

Penerapan Algoritma *Decision Tree* (C4.5) untuk Klasifikasi Minat Siswa Sekolah Dasar

Application of the Decision Tree (C4.5) Algorithm for Classifying Elementary School Students' Interests

Khairul Ummi

Universitas Potensi Utama, Jl. K.L Yos Sudarso Km.6,5 No.3A, Medan, Indonesia
Ummi12gibmie@gmail.com

ABSTRAK

Identifikasi minat siswa sekolah dasar penting dilakukan agar sekolah dan guru dapat memberikan kegiatan pembelajaran serta ekstrakurikuler yang lebih sesuai dengan potensi peserta didik. Penelitian ini menerapkan algoritma *Decision Tree* C4.5 untuk mengklasifikasikan minat siswa ke dalam tujuh kelas: Minat Akademik, Seni, Olahraga, Teknologi, Bahasa, Keagamaan, dan Kepemimpinan. Penelitian menggunakan dataset simulasi 100 siswa yang dibentuk berbasis profil kelas, dengan atribut nilai akademik, keaktifan, ekstrakurikuler, dan hobi. C4.5 dipilih karena fokus penelitian mencakup interpretabilitas aturan keputusan, bukan klaim superioritas akurasi terhadap metode ensemble. Dataset dibagi menjadi 80 data latih dan 20 data uji secara stratified. Hasil eksperimen menghasilkan 19 prediksi benar atau accuracy 95%, dengan macro precision 95,24%, macro recall 95,24%, dan macro F1-score 94,29%. Namun, karena ukuran data uji kecil dan data sintesis memiliki pola input-target yang sengaja dirancang, angka tersebut tidak dapat digeneralisasikan sebagai performa C4.5 pada sekolah nyata. Kontribusi penelitian adalah menyediakan baseline metodologis yang transparan dan aturan keputusan yang dapat ditelusuri guru. Validasi dengan data siswa nyata yang dianonimkan, cross-validation, dan perbandingan empiris dengan *Random Forest* serta *XGBoost* diperlukan pada penelitian lanjutan.

Kata Kunci : *Decision Tree*, C4.5, Klasifikasi, Minat Siswa, Sekolah Dasar, Data Mining

ABSTRACT

Identifying elementary school students' interests is important so that teachers and schools can provide learning and extracurricular activities aligned with students' potential. This study applies the *Decision Tree* C4.5 algorithm to classify student interests into seven classes: Academic, Art, Sports, Technology, Language, Religious, and Leadership Interest. The study uses a 100-student synthetic dataset constructed from class-based profiles, with academic scores, classroom activity, extracurricular activities, and hobbies as input attributes. C4.5 was selected because the study emphasizes interpretable decision rules rather than claiming predictive superiority over ensemble methods. The dataset was divided into 80 training and 20 testing records using stratified sampling. The experiment produced 19 correct predictions, yielding 95% accuracy, with macro precision of 95.24%, macro recall of 95.24%, and macro F1-score of 94.29%. However, because the testing sample is small and the synthetic data contain deliberately structured relationships between inputs and target classes, the result cannot be generalized as C4.5 performance in real schools. The contribution of this study is a transparent methodological baseline and decision rules that can be traced by teachers. Validation using anonymized real student data, cross-validation, and empirical comparison with *Random Forest* and *XGBoost* are required in future research.

Keywords : *Decision Tree*, C4.5, Classification, Student Interests, Elementary School, Data Mining

Disubmit: Info Artikel : Diterima :

Copyright © 2026 - Journal UPU. All rights reserved.

1. PENDAHULUAN

Pendidikan dasar tidak hanya berorientasi pada pencapaian nilai akademik, tetapi juga pada pengembangan potensi, bakat, karakter, dan pengalaman belajar siswa. Setiap siswa memiliki kombinasi kemampuan dan kecenderungan yang berbeda. Perbedaan tersebut dapat terlihat dari capaian mata pelajaran, aktivitas di kelas, pilihan kegiatan ekstrakurikuler, serta hobi. Pemetaan minat secara

sistematis dapat membantu guru menyusun pendekatan pembelajaran dan kegiatan pengembangan diri yang lebih tepat.

Dalam praktiknya, penentuan minat siswa masih sering dilakukan melalui pengamatan guru, wawancara, atau hasil angket sederhana. Cara tersebut tetap penting, tetapi hasilnya dapat bersifat subjektif apabila tidak didukung oleh pengolahan data yang terstruktur. Perkembangan data mining memberikan peluang untuk mengolah data akademik dan nonakademik menjadi pola yang dapat digunakan sebagai dasar pengambilan keputusan. Penelitian terkait pendidikan menunjukkan bahwa algoritma pohon keputusan dapat digunakan untuk mengidentifikasi pola prestasi, kesiapan siswa, pemilihan program studi, dan berbagai bentuk klasifikasi peserta didik ^{[1]–[6]}.

Algoritma Decision Tree C4.5 merupakan salah satu metode klasifikasi yang membentuk struktur pohon berdasarkan atribut yang paling informatif. C4.5 menggunakan entropy untuk mengukur ketidakpastian, information gain untuk mengukur pengurangan ketidakpastian, dan gain ratio untuk memilih atribut yang relatif paling baik dalam membentuk percabangan. Kelebihan utamanya adalah hasil klasifikasi dapat diterjemahkan menjadi aturan if–then yang mudah dipahami. Hal ini relevan dalam lingkungan sekolah karena guru tidak hanya membutuhkan nilai akurasi, tetapi juga membutuhkan penjelasan mengapa seorang siswa dikelompokkan ke dalam kelas minat tertentu.

Beberapa penelitian terbaru memperlihatkan penggunaan Decision Tree dan C4.5 dalam konteks pendidikan. Wicaksono dan Setiadi menggunakan C4.5 untuk memprediksi kelulusan siswa sekolah dasar berdasarkan nilai akademik ^[1]. Rizki dan Armansyah menerapkan C4.5 untuk klasifikasi calon mahasiswa dan melaporkan akurasi 85,7% ^[2]. Pratama dan Armansyah menggabungkan informasi akademik, minat, dan bakat untuk klasifikasi pilihan studi lanjut ^[3]. Penelitian lain menggunakan C4.5 untuk mengklasifikasikan hasil belajar siswa berdasarkan riwayat akademik dan minat ^[4], sedangkan Fahriah dkk. menunjukkan pengembangan C4.5 dengan seleksi fitur dan AdaBoost untuk klasifikasi minat studi ^[5].

Penelitian yang lebih baru juga memperlihatkan bahwa Decision Tree dapat digunakan untuk klasifikasi prestasi dan karakteristik peserta didik. Astuti dkk. menggunakan C4.5 untuk memprediksi keberhasilan studi mahasiswa ^[6]. Sukarto dkk. menerapkan Decision Tree untuk klasifikasi minat siswa SMA dalam melanjutkan pendidikan ^[7]. Halimatussa'diyah dan Rohman menggunakan C4.5 untuk prediksi prestasi siswa dan memperoleh akurasi 89,7% ^[8]. Moli dkk. menggunakan C4.5 untuk memprediksi minat jurusan siswa berdasarkan nilai akademik dan kecenderungan minat ^[9].

