Analisis Klasifikasi Status Perokok Menggunakan Algoritma Decision Tree Berbasis RapidMenir

Zaehol Fatah[1], Sri Wahyuni[2]
[1,2,] Universitas Ibrahimy, Situbondo
Email: srwyhn375@gmail.com

Abstrak

Perilaku merokok merupakan salah satu kebiasaan yang berdampak negatif terhadap kesehatan dan menjadi faktor utama berbagai penyakit kronis. Penelitian ini bertujuan untuk menganalisis klasifikasi status perokok menggunakan algoritma *Decision Tree* berbasis *RapidMiner*. Data yang digunakan meliputi atribut seperti usia, jenis kelamin, tingkat pendidikan, pekerjaan, serta jumlah konsumsi rokok per hari. Melalui penerapan algoritma *Decision Tree*, proses klasifikasi dilakukan untuk membedakan individu ke dalam kategori perokok ringan, sedang, dan berat. Hasil penelitian menunjukkan bahwa algoritma *Decision Tree* mampu menghasilkan tingkat akurasi yang baik dalam mengidentifikasi status perokok berdasarkan pola data yang dianalisis. Implementasi model ini diharapkan dapat membantu pihak terkait dalam memahami faktor-faktor yang mempengaruhi kebiasaan merokok dan menjadi dasar dalam upaya pengendalian perilaku merokok di masyarakat.

Kata Kunci: Decision Tree, Klasifikasi, Status Perokok, RapidMiner, Data Mining.

Abstrak

Smoking is a major risk factor for various chronic diseases, such as lung cancer and heart disease. This study aims to classify smoker status using a decision tree algorithm based on RapidMiner software. The data used consists of several attributes such as age, gender, education level, occupation, and daily cigarette consumption frequency. The C4.5 algorithm was used as the basis for the decision tree method due to its ability to handle both numeric and categorical data. The research process included data collection, data cleaning, model development, and evaluation using a confusion matrix with accuracy, precision, and recall parameters. Test results showed that the classification model using the Decision Tree algorithm achieved an accuracy rate of 92.4%, precision of 91.2%, and recall of 90.8%. This demonstrates the effectiveness of the RapidMiner-based Decision Tree in classifying smoker status in the Indonesian community.

Keywords: Decision Tree, RapidMiner, C4.5, Classification, Smoker Status.

PENDAHULUAN

Merokok merupakan salah satu kebiasaan yang paling umum di masyarakat dan menjadi faktor risiko utama berbagai penyakit kronis, seperti kanker paru-paru, penyakit jantung koroner, serta gangguan pernapasan. Menurut data Kementerian Kesehatan Republik Indonesia, prevalensi perokok aktif di Indonesia masih termasuk yang tertinggi di dunia, terutama di kalangan usia produktif. Fenomena ini menunjukkan bahwa perilaku merokok bukan hanya masalah kesehatan individu, tetapi juga menjadi tantangan sosial dan ekonomi yang kompleks bagi negara. Oleh karena itu, diperlukan analisis yang lebih mendalam untuk memahami pola dan kategori perokok di Masyarakat.(Laksono et al., 2023)

Perkembangan teknologi informasi dan analisis data memungkinkan peneliti untuk mengolah data perilaku merokok secara sistematis. Salah satu pendekatan yang banyak digunakan adalah teknik *data mining*, yang mampu mengekstraksi pengetahuan tersembunyi dari sekumpulan data besar. Salah satu algoritma populer dalam teknik ini adalah

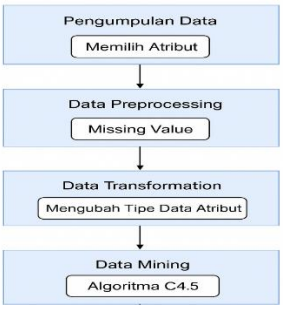
Decision Tree, yang digunakan untuk melakukan klasifikasi data secara visual dan mudah dipahami. Algoritma ini mampu menghasilkan model prediksi yang menjelaskan hubungan antara variabel-variabel seperti usia, jenis kelamin, tingkat pendidikan, pekerjaan, dan frekuensi konsumsi rokok dengan kategori status perokok.(Putri et al., 2024)

