Journal of Information System Research (JOSH)
Volume 5, No. 4, Juli 2024, pp 1042−1052
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v5i4.5643

Decision Tree C4.5 dengan Teknik Information Gain Untuk Klasifikasi Pemilihan Program Studi Tingkat Lanjut

Teddy Yogi Pratama*, Armansyah

Fakultas Sains dan Teknologi, Program Studi Ilmu Komputer, Universitas Islam Negeri Sumatera Utara, Medan
Jl. William Iskandar Ps. V, Medan Estate, Kec. Percut Sei Tuan, Kabupaten Deli Serdang, Sumatera Utara, Indonesia
Email: [1,*]eddyyogipratama@gmail.com, [2]armansyah@uinsu.ac.id
Email Penulis Korespondensi: eddyyogipratama@gmail.com
Submitted: **20/07/2024;** Accepted: **23/07/2024;** Published: **24/07/2024**

Abstrak−Tujuan penelitian ini adalah Untuk manganalisis penerapan fitur yang informatif, mengklasifikasikan data berdasarkan fitur akademis, minat dan bakat dengan Teknik Information menggunakan Decision Tree C4.5. tujuan penelitian ini adalah melakukan riset terhadap siswa dalam menentukan pemilihan program studi untuk melanjutkan pendidikan ke perguruan tinggi, dikarenakan Dalam memilih program studi untuk melanjutkan pendidikan ke perguruan tinggi, siswa sering kali mengalami kesulitan dalam menentukan pilihan program studi yang akan di pilihnnya. Penelitin mengumpulkan 140 data siswa, dengan mendistribusikan kuesioner kepada calon mahasiswa baru dan meminta nilai akademik siswa terhadap pihak sekolah, penulis memiliki 140 data yang akan digunakan dalam penelitian ini. Selanjutnya, dari 140 data tersebut, peneliti akan membaginya menjadi dua bagian, yaitu data training sebanyak 118 data dan data testing sebanyak 22 data untuk memenuhi kebutuhan dalam merancang model. Berdasarkan hasil penelitian yang dilakukan dengan menggunakan pendekatan Supervised Learning Decision Tree C4.5 dan menerapkan teknik Information Gain untuk klasifikasi pemilihan program studi tingkat lanjut, diperoleh akurasi sebesar 86%. Tingkat keberhasilan ini menunjukkan bahwa metode tersebut efektif dalam mengidentifikasi dan mengklasifikasikan program studi tingkat lanjut. Hal ini mengindikasikan bahwa penggunaan Decision Tree C4.5 yang memanfaatkan teknik Information Gain memiliki potensi besar sebagai model yang dapat membantu siswa dalam memilih program studi tingkat lanjut mereka dengan tingkat keakuratan yang memuaskan. Dengan hasil akurasi yang tinggi, metode ini dapat diandalkan untuk memberikan prediksi yang akurat dalam konteks pemilihan program studi.

Kata Kunci: Information Gain; Decision Tree C4.5; Klasifikasi

Abstract−The aim of this research is to analyze the application of informative features, classify data based on academic features, interests and talents with Information Techniques using Decision Tree C4.5. The aim of this research is to conduct research on students in determining the choice of study program to continue their education to college, because in choosing a study program to continue their education to college, students often experience difficulties in determining which study program they will choose. The research collected 140 student data, by distributing questionnaires to prospective new students and asking the school for students' academic scores, the author has 140 data that will be used in this research. Next, from the 140 data, researchers will divide it into two parts, namely 118 training data and 22 testing data to meet the needs in designing the model. Based on the results of research conducted using the Supervised Learning Decision Tree C4.5 approach and applying the Information Gain technique for classification of advanced study program selection, an accuracy of 86% was obtained. This success rate shows that the method is effective in identifying and classifying advanced study programs. This indicates that the use of Decision Tree C4.5 which utilizes the Information Gain technique has great potential as a model that can assist students in choosing their advanced study program with a satisfactory level of accuracy. With high accuracy results, this method can be relied on to provide accurate predictions in the context of study program selection.

Keywords: Information Gain; Decision Tree C4.5; Classification

1. PENDAHULUAN

Dalam memilih program studi untuk melanjutkan pendidikan ke perguruan tinggi, siswa sering kali mengalami kesulitan untuk mengidentifikasi minat dan bakat mereka. Banyak siswa tidak tahu apa yang sebenarnya mereka minati atau di mana bakat mereka berada. Situasi ini menjadi lebih sulit karena kurangnya alat evaluasi yang dapat membantu mereka menilai diri secara objektif dan komprehensif. Selain itu, nilai akademik juga sangat mempengaruhi keputusan mereka. Banyak siswa merasa tertekan untuk memilih program studi berdasarkan nilai akademik saja, tanpa mempertimbangkan apakah program tersebut sesuai dengan minat dan bakat mereka[1].

Machine learning adalah ilmu yang memungkinkan komputer berperilaku seperti manusia, di mana komputer dapat meningkatkan pemahamannya melalui pengalaman atau dengan berjalannya waktu secara otomatis [2][3]. Machine learning banyak digunakan untuk pengembangan model analisis data dengan pendekatan belajar mesin. Beberapa teknik dalam machine learning meliputi Supervised Learning, yang merupakan pendekatan klasifikasi di mana setiap data dalam kumpulan data memiliki label yang ditentukan, memungkinkan model untuk mengklasifikasikan kelas yang tidak diketahui. Unsupervised Learning, di sisi lain, sering disebut sebagai teknik clustering karena tidak memerlukan label pada data, dan hasilnya tidak mengidentifikasi contoh dalam kelas yang telah ditentukan sebelumnya. Sementara itu, Reinforcement Learning bekerja dalam lingkungan dinamis di mana model harus mencapai tujuan tanpa pemberitahuan eksplisit dari komputer ketika tujuan tersebut tercapai [4][5].