Masih terdapat ruang pengembangan pada konteks sekolah dasar, khususnya klasifikasi minat yang tidak hanya diarahkan pada pilihan jurusan, tetapi pada tujuh kategori minat yang lebih luas: akademik, seni, olahraga, teknologi, bahasa, keagamaan, dan kepemimpinan. Penelitian ini memanfaatkan kombinasi atribut nilai akademik, keaktifan, ekstrakurikuler, dan hobi untuk membangun model C4.5. Dengan demikian, penelitian diarahkan untuk menghasilkan model yang memiliki kinerja klasifikasi baik dan aturan yang dapat dipahami guru.

Rumusan masalah penelitian adalah: (1) bagaimana menerapkan algoritma Decision Tree C4.5 untuk mengklasifikasikan minat siswa sekolah dasar; (2) atribut apa yang memiliki pengaruh besar berdasarkan entropy, information gain, dan gain ratio; dan (3) seberapa baik kinerja model dalam mengklasifikasikan tujuh kelas minat berdasarkan accuracy, precision, recall, dan F1-score.

Tujuan penelitian adalah membangun model klasifikasi minat siswa menggunakan C4.5, menghitung entropy, information gain, split information, dan gain ratio, menghasilkan aturan keputusan, serta mengevaluasi kinerja model menggunakan data uji. Manfaat penelitian adalah menyediakan contoh penerapan data mining dalam pendidikan dasar serta kerangka awal yang dapat dikembangkan menjadi sistem pendukung keputusan bagi guru.

2. METODE

A. Desain Penelitian

Penelitian menggunakan pendekatan kuantitatif eksperimental berbasis data mining. Pada tahap penelitian ini, data siswa lapangan yang telah dianonimkan belum tersedia, sehingga digunakan dataset sintesis/simulasi sebanyak 100 siswa. Dataset tidak dimaksudkan sebagai representasi populasi sekolah tertentu, melainkan sebagai baseline eksperimental untuk menguji alur perhitungan C4.5, pembentukan aturan, dan evaluasi klasifikasi. Karena dataset dibentuk dari profil minat yang telah ditentukan sebelumnya, hubungan antara atribut dan label target berpotensi lebih terstruktur dibandingkan data lapangan. Oleh karena itu, hasil penelitian diposisikan sebagai bukti kelayakan metodologis, bukan bukti keunggulan C4.5 secara umum.

B. Variabel Penelitian

Atribut input terdiri atas lima nilai akademik dan tiga atribut nonakademik. Nilai akademik berada pada skala 60–100. Keaktifan terdiri atas Aktif, Cukup, dan Kurang. Ekstrakurikuler terdiri atas Pramuka, Tari, Futsal, Musik, dan Tahfiz. Hobi terdiri atas Menggambar, Membaca, Bermain Bola, dan Bernyanyi. Target terdiri atas tujuh kelas minat.

Tabel 1. Variabel penelitian dan kategori data

Variabel	Tipe	Rentang/Kategori	Peran
Matematika	Numerik	60–100	Atribut input
Bahasa Indonesia	Numerik	60–100	Atribut input
IPA	Numerik	60–100	Atribut input
Seni Budaya	Numerik	60–100	Atribut input
PJOK	Numerik	60–100	Atribut input
Keaktifan	Kategorikal	Aktif/Cukup/Kurang	Atribut input
Ekstrakurikuler	Kategorikal	Pramuka/Tari/Futsal/Musik/Tahfiz	Atribut input
Hobi	Kategorikal	Menggambar/Membaca/Bermain Bola/Bernyanyi	Atribut input
Minat	Kategorikal	7 kelas	Target

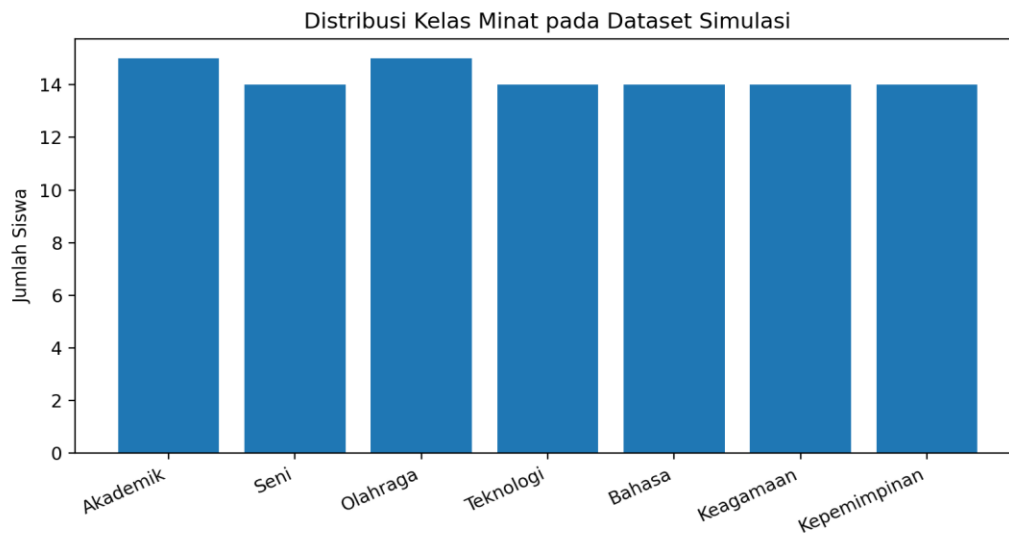
C. Pembentukan Dataset

Dataset simulasi disusun menggunakan pendekatan profile-based synthetic data. Pertama, ditetapkan tujuh kelas target, yaitu Minat Akademik, Seni, Olahraga, Teknologi, Bahasa, Keagamaan, dan Kepemimpinan, dengan distribusi 14–15 siswa per kelas. Kedua, setiap kelas diberi profil atribut yang konsisten dengan konstruk minatnya. Sebagai contoh, kelas Akademik diberi kecenderungan nilai Matematika dan IPA lebih tinggi serta pola ekstrakurikuler/hobi yang mendukung aktivitas akademik; kelas Seni diberi kecenderungan nilai Seni Budaya tinggi dan hobi Menggambar; kelas Olahraga diberi kecenderungan nilai PJOK tinggi serta ekstrakurikuler Futsal dan hobi Bermain Bola; kelas Teknologi diberi kecenderungan nilai Matematika dan IPA tinggi; kelas Bahasa diberi kecenderungan Bahasa Indonesia tinggi dan hobi Bernyanyi; kelas Keagamaan diberi ekstrakurikuler Tahfiz; sedangkan kelas Kepemimpinan diberi kombinasi nilai akademik relatif tinggi, keaktifan, dan aktivitas Pramuka. Ketiga, nilai numerik diberi variasi antar-siswa sehingga tidak seluruh anggota kelas memiliki nilai identik, sedangkan atribut kategorikal diberi variasi terbatas sebagai noise. Keempat, setiap baris diperiksa agar rentang nilai berada pada 60–100 dan kategori sesuai definisi variabel. Kelima, label Minat ditetapkan berdasarkan profil kelas yang telah dirancang, kemudian dataset dibagi secara stratified menjadi 80 data latih dan 20 data uji. Dengan prosedur tersebut, data simulasi menguji kemampuan C4.5 menemukan pola yang sengaja dibuat memiliki hubungan logis dengan target, tetapi belum merepresentasikan kompleksitas data siswa nyata. Perlu dicatat bahwa naskah awal tidak menyimpan skrip generator,

distribusi probabilitas, dan random seed; karena itu, revisi ini tidak mengklaim reproduksi numerik identik melalui proses random tertentu.

