Klasifikasi Penyakit Paru-Paru Menggunakan Data Mining Decision Tree

[1]Aviatus Sholiha, [2]Zaehol Fatah
[1,2] Sistem Informasi, Universitas Ibrahimy, Situbondo
Email : [1]aviatussholiha83@gmail.com [2]zaeholfatah@gmail.com

Abstrak

Paru-paru adalah organ vital dalam sistem pernapasan manusia yang sering kali terpapar oleh kebiasaan merokok dan bahan berbahaya lainnya, sehingga meningkatkan risiko penyakit paru-paru, terutama tuberkulosis (TB). Di Indonesia, TB merupakan masalah kesehatan serius dengan prevalensi tinggi dan menjadi penyebab kematian ketiga terbesar. Penelitian ini bertujuan untuk mengembangkan metode klasifikasi penyakit paru-paru menggunakan algoritma Decision Tree (C4.5) dan AdaBoost. Dataset penelitian mencakup data medis dari 2.149 pasien dengan atribut seperti riwayat kesehatan, hasil laboratorium, dan parameter fisiologis. Proses validasi dilakukan menggunakan teknik percentage split (80:20) dan k-fold cross-validation (k=10). Evaluasi performa algoritma berdasarkan metrik akurasi, presisi, dan recall menunjukkan hasil signifikan dalam mendeteksi penyakit paru-paru seperti asma, bronkitis, dan TB. Dengan menggunakan perangkat lunak seperti RapidMiner, penelitian ini memanfaatkan data mining untuk menemukan pola dan meningkatkan akurasi diagnosis. Hasilnya diharapkan dapat membantu tenaga medis dalam proses diagnosis yang lebih cepat dan akurat, sekaligus memberikan kontribusi dalam pengembangan metode deteksi penyakit paru-paru berbasis teknologi.

Kata Kunci: ***Data mining, Penyakit Paru-paru, Algoritma Decision Tree, Klasifikasi.***

PENDAHULUAN

Paru-Paru merupakan organ tubuh yang sangat penting untuk pernapasan. Banyak orang menggunakan paru-paru dan sistem saluran pernapasannya bukan untuk mengisap oksigen dari udara bersih, melainkan mengisap asap hasil pembakaran tembakau, cengkeh, dan bahan-bahan psikotropika berbahaya lainnya yang tidak perlu disangkal lagi merupakan racun yang merusak paruparu. Hal ini menyebabkan banyak orang yang terindikasi menderita penyakit paru-paru.(Junaidi, 2010)

Penyakit paru, terutama tuberkulosis (TB), masih menjadi masalah kesehatan di Indonesia. Menurut WHO, prevalensi TB paru adalah 130 per 100.000 penduduk, dengan 539.000 kasus baru dan 101.000 kematian setiap tahun. TB merupakan penyebab kematian ketiga setelah penyakit jantung dan saluran pernapasan. Di dunia, sekitar 9 juta kasus TB terjadi setiap tahun, dengan 3 juta kematian, terutama di negara berkembang di Asia Tenggara. Sekitar 25% kematian akibat TB sebenarnya dapat dicegah (Depkes RI, 2002).(Sedjati, 2013)

Penelitian tentang klasifikasi penyakit paru-paru memerlukan metode yang mampu mendeteksi dan mengklasifikasikan jenis penyakit secara akurat. Dataset yang digunakan dalam penelitian ini mencakup atribut-atribut seperti riwayat kesehatan pasien, hasil uji laboratorium, dan parameter fisiologis yang relevan, sehingga mampu meningkatkan akurasi deteksi dan (Sedjati, 2013)klasifikasi penyakit paru-paru. Penelitian ini membandingkan performa algoritma Decision Tree (C4.5) dan AdaBoost dalam klasifikasi penyakit paru-paru seperti asma, bronkitis kronis, dan tuberkulosis (TB). Pengujian dilakukan menggunakan dua teknik validasi: percentage split dan k-fold cross-validation. Pada teknik percentage split, dataset dibagi menjadi 80% data latih dan 20% data uji. Sementara itu, pada k-fold cross-validation, dataset dibagi menjadi 10

