Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2024 Page: 2700−2708
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6928

Prediksi Rekomendasi Pemilihan Kejuruan pada Sekolah Menengah Kejuruan Menggunakan Perbandingan Metode *Decision Tree C4.5* dan *Naïve Bayes*

Ratih Windari, Handoyo Widi Nugroho*

[1]Fakultas Ilmu komputer, Magister Teknik Informatika, IBI Darmajaya, Kota Bandar Lampung, Indonesia
Email:[1]ratihwindari87@gmail.com, [2,*]handoyo.wn.darmajaya.ac.id
Email Penulis Korespondensi: handoyo.wn.darmajaya.ac.id
Submitted: **05/02/2025**; Accepted: **25/03/2025**; Published: **26/03/2025**

Abstrak−SMK Negeri 4 Bandar Lampung menghadapi tantangan dalam membantu siswa memilih jurusan yang sesuai dengan potensi, minat, dan kemampuan mereka. Keputusan pemilihan jurusan sering kali dipengaruhi oleh faktor-faktor subjektif yang kurang transparan dan belum tentu akurat, sehingga diperlukan sistem yang dapat memberikan rekomendasi lebih akurat dan objektif, Penelitian ini mengembangkan sistem prediksi pemilihan jurusan di SMK Negeri 4 Bandar Lampung dengan menggunakan dua metode, yaitu algoritma Decision Tree C4.5 dan Naive Bayes. Sistem ini menggunakan tujuh atribut utama sebagai variabel prediksi, yang meliputi nilai matematika, bahasa Inggris, IPA, bahasa Indonesia, prestasi akademik, partisipasi dalam kegiatan ekstrakurikuler, serta kondisi buta warna. penelitian menunjukkan bahwa algoritma C4.5 menghasilkan akurasi sebesar 84.46%, sedangkan dibandingkan dengan Naive Bayes yang mencapai akurasi 92.23% dengan demikian algoritma naive bayes lebih baik . Meskipun demikian, kedua metode masih memiliki kekurangan yang dapat diperbaiki melalui optimasi parameter dan pengolahan data yang lebih mendalam. Penerapan sistem berbasis data ini diharapkan dapat meningkatkan efisiensi dalam memberikan rekomendasi jurusan yang lebih relevan di SMK Negeri 4 Bandar Lampung dan memberikan inspirasi bagi sekolah lain untuk mengadopsi pendekatan serupa guna meningkatkan mutu pendidikan.

Kata Kunci: Naive Bayes; C4.5; Confusion Matrix

Abstract−SMK Negeri 4 Bandar Lampung faces challenges in assisting students in selecting a major that aligns with their potential, interests, and abilities. The decision-making process for choosing a major is often influenced by subjective factors that lack transparency and may not be entirely accurate. Therefore, a system is needed to provide more accurate and objective recommendations. This study develops a predictive system for major selection at SMK Negeri 4 Bandar Lampung using two methods: the Decision Tree C4.5 algorithm and the Naïve Bayes algorithm. The system utilizes seven key attributes as predictive variables, including mathematics scores, English scores, science (IPA) scores, Indonesian language scores, academic achievements, participation in extracurricular activities, and color blindness condition. The study findings indicate that the C4.5 algorithm achieves an accuracy of 84.46%, whereas the Naïve Bayes algorithm outperforms it with an accuracy of 92.23%. This suggests that the Naïve Bayes algorithm is more effective for this application. Nevertheless, both methods still have limitations that can be improved through parameter optimization and more in-depth data processing. The implementation of this data-driven system is expected to enhance the efficiency of providing more relevant major recommendations at SMK Negeri 4 Bandar Lampung and serve as an inspiration for other schools to adopt similar approaches to improve education quality.

Keywords: Naive Bayes; C4.5; Confusion Matrix

1. PENDAHULUAN

Pemerintah Indonesia secara terus-menerus berupaya meningkatkan kualitas sumber daya manusia melalui kerja sama antara pemerintah pusat dan daerah dalam rangka memperbaharui sekolah menengah kejuruan (SMK). Jika jumlah industri atau perusahaan setidaknya sebanding dengan jumlah SMK, kapasitas penerimaan siswa di sekolah kejuruan dapat meningkat secara signifikan, bahkan hingga dua kali lipat. Peningkatan mutu pendidikan, khususnya kualitas lulusan SMK, berpotensi mendorong kemajuan berbagai proyek infrastruktur yang menjadi prioritas nasional. Meski demikian, penyerapan tenaga kerja bagi lulusan SMK masih tertinggal bila dibandingkan dengan lulusan sekolah menengah atas [1]. Data menunjukkan bahwa tingkat pengangguran di kalangan lulusan SMK lebih tinggi dibandingkan dengan program studi lain. Misalnya, menurut Badan Pusat Statistik, pada Agustus 2018, persentase pengangguran di kalangan lulusan SMK mencapai 11,25%, naik dari 8,92% pada Februari 2018. Salah satu penyebab utama fenomena ini adalah rendahnya mutu pendidikan yang diterima, sehingga para siswa kesulitan mengakses lapangan pekerjaan di industri setelah mereka menyelesaikan studi. Oleh karena itu, penentuan jurusan yang tepat sangat krusial untuk mendukung prestasi akademik dan kenyamanan proses belajar, karena pilihan yang kurang tepat dapat menimbulkan rasa tidak tertarik, sikap apatis, dan akhirnya menurunkan kualitas hasil belajar [2].