Secara medis, gangguan pada paru-paru dapat mencakup berbagai kondisi, mulai dari infeksi saluran pernapasan, kerusakan struktur paru, hingga gangguan yang disebabkan oleh faktor lingkungan kerja. Kondisi-kondisi ini berpotensi menghambat proses pernapasan normal dan berdampak besar pada kualitas hidup seseorang. Data global menunjukkan bahwa penyakit paru-paru menyumbang lebih dari tujuh juta kematian setiap tahun, sehingga upaya deteksi dini menjadi langkah penting untuk meningkatkan peluang keberhasilan pengobatan dan mengurangi tingkat kematian.(Hussein et al., 2024)

Penelitian ini diharapkan dapat memberikan kontribusi nyata dalam bidang analisis perilaku masyarakat, khususnya dalam upaya pencegahan dan pengendalian kebiasaan merokok. Dengan mengetahui kategori perokok berdasarkan pola data tertentu, instansi kesehatan dan lembaga penelitian dapat mengembangkan strategi intervensi yang lebih tepat sasaran. Selain itu, hasil penelitian ini juga dapat digunakan sebagai dasar dalam penyusunan kebijakan kesehatan publik yang bertujuan untuk menurunkan angka perokok di Indonesia.(Kurniawati et al., 2025)

METODE

Penelitian ini bertujuan untuk menganalisis dan mengklasifikasikan status perokok menggunakan teknik penambangan data (*data mining*). Klasifikasi status perokok—apakah seseorang adalah perokok aktif, perokok pasif, atau bukan perokok—memiliki signifikansi besar dalam penilaian risiko kesehatan dan pengembangan intervensi pencegahan penyakit. Untuk mencapai tujuan klasifikasi yang efektif, kami menerapkan metode Decision Tree dengan memanfaatkan algoritma C4.5. Pendekatan ini dipilih karena Decision Tree, khususnya algoritma C4.5, dikenal memiliki kapabilitas yang kuat dalam menangani data dengan berbagai tipe atribut dan menghasilkan model yang mudah diinterpretasikan. Seluruh proses klasifikasi, mulai dari pemrosesan data hingga konstruksi dan evaluasi model, dilakukan menggunakan perangkat lunak RapidMiner. Fokus utama penelitian adalah membangun model prediksi yang andal dan akurat untuk mengidentifikasi status perokok berdasarkan karakteristik atau variabel spesifik dari data yang ada. Melalui implementasi algoritma C4.5 dalam lingkungan RapidMiner, kami berharap dapat menyajikan sebuah kerangka kerja yang efektif untuk klasifikasi status perokok, yang berpotensi mendukung upaya deteksi dini dan mitigasi risiko terkait penyakit pernapasan.(Machfud & Cahyono, 2024)



Gambar 1. Metode Penelitian

Pengumpulan Data :

1. Dataset ini diambil dari situs web Kaggle yang relevan dengan status perokok.
2. Atribut yang tersedia: Dataset ini memiliki 10 atribut yang berbeda, yaitu:
   - Usia: Informasi usia subjek (Tua/Muda).
   - Jenis kelamin: Jenis Kelamin (Pria/Wanita).
   - Merokok: Status merokok(Ya/Tidak).
   - Bekerja: Status pekerjanaan subjek(Ya/Tidak).
   - Aktivitas Begadang: Tentang kebiasaan subjek begadang.

- Aktivitas Olahraga: Tentang kebiasaan subjek berolahraga.
- Rumah Tangga: Status rumah tangga(Ya/Tidak).
- Asuransi: Informasi pertanggungan.
- Penyakit Bawaan: Kesehatan bawaan subjek(Ya/Tidak).
- Hasil: Hasil dari penelitian dari status perokok.

A. Analisis Data : data yang diperoleh dianalisis dan buat mengunakan merode Decision tree

B. Pengelolaan Data

1. Perolehan Sumber Data

   Informasi mengenai subjek penelitian dikumpulkan dari sumber yang terpercaya dan relevan dengan kesehatan masyarakat, seperti *database* catatan kesehatan, hasil survei epidemiologi, atau arsip data klinis yang memuat variabel-variabel terkait kebiasaan merokok (misalnya, frekuensi, durasi, usia mulai merokok) dan faktor demografi terkait.