kelompok (k=10), di mana setiap kelompok bergantian menjadi data latih dan data uji. Penelitian ini mengevaluasi kinerja algoritma berdasarkan metrik akurasi, presisi, dan recall untuk menentukan algoritma dan teknik validasi yang paling optimal dalam mendeteksi penyakit paru-paru. Hasilnya diharapkan dapat memberikan kontribusi signifikan dalam membantu tenaga medis melakukan diagnosis yang lebih cepat dan akurat terhadap pasien dengan gangguan paru-paru.(Haffandi et al., 2022)

Penelitian ini menggunakan algoritma Decision Tree karena mudah dipahami oleh manusia. Decision Tree, atau pohon keputusan, adalah model prediksi berbentuk struktur pohon yang mengubah data menjadi aturan-aturan sederhana. Salah satu keunggulan utama metode ini adalah kemampuannya menyederhanakan proses pengambilan keputusan yang kompleks, sehingga solusi dapat ditemukan dengan lebih mudah dan efisien, seperti yang disebutkan oleh Asmaul dalam jurnalnya.(Baharudin & Dwi Nuryana, 2022)

METODE

Pengambilan Data

Pada bagian ini dijelaskan bagaimana penulis mengumpulkan data untuk penelitian ini. Data dikumpulkan dengan menggunakan metode studi literatur dan data sekunder yang berasal dari kumpulan data medis terkait penyakit paru-paru, yang diakses melalui platform seperti Kaggle. Kumpulan data ini mencakup informasi kesehatan dari 2.149 pasien yang diidentifikasi dengan ID unik, mencakup detail seperti demografi, riwayat gaya hidup, kondisi lingkungan, riwayat medis, hasil pengukuran klinis, gejala, dan diagnosis penyakit paru-paru. Data ini sangat relevan untuk peneliti dan ilmuwan data dalam mengeksplorasi faktor-faktor risiko, mengembangkan model prediktif, serta melakukan analisis statistik untuk mendukung diagnosis dan pengobatan penyakit paru-paru.

# No	Usia	Jenis_Kelamin	Merokok	Bekerja	Rumah_Tangga	Aktivitas_Begada...	Aktivitas_Olahraga	Asuransi	Penyakit_Bawaan
1 30.0k	Muda 51% Tua 49%	Wanita 74% Pria 26%	Aktif 51% Pasif 49%	Ya 63% Tidak 37%	Ya 51% Tidak 49%	Ya 58% Tidak 42%	Jarang 60% Sering 40%	Ada 71% Tidak 29%	Ada 65% Tidak 36%
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
20	Tua	Wanita	Aktif	Ya	Tidak	Ya	Jarang	Ada	Ada
21	Muda	Wanita	Pasif	Ya	Tidak	Ya	Sering	Tidak	Ada
22	Tua	Pria	Pasif	Tidak	Ya	Ya	Sering	Ada	Tidak
23	Tua	Pria	Aktif	Tidak	Ya	Ya	Jarang	Ada	Ada
24	Muda	Pria	Aktif	Tidak	Ya	Ya	Jarang	Ada	Tidak

Gambar 1. Dataset

Data Mining

Data Mining adalah proses analisis data besar (big data) untuk menemukan pola, hubungan, atau informasi tersembunyi yang berguna bagi pengambilan keputusan. Proses ini melibatkan teknik statistik, algoritma pembelajaran mesin (machine learning),

dan sistem basis data untuk mengidentifikasi informasi yang relevan dari dataset besar yang kompleks. Data mining dapat menganalisis data set untuk menemukan hubungan dan menyimpulkan data dengan cara yang jelas, dimana hasilnya dapat dimengerti dan berguna bagi pemilik data.(Rizki et al., 2020)

