

Penerapan Algoritma C4.5 dalam Klasifikasi Kelulusan Mahasiswa

Azwar Anas

*Fakultas Pertanian, Teknologi dan Sains
Program Studi Sistem Informasi, Universitas Graha Karya Muara Bulian
Jl. Gajah Mada Muara Bulian-Indonesia
E-Mail: azwarzayn@gmail.com*

Abstract

The advancement of information technology and computer science has encouraged various institutions, including higher education institutions, to optimize the use of data in decision-making processes. Each student possesses a unique set of data, which varies depending on their academic and non-academic abilities. The greater the number of students, the more diverse the range of values that can be observed. These data will merely appear as meaningless stacks of information if no in-depth analysis is conducted. However, when analyzed using data mining methods, this data can be transformed into new knowledge that was previously hidden. The objective of this study is to classify student data into two target categories: graduating on time and not graduating on time. The method employed is the C4.5 classification algorithm, using the Weka data mining software. The results of the study show that the data classification achieved an accuracy rate of 91.5663%, and 61 students were predicted to graduate on time out of a total of 83 classified students.

Keywords: *data mining, students, classification, the C4.5 algorithm, outcomes.*

1. Pendahuluan

Perkembangan teknologi informasi dan ilmu komputer telah mendorong berbagai institusi, termasuk perguruan tinggi, untuk mengoptimalkan pemanfaatan data dalam proses pengambilan keputusan. Salah satu bentuk pemanfaatan tersebut adalah melalui penerapan teknik *data mining*, yang bertujuan untuk mengekstraksi pola, informasi tersembunyi, atau pengetahuan baru dari kumpulan data yang besar. Di lingkungan perguruan tinggi, data akademik mahasiswa seperti nilai, indeks prestasi, keaktifan, dan status studi merupakan aset penting yang dapat dianalisis guna mendukung kebijakan akademik yang berbasis data (*data-driven decision making*).

Klasifikasi merupakan salah satu teknik dalam *data mining* yang digunakan untuk memetakan data ke dalam kelas-kelas tertentu berdasarkan atribut-atribut yang dimilikinya [1]. Teknik ini sangat berguna dalam konteks pendidikan, khususnya untuk mengidentifikasi pola karakteristik mahasiswa berdasarkan data historis, sehingga memungkinkan pihak universitas melakukan prediksi terhadap kondisi atau status mahasiswa di masa mendatang. Dengan demikian, klasifikasi dapat digunakan sebagai alat bantu dalam proses evaluasi dan perencanaan strategi pembinaan akademik.

Algoritma C4.5 merupakan salah satu algoritma klasifikasi yang populer dan banyak digunakan karena kemampuannya dalam membangun pohon keputusan (*decision tree*) secara efisien [2]. Algoritma ini mampu menangani atribut data numerik maupun kategorik, serta mengelola data yang tidak lengkap. Dengan menggunakan pendekatan *information gain ratio*, C4.5 secara sistematis membentuk struktur pohon yang dapat digunakan untuk mengklasifikasikan data baru dengan tingkat akurasi yang tinggi.

Penelitian ini bertujuan untuk mengimplementasikan algoritma C4.5 dalam klasifikasi data kelulusan mahasiswa Universitas Graha Karya Muara Bulian (UGKMB). Melalui implementasi ini, diharapkan dapat dihasilkan model klasifikasi yang dapat membantu institusi dalam mengidentifikasi kecenderungan status kelulusan mahasiswa, serta memberikan kontribusi terhadap sistem pendukung keputusan dalam lingkungan pendidikan tinggi.

Setiap mahasiswa memiliki himpunan data yang akan berbeda. Hal ini tergantung dari mahasiswa dalam kemampuan akademik maupun non akademik. Semakin banyak jumlah mahasiswa, semakin banyak pula ragam nilai yang terlihat. Data-data tersebut hanya akan tampak seperti tumpukan data yang tak bermakna,

jika tidak dilakukan suatu analisis mendalam. Namun ketika dianalisis menggunakan metode *data mining* (penambangan data), data tersebut akan tersaji dalam suatu pengetahuan baru yang selama ini tersembunyi [3].