SMK Negeri 4 Bandar Lampung merupakan salah satu institusi kejuruan yang menawarkan beragam program studi, mulai dari akuntansi, desain komunikasi visual, perhotelan, hingga pengembangan perangkat lunak, manajemen perkantoran, ritel, pariwisata, seni kuliner, serta jurusan di bidang mode dan teknologi jaringan komputer. Bagi para lulusan sekolah menengah, proses memilih jurusan di sekolah ini merupakan tahap penting yang harus disesuaikan dengan kemampuan dan minat masing-masing, karena kesesuaian tersebut sangat berpengaruh terhadap kenyamanan belajar dan keberhasilan akademik [3]. Pemilihan jurusan yang kurang tepat berpotensi menyebabkan suasana belajar yang tidak menyenangkan, menurunnya minat, dan akhirnya berdampak pada prestasi akademik.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2024 Page: 2700−2708
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6928

Dalam upaya membantu proses pemilihan jurusan yang optimal, penelitian ini mengadopsi dua metode analisis berbasis data, yaitu pendekatan pohon keputusan C4.5 dan algoritma Naive Bayes. Metode C4.5 unggul dalam menghasilkan aturan-aturan berupa pohon keputusan yang mudah dipahami, karena aturan tersebut disusun dari variabel-variabel yang relevan sehingga memberikan penjelasan logis atas prediksi yang dibuat [4]. Di sisi lain, algoritma Naive Bayes memanfaatkan prinsip probabilitas dengan mengamati pola distribusi data historis untuk menghasilkan prediksi yang akurat. Penggabungan kedua metode ini diharapkan dapat menghasilkan model prediktif yang lebih terperinci dan andal [5]. Penelitian ini bertujuan untuk meramalkan pilihan jurusan di SMK Negeri 4 Bandar Lampung dengan menerapkan metode C4.5 dan algoritma Naive Bayes. Tujuh variabel dijadikan parameter utama dalam model prediktif, yaitu nilai matematika, bahasa Inggris, sains, bahasa Indonesia, nilai umum, partisipasi dalam kegiatan ekstrakurikuler, serta kondisi buta warna, guna menentukan jurusan yang paling sesuai untuk setiap siswa. Penggunaan teknologi dalam analisis data ini diharapkan dapat menghasilkan rekomendasi yang objektif, transparan, dan akurat, dengan data siswa sebagai basis utama dalam penyusunan model.

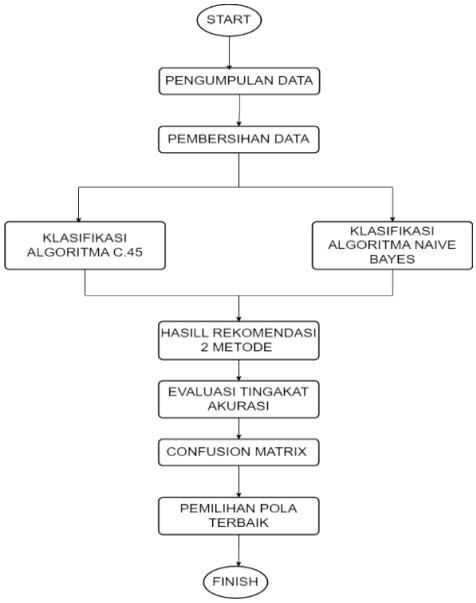
Ada beberapa penelitian sebelumnya yang mengangkat hal ini diantaranya penelitian Choirul Anam Dan Harry Budi Santoso pada Mei 2018, mendapatkan hasil algoritma C4.5 menunjukkan tingkat akurasi lebih tinggi sebesar 96.40% dibandingkan Naïve Bayes yang mencapai 95.11% [6]. Sedangkan pada penelitian lain oleh Muhammad Kamil dan Widia Choli pada September 2020, mendapatkan hasil C4.5 juga memiliki akurasi lebih baik sebesar 69.54% dibandingkan Naïve Bayes sebesar 68.38% [7]. Berikutnya ada penelitian dari Saufika Sukmawati, dkk pada Mei 2022, mendapatkan hasil C4.5 menunjukkan akurasi tertinggi sebesar 84.84%, jauh lebih unggul dibandingkan Naïve Bayes yang hanya mencapai 58.29% [8]. Terakhir penelitian dari “Bayu Sugara1, Dkk pada Maret 2019, Mendapatkan hasil untuk Naïve Bayes memiliki keunggulan dengan akurasi 73.33%, sedikit lebih tinggi dibandingkan C4.5 yang mencapai 72% [9].

Tujuan utama dari penelitian adalah membandingkan efektivitas metode C4.5 dan algoritma Naive Bayes dalam memprediksi pilihan karir yang tepat di SMK Negeri 4 Bandar Lampung. Hasil yang diperoleh diharapkan dapat memperbaiki proses pengambilan keputusan berbasis data serta meningkatkan kepuasan siswa terhadap program studi yang mereka pilih. Dengan demikian, SMK Negeri 4 Bandar Lampung dapat lebih mengoptimalkan potensi masing-masing siswa dan memberikan ruang bagi mereka untuk menentukan jalur pendidikan sesuai dengan minat dan kemampuan, sementara model yang dikembangkan pun diharapkan dapat diaplikasikan di institusi pendidikan lain guna menaikkan mutu pendidikan secara menyeluruh.

2. METODOLOGI PENELITIAN

2.1 Alur Penelitian

Ini adalah gambaran alur penelitian yang penulis lakukan yaitu dari mulai pengumpulan data sampel pada hasilnya yaitu mendapatkan pola atau atribut yang sesuai dalam pemilihan jurusan yang sesuai dengan minat dari siswa SMP ada pada Gambar 1.



Gambar 1. Alur Penelitian

2.2 Pengumpulan Data

Pengumpulan data adalah proses mengumpulkan informasi yang relevan dan diperlukan untuk suatu tujuan tertentu. Dalam konteks ini, data yang dikumpulkan berasal dari siswa SMP yang akan melanjutkan pendidikan ke SMK [10].