Tabel 2. Distribusi kelas pada dataset simulasi

Kelas Minat	Jumlah
Minat Akademik	15
Minat Seni	14
Minat Olahraga	15
Minat Teknologi	14
Minat Bahasa	14
Minat Keagamaan	14
Minat Kepemimpinan	14



Gambar 1. Distribusi tujuh kelas minat pada dataset simulasi

D. Tahapan Pengolahan Data

Tahapan penelitian terdiri dari: (1) identifikasi masalah dan penentuan tujuh kelas minat; (2) pembentukan dataset sintesis berbasis profil minat; (3) pemeriksaan rentang, konsistensi kategori, dan distribusi kelas; (4) pembagian data latih dan uji 80:20 secara stratified; (5) perhitungan entropy, information gain, split information, dan gain ratio; (6) pemilihan atribut/threshold terbaik secara rekursif dan pembentukan pohon keputusan; (7) penerjemahan jalur pohon menjadi aturan if-then; dan (8) evaluasi menggunakan confusion matrix, accuracy, precision, recall, dan F1-score. Untuk pengembangan berikutnya, dataset simulasi ini perlu digantikan atau diuji silang dengan data sekolah nyata yang telah dianonimkan serta evaluasi cross-validation.

E. Perhitungan C4.5

$$\text{Entropy}(S) = - \sum_i p_i \log_2(p_i) \quad (1)$$

$$\text{Gain}(S,A) = \text{Entropy}(S) - \sum_v (|S_v|/|S|) \text{Entropy}(S_v) \quad (2)$$

$$\text{SplitInfo}(S,A) = - \sum_v (|S_v|/|S|) \log_2(|S_v|/|S|) \quad (3)$$

$$\text{GainRatio}(S,A) = \text{Gain}(S,A) / \text{SplitInfo}(S,A) \quad (4)$$

Entropy mengukur ketidakpastian, information gain mengukur pengurangan ketidakpastian, split information mengukur informasi intrinsik suatu pembagian, dan gain ratio merupakan rasio antara gain

dan split information. Untuk atribut numerik, kandidat threshold dibandingkan dan threshold dengan gain ratio terbaik dipilih.

F. Evaluasi Model

Kinerja model dievaluasi menggunakan confusion matrix 7×7. Accuracy menunjukkan proporsi prediksi benar terhadap seluruh data uji, sedangkan precision, recall, dan F1-score dihitung per kelas dan diringkas menggunakan macro average. Karena data uji hanya berisi 20 siswa, metrik diperlakukan sebagai hasil deskriptif pada eksperimen ini. Untuk menunjukkan ketidakpastian akibat ukuran sampel yang kecil, proporsi 19/20 juga dilaporkan dengan interval kepercayaan binomial 95%.

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (5)$$

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \quad (6)$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \quad (7)$$

$$\text{F1} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (8)$$

G. Alasan Pemilihan C4.5 dan Posisi terhadap Random Forest serta XGBoost

Pemilihan C4.5 dalam penelitian ini tidak didasarkan pada klaim bahwa C4.5 selalu lebih akurat daripada Random Forest atau XGBoost. Ketiga metode sama-sama dapat digunakan untuk klasifikasi, tetapi tujuan penelitian ini tidak hanya menghasilkan prediksi, melainkan juga memperoleh aturan keputusan yang dapat dibaca dan ditelusuri oleh guru. C4.5 membentuk satu pohon keputusan dan menghasilkan percabangan berbasis atribut serta threshold yang relatif langsung diterjemahkan menjadi aturan if–then. Karakter tersebut sesuai dengan kebutuhan pemetaan awal minat siswa, ketika guru perlu mengetahui alasan sebuah prediksi dibuat. Random Forest membangun banyak pohon dan menggabungkan hasilnya sehingga umumnya lebih robust terhadap variasi data, tetapi struktur ensemble membuat penjelasan keputusan individual tidak sesederhana satu pohon C4.5. XGBoost menggunakan boosting bertahap dan optimasi ensemble yang kuat untuk performa prediktif, tetapi interpretasi langsung berupa aturan tunggal juga lebih kompleks. Oleh karena itu, C4.5 diposisikan sebagai metode yang lebih sesuai dengan fokus interpretabilitas dan demonstrasi proses pengambilan keputusan pada penelitian ini, bukan sebagai metode yang terbukti superior secara prediktif. Random Forest dan XGBoost justru relevan sebagai baseline pembandingan pada penelitian lanjutan, terutama setelah tersedia dataset nyata yang lebih besar dan evaluasi cross-validation

3. HASIL DAN PEMBAHASAN

A. DATASET SIMULASI 100 SISWA

Dataset berikut merupakan data simulasi yang digunakan dalam eksperimen. Data tidak merepresentasikan identitas atau kondisi siswa nyata.

4. Tabel A1. Dataset simulasi 100 siswa

No	Mat	B.Ind	IPA	Seni	PJOK	Aktif	Ekskul	Hobi	Minat
1	95	76	90	81	80	Aktif	Pramuka	Membaca	Akademik
2	93	84	88	76	76	Cukup	Pramuka	Membaca	Akademik
3	89	80	93	84	77	Aktif	Pramuka	Membaca	Akademik
4	92	77	92	75	79	Aktif	Pramuka	Membaca	Akademik
5	95	79	88	82	81	Aktif	Pramuka	Membaca	Akademik
6	89	84	94	82	78	Aktif	Pramuka	Membaca	Akademik
7	95	75	85	80	76	Aktif	Pramuka	Membaca	Akademik
8	90	83	89	82	80	Aktif	Pramuka	Membaca	Akademik
9	97	80	85	83	80	Aktif	Pramuka	Membaca	Akademik
10	86	78	93	84	76	Aktif	Pramuka	Membaca	Akademik
11	87	75	86	83	82	Aktif	Pramuka	Membaca	Akademik
12	96	77	92	81	80	Aktif	Pramuka	Membaca	Akademik
13	93	80	88	82	78	Cukup	Pramuka	Membaca	Akademik

