SAINTIFIK:
Jurnal Matematika, Sains, dan Pembelajarannya **Vol.12, No.1, Januari 2026, pp.121~131**
ISSN 2407-4098 (print) DOI:10.31605/saintifik.v12i1.615
ISSN 2622-8904 (online)

Klasifikasi Minat Siswa Sekolah Menengah Atas dalam Melanjutkan Pendidikan Menggunakan Metode *Decision Tree*

Made Sukarto[1], Besse Helmi Mustawinar[2]*, Yuliani[3]
[1,2,3]Program Studi Matematika, Fakultas Sains, Universitas Cokroaminoto Palopo, Indonesia
e-mail: emmy.emm92@gmail.com

Abstrak

Penelitian ini dilakukan untuk mencari faktor-faktor yang dinilai memiliki pengaruh terhadap minat siswa dalam melanjutkan studi ke tingkat perguruan tinggi, khususnya di SMA Negeri 6 Luwu Timur. Hal ini penting untuk dikaji karena minat melanjutkan studi sangat menentukan keberlanjutan pendidikan dan sumber daya manusia. Atribut dalam penelitian ini adalah biaya kuliah, beasiswa, penghasilan orang tua, status dan lokasi perguruan tinggi, rekomendasi, motivasi, serta lingkungan. Metode yang digunakan adalah teknik klasifikasi decision tree dengan algoritma C5.0 dengan menggunakan perangkat lunak WEKA untuk pengujian data dan RapidMiner untuk pembangunan model. Hasil penelitian menunjukkan bahwa faktor beasiswa, penghasilan orang tua, dan rekomendasi merupakan variabel paling signifikan dalam menentukan minat siswa. Pengujian dilakukan dengan lima metode yaitu use training set, cross validation dengan 5-folds dan 10-folds dan percentage split dengan menggunakan 60% percentage split dan 80% percentage split. Metode dengan akurasi tertinggi adalah metode 60% percentage split sebesar 87,23%. Hasil klasifikasi menunjukkan jika terdapat 18 aturan yang terdiri atas 12 aturan dengan keputusan melanjutkan studi dan 6 aturan dengan keputusan tidak melanjutkan studi. Temuan ini menegaskan bahwa dukungan finansial dan sosial berperan besar dalam membentuk minat siswa, sehingga hasil penelitian ini penting sebagai dasar kebijakan sekolah maupun perguruan tinggi dalam meningkatkan minat siswa terhadap pendidikan lanjutan.

Kata kunci– *Minat siswa, Data Mining, Decision Tree, Algoritma C5.0.*

1. PENDAHULUAN

Pendidikan tinggi memegang peranan vital dalam pembentukan sumber daya manusia (SDM) unggul yang memberikan kontribusi strategis bagi pembangunan nasional. Namun, rendahnya transisi lulusan sekolah menengah ke jenjang universitas masih menjadi tantangan serius di Indonesia. Data Badan Pusat Statistik (BPS) pada tahun 2023 menunjukkan adanya kesenjangan yang signifikan, dimana hanya sekitar 1,8 juta dari total 3,7 juta lulusan SMA/SMK yang memilih untuk melanjutkan studi, atau sekitar 49% siswa yang lanjut untuk kuliah (BPS, 2024). Fenomena ini dipicu oleh berbagai faktor termasuk di dalamnya keterbatasan finansial, motivasi internal, pengaruh lingkungan sosial, dan minimnya akses terhadap informasi pendidikan tinggi(Ayunda et al., 2024).

Keputusan seorang siswa untuk melanjutkan kuliah bukan hal yang sederhana. Ada proses panjang dimana perasaan dan logika bertentangan sebelum akhirnya minat melanjutkan studi terealisasi. Fakta lapangan menunjukkan jika faktor ekonomi, jenis kelamin, sampai sifat bawaan dari siswa itu sendiri punya andil besar dalam menentukan arah minat mereka (Makhrisa & Pradikto, 2025). Mengingat variabelnya yang sangat beragam dan kadang sulit ditebak, sekolah tidak bisa lagi hanya mengandalkan cara-cara manual untuk memahaminya. Pihak sekolah membutuhkan alat analisis yang bisa memetakan pola-pola minat tersebut secara lebih mendalam. Dengan begitu, bantuan atau arahan yang diberikan bimbingan konseling bisa benar-benar pas dan menjawab kebutuhan spesifik tiap kelompok siswa.

Di era transformasi digital, penerapan *Educational Data Mining* (EDM) menjadi solusi krusial dalam mengekstraksi pengetahuan dari data institusi pendidikan untuk mendukung pengambilan keputusan. EDM memungkinkan identifikasi pola tersembunyi dalam dataset berskala besar yang tidak dapat dilakukan secara manual. Agar proses ekstraksi ini optimal, data mentah harus melalui tahap *preprocessing*, mencakup

pembersihan dan transformasi, sehingga menghasilkan model yang akurat dan bernilai guna (Rizki, 2024). Dalam lingkup EDM, teknik klasifikasi menjadi metode yang paling banyak diimplementasikan. Salah satu algoritma yang paling menonjol adalah *Decision Tree* (Pohon Keputusan). Berbeda dengan model *machine learning* yang bersifat *black-box*, *Decision Tree* memberi keunggulan berupa interpretabilitas tinggi melalui transformasi data kompleks ke dalam aturan keputusan (*if-then rules*) yang mudah dipahami oleh praktisi pendidikan. Keunggulan ini sangat relevan dalam konteks sekolah, karena guru bimbingan konseling dapat secara langsung mengidentifikasi variabel paling dominan yang memengaruhi minat siswa.

Sejumlah studi terdahulu telah membuktikan efektivitas *Decision Tree* dalam domain pendidikan. (Rizki, 2024) menggunakan pendekatan *supervised learning* untuk mengklasifikasi profil mahasiswa baru dengan tingkat akurasi yang baik. Selanjutnya, (Pratama & Armansyah, 2024) menggali teknik *information gain* untuk memetakan pemilihan program studi berdasarkan latar belakang akademik. Dalam studi spesifik mengenai prediksi minat, (Ayunda et al., 2024) berhasil merumuskan aturan keputusan yang secara presisi membedakan kelompok siswa berminat dan tidak berminat melanjutkan studi.

Meskipun metode *Decision Tree* telah banyak diterapkan, sebagian besar riset masih menggunakan algoritma dasar seperti C4.5. Penelitian ini menggunakan Algoritma C5.0, yang merupakan pengembangan dengan keunggulan pada kecepatan pemrosesan data, penggunaan memori yang lebih efisien, serta akurasi yang lebih tajam melalui mekanisme *boosting*. Pemanfaatan C5.0 dalam mengklasifikasikan kasus kesehatan telah terbukti efektif (Nugrahani & Prapanca, 2025). Namun, eksplorasi algoritma ini dalam memetakan minat siswa SMA masih terbatas. Berdasarkan kesenjangan literatur tersebut, penelitian ini bertujuan untuk mengklasifikasikan minat siswa SMA dalam melanjutkan studi ke perguruan tinggi dengan menerapkan algoritma C5.0. Studi ini mengambil lokasi di SMA Negeri 6 Luwu Timur sebagai upaya memberikan kontribusi praktis bagi pihak sekolah dalam menyusun strategi bimbingan karir berbasis data, sekaligus memperkaya diskusi EDM di Indonesia.