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2024 Page: 2700−2708
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6928

Proses ini bertujuan untuk memahami berbagai aspek yang mempengaruhi pilihan mereka dalam menentukan jurusan, seperti nilai minat, bakat, serta faktor eksternal seperti Kesehatan mata exskul dan lainnya. Pengumpulan data dapat dilakukan melalui berbagai metode, seperti wawancara, kuesioner, atau observasi, guna memperoleh informasi yang akurat dan dapat digunakan sebagai dasar dalam pengambilan keputusan, baik oleh siswa itu sendiri maupun oleh pihak sekolah dalam memberikan bimbingan dan arahan yang tepat.

2.3 Pembersihan Data

Selanjutnya adalah tahap pembersihan data, di mana data yang tidak sesuai atau tidak relevan dihapus. Data yang dihapus adalah data yang tidak lengkap karena beberapa data siswa tidak dimasukkan ke dalam formulir. Namun, data yang digunakan sekarang sudah bersih, jadi tidak perlu membersihkan dataset lagi.[11].

2.4 Klasifikasi algoritma C4.5

Algoritma C4.5, juga dikenal sebagai "pohon keputusan", adalah evolusi dari algoritma ID3. Algoritma ini memiliki kelebihan yang mudah dipahami, fleksibel, dan menarik karena dapat divisualisasikan sebagai gambar pohon keputusan [12].

Algoritma C4.5 adalah struktur pohon di mana setiap daun menunjukkan kelas, setiap cabang menunjukkan hasil dari atribut yang diuji, dan terdapat simpul yang mendeskripsikan atribut. Algoritma C4.5 secara rekursif mengunjungi setiap titik keputusan dan memilih pembagian yang optimal sampai tidak dapat dibagi lagi. Algoritma ini melakukan ini dengan menggunakan gagasan peningkatan informasi atau pengurangan entropy.
Dengan algoritma C4.5, sebuah pohon keputusan dibuat dalam beberapa tahap. [13] yaitu:

a. Menyiapkan data pelatihan. Data pelatihan biasanya berasal dari data historis yang sudah dikelompokkan ke dalam kelas tertentu.
b. Menemukan akar pohon. Setelah nilai gain dari masing-masing atribut dihitung, akar pertama akan dipilih. Sebelum menghitung nilai gain dari atribut, nilai entropy harus dihitung. Nilai entropy dapat dihitung dengan menggunakan rumus berikut.

$$Entropy\ (S) = \sum_{i=1}^{n} - pi.log2pi \quad (1)$$

Rumus Entropy(S) digunakan untuk mengukur tingkat ketidakpastian atau keacakan dalam suatu himpunan data **S**. Secara matematis, entropy dihitung dengan menjumlahkan hasil perkalian antara probabilitas setiap kategori dalam dataset dengan logaritma basis dua dari probabilitas tersebut, kemudian dikalikan dengan negatif satu. Dalam rumus ini, **S** merupakan himpunan kasus atau kumpulan data yang sedang dianalisis, sementara **n** adalah jumlah partisi atau kategori dalam himpunan **S**. Proporsi dari setiap kategori terhadap jumlah total kasus dinotasikan sebagai **p_i**, yang menunjukkan seberapa besar bagian setiap kategori dalam dataset. Nilai logaritma basis dua dari **p_i** digunakan untuk mengukur tingkat ketidakpastian dari setiap kategori.

c. Kemudian hitung nilai gain menggunakan rumus:

$$Gain(S,A) \quad = Entropy\ (S) - \sum_{k=0}^{n} \frac{|St|}{|S|} x\ Entropy\ (St) \quad (2)$$

Rumus ini menunjukkan bahwa Information **Gain** dihitung dengan mengurangi **Entropy(S)** dari jumlah hasil perkalian antara proporsi **St** terhadap **S** dan **Entropy(St)** untuk setiap nilai atribut yang digunakan dalam pemisahan data. Secara matematis, rumus ini dapat dijelaskan sebagai berikut:

1. Entropy(S) mengukur tingkat ketidakpastian atau ketidakteraturan dalam dataset S sebelum dilakukan pemisahan berdasarkan atribut A.
2. Bagian ∑ (|St| / |S|) × Entropy(St) menghitung rata-rata entropi dari subset St yang terbentuk setelah pemisahan oleh atribut A, dengan mempertimbangkan bobot proporsi |St| / |S|.
3. Selisih antara Entropy(S) dan nilai yang dihitung dari rata-rata entropi subset menentukan Information Gain, yang menunjukkan seberapa besar pengurangan ketidakpastian setelah pemisahan menggunakan atribut A. Semakin tinggi nilai Information Gain, semakin baik atribut tersebut dalam membagi dataset, karena mengurangi ketidakpastian dalam klasifikasi.
4. Ulangi langkah kedua hingga semua rekor terkumpul.
5. Jika salah satu dari kondisi berikut terpenuhi:
   a) Semua rekaman dalam simpul N memiliki kelas yang sama;
   b) Tidak ada rekaman yang dipartisi lagi; atau.
   c) Tidak ada rekaman dalam cabang yang kosong

2.5 Klasifikasi Algoritma Naïve Bayes

Untuk memperkirakan kemungkinan keanggotaan suatu kelas, klasifikasi statistik yang dikenal sebagai klasifikasi Bayes [16] dapat digunakan. Naive Bayes—juga dikenal sebagai "idiot's Bayes", "simple Bayes", dan "independence Bayes"—adalah metode klasifikasi Bayes yang baik karena mudah digunakan, tidak membutuhkan skema estimasi parameter perulangan yang rumit, dan dapat digunakan untuk data set yang sangat besar. Sangat mudah dipahami

