KAMPUS AKADEMIK PUBLISHER
Jurnal Ilmiah Penelitian Mahasiswa
Vol.3, No.5 Oktober 2025
e-ISSN: 3025-5465; p-ISSN: 3025-7964, Hal 646-653
DOI: https://doi.org/10.61722/jipm.v3i5.1513

Penerapan Data Mining Menggunakan Algoritma C4.5 untuk Klasifikasi Penyakit Paru-Paru

Zaehol Fatah
Universitas Ibrahimy
Muhammad Faidhurrahman Wahid
Universitas Ibrahimy
Jl. KHR. Syamsul Arifin No.1-2, Sukorejo, Sumberejo, Kec. Banyuputih, Kabupaten Situbondo, Jawa Timur 68374, Indonesia
Korespondensi penulis: faidgombal.@gmail.com

Abstract. *Lung disease remains one of the leading causes of morbidity and mortality in Indonesia. Factors such as smoking habits, unhealthy lifestyles, and low awareness of respiratory health have significantly contributed to the increasing prevalence of this condition. This study aims to classify lung health conditions using the Decision Tree C4.5 algorithm, a data mining technique widely applied in medical analysis. The dataset consists of 132 respondents with various attributes, including age, gender, smoking habits, sleep patterns, and medical history. The model was validated using the percentage split method, A total of 80% of the data was allocated for training the model, while the remaining 20% was used for testing its performance.The results show that the C4.5 algorithm successfully classified lung disease risk into two categories, "Yes" (at risk) and "No" (not at risk), with a high level of accuracy. The model effectively identified the key factors contributing to lung disease risk, such as smoking habits, late-night activity, age, and insurance ownership. These findings confirm that the Decision Tree C4.5 algorithm is a reliable and efficient tool for the early detection of respiratory diseases and can support data-driven decision-making in the healthcare field.*
Keywords: *lung disease, classification, decision tree, data mining, C4.5.*

Abstrak. Penyakit paru-paru masih menjadi salah satu penyebab utama morbiditas dan mortalitas di Indonesia. Faktor-faktor seperti kebiasaan merokok, gaya hidup tidak sehat, serta rendahnya kesadaran terhadap kesehatan pernapasan telah secara signifikan meningkatkan prevalensi penyakit ini. Penelitian ini bertujuan untuk mengklasifikasikan kondisi kesehatan paru-paru menggunakan algoritma Decision Tree C4.5, yang salah satu metode dalam data mining yang memiliki penerapan luas di bidang analisis kesehatan. Dataset penelitian terdiri dari 132 responden dengan berbagai atribut, termasuk usia, jenis kelamin, kebiasaan merokok, pola tidur, dan riwayat kesehatan. Model penelitian divalidasi menggunakan metode percentage split dengan pembagian 80% data sebagai data latih dan 20% data sebagai data uji.
Dari hasil penelitian diperoleh bahwa algoritma C4.5 efektif dalam memprediksi risiko penyakit paru-paru dengan membagi data ke dalam dua kategori, yaitu "Ya" (berisiko) dan "Tidak" (tidak berisiko). Model yang digunakan dapat menentukan faktor penentu risiko seperti kebiasaan merokok, waktu tidur, usia, serta kepemilikan asuransi. Hal ini menunjukkan bahwa metode Decision Tree C4.5 memiliki potensi besar sebagai sarana pendukung deteksi dini penyakit paru-paru dan pengambilan keputusan medis berbasis data.
Kata kunci: penyakit paru-paru, data mining, klasifikasi, decision tree, C4.5.

LATAR BELAKANG

Penyakit paru-paru masih menjadi penyebab utama kematian di Indonesia, terutama akibat gaya hidup tidak sehat seperti merokok dan kurang berolahraga (Putri, 2023). Dalam penelitian Meiyanti (2020), algoritma C4.5 terbukti mampu mengklasifikasikan data pasien dengan akurasi tinggi dalam mendeteksi penyakit paru-paru. Selanjutnya, Pambudi (2024) mengungkapkan bahwa Decision Tree merupakan metode yang efisien karena mudah diinterpretasikan oleh tenaga medis dan mampu menangani data kategorikal maupun numerik.

