jantng Korner," *Informatics Eng. Res. Technol.*, vol. 1, no. 1, pp. 1–7, 2019.

[26] A. Rohman and A. Rufiyanto, "Implementasi Data Mining Dengan Algoritma Decision Tree C4 . 5 Untuk Prediksi Kelulusan Mahasiswa Di Universitas Pandaran," *Proceeding SINTAK 2019*, pp. 134–139, 2019, [Online]. Available: https://www.unisbank.ac.id/ojs/index.php/sintak/article/view/7577

[27] A. I. Putri et al., "Implementation of K-Nearest Neighbors, Naïve Bayes Classifier, Support Vector Machine and Decision Tree Algorithms for Obesity Risk Prediction," Public Research Journal of Engineering, Data Technology and Computer Science, vol. 2, no. 1, pp. 26–33, Apr. 2024, doi: 10.57152/predatecs.v2i1.1110.

[28] M. R. Anugrah, N. A. Al-Qadr, N. Nazira, and N. Ihza, "Implementation of C4.5 and Support Vector Machine (SVM) Algorithm for Classification of Coronary Heart Disease," Public Research Journal of Engineering, Data Technology and Computer Science, vol. 1, no. 1, pp. 20–25, Jul. 2023, doi: 10.57152/predatecs.v1i1.805.