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2024 Page: 2700−2708
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6928

bahkan bagi orang yang tidak terbiasa dengan teknologi klasifikasi. Teorema Bayes, yang dikenal sebagai nama ahli matematika dan menteri Prebysterian Inggris Thomas Bayes (1702-1761), adalah dasar dari klasifikasi Bayes yaitu:

$$P(H|X)=\frac{P(P|H)P(H)}{P(X)} \quad (3)$$

Dari rumus ke 3 diketehui bahwa y = data dengan kelas yang belum diketahui, x = hipotesis data y merupakan suatu kelas spesifik, P(x|y) = probabilitas hipotesis x berdasar kondisi y (posteriori probability) , P(x) = probabilitas hipotesis x (prior probability), P(y|x) = probabilitas y berdasarkan kondisi pada hipotesis x , P(y) = probabilitas dari y. Naïve bayes adalah penyederhanaan metode bayes. Teorema bayes disederhanakan menjadi:

$$P(x|y) = P(x|y)\ P(x) \quad (4)$$

Dalam algoritma Naive Bayes, teorema Bayes digunakan, yang berarti menggabungkan probabilitas prior dan probabilitas bersyarat untuk membuat rumus yang dapat digunakan untuk menghitung kemungkinan setiap klasifikasi yang mungkin. Berikut ini adalah diagram alur dari pendekatan Naive Bayes :

a. Input data training
b. Baca data training.
c. Hitung jumlah dan probabilitas, namun apabila data numerik maka:
    1. Cari nilai mean dan standar deviasi dari masing-masing parameter yang merupakan data numerik.
    2. Cari nilai probabilistik dengan cara menghitung jumlah data yang sesuai dari kategori yang sama dibagi dengan jumlah data pada kategori tersebut.
    3. Mendapatkan nilai dalam tabel mean, standar deviasi dan probabilitas

2.6 Evaluasi (Akurasi, Presisi, Recal)

Akurasi didefinisikan sebagai tingkat kedekatan antara nilai prediksi dengan nilai aktual [19]. presisi menunjukkan tingkat ketepatan atau ketelitian dalam pengklasifikasian [20]. dan recall berfungsi untuk mengukur proporsi positif aktual yang benar diidentifikasi [21]. Biasanya,

$$ACCURACY = \frac{TP+TN}{TP+TN+FP+FN}\ X\ 100\ \% \quad (5)$$

Akurasi adalah ukuran yang menunjukkan seberapa baik model klasifikasi dalam memprediksi hasil dengan benar secara keseluruhan. Akurasi dihitung dengan membagi jumlah true positive (TP) dan true negative (TN) dengan total jumlah prediksi, yang mencakup true positive (TP), true negative (TN), false positive (FP), dan false negative (FN)

$$PRESISI = \frac{TP+TN}{TP+FP}\ X\ 100\ \% \quad (6)$$

Presisi, yang juga dikenal sebagai nilai prediktif positif, mengukur seberapa banyak prediksi positif yang dibuat oleh model benar-benar sesuai dengan kenyataan. Presisi dihitung dengan membandingkan jumlah true positive (TP) dan true negative (TN) dengan jumlah total true positive (TP) dan false positive (FP)

$$RECALL = \frac{TP}{TP+FN}\ X\ 100\ \% \quad (7)$$

Recall, yang juga dikenal sebagai sensitivitas atau true positive rate, mengukur kemampuan model dalam mengidentifikasi semua kasus positif yang sebenarnya. Recall dihitung dengan membagi jumlah true positive (TP) dengan jumlah total true positive (TP) dan false negative (FN)

2.7 Confusion matrix

Cara confusion matrix bekerja adalah dengan mengolah data untuk membandingkan hasil prediksi dengan label asli. Kauras, presisi, dan ulang menghasilkan nilai dari evaluasi dengan confusion matrix. Dalam pemeriksaan klasifikasi, ada empat kemungkinan hasil dari klasifikasi data [25]. Data negatif, jika diprediksi negatif, dianggap benar negatif, dan data positif, jika diprediksi positif, dianggap benar positif. Data positif, jika diprediksi positif, dianggap benar positif [26].

Tabel 1. *confusion matrix*

Actual	**Prediction**	
	Positif	Negative
Positif	True Positif (TP)	True Negative(TN)
Negative	False Positif (FP)	False Negatif (FN)

Tabel 1 menunjukkan nilai true positive (TP) dan true negative (TN), yang masing-masing menunjukkan tingkat ketepatan klasifikasi. Semakin tinggi nilai TP dan TN, semakin baik tingkat klasifikasi dari akurasi, presisi, dan recall. Jika label prediksi keluaran bernilai benar dan nilai sebenarnya bernilai salah, itu disebut false positive

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2024 Page: 2700−2708
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6928

(FP). Sebaliknya, jika label prediksi keluaran bernilai salah dan nilai sebenarnya bernilai benar, itu disebut false negative (FN) [27].

3. HASIL DAN PEMBAHASAN

Dataset yang digunakan untuk penelitian ini berasal dari daftar siswa SMP (Sekolah Menengah Pertama), yang berisi variabel yang akan digunakan dalam penelitian ini. Dataset ini dibagi menjadi 30% sebagai data pengujian dan 70% sebagai data pelatihan.

Tabel 2. Data Mahasiswa Setiap Tahun

Nomor	variabel	Jumlah record	Kelas
1	Nilai Matekatika	1285	4 (A,B,C,D)
2	Nillai Bahasa Inggris	1285	4 (A,B,C,D)
3	Nilai IPA	1285	4 (A,B,C,D)
4	Nilai Bahasa Indonesia	1285	4 (A,B,C,D)
5	Prestasi	1285	2 (YA, TIDAK)
6	Exkull	1285	2 (YA, TIDAK)
7	Buta warna	1285	2 (YA, TIDAK)
8	Label	1285	2 (YA, TIDAK)

3.1. Analisis Data

Dalam data mining, algorithma Naive Bayes dan Decesion Tree C4.5 adalah metode klasifikasi. Satu-satunya perbedaan antara keduanya adalah bahwa algorithma Naive Bayes berfokus pada asumsi yang didasarkan pada teori Bayesian, sedangkan algorithma Decesion Tree menggunakan pohon keputusan untuk hasil akhirnya. Data yang digunakan oleh penulis adalah data PPDB SMKN 04 Bandar Lampung tahun 2017/2018, yang memiliki 1285 siswa dan 8 pilihan jurusan.