Pohon keputusan merupakan salah satu contoh pendekatan supervised learning yang menggunakan Decision Tree C4.5. Tujuan utamanya adalah untuk mengklasifikasikan dan memprediksi masa depan berdasarkan data yang ada. Algoritma C4.5 adalah pendekatan untuk membuat Decision Tree dari data pelatihan, merupakan

Journal of Information System Research (JOSH)
Volume 5, No. 4, Juli 2024, pp 1042−1052
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v5i4.5643

pengembangan dari metode ID3 [6][7]. Klasifikasi adalah proses untuk mengidentifikasi pola-pola yang mewakili dan membedakan antara kelas-kelas data [8]. Dalam konteks pengklasifikasian menggunakan decision tree, terdapat beberapa teknik yang umum digunakan, antara lain information gain, gain ratio, dan Gini index, yang bertujuan untuk menentukan pohon atau node awal [9][10][11].

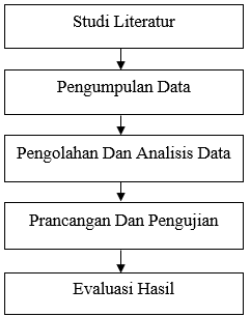
Penelitian ini bertujuan untuk menganalisis Teknik Information Gain pada pilihan keputusan Decision Tree untuk klasifikasi pemilihan program studi tingkat lanjut. Beberapa penelitian yang telah dilakukan oleh sejumlah peneliti yang memanfaatkan teknik Information Gain untuk melakukan klasifikasi, dikutip dalam studi yang memiliki permasalahan mengenai Pemilihan jurusan pada Bidang IT dengan Menggunakan Algoritma C4.5 menunjukkan bahwa penggunaan algoritma C4.5 memberikan efisiensi yang tinggi dalam proses pemilihan jurusan. Melalui pendekatan ini, dari total 301 rekaman data yang dianalisis, sebanyak 215 data diujikan. Dengan menggunakan teknik Information Gain, dengan menggunakan beberapa atribut diantaranya Pilihan jurusan, Software>=80, Jaringan >=80, Software >= Jaringan, penelitian mendapat hasil yang signifikan Hasil pengujian menunjukkan tingkat akurasi yang mencapai 96,73%, menggambarkan keefektifan metode tersebut dalam membantu pengambilan keputusan terkait penjurusan bidang IT di SMK N Manonjaya [12].

Penelitian menggunakan algoritma C4.5 dengan kriteria pemisahan berdasarkan Information Gain, untuk memilih penerima bantuan pangan non tunai, dengan menggunakan algoritma C4.5, dataset yang didapat pada tahun 2018 pada Desa Nagrak Utara Sukabumi sebanyak 130 data. klasifikasi yang dihasilkan sangat efektif dengan tingkat akurasi mencapai 91.54% [13]. Penelitian ini diharap dapat membantu siswa dalam klasifikasian pemilihan terhadap program studi tingkat lanjut terhadap siswa yang memiliki kebimbangan terhadap pilihan studi lanjutnya khususnya pada sekolah SMAS Nurul Iman.

2. METODOLOGI PENELITIAN

2.1 Kerangka Penelitian

Dalam pengembangan sebuah penelitian, struktur yang terorganisir sangat penting. Ini melibatkan pembuatan sebuah model penelitian yang menggambarkan hubungan antara variabel yang sedang diteliti. Dalam penelitian ini, pendekatan kuantitatif digunakan untuk menguji teori dengan analisis mendalam. Kerangka penelitian mencakup langkah-langkah yang diambil selama proses penelitian untuk memastikan keberlangsungan yang sistematis dan pencapaian tujuan yang diinginkan [12][14][15], dapat dilihat pada gambar 1 untuk kerangka penelitian.



Gambar 1. Kerangka Penelitian

1. **Studi Literatur:** Proses awal penelitian ini dimulai dengan studi literatur. Setelah meneliti beberapa karya tulis, diketahui bahwa penerapan Information Gain dengan Decision Tree C4.5 dalam klasifikasi memiliki signifikansi yang besar. Information Gain berperan penting dalam mengukur kepentingan relatif dari setiap fitur dalam memisahkan data. Fitur informasi yang signifikan akan dipilih untuk membangun model klasifikasi yang efisien. Penelitian ini menggunakan metode kuantitatif untuk menganalisis data numerik dan kuesioner yang dikumpulkan untuk membangun sebuah model klasifikasi.
2. **Pengumpulan Data:** Dalam penelitian ini, mengunakan Teknik triangulasi data atau penggabungan data, pengumpulan data melibatkan beberapa pendekatan, pertama data diperoleh melalui rekam akademik siswa seperti riwayat nilai, kedua penggunaan kuisoner akan menjadi instrumen tambahan. Kuisoner akan dirancang secara cermat dengan pertanyaan yang relevan dan akan di berikan untuk mengetahui minat bakat siswa. Melalui kombinasi teknik ini, diharapkan peneliti dapat memperoleh pemahaman yang komprehensif tentang faktor-faktor yang memengaruhi keputusan pemilihan program studi siswa, kuisoner dapar dilihat pada tabel 1.

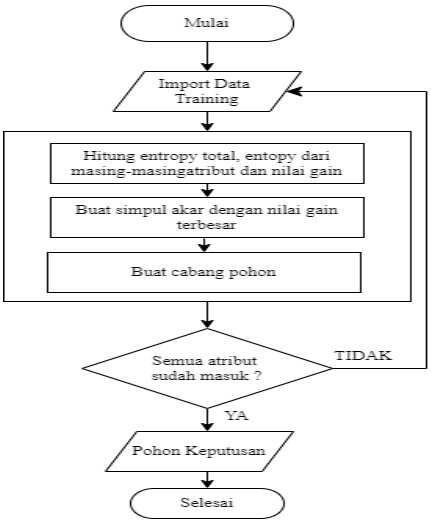
Tabel 1. Kuisoner

No	Pertanyaan	Nilai
Q1	Jenis kelamin Anda?	Pria,Wanita
Q2	Bagaimana Anda menilai status sosioekonomi Anda?	Rendah,Sedang, Tinggi

Journal of Information System Research (JOSH)
Volume 5, No. 4, Juli 2024, pp 1042−1052
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v5i4.5643

No	Pertanyaan	Nilai
Q3	Bidang studi mana yang paling Anda minati?	Ilmu alam, Ilmu social, Seni dan humaniora
Q4	Kegiatan ekstrakurikuler mana yang Anda pilih?	Olahraga, Seni dan music, Organisasi osial
Q5	Bidang karier mana yang paling Anda minati?	Teknologi dan IT, Bisnis dan manajemen
Q6	Apakah Anda pernah / sedang mengikuti program akademik khusus/kursus?	Ya, Tidak
Q7	Anda pernah mengikuti kegiatan ilmiah di luar kurikulum sekolah?	Ya, Tidak
S	program studi pilihan anda?	Sosial, Sains

