PISCES
Progressive Information, Security, Computer, and Embedded System

E-ISSN: 2986-7258
P-ISSN : 2986-724X

PISCES Volume 04 Nomor 01 Maret 2026

Comparison of Naïve Bayes and C4.5 Algorithms in Classifying the Eligibility of Smart Indonesia Card (KIP) College Scholarship Recipients

[1*]Dary Mochamad Rifqie, [2]Ahmad Faris Al Faruq, [3]Muh Alif, [4]Rendy Hidayat
[1]Jurusan Teknik Elektronika dan Teknologi Informasi, Universitas Negeri Makassar
[2]Jurusan Teknik Informatika dan Komputer, Universitas Negeri Makassar

Email: dary.mochamad.rifqie@unm.ac.id[1], ahmadfarisalfaruqq@gmail.com[2], muhaliff039@gmail.com[3], hrendy332@gmail.com[4]
*Corresponding author: dary.mochamad.rifqie@unm.ac.id

Received : 02 Januari 2026
Accepted : 28 Februari 2026
Published : 30 Maret 2026

ABSTRAK

Program Kartu Indonesia Pintar Kuliah (KIP Kuliah) bertujuan memperluas akses pendidikan tinggi bagi mahasiswa dari keluarga kurang mampu, tetapi proses seleksi penerima di sejumlah institusi masih dilakukan secara manual sehingga membutuhkan waktu relatif lama dan berpotensi menghasilkan keputusan yang kurang tepat sasaran. Penelitian ini bertujuan membandingkan kinerja algoritma Naïve Bayes dan Decision Tree C4.5 dalam mengklasifikasikan kelayakan penerima KIP Kuliah berdasarkan data kriteria mahasiswa. Penelitian menggunakan dataset sebanyak 76 sampel dengan 8 atribut independen yang dianalisis menggunakan aplikasi Orange melalui pembagian 80% data latih dan 20% data uji, dengan evaluasi berdasarkan Area Under Curve (AUC), Classification Accuracy (CA), F1-Score, Precision, Recall, dan Matthews Correlation Coefficient (MCC). Hasil pengujian menunjukkan bahwa Naïve Bayes memiliki performa lebih baik dengan AUC sebesar 0,857, CA sebesar 0,688, F1-Score sebesar 0,686, dan MCC sebesar 0,378, sedangkan C4.5 menghasilkan AUC sebesar 0,469, CA sebesar 0,500, dan MCC sebesar 0,000. Hasil tersebut menunjukkan bahwa Naïve Bayes lebih sesuai untuk mengklasifikasikan kelayakan penerima KIP Kuliah pada dataset penelitian ini dan berpotensi mendukung proses seleksi yang lebih efisien dan tepat sasaran.

Kata Kunci: C4.5, Data Mining, Klasifikasi, KIP Kuliah, Naïve Bayes

ABSTRACT

The Smart Indonesia Card for College Students (KIP Kuliah) program aims to expand access to higher education for students from economically disadvantaged families. However, the scholarship recipient selection process in several institutions still faces challenges because it is conducted manually, requiring considerable time and potentially resulting in inaccurate targeting of eligible recipients. This condition may cause students who meet the eligibility criteria to remain unidentified. This study aims to compare the performance of the Naïve Bayes and Decision Tree C4.5 algorithms in classifying the eligibility of KIP Kuliah scholarship recipients based on student criteria data. The study used a dataset consisting of 76 samples with 8 independent attributes. The analysis was conducted using the Orange application, with 80% of the data used for training and 20% for testing. The performance of both algorithms was evaluated using Area Under the Curve (AUC), Classification Accuracy (CA), F1-Score, Precision, Recall, and Matthews Correlation Coefficient (MCC). The results showed that Naïve Bayes achieved better classification performance than C4.5, with an AUC of 0.857, CA of 0.688, F1-Score of 0.686, and MCC of 0.378. In comparison, C4.5 achieved an AUC of 0.469, CA of 0.500, and MCC of 0.000. These findings indicate that Naïve Bayes is more suitable for classifying the eligibility of KIP Kuliah scholarship recipients within the dataset used in this study and has the potential to support a more efficient and accurately targeted selection process.

Keywords: C4.5, Classification, Data Mining, KIP Kuliah, Naïve Bayes

E-ISSN: 2986-7258
P-ISSN : 2986-724X

PISCES Volume 04 Nomor 01 Maret 2026

1. PENDAHULUAN

Program Kartu Indonesia Pintar Kuliah (KIP Kuliah) adalah inisiatif pemerintah yang telah berlangsung sejak 2020 hingga saat ini. KIP Kuliah didistribusikan oleh Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi melalui berbagai perguruan tinggi di seluruh Indonesia. KIP Kuliah dirancang untuk membantu mahasiswa dari keluarga kurang mampu agar dapat mengakses pendidikan tinggi [1],[2],[3]. Di Indonesia, kebutuhan akan program ini tinggi karena akses pendidikan tinggi kelompok miskin masih tertinggal, sementara biaya kuliah, keterbatasan pembiayaan, dan ketimpangan implementasi kebijakan masih menjadi hambatan utama [4],[5],[6].

Setiap perguruan tinggi menerima kuota yang berbeda-beda, disesuaikan dengan perkembangan masing-masing institusi [7]. Program ini sangat penting dalam mengatasi hambatan ekonomi yang sering kali menjadi penghalang bagi siswa berprestasi untuk melanjutkan pendidikan ke jenjang yang lebih tinggi. Melalui KIP Kuliah, mahasiswa yang berasal dari keluarga dengan ekonomi terbatas dapat memperoleh bantuan finansial yang cukup besar, yang mencakup biaya kuliah dan kebutuhan hidup sehari-hari, sehingga mereka dapat fokus pada studi mereka [8], [9].