Data mining bukanlah suatu bidang yang sama sekali baru. Salah satu kesulitan untuk mendefinisikan data mining adalah kenyataan bahwa data mining mewarisi banyak aspek dan teknik dari bidang-bidang ilmu yang dulu sudah mapan terlebih dulu. Data mining memiliki akar yang panjang dari bidang ilmu yang berbeda seperti kecerdasan buatan (artificial intelligent), machine learning, statistik, database, dan juga information retrieval.(Mardi, 2017)

Decision Tree

Decision Tree digunakan untuk mempelajari klasifikasi dan prediksi pola dari data dan menggambarkan relasi dari variabel attribut x dan variabel target y dalam bentuk pohon. Decission Tree adalah struktur menyerupai flowchart dimana setiap internal node (node yang bukan leaf atau bukan node terluar) merupakan pengujian terhadap variabel attribut, tiap cabangnya merupakan hasil dari pengujian tersebut, sedangkan node terluar yakni leaf menjadi labelnya. (Sutoyo, 2018)

Klasifikasi

Proses klasifikasi adalah membandingkan data lama (data training) dengan data baru (data testing) untuk menghasilkan kemungkinan atau prediksi berdasarkan data testing.(Hana, 2020) Sebuah model klasifikasi diuji dengan menerapkan untuk menguji data dengan nilai target dikenal dan membandingkan nilai prediksi dengan nilai-nilai diketahui. Data uji harus sesuai dengan data yang digunakan untuk membangun model dan harus dipersiapkan dengan cara yang sama. Biasanya data train dan data test berasal dari set data yang sama asalnya. Matrik tes digunakan untuk menilai seberapa akurat model dan memprediksi nilai-nilai yang diketahui.(Putra & Chan, 2018)

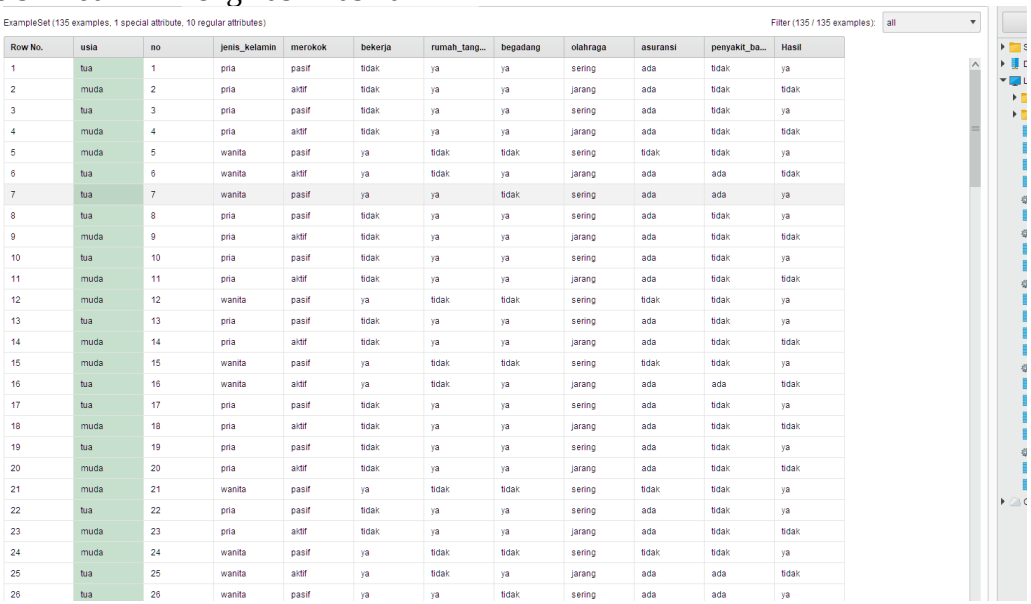
Rapid Miner

Rapidminer adalah sebuah Analisis teks yang fokus didalam pekerjaan yang dilakukan oleh RapidMiner text mining, yang melibatkan penggalian pola dari kumpulan data besar dan menggabungkannya dengan Teknik statistik, kecerdasan buatan, dan basis data.(Volume et al., n.d.) RapidMiner menawarkan antarmuka drag-anddrop yang memungkinkan pengguna untuk membangun alur kerja untuk memproses dan menganalisis data. Ini mendukung beragam sumber data, termasuk file datar, basis data, dan platform big data seperti Hadoop dan Spark. Perangkat lunak ini juga mencakup beragam operator yang sudah dibangun, yang merupakan blok bangunan dari alur kerja, yang mencakup semua tahap proses data mining, seperti pembersihan data, pemilihan fitur, dan pemodelan.(Rafi Nahjan et al., 2023)

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.