3. **Pengelolahan dan Analisis data:** Dalam penelitian ini, penulis mengumpulkan 140 data siswa, dengan mendistribusikan kuesioner kepada calon mahasiswa baru dan meminta nilai akademik siswa terhadap pihak sekolah, penulis memiliki 140 data yang akan digunakan dalam penelitian ini. Selanjutnya, dari 140 data tersebut, kami akan membaginya menjadi dua bagian, yaitu data training dan data testing, dimana data trening sebanyak 118 data, dan data testing sebanyak 22 data untuk memenuhi kebutuhan dalam merancang model. Transformasi data Diskritisasi adalah proses mengubah variabel kontinu menjadi variabel diskrit dengan membagi rentang nilai menjadi kategori diskrit. Ini membantu mengurangi kompleksitas model dan memfasilitasi penggunaan variabel kontinu dalam pembangunan pohon keputusan. Dengan kata lain, diskritisasi membagi variabel kontinu menjadi kategori, membuatnya lebih mudah diinterpretasikan oleh algoritma pembelajaran mesin seperti pohon keputusan. Misalnya, nilai dapat dibagi menjadi kategori seperti "cukup", "baik", dan "sangat baik" berdasarkan rentang nilai tertentu. Information Gain adalah teknik penting dalam algoritma Decision Tree C4.5 yang membantu memilih atribut terbaik untuk menjadi node pemisah. Ini berfokus pada seberapa baik atribut tersebut membagi data menjadi subset yang lebih homogen berdasarkan label kelas. Langkah-langkahnya mencakup menghitung entropi sebelum dan setelah pemisahan, dan kemudian menghitung perbedaan antara keduanya untuk mendapatkan Information Gain.
4. **Perancangan dan Pengujian Metode:** Dalam penelitian ini, menggunakan metode klasifikasi dengan algoritma C4.5 dan melakukan pengujian dengan menggunakan Confusion Matrix. Tahapan awal dalam pemodelan algoritma ini melibatkan persiapan data training dengan membagi dataset menjadi dua bagian, di mana 118 dari data digunakan sebagai data training dan 22 sebagai data testing. Setelah itu, langkah-langkah berikutnya adalah menentukan akar atau root dari pohon keputusan. Proses ini dimulai dengan menghitung nilai entropy total dan nilai entropy dari setiap atribut yang digunakan. Kemudian, dilakukan perhitungan nilai gain. Langkah-langkah ini diulangi hingga semua rekaman terpartisi dengan baik, dan simpul dengan nilai gain terbesar dibuat sebagai simpul pada pohon keputusan. Proses partisi pohon keputusan dapat dihentikan jika beberapa kondisi terpenuhi:
   1) semua rekaman dalam simpul N mendapatkan kelas yang sama.
   2) tidak ada atribut yang perlu dipartisi lebih lanjut.
   3) tidak ada rekaman di dalam cabang yang kosong.



Gambar 2. Flowchart Decision Tree

   1) Menghitung total data yang ada.
   2) Memilih atribut sebagai titik awal dalam pembentukan cabang.
   3) Membuat cabang untuk setiap anggota dari atribut yang dipilih.

Journal of Information System Research (JOSH)
Volume 5, No. 4, Juli 2024, pp 1042−1052
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v5i4.5643

4) Jika nilai entropy dari anggota atribut tersebut adalah nol, maka node tersebut dijadikan daun dalam pohon keputusan. Jika seluruh nilai entropy dari anggota node adalah nol, proses berhenti.
5) Jika terdapat node yang memiliki nilai entropy lebih besar dari nol, maka proses diulangi dari awal hingga semua anggota node memiliki nilai entropy nol [16][17][18][19].

5. **Evaluasi Hasil:** Evaluasi kinerja model menggunakan Confusion Matrix bertujuan untuk memperjelas sejauh mana prediksi model sesuai dengan keadaan sebenarnya. Validitas prediksi perlu diuji agar hasilnya dapat dipercaya. Confusion Matrix adalah tabel matriks yang memuat informasi tentang kelas positif dan negatif, membantu dalam memahami kualitas prediksi model [20].

Tabel 2. Confusion Matrix

Kelas Prediksi	Aktual Positif	Aktual Negative
Prediksi Positif	True Positive (TP)	False Positive (FP)
Prediksi Negatif	False Negative (FN)	True Positive (TN)

6. **Lokasi penelitian:** Penelitian dilakukan di SMA Nurul Iman untuk mengumpulkan data nilai akademik serta informasi tentang minat, bakat, dan latar belakang ekonomi siswa. Data akademik bisa diperoleh dari catatan sekolah, tetapi informasi tambahan tersebut hanya bisa didapat melalui kuesioner langsung kepada siswa. Demikian, lokasi penelitian dipilih karena memberikan akses yang sesuai untuk mendapatkan data yang diperlukan.
7. **Waktu Penelitian:** Setelah data terkumpul dan diolah, bulan kedua akan digunakan untuk merumuskan temuan dan hasil analisis ke dalam sebuah jurnal ilmiah. Proses penulisan jurnal akan melibatkan penyusunan laporan penelitian yang komprehensif serta pembahasan mendalam mengenai implikasi temuan bagi pemilihan program studi di SMA. Diharapkan bahwa pada akhir bulan kedua, jurnal ilmiah akan siap untuk diterbitkan atau diserahkan ke jurnal ilmiah terkemuka dalam bidang pendidikan.