Data mining menggunakan banyak algoritma tergantung pada jenis data dan pengetahuan yang ingin digali. *Pattern Recognition* adalah istilah tambahan untuk bidang ini [4]. Beberapa algoritma yang digunakan seperti algoritma K-Means untuk melakukan *clustering data* [5], algoritma *Market Bascet Analysis* (MBA) [6], *Frequent Pattern Tree* [7], dan Apriori [8] untuk menggali kaidah asosiasi (*Association Rule*).

Penelitian ini akan menggunakan algoritma C4.5 untuk melakukan klasifikasi data mahasiswa UGKMB. Hal ini sejalan dengan pengetahuan yang ingin dihasilkan dari bongkahan data tersebut.

Formula yang digunakan dalam algoritma C4.5 [9]:

$$\text{Entropy (S)} = \sum_{i=1}^n (-p_i) * \log_2(p_i)$$

S = himpunan kasus

n = jumlah partisi atribut S

p_i = proporsi dari s_i terhadap S

$$\text{Gain (S,A)} = \text{Entropy (S)} - \sum_{i=1}^n \frac{|S_i|}{|S|} * \text{entropy}(s_i)$$

$|S_i|$ = jumlah kasus pada partisi ke-i

$|S|$ = jumlah kasus dalam S

Beberapa penelitian terkait diantaranya:

1. Penelitian pada analisis pengadaan peralatan kantor menggunakan algoritma C4.5 berhasil membantu pengambil Keputusan dalam menentukan pengadaan alat kantor yang efektif [10].
2. Penelitian tentang klasifikasi gizi balita dengan Tingkat akurasi mencapai 95,50% [11].
3. Penelitian tentang perbandingan algoritma Naïve Bayes dan C4.5 dengan hasil lebih tinggi Tingkat akurasi Nive Bayes dibandingkan C4.5 [12].
4. Penelitian tentang klasifikasi penyakit stroke dengan hasil Tingkat akurasi mencapai 99,07% [13].
5. Penelitian tentang klasifikasi kualitas udara DKI Jakarta yang focus pada faktor yang mempengaruhi kualitas udara [14].

2. Metode Penelitian

Penelitian ini adalah penelitian pustaka (*library research*). Penulis menggunakan algoritma C4.5 karena merupakan komponen dari algoritma data mining yang sesuai dengan jenis data dan pengetahuan yang akan diproses. Weka adalah software data mining yang digunakan.

Untuk menghasilkan penelitian yang berkualitas, setiap pelaksana penelitian harus menggunakan metode yang tepat dan sesuai. Penelitian diharapkan berjalan dengan sistematis dan ilmiah berdasarkan metode yang disusun. Berikut adalah kerangka penelitian yang dilakukan.



Gambar 1. Kerangka Penelitian

Tahapan penelitian yang dilakukan sebagai berikut :