3.2. Seleksi data

Pada titik ini, pengumpulan data akan dilakukan. Data yang akan digunakan terdiri dari data PPDB di SMKN 04 Bandar Lampung tahun 2017/2018 dan data pilihan jurusan yang dipilih oleh siswa SMP yang akan mendaftar di SMKN 04 Bandar Lampung.

Tabel 3. Data Mahasiswa Setiap Tahun

Tahun PPDB	Jumlah Siswa
2017/2018	1285
Total	1285

Tabel 4 berikut menunjukkan data siswa yang telah disesuaikan dengan atribut dan akan digunakan sebagai data perbaikan. Karena banyaknya data siswa tidak lengkap, hanya beberapa yang dapat digunakan dari data ini akan dipilih. karena itu kita harus menyelesaikan pembersihan.

Tabel 4. Contoh Atribut Data Siswa

no	nama	no daftar	mat	b.ing	ind	ipa	prestasi	ekskul	buta warna
1	YUNI MARLINA	49210410040001	35	54	70	37,5	YA	YA	YA
2	ROSSA NANDA SAFITRI	49210410040002	42,5	52	74	42,5	YA	YA	YA
3	DIAN AWALIYAH	49210410040003	55	42	70	50	TIDAK	TIDAK	TIDAK
4	HALIMA TUSA'DIAH	49210410040004	85	54	62	62,5	TIDAK	TIDAK	TIDAK
5	M. ILHAM MANSIZ	49210410040005	45	52	62	50	YA	TIDAK	TIDAK
6	RESTI SINTIYA	49210410040006	35	42	65	32,5	YA	TIDAK	YA
7	WULAN SEPRIYANI	49210410040007	27,5	32	65	35	YA	YA	YA
8	FAUZAN MUAMMAR	49210410040008	25	30	65	50	TIDAK	YA	TIDAK
9	YURIKA NABILA	49210410040009	40	56	78	47,5	TIDAK	TIDAK	TIDAK
10	HELNIDA RISFA MASAYA	49210410040010	40	38	70	90	TIDAK	YA	TIDAK

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2024 Page: 2700−2708
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6928

Berdasarkan Tabel 4. penulis akan menganalisis data tersebut untuk menentukan atribut mana yang akan digunakan pada tahapan selanjutnya. Ini menunjukkan bahwa beberapa atribut dapat digunakan sebagai variable untuk menentukan jurusan yang tepat, dan beberapa atribut telah dipilih sebelumnya terlihat seperti pada tabel 5. Sebagai berikut :

Tabel 5. Data Mahasiswa Setiap Tahun

No	Atribut
1	Nilai Matekatika
2	Nillai Bahasa Inggris
3	Nilai IPA
4	Nilai Bahasa Indonesia
5	Prestasi
6	Exkull
7	Buta Warna

3.3. Data *Cleaning* (Pembersihan Data)

Selanjutnya adalah tahap pembersihan data, di mana data yang tidak sesuai atau tidak relevan dihapus. Data yang dihapus adalah data yang tidak lengkap karena beberapa data siswa tidak dimasukkan ke dalam formulir. Namun, data yang digunakan saat ini sudah bersih, jadi tidak perlu membersihkan dataset lagi.

3.4. Data Transformation (Transformasi Data)

Tidak semua format data dapat diolah oleh Google Cloud. Setelah melalui tahapan pemilihan dan perbaikan data, data masih harus melalui tahapan. Aplikasi Google Cloud Platform dapat mengolah berbagai format data, termasuk format data *.xlsx dan *.csv. Dalam penelitian ini, peneliti menggunakan data berformat *.csv untuk diolah ke dalam Google Cloud Platform.

Tabel 6. Hasil data transformation

No	MAT	B.ING	IND	IPA	PRESTASI	EKSKUL	BUTA WARNA
1	35	54	70	37,5	YA	YA	YA
2	42,5	32	74	42,5	YA	YA	YA
3	55	42	70	50	TIDAK	TIDAK	TIDAK
4	85	64	82	82,5	TIDAK	TIDAK	TIDAK
5	35	52	62	50	YA	YA	TIDAK
6	35	42	66	32,5	YA	TIDAK	YA
7	27,5	32	66	35	YA	YA	YA
8	25	30	64	50	TIDAK	YA	TIDAK
9	40	56	78	47,5	TIDAK	TIDAK	TIDAK
10	40	38	70	30	TIDAK	YA	TIDAK

3.5. Data Mining

Tahap Proses ini dilakukan pada Google Cloud dengan menggunakan Klasifikasi Naive Bayes dan Algoritma Decesion Tree C4.5. Setelah file data diolah, file diformat menjadi *.csv. Dalam bahasa ini, proses pengolahan data yang dikenal sebagai Data Mining dilakukan, yang menggunakan Google Cloud dengan bahasa phyton.

3.6. Tahap Klasifikasi

Pada tahap klasifikasi, dataset yang telah ditentukan di-import. Selanjutnya, proses pelatihan dan pengujian akan dilakukan dengan algoritma yang disarankan; proses pertama menggunakan Decesion TreeC4.5 dan proses kedua menggunakan Naive Bayes algoritme.