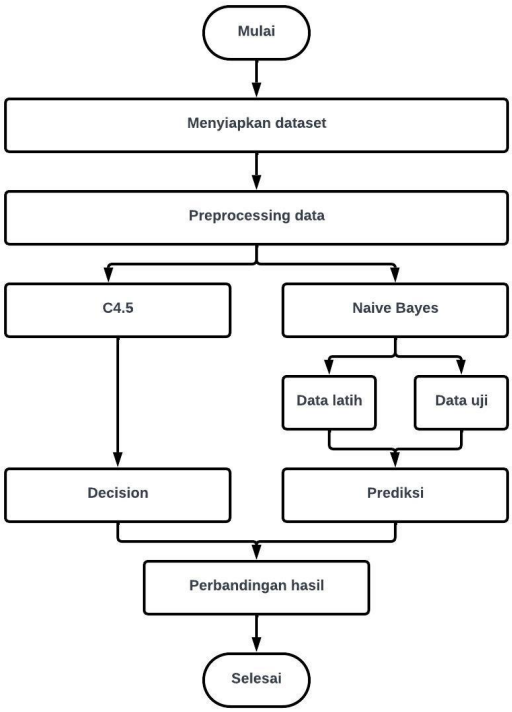
Meskipun memiliki tujuan mulia untuk membantu mahasiswa dari keluarga kurang mampu, sering kali menghadapi tantangan dalam hal penyaluran yang tepat sasaran. Salah satu masalah utama yang diidentifikasi adalah proses seleksi penerima beasiswa yang masih dilakukan secara manual dan tradisional, yang dapat menyebabkan ketidakakuratan dalam menentukan siapa yang benar-benar membutuhkan bantuan. Rema et al. mencatat bahwa proses seleksi di beberapa institusi, seperti Universitas Timor, membutuhkan waktu yang lama dan sering kali menghasilkan keputusan yang kurang tepat [10]. Ini menunjukkan bahwa ada potensi bagi mahasiswa yang tidak memenuhi syarat untuk menerima beasiswa KIP Kuliah, sementara mahasiswa yang benar-benar membutuhkan mungkin terlewatkan. Haryono juga menekankan bahwa meskipun banyak mahasiswa penerima KIP Kuliah menunjukkan prestasi akademik yang baik, tidak semua penerima beasiswa tersebut berasal dari latar belakang yang benar-benar membutuhkan [11].

Untuk mengatasi permasalahan penyaluran beasiswa KIP Kuliah yang belum tepat sasaran, implementasi teknologi data mining dapat menjadi solusi yang efektif. Berdasarkan penelitian yang dilakukan Sovia et al. (2020) menggabungkan metode K-Means dengan SAW untuk memprediksi penerima beasiswa berprestasi. Hasilnya menunjukkan bahwa kombinasi metode ini dapat meningkatkan akurasi prediksi [12]. Selain itu, penerapan algoritma K-NN untuk klasifikasi penerima beasiswa PPA dan BBM oleh Sumarlin (2015) membuktikan bahwa algoritma ini mampu memberikan akurasi tinggi dalam proses seleksi [13].

Meskipun penelitian sebelumnya telah membahas prediksi penerimaan beasiswa, tetapi sebagian besar studi cenderung berfokus pada algoritma K-Means dan K-NN tanpa mempertimbangkan algoritma C4.5 dan naïve bayes. Kajian terkini menunjukkan bahwa dalam memprediksi potensi mahasiswa penerima beasiswa, algoritma klasifikasi seperti C4.5 dan Naïve Bayes dapat digunakan untuk meningkatkan ketepatan seleksi. Algoritma C4.5, yang merupakan pengembangan dari algoritma ID3, digunakan untuk membangun pohon keputusan yang dapat membantu dalam klasifikasi data. Penelitian oleh Amalia menunjukkan bahwa penerapan algoritma C4.5 dalam konteks program pendidikan dapat membantu dalam pengambilan keputusan yang lebih objektif dan berbasis data [14]. Algoritma Naïve Bayes juga merupakan metode yang populer dalam klasifikasi. Metode ini menggunakan pendekatan probabilistik untuk menentukan kemungkinan suatu kelas berdasarkan atribut yang ada. Widaningsih mencatat bahwa Naïve Bayes memiliki akurasi yang tinggi dalam memprediksi kelulusan mahasiswa, yang menunjukkan bahwa metode ini dapat diadaptasi untuk memprediksi potensi mahasiswa dalam mendapatkan beasiswa KIP Kuliah [15]. Berdasarkan penjelasan diatas maka penelitian ini diharapkan memberikan kontribusi berharga bagi pengembangan sistem seleksi penerima beasiswa KIP Kuliah, sehingga dapat meningkatkan akurasi dan ketepatan sasaran dalam mendukung mahasiswa berprestasi dari latar belakang ekonomi yang kurang beruntung.

2. METODE PENELITIAN

Metodologi penelitian merupakan serangkaian langkah atau tahapan yang dilakukan dalam suatu penelitian, dimulai dari proses mengidentifikasi masalah hingga menyimpulkan hasil. Tahapan metodologi yang digunakan dalam penelitian ini ditampilkan pada gambar 1.



Gambar 1. Tahapan penelitian dengan metode C4.5 dan Naïve Bayes

2.1. Dataset

Dataset yang digunakan dalam penelitian ini diperoleh dari data penerima dan non-penerima beasiswa KIP Kuliah yang dikumpulkan melalui metode survei. Dataset ini mencakup data mahasiswa dari berbagai program studi, dengan total 9 atribut yang menjadi variabel independen, serta 2 label yang berfungsi sebagai variabel target.

2.2. Preprocessing

Preprocessing dalam penambangan data merupakan fase penting yang berdampak signifikan pada kualitas dan efektivitas proses penambangan. Proses ini melibatkan serangkaian langkah yang dirancang untuk membersihkan, mengubah, dan menyiapkan data mentah untuk analisis, guna memastikan bahwa data tersebut sesuai untuk operasi penambangan berikutnya. Pentingnya prapemrosesan ditegaskan oleh perannya dalam meningkatkan kualitas data, yang secara langsung memengaruhi keakuratan dan kejelasan hasil penambangan [16], berikut adalah langkah-langkah preprocessing yang dilakukan dalam penelitian ini.

a. Cleaning Data

Tahap pertama dalam preprocessing bertujuan untuk membersihkan data dari nilai yang tidak valid atau kosong.

b. Selection Atribut

Pemilihan atribut dilakukan untuk memilih atribut yang relevan dalam proses data mining. Pada metode Naïve Bayes, hanya atribut numerik yang digunakan, karena metode ini lebih optimal saat bekerja dengan data berjenis numerik.