2. METODE PENELITIAN

Dalam bidang analisis data, *decision tree* didefinisikan sebagai teknik pemodelan prediktif yang memanfaatkan diagram bercabang atau struktur hierarkis. Fungsi utamanya adalah untuk memvisualisasikan dan merumuskan keputusan berdasarkan serangkaian kondisi dan kriteria yang ditetapkan. Pohon keputusan adalah salah satu pendekatan dalam *machine learning* yang menerapkan hierarki aturan klasifikasi berurutan, yang melibatkan pembagian dataset pelatihan secara berulang (rekursif) (Ananta et al., 2024).

Algoritma C5.0 adalah pengembangan mutakhir dari algoritma berbasis *decision tree*, yang merupakan hasil penyempurnaan dari pendahulunya, yaitu Algoritma ID3 dan C4.5, yang diperkenalkan oleh Ross Quinlan pada tahun 1987 (Han et al., 2012). Algoritma C5.0 dirancang untuk memproses tipe atribut data baik yang bersifat diskrit maupun kontinu. Proses kunci dalam algoritma ini adalah seleksi atribut yang didasarkan pada perhitungan *information gain*. Atribut yang memiliki nilai *information gain* paling tinggi akan ditetapkan sebagai simpul (*node*) akar untuk cabang berikutnya. Khusus dalam pembangunan pohon keputusan menggunakan algoritma C5.0, rasio perolehan (*gain ratio)* dihitung berdasarkan nilai *entropy* dan *information gain* untuk menentukan simpul akar. Simpul yang terpilih sebagai akar adalah atribut dengan nilai *gain ratio* tertinggi yang berkorespondensi dengan entropy terkecil. Rumusan matematika untuk menghitung rasio perolehan (*gain ratio*) adalah sebagai berikut:

$$Gain\ Ratio = \frac{Gain\ (A)}{\sum_{i=i}^{m}\ Entropy\ (Aij)} \quad (1)$$

Rumus yang digunakan untuk menghitung nilai *gain* dan *entropy* adalah sebagai berikut:

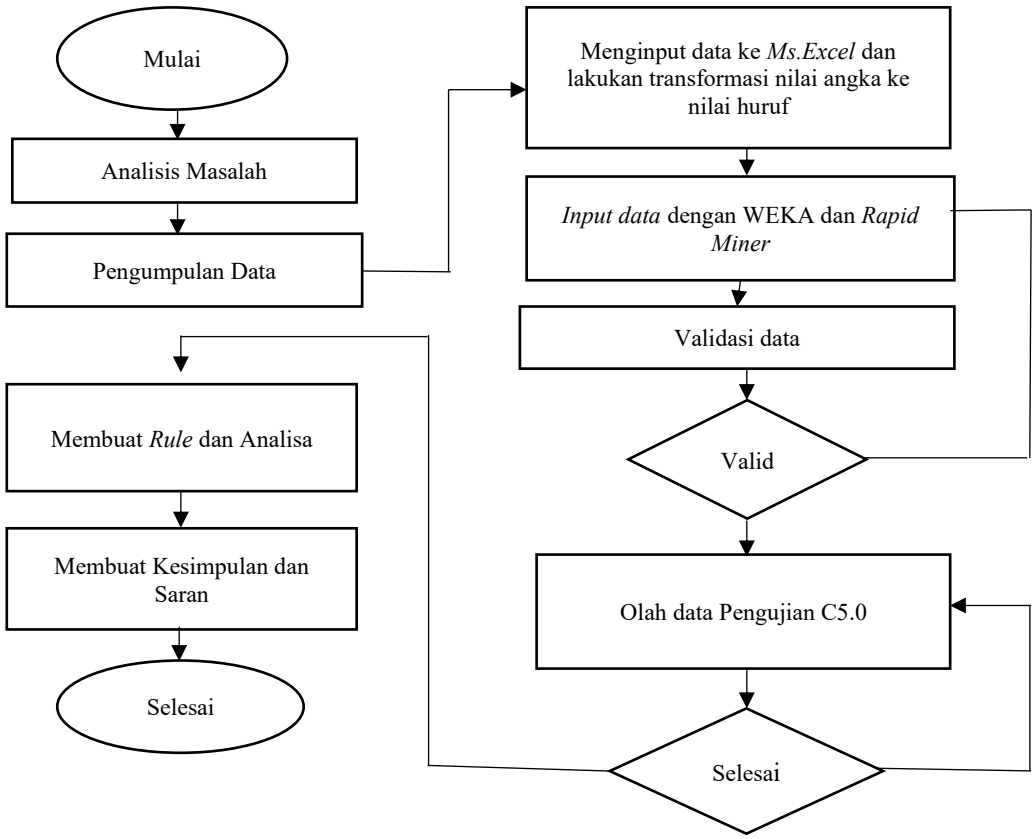
$$Entropy\ (S) = \sum_{i=1}^{n} -\ pi\ log2\ (pi) \quad (2)$$

Selanjutnya menghitung nilai dari *gain* menggunakan rumus berikut:

$$Gain\ (S, A) = entropy\ (S) - \sum_{i=1}^{n} \frac{|\ Si\ |}{|\ S\ |} Entropy\ (Si) \quad (3)$$

Dimana S = Himpunan Kasus, A = Atribut, n = Jumlah partisi atribut A, |Si| = Proporsi Si terhadap S, dan |S| = Jumlah kasus dalam S.

Berikut ini adalah prosedur penelitian yang akan digunakan dapat dilihat sebagai berikut:



Gambar 1 *Flowchart* Penelitian

Studi ini menggunakan sumber data utama yang dikumpulkan melalui penyebaran instrumen kuesioner. Kuesioner tersebut memuat berbagai atribut yang diperlukan untuk klasifikasi minat siswa dalam meneruskan studi ke jenjang perguruan tinggi. Populasi penelitian ini adalah seluruh siswa SMA Negeri 6 Luwu Timur, dengan sampel yang diambil mencakup pelajar kelas 12 dari kedua jurusan, baik Ilmu Pengetahuan Alam (IPA) maupun Ilmu Pengetahuan Sosial (IPS). Setelah pengumpulan, total data yang berhasil dihimpun berjumlah 118 entri. Variabel yang terdapat dalam dataset ini meliputi penghasilan orang tua, biaya kuliah, beasiswa, status perguruan tinggi, lokasi perguruan tinggi, motivasi, lingkungan, rekomendasi, serta variabel target (label) yaitu minat siswa, yang dapat diamati pada tabel berikut.