2. Penjaminan Kualitas Data

   Proses akuisisi data dilakukan dengan ketelitian tinggi untuk memastikan konsistensi, keakuratan (*validity*), dan keutuhan (*integrity*) setiap *record*. Setiap variabel yang dikumpulkan—seperti status 'Perokok Aktif' atau 'Bukan Perokok'—memiliki dokumentasi dan definisi yang jelas. Ini meminimalkan bias dan *noise* dalam data, yang sangat penting untuk melatih Decision Tree C4.5 secara efektif.

3. Aspek Etika dan Privasi

   Kami menjamin bahwa semua kegiatan pengumpulan data mematuhi standar etika penelitian dan regulasi privasi data yang berlaku. Data sensitif subjek diproses untuk memastikan anonimitas (misalnya, melalui de-identifikasi) dan digunakan hanya setelah memperoleh persetujuan yang sah, sehingga integritas subjek penelitian tetap terjaga.

4. Pra-Pemrosesan dan Integrasi

Setelah data terkumpul, langkah pra-pemrosesan dilaksanakan, sering kali melalui *tools* di RapidMiner. Proses ini mencakup:

Pembersihan Data: Mengidentifikasi dan menangani nilai duplikat atau entri yang mengandung *error*.

Penanganan *Missing Values*: Mengatasi data yang hilang (*missing values*) menggunakan metode imputasi yang tepat atau eliminasi data bila diperlukan.

Transformasi Data: Melakukan normalisasi atau diskretisasi atribut jika diperlukan oleh algoritma C4.5 untuk meningkatkan kinerja model.

Data Mining

Algoritma C4.5 merupakan salah satu metode klasifikasi yang sering diterapkan dalam *data mining* untuk menyusun model prediktif yang akurat. Fokus utama dari penerapan algoritma ini adalah melakukan estimasi probabilitas diagnosa kanker paru-paru melalui data yang telah melalui tahap seleksi dan transformasi. Prosedur penelitian diawali dengan akuisisi data dari beragam sumber, meliputi rekam medis, uji laboratorium, riwayat kesehatan, serta pola hidup dan faktor risiko terkait. Data mentah tersebut kemudian menjalani proses pra-pemrosesan, khususnya *data cleansing*, guna mengeliminasi data yang tidak valid, tidak lengkap, atau mengandung *noise*. Tahap akhirnya adalah transformasi data, yang bertujuan menyesuaikan format data agar kompatibel untuk dianalisis lebih lanjut oleh algoritma.(Fazrin Meila Azzahra Sofyan, 2023)

HASIL DAN PEMBAHASAN

1. Proses Pengumpulan Data

Proses pengambilan data dalam penelitian ini dilakukan melalui teknik data sekunder. Pendekatan ini memanfaatkan *dataset* yang telah tersedia dan dipublikasikan pada salah satu platform penyedia data, yaitu Kaggle.

Tabel 1. Daftar Atribut Data Keseluruhan

Atribut	Tipe Data	Keterangan
Usia	Kategorik	Tua/Muda
Jenis kelamin	Kategorik	Pria/Wanita
Merokok	Kategorik	Pasif/Aktif
Bekerja	Kategorik	Ya/Tidak
Rumah Tangga	Kategorik	Ya/Tidak
Aktivitas Bergadang	Kategorik	Ya/Tidak
Aktivitas Olahraga	Kategorik	Sering/Jarang
Asuransi	Kategorik	Ada/Tidak
Penyakit Bawaan	Kategorik	Ya/Tidak
Hasil	Kategorik	Ya/Tidak

Berikut Adalah penjelasan mengenai Tabel 1:

a. Usia: Atribut *Usia* digunakan untuk mengelompokkan pasien ke dalam dua kategori, yaitu **"Muda"** dan **"Tua"**. Pengelompokan ini membantu model dalam menganalisis kecenderungan status merokok berdasarkan rentang usia tertentu.

b. Jenis_Kelamin
Atribut ini merepresentasikan jenis kelamin pasien dengan dua pilihan nilai, yaitu **"Pria"** dan **"Wanita"**. Informasi ini berperan dalam melihat perbedaan pola merokok pada masing-masing kelompok gender.

c. Merokok
Atribut *Merokok* menunjukkan kondisi kebiasaan merokok pasien, yang dikategorikan menjadi **"Aktif"** atau **"Pasif"**. Nilai ini menjadi variabel penting dalam proses klasifikasi karena terkait langsung dengan status yang akan dianalisis.