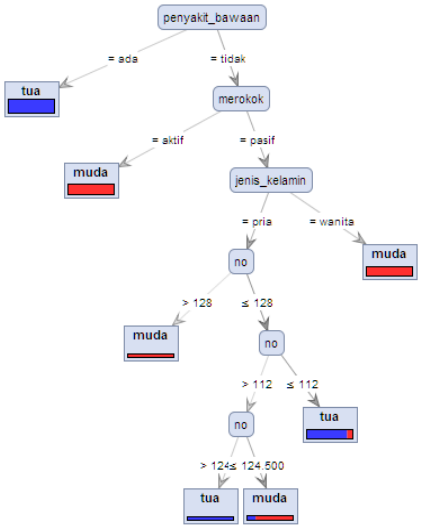


ExampleSet (135 examples, 1 special attribute, 10 regular attributes) Filter (135 / 135 examples): all

Row No.	usia	no	jenis_kelamin	merokok	bekerja	rumah_tang...	begadang	olahraga	asuransi	penyakit_ba...	Hasil
1	tua	1	pria	pasif	tidak	ya	ya	sering	ada	tidak	ya
2	muda	2	pria	aktif	tidak	ya	ya	jarang	ada	tidak	tidak
3	tua	3	pria	pasif	tidak	ya	ya	sering	ada	tidak	ya
4	muda	4	pria	aktif	tidak	ya	ya	jarang	ada	tidak	tidak
5	muda	5	wanita	pasif	ya	tidak	tidak	sering	tidak	tidak	ya
6	tua	6	wanita	aktif	ya	tidak	ya	jarang	ada	ada	tidak
7	tua	7	wanita	pasif	ya	ya	tidak	sering	ada	ada	ya
8	tua	8	pria	pasif	tidak	ya	ya	sering	ada	tidak	ya
9	muda	9	pria	aktif	tidak	ya	ya	jarang	ada	tidak	tidak
10	tua	10	pria	pasif	tidak	ya	ya	sering	ada	tidak	ya
11	muda	11	pria	aktif	tidak	ya	ya	jarang	ada	tidak	tidak
12	muda	12	wanita	pasif	ya	tidak	tidak	sering	tidak	tidak	ya
13	tua	13	pria	pasif	tidak	ya	ya	sering	ada	tidak	ya
14	muda	14	pria	aktif	tidak	ya	ya	jarang	ada	tidak	tidak
15	muda	15	wanita	pasif	ya	tidak	tidak	sering	tidak	tidak	ya
16	tua	16	wanita	aktif	ya	tidak	ya	jarang	ada	ada	tidak
17	tua	17	pria	pasif	tidak	ya	ya	sering	ada	tidak	ya
18	muda	18	pria	aktif	tidak	ya	ya	jarang	ada	tidak	tidak
19	tua	19	pria	pasif	tidak	ya	ya	sering	ada	tidak	ya
20	muda	20	pria	aktif	tidak	ya	ya	jarang	ada	tidak	tidak
21	muda	21	wanita	pasif	ya	tidak	tidak	sering	tidak	tidak	ya
22	tua	22	pria	pasif	tidak	ya	ya	sering	ada	tidak	ya
23	muda	23	pria	aktif	tidak	ya	ya	jarang	ada	tidak	tidak
24	muda	24	wanita	pasif	ya	tidak	tidak	sering	tidak	tidak	ya
25	tua	25	wanita	aktif	ya	tidak	ya	jarang	ada	ada	tidak
26	tua	26	wanita	pasif	ya	ya	tidak	sering	ada	ada	ya