Tabel 1 Dataset Penelitian

Atribut	Kategori	Jumlah	Atribut	Kategori	Jumlah
Minat	Ya	84	Lokasi PT	Jauh	85
	Tidak	34		Dekat	33
Biaya Kuliah	Mahal	64	Rekomendasi	Ya	94
	Terjangkau	54		Tidak	24
Beasiswa	Perlu	104	Motivasi	Keluarga	28
	Tidak perlu	14		Guru sekolah	62
Penghasilan orang tua	Miskin	35		Lingkungan masyarakat	18
	Rentan	8		Teman bermain	10
	Bawah	35	Lingkungan		
	Menengah	40		Mendukung	61
Status PT	Negeri	92		Tidak mendukung	57
	Swasta	26			

Tahap *preprocessing* merupakan langkah krusial dalam menyaring dan mengoptimalkan data, dengan memfokuskan perhatian hanya pada atribut yang relevan dan akan digunakan dalam proses perhitungan klasifikasi. Mengingat data yang digunakan adalah data real (data mentah), terdapat kemungkinan adanya *noise* atau ketidaksempurnaan, sehingga wajib melalui serangkaian tahapan berikut:

1. Eliminasi duplikasi data dimana langkah ini bertujuan untuk mereduksi atau menghilangkan catatan data yang bersifat redundan atau tidak diperlukan. Adanya duplikasi, yaitu entri data yang memiliki nilai identik, berpotensi menurunkan akurasi dan keandalan hasil yang diperoleh selama proses data mining.
2. Penanganan Data yang Hilang (*Missing Value*) yaitu proses pembersihan data (*data cleaning*) mencakup upaya untuk melengkapi nilai yang kosong, mengoreksi inkonsistensi data, dan mendeteksi penumpukan data. Berdasarkan pemeriksaan yang dilakukan, dataset yang digunakan dalam studi ini tidak memiliki *missing value*, sehingga memungkinkan peneliti untuk langsung melanjutkan ke tahapan berikutnya.
3. Transformasi Atribut Data dimana tahap ini melibatkan modifikasi atau reduksi atribut yang dinilai kurang signifikan pengaruhnya terhadap hasil akhir pembentukan pohon keputusan (*decision tree*). Rincian perubahan pada atribut yang digunakan dalam penelitian ini adalah sebagai berikut:

Tabel 2. Transformasi Dataset Penelitian

No.	Sebelum Transformasi	Sesudah Transformsi
1	Minat	Minat
2	Biaya Kuliah	Biaya Kuliah
3	Beasiswa	Beasiswa
4	Biaya Lain	Penghasilan Orang Tua
5	Penghasilan Orang Tua	Status Perguruan Tinggi
6	Status Perguruan Tinggi	Lokasi Perguruan Tinggi
7	Lokasi Perguruan Tinggi	Rekomendasi
8	Rekomendasi	Motivasi
9	Motivasi	Lingkungan
10	Lingkungan	
11	Lingkungan Masyarakat	
12	Peluang Kerja	

Atribut "Biaya Lain" dieliminasi karena ditemukan ketidakkonsistenan pola data (*noise*) yang cukup tinggi; dimana nilai pada atribut ini seringkali bertentangan dengan atribut "Beasiswa" dan "Minat", sehingga dapat menurunkan stabilitas prediksi model. Selanjutnya, atribut "Lingkungan Masyarakat" dihilangkan akibat adanya masalah homogenitas data. Atribut ini memiliki kemiripan informasi yang sangat tinggi dengan atribut "Lingkungan" dan sebaran nilainya cenderung seragam, sehingga tidak memiliki daya beda yang cukup untuk menjadi simpul keputusan. Terakhir, atribut "Peluang Pekerjaan" direduksi karena tidak menunjukkan pengaruh signifikan. Hal ini mengindikasikan bahwa minat melanjutkan studi pada lokasi penelitian lebih didorong oleh motivasi akademik dan dukungan sosial langsung dibandingkan pertimbangan pragmatis mengenai peluang kerja, sehingga penghapusan atribut ini dapat membantu merampingkan model agar lebih fokus pada variabel yang memiliki signifikansi statistik yang kuat.

Untuk memastikan model klasifikasi yang dihasilkan memiliki validitas dan keandalan yang tinggi, penelitian ini menerapkan strategi evaluasi ganda melalui teknik *k-Fold Cross-Validation* dan *Percentage Split.* Pemilihan *k-Fold Cross-Validation* didasarkan pada ketangguhannya dalam memberikan estimasi performa yang lebih stabil dibandingkan metode *hold-out* yang konvensional. Dengan membagi dataset menjadi k bagian yang diuji secara berulang, setiap data mendapatkan kesempatan yang sama untuk berperan sebagai data uji maupun data latih. Pendekatan ini menjadi sangat krusial untuk menekan risiko *overfitting* serta meminimalisir bias, mengingat dataset yang bersumber dari satu sekolah memiliki keterbatasan jumlah sampel. Melalui cara ini, model C5.0 yang terbentuk diharapkan memiliki daya generalisasi yang kuat pada berbagai partisi data.

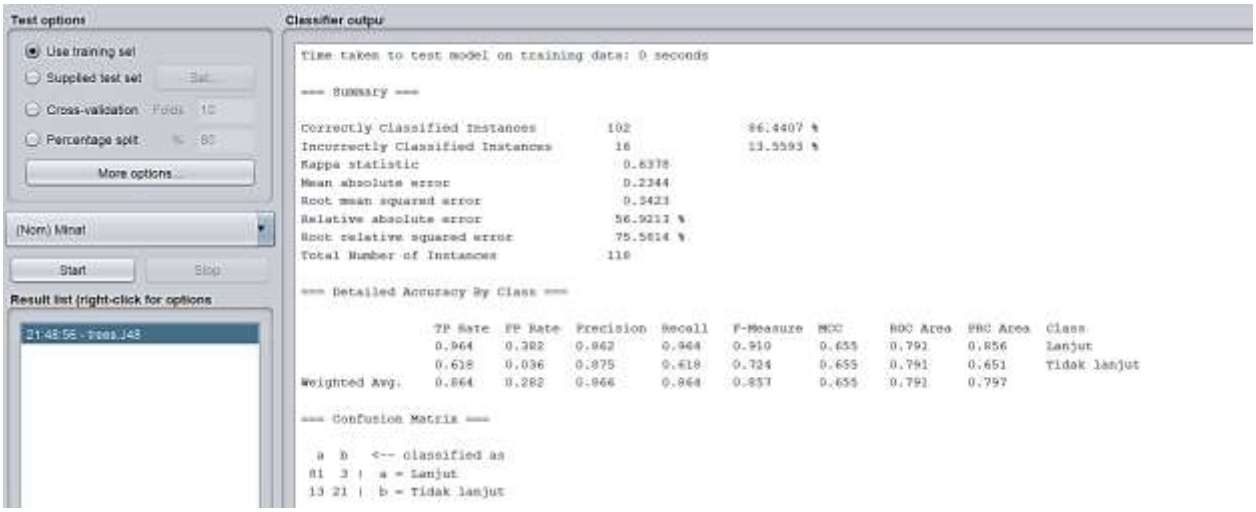
Dalam implementasinya, pengujian model dilakukan melalui tiga skema utama untuk memverifikasi konsistensi akurasi dari berbagai sudut pandang. Pertama, *Use Training Set* digunakan untuk mengukur performa awal model pada data latih. Kedua, *Cross-Validation* diterapkan guna memantau stabilitas performa algoritma. Ketiga, *Percentage Split* dilakukan untuk mensimulasikan sejauh mana model mampu memprediksi data baru yang belum pernah dikenali sebelumnya. Keseluruhan skema ini bertujuan untuk memastikan bahwa akurasi yang dihasilkan bukan merupakan hasil kebetulan, melainkan cerminan dari pola data yang konsisten.