No	Mat	B.Ind	IPA	Seni	PJOK	Aktif	Ekskul	Hobi	Minat
14	97	78	88	78	77	Aktif	Pramuka	Membaca	Akademik
15	97	76	87	80	78	Kurang	Pramuka	Membaca	Akademik
16	83	76	80	92	78	Aktif	Tahfiz	Menggambar	Seni
17	75	81	76	86	80	Aktif	Tari	Menggambar	Seni
18	83	81	79	91	76	Aktif	Tari	Menggambar	Seni
19	81	81	82	90	80	Aktif	Tari	Menggambar	Seni
20	76	76	76	97	83	Aktif	Tari	Menggambar	Seni
21	81	81	79	86	83	Aktif	Tari	Menggambar	Seni
22	83	79	81	94	83	Aktif	Tari	Menggambar	Seni
23	77	79	75	94	81	Kurang	Tari	Menggambar	Seni
24	78	82	80	97	84	Aktif	Tari	Menggambar	Seni
25	76	81	82	97	76	Aktif	Tari	Menggambar	Seni
26	83	75	84	89	77	Kurang	Tari	Menggambar	Seni
27	77	83	83	86	75	Aktif	Pramuka	Menggambar	Seni
28	81	81	81	97	82	Aktif	Tari	Menggambar	Seni
29	81	83	79	90	77	Aktif	Tari	Menggambar	Seni
30	80	77	77	77	87	Aktif	Futsal	Bermain Bola	Olahraga
31	84	84	84	80	92	Aktif	Futsal	Bermain Bola	Olahraga
32	80	79	75	79	94	Aktif	Futsal	Bermain Bola	Olahraga
33	75	84	80	83	86	Aktif	Futsal	Bermain Bola	Olahraga
34	83	75	82	83	90	Aktif	Futsal	Bermain Bola	Olahraga
35	84	80	76	81	87	Aktif	Futsal	Bermain Bola	Olahraga
36	79	81	76	81	87	Aktif	Futsal	Bermain Bola	Olahraga
37	83	81	75	77	88	Aktif	Futsal	Bermain Bola	Olahraga
38	76	83	75	84	87	Aktif	Futsal	Bermain Bola	Olahraga
39	77	75	78	79	88	Aktif	Futsal	Bermain Bola	Olahraga
40	78	84	79	83	88	Aktif	Futsal	Bermain Bola	Olahraga
41	82	76	83	76	92	Aktif	Futsal	Bermain Bola	Olahraga
42	82	81	83	81	85	Aktif	Futsal	Bermain Bola	Olahraga
43	77	79	77	82	86	Aktif	Tari	Bermain Bola	Olahraga
44	84	76	78	75	92	Kurang	Musik	Bermain Bola	Olahraga
45	96	79	90	84	76	Aktif	Musik	Menggambar	Teknologi
46	95	80	90	76	76	Aktif	Musik	Membaca	Teknologi
47	90	77	96	77	76	Aktif	Musik	Membaca	Teknologi
48	85	79	92	75	78	Aktif	Musik	Membaca	Teknologi
49	86	80	91	77	82	Aktif	Musik	Membaca	Teknologi
50	86	84	90	76	79	Aktif	Musik	Membaca	Teknologi
51	92	80	95	75	81	Aktif	Musik	Membaca	Teknologi
52	94	82	88	80	84	Aktif	Musik	Membaca	Teknologi
53	97	84	91	79	75	Aktif	Musik	Membaca	Teknologi
54	97	82	96	78	76	Aktif	Musik	Membaca	Teknologi
55	95	76	89	77	77	Aktif	Musik	Membaca	Teknologi
56	88	82	94	82	77	Aktif	Musik	Bernyanyi	Teknologi
57	93	81	90	76	83	Aktif	Musik	Membaca	Teknologi
58	97	81	86	77	77	Aktif	Musik	Membaca	Teknologi
59	79	95	79	79	76	Aktif	Musik	Bernyanyi	Bahasa
60	84	94	75	84	78	Kurang	Musik	Bernyanyi	Bahasa
61	79	88	81	76	77	Aktif	Musik	Bernyanyi	Bahasa
62	80	91	80	76	82	Aktif	Musik	Bernyanyi	Bahasa
63	81	87	80	80	77	Aktif	Musik	Bernyanyi	Bahasa
64	75	92	78	75	82	Aktif	Musik	Bernyanyi	Bahasa
65	75	89	80	80	75	Aktif	Musik	Bernyanyi	Bahasa
66	83	89	84	76	81	Aktif	Musik	Bernyanyi	Bahasa
67	77	91	75	77	82	Aktif	Musik	Bernyanyi	Bahasa
68	84	94	79	83	82	Cukup	Musik	Bernyanyi	Bahasa
69	76	97	75	79	84	Aktif	Musik	Bernyanyi	Bahasa
70	84	87	78	76	82	Aktif	Musik	Bernyanyi	Bahasa

No	Mat	B.Ind	IPA	Seni	PJOK	Aktif	Ekskul	Hobi	Minat
71	84	96	83	80	83	Aktif	Musik	Bernyanyi	Bahasa
72	83	94	78	83	79	Aktif	Musik	Bernyanyi	Bahasa
73	82	79	84	81	80	Aktif	Tahfiz	Membaca	Keagamaan
74	82	82	83	79	84	Aktif	Tahfiz	Membaca	Keagamaan
75	78	81	82	77	77	Cukup	Tahfiz	Membaca	Keagamaan
76	79	76	78	75	77	Cukup	Tahfiz	Membaca	Keagamaan
77	79	79	76	82	83	Aktif	Tahfiz	Membaca	Keagamaan
78	80	82	77	80	84	Aktif	Tahfiz	Membaca	Keagamaan
79	81	76	76	82	75	Cukup	Tahfiz	Membaca	Keagamaan
80	84	75	80	75	82	Aktif	Tahfiz	Membaca	Keagamaan
81	79	79	78	82	84	Cukup	Tahfiz	Membaca	Keagamaan
82	78	84	83	80	76	Aktif	Tahfiz	Membaca	Keagamaan
83	82	78	82	82	82	Cukup	Tahfiz	Bernyanyi	Keagamaan
84	75	75	84	81	76	Cukup	Tahfiz	Bernyanyi	Keagamaan
85	77	81	82	76	83	Cukup	Tahfiz	Membaca	Keagamaan
86	78	75	83	81	82	Aktif	Tahfiz	Membaca	Keagamaan
87	97	94	75	76	75	Aktif	Pramuka	Bernyanyi	Kepemimpinan
88	87	88	84	84	80	Cukup	Pramuka	Bernyanyi	Kepemimpinan
89	89	94	77	81	75	Cukup	Pramuka	Bernyanyi	Kepemimpinan
90	91	92	78	77	75	Cukup	Pramuka	Bernyanyi	Kepemimpinan
91	88	92	80	78	81	Aktif	Pramuka	Bernyanyi	Kepemimpinan
92	88	90	75	76	82	Aktif	Pramuka	Bernyanyi	Kepemimpinan
93	92	89	80	82	82	Cukup	Pramuka	Bernyanyi	Kepemimpinan
94	96	89	84	81	78	Aktif	Pramuka	Bernyanyi	Kepemimpinan
95	89	96	82	80	83	Aktif	Pramuka	Bernyanyi	Kepemimpinan
96	85	95	75	78	77	Aktif	Pramuka	Bernyanyi	Kepemimpinan
97	92	85	83	75	79	Cukup	Musik	Bernyanyi	Kepemimpinan
98	93	88	83	80	81	Cukup	Pramuka	Bernyanyi	Kepemimpinan
99	89	95	75	81	82	Aktif	Pramuka	Bernyanyi	Kepemimpinan
100	94	93	78	84	75	Cukup	Pramuka	Bernyanyi	Kepemimpinan

B. Karakteristik Dataset

Dataset simulasi terdiri atas 100 siswa dengan komposisi kelas relatif seimbang. Setiap siswa memiliki lima nilai akademik, indikator keaktifan, pilihan ekstrakurikuler, dan hobi. Distribusi kelas yang seimbang digunakan agar evaluasi tidak terlalu dipengaruhi oleh satu kelas mayoritas.