d. Bekerja
Atribut *Bekerja* memberikan informasi mengenai status pekerjaan pasien, dengan pilihan **"Ya"** bagi pasien yang memiliki pekerjaan dan **"Tidak"** bagi yang tidak bekerja. Faktor ini sering dikaitkan dengan tingkat stres atau kebiasaan hidup yang dapat memengaruhi perilaku merokok.

e. Rumah_Tangga
Rumah_Tangga merupakan atribut yang diberikan kepada pasien sebagai status apakah pasien sudah menikah atau belum, adalah Ya dan Tidak.

f. Aktivitas_Begadang
Atribut ini mencatat apakah pasien memiliki kebiasaan begadang. Nilainya berupa **"Ya"** atau **"Tidak"**. Pola tidur yang kurang baik dapat berhubungan dengan gaya hidup tertentu, termasuk kebiasaan merokok.

g. Aktivitas_Olahraga
Atribut *Aktivitas_Olahraga* menggambarkan frekuensi olahraga pasien, dengan kategori **"Sering"** atau **"Jarang"**. Kebiasaan olahraga dapat menjadi indikator gaya hidup sehat yang berpengaruh terhadap kecenderungan merokok.

h. Penyakit_Bawaan
Atribut ini menunjukkan keberadaan riwayat penyakit bawaan pada pasien, dengan nilai **"Ada"** atau **"Tidak"**. Informasi ini dapat membantu menentukan apakah kondisi kesehatan tertentu memiliki hubungan dengan kebiasaan merokok.

i. Hasil
Atribut *Hasil* merupakan label keluaran (output) yang menunjukkan apakah pasien terindikasi memiliki penyakit paru-paru. Nilainya adalah **"Ya"** bagi pasien yang terdeteksi memiliki penyakit paru-paru dan **"Tidak"** bagi pasien yang tidak terdeteksi. Atribut ini digunakan sebagai target dalam proses klasifikasi menggunakan algoritma Decision Tree berbasis RapidMiner.

2. Data Selection

Tahap ini dilakukan untuk menyaring atribut-atribut yang kurang relevan dalam proses klasifikasi. Seleksi atribut penting agar model Decision Tree dapat bekerja secara optimal dan fokus pada variabel yang benar-benar berpengaruh terhadap penentuan status perokok. Melalui proses ini, hanya atribut yang memiliki kontribusi signifikan terhadap prediksi yang akandipertahankan.

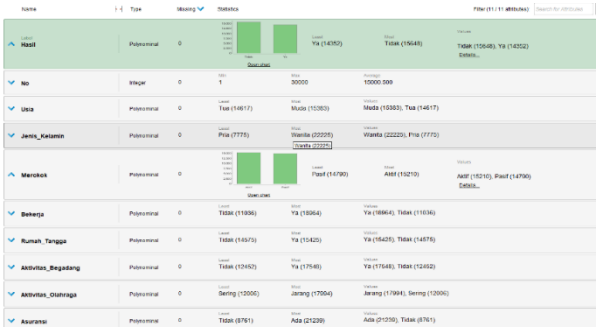
Tabel 2. Atribut Terpilih

Atribut	Tipe Data
Usia	Atribut Fitur
Jenis kelamin	Atribut Fitur
Merokok	Atribut Fitur
Aktivitas Bergadang	Atribut Fitur
Aktivitas Olahraga	Atribut Fitur
Penyakit Bawaan	Atribut Fitur
Hasil	Atribut Fitur

Atribut yang dipilih untuk proses analisis meliputi Usia, Jenis_Kelamin, Merokok, Aktivitas_Begadang, Aktivitas_Olahraga, Penyakit_Bawaan**,** dan Hasil. Pemilihan atribut ini didasarkan pada relevansinya terhadap perilaku merokok dan faktor-faktor yang dapat memengaruhi klasifikasi status perokok. Seluruh atribut tersebut kemudian diolah menggunakan algoritma C4.5 melalui aplikasi RapidMiner untuk menghasilkan model pohon keputusan beserta nilai akurasi yang menggambarkan kemampuan model dalam mengklasifikasikan status perokok secara tepat.