Untuk optimalisasi hasil, penelitian ini mengintegrasikan dua perangkat lunak yang berbeda. Tahap Seleksi Atribut dilakukan di WEKA dengan mengombinasikan metode *Forward Selection* dan *10-Fold Cross-Validation*. Prosedur ini bekerja secara bertahap, dimulai dari himpunan kosong lalu menambahkan atribut satu per satu berdasarkan kontribusi nyatanya terhadap performa model. Langkah ini sangat efektif untuk mereduksi dimensi data dan membuang atribut yang redundan secara objektif. Setelah diperoleh subset atribut yang paling signifikan, tahap Pemodelan Klasifikasi dilanjutkan menggunakan RapidMiner. Perangkat lunak ini dipilih untuk mengonstruksi pohon keputusan dengan algoritma C5.0 karena keunggulannya dalam visualisasi model yang intuitif serta fleksibilitasnya dalam ekstraksi aturan (rules) keputusan. Sinergi antara analisis statistik di WEKA dan efisiensi pemodelan di RapidMiner diharapkan mampu menghasilkan model yang tidak hanya akurat secara teknis, tetapi juga mudah diinterpretasikan oleh pihak sekolah dalam mengambil kebijakan.

3. HASIL DAN PEMBAHASAN

3.1 Metode Use data training

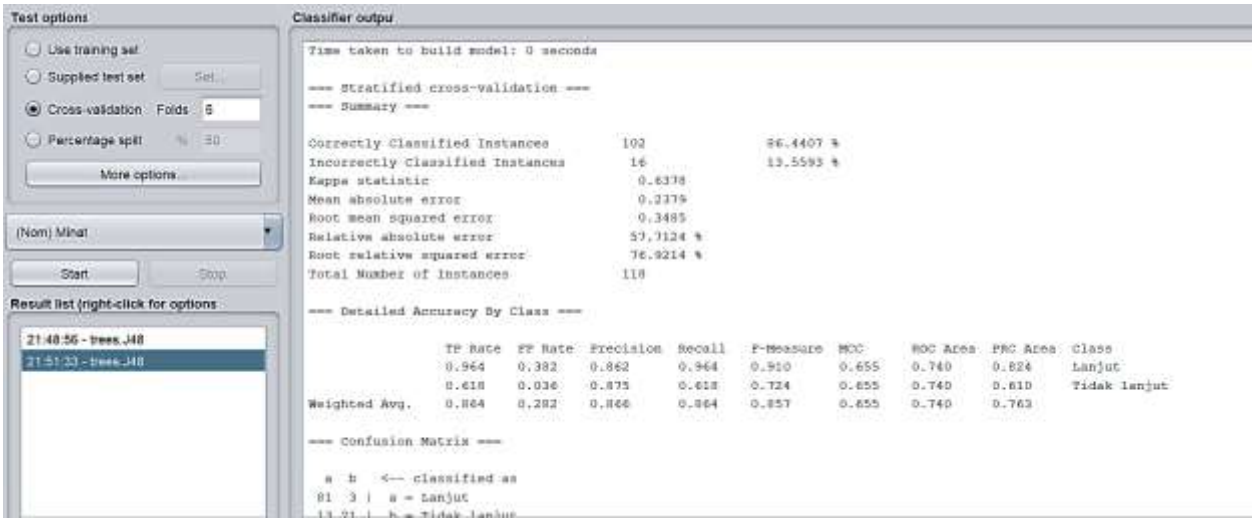
Hasil uji menggunakan metode *Use data training set* menunjukkan bahwa sebanyak 102 data dengan prediksi benar yang memiliki tingkat akurasi kebenaran sebesar 86,44% dan 16 data dengan prediksi salah dengan tingkat persentase sebesar 13,55% dengan waktu uji selama 0 detik yang dapat dilihat pada Gambar 2 berikut.



Gambar 2 Hasil uji menggunakan *Use data training*

3.2 Metode Cross-validation

Untuk mengevaluasi kinerja model klasifikasi yang dikembangkan, peneliti menggunakan teknik pengujian *Cross-Validation*. Metode ini diterapkan untuk menilai model secara lebih objektif dengan cara memecah dataset menjadi beberapa kelompok bagian (*folds*). Melalui pembagian ini, setiap segmen data memperoleh peluang yang setara untuk berperan sebagai data pelatihan maupun data pengujian. Dalam studi ini, proses pengujian *Cross-Validation* dilakukan dengan variasi sebesar 5 *fold* dan 10 *fold*. Hasil dari pengujian menunjukkan tingkat akurasi yang memuaskan. Dari total 118 data yang diujikan, model berhasil memprediksi dengan tepat sebanyak 102 data, menghasilkan persentase akurasi sebesar 86,44%. Sementara itu, jumlah data yang diprediksi keliru adalah 16, setara dengan persentase kesalahan 13,55%. Hasil uji tersebut dapat dilihat pada Gambar 3 berikut.

