JUSTEK : JURNAL SAINS DAN TEKNOLOGI
http://journal.ummat.ac.id/index.php/justek
ISSN 2620-5475
Vol. 8, No. 3, September 2025, Hal. 232-240

Klasifikasi Penyakit Kanker Paru-Paru Menggunakan Metode Decision Tree C4.5

[1]Khorlis Jainudin, [2*]Asrul Abdullah, [3]Sucipto
[1]Teknik Informatika, Universitas Muhammadiyah Pontinak, Indonesia
Email: khorlisjainudin346@gmail.com, asrul.abdullah@unmuhpnk.ac.id, sucipto@unmuhpnk.ac.id

ARTICLE INFO

Article History:

Diterima : 03-06-2025
Disetujui : 04-09-2025

Keywords:
Classification
Confusion Matrix
Data Mining
Decision Tree C4.5
Lung Cancer

ABSTRACT

Abstract: *The incidence of lung cancer in Indonesia has shown a significant increase, positioning the country as the eighth highest in Southeast Asia, with a growth rate of 10.85% over the past five years. A considerable number of lung cancer cases remain undiagnosed at earlier stages due to difficulties in detection, which contributes to the high mortality rate associated with this disease. Consequently, there is a need for a relatively efficient and straightforward technique to uncover knowledge, patterns, and interrelationships among data. The objective of this study is to develop a classification model for lung cancer using the C4.5 decision tree method and to evaluate its predictive performance. The methodology comprises several stages, including data preprocessing, exploratory data analysis (EDA), handling of missing values, identification of duplicate records, assessment of feature correlations, separation of features and target variables, partitioning of data into training and testing sets, model implementation, and performance evaluation through a confusion matrix. The experimental results demonstrate that the proposed model achieves a recall of 90%, a precision of 86%, an F1-score of 88%, and an overall accuracy of 89%. These findings indicate that the C4.5 decision tree method is effective in classifying lung cancer cases and holds potential as a reliable approach in medical data analysis for early detection and diagnosis.*

Abstrak: Angka kejadian kanker paru-paru di Indonesia setiap harinya meningkat pesat hingga menduduki peringkat ke-8 di Asia Tenggara dan meningkat sebesar 10,85 persen dalam lima tahun terakhir. Banyak kasus kanker paru-paru yang tidak dapat ditemukan lebih awal karena susah terdeteksi sehingga mengakibatkan tingginya tingkat kematian yang disebabkan karena kanker paru-paru. Untuk itu dibutuhkan suatu teknik yang relatif cepat dan mudah untuk menemukan pengetahuan, pola dan relasi antar data. Tujuan Penelitian ini adalah merancang model klasifikasi kanker paru-paru menggunakan metode *decision tree C4.5* serta mengetahui akurasi yang dihasilkan. Metode ini melalui beberapa tahapan, antara lain mulai *dari input data*, EDA, menangani *missing value,* mengecek *duplicate data*, mengecek korelasi antar fitur, pemisahan fitur dan target, membagi data latih dan data uji, implementasi metode, dan evaluasi model menggunakan *confusion matrix.* Hasil confusion matrix pada recall 90%, presisi 86% dan F1-score 88%, dan akurasi 89% yang menunjukkan bahwa metode decision tree c4.5 baik dalam melakukan klasifikasi penyakit kanker paru-paru.

https://doi.org/10.31764/justek.vXiY.ZZZ

A. LATAR BELAKANG

Dalam era digital, data menjadi salah satu aset terpenting bagi organisasi dan individu. Data adalah sumber informasi yang digunakan untuk membuat keputusan penting, menghasilkan nilai tambah bagi perusahaan, dan mendorong inovasi (Suryawijaya, 2023). Data adalah fakta mengenai objek. Data dinyatakan dengan nilai (angka, deretan karakter, atau simbol). Menurut pendapat lainnya, data adalah fakta tentang sesuatu di dunia nyata yang dapat direkam dan disimpan pada media komputer (Fatimatuz Zahroh & Nur Rahmawati, 2024).

Kemajuan pesat dalam bidang penambangan data (*data mining)* tak lepas dari perkembangan ilmu pengetahuan dan teknologi yang terus dikembangkan. Data mining merupakan suatu proses yang bertujuan untuk menemukan hubungan, pola, dan kecenderungan yang signifikan dalam sekumpulan data. Proses ini melibatkan penggunaan teknik pengenalan pola, seperti teknik statistik dan matematika. (Anugrah Pratama et al., 2023).