Received Agustus 30, 2025; Revised September 30, 2025; Oktober 18, 2025
*** Muhammad Faidhurrahman Wahid , faidgombal.@gmail.com**

Penerapan algoritma C4.5 menggunakan perangkat lunak seperti RapidMiner juga dilakukan oleh Sofyan (2023), yang menunjukkan hasil klasifikasi akurat terhadap data medis. Sementara itu, Budiyono (2024) menegaskan bahwa teknik klasifikasi seperti C4.5 memberikan hasil lebih baik dibandingkan algoritma probabilistik sederhana seperti Naïve Bayes.

Christian dan Sumanto (2025) berpendapat bahwa integrasi teknik pembelajaran mesin dalam sistem kesehatan mampu meningkatkan efektivitas diagnosis dini. Di sisi lain, Kurniawan (2024) menunjukkan bahwa metode Random Forest dapat mendeteksi penyakit paru-paru dengan akurasi lebih tinggi, meskipun interpretasinya lebih kompleks.

Selain itu, Putri (2023) menemukan adanya hubungan signifikan antara kebiasaan merokok dan tingkat keparahan penyakit paru obstruktif kronis (PPOK). Hal ini menegaskan pentingnya analisis faktor gaya hidup dalam penelitian penyakit paru-paru. Terakhir, Sholiha (2025) menambahkan bahwa kombinasi Decision Tree dengan teknik ensemble learning dapat meningkatkan akurasi klasifikasi pada kasus penyakit paru-paru.

Oleh karena itu, penelitian ini menitikberatkan pada pemanfaatan algoritma Decision Tree C4.5 dalam mengelompokkan penyakit paru-paru menggunakan data sosial dan gaya hidup pasien sebagai variabel analisis.

KAJIAN TEORITIS

Pambudi (2024) menyatakan bahwa data mining adalah suatu pendekatan analitis yang bertujuan untuk mengekstraksi pola serta keterkaitan yang tersembunyi dari kumpulan data berukuran besar melalui penerapan metode statistik dan teknik pembelajaran mesin. Dalam konteks kesehatan, penerapan data mining berperan penting dalam proses diagnosis maupun prediksi penyakit melalui analisis data pasien.

Salah satu algoritma yang umum diterapkan dalam klasifikasi medis ialah Decision Tree, khususnya varian C4.5, karena mampu menghasilkan model berbentuk aturan yang mudah diinterpretasikan (Meiyanti, 2020). Secara konseptual, algoritma C4.5 bekerja dengan memilih atribut yang memiliki nilai gain ratio tertinggi untuk dijadikan simpul utama dalam pembentukan struktur pohon keputusan (Sofyan, 2023).

Menurut Budiyono (2024), algoritma C4.5 unggul karena dapat mengolah data bertipe numerik maupun kategorikal secara bersamaan. Sementara itu, Kurniawan (2024) menambahkan bahwa model Decision Tree dapat dikembangkan lebih lanjut melalui pendekatan ensemble seperti Random Forest guna meningkatkan akurasi hasil klasifikasi.

Dalam penelitian *Christian dan Sumanto (2025)*, algoritma *machine learning* digunakan untuk prediksi penyakit paru-paru dengan hasil yang menunjukkan peningkatan akurasi

signifikan. Pendekatan serupa dilakukan oleh *Sholiha (2025)* dengan menggunakan kombinasi Decision Tree dan *AdaBoost* untuk meningkatkan stabilitas hasil klasifikasi.

METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif prediktif dengan jenis penelitian terapan. Data diperoleh dari dataset sekunder yang bersumber dari platform Kaggle, yang berisi informasi medis dan sosial pasien. Dataset tersebut mencakup atribut seperti usia, jenis kelamin, kebiasaan merokok, aktivitas begadang, status pernikahan, dan riwayat penyakit bawaan. Data ini dipilih karena memiliki struktur yang sesuai untuk penerapan algoritma klasifikasi berbasis *Decision Tree*. Proses analisis dilakukan menggunakan perangkat lunak RapidMiner dengan metode *percentage split*, di mana 80% data digunakan untuk pelatihan dan 20% untuk pengujian (*Sofyan, 2023*).

Jenis dan Desain Penelitian

Desain penelitian berupa klasifikasi berbasis *supervised learning* menggunakan algoritma Decision Tree C4.5 (*Meiyanti, 2020*). Dataset dibagi menjadi dua bagian menggunakan metode *percentage split* dengan 80% data latih dan 20% data uji (*Sofyan, 2023*).