Tabel 3. Distribusi kelas pada dataset

Kelas	Jumlah	Proporsi
Minat Akademik	15	15%
Minat Seni	14	14%
Minat Olahraga	15	15%
Minat Teknologi	14	14%
Minat Bahasa	14	14%
Minat Keagamaan	14	14%
Minat Kepemimpinan	14	14%

C. Entropy Awal

Data latih terdiri atas 80 siswa dan data uji terdiri atas 20 siswa. Entropy awal data latih adalah 2,8060. Nilai ini menggambarkan ketidakpastian awal sebelum pemisahan atribut. Karena kelas relatif seimbang, entropy berada cukup tinggi. Namun, perlu ditekankan bahwa keseimbangan kelas pada dataset simulasi tidak berarti ukuran sampel sudah memadai untuk mengestimasi performa pada populasi siswa yang lebih luas.

D. Perhitungan Gain dan Gain Ratio

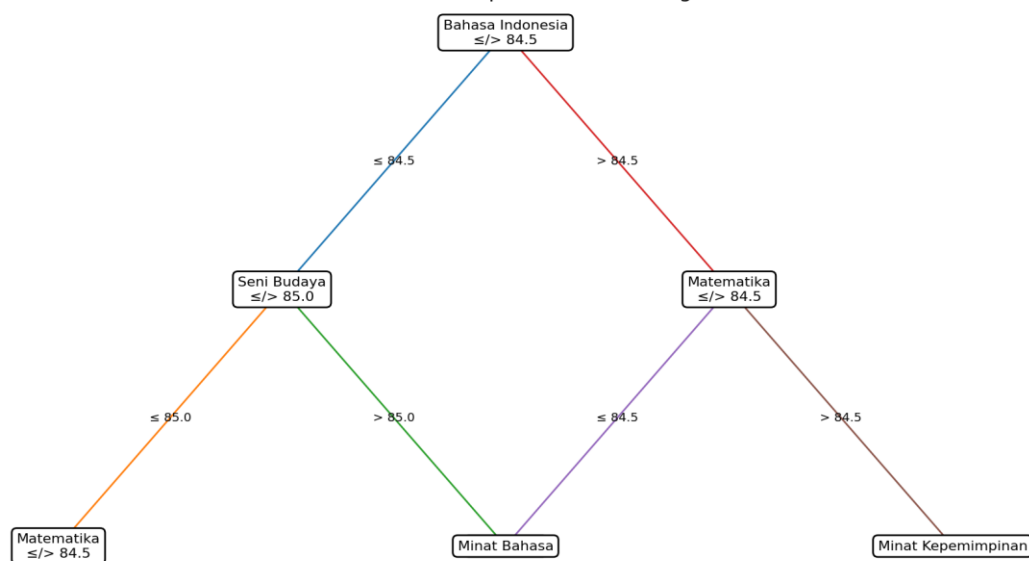
Tabel 4. Perbandingan gain, split information, dan gain ratio

Atribut	Information Gain	Split Information	Gain Ratio
Matematika	0.9837	0.9837	1.0000
Bahasa Indonesia	0.8485	0.8485	1.0000
IPA	0.8655	0.8655	1.0000
Seni Budaya	0.5777	0.5777	1.0000
PJOK	0.6098	0.6098	1.0000
Keaktifan	0.2878	0.9135	0.3150
Ekstrakurikuler	1.9845	2.2279	0.8908
Hobi	1.6898	1.8710	0.9032

Hasil perhitungan menunjukkan bahwa Ekstrakurikuler memiliki gain ratio 0,8908 dan Hobi 0,9032 pada pemeriksaan atribut kategorikal, sedangkan beberapa atribut nilai memiliki threshold yang sangat informatif. Pada pohon yang terbentuk dari keseluruhan proses rekursif, Bahasa Indonesia menjadi pemisah awal dengan threshold 84,5, kemudian cabang menggunakan Seni Budaya dan Matematika. Pemilihan node pada C4.5 dilakukan secara lokal pada setiap subset, sehingga atribut yang terbaik pada root belum tentu menjadi atribut terbaik pada semua cabang.

E. Pohon Keputusan

Visualisasi Pohon Keputusan C4.5 (Ringkas)



Gambar 2. Visualisasi ringkas pohon keputusan C4.5

Pohon keputusan tidak selalu menggunakan seluruh atribut pada setiap jalur. Atribut yang memberikan pemisahan paling informatif diprioritaskan. Struktur ini membuat model lebih ringkas dan memungkinkan aturan klasifikasi diterjemahkan ke dalam bentuk if-then.

F. Aturan Keputusan

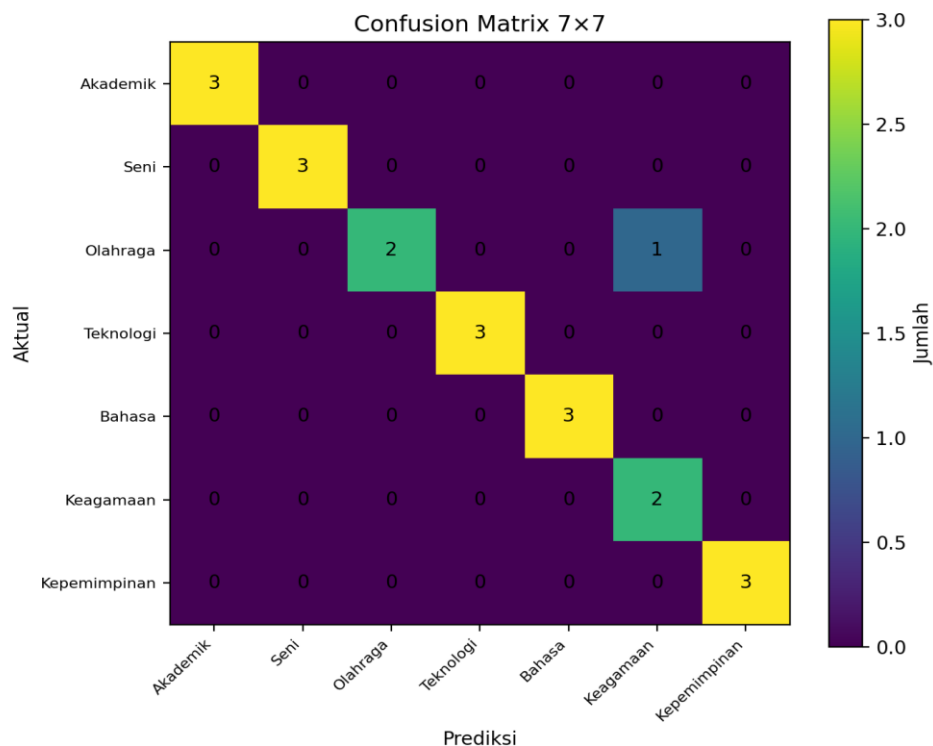
1. Jika Bahasa Indonesia > 84,5 dan Matematika > 84,5, maka kecenderungan klasifikasi mengarah ke Minat Bahasa atau Minat Kepemimpinan berdasarkan pemisahan lanjutan.

2. Jika Bahasa Indonesia $> 84,5$ dan Matematika $\leq 84,5$, maka kecenderungan mengarah ke Minat Bahasa.
3. Jika Bahasa Indonesia $\leq 84,5$ dan Seni Budaya > 85 , maka kecenderungan mengarah ke Minat Seni.
4. Jika Bahasa Indonesia $\leq 84,5$ dan Seni Budaya ≤ 85 , maka model menggunakan pemisahan lanjutan.
5. Cabang dengan PJOK tinggi dan ekstrakurikuler Futsal menunjukkan pola yang kuat menuju Minat Olahraga.
6. Cabang dengan ekstrakurikuler Tahfiz dan pola nilai yang sesuai mengarah pada Minat Keagamaan.
7. Gabungan nilai Matematika/IPA tinggi dengan pola aktivitas tertentu dapat mengarah pada Minat Akademik atau Minat Teknologi.
8. Aturan di atas merupakan ringkasan interpretatif. Pada implementasi lapangan, aturan lengkap sebaiknya diekspor langsung dari model setelah dataset aktual diproses.

G. Confusion Matrix

Tabel 5. Confusion matrix 7×7

Aktual\ Prediksi	Akademik	Seni	Olahraga	Teknologi	Bahasa	Keagamaan	Kepemimpinan
Akademik	3	0	0	0	0	0	0
Seni	0	3	0	0	0	0	0
Olahraga	0	0	2	0	0	1	0
Teknologi	0	0	0	3	0	0	0
Bahasa	0	0	0	0	3	0	0
Keagamaan	0	0	0	0	0	2	0
Kepemimpinan	0	0	0	0	0	0	3



Gambar 3. Confusion matrix 7×7

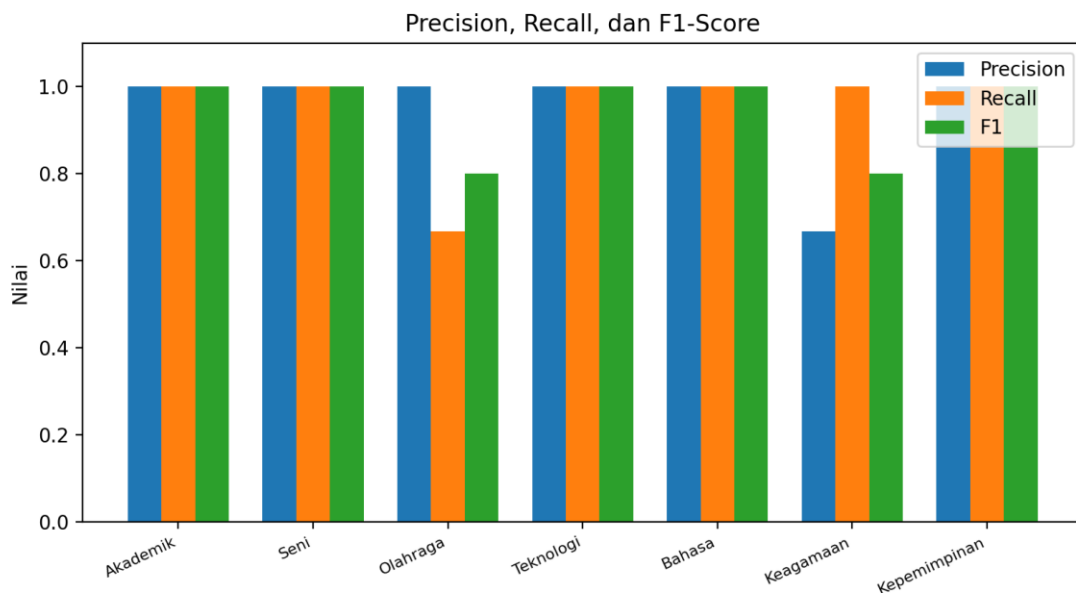
Dari 20 data uji, sebanyak 19 data diklasifikasikan benar dan satu data salah klasifikasi, yaitu satu anggota Minat Olahraga diprediksi sebagai Minat Keagamaan. Accuracy menjadi $19/20 = 95\%$. Secara statistik, dengan interval kepercayaan binomial 95%, proporsi keberhasilan 19 dari 20 observasi berada pada kisaran sekitar 75,1%–99,9%. Rentang yang lebar ini menunjukkan bahwa angka 95% perlu dibaca sebagai hasil eksperimen pada sampel uji kecil, bukan sebagai estimasi pasti performa C4.5 pada populasi siswa.

H. Precision, Recall, dan F1-Score

Tabel 6. Hasil evaluasi klasifikasi per kelas

Kelas	Precision	Recall	F1-Score	Support
Akademik	100.00%	100.00%	100.00%	3
Seni	100.00%	100.00%	100.00%	3
Olahraga	100.00%	66.67%	80.00%	3
Teknologi	100.00%	100.00%	100.00%	3
Bahasa	100.00%	100.00%	100.00%	3
Keagamaan	66.67%	100.00%	80.00%	2
Kepemimpinan	100.00%	100.00%	100.00%	3

Macro precision sebesar 95,24%, macro recall sebesar 95,24%, dan macro F1-score sebesar 94,29%. Sebagian besar kelas memperoleh precision dan recall 100%. Minat Keagamaan memiliki precision 66,67% karena satu prediksi dari Minat Olahraga masuk ke kelas tersebut, tetapi recall-nya 100% karena seluruh anggota Minat Keagamaan pada data uji berhasil dikenali. Nilai-nilai ini konsisten dengan confusion matrix, tetapi support tiap kelas hanya 2–3 data sehingga perubahan satu prediksi dapat mengubah metrik secara besar.



Gambar 4. Precision, recall, dan F1-score per kelas

I. Pembahasan

Penggunaan kombinasi atribut akademik dan nonakademik dapat membantu membentuk klasifikasi minat yang lebih informatif. Temuan ini sejalan dengan penelitian pendidikan berbasis Decision Tree yang menggunakan nilai akademik, minat, dan karakteristik siswa sebagai atribut prediktif ^{[1]–[4]}. Perbedaannya, penelitian ini memetakan tujuh kategori minat pada konteks sekolah dasar.

Atribut ekstrakurikuler dan hobi secara langsung menggambarkan aktivitas yang dipilih siswa, sedangkan nilai akademik membantu menunjukkan kecenderungan kemampuan. Penggabungan keduanya dapat mengurangi ketergantungan pada satu indikator. Penelitian tentang pemilihan program studi juga menunjukkan manfaat penggabungan informasi akademik, minat, dan bakat ^{[3],[9]}.

Accuracy 95% menunjukkan bahwa C4.5 bekerja dengan baik pada 20 data uji dari dataset simulasi yang digunakan dalam eksperimen. Namun, hasil ini tidak cukup untuk menyimpulkan bahwa C4.5 memiliki performa baik secara umum. Selain ukuran data uji yang kecil, dataset dibangun berdasarkan profil kelas yang sengaja memiliki hubungan logis dengan atribut input. Kondisi tersebut dapat membuat pola klasifikasi lebih mudah ditemukan dibandingkan data sekolah nyata yang mungkin mengandung noise, data hilang, ketidakseimbangan kelas, ambiguitas minat, dan hubungan antaratribut yang lebih kompleks. Karena itu, temuan utama penelitian ini sebaiknya dipahami sebagai baseline metodologis dan bukti bahwa alur C4.5 dapat diimplementasikan pada skenario klasifikasi tujuh kelas, bukan sebagai estimasi generalisasi model.

Interpretabilitas menjadi alasan metodologis utama pemilihan C4.5. Guru dapat menelusuri jalur pohon dan menghubungkan prediksi dengan atribut yang digunakan pada setiap percabangan. Keunggulan ini relevan ketika sistem digunakan sebagai alat bantu keputusan, karena hasil prediksi perlu dapat dijelaskan dan dikaji kembali oleh guru. Random Forest dan XGBoost berpotensi memberikan performa prediktif yang lebih kuat pada dataset nyata, tetapi keduanya menggunakan struktur ensemble yang lebih kompleks. Oleh sebab itu, penelitian ini tidak menyatakan C4.5 lebih unggul dari kedua metode tersebut; kontribusinya adalah menunjukkan kesesuaian C4.5 untuk skenario yang menempatkan keterbacaan aturan sebagai kebutuhan utama.

Perbandingan dengan Random Forest dan XGBoost belum dilakukan pada penelitian ini, sehingga tidak tersedia dasar empiris untuk menyatakan metode mana yang paling akurat. Keduanya perlu dijadikan baseline pada penelitian lanjutan dengan dataset yang sama, preprocessing yang sama, pembagian data yang sama, dan skema cross-validation yang identik. Perbandingan tersebut dapat mencakup accuracy, macro precision, macro recall, macro F1-score, ukuran model, waktu pelatihan, serta tingkat interpretabilitas. Selain itu, model sebaiknya diuji pada data dari sekolah yang berbeda untuk menilai generalisasi. Model juga masih menggunakan satu label target, padahal seorang siswa dapat memiliki beberapa minat sekaligus.

Secara praktis, model lebih tepat digunakan sebagai alat bantu pemetaan awal, bukan sebagai pengganti penilaian guru atau hasil angket minat. Guru dapat membandingkan prediksi model dengan observasi, hasil angket, dan konteks perkembangan siswa. Keputusan akhir mengenai kegiatan pengembangan siswa tetap perlu mempertimbangkan konteks individu dan pertimbangan profesional guru.

J. Perbandingan dengan Penelitian Terdahulu

Tabel 7. Perbandingan penelitian

Penelitian	Konteks	Metode	Temuan	Perbedaan
Wicaksono & Setiadi ^[1]	SD, kelulusan	C4.5	Prediksi kelulusan	Target penelitian adalah minat
Rizki & Armansyah ^[2]	Calon mahasiswa	C4.5	Akurasi 85,7%	Objek dan jenjang berbeda
Pratama & Armansyah ^[3]	Pilihan studi lanjut	C4.5	Akademik, minat, bakat	Penelitian ini tujuh kelas minat SD
Pinenda dkk. ^[4]	SMP, hasil belajar	C4.5	Klasifikasi hasil belajar	Target bukan minat
Fahriah dkk. ^[5]	Minat studi	C4.5 + AdaBoost	Akurasi meningkat	Penelitian ini C4.5 dasar
Sukarto dkk. ^[7]	Minat pendidikan SMA	Decision Tree	Klasifikasi minat	Penelitian ini jenjang SD

Halimatussa'diyah & Rohman ^[8]	Prestasi siswa	C4.5	Akurasi 89,7%	Target berbeda
Moli dkk. ^[9]	Minat jurusan SMA	C4.5	Prediksi minat jurusan	Penelitian ini tujuh dimensi minat

K. Kontribusi Penelitian

- Merancang kerangka klasifikasi tujuh kelas minat siswa sekolah dasar yang memadukan indikator akademik dan nonakademik.
- Menjelaskan secara eksplisit prosedur pembentukan dataset sintesis berbasis profil kelas, sehingga sumber keteraturan pola dan potensi keterbatasan generalisasi dapat dinilai.
- Menunjukkan proses perhitungan C4.5 melalui entropy, gain, split information, dan gain ratio serta menghasilkan confusion matrix dan metrik evaluasi.
- Menunjukkan nilai praktis interpretabilitas C4.5 melalui aturan keputusan yang dapat dibaca guru, sekaligus menempatkan Random Forest dan XGBoost sebagai baseline penting untuk pengujian komparatif berikutnya.
- Menyediakan baseline metodologis yang dapat divalidasi dan diperluas menggunakan data sekolah aktual yang telah dianonimkan, cross-validation, dan pengujian lintas sekolah.

L. Keterbatasan Penelitian

- Dataset hanya 100 siswa dan bersifat simulasi; data sekolah nyata yang telah dianonimkan belum tersedia pada tahap penelitian ini.
- Prosedur pembentukan data menggunakan profil kelas yang sengaja berkaitan dengan label target, sehingga pola input-target dapat lebih teratur daripada data lapangan.
- Naskah awal tidak mendokumentasikan skrip generator, distribusi probabilitas, dan random seed, sehingga reproduksibilitas numerik dataset secara identik masih terbatas.
- Belum dilakukan cross-validation, pengujian pada sekolah berbeda, maupun perbandingan empiris dengan Random Forest dan XGBoost.
- Model menggunakan satu label target untuk setiap siswa dan belum memasukkan faktor seperti hasil angket terukur, kehadiran, data longitudinal, atau kemungkinan multi-label.