Penerapan data mining dalam bidang kesehatan telah mengalami perkembangan pesat dalam beberapa tahun terakhir. Teknologi ini telah terbukti efektif dalam berbagai aplikasi medis, seperti diagnosis penyakit, prediksi risiko kesehatan, analisis citra medis, dan pengoptimalan protokol pengobatan. Data mining telah terbukti efektif dalam mengidentifikasi pola-pola tersembunyi dalam data medis yang kompleks, yang seringkali sulit dideteksi melalui metode konvensional (Lestari & Homaidi, 2024). Penggunaan data mining untuk membantu riset dan penanganan kanker saat ini juga sangat meningkat (Marzuq et al., 2023).

Kanker telah menjadi penyebab paling umum kematian manusia dalam beberapa tahun terakhir. Menurut WHO, kanker menyebabkan 9,6 juta kematian di seluruh dunia pada tahun 2018. Angka kejadian kanker paru-paru di Indonesia setiap harinya meningkat pesat hingga menduduki peringkat ke-8 di Asia Tenggara dan meningkat sebesar 10,85 persen dalam lima tahun terakhir. (Jatnika Fahmi Idris et al., 2024). Merokok adalah penyebab utama kanker paru-paru, namun perokok pasif kecil kemungkinannya berisiko terkena kanker paru-paru (Aktalina, 2022).

Banyak kasus kanker paru-paru yang tidak dapat ditemukan lebih awal karena susah terdeteksi sehingga mengakibatkan tingginya tingkat kematian yang disebabkan karena kanker paru-paru. Guna membantu diagnosa awal pada penyakit kanker paru-paru, proses klasifikasi dapat membantu pola kanker paru-paru ditemukan. Salah satu metode yang dapat digunakan untuk melakukan klasifikasi adalah Decision Tree C4.5 (Widya et al., 2023)

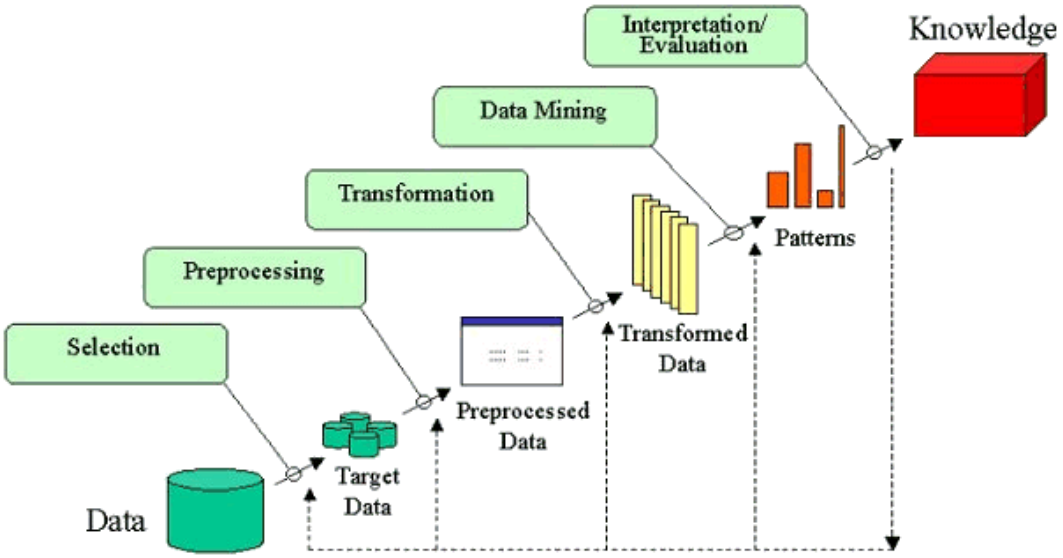
Beberapa studi telah menunjukkan bahwa *Decision Tree C4.5* memiliki akurasi yang baik dalam melakukan klasifikasi. Studi (Nasrullah, 2021) yang telah dilakukan untuk data penjualan dan stok barang di UD. Cipta Karya Gorontalo maka diperoleh akurasi sebesar 90% dan nilai AUC 0.709. Dimana nilai ini termasuk dalam Fair Classification (Klasifiksi yang cukup). Sehingga dapat disimpulkan bahwa model klasifikasi data mining menggunakan Algoritma Decision Tree C4.5 akurat dalam melakukan klasifikasi untuk produk laris. Studi lain yang dilakukan oleh (Taufik & Jatmika, 2021) dapat menghasilkan informasi berupa klasifikasi keberhasilan pengiriman barang dimana dari