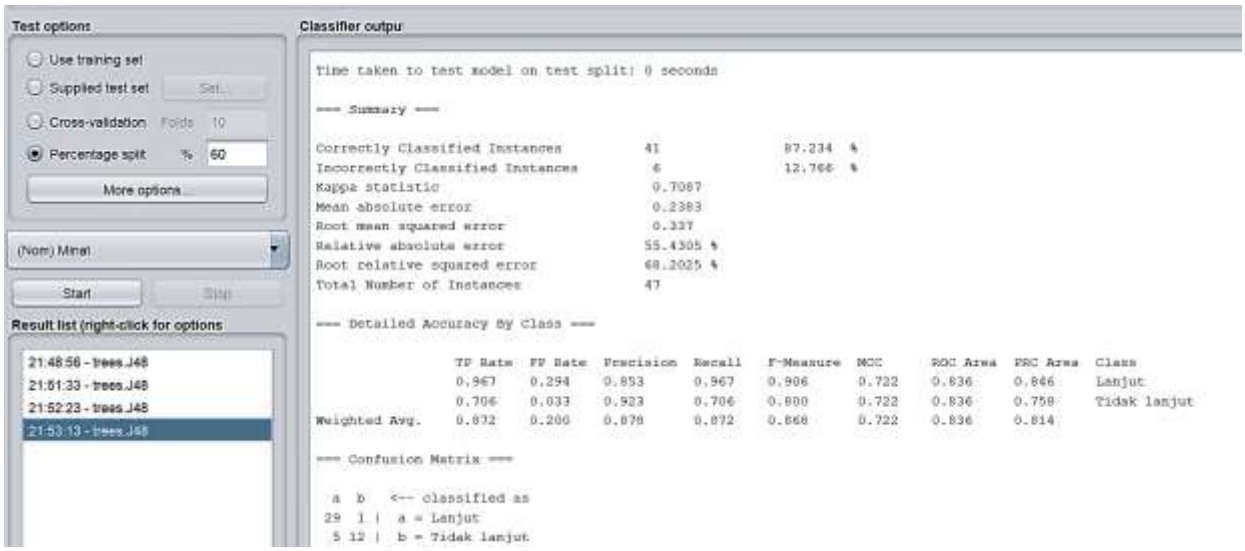


Gambar 3 Hasil uji menggunakan *5 fold* dan *10 fold Cross validation*

3.3 Metode Percentage split

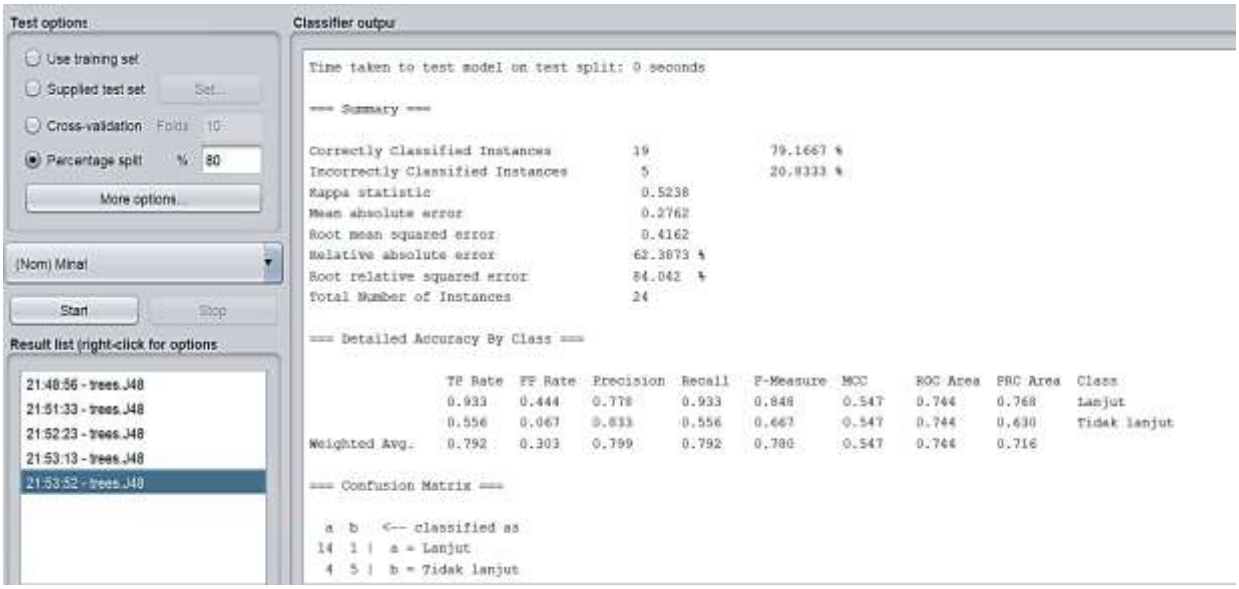
Pengujian Pembagian Persentase adalah suatu prosedur evaluasi model yang memisahkan himpunan data menjadi dua bagian dengan perbandingan persentase yang telah ditentukan. Data yang ada akan dikelompokkan menjadi set data untuk pelatihan (*training*) dan set data untuk pengujian (*testing*). Dalam penelitian ini, pembagian data direncanakan menggunakan dua skema, yaitu 60% dan 80% (yang berarti 60:40 dan 80:20).

Pada skema pertama, data dibagi dengan alokasi 60% digunakan sebagai data pelatihan dan 40% sebagai data pengujian. Hasil evaluasi model memperlihatkan bahwa terdapat 42 prediksi yang benar, mencapai tingkat akurasi sebesar 87,23%. Sementara itu, ditemukan 6 prediksi yang keliru, merepresentasikan persentase kesalahan sebesar 12,76%. Hasil pengujian tersebut dapat dilihat pada Gambar 4 berikut.



Gambar 4 Hasil uji menggunakan *60% Percentage split*

Uji selanjutnya dilakukan dengan membagi data menjadi 2 bagian yaitu sebanyak 80% sebagai *data training* dan 20% sebagai *data testing*. Hasil uji menunjukkan dari 118 data yang diuji sebanyak 19 data dengan prediksi benar yang memiliki tingkat akurasi sebesar 79,16% sementara 5 data lainnya diprediksi salah dengan tingkat akurasi sebesar 20,83% dengan waktu uji selama 0 detik. Hasil uji tersebut dapat dilihat pada Gambar 5 berikut.



Gambar 5 Hasil uji menggunakan *80% Percentage split*

Berdasarkan hasil pengujian model yang telah dilakukan menggunakan perangkat lunak WEKA, performa algoritma C5.0 diukur melalui beberapa skema evaluasi untuk memastikan konsistensi akurasinya. Data hasil pengujian tersebut disajikan secara rinci pada Tabel 3.