1. **Penentuan Ruang Lingkup Penelitian**
Untuk memastikan bahwa penelitian terkonsentrasi pada masalah yang akan dibahas, langkah pertama adalah menentukan ruang lingkup dan batasan masalah.
2. **Analisis Masalah**
Analisis masalah dilakukan untuk menetapkan masalah penelitian atau batasannya. Proses penggunaan algoritma C4.5 untuk mengklasifikasikan data mahasiswa dilakukan pada tahap ini.
3. **Penentuan Tujuan**
Setelah masalah dipahami, tujuan penelitian ini ditetapkan. Tujuannya adalah untuk menjawab masalah-masalah sebelumnya. Seperti menggali tumpukan data mahasiswa untuk kemudian dihasilkan suatu pengetahuan.
4. **Kajian Literatur**
Untuk mencapai tujuan penelitian, berbagai referensi yang terkait dengan topik penelitian diperiksa. Referensi ini kemudian dipilih sesuai dengan kebutuhan penelitian dan dipilih dari jurnal ilmiah, buku, dan sumber internet yang relevan.
5. **Pengumpulan Data**
Untuk memahami masalah yang ada, proses pengumpulan data dimulai dengan melakukan observasi langsung pada objek penelitian. Studi kepustakaan dilakukan dengan membaca berbagai buku yang relevan dengan masalah penelitian. Data utama yang digunakan adalah mahasiswa.
6. **Penentuan Teknik yang digunakan**
Penentuan teknik pengolahan data adalah langkah yang dilakukan dalam mengolah data. Dalam penelitian ini, penulis menggunakan algoritma C4.5.
7. **Implementasi Sistem**
Selanjutnya, implementasi dilakukan pada *software data mining* Weka.
8. **Pengujian Sistem**
Untuk membandingkan perhitungan manual dan komputerisasi pada *software*, pengujian sistem dilakukan sebagai berikut:
 - a. Menguji data mahasiswa dengan formula algoritma C4.5.
 - b. Kemudian, menggunakan algoritma C4.5 sebagai metode analisis untuk menerapkan data pada *software* Weka.
 - c. Terakhir, membandingkan data yang diolah secara manual dan yang diolah secara komputerisasi. Jika hasilnya tidak jauh berbeda, hasilnya dapat dianggap benar.

3. Hasil Penelitian

3.1. Data Mahasiswa

Penulis menggunakan data mahasiswa yang telah pada Program Studi Manajemen UGKMB. Variabel yang digunakan adalah Nama Mahasiswa, IPK, SKS, Kehadiran, Beasiswa, Organisasi dan Status Kelulusan. Khusus nama mahasiswa pada tabel hanya diisi dengan huruf abjad dan tidak diproses dalam algoritma C4.5.