Gambar 2. Data Penyakit Paru-paru

Berikut gambar dibawah ini adalah hasil pohon decition tree dan persentase klasifikasi



Gambar 3. Pohon keputusan untuk Penyakit Paru-Paru

Tree

```
penyakit_bawaan = ada: tua {tua=45, muda=0}
penyakit_bawaan = tidak
|   merokok = aktif: muda {tua=0, muda=30}
|   merokok = pasif
|   |   jenis_kelamin = pria
|   |   |   no > 128: muda {tua=0, muda=3}
|   |   |   no ≤ 128
|   |   |   |   no > 112
|   |   |   |   |   no > 124.500: tua {tua=2, muda=0}
|   |   |   |   |   no ≤ 124.500: muda {tua=1, muda=4}
|   |   |   |   no ≤ 112: tua {tua=24, muda=3}
|   |   jenis_kelamin = wanita: muda {tua=0, muda=23}
```

Gambar 4. Description

Perhitungan :

1. Root Node (penyakit_bawaan = ada atau tidak):
   - Jika penyakit_bawaan = ada, maka:
     - Klasifikasi: Tua (45 data).
     - Tidak ada data untuk kategori Muda.
   - Jika penyakit_bawaan = tidak, maka analisis dilanjutkan ke node berikutnya.
2. Node merokok (aktif atau pasif):
   - Jika merokok = aktif, maka:
   - Klasifikasi: Muda (30 data).
   - Tidak ada data untuk kategori Tua.
   - Jika merokok = pasif, maka dilanjutkan ke jenis_kelamin.
3. Node jenis_kelamin untuk merokok = pasif:
   - Jika jenis_kelamin = pria, maka dilanjutkan ke node berikutnya.
   - Jika jenis_kelamin = wanita, maka:
     - Klasifikasi: Muda (23 data).
     - Tidak ada data untuk kategori Tua.

4. Node jenis_kelamin = pria (No ≤ 128 atau No > 128):
    - Jika No > 128, maka:
        - Klasifikasi: Muda (3 data).
        - Tidak ada data untuk kategori Tua.
    - Jika No ≤ 128, maka:
        - Jika No > 112, maka:
        - Jika No > 124.500, maka:
        - Klasifikasi: Tua (2 data).
        - Tidak ada data untuk kategori Muda.
        - Jika No ≤ 124.500, maka:
        - Klasifikasi: Muda (4 data).
        - Ada 1 data untuk kategori Tua.
        - Jika No ≤ 112, maka:
        - Klasifikasi: Tua (24 data).
        - Ada 3 data untuk kategori Muda.

Perhitungan Total Data:

- Tua:
    - 45 (penyakit_bawaan = ada) + 2 (No > 124.500) + 1 (No ≤ 124.500) + 24 (No ≤ 112) = 72 data.
- Muda:
    - 30 (merokok = aktif) + 3 (No > 128) + 4 (No ≤ 124.500) + 23 (jenis_kelamin = wanita) = 60 data.

Total data = 72 (Tua) + 60 (Muda) = 132 data.

KESIMPULAN

Dalam Penelitian ini dapat disimpulkan bahwa penyakit paru-paru, termasuk tuberkulosis (TB), masih menjadi masalah kesehatan serius di Indonesia dan dunia, terutama di negara berkembang. Klasifikasi penyakit paru-paru menggunakan algoritma seperti Decision Tree dan AdaBoost terbukti efektif dalam menganalisis data medis untuk membantu tenaga medis dalam diagnosis yang lebih cepat dan akurat. Penggunaan metode seperti percentage split dan k-fold cross-validation dalam pengujian model meningkatkan keakuratan prediksi, dengan atribut dataset seperti riwayat kesehatan, hasil uji laboratorium, dan parameter fisiologis yang relevan. Decision Tree dipilih karena kemampuannya menyederhanakan proses pengambilan keputusan dan memprediksi data baru berdasarkan pola data sebelumnya.