Waktu dan Tempat

Proses pengumpulan dan pengolahan data dilakukan secara daring pada Mei hingga Juni 2025. Data diperoleh dari platform Kaggle yang menyediakan dataset medis terkait penyakit paru-paru. Proses klasifikasi dilakukan menggunakan perangkat lunak RapidMiner Studio yang memiliki antarmuka visual untuk pemodelan, analisis, dan evaluasi dalam penerapan teknik *data mining* (Nahjan, Heryana, & Voutama, 2023).

Variabel Penelitian

- **Variabel bebas**: usia, jenis kelamin, kebiasaan merokok, status bekerja, status rumah tangga, aktivitas begadang, aktivitas olahraga, kepemilikan asuransi, dan riwayat penyakit bawaan.
- **Variabel terikat**: hasil klasifikasi penyakit paru-paru (Ya atau Tidak).

Teknik Pengumpulan dan Pengelolahan Data

Data penelitian diperoleh melalui studi pustaka dan pemanfaatan data sekunder dari Kaggle. Dataset mencakup 566 pasien dengan atribut demografi, gaya hidup, kondisi lingkungan, riwayat medis, gejala, hasil pemeriksaan klinis, hingga diagnosis akhir. Tahap pra-pemrosesan dilakukan meliputi pembersihan data, pengkodean atribut numerik dan kategorikal, serta normalisasi nilai. Dataset dibagi menjadi 80% data latih dan 20% data uji menggunakan metode

percentage split (Witten, Frank, & Hall, 2011). Langkah ini dilakukan untuk memastikan kualitas data optimal dalam klasifikasi dan analisis.

# No	Usia	Jenis_Kela...	Merokok	Bekerja	Rumah_Ta...	Aktivitas_...	Aktivitas_...	Asuransi	Penyakit_B...
1	Tua	Pria	Pasif	Tidak	Ya	Ya	Sering	Ada	Tidak
2	Tua	Pria	Aktif	Tidak	Ya	Ya	Jarang	Ada	Ada
3	Muda	Pria	Aktif	Tidak	Ya	Ya	Jarang	Ada	Tidak
4	Tua	Pria	Aktif	Ya	Tidak	Tidak	Jarang	Ada	Ada
5	Muda	Wanita	Pasif	Ya	Tidak	Tidak	Sering	Tidak	Ada
6	Muda	Wanita	Pasif	Ya	Tidak	Tidak	Sering	Tidak	Ada
7	Tua	Wanita	Pasif	Tidak	Ya	Tidak	Sering	Tidak	Tidak
8	Muda	Pria	Aktif	Tidak	Ya	Ya	Sering	Tidak	Tidak
9	Tua	Wanita	Aktif	Ya	Ya	Ya	Jarang	Ada	Ada
10	Muda	Wanita	Pasif	Ya	Tidak	Ya	Jarang	Ada	Ada
11	Tua	Wanita	Pasif	Ya	Ya	Tidak	Sering	Ada	Ada
12	Tua	Wanita	Aktif	Tidak	Ya	Tidak	Jarang	Ada	Tidak
13	Muda	Pria	Aktif	Tidak	Ya	Ya	Jarang	Ada	Tidak
14	Tua	Wanita	Aktif	Ya	Tidak	Ya	Jarang	Ada	Ada
15	Muda	Wanita	Pasif	Ya	Tidak	Ya	Sering	Tidak	Ada
16	Muda	Wanita	Pasif	Ya	Tidak	Ya	Jarang	Ada	Ada
17	Tua	Wanita	Pasif	Ya	Ya	Tidak	Sering	Ada	Ada
18	Tua	Wanita	Aktif	Tidak	Ya	Tidak	Jarang	Ada	Tidak
19	Muda	Pria	Aktif	Tidak	Ya	Ya	Jarang	Ada	Tidak

Gambar 1. Dataset

Algoritma Decision Tree C4.5

Algoritma Decision Tree C4.5 bekerja dengan memilih atribut terbaik berdasarkan nilai *gain ratio* untuk setiap cabang pohon keputusan. Atribut yang memberikan informasi tertinggi akan digunakan sebagai *root node*. Proses ini terus berlangsung hingga data terbagi secara optimal atau tidak ada atribut tersisa. Metode ini banyak digunakan dalam dunia medis karena mampu memberikan kejelasan dalam pengambilan keputusan serta mudah diinterpretasikan (Han, Kamber, & Pei, 2012; Sutoyo, 2018).

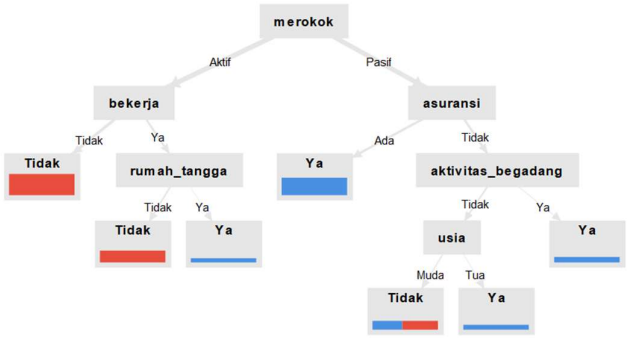
HASIL DAN PEMBAHASAN

Gambar di bawah ini menunjukkan proses analisis data pada kasus penyakit paru-paru menggunakan perangkat lunak analisis data. Tampilan tersebut menggambarkan alur kerja yang terdiri dari beberapa tahapan penting. Proses dimulai dengan membaca data medis pasien, seperti riwayat kesehatan, hasil laboratorium, dan gejala yang dialami, yang diolah menggunakan format data tertentu (misalnya, Microsoft Excel). Tahapan selanjutnya mencakup transformasi data, seperti mengonversi nilai numerik (misalnya, kadar oksigen atau jumlah sel darah) menjadi kategori nominal untuk memudahkan analisis. Setelah itu, model klasifikasi dibuat untuk mengidentifikasi jenis penyakit paru-paru, seperti pneumonia, asma, atau tuberkulosis. Koneksi antar tahapan menunjukkan alur data yang mengalir dari satu proses ke proses lainnya. Selain itu, komponen evaluasi seperti cross-validation digunakan untuk mengukur dan memastikan kinerja model dalam menganalisis dan mengklasifikasikan penyakit paru-paru secara akurat.

Row No.	hasil	no	usia	jenis_kelamin	merokok	bekerja	rumah_tangga	aktivitas_be...	aktivitas_ola...	asuransi	penyakit_ba...
1	Ya	1	Tua	Pria	Pasif	Tidak	Ya	Ya	Sering	Ada	Tidak
2	Tidak	2	Tua	Pria	Aktif	Tidak	Ya	Ya	Jarang	Ada	Ada
3	Tidak	3	Muda	Pria	Aktif	Tidak	Ya	Ya	Jarang	Ada	Tidak
4	Tidak	4	Tua	Pria	Aktif	Ya	Tidak	Tidak	Jarang	Ada	Ada
5	Ya	5	Muda	Wanita	Pasif	Ya	Tidak	Tidak	Sering	Tidak	Ada
6	Tidak	6	Muda	Wanita	Pasif	Ya	Tidak	Tidak	Sering	Tidak	Ada
7	Ya	7	Tua	Wanita	Pasif	Tidak	Ya	Tidak	Sering	Tidak	Tidak
8	Tidak	8	Muda	Pria	Aktif	Tidak	Ya	Ya	Sering	Tidak	Tidak
9	Ya	9	Tua	Wanita	Aktif	Ya	Ya	Ya	Jarang	Ada	Ada
10	Ya	10	Muda	Wanita	Pasif	Ya	Tidak	Ya	Jarang	Ada	Ada
11	Ya	11	Tua	Wanita	Pasif	Ya	Ya	Tidak	Sering	Ada	Ada
12	Tidak	12	Tua	Wanita	Aktif	Tidak	Ya	Tidak	Jarang	Ada	Tidak
13	Tidak	13	Muda	Pria	Aktif	Tidak	Ya	Ya	Jarang	Ada	Tidak
14	Tidak	14	Tua	Wanita	Aktif	Ya	Tidak	Ya	Jarang	Ada	Ada
15	Ya	15	Muda	Wanita	Pasif	Ya	Tidak	Ya	Sering	Tidak	Ada
16	Ya	16	Muda	Wanita	Pasif	Ya	Tidak	Ya	Jarang	Ada	Ada
17	Ya	17	Tua	Wanita	Pasif	Ya	Ya	Tidak	Sering	Ada	Ada
18	Tidak	18	Tua	Wanita	Aktif	Tidak	Ya	Tidak	Jarang	Ada	Tidak
19	Tidak	19	Muda	Pria	Aktif	Tidak	Ya	Ya	Jarang	Ada	Tidak

Gambar 2. Data Penyakit Paru-Paru

Berikut gambar di bawah ini menunjukkan hasil dari pemodelan menggunakan algoritma *decision tree* beserta persentase klasifikasi yang dihasilkan, yang mencerminkan tingkat akurasi dan distribusi keputusan berdasarkan atribut yang dianalisis.



Gambar 3. Pohon Keputusan untuk Penyakit Paru-Paru

Tree

```
merokok = Aktif
|   bekerja = Tidak: Tidak {Ya=0, Tidak=180}
|   bekerja = Ya
|   |   rumah_tangga = Tidak: Tidak {Ya=0, Tidak=97}
|   |   rumah_tangga = Ya: Ya {Ya=29, Tidak=0}
merokok = Pasif
|   asuransi = Ada: Ya {Ya=147, Tidak=0}
|   asuransi = Tidak
|   |   aktivitas_begadang = Tidak
|   |   |   usia = Muda: Tidak {Ya=32, Tidak=38}
|   |   |   usia = Tua: Ya {Ya=35, Tidak=0}
|   |   aktivitas_begadang = Ya: Ya {Ya=40, Tidak=0}
```

Gambar 4. Description

Tabel 1. Klasifikasi Tidak

Sumber Jalur	Jumlah Data
merokok = Aktif → bekerja = Tidak	180
merokok = Aktif → bekerja = Ya → rumah_tangga = Tidak	97
merokok = Pasif → asuransi = Tidak → aktivitas_begadang = Tidak → usia = Muda	38
Total Tidak	**315**

Tabel 2. Klasifikasi Iya

Sumber Jalur	Jumlah Data
merokok = Aktif → bekerja = Ya → rumah_tangga = Ya	29
merokok = Pasif → asuransi = Ada	147
merokok = Pasif → asuransi = Tidak → aktivitas_begadang = Tidak → usia = Tua	35
merokok = Pasif → asuransi = Tidak → aktivitas_begadang = Ya	40
Total Ya	**251**

Pohon keputusan ini melakukan klasifikasi terhadap dua kategori target, yaitu “Ya” dan “Tidak”, berdasarkan sejumlah atribut seperti merokok, bekerja, rumah_tangga, asuransi, aktivitas_begadang, dan usia. Proses klasifikasi dimulai dari atribut pada simpul akar (root node), yaitu merokok, kemudian menelusuri cabang ke bawah sesuai dengan nilai setiap atribut hingga mencapai simpul daun yang memuat hasil klasifikasi serta jumlah data yang memenuhi kondisi tersebut.

Berdasarkan hasil perhitungan, klasifikasi “Tidak” lebih mendominasi pada kelompok individu yang merokok secara aktif, terutama pada mereka yang tidak bekerja atau bekerja namun tidak berstatus rumah tangga. Sementara itu, klasifikasi “Ya” lebih sering muncul pada kelompok perokok pasif, khususnya individu yang memiliki asuransi, sering melakukan aktivitas begadang, serta berusia lebih tua. Temuan ini menunjukkan bahwa status perokok pasif, kepemilikan

asuransi, dan kebiasaan begadang memiliki korelasi yang signifikan terhadap kategori "Ya" dalam hasil klasifikasi.

Secara keseluruhan, hasil klasifikasi menunjukkan bahwa sebanyak 315 data termasuk dalam kategori "Tidak", sedangkan 251 data termasuk dalam kategori "Ya", dengan total keseluruhan 566 data. Perbedaan jumlah antara kedua kategori ini menggambarkan adanya kecenderungan dominan pada karakteristik tertentu dalam populasi data yang dianalisis.

KESIMPULAN DAN SARAN

Dari hasil penelitian diketahui bahwa penerapan algoritma C4.5 pada Decision Tree efektif dalam memprediksi kondisi kesehatan paru-paru. Model tersebut mengklasifikasikan responden menjadi dua kelompok, "Ya" (berisiko) dan "Tidak" (tidak berisiko), dengan mempertimbangkan variabel seperti perilaku merokok, jam tidur, usia, status pernikahan, dan kepemilikan asuransi.