Tabel 3 Hasil *Data Testing Decision Tree*

Model Evaluasi	Akurasi	Jumlah kelas	Persentase
Use Training Set	Diklasifikasikan dengan Benar	102	86,44%
	Diklasifikasikan dengan Salah	16	13,55%
5 Fold Cross Validation	Diklasifikasikan dengan Benar	102	86,44%
	Diklasifikasikan dengan Salah	16	13,55%
10 Fold Cross Validation	Diklasifikasikan dengan Benar	102	86,44%
	Diklasifikasikan dengan Salah	16	13,55%
60% Percentage Split	Diklasifikasikan dengan Benar	41	87,23%
	Diklasifikasikan dengan Salah	6	12,76%
80% Percentage Split	Diklasifikasikan dengan Benar	19	79,16%
	Diklasifikasikan dengan Salah	5	20,83%

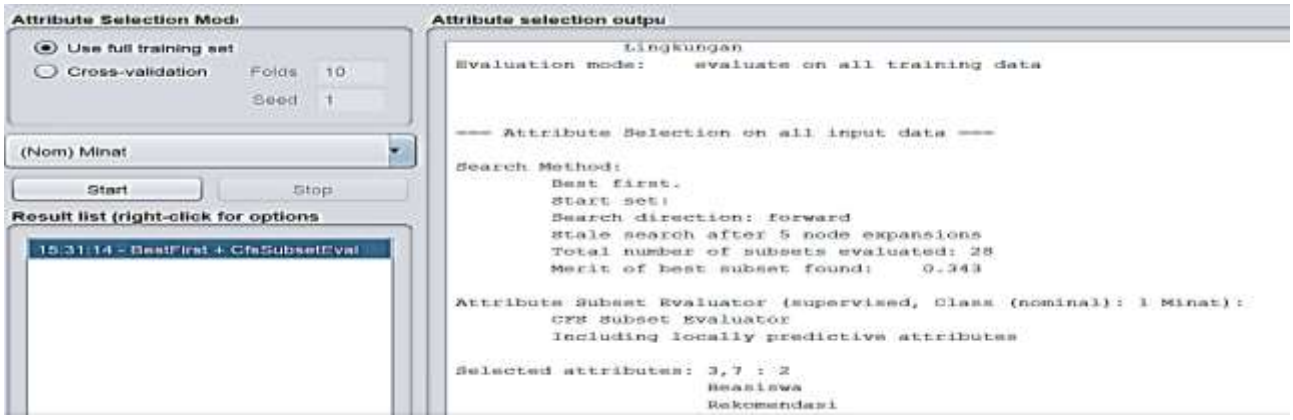
Berdasarkan dari tabel 2, model C5.0 menunjukkan stabilitas tinggi dengan akurasi konsisten 86,44% pada pengujian *Training Set* hingga *10-Fold Cross Validation*. Hal ini membuktikan bahwa pohon keputusan yang terbentuk mampu menangkap struktur informasi secara mendalam. Untuk skema 60% *Percentage Split* mencapai akurasi tertinggi sebesar 87,23%, sementara skema 80:20 menurun ke 79,16%. Penurunan ini dipicu oleh terbatasnya jumlah data uji yang membuat model lebih sensitif terhadap kesalahan klasifikasi. Meski demikian, rata-rata akurasi di atas 80% mengonfirmasi bahwa algoritma ini sangat handal untuk memprediksi minat studi lanjut siswa di SMA Negeri 6 Luwu Timur. Jika dibandingkan dengan hasil riset sebelumnya oleh (Ayunda et al., 2024) yang mencatat akurasi sebesar 80% dengan menggunakan algoritma C4.5, performa algoritma C5.0 dalam penelitian ini menunjukkan hasil yang sangat kompetitif, dengan nilai 87,23%. Selisih akurasi ini kemungkinan besar dipengaruhi oleh keunggulan intrinsik algoritma C5.0 yang memang dirancang untuk lebih efisien dan akurat dalam menangani data klasifikasi dibandingkan versi terdahulu.

Selanjutnya, adalah optimalisasi melalui seleksi atribut dilakukan untuk mengidentifikasi variabel yang paling relevan dalam memprediksi minat siswa. Proses seleksi ini dijalankan menggunakan perangkat lunak WEKA dengan menerapkan dua metode utama, yaitu: 1) *Forward Selection*, dan 2) *Forward Selection yang terintegrasi dengan Cross-Validation*. Langkah ini diambil untuk memastikan bahwa hanya atribut

dengan kontribusi signifikan yang digunakan dalam pembentukan model, sehingga meningkatkan efisiensi dan akurasi klasifikasi akhir.

3.4 Metode Forward Selection

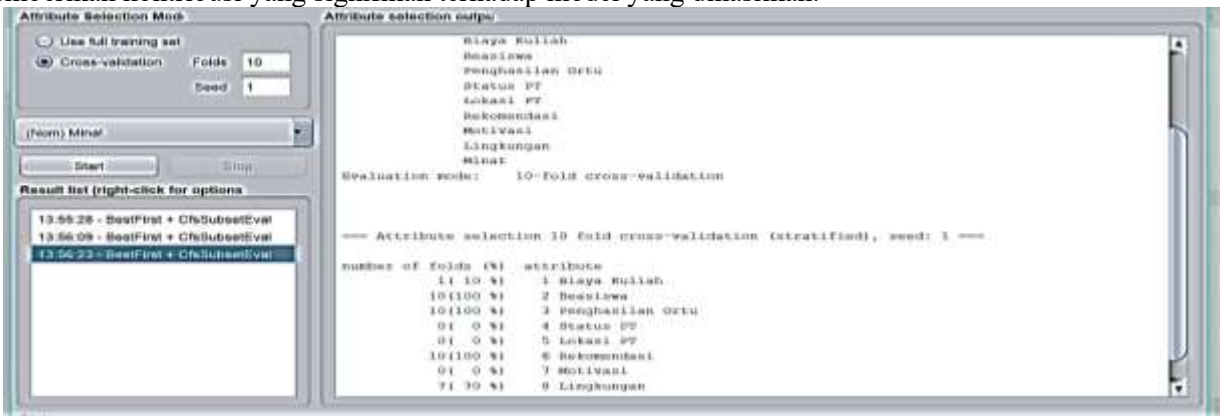
Pemilihan fitur dilaksanakan melalui metode *Forward Selection*. Prosedur ini merupakan tahapan yang dimulai *tanpa* memasukkan atribut apa pun ke dalam model, kemudian secara bertahap menambahkan atribut yang dinilai paling berkontribusi dalam meningkatkan kinerja klasifikasi. Tujuannya adalah untuk mengidentifikasi subsetatribut yang dianggap paling signifikan pengaruhnya terhadap hasil klasifikasi akhir. Hasil dari pengujian yang telah dijalankan memperlihatkan bahwa dua atribut, yaitu Beasiswa dan Rekomendasi, memiliki dampak yang sangat signifikan pada pembentukan model pohon keputusan terkait minat siswa untuk melanjutkan pendidikan ke jenjang universitas. Detail temuan ini dapat diamati pada gambar berikut.



Gambar 6 Hasil seleksi dengan metode *Forward selection*

3.5 Forward Selection dengan Cross Validation

Uji yang dilakukan untuk memperoleh subset atribut yang paling relevan dan stabil dalam memprediksi variabel target. Hal ini dilakukan untuk memastikan bahwa atribut yang terpilih benar-benar memberikan kontribusi yang signifikan terhadap model yang dihasilkan.