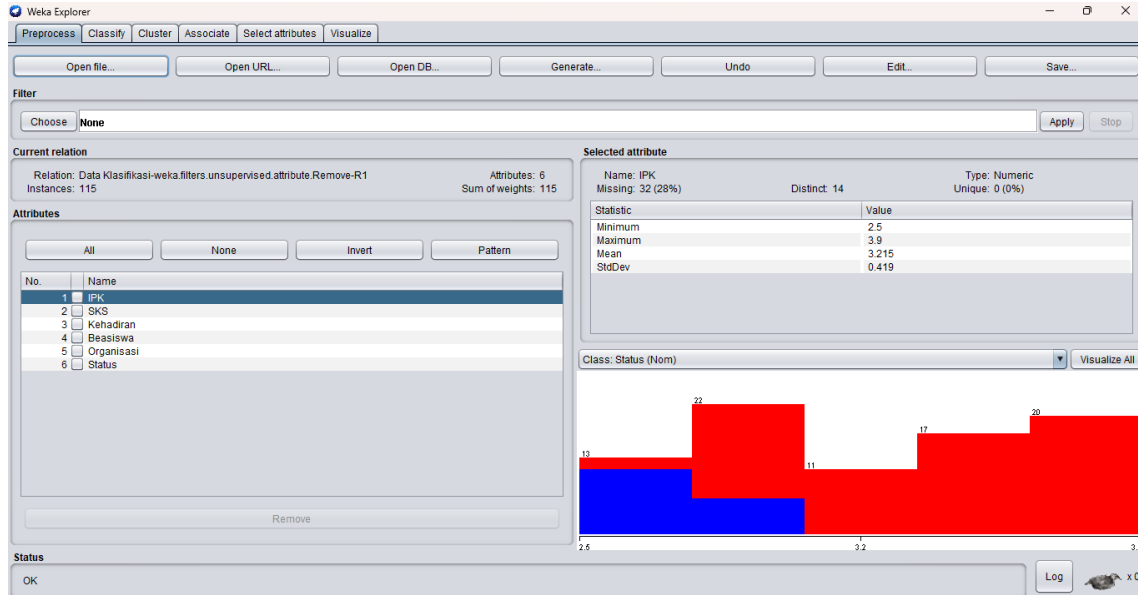
Tabel 1. Data Mahasiswa

No.	Nama	IPK	SKS	Kehadiran	Beasiswa	Organisasi	Status Kelulusan
1	A	2.5	126	Baik	Tidak	Ya	Tidak Tepat Waktu
2	B	3	137	Kurang	Tidak	Ya	Tepat Waktu
3	C	3.4	134	Baik	Tidak	Ya	Tepat Waktu
4	D	2.76	126	Cukup	Tidak	Ya	Tepat Waktu
5	E	2.89	128	Cukup	Ya	Ya	Tepat Waktu
6	F	3.4	132	Cukup	Ya	Tidak	Tepat Waktu
7	G	3	123	Baik	Tidak	Tidak	Tepat Waktu
8	H	3.8	127	Baik	Tidak	Tidak	Tepat Waktu
9	I	3.9	124	Kurang	Tidak	Tidak	Tepat Waktu
10	J	3.1	120	Kurang	Tidak	Ya	Tepat Waktu
11	K	2.54	139	Baik	Ya	Tidak	Tepat Waktu
12	L	3.77	132	Cukup	Tidak	Tidak	Tepat Waktu
13	M	3.4	123	Cukup	Tidak	Tidak	Tepat Waktu
14	N	3.21	139	Cukup	Tidak	Tidak	Tepat Waktu
15	O	3	123	Baik	Tidak	Tidak	Tepat Waktu
16	P	3.67	126	Baik	Tidak	Tidak	Tepat Waktu
17	Q	3.36	140	Baik	Tidak	Tidak	Tepat Waktu
18	R	2.84	128	Baik	Tidak	Ya	Tidak Tepat Waktu
19	S	2.5	121	Kurang	Ya	Ya	Tidak Tepat Waktu
20	T	3	140	Baik	Ya	Ya	Tepat Waktu
21	U	3.4	140	Baik	Ya	Ya	Tepat Waktu
22	V	2.76	121	Baik	Tidak	Ya	Tidak Tepat Waktu
23	W	2.89	136	Baik	Tidak	Ya	Tidak Tepat Waktu
24	X	3.4	121	Cukup	Tidak	Tidak	Tepat Waktu
25	Y	3	128	Cukup	Ya	Tidak	Tepat Waktu
26	Z	3.8	123	Baik	Tidak	Tidak	Tepat Waktu
27	AA	3.9	130	Baik	Tidak	Tidak	Tepat Waktu
83	AB	3.1	134	Baik	Tidak	Tidak	Tepat Waktu

Tabel 1 di atas menyajikan data mahasiswa dengan atribut IPK, SKS, Kehadiran, Beasiswa, Organisasi, dan Status Kelulusan sebagai target. Data tersebut akan diproses lebih lanjut dengan menggunakan fungsi algoritma C4.5.