3. HASIL DAN PEMBAHASAN

3.1 Analisa data

Adapun data yang digunakan dalam membentuk sebuah pohon keputusan adalah berupa data nilai akademik, minat, bakat, dan tingkatan ekonomi siswa yang dapat kita lihat pada tabel di bawah. Setelah data tersebut diolah, data tersebut kemudian dibagi menjadi data pelatihan (training data) dan data pengujian (testing data). Data pelatihan digunakan untuk membangun model pohon keputusan, sedangkan data pengujian digunakan untuk mengevaluasi kinerja model yang telah dibangun. Data setelah di olah dan diset menjadi data training pada tabel 3 dan 4 dan data testing pada tabel 5 dan 6

Tabel 3. Data Training

NO	PAI	PPKN	B.INDO	MM	SEJARAH	B.INGG	SENI	PENJAS
1	SANGAT BAIK	SANGAT BAIK	SANGAT BAIK	SANGAT BAIK	SANGAT BAIK	SANGAT BAIK	SANGAT BAIK	SANGAT BAIK
2	SANGAT BAIK	SANGAT BAIK	SANGAT BAIK	SANGAT BAIK	SANGAT BAIK	BAIK	SANGAT BAIK	SANGAT BAIK
3	SANGAT BAIK	SANGAT BAIK	SANGAT BAIK	SANGAT BAIK	SANGAT BAIK	SANGAT BAIK	SANGAT BAIK	SANGAT BAIK
4	SANGAT BAIK	SANGAT BAIK	SANGAT BAIK	SANGAT BAIK	SANGAT BAIK	SANGAT BAIK	SANGAT BAIK	SANGAT BAIK
……	……	……	……	……	……	……	……	……
118	SANGAT BAIK	SANGAT BAIK	SANGAT BAIK	BAIK	SANGAT BAIK	SANGAT BAIK	SANGAT BAIK	BAIK

Tabel 4. Lanjutan Data Training

NO	1	2	3	4	5	6	7	8
1	PRIA	SEDANG	SENI DAN HUMANIORA	ORGANISASI SOSIAL	BISNIS DAN MANAJEMEN	TIDAK	TIDAK	sosial
2	WANITA	SEDANG	SENI DAN HUMANIORA	OLAHRAGA	BISNIS DAN MANAJEMEN	YA	TIDAK	sosial
3	WANITA	SEDANG	ILMU ALAM	OLAHRAGA	BISNIS DAN MANAJEMEN	TIDAK	TIDAK	sains

Journal of Information System Research (JOSH)
Volume 5, No. 4, Juli 2024, pp 1042−1052
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v5i4.5643

NO	1	2	3	4	5	6	7	8
4	WANITA	SEDANG	ILMU ALAM	SENI DAN MUSIK	TEKNOLOGI DAN IT	TIDAK	TIDAK	sains
…	……	……	……	……	……	……	……	……
…								
118	WANITA	RENDAH	SENI DAN HUMANIORA	ORGANISASI SOSIAL	BISNIS DAN MANAJEMEN	TIDAK	TIDAK	sains

Tabel 5. Data Testing

NO	PAI	PPKN	B.INDO	M.M	SEJARAH	B.INGG	SENI	PENJAS
1	SANGAT BAIK	SANGAT BAIK	SANGAT BAIK	SANGAT BAIK	SANGAT BAIK	SANGAT BAIK	SANGAT BAIK	SANGAT BAIK
2	SANGAT BAIK	BAIK	BAIK	BAIK	BAIK	BAIK	BAIK	BAIK
3	SANGAT BAIK	SANGAT BAIK	SANGAT BAIK	SANGAT BAIK	SANGAT BAIK	SANGAT BAIK	SANGAT BAIK	SANGAT BAIK
4	SANGAT BAIK	CUKUP	BAIK	BAIK	BAIK	BAIK	SANGAT BAIK	SANGAT BAIK
……	……	……	……	……	……	……	……	……
22	BAIK	BAIK	BAIK	BAIK	BAIK	BAIK	BAIK	BAIK

Tabel 6. Lanjutan Data Testing

NO	1	2	3	4	5	6	7	8
1	WANITA	SEDANG	SENI DAN HUMANIORA	SENI DAN MUSIK	BISNIS DAN MANAJEMEN	TIDAK	TIDAK	sosial
2	WANITA	SEDANG	ILMU SOSIAL	OLAHRAGA	BISNIS DAN MANAJEMEN	TIDAK	TIDAK	sains
3	PRIA	SEDANG	ILMU SOSIAL	OLAHRAGA	BISNIS DAN MANAJEMEN	TIDAK	TIDAK	sosial
4	PRIA	SEDANG	ILMU ALAM	OLAHRAGA	TEKNOLOGI DAN IT	TIDAK	TIDAK	sains
…	……	……	……	……	……	……	……	……
…								
22	PRIA	SEDANG	ILMU SOSIAL	OLAHRAGA	BISNIS DAN MANAJEMEN	TIDAK	TIDAK	sosial

3.2 Analisis menggunakan algoritma C4.5

Adapun proses analisis dimulai dengan menghitung entropi awal dari dataset, yang berfungsi untuk mengukur tingkat ketidakpastian atau heterogenitas dalam data sebelum dilakukan pemisahan. Selanjutnya, information gain dihitung untuk setiap atribut. Information gain ini mengukur seberapa besar penurunan entropi yang terjadi setelah data dipisahkan berdasarkan atribut tersebut. Atribut dengan nilai information gain tertinggi dipilih sebagai node internal pada pohon keputusan, karena atribut ini dianggap paling informatif dalam memisahkan data ke dalam kelas-kelas yang berbeda.

Berikut adalah perhitungan manual untuk mencari Node 1:

Menghirung entropy

$$\text{Entropy(total)} = \sum_{i=1}^{n} -\frac{58}{103}\log_2\left(\frac{58}{118}\right) - \frac{60}{118}\log_2\left(\frac{60}{118}\right) = 0.997928$$

$$\text{Entropy(PAI(Sangat Baik))} = \sum_{i=1}^{n} -\frac{54}{105}\log_2\left(\frac{54}{105}\right) - \frac{51}{105}\log_2\left(\frac{51}{105}\right) = 0.9994111$$

$$\text{Entropy(PAI(Baik))} = \sum_{i=1}^{n} -\frac{4}{13}\log_2\left(\frac{4}{13}\right) - \frac{9}{13}\log_2\left(\frac{9}{13}\right) = 0.8904916$$

$$\text{Entropy(PKN(Sangat Baik))} = \sum_{i=1}^{n} -\frac{32}{67}\log_2\left(\frac{32}{67}\right) - \frac{35}{67}\log_2\left(\frac{35}{67}\right) = 0.9985533$$

$$\text{Entropy(PKN(Baik))} = \sum_{i=1}^{n} -\frac{26}{50}\log_2\left(\frac{26}{50}\right) - \frac{24}{50}\log_2\left(\frac{24}{50}\right) = 0.9988455$$