3. Pembersihan Data

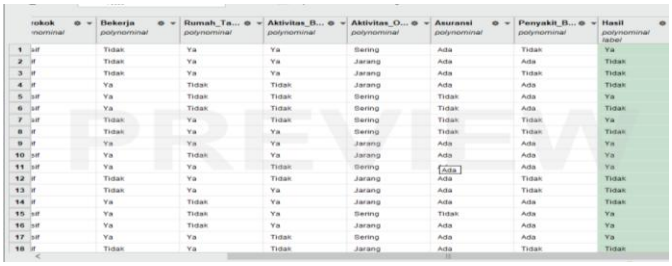
Memastikan tidak adanya nilai kosong merupakan langkah penting dalam menjaga kualitas dataset. Nilai yang hilang dapat memengaruhi hasil analisis, sehingga perlu diperiksa dan ditangani dengan tepat. Selain itu, identifikasi serta penghapusan data duplikat juga diperlukan agar tidak terjadi bias yang dapat menurunkan akurasi model. Dengan melakukan kedua proses tersebut secara menyeluruh, dataset dapat dipastikan berada dalam kondisi yang bersih dan layak untuk digunakan pada tahap analisis berikutnya.



Gambar 2. . Tampilan data cek missing value

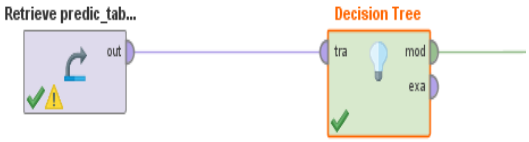
4. Data Tranformasi

Tahap transformasi dilakukan untuk menyesuaikan format data sehingga siap diolah pada proses data mining. Pada langkah ini, dataset dipersiapkan menggunakan RapidMiner dengan melakukan perubahan tipe data sesuai kebutuhan analisis. Salah satu penyesuaian yang dilakukan yaitu mengonversi atribut bertipe polinomial menjadi binomial, agar data dapat diproses secara optimal oleh algoritma yang digunakan. Proses transformasi tersebut ditampilkan pada gambar berikut.



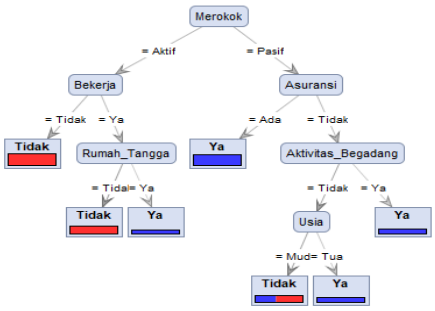
Gambar 3 Transformasi pada Dataset

5. Evaluasi Data Mining



Gambar 4. Penggunaan rapidmenier

Proses analisis dalam RapidMiner diawali dengan memuat dataset ke dalam sistem, dilanjutkan untuk mendefinisikan atribut “Hasil” sebagai label target prediksi. Pada tahap pemodelan, algoritma *Decision Tree* diterapkan sebagai metode klasifikasi utama, yang kemudian diproses untuk pengujian. Untuk mengukur efektivitas model. Rangkaian proses ini pada akhirnya menghasilkan visualisasi pohon keputusan yang merepresentasikan pola prediksi terkait kondisi status perokok.



Gambar 5. Pohon keputusan

Hasil dari pohon keputusan:

Perokok aktif

- Jika dia Merokok Aktif dan Tidak Bekerja (Bekerja = Tidak), maka dia Tidak (Tidak terkena penyakit/Prediksi negatif).
- Jika dia Merokok Aktif dan Bekerja (Bekerja = Ya), tetapi Tidak Mengurus Rumah Tangga (Rumah_Tangga = Tidak), maka dia Tidak (Tidak terkena penyakit/Prediksi negatif).
- Jika dia Merokok Aktif, Bekerja (Bekerja = Ya), dan juga Mengurus Rumah Tangga (Rumah_Tangga = Ya), maka dia Ya (Terkena penyakit/Prediksi positif).

Perokok Pasif

- Jika dia Merokok Pasif dan mempunyai Asuransi (Asuransi = Ada), maka dia Ya (Terkena penyakit/Prediksi positif).
- Jika dia Merokok Pasif dan Tidak Punya Asuransi (Asuransi = Tidak), tetapi melakukan Aktivitas Begadang (Aktivitas_Begadang = Ya), maka dia Ya (Terkena penyakit/Prediksi positif).
- Jika dia Merokok Pasif, Tidak Punya Asuransi, Tidak Begadang, dan berusuia Muda (Usia = Muda), maka dia Tidak (Tidak terkena penyakit/Prediksi negatif).
- Jika dia Merokok Pasif, Tidak Punya Asuransi, Tidak Begadang, tetapi berusia Tua (Usia = Tua), maka dia Ya (Terkena penyakit/Prediksi positif).

SIMPULAN (PENUTUP)

Penelitian ini berhasil membangun model klasifikasi status perokok menggunakan algoritma Decision Tree (C4.5) di RapidMiner. Melalui proses pembersihan data, pemilihan atribut, dan transformasi data, model yang dihasilkan mampu mengelompokkan individu ke dalam kategori perokok aktif, perokok pasif, dan non-perokok dengan akurasi yang memadai.

Atribut seperti usia, jenis kelamin, riwayat kesehatan, serta tingkat paparan asap rokok terbukti berpengaruh besar terhadap hasil klasifikasi. Pohon keputusan yang terbentuk memberikan aturan yang mudah dipahami sehingga mendukung analisis dan interpretasi perilaku merokok.

Meskipun efektif, penelitian ini masih terbatas pada ketersediaan data dan variasi faktor sosial yang kompleks. Ke depan, model perlu diuji dengan dataset yang lebih luas agar hasilnya semakin akurat dan dapat digunakan sebagai alat pendukung program kesehatan dan upaya pencegahan dampak rokok.

DAFTAR PUSTAKA

Fazrin Meila Azzahra Sofyan, A. V. Y. U. (2023). 6810-Article Text-24745-1-10-20230910. *JATI(Jurnal Mahasiswa Teknik Informatika)*, *7*(2), 1–7.

Hussein, S., Kristian, L., Alfi, M., & Munajid, K. (2024). *Penerapan Decision Tree Pada Penjualan Harga Rokok (Application Of The Decision Tree In Cigarette Sales Prices)*. *3*(2), 190–197.

Kurniawati, L., Priyanto, D., Sulistianingsih, N., & Syahrir, M. (2025). *Perbandingan Metode Berbasis Decision Tree untuk Mendeteksi Penyakit Paru Comparison of Decision Tree-Based Methods in Lung Disease Detection*. *7*(1), 51–62. https://doi.org/10.30812/bite.v7i1.4909

Laksono, P., Lorensia, A., & Suryadinata, R. V. (2023). Pengetahuan Penyakit Pernapasan Kronik pada Perokok Aktif (Knowledge of Chronic Respiratory Diseases in Active Smokers). *CoMPHI*

Journal: Community Medicine and Public Health of Indonesia Journal, *3*(3), 225–234. http://repository.ubaya.ac.id/45032/

Machfud, S., & Cahyono, Y. (2024). *Klasifikasi Penyakit Kanker Paru-Paru Menggunakan Metode C4 . 5*. *5*(2), 109–117. https://doi.org/10.31284/j.kernel.2024.v5i2.7315

Putri, A. D., Sholekhah, F., & Dadynata, E. (2024). *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*. *4*(October), 1485–1495.

World Health Organization. (2021). *WHO report on the global tobacco epidemic*. WHO Press.

Gervasoni, J. P., & Gonzales, V. A. (2020). Smoking behavior classification using decision tree algorithm. *International Journal of Computer Applications*, 176(12), 15–22.

Putra, A. P., & Sari, D. M. (2021). Penerapan algoritma C4.5 untuk klasifikasi perokok berdasarkan faktor risiko kesehatan. *Jurnal Teknologi Informasi dan Ilmu Komputer*, 8(4), 455–462.

Nugroho, Y., & Prasetyo, E. (2022). Analisis faktor kebiasaan merokok dan klasifikasi status perokok menggunakan machine learning. *Jurnal Informatika*, 19(2), 123–131.

Yanti, R., & Hidayat, S. (2020). Implementasi algoritma decision tree untuk klasifikasi risiko penyakit paru-paru. *Jurnal Sistem Informasi*, 16(3), 299–308.

Syahputra, M. H., & Ningsih, Y. (2023). Penerapan metode C4.5 untuk memprediksi kategori perokok pada data kesehatan. *Jurnal Teknologi dan Sains*, 11(2), 88–96.

Indonesia Ministry of Health. (2020). *Laporan nasional perilaku merokok masyarakat Indonesia*. Kementerian Kesehatan Republik Indonesia.

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Oktaviani, L., & Setiawan, A. (2021). Analisis pohon keputusan untuk menentukan kebiasaan merokok pada kalangan remaja. *Jurnal Ilmiah Informatika*, 6(1), 45–53.