3.2. Analisis Algoritma C4.5

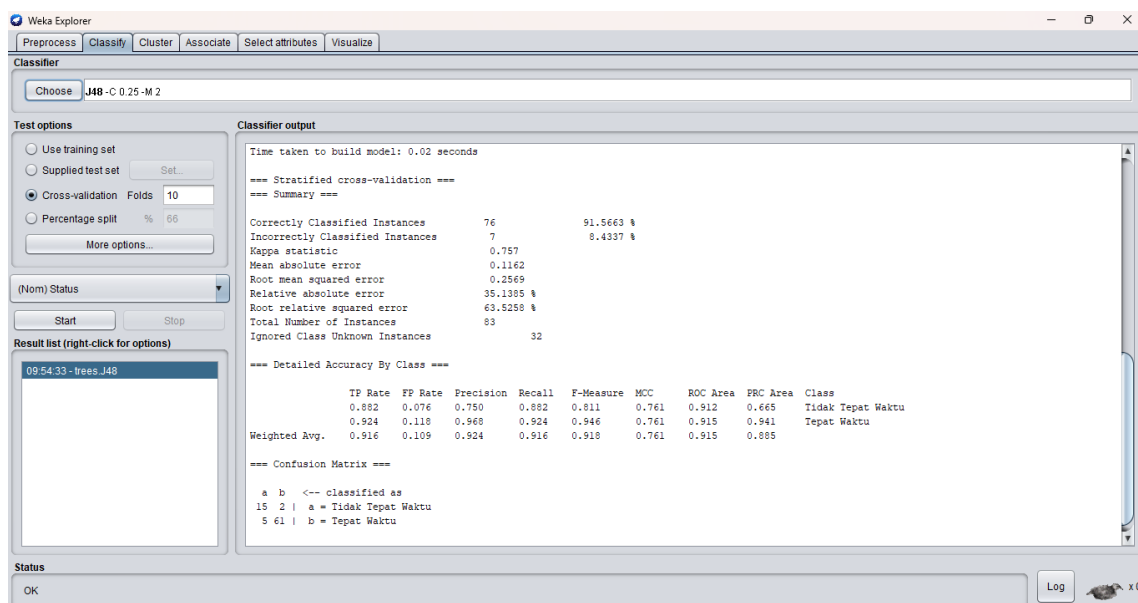
Data pada tabel 1 di atas, terlebih dahulu dibuat dalam format CSV (comma delimited). Tampilkan data pada software Weka sebagaimana gambar berikut.



Gambar 2. Input Data Mahasiswa

Pada gambar 2 di atas, enam variabel ditampilkan dan satu variabel target yaitu status, yang akan menjadi Keputusan apakah lulus tepat waktu atau tidak tepat waktu.

Kemudian pilih menu Classify untuk menerapkan fungsi klasifikasi. Terdapat beberapa algoritma, namun dalam hal ini pilih J48 sebagai implementasi dari algoritma C4.5, sebagaimana gambar berikut.