Journal of Information System Research (JOSH)
Volume 5, No. 4, Juli 2024, pp 1042−1052
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v5i4.5643

$$\text{Entropy}(\text{PKN}(\text{Cukup})) = \sum_{i=1}^{n} -\frac{0}{1}\log_2\left(\frac{0}{1}\right) - \frac{1}{1}\log_2\left(\frac{1}{1}\right) = 0$$

$$\text{Entropy}(\text{B. Indo}(\text{Sangat Baik})) = \sum_{i=1}^{n} -\frac{41}{81}\log_2\left(\frac{41}{81}\right) - \frac{40}{81}\log_2\left(\frac{40}{81}\right) = 0.9998901$$

$$\text{Entropy}(\text{B. Indo}(\text{Baik})) = \sum_{i=1}^{n} -\frac{17}{36}\log_2\left(\frac{17}{36}\right) - \frac{19}{36}\log_2\left(\frac{19}{36}\right) = 0.9977725$$

$$\text{Entropy}(\text{B. Indo}(\text{Cukup})) = \sum_{i=1}^{n} -\frac{0}{1}\log_2\left(\frac{0}{1}\right) - \frac{1}{1}\log_2\left(\frac{1}{1}\right) = 0$$

Menghitung nilai Gain information

$$\text{Gain}(\text{PAI}) = 0.9997928 - \left(\frac{105}{118}\,0.9994111 + \frac{13}{118}\,0.8904916\right) = 0.012381$$

$$\text{Gain}(\text{PPKN}) = 0.9997928 - \left(\frac{67}{118}\,0.9985533 + \frac{50}{118}\,0.9988455\right) = 0.0095780$$

$$\text{Gain}(\text{B. INDO}) = 0.9997928 - \left(\frac{81}{118}\,0.9998901 + \frac{36}{118}\,0.9977725\right) = 0.0090224$$

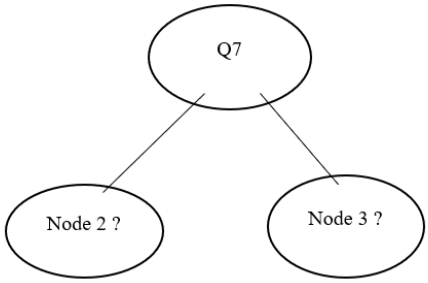
Tabel 7. Hasil Perhirungan Dari Node 1

NODE		Keterangan	Jumlah Kasus	Main SAINS	Main SOSIAL	Entropy	Gain
1		TOTAL	118	58	60	0.9997928	
	PAI	SANGAT BAIK	105	54	51	0.9994111	
		BAIK	13	4	9	0.8904916	
							0.0123813
	PPKN	SANGAT BAIK	67	32	35	0.9985533	
		BAIK	50	26	24	0.9988455	
		CUKUP	1	0	1	#NUM!	
							0.0095780
	B.INDO	SANGAT BAIK	81	41	40	0.9998901	
		BAIK	36	17	19	0.9977725	
		CUKUP	1	0	1	#NUM!	
							0.0090224
	MM	SANGAT BAIK	77	37	40	0.9989047	
		BAIK	40	21	19	0.9981959	
		CUKUP	1	0	1	#NUM!	
							0.0095936
	SEJARAH	SANGAT BAIK	99	47	52	0.9981592	
		BAIK	18	11	7	0.9640788	
		CUKUP	1	0	1	#NUM!	
							0.0152912
	B.ING	SANGAT BAIK	81	39	42	0.9990103	
		BAIK	36	19	17	0.9977725	
		CUKUP	1	0	1	#NUM!	
							0.0096263
	SENI	SANGAT BAIK	88	44	44	1.0000000	
		BAIK	30	14	16	0.9967916	
							0.0006085
	PENJAS	SANGAT BAIK	82	41	41	1.0000000	
		BAIK	36	17	19	0.9977725	
							0.0004724
	Q1	PRIA	46	27	19	0.9780710	
		WANITA	72	31	41	0.9860400	
							0.0168593
	Q2	TINGGI	2	2	0	#NUM!	
		SEDANG	105	48	57	0.9946938	
		RENDAH	11	7	4	0.9456603	
							0.0265291
	Q3	IPA	32	19	13	0.9744894	
		IPS	43	23	20	0.9964860	
		SENI	43	16	27	0.9522656	
							0.0253862
	Q4	OLAHRAGA	40	22	18	0.9927745	

Journal of Information System Research (JOSH)
Volume 5, No. 4, Juli 2024, pp 1042−1052
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v5i4.5643

NODE		Keterangan	Jumlah Kasus	Main SAINS	Main SOSIAL	Entropy	Gain
		SOSIAL	41	21	20	0.9995708	
		SENI	37	16	21	0.9867867	
							0.0065344
	Q5	BISNIS	88	40	48	0.9940302	
		TEKNOLOGI IT	30	18	12	0.9709506	
							0.0116303
	Q6	YA	22	14	8	0.9456603	
		TIDAK	96	44	52	0.9949848	
							0.0140040
	Q7	YA	16	12	4	0.8112781	
		TIDAK	102	46	56	0.9930555	
							0.0313851

Dari hasil perhitungan Tabel 7 hasil Perhitungan dari node 1, fitur Q7 memiliki nilai perhitungan yang paling tinggi berdasarkan kriteria nilai Gain. Oleh karena itu, fitur Q7 dipilih sebagai node awal dalam pohon keputusan. Q7 memiliki pengaruh terbesar dalam memisahkan data secara optimal pada tahap pertama pembentukan pohon keputusan pada gambar 3 node awal.