Dari hasil analisis pohon keputusan, diperoleh bahwa kategori "Tidak" lebih banyak ditemukan pada individu yang merokok secara aktif, terutama pada mereka yang tidak bekerja atau belum berstatus rumah tangga. Sebaliknya, kategori "Ya" lebih sering muncul pada kelompok perokok pasif, khususnya yang memiliki asuransi, sering begadang, dan berusia lebih tua. Temuan ini menunjukkan bahwa kebiasaan begadang, usia lanjut, serta paparan asap rokok pasif merupakan faktor yang memiliki pengaruh signifikan terhadap peningkatan risiko penyakit paru-paru.

Secara keseluruhan, penerapan algoritma C4.5 terbukti efektif digunakan untuk deteksi dini penyakit pernapasan dan berpotensi dikembangkan sebagai sistem pendukung keputusan berbasis data di bidang kesehatan. Penelitian selanjutnya disarankan untuk menggunakan dataset yang lebih besar dan beragam, serta menambahkan variabel seperti riwayat paparan polusi, aktivitas olahraga, dan pola konsumsi harian agar hasil klasifikasi menjadi lebih akurat. Selain itu, model ini dapat dibandingkan dengan algoritma lain seperti Random Forest, Naïve Bayes, atau Support Vector Machine guna memperoleh performa klasifikasi yang lebih optimal. Implementasi hasil penelitian ini juga dapat diarahkan pada pengembangan aplikasi kesehatan berbasis teknologi untuk membantu tenaga medis maupun masyarakat dalam mendeteksi risiko penyakit paru-paru secara dini, disertai edukasi mengenai pentingnya gaya hidup sehat sebagai langkah preventif.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada dosen pembimbing yang memberikan bimbingan, arahan, dan motivasi yang telah diberikan selama proses penelitian ini. Ucapan terima

kasih juga penulis sampaikan kepada kedua orang tua yang senantiasa memberikan doa, dukungan, serta semangat, sehingga penelitian ini dapat terselesaikan dengan baik.

DAFTAR REFERENSI

Budiyono, P. (2024). Penerapan algoritma *Naïve Bayes* untuk prediksi penyakit paru-paru. *Jurnal Teknologi dan Sistem Informasi*, *10*(2), 55–62.

Christian, A., & Sumanto. (2025). Analisis *machine learning* untuk prediksi penyakit paru-paru menggunakan *Random Forest*. *Jurnal Widya Informatika*, *7*(1), 45–53.

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Waltham, MA: Morgan Kaufmann.

Kurniawan, D. (2024). Deteksi dan prediksi cerdas penyakit paru-paru dengan algoritma *Random Forest*. *Jurnal Sains Komputer*, *6*(4), 101–108.

Meiyanti, A. (2020). Klasifikasi diagnosa penyakit paru-paru pada Klinik Raditya Medical Center dengan metode algoritma C4.5. *Jurnal Teknologi dan Ilmu Komputer*, *8*(3), 12–19.

Pambudi, R. (2024). Klasifikasi penyakit paru-paru menggunakan metode *Decision Tree*. *Jurnal Ilmiah Komputer dan Sistem Informasi*, *12*(4), 2397–2402.

Putri, N. S. D. (2023). Hubungan antara kebiasaan merokok terhadap tingkat keparahan penyakit paru obstruktif kronis (PPOK). *Jurnal Kesehatan Respirasi*, *5*(2), 112–118.

Sholiha, A., & Fatah, Z. (2025). Klasifikasi penyakit paru-paru menggunakan *data mining Decision Tree*. *JAMASTIKA*, *4*(1), 45–52.

Sofyan, F. M. A. (2023). Penerapan algoritma C4.5 untuk prediksi penyakit paru-paru menggunakan RapidMiner. *Jurnal Sains Komputer dan Teknologi Informasi*, *11*(2), 247–254.

Sutoyo, I. (2018). Implementasi algoritma Decision Tree untuk klasifikasi data peserta didik. *Jurnal Pilar Nusa Mandiri*, *14*(2), 217–223.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques* (3rd ed.). Burlington, MA: Morgan Kaufmann.