data training yang digunakan dengan jumlah 100 dapat dibangun sebuah decision tree yang menghasilkan 9 rules. Berdasarkan hasil pengujian dengan algoritma C4.5 menggunakan tools Rapid Miner yang diukur tingkat akurasinya menggunakan pengujian confusion matrix dan kurva ROC didapat nilai akurasi sebesar 93 persen dan menghasilkan nilai AUC (Area Under Curve) sebesar 0.739 dengan nilai akurasi klasifikasi Cukup (Fair classification). Data mining dengan algoritma C4.5 dapat diimplementasikan untuk mengklasifikasian keberhasilan dan kegagalan pengiriman. Hal ini dapat menjadi rekomendasi bagian distribusi dalam memilih jasa logistik untuk pengiriman barang berdasarkan alamat pengiriman, dan kategori barang agar resiko kegagalan pengiriman bisa dikurangi. Studi lain yang dilakukan oleh (Sartika & Yupianti, 2020) didapat informasi bahwa sistem klasifikasi penyakit tiroid mengguna kan algoritma C4.5 yang dibuat telah sesuai dengan tahapan atau urutan proses yang se mestinya sehingga dengan adanya sistem ini, maka RSUD Hasa nuddin Damrah Manna dapat terbantu dalam melakukan diagnosa penyakit yang disebab kan oleh kelenjar tiroid dan memperfmudah menampilkan informasi diag nosa penyakit pasien berdasarkan gejala yang dialami atau yang telah dipilih di dalam system. Kemudian studi yang dilakukan oleh (Fitriani et al., 2020) dapat diketahui metode terbaik dalam klasifikasi kelayakan marketing. Untuk mengukur kinerja model digunakan confusion matrix dan kurva ROC, dan diketahui bahwa dapat disimpulkan bahwa nilai akurasi yang didapat model algoritma C4.5 adalah 91,10%. Studi yang dilakukan oleh (Kresimo Negoro et al., 2022). Beberapa penelitian terkait tersebut dengan berbagai objek klasifikasi yang berbeda tidak menggunakan metrik yang sama dengan jumlah data bervariasi dan cenderung terlalu sedikit sehingga data awal penulis pada penelitian memiliki originalitasnya sendiri dan dapat diharapkan memberikan hasil prediksi yang lebih baik dari penelitian-penelitian sebelumnya.

Gap analisis penelitian ini dengan penelitian metode *Decision Tree C4.5* lainnya terletak pada penerapan metode *Decision Tree C4.5* dalam konteks klasifikasi penyakit kanker paru-aru. Diharapkan melalui penelitian ini, dapat lebih mendorong semua pihak terkait dalam memahami dan memanfaatkan potensi data dalam melakukan diagnosa awal penyakit kanker paru-paru. Penelitian ini bertujuan untuk merancang model klasifikasi penyakit kanker paru-paru. Hasil penelitian ini diharapkan dapat menjadi pengetahuan dalam melihat pola-pola tanda munculnya penyakit kanker paru-paru.

B. METODE PENELITIAN

Metode penelitian merupakan cara ilmiah yang dilakukan untuk mendapatkan data dengan tujuan tertentu. Cara ilmiah berarti kegiatan yang dilandasi dengan metode keilmuan. Dalam hal ini melakukan kombinasi pendekatan rasional yang mengedepankan teoritik dengan pendekatan empiris yang membuktikan pembuktian di lapangan. Metode Penelitian ini menggunakan metode *knowledge discovery in databases (KDD)*. Knowledge Discovery in Databases (KDD) adalah penerapan metode scientific pada data mining (Petra Valentino & Siska Narulita, 2023).