Gambar 3. Node awal

Tabel 8. Hasil Perhitungan Dari Node 2

NODE		Keterangan	Jumlah Kasus	Main SAINS	Main SOSIAL	Entropy	Gain
2	Q7 TIDAK	TOTAL	102	46	56	0.9930555	
	PAI	SANGAT BAIK	89	42	47	0.9977221	
		BAIK	13	4	9	0.8904916	
							0.0090000
	PPKN	SANGAT BAIK	59	26	33	0.9898221	
		BAIK	42	20	22	0.9983637	
		CUKUP	1	0	1	#NUM!	
							0.0094204
	B.INDO	SANGAT BAIK	74	35	39	0.9978913	
		BAIK	27	11	16	0.9751191	
		CUKUP	1	0	1	#NUM!	
							0.0109754
	MM	SANGAT BAIK	65	28	37	0.9861261	
		BAIK	36	18	18	1.0000000	
		CUKUP	1	0	1	#NUM!	
							0.0117006
	SEJARAH	SANGAT BAIK	85	37	48	0.9878854	
		BAIK	16	9	7	0.9886994	
		CUKUP	1	0	1	#NUM!	
							0.0147276
	B.INGG	SANGAT BAIK	69	30	39	0.9876925	
		BAIK	32	16	16	1.0000000	
		CUKUP	1	0	1	#NUM!	
							0.0111851
	SENI	SANGAT BAIK	74	32	42	0.9867867	
		BAIK	28	14	14	1.0000000	
							0.0026416

Journal of Information System Research (JOSH)
Volume 5, No. 4, Juli 2024, pp 1042−1052
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v5i4.5643

NODE		Keterangan	Jumlah Kasus	Main SAINS	Main SOSIAL	Entropy	Gain
	PENJAS	SANGAT BAIK	69	31	38	0.9925631	
		BAIK	33	15	18	0.9940302	
							0.0000177
	Q1	PRIA	37	18	19	0.9994730	
		WANITA	65	28	37	0.9861261	
							0.0020879
	Q2	TINGGI	1	1	0	#NUM!	
		SEDANG	96	42	54	0.9886994	
		RENDAH	5	3	2	0.9709506	
							0.0149192
	Q3	IPA	25	13	12	0.9988455	
		IPS	38	19	19	1.0000000	
		SENI	39	14	25	0.9418285	
							0.0155805
	Q4	OLAHRAGA	33	15	18	0.9940302	
		SOSIAL	36	17	19	0.9977725	
		SENI	33	14	19	0.9833762	
							0.0011514
	Q5	BISNIS	78	34	44	0.9881108	
		TEKNOLOGI IT	24	12	12	1.0000000	
							0.0021472
	Q6	YA	17	11	6	0.9366674	
		TIDAK	85	35	50	0.9774178	
							0.0224294

Tabel 9. Hasil Perhitungan Dari Node 3

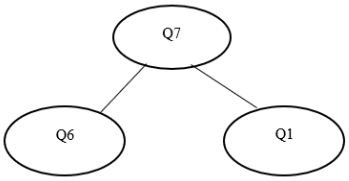
NODE		Keterangan	Jumlah Kasus	Main SAINS	Main SOSIAL	Entropy	Gain
3	Q7 YA	TOTAL	16	12	4	0.8112781	
	PAI	SANGAT BAIK	16	12	4	0.8112781	
		BAIK	0	0	0	#DIV/0!	
							0.0000000
	PPKN	SANGAT BAIK	8	6	2	0.8112781	
		BAIK	8	6	2	0.8112781	
		CUKUP	1	0	1	#NUM!	
							0.0000000
	B.INDO	SANGAT BAIK	7	6	1	0.5916728	
		BAIK	9	6	3	0.9182958	
		CUKUP	0	0	0	#DIV/0!	
							0.0358799
	MM	SANGAT BAIK	12	9	3	0.8112781	
		BAIK	4	3	1	0.8112781	
		CUKUP	0	0	0	#DIV/0!	
							0.0000000
	SEJARAH	SANGAT BAIK	14	10	4	0.8631206	
		BAIK	2	2	0	#NUM!	
		CUKUP	1	0	1	#NUM!	
							0.0560476
	B.INGG	SANGAT BAIK	12	9	3	0.8112781	
		BAIK	4	3	1	0.8112781	
		CUKUP	0	0	0	#DIV/0!	
							0.0000000
	SENI	SANGAT BAIK	14	12	2	0.5916728	
		BAIK	2	0	2	#NUM!	
							0.2935644
	PENJAS	SANGAT BAIK	13	10	3	0.7793498	
		BAIK	3	2	1	0.9182958	
							0.0058759
	Q1	PRIA	9	9	0	#NUM!	

Journal of Information System Research (JOSH)
Volume 5, No. 4, Juli 2024, pp 1042−1052
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v5i4.5643

NODE		Keterangan	Jumlah Kasus	Main SAINS	Main SOSIAL	Entropy	Gain
		WANITA	7	3	4	0.9852281	
							0.3802408
	Q2	TINGGI	1	1	0	#NUM!	
		SEDANG	9	6	3	0.9182958	
		RENDAH	6	5	1	0.6500224	
							0.0509783
	Q3	IPA	7	6	1	0.5916728	
		IPS	5	4	1	0.7219281	
		SENI	4	2	2	1.0000000	
							0.0768188
	Q4	OLAHRAGA	7	7	0	#NUM!	
		SOSIAL	5	3	2	0.9709506	
		SENI	4	2	2	1.0000000	
							0.2578561
	Q5	BISNIS	10	6	4	0.9709506	
		TEKNOLOGI IT	6	6	0	#NUM!	
							0.2044340
	Q6	YA	5	3	2	0.9709506	
		TIDAK	11	9	2	0.6840384	
							0.0375796

Untuk node kedua, hasil perhitungan pada Tabel 8 hasil Perhitungan dari node 2 menunjukkan bahwa fitur 'Q6' adalah yang terbaik dalam memisahkan data pada tahap ini. Oleh karena itu, 'Q6' dipilih sebagai node kedua dalam pohon keputusan. Ini mengindikasikan bahwa setelah pemisahan berdasarkan 'Q7', fitur 'Q6' memberikan pemisahan data yang paling informatif berikutnya.