3.7. Hasil Rekomendasi

Hasil Hasil rekomendasi Naive Bayes yang dihasilkan oleh Google Colab (Phyton) dan rancangan proses untuk prediksi data penjurusan siswa menggunakan metode Naive Bayes ditunjukkan pada Gambar 2 di bawah ini.

```
Akurasi Naive Bayes: 92.23%
Precision Naive Bayes: 100.00%
Recall Naive Bayes: 57.75%
```

Gambar 2. Hasil Naïve Bayes

3.8. Hasil Rekomendasi

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2024 Page: 2700−2708
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6928

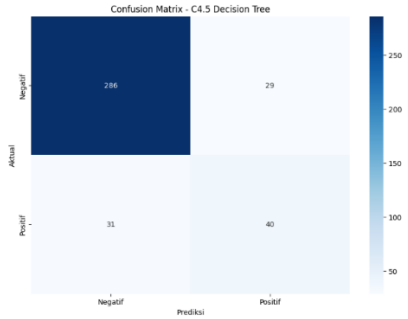
Hasil rekomendasi Decesion TreeC4.5 menggunakan kolaborasi Google (Phyton) digambarkan pada Gambar 3. Rancangan proses klasifikasi data penjurusan siswa menggunakan metode decesion tree C4.5 ditunjukkan di sini.

```
Akurasi C4.5 Decision Tree: 84.46%
Precision C4.5 Decision Tree: 57.97%
Recall C4.5 Decision Tree: 56.34%
```

Gambar 3. Decesion Tree C45

3.9. Confusion Matrix C4.5

Confusion Gambar 4 berikut menunjukkan proses yang dirancang untuk mengklasifikasikan data penjurusan siswa menggunakan metode decesion tree C4.5, yang menggunakan kolaborasi Google (Phyton).



Gambar 4. *Confusion Matrix Decesion Tree C4.5*

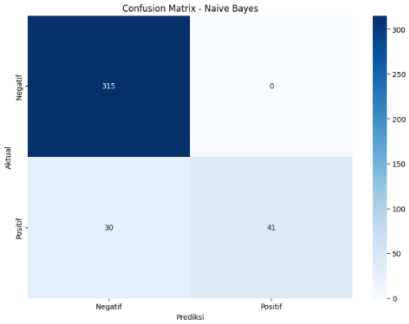
Confusion Matrix digunakan dalam model C4.5 Decision Tree untuk mengevaluasi performa klasifikasi terhadap dua kelas, Negatif dan Positif. Hasil matrix menunjukkan bahwa model berhasil mengklasifikasikan 286 data dengan benar sebagai Negatif (True Negative/TN), tetapi 29 data yang sebenarnya Negatif tetapi salah diklasifikasikan sebagai Positif (False Positive/FP), dan 40 data yang sebenarnya Positif berhasil diprediksi dengan benar (True Positive/TP), sehingga hasilnya tidak sesuai dengan prediksi sebelumnya.

Berdasarkan perhitungan akurasi, model ini memiliki tingkat akurasi sekitar 84,4%, yang menunjukkan bahwa secara keseluruhan model mampu mengklasifikasikan data dengan cukup baik. Namun, dalam hal presisi untuk kelas Positif, model hanya memiliki tingkat presisi sekitar 58%, yang berarti hanya 58% dari semua prediksi Positif yang benar. Selain itu, model memiliki recall sebesar 56%, yang menunjukkan bahwa dari seluruh data yang benar-benar Positif, hanya 58% Oleh karena itu, skor F1, yang menghitung keseimbangan antara recall dan presisi, berada di 57%.

Hasilnya adalah bahwa model lebih baik dalam mengenali kelas Negatif daripada kelas Positif. Namun, karena jumlah kelas Negatif Palsu dan Positif Palsu yang cukup tinggi, model masih memiliki kelemahan dalam mendeteksi data Positif dengan akurat. Penyesuaian ambang, balancing data, atau penggunaan teknik lain seperti oversampling dan tuning hyperparameter diperlukan jika tujuan utama klasifikasi ini adalah untuk menemukan kasus Positif yang lebih baik. Selain itu, model dapat ditingkatkan melalui eksplorasi fitur tambahan atau penggunaan algoritma yang lebih kompleks.

3.10. Confusion Matrix Naive Bayes

Dengan menggunakan Google Colab (Phyton), confusion matrix Naïve Bayes digunakan. Rancangan proses untuk prediksi data penjurusan siswa menggunakan metode naïve bayes ditunjukkan pada Gambar 5 di bawah ini.



Gambar 5. *Confusion Matrix Naive Bayaes*

Konfusion Matrix dari model Naive Bayes menunjukkan hasil klasifikasi terhadap dua kelas, Negatif dan Positif. Hasilnya menunjukkan bahwa model berhasil mengklasifikasikan 315 data dengan benar sebagai Negatif (True Negative/TN), tanpa kesalahan dalam memprediksi data Negatif sebagai Positif (False Positive/FP = 0), 30 data

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2024 Page: 2700−2708
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6928

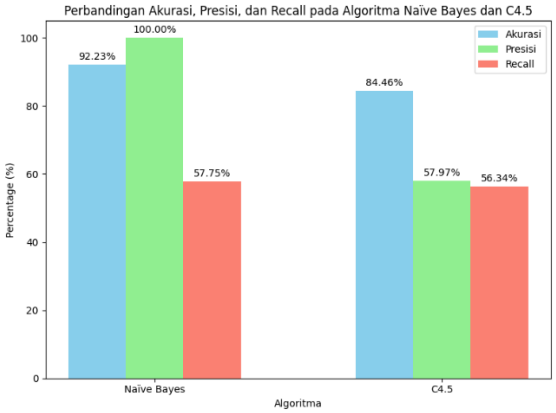
yang sebenarnya Positif tetapi salah diklasifikasikan sebagai Negatif (False Negative/FN), dan 41 data yang benar diklasifikasikan sebagai Positif.

Sebagai hasil dari perhitungan evaluasi, model ini menunjukkan akurasi sekitar 92,2%, menunjukkan bahwa secara keseluruhan, itu melakukan pekerjaan yang cukup baik untuk klasifikasi. Selain itu, model memiliki presisi sempurna untuk kelas Negatif (100%) karena tidak ada data Negatif yang salah diklasifikasikan sebagai Positif. Namun, presisi model untuk kelas Positif sekitar 57,8%, yang menunjukkan bahwa dari semua prediksi Positif, hanya 57,8% yang benar. Recall untuk kelas Positif juga sekitar 57,7%, yang menunjukkan bahwa model masih kesulitan mengidentifikasi semua data Positif dengan benar. Selain itu, skor F1-nya sebesar 57,7% menunjukkan keseimbangan antara recall dan presisi.