Gambar 1 Tahapan KDD

Tahapan *Knowledge Discovery in Databases* (KDD) sebagai berikut

1) Data Selection

Data yang ada akan dilakukan seleksi data dan atribut yang akan digunakan untuk proses selanjutnya. Seleksi data dari sekumpulan data operasional dilakukan sebelum tahap penggalian informasi. Data hasil seleksi akan digunakan untuk proses data mining, dan disimpan dalam berkas yang terpisah dari data operasional (Naldy & Andri, 2021)

2) Preprocessing

Sebelum proses data mining dapat dilaksanakan, perlu dilakukan proses cleaning pada data yang menjadi fokus Knowledge Discovery in Database (KDD). Proses cleaning mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak. Juga dilakukan proses enrichment, yaitu proses "memperkaya" data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk Knowledge Discovery in Database (KDD), seperti data atau informasi eksternal lainnya yang diperlukan (Mardi, 2017)

3) Transformation

Tahap transformasi data yaitu tahapan mengubah tipe data agar data dapat diolah sesuai kebutuhan sehingga data siap untuk diproses [21]. Pada tahap ini dilakukan transformasi dari tipe data non-numerik menjadi data numerik (Gustipartsani et al., 2023).

4) Data Mining

Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu (Hikmah et al., 2019)

5) Interpretation/ Evaluation

Dari proses KDD yang sudah dilakukan sebelumnya, penulis menginterpretasi dengan membuat visualisasi hasil pengolahan data (Wahidah et al., 2022)

C. HASIL DAN PEMBAHASAN

Penelitian ini dilakukan melalui tahapan-tahapan yang meliputi pengumpulan data, pemodelan menggunakan metode decision tree c4.5, evaluasi model dan interpretasi hasil.

1. Pengumpulan Data

Data yang digunakan pada penelitian ini merupakan data yang bersumber dari *Kaggle*. Tabel 1 menyajikan beberapa data pada dataset yang digunakan dalam penelitian ini.

Tabel 1. Dataset

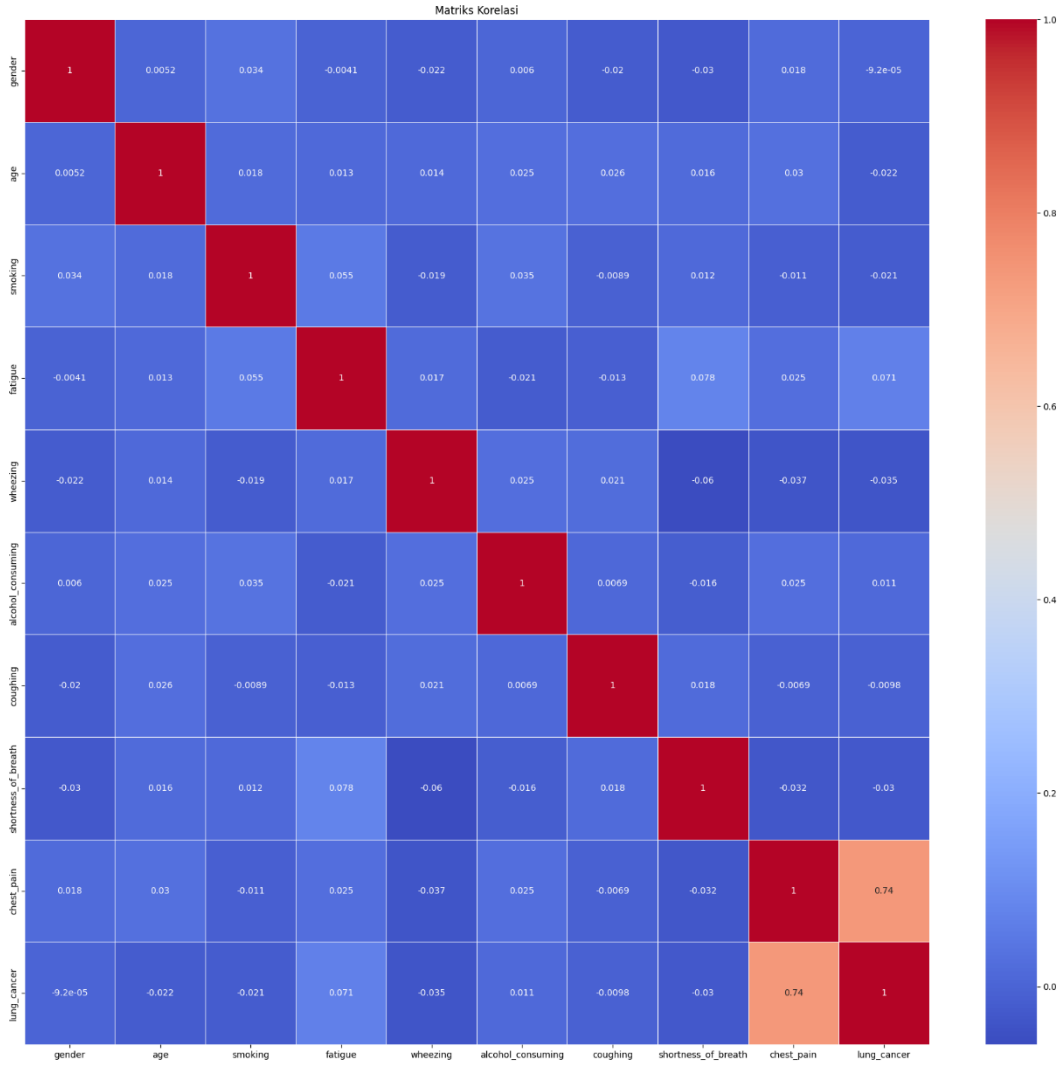
No.	Atribut	Keterangan	Jenis Data
1.	Age	Usia Pasien	Numerik
2.	Gender	Jenis Kelamin (Laki-Laki/ Perempuan)	Ordinal
3.	Smoking	Merokok (Ya/Tidak)	Kategorikal
4.	Chest Pain	Nyeri Dada (Ya/Tidak)	Kategorikal
5.	Coughing	Batuk (Ya/Tidak)	Kategorikal
6.	Fatigue	Kelelahan (Ya/Tidak)	Kategorikal
7.	Shortness Breath	Sesak Napas (Ya/Tidak)	Kategorikal
8.	Wheezing	Mengi (Ya/Tidak)	Kategorikal
9.	Lung Cancer	Kanker Paru-Paru (Berisiko/Tidak Berisiko)	Kategorikal