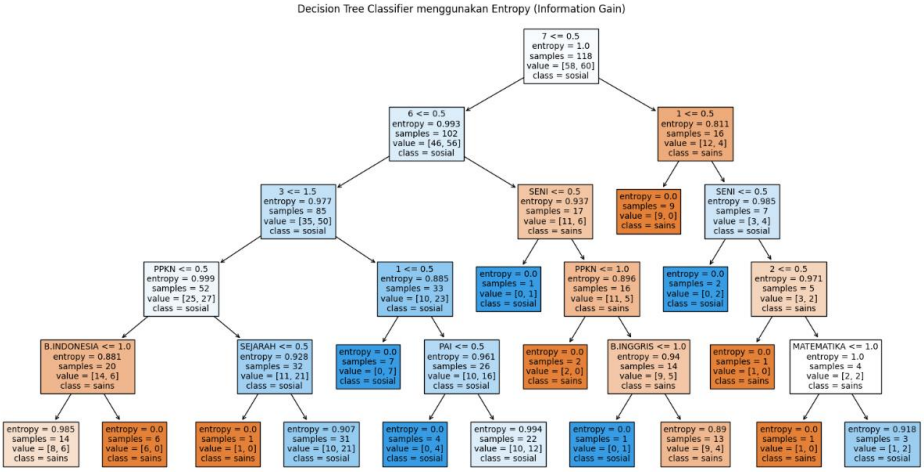
Selanjutnya, pada node ketiga, perhitungan pada Tabel 9 hasil Perhitungan dari node 3 mengindikasikan bahwa fitur 'Q1' adalah yang terbaik dalam memisahkan data lebih lanjut. Dengan demikian, 'Q1' dipilih sebagai node ketiga dalam pohon keputusan. Ini menunjukkan bahwa setelah mempertimbangkan pemisahan berdasarkan 'Q7' dan 'Q6', fitur 'Q1' memberikan informasi tambahan yang signifikan untuk pemisahan data pada tahap ini.



Gambar 4. ke 2 dan ke 3

3.3 Hasil decision tree

Berikut adalah hasil decision tree dengan menggunakan Gain Information sebagai proses pembelajaran model menggunakan data training yang terdiri dari 118 data. Hasil pembelajaran padamodel ini memberikan hasil berupa pohon keputusan dengan konfigurasi yang sesuai dengan perintah yang dimasukkan pada gambar 5.



Gambar 5. Hasil decision tree

Journal of Information System Research (JOSH)
Volume 5, No. 4, Juli 2024, pp 1042−1052
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v5i4.5643

3.4 Pengujian

Setelah pembelajaran model selesai, langkah berikutnya adalah melakukan pengujian model. Proses pengujian ini dilakukan dengan menggunakan 22 data uji, dan selanjutnya akan dilakukan evaluasi performa model. Evaluasi performa model dilakukan melalui perhitungan confusion matrix. Berikut Pada Tabel 10 adalah hasil evaluasi akurasi model menggunakan data uji.

Tabel 10. Convision Matrix

Kelas Prediksi	Aktual Negative	Aktual Positif
Prediksi Negatif	9	2
Prediksi Positif	1	10

```
Confusion Matrix:
[[ 9  2]
 [ 1 10]]
Accuracy: 0.86
```

Gambar 6. Hasil pengujian python

Berdasarkan tabel 10, nilai akurasi pada data uji dapat dihitung dengan menggunakan rumus berikut:

$$\text{Akurasi} = \frac{\text{TP+TN}}{\text{TP+FN+FP+TN}}$$

$$\text{Akurasi} = \frac{10+9}{10+1+2+9} = 0.86$$

Nilai akurasi model yang diperoleh dari penelitian ini adalah sebesar 86%. Hal ini menunjukkan bahwa performa model yang dibangun cukup baik dan mampu memberikan hasil prediksi yang cukup akurat pada data yang belum diketahui labelnya. Selain itu, akurasi yang tinggi ini menunjukkan bahwa model tersebut memiliki kemampuan yang andal dalam mengklasifikasikan data dengan benar, yang merupakan indikator bahwa pendekatan yang digunakan dalam pelatihan model ini berhasil. Dengan demikian, model ini dapat digunakan sebagai alat prediksi yang efektif untuk aplikasi lebih lanjut.

4 KESIMPULAN

Berdasarkan hasil penelitian yang dilakukan dengan menggunakan pendekatan Supervised Learning Decision Tree C4.5 dan menerapkan teknik Information Gain untuk klasifikasi pemilihan program studi tingkat lanjut, diperoleh akurasi sebesar 86%. Tingkat keberhasilan ini menunjukkan bahwa metode tersebut efektif dalam mengidentifikasi dan mengklasifikasikan program studi tingkat lanjut. Hal ini mengindikasikan bahwa penggunaan Decision Tree C4.5 yang memanfaatkan teknik Information Gain memiliki potensi besar sebagai model yang dapat membantu siswa dalam memilih program studi tingkat lanjut mereka dengan tingkat keakuratan yang memuaskan. Dengan hasil akurasi yang tinggi, metode ini dapat diandalkan untuk memberikan prediksi yang akurat dalam konteks pemilihan program studi. Selain itu, penelitian ini menyoroti keefektifan teknik Information Gain dalam membangun model klasifikasi yang kuat. Potensi aplikasi dari model ini mencakup berbagai aspek dalam proses pemilihan program studi, mulai dari membantu siswa dalam pengambilan keputusan hingga memberikan rekomendasi yang didasarkan pada data historis yang akurat. Penelitian ini masih banyak kekurangan Untuk pengembangan model ini, disarankan agar pengembang menambahkan beberapa variabel jurusan memperbanyak jurusan yang dapat dipilih. Selain itu, untuk mencapai akurasi yang lebih tinggi, sangat diharapkan agar para pengembang menggunakan jumlah data yang lebih banyak lagi. Semakin banyak data sampel yang digunakan, semakin besar pula kemungkinan model menghasilkan prediksi yang lebih akurat.