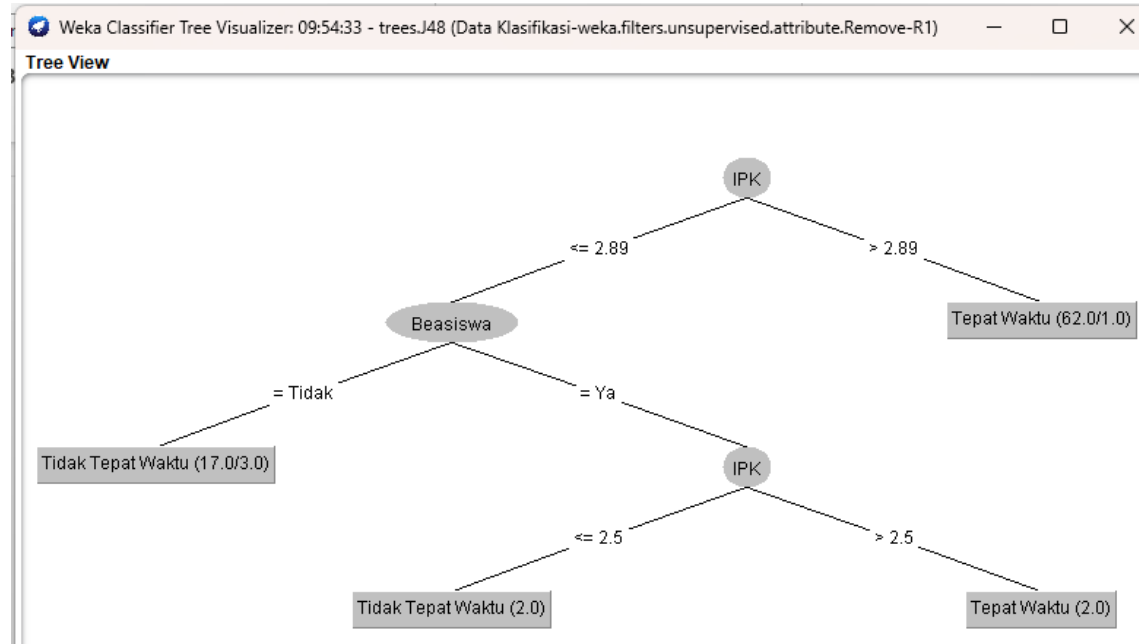


Gambar 3. Implementasi Klasifikasi Algoritma C4.5

Pada gambar di atas, pada bagian correctly classified instance sebesar 91,5663% yang menunjukkan tingkat akurasi data klasifikasi. Kappa statistic sebesar 0,757 merupakan alat ukur kesepakatan prediksi dan

kebenaran, semakin mendekati 1 (satu) maka semakin baik. Pada bagian confusion matrix terlihat bahwa 15 (lima belas) orang mahasiswa tidak tepat waktu dan masuk klasifikasi tidak tepat waktu (benar). 2 (dua) orang tidak tepat waktu namun masuk klasifikasi tepat waktu (salah). 5 (lima) orang tidak tepat waktu namun masuk klasifikasi tepat waktu (salah). 61 (enam puluh satu) orang akan tepat waktu dan masuk klasifikasi tepat waktu (benar).

Kemudian buat pohon Keputusan berdasarkan gambar sebelumnya sebagaimana gambar berikut.



Gambar 4. Pohon Keputusan Algoritma C4.5

Berdasarkan gambar 4 diatas, terlihat variabel IPK sebagai cabang pertama diikuti dengan variabel beasiswa. Jika $IPK > 2,89$ maka akan tepat waktu menyelesaikan studinya. Jika $IPK \leq 2,89$ dan tidak mendapat beasiswa, maka tidak akan tepat waktu menyelesaikan studinya. Jika mendapat beasiswa dan $IPK \leq 2,5$ maka tidak akan tepat waktu menyelesaikan studinya. Jika mendapat beasiswa dan $IPK > 2,5$ maka akan tepat waktu menyelesaikan studinya.

4. Penutup

Berdasarkan pembahasan di atas, dengan menggunakan data mahasiswa yang berjumlah 83 (delapan puluh tiga) orang berhasil mengklasifikasikan mahasiswa baik tepat waktu maupun yang tidak tepat waktu menyelesaikan studinya. Tingkat akurasi yang tinggi yakni mencapai 91,5663% terbukti dari 61 (enam puluh satu) orang mahasiswa diprediksi akan tepat waktu menyelesaikan studinya dan masuk kategori benar. Artinya mahasiswa tersebut hampir dipastikan menyelesaikan studinya tepat waktu. Sementara 15 (lima belas) orang dipastikan tidak akan tepat waktu menyelesaikan studinya.

Daftar Pustaka

- [1] Yusuf Maulana, Riki Winanjaya, and Fitri Rizki, "Penerapan Data Mining dengan Algoritma C4.5 Dalam Memprediksi Penjualan Tempe," *Bull. Comput. Sci. Res.*, vol. 2, no. 2, pp. 53–58, 2022, doi: 10.47065/bulletincsr.v2i2.163.
- [2] I. Anggraeni and S. Andriani, "3562-9606-1-Pb," vol. 18, no. 2, pp. 62–68, 2021.
- [3] B. Hasmaulina, "Penerapan Data Mining Untuk Membentuk Kelompok Belajar Menggunakan Metode Clustering Di SMK Negeri 3 Seluma," *JUKOMIKA (Jurnal Ilmu Komput. dan Inform.)*, vol. 4, no. 2, pp. 57–71, 2022, doi: 10.54650/jukomika.v4i2.368.
- [4] G. Gustientiedina, M. H. Adiya, and Y. Desnelita, "Penerapan Algoritma K-Means Untuk Clustering Data Obat-Obatan," *J. Nas. Teknol. dan Sist. Inf.*, vol. 5, no. 1, pp. 17–24, 2019, doi:

- 10.25077/teknosi.v5i1.2019.17-24.
- [5] N. T. Suswanto, P. Chyan, and V. Putri, "Using K-Means Algorithm to Investigate Community Behavior in Treating Waste toward Smart City," vol. 11, no. 4, 2021.
 - [6] A. Anas, Azwar; Jermawinsyah Zebua, "Implementasi Data Mining Menggunakan Algoritma Apriori Untuk Meningkatkan Pola Penjualan Obat," *J. Ilm. MEDIA SISFO*, vol. 16, no. 1, pp. 54–61, 2022, doi: 10.35957/jatisi.v7i2.195.
 - [7] B. S. Hasugian, "Penerapan Metode Association Rule Untuk Menganalisa Pola Pemakaian Bahan Kimia Di Laboratorium Menggunakan Algoritma FP-Growth (Studi Kasus di Laboratorium Kimia PT . PLN (Persero) Sektor Pembangkitan Belawan Medan) Buyung Solihin Hasugian Universitas," *J. Ilmu Komput. dan Inform.*, vol. 6341, no. November, pp. 56–69, 2019.
 - [8] D. Ai, H. Pan, X. Li, Y. Gao, and D. He, "Association rule mining algorithms on high-dimensional datasets," *Artif. Life Robot.*, vol. 23, no. 3, pp. 420–427, 2018, doi: 10.1007/s10015-018-0437-y.
 - [9] A. Prasetyo and A. Wahyono, "Implementasi Algoritma C4.5 dalam Klasifikasi Kondisi Ekonomi Warga Kabupaten Boyolali," *Gener. J.*, vol. 8, no. 2, pp. 121–131, 2024, doi: 10.29407/gj.v8i2.22827.
 - [10] A. Iftitah and R. Setyadi, "Penerapan Algoritma C4.5 Untuk Analisis Pengadaan Peralatan dan Mesin Kantor," *J. Inf. Syst. Res.*, vol. 4, no. 2, pp. 434–442, 2023, doi: 10.47065/josh.v4i2.2673.
 - [11] R. A. A. Sri Lestari, "Penerapan Algoritma C4.5 Pada Klasifikasi Status Gizi Balita di Posyandu Desa Sukalilah Cibatuk Kabupaten Garut Jawa Barat," *J. Wahana Pendidik.*, vol. 8, no. 10, pp. 116–125, 2022.
 - [12] F. Solikhah, M. Febianah, A. L. Kamil, W. A. Arifin, and Shelly Janu Setyaning Tyas, "Analisis Perbandingan Algoritma Naive Bayes Dan C4.5 Dalam Klasifikasi Data Mining Untuk Memprediksi Kelulusan," *Tematik*, vol. 8, no. 1, pp. 96–103, 2021, doi: 10.38204/tematik.v8i1.576.
 - [13] R. E. Pambudi, Sriyanto, and Firmansyah, "Klasifikasi Penyakit Stroke Menggunakan Algoritma Decision Tree C4.5," *Ijccs*, vol. x, No.x, no. x, pp. 1–5, 2022.
 - [14] M. Astriyani, I. N. Laela, D. P. Lestari, L. Anggraeni, and T. Astuti, "Analisis Klasifikasi Data Kualitas Udara DKI Jakarta Menggunakan Algoritma C4.5," *JuSiTik J. Sist. dan Teknol. Inf. Komun.*, vol. 6, no. 1, pp. 36–41, 2023, doi: 10.32524/jusitik.v6i1.790.