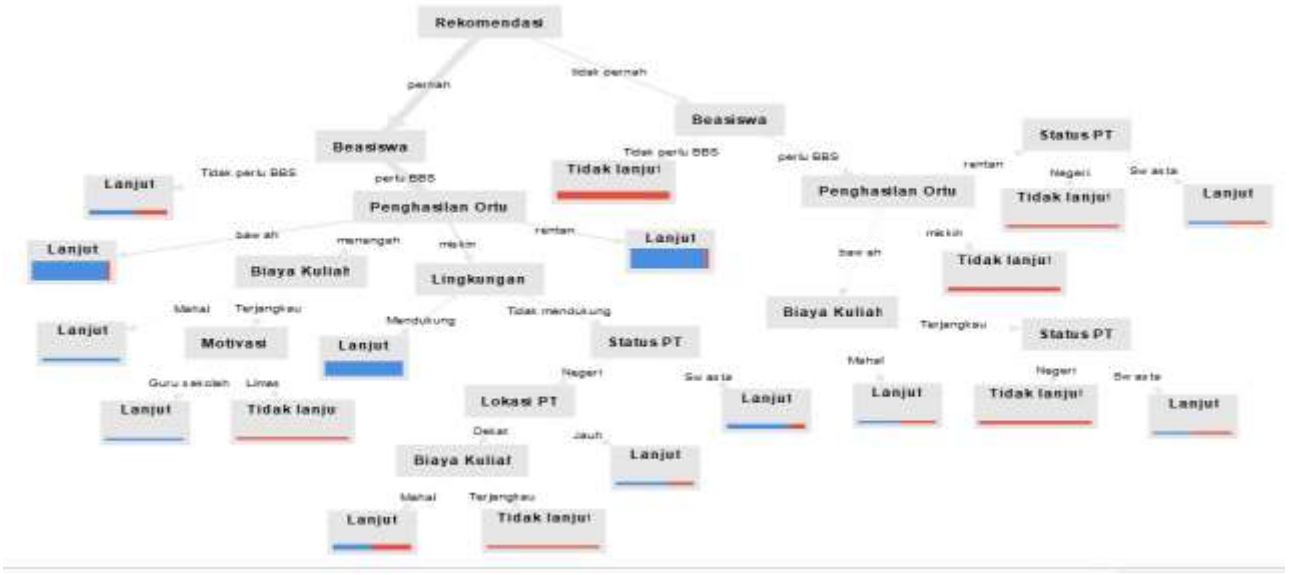


Gambar 7 Hasil seleksi metode *Forward Selection* dengan *Cross Validation*

Proses seleksi dilakukan dengan menggunakan teknik *Cross validation* dengan *10-Folds seed* sebesar 1. Hasil atribut penelitian yang diuji bahwa atribut biaya kuliah memiliki pengaruh pada model sebesar 10%, atribut lingkungan 70%, atribut beasiswa, penghasilan orang tua, dan rekomendasi memiliki pengaruh 100%. Atribut-atribut hasil seleksi yang telah terpilih melalui proses tersebut, kemudian digunakan untuk keperluan data mining pada *decision tree* dengan menggunakan Algoritma C5.0 untuk mengklasifikasi minat siswa dalam melanjutkan studi ke jenjang perguruan tinggi.

Untuk memperoleh model *decision tree* yang akan digunakan untuk mengklasifikasi minat siswa, proses pembuatan model dilakukan dengan menggunakan perangkat lunak *Rapid Miner*. Berdasarkan hasil

analisis yang tersebut, diketahui bahwa terdapat 18 aturan (*rules)* dimana 12 aturan menyatakan bahwa siswa lanjut studi dan 6 diantaranya tidak lanjut studi ke tingkat perguruan tinggi seperti pada Gambar 7 berikut.



Gambar 7 Hasil klasifikasi *Decision tree* Algorittma C5.0

Adapun hasil dari klasifikasi pada pohon keputusan tersebut adalah sebagai berikut.

1. Jika Rekomendasi = pernah, Beasiswa = tidak perlu maka hasil keputusan Lanjut.
2. Jika Rekomendasi = pernah, Beasiswa = perlu, Penghasilan Orang Tua = bawah, maka hasil keputusan Lanjut.
3. Jika Rekomendasi = pernah, Beasiswa = perlu, Penghasilan Orang Tua = menengah, Biaya kuliah = mahal, maka hasil keputusan Lanjut.
4. Jika Rekomendasi = pernah, Beasiswa = perlu, Penghasilan Orang tua = menengah, Biaya kuliah = terjangkau, Motivasi = guru sekolah maka diperoleh hasil keputusan Lanjut.
5. Jika Rekomendasi = pernah, Beasiswa = perlu, Penghasilan Orang Tua = miskin, Lingkungan = Mendukung maka diperoleh hasil keputusan Lanjut.
6. Jika Rekomendasi = pernah, Beasiswa = perlu, Penghasilan Orang Tua = miskin, Lingkungan = tidak mendukung, Status PT = Negeri, Lokasi PT = dekat, Biaya kuliah = mahal, maka diperoleh hasil keputusan Lanjut.
7. Jika Rekomendasi = pernah, Beasiswa = perlu, Penghasilan Orang Tua = miskin, Lingkungan = tidak mendukung, Status PT = Negeri, Lokasi PT = dekat, maka diperoleh keputusan Lanjut.
8. Jika Rekomendasi = pernah, Beasiswa = perlu, Penghasilan Orang Tua = miskin, Lingkungan = tidak mendukung, Status PT = swasta maka diperoleh keputusan Lanjut.
9. Jika Rekomendasi = pernah, Beasiswa = perlu, Penghasilan Orang Tua = rentan, maka diperoleh hasil keputusan Lanjut.
10. Jika Rekomendasi = tidak pernah, Beasiswa = perlu, Penghasilan Orang Tua = bawah, Biaya Kuliah = mahal, maka diperoleh keputusan Lanjut.
11. Jika Rekomendasi = tidak pernah, Beasiswa = perlu, Penghasilan Orang Tua = bawah, Biaya kuliah = terjangkau, Status PT = Swasta, maka diperoleh keputusan Lanjut.
12. Jika Rekomendasi = tidak pernah, Beasiswa = perlu, Penghasilan Orang Tua = rentan, Status PT = Swasta, maka diperoleh keputusan Lanjut.