Hasil analisis data menunjukkan bahwa faktor seperti riwayat penyakit bawaan, kebiasaan merokok, dan jenis kelamin pasien berkontribusi dalam klasifikasi penyakit paru-paru. Dari total 132 data yang dianalisis, 72 data terklasifikasi sebagai kategori "Tua," sementara 60 data masuk dalam kategori "Muda." Studi ini memberikan kontribusi signifikan dalam pengembangan alat diagnosis berbasis data mining untuk mendukung pengambilan keputusan medis.

UCAPAN TERIMA KASIH

Ucapkan terima kasih diberikan kepada semua pihak yang telah membantu dalam penyusunan jurnal ini, terutama kepada pembimbing Zaehol Fatah, M.Kom, keluarga dan sahabat-sahabat seperjuangan yang telah memberikan dukungan motivasi dan nasehat. Semoga jurnal ini dapat memberikan manfaat bagi pembaca.

DAFTAR PUSTAKA

Baharudin, M. N., & Dwi Nuryana, I. K. (2022). Implementasi Algoritma Decision Tree untuk Klasifikasi Surat pada Aplikasi Mobile E-Surat Dinas Komunikasi dan Informatika Kota Kediri Berbasis Android. *Journal of Informatics and Computer Science (JINACS)*, *4*(01), 76–85. https://doi.org/10.26740/jinacs.v4n01.p76-85

Haffandi, M. Y., Haerani, E., Syafria, F., & Oktavia, L. (2022). Klasifikasi Penyakit Paru-Paru Dengan Menggunakan Metode Naïve Bayes Classifier. *Jurnal Teknik Informasi Dan Komputer (Tekinkom)*, *5*(2), 176. https://doi.org/10.37600/tekinkom.v5i2.649

Hana, F. M. (2020). Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma Decision Tree C4.5. *Jurnal SISKOM-KB (Sistem Komputer Dan Kecerdasan Buatan)*, *4*(1), 32–39. https://doi.org/10.47970/siskom-kb.v4i1.173

Junaidi, I. (2010). *Penyakit Paru dan Saluran Napas*.

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

Putra, P. P., & Chan, A. S. (2018). Pengembangan Aplikasi Perhitungan Prediksi Stock Motor Menggunakan Algoritma C 4.5 Sebagai Bagian dari Sistem Pengambilan Keputusan (Studi Kasus di Saudara Motor). *INOVTEK Polbeng - Seri Informatika*, *3*(1), 24. https://doi.org/10.35314/isi.v3i1.296

Rafi Nahjan, M., Nono Heryana, & Apriade Voutama. (2023). Implementasi Rapidminer Dengan Metode Clustering K-Means Untuk Analisa Penjualan Pada Toko Oj Cell. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *7*(1), 101–104. https://doi.org/10.36040/jati.v7i1.604

Rizki, M., Devrika, D., Umam, I. H., & Lubis, F. S. (2020). Aplikasi Data Mining dalam Penentuan Layout Swalayan dengan Menggunakan Metode MBA. *Jurnal Teknik Industri: Jurnal Hasil Penelitian Dan Karya Ilmiah Dalam Bidang Teknik Industri*, *5*(2), 130. https://doi.org/10.24014/jti.v5i2.8958

Sedjati, F. (2013). Balai Pengobatan Penyakit Paru-paru (BP4) Yogyakarta. *Universitas Ahmad Dahlan*. https://adoc.pub/fitria-sedjati-fakultas-psikologi-universitas-ahmad-dahlan-j.html

Sutoyo, I. (2018). Implementasi Algoritma Decision Tree Untuk Klasifikasi Data Peserta Didik. *Jurnal Pilar Nusa Mandiri*, *14*(2), 217. https://doi.org/10.33480/pilar.v14i2.926

Volume, X., Nomor, X., Tahun, B., Sari, Y., & Fatah, Z. (n.d.). *Gudang Jurnal Multidisiplin Ilmu Klasifikasi Penyakit Alzheimer Menggunakan Data Mining Decision Tree*. *X*, 1–6.