4. KESIMPULAN

- Algoritma Decision Tree C4.5 dapat diterapkan untuk membangun model awal klasifikasi minat siswa sekolah dasar ke dalam tujuh kelas, yaitu Minat Akademik, Minat Seni, Minat Olahraga, Minat Teknologi, Minat Bahasa, Minat Keagamaan, dan Minat Kepemimpinan. Pada dataset simulasi 100 siswa, model menghasilkan 19 prediksi benar dari 20 data uji dengan accuracy 95%, macro precision 95,24%, macro recall 95,24%, dan macro F1-score 94,29%.
- Angka 95% tersebut harus ditafsirkan secara terbatas sebagai hasil pada eksperimen simulasi. Dengan hanya 20 data uji, interval kepercayaan binomial 95% untuk proporsi 19/20 berada sekitar 75,1%–99,9%. Selain itu, data sintesis dibentuk dari profil minat yang sengaja memiliki hubungan logis dengan atribut input. Dengan demikian, penelitian ini tidak memberikan dasar yang cukup untuk menyatakan bahwa C4.5 memiliki performa baik secara umum atau lebih unggul daripada Random Forest dan XGBoost.
- Kontribusi utama penelitian terletak pada penyusunan baseline klasifikasi tujuh kelas yang transparan, demonstrasi perhitungan C4.5, identifikasi atribut yang informatif, dan penyediaan aturan keputusan yang mudah ditelusuri. Dari sisi metodologis, hasil ini memperlihatkan bahwa C4.5 sesuai ketika interpretabilitas menjadi kebutuhan utama dalam pemetaan awal minat siswa. Namun, kesesuaian tersebut berbeda dari klaim superioritas prediktif; Random Forest dan XGBoost tetap perlu diuji secara empiris sebagai metode pembandingan pada dataset nyata.
- Penelitian berikutnya perlu menggunakan data siswa nyata yang telah dianonimkan dengan persetujuan dan prosedur etis yang sesuai, memperbesar jumlah sampel, mendokumentasikan proses pembentukan dan pembersihan data, menerapkan stratified k-fold cross-validation, menguji

generalisasi lintas sekolah, serta membandingkan C4.5, Random Forest, dan XGBoost dengan protokol evaluasi yang sama. Jika data menunjukkan bahwa satu siswa dapat memiliki lebih dari satu minat, pendekatan multi-label juga perlu dipertimbangkan

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Universitas Potensi Utama yang telah banyak memberikan masukan dan saran dalam penyelesaian karya tulis ilmiah ini.

REFERENSI

- [1]. A. Wicaksono and T. Setiadi, "Penerapan klasifikasi decision tree (C4.5) untuk memprediksi kelulusan siswa Sekolah Dasar di Kecamatan Juai," *Jurnal Format*, vol. 12, no. 2, pp. 151–157, 2023, doi: 10.22441/format.2023.v12.i2.008.
- [2]. P. F. Rizki and Armansyah, "Pendekatan supervised learning Decision Tree C4.5 untuk klasifikasi calon mahasiswa baru," *The Indonesian Journal of Computer Science*, vol. 12, no. 6, 2023, doi: 10.33022/ijcs.v12i6.3595.
- [3]. T. Y. Pratama and Armansyah, "Decision Tree C4.5 dengan teknik information gain untuk klasifikasi pemilihan program studi tingkat lanjut," *Journal of Information System Research*, vol. 5, no. 4, pp. 1042–1052, 2024, doi: 10.47065/josh.v5i4.5643.
- [4]. B. D. P. Pinenda, T. Sutrisno, and B. Mulyawan, "Klasifikasi hasil belajar siswa menggunakan metode C4.5 berdasarkan riwayat akademik di SMP XYZ," *Jurnal Ilmu Komputer dan Sistem Informasi*, vol. 11, no. 2, 2023, doi: 10.24912/jiksi.v11i2.26004.
- [5]. S. Fahriah, N. D. Kamarudin, L. Triyono, and A. Rizaldy, "The addition of AdaBoost to the use of the C4.5 algorithm to improve the accuracy of classification of study interests," *Indonesian Journal of Information Systems*, vol. 6, no. 2, pp. 130–139, 2024, doi: 10.24002/ijis.v6i2.7588.
- [6]. A. Astuti, D. Maryono, and F. Liantoni, "Prediction of PTIK students' study success in the first year using the C4.5 algorithm," *Journal of Soft Computing Exploration*, vol. 5, no. 1, 2024, doi: 10.52465/josce.v5i1.237.
- [7]. M. Sukarto, B. H. Mustawinar, and Y. Yuliani, "Klasifikasi minat siswa Sekolah Menengah Atas dalam melanjutkan pendidikan menggunakan metode Decision Tree," *SAINTIFIK*, vol. 12, no. 1, pp. 121–131, 2026, doi: 10.31605/saintifik.v12i1.615.
- [8]. U. Halimatussa'diyah and A. Rohman, "A prediksi prestasi siswa menggunakan algoritma Decision Tree C4.5," *Jurnal Algoritma*, vol. 23, no. 1, pp. 51–58, 2026, doi: 10.33364/algoritma/v.23-1.3111.
- [9]. K. M. Moli, M. Syawaludin, H. Nabilah, N. Putri, Qudsiya, A. Irianto, and A. Putri, "Prediksi minat jurusan siswa di SMAN 1 Manokwari menggunakan algoritma Decision Tree," *Jurnal Sains Informatika Terapan*, vol. 4, no. 2, 2025, doi: 10.62357/jsit.v4i2.583.
- [10]. I. M. R. P. D. Udayana, G. A. V. M. Giri, I. P. G. H. Suputra, and A. A. I. N. E. Karyawati, "Optimasi algoritma Decision Tree dengan seleksi fitur dalam klasifikasi prestasi akademik siswa sekolah," *JELIKU*, vol. 13, no. 4, 2025, doi: 10.24843/JLK.2025.v13.i04.p06.
- [11]. T. Suprapti, B. Nurhakim, B. W. A. Hermina, and V. A. S. Simbolon, "A Decision Tree model with grid search optimization for scholarship recipient classification," *Jurnal Teknik Informatika (JUTIF)*, vol. 6, no. 5, 2025, doi: 10.52436/1.jutif.2025.6.5.5235.
- [12]. D. Hartama and N. Amalya, "Perbandingan algoritma Decision Tree, ID3, dan Random Forest dalam klasifikasi faktor-faktor yang mempengaruhi karier mahasiswa Ilmu Komputer," *Jurnal Indonesia: Manajemen Informatika dan Komunikasi*, vol. 6, no. 1, pp. 72–80, 2025, doi: 10.35870/jimik.v6i1.1113.

- [13]. I. Akbar, I. S. Samad, R. Rahmat, and S. Rosmiana, "Data mining analysis of K-means algorithm and Decision Tree for early detection of students at risk of dropping out," *INISTA*, vol. 7, no. 2, 2025, doi: 10.20895/inista.v7i2.1630.
- [14]. E. Desvazulinda, M. Iqbal, and M. S. Novelan, "Analysis of the Decision Tree (C4.5) and Random Forest algorithms to determine student eligibility for final project assignments based on academic requirements," *Jurnal Sistem Komputer dan Informatika (JSON)*, vol. 7, no. 4, 2026, doi: 10.30865/json.v7i4.9947.
- [15]. A. H. Ibrahimy and Z. F. Ibrahimy, "Decision Tree application of Decision Tree C4.5 in predicting the best predicate in Madrasah Ta'hiliyah Ibrahimy," *JITTER: Jurnal Ilmiah Teknologi dan Komputer*, vol. 5, no. 3, 2024, doi: 10.24843/JTRTI.2024.v05.i03.p08.
- [16]. B. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
- [17]. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016, doi: 10.1145/2939672.2939785