Berdasarkan 12 aturan keputusan yang dihasilkan oleh algoritma C5.0, terlihat bahwa variabel Rekomendasi menempati posisi sebagai root node (akar utama). Hal ini menunjukkan bahwa faktor sosial memiliki pengaruh yang lebih dominan dibandingkan variabel ekonomi dalam membentuk niat siswa untuk melanjutkan studi. Secara lebih mendalam, pengaruh sosial ini terlihat pada interaksi antara variabel ekonomi dan dukungan lingkungan. Sebagai contoh, pada Aturan 5 dan 6, siswa dengan kondisi ekonomi "miskin" tetap menunjukkan keputusan "Lanjut" selama terdapat variabel Lingkungan yang mendukung. Sebaliknya, ketika

lingkungan sosial dianggap tidak mendukung, siswa cenderung mencari kompensasi keamanan pada variabel institusional seperti Status Perguruan Tinggi (Negeri) dan Lokasi yang dekat untuk meminimalisir risiko kegagalan. Temuan ini menegaskan bahwa dukungan dari orang tua, guru, dan teman sebaya berperan sebagai buffer yang mampu mereduksi hambatan finansial yang dihadapi siswa.

Temuan pola klasifikasi ini memiliki implikasi kebijakan yang strategis bagi institusi pendidikan. Bagi SMA Negeri 6 Luwu Timur, pihak sekolah disarankan untuk mengoptimalkan peran bimbingan konseling melalui pendekatan yang lebih dialogis dan inklusif dengan melibatkan orang tua. Mengingat variabel "Rekomendasi" sangat krusial, sekolah dapat menyelenggarakan program Parenting Career Day untuk menyamakan persepsi antara aspirasi siswa dengan dukungan keluarga, khususnya bagi kelompok ekonomi rentan. Sementara bagi perguruan tinggi, strategi sosialisasi harus lebih menekankan pada aksesibilitas informasi beasiswa dan keterjangkauan biaya guna menarik minat siswa yang memiliki motivasi tinggi namun terkendala persepsi biaya kuliah yang mahal.

Meskipun penelitian ini memberikan kontribusi penting dalam pemetaan minat siswa berbasis data, terdapat beberapa keterbatasan yang perlu dipertimbangkan. Pertama, penelitian ini terbatas pada lingkup satu sekolah, sehingga kemampuan generalisasi temuan terhadap populasi siswa SMA secara nasional atau di wilayah dengan karakteristik sosial-ekonomi yang berbeda mungkin terbatas. Validitas eksternal model ini perlu diuji lebih lanjut dengan melibatkan dataset yang lebih luas dan bervariasi dari berbagai institusi pendidikan. Meskipun demikian, model ini dapat menjadi pilot project atau kerangka kerja bagi sekolah lain dengan karakteristik serupa dalam melakukan profiling siswa. Kedua, terdapat keterbatasan terkait aspek data. Variabel yang digunakan dalam penelitian ini terbatas pada faktor sosial dan ekonomi. Faktor eksternal lain, misalnya pengaruh media sosial yang mungkin memengaruhi minat siswa belum sepenuhnya terakomodasi dalam model. Selain itu, asumsi bahwa data kuesioner mencerminkan kondisi objektif siswa sangat bergantung pada kejujuran responden saat pengisian instrumen. Penelitian selanjutnya disarankan untuk memperluas sampel serta menambahkan variabel psikografis yang lebih mendalam guna meningkatkan akurasi dan cakupan model klasifikasi.

4. KESIMPULAN

Penelitian ini menyimpulkan bahwa implementasi *data mining* menggunakan algoritma C5.0 berhasil membangun model prediktif yang akurat untuk mengidentifikasi minat siswa dalam melanjutkan pendidikan ke perguruan tinggi. Temuan menunjukkan bahwa variabel Beasiswa, Penghasilan Orang Tua, dan Rekomendasi merupakan faktor kunci yang paling signifikan dalam memengaruhi keputusan siswa. Model ini menghasilkan 18 aturan keputusan, yang terdiri dari 12 pola untuk keputusan "Lanjut" dan 6 pola untuk "Tidak Lanjut". Melalui berbagai skema pengujian, metode *60% Percentage Split* mencatatkan kinerja terbaik dengan akurasi mencapai 87,23%. Meskipun model ini memiliki akurasi yang tinggi, penelitian ini memiliki keterbatasan karena hanya berfokus pada satu sekolah (SMA Negeri 6 Luwu Timur), sehingga generalisasi temuan untuk populasi yang lebih luas memerlukan pengujian lebih lanjut pada dataset yang lebih bervariasi. Secara praktis, hasil ini dapat digunakan oleh pihak sekolah untuk memperkuat peran bimbingan konseling, khususnya dalam mengomunikasikan peluang beasiswa kepada keluarga dengan ekonomi rentan. Bagi perguruan tinggi, model ini memberikan dasar strategis untuk merancang program sosialisasi yang lebih tepat sasaran guna meningkatkan minat studi lanjut siswa berdasarkan profil sosial-ekonominya.

DAFTAR PUSTAKA

Ananta, A., Wulandari, N., Mustawinar, B. H., & Putri, F. G. (2024). Klasifikasi Pembelian Produk Rumah Tangga Melalui Metode Naïve Bayes. *Indonesian Journal of Material and Applied Physics*, *1*(1), 16–21.

Ayunda, Y. S., Hartama, D., Lubis, M. R., Gunawan, I., & Rafai, M. (2024). Analisis Pola Minat Siswa Lulusan SMU/SMK Untuk Melanjutkan Kuliah dengan Menggunakan Algoritma C4.5. *TIN: Terapan Informatika Nusantara*, *4*(9), 581–595. https://doi.org/10.47065/tin.v4i9.4880

BPS. (2024). *Data Jumlah Sekolah, Guru, dan Murid Sekolah Menengah Atas Tahun 2023/2024*. Badan Pusat Statistika. https://bps.go.id.

Han, J., Kamber, M., & Pei, J. (2012). Data mining: concepts and techniques. *Choice Reviews Online*, *49*(06), 49–3305. https://doi.org/10.5860/choice.49-3305

Makhrisa, R., & Pradikto, S. (2025). Analisis Peran Lingkungan Sosial Terhadap Minat Peserta Didik dalam Memilih Pendidikan Tinggi. *Jurnal Kajian Dan Penelitian Umum*, *3*(1), 78–98. https://doi.org/10.47861/jkpu-nalanda.v3i1.1503

Nugrahani, N., & Prapanca, A. (2025). Implementasi Algoritma C5.0 Pada Klasifikasi Status Gizi Balita di Kecamatan Ponorogo. *Journal of Informatics and Computer Science (JINACS)*, *6*(04), 1089–1098. https://doi.org/10.26740/jinacs.v6n04.p1089-1098

Pratama, T. Y., & Armansyah, A. (2024). Decision Tree C4.5 dengan Teknik Information Gain Untuk Klasifikasi Pemilihan Program Studi Tingkat Lanjut. *Journal of Information System Research (JOSH)*, *5*(4), 1042–1052. https://doi.org/10.47065/josh.v5i4.5643

Rizki, P. F. (2024). Pendekatan Supervised Learning Decision Tree C4.5 untuk Klasifikasi Calon Mahasiswa Baru. *Indonesian Journal of Computer Science*, *12*(6). https://doi.org/10.33022/ijcs.v12i6.3595