2. Pemodelan Decision Tree C4.5

Setelah dataset telah selesai disiapkan, selanjutnya dilakukan pemodelan dengan menggunakan algoritma Decision Tree C4.5. Algoritma Decision Tree C4.5 memiliki beberapa perubahan dan peningkatan yang membuatnya lebih efektif dalam menangani masalah yang lebih kompleks (Nazifah & Prianto, 2023).

3. *Correlation Matrix*

Correlation matrix adalah metrik digunakan untuk mengecek korelasi antara setiap fitur pada *dataset.*



4. Hasil Pengujian

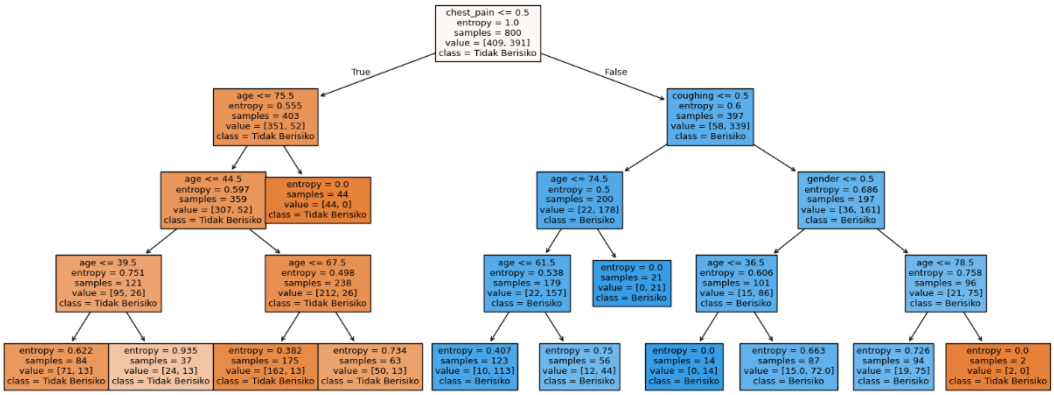
Hasil pengujian performa pada model klasifikasi penyakit kanker paru-paru menggunakan *confusion matrix.* Pengujian memiliki 4 metrik evaluasi, yaitu akurasi, presisi, sensitivitas, dan *F-1 score*. Hasil pengujian dapat dilihat pada Tabel 5.1