2.3. Naïve Bayes

Naive Bayes adalah kumpulan algoritma probabilistik yang berlandaskan pada teorema Bayes dan sering digunakan untuk tugas klasifikasi dalam pembelajaran mesin. Prinsip utama dari Naive Bayes adalah asumsi independensi bersyarat antara fitur berdasarkan label kelas, yang berarti keberadaan satu fitur dalam suatu kelas tidak dipengaruhi oleh keberadaan fitur lainnya. Meskipun asumsi ini cukup kuat, pengklasifikasi Naive Bayes terbukti mampu memberikan performa yang sangat baik dalam berbagai aplikasi, bahkan sering kali sebanding dengan model yang lebih kompleks [17], [18].

$$P(H \setminus X) = (P(X \mid H)\, P(H))/(P(X)) \quad (1)$$

Keterangan:

X : Data dengan class yang belum diketahui

H : Hipotesis data X merupakan suatu class spesifik

P(H|X) : Probabilitas hipotesis H berdasarkan kondisi x (posteriori prob.)

P(H) : Probabilitas hipotesis H (prior prob.)

P(X|H) : Probabilitas X berdasarkan kondisi tersebut

P(X) : Probabilitas dari X

2.4. Algoritma C4.5

Algoritma C4.5 merupakan metode dalam pengolahan data yang digunakan untuk klasifikasi dan membuat keputusan, yang merupakan peningkatan dari algoritma ID3. Algoritma ini dirancang untuk menghasilkan pohon keputusan yang dapat digunakan dalam berbagai aplikasi, termasuk prediksi dan klasifikasi data. C4.5 memiliki kemampuan untuk menangani atribut yang bersifat kontinu maupun diskrit, sehingga menjadikannya fleksibel dalam berbagai konteks analisis data [19].

Salah satu keunggulan utama dari algoritma C4.5 adalah kemampuannya dalam mengurangi kompleksitas pohon keputusan melalui proses pemangkasan (pruning). Proses ini bertujuan untuk menghindari overfitting, yaitu kondisi di mana model terlalu kompleks dan tidak dapat menggeneralisasi data baru dengan baik. Penelitian menunjukkan bahwa penggunaan algoritma genetik untuk pemangkasan dapat meningkatkan akurasi model C4.5 dengan mengoptimalkan faktor kepercayaan [20]. Tahapan rumus yang digunakan dapat dilihat di bawah ini.

$$Entropy\ (S) = \sum_{i=0}^{n} - pi * log2\ pi \quad (2)$$

Keterangan :

S : Himpunan kasus

n : Jumlah Partisi S

pi : Jumlah kasus pada partisi ke – I

Untuk mendapatkan nilai gain, rumus berikut digunakan.

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

2.5. Evaluasi Hasil

Penelitian ini melibatkan perbandingan dua algoritma klasifikasi, yaitu C4.5 dan Naïve Bayes, yang digunakan untuk memprediksi potensi mahasiswa dalam menerima beasiswa KIP Kuliah. Perbandingan ini dilakukan untuk menentukan algoritma yang lebih efektif dalam menghasilkan prediksi dengan akurasi yang tinggi dan kesalahan yang rendah. Perbandingan dilakukan dengan mengukur beberapa metrik evaluasi seperti Akurasi, Presisi, Recall, F1-Score, Train, AUC, CA, dan MCC.

3. HASIL DAN PEMBAHASAN

3.1. Dataset

Variabel data penelitian yang digunakan dalam penelitian ini disajikan pada Tabel 1.

Tabel 1. Atribut dan value dataset Beasiswa KIP Kuliah

Atribut	Value
Inisial	Nama
Umur	18, 19, 20, 21, 22, 23
Pekerjaan orang tua	PNS, TNI/Polisi, Petani, Wiraswasta, Buruh, Karyawan Swasta, IRT, Nelayan, Meninggal
Pendapatan orang tua	Rp1.000,000, Rp1.000,000 – Rp3.000,000, Rp.3.000,000 – Rp 5.000,000, Rp5.000,000
Jumlah tanggungan	1-2 orang, 3-4 orang, Lebih dari 4 orang
Nilai akhir ijazah SMA	75, 75 sampai 85, 85
Pernah menerima beasiswa	Ya, Tidak
Memiliki Prestasi	Ya, Tidak
Menerima KIP kuliah	Ya, Tidak

3.2. Dataset

Setelah melalui tahap praproses, terdapat total 76 data yang tersisa. Dari jumlah tersebut, 38 data termasuk dalam class penerima Beasiswa KIP, sementara 38 lainnya termasuk dalam class non-penerima. Dari 9 atribut yang tersedia, hanya 8 atribut yang digunakan dalam analisis, seperti yang ditunjukkan pada tabel 2.

Tabel 2. Sampel dataset Beasiswa KIP Kuliah

Umur	Pekerjaan orang tua	Pendapatan orang tua	Jumlah tanggungan	Nilai rata-rata SMA	Pernah menerima beasiswa	Memiliki Prestasi	Menerima KIP kuliah
21	Petani	< Rp1.000.000	3-4 orang	> 85	Tidak	Tidak	Ya
21	PNS	> Rp5.000.000	3-4 orang	> 85	Tidak	Ya	Tidak
21	Petani	< Rp1.000.000	3-4 orang	> 85	Tidak	Tidak	Ya
21	PNS	> Rp5.000.000	3-4 orang	> 85	Tidak	Ya	Tidak
20	Petani	< Rp1.000.000	1-2 orang	> 85	Tidak	Ya	Ya

3.3. Implementasi Algoritma Klasifikasi

1. Naïve Bayes

Langkah pertama dalam proses analisis adalah membagi dataset menjadi dua bagian, yaitu data latih yang mencakup 80% dari total data, dan data uji yang mencakup 20%. Pembagian ini bertujuan untuk memastikan bahwa model dapat dilatih menggunakan sebagian besar data yang tersedia, sementara sisanya digunakan untuk menguji performa model dalam memprediksi data yang belum pernah dilihat, sampel data latih dan data uji sebagai berikut.