Dari hasil ini, dapat disimpulkan bahwa model Naive Bayes sangat baik dalam mengenali data Negatif, tetapi kurang optimal dalam mengenali data Positif karena ada cukup banyak False Negative. Jika fokus klasifikasi adalah meningkatkan deteksi data Positif, model dapat diperbaiki dengan membagi data, mengoptimalkan fitur, atau meneliti teknik yang lebih kompleks.

3.11. Evaluasi

Pada titik ini, kami dapat menguji akurasi dan recall dari masing-masing metode. Tingkat akurasi naive bayes adalah 92.23%, dan Decesion Tree C4.5 adalah 84.46%.



Gambar 6. Grafik Perbadingan Kedua Algoritma

Kemudian kita bisa melihat hasil komparasi dari kedua metode tersebut bisa kita lihat akurasi, presisi, dan recal dari yang tersaji pada tabel 7. sebagai berikut

Tabel 7. Komparasi *Naive Bayes* dan *Decision Tree* C4.5

Nomor	Algoritma	Akurasi	Presisi	Recall
1	**Naïve Bayes**	**92.23%**	**100.00%**	**57.75%**
2	C4.5	84.46%	57.97%	56.34%

Dari Tabel 7 komparasi tersebut bisa kita simpulkan bahwa metode Algortima *Nive Bayes* jauh lebih baik di bandingkan dengan metode algoritma *Decision Tree C4,5* dalam menentukan pemilihan jurusan yang berada di SMKN 4 Bandar Lampung, dengan hasil yang tinggi di bagian algortima *Naive Bayes* dengan hasil *Akurasi Presisi Dan Recal* yang sama besar sedangkan untuk Algoritma *Decision Tree C4.4* hannya di bagian akurasi yang mendapat hasil yang bagus, akan tetapi di bagian recall dan presisi masih jauh lebih baik di Algoritma *naive bayes* .

4. KESIMPULAN

Berdasarkan Berdasarkan hasil evaluasi, algoritma Naïve Bayes dan C4.5 (Decision Tree) menunjukkan perbandingan performa dengan keunggulan tipis di pihak Naïve Bayes. Algoritma Naïve Bayes memiliki tingkat akurasi, presisi, dan recall masing-masing sebesar 92,23%, 100,00%, dan 57,75%, yang menunjukkan bahwa model ini mampu membuat prediksi benar untuk 92% dari total data. Berdasarkan Matriks Konflik, terlihat bahwa algoritma Naïve Bayes sering melakukan kesalahan. Sementara itu, algoritma C4.5 Decision Tree mencatatkan performa yang lebih rendah dibandingkan Naïve Bayes, dengan akurasi 84,46%, presisi 57,97%, dan recall 56,34%. Berdasarkan Confusion Matrix, model ini masih memiliki kelemahan dalam mengklasifikasikan kedua kelas, dengan False Negative sebanyak 31 dan False Positive sebanyak 29. Hal ini menunjukkan bahwa model masih sering melakukan kesalahan dalam mengenali data Positif dan Negatif. Secara keseluruhan, Naive Bayes lebih akurat dan presisi daripada C4.5, terutama dalam mengenali data negatif dengan sangat baik. Tetapi kedua model masih perlu dioptimalkan lebih lanjut jika tujuan klasifikasi adalah untuk meningkatkan deteksi data positif. Oleh karena itu, untuk meningkatkan kualitas prediksi dan performa model, diperlukan tindakan tambahan seperti optimasi parameter, balancing data, pengolahan fitur yang lebih baik, atau eksplorasi algoritma lainnya.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2024 Page: 2700−2708
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6928

REFERENCES

[1] F. Agustina, A. T. Sumpala, and A. Arysespajayadi, "SPK Pemilihan Jurusan Siswa Baru Menggunakan Metode AHP dan MOORA Pada SMKN 1 Kolaka," *Jurnal Sains Dan Informatika*, vol. 7, no. 1, pp. 87–96, 2021.

[2] M. Ridwan, F. Badri, and B. M. Basuki, "Rancang Bangun Sistem Pendukung Keputusan Penjurusan Mahasiswa Teknik Elektro Unisma Menggunakan Metode Analytical Hierarchy Process (AHP)," *SCIENCE ELECTRO*, vol. 17, no. 1, 2024.

[3] N. S. Atmaja, "Attribution-NonCommercial 4.0 International. Some rights reserved Sistem Pendukung Keputusan Sistem Pendukung Keputusan Pemilihan Jurusan Menggunakan Metode PROMETHEE (Studi Kasus: SMK Negeri 6 Medan)," *J. Nas. Inform. dan Teknol. Jar*, vol. 5, no. 2, pp. 75–84, 2021.

[4] A. F. O. Pasaribu, "Analisis Pola Menggunakan Metode C4. 5 Untuk Peminatan Jurusan Siswa Berdasarkan Kurikulum (Studi Kasus: Sman 1 Natar)," *Jurnal Teknologi Dan Sistem Informasi*, vol. 2, no. 1, pp. 80–85, 2021.

[5] P. P. Putra and A. S. Chan, "Pengembangan Aplikasi Perhitungan Prediksi Stock Motor Menggunakan Algoritma C 4.5 Sebagai Bagian dari Sistem Pengambilan Keputusan (Studi Kasus di Saudara Motor)," *INOVTEK Polbeng-Seri Informatika*, vol. 3, no. 1, pp. 24–33, 2018.

[6] C. Anam and H. B. Santoso, "Perbandingan Kinerja Algoritma C4.5 dan Naive Bayes untuk Klasifikasi Penerima Beasiswa," 2018.