REFERENCES

[1] Reynaldo, B. Mulyawan, and T. Sutrisno, “Rekomendasi Pemilihan Program Studi Tarumangara Menggunakan Metode,” J. Komput. dan Inform., vol. 15, no. 1, pp. 326–333, 2020.
[2] P. D. Kusuma, Machine Learning Teori,Program,Dan Studi Kasus. Jl.Rajawali, G.Elang 6, No 3, Drono, Sardonoharjo,Ngaglik, Sleman Jl.Kaliurang Km.9,3-Yogyakarta 55581: Grup Penerbitan CV BUDI UTAMA, 2020.
[3] A. Yuliana and D. B. Pratomo, “Memprediksi Kepuasan Mahasiswa Terhadap Kinerja Dosen Politeknik TEDC Bandung,” Semnasinotek 2017, pp. 377–384, 2017.
[4] A. Roihan, P. A. Sunarya, and A. S. Rafika, “Pemanfaatan Machine Learning dalam Berbagai Bidang: Review paper,” IJCIT (Indonesian J. Comput. Inf. Technol., vol. 5, no. 1, pp. 75–82, 2020, doi: 10.31294/ijcit.v5i1.7951.
[5] Dhea Halimah, Muhammad Ridwan Lubis, and Widodo Saputra, “Algoritma C4.5 Untuk Menentukan Klasifikasi Tingkat Pemahaman Mahasiswa Pada Matakuliah Bahasa Pemrograman,” J. Tek. Mesin, Ind. Elektro Dan Inform., vol. 1, no. 3, pp. 24–38, 2022, doi: 10.55606/jtmei.v1i3.534.
[6] D. Hartama and K. D. R. Sianipar, “Penerapan Algoritma C4.5 Untuk Analisa Tingkat Keberhasilan Mahasiswa Dalam Pembelajaran Praktikum di Masa Pandemi,” J. Comput. Syst. Informatics, vol. 4, no. 1, pp. 128–134, 2022, doi: 10.47065/josyc.v4i1.2584.
[7] E. E. Barito, J. T. Beng, and D. Arisandi, “Penerapan Algoritma C4.5 Untuk Klasifikasi Mahasiswa Penerima Bantuan Sosial Covid-19,” J. Ilmu Komput. dan Sist. Inf., vol. 10, no. 1, 2022, doi: 10.24912/jiksi.v10i1.17819.

Journal of Information System Research (JOSH)
Volume 5, No. 4, Juli 2024, pp 1042−1052
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v5i4.5643

[8] F. F. Kusuma, "Penerapan Data Mining Untuk Akurasi Analisis Cuaca di Australia Menggunakan Algoritma J48 Decision Tree," J. Comput. Sci. Inf. Syst. J-Cosys, vol. 3, no. 2, pp. 65–68, 2023, doi: 10.53514/jco.v3i2.396.
[9] A. Baktiar, "Decission Tree Sebagai Metode Penentuan Penjurusan Perguruan Tinggi Berdasarkan Minat Dan Bakat Melalui Data Raport Dengan Uji Algoritma C4.5 (Studi Kasus di SMKN 1 Donorojo Pacitan)," J. PILAR Teknol. J. Ilm. Ilmu Ilmu Tek., vol. 7, no. 1, pp. 40–45, 2022, doi: 10.33319/piltek.v7i1.110.
[10] M. Yunus, H. Ramadhan, D. R. Aji, and A. Yulianto, "Penerapan Metode Data Mining C4.5 Untuk Pemilihan Penerima Kartu Indon[1] M. Yunus, H. Ramadhan, D. R. Aji, and A. Yulianto, 'Penerapan Metode Data Mining C4.5 Untuk Pemilihan Penerima Kartu Indonesia Pintar (KIP),' Paradig. - J. Komput. dan Inform., vol.," Paradig. - J. Komput. dan Inform., vol. 23, no. 2, 2021.
[11] R. Haqmanullah Pambudi and B. Darma Setiawan, "Penerapan Algoritma C4.5 Untuk Memprediksi Nilai Kelulusan Siswa Sekolah Menengah Berdasarkan Faktor Eksternal," J. Pengemb. Teknol. Inf. dan Ilmu Komput., vol. 2, no. 7, pp. 2637–2643, 2018.
[12] E. D. Sri mulyani, I. Agustina, R. Ismanto, S. J. Susanto, S. S. Adhi, and T. S. Rismayanti, "Klasifikasi Pola Penjurusan Bidang It Menggunakan Algoritma C4.5 Studi Kasus Smkn Manonjaya," TEKTRIKA - J. Penelit. dan Pengemb. Telekomun. Kendali, Komputer, Elektr. dan Elektron., vol. 6, no. 1, p. 1, 2022, doi: 10.25124/tektrika.v6i1.2254.
[13] R. A. Saputra, S. Wasiyanti, and D. Pribadi, "Information Gain Pada Algoritma C4.5 Untuk Klasifikasi Penerimaan Bantuan Pangan Non Tunai (Bpnt)," Indones. J. Bus. Intell., vol. 4, no. 1, p. 25, 2021, doi: 10.21927/ijubi.v4i1.1757.
[14] R. Triyandika, "Penerapan Metode Decision Tree Dengan Algoritma C4.5 Untuk Klasifikasi Penyakit Jantung," p. 1, 2022.
[15] A. F. A. Rahman, Sorikhi, and S. Wartulas, "Prediksi Kelulusan Mahasiswa Menggunakan Algoritma C4.5 (Studi Kasus Di Universitas Peradaban)," Ijir, vol. 1, no. 2, pp. 70–77, 2020.
[16] M. A. Abdillah, A. Setyanto, and S. Sudarmawan, "Implementasi Decision Tree Algoritma C4.5 Untuk Memprediksi Kesuksesan Pendidikan Karakter," Respati, vol. 15, no. 2, p. 59, 2020, doi: 10.35842/jtir.v15i2.349.
[17] M. Solehuddin, W. A. Syafei, and R. Gernowo, "Metode Decision Tree untuk Meningkatkan Kualitas Rencana Pelaksanaan Pembelajaran dengan Algoritma C4.5," J. Penelit. dan Pengemb. Pendidik., vol. 6, no. 3, pp. 510–519, 2022, doi: 10.23887/jppp.v6i3.52840.
[18] U. Enri, "Penerapan Algoritma C4.5 Dalam Pemilihan Program Studi Fakultas Ilmu Komputer (Studi Kasus Sekolah Menengah Atas Negeri 1 Tambun Utara)," J. Rekayasa Inf., vol. 7, no. 1, pp. 1–7, 2018.
[19] P. P. Haryoto, H. Okprana, and I. S. Saragih, "Algoritma C4.5 Dalam Data Mining Untuk Menentukan Klasifikasi Penerimaan Calon Mahasiswa Baru," TIN Terap. Inform. Nusant., vol. 2, no. 5, pp. 358–364, 2021.
[20] Suherman, M. Purnamasari, and F. D. Hastuti, "Klasifikasi Siswa Berdasarkan Mata Pelajaran Lintas Minat Menggunakan Metode Decision Tree C4.5," J. Sist. Inf., vol. 8, no. 08, pp. 141–149, 2021.