Tabel 3. Sampel data latih

Umur	Pekerjaan orang tua	Pendapatan orang tua	Jumlah tanggungan	Nilai rata-rata ijazah SMA	Pernah menerima beasiswa	Memiliki Prestasi	Menerima KIP kuliah
21	Petani	< Rp1.000.000	1-2 orang	75 - 85	Tidak	Tidak	Ya
21	Wiraswasta	< Rp1.000.000	> 4 orang	> 85	Ya	Ya	Ya
21	PNS	Rp3.000.001 - Rp5.000.000	3-4 orang	> 85	Tidak	Tidak	Tidak
23	Pengusaha	< Rp1.000.000	3-4 orang	> 85	Tidak	Tidak	Tidak
22	IRT	< Rp1.000.000	> 4 orang	> 85	Ya	Ya	Ya

E-ISSN: 2986-7258
P-ISSN : 2986-724X

PISCES Volume 04 Nomor 01 Maret 2026

Tabel 4. Sampel data uji

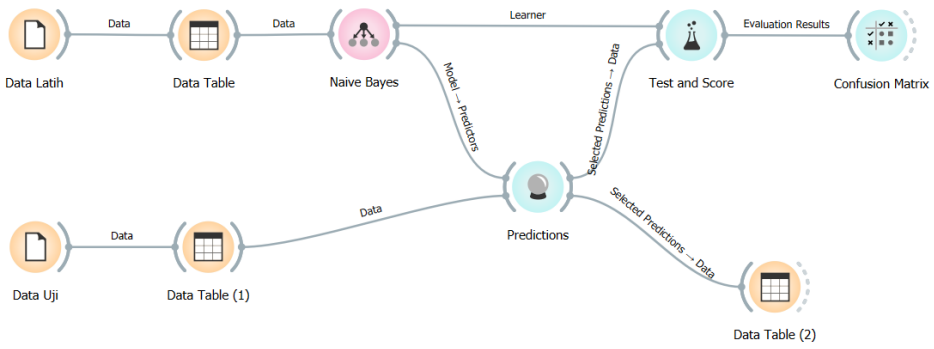
Umur	Pekerjaan orang tua	Pendapatan orang tua	Jumlah tanggungan	Nilai rata-rata ijazah SMA	Pernah menerima beasiswa	Memiliki Prestasi	Menerima KIP kuliah
19	Wiraswasta	< Rp1.000.000	3-4 orang	75 – 85	Ya	Tidak	Ya
21	PNS	Rp3.000.001 – Rp5.000.000	1-2 orang	> 85	Tidak	Ya	Tidak
20	PNS	Rp3.000.001 – Rp5.000.000	3-4 orang	75 – 85	Tidak	Tidak	Tidak
21	PNS	Rp3.000.001 – Rp5.000.000	> 4 orang	75 – 85	Tidak	Ya	Tidak
23	Wiraswasta	< Rp1.000.000	3-4 orang	75 – 85	Tidak	Tidak	Ya

Sebelum dilakukan analisis, data terlebih dahulu diubah ke dalam format numerik untuk memastikan kompatibilitas dengan algoritma yang digunakan, sehingga proses pengolahan data dapat berjalan dengan lebih akurat dan efisien, Seperti pada tabel 5.

Tabel 5. Atribut dataset dalam bentuk numerik

Atribut	Value	Numerik
Umur	18, 19, 20, 21, 22, 23	18, 19, 20, 21, 22, 23
Pekerjaan orang tua	PNS, TNI/Polisi, Petani, Wiraswasta, Buruh, Karyawan Swasta, IRT, Nelayan, Meninggal	4, 7, 6, 0,
Pendapatan orang tua	< Rp1.000,000, Rp1.000,000 – Rp3.000,000, Rp.3.000,000 – Rp 5.000,000, Rp5.000,000	0, 2, 3, 1
Jumlah tanggungan	1-2 orang, 3-4 orang, Lebih dari 4 orang	0, 1, 2
Nilai akhir ijazah SMA	75, 75 sampai 85, 85	1, 0, 2
Pernah menerima beasiswa	Ya, Tidak	1, 0
Memiliki Prestasi	Ya, Tidak	1, 0
Menerima KIP kuliah	Ya, Tidak	1, 0

Selanjutnya, data akan dianalisis menggunakan perangkat lunak Orange. Proses analisis ini akan mengikuti kerangka kerja yang telah dirancang sebelumnya, seperti yang ditunjukkan pada gambar 2.



Gambar 2. Framework algoritma naïve bayes

Dengan menggunakan metode Naïve Bayes, model menghasilkan akurasi prediksi sebesar 50%, sebagaimana ditunjukkan pada tabel 6.

Tabel 6. Confusion matrix naïve bayes

		Prediksi		
		Tidak	Ya	∑
Aktual	Tidak	6	2	8
	Ya	6	2	8
	∑	12	4	16

Berdasarakan sampel data hasil prediksi algoritme Naïve Bayes yang disajikan pada tabel 7 menujukkan bahwa baris pertama, algoritma berhasil memprediksi siswa sebagai penerima KIP Kuliah (True Positive). Hal ini didukung oleh kondisi ekonomi siswa yang sulit, dengan pendapatan orang tua kurang dari Rp1.000.000 dan jumlah tanggungan keluarga besar (3-4 orang), meskipun siswa tidak memiliki prestasi.

Pada baris kedua, algoritma memprediksi siswa tidak layak menerima KIP Kuliah (False Negatif), meskipun kenyataannya siswa diterima. Kesalahan ini kemungkinan terjadi karena pendapatan orang tua dalam kategori menengah rendah (Rp1.000.000 - Rp3.000.000) dan jumlah tanggungan yang kecil (1-2 orang), sehingga sinyal ekonomi menjadi kurang mendukung, meskipun siswa memiliki nilai ijazah tinggi dan prestasi.Pada baris ketiga, algoritma kembali berhasil memprediksi siswa sebagai bukan penerima KIP Kuliah (Benar Negatif). Pendapatan orang tua yang berada pada kisaran Rp3.000.001 - Rp5.000.000 dan jumlah tanggungan kecil menjadi faktor utama yang menurunkan kelayakan siswa, meskipun siswa memiliki nilai akademik tinggi dan prestasi.

Sebaliknya, pada baris keempat, algoritma salah memprediksi siswa sebagai penerima KIP Kuliah (False Positive). Kesalahan ini terjadi karena algoritma memberikan bobot besar pada pendapatan rendah (< Rp1.000.000) dan jumlah tanggungan besar (3-4 orang), tetapi tidak memperhatikan faktor lain seperti riwayat beasiswa, yang ternyata tidak dimiliki siswa ini.

Pengujian ini menggunakan dataset yang berukuran relatif kecil (76 instance). Algoritma Naïve Bayes menghitung posterior kelas dari prior kelas dan likelihood fitur, tetapi komputasinya disederhanakan karena evidence tidak perlu dihitung eksplisit untuk tiap kelas dan likelihood gabungan dipecah menjadi hasil kali marginal kondisional [21],[22]. Penyederhanaan ini membuat pembelajaran menjadi cepat, tractable, dan hemat sampel karena setiap fitur dapat diestimasi sendiri-sendiri, bukan sebagai sistem dependensi multivariat penuh [21],[22],[23]. Literatur tinjauan juga melaporkan bahwa Bayesian classifier sering mengungguli alternatif pada dataset kecil berukuran ratusan hingga ribuan contoh, dan bahwa independensi atribut tidak harus benar-benar terpenuhi agar Naïve Bayes tetap optimal dalam kondisi zero-one loss tertentu [22]. Ini menjelaskan mengapa Naïve Bayes dapat tetap kompetitif walau asumsi independensi jelas disederhanakan.

Tabel 7. Sampel data hasil prediksi algoritma naïve bayes

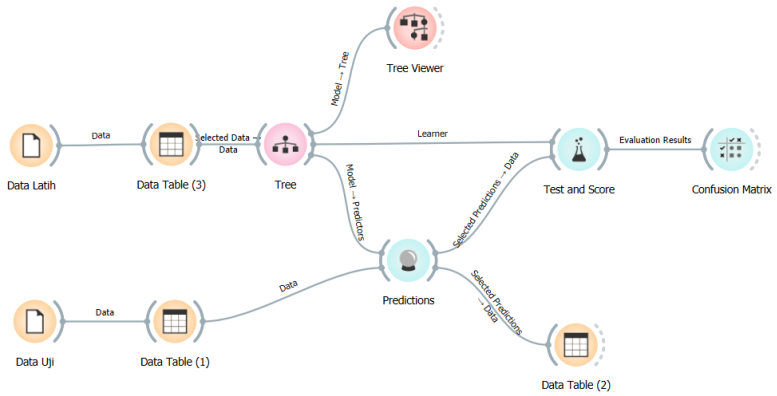
Label Actual	Label Hasil	Umur	Pekerjaan orang tua	Pendapatan orang tua	Jumlah tanggungan	Nilai rata-rata ijazah SMA	Pernah menerima beasiswa	Memiliki Prestasi
Ya	Ya	19	Wiraswasta	< Rp1.000.000	3-4 orang	75 - 85	Ya	Tidak
Ya	Tidak	21	PNS	Rp1.000.000 - Rp3.000.000	1-2 orang	> 85	Ya	Ya
Tidak	Tidak	21	PNS	Rp3.000.001 - Rp5.000.000	1-2 orang	> 85	Tidak	Ya
Tidak	Ya	18	Petani	< Rp1.000.000	3-4 orang	75 - 85	Tidak	Ya

E-ISSN: 2986-7258
P-ISSN : 2986-724X

PISCES Volume 04 Nomor 01 Maret 2026

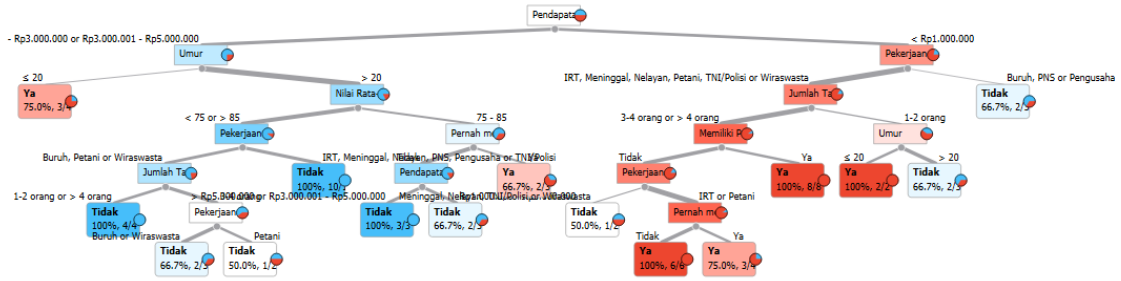
2. Algoritma C4.5

Selanjutnya, data dianalisis menggunakan algoritma C4.5. Proses analisis ini mengikuti langkah-langkah yang telah dirancang sebelumnya, seperti pada gambar 3.



Gambar 3. Framework algoritma C4.5

Hasil analisis menggunakan algoritma C4.5 dapat dilihat pada gambar 4, yang dimana algotima ini menghasilkan pohon keputusan dengan atribut utama Pendapatan sebagai akar, yang menunjukkan bahwa pendapatan memiliki pengaruh terbesar dalam menentukan klasifikasi. Dari atribut ini, data dibagi menjadi beberapa kelompok berdasarkan rentang pendapatan, yang kemudian dianalisis lebih lanjut menggunakan atribut tambahan seperti Umur, Pekerjaan, Nilai Rata-rata, dan Jumlah Tanggungan. Setiap cabang pohon mencerminkan aturan keputusan yang spesifik, sedangkan daun pohon merepresentasikan hasil klasifikasi, yaitu apakah seseorang termasuk dalam kategori "Ya" (menerima beasiswa KIP) atau "Tidak" (tidak menerima), disertai dengan persentase data yang mendukung keputusan tersebut.



Gambar 4. Decision Tree

Dengan menggunakan metode Naïve Bayes, model menghasilkan akurasi prediksi sebesar 43%, sebagaimana ditunjukkan pada tabel 8.

Tabel 8. Confusion matrix algoritma C4.5

		Predicted		
		Tidak	Ya	∑
Actual	Tidak	5	3	8
	Ya	6	2	8
	∑	11	5	16

Pada tabel 9 menujukkan bahwa baris pertama, algoritma berhasil memprediksi dengan benar bahwa siswa layak menerima KIP Kuliah (Benar Positif). Hal ini didukung oleh pendapatan orang tua yang rendah (Rp1.000.000 - Rp3.000.000), jumlah tanggungan besar (>4 orang), nilai ijazah tinggi (>85), dan riwayat penerimaan beasiswa. Sebaliknya, pada baris kedua, algoritma salah memprediksi bahwa siswa tidak layak menerima KIP Kuliah (False Negatif). Meskipun pendapatan orang tua sangat rendah (< Rp1.000.000) dan ada riwayat beasiswa, algoritma tampaknya terpengaruh oleh nilai akademik yang hanya cukup (75-85) serta ketiadaan prestasi.

Pada baris ketiga, algoritma kembali memprediksi dengan benar bahwa siswa tidak layak menerima KIP Kuliah (Benar Negatif). Pendapatan orang tua yang cukup tinggi (Rp3.000.001 - Rp5.000.000) menjadi faktor dominan, meskipun siswa memiliki prestasi dan jumlah tanggungan yang besar (>4 orang). Namun, pada baris keempat, algoritma salah memprediksi siswa sebagai penerima KIP Kuliah (False Positif). Kesalahan ini mungkin disebabkan oleh pendapatan orang tua yang sangat rendah (< Rp1.000.000) dan jumlah tanggungan besar (3-4 orang), sementara atribut lain, seperti ketiadaan riwayat beasiswa, kurang diperhitungkan oleh algoritma.

Penyebab turunnya performa C4.5 dipengaruhi beberapa hal, salah satunya adalah kasus overfitting. Literatur terdagulu secara konsisten menyebut ukuran sampel kecil, noise, dan over-learning sebagai penyebab utama overfitting, yakni kondisi ketika akurasi latih tinggi tetapi performa pada data tak terlihat menurun [24]. Teori dan studi empiris juga menunjukkan bahwa ketika kedalaman pohon atau dimensi kovariat meningkat sementara sampel terbatas, minimisasi risk empiris pada pohon dapat overfit dan tidak lagi menghasilkan risiko prediksi minimum [25], [26]. Overfitting pada decision tree didefinisikan justru sebagai kegagalan mentransfer performa dari data latih ke data tak terlihat, dan beberapa studi menekankan bahwa pohon kompleks cenderung salah klasifikasi pada pengujian karena generalisasi yang buruk [24], [27].

Tabel 9. Sampel data hasil prediksi algoritma C4.5

Label Actual	Label Hasil	Umur	Pekerjaan orang tua	Pendapatan orang tua	Jumlah tanggungan	Nilai rata-rata ijazah SMA	Pernah menerima beasiswa	Memiliki Prestasi
Ya	Ya	20	Wiraswasta	Rp1.000.000 - Rp3.000.000	> 4 orang	> 85	Ya	Ya
Ya	Tidak	19	Wiraswasta	< Rp1.000.000	3-4 orang	75 - 85	Ya	Tidak
Tidak	Tidak	21	PNS	Rp3.000.001 - Rp5.000.000	> 4 orang	75 - 85	Tidak	Ya
Tidak	Ya	18	Petani	< Rp1.000.000	3-4 orang	75 - 85	Tidak	Ya

3. Perbandingan Hasil

Pada tabel 10 menujukkan perbandingan kinerja dua algoritma, yaitu Naïve Bayes dan C4.5, berdasarkan berbagai metrik evaluasi.

Tabel 10. Perbandingan kinerja Naïve Bayes dan C4.5

Model	Train	AUC	CA	F1	Prec	Recall	MCC
Naïve Bayes	0.005	0.857	0.688	0.686	0.690	0.688	0.378
C4.5	0.004	0.469	0.500	0.467	0.500	0.500	0.000

Evaluasi kinerja model klasifikasi dilakukan dengan membandingkan metrik Area Under Curve (AUC), Classification Accuracy (CA), F1-Score, Precision, Recall, dan Matthews Correlation Coefficient (MCC) antara algoritma Naïve Bayes dan Decision Tree C4.5. Berdasarkan hasil pengujian, algoritma Naïve Bayes mencatatkan nilai AUC sebesar 0,857, yang tergolong dalam kategori pemisahan kelas sangat baik karena berada di atas batas ideal 0,8. Sebaliknya, algoritma C4.5 hanya mencapai nilai AUC 0,469, yang

menunjukkan bahwa kemampuan pemisahan kelasnya tidak lebih baik daripada tebakan acak. Pada metrik akurasi klasifikasi (CA), Naïve Bayes mendokumentasikan tingkat ketepatan sebesar 0,688 atau mendekati ambang batas akurat 0,7. Sementara itu, C4.5 hanya memperoleh nilai CA 0,500, yang menandakan bahwa separuh dari seluruh prediksi yang dihasilkan mengalami kesalahan.

Keseimbangan antara presisi dan sensitivitas model juga tecermin melalui nilai Precision, Recall, dan F1-Score. Naïve Bayes menunjukkan performa yang stabil dengan Precision sebesar 0,690 dan Recall sebesar 0,688. Keseimbangan ini menghasilkan nilai F1-Score sebesar 0,686, yang melampaui kriteria baik (di atas 0,6). Nilai tersebut membuktikan kemampuan Naïve Bayes dalam meminimalkan kesalahan False Positive (mencegah pemberian beasiswa kepada pihak yang tidak berhak) sekaligus False Negative (mencegah mahasiswa tidak mampu terlewat dari bantuan). Di sisi lain, C4.5 hanya mencatatkan nilai 0,500 untuk Precision dan Recall, serta nilai F1-Score yang rendah sebesar 0,467.

Penilaian kualitas klasifikasi secara menyeluruh ditunjukkan oleh metrik Matthews Correlation Coefficient (MCC). Naïve Bayes memperoleh nilai MCC sebesar 0,378, yang berada di atas ambang batas 0,3 dan mengonfirmasi adanya korelasi positif yang kuat antara hasil prediksi dengan kondisi aktual. Sebaliknya, algoritma C4.5 mencatatkan nilai MCC sebesar 0,000, yang menegaskan tidak adanya korelasi linier antara prediksi model dengan data sebenarnya. Kontras performa yang signifikan ini dipengaruhi oleh ukuran dataset yang terbatas (76 sampel). Naïve Bayes berbasis pendekatan probabilistik yang efisien pada dataset berskala kecil, sedangkan C4.5 cenderung mengalami overfitting saat membentuk pohon keputusan pada data yang terbatas, sehingga kehilangan kemampuan generalisasi ketika diuji pada data uji.

Meskipun algoritma Naïve Bayes menunjukkan performa yang unggul, penelitian ini masih memiliki beberapa keterbatasan yang perlu diperhatikan. Pertama, keterbatasan ukuran dataset yang hanya mencakup 76 sampel data tergolong relatif kecil untuk penelitian data mining, sehingga kurang optimal saat diterapkan pada algoritma berbasis pohon seperti C4.5 yang membutuhkan volume data lebih besar. Kedua, keterbatasan cakupan atribut yang digunakan belum memuat variabel non-ekonomi tambahan, seperti kondisi fisik tempat tinggal, daya listrik, atau jumlah anggota keluarga yang sedang menempuh pendidikan dasar dan menengah. Ketiga, pengujian ini terbatas pada dua algoritma klasifikasi (Naïve Bayes dan C4.5) serta belum menerapkan teknik hyperparameter tuning, cross-validation, atau algoritma ensemble (seperti Random Forest atau XGBoost). Oleh karena itu, penelitian selanjutnya disarankan untuk menambah jumlah sampel data, memperluas variasi atribut kriteria, serta menerapkan metode optimasi guna meningkatkan akurasi dan stabilitas generalisasi model secara keseluruhan.

4. KESIMPULAN DAN SARAN

Berdasarkan hasil evaluasi dan komparasi kinerja model, dapat disimpulkan bahwa algoritma Naïve Bayes memiliki performa klasifikasi yang jauh lebih unggul dibandingkan algoritma C4.5 dalam memprediksi kelayakan penerima beasiswa KIP Kuliah. Algoritma Naïve Bayes mencatatkan nilai Area Under Curve (AUC) sebesar 0,857, yang menunjukkan kemampuan sangat baik dalam membedakan kelas penerima dan non-penerima beasiswa. Selain itu, Naïve Bayes unggul pada seluruh metrik evaluasi lainnya, yaitu Classification Accuracy (CA) sebesar 0,688, F1-Score sebesar 0,686, Precision sebesar 0,690, dan Recall sebesar 0,688, yang mencerminkan keseimbangan serta akurasi prediksi yang tinggi. Sebaliknya, algoritma C4.5 menunjukkan kinerja yang kurang optimal dengan nilai AUC sebesar 0,469, CA sebesar 0,500, serta Matthews Correlation Coefficient (MCC) sebesar 0,000 yang mengindikasikan tidak adanya korelasi antara hasil prediksi dengan kondisi aktual. Dengan demikian, algoritma Naïve Bayes lebih tepat, efisien, dan dapat diandalkan untuk diimplementasikan dalam sistem pendukung keputusan seleksi penerima beasiswa KIP Kuliah.

REFERENSI

[1] Novitasari, S., Siswanto, D., & Ambarwati, A. (2025). Improving Educational Equity through the Smart Indonesia Card for College: Policy Implementation and Challenges. Jurnal Public Policy. https://doi.org/10.35308/jpp.v11i1.10413

[2] Purwanto, R., Mahardika, F., & Faiz, M. N. (2025). A Comparative Analysis of KIP-K Acceptance Prediction Based on School Type Using XGBoost, Random Forest, and SVM-RBF: Evaluation Through Accuracy and Data Visualization. *Journal of Innovation Information Technology and Application (JINITA)*. https://doi.org/10.35970/10.35970/jinita.v7i2.2948

[3] Hadi, Nofiar, et al. "Decision Support System Model for College Kartu Indonesia Pintar (KIP) Scholarship Recipients Using the C4.5 Decision Tree Method and Simple Additive Weighting (SAW)." JURNAL SISFOTEK GLOBAL, 2024, https://doi.org/10.38101/sisfotek.v14i2.14395.

[4] Fadhil, Ihsan, and Amra Sabic-El-Rayess. "Providing Equity of Access to Higher Education in Indonesia: A Policy Evaluation." vol. 3, 2020, pp. 57-75, https://doi.org/10.23917/ijolae.v3i1.10376.

[5] Haruna, Dennis, et al. "ANALYSIS OF EDUCATION FINANCING STRATEGIES TO IMPROVE ACCESSIBILITY IN HIGHER EDUCATION." ICMIE Proceedings, 2025, https://doi.org/10.30983/icmie.v1i1.47.

[6] Balapradana, Muhamad Aryasandha, and Jejen Hendar. "Kebijakan Skema Student Loan sebagai Upaya Peningkatan Aksesibilitas Pendidikan Tinggi Mahasiswa Kurang Mampu." Bandung Conference Series: Law Studies, 2025, https://doi.org/10.29313/bcsls.v5i2.18509.

[7] I. Arfyanti, M. Fahmi, and P. Adytia, "Penerapan Algoritma Decision Tree Untuk Penentuan Pola Penerima Beasiswa KIP Kuliah," bits, vol. 4, no. 3, Dec. 2022, doi: 10.47065/bits.v4i3.2275

[8] A. Amin, R. N. Sasongko, and A. Yuneti, "Kebijakan Kartu Indonesia Pintar Untuk Memerdekakan Mahasiswa Kurang Mampu," Journal of Administration and Educational Management (Alignment), vol. 5, no. 1, pp. 98–107, 2022, doi: 10.31539/alignment.v5i1.3803

[9] M. Sompa, "Clustering Tingkat Ekonomi Mahasiswa Calon Penerima Kartu Indonesia Pintar (KIP) Kuliah Metode K-Means," Jurnal Ilmiah Ilmu Komputer Banthayo Lo Komputer, vol. 1, no. 2, pp. 65–71, 2022, doi: 10.37195/balok.v1i2.175.

[10] Y. O. L. Rema, Y. P. K. Kelen, and S. S. Manek, "Penepan Metode Simple Additive Weighting Untuk Menentukan Mahasiswa Penerima Beasiswa Bidik Misi Di Universitas Timor," Saintekbu, vol. 14, no. 01, pp. 45–50, 2022, doi: 10.32764/saintekbu.v14i01.1921.

[11] E. Haryono, "Analisis Pengaruh Beasiswa Kip Terhadap Prestasi Mahasiswa Iai Al Muhammad Cepu Dengan Pendekatan Variabel Dummy," Riemann Research of Mathematics and Mathematics Education, vol. 6, no. 2, pp. 35–43, 2024, doi: 10.38114/y5tj1y42.

[12] R. Sovia, E. P. W. Mandala, and S. Mardhiah, "Algoritma K-Means Dalam Pemilihan Siswa Berprestasi Dan Metode SAW Untuk Prediksi Penerima Beasiswa Berprestasi," Jurnal Edukasi Dan Penelitian Informatika (Jepin), vol. 6, no. 2, p. 181, 2020, doi: 10.26418/jp.v6i2.37759.

[13] S. Sumarlin, "Implementasi Algoritma K-Nearest Neighbor Sebagai Pendukung Keputusan Klasifikasi Penerima Beasiswa PPA Dan BBM," Jurnal Sistem Informasi Bisnis, vol. 5, no. 1, 2015, doi: 10.21456/vol5iss1pp52-62.

[14] A. Amalia, "Implementasi Algoritma C4.5 Dan Naïve Bayes Dalam Pengambilan Keputusan Untuk Program Indonesia Pintar (Pip) Di Sekolah Dasar Negeri 04 Majalangu," Jati (Jurnal Mahasiswa Teknik Informatika), vol. 8, no. 2, pp. 1889–1896, 2024, doi: 10.36040/jati.v8i2.8311.

[15] M. Vasumathy, "Exploring Data Mining Applications and Techniques: A Comprehensive Research Survey," International Journal of Membrane Science and Technology, vol. 10, no. 3, pp. 2909–2919, 2023, doi: 10.15379/ijmst.v10i3.2736.

[16] Dr. Buthynna Fahran, Dr. Mohammed Najm, and Mustafa Abdulsamea Abdulhamed, “Attack Detection Based on Data Mining Techniques,” International Journal of Science and Research (Ijsr), vol. 6, no. 12, pp. 1547–1551, 2017, doi: 10.21275/art20179053.

[17] V. Robles, P. Larrañaga, J. M. Peña, E. Menasalvas, and M. S. Pérez, “Interval Estimation Naïve Bayes,” pp. 143–154, 2003, doi: 10.1007/978-3-540-45231-7_14.

[18] K. M. Al-Aidaroos, A. A. Bakar, and Z. Othman, “Naïve Bayes Variants in Classification Learning,” 2010, doi: 10.1109/infrkm.2010.5466902.

[19] E. P. K. Orpa, E. F. Ripanti, and T. Tursina, “Model Prediksi Awal Masa Studi Mahasiswa Menggunakan Algoritma Decision Tree C4.5,” Jurnal Sistem Dan Teknologi Informasi (Justin), vol. 7, no. 4, p. 272, 2019, doi: 10.26418/justin.v7i4.33163.

[20] M. M. Mijwil and R. A. Abttan, “Utilizing the Genetic Algorithm to Pruning the C4.5 Decision Tree Algorithm,” Asian Journal of Applied Sciences, vol. 9, no. 1, 2021, doi: 10.24203/ajas.v9i1.6503.

[21] Blanquero, R., et al. "Variable selection for Naïve Bayes classification." ArXiv, vol. abs/2401.18039, 2021, https://doi.org/10.1016/j.cor.2021.105456.

[22] Wickramasinghe, I., and Harsha Kalutarage. "Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation." Soft Computing, vol. 25, 2020, pp. 2277 - 2293, https://doi.org/10.1007/s00500-020-05297-6.

[23] Chen, Hong, et al. "Improved naive Bayes classification algorithm for traffic risk management." EURASIP Journal on Advances in Signal Processing, vol. 2021, 2021, https://doi.org/10.1186/s13634-021-00742-6.

[24] Zhang, Likun, and Wei Ma. "Covariance-Driven Regression Trees: Reducing Overfitting in CART." ArXiv, vol. abs/2601.07281, 2026, https://doi.org/10.48550/arxiv.2601.07281.

[25] Zhang, Likun, and Wei Ma. "Covariance-Driven Regression Trees: Reducing Overfitting in CART." ArXiv, vol. abs/2601.07281, 2026, https://doi.org/10.48550/arxiv.2601.07281.

[26] Tan, Yan Shuo, et al. "A cautionary tale on fitting decision trees to data from additive models: generalization lower bounds." 2021, pp. 9663-9685, https://doi.org/10.48550/arxiv.2110.09626.

[27] Lazebnik, T., and S. Bunimovich-Mendrazitsky. "Decision tree post-pruning without loss of accuracy using the SAT-PP algorithm with an empirical evaluation on clinical data." Data Knowl. Eng., vol. 145, 2023, pp. 102173, https://doi.org/10.1016/j.datak.2023.102173.