Tabel 2 Metrik Evaluasi

No.	Metrik Evaluasi (Entropy)	Hasil (%)
1.	Akurasi (*Accuracy*)	89%
2.	Presisi (*Precision*)	86%
3.	Sensitivitas (*Recall*)	90%
4.	*F1-Score*	88%

5. Hasil Pohon Keputusan

Hasil pohon memberikan informasi mengenai pola aturan dalam menentukan klasifikasi pada data. Hasil pohon keputusan dapat dilihat pada gambar 5.9.



Gambar 2 Pohon Keputusan

Berdasarkan pohon keputusan (*decision tree)* untuk klasifikasi penyakit kanker paru-paru, dapat dilihat bahwa atribut yang paling berpengaruh dalam proses klasifikasi adalah atribut *chest_pain*. Dari pohon keputusan pada gambar 5.10 didapatkan aturan (rule) untuk klasifikasi penyakit kanker paru-paru. aturan yang didapat sebagai berikut:

D. SIMPULAN DAN SARAN

Dari hasil penelitian dapat ditarik Kesimpulan yaitu penelitian yang dilakukan melalui beberapa tahapan mulai dari input data, EDA, menangani missing value, mengecek duplicate data, mengecek korelasi antar fitur, pemisahan fitur dan target, membagi data latih dan data uji, implementasi metode, dan evaluasi model. Hasil dari pengujian metode decision tree C4.5 serta tahapan model dievaluasi untuk mengetahui pengaruh setiap tahapan terhadap kemampuan model dalam melakukan klasifikasi. Metrik evaluasi yang digunakan berupa hasil dari confusion matrix. Hasil confusion matrix pada recall 90%, presisi 86% dan F1-score 88%, dan akurasi 89% yang menunjukkan bahwa metode *decision tree c4.5* baik dalam melakukan klasifikasi penyakit kanker paru-paru.

Penelitian selanjutnya dapat menggunakan *tunning hyperparameter* untuk meningkatkan performa model klasifikasi.

REFERENSI

Aktalina, L. (2022). Edukasi tentang Merokok sebagai Faktor Risiko Terjadinya Kanker Paru pada Masyarakat di Kelurahan Pelawi Utara Kecamatan Babalan Kabupaten Langkat. *Jurnal Pengabdian Masyarakat*, *1*(2), 1–8.

Anugrah Pratama, D., Rizal Mutaqin, I., & Rafael Manuela, K. (2023). Analisis Terjadinya Kanker Paru-Paru Pada Pasien Menggunakan Decision Tree: Penerapan Algoritma C4.5 Dan RapidMiner Untuk Menentukan Risiko Kanker Pada Gejala Pasien. *Jtmei*, *2*(4), 156–170. https://doi.org/10.55606/jtmei.v2i4.3004

Fatimatuz Zahroh, & Nur Rahmawati. (2024). Analisis Sistem Pengolahan Data Berbasis Web pada Badan Pusat Statistik Kabupaten X. *Manufaktur: Publikasi Sub Rumpun Ilmu Keteknikan Industri*, *2*(1), 01–13. https://doi.org/10.61132/manufaktur.v2i1.119

Fitriani, E., Aryanti, R., Saepudin, A., & Ardiansyah, D. (2020). Penerapan Algoritma C4.5

Untuk Klasifikasi Penempatan Tenaga Marketing. *Paradigma - Jurnal Komputer Dan Informatika*, *22*(1), 72–78. https://doi.org/10.31294/p.v22i1.6898

Gustipartsani, K., Rahaningsih, N., Dana, R. D., Mustafa, I. Y., Studi, P., Informatika, T., Studi, P., Akuntansi, K., Studi, P., Informatika, M., Perhotelan, P. S., Pariwisata, P., Internasional, P., Cirebon, K., & Barat, J. (2023). *Data Mining Clustering Menggunakan Algoritma K-Means Pada*. *7*(6), 3595–3601.