[7] M. Kamil, W. Cholil, "Perbandingan Algoritma C4.5 dan Naive Bayes Pada Lulusan Tepat Waktu Mahasiswa," *JURNAL INFORMATIKA*, vol. 7, no. 2, pp. 97–106, 2020

[8] S. Sukmawati, H. Februariyanti, A. Jananto, "Perbandingan Algoritma C 4.5 Dan Algoritma Naive Bayes Untuk Klasifikasi Pekerja Migran Indonesia," *Jurnal Informatika, Manajemen dan Komputer*, vol. 14, no. 1, 2022.

[9] B. Sugara, D. Adidarma, and S. Budilaksono, "Perbandingan Akurasi Algoritma C4. 5 dan Naïve Bayes untuk Deteksi Dini Gangguan Autisme pada Anak," *Jurnal IKRA-ITH Informatika*, vol. 3, no. 1, pp. 119–128, 2019.

[10] P. C. Susanto, D. U. Arini, L. Yuntina, J. P. Soehaditama, and N. Nuraeni, "Konsep Penelitian Kuantitatif: Populasi, Sampel, dan Analisis Data (Sebuah Tinjauan Pustaka)," *Jurnal Ilmu Multidisplin*, vol. 3, no. 1, pp. 1–12, 2024.

[11] H. Syah and A. Witanti, "Analisis Sentimen Masyarakat Terhadap Vaksinasi Covid-19 Pada Media Sosial Twitter Menggunakan Algoritma Support Vector Machine (Svm)," *Jurnal Sistem Informasi Dan Informatika (Simika)*, vol. 5, no. 1, pp. 59–67, 2022.

[12] M. Kamil and W. Cholil, "Analisis Perbandingan Algoritma C4. 5 dan Naive Bayes pada Lulusan Tepat Waktu Mahasiswa di Universitas Islam Negeri Raden Fatah Palembang," *Jurnal Informatika*, vol. 7, no. 2, pp. 97–106, 2020.

[13] R. Pratama, B. Huda, E. Novalia, and H. Kabir, "Perbandingan Algoritma C4. 5 dan Naïve Bayes dalam Menentukan Persediaan Stok," *METIK JURNAL*, vol. 6, no. 2, pp. 115–122, 2022.

[14] C. R. A. Nugroho and T. Kristiana, "Penerapan Algoritma C4. 5 Untuk Kepuasan Pelanggan Toko Online Parfume Chantik," *Jurnal Algoritme*, vol. 3, no. 1, pp. 10–21, 2022.

[15] F. F. Harryanto and S. Hansun, "Penerapan Algoritma C4. 5 untuk Memprediksi Penerimaan Calon Pegawai Baru di PT WISE," *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, vol. 3, no. 2, pp. 95–103, 2017.

[16] F.-J. Yang, "An implementation of naive bayes classifier," in *2018 International conference on computational science and computational intelligence (CSCI)*, IEEE, 2018, pp. 301–306.

[17] R. Jiandi, "Implementasi Algoritma C4. 5 untuk Prediksi Potensi Mahasiswa Sebagai Pengurus Organisasi Menggunakan Data Hasil PAPI KOSTICK (Studi Kasus: Universitas Multimedia Nusantara)," *Universitas Multimedia Nusantara, Tangerang*, 2016.

[18] N. Mona, "Konsep isolasi dalam jaringan sosial untuk meminimalisasi efek contagious (kasus penyebaran virus corona di Indonesia)," *Jurnal sosial humaniora terapan*, vol. 2, no. 2, p. 12, 2020.

[19] M. C. Wijanto, "Sistem pendeteksi pengirim tweet dengan metode klasifikasi naive Bayes," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 1, no. 2, 2015.

[20] L. Qadrini, A. Seppewali, and A. Aina, "Decision tree dan adaboost pada klasifikasi penerima program bantuan sosial," *Jurnal inovasi penelitian*, vol. 2, no. 7, pp. 1959–1966, 2021.

[21] R. N. Hidayat, L. M. Sabri, and M. Awaluddin, "Analisis desain jaring GNSS berdasarkan fungsi presisi (studi kasus: titik geoid geometri Kota Semarang)," *Jurnal Geodesi Undip*, vol. 8, no. 1, pp. 48–55, 2019.

[22] A. Primajaya and B. N. Sari, "Random forest algorithm for prediction of precipitation," *Indonesian Journal of Artificial Intelligence and Data Mining*, vol. 1, no. 1, pp. 27–31, 2018.

[23] M. M. Baharuddin, H. Azis, and T. Hasanuddin, "Analisis Performa Metode K-Nearest Neighbor Untuk Identifikasi Jenis Kaca," *ILKOM Jurnal Ilmiah*, vol. 11, no. 3, pp. 269–274, 2019.

[24] M. R. A. Yudianto, K. Kusrini, and H. Al Fatta, "Analisis Pengaruh Tingkat Akurasi Klasifikasi Citra Wayang dengan Algoritma Convolitional Neural Network," *(JurTI) Jurnal Teknologi Informasi*, vol. 4, no. 2, pp. 182–191, 2020.

[25] D. Normawati and S. A. Prayogi, "Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter," *J-SAKTI (Jurnal Sains Komputer dan Informatika)*, vol. 5, no. 2, pp. 697–711, 2021.

[26] H. Azis, P. Purnawansyah, F. Fattah, and I. P. Putri, "Performa Klasifikasi K-NN dan Cross Validation Pada Data Pasien Pengidap Penyakit Jantung," *ILKOM Jurnal Ilmiah*, vol. 12, no. 2, pp. 81–86, 2020.

[27] L. Qadrini, A. Seppewali, and A. Aina, "Decision tree dan adaboost pada klasifikasi penerima program bantuan sosial," *Jurnal inovasi penelitian*, vol. 2, no. 7, pp. 1959–1966, 2021.