Hikmah, N., Ariyanti, D., & Sugesti, M. (2019). Penerapan Teknik Data Mining untuk Clustering Armada pada PT. Siaga Transport Indonesia Menggunakan Metode k-Means. *Explore*, *9*(1), 8. https://doi.org/10.35200/explore.v9i1.116

Jatnika Fahmi Idris, Rafid Ramadhani, & Muhammad Malik Mutoffar. (2024). Klasifikasi Penyakit Kanker Paru Menggunakan Perbandingan Algoritma Machine Learning. *Jurnal Media Akademik (JMA)*, *2*(2). https://doi.org/10.62281/v2i2.145

Kresimo Negoro, N., diana, M., Izul Ula, M., & Dwi Insani, F. (2022). Analisis Kebakaran pada Hutan dan Lokasi Lahan di Provinsi Riau Menggunakan Metode C4.5. *Jurnal Informatika Universitas Pamulang*, *7*(1), 107–114. http://openjournal.unpam.ac.id/index.php/informatika

Lestari, I. I., & Homaidi, A. (2024). *Gudang Jurnal Multidisiplin Ilmu Komparasi Algoritma Naive Bayes Dan Random Forest Pada Klasifikasi Kanker Payudara*. *2*, 778–785.

Mardi, Y. (2017). Data Mining : Klasifikasi Menggunakan Algoritma C4.5. *Edik Informatika*, *2*(2), 213–219. https://doi.org/10.22202/ei.2016.v2i2.1465

Marzuq, R. D., Wicaksono, S. A., & Setiawan, N. Y. (2023). Prediksi Kanker Paru-Paru menggunakan Algoritme Random Forest Decision Tree. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *7*(7), 3448–3456.

Naldy, E. T., & Andri, A. (2021). Penerapan Data Mining Untuk Analisis Daftar Pembelian Konsumen Dengan Menggunakan Algoritma Apriori Pada Transaksi Penjualan Toko Bangunan MDN. *Jurnal Nasional Ilmu Komputer*, *2*(2), 89–101. https://doi.org/10.47747/jurnalnik.v2i2.525

Nasrullah, A. H. (2021). Implementasi Algoritma Decision Tree Untuk Klasifikasi Produk Laris. *Jurnal Ilmiah Ilmu Komputer*, *7*(2), 45–51. https://doi.org/10.35329/jiik.v7i2.203

Nazifah, N., & Prianto, C. (2023). Decision Tree Algoritma C4.5 dengan algoritma lainnya: Sistematic Literature Review. *Jurnal Informatika Dan Teknologi Komputer*, *04*(https://ejurnalunsam.id/index.php/jicom/), 57–64. https://ejurnalunsam.id/index.php/jicom/

Petra Valentino, & Siska Narulita. (2023). Performansi Algoritma Decision Tree (C4.5) untuk Prediksi Penyakit Jantung. *Jurnal Cakrawala Informasi*, *3*(2), 18–24. https://doi.org/10.54066/jci.v3i2.349

Sartika, D., & Yupianti. (2020). Klasifikasi Penyakit Tiroid Menggunakan Algoritma C4 . 5. *Journal of Science and Technology*, *13*(1), 71–76.

Suryawijaya, T. W. E. (2023). Memperkuat Keamanan Data melalui Teknologi Blockchain: Mengeksplorasi Implementasi Sukses dalam Transformasi Digital di Indonesia. *Jurnal Studi Kebijakan Publik*, *2*(1), 55–68. https://doi.org/10.21787/jskp.2.2023.55-68

Taufik, G., & Jatmika, D. (2021). *Penerapan Algoritma C4 . 5 Untuk Klasifikasi*. 12–26.

Wahidah, A. R., Bachtiar, Y., & Wulan, R. (2022). Sistem Pendukung Analisa Key Performance Indicator (KPI) Menggunakan Metode Data Mining Berbasis Web Python Programming. *JRKT (Jurnal Rekayasa Komputasi Terapan)*, *2*(03), 151–158. https://doi.org/10.30998/jrkt.v2i03.7971

Widya, H., Surya Putra, N., Atina, V., & Maulindar, J. (2023). Penerapan Algoritme Decision Tree Pada Klasifikasi Penyakit Kanker Paru-Paru. *Jurnal Ilmiah Teknik Informatika Dan Sistem Informasi*. https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer,